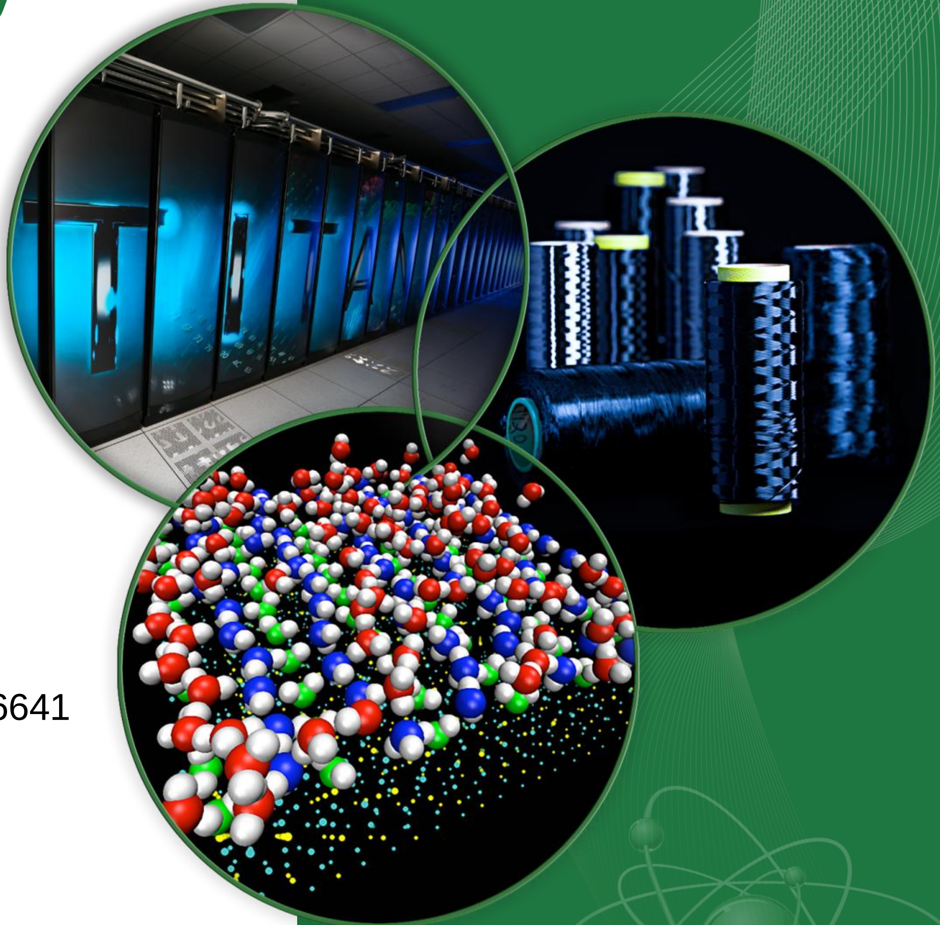
Machine Learning for Data-Driven Discovery

Thoughts on the Past, Present and Future



Sreenivas Rangan Sukumar, PhD

Data Scientist and R&D Staff
Computaional Data Analytics Group
Computational Sciences and Engineering Division
Oak Ridge National Laboratory

Email: sukumarsr@ornl.gov Phone: 865-241-6641

Tomorrow: Experience with Data Parallel Frameworks

Food-for-thought towards the exascale data analysis supercomputer

ORNL is managed by UT-Battelle
for the US Department of Energy

Today's Outline

- Scalable Machine Learning
  – Recent Advances and Trends
- State of the Practice
  – Philosophy, Engineering, Process, Paradigms
- Are we there yet ?
  – If yes, how so ?
  – If not, why not ?
- Concluding Future Thoughts
- Offline Debate and Discussion

Machine Learning

Given examples of a function *(x, f(x))*, **Predict** function *f(x)* for new examples *x*

Columns

Rows

	f_1	f_2	…	…	…	f_d	Labels
1							c_1
2							c_1
3							c_3
.							.
.							.
.							.
.							c_2
.							c_k
N							c_k

Machine Learning in the Big Data Era

Just in case you missed….

	1990 – 2000s	2010-Present	Insight
Assumption	A model exists. Better data will reveal the beautiful model. (Knowing “why” is important)	A model may not exist, but find a model anyway. (“Why” is not as important)	Dilemma: Better data or better algorithms.
Complexity of data	N ~ $\mathbf{O}(10^2)$, d ~ $\mathbf{O}(10^1)$ (e.g. IRIS data) k ~ $\mathbf{O}(1)$	N ~ $\mathbf{O}(10^6)$ d ~ $\mathbf{O}(10^4)$ (e.g. ImageNet) k ~ $\mathbf{O}(10^4)$	Volume, Velocity, Variety and Veracity have all increased several orders of magnitude.
Data – Model Relationship	Model abstracts data $\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$	Data is the model $f(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h_i} G\left(\frac{x - x_i}{h_i}\right)$	Models aggregated data. It is not anymore about the average. It is about every individual data point.
Model Parameter Complexity (e.g. Size of Neural Network)	$\mathbf{O}(10^3)$	$\mathbf{O}(10^{10})$ $\mathbf{O}(10^8)$ to $\mathbf{O}(10^{10})$ in months.	10-billion parameter network learned to recognize cats from videos.
Accuracy, Precision, Recall **e.g. Face Recognition** **Visual Scene Recognition**	~ 70% was accepted Not possible	~95% is the norm ~10% is the best result to date.	Big Data also means Big Expectations.
Computing Capability **Personal Computing** **High Performance Computing**	1 core, 256MB RAM, 8GB disk 1000 cores, 1 teraflops	16 cores,64 GB RAM,2TB disk 3 million cores, 34 petaflops	Commercial tools are keeping pace with the PC market and not HPC market.
Number of Dwarves !	**7**	**13**	**Big Data Magic: Dwarves are doubling.**

Today's Talk

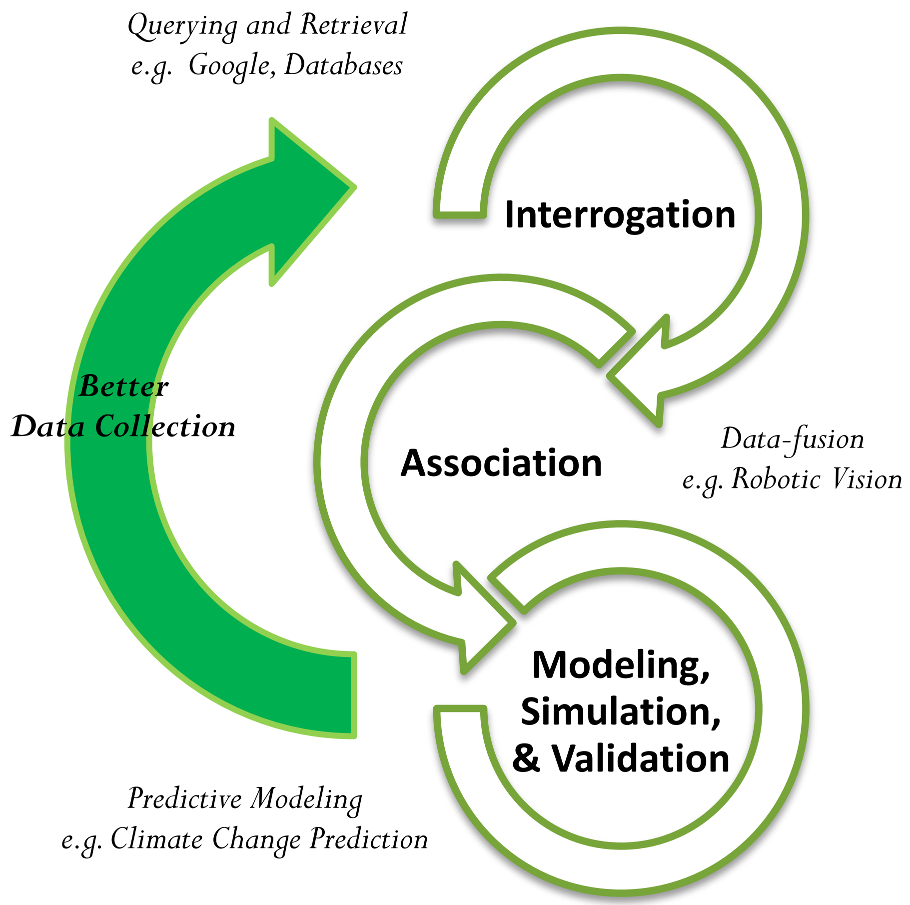
'Compute' is scaling up commensurate the 'data'. Is machine learning keeping pace with the data and compute scale-up ?

- If Yes : How so ?
- If Not : Why not ?

Oak Ridge National Laboratory

Scalable Machine Learning: Philosophy

The Lifecycle of Data-Intensive Discovery

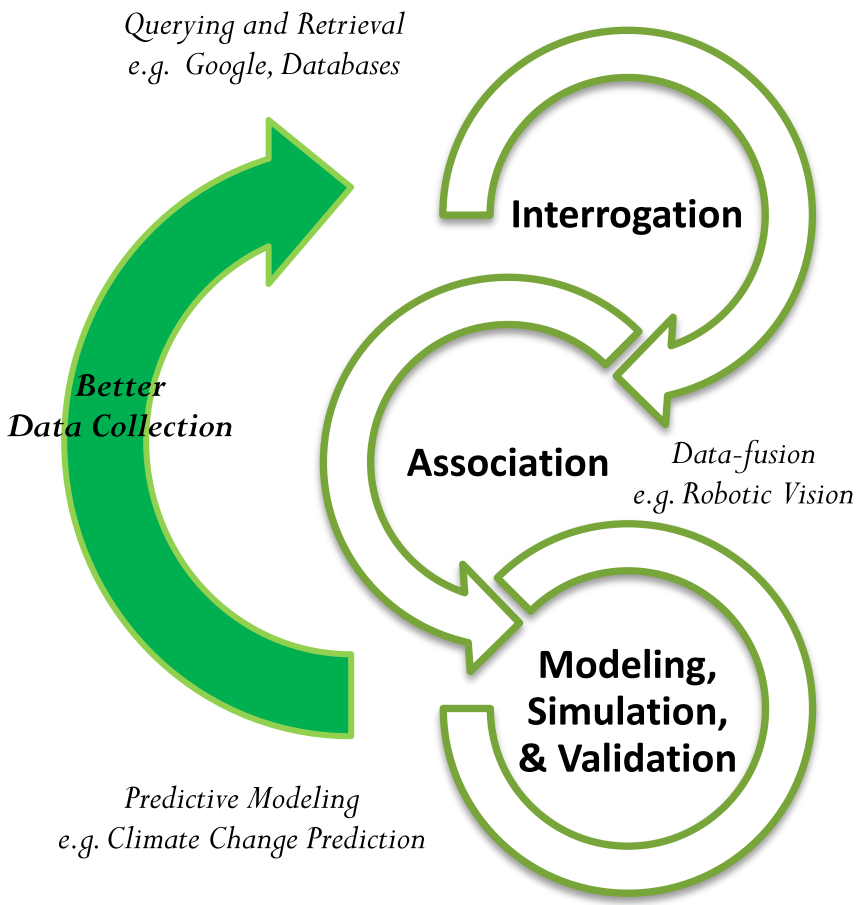


Off-the-shelf Parallel Hardware

- **Custom ICs**
  - e.g. FPGAs, Adapteva, Rasberry Pi)
- **Customized Processing**
  - E.g. Nvidia GPGPUs, YarcData Urika
- **Multi-core HPC**
  - e.g. (Cray XK, Cray XC, IBM Blue Gene)
- **Virtual clusters / Cloud computing**
  - e.g. Amazon AWS, SAS (PaaS, + SaaS)

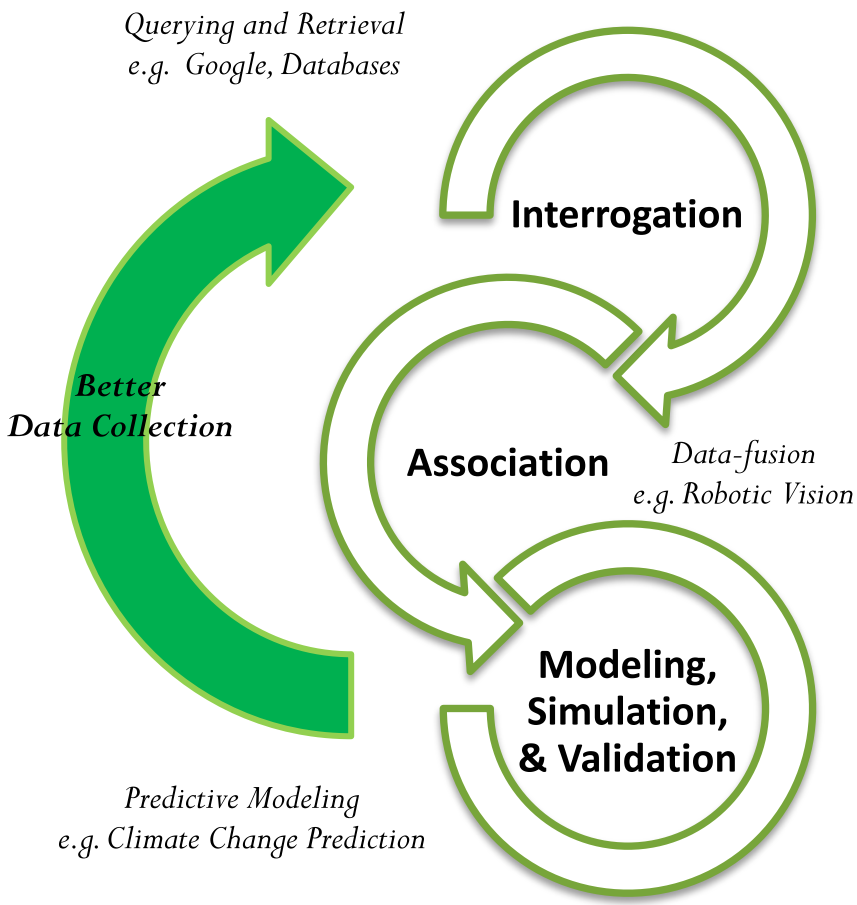
Scalable Machine Learning: Philosophy

The Lifecycle of Data-Intensive Discovery

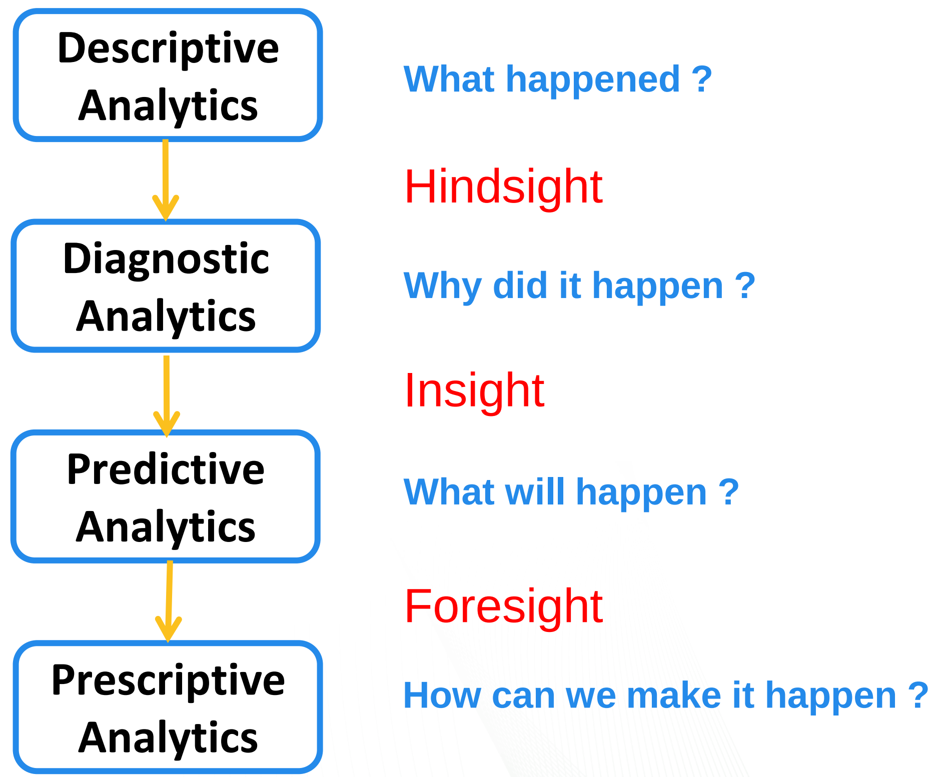


Scalable Machine Learning: Discovery Process

The Lifecycle of Data-Intensive Discovery



Data-Driven Discovery Process



Concept adapted from Gartner's Webinar on Big Data

OAK RIDGE National Laboratory

Scalable Machine Learning: System Engineering

Data-Driven Discovery Process

Descriptive Analytics — What happened ?

Hindsight

Diagnostic Analytics — Why did it happen ?

Insight

Predictive Analytics — What will happen ?

Foresight

Prescriptive Analytics — How can we make it happen ?

Concept adapted from Gartner's Webinar on Big Data

- **Staging for Predictive Modeling**
  - Extract, Transform, Load
  - Data Pre-processing
  - Feature Engineering
- **Predictive Modeling**
  - Rule-base extraction
  - Pairwise-similarity (Distance Computation)
  - Model-parameter estimation
  - Cross validation
- **Inference/ Model Deployment**
  - Data is model ? Model is data ?
  - Adaptive model ? Reinforcement ?

Scalable Machine Learning: Production

- **Staging for Predictive Modeling**
  - Extract, Transform, Load
  - Data Pre-processing
  - Feature Engineering

Disk Intensive

File processing and repeated retrieval best done in massively parallel file systems or databases

- **Predictive Modeling**
  - Rule-base extraction
  - Pairwise-similarity (Distance Computation)
  - Model-parameter estimation

Disk, Memory and Compute Intensive

Typically computing an aggregate measure, vector product, a kernel function etc.

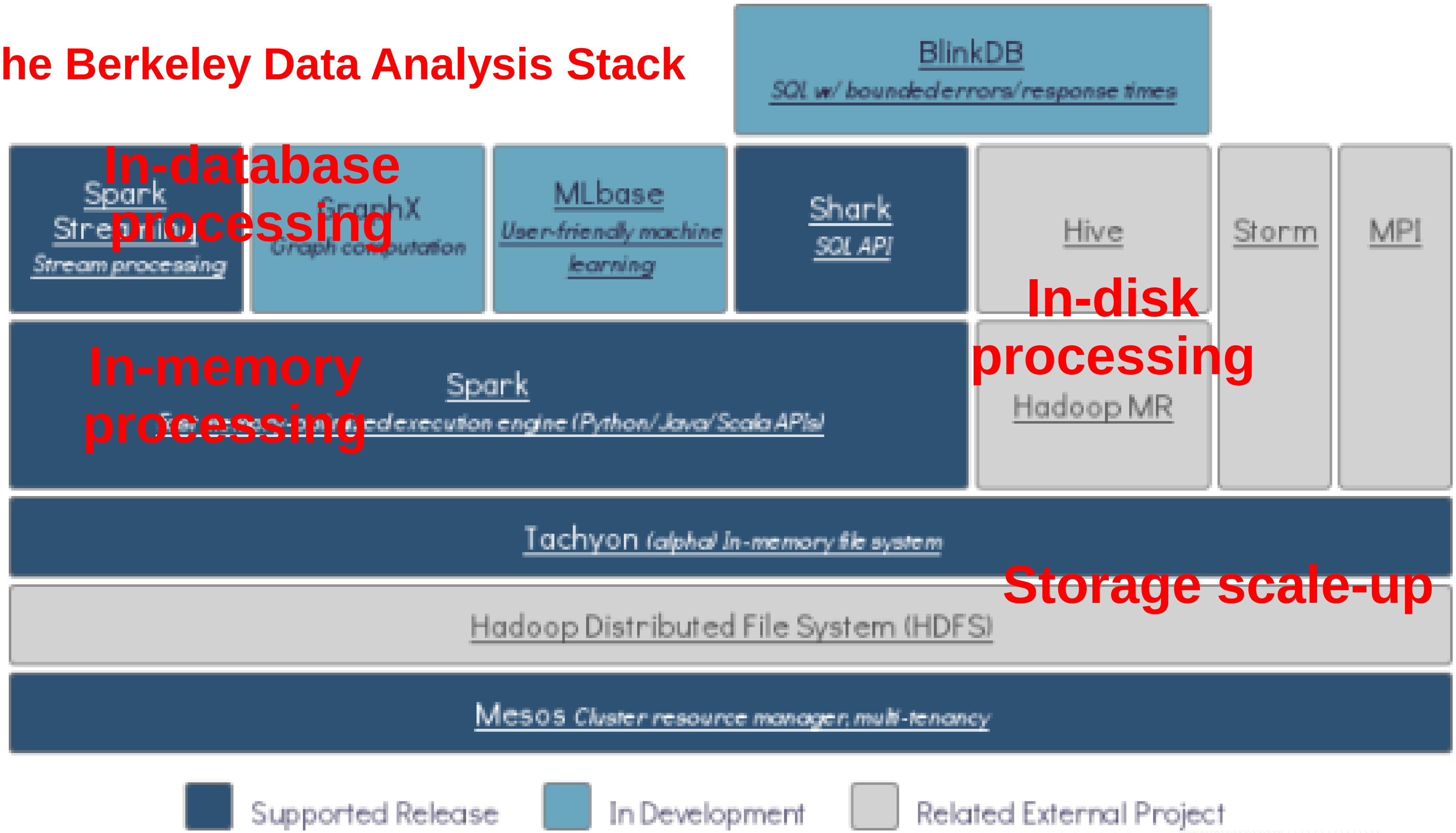
- **Inference/ Model Deployment**
  - Cross-validation
  - Data is model ? Model is data ?
  - Adaptive model ? Reinforcement ?

Memory + Compute Intensive

Real-time requirements

Scalable Machine Learning: Bleeding Edge

The Berkeley Data Analysis Stack



This is tremendous progress….

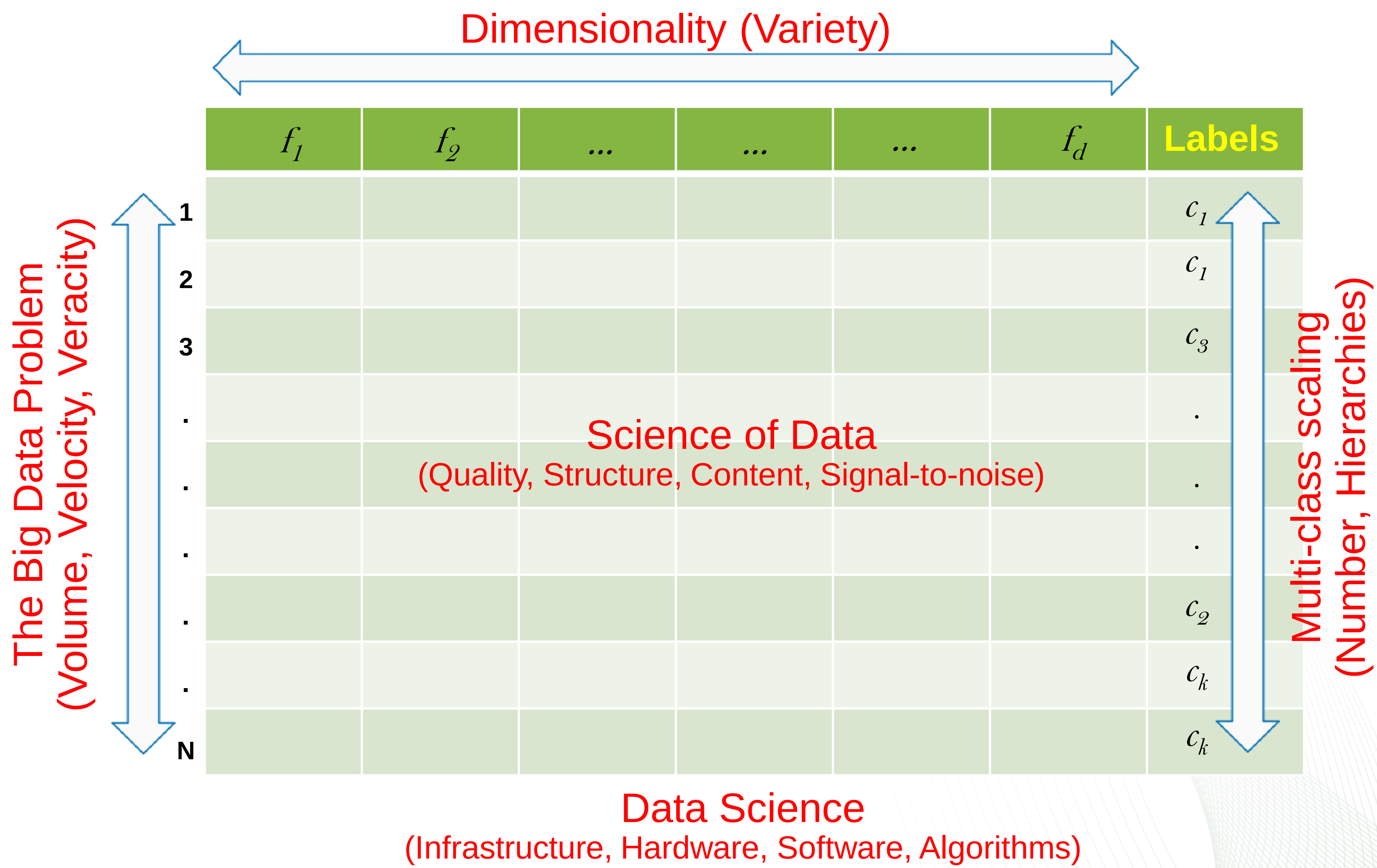
But...

Is machine learning keeping pace with the data and compute scale-up ?

- If Yes : How so ?
- If Not : Why not ?

The 5 Challenges of Scalable Machine Learning

Given examples of a function *(x, f(x))*, **Predict** function *f(x)* for new examples *x*



Challenge #1: Data Science

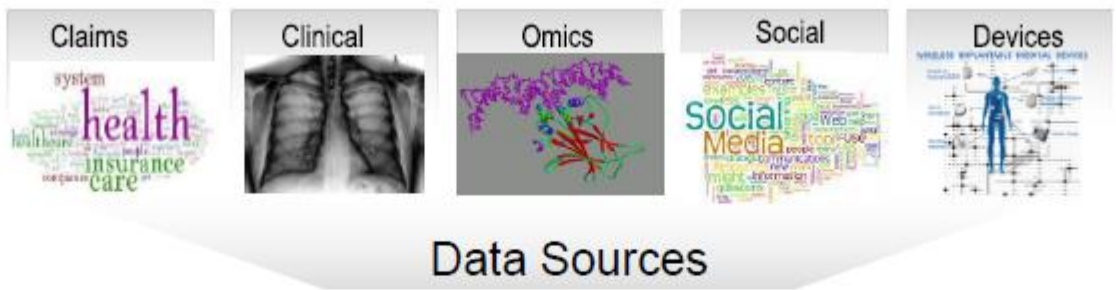
Systems	Data		Compute		Analysis	
Infrastructure	**Management**	**Structure**	**HPC**	**Programming**	**Algorithms**	**Viz**
Design Operations Management	• Quality • Privacy • Provenance • Governance	• Matrix/ Table • Text, Image, Video • Graphs • Sequences • Spatiotemporal • Schema	• TITAN • CADES • Cloud • Urika • Hadoop	• OpenMP/MPI • CUDA/ OpenML • RDF/SPARQL • SQL • Map-Reduce	• In-database • In-memory • In-situ	• HCI • Interfaces • Viz-Analytics
Architecture Design Operations					**Theory**	
Databases SQL NoSQL Graph					• Design • Scalability • V&V	

- Performance of "algorithm" dependent on architecture.
  - Most data scientists/algorithm specialists are used to in-memory tools such as R, MATLAB etc.
  - Existing cloud-based solutions are designed for high performance storage and not high-performance compute or in-memory operations.
  - Steep learning curve towards programming "new" innovative algorithms. Too many options without guiding benchmarks.

Challenge #2: Science of Data

- Data-science is not the same as “science of data”
  - Is the process of understanding characteristics of data before applying/designing a machine-learning algorithm.

SNR > 1 SNR >>> 1 SNR << 1 SNR >> 1



  - Data characterization – (Avoid using machine learning as a black box)
    - Signal-noise-ratio , bound on noise
    - i.i.d sampling assumptions
    - stationarity, randomness, ergodicity, periodicity
    - Generating models behind data

Challenge #3: The N-d-k problem

- The Big Data Problem
  - The future is unstructured.
    - Text, images, videos, sequences
- Algorithms and infrastructure expected to handle Big Data – i.e., **increasing N, d and k.**
  - Feature engineering and requires automation.
    - Self-feature extracting methodologies encouraged.
  - Traditional (pain staking) pipeline of SMEs creating features from the data will fail or transform into a collaborative-parallel effort.
  - Increasing *N* does not imply increasing information content. (Samples can still be good if not better than all of the data statistically.)
  - There can be hierarchies within the N-d-k dimensions.

Challenge #4: The N-d-k problem (d)

- Traditional algorithms assume N >> d and d > k
  - Most tools available today scale well for increasing N.

	single	multi
LWLR	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
LR	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
NB	$O(mn+nc)$	$O(\frac{mn}{P}+nc\log(P))$
NN	$O(mn+nc)$	$O(\frac{mn}{P}+nc\log(P))$
GDA	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
PCA	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
ICA	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
k-means	$O(mnc)$	$O(\frac{mnc}{P}+mn\log(P))$
EM	$O(mn^2+n^3)$	$O(\frac{mn^2}{P}+\frac{n^3}{P'}+n^2\log(P))$
SVM	$O(m^2n)$	$O(\frac{m^2n}{P}+n\log(P))$

Time-complexity analysis

Data Sets	samples (m)	features (n)
Adult	30162	14
Helicopter Control	44170	21
Corel Image Features	68040	32
IPUMS Census	88443	61
Synthetic Time Series	100001	10
Census Income	199523	40
ACIP Sensor	229564	8
KDD Cup 99	494021	41
Forest Cover Type	581012	55
1990 US Census	2458285	68

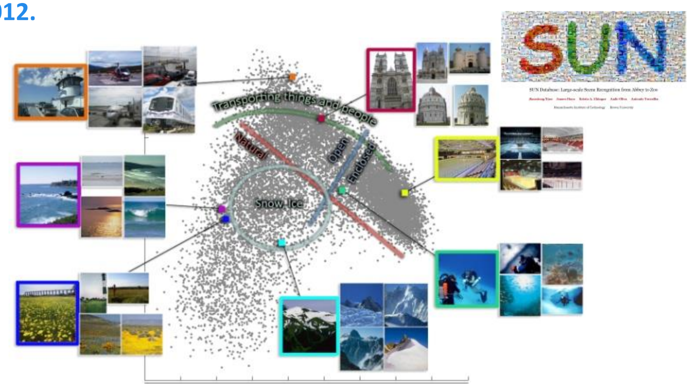
Data characteristics

[Chu et al., NIPS 2007]

  - Not so much for increasing *d* or *k*
    - [Donoho, 2000] – The curse and blessings of dimensionality
    - Methods are emerging : Multi-task learning, Spectral Hashing etc.

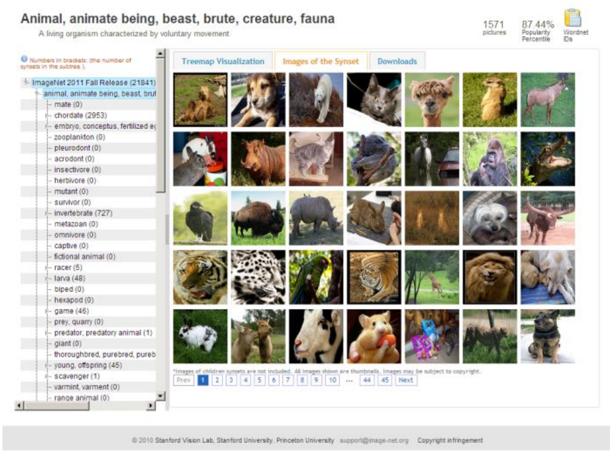
Challenge #5: The N-d-k problem (k)

Hays et al., "SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes", CVPR 2012.



Hasegawa et al., "Online Incremental Attribute-based Zero-shot Learning", CVPR 2012

Berg et al., "What Does Classifying More Than 10, 000 Image categories Tell Us?", ECCV 2010.



- **What happens when K= K + 1 ? (adding a new class)**
  - Engineered features may not be good enough.
  - Trained model has to relearn from the entire feature set without guarantees on accuracy.

OAK RIDGE National Laboratory

Concluding Thoughts

What can be scaled up ?

Compute

Storage
Memory
Cores

I/O ?
Network ?

What aspect of data that needs scale up ?

Dimensions of Big Data
Software

Volume
Hadoop, MPP, Spider
Archival | Reports | Discovery

Velocity
Streaming | Batch

Variety
SQL | NoSQL | Graph

What aspect of algorithm that needs scale up ?

Analytical Requirements
Algorithms

Programming — MapReduce, MPI, Threads
Data-Parallel | Task-parallel

Complexity of Algorithms
Linear | Iterative | > $O(N^2)$

Compute on Data
Retrieval | Machine Learning

Speed of Execution
Real-time | Feasibility

- Future
  - We need benchmarks before me make big investments. (Fox et al., 2014)

Concluding Thoughts

- Storage/Memory and Memory/Compute Ratios that are critical for machine learning are smaller than Storage/Compute Ratio.
- Associative memory and cognitively-inspired architectures may prove better than the Von-Neuman “store-fetch-execute paradigm”.
  - May be time to redesign from scratch.
- The machine learning algorithms that scale all use either data-parallelism or the “dwarves of parallel computing in some form”.
  - Encouraging because – gives us an intuition to build custom “hardware” for learning algorithms.
- We have done well so far by treating – “Analysis as a retrieval problem” – We can do better.

Thank You

- Questions ?