HOT CHIPS 33

Architectural challenges: AI Chips, Decision Support and High Performance Computing

Dr. Dimitri Kusnezov

U.S. Department of Energy

August 24, 2021

Disclaimer: These views are mine and not of the DOE

Preliminary reflections

"Scientific development is like Darwinian evolution, a process driven from behind rather than pulled toward some fixed goal towards which it grows ever closer."

Thomas Kuhn, in The Trouble with the Historical Philosophy of Science, 1992

While Science and Technology are inexorably intertwined, they are distinct, and can evolve independent of each other.

- *Quantum technologies today rely on the theories developed a century ago (1920s). [Science leading Technology]*
- *The Steam Engine, perhaps the most transformational technology of its period, was not scientifically understood for roughly a century following its invention (though science helped with the building blocks). [Technology leading Science]*

We shouldn't confuse discovery and invention. Both are important and need separate consideration.

DOE: Department of "Everything"
Why do we care about chips and architectures?

- **We are a mission agency**: *we have problems to solve, often on schedules.*
- **We have tough requirements***: US Energy Sector, Cyber over many unclassified and classified networks, Nuclear Security, Emergency Preparedness and Response, Intelligence, Strategic Petroleum Reserve, Loan Program, Environmental clean-up,…*
- **We create and maintain tools and people to work these challenges:** *17 national labs, world class user facilities, worlds fastest supercomputers,..*
- **We solve problems no one else can.**
- **We worry about rare and unlikely eventualities.**
- **We know how to protect data and information.**
- **We are a go-to agency for informing crises and urgent decisions.**
- **Today AI based methods, which are still nascent and narrowly applied, are providing means to innovate and impact everything including health, science, environment, energy, and climate.**

We turn to science and technology to make predictions to inform decisions & AI is increasingly becoming a part of this. But the path we are on needs to evolve.

Decisions & Discovery

- Aside from the sets of problems we can readily find our way through with today's and tomorrow's (near term) technologies and approaches, the most challenging things we face in the future are:
  a) Providing actionable options for increasingly complex, high consequence decisions via science-based predictions;
  b) Reinventing scientific discovery.

- We operate beyond market and take a long view to solving problems.

- Today's high-performing computer path won't get us there. Neither will the AI path we are on. Its also not about 'faster'.

Decisions are based on the questions that emerge, sometimes urgently:

What actions are needed and when?

What is your confidence?

What does it mean?

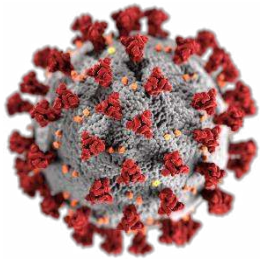
What are the risks?

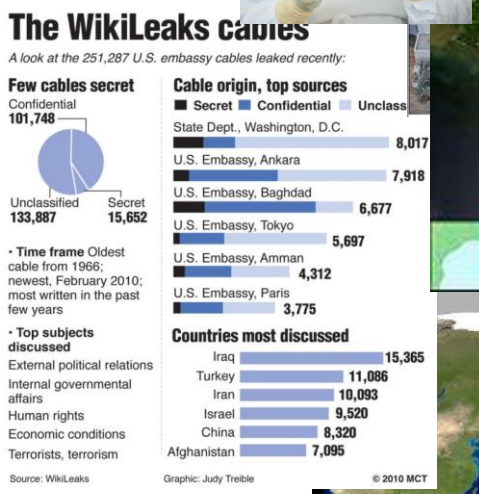
What happened?

Can it happen again?

What are the options?

…





There are no do-overs if you are wrong, and no deep data to start from.

Are they the right questions?

Are the right people asking?

Are we positioned to answer them?

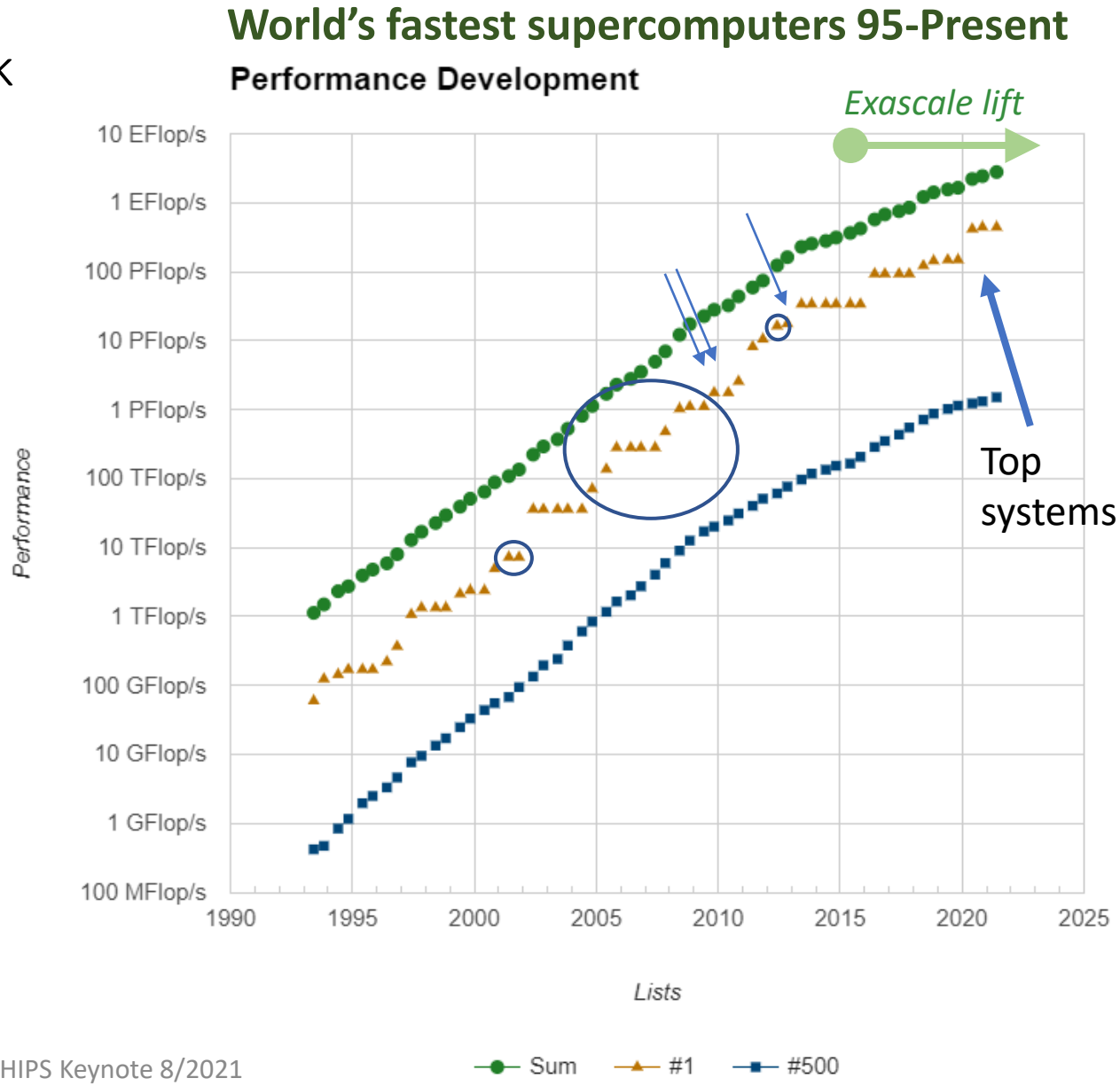
The questions are often imprecise, ill-posed, but we still have to do what we can.

My lens is based on a particular experience:

To meet nuclear security mission needs, took risks on architectures and chips to make the predictions we required for decisions:

- First to petaflop (ahead of DARPA effort by 3 years);
- Planned for >10^6 processor systems in early '00s and delivered in '12;
- Launched graphics co-processor system (using PS3 chips) in '07 at petascale; planned in early '00s;
- Created novel architectures and novel chips delivering world's fastest systems (circled), sometimes with top 2 or (3 out of 5) maintaining US leadership for years;
- Novel NIC chip kept Cray viable and transitioned them from vector systems to XT series;
- Delivered (and exceeded) on DOE's 10yr grand challenge to deliver a performance increase of 10^4 -10^6
- …

There are directions we should be pushing beyond our current heterogeneous designs.



Lets take a step back and examine how we have learned to address complex questions:

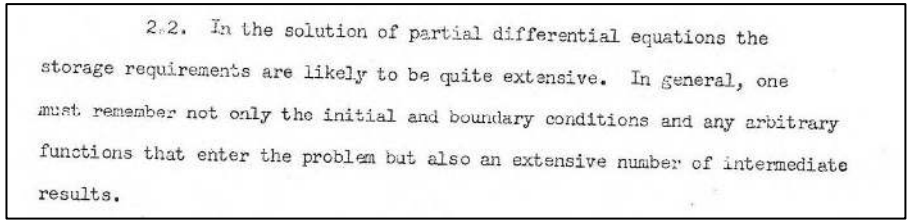
- Traditional paths are to model or to measure. These are based on how we have learned to create understanding and emerged from the Scientific Revolution.
- Several competing factors shaped the Scientific Revolution[§] *(1572-1704) :*
  1. The culture of science was created with key ingredients we use today: discovery, originality, progress, authorship and their practice.
  2. Printing press: knowledge dissemination, communities of expertise.
  3. Measuring instruments: telescopes, microscopes, barometers, prisms,.. ← *discipline of Data*
  4. New Theories: Newton, Galileo, Kepler, Pascal,.. ← *discipline of Theory*
  5. Language of facts.
- The societal problems and big questions of the time were understanding ballistics and longitude/time-keeping for navigation. The scientific approach led to their solution.
- The Scientific Revolution, which was not appreciated at the time, defined and became the model of modernity.
- It shaped the world for the next centuries, including the industrial revolution and bringing us to our technological society today.
- *This has bearing on where we are today.*

[§] D. Wootton, " The Invention of Science" (2016)

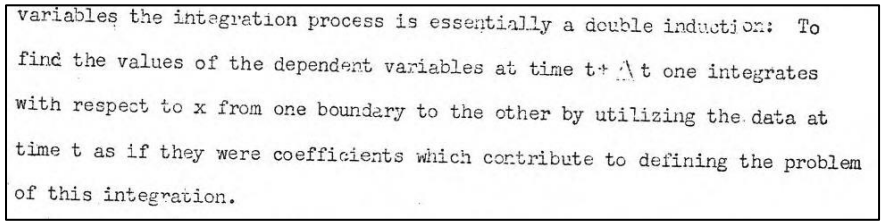
Most of our efforts have been focused on the modeling side for past 70 years

- Our roots in computing are in the Manhattan Project and the AEC, DOE's predecessor.
- John von Neumann, among others, helped craft computer architectural concepts based on solving partial differential equations:

"Preliminary Discussion of the Logical Design of an Electronic Computing Instrument", Burks, Goldstine, von Neumann, (1946)



2.2. In the solution of partial differential equations the storage requirements are likely to be quite extensive. In general, one must remember not only the initial and boundary conditions and any arbitrary functions that enter the problem but also an extensive number of intermediate results.



variables the integration process is essentially a double induction: To find the values of the dependent variables at time t+Δt one integrates with respect to x from one boundary to the other by utilizing the data at time t as if they were coefficients which contribute to defining the problem of this integration.

- When we propose our next supercomputers, we provide performance benchmarks that are based on solving many traditional classes of problems rooted in this approach.
- Creating several architectural paths over the past 20 yrs was deliberate in attacking the specific modeling problems.
- But this approach, even with our heterogeneous architectures, is not closing the gaps on how we deal with growing richness of data.

How we make decisions in simulations has defined how we develop the supporting computer technologies

Our leading edge high performance computing (HPC) has been designed for, and enable exploration of, theories and models well beyond what humans can analytically compute.

Computing has allowed us to push the limits of taking all of our theoretical understanding of the world to make predictions and to test them.

- *The most complex theories today cannot be solved any other way.*
- *Our largest computer simulations stress our largest systems.*
- *Our most complex models can be more than 15 orders of magnitude in length and perhaps 10 orders in time.*

Two points:

- *Decisions/Actions are based on trust in the resulting predictions and this lives outside the space of models for many complex issues.*
- *Architectures have historically been developed to solve equations and not to deal with data at today's and tomorrow's scale.*

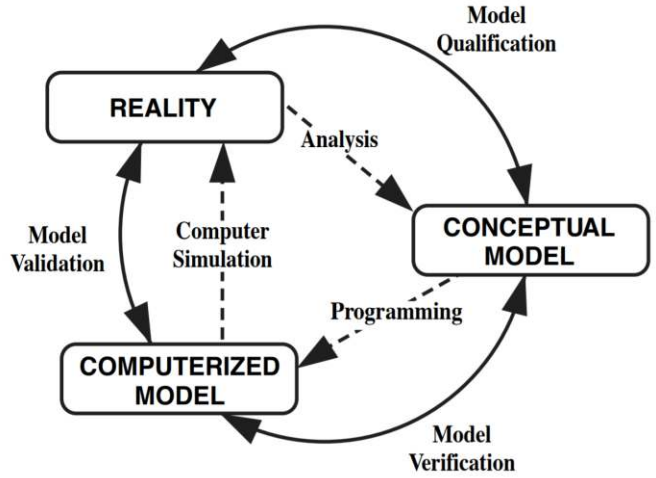
We have worked hard to develop approaches to make decisions through “Uncertainty Quantification” (UQ)

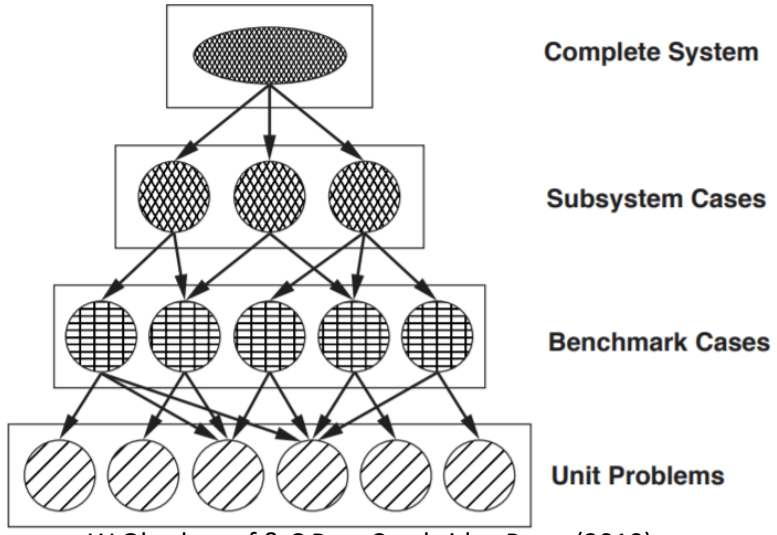
So how do you provide guarantees that what you predict is actionable?

- We focus on pragmatic solutions since urgencies and schedules can drive decisions.
- Trust is not an afterthought and has to be built in at every step.
- Computer architectures have been designed and built around this approach now for many years.

We do this well, but it is becoming increasingly strained since:

- The historical approach cannot accommodate the volumes of both simulation data and experimental data.
- It is based on small amounts of data touching simulations at key points.





e.g. see W Oberkampf & C Roy, Cambridge Press (2010)

Decision support is not just post-processing. It must be built into every aspect of the problem.

UQ is not a statistical package we add onto simulations. Things do not add in quadrature.

These problems are at least NP hard.

There are many intangibles that have to be accounted for, e.g.

- *Surrogate materials in simulated conditions – no metric for how 'close' they are to the desired environments and materials.*
- *Calibrations, phenomenology, uncontrolled approximations, material properties, discretization approaches,… (Think of GPS without General Relativity, Epicycles, Newton and ballistics,…)*

We have experts in the loop at every scale, we use families of different approaches etc.

We strain the limits of Supercomputer architectures to meet the demands of making predictions.

We validate our abilities on broader classes of problems.

What we have developed here over the years for likely the most complex simulations of systems, has no counterpart in the AI/ML world.

AI and ML are becoming important to our day-to-day mission, business and operational needs

Increasingly data rich domains that touch just about everything we do, with a range of complexity from straight forward to unsolved. For instance:

- *Clean energy & Climate resilience*
- *Effective all-hazard response to energy sector disruptions*
- *Nuclear non-proliferation*
- *Classification/declassification*
- *Physical security*
- *Methane leak detection*
- *Infrastructure: Surge, line slack, security*
- *Environmental remediation*
- *Future of Scientific Discovery*
- *Grid reliability and resilience*
- *Oil & Gas*
- *Potential sorting of multimodal data sources*
- *Nuclear emergencies*
- *Nuclear Security*
- *Future cybersecurity/smart grid/wireless nexus*
- *Safety in high-hazard operations*

Accelerating discovery cycles is a natural step

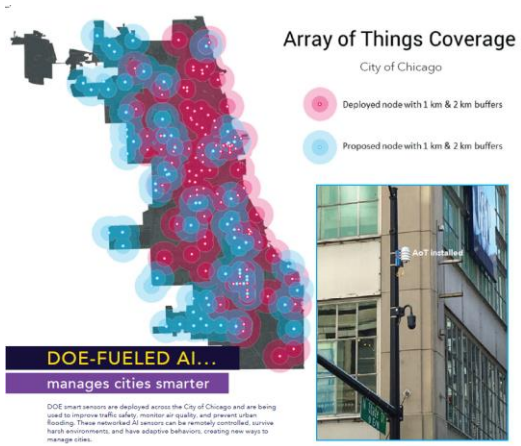
An end-to-end smarter processes will accelerate discovery rates due to faster experimental cycles:

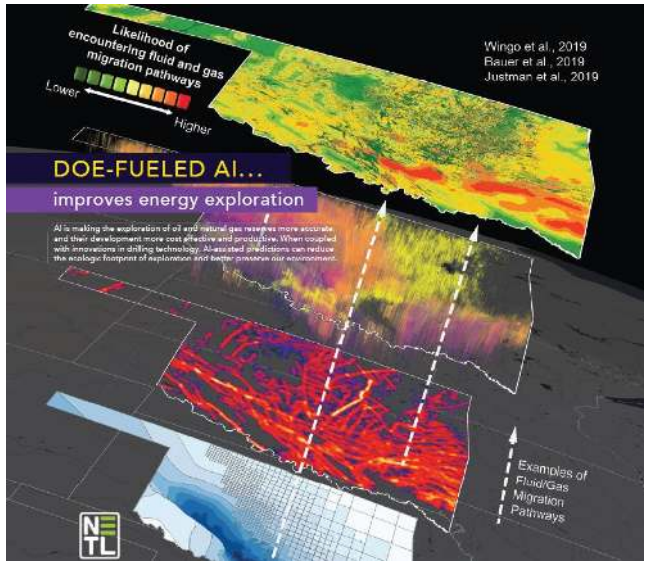
- Accelerated Discovery: *Anomaly detection in events, images…, source/transient classification*
- Interpretation of Measurements
- Using learned models and the latent space to augment data: *ML based models with UQ – observations, measurements, interpretation of simulations and experimental outcomes*
- Acceleration of discovery rates due to faster experimental cycles: *Targeted search, optimization, automation*
- Smart facilities and instruments: *Semi-autonomous science driven by active learning loops*
- Simulation + AI hybrids data
- Accessible and Integrated Knowledge bases: *New interfaces to the literature, data and models*
- AI everywhere, smart processes, smart data, smart simulations

There is a place here for many types of AI technologies discussed at HOTCHIPS. It is a first step.

Larger societal problems will need something more

- We have to come up with options that protect us against natural and man-made disasters.
- Complex issues with impact on people have to be informed by scenarios that we hope never happen.
- For many problems, the solution does not lie just in the data:
  - Energy
  - Climate & Climate resilience
  - Energy/Water
  - Environment
  - Nuclear and National Security
- Just simulating faster wont help here.





We look for willing partners who see beyond the near term horizon

- Our challenges are not transactional. If they were, life would be easy.

DoE Strategic Partnership with SambaNova Systems

U.S. DEPARTMENT OF ENERGY

SambaNova SYSTEMS

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

Department of Energy and Cerebras Systems Partner to Accelerate Science with Supercompute Scale Artificial Intelligence

Argonne National Laboratory and Lawrence Livermore National Laboratory Are First Labs Announced in Multi-laboratory, Multi-year Collaboration with Cerebras

September 17, 2019 09:00 AM Eastern Daylight Time

LOS ALTOS, Calif.--(BUSINESS WIRE)--Cerebras Systems, a company dedicated to accelerating artificial intelligence (AI) compute, today announced a partnership with the U.S. Department of Energy (DOE) to advance the massive deep learning experiments being pursued at its laboratories for basic and applied science and medicine with supercomputer-scale AI. Argonne National Laboratory and Lawrence Livermore National Laboratory are the first labs announced in Cerebras' multi-year, multi-laboratory partnership, with more to follow in the coming months. The partnership comes on the heels of Cerebras' introduction of

cerebras

CEREBRAS SYSTEMS

Release Summary

Department of Energy and Cerebras Systems Partner to Accelerate Science with

groq

GRAPHCORE

Back to the main transformative factors that shaped the scientific revolution

1. *The culture of science was created with key ingredients we use today: discovery, originality, progress, authorship and their practice*
2. *Printing press: knowledge dissemination, communities of expertise*
3. Measuring instruments: telescopes, microscopes, barometers, prisms,.. ← *discipline of Data*
4. New Theories: Newton, Galileo, Kepler, Pascal,.. ← *discipline of Theory*
5. *Language of facts*

Since the 1950s, we have pushed the development of HPC architectures focused on advancing the discipline of Theory. HPC systems have been built and honed to enable exploration of theories and models well beyond what humans can analytically compute.

> We moved away from pen to paper calculations, tables of evaluated functions and learning approximation techniques to aggregating models across many length and time scales to simulate things never thought possible.

Today, AI Chips, Data, sensors, storage, measurements, and ML tools and approaches are enabling exploration of data and information well beyond what humans can comprehend, advancing the discipline of Data.

So we are at a place where #3 and #4 have moved beyond human abilities but Decisions/Actions based on trust in predictions requires both of these simultaneously. Today's suites of chips and architectures were not designed for this.

Decisions and Discovery: Next steps?

What endured from the Scientific Revolution was not specific laws or data or techniques. Rather it was the structured discipline of how we approach discovery. And this led to a fundamental transformation of society to an industrial one and now a technological society. Where do we go from here?

Complex decisions and scientific discovery in our era where both theory and experiment can be done far beyond what humans can comprehend requires us to develop the means to explore the fullness of mathematics and theory/simulation in the richness of all available data.

For discovery, this intersection is where great lasting ideas emerged over the centuries, but we had the luxury of depending on unique individuals to be these intersection points.

For decision making, it is the only way to make increasingly complex predictions that have consequences.

From chip designs to computer architectures, we have been caught in the paradigm created and driven during the Cold War.

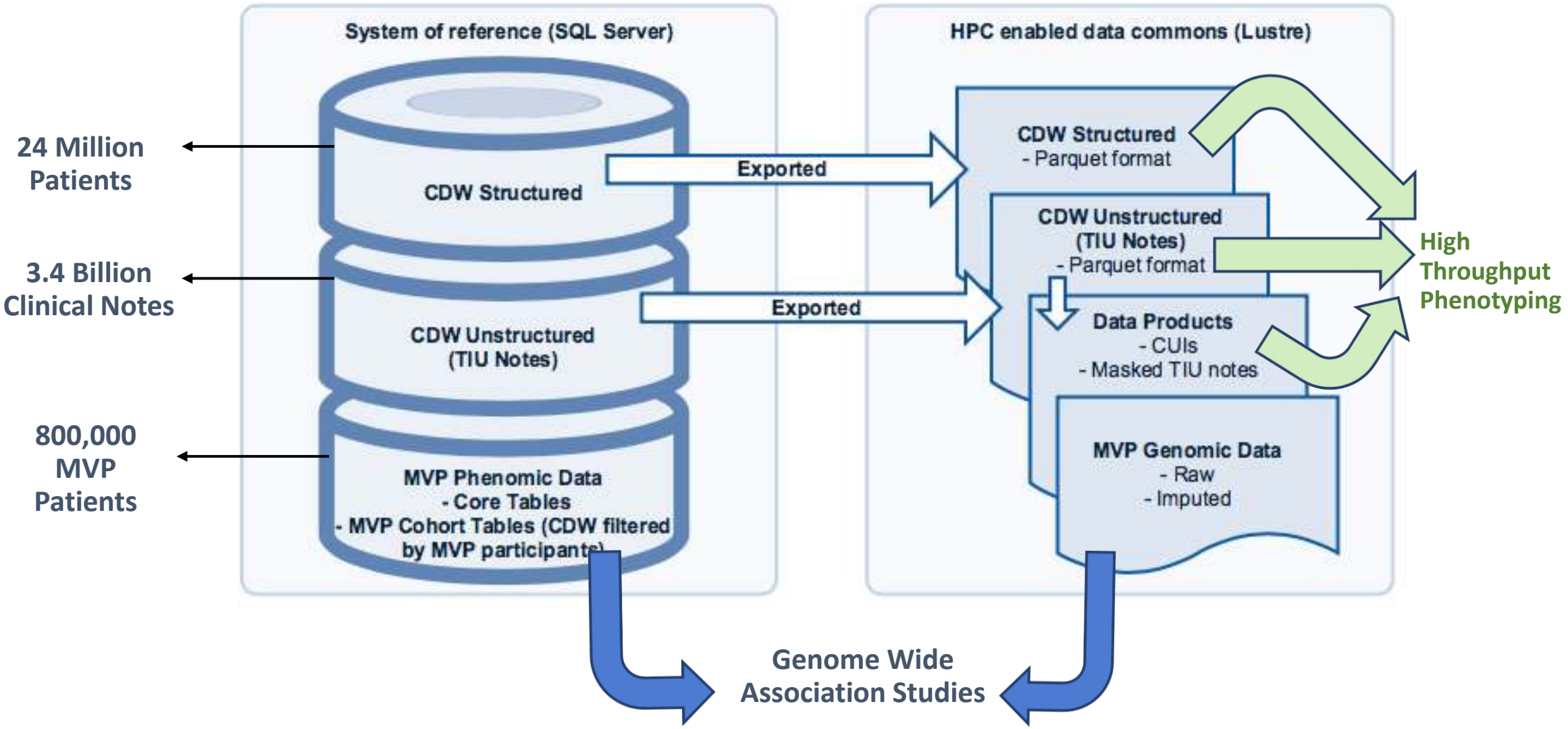
Today's novel AI chips are a breath of fresh air and offer opportunities to depart from where we have been, but they focus, as one should expect, on near term market opportunities and specific application sets.

Getting to where we want to go will require some deep reflection on how and why we do what we do.

Using health data to break the hold of our past

- Force the rethinking of traditional paradigms by challenging ourselves with qualitatively new classes of prediction and a richness of data.
- We worry about rare events - Our tools need a pedigree, and developing and testing on complex data where we can evaluate robustness and test against actual outcomes is important.
- Use the qualities of data to change how we think of many of our traditional approaches from architectures to Uncertainty Quantification (UQ) to codes and data.
- Align with where next economic drivers could provide most amplification
- Partners include: *Federal Agencies; Philanthropy (Weill Family Foundation), Academia, Pharma (GSK…), Tech Sector , Foreign Governments(Norway,…)*
- **Veterans issues we are exploring**:
  - ❑ Suicide Risk, Prostate Cancer, Cardiovascular Disease, Polypharmacy, Opioids.
  - ❑ Rich multimodal data ranging from genomics to EHRs, images, notes,…
- **Traumatic Brain Injury (TBI)**
  - ❑ With the TRACK-TBI consortium, also connections to sports, military and civilian data.
- **Many other data sources as well related to cancer (population, patient, molecular), drugs discovery,…**

DOE-VA: Snapshot of the MVP CHAMPION Data Repository



Quick update on Semiconductors

Recognition on the Hill of importance of US semiconductor industry.

Most significant bill that became law:

> The 2021 NDAA or in its fuller title the "William M. (Mac) Thornberry National Defense Authorization Act for 2021".

Includes:

> TITLE XCIX—CREATING HELPFUL INCENTIVES TO PRODUCE SEMICONDUCTORS FOR AMERICA ("CHIPS for America") [page 1456 of 1480]

and

> TITLE XCII—COMMUNICATIONS MATTERS [page 1398]
>
> *Authorization for activities but no funding.*

Senate: USICA Bill. Proposes funding levels. Passed 68-32. Over $50B for semiconductors of $250B that can help as well.

House: Require reconciliation process and development of companion bill(s) to set funding levels. This is still pending.

> *(more detailed discussion at the HPC User Forum 9/9/21 hpcuserforum.com/events/html)*

Summary

- Many of the societal problems we face today do not have the benefit of being data rich for the situations we worry about: in climate, energy, health, security.
- They also live outside most of the market forces that drive technology development.
- We have to look beyond our current horizons and balance the grip of near term pressures with these larger opportunities.
- Exploring new paradigms will require risk that many are not willing to do, but the benefits are societal.
- As with the Steam Engine, AI (as defined today) is a means to get to a place where we can achieve far beyond our own limitations. It is not and endpoint but rather a step towards this evolution.

Special thanks: Dr. Bambi DeLaRosa (Micron)