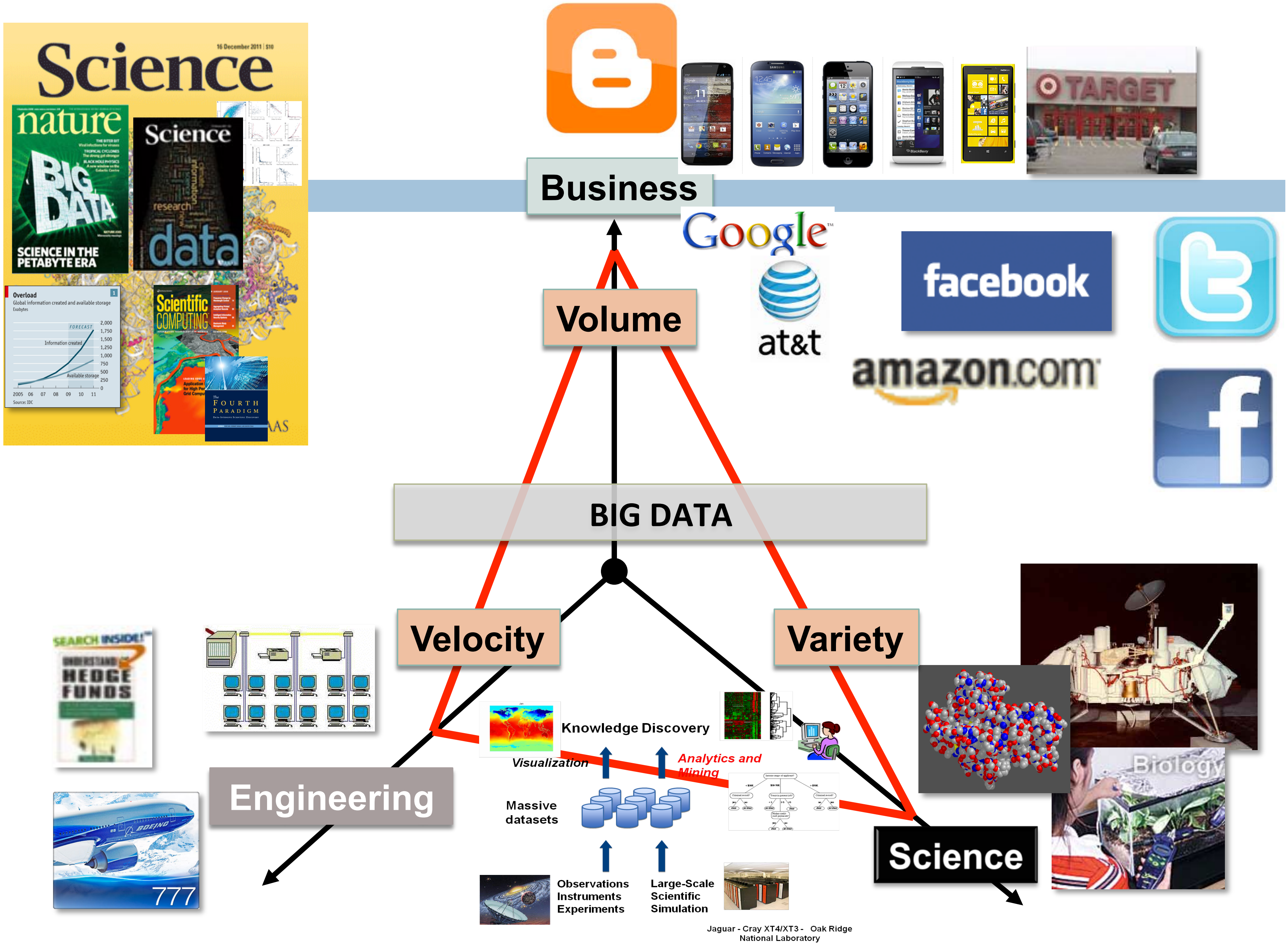
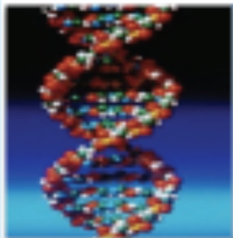
Big Data, Exascale Systems and Knowledge Discovery – The Next Frontier for HPC

Alok Choudhary
Henry and Isabel Dever Professor
of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

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

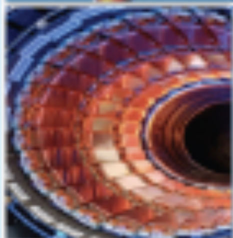


Business
Volume
BIG DATA
Velocity
Variety
Engineering
Science
Knowledge Discovery
Visualization
Analytics and Mining
Massive datasets
Observations Instruments Experiments
Large-Scale Scientific Simulation
Jaguar - Cray XT4/XT3 - Oak Ridge National Laboratory
Google
at&t
facebook
amazon.com
TARGET
Science
nature
BIG DATA
SCIENCE IN THE PETABYTE ERA
data
Scientific COMPUTING
The FOURTH PARADIGM
Overload
Information created
Available storage
HEDGE FUNDS
BOEING
777
Biology



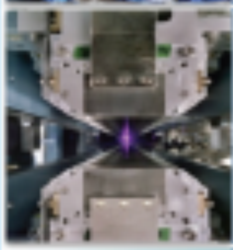
Genomics

Data Volume increases to 10 PB in FY21



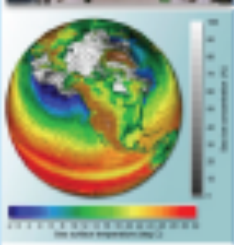
High Energy Physics

(Large Hadron Collider) 15 PB of data/year



Light Sources

Approximately 300 TB/day



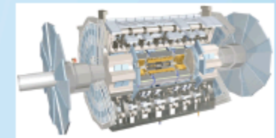
Climate

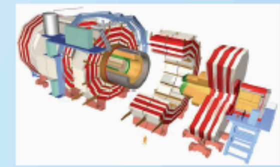
Data expected to be hundreds of 100 EB

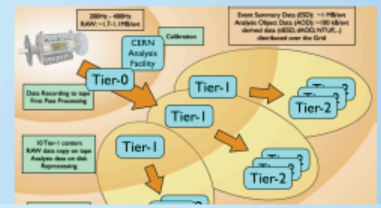
Source: Bill Harrod, SC12 plenary presentation

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

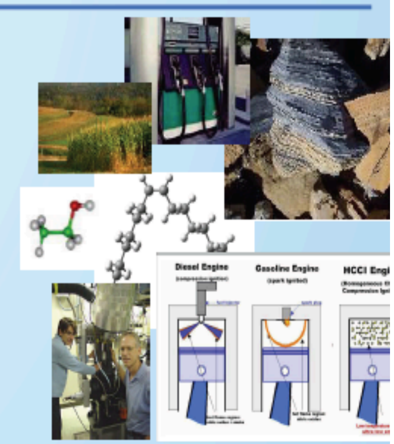


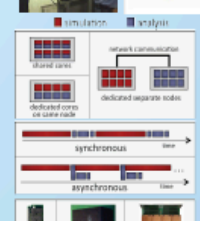




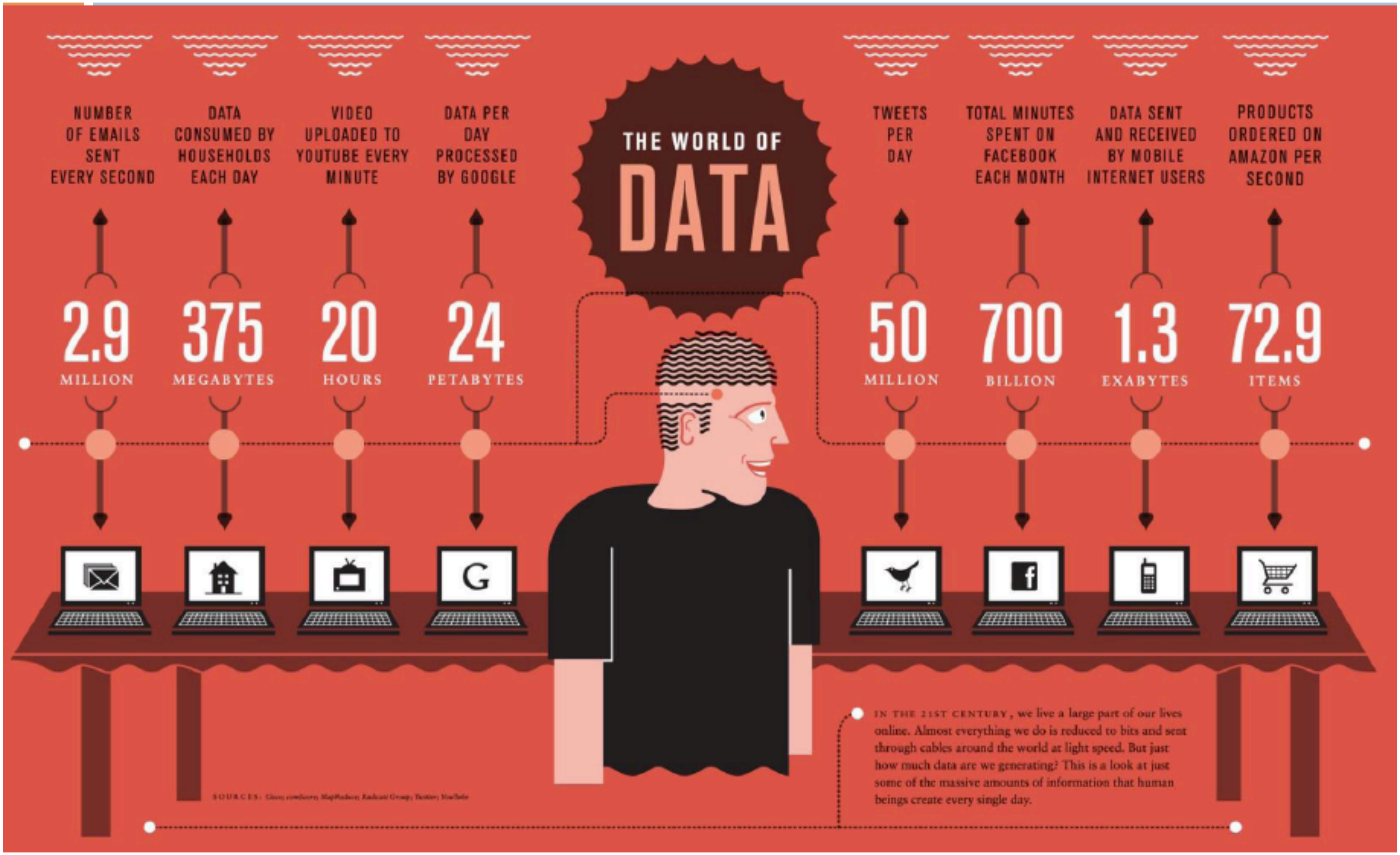
Data Challenges in Large-Scale Simulations: S3D Combustion code exemplar

- 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





Big Data …



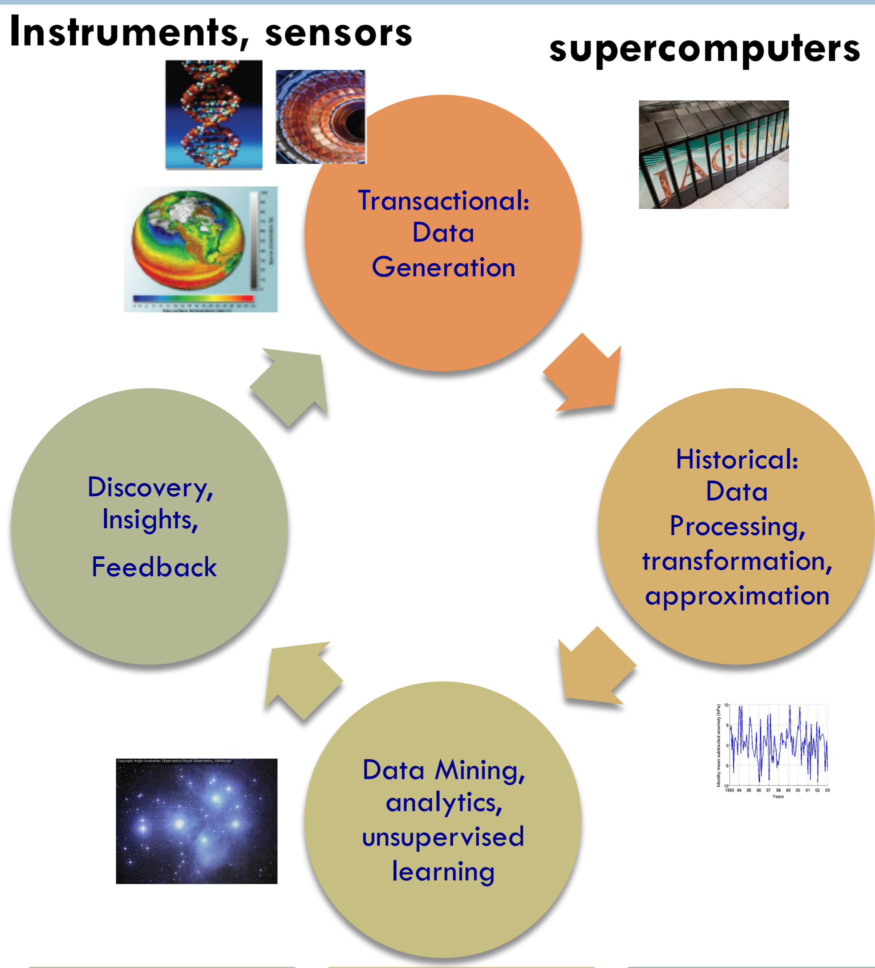
THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
f
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
SOURCES:

BIG DATA?

Wikipedia says; "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." ☺

How should we think about big data?

Knowledge Discovery Life-Cycle: Transactional to Relationships – Current to Historical



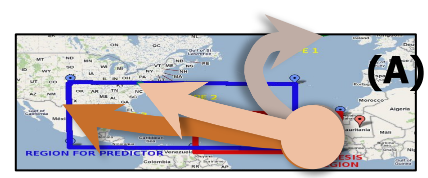
Data Management

Data Reduction, Query

Data Visualization

Data Sharing

Historical data → Learning Models → Trigger/ questions → Predict



“Data intensive” vs “Data Driven”

Data Intensive (DI)

- Depends on the perspective
  - Processor, memory, application, storage?
- An application can be data intensive without (necessarily) being I/O intensive

Data Driven (DD)

- Operations are driven and defined by data
  - BIG analytics
    - Top-down query (well-defined operations)
    - Bottom up discovery (unpredictable time-to-result)
  - BIG data processing
  - Predictive modeling
- Usage model further differentiates these
  - Single App, users
  - Large number, sharing, historical/temporal

Very few large-scale applications of practical importance are NOT Data Intensive

In Extreme Scale Science domain, we typically focus on “Transactional” thinking

Understanding Climate Change

- **CO2 levels hit new peak at key observatory**

CNN U.S.

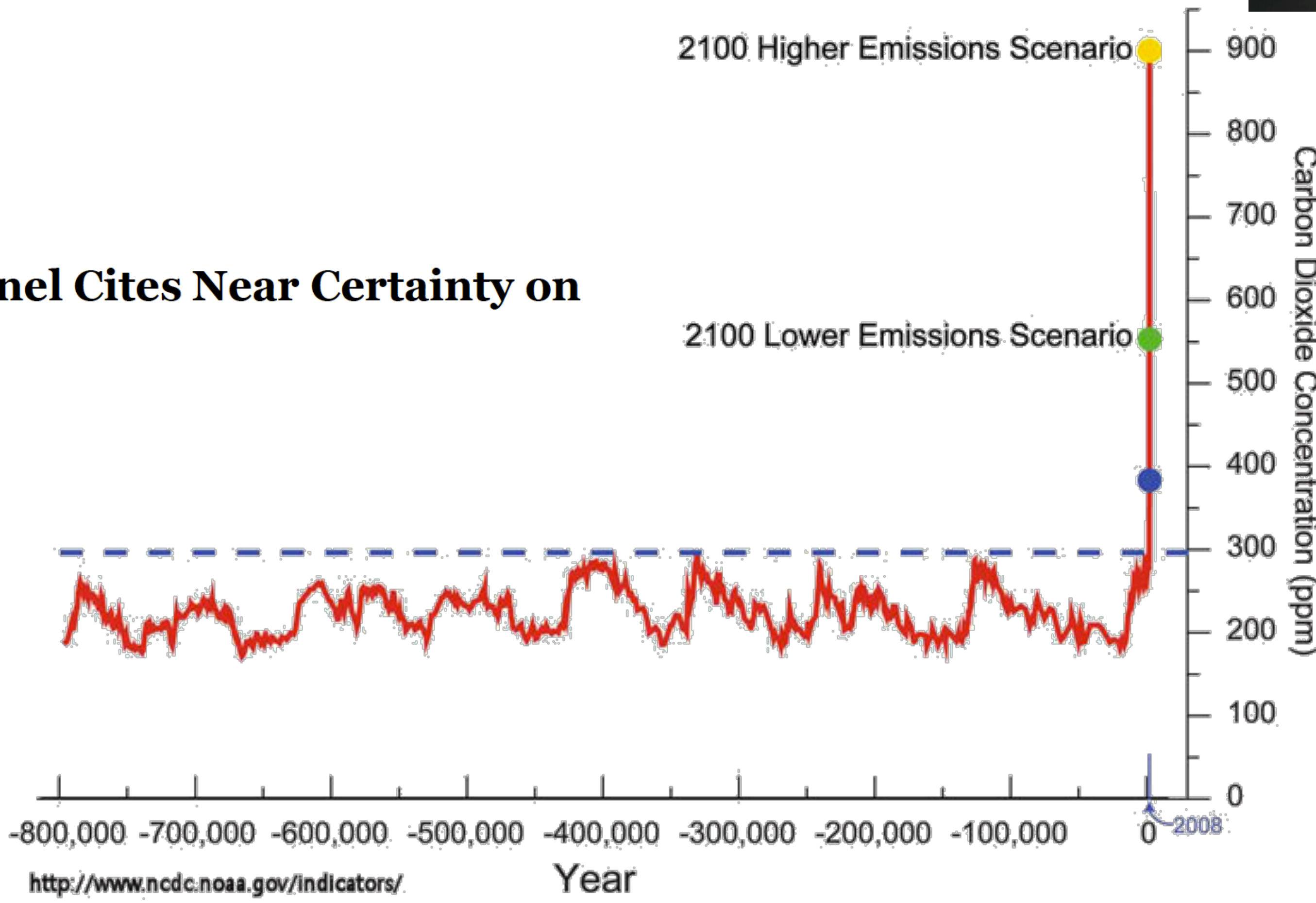
NOAA Satellite and Information Service
National Environmental Satellite, Data, and Information Service (NESDIS)

National Climatic Data Center
U.S. Department of Commerce

The New York Times

August 19, 2013

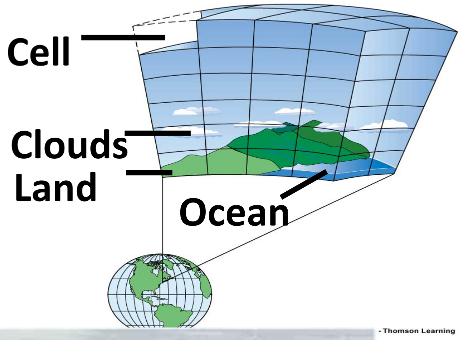
Climate Panel Cites Near Certainty on Warming

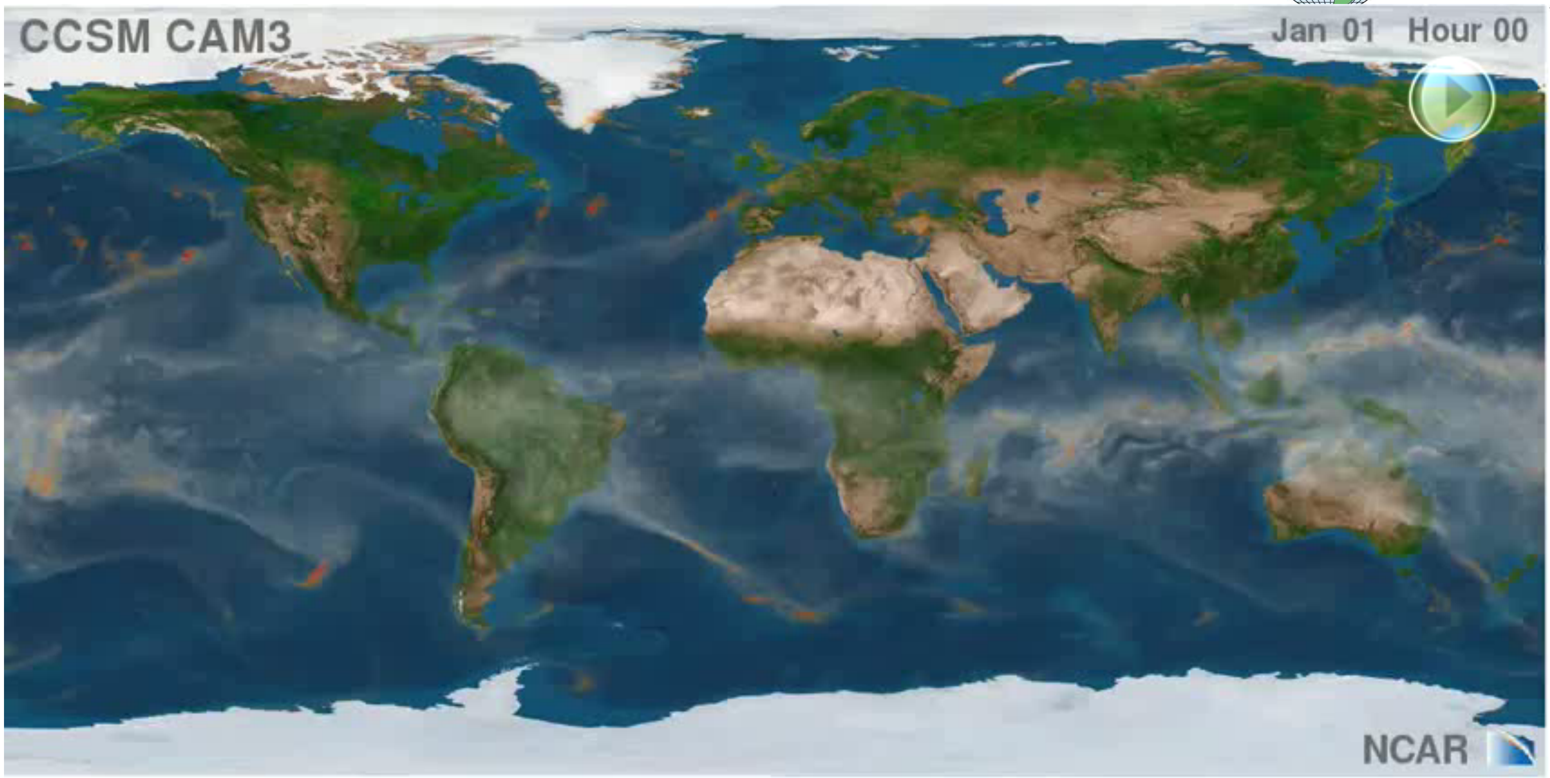


Understanding Climate Change – Physics-Based Approach

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

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





Understanding Climate Change - Physics Based Approach

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

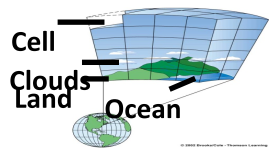


Figure Courtesy: NCAR

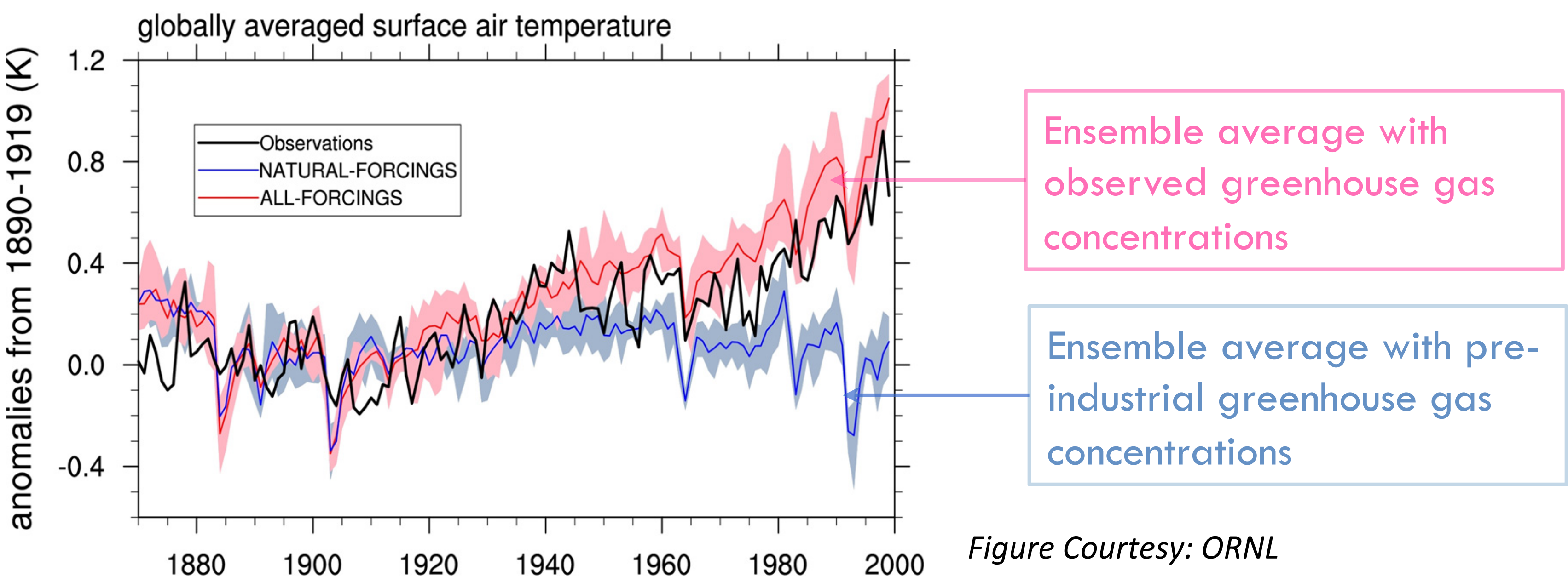
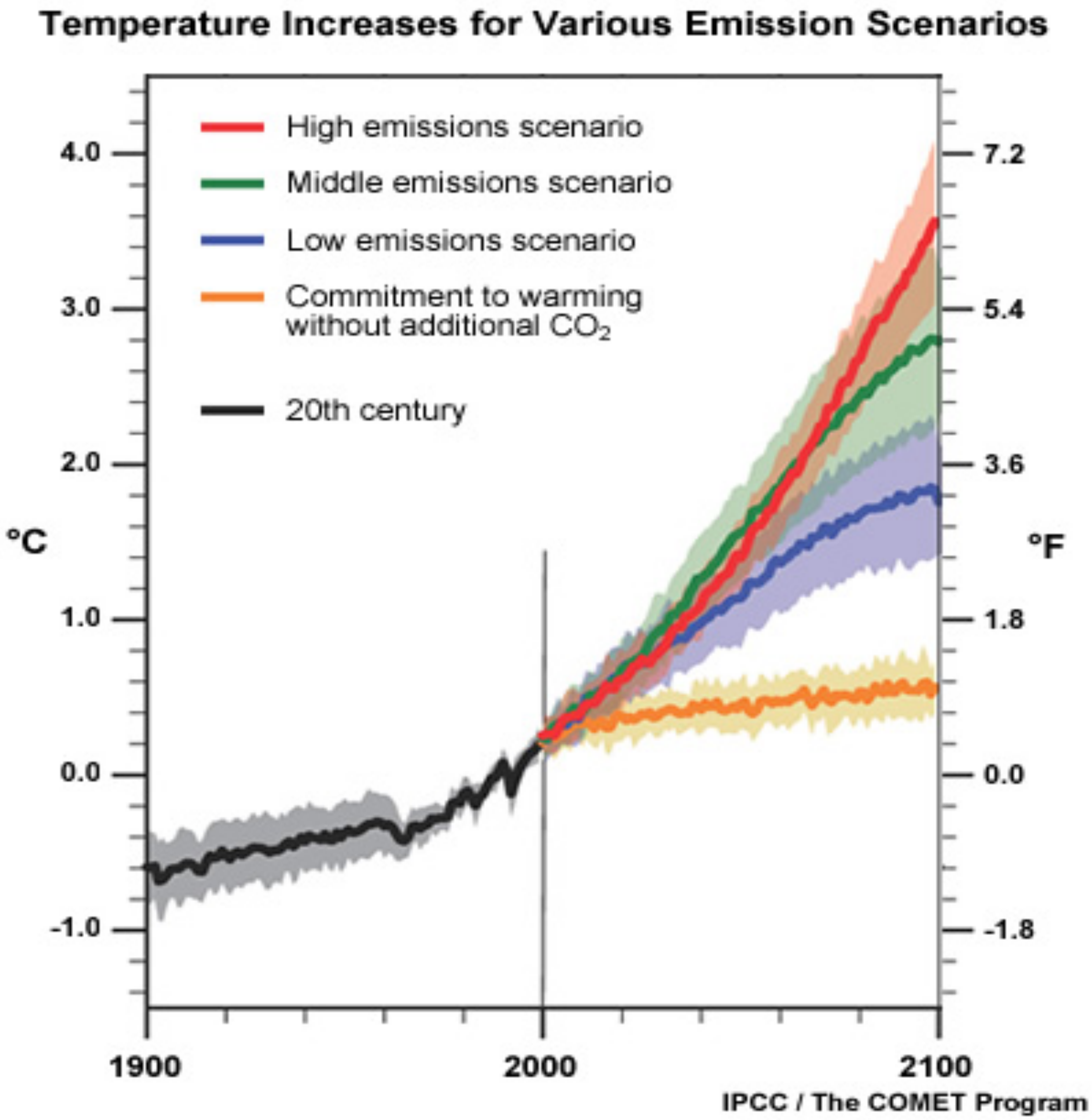


Figure Courtesy: ORNL

Understanding Climate Change - Physics Based Approach

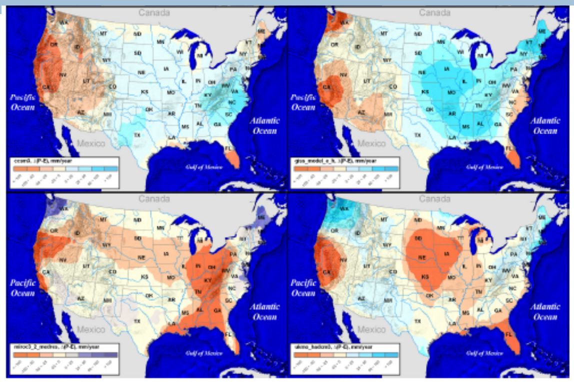


Projection of temperature increase under different **Special Report on Emissions Scenarios** (SRES) by 24 different GCM configurations from 16 research centers used in the **Intergovernmental Panel on Climate Change** (IPCC) 4th Assessment Report.

Physics based models are essential but insufficient

- 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

Disagreement between IPCC models



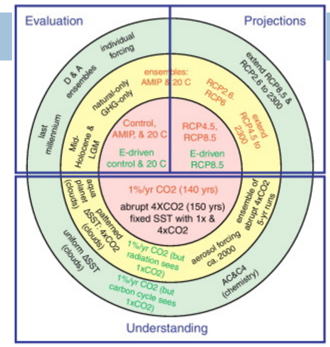
Regional hydrology exhibits large variations among major IPCC model projections

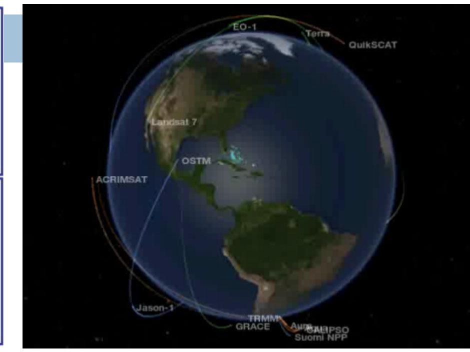
"The sad truth of climate science is that the most crucial information is the least reliable"
(Nature, 2010)

Physics based models

Low uncertainty	High uncertainty
Temperature	Hurricanes
Pressure	Extremes
Large-scale wind	Precipitation

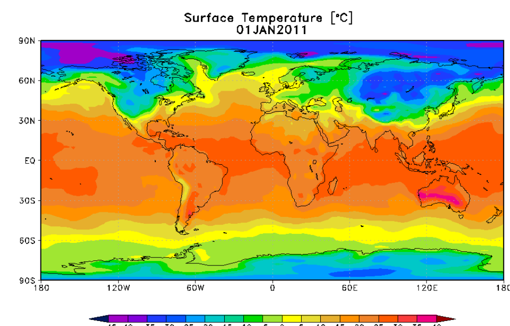
Data-Driven Knowledge Discovery in Climate Science





Transformation from Data-Poor to Data-Rich

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



A new and transformative 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."

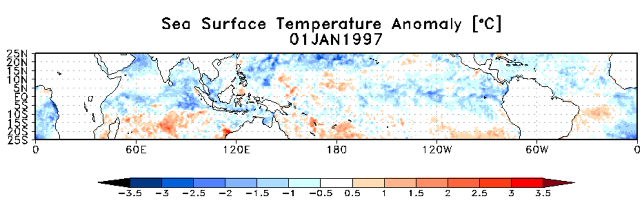
(Nature Climate Change, Oct 2012)

Need for data driven discovery

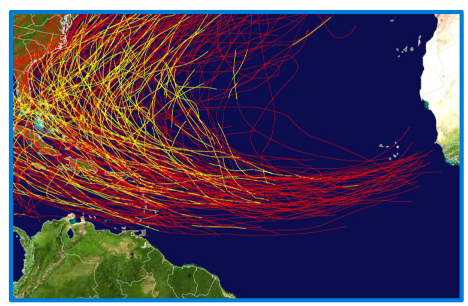
Physics based models

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

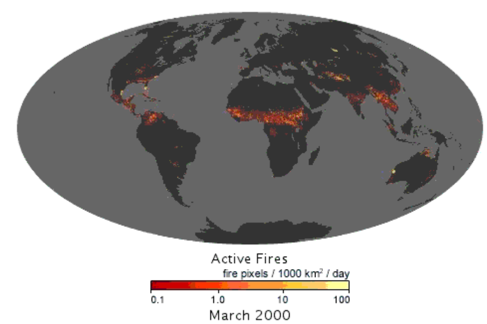
Global sea surface temperatures



Atlantic hurricanes



Global fires



Data Mining, Analytics and Actionable Insights?

A Poem

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

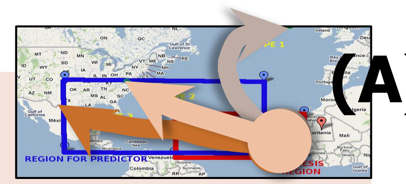
Validate Hypothesis

- Nuclear Reaction happens under these conditions
- Did combustion occur at the expected parameter values
- I think this location contains a black hole

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?

The Unknown

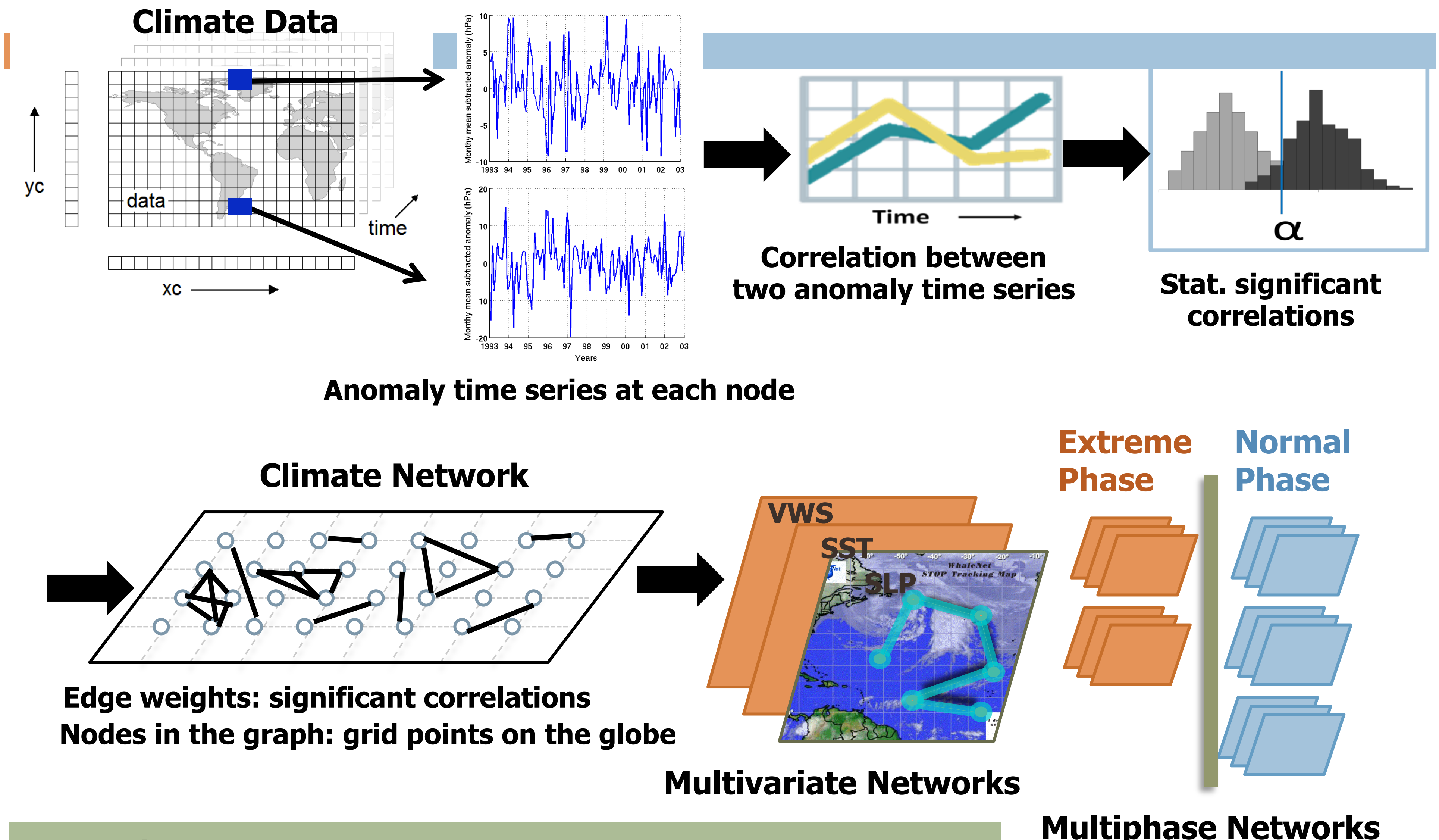
As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.

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.

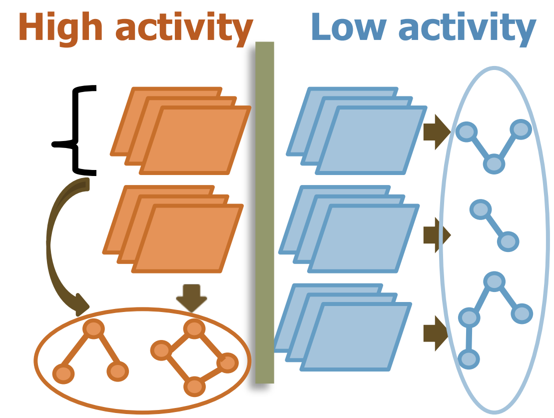
End-to-End: From Transactional analytics to relationship mining



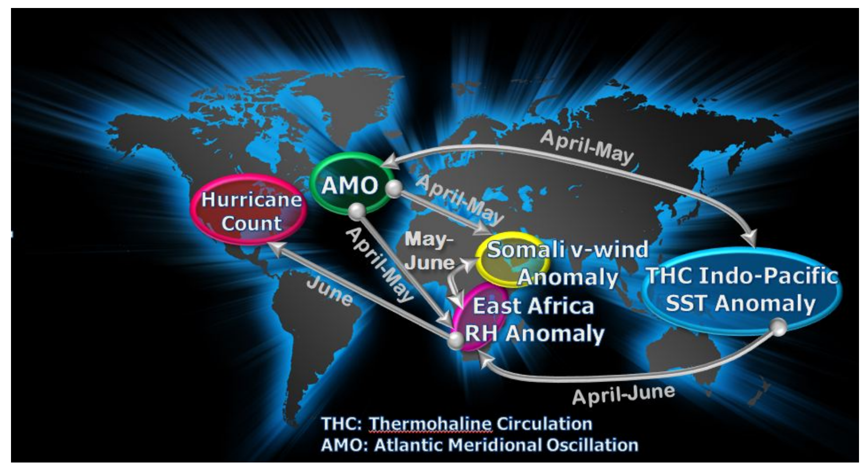
CMIP3 → CMIP5 => Climate BIG DATA : 10s of TBs to 10s of PBs

Relationship mining: Seasonal hurricane activity

- Contrast-based network mining for discriminatory signatures
- Novel dynamic graph clustering for dense directed graphs
- Statistically robust methodology for automatic inference of modulating networks

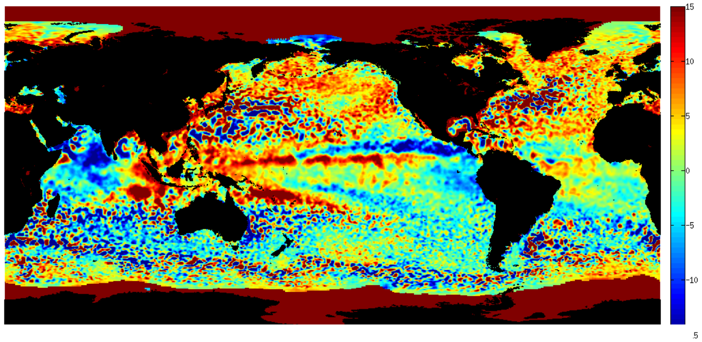


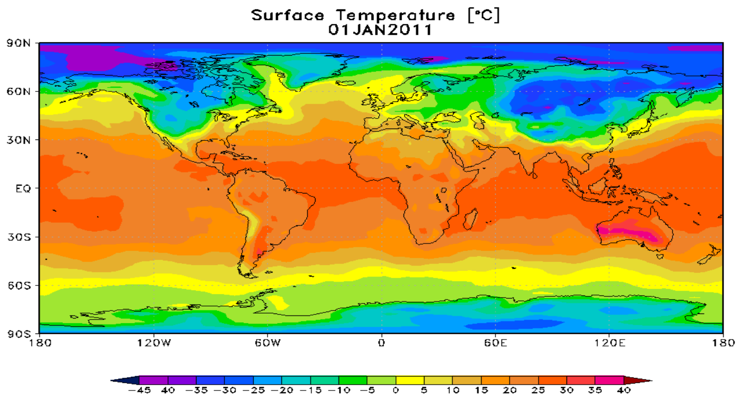
- Improved forecast skill for seasonal hurricane activity
- Discovered key factors and mechanisms modulating NA hurricane variability
- Discovered novel climate index with much improved correlation with NA hurricane variability: 0.69 vs 0.49

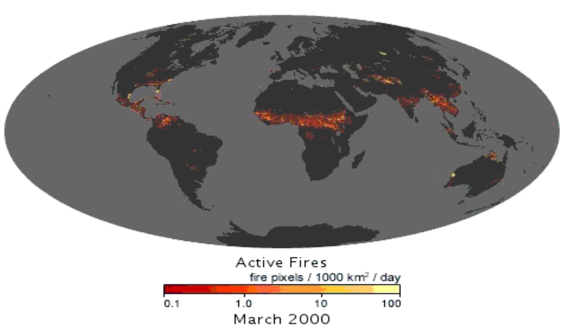


NSF News, DOE Research News, Science360
Sencan et al. *IJCAI* (2011)
Pendse et al. *SIAM SDM* (2012)
Chen et al. *Data Mining & Knowledge Discovery* (2012)
Chen *et al.* *SIAM SDM* (2013)
Chen *et al.* *IJCAI* (2013)
Semazzi *et al.* in review at journal (2013)

Challenges in data driven analysis



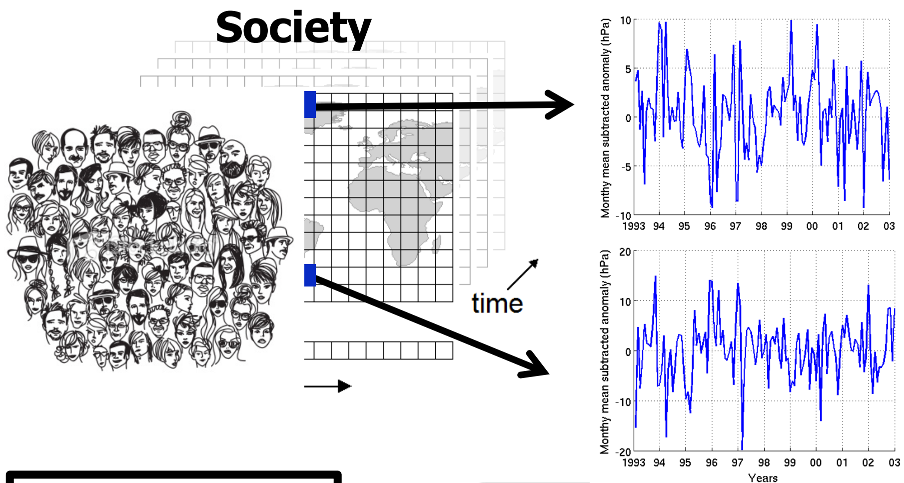




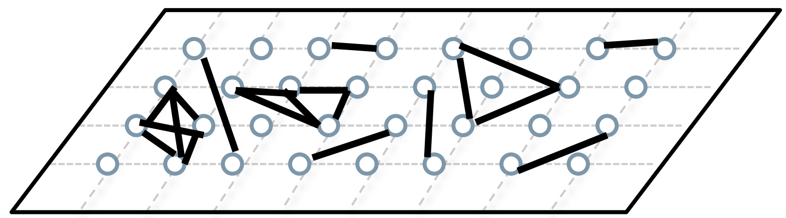
- Complex dependence
  - Non-IID
  - Spatio-temporal correlation
  - Long memory in time
  - Long range dependence in space
  - Nonlinear relationships
- Data characteristics
  - Heterogeneous, Multivariate
  - Heavy Tailed Distributions
  - Noisy, incl. low frequency variability
  - Paucity of training data
- Complex processes
  - Evolutionary
  - Multi-scale in space and time
  - Non-stationary

From Science to Social

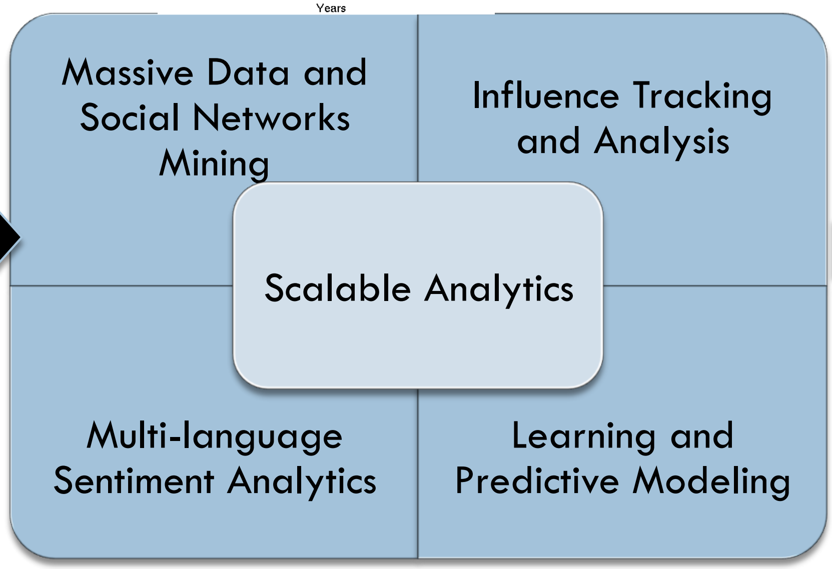
- People/Customers/fans are interacting points in space-time
- Similarity of interests defines communities
- Communication across globes defines networks



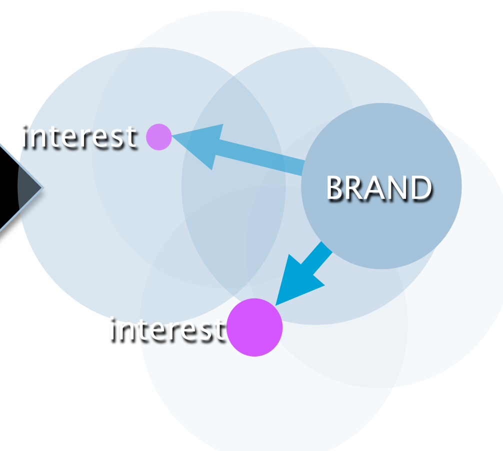
Activity/interaction based Network



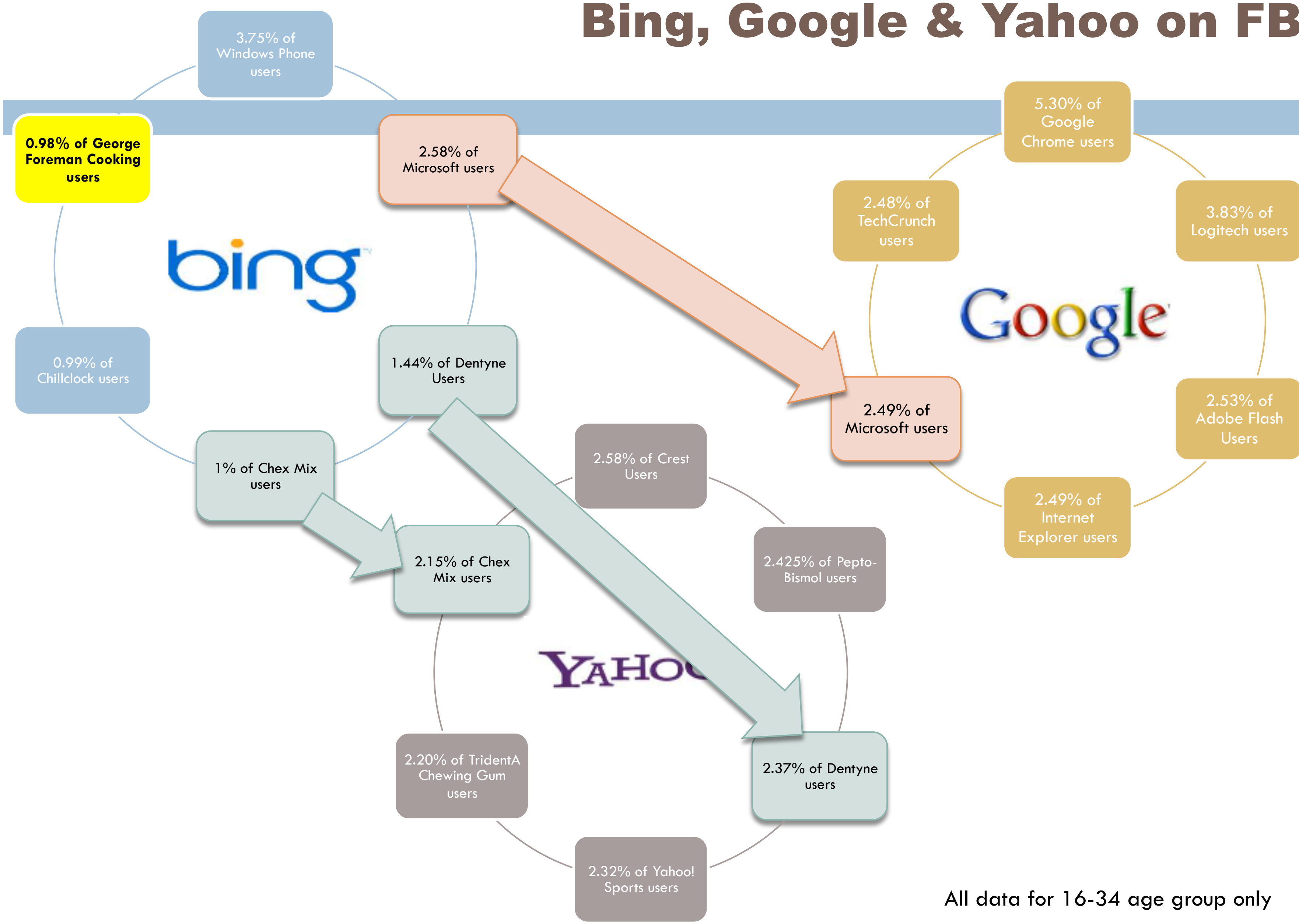
Edge weights: significant interactions/influence
Nodes in the graph: people/brands/…



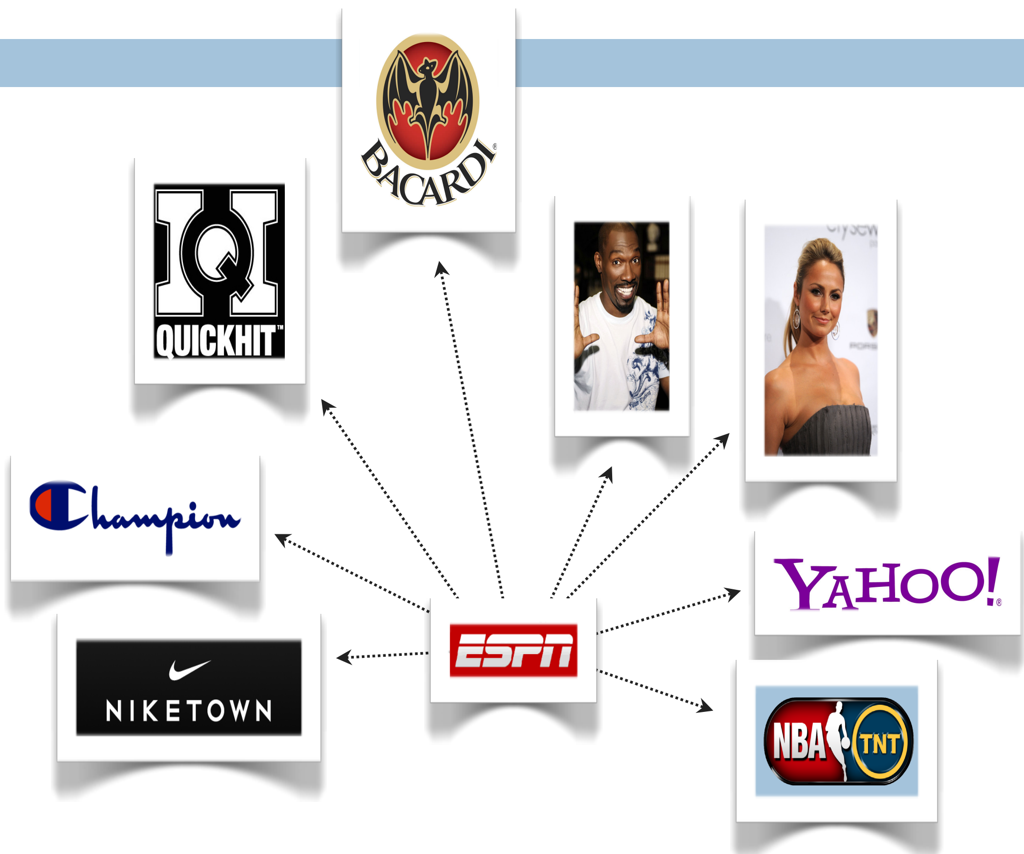
Action-Based Connections



Top Associations by Fans For Bing, Google & Yahoo on FB



All data for 16-34 age group only



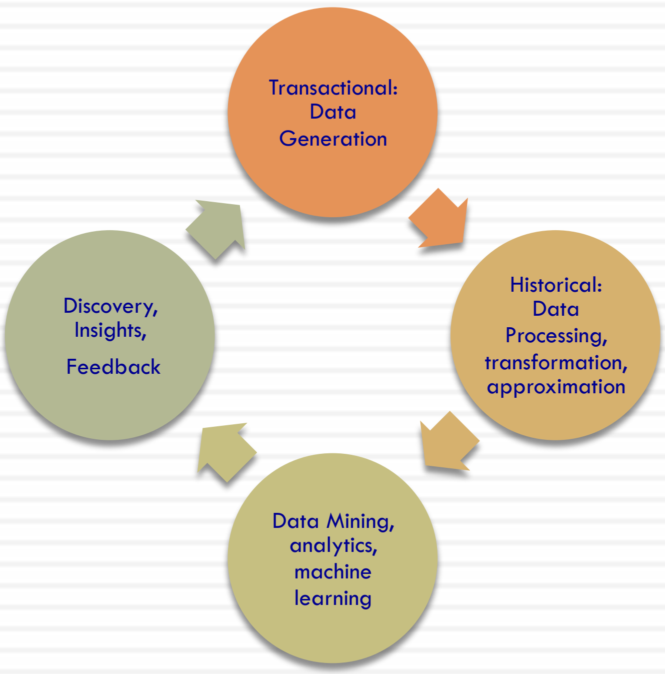
Affinity Mapping

DISCOVER CONNECTIONS

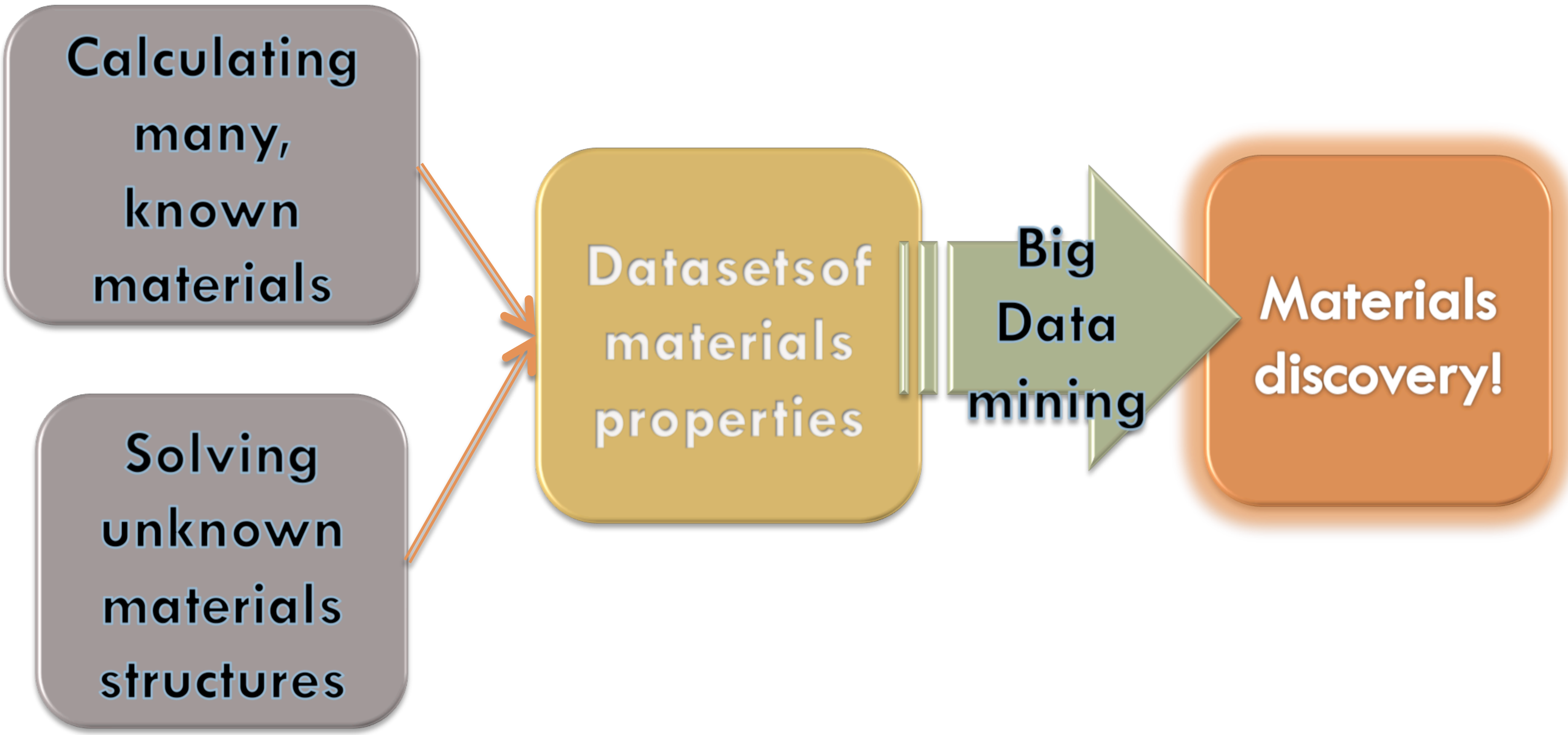
- Track engagement patterns
- Unlimited mapping
- Surprises will ensue

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

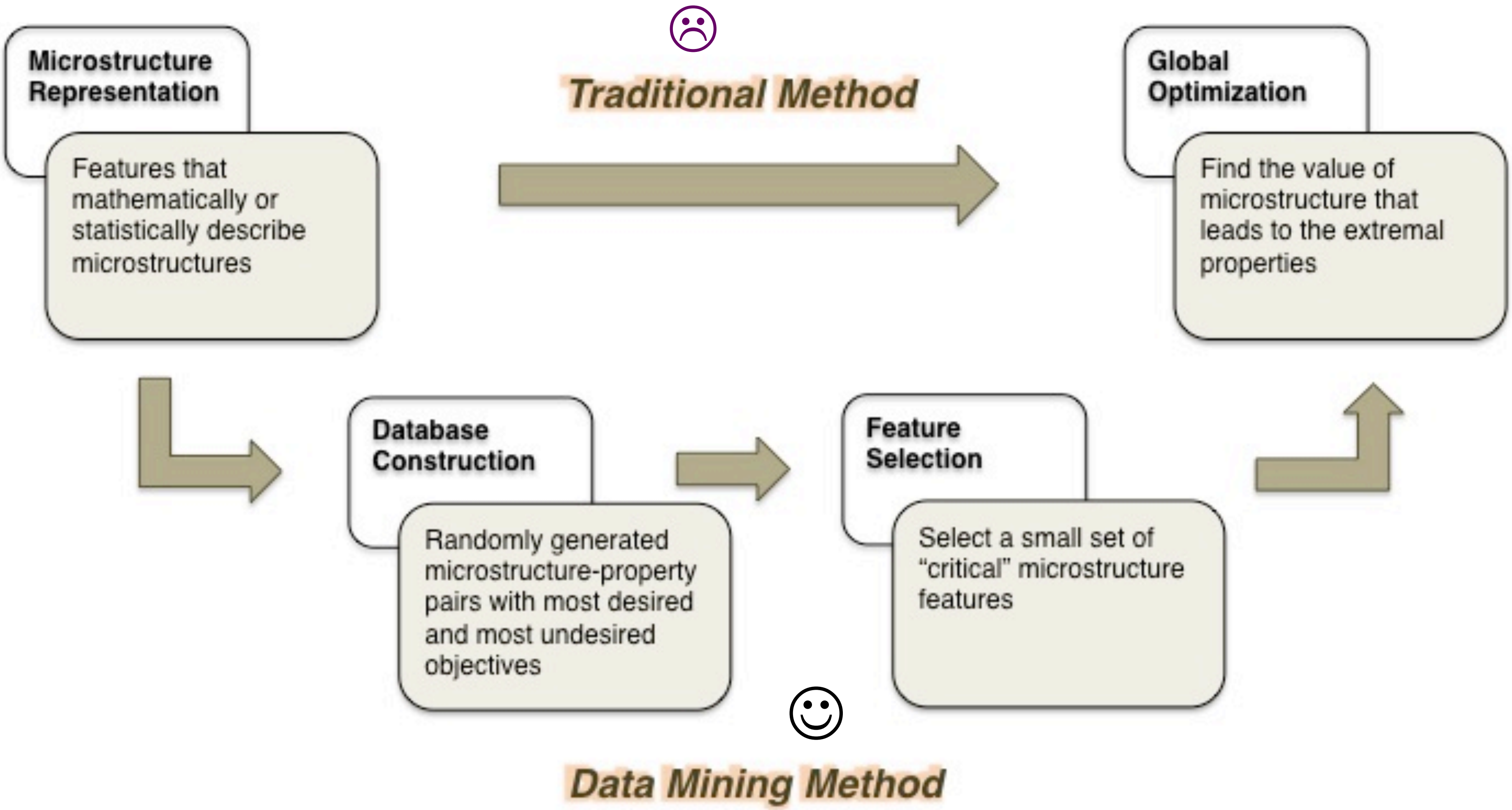
MATERIAL SCIENCE: A "DATA DRIVEN DISCOVERY" WORTH A THOUSAND SIMULATIONS?



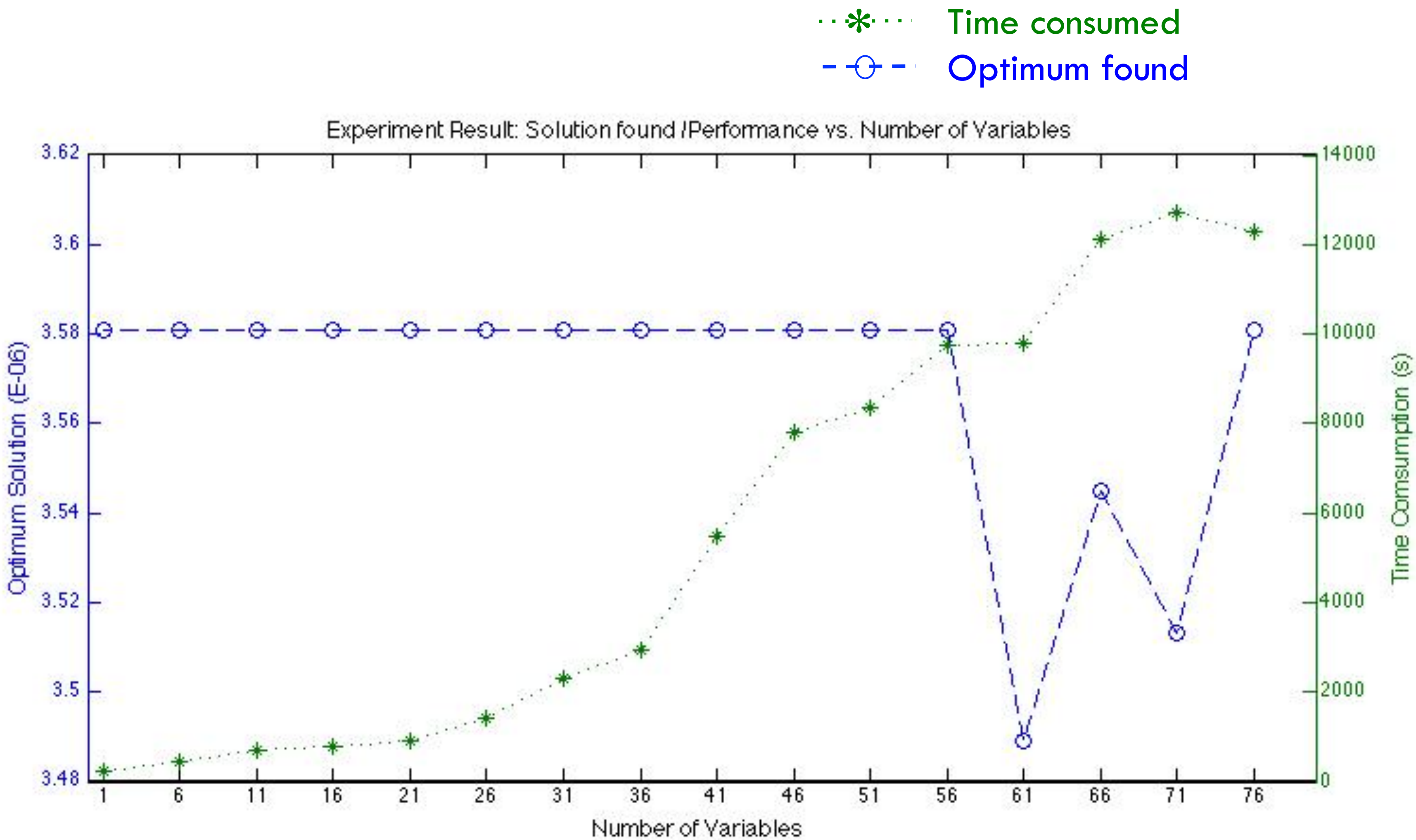
Discovery of stable compounds



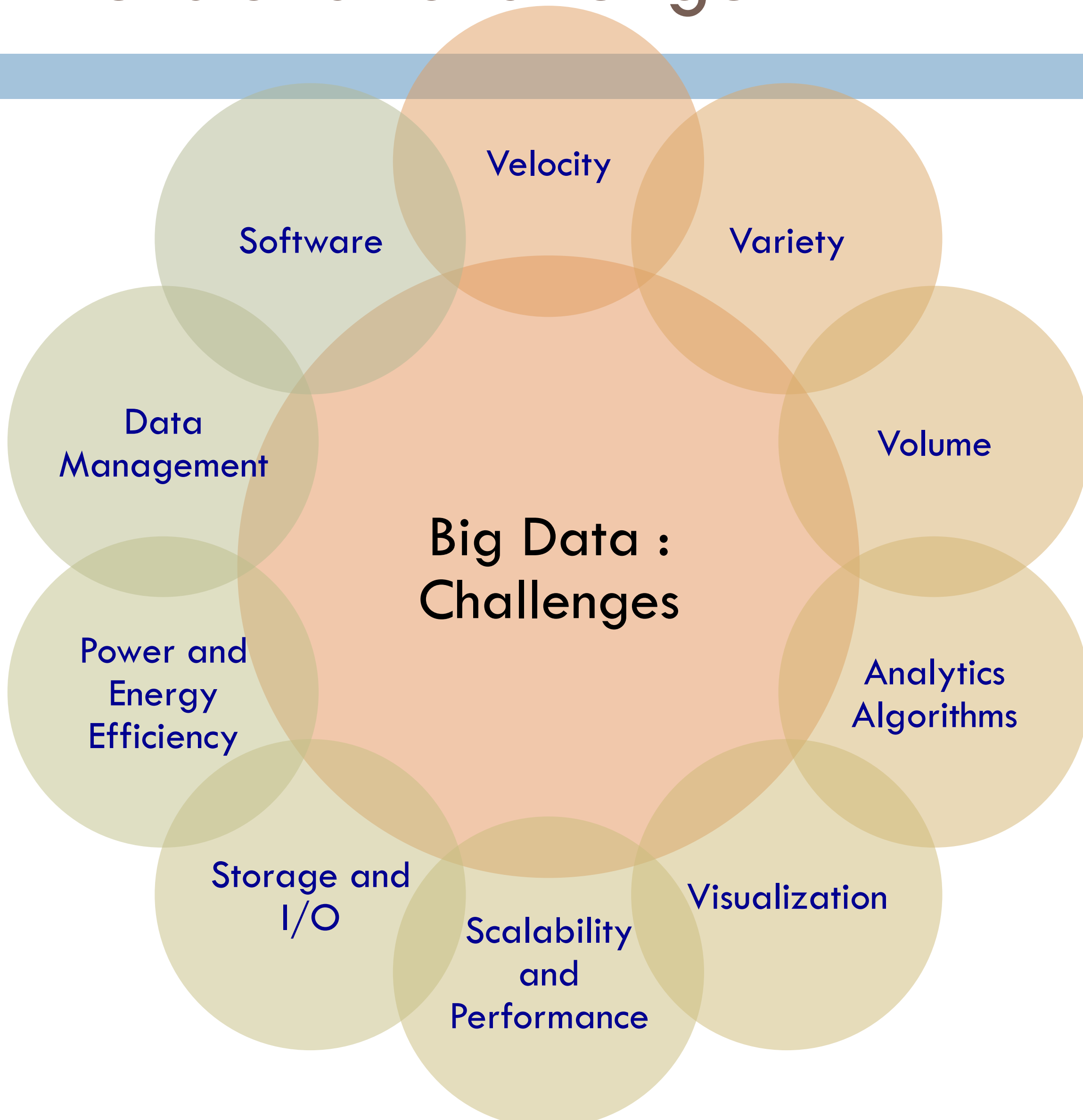
Structure-Property Optimization – Try optimization for 10^3 dimensions



Accelerating Time to Insights

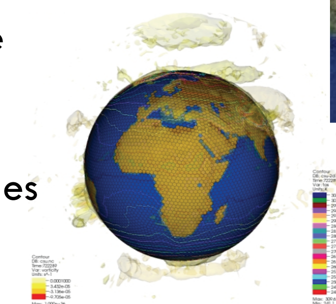


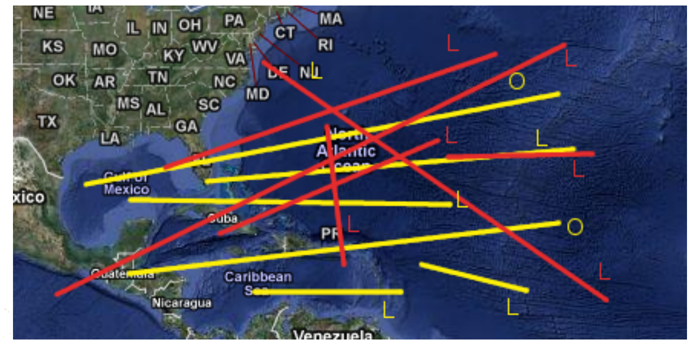
Extreme Computing + Big data : Not a single dimensional challenge



The Growth of Complexity → Need for Simplicity

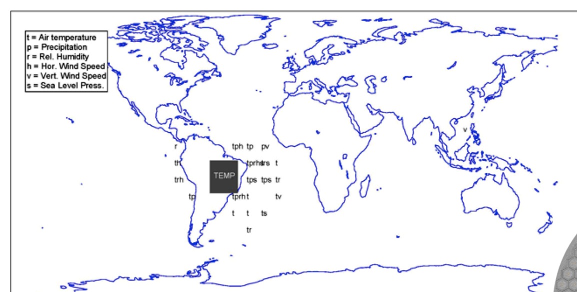
- **Higher spatial or temporal resolution**
  - extremes analysis
  - Network-based prediction
  - Estimation of spatiotemporal dependence
- **Higher data dimensionality**
  - Bayesian analysis of multi-model ensembles
  - Sampling-based statistical methods
  - Multivariate quantile analysis
- **Greater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





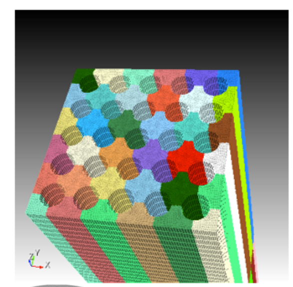
Significant correlations for hurricane prediction

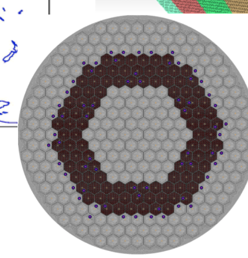
(Sencan, Chen, Hendrix, Pansombut, Semazzi, Choudhary, Kumar, Melechko, and Samatova, 2011)



Prediction of land climate using ocean climate variables

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



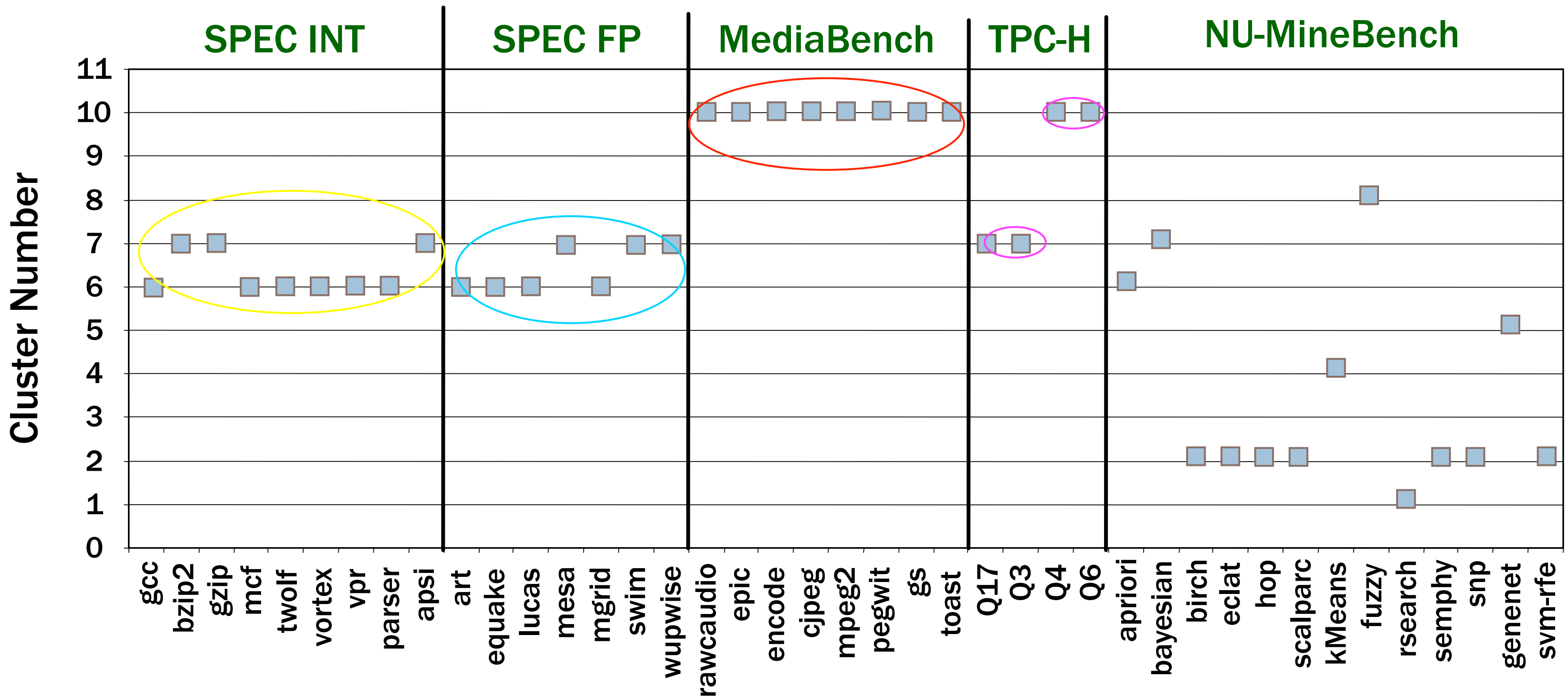


Right Computing infrastructure? What characteristics do typical analytics functions have?

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

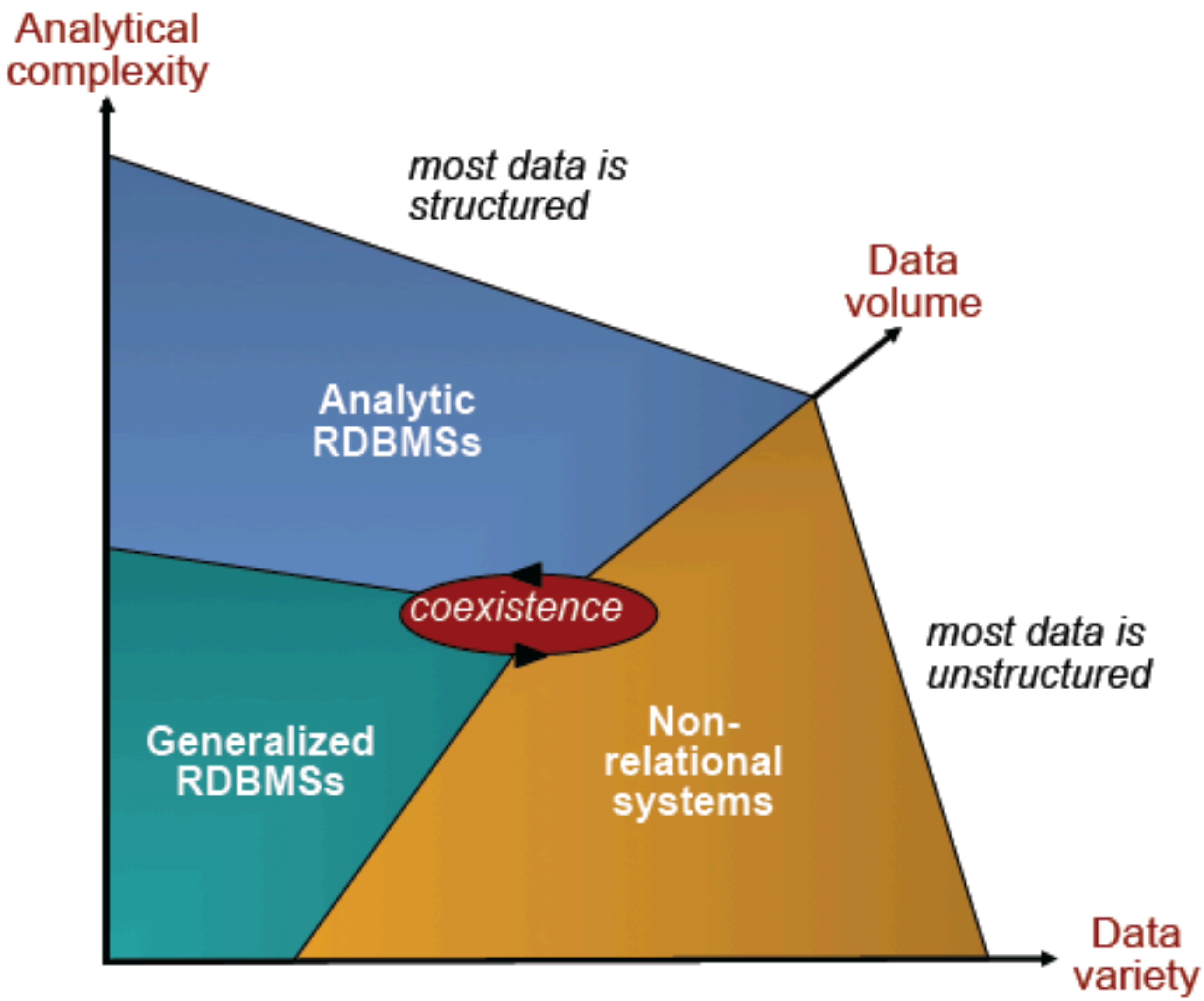
† The numbers shown here for the parameters are values per instruction

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

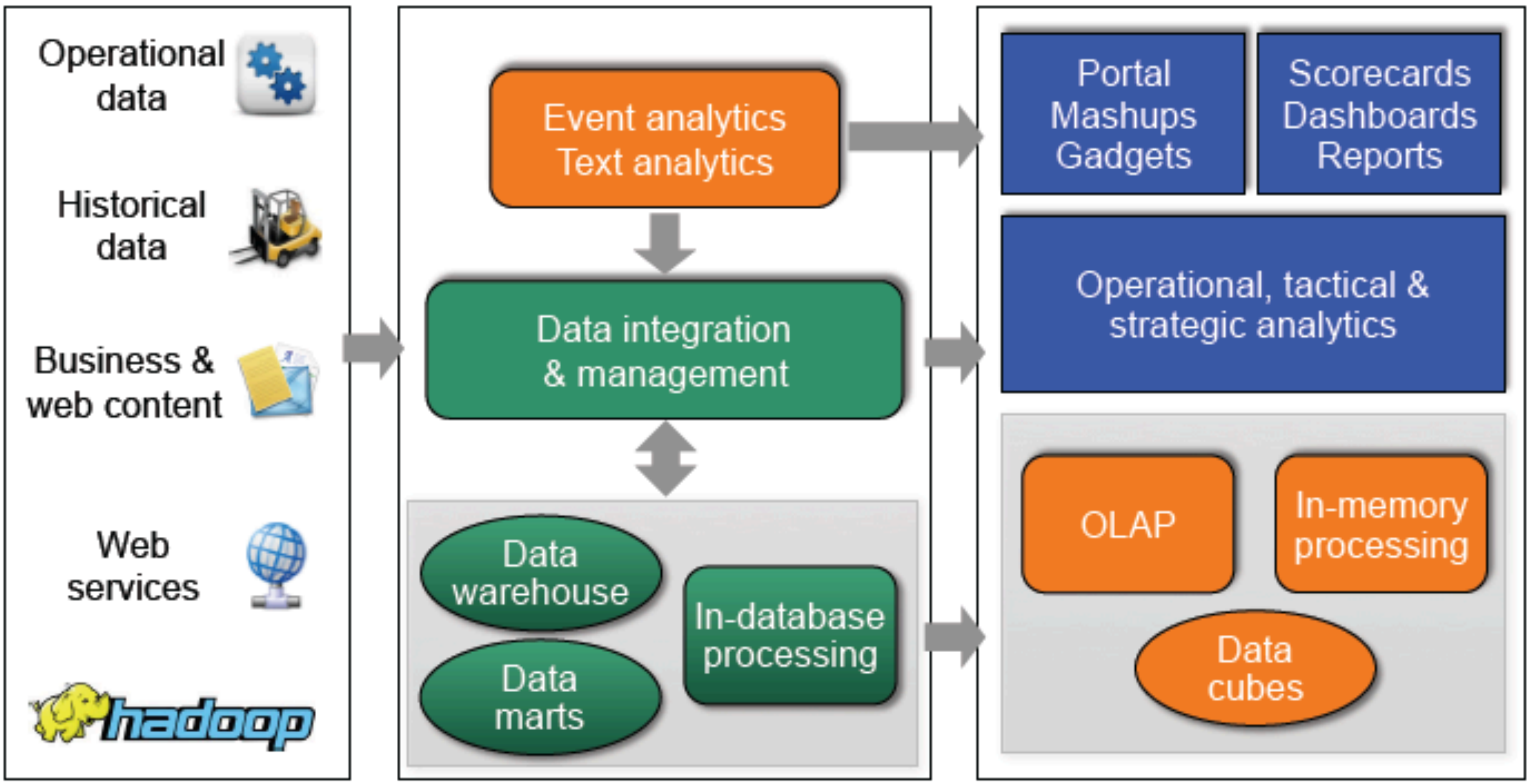


Clear Implications on architecture, modes, memory hierarchy and other components
Identify similarities and design for co-existence

Big Data: Generalization and Optimizations



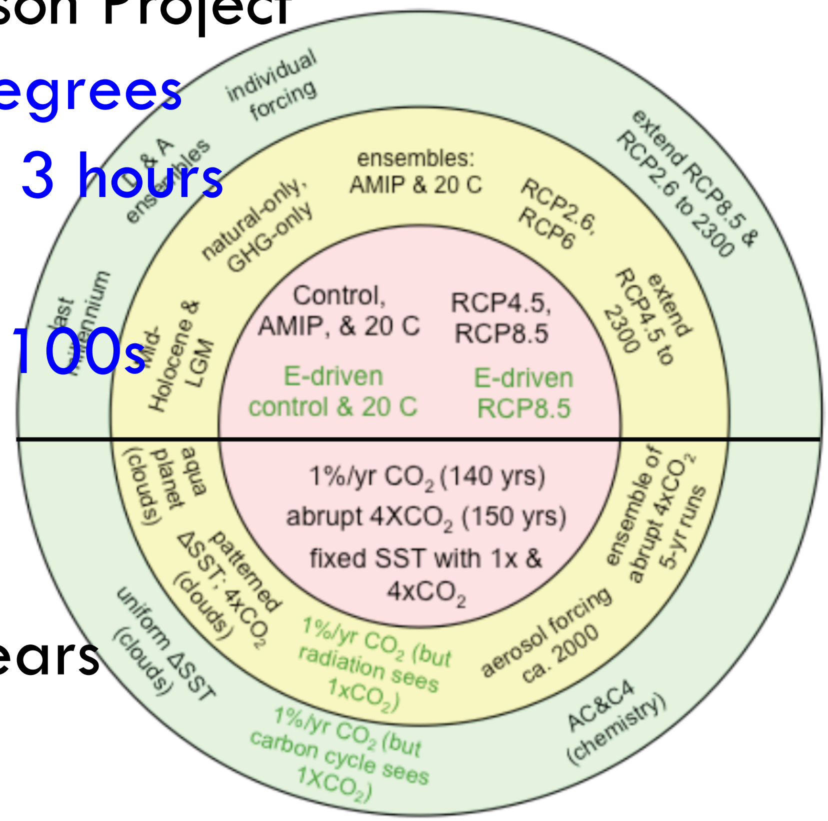
Data → Information → Insights → Actions



Create a suite of Mini - Analytics Apps?

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

- 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

An instrument and a discovery engine

Millions of cores

Each core is like a sensor

Each core generates data based on a model

…Millions of cores

Each core can be a data processor/analyst

Extreme scale system can be a discovery engine

NO other type of sensor can claim this capability!

Data Analytics Algorithms – 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, Mahalanobis distance, Jaccard coefficient, Tanimoto coefficient, shortest paths
Cosmology	Halos and sub-halos	SUBFIND, AHF

Thank You!

Alok Choudhary
John G. Searle Professor
Dept. of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu