Big Data + Big Compute = An Extreme Scale Marriage for Smarter Science?

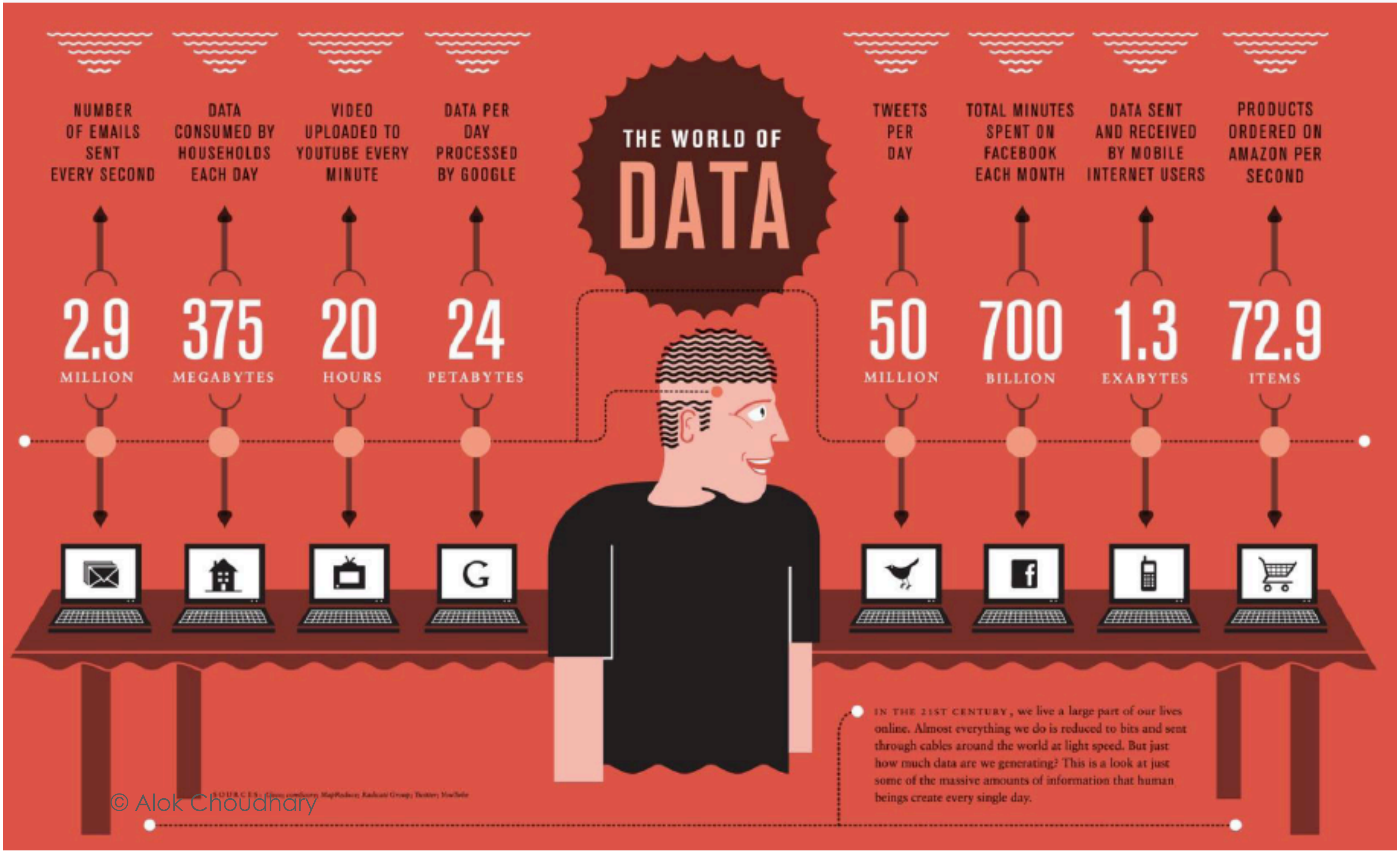
Alok Choudhary
Henry and Isabel Dever Professor
EECS and Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

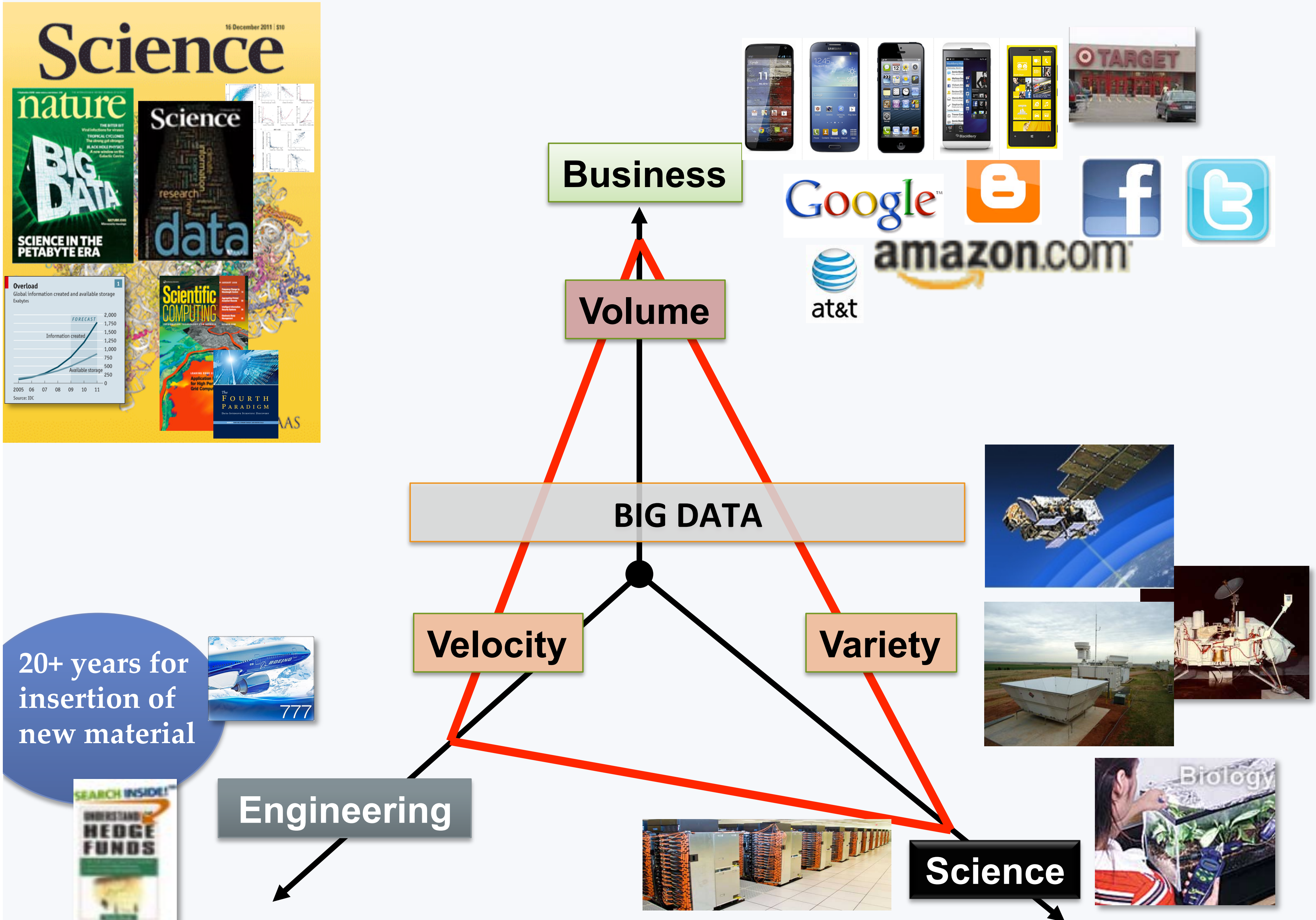
Founder and President
Voxsup Inc: A Big Data Science Company
+1 312 515 2562
alok@voxsupinc.com

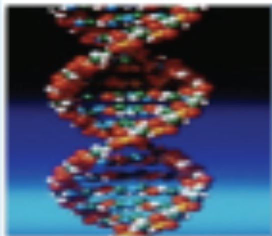
SC 2013 , November 21, 2013

National Science Foundation
WHERE DISCOVERIES BEGIN
ACKNOWLEDGEMENTS
U.S. DEPARTMENT OF ENERGY

Big Data …Popular View.. Streaming..

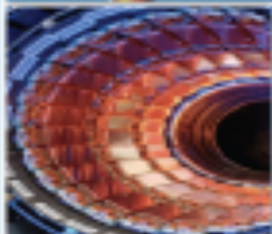



Science
nature
BIG DATA
SCIENCE IN THE PETABYTE ERA
Science
data
Overload
Scientific COMPUTING
The FOURTH PARADIGM
Business
Volume
BIG DATA
Velocity
Variety
Engineering
Science
20+ years for insertion of new material
777
HEDGE FUNDS
TARGET
Google
amazon.com
at&t
Biology



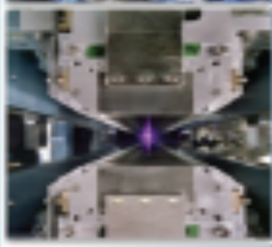
Genomics

Data Volume increases to 10 PB in FY21



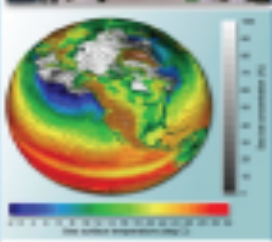
High Energy Physics (Large Hadron Collider)

15 PB of data/year



Light Sources

Approximately 300 TB/ day



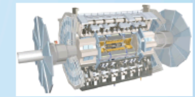
Climate

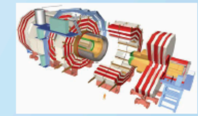
Data expected to be hundreds of 100 EB

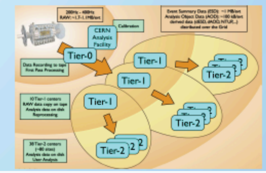
Source: Bill Harrod, SC12 plenary presentation

Data Challenges in High Energy Physics: Large Hadron Collider exemplar

- ATLAS and CMS detectors generate analog data at rates equivalent to 1PB/second
- Output rate after *data reduction* is 1GB/second ~ 10PB/year
- Storage of cumulative derived data, simulated data, replicated data is currently ~ 100PB, and is rapidly increasing
- Workflow: homogeneous community of physicists access read-only shared data using the Worldwide LHC Computing Grid

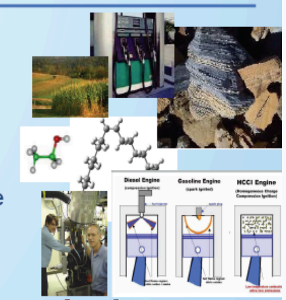


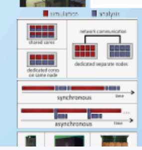




Data Challenges in Large-Scale Simulations: S3D Combustion code exemplar

- Goal: simulate turbulence-chemistry interaction at conditions that are representative of realistic systems
  - High pressure
  - Turbulence intensity
  - Turbulent length scales
  - Sufficient chemical fidelity to differentiate effects of fuels
- Exascale simulation will require 3PB of memory, and will generate 400PB of raw data (1PB every 30 minutes)
- Workflow challenges include co-design for simulation and in-situ analyses





http://science.energy.gov/~/media/ascr/ascac/pdf/reports/2013/ASCAC_Data_Intensive_Computing_report_final.pdf

Thinking about BIG DATA?

• • •

Wikipedia Definition; "Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications." ☺

Many think big data processing is..

Drinking from a Firehose..

To quench the thirst..

"Data intensive" vs "Data Driven"

Data Intensive (DI)

- Perspective Driven
  - Processor, memory, application, storage?
- An application can be data intensive without being I/O intensive

Data Driven (DD)

- (Big) Data Analytics
  - Top-down query
  - Bottom up discovery (unpredictable TTR)
  - Predictive modeling
- Usage model differences

DD is Not only about "What you Know", It is ALSO about "What else you may know"... and faster

The Engagement?

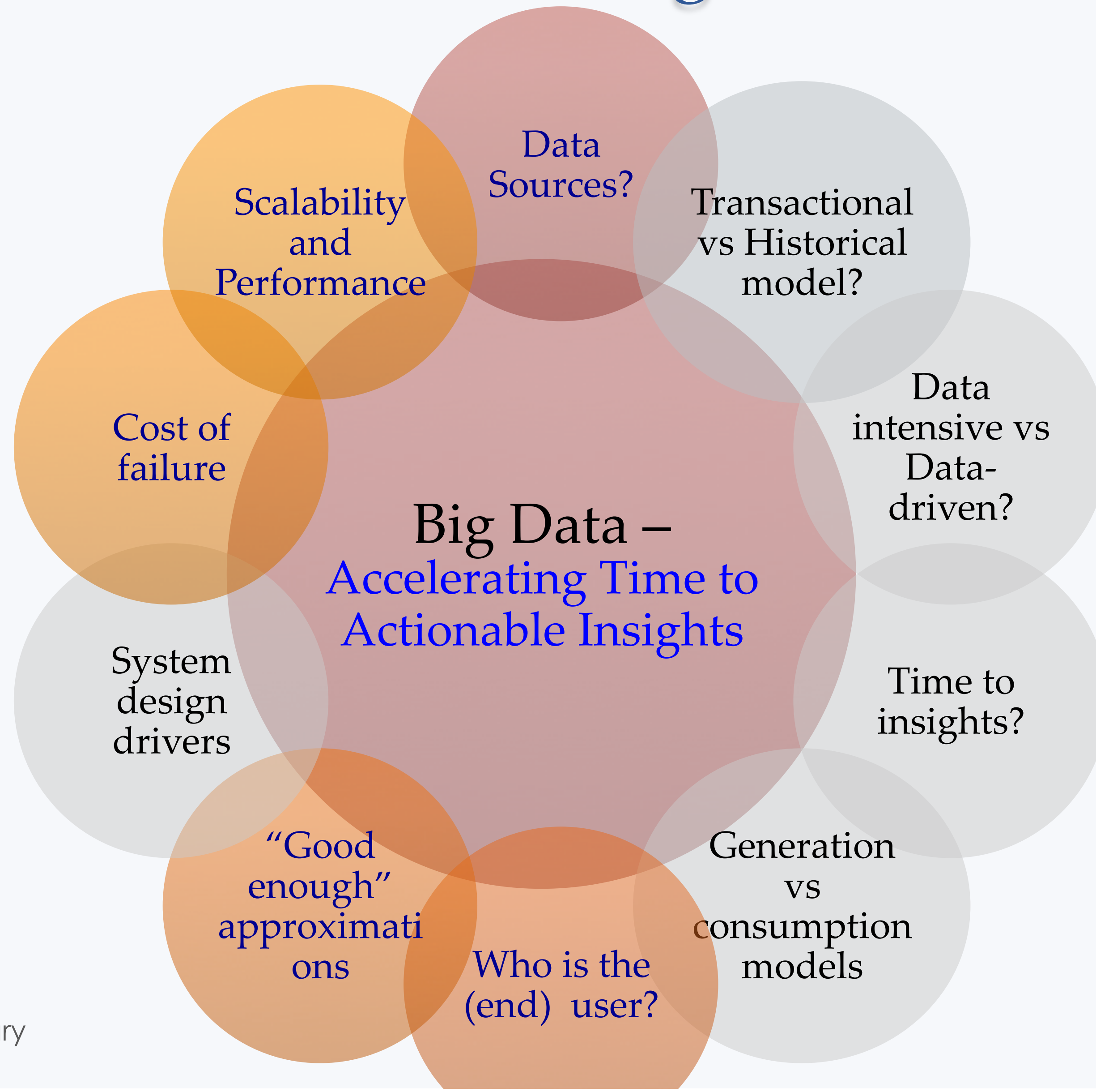
enables

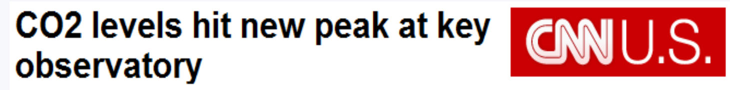
Data Intensive Techniques in Big Compute

Data Driven Computing at Scale

HW/SW design feedback

…True Promise - Accelerating Time to Actionable Insights

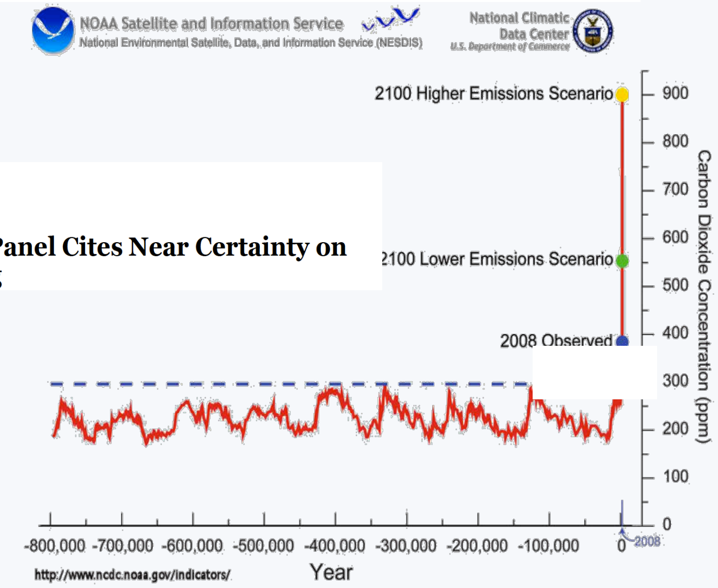



CO2 levels hit new peak at key observatory

The New York Times

August 19, 2013

Climate Panel Cites Near Certainty on Warming

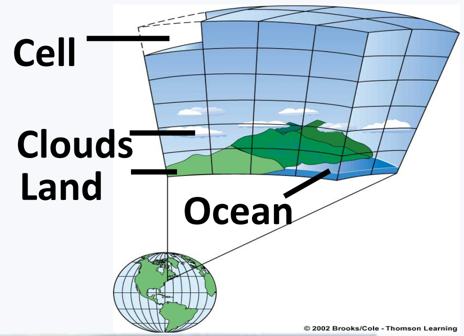


Understanding Climate Change Exemplar

A Case for Big Compute + Big Data Science

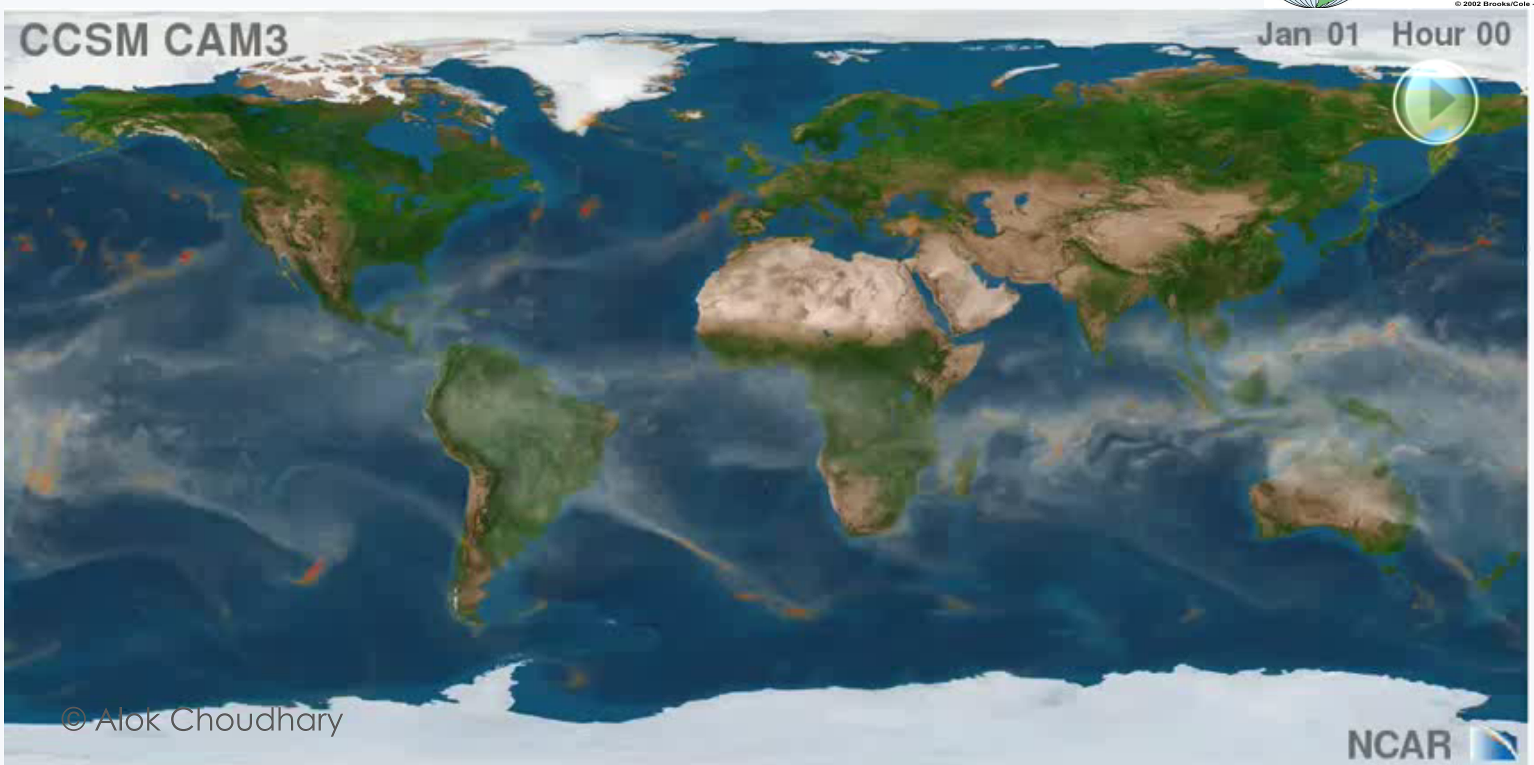
Understanding Climate Change – DI - Physics-Based Approach (Simulation → Data Generator)

General Circulation Models: Mathematical models with physical equations based on fluid dynamics

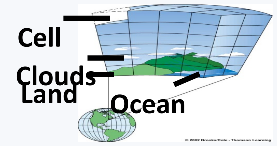


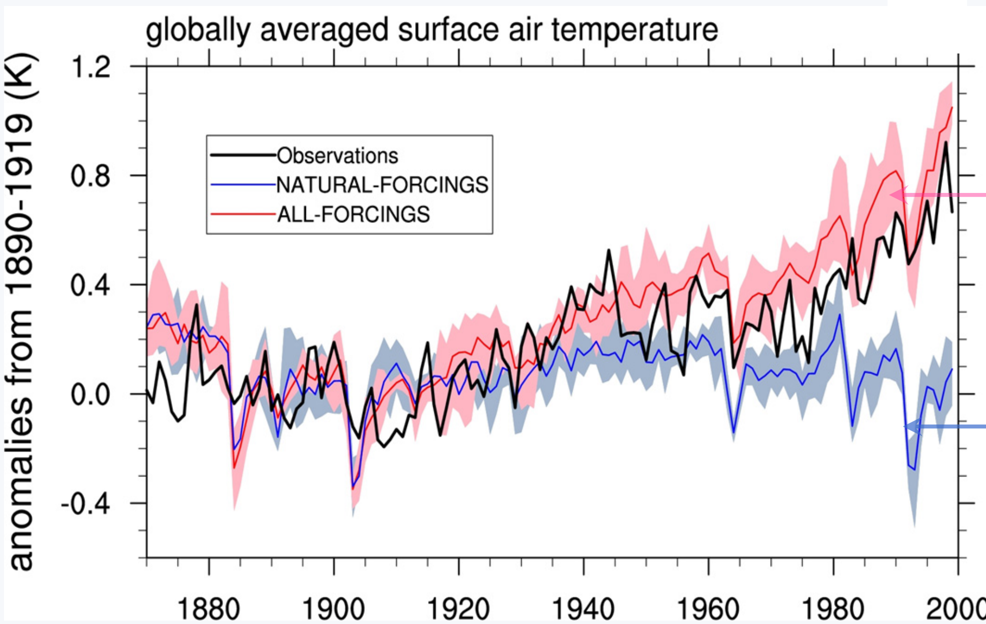
Parameterization and non-linearity of differential equations are sources for uncertainty!

Figure Courtesy: NCAR



Understanding Climate Change – (Simulation) Physics Based Approach…





Ensemble average with observed greenhouse gas concentrations

Ensemble average with pre-industrial greenhouse gas concentrations

Figure Courtesy: ORNL

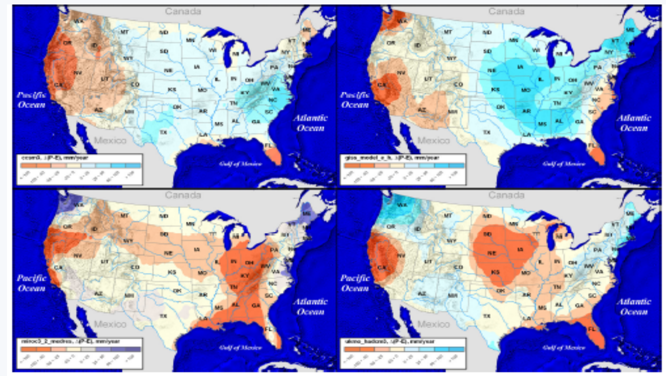
Simulation + data-driven science ☺

Physics based models are essential but Limited

- Relatively reliable predictions at global scale for ancillary variables such as temperature
- Least reliable predictions for variables that are crucial for impact assessment such as regional precipitation

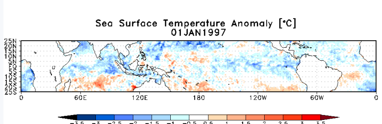
"The sad truth of climate science is that the most crucial information is the least reliable"
(Nature, 2010)

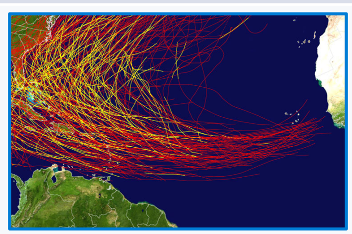
Disagreement between IPCC models

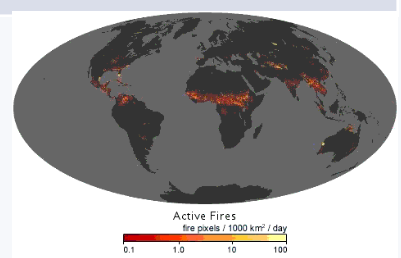


Regional hydrology exhibits large variations among major IPCC model projections

Low uncertainty	High uncertainty	Out of scope
Temperature	Hurricanes	Fires
Pressure	Extremes	Malaria outbreaks
Large-scale wind	Precipitation	Landslides

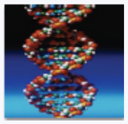


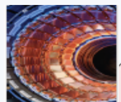


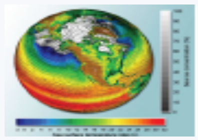


Data Driven Science – Operational to Strategic

Instruments, sensors



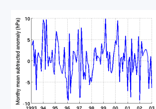




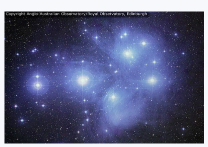
supercomputers

Transactional: Data Generation

Historical: Data Processing, transformation, approximation



Data Mining, analytics, unsupervised learning



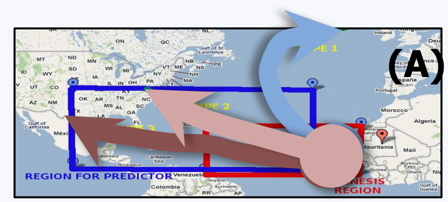
Discovery, Insights, Feedback

Data Management

Data Reduction, Query

Data Visualization

Data Sharing



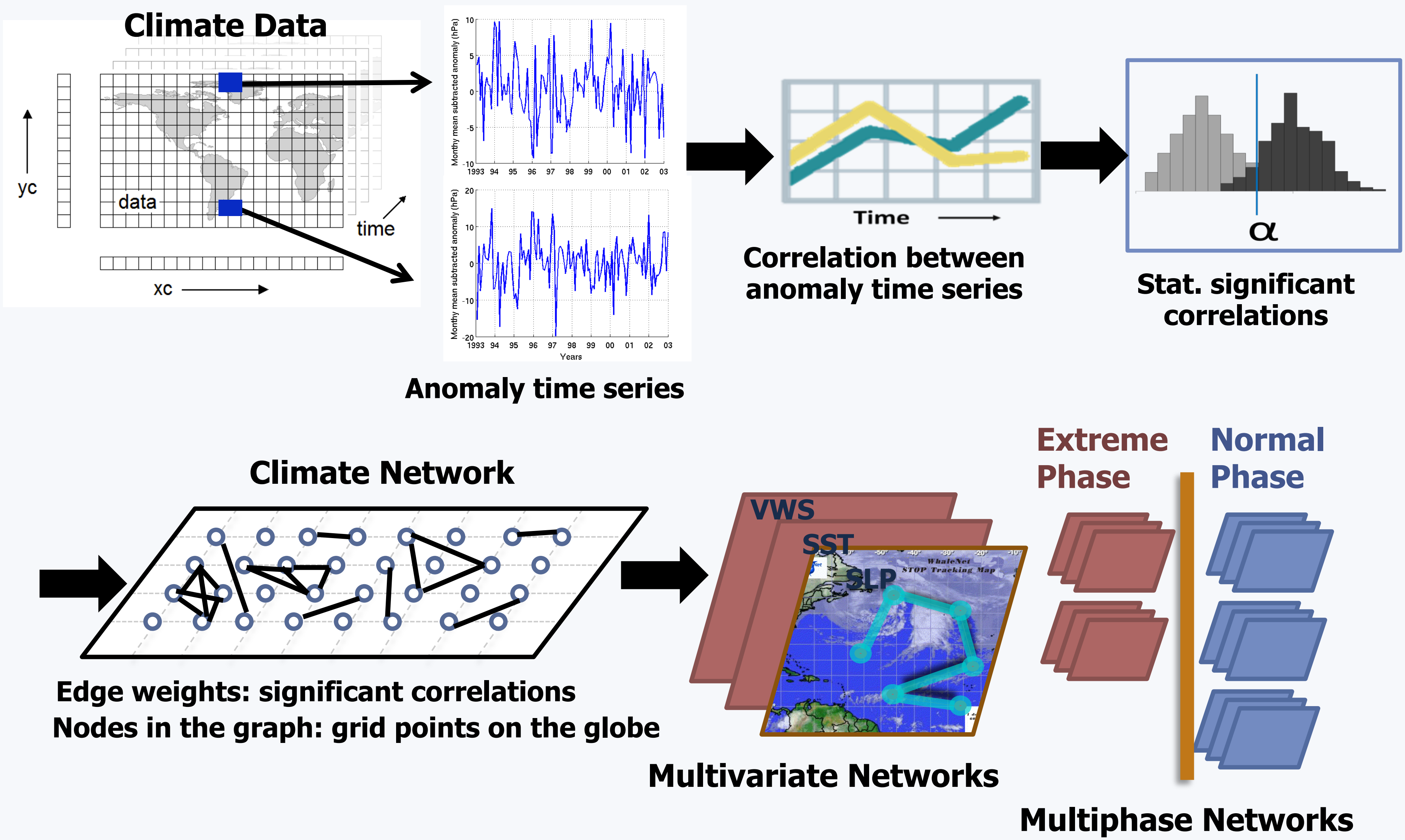
Historical data

Learning Models

Trigger/ questions

Predict

Transactional analytics to Data- Driven Science



Data Driven Science : Thinking about Analytics?

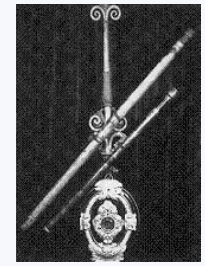
. . .

- Makes use of wealth of historical observational and simulation data
- Accelerate Time-to-Discovery and Actionable Insights

Requires Understanding Analytics Algorithms and SW

The Unknown

As we know,
There are known knowns.
There are things we know we know.



Conventional Wisdom

- High Humidity results in outbreak of Meningitis
- Customers switch carriers when contract is over

Validate Hypothesis

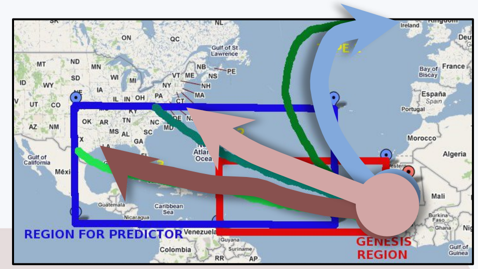
- Nuclear Reaction happens under these conditions
- Did combustion occur at the expected parameter values

e.g., Statistics, Query, Transformation, Viz

The Unknown

As we know,
There are known knowns.
There are things we know we know.

We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.



Top-Down Discovery - We know the question to ask

- Will this hurricane strike the Atlantic coast?
- What is the likelihood of this patient to develop cancer
- Will this customer buy a new smart phone?

Predictive Modeling…; e.g., SVM, Decision Trees

The Unknown

As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.

But there are also unknown unknowns,
The ones we don't know
We don't know.

Bottom up Discovery - We don’t know the question to ask

- Wow! I found a new galaxy?
- Switch C fails when switch A fails followed by switch B failing
- On Thursday people buy beer and diaper together.
- The ratio K/P > X is an indicator of onset of diabetes.

Relationship Mining, Clustering etc.. - ARM

The Unknown Unknown

Strong Affinity

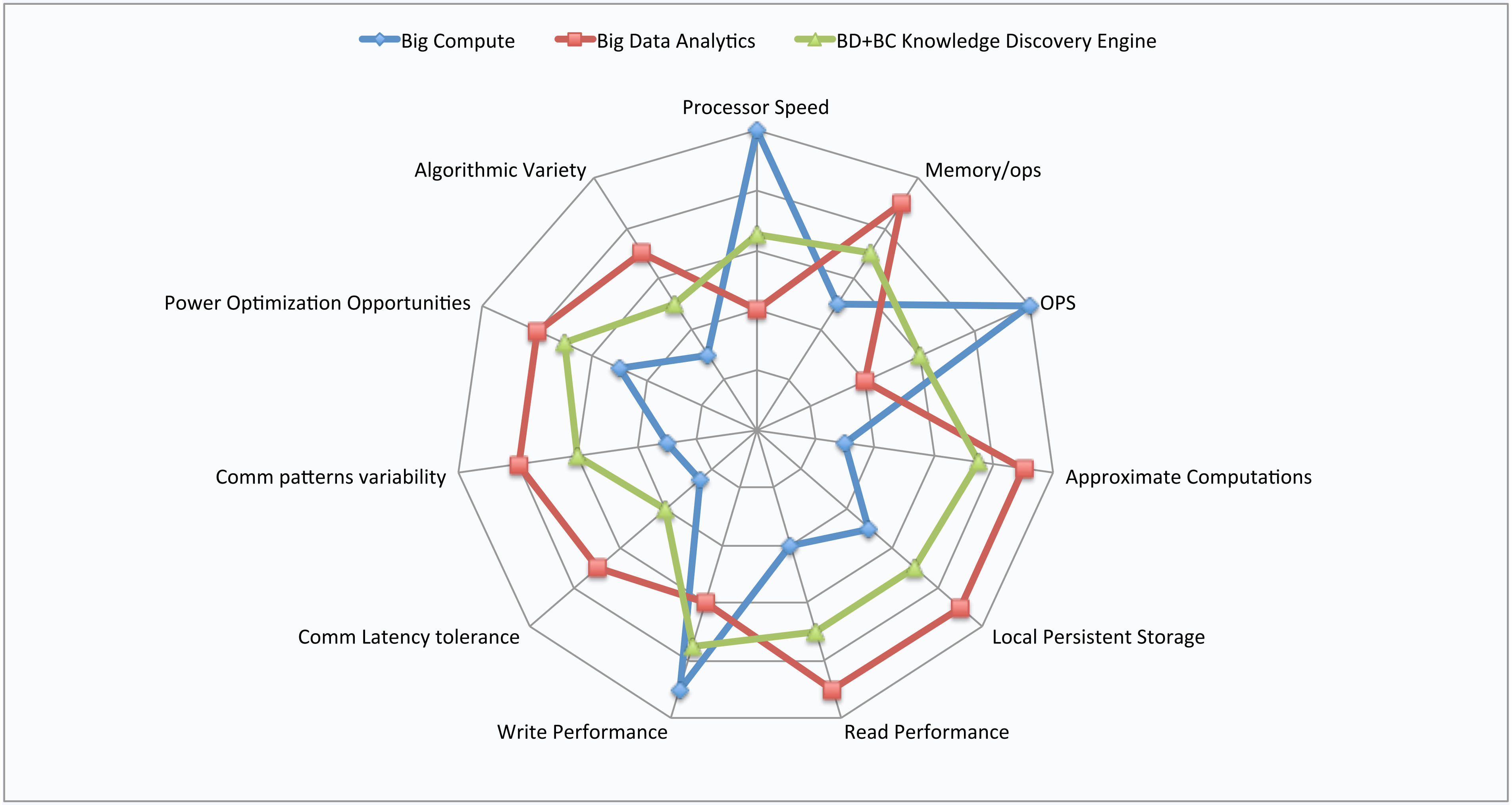
What Else you may find!

Big Compute + Big Data

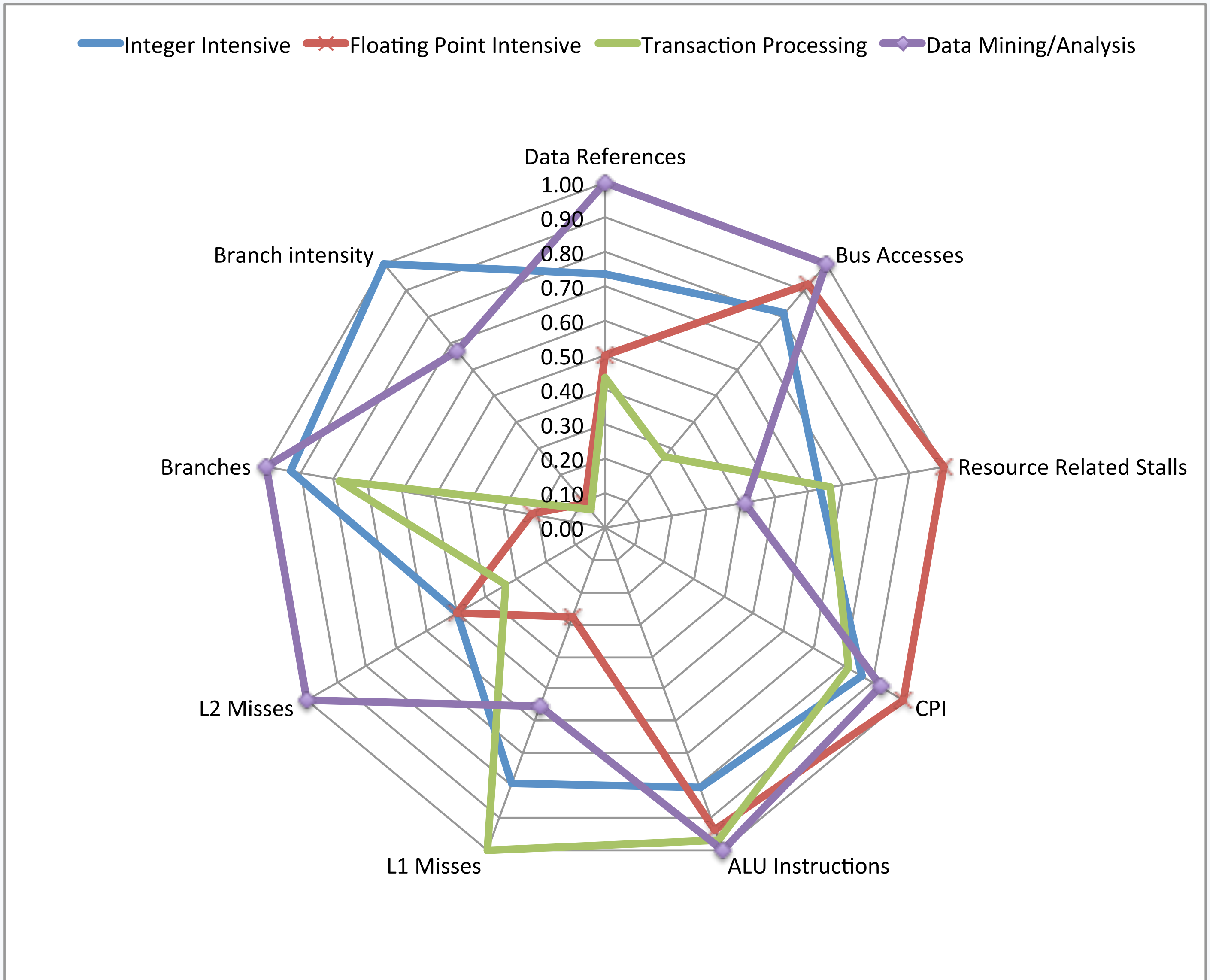
The HW/SW Design Goals?

Big Compute	Big Data
Time to Compute	Time to Insight
Speed of Data Output	Speed of data Ingestion
(Typically) Model Driven	(Typically) Data-Driven
End Consumer – (Typically designer of algorithms and SW (scientist)	End consumer != Designer of Algorithms or scientist
Performance Metrics – FLOPS	Performance Metric – Many
(Mostly) Latency Intolerance	(Mostly) Latency Tolerant
Fault-tolerance important?	Fault-tolerance : central
Top-Down Design	Bottom-up Design

Big Compute + Big Data Analytics = A Knowledge Discovery Engine?



Computation Characteristics



Extreme-scale System: An instrument and a discovery engine

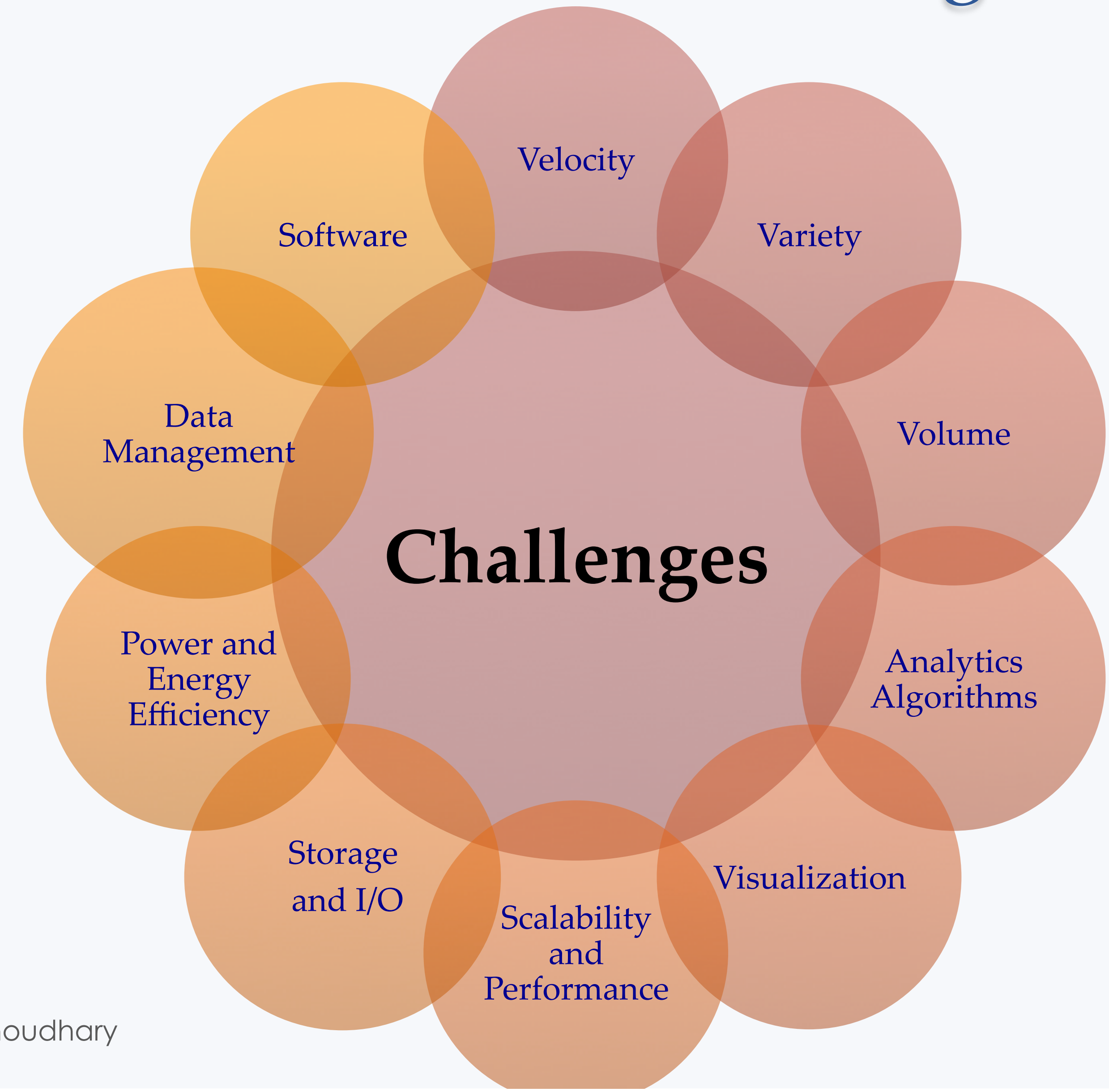
Millions of cores

Each core is a data generator

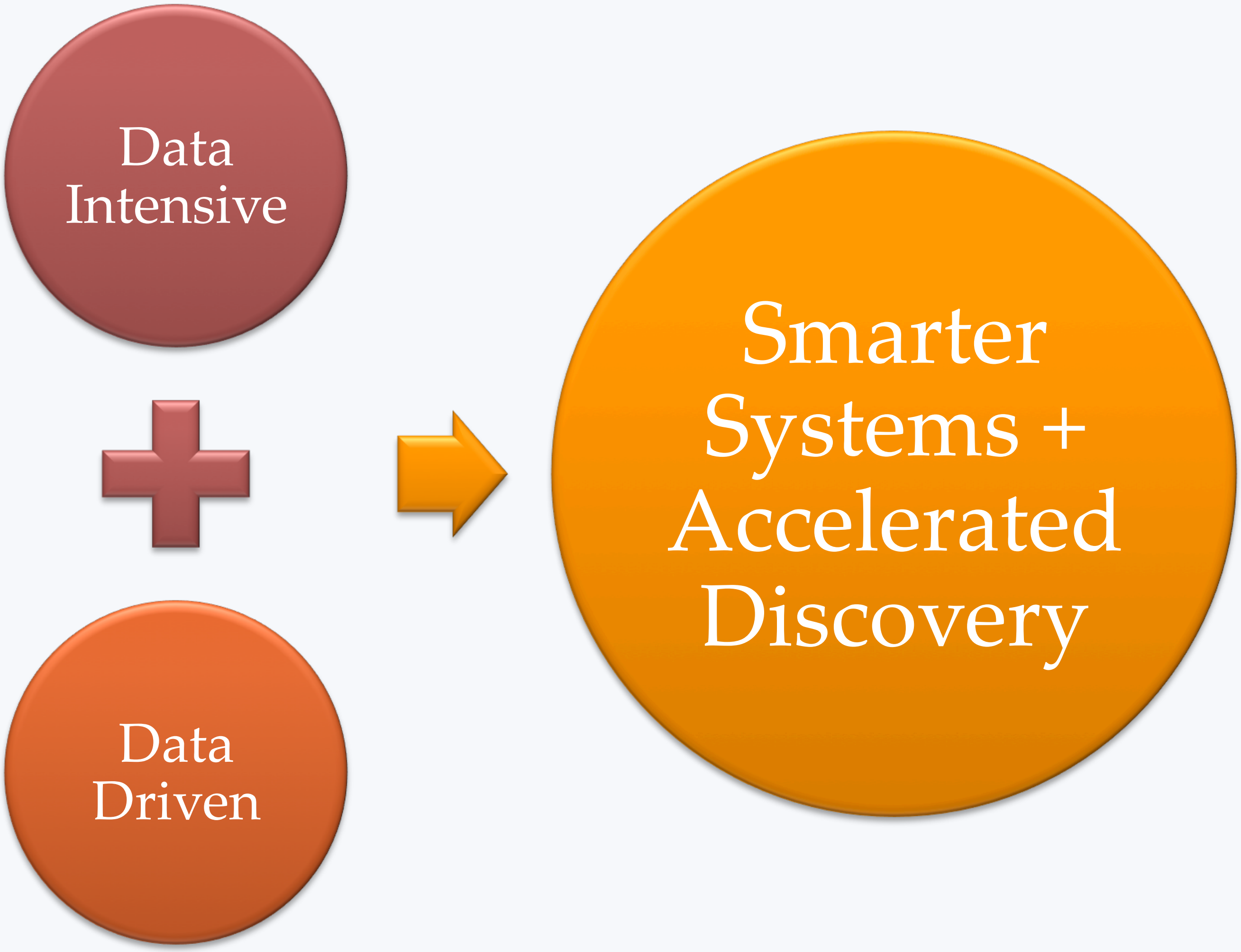
…A core is a data processor/analyst

Extreme scale system is a discovery engine

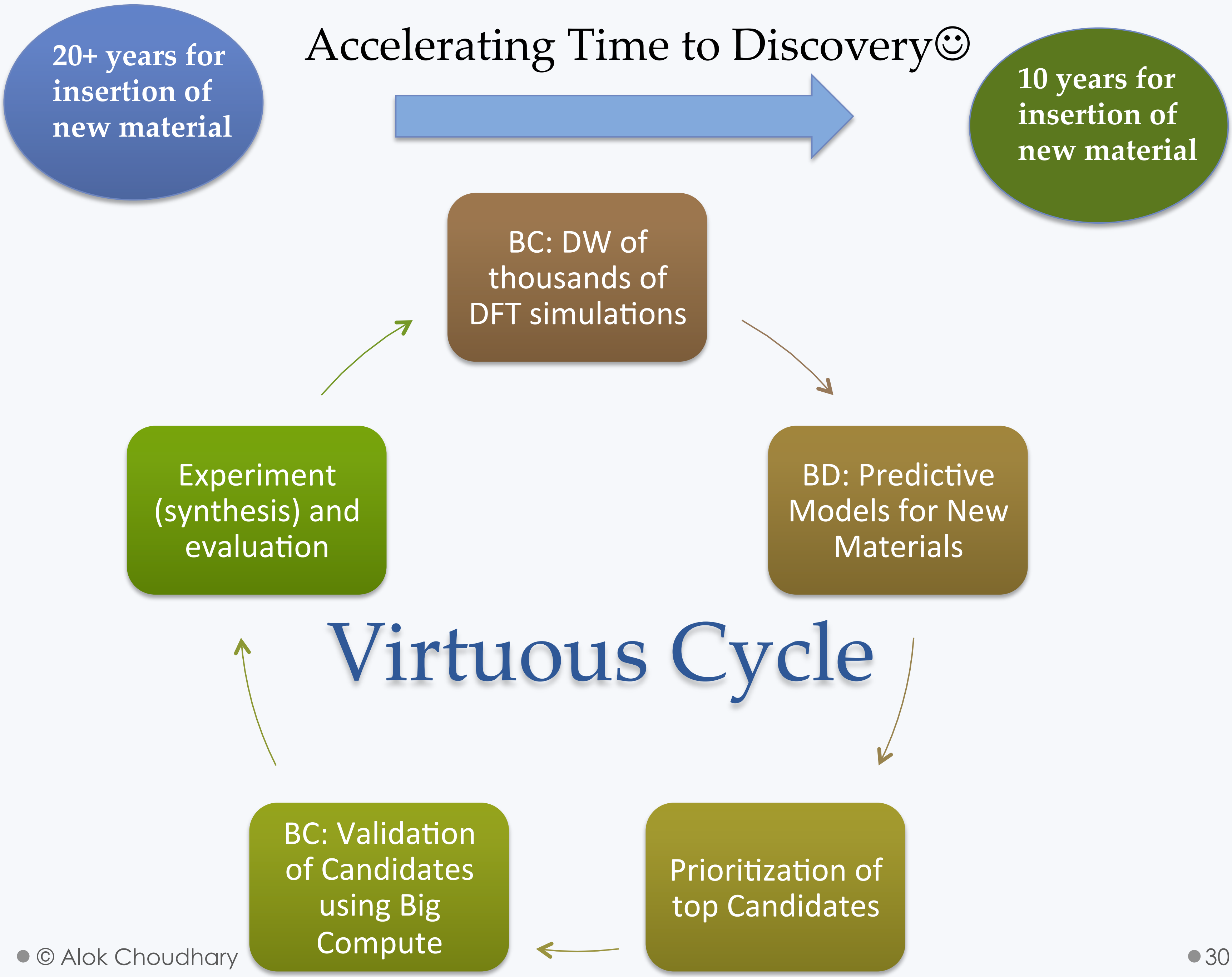
Big Compute + Big Data : Not a single dimensional challenge



Big Data + Big Compute Strategy



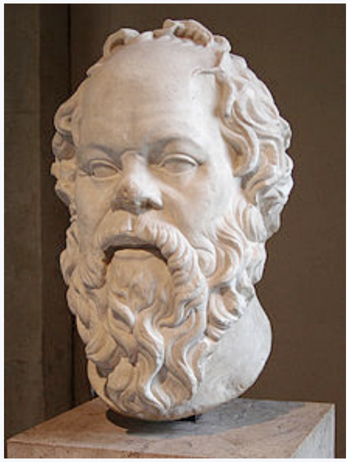
Accelerating Time to Discovery☺



Who Knew?

The Unknown
As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know.

—Feb. 12, 2002, Department of Defense news briefing by Donald Rumsfeld



469-399 BC

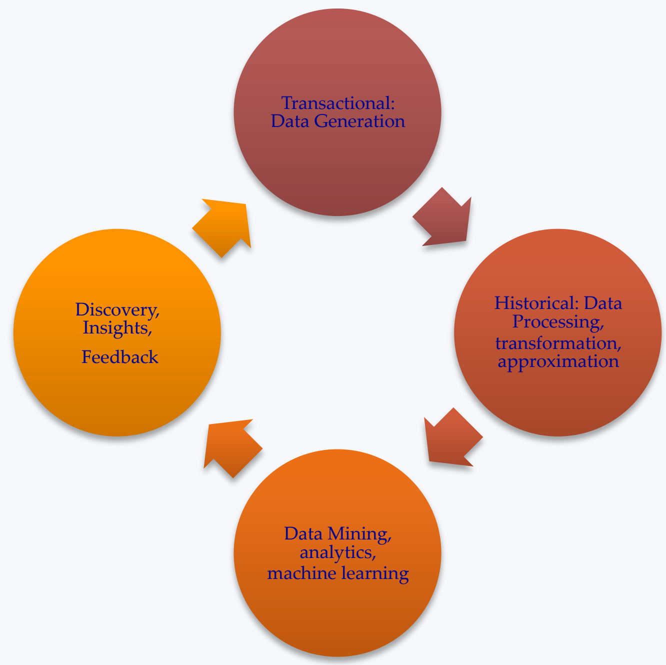
Thank You!

Alok Choudhary
Dept. of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu
312 515 2562

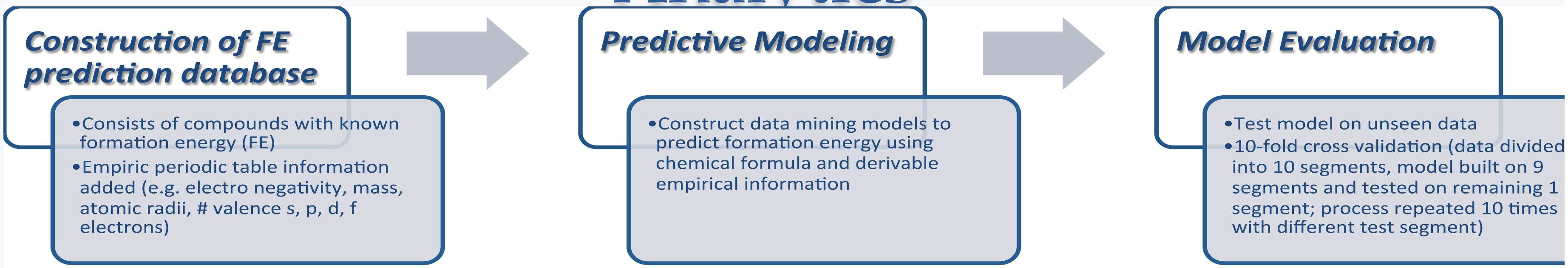
A different way of thinking: Extreme Computing + Big data analytics => Accelerating Discovery

MATERIAL SCIENCE: A "DATA DRIVEN DISCOVERY" WORTH A THOUSAND SIMULATIONS?

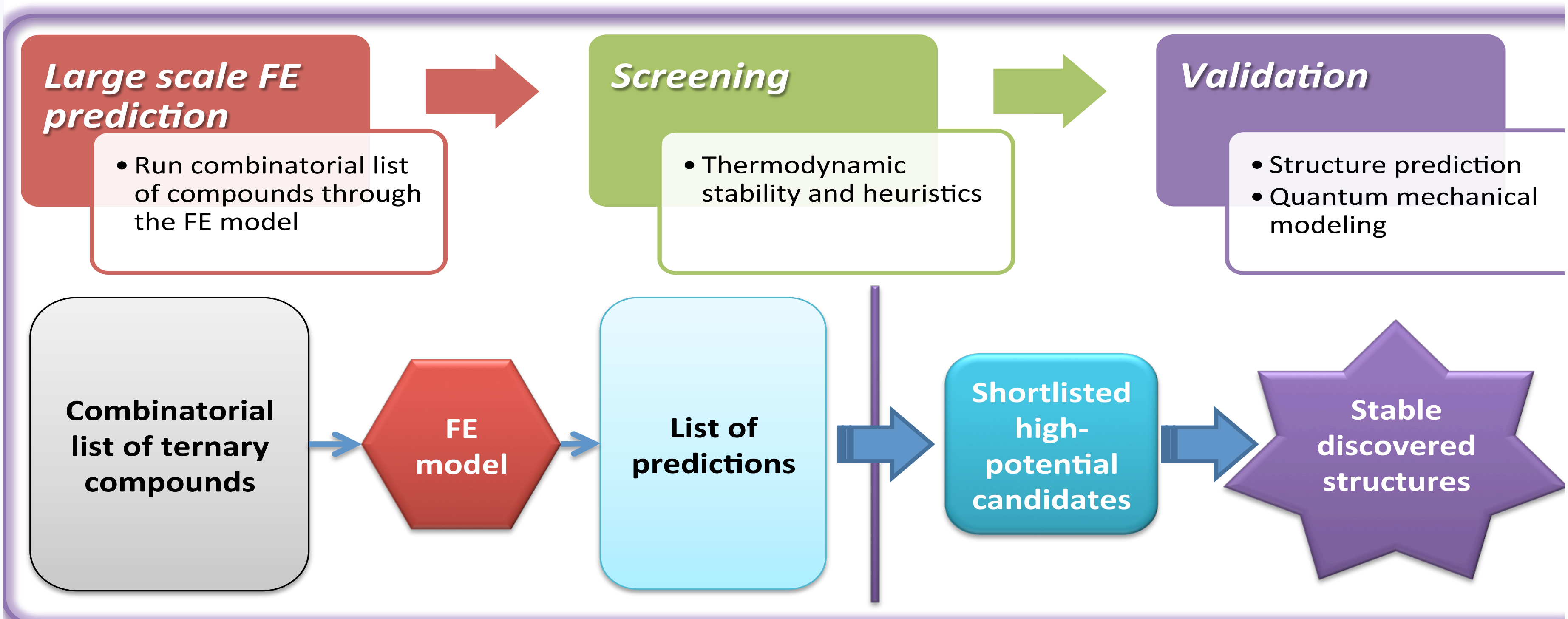
• • •



Discovering Materials : Simulations → Analytics

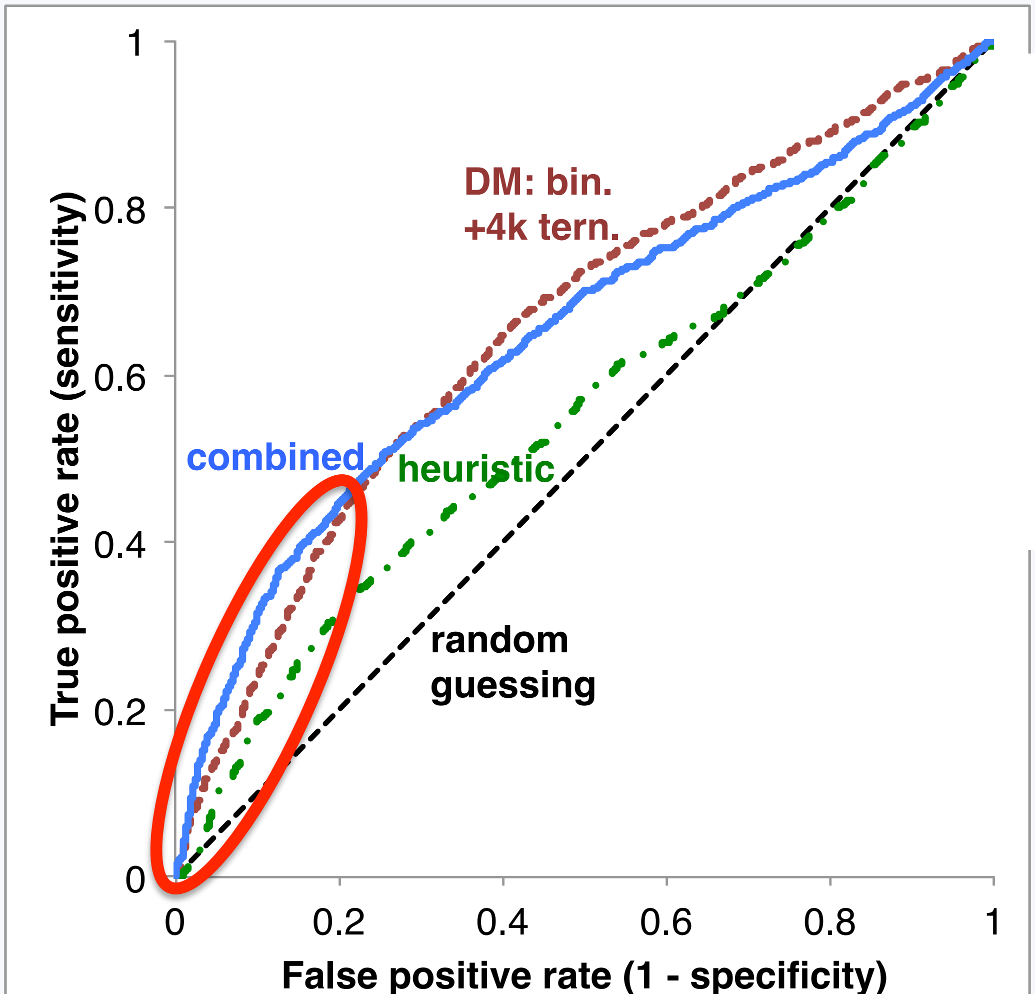


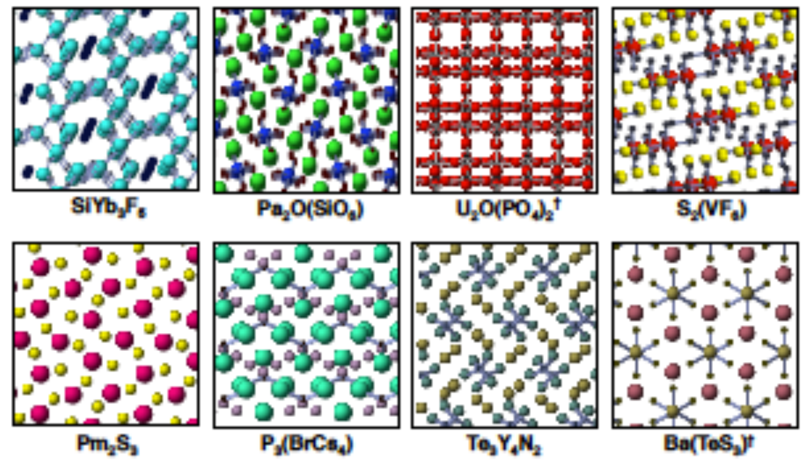
(a)



(b)

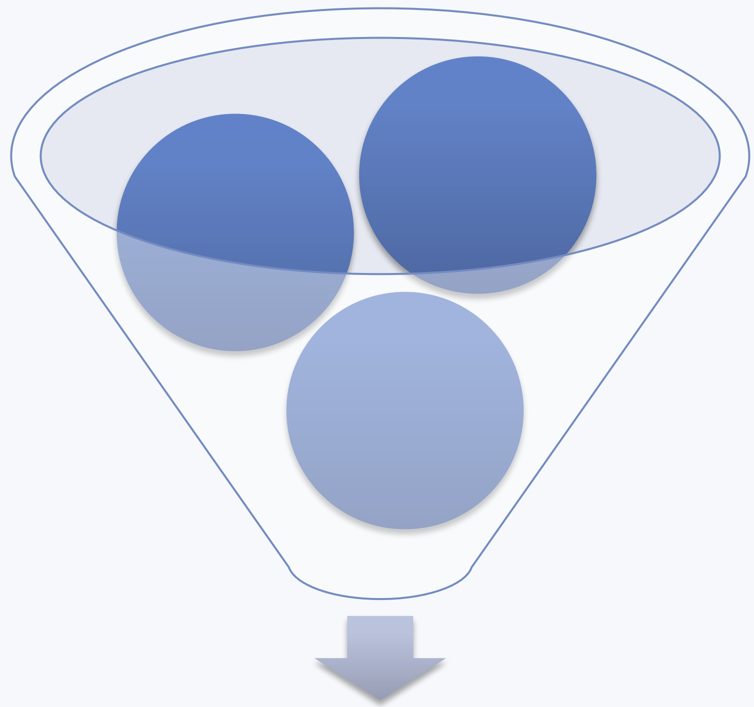
Ranking – Approximation is good enough for ranking ☺ (closing the loop)



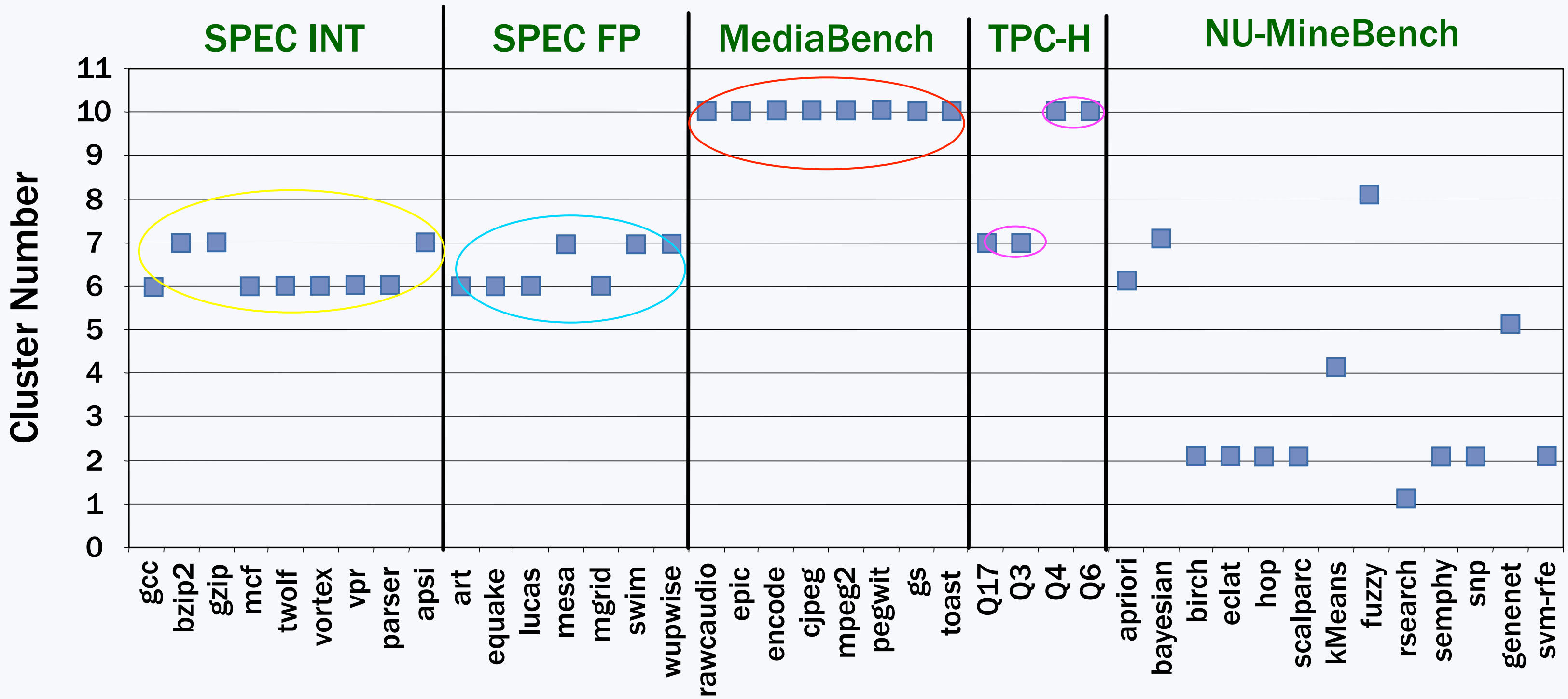


† indicates a model prediction associated with a known stable ternary compound that had was absent from DFT thermodynamic database; the prediction is thus confirmed, but no crystal structure search was necessary.

Appendix



Data Analytics/Mining applications: Do they have different characteristics?



Clear Implications on architecture, modes, memory hierarchy and other components. Identify similarities and design for co-existence

Analytics Apps Algorithms and Kernels…?

Analytics Algorithms	Top 3 Kernels			Σ (%)
	Kernel 1 (%)	Kernel 2 (%)	Kernel 3 (%)	
K-means	Distance (68)	Center (21)	minDist (10)	99
Fuzzy K-means	Center (58)	Distance (39)	fuzzySum (1)	98
BIRCH	Distance (54)	Variance (22)	Redist (10)	86
HOP	Density (39)	Search (30)	Gather (23)	92
Naïve Bayesian	probCal (49)	Variance (38)	dataRead (10)	97
ScalParC	Classify (37)	giniCalc (36)	Compare (24)	97
Apriori	Subset (58)	dataRead (14)	Increment (8)	80
Eclat	Intersect (39)	addClass (23)	invertC (10)	72
SVMlight	quotMatrix (57)	quadGrad (38)	quotUpdate (2)	97

Data Analytics – Broad Impact => Accelerating Discoveries

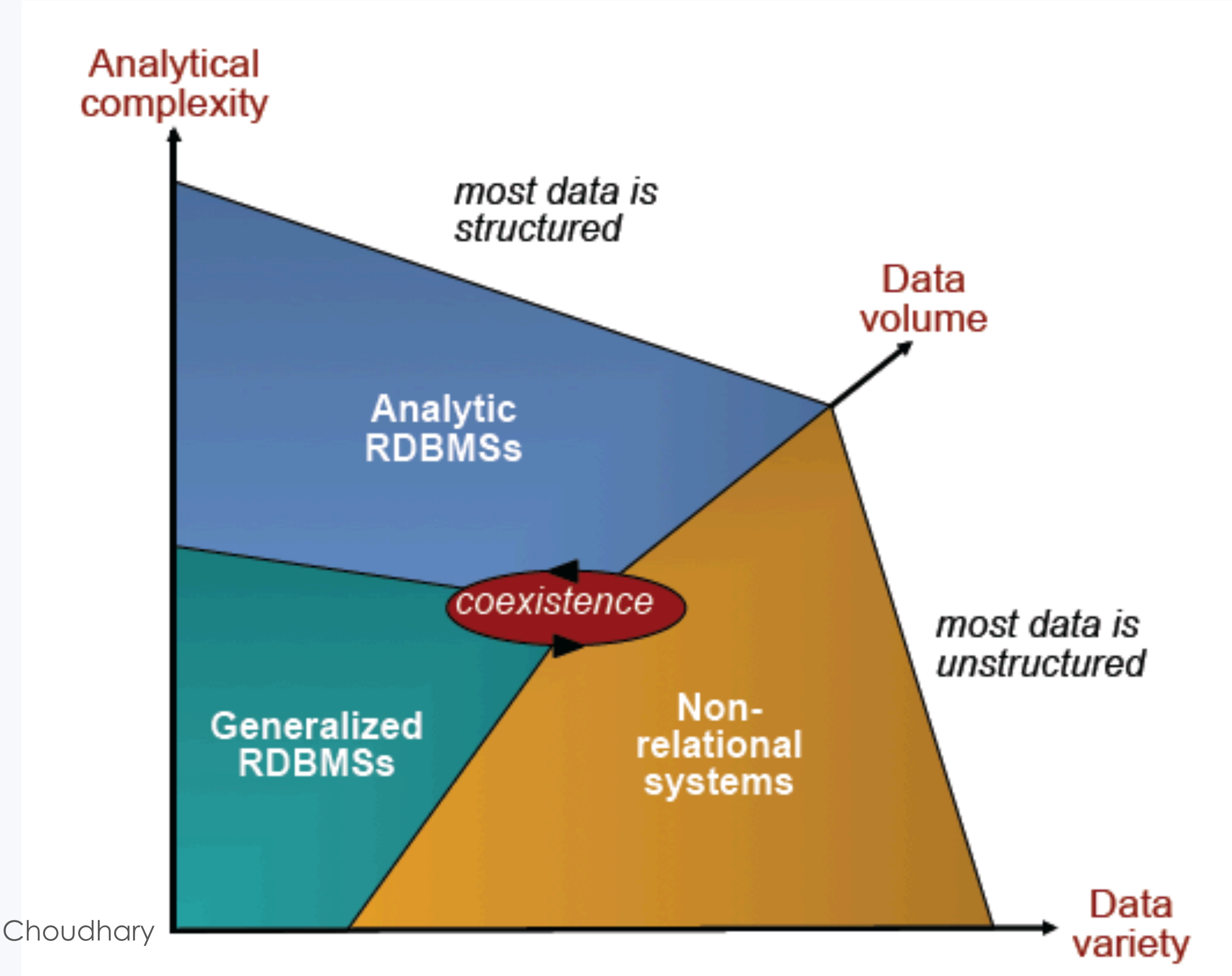
Illustrative Applications	Feature, data reduction, or analytics task	Data analysis kernels
Chemistry, **Climate,** Combustion, Cosmology, Fusion, Materials science, Plasma	Clustering	k-means, fuzzy k-means, BIRCH, MAFIA, DBSCAN, HOP, SNN, Dynamic Time Warping, Random Walk
Biology, **Climate**, Combustion, Cosmology, Plasma, Renewable energy	Statistics	Extrema, mean, quantiles, standard deviation, copulas, value-based extraction, sampling
Biology, **Climate,** Fusion, Plasma	Feature selection	Data slicing, LVF, SFG, SBG, ABB, RELIEF
Chemistry, Materials science, Plasma, **Climate**	Data transformations	Fourier transform, wavelet transform, PCA/SVD/EOF analysis, multidimensional scaling, differentiation, integration
Combustion, **Earth science**	Topology	Morse-Smale complexes, Reeb graphs, level set decomposition
Earth science	Geometry	Fractal dimension, curvature, torsion
Biology, **Climate,** Cosmology, Fusion	Classification	ScalParC, decision trees, Naïve Bayes, SVMlight, RIPPER
Chemistry, **Climate**, Combustion, Cosmology, Fusion, Plasma	Data compression	PPM, LZW, JPEG, wavelet compression, PCA, Fixed-point representation
Climate	Anomaly detection	Entropy, LOF, GBAD
Climate, Earth science	Similarity / distance	Cosine similarity, correlation (TAPER), mutual information, Student's t-test, Eulerian distance,

Right Computing infrastructure? What characteristics do typical analytics functions have?

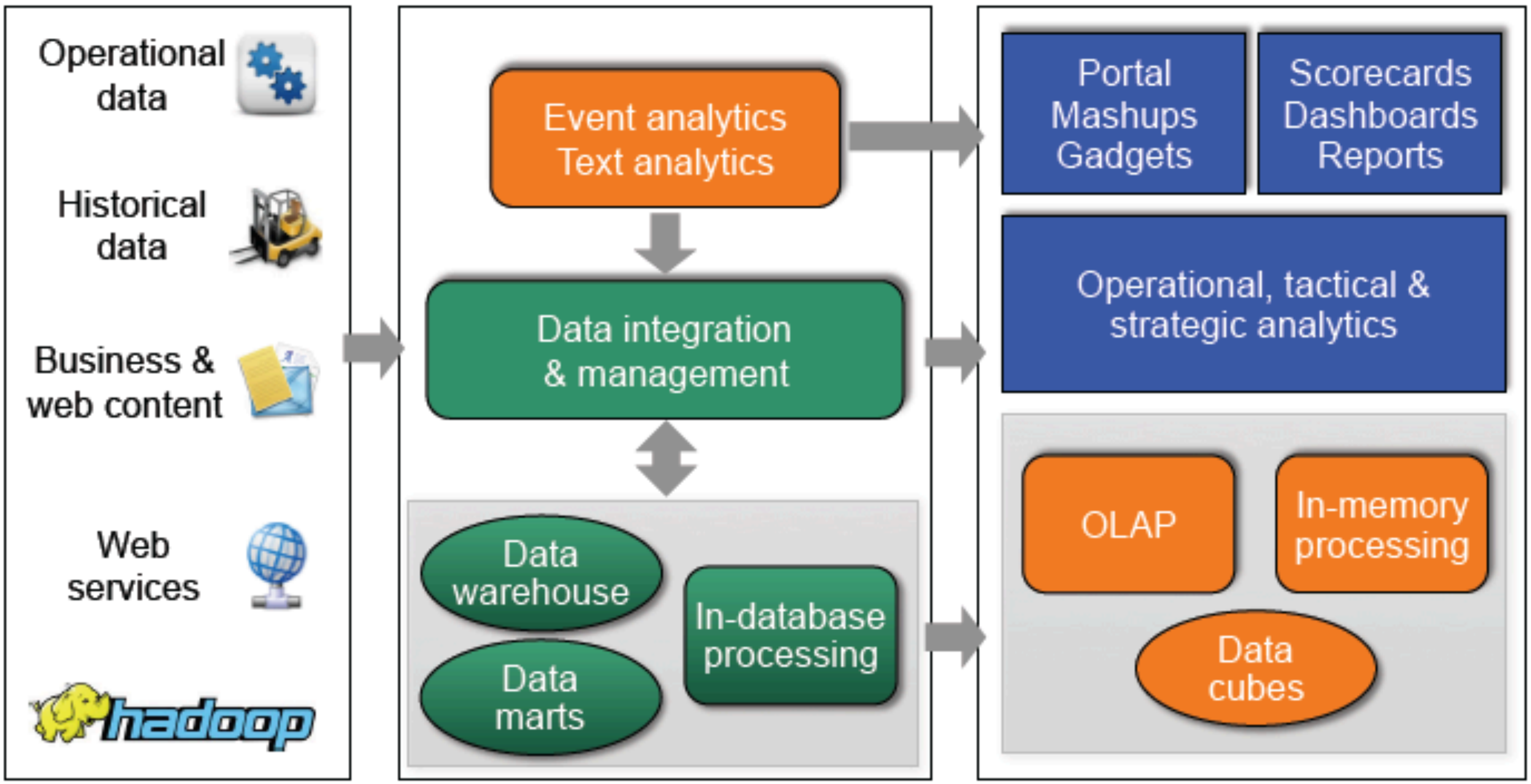
Parameter†	Benchmark of Applications				
	SPECINT	SPECFP	MediaBench	TPC-H	MineBench
Data References	0.81	0.55	0.56	0.48	1.10
Bus Accesses	0.030	0.034	0.002	0.010	0.037
Instruction Decodes	1.17	1.02	1.28	1.08	0.78
Resource Related Stalls	0.66	1.04	0.14	0.69	0.43
CPI	1.43	1.66	1.16	1.36	1.54
ALU Instructions	0.25	0.29	0.27	0.30	0.31
L1 Misses	0.023	0.008	0.010	0.029	0.016
L2 Misses	0.003	0.003	0.0004	0.002	0.006
Branches	0.13	0.03	0.16	0.11	0.14
Branch Mispredictions	0.009	0.0008	0.016	0.0006	0.006

† The numbers shown here for the parameters are values per instruction

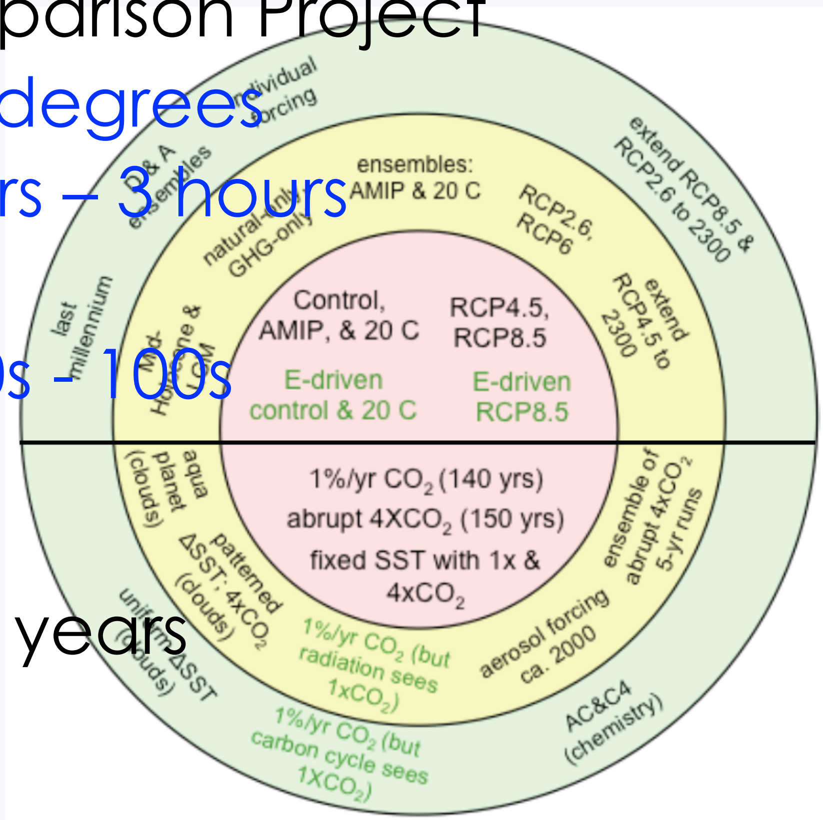
Big Data: Generalization and Optimizations



Data → Information → Insights → Actions

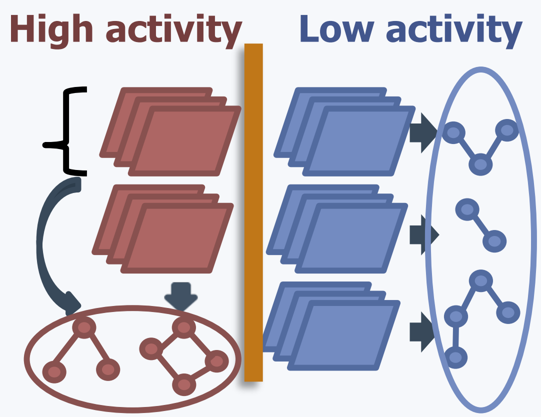


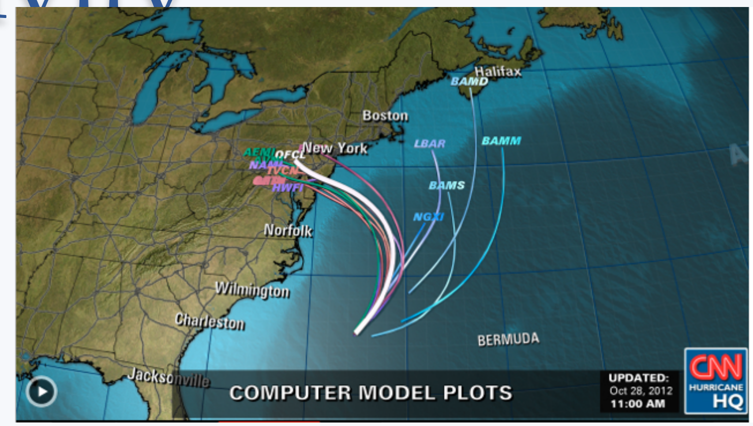
- Coupled Model Inter comparison Project
- Spatial resolution: 1 – 0.25 degrees
- Temporal resolution: 6 hours – 3 hours
- Models: 24 - 37
- Simulation experiments: 10s - 100s
  - Control runs & hindcast
  - Decadal & centennial-scale forecasts
- Covers 1000s of simulation years
- 100+ variables
- 10s of TBs to 10s of PBs

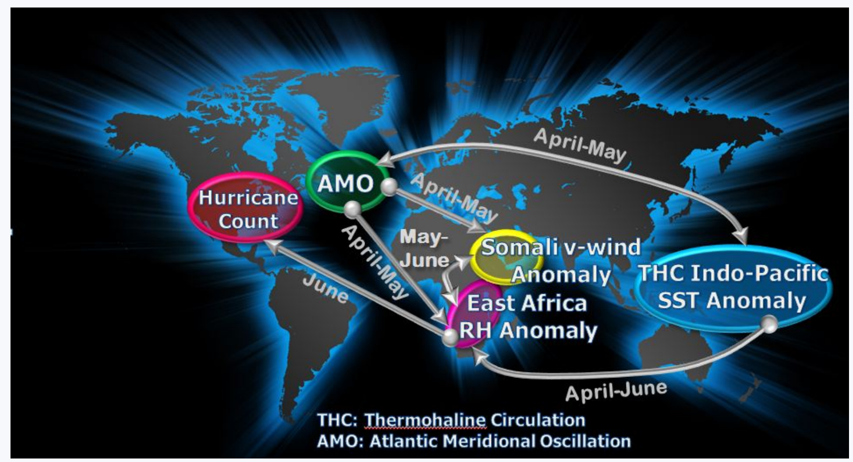


Summary of CMIP5 model experiments, grouped into three tiers

Relationship mining: Seasonal hurricane activity







- Contrast-based network mining for discriminatory signatures
- Novel dynamic graph clustering for dense directed graphs

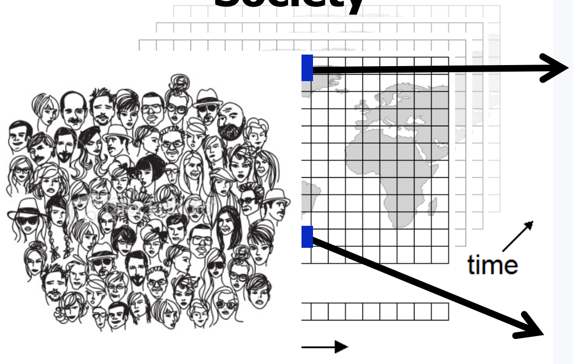
- Improved forecast skill for seasonal hurricane activity
- Discovered key factors and mechanisms modulating NA hurricane variability

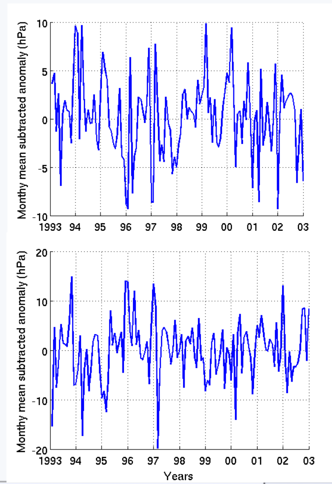
NSF News, DOE Research News, Science360
Sencan et al. *IJCAI* (2011)
Pendse et al. *SIAM SDM* (2012)
Chen et al. *Data Mining & Knowledge Discovery* (2012)
Chen *et al.* *SIAM SDM* (2013)
Chen *et al.* *IJCAI* (2013)
Semazzi *et al.* in review at journal (2013)

From Science to Business + Social

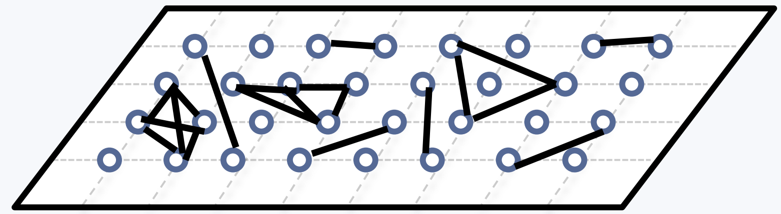
- People/Customers/fans are interacting points in space-time
- Similarity of interests defines communities
- Communication across globes defines networks

Society





Activity/interaction based Network



Edge weights: significant interactions/influence
Nodes in the graph: people/brands/…

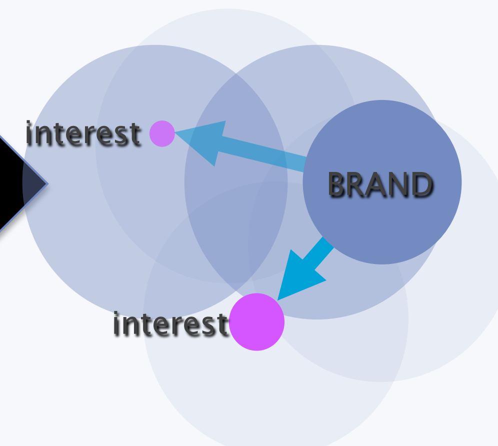
facebook
twitter
LinkedIn
Google+
Blogger
YouTube Broadcast Yourself™

Massive Data and Social Networks Mining | Influence Tracking and Analysis

Scalable Analytics

Multi-language Sentiment Analytics | Learning and Predictive Modeling

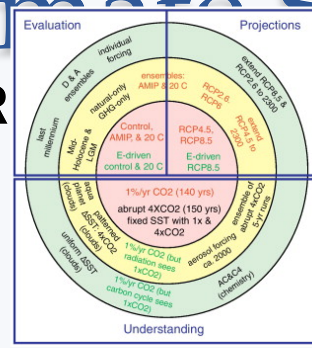
Action-Based Connections

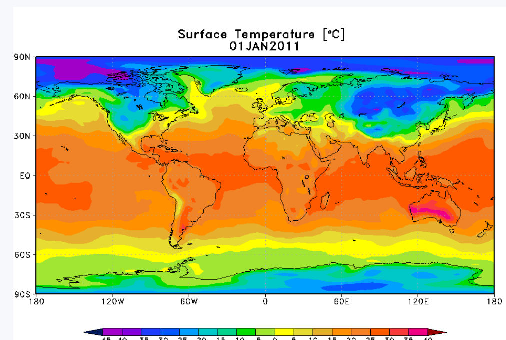


Data-Driven Knowledge Discovery in Climate Science

Transformation from Data-Poor to Data-R

- Sensor Observations
- Reanalysis Data
- **Model Simulations**





A data-driven approach that:

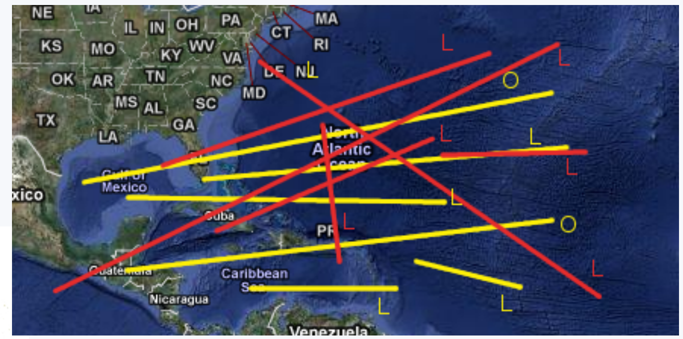
- Makes use of wealth of observational and simulation data
- Advances understanding of climate processes
- Informs climate change impacts and adaptation

"Climate change research is now 'big science,' comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics."

(Nature Climate Change, Oct 2012)

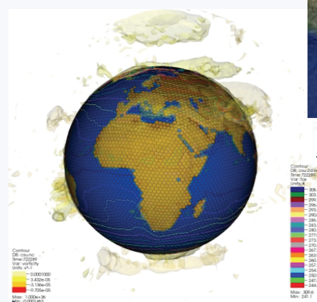
The Growth of Complexity ➔ Need for Simplicity

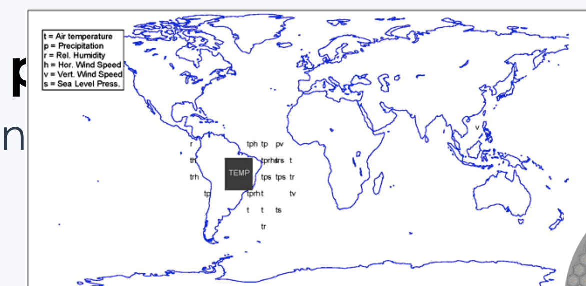
- **Higher spatial or temporal resolution**
  - extremes analysis
  - Network-based prediction
  - Estimation of spatiotemporal dependence
- **Higher data dimensionality**
  - Bayesian analysis of multi-model ensembles
  - Sampling-based statistical methods
  - Multivariate quantile analysis
- **Greater complexity per data p**
  - Estimation of complex dependen structures
  - Handling non-stationarity
  - Multi-resolution analysis
- **Shorter response time**
  - Interactive hypothesis testing



Significant correlations for hurricane prediction

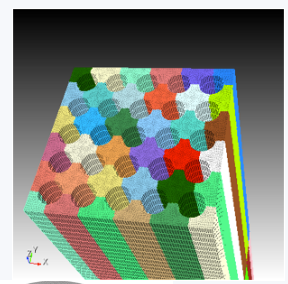
(Sencan, Chen, Hendrix, Pansombut, Semazzi, Choudhary, Kumar, Melechko, and Samatova, 2011)

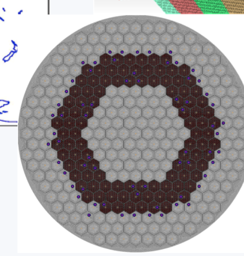




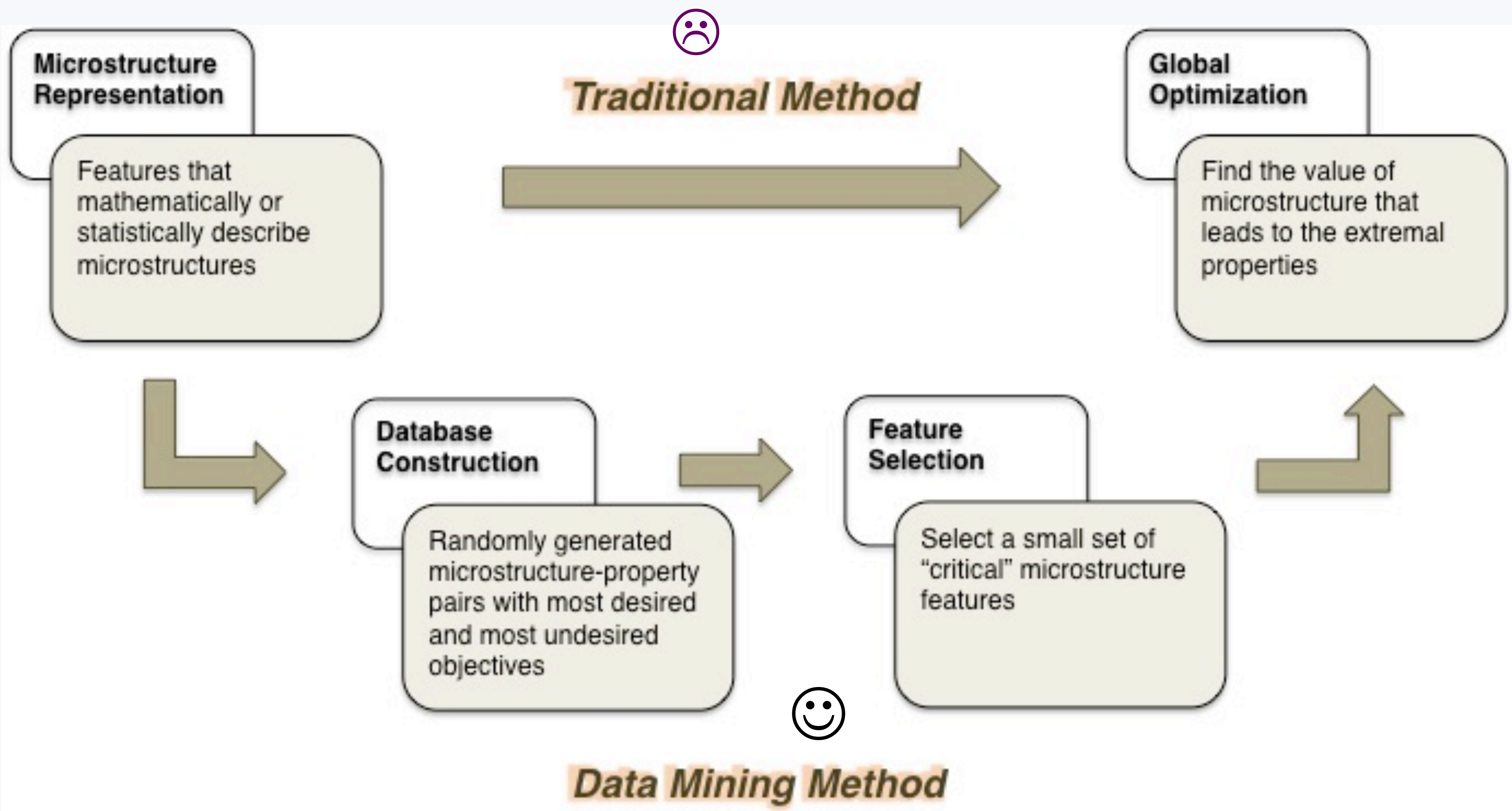
Prediction of land climate using ocean climate variables

(Chatterjee, Steinhaeuser, Banerjee, Chatterjee, and Ganguly, 2012)

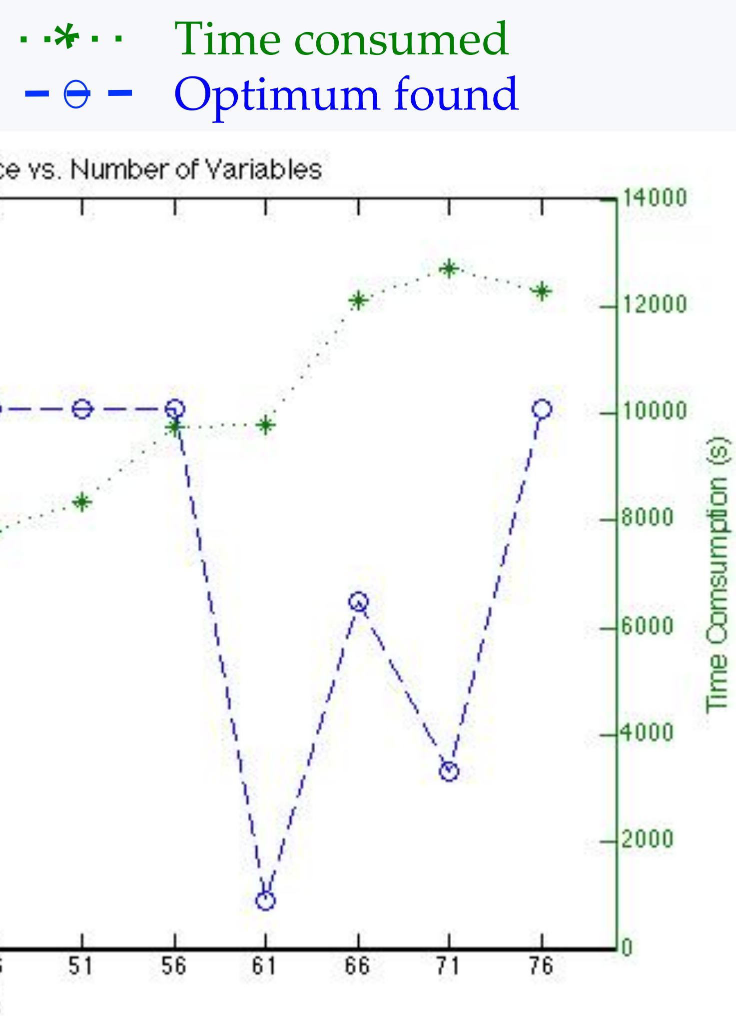




Structure-Property Optimization – Try optimization for 10^3 dimensions



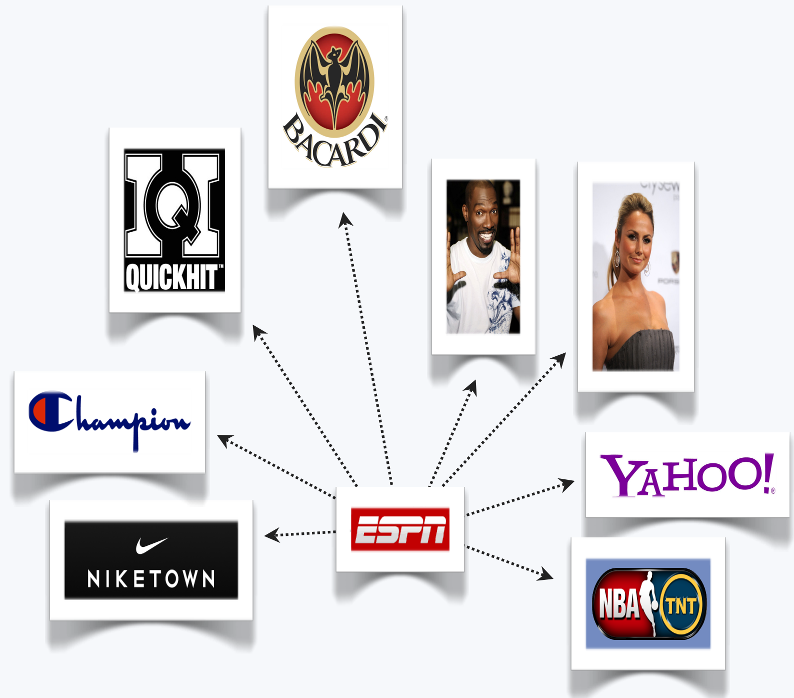
Accelerating Time to Insights




Wedding

Actionable Insights? Unknown-Unknown

Top Brand Affinity



Affinity Mapping