TTI VANGUARD
ADVANCED TECHNOLOGY RESEARCH MEMBERSHIP

[next] HIGHLIGHTS

December 3–4, 2019 • San Francisco

prepared by Nancy Kleinrock nk33@cornell.edu

The Future of Supercomputing—Dr. Jack Dongarra, University of Tennessee and Oak Ridge National Laboratory

- The persistent goal of high-performance computing has been simulation—now the third pillar of science, augmenting the traditional practices of experimentation and theory.
  - "We do simulation when things are too difficult to build," says Jack Dongarra of the University, of Tennessee, Oak Ridge National Laboratory, and the University of Manchester, who has long been steeped in the supercomputing culture, as evidenced by his 1993 cofounding role of the TOP500 list.
  - Twice yearly, the world's most powerful machines compete to outperform one another on the LINPACK benchmark, which entails solving, to 64-bit accuracy, a system of linear equations (a dense nonsymmetric matrix) through Gaussian elimination using partial pivoting.
    - Currently, China's designs dominate the list, with 224 entries, with the United States a distant second with 115.
    - Still, Oak Ridge's Summit occupies the top spot and the United States (#1, 2, 5, 7, 10) dominates the cream of the crop; China (#3, 4), Switzerland (#6), Japan (#8), and Germany (#9) round out the top ten.
  - Dongarra discusses currently dominant machine architectures, plus what he expects to shine as performance advances from petaFLOPS to exaFLOPS, as is expected in the coming couple of years (China anticipates crossing the exascale threshold in 2012–2022; the United States, in 2022–2023).
    - Notably, he shares his thoughts on how big data, machine learning, cloud computing, and the rise of computing at the edge will adjust the focus of HPC going forward.
- HPC overview:
  - Why supercomputers?
    - Simulation via supercomputer solves otherwise intractable scientific/engineering problems that are too difficult, expensive, dangerous, or slow to address with conventional methods.
      - "Crashing birds into the engine of jet aircraft is an experiment that we hope not to do very often," says Dongarra, by way of example, "but we can simulate that on a computer."
      - Similarly, HPC rooted in physical laws—perhaps augmented by machine learning and big data—can help us understand climatic or astronomical phenomena, which occur on a timeframe and size scale incompatible with experimentation.
    - In the United States, the Department of Energy's multibillion-dollar Exascale Computing program has six areas of interest—national security, energy, economic security, scientific, Earth system, healthcare—encompassing 24 distinct applications.
  - What constitutes a TOP500 contender?
    - Currently, every TOP500 machine operates in the petaFLOPS regime (>10^{15} floating point operations per second), but none can claim the follow-on milestone of 10^{18}.
    - Three architectures dominate the HPC landscape:
      - cluster of commodity processors (almost exclusively Intel)
      - cluster of commodity processors, augmented by accelerators—notably, GPUs—to boost the speed of computation (Nvidia is dominant)
      - cluster of lightweight cores (currently only two TOP500 systems have this architecture).
    - Generally, each chip/socket node contains many cores (and could be connected to one or more GPUs), each board contains many nodes, each cabinet contains many boards (tightly wired to limit latency), each machine consists a room full of cabinets intra- and interconnected through switches (e.g., Mellonox); memory is shared across nodes/chips/boards.
      - No longer does a private firm like Cray build and sell stylish special-purpose machines; the market does not support that business model.

- No one architecture is unequivocally best: Summit (#1, ~2.4M cores)) is a cluster/accelerator machine (149 petaFLOPS); TaihuLight (#2, ~10.6M cores) uses lightweight cores; Frontera (#5, ~450K cores), at the Texas Advanced Computing Center (which TTI/Vanguard visited during the 2016 Austin field trip), is a processor cluster only, but each core packs a greater punch.
  - For comparison, #500 achieves 1.14 petaFLOPS and Dongarra's laptop ekes out 166 gigaFLOPS (six orders of magnitude slower than Summit), equaling the performance of 1996's #1 machine.
    - As phenomenal as this progress has been, it could have been better had contenders continued to replace machines every 7.5 months (on average), as was the case pre-2008, whereas 15 months is the current norm.
  - Regardless the architecture, the speed of data movement is the most critical factor in determining the performance of a modern supercomputer.
    - "We're overprovisioned for floating points," he says. "It's the data movement that kills things."
    - Further exposing the perils of data movement, TOP500 machines are also ranked by performance on the High Performance Conjugate Gradients (HPCG) benchmark, which addresses the sparse-matrix regime.
      - Specifically, "it solves the problem that arises from the discretization of a 3-D PDE leading to a solution that requires an iterative solver," says Dongarra: "A relatively easy solver, but it showcases the limitations of HPC today."
      - Ranking entrants not by LINPACK performance but by HPCG performance shows just how poorly tuned today's best supercomputers are for addressing the problems of sparse matrices: The best—again, Summit—achieves only 2.9 petaFLOPS (1.5% of its theoretical peak).
      - Such problems require a great deal of data movement, which kills performance.
    - Dongarra has high hopes for Japan's incipient exascale Fugaku supercomputer, which is designed to optimize for bandwidth.
  - Also not to be discounted is that the extreme number of cores of TOP500 machines require a commensurate use of software parallelism.
    - "The DoE has this mantra of codesign between the hardware, the software, and the applications," says Dongarra, "but the reality is that it doesn't happen much."
- Although the TOP500 list focuses on benchmark performance, competition for the ACM's annual Gordon Bell Prize rests with excellence in practical applications of HPC, therefore ensuring that utile machines receive recognition, not only those capable of performing benchmarking "stunts."
- Some details on the current #1 machine:
  - The hardware of Oak Ridge's Summit is as follows:
    - Each of its 4608 nodes consists of two 22-core IBM POWER9 processors, paired with six Nvidia Tesla V100 GPUs, 608 GB of fast memory and 1.6 TB of NVMe memory; the processors account for 2.3% of performance, and the GPUs, 97.7%.
    - The system as a whole has 27,648 GPUs (street value of each, $10K), dual-rail Mellanox EDR InfiniBand network, and a 250 PB IBM Spectrum Scale file system transferring data at 2.5 TB/s.
  - It not only achieves 149 petaflop/s on LINPACK, but also clocks 200 petaflop/s for modeling and simulation (i.e., in computations important to the scientific community), and its peak performance to date is a blazing 3.3 exaflop/s for 16-bit floating-point computation, which Nvidia built into its hardware for the firm's core market of data analytics, machine learning, and artificial intelligence applications.
    - Note that GPU price has recently experienced a major uptick from the high demand for machine-learning community; "We couldn't afford this machine today," laments Dongarra.
  - Oak Ridge's machine notably achieves 74% of theoretical peak performance on LINPACK; compare with mean peak performance across the TOP500 of 1.5%.
    - This latter value is in line with that of Summit's IBM processors, highlighting both the acceleration advantage of incorporating GPUs and the crucial need to minimize latency from data movement, variously through fast switching, architectural design, and parallel software optimization.
      - "Your application had better be hybrid to use both the CPU and the GPU, otherwise you will get very little performance out of this machine," says Dongarra.

- Supercomputing success does not come cheap when it comes to energy: Summit's LINPACK computation consumed 11.1 MW; annualized, assuming continual computation, the energy cost is roughly $11.1M, not to mention the cost of operation, maintenance, programming, and so forth.
- Uptime reliability is high (>90%), and if a node fails, the remainder keep chugging away such that computations continue unabated as long as they weren't relying on the node in question.
  - "A failure requires a checkpoint and restart," concedes Dongarra. "We don't have any additional fault tolerance built into the applications."
- What is the global distribution of supercomputing prowess?
  - As both the top consumer and producer of supercomputers, China is throwing its weight behind supercomputing and increasingly uses domestic chips; this should serve as a warning to U.S. manufacturers, given that a Chinese firm has licensed Intel's x86 architecture.
    - "If you outcompute, you can outdo the science [of global competitors]," says Dongarra.
      - Note that China's rise has been sudden, having occupied zero TOP500 spots in 2001.
      - Today, China produces the lion's share of TOP500 computers, serving as integrator for 327 of them, while the United States accounts for just 123; as recently as 2014, upwards of 400 were produced in the States, evidence of a marked shift.
    - As importantly, Dongarra worries about the national security threat were the U.S. eventually forced to default to the use of Chinese-made supercomputers.
      - 2019 production still used Intel parts, but this will change directly.
    - Toward this end, the U.S. federal supercomputer budget is robust, with funds allocated by both the DoE and the Office of Science.
      - To wit, in the prior fiscal year, the DoE allocated $1.8B for the hardware alone for three new systems targeting the goal of exascale computation, each at a different national lab, and the DoE is making a comparable investment in software/applications:
        - Argonne will house the first U.S. supercomputer expected to reach the exascale mark, an Intel/Cray machine to be dubbed Aurora;
        - Oak Ridge will replace Summit with a new machine, Frontier, comprising AMD hardware integrated by Cray;
        - Lawrence Livermore's El Capitan will also use Cray as its integrator, although the hardware base is yet to be finalized.
    - China's monetary investment is on par with that of the United States, and the European Union and Japan are also firmly in the race toward exascale, with Russia less of a factor/threat; in fact, Japan and China are top contenders for first-in success.
    - "There is a rumor that there are two machines in China that are faster than the [Summit] machine, and those are built by Sugon," says Dongarra, "but those machines are not being disclosed for fear of retribution by the U.S. government, which has imposed embargoes prohibiting U.S. parts from being integrated into systems. What that results in, unfortunately, is China investing more in its own processors."
    - It would be naïve to believe that China's leadership is not a national security threat to the United States; surely both nations use their supercomputing prowess for scientific advancement, but also to simulate and certify weapons systems.
  - Every supercomputer is physically in some country, but roughly half of them are owned and operated by industry players.
    - "Industry gets it," says Dongarra. "This provides a strategic advantage; [firms] can outcompete their competitors by using supercomputers to drive their work."
- What's next for HPC?
  - Following the expected achievement of exascale on the TOP500 benchmark during the 2021–2013 timeframe, technology trends—namely the approaching end of Moore's Law—suggest that wholly new models of computation will be necessary to move performance forward.
    - Most likely the near term (coming decade) will benefit from investment in more efficient architectures or ways to package transistors: "things that relate to reconfigurable computing, dark silicon, and 3-D stacking," says Dongarra.
    - The likes of carbon nanotubes and spintronics will receive funding to enhance current devices.
    - Neuromorphic, analog, adiabatic reversible, and quantum computing are novel computing models that could move the field forward, each with their own set of relevant applications.
      - When each will arrive in a useful form is not entirely clear.
        - "Quantum computing is a great thing. It is the future of computing, and it may always be the future of computing," he quips. "It's not going to happen in my lifetime, I would say. We're

not going to see quantum computers replace the types of computers we have." Instead, they will function as accelerators, boosting performance for applicable use cases, but not for solving the partial differential equations important to the DoE.

- But the greatest transition in the HPC arena could be its application to alternative types of computation, incorporating not only the solution of conventional PDEs (i.e., dense-matrix problems), but also using machine learning to gain deeper insight into scientific matters, such as climate, biology, drug design, epidemiology, materials, cosmology, and high-energy physics.
  - It is not an overstatement to say that machine learning is changing science.
  - The hardware and software of HPC must change accordingly.
  - "The Department of Energy is going to invest in this," says Dongarra. "They are going to make an investment into AI to the tune of $3B–4B over the next ten years.
- Edge computing adds yet another wrinkle to the landscape of scientific computing, with data being fed up the chain for processing from edge devices to supercomputers.
- Taken together, increasingly the edge provides data that machine learning refines into models; supercomputers incorporate the outcomes into their simulations, and results filter back down both to iterate models and to inform actions at the edge.
  - "It is a flow in both directions," says Dongarra. "Information is going to be pushed from the edge to the high-performance systems, and once at the high-performance systems machine learning will help push new algorithms out to the edge to help understand and collect data in a more effective way."
- In discussions of edge computing and machine learning, the topic of the cloud generally surfaces, as well.
  - In this regard, Dongarra does not believe that, say, Amazon Web Services will supplant the need for dedicated supercomputers.
    - "Studies have shown that [the cloud] is not cost-effective in terms of dollars per cycle, over the long run you don't have control over the machine, there are issues about security of data, and so on," he says. "There are a number of issues that need to be overcome, but it could provide significant computing for another form of computing. Still, the machines at the top of the TOP500 list are not going to be replaced by a cloud service."

Technological Interventions in Climate—Ms. Kelly Wanser, SilverLining

- The climate is changing: It is getting hotter, and just as a human body can handle the heat stress from a fever—until it no longer can—so too can Earth's ecosystems continue to function in the presence of trapped heat energy—until they no longer can.
  - The science is not yet settled on when the tipping point will come, but the UN's International Panel on Climate Change's projections are clear that doing nothing to mitigate the problem will be catastrophic.
    - Business as usual would warm the planet by more than 4°C over the preindustrial baseline temperature, and currently negotiated targets of emissions reduction would achieve a rise of just under 3°C, but the IPCC now projects that only a maximum rise of 1.5°C would be safe.
      - "Right now we are *not* in a safe place," says Wanser.
    - Early effects of warming include increased severity and frequency of fires, floods, and severe storms, not to mention the looming threat to coastal regions from sea-level rise as the Arctic melts.
    - Individual insults to the planet become self-perpetuating, for instance melting arctic ice reduces the range of the albedo effect, which leads to further warming and therefore additional melting, eventually releasing stored methane, which would kick warming into yet a higher gear.
    - "The kind of state changes that can happen are very nonlinear and are generally thought to be irreversible," says Wanser. "We have climate refugees in America already: from Puerto Rico, from Barbados, and they have evacuated one of the Florida Keys already," not to mention the lobsters moving north from the Gulf of Maine into cooler waters.
  - The focus of her nonprofit firm SilverLining is to identify thresholds that must not be crossed if catastrophe is to be avoided, and to promote research to counter near-term (decadal) climate risks.
  - Wanser discusses the need for immediate and dedicated action to moderate warming until large-scale efforts to remove carbon from the atmosphere can come online 30–50 years from now and the bulk elimination of emissions can be effected over the course of the coming century.

- The specific geoengineering method she promotes is cloud brightening by spraying seawater into clouds to increase their ability to reflect sunlight.

- "The problem we have now is that none of our solutions—reducing emissions or removing carbon from the atmosphere—operate to reduce heat stress on the [10–20-year] timeframe," says Wanser; hence the need for drastic short-term action.
  - She turns to the National Academy of Sciences 2015 report—Climate Intervention: Reflecting Sunlight to Cool Earth—for practicable large-scale strategies quickly to quell temperature rise in the short term, specifically albedo modification strategies.
  - The study considered many such strategies, narrowing them down to two—the introduction of stratospheric aerosols to reflect sunlight back into space and marine cloud brightening to enhance the reflection of sunlight from lower altitudes—recommending no immediate deployment at scale, but rather small-scale experiments to test efficacy and unintended consequences without posing significant risk.
  - The phenomenon of cloud brightening is readily visible in satellite photos of cloud banks off the western cost of the United States, with clear streaks of white above the paths of oceangoing ships caused by their particulate emissions rising to the level of the clouds and causing them to reflect more brightly in an upward direction.
    - This particular cause of cloud brightening is dirty anthropogenic pollution, which has a mixed effect on the climate: the carbon emissions from diesel trap heat, but the particulates enhance clouds' reflectivity.
      - "We are already cooling the planet by putting particles in clouds at scale," says Wanser; the problem is that the mixed effect from the dirty source confounds the ability for scientists to understand the relative balance of the helpful and deleterious contributions.
        - "We're pretty sure we understand how the warming effect works for greenhouse gases and other things, but there is a lot of uncertainty about how the cooling effect works," she says; it is hypothesized that the cooling is equal in absolute value to the warming that we have experienced up to now.
        - "The 'cloud aerosol effect' is one of the highest priorities in all of climate research," she says.
  - Wanser has worked with researchers at the University of Washington and PARC to nebulize seawater into tiny droplets as a clean strategy for cloud brightening, with sea spray's salt particles serving as cloud-brightening particulates.
    - The team's novel nozzle generates 3T 80–100-nm droplets per second.
      - "That nozzle was invented by a set of retired researchers from Silicon Valley," says Wanser. "They have been working on it for six years for the sake of their grandchildren."
    - Millions in funding are necessary before testing will be possible at scale.
    - "The theory is that if you were to brighten marine clouds in susceptible regions—about 10–20% of those clouds, or something like 3–5% of the surface area of the ocean—you might be able to offset a doubling of CO_2, or all of the warming that we have," says Wanser.
      - The word *might* in Wanser's statement is relevant; it embodies the scientific uncertainty about the strength of the effect.
      - What is known is that the effect of particulate introduction into clouds is short term, lasting only a few days, and localized.
        - While this implies the need for ongoing action, it similarly suggests an escape hatch from unintended consequences; the consequences will occur, but like the particulates will quickly dissipate.
        - Potential regional applications include cooling the air and underlying water of the Great Barrier Reef to sustain the health of that vital ecosystem, or to do similarly in advance of hurricane season to lessen its impact.
  - Wanser emphasizes that the use of particulates for cooling has a firm basis in fact, albeit under circumstances beyond human control: a volcanic eruption spews a large quantity of particulates into the atmosphere yielding broadscale cooling.
    - "When Mount Pinatubo erupted in 1991, it had the effect of cooling the entire planet by 0.5°C for almost two years," she says. "This was measured, observed, and documented"; among those observations was a substantial recovery of arctic ice during the subsequent year.
    - As such, controlled injection of particulates into the stratosphere warrants study.
- Her optimistic outlook: "It is possible to maintain surface conditions that are more similar to now, even with our trajectory of emissions."

- While more study is needed, current knowledge suggests a path forward rooted in an engineering mindset: “What do I need to measure this system? What do I need to predict this system? And what kind of controls do I need on management tools that will interact with it to maintain its state?”
  - That is, Wanser views the most complex system on Earth the same way she would view any nonlinear system that she wished to have maintain stasis.
  - As is customary, when modeling a complex system, the more data, the better, which leads the SilverLining team to push for additional federal funding in the areas of climate prediction and observation, including an all-hands-on-deck, cooperative push for computational simulation to inform decision making.
  - To make good forecasts requires a good baseline, which is currently lacking for the chemistry of the stratosphere.
    - Three clearly identified, but poorly understood, risks:
      - inducing chemical changes to the ozone layer by injecting the likes of sulfates or calcium carbonate;
      - potential effects on the cloud layer as stratospheric particles descend over time;
      - “changing the dynamics of the way the atmosphere works at a very large scale.”
  - With climate scientists reaching the consensus around the unsurvivable extent of the known risks of global warming, the unknown risks of large-scale, purposeful tinkering with the global system are only now being deemed appropriate to take.
    - Still, the temporal constraint associated with either adding particulates to clouds to brighten them or to the stratosphere to shield sunlight more broadly should be considered a feature, not a bug: “This is a system with a high degree of control,” says Wanser. “You turn it on, you turn it off; you dial it up, you dial it down.”
    - While the National Academy’s 2015 report recommended further study, rather than spewing anything and everything into the air and beyond, action—once taken—can be readily undone.
    - Two studies will generate data to feed into models relevant at larger scales:
      - a Harvard-based proposed controlled experiment on stratospheric sunlight reflection: balloon-based release of a mere handful of liters of different kinds of materials into the stratosphere, studying the chemistry and dispersion around the site of release;
      - a University of Washington-based experiment on cloud brightening: small-scale emissions of salt spray into clouds to study the dynamics and dispersal of benign aerosolized particulates prior to performing an experiment of sufficient scale—100 x 100 km—to observe a brightening effect.
- When considering geoengineering more generally, as with any global policy there will be winners and losers; as one TTI/Vanguard participant puts it, “Some Pacific islanders are under threat, but farmers in Scandinavia and Siberia are benefitting [from a warming climate].”
  - Wanser concedes that the threshold for safety might be fluid, but there is an existence proof for global viability at preindustrial carbon levels, making that a useful target.
    - “I don’t think there are any winners with this kind of risk, because it is incredibly disruptive to everything,” she says, speaking of the specter of a warming planet.
    - An associated risk to the developing world, however, is to impose restrictions on modernization due to attendant greenhouse gas emissions.
      - Wanser recognizes that technological innovation can overcome this risk as less developed societies leapfrog the global north by adopting renewables-based distributed-energy systems and other sustainable practices.
  - Still, any avenue to address global warming will inherently have to be international in nature.
    - Weather modification modeling efforts are underway in various nations, with China making the greatest investment.
    - “As conditions worsen, people are likely to look at these research questions harder,” says Wanser, who would like to see U.S. expertise leading the way.
- “Failure is not an option,” concludes Wanser.

Security of 4G and 5G Cellular Networks—Dr. Elisa Bertino, Purdue University

- Unambiguously, cellular networks have risen to the status of critical infrastructure; as such, it is hoped that the attack surface is minimal.
  - Yet, using a framework developed at Purdue University, Elisa Bertino and her colleagues have discovered a range of vulnerabilities that bad actors could use to target wireless users.

        - For instance, one class of vulnerability in the cellular paging protocol present in both the 4G and 5G wireless specifications, when exploited to its fullest, permits an adversary to localize the victim—albeit only in a coarse-grained manner—spoof the victim's identity with fabricated paging messages, execute denial-of-service attacks, and even extract the device's persistent identity (i.e., the international mobile subscriber identity (IMSI)).
        - Of course, before publishing a vulnerability, Bertino informs relevant carriers or device manufacturers to give them an opportunity to first implement a fix.
    - Bertino offers a reminder of the importance of wireless networks and the added potential vulnerability of 5G, she also describes these attack methods against various stages of any cellular communication, and—in the contribution that she deems most important—she describes her systematic framework for adversarially analyzing the network specification in an effort to uncover additional problems this vital infrastructure.
        - The framework has been formalized in a tool Bertino dubs LTEInspector.
- Having been steeped in 4G wireless for a decade, most members of the general public feel lost when out of range of a cell tower—sometimes literally, due to reliance on the always-on availability of mapping and other apps.
    - 5G promises yet more, enabling innovative low-latency applications in industry sectors as diverse as automotive, entertainment, healthcare, and energy.
    - New capabilities will require new technologies, including MIMO-based beam forming, small cells, and duplex communication, operating in the millimeter-wave regime both to expand the effective spectrum and to increase data rates.
    - "We want to provide enhanced bandwidth, while at the same time reducing energy consumption and providing all the needed connectivity to many novel applications, including to IoT systems and devices," says Bertino.
    - As the breadth, importance, and temporal reliance of the resources the wireless network enables become increasingly embedded in the day-to-day workings of society, any network vulnerability that might lead to an interruption in service (DoS) or invade users' privacy is serious indeed.
- The expectation was that the 5G protocol would eliminate the potential for IMSI-related attacks, but although IMSI is encrypted on the device, for one U.S. carrier this key identifier can still be exposed and intercepted.
    - A primary purpose of paging is to establish a dedicated communication channel between the phone and the base station; to appropriately balance energy consumption with quality-of-service, each generation of wireless protocols dictates a specific duration between paging (polling) incidents.
    - The ToRPEDO attack exploits this fixed interval by first initiating a rapid series of phone calls to the device and using this increased traffic—unique patterns in its paging messages—to deduce the base station to which it is then connected, opening the opportunity to launch a stingray attack by impersonating that base station and wreak havoc by intercepting or injecting the device's communications.
        - This vulnerability exploited by this somewhat tortured acronym of an attack—TRacking via Paging mEssage DistributiOn(!)—exists within the 4G and 5G data standards, and would need to be addressed above the carrier level.
    - On 4G, and only on the ill-protected network, the PIERCER attack (Persistent Information ExposuRe by the CorE netwoRk) furthermore makes it possible for an attacker to associate a victim's phone number and IMSI to track the victim's location.
        - A fix to PIERCER would pertain only to the vulnerable network, but this attack type is an example of exploiting the need for backward compatibility as new generations of wireless proceed.
            - "We have shown several attacks in which the attacker can force a phone to go back to the 3G protocols," says Bertino, "and then be far more vulnerable."
- Of course, paging is not the only action taken over cellular networks; the core actions within 4G are attach, detach, paging, handover, VoLTE, and SMS.
    - The scope of Bertino's research addresses vulnerabilities in the attach, detach, and paging steps.
- Several challenges must be overcome to develop a systematic strategy to uncover vulnerabilities.
    - Stateful procedures and multiple participants:
        - For instance, when a piece of user equipment attempts to connect to the network, it first sends a message, followed by acknowledgement from the core network.
            - "This is a stateful procedure," says Bertino, reliant on the sequential tracking of each step and each participant as the communication proceeds through the various network layers.
    - Lack of formal specification of 4G LTE:

        - Specification documents amount to thousands of pages of text written in natural language, sometimes with embedded ambiguity and often ill-suited to effective textual data mining.
    - Closed system:
        - The standards themselves are, of course, public, but their implementations tend not to be, with the various carriers choosing among options within the specs and not disclosing each selection.
            - "Sometimes implementation can introduce additional layers," says Bertino. "If we could go into the implementation, we could discover more vulnerabilities."
    - Legal barrier:
        - The cellular bands comprise licensed spectrum that is not freely accessible to researchers for experimentation to search for vulnerabilities and test out exploits.
    - To address this collection of challenges, Bertino focused on an abstract specification of protocols manually constructed from the standard.
- Basic architecture and operation of an LTE network:
    - The network covers a tracking area outfitted with a collection of base stations (most generally, eNodeB (typically, cell towers)); the user's device (formally, user equipment (UE)) connects to it, and in turn the eNodeB communicates with the evolved packet core (EPC), consisting of the mobile management entity (MME) responsible for paging with idle UEs, allocating temporary identities to UEs, enforcing roaming restrictions, authenticating SIM cards via the home subscriber server (HSS), and the like; after performing its various internal checks, EPC connects the user out to the broad Internet.
    - Consider the process of the action *attach*.
        - As a user carries a device into a new cellular area, the UE by default establishes a connection setup with the closest eNodeB.
        - Via its just established connection, the UE sends an attach request to the EPC that includes the SIM's IMSI and perhaps the device's international mobile equipment identity (IMEI), along with the security capabilities of the UE (e.g., the UE's encryption algorithm, which is relevant for backward compatibility).
        - Initiating a challenge–response protocol, the network accepts the attach, issues a TMSI (temporary mobile subscriber identity), and sends an authentication request to the UE, to which its SIM must properly respond.
            - Note that if an attacker can suss out the TMSI, it can track the movement of the phone from one eNodeB to another.
                - Although the cellular standard specifies the frequent reassignment of TMSIs to protect against such attacks, network operators do not faithfully do so.
        - To complete attachment, the EPC confirms the authentication, and the EPC and UE negotiate on and confirm a security mode to adopt for the remainder of the communication.
    - Paging, in contrast, begins with a request from EPC's MME to the eNodeB, which in turn passes it along to the UE.
    - When the channel is no longer required, a detach request and acceptance pass between the UE and MME.
- LTEInspector:
    - To explore potential vulnerabilities in cellular communications, Bertino uses the well-understood Dolev-Yao model of interactive cryptographic protocols and considers the diverse adversarial actions of intercepting (eavesdropping), dropping, modifying, or interjecting messages, even when the adversary lacks the UE's private key (as can be true for a powerful adversary).
        - The Dolev-Yao model benefits from automatic tools (e.g., ProVerif, Tamarin) to analyze protocols for vulnerabilities.
        - Use of automatic verification tools requires understanding the characteristics of the properties being verified, such as the temporal ordering of events, crypto constructs, and mathematical predicates, such as reliance on linear integer arithmetic.
            - Example: The presence of misordered requests and confirmations suggest an attack in progress.
        - Model checkers serve as powerful tools: "You provide as input a model of your protocol, and properties that you want your system to have," says Bertino. "What the model checker will do is analyze your specification and tell you if certain properties that you want to have are violated. If they are violated, it will give you a counterexample—a sequence of steps that will lead you to the violation of the property."
            - Model checkers are particularly useful for evaluating temporal traces and linear arithmetic.

- Cryptographic protocol verifiers similarly evaluate the integrity of cryptographic constructs.
- LTEInspector combines model checkers and cryptographic protocol verifiers into a single tool for network evaluation; it has successfully uncovered multiple network protocol and implementation vulnerabilities.
  - LTEInspector models the communicating UE and the EPC, each as a finite state machine coexisting in a Dolev-Yao-style adversarial environment.
    - Messages exchanged between the UE and EPC represent valid state transitions.
    - "We obtain the model that is instrumented with the adversary," says Bertino. "We give this model as input to the model checker."
      - Developing the abstraction of the LTE model was no small feat, involving the translation of substantial domain knowledge—for both the UE and EPC and their various interactions—into propositional logic (i.e., an expert system), and abstracting away cryptographic constructs.
      - Following over a year of model development, Bertino openly released the XML-formatted model.
    - If the model checker identifies violations of the properties of the cellular standard, the model checker's counterexample is then analyzed by the cryptographic protocol verifier to assess whether the proposed attacks are possible given the levels of encryption implemented by actual network operators.
    - If the verifier concludes an attack is possible, Bertino's lab goes the extra step to carry out an experimental attack in-house to demonstrate its (nefarious) practicality.
      - "We couldn't try the attacks on the real cellular network, because it is forbidden, so we created our own malicious eNodeB using some publicly available software," says Bertino, noting that in-the-wild attacks can be considerably more sophisticated.
  - LTEInspector has, to date, discovered ten new attacks along with nine attacks that had been identified previously.
    - Four of the newly discovered attacks prey on the attach procedure, five on paging (including ToRPEDO), and one on detach, with denial of service and coarse-grained location tracking being the most frequent impacts of note.
    - In the overwhelming number of cases, the vulnerability was in the 3GPP wireless specification itself, rather than poor implementation of the standard by either a carrier (as with PIERCER) or a device maker.
  - At the request of a cellular operator, Bertino has also used LTEInspector to verify its implementation of 4G LTE.
- Bertino's lab is now adapting LTEInspector to the analysis of 5G.
- Understanding the problems is important both in its own right and to improve the standards; it is also a first step in developing defenses to potential exploits.
  - Bertino is engaging in initial work toward this end:
    - On the millisecond timescale, eNodeBs broadcast master information blocks (MIBs) and system information blocks (SIBs) without the benefit of digital signatures, which opens the door to UEs connecting to fake base stations.
      - To close this vulnerability, she observed that the 3GPP standard offers the use of either the TESLA symmetric broadcast authentication protocol or public key encryption.
        - "We [considered] what happens if we try a very simple PKI, and we found a lot of complicated requirements," says Bertino: a need to minimize signature size (to save bandwidth), generation time (due to MIB and SIM broadcast frequency), and verification time (to reduce UE energy expenditure).
    - Her practical solution includes PKI-level optimization in the form of a lightweight certificate, protocol-level optimization by only authenticating SIB once through signature aggregation for SIB1 and SIB2, and cryptographic scheme-level optimization by precomputing using BGLS encryption and structure-free and compact real-time authentication (SCRA).

AI, Risk, and Dynamic Flood Mapping—Ms. Bessie Schwarz, Cloud to Street

- Water is both a blessing and a curse: we need it to live, but inundation can be equally deadly.
  - Knowing where water is overrunning—or likely to overrun—populated areas or farmland can prove lifesaving and is also vital to planning where to live, build, and plant, as well as how to appropriately insure against the hazard of water.

- "What we have is a cloud platform that leverages satellites, AI, and community intelligence in order to map floods as they are happening around the globe," says Bessie Schwarz of her nonprofit Cloud to Street, "and then to go back in time to analyze flood risk across the world instantaneously."
- Her team of scientists and AI practitioners, in partnership with experts in insurance and disaster relief, are reinventing the absorption of disaster risk, particularly with the increased incidence of climate change-induced disasters.
- Schwarz shares informative stories from Mozambique to Congo, from the New York metro area to Houston, and beyond, when describing the impact flooding has on lives and economies, and explains the benefits Cloud to Street can offer in terms of targeted evacuation alerts, accurate pricing of insurance, and fast response times and payments in the wake of a disaster.
- "Risk is growing, becoming more extreme and unpredictable, but we can distribute it more equitably," she says.

- Case studies:
  - Beira, Mozambique, 2000 and 2019: In 2000, a flood devastated the region, leaving millions of people homeless; Beira was similarly impacted in March 2019 as Cyclone Idai let loose with torrential rains and a 14' storm surge, destroying the city and its environs.
    - "Even with 20 years of scientific innovation, new technology, and development work, there was basically no increased resilience in this community," says Schwarz. "A lot of this is a technology problem, but a lot of this is where we are putting our investments. We could be enabling people to [have] that resilience."
  - Hurricanes Sandy, 2012, and Harvey, 2017: When these storms respectively hit the U.S. eastern seaboard and the Houston area, many of the damaged homes (10K in NY and 100K in TX) were outside the FEMA flood maps and therefore underinsured.
    - "In Harvey, 80% of the loss was not insured," says Schwarz. "FEMA says that most of the country has inaccurate, outdated, or nonexistent flood maps."
  - Impfondo, Republic of Congo, 2017: This town of 5000 people suffered a moderate flood, impacting not only residents but also regional production, but the outside government remained unaware of the situation for three weeks, necessitating a natural disaster declaration and sizeable UN relief operation.
    - In the wake of this catastrophe, to alleviate impacts of future events, Cloud to Street initiated a Congo-wide flood information system that is perhaps the most sophisticated in the world.
      - "We are scanning every inch of the country every day with a bunch of different satellites, looking for any anomalous flooding and alerting them immediately," says Schwarz.
    - This proved invaluable when, in late 2018, 16K–17K people from neighboring Democratic Republic of the Congo crossed the river into the region of Mokotipoko and settled into four refugee camps.
    - The combinations of ambient conditions and flood history indicated that one of the camps, housing 6K people, was not only at severe flood risk, but indeed was starting to flood.
      - The refugees were successfully relocated to the safer camps, which just a year later are also deemed at severe risk.
  - The Atlantis Resort, Bahamas, Hurricane Dorian, 2019: As the storm was approaching, Goldman Sachs fretted about the potential fate of its significant asset, the Atlantis.
    - Visual images only showed the menacing storm system, but satellite-based radar peered through that layer to reveal the still-intact buildings and unflooded grounds of the resort.
      - "Goldman was trying to call the front desk [to get the ground truth]," says Schwarz. "But if it's Houston during Harvey, are you going to call the thousands of mortgages you have?"
  - Globally, 70% of disaster losses are uninsured with most of the uncovered risk (90–99%) existing in low- and middle-income nations, putting an enormous strain on societies least able to bounce back from disruption.
  - Moreover, the mean protection gap between losses from disaster and covered losses is expected to grow considerably from the 2010 value of $70B as the severity of incidents increases with the changing climate.
- Why do flood-prone communities lack resilience? In large part because nearly all (97%) of the humanitarian aid for disaster relief allocated in the developing world is spent post-disaster, rather than in preparation for the inevitable; still, the annual aid outlay of $2.2B only covers 8% of the actual losses such nations incur.
  - "Prevention is much more cost-effective than being reactionary," says Schwarz.
  - The goal of Cloud to Street's work is to flip this model.

- Currently, flood mapping is a very manual science, reliant on considerable field work and a large number of distributed river and weather gauges that collectively generate Hydromet data.
  - "You take all that data and run simulations," she says. "What's the five-year flood? What's the 5000-year flood?"
    - Models become outdated whenever an on-the-ground change occurs, such as the construction of a new bridge or sewer system, or if the course of a river changes because of a flood or construction project since the development of the most recent flood map.
  - Hydromet systems are not only expensive to install, but also to maintain, particularly in strife-ridden parts of the world.
    - Example: The Republic of Congo's robust Hydromet system consisted of roughly 80 stream gauges in the 1980s; all have been wiped out due to conflict, and only a half dozen have been rebuilt.
      - The data that does exist is maintained in a handwritten ledger, updated only monthly.
        - "This type of recordkeeping is much more common than you might think," says Schwarz.
    - As often as not, global flood models offer poor predictions.
    - Example: Not only did six Sudanese flood models disagree with one another, all failed to align with an actual flood event.
      - In fact, Schwarz claims "zero-percent accuracy" to be the norm across East Africa.
  - Finally, damage assessment involves the local presence of insurance adjusters, who go home to home—a slow and expensive process.
- In the meantime, for more than four decades satellites have been circling the planet, remotely capturing relevant data.
  - The confluence of an increasing number of satellites, extensive temporal coverage, high spatial resolution, and powerful computing power here on Earth has created the opportunity to develop high-quality flood models that can incorporate real-time data when a locality is under threat or has just suffered a severe storm.
  - In conjunction with Google and the Dartmouth Flood Observatory, Cloud to Street has incorporated every major flood event worldwide since 1985 into a web-based visualization platform.
    - Input data amounted to more than 4M images, collectively for 4500 floods, from MODIS (250-m resolution, to visualize large farms), Landsat (30 m, small farms), and Sentinel-1 (10 m, roads; radar) satellites.
    - This web-based platform makes it easy to search, view, and download flood maps.
  - The use of satellite imagery is tougher than it might seem.
    - Any given image of an inundation can be partially obscured by clouds, include water with diverse colorations (tan, green, blue), and other complicating features.
      - "This is where machine learning and the physics-based algorithms come in to try to understand [scenes] across millions of images," says Schwarz.
      - Note that cloud cover confounds the utility of an optical camera to see what lies beneath, but does not disrupt the signal captured by satellite-based radar.
    - Cloud to Street leverages all available satellite data—public (NASA, ESA) and private (PlanetScope (3–5 m resolution, to visualize small roads), WorldView (1.3 m, urban buildings), ICEYE (1 m; radar))—processing it in Amazon's cloud and integrating ground truth verification from field agents to distinguish between normal and abnormal extent of water.
      - Social media, crowdsourcing, and urban surveillance cameras can all contribute to verification; Hurricane Harvey benefited from drone-collected imagery, as well.
      - "While we have a box that's flying around Earth trying to tell you that it's flooding at your feet, you often know if it is flooding at your feet," she says.
        - Environmental sensors are only moderately useful during flood events: "The problem with those is that they are point-based, and flooding is so localized that it can be flooding on one side of the street but not on the other," says Schwarz.
  - In the midst of a disaster, the information deemed most valuable to people in its path is a simple text message that directs them to evacuate now, prepare to evacuate, or remain safely in place.
    - "We have really fancy, cool dashboards, millions and millions of pixels, and fancy AI tools, but what people want is a text message," says Schwarz.
      - Unfortunately, some people ignore even an overt directive to evacuate.
        - Research has shown that in the United States social vulnerability to flood risk rises for people who fit into particular profiles:

        - Property damage tends to be worse for those with low income, women-headed households, and Southerners;
        - Fatalities tend to cluster in rural areas, for people who are under 5 or over 65 years of age, or have a disability, especially during a strong-to-moderate flood.
      - "Two communities hit by the same exact disaster will have very different amounts of loss," she says.
    - Governmental agencies can greatly benefit from a deeper exploration of the Cloud to Street resources in the midst of an event.
      - Example: During the July 2015 floods of Chennai, India, the dashboard offered neighborhood-level predictions of how many people would likely be impacted.
        - This degree of precision informed rescue crews with a highly targeted plan of action.
        - In this case, the best street-level data came from social media reports, despite the acknowledged accuracy challenges when using such resources.
  - Prediction of the potential impact of future events is of particular value for planning purposes—for governments, for insurers, for lenders, and for individuals and businesses in the path of a disaster.
    - Cloud to Street combines anticipated precipitation levels, soil saturation, and the corpus of flood history with machine learning to predict the impact of a storm as it approaches.
      - Additional detail could improve models, in particular detailed elevation profiles to simulate water flow and predict water depth.
        - "Depth maps are really hard to get," says Schwarz, "but we are on the edge of getting some depth [information] from the satellites without the depth map."
    - More generally, the combination of multiple data sources of local and temporal relevance are stacked to create a longitudinal sense of flood risk to assess how many people, over what geographic extent, are at risk of being exposed to flood every 2, 5, 10, 15, or 30 years.
      - Schwarz's firm has performed such an assessment for East Africa, deeming 1.4M people to be at severe flood risk by midcentury.
- Among the positive applications of Cloud to Street's platform, parametric insurance could prove one of the most impactful.
  - The first goal is to reduce the cost of pricing premiums through the use of instantaneous flood frequency methods.
  - Second, real-time observations at street- and even building-level granularity can largely do away with the need for field appraisers; instead, automatic payments can be dispatched to insurance customers recognized to be in impacted zones.
    - Instantaneous, digitally delivered insurance funds are not just a convenience; it can mean the difference between life and death for a family otherwise lacking resources to evacuate to safety.
  - Schwarz's vision is not limited to vulnerable places in developing nations; parametric insurance for the U.S. market is also firmly on her mental radar.
    - Access to her flood risk analysis data and parametric insurance would have properly evaluated the situation in Houston, rather than 80% of the damage impacting uninsured homes outside of the FEMA floodplain.
      - "What if we use parametric insurance to power access to capital everywhere?" she envisions.
  - "Let's just detect from the sky, and let's use the community in a way that can be scientifically verifiable to say whether it is flooding at their feet," says Schwarz. "We envision a world in which the quality of information for people in California experiencing floods is dependent on people in Tamil Nadu, India, because it's creating a better algorithm for both of them."
  - Not surprisingly, conventional insurance companies are of two minds with respect to an upstart like Cloud to Street.
    - Lacking good data of their own, many insurance firms are moving away from covering floods and other hard-to-predict perils in this time of a warming planet.
      - "They don't have great solutions to climate change," she says, because of the mismatch between the industry's backward-looking modeling strategy and the knowledge that the future will differ, but in uncertain ways.
    - At the same time, insurers are always looking to diversify risk by entering new markets.
    - Moreover, the cost-cutting and transparency inherent in parametric insurance's remote appraisal entices them.
      - Payouts can be preestablished according to clear metrics that are unlikely to be litigated after the fact: "If this pixel becomes blue, you will get a payout; if it doesn't become blue you won't."
      - Some insurers are already dipping a toe into the waters of parametric drought insurance.

- "[Current flood profiles due to climate change] should not be confused with normal change," says Schwarz, "but like any good disaster, it shouldn't be wasted. We should use this moment to do things differently and rethink how we absorb the risk that we have today."

World War Zero—Dr. Saul Griffith, Otherlab

- Otherlab's Saul Griffith's story is one part alarmism, one part optimism, and, if he is right, it is all parts realism; the topic is climate change, and the solution is soup-to-nuts electrification and the production of enough backup battery storage to enable renewables to manage the load.
  - "I'm an energy nerd," he says, "so I think about the solutions to climate change, not the problem."
  - Of course, consideration of solutions requires a deep understanding of the problem.
    - He has therefore updated previous Sankey diagrams of U.S. energy flow with an intensely granular version.
  - Griffith's analysis, combined with the overwhelming consensus of climate science, has led to his conclusion that to avert a climate catastrophe there is no time to spare—i.e., *zero* years—to go all-in on doing away with a fossil fuel-based economy and converting to one rooted in renewables-based electricity.
    - "The IPCC, in its 1.5°C report of 2018, said that we have until 2030 to reduce emissions to 55% to stay on a trajectory compatible with a 1.5°C world," reports Griffith; this means shifting systems *now*, not pondering the matter for yet another decade, when it will be too late.
      - Tucked away in Appendix Y4 of a report entitled "Restoring the quality of our environment," the Environmental Pollution Panel of the President's Science Advisory Committee first sounded the greenhouse gas alert in 1965; far too little has been done in the intervening half century.
  - To keep the planet habitable will take the kind of commitment that the United States has made previously in times of crisis (New Deal, Arsenal of Democracy, Manhattan Project) or hope, wonder, and a national security-based need to lead the world into the future (Apollo Project).
  - Today, the need is dire, hence the alarmism, but the future can be bright if we invest in it.
    - Griffith does not preach austerity—no pleas for lowering thermostats and wearing sweaters, as Jimmy Carter did during the energy crisis that instigated the first U.S. energy flow diagram in 1976; we all love our appliances, wide-screen TVs, gadgets, and no-sweaters lifestyle too much.
    - Instead, he paves a road toward electrification of everything from transportation to HVAC and will be driving his recently purchased and soon-to-be-electrified 1961 Lincoln Continental along it—a car and conversion that serves as an able metaphor for the transition we now face in which the efficiency of electricity overcomes the excesses of the lifestyle we have come to expect.
    - The future Griffith wants us to adopt is one without sacrifice: "I give you your SUV, it's just electric; I give you your McMansion, but it has solar panels on the roof and a heat pump in the basement instead of a furnace; and I give you all the rest of the trappings of the American way of life. I just electrify the whole economy, and magically instead of needing 98 units of energy you need about 42."
- Energy sources, uses, waste, and recovery/reuse:
  - With incredible detail—0.1% resolution—Griffith acquired and organized the nation's energy data into a Sankey diagram that reveals that current systems of production and use across the transportation, industrial, residential, commercial, and government sectors incur a net waste of more than half the input energy (56%).
  - Conversely, even with the same supply side mix of fossil and renewable fuels as exists today, use-side electrification of all societal sectors would shrink the waste energy to less than 10%—a remarkable benefit from executing a single, easy-to-articulate strategy, that would ideally enjoy rapid implementation even as renewable energy production continues to ramp up and fossil fuel-based production wanes.
    - Improvements come largely from the efficiency gains of electric motors over gas- or diesel-powered motors; for instance, the 1976 diagram assumed autos were 20% efficient, whereas electric cars can perform at 80–90% efficiency.
      - "You don't need 15–18% of U.S. energy flow if we just electrified all ground transportation," says Griffith.
    - An alternative metric is the energy currently expended to find, mine, refine, and deliver fossil fuels; his Sankey diagram reveals that 2% of energy flow is for exploration and extraction, 4% to refine oil into gasoline 1% to maintain the line pressure of the nation's natural gas pipelines, and so forth.

        - "About 10% of our energy flow goes away if we don't have to make fossil fuels anymore," he says.
    - However, if it were easy, the changeover would already have taken place.
        - The challenge is that legacy systems exist across all sectors: gasoline-sucking vehicles clog the roadways, jet fuel is a petroleum product, natural gas and oil cool and heat buildings and underpin industrial processes, and so forth.
        - All of these have to shift to electricity—and fast—yet, for instance, most new vehicles still rely exclusively on gasoline.
            - If you just purchased a new gas-powered SUV, will you put out tens of thousands of additional dollars next year for an electric one (with limited driving range and a catch-as-catch-can recharging infrastructure)? Or if you just replaced your home's natural gas furnace, will you convert to an electric heat pump, just because it's the right thing to do?
            - Likely not; instead, you will wait until the item needs replacing before even considering whether a carbon-free version makes sense.
        - "If we, with perfect execution, never make a machine that makes carbon dioxide again, and at the replacement time of every car, every furnace, every powerplant do a zero-carbon one, we will only get a middling climate outcome with a reasonable probability of already going over the climate tipping points," warns Griffith.
- Griffith sees a way out, and we must take it. Humanity has no choice: "Go big, or we'll have no home," he says. "It's a climate emergency!"
    - Therefore, society should treat it as an emergency and implement emergency economics by replacing every machine at its end of life with a clean alternative "with 100% certainty," he says.
        - "Normal market rules can't apply"; nor did they during the Depression or Manhattan Project.
    - Consider an earlier prior emergency—WWII—when American industrialist Henry J. Kaiser took on the yeoman's task of converting his various shipyards to the production of nearly 750 replicates of a single model, the Liberty ship, for the merchant marines to assist the war effort (2710 were built in all, by Kaiser and other shipbuilders).
        - Wartime instigated similar effort in aviation—"I think they were building two airplanes in 1939 and 35K in 1944," says Griffith.
    - To bring industrialists on board, the U.S. government offered a patriotism incentive of cost-plus-7%, perhaps not the markup they might otherwise have gleaned, but the volume more than made up for it.
    - Of course, the specifics of the need are different now than then: "In World War II, it was planes, tanks, Liberty ships, bombs, and bullets," says Griffith. "For World War Zero, it needs to be EVs, solar panels, wind turbines, heat pumps, and batteries—that's roughly everything you need to solve the problem."
        - Using batteries as an example, and estimating that annual production of 90B Li–ion rechargeable cells would electrify the whole of the U.S. auto fleet in 20 years, Griffith makes the case that production on this scale is, indeed, possible.
            - Consider that annual global production of Lego pieces is 9B and that of bullets is 90B.
                - Setting aside the predictable commentary on the state of a species that makes an order-of-magnitude more bullets than Legos, it is useful to note that a bullet has the same size and complexity as an 18650 battery.
                - Relevant scaleup is both feasible and achievable, believes Griffith, with well-aligned incentives.
    - Although the detail he has provided pertains to battery production, Griffith is a proponent of the "yes, and" mindset, open to all beneficial technologies, but sees electrification as the best to kick off the transition that must ensue.
        - "Electrification is the only efficiency that you need," he says, noting that, to him, electrification encompasses not only solar, wind, and hydropower, but also nuclear, hydrogen, "and all of the other yes, ands," which include carbon-zero biofuels for not-yet-electrified systems that remain on a liquid-fuel diet, such as flight.
    - Naysayers contend that the conversion necessary to keep the planet viable will cost trillions; Griffith, in contrast, recognizes that conversion will *save* trillions; what is necessary is to jump-start the process with appropriate financing and economic incentives.
        - In light of the quick payback for the production of solar cells (6–9 months to generate the energy expended to manufacture) and especially wind turbines (three months), Griffith estimates that,

were the 10% of energy flow currently devoted to producing fossil fuels instead diverted to renewables, it would fully cover the manufacturing needs for the zero-carbon changeover.
    - "If you took the price of Australian solar, the price of California electric vehicles, and the price and polices of Germany's heat pumps and electrification of heat, and you made that a financeable package, today you would save every home in America $1000 a year," says Griffith, which amounts to slashing personal energy budgets by 25%.
        - So the future is here, but it is not evenly distributed; consider that installed Australian solar is US$1.20/W, but in California it is $3.20/W, with a significant chunk of the difference going toward California's permitting and regulations.
            - That is, a straightforward set of policy changes would accelerate the transition.
- "We can do it, it will save us all, and it will save us trillions," says Griffith, who believes the economic solution is hardly rocket-fuel science.
    - Long-term financing (e.g., home loans, car loans) made possible earlier step-function changes in personal ownership; so too with the current challenge.
    - Utility companies purchase infrastructure at 2–3% interest, but a homeowner must shell out much for higher rates to finance rooftop solar or purchase a car.
    - "If we could put together a package that considers the solar on their rooftop, the battery in their car, and the furnace inside their home as the infrastructure—with infrastructure financing—because it will all be connected to the broader [electric] infrastructure, we would do even better than saving $1000/year for the average family," he says. "We may just be an interest rate away from solving climate change, at least in America."
- Griffith is quick to recognize that buy-in for his path to winning World War Zero will be hard to achieve given the current political climate, which is fueled in part by targeting fuel sources (and the types of jobs and corporate profits they represent) against one another.
    - To generate not only unity but available cash, he is in favor of governments buying back petroleum firms' balance sheets and having them instead serve as finance companies providing loans for the new infrastructure—a win–win.
        - "To solve climate change, we have to have some big, crazy ideas," he says. "The big compromise between left and right might be that we just finance the existing energy companies—which are really only banks anyway, since they don't do the mining themselves or the refining themselves, but are just banks that coordinate it. So let's give them their balance sheets back and let them finance the deployment of renewables to break the insane deadlock on that issue."
- And what of the needs and challenges of the developing world?
    - Just as they leapfrogged over the developed world by installing wireless communications without having wired infrastructure previously in place, regions that currently lack reliable electricity infrastructure or an entrenched fossil fuel infrastructure (e.g., pipelines) are already foregoing centralized, power plant-based production and distribution in favor of distributed generation, with solar as an important facet.
        - "I just got back from Kenya," says Griffith. "They don't have 4.4M miles of natural gas pipeline. Without that legacy infrastructure, for Kenya to go straight to a solution that looks like rooftop solar and electric vehicles, they can completely skip all the problems."
            - His fear is that Kenya is currently in negotiations with China to buy nuclear infrastructure at 12% interest; "Given that its neighbors are Sudan and Somalia, what could possibly go wrong?"
    - Like the developed world, what the developing world needs most are financing instruments: "We have a capital excess [and capable industrialists] in the developed world and a labor excess in the developing world, and we need to solve that balance to make it work out for everyone," says Griffith.

AI Everywhere—Even at the Edge—Mr. Jags Kandasamy and Dr. Sek Chai, Latent AI

- The history of modern computing began with on-premises machines, progressed to data centers as storage and computational needs increased, and then to the cloud infrastructure, now a cornerstone of the computing landscape, although it began somewhat accidentally when Amazon trialed monetizing its excess capacity.
    - The growth of interest in the cloud ballooned, in part, due to the buildout of edge devices (e.g., IoT), which generated data but lacked adequat onboard resources to compute over them.
    - "With all of this," says Latent AI's Jags Kandasamy, "networking had to catch up" to carry the data to/from the computational resource.

        - Each new generation of wireless communication offers increased bandwidth and reduced latency, but the rate of data generation never ceases to grow.
        - "Will we become victims of induced demand," he asks, "as we open up more bandwidth and more devices come online?"
    - A way out of this conundrum is to compute at the edge, rather than transfer the data to the site of computation.
        - This begs the question of what *is* the edge: Is it the cloud infrastructure? the edge data center? the micro data center? the edge device itself?
            - Context-dependent definitions vary, and Kandasamy and his cofounder Sek Chai allow for all of these to fall under the umbrella of the edge, with the commonality being the need—at least in some circumstances—to perform processing in place, rather than engaging in large-scale data transfer for the sake of computation.
            - For example, with microcontrollers and communication capabilities embedded in all manner of household, environmental, and industrial objects—both big and small, stationary and mobile, vital and incidental—ongoingly collecting analog information from the environment and tucking it away in digital form, what is clear is that the volume of data they passively collectively generate has become too great to transfer off every device to be processed elsewhere (and the problem is going to worsen exponentially going forward).
    - The solution that the Latent AI team offers involves training neural networks at variable levels of precision, to perform useful inference where the data resides.
        - Latent AI is a young firm, just a year old, spun out from SRI with seed funding from Steve Jurvetson's Future Ventures, and based on DARPA-funded technology developed by Chai.
    - Kandasamy and Chai discuss the biological inspiration for their highly efficient approach and provide an example of it in action.
- Kandasamy and Chai's prototype example is real-time language understanding, albeit in a narrow domain.
    - For Comcast, 22M of its video customers use the provider's Xfinity Voice Remote service to change the volume setting, search for shows, and so forth, totaling 1B inferences during the past year.
        - "Every time you call for a channel change, ask for the weather, or whatever, it goes up to Comcast's data center, it gets the inference done, and your channel changes or your information comes up," says Kandasamy.
        - With an expectation of 75B IoT devices by 2025 requesting 3.4T inferences annually, the data transfer load could overwhelm the network and the computational resource.
        - In fact, Google estimates that a mere two minutes per day of active interaction with Google Assistant by every Android user would necessitate a doubling of the firm's data center capacity.
- Chai explains their radically different strategy to perform active inference over data at the edge.
    - Biology offers a useful model; notably, the brain processes a wide variety of sensory inputs but operates at the very low power of roughly 20 W.
        - Moreover, of the neuronal connections, only a small sliver are active at any point.
    - Chai carries over this principle of sparse activations into an artificial neural network.
        - Instead of cleaving to the conventional strategy of training a deep neural network (DNN) by activating every neuron in all layers, he instead trains it to fire at only a fraction of its potential through dynamic throttling when doing so would be beneficial.
        - "You can train a neural network today to run at 100% of its capacity—at its highest performance of accuracy," says Chai. "But the same neural network can be trained such that you run at maybe 60% or 30%. You might lose 1% or 2% in terms of accuracy, but you gain in terms of the [improvement to] latency and power efficiency that you get."
        - The question is which 30%? Just like the visual cortex, which attends to movement first and only the identification of what is moving afterward, it is possible to train a DNN to preferentially focus on attributes deemed most important.
    - A second efficiency-gaining strategy when computing at the edge is to lower the DNN's memory footprint, recognizing that there will be a trade-off between bit precision and accuracy.
        - Instead of taking advantage of the GPU's full capacity to deliver 64-bit floating-point precision during the training process, the Latent AI thrust is to decrease the target precision.
            - "We have trained neural networks down to 8-bit [precision] with less than 1% loss in accuracy," says Chai, "and we have folks who have trained things down to binary."
        - Chai also reports a gain in training efficiency with a compressed-bit network: "We find that we can train the network faster," he says. "With a 5-bit rather than a 32-bit network, we are guiding the

training much more closely such that it doesn't oscillate so much and converges faster. This is a good side effect, in addition to being much more efficient for the target platform."
    - These two design parameters—dynamic throttling and lowering the memory footprint—can be combined during the DNN training process to yield orders-of-magnitude improvement in efficiency when inferring over such models at the edge.
        - To date, the Latent AI team has successfully achieved a combined three-fold improvement in throttling, ten-fold in compression, and 25-fold in compute, totaling 750-fold efficiency gain.
    - Kandasamy emphasizes that the Latent AI training strategy is general purpose, not task specific: "It's across computer vision, it's across natural language processing, it's across time series data—any use case that you want to run."
- "When the neural network model becomes so small—we're talking kilobytes for this neural network—you can choose to move the AI model itself to the data, rather than shuffling all the data to the cloud to do the processing," says Chai.
    - Moreover, there is no need for a one-size-fits-all degree of modeling accuracy; that is, the definition of the edge can remain fluid, with high-fidelity, 64-bit floating-point models residing in the cloud for circumstances requiring high accuracy and reduced model operating on constraint-limited IoT devices.
        - "You can push this model up and down the infrastructure to be able to scale it," says Chai.
- Case study: Kandasamy describes the performance of a DNN trained to run on a 2x1-mm voice AI chip such that the chip activates its device when detecting one of several wakeup words.
    - "When [our customer] gave us the five-word wakeup model, it was 2 MB in size," he says. "We used our technology to compress it to 150 kB to run it on a 4-bit processor."
    - In practical terms, despite a limited vocabulary size, such a chip could bring voice capability to any device: "You can put it on your glasses, you can put it on your microwave, you can put it on your toaster," says Kandasamy. "I don't know why you would need it, but that is what this is enabling."
        - And the vocabulary is soon to expand, with this client slashing its timeline to release a second-generation product that will accommodate 30 words without increasing the chip's physical size.
- A side issue, separate from computing over data, is whether and how to store the data generated by the myriad sensors.
    - As suggested above, even as the amount of data on any given device compounds, in many cases most of it is uninteresting (e.g., a security camera that observes just a few seconds of interesting but randomly occurring visuals each day).
        - Anomaly detection algorithms—i.e., trained DNNs—would decide what makes the cut in this instance; if small enough, these can run on the device.
        - "Because you have smart capabilities at the edge, you can choose to store what's important," says Chai, although even with selection and compression, over time it becomes unwieldy and difficult to search unless it is well labeled from the get-go—a task that too often goes undone.
        - Compound this to 75B devices, and the necessary storage space explodes.
        - Kandasamy poses the question, "Are we going to be data hoarders, or are we going to be smart about what we hoard?"

A Conversation with Laurie Yoler—Ms. Laurie Yoler, Playground Global

- In an interview, TTI/Vanguard's Steven Cherry guides Playground Global's Laurie Yoler on a tour through companies on whose board she has sat, while also venturing into topics as varied as corporate use of personal data, the future of work, and corporate–startup collaborations.
- Specific firms:
    - Tesla (founding board member):
        - From its outset, energy storage has been an area of focus for Tesla, in significant part because of Yoler's contemporaneous work on networking products and high-performance computing.
            - In the early 2000s, the then-new concept of using commodity servers as atomic components of large computational clusters translated, for her, into the potential for using commodity batteries as atomic units in large-scale energy storage systems.
        - In 2002, with General Motors busy physically crushing its own innovative, for-lease-only EV1 vehicles, Tesla's initial decision makers determined to offer an alternative to GM's disappointed leasees by creating a simple-drivetrain-based electric vehicle of their own, while also pursuing battery backup systems for renewably produced electricity, whether mounted on the wall or rolling down the highway.

        - "It really was GM and high-performance computers that got us thinking about high-performance electric sports cars," says Yoler.
        - Nevertheless, when approaching venture capitalists, they were laughed out of the room, due to the miserable track record of every upstart car company since Ford first went public.
        - Despite having manufactured roughly three-quarters of a million cars to date, Tesla's most recent quarterly production (96K, Q3 2019) pales in comparison to the likes of Toyota (2.3M), Ford (1.2M), or BMW (613K).
        - Still, Yoler is bullish, noting that Tesla's new plant in China should boost the firm's standings on the manufacturing front before other automakers catch up with its AI prowess.
            - "I am hoping on both Tesla and the rest of the auto manufacturers to move faster to electric powertrains for all of the wonderful energy reasons that we've heard about," she says, referring to Saul Griffith's approach to winning World War Zero.
            - "I think AI is moving in some places faster than I had expected, but in many places more slowly—commonsense reasoning and generalized, horizontal AI," she says, harkening back to her early AI career in the 1980s.
            - Both for AI and electric vehicles, one major barrier to progress is social acceptance, while another is the capital infused into moving the field forward.
    - Zoox (founding board member):
        - Very optimistic about fully autonomous vehicles, Yoler is excited about the progress since Zoox's 2015 inception; it now has roughly 1000 employees, largely hardware and software engineers, pushing toward solving the urban driverless-car problem.
            - "We have had cars testing and driving in San Francisco in the most crazy urban settings, including six-way blind intersections with all kinds of pedestrians and dogs and children—and going through the tunnel that is right near this hotel—fully autonomously for quite some time," she says.
            - While recognizing that "AI won't solve 100% of anything, and it is not a great goal to think that we are going to get there," an autonomous vehicle that runs into a jam can pull over to the curb and await teleoperation management by a remote human fixer.
            - The bigger obstacles for driverless cars are regulatory and financial: A very safe vehicle requires a large investment in not only AI but also sensors (radar, LIDAR, time-of-flight, as well as cameras), increasing the vehicle's pricetag.
                - "Sensors are available, but many of them are very expensive," says Yoler.
        - For her, the most compelling use case for autonomous vehicles is to satisfy the joint safety and transportation needs of distracted, elderly, drunk, drowsy, or otherwise unsafe drivers.
    - Bose (board member):
        - "If you are selling to consumers, you have to continue to innovate and have new products that they are excited about," says Yoler.
    - The Bose product that currently excites her most is Bose Frames, sunglasses with miniaturized Bose speakers built into the temples, delivering personal sound in a stylish package—which, incidentally, sold out within two hours of the 2019 South by Southwest kickoff.
        - "I use them to walk my dog, I listen to audiobooks, I can take phone calls as I'm running through the airport with bags, and I took two board calls on these just last night," she says. "There can be many people around me, but nobody [else] can hear."
    - The company is also developing products for the hearing impaired, such as the in-ear Hearphones, which use active noise cancellation to filter out background noise and enhance the clarity of a nearby conversational partner.
        - "I get very excited about big companies that see engineering and innovation as core differentiators and that are trying to do the next thing—otherwise, they are not going to be around."
- General topics:
    - Personal data use by platforms:
        - Cherry notes that Google Maps, Gmail, and Chrome are such good products that people willingly give away their most personal data for the pleasure of using them; similarly, with Apple and Amazon. He worries that allegiance to tech giants crowds out innovation elsewhere.
        - Yoler is heartened that some governments are responding to the matter by legislating the need for consumers to have a degree of control over data that pertains to them.

        - Hearkening back to her time at Visa, which faced the same concerns of pervasive access to consumers' financial and purchasing data, then—as now—regulation on data use and sharing tamped down overstepping by the firm.
    - Today the discussion has expanded to the possibility of firms being required to pay consumers for the use of their data.
- Future of work:
    - Cherry, raising the matter that AI—whether in autonomous vehicles, in medical or legal settings, or elsewhere—is poised to displace large number of workers in the near future, asks for Yoler's opinion on the changing world of work.
        - Her sister, a radiologist, has heard all the dire predictions about this specialty giving way to AI pattern matching, but is actually enthusiastic about the diagnostic progress AI is ushering in.
            - As a second-level radiologist, the vast majority of cases escalated for her oversight needn't have been; she expects AI wouldn't make the same kinds of errors as her junior colleagues.
            - Moreover, with all the radiology-as-profession doomsaying, practitioners in the field are becoming scarce, and AI is now all the more necessary to keep up with the workload.
            - "Bring it on," says Yoler's sister.
    - Similarly, there is a shortage of truck drivers, so letting vehicles manage themselves on the highways with human drivers taking over for city streets and loading/unloading tasks would also be a boon.
    - Yes, there will be workplace disruption and employee displacement, but the net outcome will be favorable, in Yoler's opinion.
    - Creative workforce management also makes a positive difference, as Yoler reports regarding consumer goods firm Church and Dwight (e.g., Arm & Hammer, Water Pik, Trojan, various toothpastes), on whose board she sits.
        - Seeking to bring all of its employees up to speed regarding digital technologies, it very successfully paired employees with longer tenure in the firm and newer (and largely younger) hires to provide mutual mentorship.
        - Not all firms exhibit this degree of introspection, causing Yoler to worry about those that do their employees a disservice on the development side.
            - Believing that continuing education is essential for employees at all levels of a company, Yoler regularly takes night classes in all manner of subject areas.
- Corporate–startup collaborations:
    - "Startups desperately need large companies," says Yoler. "Besides needing them for money, they need them for validation, they need them as customers, they need them to run pilots, and they need them just to talk about what their requirements are. If large companies aren't going to be [startups'] customers in the future, they aren't going to be successful."
    - On the flipside, corporations need startups to provide innovation, but many struggle with how to make useful connections, toying with attempting to do so at the university level, setting up an incubator or a venture group, or setting up an office in the Bay Area where "two guys sit and meet with startups all day long—and how are they bringing that knowledge back to the mothership?" she poses.
        - Supporting innovators within the company and discovering and connecting with outside innovators are both essential for a large firm to remain relevant.
    - "I don't care what your sensors are, but you had better have sensors in all of the interesting and disruptive areas that are starting to come on the horizon, because it is easy to pooh-pooh them," cautions Yoler.

Active Inference and Artificial Curiosity—Dr. Karl Friston, Wellcome Trust Centre for Neuroimaging and University College London

- Free energy minimization is a tried-and-tested tool of theoretical physicists; with that principle in hand, much of understanding the world lies only in mathematical manipulation.
    - Couch this in the language of information theory, and it becomes possible to incorporate the uncertainty and statistical nature of machine learning into the formulation.
    - First some basics: The (negative of) free energy can be expressed as the difference between complexity (the bounds of a system) and accuracy (its evidence), and is always greater than the

evidence expressed as the negative logarithm of the joint probability over the model and its set of temporally ordered outcomes:
    - $-F = D[Q(s_\tau)||P(s_\tau|o_\tau)] - E_{Q(s\tau)}[\ln P(o_\tau|s_\tau)] \geq -\ln P(o_\tau|m)$.
  - "To minimize the free energy [*F*], or maximize its negative, I will set that bound [*D*] to zero, and then this now becomes an approximation to the negative evidence [*E*], by which I mean some data given a generative model of how those data were generated," says Karl Friston of the Wellcome Trust Centre for Neuroimaging and University College London. (*P* is the probability operator; *Q*, represents posterior beliefs; o_τ and s_τ are the outcome and state, respectively, at time τ; *m* denotes the model.)
  - Friston offers up a game for sussing out rules to serve as motivation for the importance of generative models: "With a formulation based on a generative model, you have for free explainable and interpretable AI," he says. "It is there at the beginning; you have built that in."
    - This is true whatever the specifics of the generative model, but "because you are starting from that you always have an interpretable and explainable way of making inferences and organizing and sampling your data," says Friston.
  - Put another way, the generative model permits the compression of the evidence present in the data into the form of a hypothesis.
    - "You can think of this free energy—or log of evidence, given a model of how the data were generated—as the difference between the accuracy and the complexity," says Friston. "What [the free energy expression, above] says is that you want the most accurate account of your data in the simplest way possible. Mathematically, it corresponds to the difference in terms of probability distributions of your posterior beliefs, having seen some data, to your prior beliefs based on the states of the world that were generating those data."
  - The physicists' approach, therefore, provides a shortcut to design optimality specified in a single objective function, namely minimization of the variational free energy, and moreover the generative model offers up a principled way to explore the data—"not with big data, but sparse data," says Friston. "You don't want lots of data, but the right kinds of data, given what you think you know about the world that is generating those data."
    - Built into this energy minimization formulation is the principle of Occam's razor—i.e., simplicity wins, all else equal.
- And he dives in:
- First a game to measure the Bayesian reasoning ability of the TTI/Vanguard community; the objective is to determine the abstract rule—the generative model (*m*)—governing the placement of colored dots, given the prior that the correct outcome depends on the color of the central dot.
  - Outshining other groups (relative to group size) Friston has tested, the TTI/V audience required just four instances of triplets of red, blue, and green dots to determine that the answer to a blue/red/green central dot is the dot on the right/left/center, respectively.
    - The first several test cases: red/blue/green, green; green/red/red, green; red/red/red, red; green/green/blue, green; red/green/green, green; blue/green/blue, green; blue/blue/red, red.
    - In fact, his expectation was to present eight examples before the group determined the rule.
  - Note that there is no deterministic answer to the number of instances for complete knowledge of the rule using approximate Bayesian inference embedded in the process of variational free energy minimization; "There is no truth," says Friston, "just the explanation that has the greatest evidence, which is operationally is the most accurate and the most simple."
  - Importantly, successful rule determination entails understanding the very nature of rules.
    - "You know what a rule is; well, you know what I mean by a rule," he says. "You know that there is some symmetry there—some invariance there—that you could exploit, which means that the space of hypotheses that you could explore in terms of that likelihood mapping is vastly reduced, so that you can step through a number of hypotheses and say, 'Oh, that fits! It accounts for all of these exemplars.'"
- Next, apply the perspective of free energy minimization to inferences over data, specifically active inference, which involves taking in observations of the world and using generative models to make sense of those observations based on prior beliefs developed through prior iterations of the process.
  - Various timescales apply to active inference and self-evidencing:
    - short-term (observation), for inferring rapidly changing states of the world;
    - medium-term, for learning model parameters;
    - long-term, for learning model structure.
  - "Not only do you have learn the best parameters—that is, the free energy-minimizing parameters of any given generative model—but you also have to learn the structure of the generative model if you

are working with a deep-learning neural network," says Friston. "How many layers, how many parameters, what kind of factorization comes into play, do you have symmetries or translational invariance in convolutional networks? All of these structural aspects should all be chosen to minimize this variational free energy."

- He fleshes this out with an example: a hungry owl that sets out to hunt for prey.
  - Implicit in the activity of hunting is exploration—in this case of the physical world, perhaps for mice in a prairie grassland.
    - "Why do you explore?" poses Friston. "Essentially, to reduce your uncertainty about the world in which you are operating, but uncertainty is an attribute of your belief, not an attribute of the world."
    - This suggests the expression of (Bayesian) beliefs as probability distributions, and the objective function governing the predator's actions not being a function of the world after being acted on by the owl, but rather of the state of the world the owl observes from on the wing (i.e., Q in the free energy equation).
  - This setup suggests two alternative approaches to optimization:
    - Bellman's optimality principle, where optimal action depends on states of the world.
      - Operationally, this approach draws on optimal control theory, dynamic programming, deep reinforcement learning and expected utility, Bayesian decision theory, state–action policy iteration, and the like.
    - Or Hamilton's principle of least action, where optimal action depends on *beliefs* about states and on subsequent action.
      - Operationally, this approach draws on the free energy principle, active inference, artificial curiosity and intrinsic motivation (in robotics parlance), optimal Bayesian design, sequential policy optimization, partially observable Markov decision processes, and the like.
    - Only Hamilton's principle addresses hunting, however, since exploration of the physical (or state) space is rooted in beliefs-based exploration before action (swooping down on prey) becomes possible.
  - Active inference involves the iteration of a two-step process: first, policy selection, based on Bellman's principle, updates the state of the world, which in turn feeds into new observations that underpin perceptual inference performed using Hamilton's principle; outcomes then serve as the basis for additional policy selection, and the cycle continues until achieving acceptable convergence.
    - "We actually infer what we would do *if* we were self-evidencing," says Friston, "because we have written down the imperative, which is to minimize free energy or maximize evidence. We associate the probability of a particular policy with its tendency to increase the log of evidence as a consequence of its action. Now we have the concept of 'expected free energy' given a sequence of action."
    - He assigns the expected inaccuracy to ambiguity, and the complexity cost to risk.
      - The task of minimizing free energy then becomes one of jointly minimizing risk and ambiguity, when averaging over outcomes due to future actions.
    - The relationship between the probabilistic approach to uncertainty and the state space approach to optimization is the exclusion/inclusion of the uncertainty term in the free energy equation, $E_{Q(s\tau)}[\ln P(o_\tau|s_\tau)]$; that is, purporting that $E_Q[\ln P(s_\tau)] = E_Q[\ln P(o_\tau)]$.
      - Considered more intuitively from a statistical perspective, the expected free energy is the sum of the expected value considered in Bayesian optimal decision making and the information gain of Bayesian optimal design; the two processes taken together amount to active inference.
        - "You want to forage around for information to shrink your uncertainty, so you are more confident about the states of the world," says Friston. "Once you have done that, you can make your move to go and secure your preferred outcome. This is very much like the hunter doing the predation."
- Friston then simulates elucidating the rules of the color game using the principles of variational free energy minimization, active inference, and self-evidencing; as such, he demonstrates novelty and epistemic affordance.
  - "This synthetic agent has to look around and choose where to look, so it does its epistemic foraging to choose the data that will reduce its uncertainty about what caused the sensory input, and then it makes a decision in exactly the way we have theoretically described," he says. "All the heavy lifting done in this particular kind of problem is all in the likelihood mappings between hidden/latent states

out there that you cannot observe and the outcomes that you can observe. If you can get that structure right, you can start to understand the rules and contingencies that govern your sensory data and the consequences of your actions will play out. This is where the rule lives."
    - That is, the rule—the generative model—is rooted in the context-sensitive mappings between the state of the system (the positions of the colored dots) and the outcome (look right, left, or center if the center dot is blue, red, or green). Suss this out, and the world makes sense.
  - When feeding the game fed into a Bayes-optimized simulation, 15 examples were required to reliably infer the rule.
    - "The confidence level is consistently high after about 15 data points or observations, after which the effective free energy is consistently low, and there is 100% [confidence]," says Friston. "You stop exploring and start being quite exploitative, going straight for the answer."
    - As clever humans, with a predetermined linguistic and epistemic understanding of the nature of rules, TTI/Vanguard audience members reached this point in far fewer iterations than the machine.
      - "What I think is a rule is something very specific to me, being a human being," he says. "We have created a world of structures and concepts and memes in a way that makes our predictions and our self-evidencing much more efficient and much easier. Not only do we go out and get the right kind of sparse data that resolves our uncertainty, but we make that job easier by arranging the environment and using, literally, language and discourse to make the self-evidencing as efficient as it can possibly be."
  - Just as new observations of the system's states feed into perceptual inference to improve the basis of the next step in policy selection, so too do they advance the parametric learning pertaining to the generative model itself.
    - "That means that I will choose those behaviors that reduce my uncertainty about what will happen if I [take some action]," he says. "This is, in effect, artificial novelty."
- Gary Marcus offers a word of caution regarding Friston's approach to adaptive inference:
  - "As a theory of AI, this seems too abstract to me," says Marcus. "I'm sympathetic to it—you are trying to optimize the right things—but if I were building, say, a driverless-car system, I wouldn't know where to begin in terms of how I actually do the real-world counterfactuals: What would have happened if I pressed the gas here?"
  - "On the psychology side," Marcus continues, "people do all kinds of nonoptimal things all the time, and you didn't engage in the Kahneman-Tversky literature."
  - Friston appreciates Marcus's points, but as a neuroscientist/physicist, his theory is intended for neither AI nor psychology, but rather as a means to formalize self-organization.
    - Nevertheless, he sees an AI application, using Marcus's example as a launching pad:
      - An autonomous car, fitted with myriad sensors, could use adaptive inference to learn a generative model for where best to point those sensors, since this problem reduces to seeking a low-dimensional model from a large yet sparsely relevant dataset; "The whole imperative here is complexity minimization," says Friston: "finding low-dimensional manifolds that provide the best and simplest account in the high-dimensional data."
        - "This is the active vision problem," he says. "To get that working properly is all in the generative model, although I think we are many years away from putting this into an autonomous vehicle."
    - Regarding psychology, whether behavior is to be considered rational/irrational depends on the individual's prior beliefs; that is, the person's internal rules for making sense of the world.
      - Friston's models offer a means to generate such rules from a person's actions and thereby establishing a basis for understanding—and, when appropriate, for cognitive adjustment.

Building Artificial Intelligence We Can Trust—Dr. Gary Marcus, New York University

- Deep learning might be all the rage, and in fact is often deemed synonymous with artificial intelligence, but New York University's Gary Marcus provides a pointed reminder that AI comprises much more than just deep learning, and in fact deep learning is but a subset of AI's subgenre of machine learning.
  - "Machine learning is about getting machines to figure out things about the world on their own, rather than us teaching them," says Marcus.
  - Not to be discounted are the nonmachine-learning strategies of planning, reasoning, search, and knowledge representation, nor other machine-learning categories, including probabilistic learning, decision trees, and genetic algorithms, that share the shelf of tools with deep learning.

- Should we believe deep-learning pioneer Andrew Ng when he tweets *If a task takes you less than one second of thought, a machine can probably do it* or when Google's Sundar Pichai says that AI is "more profound than electricity or fire"? No, believes Marcus—or at least not yet.
  - "Don't leave home without fire or electricity, but you can do without AI for now," he says.
- However, neither does Marcus buy into Elon Musk's stated attitude that with artificial intelligence "we are summoning the devil"—despite also estimating that Tesla owners will be able to tune out entirely and let Autopilot take over by mid-2020, even in light of the occasional Autopilot-induced death involving people lulled into inattentiveness while behind the wheel.
- So, which is it: Will AI be a savior or a deadly menace?
  - Marcus returns to TTI/Vanguard to express his short-term pessimism, but long-term optimism, regarding AI and, importantly, to inject a dose of realism into the debate around deep learning in particular and AI more generally.
  - "I don't think AI is impossible," he says. "What I do think we need to do is temper our expectations."

- Strengths and limitations of deep learning, which is the most popular AI tool currently:
  - "Deep learning is a very good tool, but it is not perfect," says Marcus. "It is not anywhere near perfect."
  - Deep learning's basics:
    - Provide a deep neural network (i.e., a neural net with a large number of layers) with enough examples (i.e., with big data in a narrow regime) and it will generate complex statistical correlations to effectively categorize additional instances of data within that regime.
      - "Early neural networks had an input layer and an output layer," says Marcus. "Then it was discovered that you could add a hidden layer to make a three-layer network. Now we can train neural networks with over 100 layers. That's what deep learning means: a deep stack of these layers. [The name] makes it sound conceptually deep, but it's not."
  - Deep learning's upsides:
    - Photo tagging is a widely accepted application of deep learning: "I used Apple's tool to help me find a picture I took of a giraffe five years ago," says Marcus. "It took two seconds, and I was really happy."
    - Speech recognition is another area of deep learning's success, but language understanding is not: "Siri can now mostly transcribe my requests, but it doesn't mean it has any idea what to do with my request," says Marcus. "Usually it sends it to a webpage that is vaguely relevant to what I asked."
    - Game playing, both board games and videogames, are other areas in which deep learning has come to shine.
  - Deep learning's downsides:
    - As suggested by the Siri example above, deep neural nets can also fall flat, as is also clear from these image classifier examples:
      - One trained on many photos of elephants "learns the textures of elephants," says Marcus, but not that an elephant has one trunk, two big floppy ears, four thick legs, a familial society, and so forth.
        - If trained only on well-lit elephant photos, it could well misidentify a silhouetted elephant on the beach at sunset (because photos in training sets of silhouetted figures in such scenes tend to be of people).
      - Another classifier misidentified, with high confidence, a crashed school bus laid flat across a snowy road as a snowplow, due to similarities in pixel texture and scene.
      - Individually, a classifier correctly identified a banana and a toaster, but put a small sticker of the toaster next to the full-sized banana, and the system honed in on the toaster, without recognizing that it much too small to accept an actual slice of bread.
    - Move outside of deep learning's wheelhouse of pattern matching-based classifiers, and AI's list of successes dwindles.
      - "Language understanding is something that we really haven't made any genuine progress on at all," says Marcus. "On making AI flexible and adapting to unusual circumstances, which is what I think is the heart of intelligence, we haven't made progress since the 1950s."
        - Understanding the world? No. Driving has proven difficult, and there is no reason to believe that medical diagnosis will be given over to AI without considerable human oversight anytime soon.

- ELIZA, the 1966 computer simulation of a psychotherapist, extracted keywords from users' input to fit into preprogrammed leading questions or empathetic statements—quite a development for its time for text-generating software.
  - Fast-forward to today, and Musk's OpenAI text generator GPT-2 offers what appear to be cogent answers to interview questions posed by an *Economist* journalist, until one looks at the fine print and realizes that the human selected among five generated responses to each question, choosing the "most coherent or funniest."
    - "That's cherry picking and completely misrepresents what the system can do," argues Marcus. "If you actually talk to it, you will not be able to replicate—to use a word that has become popular in science, now, after far too long—what [the journalist] got, unless you replicate his methodology and use your *human* measure of coherence."
    - This is not just Marcus's supposition; he did the experiment, asking GPT-2 the following question: *Two lemurs walk on a road and another joins in. The total number of lemurs on the road is ___?* Numerical answers jumped around among 1, 6 (morphing into 3, 6, 4, 7, 6, 9, 9 (again), 10, 11, 11 (again), 12, 13, 14, 14 (again)), and "not 100 as claimed, but about 80 or so," some with lengthy explanatory verbiage, but never resorting to actual 2 + 1 = 3 arithmetic.
      - This GPT-2 example, along with others equally nonsensical, demonstrated that the AI's fault is not only innumeracy but also deep lack of context and understanding.
        - GPT-2 wasn't reasoning, but rather cutting and pasting somewhat-relevant text from Reddit.
- Marcus has long held his contrarian view of deep learning, having written a *New Yorker* article in 2012, entitled "Is deep learning a revolution in artificial intelligence?"
  - His answer then, as now, is that deep learning is ill-equipped to represent causal relationships, perform logical inference, or integrate abstract knowledge; in essence, AI won't be worthy of the name until systems incorporate artificial approaches to the many different ways that humans observe, think, and reason.
    - "There has been lots of progress [since 2012] in narrow AI, but I don't think there has been any progress in the [other] areas I mention," says Marcus.
- Further exploring the matter of numeracy, recent research continues to expose deep-learning systems as thoroughly hopeless when it comes to mathematical reasoning; in fact, a paper this year touted a neural net's ability to achieve 90% accuracy with arithmetic; i.e., claiming success with the result that 1 + 1 +1 +1 + 1 +1 +1 = 6 (count 'em: there are seven 1's); "When they put in parentheses, accuracy dropped to 50%," condemns Marcus, although the paper reports improved results when combining larger—and, notably—different integers.
- Marcus performed his own experiment (in 1998), by training a parallel distributed processing (PDP) auto-associator network on binary numbers; that is, the output number is always the same as the input—the simplest of functions.
  - This task poses problems for backpropagation networks.
  - The network failed to generalize this "universally quantified one-to-one mapping" to odds (binary strings ending in 1), when trained only on even numbers (binary strings ending in 0).
    - "It will generalize to some other even numbers, but it will never generalize to the right-most bit," says Marcus, "although that is a lawful thing, statistically speaking, for the network to do, but it is the wrong inference from the human perspective."
  - The greatest downside of using deep-learning networks—earlier dubbed multilayer perceptrons—as models of cognition is their inability to stray from the bounds of their training space, whereas "humans can freely generalize from restricted data," Marcus wrote in 1998.
- "The problem [of ungeneralizability] is still here," he says. "The reality is that deep learning works best in a regime of big data, and it works worst with unusual cases. You move out to the tail and it just doesn't work very well anymore."
  - AI is about more than perception, the only realm in which deep learning stands a chance.
  - In contrast, consider the structural differences between life and games like Go, chess, and Shogi.
    - The games, like life (as distinct from Life, the game) are highly complex, but life cannot be perfectly simulated, even with unlimited resources and data, and decisions must be made with imperfect information; arbitrary input parameters can be consequential in life, rather than the immediate state of the game board, and the uniqueness of situations in life mean that only modest amounts of data are available per task.

        - "Life is very complex, just like Go is, but you have imperfect information, and you can't simulate it perfectly to gather lots of perfectly valid data," says Marcus. "You only get a limited amount of data for most of the tasks that you do in life."
        - That is, the tools of deep learning do not pertain to most that occurs in life.
        - As is always true with applied mathematics, the problem must be well matched to the tool; for deep learning, this translates to adherence to relevant statistical assumptions—such as mutual independence and normality of observations, and similarity of training and test data.
            - Deviate, and perception yields untrustworthy outputs.
    - This underpins the dangerous current state of AI: "It has a trust problem. We are relying on AI more and more, but it hasn't yet earned our confidence. It doesn't deserve it yet," says Marcus. "You take a system that can't tell an elephant in silhouette and you apply the same techniques to some real-world problem about deciding who should get a job or recognizing faces for crime or whatever, you have to be really, really careful."
- How, then, might we climb out of this AI hole? Marcus has two recommendations:
    - Raise the bar:
        - Instead of being satisfied with statistical approximation (i.e., the correlative process that underpins deep learning), the AI community should focus research attention on tasks that demand reliability.
            - "I like to use the example of, if you built an eldercare system that lifts Grandpa into bed 80% of the time and drops him one in five times, you are going to go out of business," says Marcus.
        - Instead of being satisfied with closed-end problems (e.g., structured games), the AI community should focus research attention on tasks relevant to the open-ended real world.
    - Look at human cognition for inspiration:
        - Humans learn language, and humans learn about how the world works—machines do not yet.
        - To address this need, Marcus—in conjunction with cofounders Rodney Brooks, Mohamed Amer, Anothony Jules, and Henrik Christensen—has launched the company Robust AI with the goal of "building the world's first industrial-grade cognitive engine, a hybrid deep comprehension-driven platform to allow robots to act intelligently in the dynamic, open-ended real world."
            - "The question that we are trying to solve is how do you build a high-level operating system for robots that allows robots to deal with the open-ended real world, to deal with the unexpected, to give them situational awareness and the ability to anticipate the consequences of their actions," he clarifies.
- That is, perceptual cognition, as instantiated in deep learning, is only one tool among many; people exhibit not only perception and planning, but also common sense, analogy, and language—human intelligence is multifaceted.
    - Steven Pinker's research on language, to which Marcus contributed, suggests that natural-language understanding inherently relies on both a rules-based system and a neural network in combination.
    - "We need to have new tools," says Marcus. "Why shouldn't we bring symbol manipulation—by which I mean operations over variables, like you see in algebra or lines of computer code—back into the fold? It doesn't have to be binary [deep learning *or* symbols]."
        - The key concepts of this approach are variables, binding, and operations over those variables.
            - "When a human learns the identity function $f(x) = x$, they don't care which examples they have applied it to before, they just know that is the rule and I will generalize it."
        - Marcus does not claim that symbol manipulation itself constitutes learning, but as a machinery for representation and computation it offers an orthogonal information-processing strategy to deep learning; the two can be used in conjunction, each where it is best suited.
    - "We can't get to AI we can trust by relying on deep learning alone," he says. "It's good for some learning, but poor for abstraction. Classical AI—symbol manipulation [including Doug Lenat's monumental work with Cyc]—is not going to get us to robust AI either; it is good for abstraction, but it is not particularly good at statistical learning. What we really need are hybrid models that bring together these two traditions."
        - Note that Lenat would likely argue that Cyc offers many counterfactual examples to Marcus's claim of limitation, despite Marcus's criticism that it lacks a natural-language front end.
    - The primary point is that learning is a multifaceted task, involving not only big data and number-crunching, but also "exploration, problem solving, and an intuitive understanding of physics," says Marcus.
        - That Marcus requires an entire talk—and an entire book—to make his point about the need for a multipronged approach to AI might surprise, but personal experience launched him on his quest to get the word out.

        - Despite what Marcus believes is a concerted attack by AI luminaries (e.g., Yann LeCun, Dan Brickley) against not only the need for hybrid models but also against himself personally, he adheres to an argument the connectionists should appreciate, namely that the ability to wield both a hammer *and* a screwdriver—i.e., distinct tools toward a common end—is better than only one or the other.
    - Echoing the cognitive science-based refrain on which Lenat has built his career and company, Marcus says that deep understanding—as opposed to deep learning—must begin with common sense.
        - An example of how commonsense reasoning goes beyond mere perception to assist humans in categorization tasks: A yarn feeder is a container with a hole. Place a ball of yarn in the container, feed the end of the skein through the hole, and the yarn remains untangled for the duration of the craft project. A wooden box, a silk bag, or a ceramic bowl (even one decorated as a cat head) serves this purpose—"In five seconds I can explain it to you and you will understand what it."—but deep learning would be hard pressed to categorize them as functionally equivalent.
            - "Once you recognize [the concept of a yarn feeder], because you have an unconscious causal theory of how the world works, you can now recognize another one in which every pixel is different," says Marcus. "Instead of memorizing pixels, you are recognizing ideas."
- General artificial intelligence will not arrive until systems can "set their own problems—in part by doing active learning to determine which problems are interesting—and have good causal understanding of the world through commonsense reasoning," says Marcus.
    - To do so will, in turn, require some innate knowledge of the world; Immanuel Kant recognized this in the 18th century, with his argument of space, time, and causality being innate structural features underpinning understanding of the world—that there is room for both rationality (symbolic reasoning) and empiricism (perceptual learning).
    - Marcus expands Kant's terse list with additional fundamental capacities that evolution has equipped humans with:
        - representations of objects; structured, algebraic representations; operations over variables; type–token distinctions; capacity to represent sets, locations, paths, trajectories, and enduring individuals; a way to represent an object's affordances; spatiotemporal contiguity and conservation of mass; causality; translational invariance; and capacity for cost–benefit analysis.
    - He adds to these additional innate human endowments to draw on to solve particular adaptive problems:
        - intuitive mechanics; intuitive biology; number; habitat selection; danger, fear, caution, phobia; mental maps for large territories; food; contamination; monitoring of well-being; intuitive psychology; mental Rolodex; self-concept; justice; kinship, nepotism, parenting; mating, sexual attraction, love.
    - That is, every human comes with a hot mess of potential; "We need to have similar innateness in our machines," says Marcus.
    - In the meantime, humankind should not worry about domination by embodied AI (robots), as per the performance of top entrants in the DARPA Robotics Challenge, which regularly toppled over and otherwise failed to perform as intended (and meticulously practiced).
        - "Don't worry about the robot coming for us anytime soon, but if they do, close the door; hide behind a bus, dress like a bus, or hide behind a shiny toaster; keep a pack of psychedelic stickers on hand, and a giant fan, and banana peels to throw on the floor; talk in a noisy room with a foreign accent; in the worst case, climb some stairs; oh, a couple of robots can do that, so climb a tree—none of them can," says Marcus; any of these will totally flummox a robot.
    - But, also in the meantime, all manner of robots and unembodied AIs (virtually) roam the world, many overly dependent on deep learning without backup from deep comprehension.
        - From this perspective, we *should* be afraid:
            - Today, we have come to rely heavily on AI without it yet being trustworthy.
            - Eventually, AI will be both powerful and trustworthy, helping society solve its most challenging problems, but that day is not yet here.
            - "Now that AI is out of the box, we *must* push the field forward," he emphasizes, and, donning his psychologist's hat, "If we want to build machines that are as smart as small people, we should start by studying small people."

Information Retrieval for Digital Commerce and Digital Workplaces—Mr. Will Hayes, Lucidworks

- Who would have thought enterprise search would still be an area in need of creative solutions?
  - Will Hayes of Lucidworks lays out a case for upping the AI game when applied to the problem of enterprise search.
  - Even as consumer-grade search has grown into an enjoyable experience—starting the day with Siri delivering the weather forecast, Google Assistant displaying local restaurants in Maps when sharing an expression of hunger with a texting partner, or Netflix suggesting the next TV show or movie you're sure to love—enterprise search remains mired in old-school interfaces teeming with blue links, even though more useful and pleasing options exist.
  - Hayes describes the need for enterprise search to not only capture, but also to predict users' intent, thus personalizing search results, sometimes zeroing in on just the right document and other times presenting a range of options to stimulate exploration, depending on the user's contextual need.
  - Search results might arrive in the form of documents or as potential collaborators, but they should always enhance productivity and improve the experience of work.
- When Hayes thinks about AI, it is not in the context of the fear, uncertainty, and doubt, that the dark side of AI can engender; nor is it in the context of overhyped optimism or, on the flipside, cynicism.
  - Instead, "the reality is that we have been experiencing—in the most practical sense—artificial intelligence for probably the past 20 years through search technology."
    - As AI has improved, so too has search.
    - "Over the past ten years, we have come to see how much smarter search engines have become in interpreting our intent—interpreting users," says Hayes. "This has allowed us to unlock a lot more value from the world's information and a lot more value from data in general. It has been search technology that has been activating AI within the enterprise and within our daily lives."
    - Natural-language processing, machine learning, computer vision, and deep learning all derived from the quest for better search.
  - His prior firm, Splunk, initially focused on enveloping enterprise data with Google-like analytics "to give system administrators a better way to parse through logs," says Hayes.
    - "What we didn't anticipate was what democratization of machine information could do for our customers," he says, "even though the intention of the tool was just to replace grep within the terminal."
    - The fact that Splunk has grown into a $19B company exemplifies the power of how a natural interface into data can bring value to users.
  - Hayes has learned that the more intuitive the search process, the more it engages users and the better the outcomes.
    - "As we learn more and more what a user's desire is and how we can delight the user, we create this really productive experience," he says.
    - Indeed, natural-language processing, machine learning, computer vision, and deep learning all derived from the quest for better search.
  - Now there is danger in getting search wrong—in not meeting the expectations of the searcher.
    - It is not only members of the young, tech-savvy generation who have come to expect search to be an engaging experience; we all have.
    - This is as true when it comes to engaging people in the workplace as it is in their personal lives.
      - "Those expectations are there," says Hayes. "What makes folks frustrated is the inability to be productive."
    - It wasn't so long ago that asking Google for the score of an ongoing game or for the weather in some particular locale would return a list of links that the user had to sort through for an answer; although it isn't always right, the search giant has generated an expectation of a clean answer to even a moderately ambiguous question.
      - Lucidworks strives to offer the same experience to workers when asking questions related to enterprise data.
- "I challenge people to think about digital transformation a little bit differently [than in the past]," proposes Hayes. "Now we need to think about what does it mean to be successful in this digital era—to meet those expectations that our users have when they come into work?"
  - While details of digital infrastructure matter at some level, the end user in the enterprise—the one using the enterprise search tool—doesn't care a whit about the data per se; instead they care about how it can help them do their jobs.

    - "I argue that the reason that big data is failing is that we are looking at it from the wrong perspectives," he says. "We're looking at it from the infrastructure, we are looking at it as just a bunch of bits and bytes sitting on disk, and we're trying to pretend that these things are actually exciting."
        - But digital success is not about bits and bytes; it's about their meaning and their accessibility for, relevance to, and beneficial contribution to the job the user faces in the moment.
            - "In this shift from thinking about digital transformation to digital success, we have to start thinking about people," says Hayes. "There will always be a human being on the other end [of the digital product] who has an expectation, who is trying to accomplish a task, who wants to be productive and successful in what they are doing."
- "Think about the moment," implores Hayes: What can the tool bring to a user in the moment when solving a pending problem? Satisfying this need is the goal.
    - While digital design remains important—Is the shade of blue appealing? Is the linked object visually obvious and easy to click?—"the way you delight a user is meeting that expectation."
    - Instead of focusing on applications, which Hayes describes as "an instantiation of the user, a workflow to present them with information to get them to some kind of end goal," he instead zeroes in on the intention of the user at every given moment and meeting the immediate need.
        - Users have distinct needs when clicking on a notification within an email versus when opening a customer relational database, and responses to such actions must differ accordingly.
    - An example of success in this regard is Uber: From the perspective of a traveler wishing to go from airport to hotel, no longer is it necessary to acquire a physical map from AAA, discern a route, and acquire a vehicle; just set the endpoint location, request a driver, hop in, and enjoy the journey.
        - Mapquest made strides in the early days of digital by planning a route, but travelers still had to print the map, rent a car, and navigate their way through a strange city.
        - "Uber is digital success," says Hayes. "It shapes experiences."
    - Similarly, "Reddit distills the entire Internet into a single delightful experience" by making curated content instantly discoverable and sharable.
- To succeed in today's environment, an enterprise must rely not only on digital tools per se, but also the institutional knowledge, corporate culture, and human expertise that makes that firm unique.
    - "It's about leveraging the knowledge [within] a person," says Hayes. "How do we make them more effective. We can't simply have machines making decisions for us; the job of the machine is to present us with choices—to simplify."
    - Retailers have figured this out with respect to their customer-facing platforms: put an item in the digital shopping cart, and suggestions pop up, not only for similar items (another pair of brown shoes), but also coordinating items (an insole or a brown belt).
        - Attempts sometimes go awry; Amazon's massive and diverse customer base and sales volume can yield odd suggestions, such as a pair of shoes when shopping for a lawnmower, or repeated offerings of toilet seats once one has been purchased, leading one customer to complain that the purchase represented a one-time need, not a persistent obsession.
- The enterprise environment is susceptible to overprovisioning in this way, as well.
    - "In the workplace, we have so much potential because we understand our users—we know what their roles are, we know what departments they work in, we typically know what tasks they perform throughout the day—but our job is not to simply automate those things, because if we do we are going to start promoting toilet seats," analogizes Hayes. "We need to use these technologies to put choices in front of people so we can begin to understand what those interactions and end goals are really about."
        - The feedback to the enterprise from actions employees take within their workplace applications constitute a wealth of data that can serve as a training set for machine learning to improve the choices presented to workers going forward.
        - That is not to say the task of building user intent models is easy; it is not, particularly if the developers of relevant systems fail to keep top of mind that the aim is always to satisfy the end users—the workers in the enterprise, people with in-the-moment intentions, goals, and context.
    - When presenting his vision—and Lucidworks' solution to customers—Hayes sometimes receives the argument that Google, applied to the enterprise, should suffice, to which he counters that Google does not and cannot know each firm as well as the firm knows itself.
        - As such, he recognizes the journey inherent in application development, a journey that indeed begins with data and data structures, but does not stop there.

        - Stage 1 is data operationalization; stage 2 is to refine the relevancy; stage 3 involves personalizing results; the final stage, stage 4, is the instantiation of intelligence, when the system reliably predicts user intent.
            - Success within each of these stages is relative; for instance, relevancy could be measured as term frequency of text strings or be more refined, as is when the system learns and responds to modes of user interaction.
                - "We continue to operationalize not just the content, but all of the data produced by these applications, and try to get to the next level of personalization by taking these relevancy models and starting to look at them on the individual basis," says Hayes.
                - Part of the process is to establish a profile for each worker, in a manner similar to how a commerce company builds customer profiles that reflect each customer's tastes, preferences, and shopping habits.
    - Such an initiative need not involve "massive teams of data scientists or massive amounts of training data," ensures Hayes. "In a lot of cases, it just requires understanding a user's intent and then analyzing a user's behavior, and then marrying those two things," just as the best consumer sites do to make the user experience exceptional.
    - "I like to call this the DNA of great experience," he says. "That experience, today, is the most important thing for our users, the most important thing to make folks successful. And, search, I think, is one of these unique paradigms that really intersect people and technology."
- Two open-source technologies—SOLr and Spark—underpin the Lucidworks enterprise solution, Latent AI.
    - Hayes's firm both benefits from and contributes to the community effort.
        - "We are really grateful to have so many great folks out there in the community that we don't require the same teams that the Facebooks, Amazons, and Teslas have deployed," he says. "We can do things within our working groups, with bars coming down much lower as these technologies become much more attainable. Open-source is definitely a key to accelerating the adoption [of the potential of search]."
- When building the Latent AI search engine, Hayes looked to the Amazon user experience as a model.
    - The effort began seven or eight years ago, just when machine learning began to become generally accessible, by building a platform rooted in machine learning with search as the target application.
    - "We can help with things like question–answering and understanding," he says, noting that among his 400 enterprise clients are retailers like REI, Home Depot, and Lowe's.
        - These receive customer-facing solutions, while for others the end user is the employee.
    - "It is 'operationalized AI' using practical data science," says Hayes.
        - The platform ingests data from all available enterprise sources—documents downloaded, messages sent/received, items ordered, and so forth—and from these infers intent.
            - As expected, there is a tradeoff within the enterprise setting between privacy and utility.
                - "In some industries, it is regulated, but in other companies the question comes down to to what factor can I exchange creepiness for value," he says. "It is the same thing as in our personal lives."
        - The machine-learning models and the integration between the search engine and the machine-learning system—collectively dubbed Fusion—are proprietary to Lucidworks, although much of the rest of the platform is maintained as open source.
- Case study: an internal question–answering platform for a large, venerable wealth management firm.
    - The firm recognizes the value inherent in its long-term data, experience, and knowledge, and applies the new platform to convert these into a competitive asset to fend off the myriad upstarts attempting to eat its lunch.
        - "Through natural-language processing and query understanding we have been able to take huge corpuses of information—we're talking terabytes of research and content—and start to pull and summarize pieces of that to answer questions that a financial advisor might have," says Hayes, to retain the company's talent.
            - "We provide an employee data profiles that help understand the way people interact," says Hayes, which in turn "helps make decisions [pertinent] to those moments. When you log into a portal, we know immediately that you have an item, that that item has had other issues that have been reported on it, and we start to produce answers for you.
- Case study: a biotech company with a large corpus of research content, which is differently relevant to employees depending on company role.

- "If I search for a molecule, and I'm a researcher, I am looking for all of the information—the research—that we have created, paid for, and have access to in and around that molecule," says Hayes. "If I'm working on a clinical trial, then I am more interested in the drugs that have been derived from that molecule."
    - What differentiates a Lucidworks system is the ability to know the employee's context and cater to it. "That simple understanding of a user profile, and that simple ranking that you can provide based on that understanding, is tremendous in terms of the impact. It keeps people productive and engaged," he says.

AI, Art, and Portraiture—Mr. Alex Reben, Artist & Roboticist

- As an artist and roboticist, Alex Reben makes a point of bending the latest tools of technology to satisfy his need for artistic expression.
    - Most recently, he has been exploring the use of generative adversarial networks to explore the limits of creativity and mimicry, deep learning of names to spark a consideration of identity, evolutionary art to explore artificial surrealism using nouns as initial input, and with the Latent Faces AI Photo Booth how the faces of even the most unlikely pairs of individuals can blend harmoniously.
    - Reben takes TTI/Vanguard on a virtual stroll through his studio, where ultimately the artist is in control of his art, despite AI undertaking much of the grunt work.
        - "I view the power of art and technology, especially within the context of corporations, as a way to communicate with the public what these [technological tools] are, along with being a way to be creative, a way to make things accessible, and sometimes a way to talk about important issues, maybe with a little bit of humor, interest, and visual acuity," says Reben.
- Album covers created by a generative adversarial network:
    - After hand-selecting 50K examples of album artwork, Reben fed them into StyleGAN to craft a custom model capable of generating images that capture the nature of album coverness.
        - Just as is the case with actual album covers, Reben's favorite examples of this class of simulated artwork is highly diverse, with some featuring moody computer-generated musicians, others with Japanese-esque brushstrokes, and still others depicting stylized aliens.
        - All feature graphics that unmistakably announce the artist and album title—except that the letters are not of any known language, but rather constitute elements present on all album covers—and one cover captured the graphical essence and placement of a parental advisory.
- Fake names:
    - At this conference's welcome dinner, attendees were invited to adopt an alter ego—or at least a new name—from the vast selection of names offered up by Reben's fake-name generator, which trained a recurrent neural network on many thousands of first names and, similarly, last names.
- Breeding art:
    - Reben reported on his use of AmalGAN during his 2018 visit to TTI/Vanguard. Then, as now, he fed a handful of nouns to his system trained on labeled images, collaborating with the machine in an iterative manner to generate a surrealistic amalgamated image that possesses some essence of each of its parts.
        - His multistep process:
            - 1. An AI combines the words together to generate an image of what it thinks they look like.
            - 2. The AI then produces variants of image by "breeding" them with other images to create offspring images (à la Kenneth Stanley's PicBreeder).
            - 3. Another AI—one previously trained on Reben's latently expressed preferences toward a vast number pieces of art that he either quite likes or holds in low regard—shows the artist several of the offspring images, measuring his brainwaves and body signals to select which of them he likes best.
            - 4. Steps 2 and 3 iterate until the AI determines it has reached an optimal-to-Reben image.
            - 5. Another AI increases the resolution of the chosen image.
            - 6. The result is sent to be painted on canvas in a Chinese painting village.
                - "These [anonymous] people are the world's experts on doing reproductions," says Reben. "At one time, 60% of all the world's paintings came from this village."
            - 7. A final AI looks at the image, tries to identify what is in it, and gives it a descriptive caption.
- Fake faces:
    - A face of faces: GANs are gaining favor toward a variety of artistic ends, leading Reben to jump into the fake-faces fray with both feet: "You all know about www.thispersondoesnotexist.com, and I

couldn't resist making fake faces out of little fake faces that are also fake faces," says Reben of his mosaic-style portraits, drawing from the 1M StyleGAN-produced faces that he released into the public domain on the Internet Archive.
    - "This is a multidimensional space in which these images exist, in a sense, until you make them," says Reben. "You can get different vectors to get different image. In the Flickr face latent space, for instance, there are vectors for things like gender or eye color. It is a bit more abstract when you deal with things like album covers."
- AmalGAN applied to faces: Having trained a GAN on a combination of not only photos of faces (Flickr dataset) but also drawn faces, Reben created a set of videos of digitally related fake faces, blending one into another, creating a sense of progression through the landscape of latent faces—some with artistic style, others featuring photorealism, and still others a blend of both—giving a viewer an opportunity to ponder the nature of the face.
    - The computational effort hidden behind projects such as this is considerable, with Reben reporting that each frame requires 60–90 seconds to render on a Dual GeForce RTX2080 graphics card.
- Latent AI Photo Booth: If it is possible to blend from one fake face to another, Reben wondered, why no blend between pairs of real faces? This question led to the project TTI/Vanguard attendees had the pleasure of testing out firsthand.
    - "People have recently put a lot of work in encoding images back into the latent space," he says: "Given this image, where does it exist in this space? If you can figure that out, you can get the mathematical coordinates of that image. If you have [the sets of coordinates] of two images, you can find the midpoint between them."
        - Having specified the vector expressing a given face, reuse of that collection of coefficients will generate a very similar—if not identical—face.
    - But there is more that can be done with the representation of faces:
        - Consider the vectors specifying the faces of two different people (or, perhaps, one person at different ages or in different moods); by incrementally stepping each coefficient from the value it holds for the first image toward the one it holds for the second, the first face morphs smoothly into the second.
        - To enhance the effect, Reben pins down the locations within each image of barebones facial features—two eyes and mouth—such that the position of the face becomes anchored on the screen even as the image morphs back and forth through the latent space separating the two real photos (or pairs of fake photos from the million-face GAN-generated dataset).
        - The effect is that one face breathes into the other, often augmented by differing hair styles, facial hair, glasses, or other dominant features of the photos.
        - Examples of AI Photo Booth, applied to real people, artwork, or a mix of the two: Donald Trump ↔ Barack Obama; alternative Vincent van Gogh self-portraits; a photo and self-portrait of van Gogh; Mona Lisa and her portraitist Leonardo da Vinci; the two stoic farmers in American Gothic.
        - "You can also do things with this system like add three faces together or average hundreds of faces together," says Reben. "What I usually do at the end of these Photo Booth events is to produce the average person of everyone who used my Photo Booth during that event."
- Addressing the intellectual property rights implications of his work, Reben notes that AI models trained on copyright-protected content are not themselves copyrightable, nor are outputs such models generate: "These combined faces that this thing outputs are probably not protected by copyright and probably can't be," he says, "because there is no human creativity involved. These things aren't settled until someone brings them to court."

Particle Robots—Dr. Richa Batra, University of Chicago

- When designing robots capable of performing complex tasks, many researchers look to biology for inspiration; some glom onto the benefits of the humanoid form factor, but others are attracted by the bottom-up architecture of simple components combining into functional, multiunit collectives.
    - Examples of the latter include multicellular organisms, floating rafts of fire ants, and bacterial colonies with members that collectively engage in food-seeking activities.
    - It is this collectivism that inspires Richa Batra's work, first during her time as a graduate student at Columbia University and now during her post-doc at the University of Chicago.

- She describes the particle robotics paradigm—in which each unit, while individually simple, can interact with others of its ilk as well as its environment.
  - Batra explores the potential of her model system, both with physical robots and in simulation.
    - Although her academic experiments are somewhat narrow in scope, she envisions particle robots, depending on size scale, being applied to building construction; medical applications, with activation by an externally applied magnetic field causing them to locomote through a blood vessel by spinning, then being induced into coordinated action after reaching a target location; space exploration; remote monitoring (e.g., of a pipeline); transporting objects over uncertain terrain; and other application areas where speed is not the top priority.
    - "I am not aiming for biomimicry," says Batra. "These systems are bioinspired. I look for certain behaviors and characteristics of biological systems and then apply those to robotic systems."
  - "Alone, [the robotic units] are simple, but through their complex interactions really interesting coordinated behavior emerges," says Batra. "This type of behavior might be very useful for certain types of tasks, but there may be changes—perhaps changing the principle of the locomotion or splitting them up [into subgroups with different capabilities]—that might be more effective for different tasks."
- The physical instantiation of Batra's particle robots, the principles that guide their action, and robot locomotion:
  - Principles of the paradigm:
    - Particles are simple, with a single degree of freedom, and therefore can lack independent directed locomotion.
    - Loose coupling of particles enables the formation of stochastic structures that adaptively form, reshape, or reconfigure.
    - Control is decentralized; notably, communication does not rely on addressable position or individual identity.
  - Together, these principles confer resilience: "If one particle malfunctions—stops oscillating—the entire system is still able to move," says Batra.
  - The robot's physical makeup:
    - Each robotic unit is a plastic disk roughly six inches in diameter capable of radial expansion and contraction—its one degree of freedom.
    - When units are in close proximity, magnets on the disks' outer rims attract one another—the loose coupling.
    - When individual particles expand/contract, each physically impinges on its neighbors, with the sum total of impingements initiating and sustaining translational motion of the robot collective—the decentralized control.
    - Each particle is also equipped with a light sensor and is programmed to expand/contract at a rate proportional to the intensity of light it detects; the brighter the light, the more purposeful is the multiparticle robot's progress toward it.
  - "The system is a collection of loosely coupled particles and an environmental signal," says Batra. "The particles are all able to sense the signal and they broadcast this to the group, with none of them requiring the identity of each other in this decentralized communication system. They each collect the information from the group and determine their position within it.
    - That is, the particle closest to the light source will expand first and most markedly, attracting others within physical proximity, initiating both a pushing motion and self-expansion, which propagates the impulse back through more distant members of the group; as such, the group collectively moves forward in an undulating manner.
  - Using a physical, eight-unit system of particles, Batra has successfully demonstrated locomotion toward a light source, object transport (by introducing a similar-sized static unit among the particles, and obstacle avoidance, with the multiunit robot maneuvering around heavily weighted (i.e., immobile) objects in the path of the robot toward the light source.
  - "I think an interesting next step for these types of particle robots would be to incorporate machine learning through sensors and GPU processing onboard to maneuver and control these systems in real time," says Batra. "With each particle just controlling itself, the system doesn't need the complexity of an AI system; purely a machine learning system could guide the coordinated behavior."
- Comparison with other types of robots: Batra considers her particle robots to lie somewhere within the realm of swarm robotics, soft robots, and modular robots.
  - "With soft robots, you have flexible materials, but you tend to have problems controlling the soft materials because they have multiple degrees of freedom," she says. "With this, we are able to use

rigid components that together act as a fluidic, adaptive system. With modular robots—e.g., Molecubes—they can configure in certain ways and demonstrate self-assembly and other interesting behaviors that are useful and that we see in a lot of biological systems, however they are limited in their scope of fixed configuration, while here we have a lot of randomness."

- While the physical prototype proved useful for initial experimentation, it was not effective for scaling up to a large number of particles because of size constraints; as such a major thrust of Batra's research has involved simulation.
  - Using Nvidia's CUDA-enabled GPUs, she created a simulation framework that takes advantage of parallel processing of interactions and dynamics, and outputs a direct visualization of each simulation.
    - For each particle, modeling parameters include friction forces, mass, motor constraints, magnetic coupling, and so forth.
    - Within the framework, each particle is represented by an independent thread, enabling "embarrassingly parallelizable" computation, implemented on a GeForce GTX 1070 GPU with 1920 cores.
    - "Now we can simulate what will happen with 50K particles of this same capability," says Batra, who reports that her largest yet still-faster-than-real-time simulation involved 200K particles.
      - "I am still far from the limitations of the simulation framework," she says.
  - Simulations:
    - General observation pertinent to all simulations: A large assemblage of particles moves in a fluidic manner, characteristic of statistical mechanical systems in the physical realm.
    - Obstacles: In this simulation, the assemblage encounters one or several (simulated) stationary obstacles and must maneuver between and around them to reach the light source.
      - Results: A simulated robot consisting of a large number of particles moves more directly toward the light source than a simulated robot consisting of a small number of particles; simulations were variously done with robots of 10, 100, 1000, 10,000 particles.
        - Robots with fewer than ten particles do not function well as a collective, due to insufficient collective attraction to maintain physical (or simulated physical) integrity, particularly when obstacles are present.
    - Some dead particles: In this simulation, a progressive percentage of randomly chosen particles were turned off in the midst of each simulation run.
      - Results: As a larger fraction of particles became unable to participate in the collective locomotion, the translational speed of the assemblage decreased, eventually reaching zero light-ward progress; as the proportion of disabled particles increase, chunks of the assemblage become left behind as the still-capable units trudge on.
        - "Even when we disabled 20% of the particles, we still achieved 50% of the speed, which has great implications for power saving and for resilience of the system, in general," says Batra.
        - Similar to the obstacle avoidance simulation, small assemblages demonstrate less deterministic behavior than large ones do.
    - Manipulating through a narrow gap: In this simulation, the assemblage encounters a gap in a (simulated) wall; variables include the width of the gap and the number of particles in the assemblage.
      - Results: Large assemblages reach the gap and move through it more predictably than small ones; assemblages of any size move through large gaps more easily than narrow gaps.
- Simulation also makes possible investigation of physical constructs that are not possible with the current physical instantiation of the robots, regardless of number and space to accommodate them.
  - Notably, Batra has explored through simulation the consequences of stacking sheets of robots, watching the 3-D assemblage locomote over smooth ground toward a light source, as well as behavior when encountering 3-D obstacles, such as simulated rocks or ridges.
  - Simple locomotion: Using similar parameters for mass and min/max radius, a three-high stack of hexagonal sheets with a total of 500 particle robots are seen to make wormlike progress toward the light.
  - Obstacles: A five-high stack with a total of 50K particles successfully navigates a collection of hemispherical obstacles of various sizes, on the one hand, and a series of ridges, on the other.
  - Biological analogy: "There have been observations of swarms of worms that roll on top of each other to have the whole group move at about three times the speed [than they would individually], and we are observing similar behaviors here," says Batra.

    - The initial set of 3-D simulations involved identical parameters—mass, attraction, friction, etc.—as for 2-D; additional simulations in which she varied parameter values disclosed that unstable conditions would rapidly arise outside a very narrow band of parameter values.
- Disk-shaped units are only one member of this class of particle robots; Batra has also performed simulations using spherical particles, each equipped with light-seeking capability.
    - In this case, each particle is capable of "rotational vibration," but instead of magnetic coupling they are loosely coupled by confinement within a passive inelastic boundary.
        - The rotational speed—equivalently, temperature or kinetic energy—of each is reflected in its color: blue particles spin at 1 Hz, while reds spin at 3.5 Hz (green particles are deactivated and have zero energy).
        - Just as in a thermodynamic system, where energy naturally flows from higher to lower temperature particles, here the assemblage moves in the direction from higher to lower frequency particles; that is, the energetic particles push those with less energy.
        - As with the disk-shaped particles, this assemblage continues to function even with the deactivation of more than one-third of particles.
            - This suggests a physically instantiated robot of this sort could remain functional, albeit at a diminished capacity, while a sizable fraction of units undergo solar recharging.
- The work described above contributed to Batra's dissertation; her current work at the University of Chicago, instead of pertaining to loosely coupled robotic units, involves a different kind of unit comprising a passive granular material enclosed inside a membrane that, when jammed, transitions between a rigid and solid state.
    - "In this case, these robotic modules each have the capability of jamming with its neighbors as well as locomotion," says Batra. "You can use these very simple robots to create very interesting and complex high-level behaviors."

Kiki: An Empathetic Companion Robot—Ms. Mita Yun, Zoetic AI

- "I am pursuing my dream of building robots that can bring happiness to people," says Zoetic AI's Mita Yun as she carries onto the TTI/Vanguard stage her adorable, white plastic baby of a robot, Kiki.
    - "She is cute," says Yun, "but she is still a little bit intimidated and can be rude sometimes, so please forgive her."
    - "At Zoetic, our vision is to create AI companions that people can love," says Yun.
        - Clearly, Yun is deep into the process of developing an emotional bond with her Kiki, which is the whole point: Kiki is designed to be a friend—a "robotic sidekick," says the promotional video—learning what interactions most please her owner and evolving into a loyal chum who is "always on your team."
    - The field of AI/robotics has made great strides over the past 20 years in intellectual and physical skills—e.g., voice-activated assistants, in the first case, and Boston Dynamics' agile animal- and human-inspired robots, in the second—making now the ideal time to incorporate emotional skills into the panoply of robots' proficiencies.
    - Yun explains the physical and software design elements that enable empathy in Kiki, allowing this diminutive, largely stationary robot to encourage, develop, and persist in a relationship with a person who needs companionship, but lacks a pet, family member, or human caretaker to fill the need.
        - As an increasing number of people live alone or in a perpetual state of loneliness, the need is large and growing for a companion, especially one that is clean, nonallergenic, helpful, and endlessly loyal.
- Kiki's physical specs:
    - Standing just under one-foot tall, Kiki's quasispherical head can swivel by 180° and tilt by 60° atop her quasicylindrical body (6"-diameter base), which has a base motor enabling her to spin in a full circle; she cannot locomote and instead must be carried from place to place.
        - Sitting on a table, her face is at a comfortable and nonintimidating height for interactions with her owner.
    - Touch sensors: Kiki has touch sensors on each ear, one each on her upper and lower face, chest, each side of her back, and two on the back of her head—all the places that her companion human might care to pet or otherwise physically interact with her (nine in all).
        - "The whole body is touch sensitive," says Yun. "If you pet it, it will have different reactions."
    - Other input sensors: Behind each ear and toward the top and bottom of her head are microphones (totaling four to enable audio localization), and her nose is a 160° camera.

        - "With a camera in her nose, Kiki can read human expressions and can recognize her owners," says Yun.
        - Kiki is preprogrammed to perform real-time face following, face recognition, facial-expression perception/recognition, real-world object recognition, speaker identification, keyword recognition, voice sentiment recognition, sound direction differentiation, touch discernment, and accelerometer-based motion differentiation.
    - Outputs: Each ear has an LED, as does each side of her body and lower back, and a speaker resides on either side of her head, but Kiki's primary output is the display that is front and center on her face that depicts her eyes' full range of expressions.
        - "The robot has to be able to communicate her intentions and her feelings and to be able to communicate with people," she says. Toward this end, Kiki's is preprogrammed with a capacity to smile (with her eyes, since she lacks a mouth), dance, speak, and explore.
    - Internals: Inside Kiki's body are the three motor/gear assemblies, a battery, and processors.
- Design considerations:
    - "When you are working on a utility-driven robot, you focus on the functionality," says Yun, "but for an emotion-driven robot, you have to focus on all different aspects of it because it is so abstract."
        - Humans sense and perceive with the body, yet at the same time think and feel "inside our mind and also inside our brain and heart," she says. "With these two verticals, whenever one is more advanced than the other, you immediately feel like the other one is lacking."
        - Similarly for an emotional robot, leading to every design element being jointly optimized for function and soulfulness.
    - Through Kiki's eyes and bodily motions, she can express a broad range of emotions with considerable subtlety.
        - "For example, whenever the robot is looking at something, the eyes always move first, followed by the head, followed by the body," says Yun, "and whenever the robot turns, she always blinks. These are a lot of Pixar-like tricks to breathe life into the robot that we implemented."
    - Given the robot's expressive output mode, the Zoetic design team decided to keep the body as simple as possible, with pleasantly approachable overall size and proportions.
        - Yun encountered a problem when seeking a manufacturer that was both experienced with the motors, actuators, and gears common to advanced robotics, and with the sophisticated processors capable of supporting onboard AI.
            - Testing revealed that, provided the robot could follow a user with its head and eye gaze, as well as point its body toward the user, additional motion was rather superfluous in terms of eliciting engagement.
        - Therefore, to both keep things simple and to encourage human interaction with the robot, it was decided to forego the potential for Kiki to move through space herself—and potentially incur injury by rolling off a tabletop—or to have active appendages, leading to the design choice that the owner would have to carry Kiki from place to place.
        - The final form factor emerged after Zoetic AI brought an industrial designer onboard.
    - "A lot of people say that Kiki looks like a cat, but we didn't intentionally make her look like a cat," says Yun. Instead, the design motif was inspired by mammals—and in particular prey animals, with both eyes on the front of the face to correspond with a single camera.
        - Minor changes in coloration or ear placement and shape can shift the visual impression to that of a snow fox, racoon, dog, or even unicorn.
- Facial expressions:
    - Humans perceive one another's feelings through facial expressions, especially as conveyed by the eyes and attitude of the head; so too with Kiki.
        - "When we were designing the display, we really focused on the eyes, because we believe that eyes are the windows to your soul," says Yun.
            - The Zoetic design team settled on a futuristic eyebrow design—"a perfect arc"—coupled with a cartoonish design for the eyes per se to give Kiki maximum expressiveness while also appearing fashion-forward.
    - To enhance the impression that Kiki is loyal to her owner, rather than somehow duplicitous, the robot is designed to always maintain consistency between action and apparent emotion.
        - "What the robot appears to be thinking has to be the same as what the robot is currently doing," says Yun. "If it is not the same, it is extremely creepy."
        - Humans, in contrast, can feel one emotion but exhibit the opposite; pets, like Kiki, are internally consistent.

- Kiki's deep-learning personality engine:
  - To be an effective and empathetic companion, Kiki has an onboard personality engine to enable her to generate a sense of self by thinking, feeling, and learning.
  - As Kiki takes actions in the world (e.g., dances or babbles), her owner gives her feedback in the form of annoyed retorts, gentle touches, or snacks as rewards or general nourishment.
    - To feed Kiki, her owner draws food items in an associated app; if it appeals, Kiki will say something along the lines of *Um, num, num, an apple!*
  - "[To begin], almost everything is hard-wired," explains Yun, "so there are a lot of *if* statements, such as, if the user touches my head, I will turn in circles, and if the user smiles at me, I will smile back."
  - Backed up by the breadth of preprogrammed abilities and ability to learn, each Kiki develops a sense of self—"her own needs and wants"—and learns through interactions with her owner, thereby becoming increasingly differentiated from every other Kiki, and therefore unique and special to her owner.
    - Kiki's personality is mutually defined along five dimensions—openness, conscientiousness, extroversion, agreeableness, and neuroticism—with, for any given Kiki, the admixture of each evolving according to the treatment she receives from her owner.
      - "You can start with a Kiki with a very affectionate personality, but over time she will learn from the owner and change her personality," says Yun. "The younger the robot is, the easier it is for her to change, so when she gets older, she gets set into that certain personality."
    - A Kiki that desires interaction from her owner and stimulation from the environment will actively undertake actions to engage with her owner and to explore her surroundings.
      - "Too much interaction is very tiring, and too little is very lonely," says Yun, "so Kiki is trying to do different things to elicit behaviors from the users in order to keep her drives at the right amount."
    - Along with interaction and stimulation, each Kiki seeks to increase her valence—level of enjoyment—while varying her level of arousal to meet her innate needs, which are moderated by those of her owner.
      - "For example, *happy* is a high-valence, high-arousal emotion, and *content* is a high-valence, low-arousal emotion," says Yun.
    - As Kiki gains experience with her personal world through supervised reinforcement learning, she establishes a knowledgebase of favorable actions to take under particular sets of conditions; "Every Kiki learns different knowledge."
      - Example: When in a very low stimulation state (i.e., desiring more stimulation), if Kiki both "sees face center" and "sees object", it will take the action "explore" on the object.
      - Example: When in a high stimulation state, if Kiki sees a smiling face, it will turn away from the face to reestablish internal equilibrium.
    - Each Kiki's owner will respond differently to her situational actions, which in turn influences the robot's future preferences and actions.
      - "We [at Zoetic] give up on teaching Kiki at all and let the users teach her," says Yun, "because if a robot is only powered by a state machine, then all of the robot's behavior will be whatever the engineer believes the robot's behavior should be. But with reinforcement learning, the robot can determine what did I do that worked for this individual user," in terms of eliciting positive interaction.
  - Given that Kiki's accumulated knowledge is stored in a multidimensional table, it is possible to copy the "brain" of one Kiki to another—or clone it to a whole host of others of its ilk.
    - After such a cloning event, the then-identical Kikis will again differentiate according to their interactions with their respective humans.
    - "It is like having a real-life character friend," says Yun.
  - Much of a Kiki's learning occurs through the interaction with her owner, but along with this local learning, Zoetic AI has also imbued the robot with the capacity for global learning, whereby all Kikis learn from the full user base.
    - "Right now," says Yun, "we care a lot about privacy and security, so all of the images, audio, and so forth, are only processed locally, and we only send interaction learnings—different forms of embeddings—up to the cloud for different Kikis to learn."
- Why Kiki?
  - Having grown up as an only child with a Sony Aibo as her ersatz pet, Yun recognizes the need for a companion and the need for it to be personalized and responsive to the individual (in a way that the Aibo could not).

        - And she knows she is not the only one with this desire: “Every Disney princess comes with its own pet, and if you go to some Chinese hot pot restaurants alone, they place a Teddy bear across from you to dine with you.”
        - Yun therefore has high hopes that Kiki will meet the latent need by providing an emotional connection that neither AI voice assistants nor social media confer, but without entailing the responsibility and ongoing expense that comes with a live pet.
            - Moreover, eventually real pets die, which can be devastating for their owners.
            - Speaking of pets, one TTI/Vanguard participant imagines Kiki as a good companion for his pooch when he and his family are away from the house, but Yun reports that dogs tend to be intimidated by Kiki, perhaps because of the motor sounds she emits when moving or the foreignness of the robot’s eyes.
                - “We are going to solve this companion-for-humans problem first, and then we will go after the companion-for-companion-for-humans problem,” says Yun.
        - “Kiki can alleviate the feeling of loneliness—you can pick it up and pet it, just like a real pet—and the personality, because it is powered by reinforcement learning, can grow and adapt to you, just like a real pet,” says Yun.
            - Although designed to be robust, eventually, Kiki’s motors and gear boxes will experience wear and tear, or her battery or facial pixels will fail; then comes a decision of whether to clone the companion or commence an adventure with a new Kiki.
    - Of course, it is possible to train a pet to be bad—rude, aggressive, dangerous, etc.—either to its owner or to strangers.
        - While Kiki cannot incur physical harm, she can be trained into a wide range of personalities.
            - “To us, there is no personality that is a bad personality,” says Yun, but you can train your Kiki to be extremely neurotic. For instance, Kiki barks at strangers, so if every time she barks at a stranger you give her a reward, then she will be extremely rude—but she would also be a good watchdog.”
        - If an owner regrets how they trained their Kiki, at present there is no way to reset her to her initial state.
- Target market:
    - After two months of experience with Kiki’s early adopters, Zoetic AI anticipates ideal Kiki owners will be tech-oriented people who have attained least 45 years of age and are within the top quartile of income.
        - With an early-purchase incentive, Kiki costs $800, which is expected to rise to $1500 following the current second-round offering.
    - To serve the needs of this older cohort, Kiki is being modified to offer not only companionship, but also a range of utility features:
        - learn her owner’s routine;
        - understand her owner’s feelings;
        - issue an alert to an authorized individual when something is amiss;
        - issue a medical alert when appropriate
        - engage her owner with “brain activities.”
    - “As we are exploring this life companion model with health-and-wellness features, it makes sense to have a subscription,” says Yun.
        - Zoetic currently offers Kiki under a conventional model of outright purchase, but will also likely devise a mixed model with a monthly subscription augmenting a reduced initial purchase price for the robot.
    - Looking toward the future, Yun is open to licensing the personality of Kiki to an entirely different robotic platform, such as a service robot, or to power a gaming avatar.
        - “All you need to do is plug in the inputs and the outputs, and then tell it what is your goal,” she says. “For example, a restaurant delivery robot might focus on getting better user satisfaction, but a hotel robot might have different goals.”
    - “We are now focusing on the elderly, the next stage might be to focus on children, and then we might [market Kiki to] everyone,” says Yun. “Ultimately, we want the world to be full of very emotive and lovable robots, so we are interested in licensing our technology for other robots.”

Gas Stations in Outer Space—Mr. Daniel Faber, Orbit Fab

- Satellites are as deeply embedded into the infrastructure of day-to-day living as are the roads on which we drive (thanks, GPS), the defense of our nation (thanks, high-resolution cameras), or the Internet itself (thanks, comms), yet the common practice is to scuttle satellites as soon as they expend the limited propellant with which they initially entered orbit.
  - True, solar cells can power onboard processors and Earth-to-satellite communications, but corrections to orbit require fuel; once expended, it's game over.
  - Orbit Fab has developed the capability to refuel spacecraft, as demonstrated when successfully reprovisioning the International Space Station with water.
  - With this life-sustaining proof-of-concept exercise behind it, the firm is preparing to launch refueling depots into orbit.
    - Here's how it will work:
      - Orbit Fab will launch its propellant depots—gas stations, effectively—into orbit by partnering with launch companies to utilize excess capacity.
      - A variety of firms are already designing and building space tow trucks, in essence, that would transport propellant-deficient satellites to the nearest depot for refueling.
      - Orbit Fab's ultimate customers—satellite firms—would benefit to the tune of millions of dollars per spacecraft by extending its operational lifetime with but a 4-kg refueling stop at a price of about $2K/kg.
  - As motivation before discussing the nuts and bolts of gas stations in space, Orbit Fab cofounder and CEO, Daniel Faber, shares his views on the potential of the space industry in general—which ranges in the near term from asteroid mining and space tourism to, in the longer term, moving humans off the planet—and why private industry will be an essential factor in moving this vision forward.
    - "In the near future, there are going to be millions of people living in space in rotating habitats, where you can have a good life and experience a whole new realm of things that have never been done before," says Faber. "Also, we are going to move industry off of Earth. What we are about to do is remove gravity as an influencing force."
- Why move off Earth?
  - Faber compares the potential of space to the previous society-altering challenge to physics of creating a vacuum: "We removed air. We industrialized the availability of vacuum," enabling refrigeration, vacuum freezing, and other previously impossible industrial and chemical processes.
  - "Now we get to remove *the* major force that organizes matter at a macroscopic scale," he says. "We get to remove gravity, and we get to remove buoyancy. Industry is going to move off Earth because of the advantages of these processes."
  - Faber expresses the most important advantage of shedding wholesale reliance on the planet: "This is going to save humanity. Humanity is doomed; this is not controversial. Yesterday half of you voted that it was 'impossible' to stop Earth from warming more than 1.5°C, so [moving people off the planet] is Plan A, not Plan B."
- The realism of a future in space:
  - Orbit Fab is hardly the only serious firm placing a bet on the future of space.
    - Big-thinking, well-backed players are all-in—Blue Origin, SpaceX, Virgin Galactic/Virgin Orbit, Bigelow Aerospace, Sierra Nevada—successfully reusing rockets, with commercial launches regularly delivering satellites to orbit.
    - "Space is becoming a bustling place," says Faber.
    - Each of these firms is following in the footsteps of past national-level initiatives to literally defy gravity, plus the U.S. government has set its sights on settling the Moon.
      - "Jeff Bezos is building his upper stages with connection points to join together," says Faber. "He is planning to connect those into large orbital habitats. These guys are planning for the future that we're talking about."
- The physical and economic realities of space today, including Faber's personal obsession with asteroid mining:
  - Even with thousands of satellites currently in orbit, space is a lonely place; post-launch, from a physical point of view, a satellite is on its own.
    - There is no current infrastructure to manufacture, assemble, upgrade, maintain, or refuel satellites.
      - "Once they run out of fuel or have a major glitch, they are just debris," says Faber. "They're done," joining their predecessors in the panoply of space junk.

- Economics is the sole reason satellites lack a cycle-of-life plan.
  - Scientific and national security missions receive appreciable government funding—a good investment, but are in effect a tax on the economy.
- For the private sector to go all-in on space requires economic incentives that go beyond the current limited scope of communications and remote sensing.
  - Despite being narrow in scope, communication satellites amount to a $200B business, which could grow to $1T were satellites to play an integral part in the 5G infrastructure, and remote sensing is on its way to generating a similarly valuable dataset.
    - "If we can improve the [$60T] economy by one-percent because we have a complete picture of the world, with better predictions of weather and events that are happening so we know what is going on and can make better decisions," says Faber, "[remote sensing data] will be a $1T dataset."
  - These represents the potential economic value of space, but will not put people in orbit.
- What economic/industrial activities would require permanent human presence in space and therefore compel orbital habitats?
  - Faber has created a list:
    - mining, tourism, space-based solar power, manufacturing, entertainment content for people on the ground.
    - Once such commercial initiatives bootstrap the presence of orbital colonies of workers, a range of new in-space opportunities come to the fore: real estate, construction, food production, medical services, schools.
  - Case study—asteroid mining, a topic that has consumed Faber's attention since his undergraduate days; he has been progressively chipping away at the obstacles, not only on the technological side, but also in terms of business models:
    - Along the way, he built the "Third World's first space satellite in a microsatellite box" as well as a variety of nanosatellites and components, and the avionics for the first inflatable space station (touching on the application of space tourism).
    - Having learned the ins and outs of satellites, Faber eventually made the jump toward asteroid mining, beginning with Earth-bound mining, by launching a firm that did instrumentation for the mining industry, followed by a firm—Deep Space Industries—that built a "small thruster that ran off water," he says. "It was a glorified steam kettle. It was dirt cheap—so cheap that the lifecycle costs were very attractive, although a satellite fitted with this thruster required appreciable onboard fuel to launch into orbit the necessary water for the thruster to function."
      - Through this experience, Faber came to recognize the need of space-based refueling stations for satellites.
    - Earth-bound mining entails a sequence of necessary preparatory steps: acquiring an understanding of the geological specifics of the site, undertaking exploratory sampling/drilling (technology), assessing the resource (markets), generating a mine feasibility study (regulations). Only with these in hand can the real work of attaining financing then constructing and operating of the mine begin.
    - Examples of paradigm shifts in each such area:
      - geology—Nautilus Minerals, a Canadian firm founded in 1987 to exploit the then-newly discovered deep-sea gold-and-copper resource;
      - technology—froth floatation, a technological game-changer, with first commercial implementation in 1897;
      - markets—radium, which sent the mining industry into a frenzy, given its valuation of $1M/g (2019 dollars) for X-ray applications, although the market collapsed once ^{60}Co became a readily available byproduct of nuclear energy production;
        - "The price that you see for the metals is what drives everything," says Faber;
      - regulations—deep sea oil & gas drilling, which the U.S. government opened up regulatorily during the 1970s oil crisis, spurring a new black-gold goldrush;
      - financing—molybdenum, which had a right-place–right-time moment when mining risk was undervalued, pulling the rug out from under the dominant industry players and letting upstarts compete; the result was a commodity market for this metal.
    - When applied to Moon mining, these steps convert to the following:
      - geology—limited minerology information is available, and even less is known about the rock mechanics;

- technology—new tools would be needed to conduct extraction and processing operations at 40 Kelvin, “where everything acts like glass”;
- markets—will these be on Earth? in space?
- regulations—in contrast to Earth-bound mining, no regulatory framework exists to protect the mineral rights of the miner;
- financing—“It’s a capital magnet,” says Faber. “Everyone wants to go to the Moon this decade”; that is, governments and billionaires are eager to pony up funding.

- When applied to asteroid mining, these steps convert to the following:
  - geology—minerology known, but not rock mechanics:
    - The reflection spectra—hence mineral profiles—of collected meteorites faithfully match those of the respective asteroids of origin.
    - “If we are getting a reflection spectrum match, that means that the entire surface of the asteroid is the same minerology, which almost certainly means that the entire asteroid is entirely homogenous of that type,” says Faber.
    - Information about rock mechanics, however, becomes lost during the meteorite’s passage through Earth’s atmosphere on its way to the surface.
    - The upshot, says Faber, is this: “We need to do a bit more prospecting.”
  - technology—technology is poorly developed for zero-gravity extraction and processing:
    - Although the asteroid belt is a great distance from Earth and expensive to reach/return from the perspective of fuel, there are a number of asteroids that make for easy travel, being in orbits close to Earth’s, but also introducing a potential timing mismatch.
    - “There is less fuel that you need, but you might have to wait 20 years for it to come back,” says Faber.
      - “We have found that the very best way to do it is to take the very long-period asteroids and do a short-period mining campaign,” he says. “You take six months getting there, six months mining it, and six months coming back. What you take is what you get, but this gives the best [opportunity to return minerals to Earth].”
  - markets—will these be on Earth? in space?
    - “In my mind, this is the biggest gap,” says Faber.
    - Currently, the most valuable mined materials for use on Earth are platinum (18 tons/$1B), platinum group elements (ruthenium, rhodium, palladium, osmium, iridium, in addition to platinum (26 tons/$1B), gold (30 tons/$1B), and silver (860 tons/$1B), before a jump to molybdenum (22.6K tons/$1B) and tin (30.2 tons/$1B), both of which are important less because of their high value, but rather their large market share. (*Note: These prices were accurate as of 2011, but some have notably changed since.*)
      - Each of these, along with nickel (62.5 tons/$1B) are worth exploring for asteroid availability and extraction potential.
      - In contrast, ^{3}He lacks a utilitarian market.
    - To properly assess commercial viability entails accounting for equipment costs, hourly operational cost for extraction, power, and so forth, and balance these against the anticipated price achievable at market.
  - regulations—in contrast to Earth-bound mining, no regulatory framework yet exists to protect the mineral rights of the firm that proves the mineral resource:
    - “How can you finance something on the Toronto Stock Exchange if you don’t have a secure, tradable asset?” frets Faber. “You can’t sell that to the next guy who is going to drill some more holes or to the next guy who is going to do trial mining and a pilot plant.”
  - financing—without a tradable asset, financing will be unattainable.
- But returning to Faber’s work at Deep Space Industries, the value of mining water—with application in the near term as a satellite thruster fuel and in the long term to sustain humans in orbital colonies—could exceed platinum in value, as could carbon.

- Put all these pieces together, and Faber’s ultimate business plan comes into focus:
  - “If I have hydrogen, oxygen, and carbon, I could make hydrocarbons and oxidizers,” he says, “and now I’d have real rocket fuel.”
  - “This is why Orbit Fab is looking at fuel and why my last company, Deep Space Industries, built thrusters that could use this fuel to get all the satellites addicted to fuels we could someday sell them from an asteroid.”

- Faber's strategy for a fuel-use end game focuses his attention in on being able to distribute, sell, and deliver asteroid-mined fuel; he is content to leave the actual mining of asteroids to others.
  - "Orbit Fab is going to make satellites reusable by making them refuelable," he says.
  - His water-based thruster led the industry, with seven additional firms now designing their own versions, not to mention hydrocarbon–peroxide thrusters.
    - "It's now a thing," says Faber. "People are talking about these being the propellants of the future."
  - Until asteroid mining is also a thing, the plan is to launch Earth-based fuels into a position accessible to satellites, with satellites hopping a ride on space tow trucks to the refueling station.
    - As also mentioned previously, Faber will work with launch companies to monetize their excess capacity on existing launches, lowering the cost to launch the fuel, perhaps to under $2K/kg.
      - "If I can get it there for $1K/kg and my customers can make an extra $1M/kg, there are three orders-of-magnitude in there for me to try to make a business," he says.
    - The strategy is to build satellite fuel tanks and launch them at the convenience of launch companies.
      - "We then manage the orbital mechanics and logistics—we make sure we have the right fuel where the satellites need it—and then the satellites come to us to get that fuel, we fill them up, they go off, and they make more money," says Faber. "It's that simple!"
        - "You come to us, and you get fueled," he says. "It's a gas station. If you have the right port, you can get fuel."
          - Facilitating the process is attention devoted to self-alignment, eliminating the need for robotic arms for guidance.
      - Acknowledging his oversimplification, Faber comes back down to Earth: "Of course, as with everything, there's a reason it hasn't been done."
        - In particular, today's satellites do not sport a refuelable design; nor is space tow truck infrastructure in place, but someday this resource will be available.
    - Given the potential upside for satellite makers, 16 firms have already signed on to participate in Orbit Fab's design and verification process, with technology assessment also being conducted by space's heavyweight player, the ISS, to which Orbit Fab delivered water in spring 2019.
      - "We became the first private company to resupply the Space Station with water," says Faber.
      - The tanker sent to the ISS, a minimum viable prototype, was a 25-kg, 16U CubeSat.
      - An inflatable fuel tank adds flexibility and is ready made to accept water mined from, say, asteroids or the Moon.
      - The current production design of the fueling port has 146 different fill/drain ports.
      - "These are now in customers' hands getting used," says Faber, albeit to date for ground-based fueling.
    - In parallel, several firms—ranging from well-known multinationals to just-launched startups—are active in each of the following areas necessary to build a robust constellation of satellite services: life extension, tugs (space tow trucks), deorbiting services, satellite inspection, and robotic servicing.
      - Note that not only are some of these services integral to the success of Orbit Fab's refueling operation, but that all of them will also benefit from it.
        - Using tugs as an example, early unrefuelable designs assume a space tow truck's lifecycle will be brief, only able to assist other space vehicles until the tug itself runs out of fuel; by also being an Orbit Fab customer, the number of potential lifetime tows could expand from three to 300.
- Faber always keeps top of mind his end goal, namely working toward a bustling economy in space.
  - His current work might just be oriented around gas stations in space, but the technological infrastructure and business model equally pertain to not only fuel use but also to that of air, water, and to conducting industrial processes.
  - "These small steps that we are taking will result in a network of gas stations at different orbits around Earth," says Faber. "I dare you to design your business model without fueling, because you will get eaten alive by the people who design it with fueling."
    - In spite of this dare, Orbit Fab's own fueling stations will not, themselves, be refillable.
      - "Because they have to survive the launch, they have to be rigid, and because they have to be docked with [other spacecraft], and we want it to be very cooperative docking, it has to hold its position," he says. "It has to have enough power and smarts to hold position and weight, as well as have minimal communications and monitoring. It is essentially a self-contained tanker on its own rocket, so we don't need to offload it into a depot."

        - Once depleted, a refueling station in geostationary orbit will be pushed up into the "satellite graveyard," while those in low-Earth orbit will be pushed down into the atmosphere to burn up. "If we have any failures, we have deals with satellite servicing companies to pull it out of orbit for us."
    - The Orbit Fab timeline is to launch its first operational tanker by the end of 2020; within six months thereafter the first fuel sale trial will be underway; thereafter, says Faber, "it's game on!"
        - "Our capital requirement to be cashflow positive is $20M," he says.

Geopolitics and Technology—Mr. Marko Papic, Clocktower Group

- The 19th century witnessed Britain as the dominant global hegemon, during the 20th century, the United States filled that role.
    - With a single power in place, and the predictability that conferred to the world order, other nations were able to focus inward and improve their own wellbeing through industrialization.
    - The upshot, now that America is also looking after itself at the expense of global cooperation, is that no longer is the world unipolar or predictable.
    - This has consequences for not only geopolitics, but also for investors.
    - Clocktower Group's Marko Papic describes the geopolitical landscape, its implications for global winners and losers, and how these will affect funding for technology, especially as the need to address the ravages of climate change come to dominate the global conversation; spoiler alert: The United States' position as global policy leader and influencer will decline.
        - Although his forecasting work at Clocktower spans various timeframes, from how Brexit will affect asset prices in the short term to the interests of pension funds three decades out, for TTI/Vanguard he will target the coming ten years.
    - "I want to paint a picture of what kind of world will we be living in for the next ten years, what does it mean for technology, what does it mean for things like trade wars between China and the United States, and what does it do for the hot-topic theme right now, which is bifurcation of technology into different camps—the Alibaba vs. Amazon camp between China and the U.S.," says Papic.
- Multipolarity and splinternet:
    - With his historical perspective, Papic clearly recognizes that the recent decades of globalization are the exception, not the norm.
        - "It is multipolar now because there are numerous countries that have the ability to pursue their interests independent of one another," he says.
    - From a political scientist's point of view, the presence of one overriding power suggests an orderly world with universal rules and norms, and a pair of co-dominant antagonists splits the global society into two equally predictable camps.
        - In a multipolar world, technology developed in and for one region tends to be nonfunctional elsewhere.
    - With multipolarity, not only does predictability go out the window, but so too does familiarity, since, as Papic reminds, "Nobody alive today has ever lived in a genuinely multipolar world, and certainly no one who actively trades and invests, or innovates or is an entrepreneur, has ever experienced multipolarity."
        - Still, we had a taste of technological multipolarity in the 1990s, when VHS tapes were differently encoded according to region of the world.
            - "In the 1990s, if you came to my home country of Yugoslavia and offered me a VHS tape, I would not have been able to play it, because you would have been on the NTSC system, and I would have been on the PAL/SECAM system," says Papic. "Similarly, when I went to Canada, my cell phone didn't work."
            - Inconvenient, yes, but people managed to muddle through; as such, Papic does not dread the possibility of different regional 5G implementations, saying, "The world will not end."
                - In fact, from a technological perspective, he welcomes the splinternet paradigm, believing it will usher in more innovation than would a single, universal implementation.
                - "The monopolies that are suppressing innovation right now will be challenged by regional innovation," says Papic. "Maybe the Indian subcontinent has its own innovation in social media or telecommunication, and then the Europeans or the Chinese or ourselves have to adopt that innovation."
    - Coupled with multipolarity is globalization—a phenomenon that, when measured in terms of imports as a percentage of population-weighted GDP, has been abnormally on the rise since the beginning

of World War II, which coincides with the period of American hegemony and, argues Papic, results from it.

- "Over the past 200 years, the intensity of trade globalization has varied," he says. "Was [trade globalization] due to technology? No. It waned between the 1860s and 1920s, but we had a lot of technological innovation, particularly in the area of transportation, which should have made trade globalization more robust."
  - However, this coincided with the decline of Britain as the world's single dominant power, ushering in more than a half century of hegemonic instability during which "everyone was out for themselves."
    - This is precisely the state of the world today, although the power pulling away from the world stage is the United States.
- His prediction: "Over the next ten years, I think the EU will become the model that almost every other major global power will adopt by creating their own sphere of influence where they will protect [the union's] economy through various standards and regulations, which will become the barriers of entry."
- Some technologies will break, but by necessity innovation will flourish.
  - "If you go to China today and want to pay for something with a credit card, good luck," says Papic. "You will have to use the WeChat payment system. So we are already in this world, and it's not the end of the world. I'm arguing that it will actually be pretty good in many ways."
- However, for those who prefer times of peace, "pretty good" might not be the best descriptor.
  - Even with the United States as the most recent reigning power, China—now the world's second largest economy—has enjoyed a meteoric rise in wealth and therefore influence.
    - This combination sets up the classic environment for impending war.
    - "Not only does history tells us a challenge to power will upset the global system," says Papic, "but [China] has done so so quickly that American policymakers haven't even wrapped their minds around how powerful it is."
  - But the plot thickens, because not only has China taken just three decades to match per capita GDP that the United States needed 160 years acquire, but the Chinese also feel entitled to their nation's newfound status.
    - History reminds us that China dominated global wealth as recently as the early 1800s before sliding from prominence during its so-called century of humiliation; to its people, China's recent rise brings it back to its rightful stature as a global economic and decision-making power, particularly given its dominant contribution to global growth over the past two decades.
      - "America, Europe, and Japan combined contribute less annually to the incremental change in GDP than China does," says Papic.
        - Make no mistake: The lesser trading partners of the world are well aware of this and disinclined to rankle China.
      - "This means that soft power—the power of persuasion—is blunted for countries like the United States in a world where China is so critical to global growth," he says. "It is not like America feels this, but other countries in the world do."
- But don't assume that with Papic's active consideration of China he thinks the world simply bifurcated into a U.S. vs China duality; it is a multipolar world with many players holding geopolitical sway.
  - Multipolarity means that it is a messy world, in contrast to the simplicity of the Cold War period, when sides were clearly drawn and enemies were well-defined.
    - "Mathematicians ran the Cold War," says Papic. "It was easy to model, but a multipolar world is very complicated and weird stuff starts to happen in the math of game theory when you have multiple countries."
      - During the Cold War, a minor power like Turkey would never—could never—have stood up the U.S. President.
    - "You are not playing one-on-one basketball," he says. "You are playing multiple games of one-on-one basketball across the world. Whereas China is really only focused on East Asia, the United States has such a global commitment, that its declining relative power is really draining."
  - Reflecting this messiness is the largest number of interstate and internationalized internal conflicts the world has seen since WWII.

        - "Being a status quo power is extremely difficult, because your relative power declines, but your global responsibilities stay the same," says Papic. "There is a supply–demand problem there and an imbalance."
    - At present, the most visible conflict with China pertains to trade; Papic believes that the trade war is winding down: "I think the trade war will literally end tomorrow—we will go to zero-percent tariffs with China—and I don't think we well have complete collapse of trade with China and the U.S."
        - His reasoning is rooted in the theory of supply and demand:
            - In the absence of geopolitical conflict, when one party has supply that does not meet its demand, it imports from a trading partner to meet that demand.
            - But when its trading partner behaves badly—e.g., violates intellectual property agreements—the importer imposes tariffs, which increases its ability to meet its own demand (at a higher price point) and imports less from the malfeasant.
            - However, with multipolarity, the nation with excess supply can successfully sell to an array of other nations; even when the in-demand nation tells its allies not to buy, they may well not heed the request.
                - Again, history bears this out, with enemies avidly engaging in trade, even in the lead up to both World Wars (UK–Germany pre-WWI, and U.S.–Japan pre-WWII).
                    - "The reason that the British merchants and government didn't ban trade with Germany is that they knew that if they did they would lose that potential revenue source to France," says Papic. "It is literally what is happening in the news flow right now," although the central player is China's Huawei, which—having been locked out of using U.S. chips—is now building its Mate 30 smartphone with chips from Dutch firm NXP Semiconductors.
            - "This conflict won't play out with tariffs," he says. "It will play out with imperialism—not imperialism where we go out and enslave other countries, but where we create regional blocks."
            - "This is why the EU goes from a really bad idea to a model that is replicated around the world," says Papic, underscoring his belief that economies of scale—economic blocs—matter more than ever in a deglobalized world. (As such, with Brexit, the UK will find itself on the wrong side of history, whereas China is building "an Asian economic co-prosperity sphere," but the United States, having stepped away from the Trans-Pacific Partnership will be left on the outside.)
- Government spending:
    - A second major thrust Papic sees on the near-term horizon is government spending on technology.
        - Look under the hood of your favorite tech toy—say, the iPhone—and the lion's share of its smarts derives from government-funded R&D.
        - "The good news is that there does seem to be some connection between federal R&D spending and total factor productivity (which is the part of productivity that can't be explained by either capital or labor; it's basically technology)," he says.
            - Over the past decade, total factor productivity in the United States has been very low, but Papic anticipates not only U.S. but other Western nations will crawl out from beneath the post-recession austerity mindset and be ready to loosen their purse strings.
        - He credits the outrage at income inequality as the impetus for what he anticipates will be "a massive shift in politics in America" over in the next ten years, instigating increased government spending, regardless who is inaugurated as President in January 2021.
            - "My view is that in order to understand policies over a decade, you have to pay attention to the median voter, and the median voter in the U.S. is shifting to the left," he says.
            - Rather than further blowing out the budget balance, the money for additional spending will necessarily come from increased income taxes as the populace demands that the wealthy pay their fair share.
                - Papic predicts that U.S. taxpayers over the coming decade will see some combination of a wealth tax, higher income taxes generally, and increased capital gains tax in particular (perhaps with all capital gains losing the special status they currently enjoy).
                    - "This is bad if you are an investor in U.S. public markets," he says, "but the good news is that the government is going to spend a lot of money, including on innovation, including on big projects."

- From a global perspective, a big, moonshot-scale projects is surely needed to address climate change, even if the United States fails to lead the effort due to the partisan view of its threat.
  - From a geopolitical perspective, Germany is leading the way, with its Green Party experiencing a surge in support over the past year alone (from ~15% to ~25%).
  - "If this continues over the next 12–24 months, it is quite likely that the Europeans are going to start issuing what they will call Green Bonds, which are just the eurobonds that they should have issued five or six years ago, but they will suddenly have a normative reason to not be austere," says Papic. "That will be very interesting for innovation over the next ten years. Whether you believe in climate change or not is irrelevant, because it will be a driver of government spending that will drive the next leg of innovation."
    - The United States is poised to fall behind the rest of the world when it comes to climate-relevant technological innovation—technologies that will define the future of energy and transportation from 2030 onward—unless there is a sea change in attitude among climate change deniers.

Disinformation Campaigns around the World—Ms. Renee DiResta, New Knowledge

- According to a 2019 Pew Research survey, more than half of all U.S. adults (55%) turn to social media on a frequent basis for news, despite 88% recognizing that the site(s) they rely on control the news one sees and despite 62% considering this a disservice.
  - Still, not only do people keep coming back for more, they also believe what they read, especially stories and points of view shared within the confines of groups they have joined.
  - This is a problem, especially when the likes of Russia's Internet Research Agency are purposefully establishing Facebook groups to foster the spread of disinformation.
  - Renee DiResta, now at New Knowledge and fresh from a stint researching and writing "The tactics and tropes of the Internet Research Agency" for the Senate Select Committee on Intelligence, engages with TTI/Vanguard's Steven Cherry to speak about the dangers to American society as disinformation percolates through social media platforms, people lose track of the origin of the news they read, and the platforms steer folks toward increasingly nefarious and conspiratorial content, all because conflict proves engaging, and engagement increases time on site and therefore the number of ad impressions.
- The Internet is a wealth of information—some accurate, and some inaccurate; some communicated sincerely, and some knowingly wrong.
  - DiResta distinguishes among different types of incorrect information:
    - misinformation—things that are inadvertently wrong;
    - disinformation—incorrect information conveyed with an intent to deceive;
    - propaganda—information that is tailored for a particular audience with persuasion as the goal.
- Back in the days before Facebook, Twitter, and YouTube, when the open Web was still decentralized, there was plenty of erroneous information of all three sorts online, but purveyors lacked a ready means for audience consolidation, targetability, or amplification.
  - The difference today is not only the ability but also "the propensity for platforms to colonize the Internet," says DiResta. "With the ubiquity of the Like and Share buttons, the platform is instrumenting the entire Web, creating a social layer over it for everything."
    - This enables social media platforms to gather information about users both on- and off-platform: "The facilitation of you being able to share content more easily trickles back to them having a better sense of who you are," she says.
      - The impetus is two-pronged: to feed content to users that keeps them on the platform and engaged longer, and to better target users with relevant ads—ads that would increase the platform's revenue because advertisers knew they would reach their target customers.
  - As Facebook's News Feed has evolved, what began in 2006 as a stream of friends' Facebook activities eventually became littered with news articles and sponsored advertisements.
    - News Feeds became crowded spaces, spurring Facebook to curate what would rise to the top of the Feed and what would become buried so far down-screen that users would never see it.
      - "As more and more content appears on the platform, they have to curate more heavily to decide what you are going to see," says DiResta, "so you start to see things that the platform deems engaging."
      - Posts by friends about important personal moments in their lives fit this bill, but so too do does propaganda, which by design is constructed to engage.

        - "There is the needle model of propaganda," says DiResta: "If you poke someone at the right time with a message, they will be receptive to it and you will change their mind," especially when the message arises repeatedly and when it comes from trusted sources.
            - How propaganda plays out on Facebook: "Originally you came to Facebook with your social graph. You became [Facebook] friends with people you knew in the real world or maybe with two degrees of separation. But what the platform did was to create an architecture that put people into groups, and once you are in these groups, there is a remarkable repetition of the kind of content you see and the kind of people that you engage with. This is where the idea of the echo chamber begins to come about."
- When a user has many friends spread across numerous interlocking interest groups, that person is likely to remain deeply engaged with the platform—plus-one for Facebook.
    - Moreover, that user is likely to be susceptible to following "friends" to additional interest groups and therefore be exposed to more sources of propaganda—plus-one for the Internet Research Agency (or other agent of intentional and perhaps pernicious change).
        - How the progression might work:
            - A woman who is a mother joins a moms' group.
            - She clicks a Like button for Birkenstocks somewhere on the Web.
            - Facebook's recommendation engine perceives a correlation; moms with a "crunchy" profile, as DiResta puts it, are somewhat more apt to also engage with an antivaccine group.
            - Posts from the antivaxxers begin to show up increasingly often in this woman's News Feed.
                - "If you don't click, no harm–no foul, but if you do click, you have just made a new group of friends, and because you begin to engage with the content or communicate with those people, you have communicated something to the algorithm," she says, "which allows it to push you more into these groups and spaces that you might not have known you had an interest in."
            - Before long, this mom—having demonstrated an inclination to believe conspiracy theories about vaccination—is now seeing Pizzagate posts in her News Feed.
        - "This is how you go down the bizarre path from the moms' group to QAnon," says DiResta.
- Clearly, Facebook groups serve an important role for propagandists and purveyors of disinformation.
    - Not only do the platform's algorithms encourage people to join particular groups, once there the top-of-page posts are those deemed algorithmically most engaging; often these are flame wars.
        - "It doesn't matter what the content is," says DiResta. "What matters is that people are communicating, so that [other users] want to get in and join in, too."
    - As Cherry points out, this opens the door for bad actors to establish their own narrative, especially since anyone is free to form a Facebook group.
        - "The propaganda has always evolved to fit the information architecture of the day," says DiResta. "The information architecture of today lends itself to this very participatory, peer-to-peer type of content creation."
        - It is no secret that, despite the platform's real-name policy, many millions of Facebook accounts are pseudonymous—sometimes for valid reasons of self-protection, but sometimes to take on an identity to ingratiate oneself into a target group.
        - Evidence indicates that long before the 2016 U.S. elections, the Internet Research Agency, at the behest of the Russian political interests, began to use its troll farms to set up Facebook groups, develop relationships with group members, and eventually evolve the focus of such groups to further Donald Trump's electability.
            - "We saw the creation of fake pages and fake-persona accounts," says DiResta. "They created the pages to appeal to particular communities and created the pages as members of that community. All of the messages—all of the content—were inflected as if they were authentic speakers, themselves."
                - That is, Russian trolls were masquerading as descendants of Confederate soldiers, anti-Hillary Clinton African Americans, and so forth.
                - "It wasn't an anti-Hillary Clinton message," clarifies DiResta. "It wasn't *You shouldn't like Hillary Clinton*, but it was *As members of group X, we do not like Hillary Clinton*, always reinforcing this in-group identity and reinforcing and entrenching beliefs that people already had before they got there by just nudging them a little bit in a particular direction."
    - The closed nature of Facebook and the other major social media platforms is a challenge for DiResta and other members of the research community.

        - Tools might enable her to see how many people viewed a particular YouTube video, for instance, but not to open a window into subsequent activity—that is, how the video influenced views or actions.
            - "A common critique of the Russia propaganda issue is that it is overblown because they just targeted people who already believed this," she says. "To some extent that is true, but we don't have a very strong sense of how exposure to this kind of material is impacting people. The platforms do [have access to this data], but the people on the outside don't."
- Might there come a time when the general populace becomes desensitized to messaging strategies that are, when laid bare, so clearly manipulated?
    - The numbers reflected by the Pew survey at the start of this piece suggest otherwise: People are aware of the problem, are not happy about it, but continue to engage with social media platforms as principal sources of news.
        - Research shows that young people are more discerning in this regard than older folks, who are consistently more likely to be taken in—and exploited—by extreme groups on the Internet.
            - A possible educational thrust might be to use *AARP The Magazine* to engage in digital literacy with those who need it most: "Offer guidelines on how not to get scammed, how not to be taken in, how you should fact-check, and how you should think about the information you are receiving," suggests DiResta.
    - "A degree of trust in the government and a degree of trust in the information ecosystem are needed for a cure," says DiResta, but trust in both is currently low.
        - "We are instead thinking about this in terms of inoculating—preventing people from falling into these groups," she says.
            - Mechanisms could include forbidding social media platforms from recommending extreme propaganda groups to users: "They'd be allowed to exist, but not promoted."
            - This, of course, raises the challenge of how to properly balance freedom of expression in the United States with the dangerous influence extreme communities seek to exert—a challenge that is complicated by questionable moral leadership at the apex of government, with cynical falsehoods and conspiracy theories being actively fostered, encouraged, and repeated.
                - "How do you live in a state of perpetual descensus?" she poses. "How do you live in an environment in which people move in almost completely different realities?"
                - "It makes me think that other parts of the world stand a better change [for inoculation]," says DiResta.
        - Facebook is currently promoting a purported solution that DiResta fears will push the problem under the table, rather than help to solve it.
            - "With the trend for people to be in small groups, Facebook is going to push WhatsApp, which is not only small, but also encrypted," she says, "which means that nobody on the outside sees what is going on in those groups."
                - DiResta contrasts Facebook's move with Reddit's approach, which she finds more palatable: "Reddit will moderate if it finds a group actively harassing another group. They tend to have a better sense of what's happening, because they have a smaller ecosystem to manage than Facebook."
                    - This raises the question of scale, and whether Facebook's network is too large for the good of society.
        - More generally, the amorality of Facebook's algorithms does not promote societal well-being; even though, with profitability as Facebook's prime metric, the algorithms could arguably be deemed fair.
            - "There is no determination where the algorithm stops and thinks whether to put people in the QAnon group and whether it is a beneficial thing for that person to be in that group, or whether it is a beneficial thing for to push a new mom who might have concerns about vaccination into the antivaccine group," says DiResta. "That's where you start to get at the question of what happens when you have a system that is based in a very amoral way, purely around gameable metrics."
- Even having established both social media and political landscapes as important players in the current disinformation snafu, DiResta does not exonerate conventional media from playing a part.
    - In particular, the consolidation of media into a mere handful of powerful conglomerates exacerbates the problem.
        - Example: Sinclair Broadcast Group's policy of must-run segments and scripted talking points echoed by local broadcasters from coast to coast.

- For a lot of the social campaigns, the goal is to get it trending, get it viral, and get it up to Hannity or Maddow late at night, so that the entirety of the television-watching public sees it," she says.
    - And the migration goes both ways, with the News Feed percolating the most scandalous or polarizing news to the top of the page, rather than that which is either the most recent or objectively the most important.
- At the same time, with news coming to people from Google Assistant or Apple News or Facebook's News Feed, many no longer bother to pay attention to the actual source of their news.
    - "I think it is really easy even for educated people to lose sight of what the source is," says Cherry. "One person's tabloid is another person's canonical source, and vice versa for their political mirror."
    - Did the article first appear in the *Washington Post* or on Breitbart? Was the story from NPR or RT?
        - To its credit, in the wake of the 2016 U.S. election, Facebook has implemented page transparency features that make plain the source of an article, yet either too few people engage with the tool, or too few care to discriminate when it comes to stories or opinions toward which they are inclined.
        - To some extent, such tendencies are innate; as one TTI/Vanguard participant puts it, "There is a tendency to think something is the truth when you hear it. Psychologically, we don't question something well enough, and also people don't like changing their minds."
        - DiResta raises the potential of strategies used for cult deprogramming, notably the influence of redirecting beliefs by people within an individual's true sphere of trust, particularly by those who were previously caught within the web of a conspiracy campaign but have come out the other side and can speak truth with empathy.
- Harkening back to matters pertaining to Russia, although the fact of its government's direct influence/control over the activities of the Internet Research Agency is not ironclad—IRA is a private firm tightly linked to Russian oligarch Yevgeny Prigozhin—there is no such plausible deniability for activities conducted by the nation's Main Intelligence Directorate (GRU).
    - "The GRU has run influence operations on social platforms since about 2012," says DiResta. "They have failed as social propagandists, but they have incredible hacking capabilities."
        - (This comes as no surprise to anyone familiar with the TTI/Vanguard presentation, Spy versus Spy, by Eric Haseltine and Charles Gandy in September 2019.)
    - "While the GRU hacked the Democratic National Committee and hacked the Hillary Clinton campaign—two examples of hack-and-leak operations that they did in the United States—they have also executed that strategy in a number of places worldwide," reports DiResta.
        - The Baltics and Montenegro are other recent targets; in the latter, the hack was followed up by a planned assignation attempt on that nation's prime minister [Milo Dukanovic] on the day of the election in question.
            - "The information ecosystem is but one tool in an arsenal," she emphasizes. "First they tried to ensure the guy didn't get elected [in Montenegro], and then they took things in another direction."
        - Russia has also been going all-in on infiltrating African politics by hiring locals on the ground in a variety of countries to disseminate politically impactful falsehoods.
- China, in contrast, approaches its spheres of influence differently.
    - "There are geopolitical goals, and cultural and military competencies, that make [China] look quite different" says DiResta.
    - The Asian giant has launched what she considers a surprisingly sloppy response to the months of antigovernment protests in Hong Kong, and it has not responded as forcefully toward Taiwan as many would have anticipated.
    - "We can say that there is a substantial amount of propaganda, we can say that there are accounts that behave in anomalous ways, we can say that there are fake accounts and bots involved in the conversation that put out narratives in line with what we would expect from China, but making that attribution is not a thing that anyone has been able to do yet," says DiResta.
- And now the United States is in the throes of the 2020 election season.
    - DiResta's comments about the challenge of attributing online actions to China—and the generalization of this challenge to other nations or internal purveyors of disinformation—will be a major issue between now and November 2020.
        - If Russia carries over its African tactics to its influence over the U.S. election, attribution will be particularly difficult.

    - "The question becomes where is our hyperpartisan content originating from, who is paying for it, and how is it being pushed out to people," says DiResta. "The platforms are very reluctant to take down domestic political speech because they prioritize preserving viewpoints, and that is one of the challenges that we face when we don't have that attribution."
- The policies of the various platforms on political advertisements do not follow a single model, with some, in fact, in opposition to their own terms of operation.
    - "Google changed its policy a couple of days ago, Twitter said it is not accepting certain types of political ads, and Facebook said it is not going to fact-check political ads," says DiResta. "Facebook's political ad policy conflicts with its own misinformation policy. If the content is posted organically, it is subject to fact-checking and is subject to their remove, reduce, inform framework to help people understand the organic content, but if it is produced by a politician as a paid political ad, they have chosen not to put that through the same policy as the organic content. In my opinion, that is a terrible decision."
- She anticipates another significant hack-and-leak operation at some point before next year's election, in large part because of how hungrily the media laps up such things, as well as "nebulous attribution of the stuff that does go viral."