Big Data + Extreme-scale

Time to Compute → Actionable Insights

Alok Choudhary

John G. Searle Professor

Dept. of Electrical Engineering and Computer Science

and Professor, Kellogg School of Management

Northwestern University

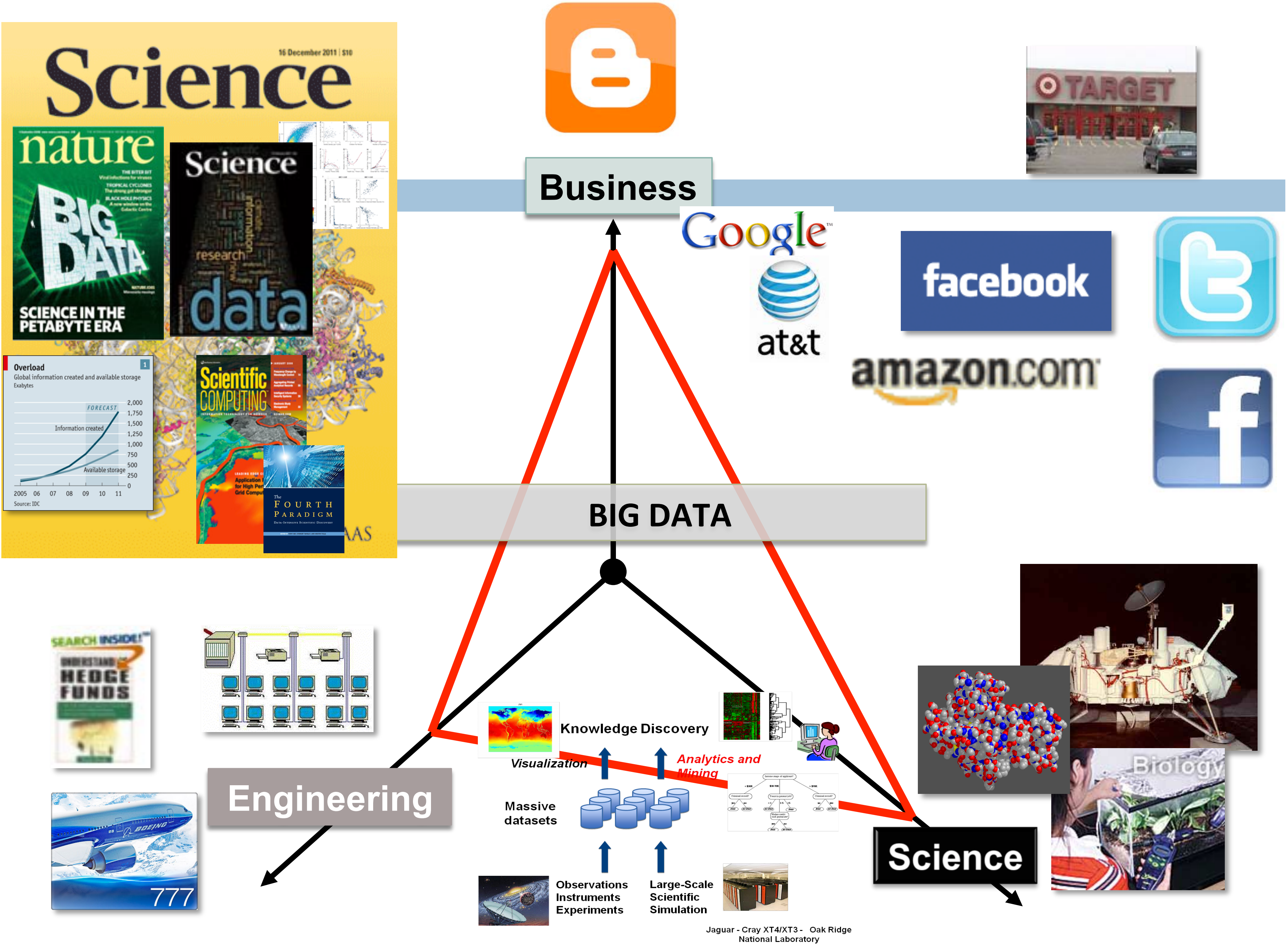
choudhar@eecs.northwestern.edu

National Science Foundation
WHERE DISCOVERIES BEGIN

ACKNOWLEDGEMENTS

U.S. DEPARTMENT OF
ENERGY

BIG DATA?



Business

BIG DATA

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

“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

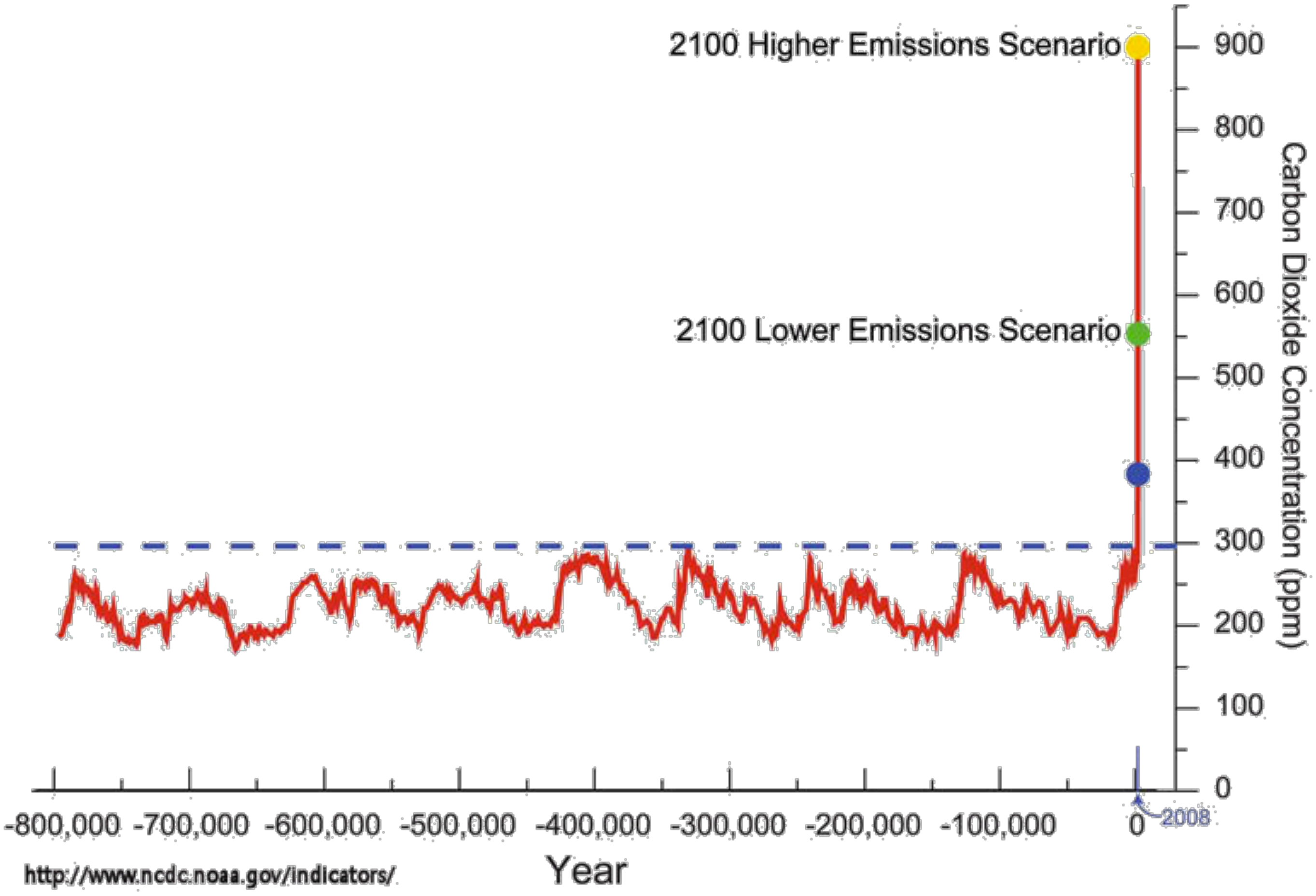
Understanding Climate Change

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

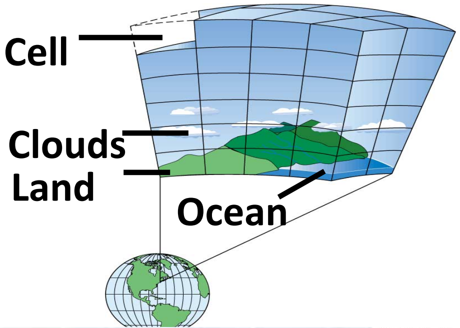


http://www.ncdc.noaa.gov/indicators/

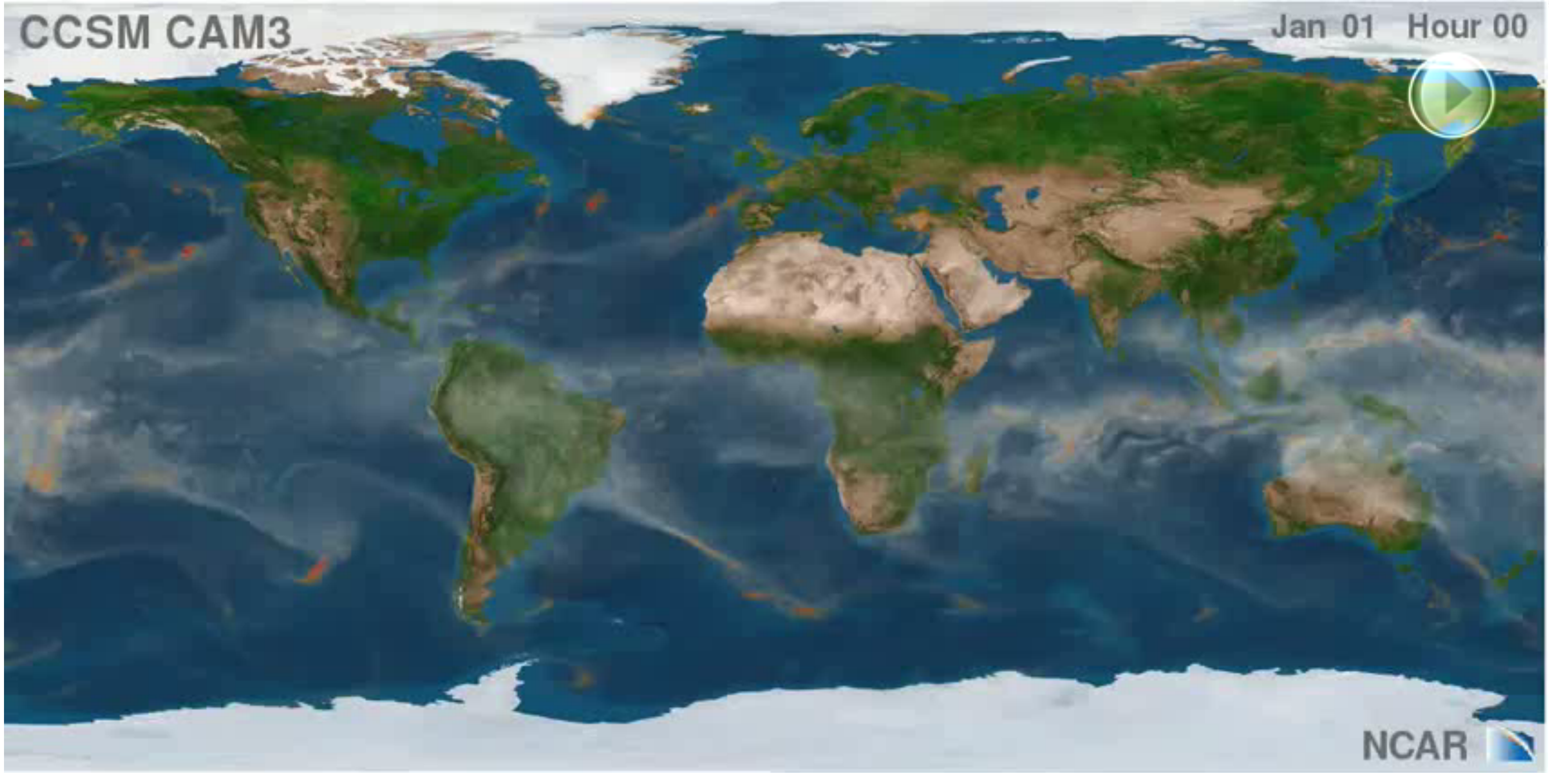
Understanding Climate Change – Physics-Based Approach

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

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

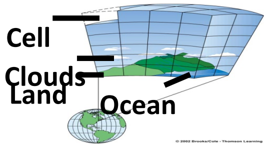


- Thomson Learning



Understanding Climate Change - Physics Based Approach

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



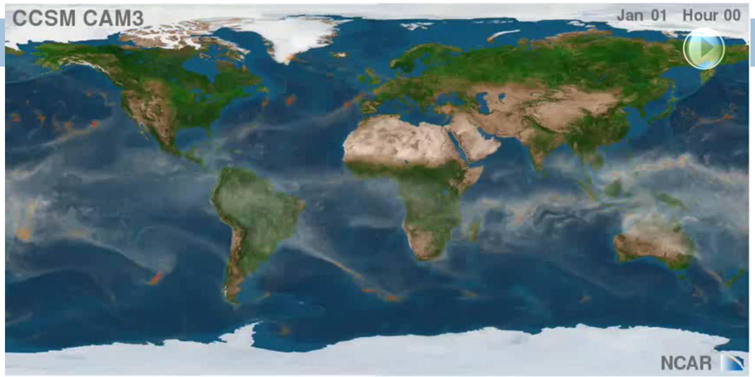
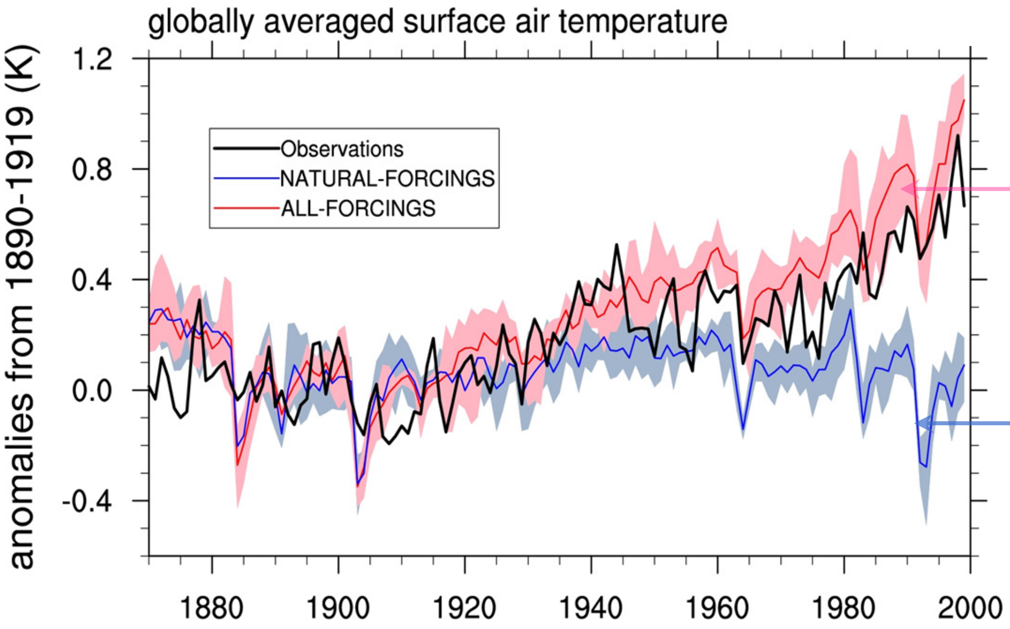


Figure Courtesy: NCAR



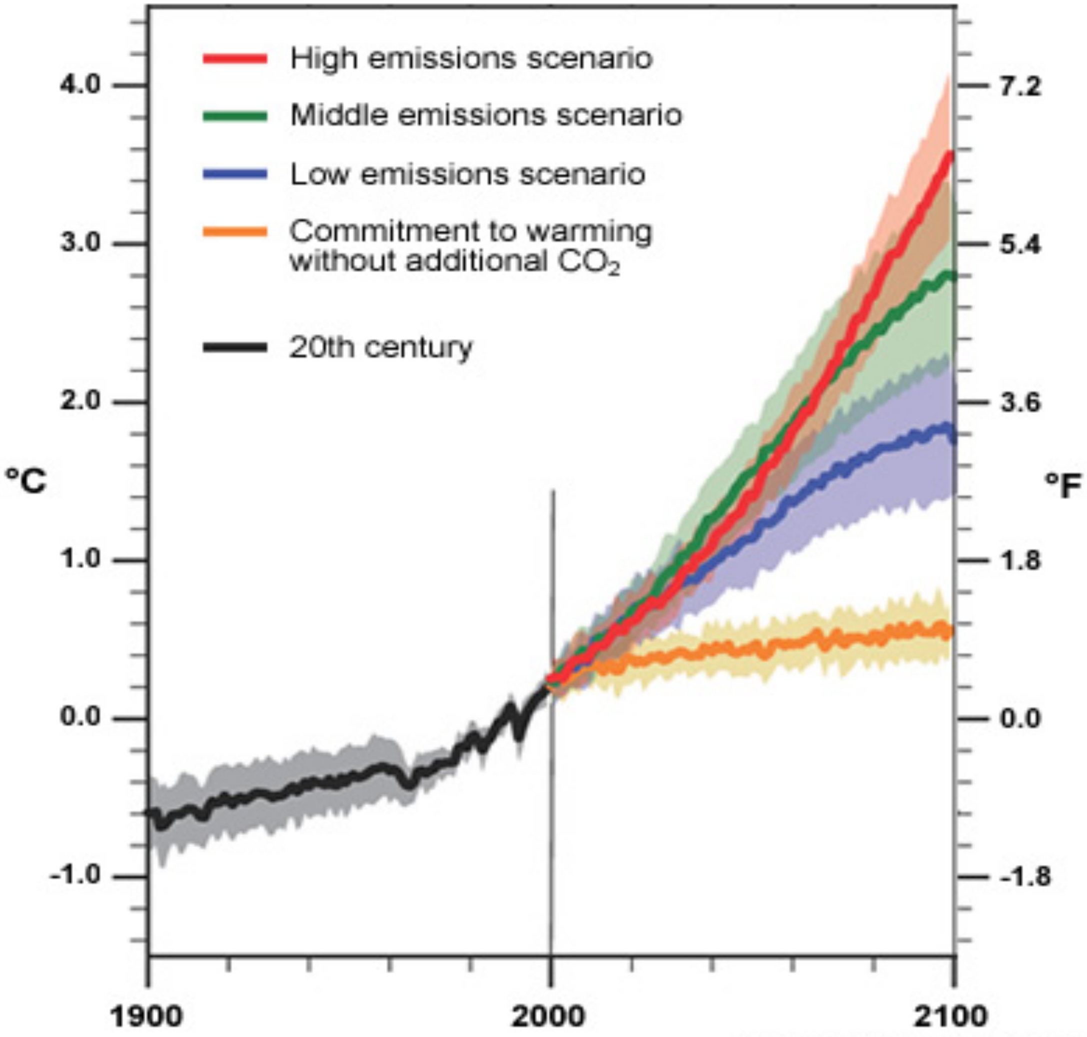
Ensemble average with observed greenhouse gas concentrations

Ensemble average with pre-industrial greenhouse gas concentrations

Figure Courtesy: ORNL

Understanding Climate Change - Physics Based Approach

Temperature Increases for Various Emission Scenarios



IPCC / The COMET Program

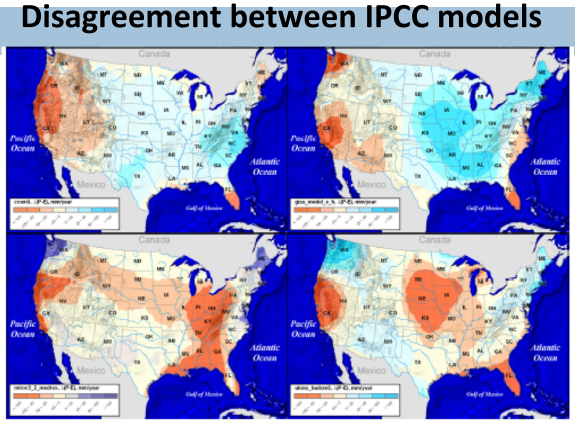
Projection of temperature increase under different **Special Report on Emissions Scenarios** (SRES) by 24 different GCM configurations from 16 research centers used in the **Intergovernmental Panel on Climate Change** (IPCC) 4th Assessment Report.

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

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

Disagreement between IPCC models

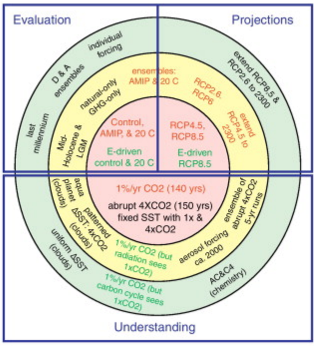


Regional hydrology exhibits large variations among major IPCC model projections

Physics based models

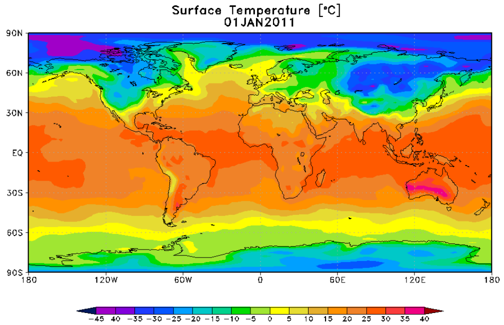
Low uncertainty	High uncertainty
Temperature	Hurricanes
Pressure	Extremes
Large-scale wind	Precipitation

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

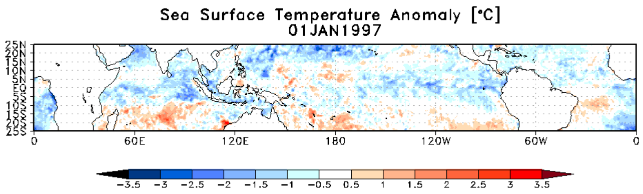
> "Climate change research is now 'big science,' comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics."
>
> **(Nature Climate Change, Oct 2012)**

Need for data driven discovery

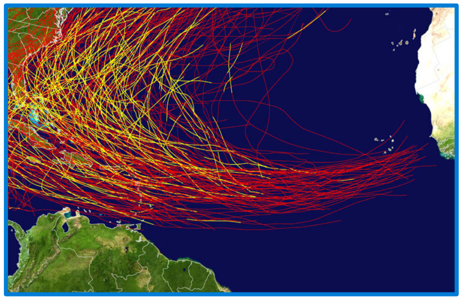
Physics based models

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

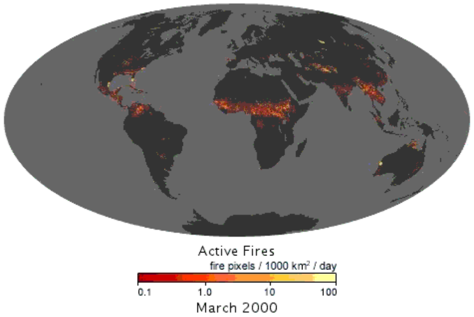
Global sea surface temperatures



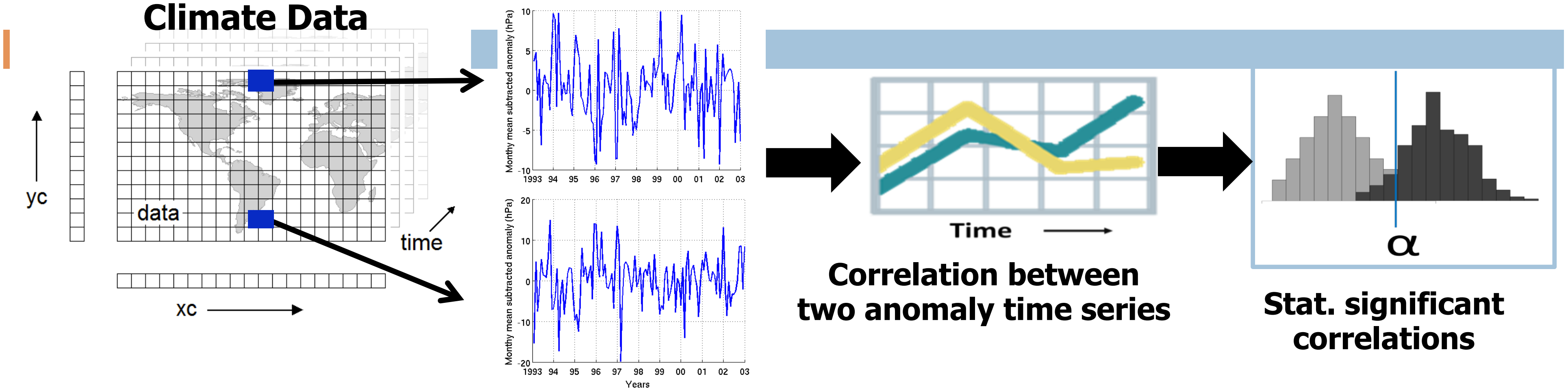
Atlantic hurricanes



Global fires



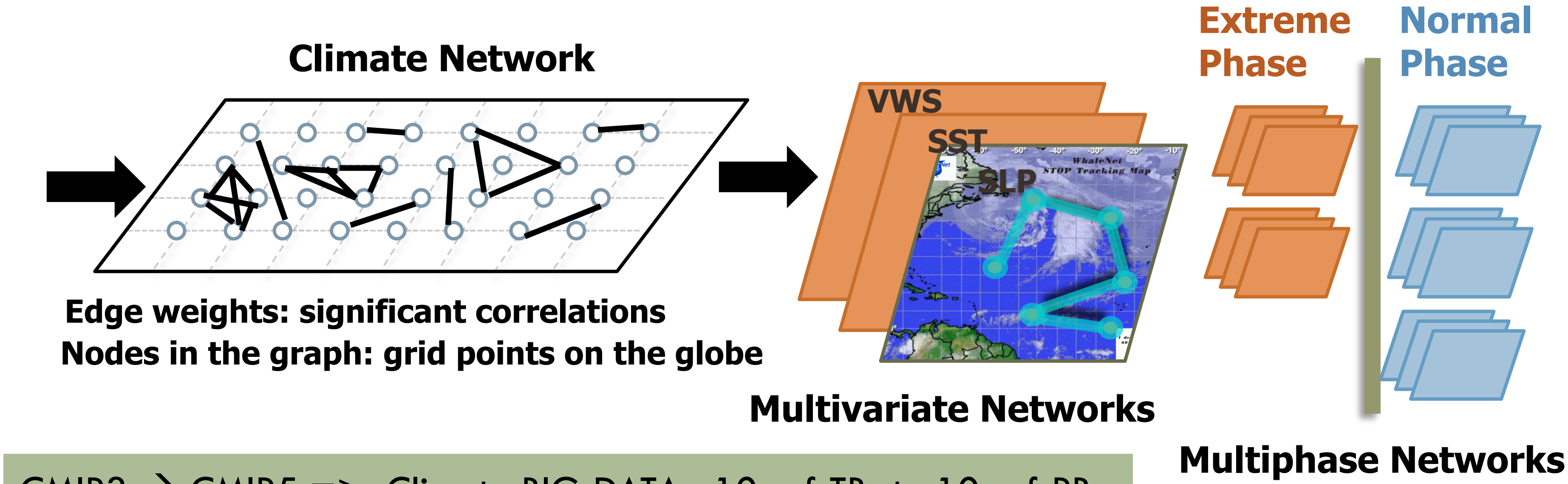
End-to-End: From Transactional analytics to relationship mining


Climate Data

Correlation between two anomaly time series

Stat. significant correlations

Anomaly time series at each node


Climate Network

Edge weights: significant correlations
Nodes in the graph: grid points on the globe

Multivariate Networks

Extreme Phase

Normal Phase

Multiphase Networks

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

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

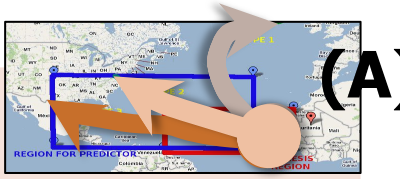
Validate Hypothesis

- 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

Copyright Anglo-Australian Observatory/Royal Observatory, Edinburgh

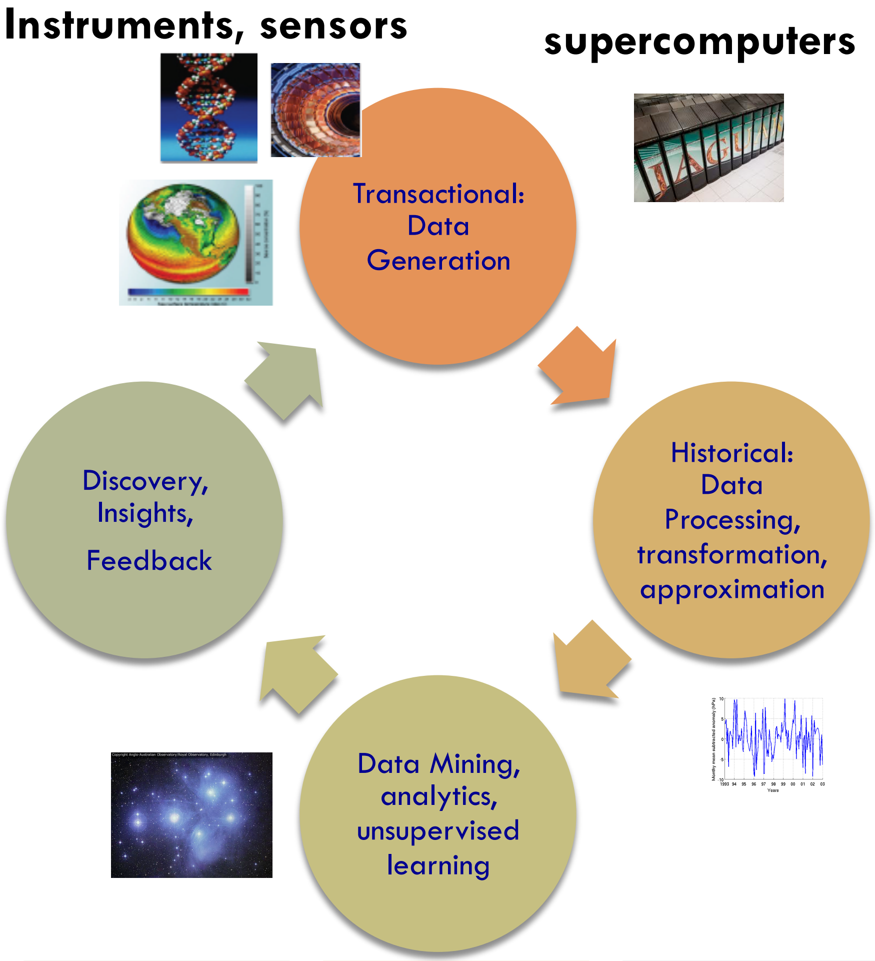
- Wow! I found a new galaxy?
- Switch C fails when switch A fails followed by switch B failing
- On Thursday people buy beer and diaper together.
- The ratio K/P > X is an indicator of onset of diabetes.

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.

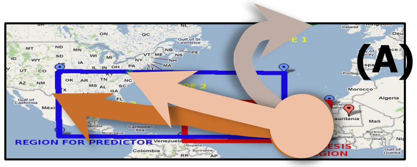
—Feb. 12, 2002, Department of Defense news briefing by Donald Rumsfeld

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


Instruments, sensors
supercomputers
Transactional: Data Generation
Historical: Data Processing, transformation, approximation
Data Mining, analytics, unsupervised learning
Discovery, Insights, Feedback

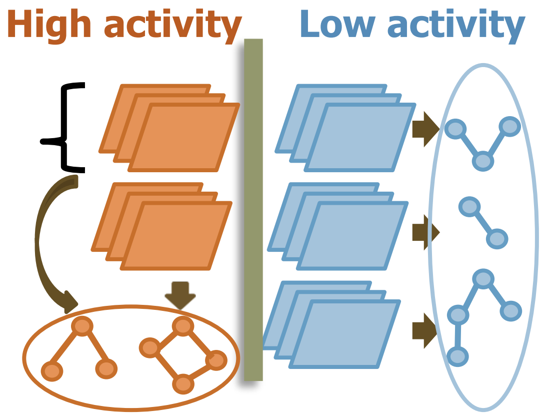
- Data Management
- Data Reduction, Query
- Data Visualization
- Data Sharing

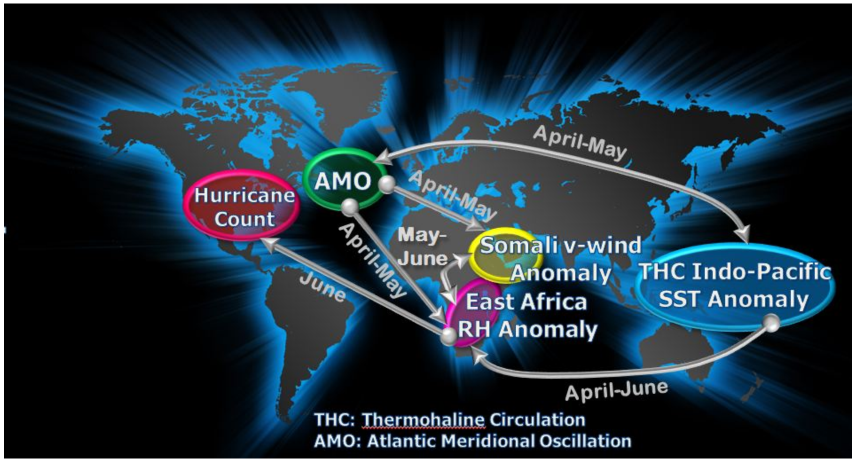
Historical data → Learning Models → Trigger/ questions → Predict



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)

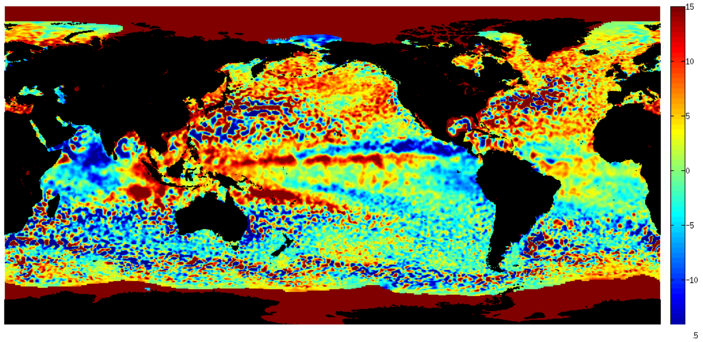
Chen et al. *Data Mining & Knowledge Discovery* (2012)

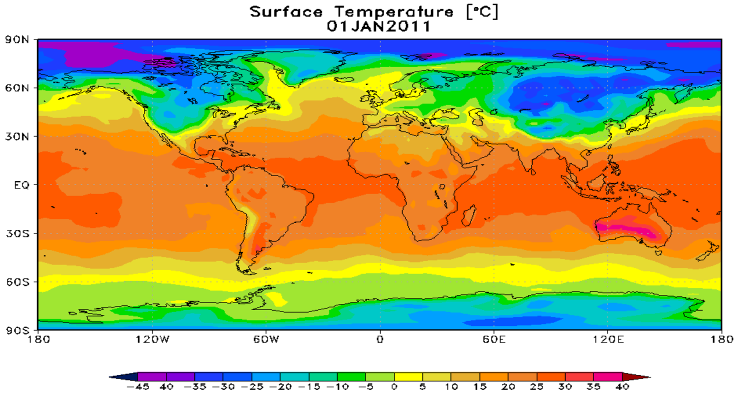
Chen *et al. SIAM SDM* (2013)

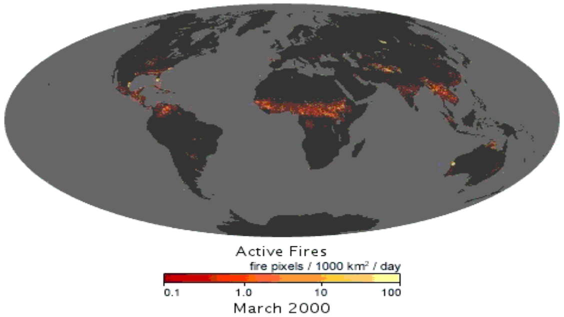
Chen *et al. IJCAI* (2013)

Semazzi *et al.* in review at journal (2013)

Challenges in data driven analysis



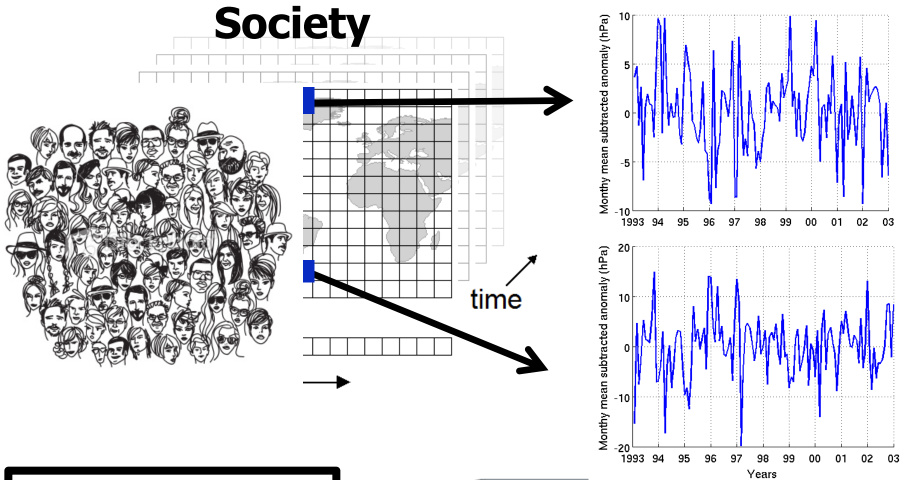




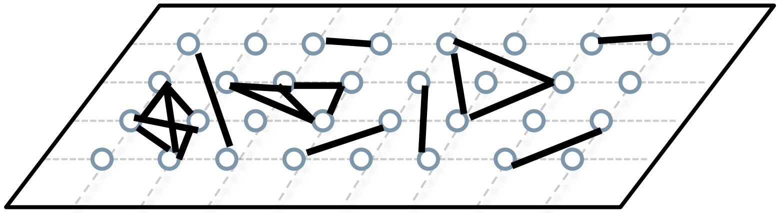
- Complex dependence
  - Non-IID
  - Spatio-temporal correlation
  - Long memory in time
  - Long range dependence in space
  - Nonlinear relationships
- Data characteristics
  - Heterogeneous, Multivariate
  - Heavy Tailed Distributions
  - Noisy, incl. low frequency variability
  - Paucity of training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

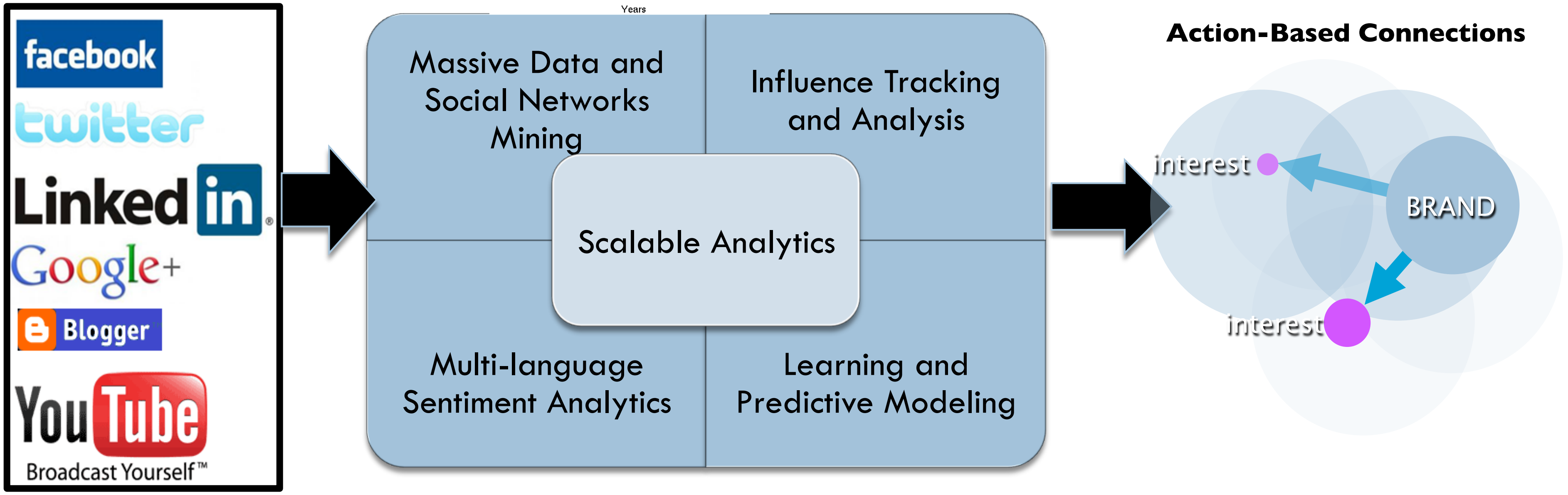


Activity/interaction based Network



Edge weights: significant interactions/influence

Nodes in the graph: people/brands/…



Massive Data and Social Networks Mining

Influence Tracking and Analysis

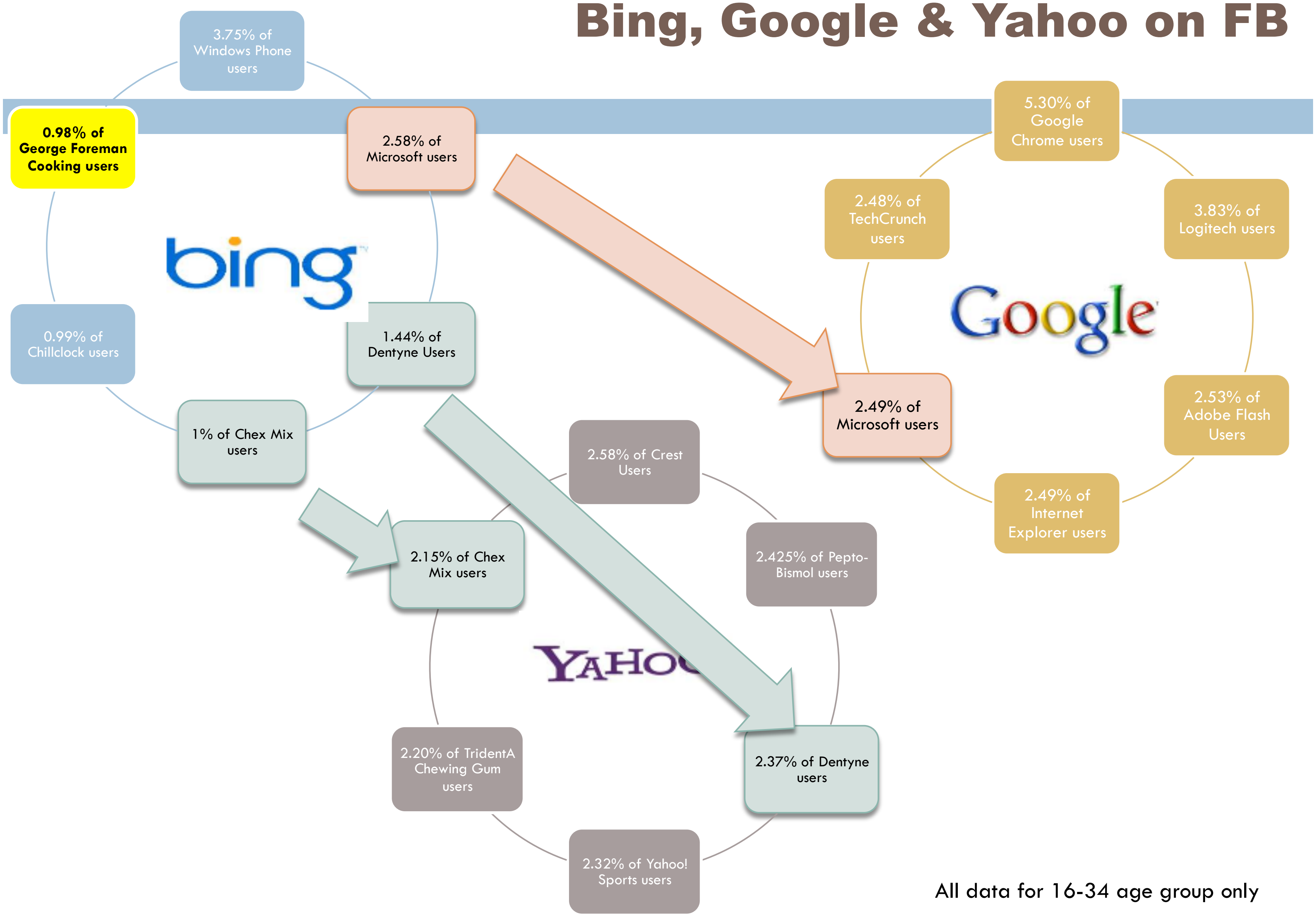
Scalable Analytics

Multi-language Sentiment Analytics

Learning and Predictive Modeling

Action-Based Connections

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



3.75% of Windows Phone users

2.58% of Microsoft users

1.44% of Dentyne Users

1% of Chex Mix users

0.99% of Chillclock users

0.98% of George Foreman Cooking users

5.30% of Google Chrome users

3.83% of Logitech users

2.53% of Adobe Flash Users

2.49% of Internet Explorer users

2.49% of Microsoft users

2.48% of TechCrunch users

2.58% of Crest Users

2.425% of Pepto-Bismol users

2.37% of Dentyne users

2.32% of Yahoo! Sports users

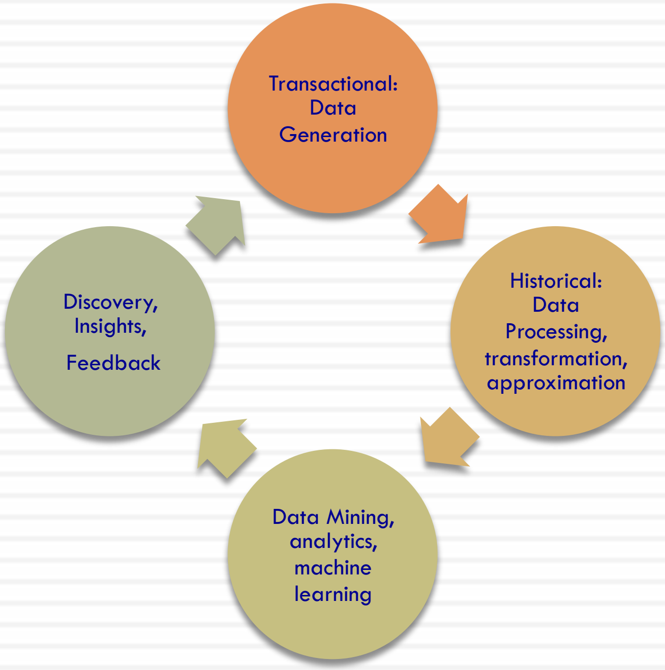
2.20% of TridentA Chewing Gum users

2.15% of Chex Mix users

All data for 16-34 age group only

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

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



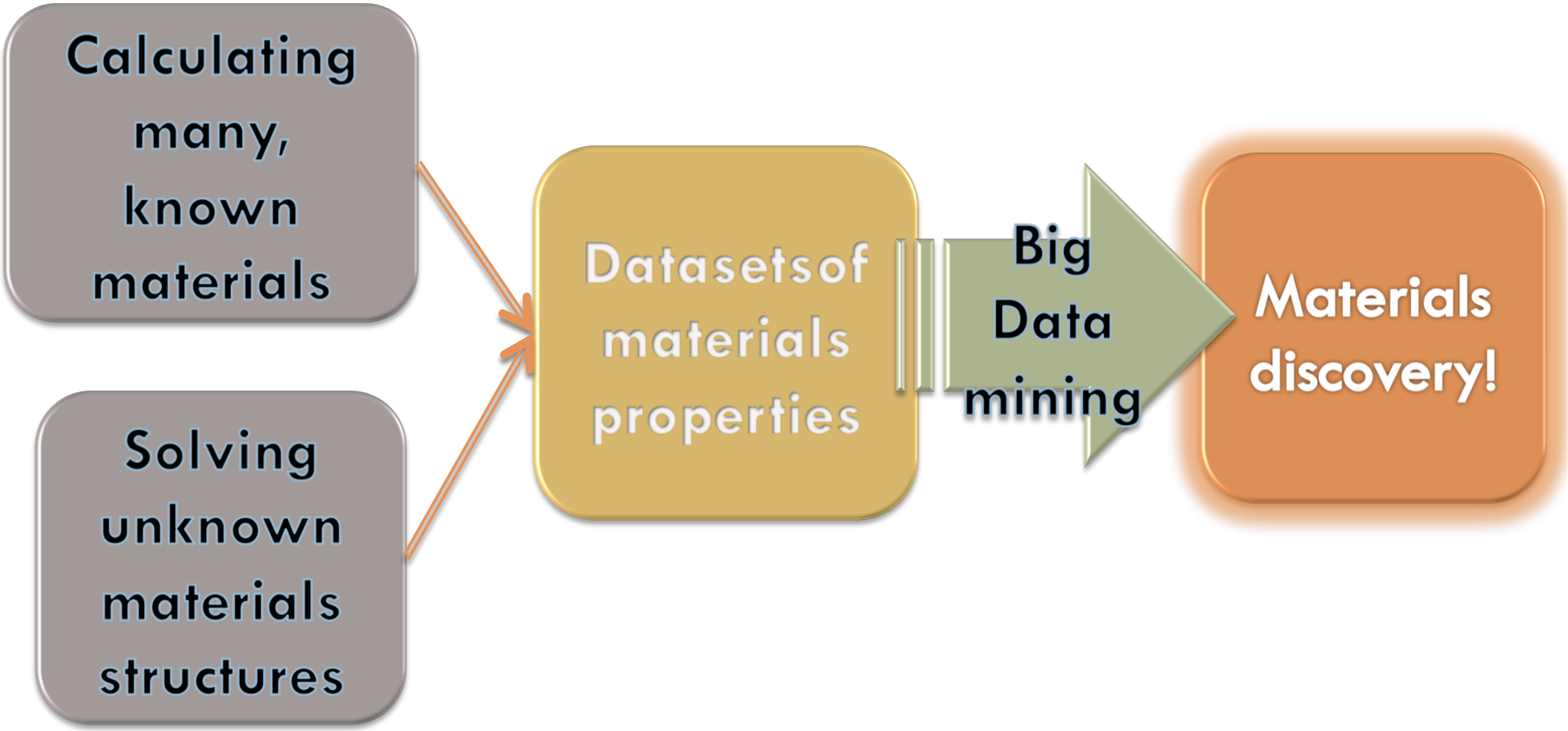
Transactional: Data Generation

Historical: Data Processing, transformation, approximation

Data Mining, analytics, machine learning

Discovery, Insights, Feedback

Discovery of stable compounds



Calculating many, known materials

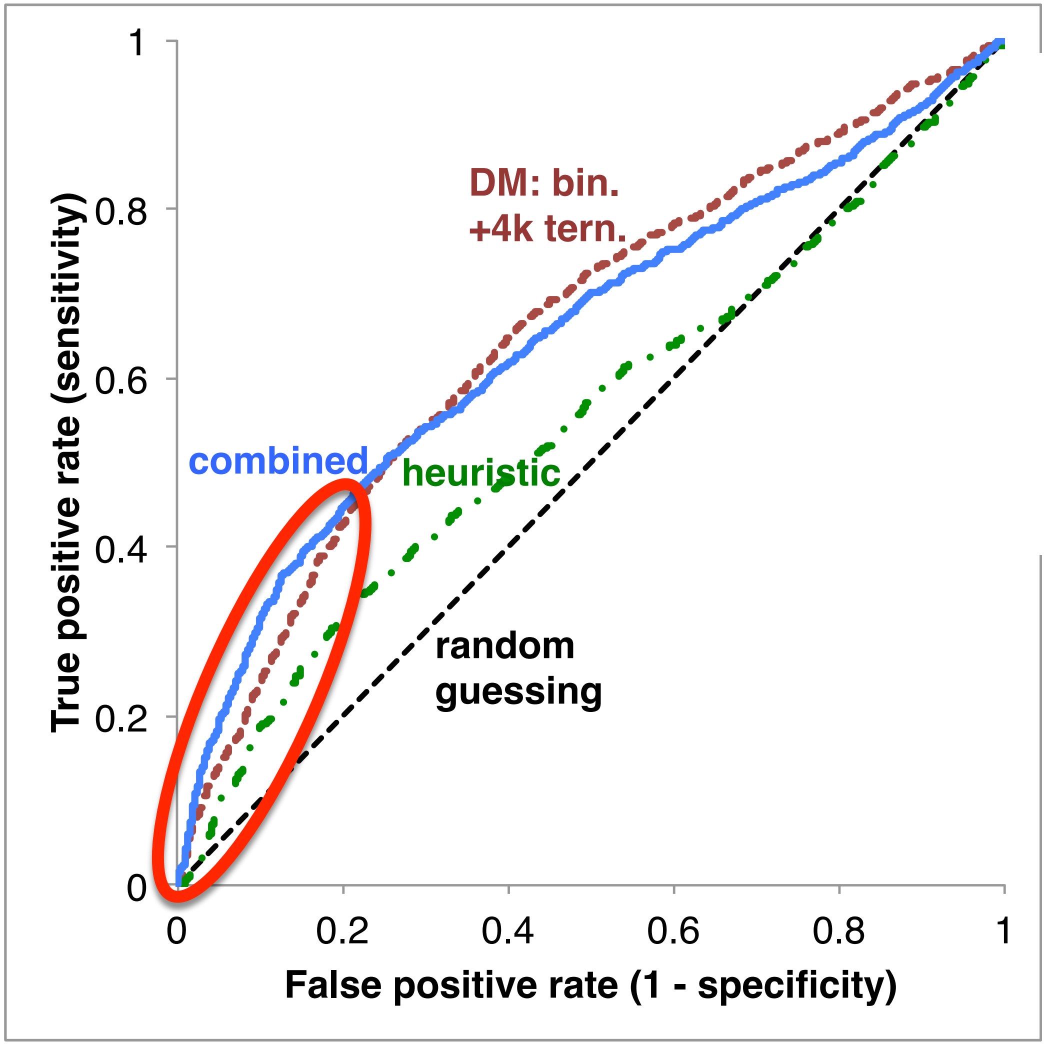
Solving unknown materials structures

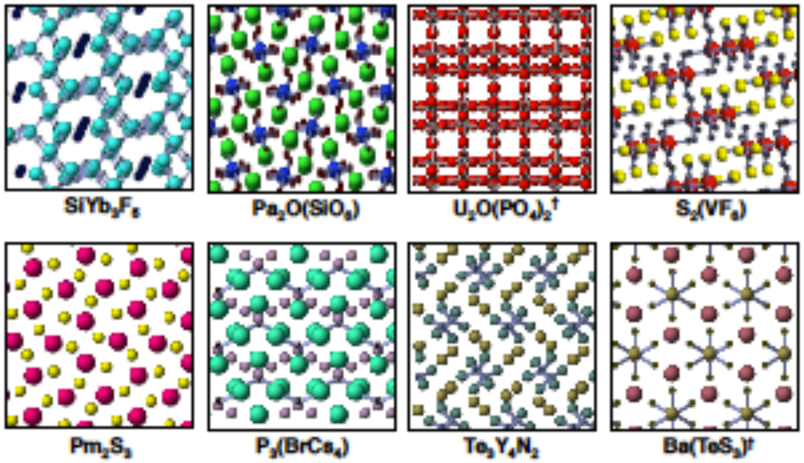
Datasetsof materials properties

Big Data mining

Materials discovery!

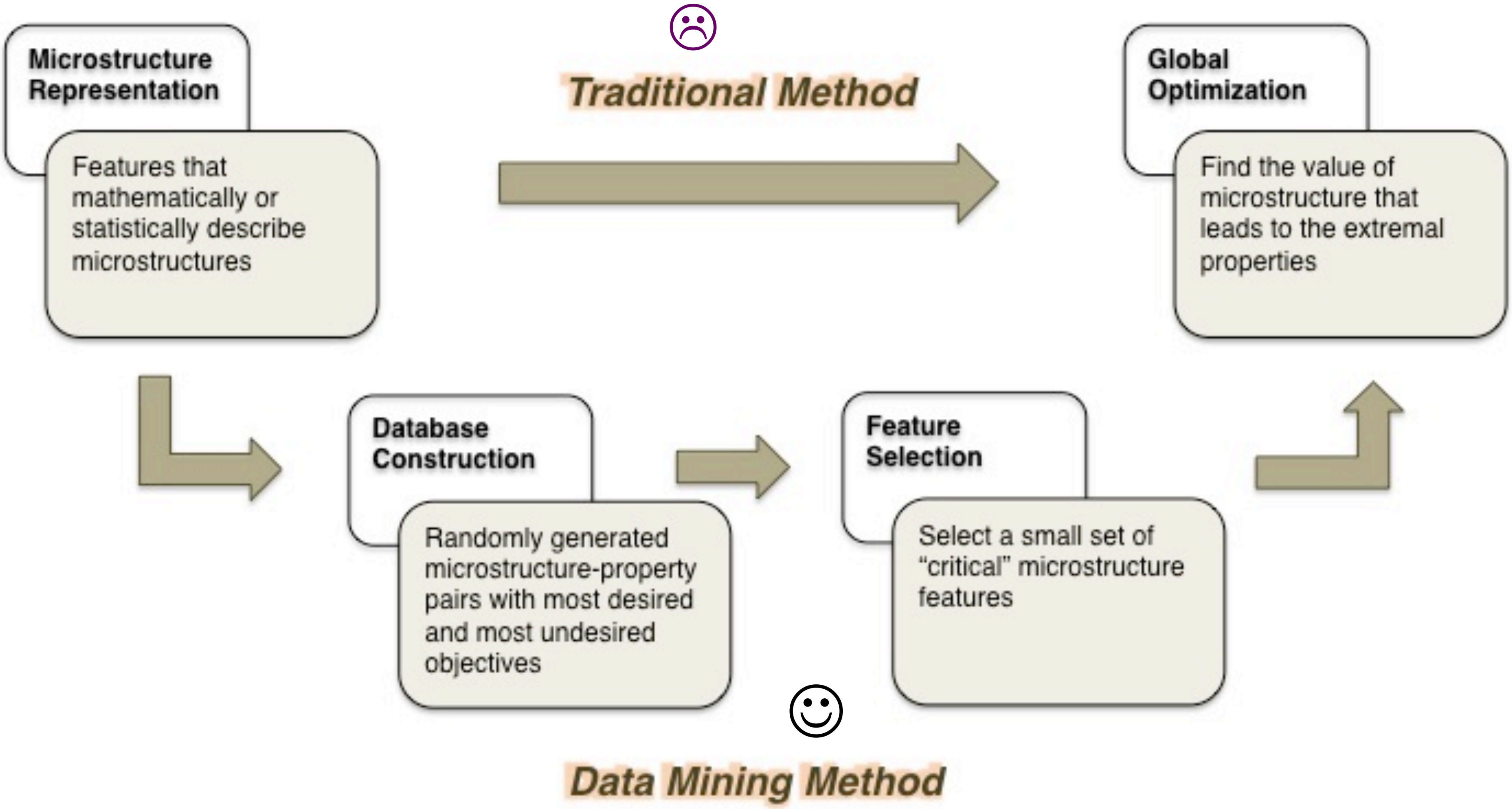
Ranking – Approximation is good enough for ranking ☺ (closing the loop)





† 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.

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



☹

Traditional Method

Microstructure Representation

Features that mathematically or statistically describe microstructures

Database Construction

Randomly generated microstructure-property pairs with most desired and most undesired objectives

Feature Selection

Select a small set of "critical" microstructure features

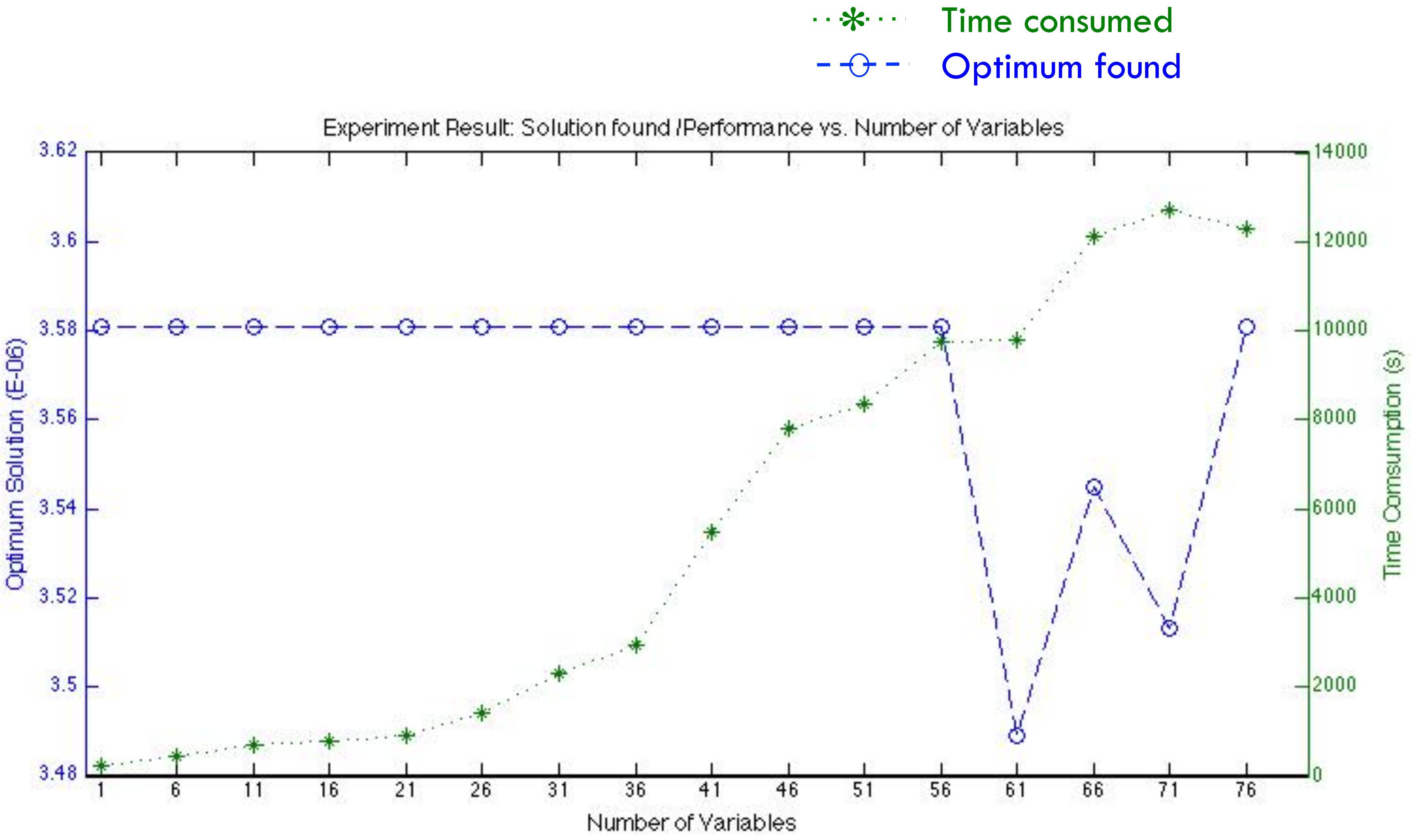
Global Optimization

Find the value of microstructure that leads to the extremal properties

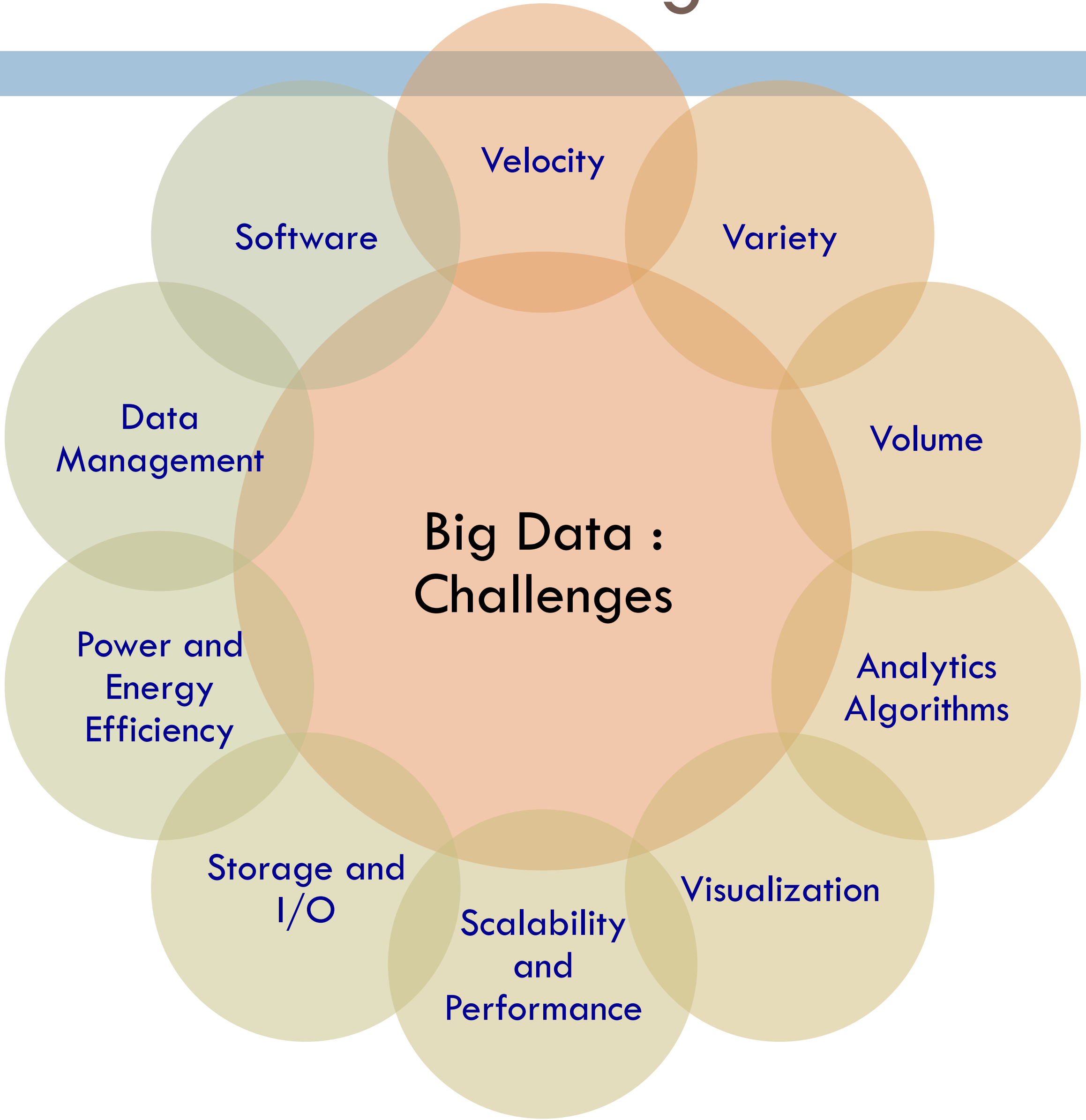
☺

Data Mining Method

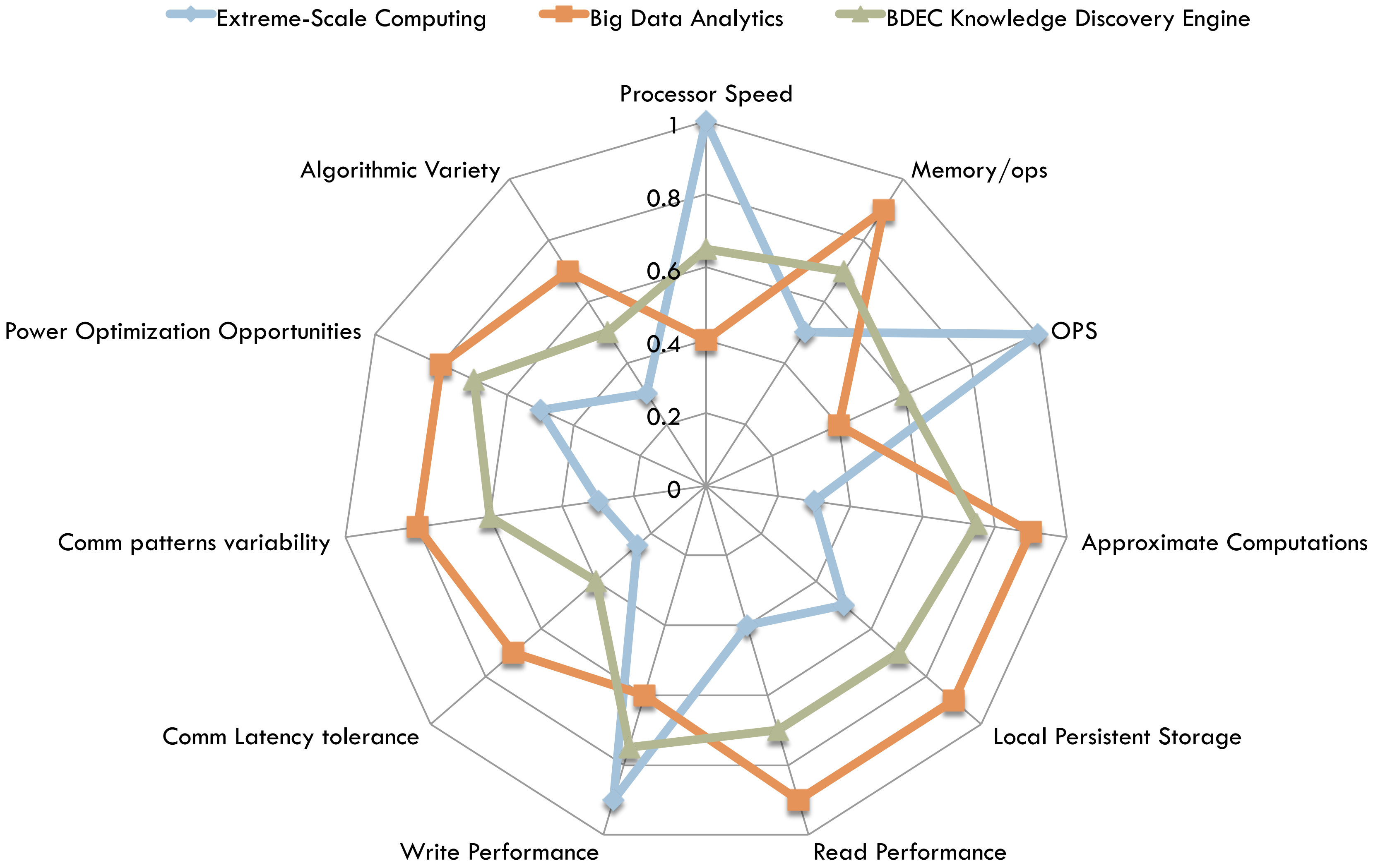
Accelerating Time to Insights



Extreme Computing + Big data : Not a single dimensional challenge



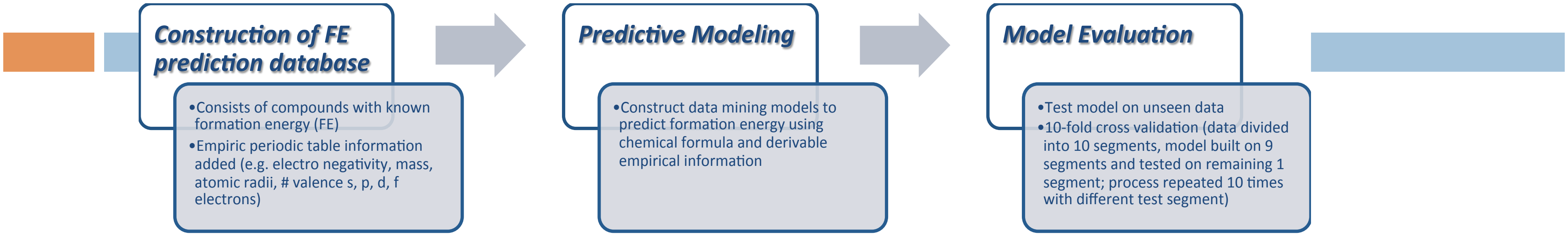
Extreme Computing + Big Data Analytics = A Knowledge Discovery Engine?



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

Discovering Materials : Simulations → Analytics



Construction of FE prediction database

- Consists of compounds with known formation energy (FE)
- Empiric periodic table information added (e.g. electro negativity, mass, atomic radii, # valence s, p, d, f electrons)

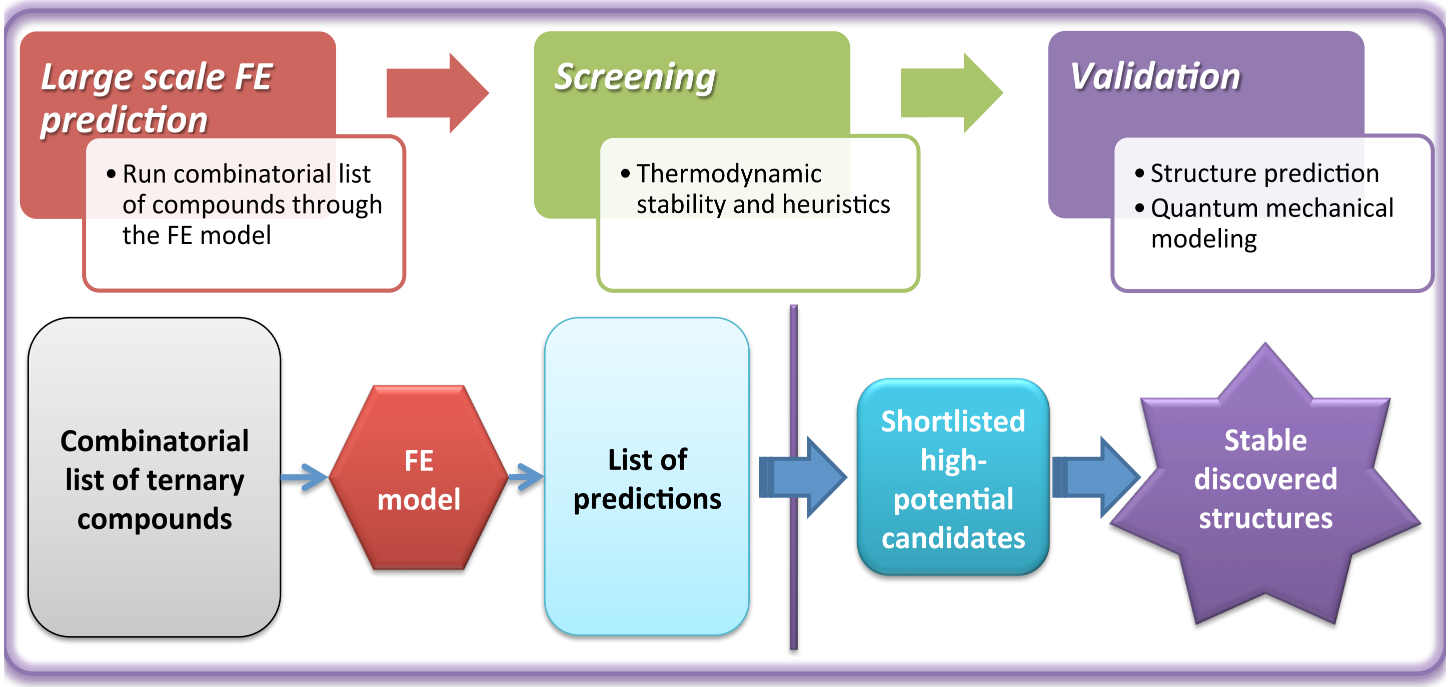
Predictive Modeling

- Construct data mining models to predict formation energy using chemical formula and derivable empirical information

Model Evaluation

- Test model on unseen data
- 10-fold cross validation (data divided into 10 segments, model built on 9 segments and tested on remaining 1 segment; process repeated 10 times with different test segment)

(a)



Large scale FE prediction

- Run combinatorial list of compounds through the FE model

Screening

- Thermodynamic stability and heuristics

Validation

- Structure prediction
- Quantum mechanical modeling

Combinatorial list of ternary compounds
FE model
List of predictions
Shortlisted high-potential candidates
Stable discovered structures

(b)

Climate Change → Analytics Challenges

Process Understanding	Extreme Events - Heat Waves - Rainfall Extremes - Droughts - Hurricanes Model Evaluation Downscaling - Statistical - Dynamical Ocean-Atm.-Land Interactions	Change Detection - Abrupt vs. Gradual - Point vs. Regions/Intervals - Change in Extremes Spatio-Temporal Classification Sparse/High-Dim. Methods Causal Relationships Networks/Graphs HPC	Computational Innovations
Understanding Climate Change			