# Big Data's Biggest Needs- Deep Analytics for Actionable Insights

**Alok Choudhary**

**John G. Searle Professor**

Dept. of Electrical Engineering and Computer Science

and Professor, Kellogg School of Management

Northwestern University

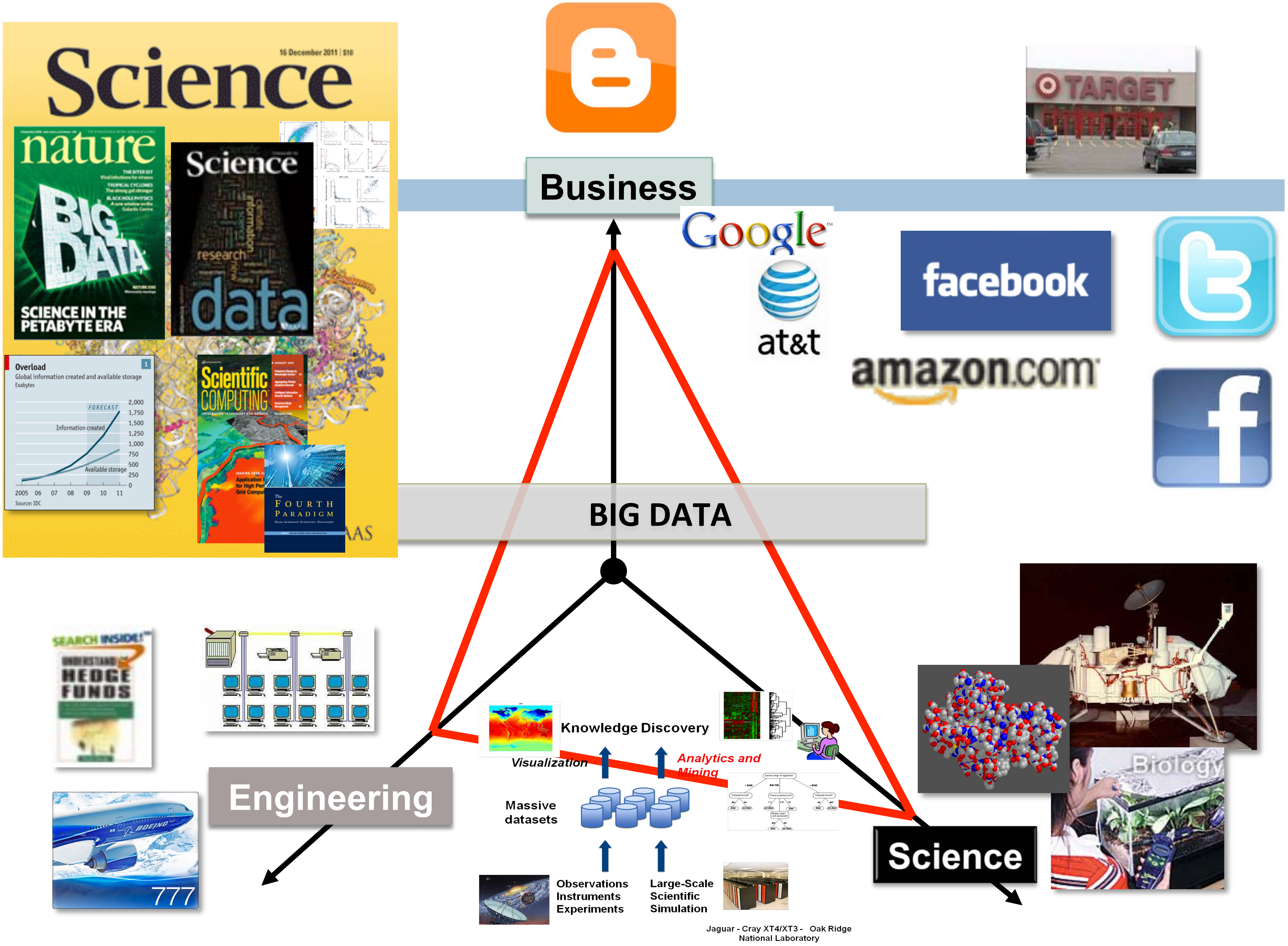
choudhar@eecs.northwestern.edu

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**

# Data Mining, Analytics and Actionable Insights?

Time to Compute → Time to 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

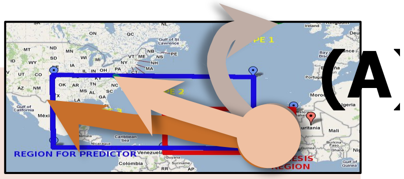
Validate Hypothesis

- 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

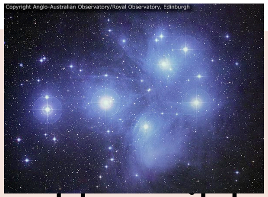
- Will this hurricane strike the Atlantic coast?
- What is the likelihood of this patient to develop cancer
- Will this customer buy a new smart phone?

# 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

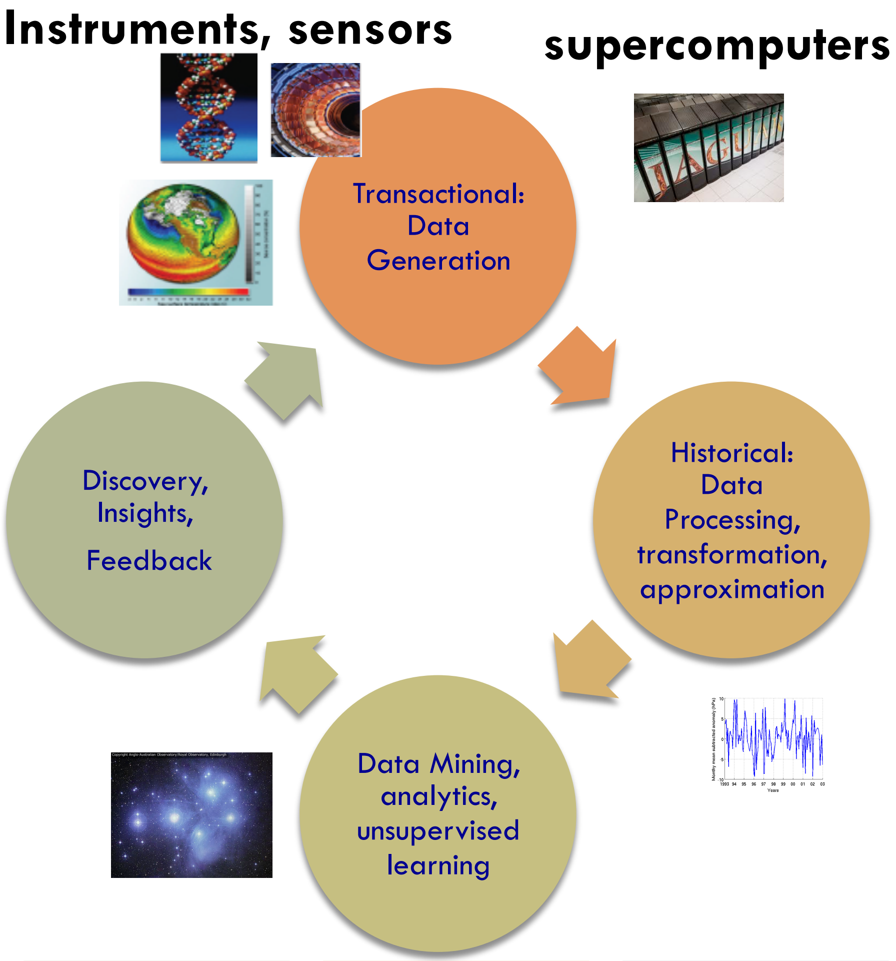
- Wow! I found a new galaxy?
- Switch C fails when switch A fails followed by switch B failing
- On Thursday people buy beer and diaper together.
- The ratio K/P > X is an indicator of onset of diabetes.

# 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

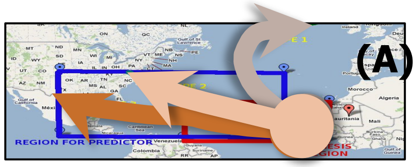
Data Management

Data Reduction, Query

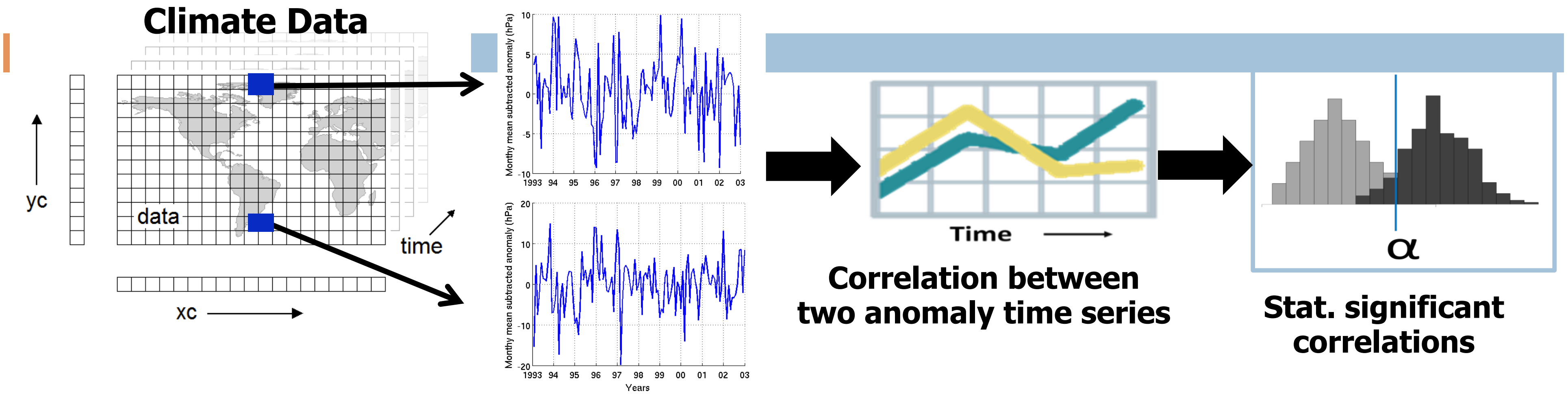
Data Visualization

Data Sharing

Historical data → Learning Models → Trigger/ questions → Predict



# From multi-dimensional data analytics to relationship mining



Correlation between two anomaly time series

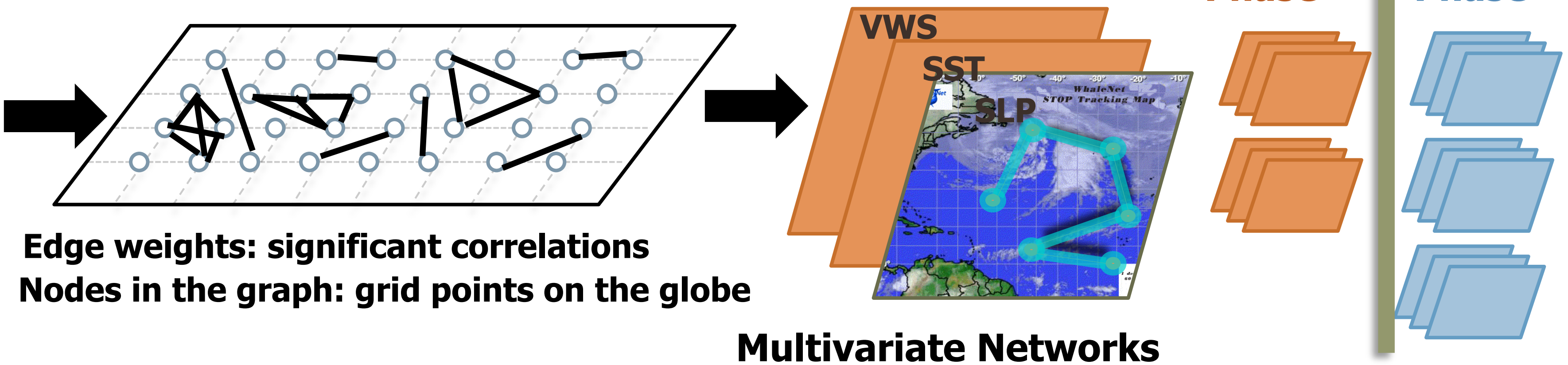
Stat. significant correlations

Anomaly time series at each node

Extreme Phase

Normal Phase

## Climate Network



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

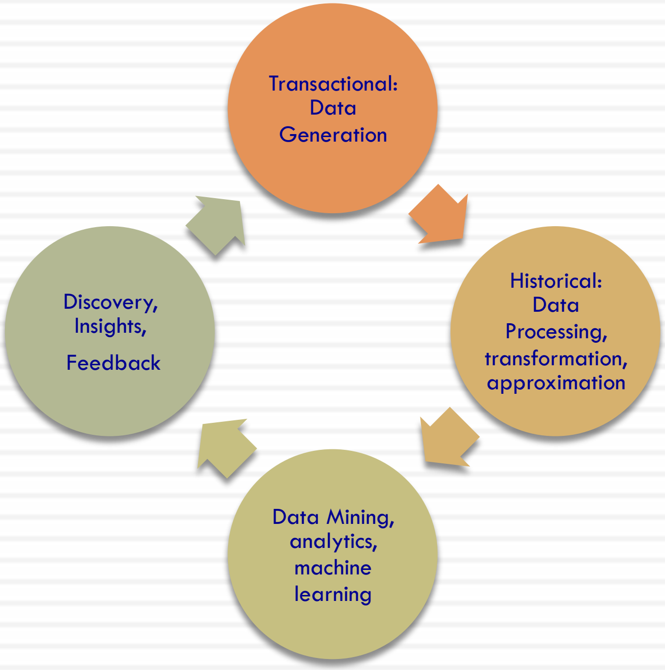
Multivariate Networks

Multiphase Networks

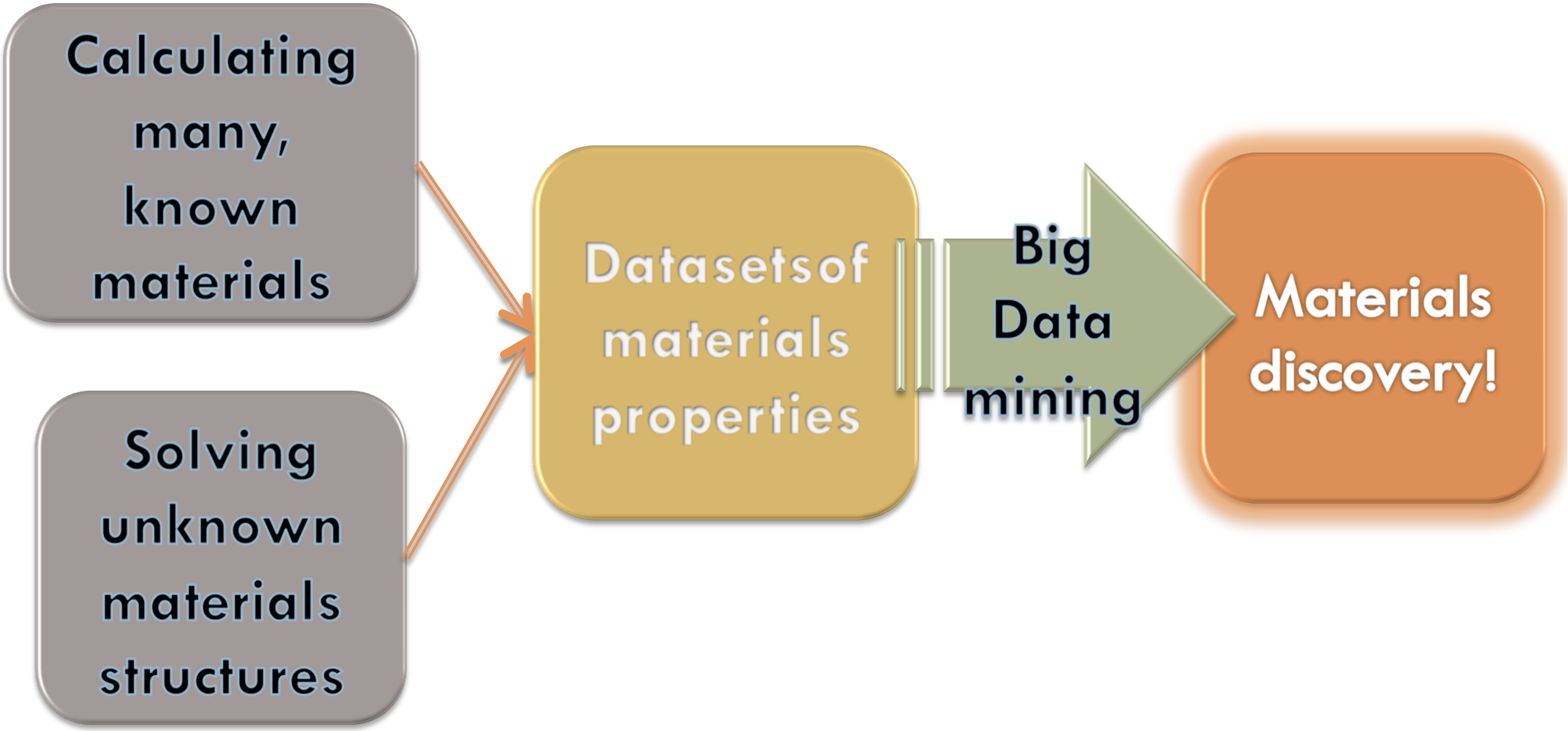
CMIP3 → CMIP5 => Climate BIG DATA : 10s of TBs to 10s of PBs

# 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

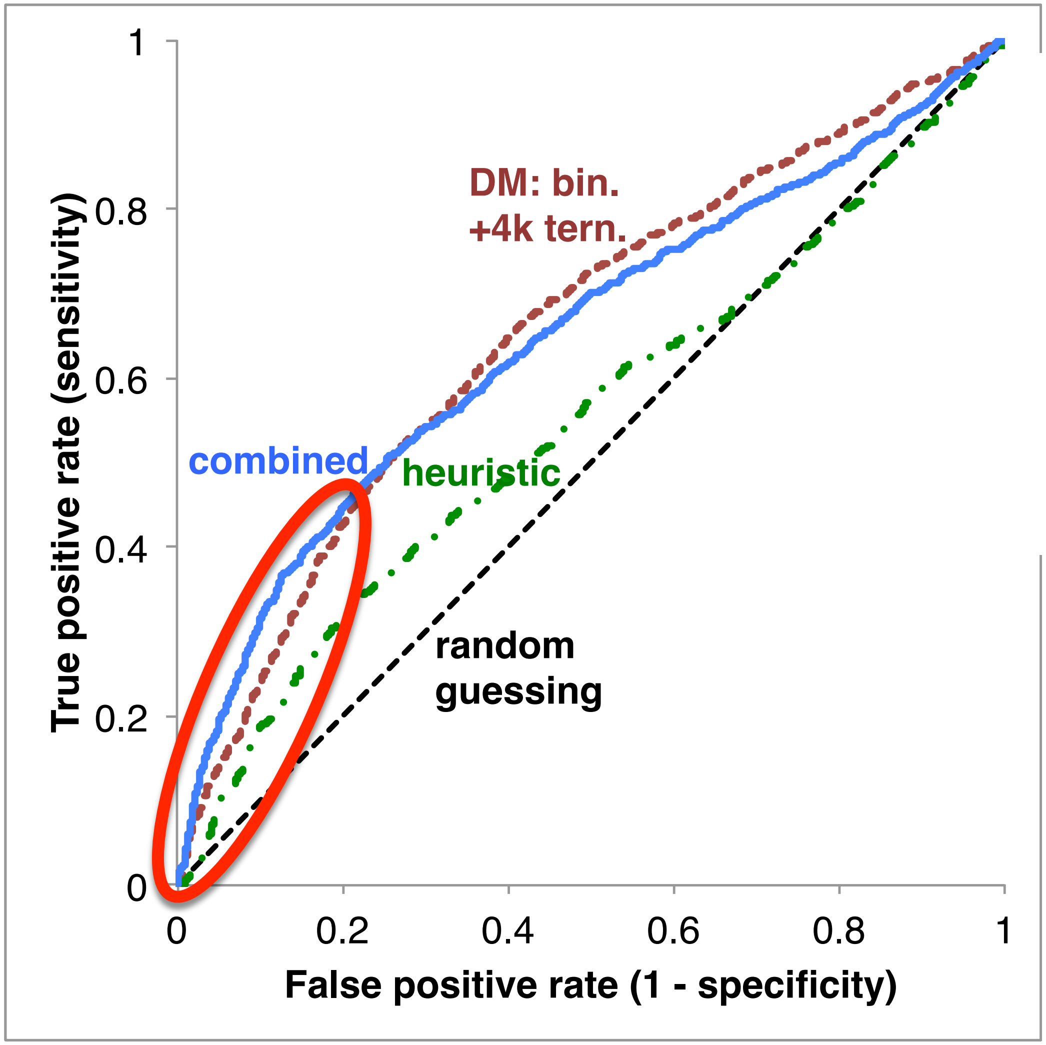
Solving unknown materials structures

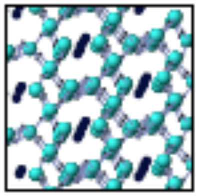
Datasetsof materials properties

Big Data mining

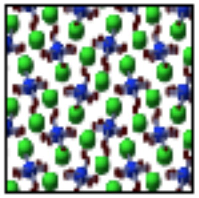
Materials discovery!

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

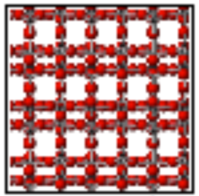




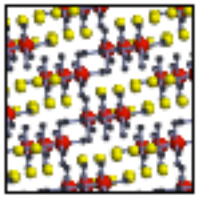
$SiYb_3F_5$



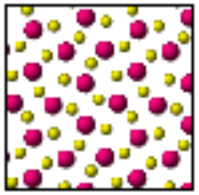
$Pa_2O(SiO_4)$



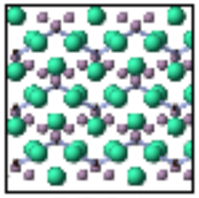
$U_2O(PO_4)_2$†



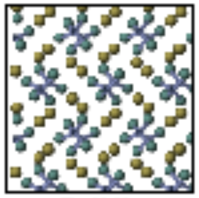
$S_2(VF_6)$



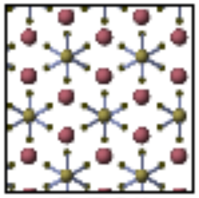
$Pm_2S_3$



$P_3(BrCs_4)$



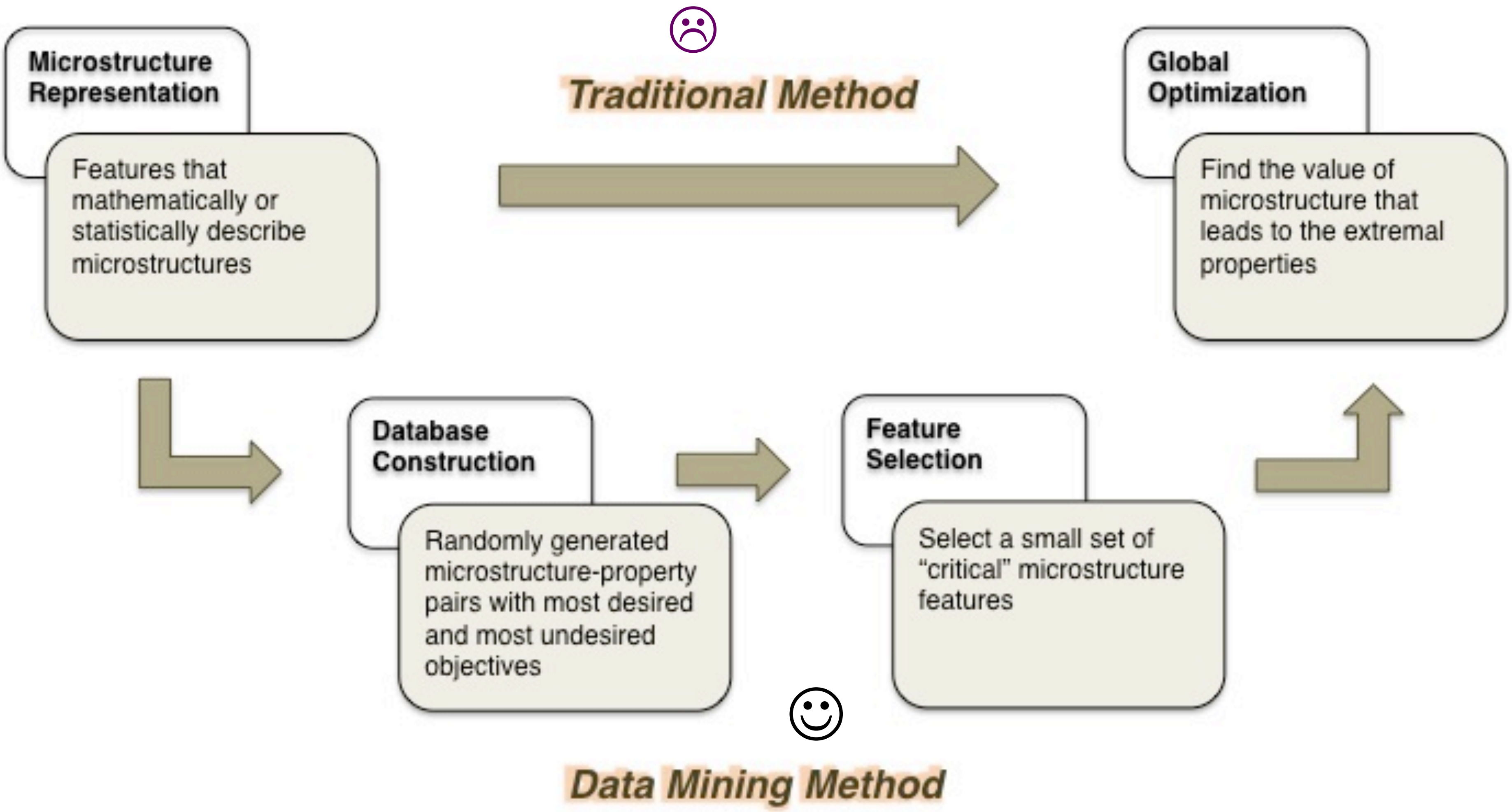
$Te_3Y_4N_2$



$Ba(TeS_3)$†

† 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



**Microstructure Representation**

Features that mathematically or statistically describe microstructures

☹

*Traditional Method*

**Global Optimization**

Find the value of microstructure that leads to the extremal properties

**Database Construction**

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

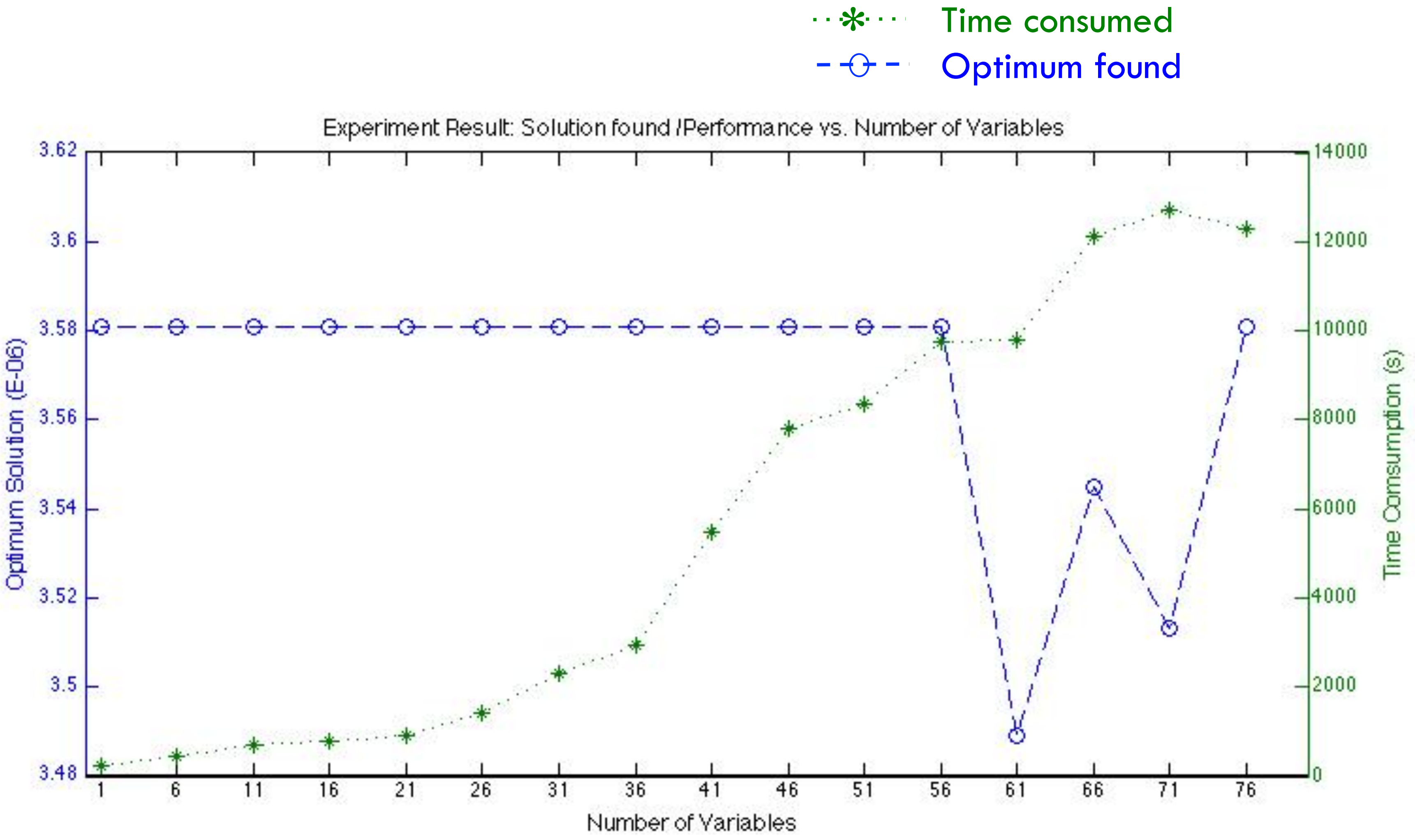
**Feature Selection**

Select a small set of "critical" microstructure features

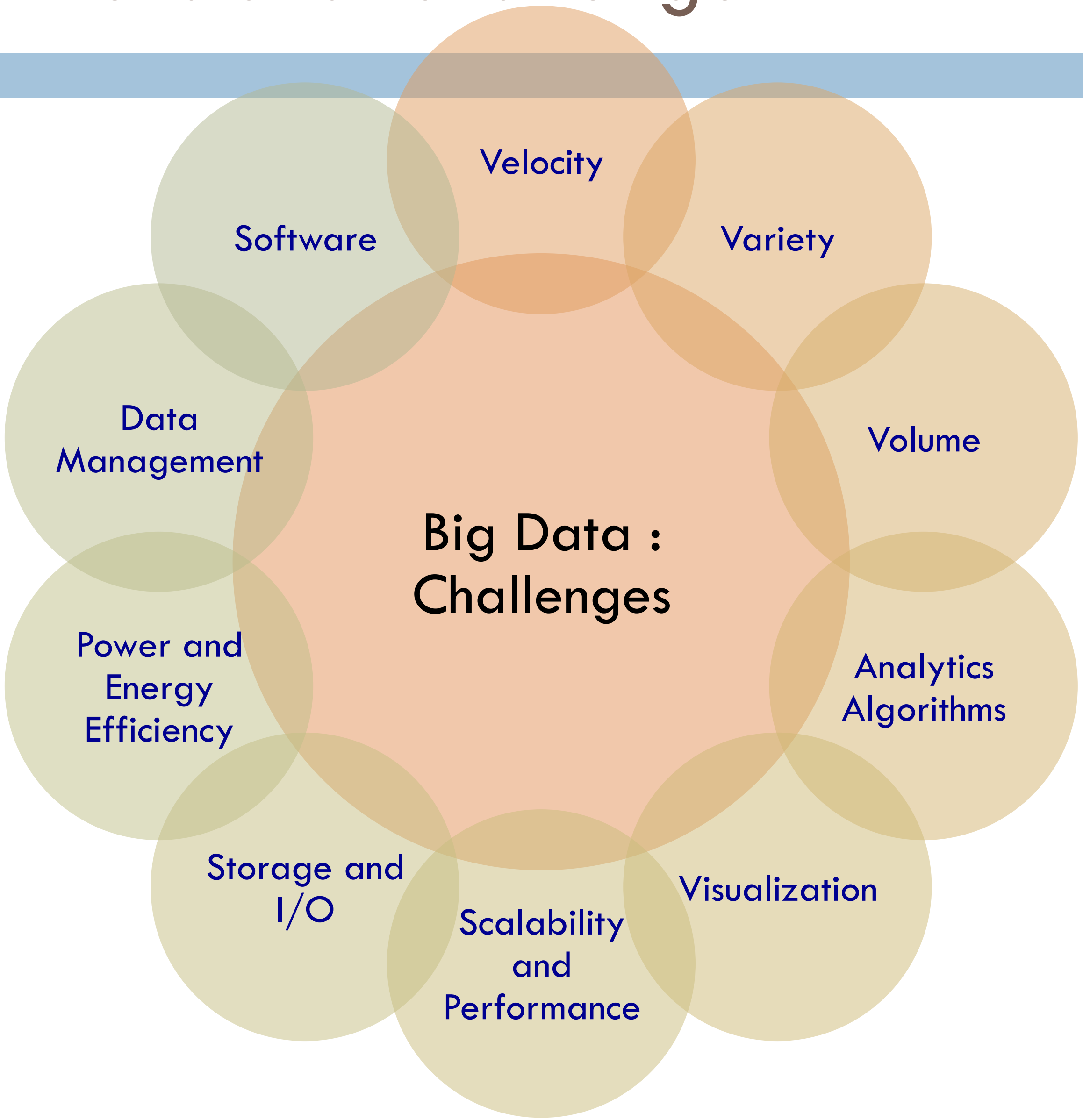
☺

*Data Mining Method*

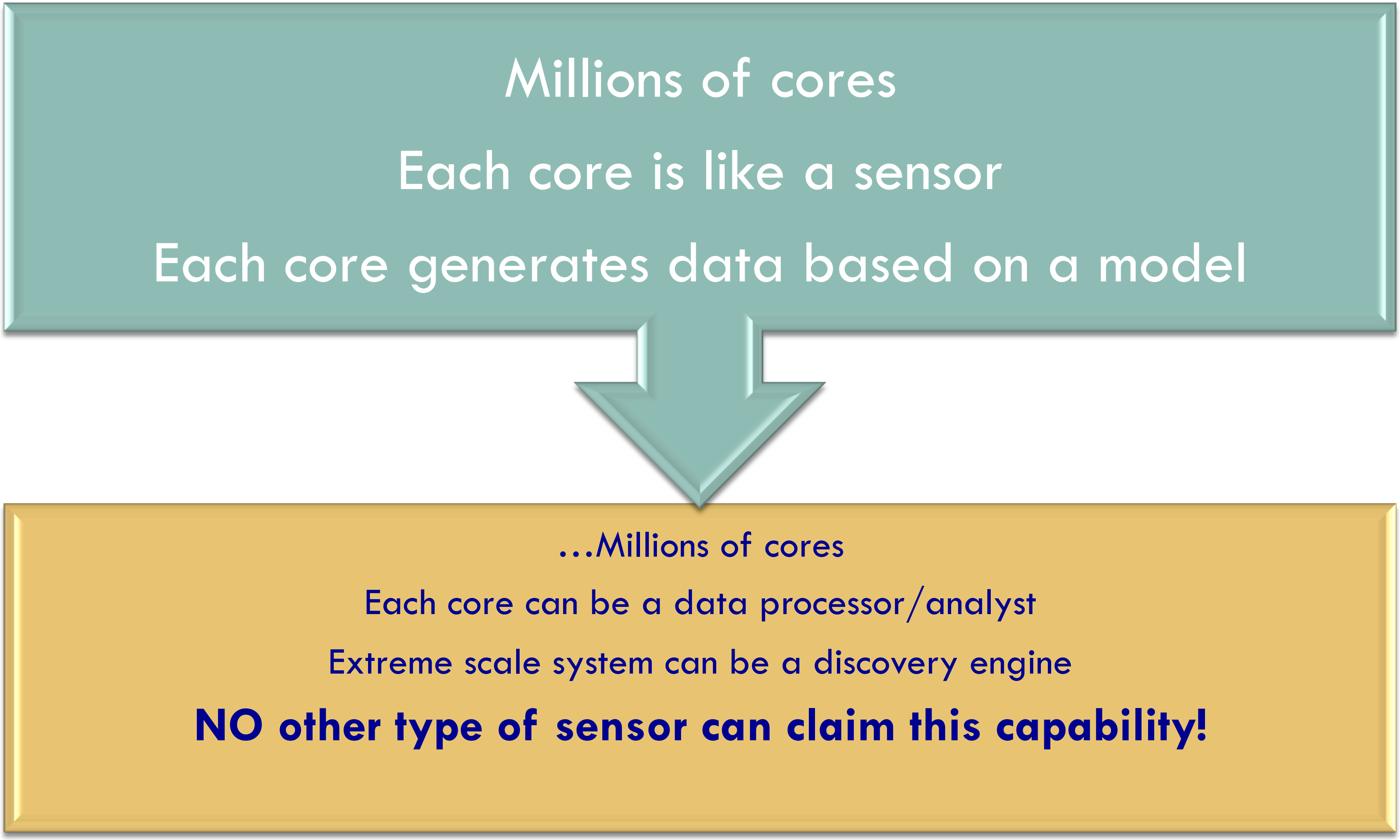
# Accelerating Time to Insights



# Extreme Computing + Big data : Not a single dimensional challenge



# An instrument and a discovery engine



Millions of cores

Each core is like a sensor

Each core generates data based on a model

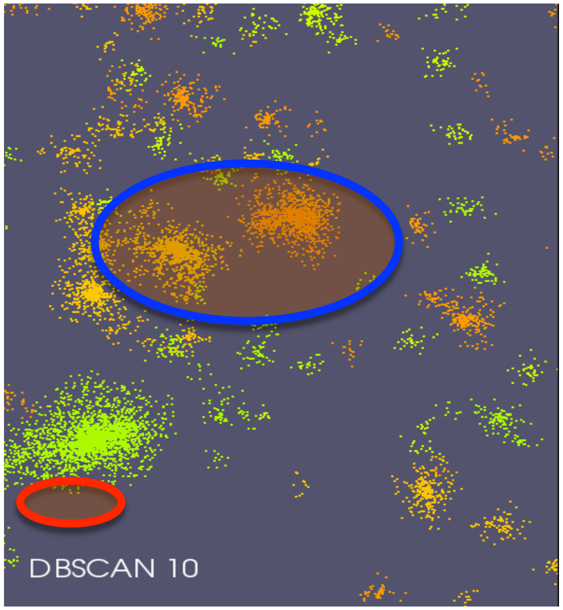
…Millions of cores

Each core can be a data processor/analyst

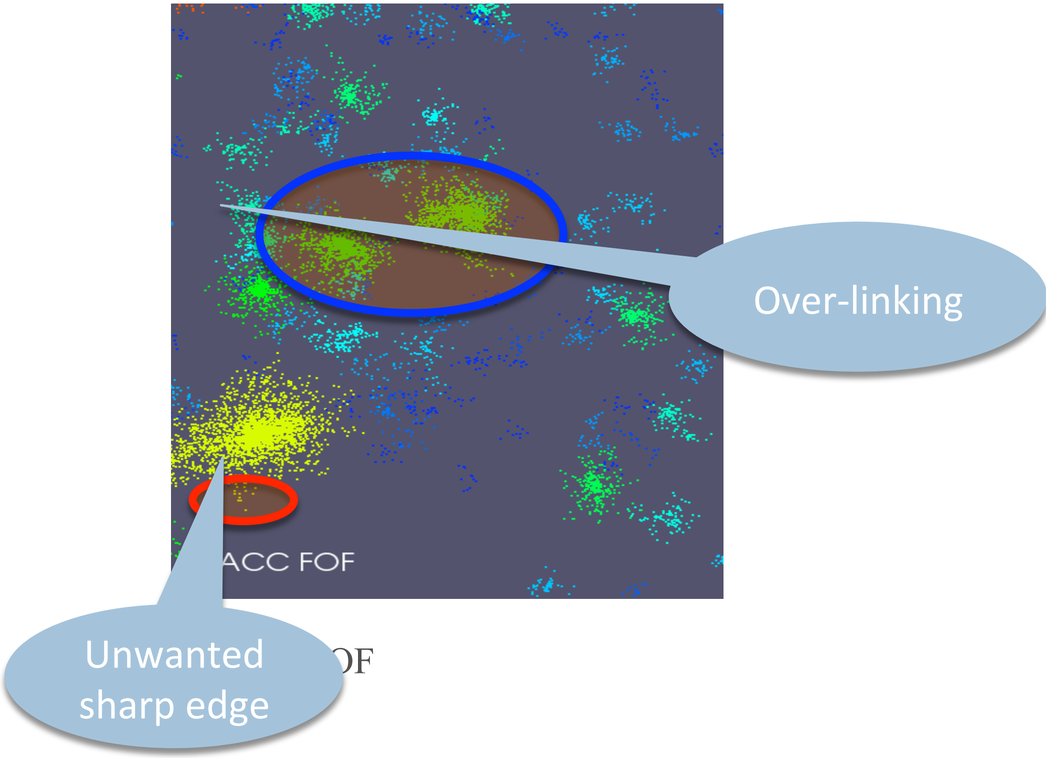
Extreme scale system can be a discovery engine

**NO other type of sensor can claim this capability!**

# BDEC: Can we do this type of analytics in-situ?



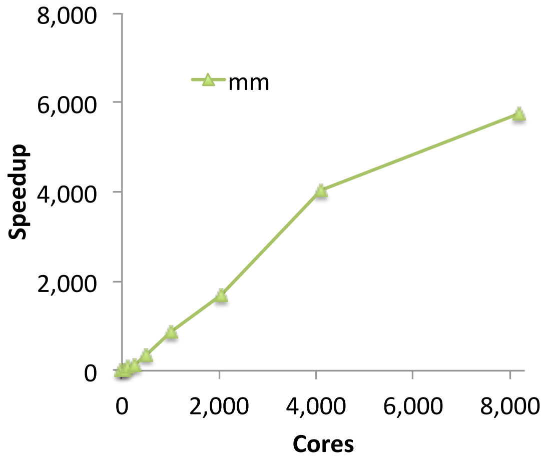
Scalable DBSCAN+



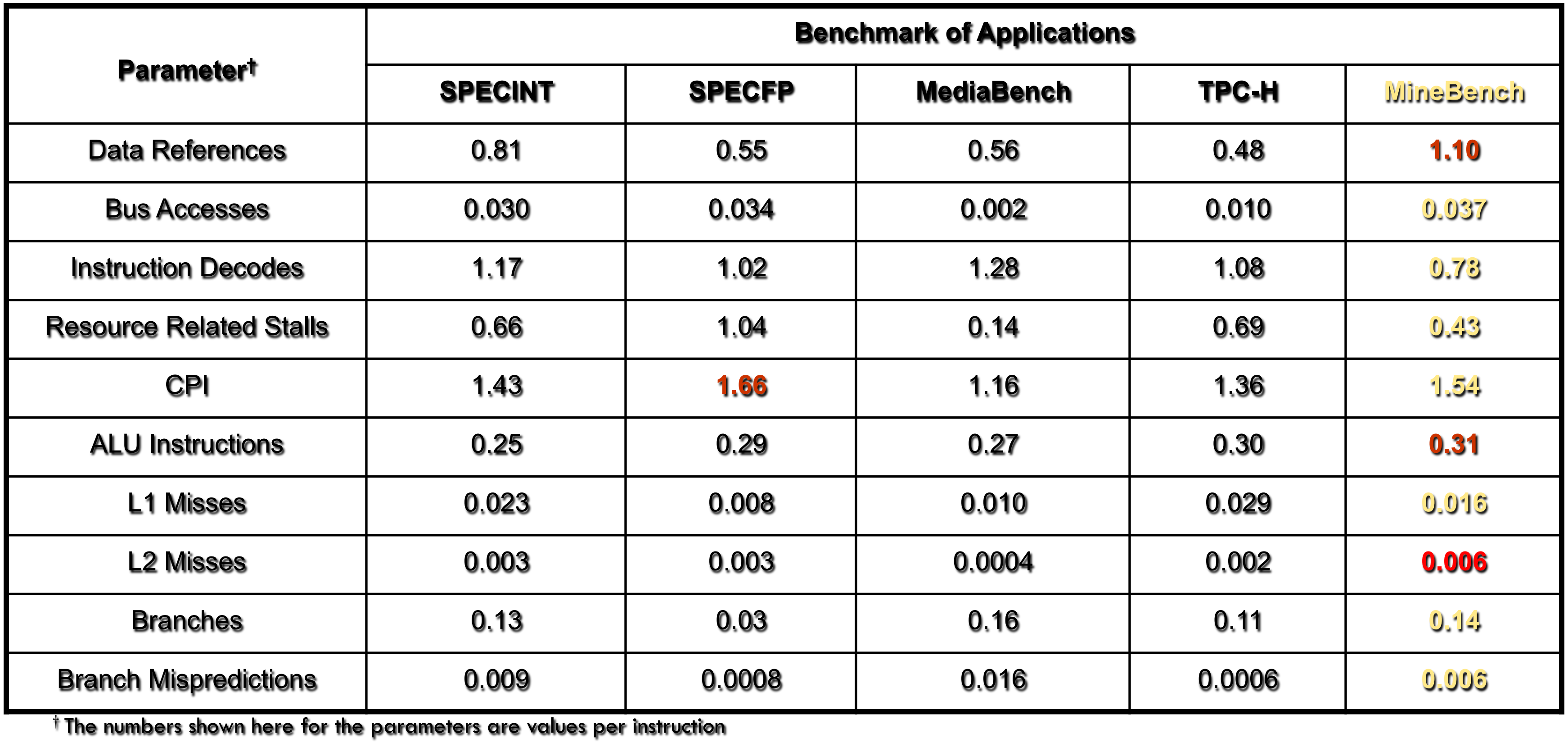
OF

Identifying arbitrary shaped structures using astrophysics data (http://arxiv.org/abs/1203.3695)

- ❑ Climate, Astronomy, Biology, Earth science
- ❑ Advanced data structure to break the inherent sequential data access order of DBSCAN
- ❑ Scalable DBSCAN identifies the clusters without sacrificing the quality of the solution
- ❑ Strong scaling on astrophysics datasets



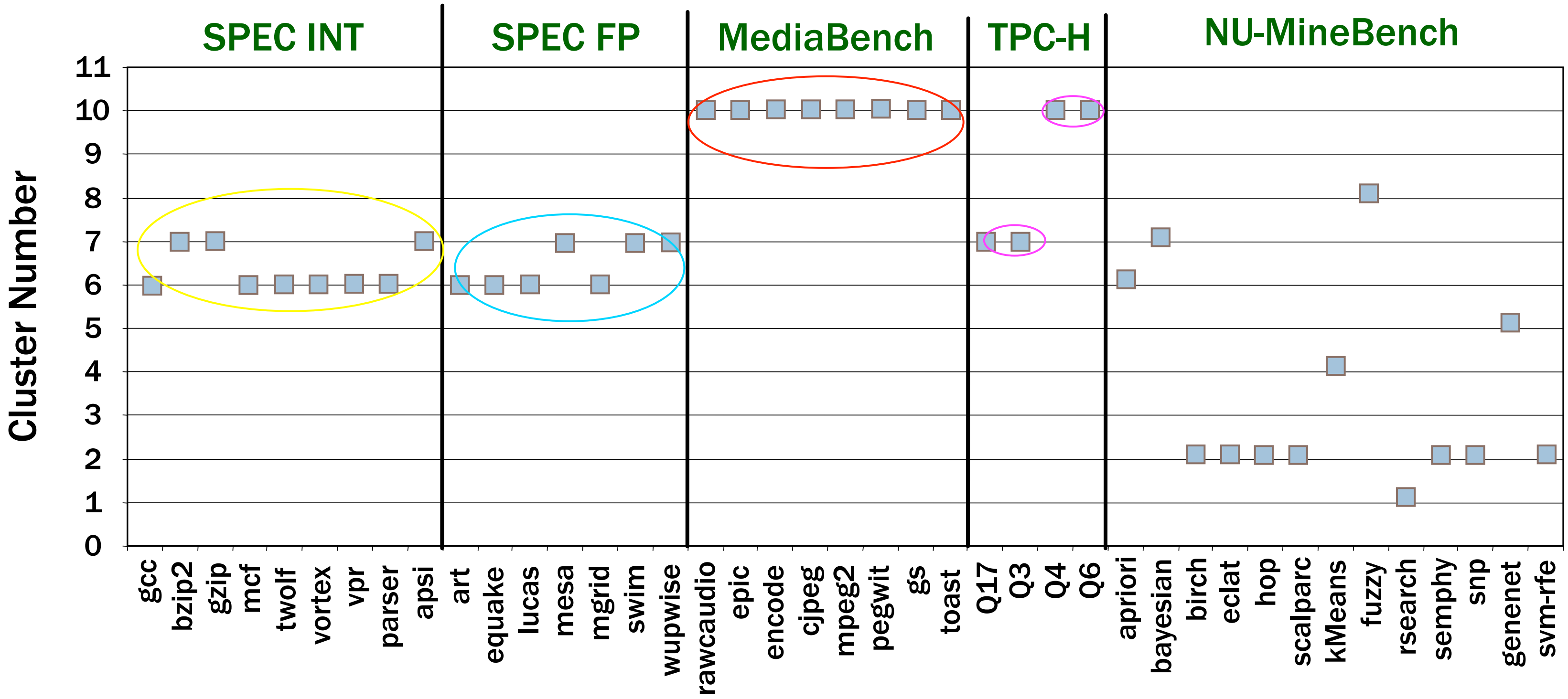
# Right Computing infrastructure? What characteristics do typical analytics functions have?



| 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?

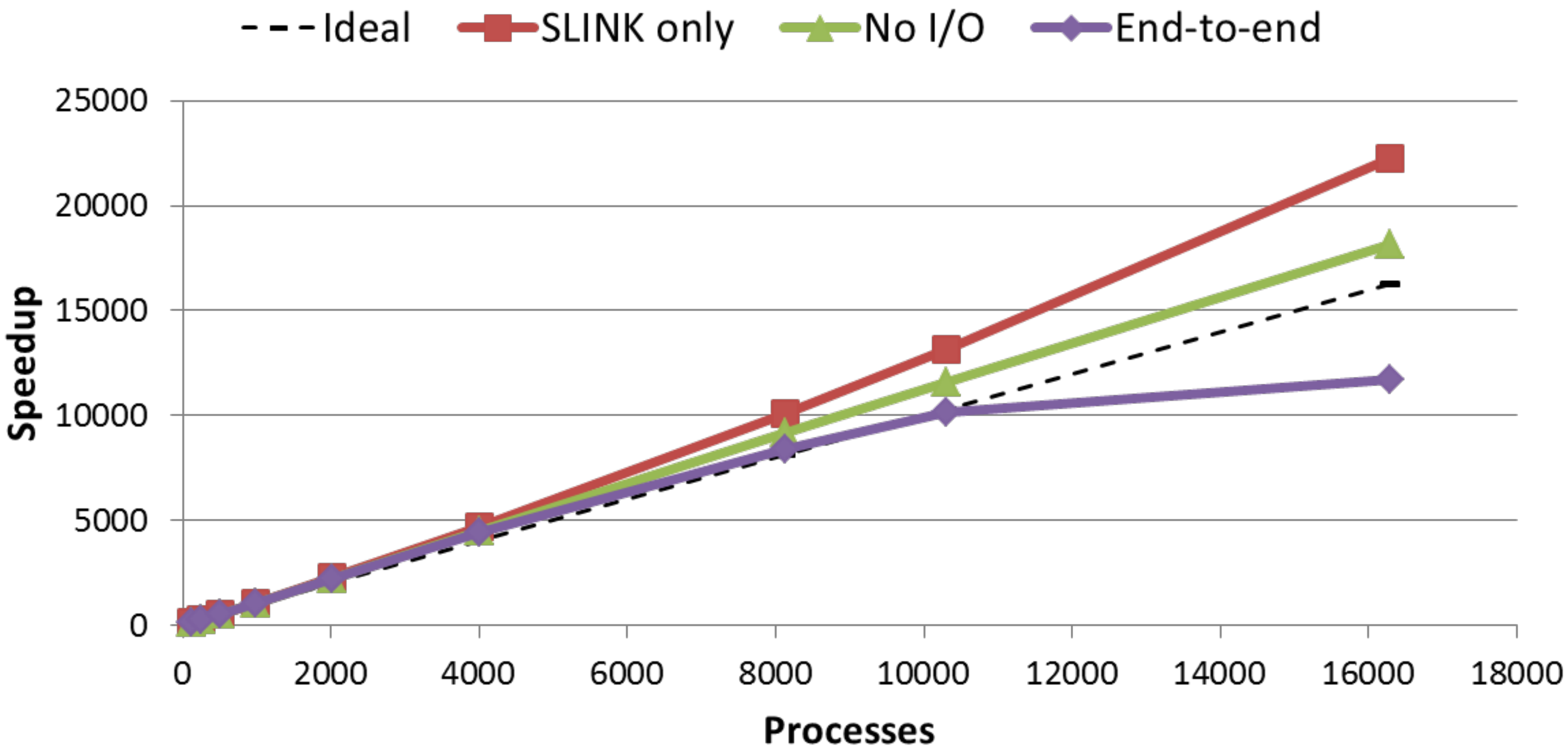


Clear Implications on architecture, modes, memory hierarchy and other components

Identify similarities and design for co-existence

# Develop scalable versions – Pay attention to I/O : Particularly reads

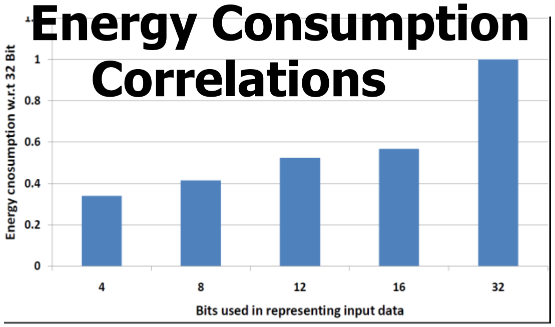
- **Parallel hierarchical clustering**
  - Speedup of 18,000 on 16k processors
  - I/O significant at large scale



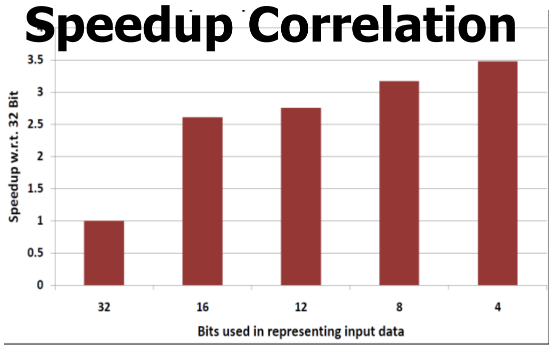
# Good News: Approximation is a TOP Option in analytics => Power aware data analytics

## Power-aware analytics

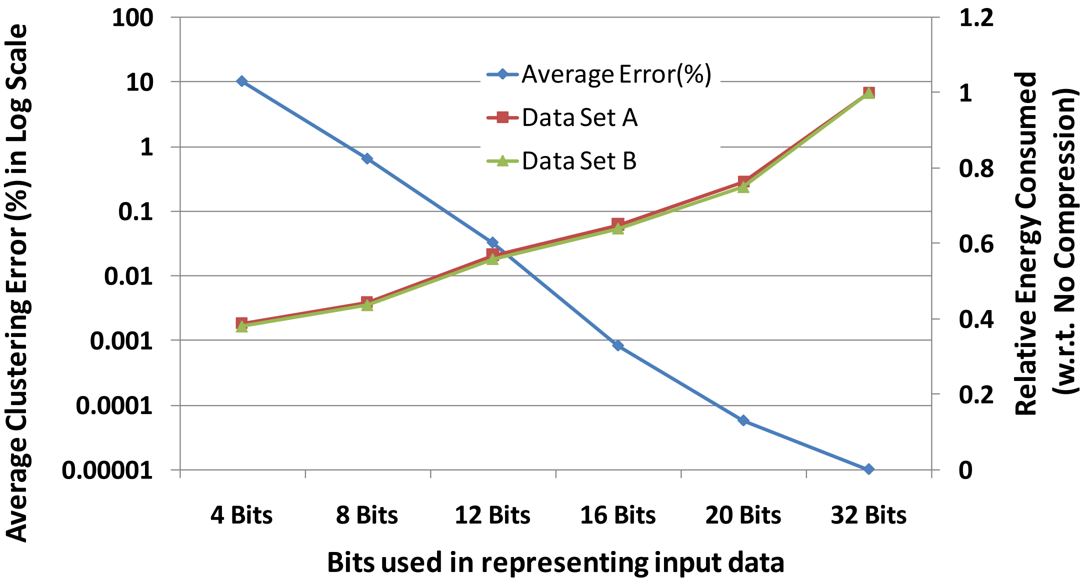
- Reduced bit fixed-point representations
- Pearson correlation
  - 2.5-3.5 times faster
  - 50-70% less energy
- K-means
  - ~44% less energy with an error of only 0.03% using 12-bit representation

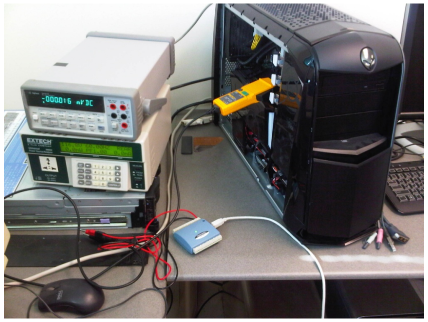


Energy Consumption Correlations

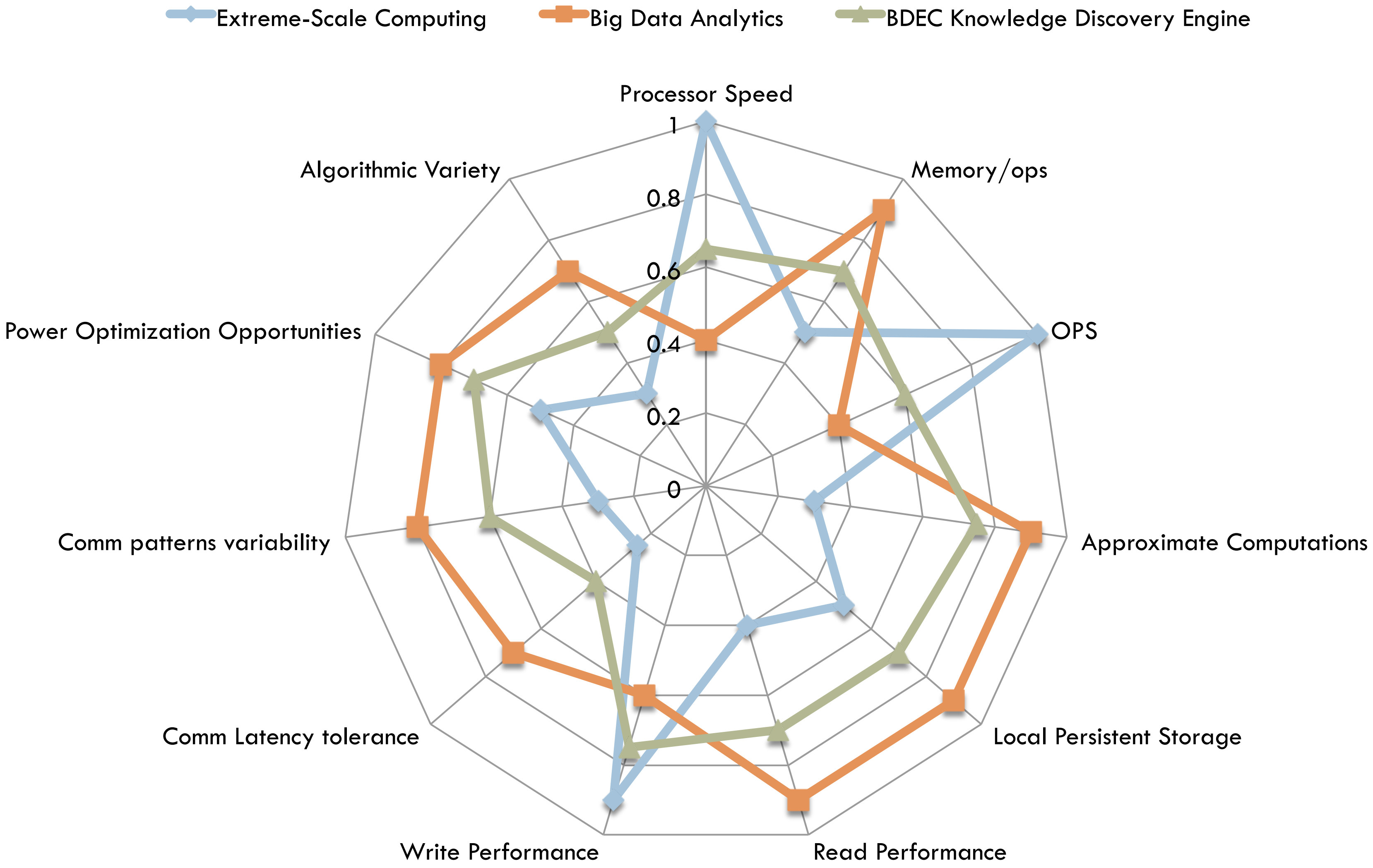


Speedup Correlation





# Extreme Computing + Big Data Analytics = BDEC 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