



The Transformation of Science with HPC, Big Data, and AI

Jeff Kirk

HPC & AI Technology Strategist, Dell EMC

Oct 17, 2017

j.kirk@dell.com

DELLEMC

Context & Topics

- This will be a mostly oral presentation, ranging from the history of supercomputing to the future of computational science integrating.
- Topics to be covered include
 - **Simulation**-based science and **supercomputing**
 - **Big Data, data analytics**, and data-driven science
 - **IoT**: even more data
 - **AI, machine learning, deep learning**
 - Putting it all together: it's about getting science done

A Short Digression...

From 1930's until 2010's:

CPU (and GPU)
+
Programming

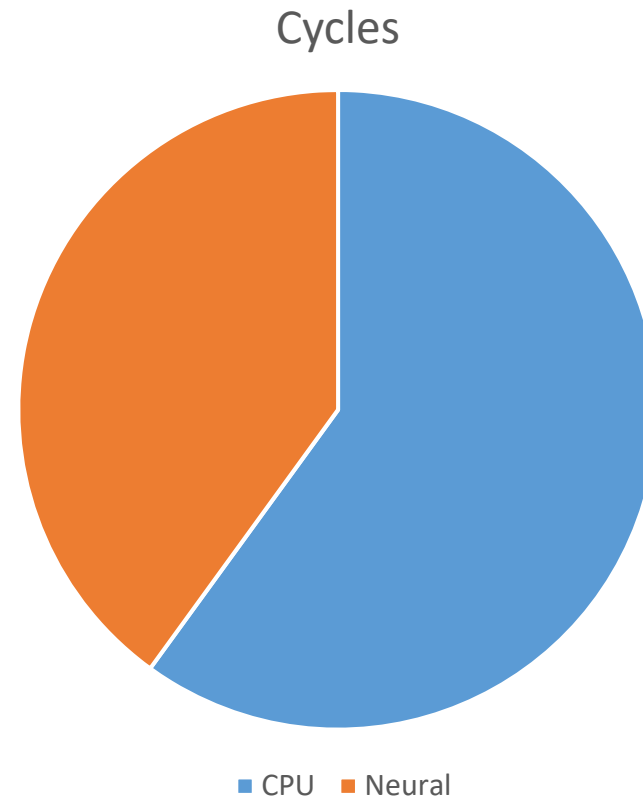
Recently Added:

Neural
Networks
+
Learning

My Big Prediction


In 10 years:

Neural Networks will represent about 40% of all compute “cycles”



Simulation-based Science, Supercomputing, HPC

- Computational science has been mostly about simulation for its ~70 year history
 - Solving theoretical equations forward in time to make predictions
 - Comparing those simulations outputs to observations, to test theories
 - Revising/improving theories/equations to make more accurate predictions, improve understanding of nature
- The human body, our world, and the universe are far bigger ‘models’ than any computer can solve *directly*, hence supercomputing
 - First supercomputers had much higher performance than ‘regular’ computers: Crays and CDCs with very different processors
 - Over time, performance and price of microprocessors won out
 - **Supercomputers became scalable systems of microprocessors, eventually clusters of commodity servers**



THE GRAND CHALLENGE EQUATIONS

$$B_i A_i = E_i A_i + \rho_i \sum_j B_j A_j F_{ji} \quad \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad \vec{F} = m \vec{a} + \frac{d\vec{m}}{dt} \vec{v}$$

$$dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV \quad \nabla \cdot \vec{D} = \rho \quad Z = \sum_j g_j e^{-E_j/kT}$$

$$F_j = \sum_{k=0}^{N-1} f_k e^{2\pi i j k/N} \quad \nabla^2 u = \frac{\partial u}{\partial t} \quad \nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J} \quad P(t) = \frac{\sum_i W_i B_i(t) P_i}{\sum_i W_i B_i(t)}$$

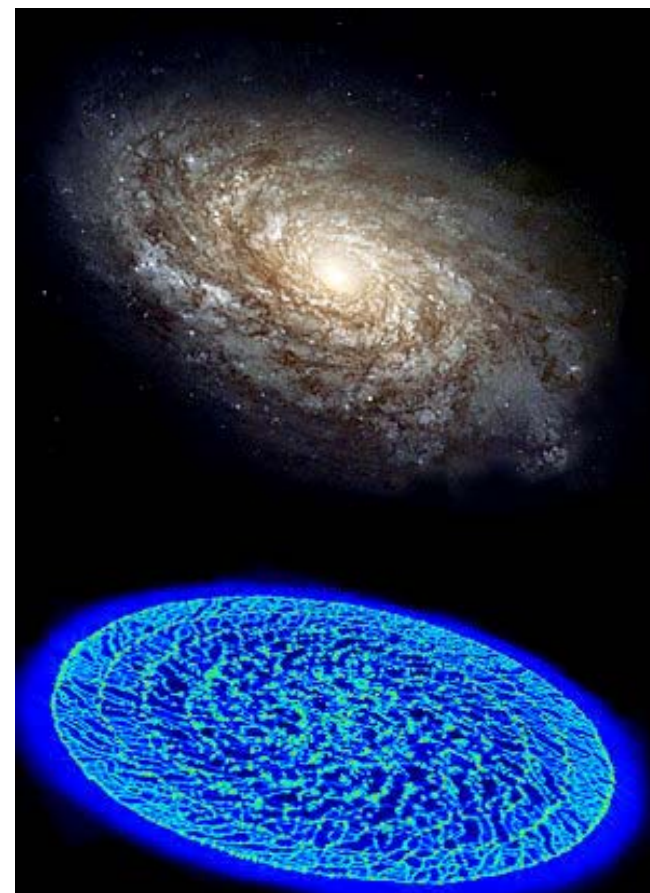
$$P_{n+1} = r p_n (1 - p_n) \quad \nabla \cdot \vec{B} = 0 \quad -\frac{\hbar^2}{8\pi^2 m} \nabla^2 \Psi(x, t) + V \Psi(x, t) = -\frac{\hbar}{2\pi i} \frac{\partial \Psi(x, t)}{\partial t} \quad -\nabla^2 u + \lambda u = f$$

$$\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla) \vec{u} = -\frac{1}{\rho} \nabla p + \gamma \nabla^2 \vec{u} + \frac{1}{\rho} \vec{F} \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f$$

• NEWTON'S EQUATIONS • SCHRÖDINGER EQUATION (TIME DEPENDENT) • NAVIER-STOKES EQUATION •
 • POISSON EQUATION • HEAT EQUATION • HELMHOLTZ EQUATION • DISCRETE FOURIER TRANSFORM •
 • MAXWELL'S EQUATIONS • PARTITION FUNCTION • POPULATION DYNAMICS •
 • COMBINED 1ST AND 2ND LAWS OF THERMODYNAMICS • RADIOSITY • RATIONAL B-SPLINE •

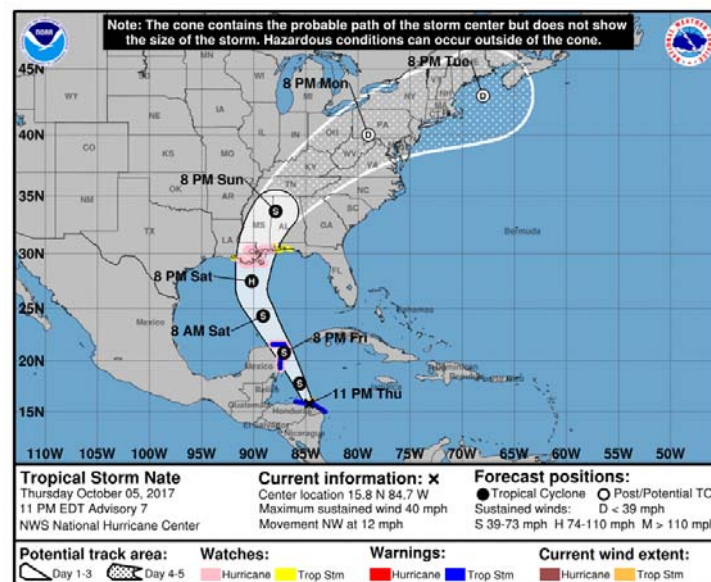
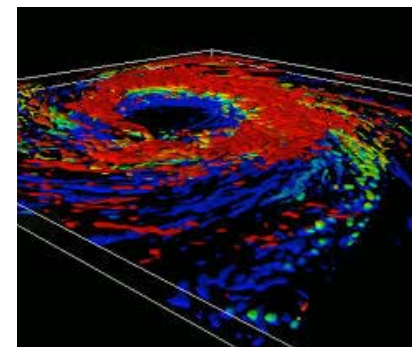
Gravity & Orbits: Becomes a Supercomputing Problem

- Do we need a supercomputer to calculate the orbit of a moon around a planet? Not really
- Do we need a supercomputer to calculate the orbits of 100 billion stars in a galaxy? And of all the motions of the interstellar medium? Yes
 - In fact, we still cannot do it directly: we use algorithms with approximations, that we hope are accurate enough
- More computing power, memory, etc. enables us to solve huge problems more accurately (usually), more quickly
- Exascale computing is not just a 'machoflops' goal or for bragging rights: it's needed for solving real, complex, important problems in weather/climate, healthcare, neuroscience, energy and more.



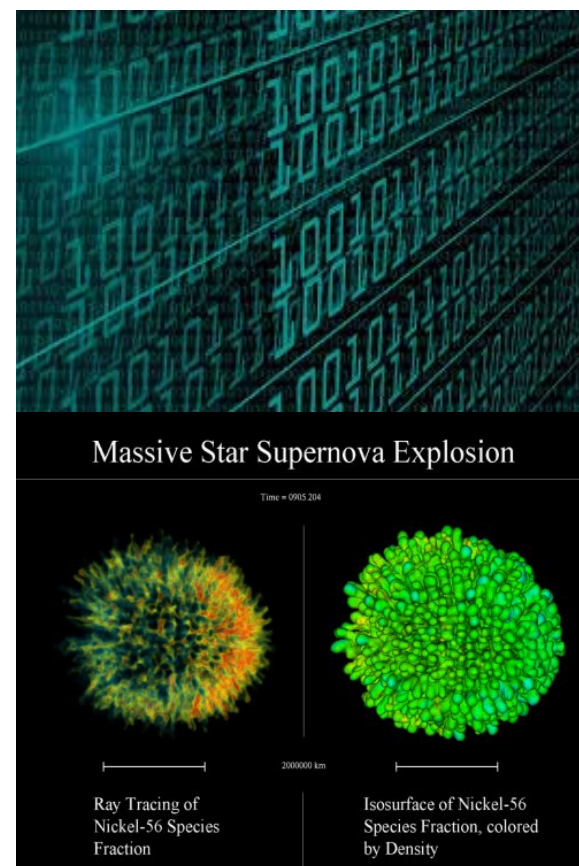
Hurricanes: Supercomputing Is Crucially Important, but Not Perfectly Accurate

- Hurricane modeling and simulation can only be done with sufficient predictive performance on large supercomputers
 - Otherwise, simulations will take longer than real time, and results will be too inaccurate
- However, non-linear dynamics means simulated answers accuracy decreases in time
- Thus, good predictions require
 - *as much data as possible*
 - *many simulations, to produce a higher accuracy through statistical probability*
 - *ongoing data assimilation*
- Hurricane projected tracks have not been great this season—it's a very difficult problem!



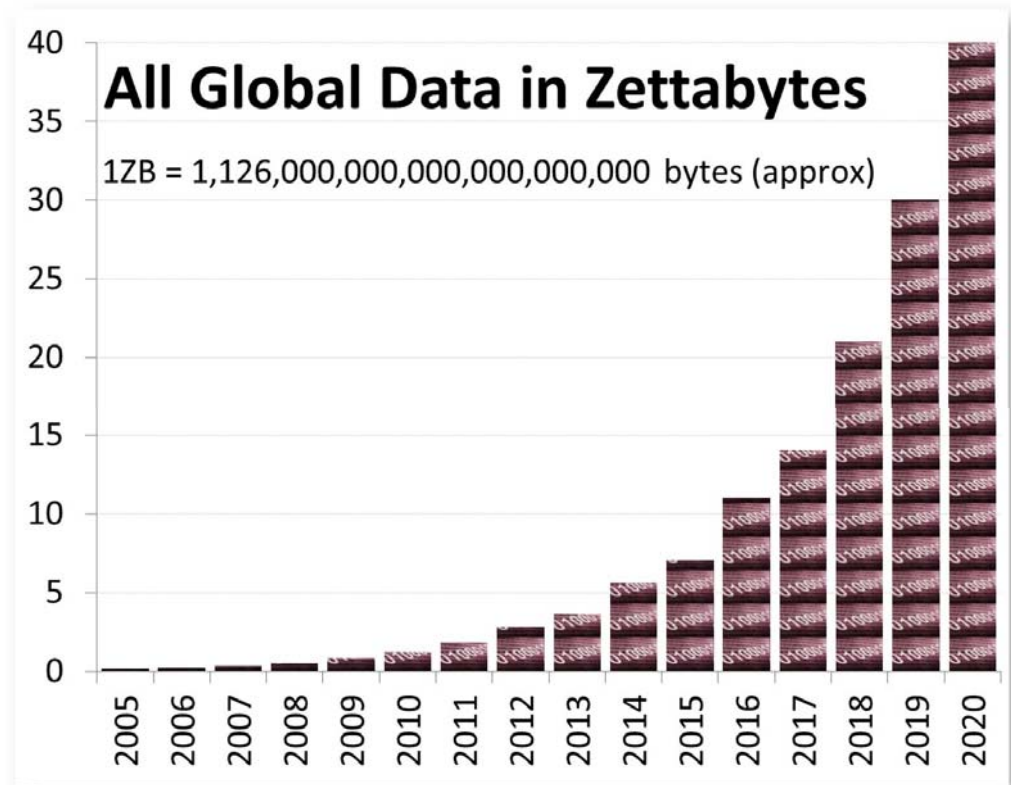
Note: Computational Science Has Always Had Big Data, and Performed Data Analysis

- Simulations are always compared to observational/experimental data
- We analyze the output data of simulations to compare accuracy to observational data, and this output can be very 'big data' based on simulation resolution and timesteps
- Most simulation data analysis is visualization of output
 - But there is also lots of statistical measurement of characteristics, feature scales, etc.



The Rise of 'Big Data' and Data Analytics

- Technology has enable production, collection & analysis of massive data
 - More computers, more plentiful
 - Ubiquitous sensors
 - **More capable instruments**
 - Smartphones, cameras
 - Social media, email
 - Etc.
- Business are harvesting this data to improve products, services, sales
- Science is producing and using data to treat disease, understand weather, etc.

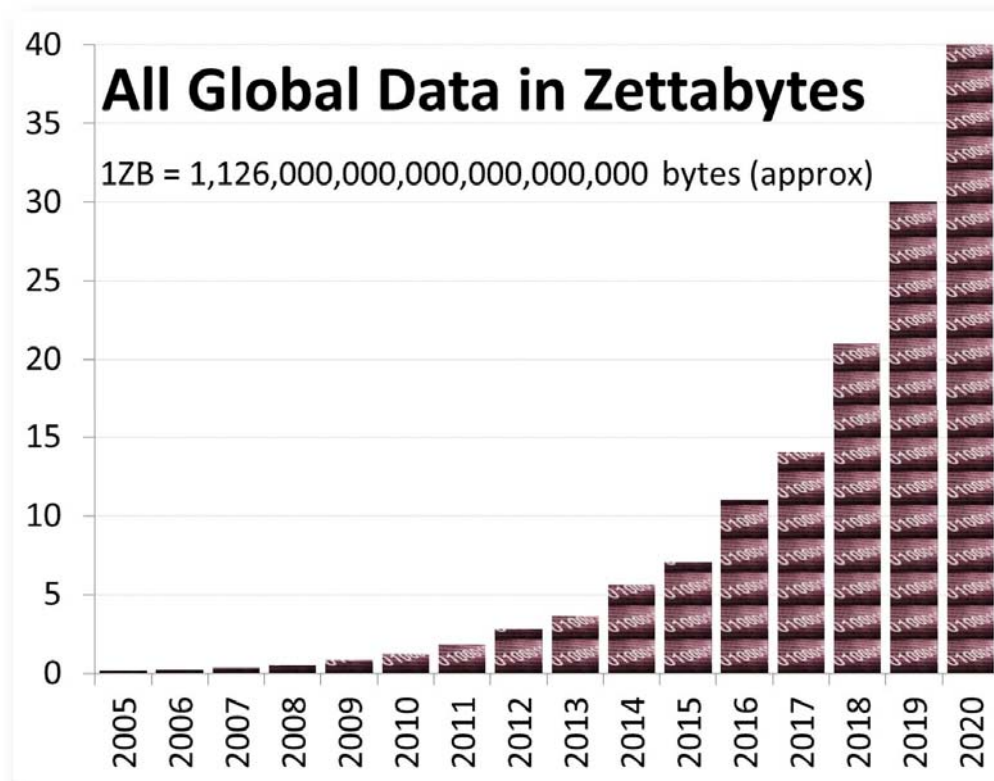


United Nations Economic Commission for Europe

DELL EMC

The Rise of Big Data, Data Analytics

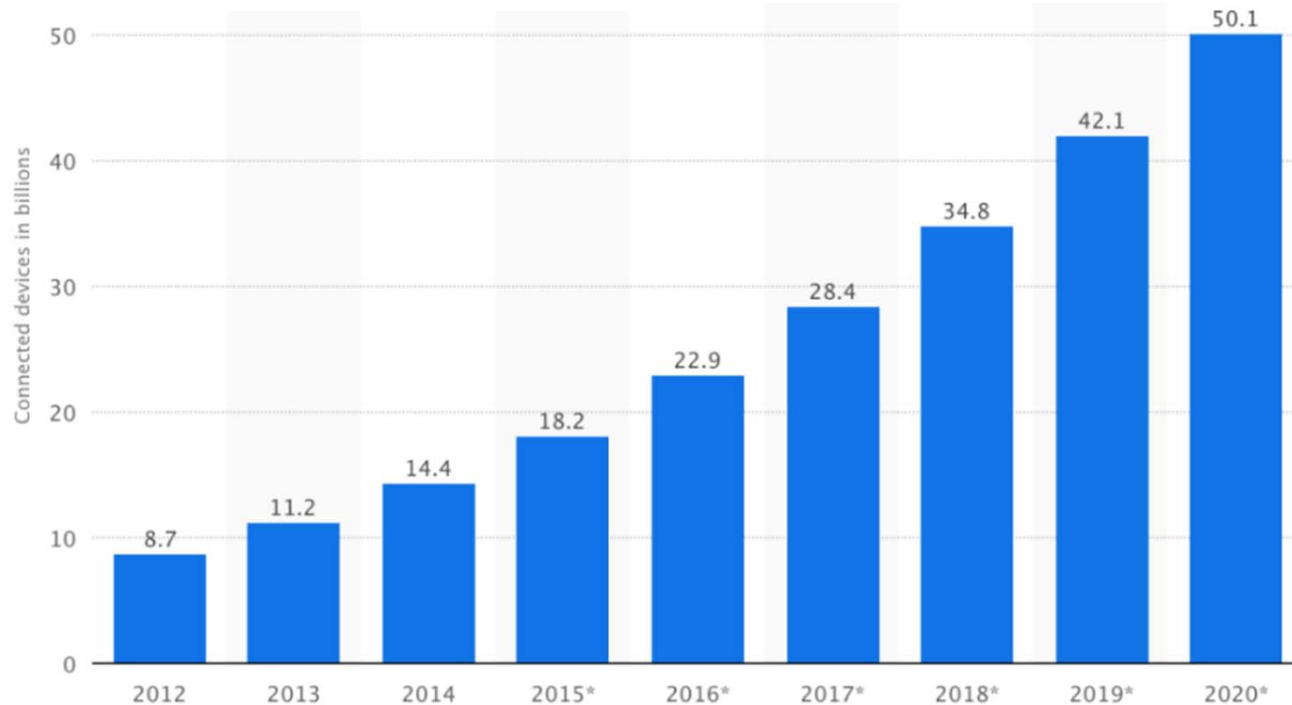
- Technology has enable production, collection & analysis of massive data
 - More computers, more plentiful
 - Ubiquitous sensors
 - More capable instruments
 - Smartphones, cameras
 - Social media, email
 - Etc.
- Science is producing and using data to treat disease, understand weather, etc.
- **Harvard Business Review claims (2012) that Data Scientist is the Sexiest Job of 21st Century!**



United Nations Economic Commission for Europe

DELL EMC

The Internet of Things (IoT): Tens of Billions of Connected Devices



Internet of Things (IoT): number of connected devices worldwide from 2012 to 2020 (in billions)

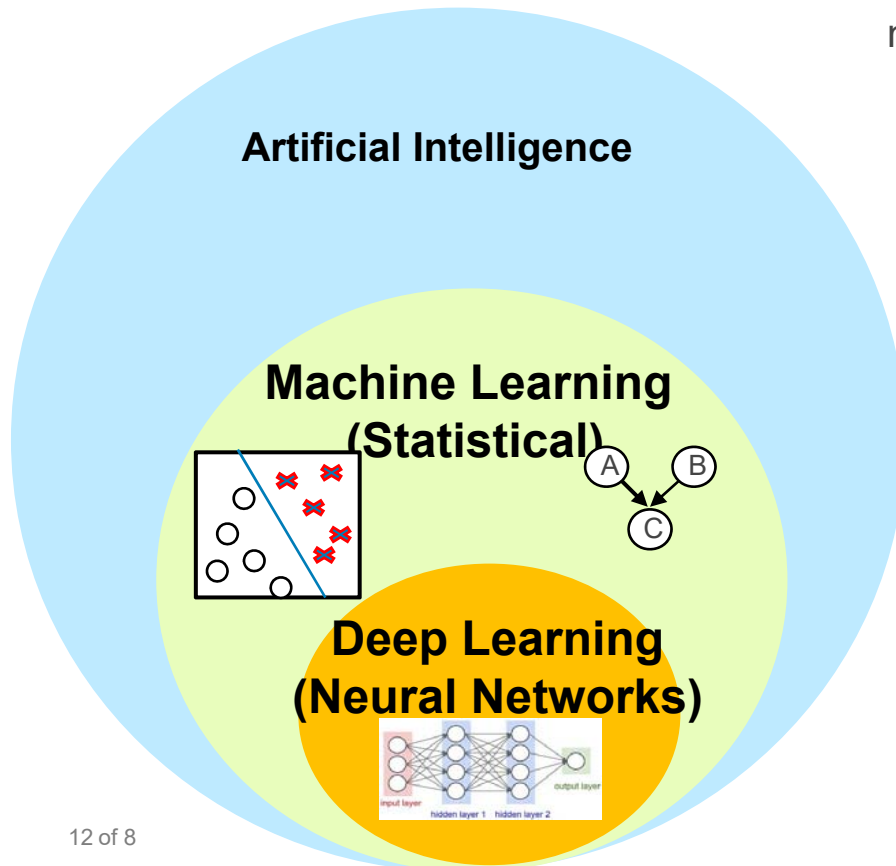
AI: Machine Learning & Deep Learning are Transformative, Disruptive Analytics

Artificial Intelligence is the broader concept of machines being able to carry out tasks in a way that we would consider “smart”.

Machine Learning is a approach in which we enable computers make decisions without explicit programming: ***we use data to train a model which can then be used to make inferences, predictions, etc. (probabilities).***

Deep Learning is an area of Machine Learning that uses neural net algorithms for training the model using massive data.

A **Neural Network** is a computer system designed to work by classifying information in the same way a human brain does. It can be taught to recognize, for example, images, and classify them according to elements they contain.



Tech Leaders Are Proclaiming Disruption, Revolution...

“A breakthrough in machine learning would be worth ten Microsofts” – Bill Gates

“AI is the most far-reaching technological advancement in our lifetime. It changes every industry, every company, everything.” – Jen-Hsun Huang, Nvidia CEO

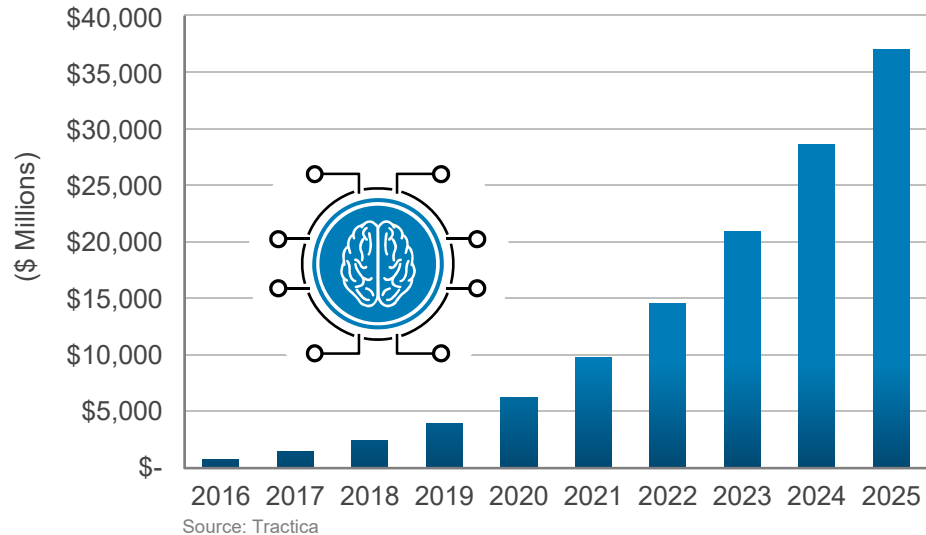
“Just as electricity 100 years ago transformed industry after industry after industry, I think AI powered by deep learning will now do the same... It’s hard to think of an industry that will not be transformed by AI in the next decade.” – Andrew Ng, former Baidu Chief Scientist

“Smart machine technologies will be the most disruptive class of innovations over the next 10 years due to their computational power, scalability in analyzing large-scale data sets, and rapid advances in neural networks.” – Gartner Report

Cognitive computing will become “the largest consumer of computing cycles by 2020” – Rob High, IBM Watson CTO

Analysts, Experts, & Our CEO Agree on AI Impact

Artificial Intelligence Revenue, World Markets: 2016-2025



“IBM Invests \$240 Million Into AI Research Lab With MIT As It Struggles In AI Battle”

“The initiative's first public investment: lead investor in a \$10.5 million funding round for Seattle startup Algorithmia”



“Revenues for Cognitive Solutions to \$4.6 billion last quarter.”

*By 2019, **startups will overtake Amazon, Google, IBM and Microsoft** in driving the artificial intelligence economy*

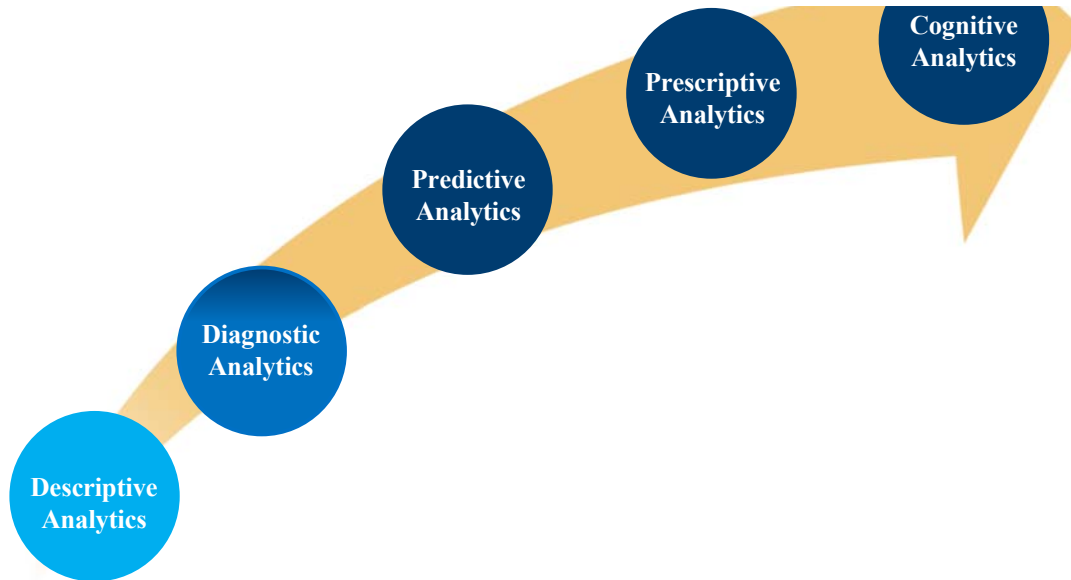


“When you get all this data coming in from hundreds of billions of connected devices and apply to that artificial intelligence, it’s almost like a fourth industrial revolution and an incredible opportunity for companies to re-imagine themselves in this digital age... The last 30 years have been incredible in IT, but the next 30 years will make it look like child’s play.”

– Michael Dell, Sapphire SAP Conference May 16th 2017



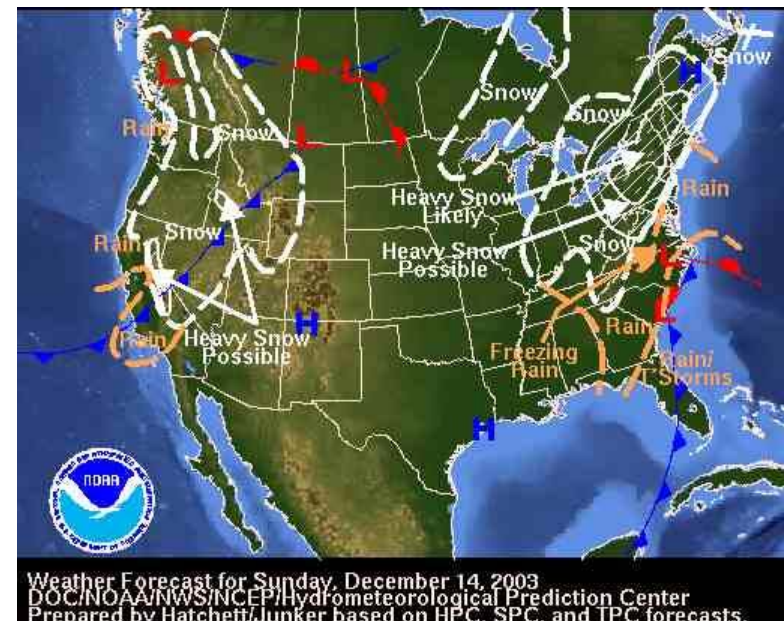
The Evolution of analytics



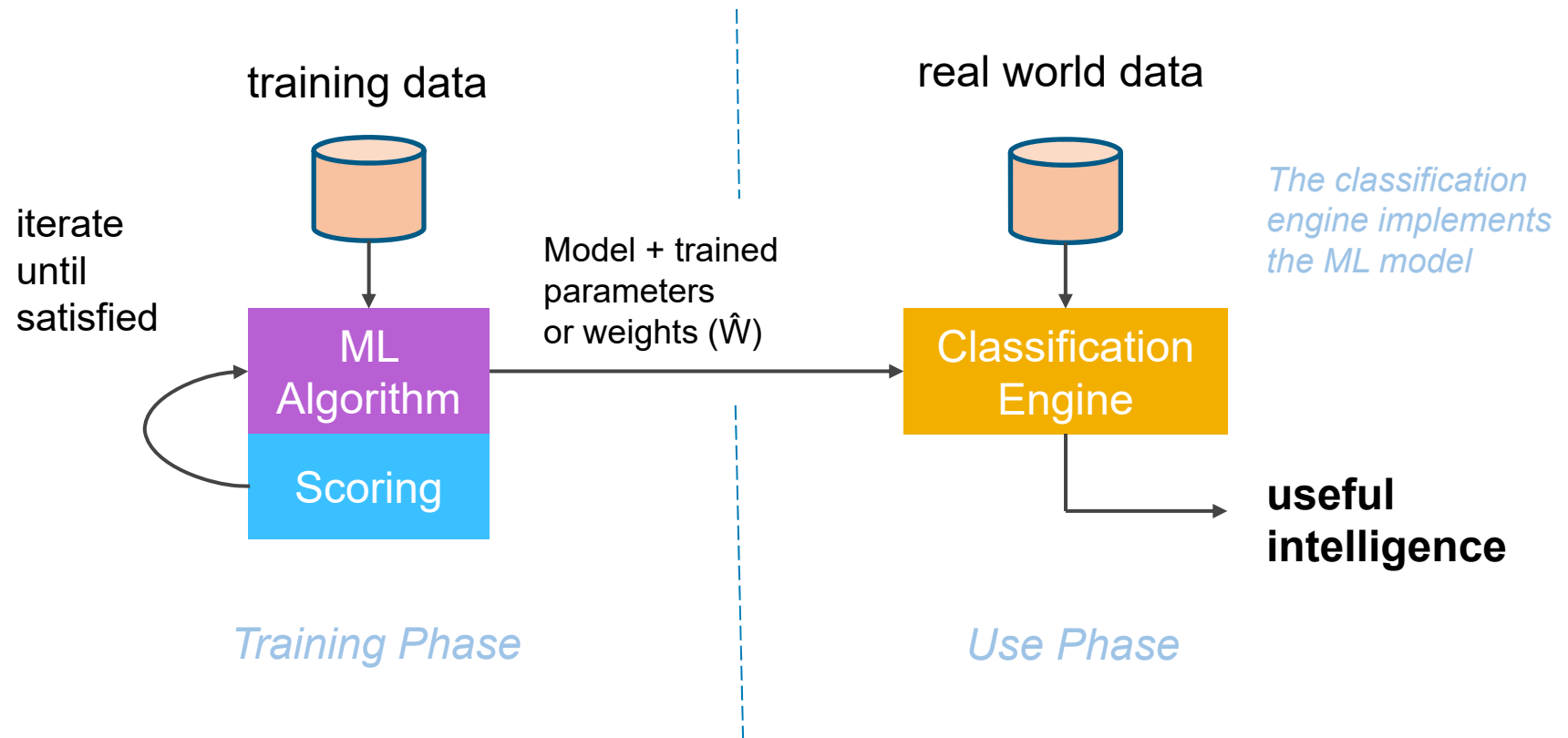
Background: Intelligence from Processing

Machine Learning – condensing data into a high-dimensional probability model to be used for:

- **CLASSIFICATION** – using the model to label or tag the data
- **INFERENCE** – using the model to deduce probable inputs given some outputs
- **JUDGEMENT** – summarizing the content of the probability model
- **PREDICTION** – using the model to deduce probable outputs given some inputs



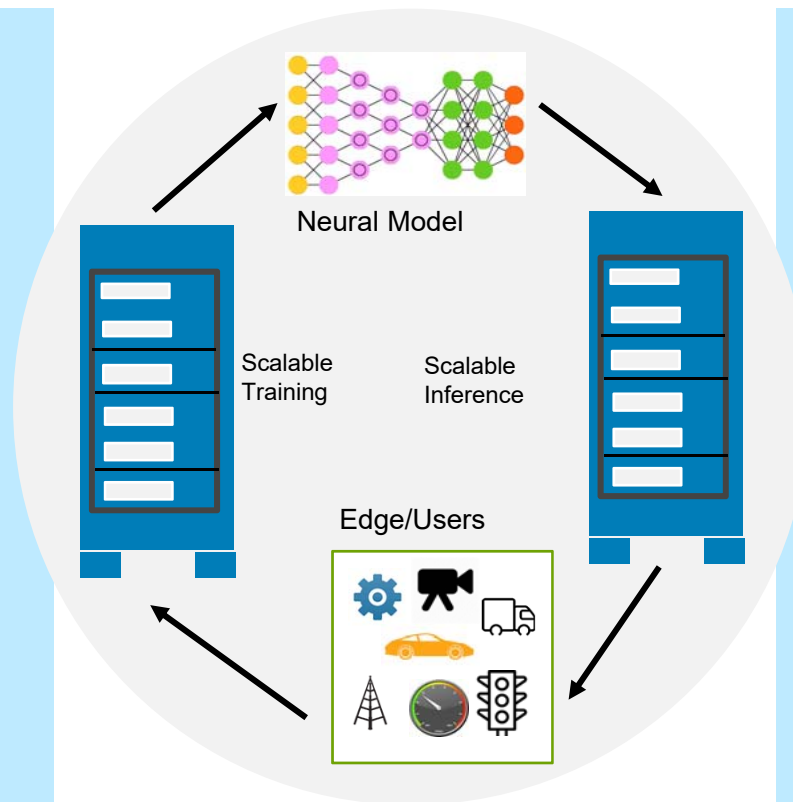
Background: The Machine Learning Process



Deep Learning – Train Model, Then Inference Against

Training

- **Computationally Intensive: massive data, massive computations in neural net**
- Billion of Tflops per training run (train a model)
- Can sacrifice precision (e.g. FP32, FP16) for more performance



Inference

- Less computationally intensive, but still must be fast (and often low power)
- Doesn't require high precision math, so can use accelerators like GPU & FPGA with INT8 support.
- Can also be run in Xeon & Xeon-Phi based systems.

Background: Machine Learning Requires Matrix Math

Math Matrix form of RSS* (minimize the error)

$$\mathbf{y} = \mathbf{H} \mathbf{W} + \boldsymbol{\epsilon}$$

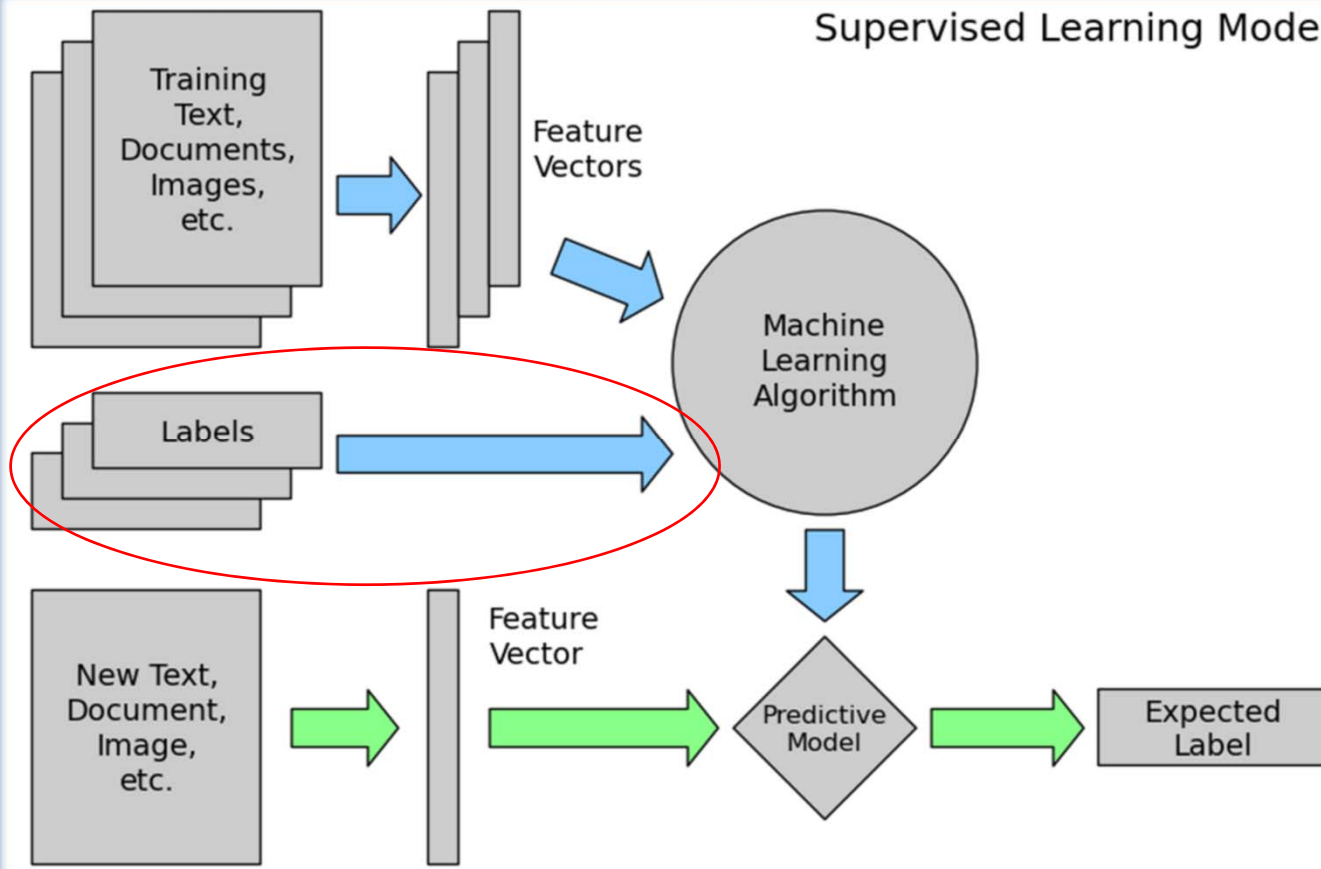
y = true value
H x w = predicted
ε = error

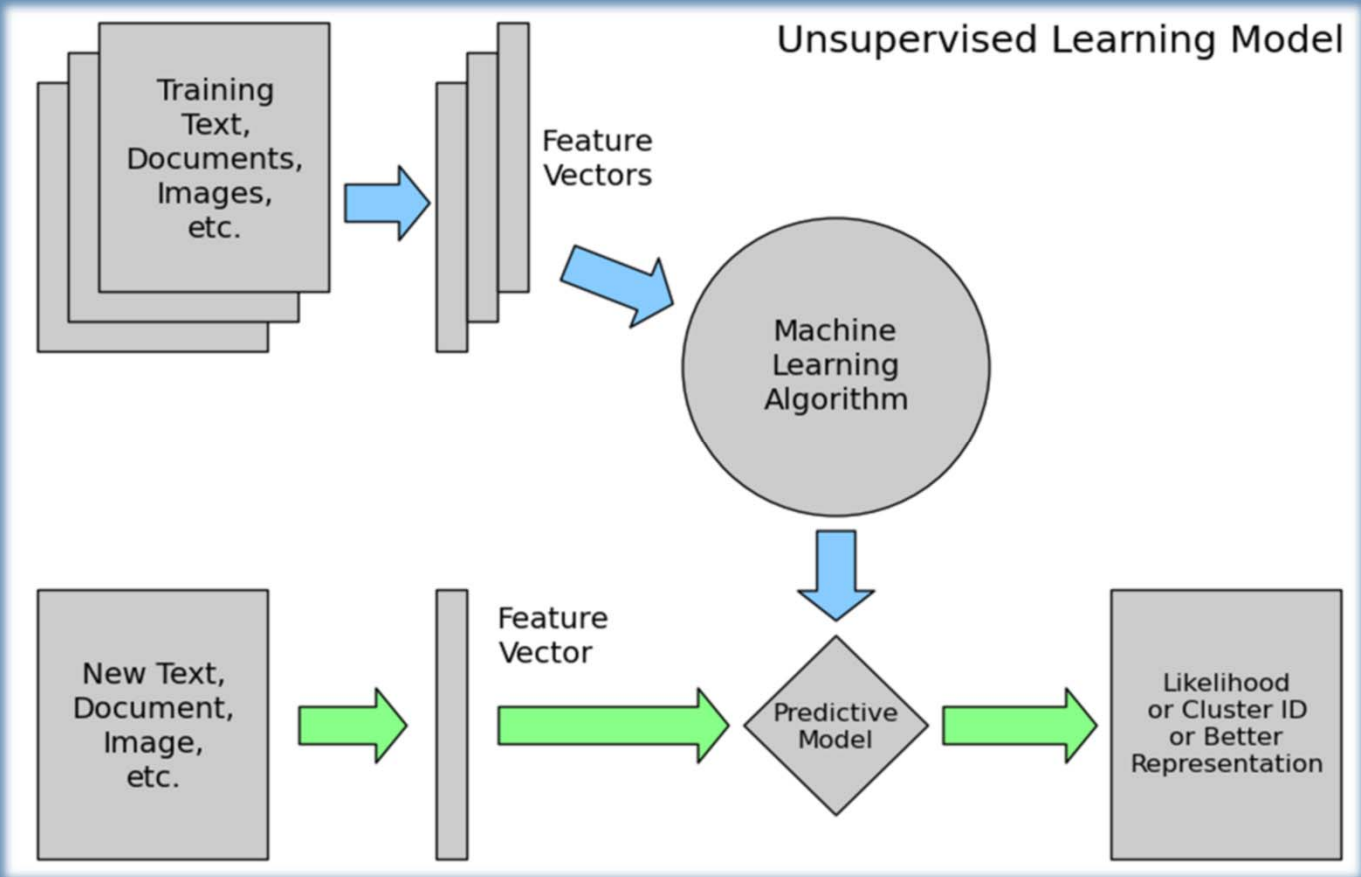
**W = parameters
or coefficients**

H = training data

*residual sum of square

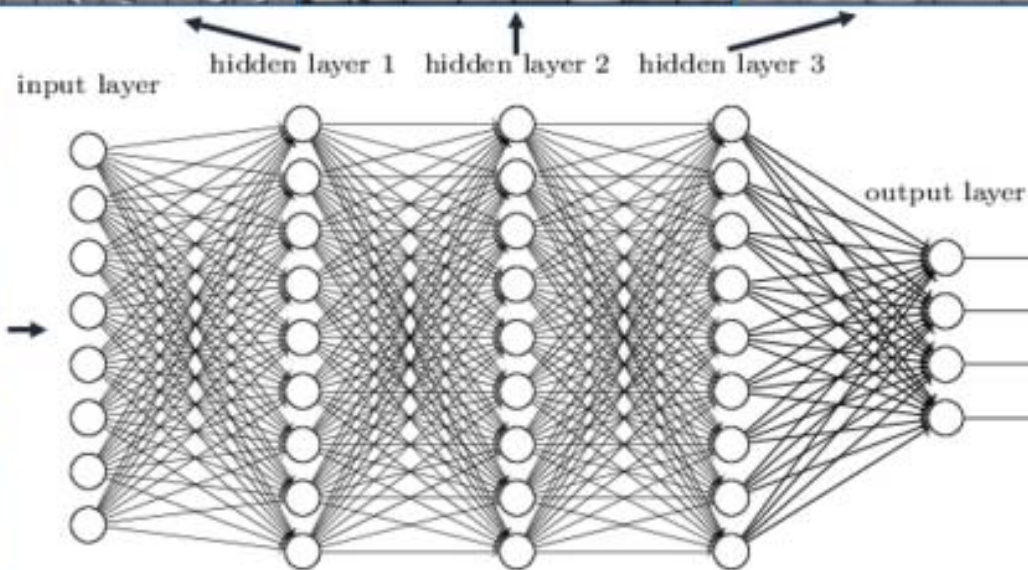
Supervised Learning Model





Deep Neural Networks Learn Features in Layers

Deep neural networks learn hierarchical feature representations



Deep Learning Score Card

Pro:

- Enables learning of features rather than hand tuning
- Impressive performance gains
 - Computer vision
 - Speech recognition
 - Some text analysis
- Potential for more impact

computational cost + so many choices
=
incredibly hard to tune

Con:

- Requires a lot of data for high accuracy
- Computationally really expensive
- Extremely hard to tune
 - Choice of architecture
 - Parameter types
 - Hyperparameters
 - Learning algorithm ...

Plus,

not analytic,
i.e. often
little insight
to the
weights





Background: Math Performance is Key

- Most of the recent performance gains by GPUs and KNM is due to precision optimizations:





Precision Evolution: 64-bit DP → 32-bit SP → 16-bit HP → Some 8-bit

- But there is one more optimization step: **specialized silicon**
 - Special 16 bit precision enhancements
 - Better internal network, i.e. graph support and more connectivity
 - Better use of memory
 - Inference --> lower bit precision (4 or even 1 bit)

Importance of AI in Enterprise, Government

Healthcare & Life Sciences	Financial Services	Government	Manufacturing
<ul style="list-style-type: none"> Alerts and diagnostics from real-time patient data Disease identification and risk stratification Patient triage optimization Proactive health management Healthcare provider sentiment analysis 	<ul style="list-style-type: none"> Risk analytics and regulation Customer segmentation Cross-selling and up-selling Sales and marketing campaign management Credit worthiness evaluation 	<ul style="list-style-type: none"> Cyberattack intrusion attempt detection, analysis Smart power, transportation design for resiliency Terrorist threat prediction Socioeconomic trends and population planning 	<ul style="list-style-type: none"> Predictive maintenance or condition monitoring Warranty reserve estimation Prosperity to buy Demand forecasting Process optimization 

IDC: in 2018, 75% of enterprise & ISV development will include AI/ML in at least one application

Retail	Energy	Transportation	Travel & Hospitality
<ul style="list-style-type: none"> Predictive inventory planning Recommendation engines Upsell and cross-channel marketing Marketing segmentation and targeting Customer ROI and lifetime value 	<ul style="list-style-type: none"> Power usage analytics Seismic data processing Carbon emissions and trading Customer specific pricing Energy demand and supply optimization 	<ul style="list-style-type: none"> Vehicle crash avoidance systems Smart traffic routing Public transportation planning for maximum mobility Smart service vehicles for optimal routes, autonomous navigation 	<ul style="list-style-type: none"> Aircraft scheduling Dynamic pricing Social media-consumer feedback and interaction analysis Consumer complaint resolution Traffic patterns and congestion management 

Can AI Replace Simulation?

- AI (and data analytics in general) can provide answers where there is plentiful, accurate data, but no, or insufficient, theoretical understanding for predictive simulation (business, social models, some healthcare issues, etc.)
- AI (and data analytics in general) can supplement simulation where theory is not entirely sufficient, or where historical data is plentiful but real-time data is too sparse (perhaps in weather?)
- ***More commonly, AI will augment simulation in guiding simulations and in data analysis, and new supercomputers are being designed with this in mind***

SCIENTISTS ENLIST SUPERCOMPUTERS, MACHINE LEARNING TO AUTOMATICALLY IDENTIFY BRAIN TUMORS

Dr. George Bhiros and team at UT Austin using TACC supercomputers

Retooled Aurora Supercomputer Will Be America's First Exascale System

System will now be designed for simulation, Big Data and AI

Three Pillar View (DOE)

- | | | |
|---|--|--|
| <ul style="list-style-type: none">• Simulation• 64-bit floating point• Memory bandwidth• Random Memory access• Sparse Matrices• Scale limited by fabric• Low Latency, high BW• Distributed memory jobs• Output is Data | <ul style="list-style-type: none">• Big Data• 64-bit and integer• Analysis pipelines• Access to Data Bases• MapReduce/Spark• Fabric less important• Large data in and out• Millions of jobs• Output is Data | <ul style="list-style-type: none">• Deep Learning• 16-bit floating point• FMAC enhanced 16-bit• Fast local store (NVMe)• Frameworks (Tensorflow)• Fabric is an open question• Reuse of training data• Training dominates• Output is model+weights |
|---|--|--|



Smart Cities: One Example of Bringing it All Together

- Smart cities efforts not focus mostly on data analytics of sensor data (IoT)
- Smart cities efforts in the future will include very large scale data analytics, AI, and simulation
- E.g. smart transportation
 - Smart traffic level 1: change signals based on current traffic
 - Smart traffic level 2: use historical data for day, time, etc. to adjust for patterns
 - Smart traffic level 3: include simulations of events, weather, etc. to further refine signal optimization
- Smart cities in the future will use predictive analytics **and** predictive simulations to optimize traffic, safety, energy and water utilization, and more

28 of 8



The Future of AI and Machine Learning

Neuromorphic Computing

One example:

- spintronic oscillators
- 10 nanometers, 1000x denser than human brains
- 1000x faster than human brains
- Very low power (compared to today's CPU and GPU)
- One oscillators used to recognize human speech!

The screenshot shows the Nature journal website interface. At the top, the 'nature' logo is displayed with the tagline 'International weekly journal of science'. A search bar is located in the top right corner. Below the logo, a navigation menu includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. The main content area features a breadcrumb trail: Archive > Volume 547 > Issue 7664 > Letters > Article. The article title is 'Neuromorphic computing with nanoscale spintronic oscillators', categorized as a 'LETTER'. The authors listed are Jacob Torrejon, Mathieu Riou, Flavio Abreu Araujo, Sumito Tsunegi, Guru Khalsa, Damien Querlioz, Paolo Bortolotti, Vincent Cros, Kay Yakushiji, Akio Fukushima, Hitoshi Kubota, Shinji Yuasa, Mark D. Stiles & Julie Grollier. The article is dated 27 July 2017. On the right side, there is an 'Editor's summary' section with a 'العربية' (Arabic) language option, followed by 'Associated links' and 'Related audio' which includes a player for Julie Grollier's explanation.

Science is Understanding, not Computing Trends

- Always have to match observations (nature, reality)
- Large-scale problems will always require more performance, accuracy: more powerful systems, more data, etc.
- Complex problems will often require mix of simulation and data analytics, increasingly including AI
- Science/research is about understanding
- ***The techniques and technologies we use to increase our understanding of complex problems are themselves an area of research as well...***

D~~E~~LL EMC