Graph Data Analytics at Scale:
Data Science Perspective

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

OAK RIDGE National Laboratory

SIAM CSE, March 14, 2015

NC STATE UNIVERSITY

Talk Slide Contributors

- David Bader, GeorgiaTech
- Christos Faloutsos and his research team, CMU
- Erik Demaine, MIT
- Vipin Kumar and his research team, UMN
- Michael Langston, UTK
- Blair Sullivan, NC State
- PhD Students who conducted research

Logistics: Guidelines, Assumptions, Homework

From SIAM CSE Organizers

- **Broadly accessible**
- **CSE applications**
- **Technical insights**
- **Technical challenges**
- **Solving the challenges**
- **Future perspective**
- **Student training**

Assumptions: ***Audience has little knowledge about graph analytics but infinite intelligence.***

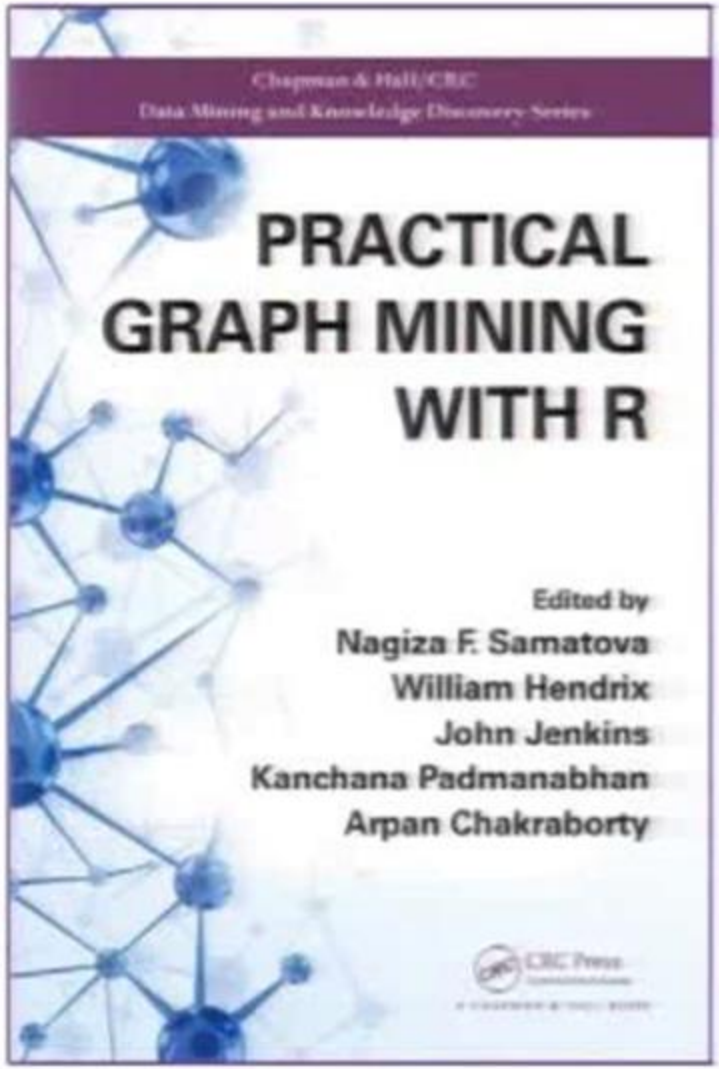
Homework: Email me (samatova@csc.ncsu.edu) questions/suggestions.

Textbook Written Entirely by NC State Computer Science Students

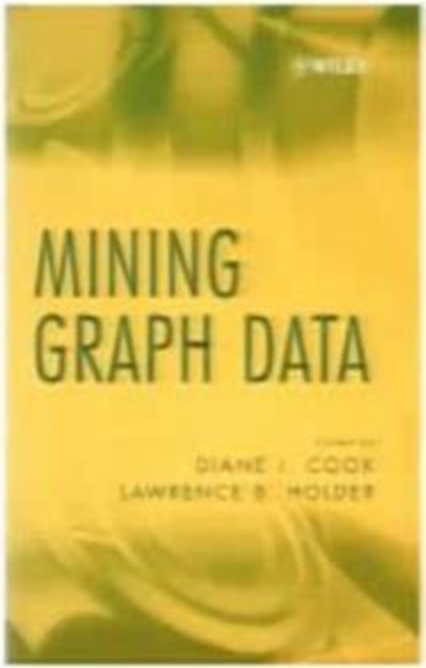
- ~40 graduate students
- ~20 undergraduate students
- Co-editors: PhD students

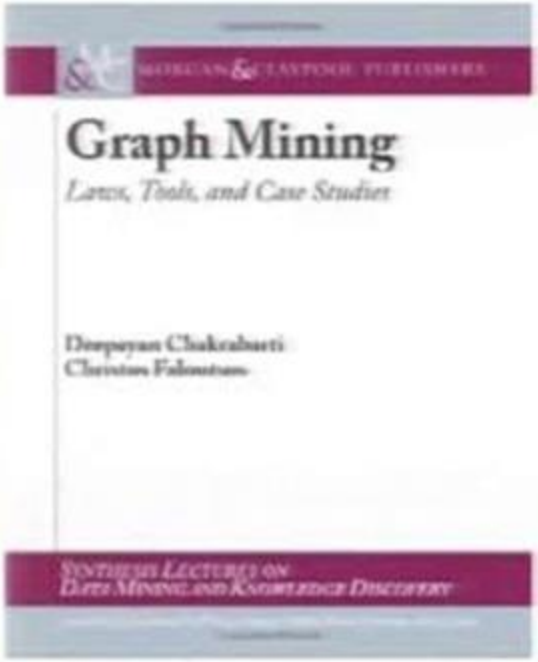
Proceeds to benefit the NC State Department of Computer Science: to sponsor students' involvement in research

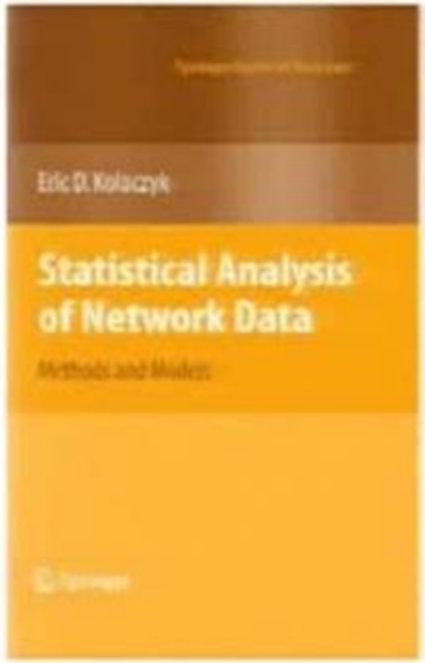
http://www.crcpress.com/product/isbn/9781439860847

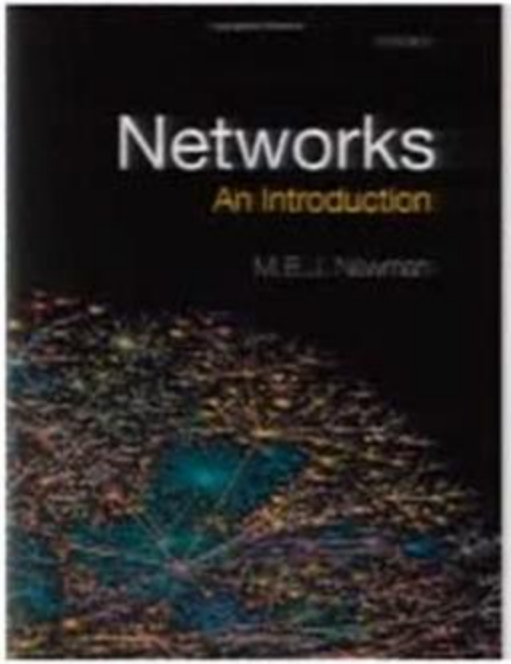


Other books...





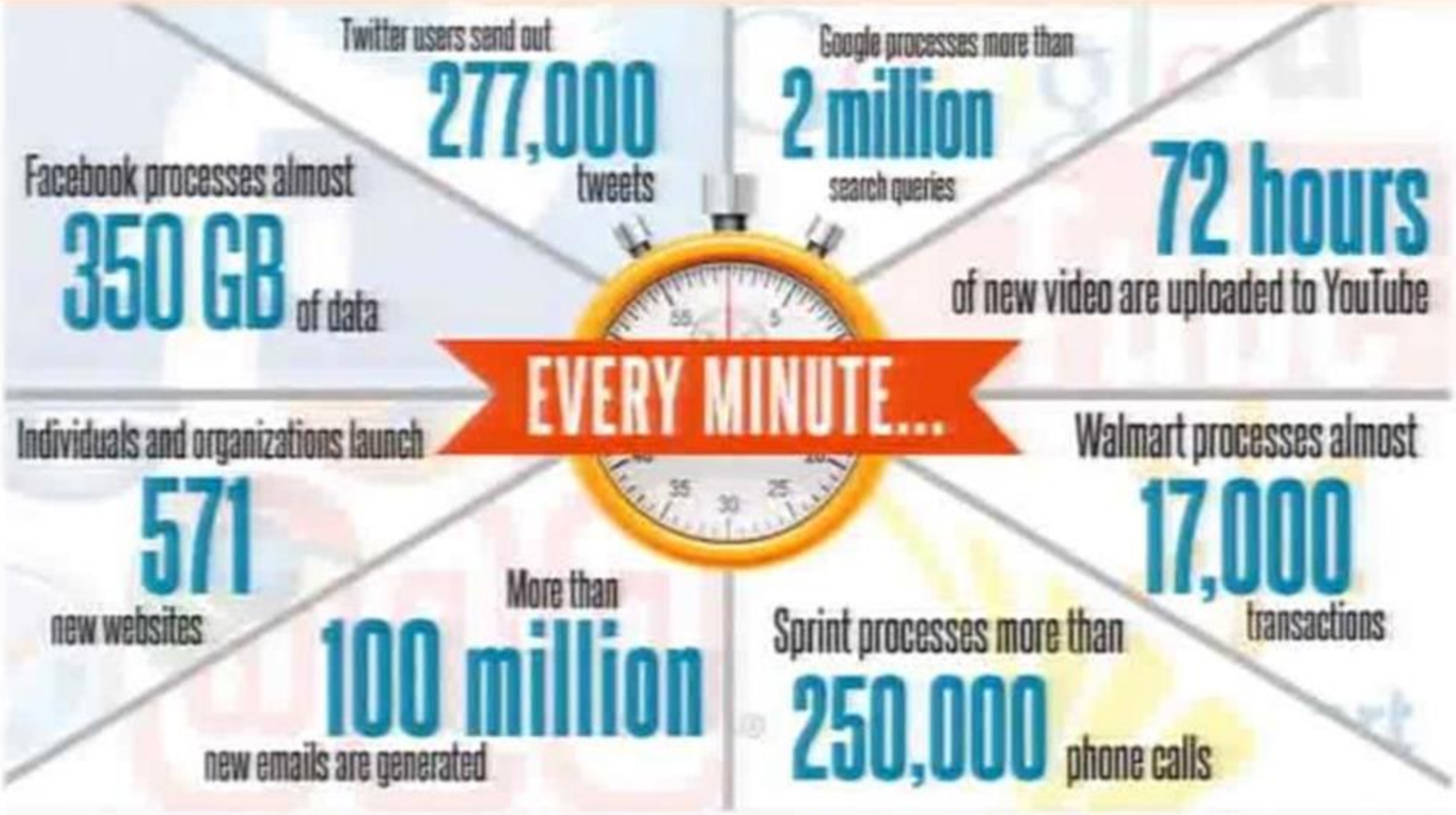




Graphs ≡ Networks

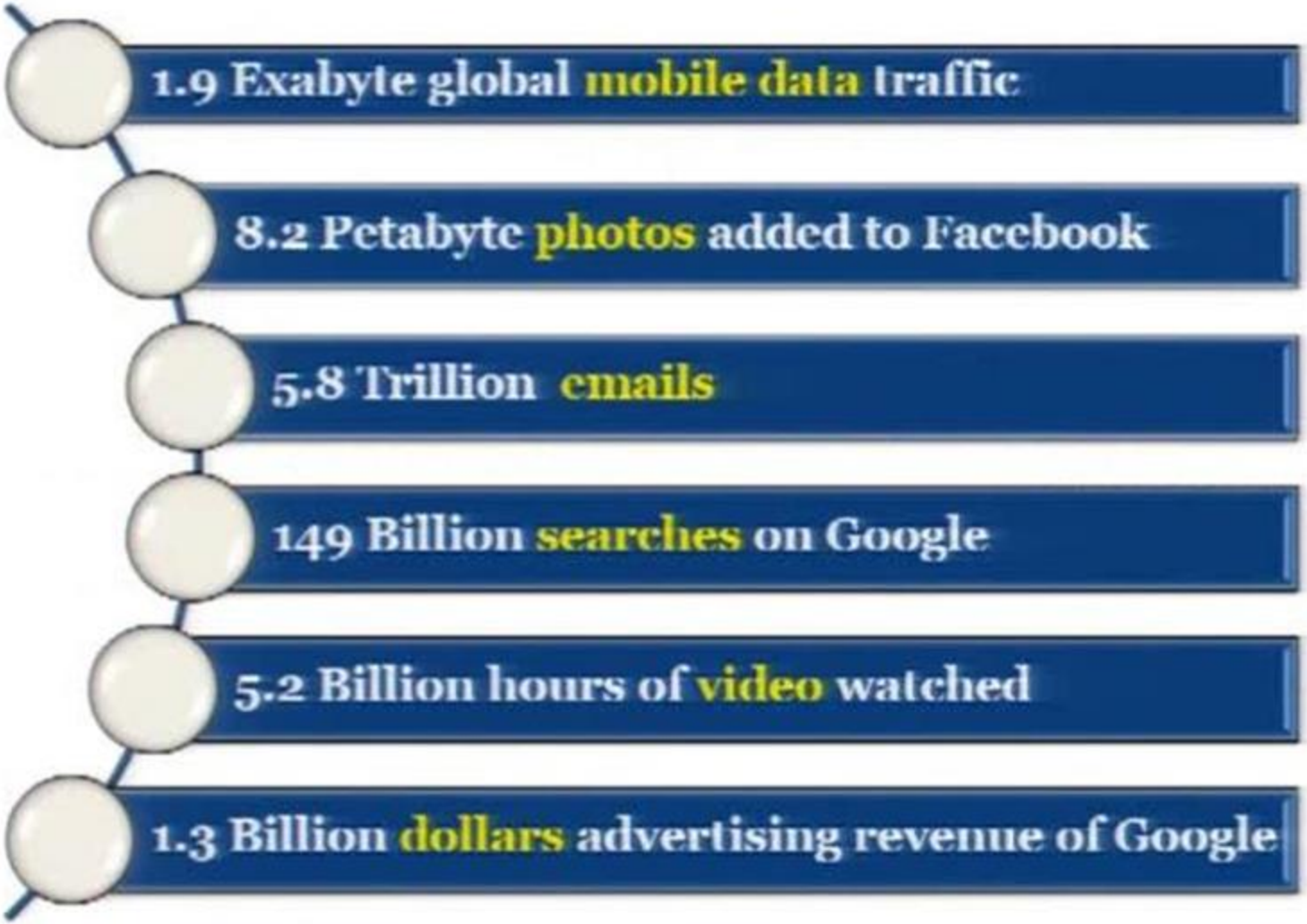
2,500,000,000,000,000,000

In 2012, we created 2.5 quintillion bytes of data **every day.**



http://removeandreplace.com/2013/03/13/how-much-data-is-on-the-internet-and-generated-online-every-minute/

2013 Facts Hunt per Month



http://www.factshunt.com/2014/01/world-wide-internet-usage-facts-and.html

How Big is the Internet?

- **14.3 Trillion - Webpages, live** on the Internet.
- **48 Billion** - Webpages indexed by Google.Inc.
- **14 Billion** - Webpages indexed by Microsoft's Bing.
- **672 Exabytes - 672,000,000,000 Gigabytes (GB)** of accessible data.
- **43,639 Petabytes** - Total **World-wide Internet Traffic** in 2013.

- **Over 1 Yotta-byte - Total data stored on the Internet**

1 Yotta-byte = 10^{24} =

1,000,000,000,000,000,000,000,000 Bytes!

http://www.factshunt.com/2014/01/world-wide-internet-usage-facts-and.html

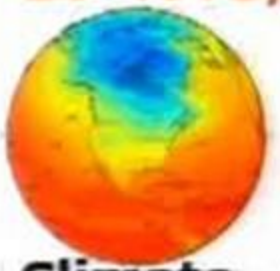
How Big is Scientific Data?

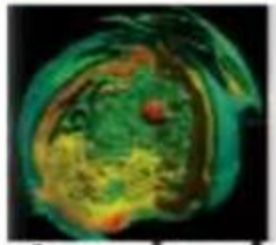

Ecology


Biology

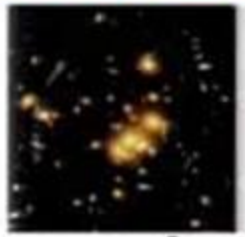
1PB/year

20-40TB/simulation


Climate


Astrophysics

30TB/day


Cosmology

My laptop:
512 GB (GigaBytes) – $512*10^9$ Bytes

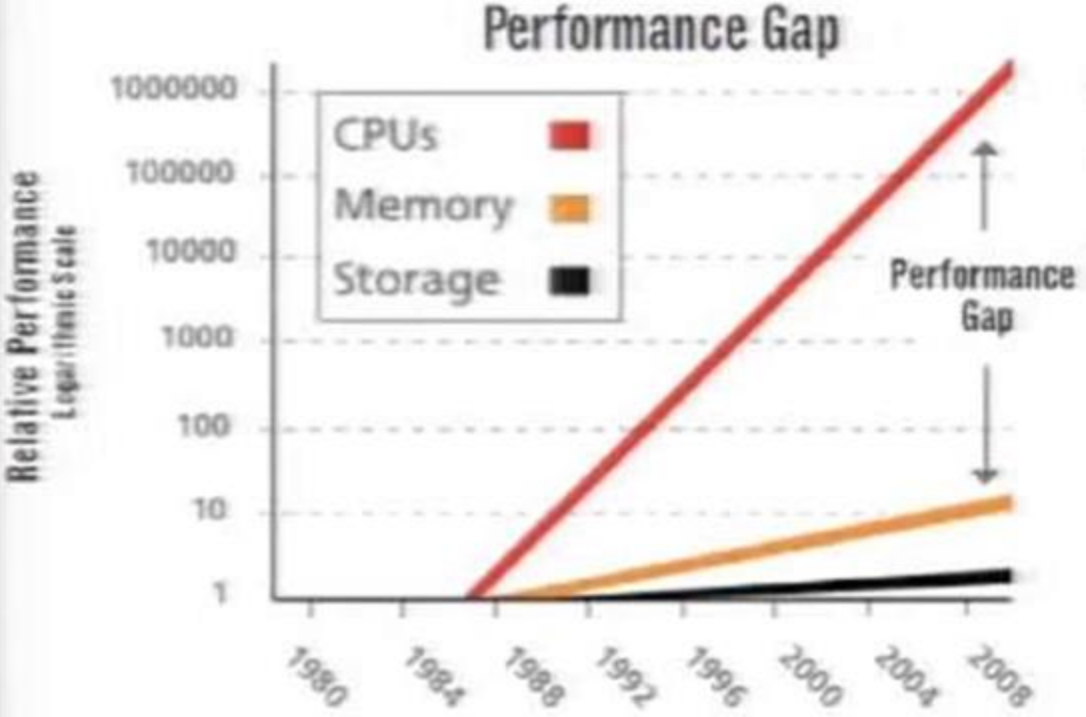
1 TB (TeraByte) – 10^{12} Bytes
1 PB (PetaByte) – 10^{15} Bytes

How to Move and Access the Data?

Technology trends are a rate limiting factor.

With the current trends in technology, most of these data will NEVER be touched!

Data *doubles* every 9 months;
CPU – 2 years; Memory – 6 years.



Naturally distributed
but effectively immovable

Streaming/Dynamic
but not re-computable

http://www.fusionio.com/load/-media-/1952q9/imagesLibrary/power1.png

What Analysis Methods to Use?

Analysis methods fail for a few gigabytes.

Method Complexity:

Calculate means	$O(n)$
Calculate FFT	$O(n \log(n))$
Calculate SVD	$O(r \cdot c)$
Clustering algorithms	$O(n^2)$

Data size n	Algorithm Complexity		
	n	$n\log(n)$	n^2
100B	10^{-10} sec.	10^{-10} sec.	10^{-8} sec.
10KB	10^{-8} sec.	10^{-8} sec.	10^{-4} sec.
1MB	10^{-6} sec.	10^{-5} sec.	1 sec.
100MB	10^{-4} sec.	10^{-3} sec.	3 hrs
10GB	10^{-2} sec.	0.1 sec.	3 yrs.

If n=10GB, then what is $O(n)$ or $O(n^2)$ on a teraflop computers?

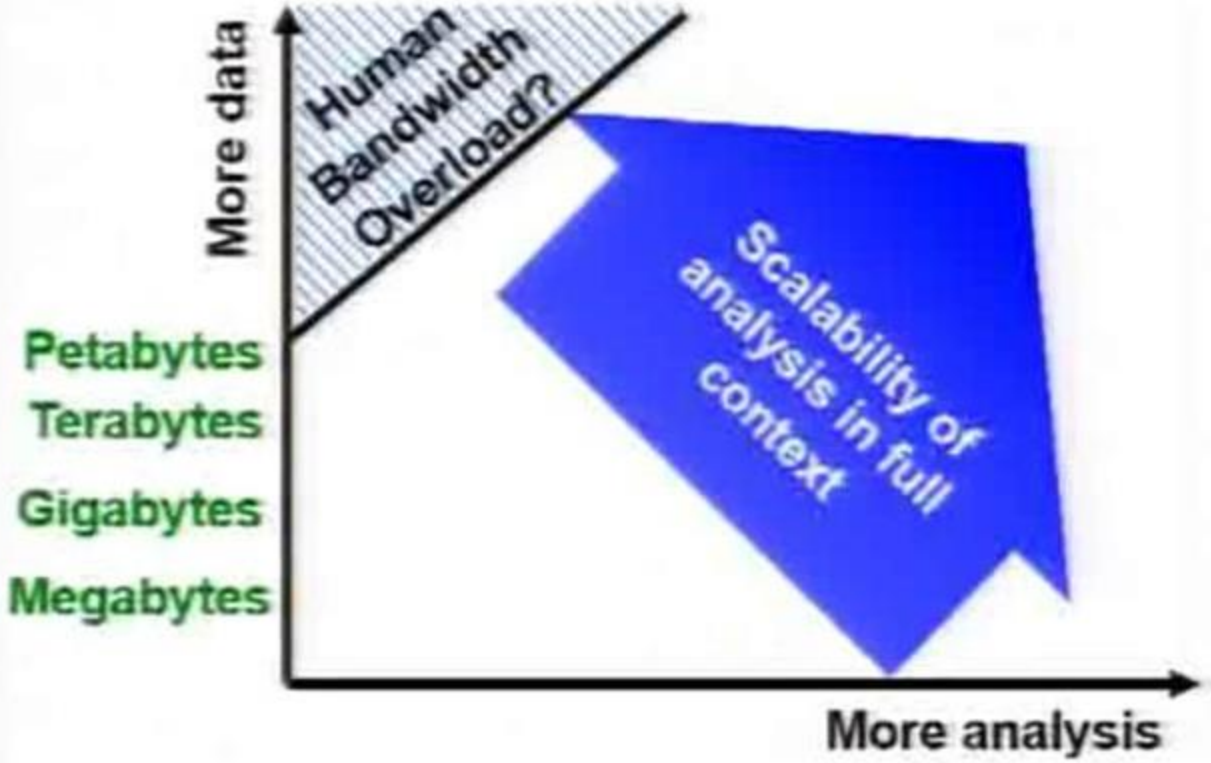
1GB = 10^9 bytes
1Tflop = 10^{12} op/sec

For illustration chart assumes 10^{-12} sec. (1Tflop/sec) calculation time per data point

How to Make Sense of Data?

Know Your Limits & Be Smart!

Not humanly possible to browse a petabyte of data.
Analysis must reduce data to quantities of interest.



Computations:
Must be smart about which probe combinations to see!

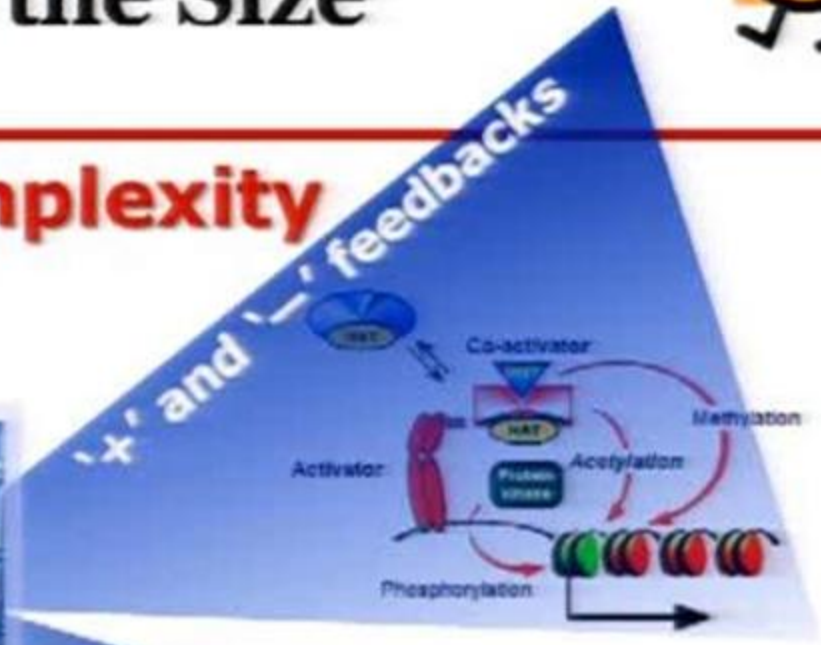
Physical Experiments:
Must be smart about probe placement!

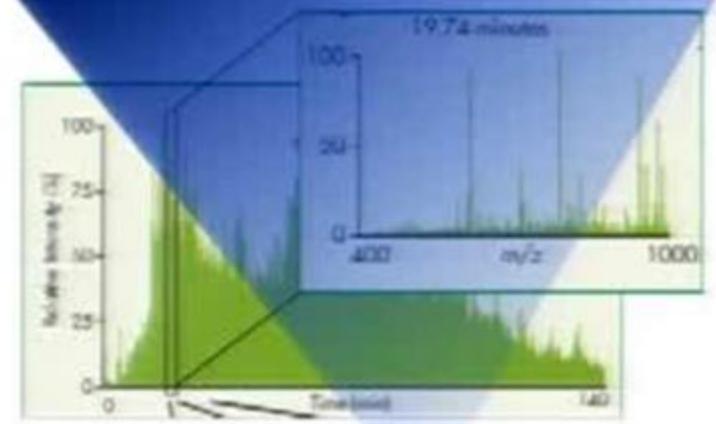
To see 1 percent of a petabyte at 10 megabytes per second takes:

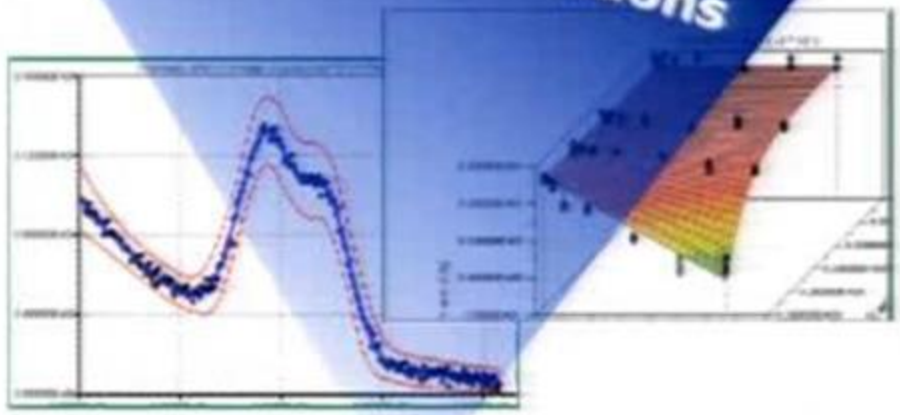
35 8-hour days!

It is not just the Size

– but the Complexity

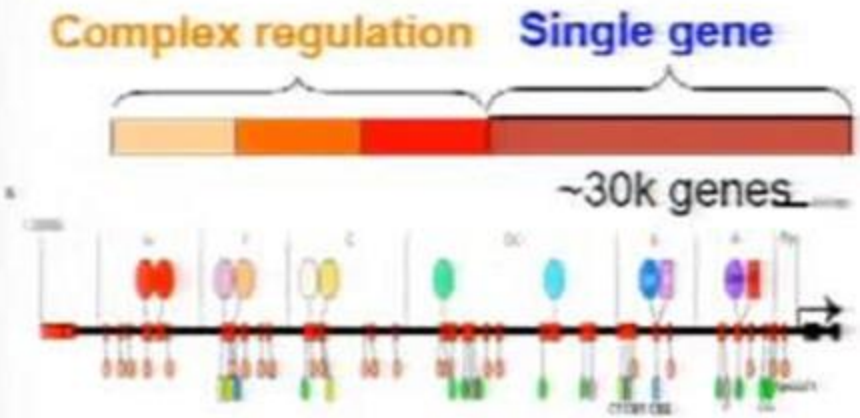






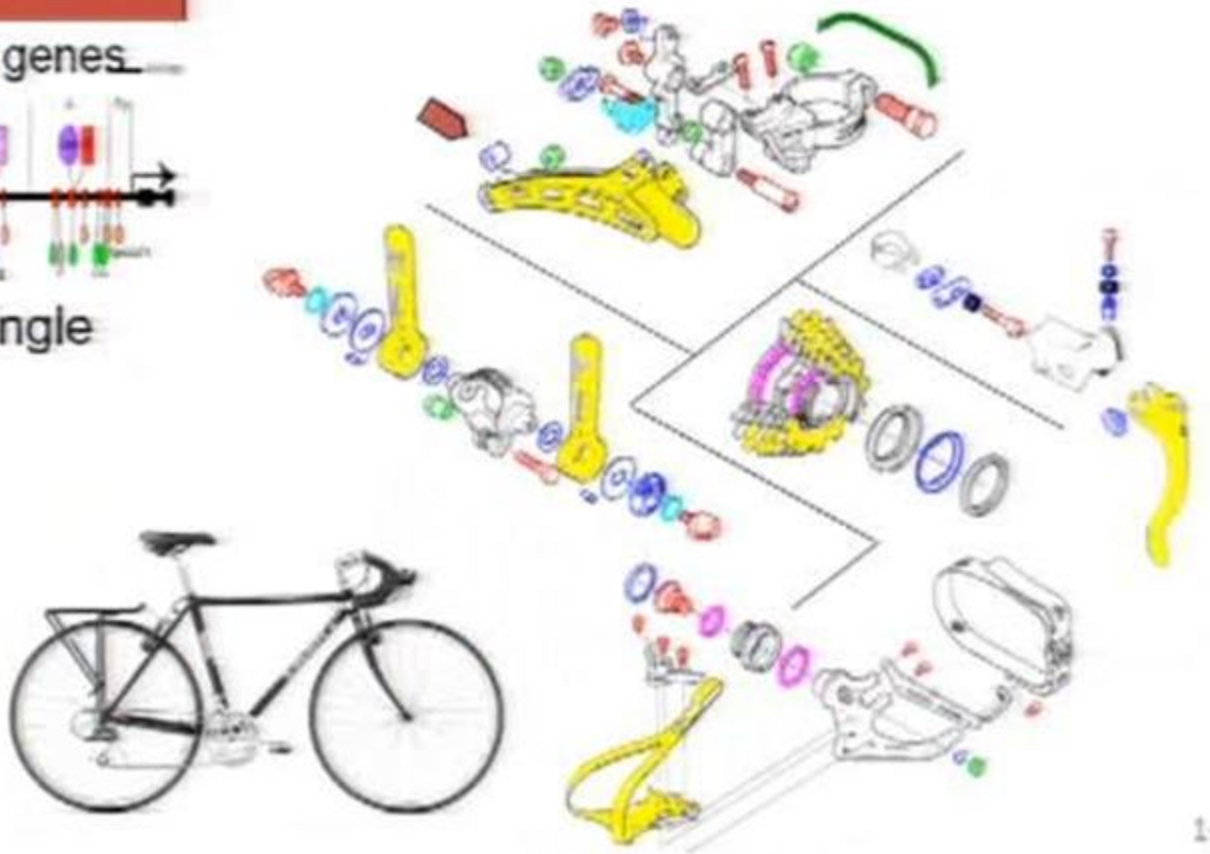
Data Describes Complex Patterns/Phenomena

How to untangle the riddles of the complexity?



50 trans elements control single gene expression

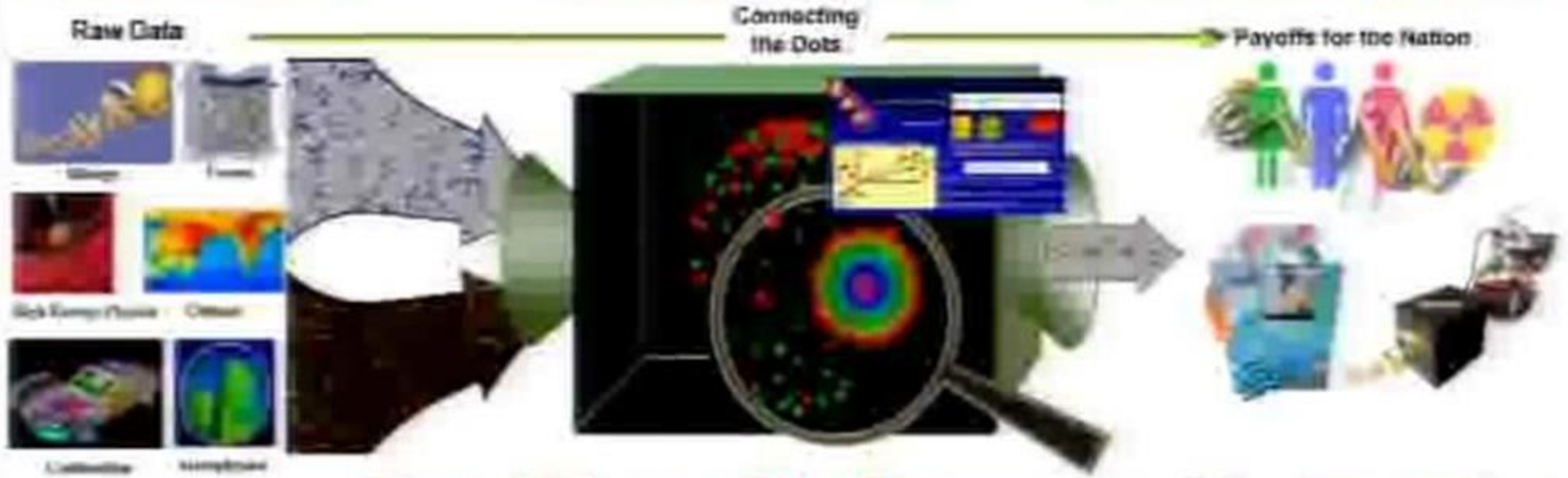
Analytical tools that **find the "dots"** from data significantly reduce data.



The Data Science Challenge

Challenge: How to **"connect the dots"** to answer important questions?

Finding the Dots | **Connecting the Dots** | **Understanding the Dots**



Massive Data

Climate:
5-10 Petabytes/year
Fusion:
1000 Megabytes/2 min

Data Science Challenges

- Huge dimensional space
- Combinatorial challenge
- Complicated by noisy data
- Requires high-performance networks, disks, computers,...

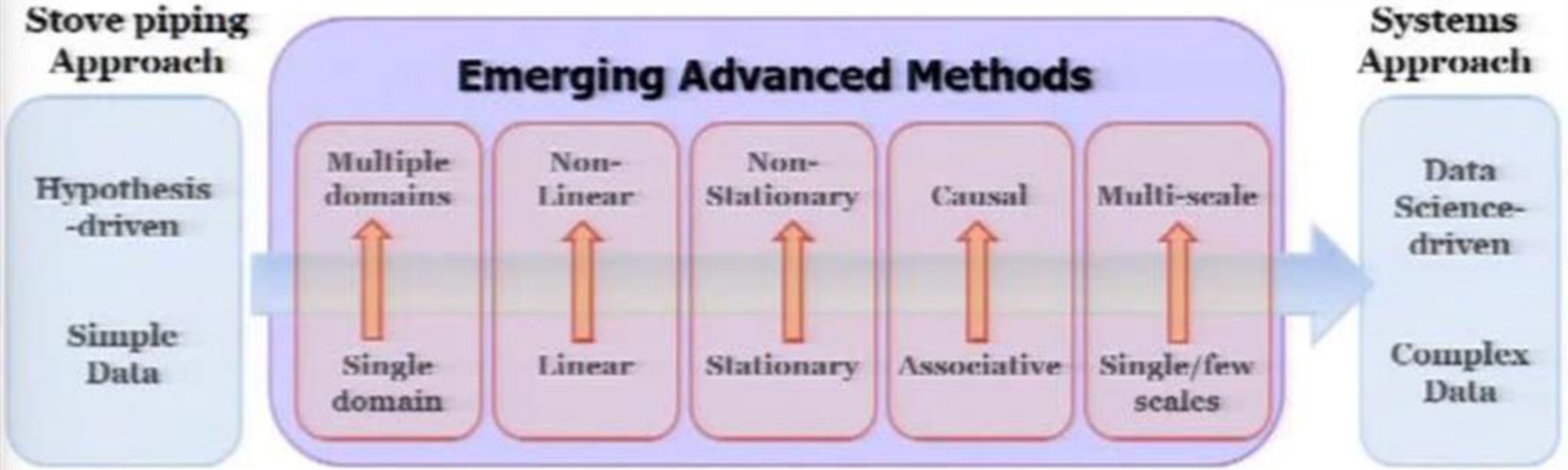
Providing Data-Driven Actionable Decisions

- Produce bioenergy
- Stabilize CO_2
- Clean toxic waste

Challenges of Data Science

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation and Other Ethical Issues
- Streaming/Dynamic/Distributed Data

Trends in Data Science Research



Application Drivers

- Health risk-aware energy systems
- Water management policies as climate change adaptation
- Environmentally influenced human diseases
- Sustainable local foods & health in changing climate
- Adaptation to climate extremes at regional and decadal scales

Big Data: The next frontier for leadership

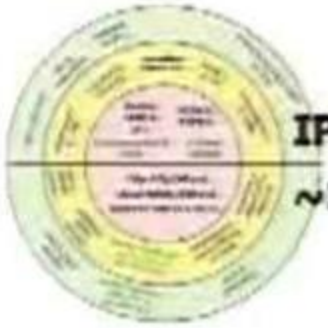
In this "Era of Big Data" it is obvious that *whoever controls the data-to-knowledge transformation controls the science & technology.*

From data-poor to data-rich transformation

$100 billion data management and analytics sector economy, growing at almost 10 percent per year, twice as fast as the software sector as a whole
(Cukier 2010)

"The world ... has changed ... data-intensive science [is] so different that it is worth distinguishing [it] ... as a new, fourth paradigm for ... exploration."

--Jim Gray



IPCC Simulations:

~2PB of data expected

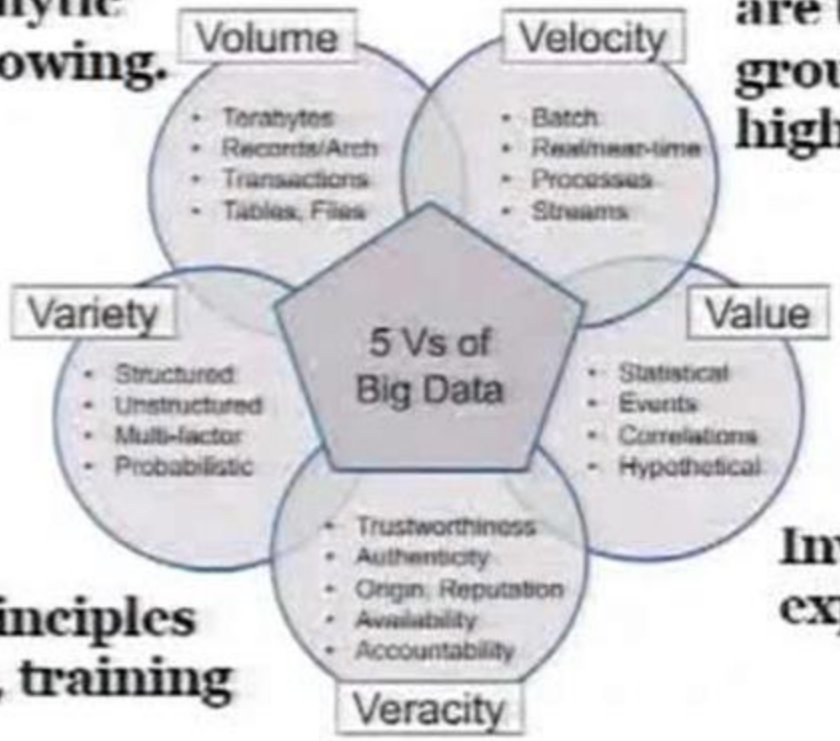
McKinsey & Co.

The United States alone faces a shortage of 140,000-190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data.

Implications of 5Vs of Big Data

Demands for more sophisticated analytic workflows are growing.

Powerful analytic techniques are the privilege of an elite group of experts, who are in high demand.



Investments to train experts are high.

Fundamental principles (e.g., overfitting, training bias) are easily overlooked.

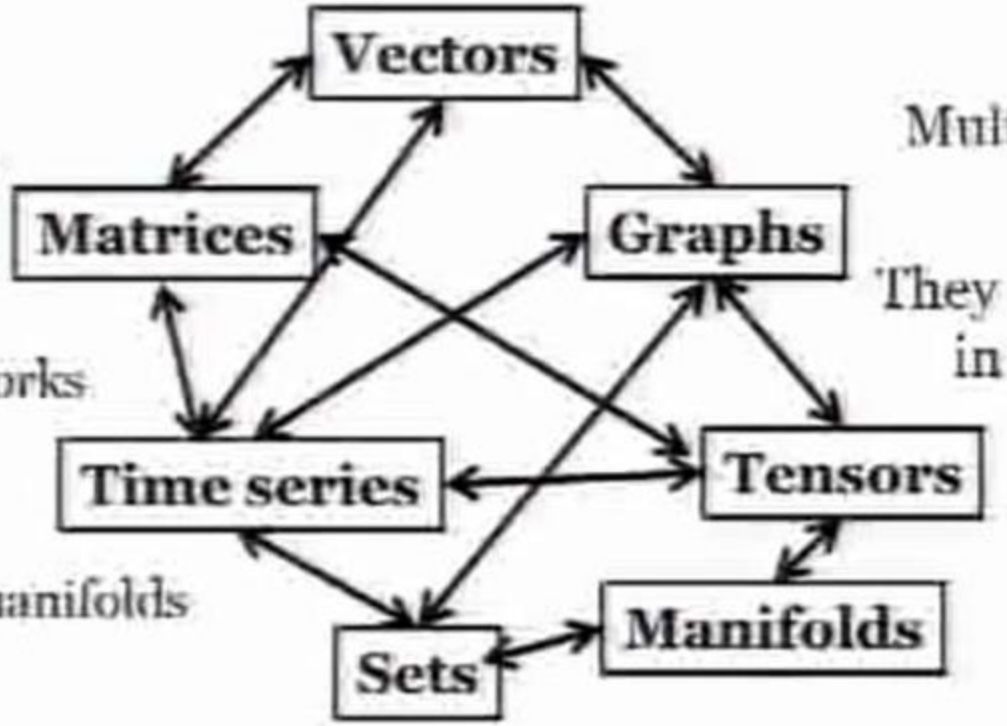
Improper use of analytic techniques leads to erroneous or flawed business decisions or scientific discoveries.

How to represent data mathematically?

Data Object & its Features→ Data Model

Different techniques require different *abstractions for data representation*:

- Scalar
- Points
- Vectors
- Vector spaces
- Matrices
- Sets
- Graphs, networks
- Tensors
- Time series
- Topological manifolds
- ...



Not one hat fits all

Multiple representations are needed

They are related but often in complementary way

Mathematical Data Representation (Data Model)

Big Data problems need Graph Analysis

Health Care	• Finding outbreaks, population epidemiology
Social Networks	• Advertising, searching, grouping, influence
Intelligence	• Decisions at scale, regulating algorithms
Systems Biology	• Understanding interactions, drug design
Power Grid	• Disruptions, conversion
Simulation	• Discrete events, cracking meshes

Graphs are a unifying motif for data analysis.

Source: David A. Bader

Why Would Graph Data Analytics Matter?

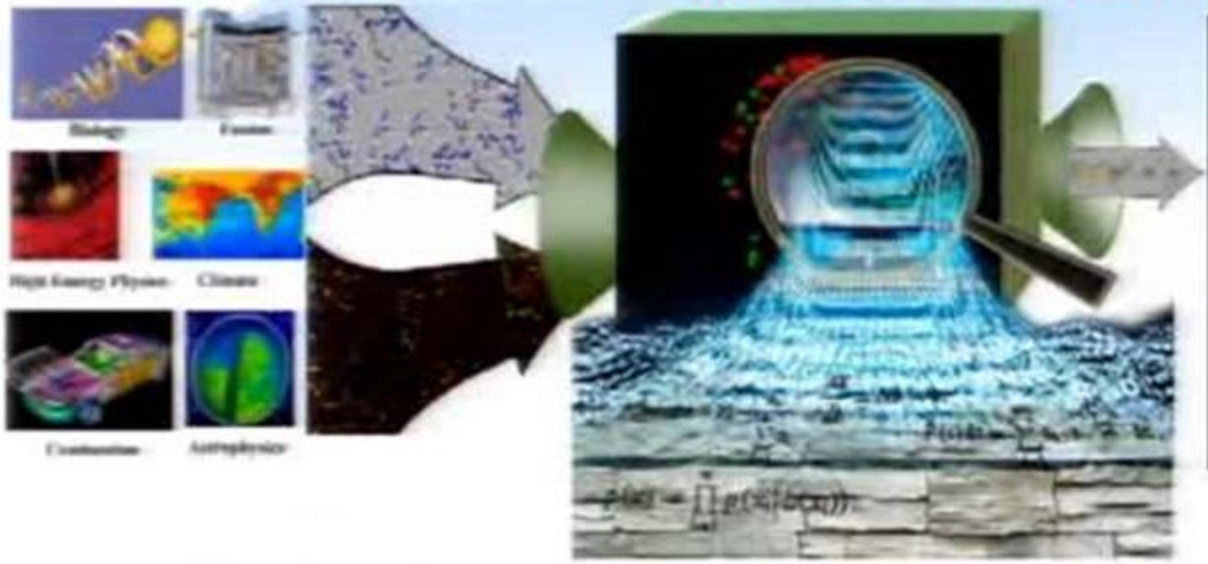
Enables solving many large-scale data problems

Draws ideas from many fields:

- Machine Learning and AI
- Data Mining
- Pattern Recognition
- Database systems
- Statistics
- Mathematics
- Graph Theory

- **Data reduction**
- **Capture inter-connections**
- **Powerful data model**

Data-intensive Apps → **Graph Data Analytics** → **Actionable Intelligence**



- Emergency response to weather extremes
- Real-time fraud preventions
- Cyberattack detection
- Health care monitoring & forecasting patient's state
- ....

The taste of Graph Data Analytics…

Like a dim-sum meal

Graph Data Analytics

PART I: STATIC SINGLE GRAPH

The Curse & Blessings of Intractability

Alan Turing, 1936

"Problems that would otherwise be impossible to solve can now be computed, as long as one settles for what happens on the average." – J. F. Traub and H. Woz´niakowski

Stanislaw M. Ulam
Nicholas Metropolis

One of the mathematical achievements of the last century was the idea that mathematical problems maybe undecidable, non-computable, or intractable (Turing, Godel, Church).

Many data-related problems are intractable computationally, even on a supercomputer.

Blessings of Intractability

- *Randomization-based feasibility*
- *Average-case complexity*

"Breaking Intractability," SCIENTIFIC AMERICAN, January 1994

Examples: Intractable Problems

Many problems (discovery of disease genes, network homology, protein 3-d structure matching) are computationally intractable.

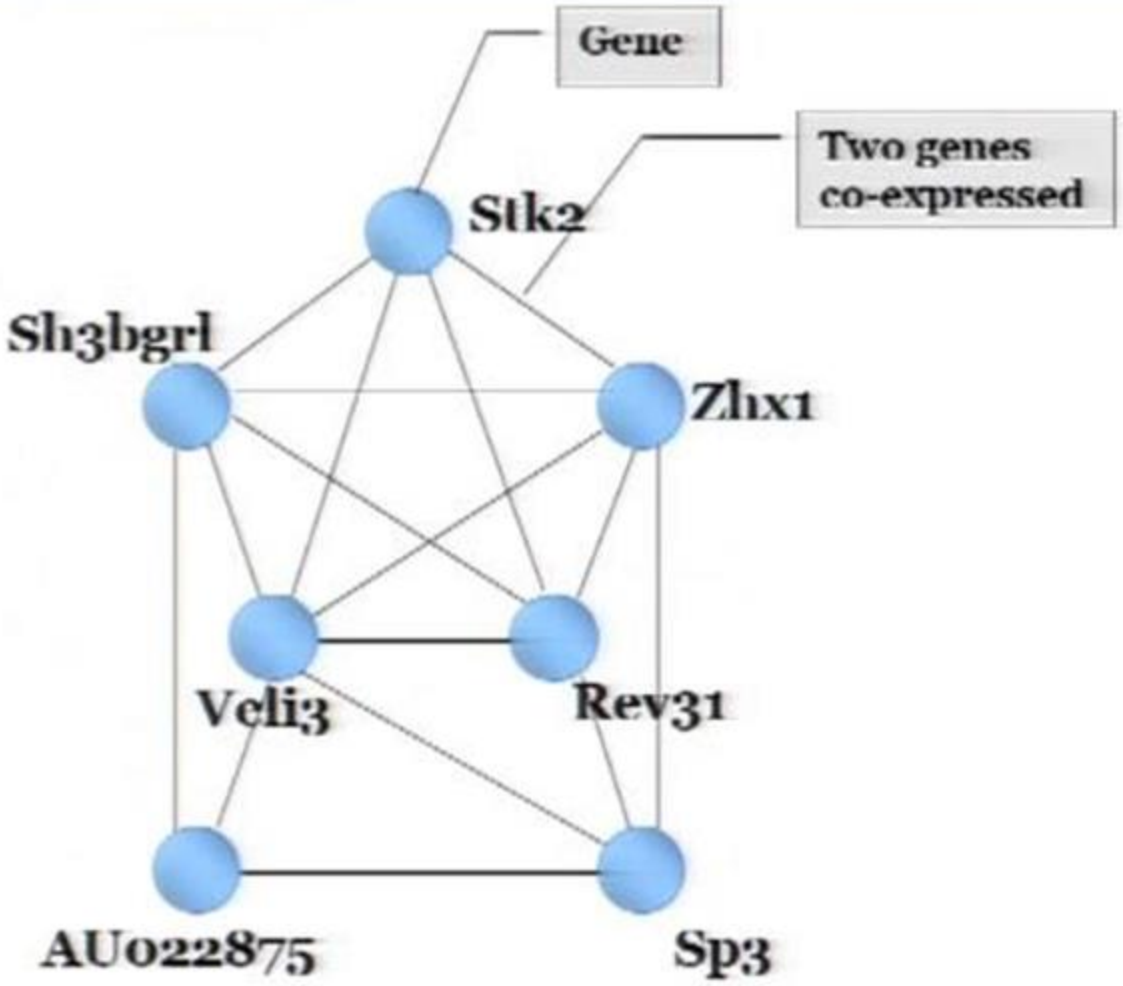
Routine Questions:

- Identify a minimum set of genes to knockout to disrupt a set of biochemical pathways (**Minimum Vertex Cover**);
- Discover pathways that are similar to a given signaling pathway (**Sub-graph Isomorphism**);
- Find all co-expressed gene clusters in microarray data (**Maximal Clique**);

These questions are *NP*-hard problems!

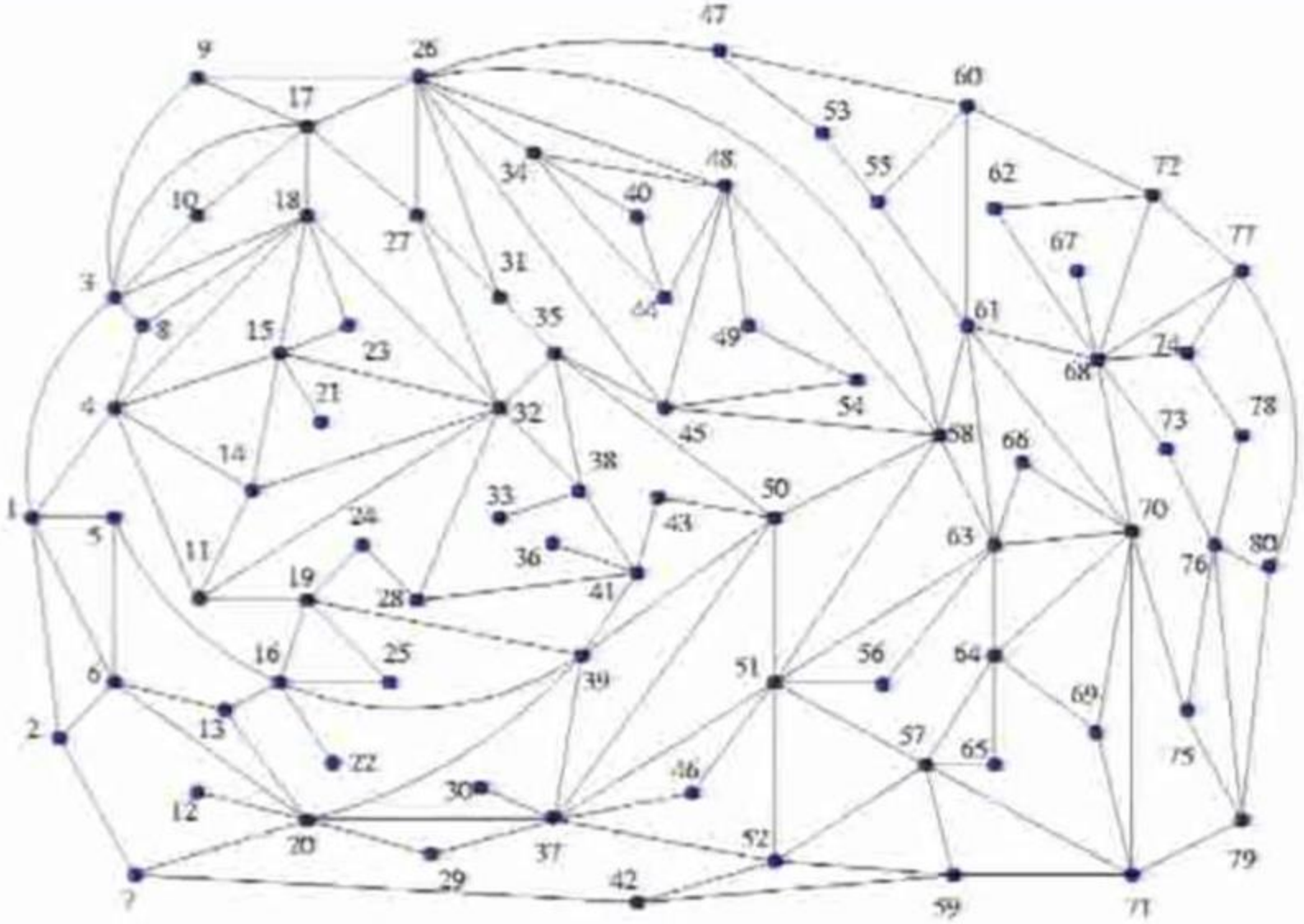
Graphs do not even have to be BIG!!!

Example: Maximum Clique Problem



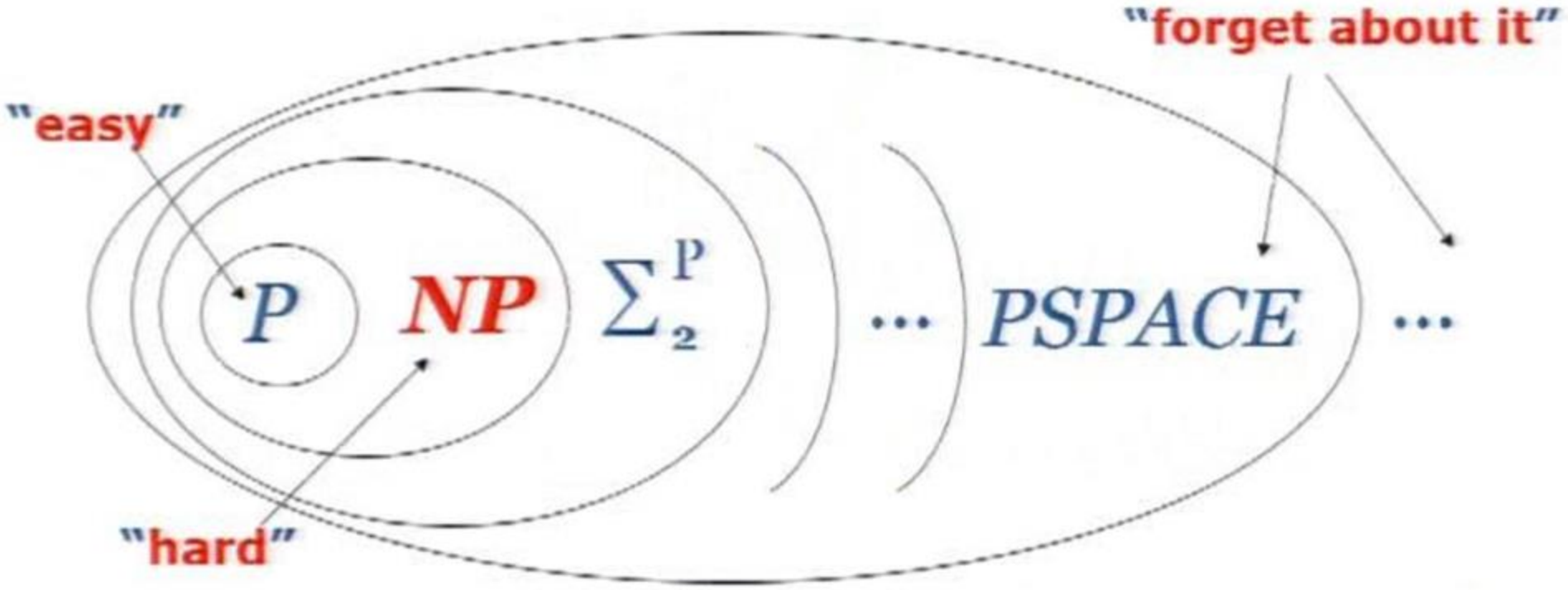
Genes from Mus musculus brain

Does this graph contain a 4-clique?



Classic Complexity Theory

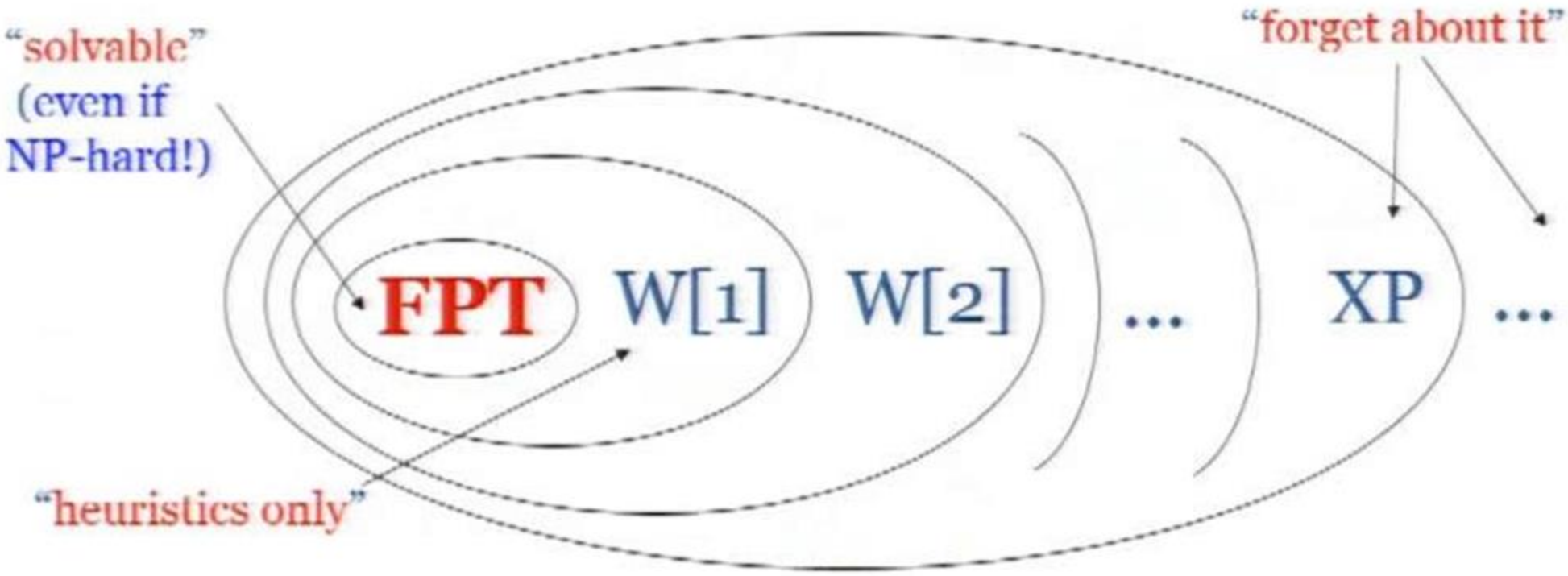
The Classic View:



"About ten years ago some computer scientists came by and said they heard we have some really cool problems. They showed that the problems are NP-complete and went away!"

Parameterized Complexity Theory

Hence, the Parameterized View:



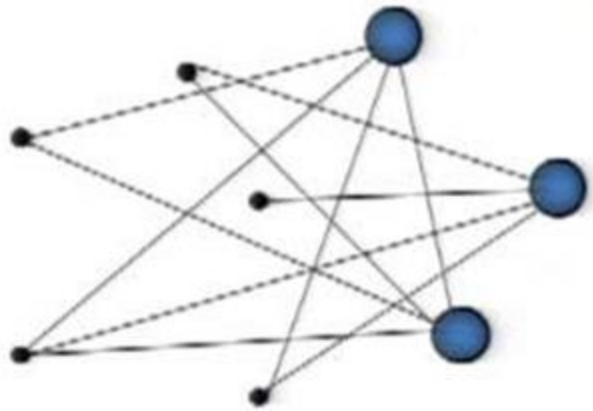
Pioneering work of Fellows and Langston

NP-hard ≠ FPT

Not every NP-hard problem is FPT. Many non-FPT problems can be reduced to an FPT one in polynomial time.

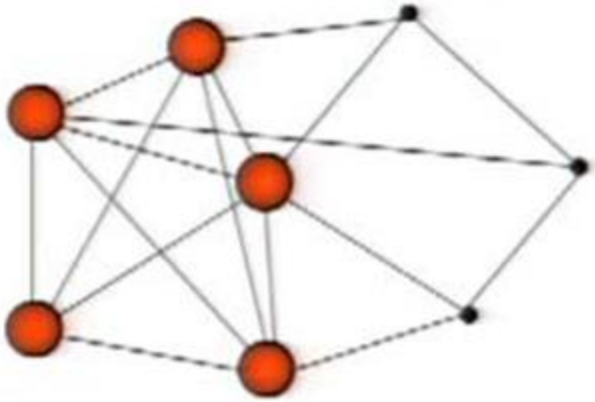
Minimum Vertex Cover is FPT

- Reduce in *polynomial* time
- Via *dual* graph



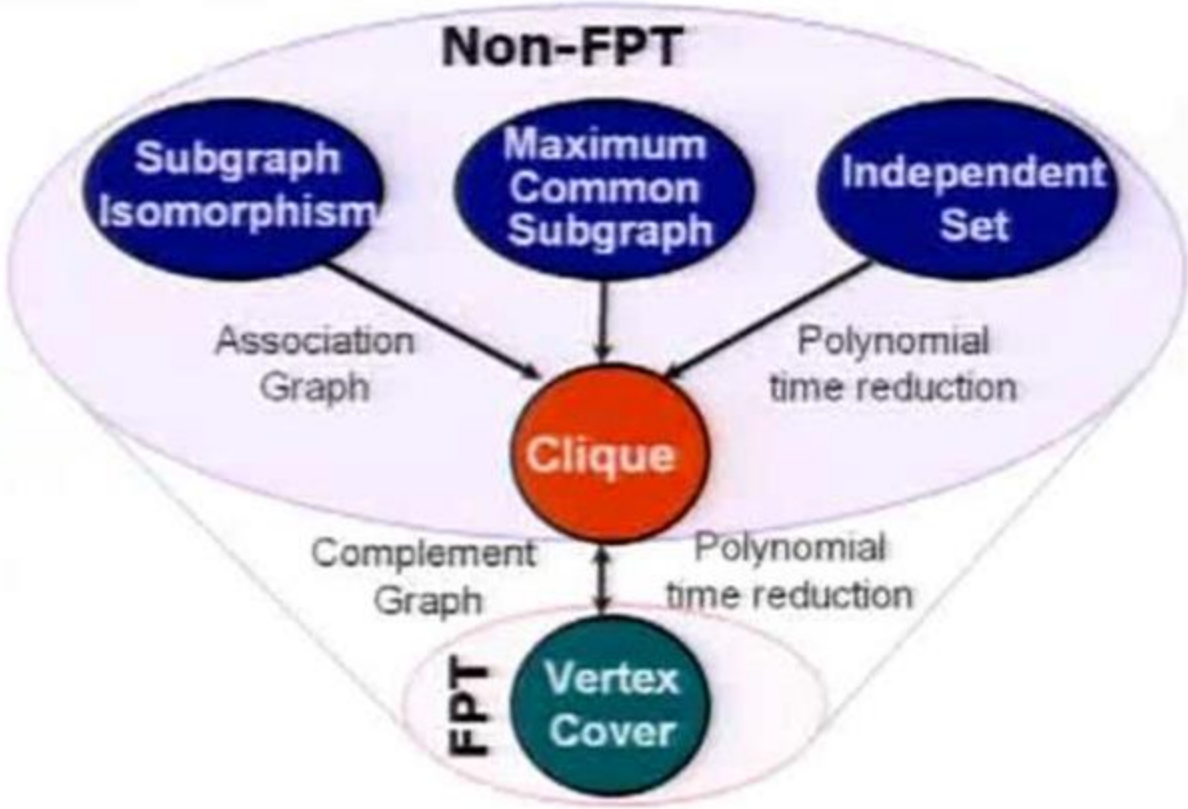
Minimum set of vertices that cover ALL edges in the graph

Maximum Clique is not FPT



Solving Many NP-hard Problems

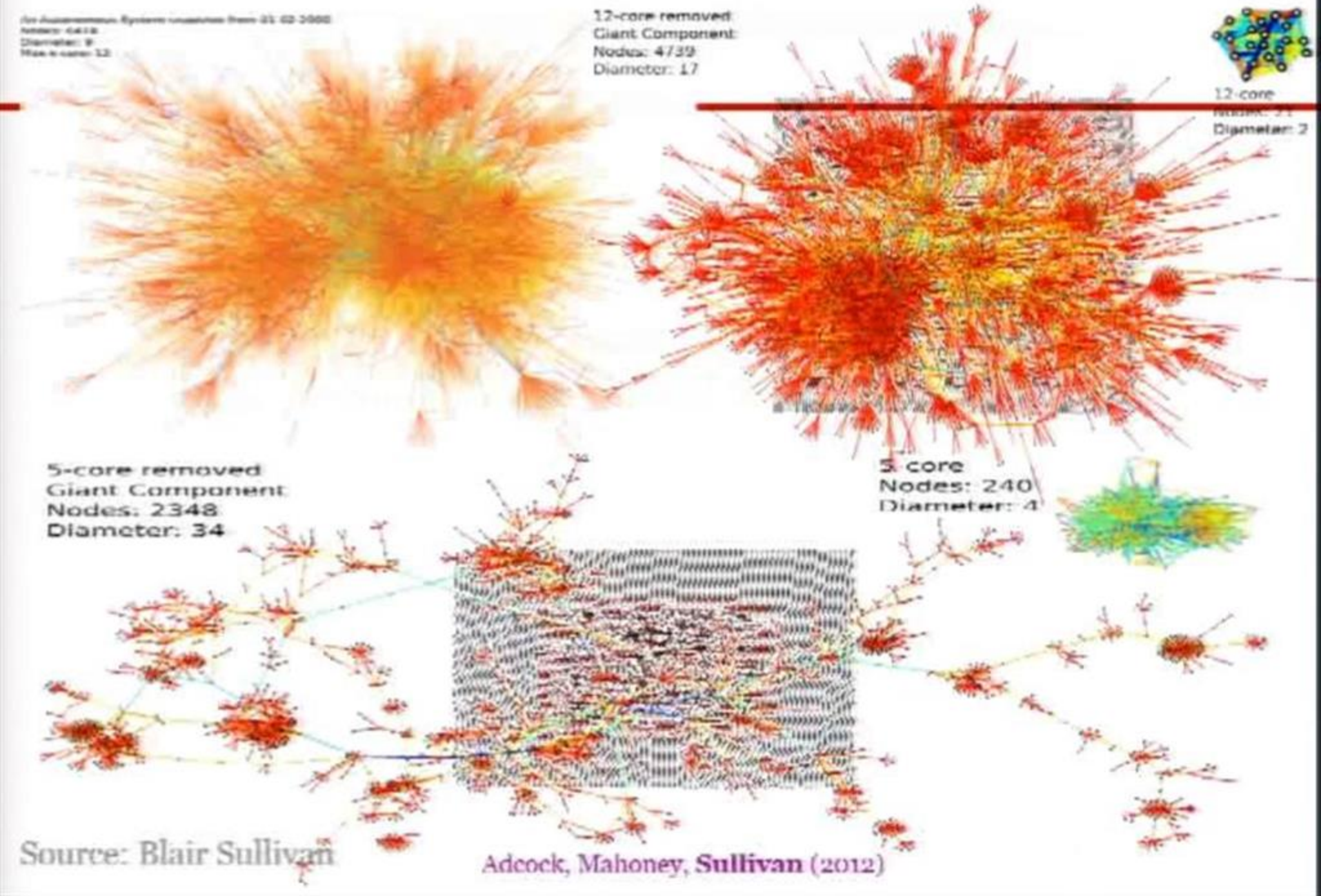
Focus on a few optimized core algorithms and efficiently solving many other NP-hard non-FPT graph problems.



Take-away Message

- **Many real-world problems are reduced to problems on graphs**
- **While graph problems are often computationally intractable (due to NP-hard nature), recent advanced in Fixed Parameter Tractability (FPT) offer practical solutions**

Hints of Structure: k-cores expose tree-like backbone



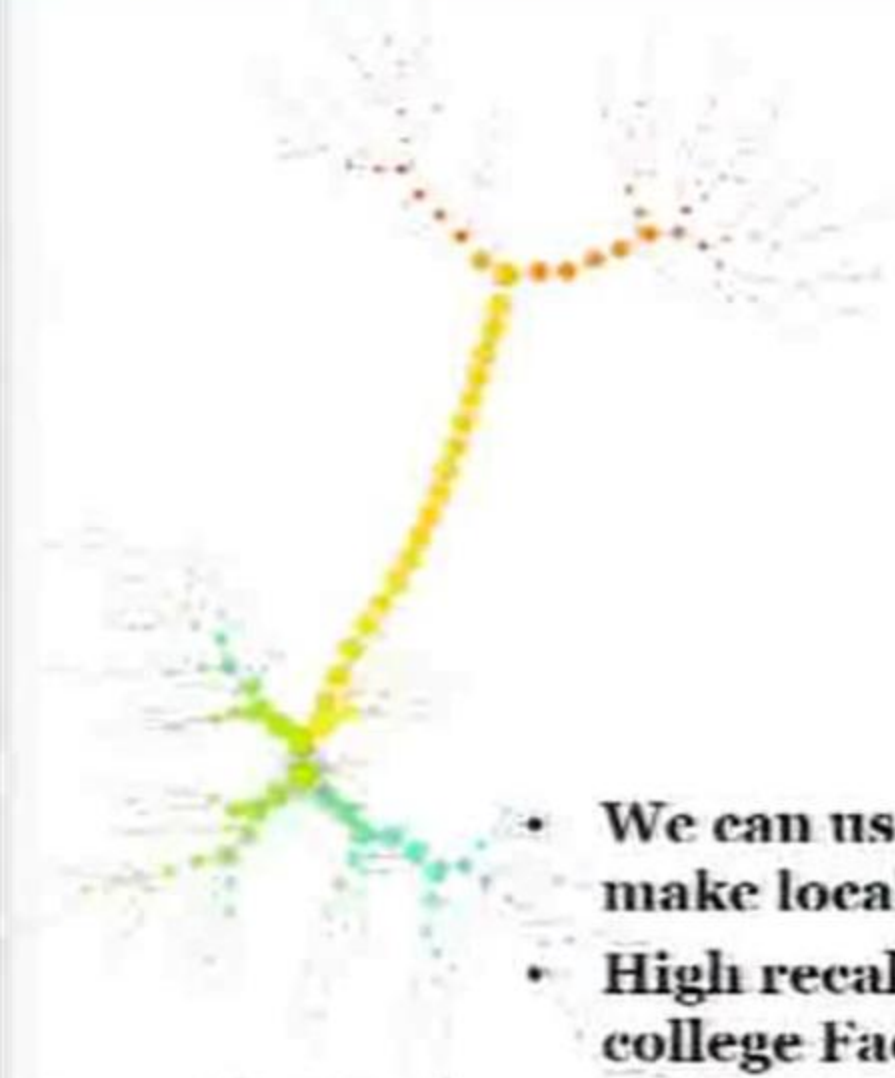
Source: Blair Sullivan

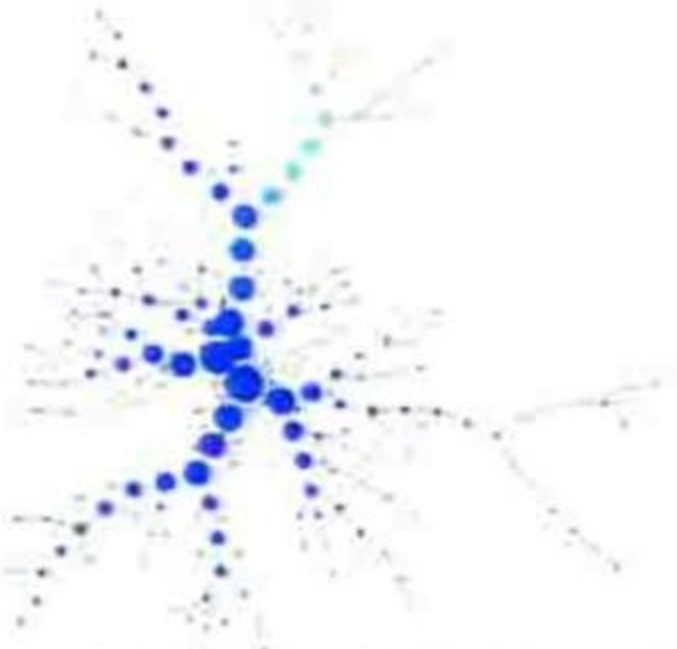
Adcock, Mahoney, **Sullivan** (2012)

Hints of Structure:

Tree decompositions respect ground-truth communities

Adcock, Mahoney, Sullivan (2014)





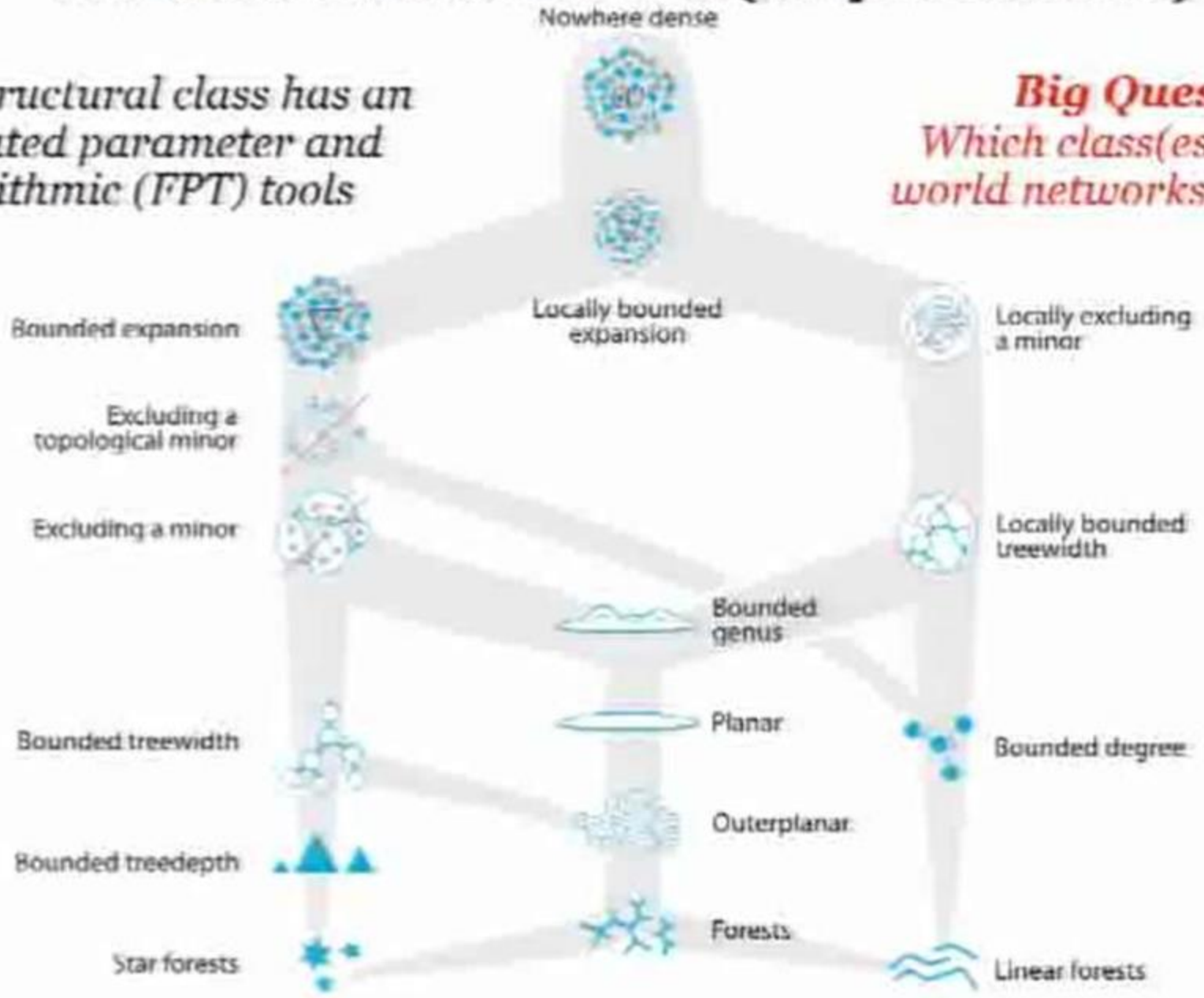
- We can use framework of tree decompositions to make localization algorithmic
- High recall when identifying communities in college Facebook networks (heat-maps indicate concentration of selected community)

Source: Blair Sullivan

What's the best-fit (exploitable)

Each structural class has an associated parameter and algorithmic (FPT) tools

Big Question: *Which class(es) do real-world networks belong to?*

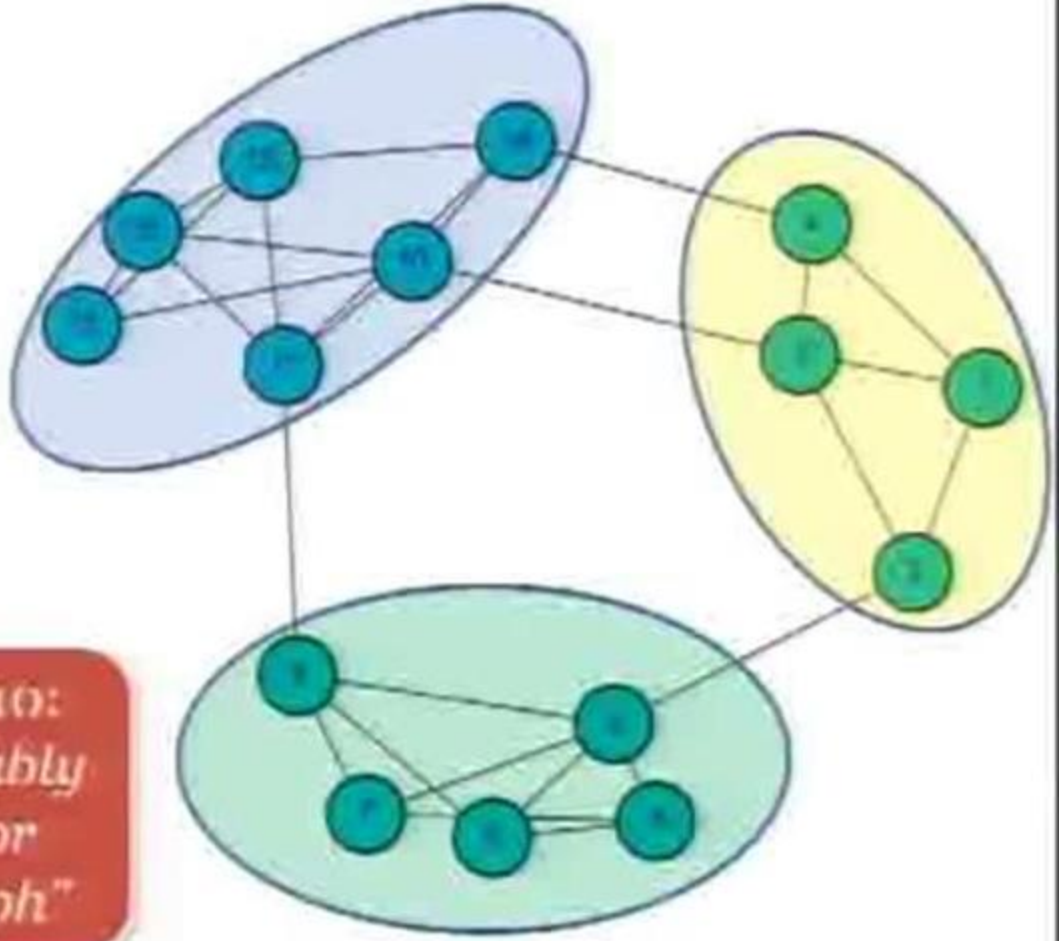


Source: Blair Sullivan

Graph Mining Task: Community Detection

- Find groups of highly connected nodes that have few inter-group connections

- Many different types:
  - Disjoint
  - Overlapping
  - Weighted
  - Hierarchical



Fortunato in Physics Reports, 2010:
"... groups of vertices which probably share common properties and/or play similar roles within the graph"

Disjoint Communities: Modularity

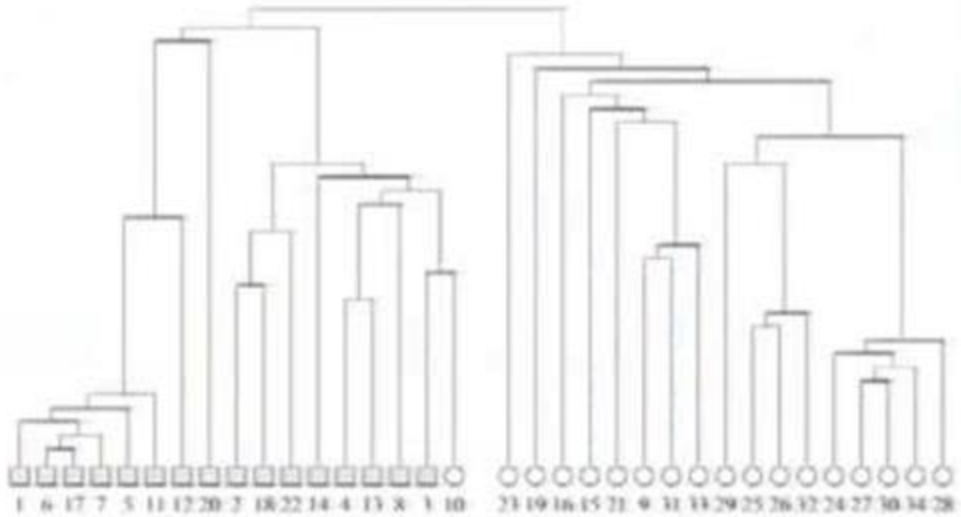
- **Modularity** measures the fraction of edges that fall within the communities minus the expected value if the edges fall at random

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{i,j} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

- $\delta(C_i, C_j) = 0$ or 1 depending if i and j are in the same community
- k_i = degree of node i

Greedy Algorithm:

- Initialize each vertex in its own community
- Merge two communities that results in highest modularity gain
- Repeat and select best level



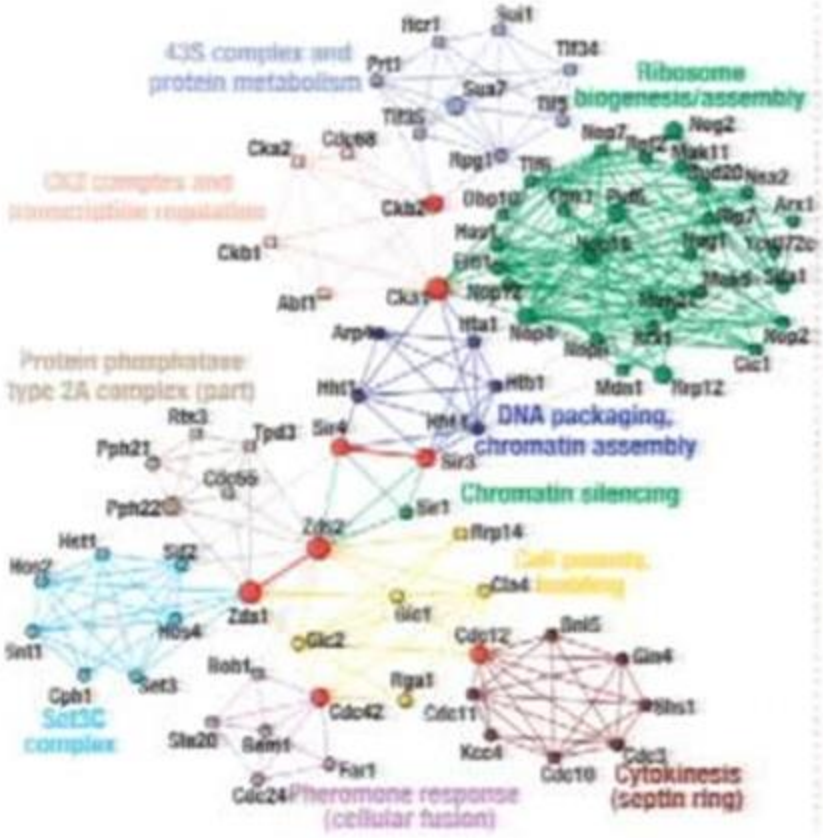
Physical review E 69.6 (2004): 066133

Overlapping Communities: Clique Percolation

- **Motivation:** Nodes commonly belong to more than one cohesive group
- **Intuition:** communities usually consist of several complete subgraphs

Algorithm:
- Find all cliques of size k
- Merge cliques that share k-1 edges

- **Exponential run-time ☹**



Nature 435.7043 (2005): 814-818

(Overlapping) Communities: Label Propagation

- **Motivation:** optimizing for some community metric (e.g., modularity, density) can be slow
- **Intuition:** use network structure to guide community detection process
- **Linear run-time** ☺

SLPA Algorithm:

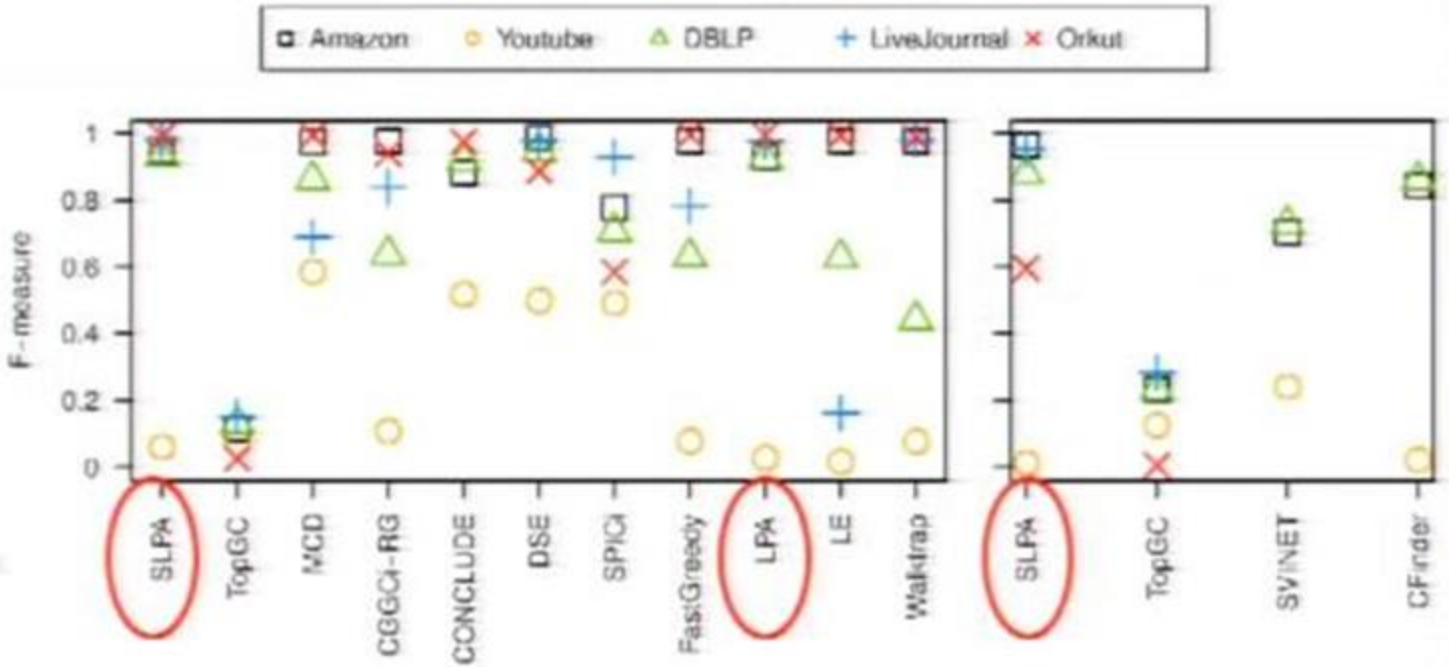
- Initialize each node w/ a unique label
- At each step, every node takes the most popular label of its neighbors

Physical Review E 76.3 (2007): 036106

A Survey & Empirical Evaluation against Ground Truth Communities

SLPA did not identify communities with good clustering coefficient, but yielded the communities most similar to the ground-truth.

These results show that goodness metrics and performance metrics are not equivalent: communities with "good" structural properties do not necessarily yield good performance metrics.



Algorithms were empirically compared using goodness metrics that measure the structural properties of the identified communities, as well as performance metrics that evaluate these communities against the ground-truth

Steve Harenberg, et al: Community detection in large-scale networks: A survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*: 1939-0068, 2014.

Take-away Message

- **Community detection is a ubiquitous graph data analytics task.**
- **However, community definition is problem-dependent.**
- **Evaluating community detection methods requires ground-truth communities for the problem at hand.**
- **Too many communities could be detection: *which ones are relevant to the problem at hand*?**

Coupling Data-driven with Analyst-guided Discovery

Challenge: To organically consolidate the two complementary strategies: ***data-driven*** and ***analyst-guided***.

Data-driven methods: Aim to find *all* the solutions to the target problem that satisfy the objective function(s) and constraints:

- Find novel and unexpected patterns or states, i.e., an Aha moment,
- Better for finding strong and most probable signals in the data, and
- Less robust to finding rare patterns, esp. in the presence of noise or the complex mixture of signals.

Analyst-guided methods: Bring user's expert knowledge to guide the search to the regions of interest:

- Find solutions that are most relevant to the target problem at hand,
- Significantly reduce space of solutions for manual examination,
- Perform faster computation, and
- Find rare, less expected patterns

The user-guided approach completes the data-driven approach for "what-if" imagination.

Query-Driven Community Detection

Challenges

Traditional graph mining algorithms (e.g., anomaly detection, dense substructures, frequent subgraphs, etc.) can be prohibitively compute-/memory-intensive

Graph databases provide scalable solutions using index structures yet have a limited selection of problems they can solve (e.g., traversal-based queries and exact subgraph matching).

Technology

Pre-computed index of partially enumerated search space

- Utilize index for subsequent queries to improve query response time and reduce redundant computations

Out-of-core algorithms that operate on disk resident graphs, fetching parts into memory as necessary

Result/Impact

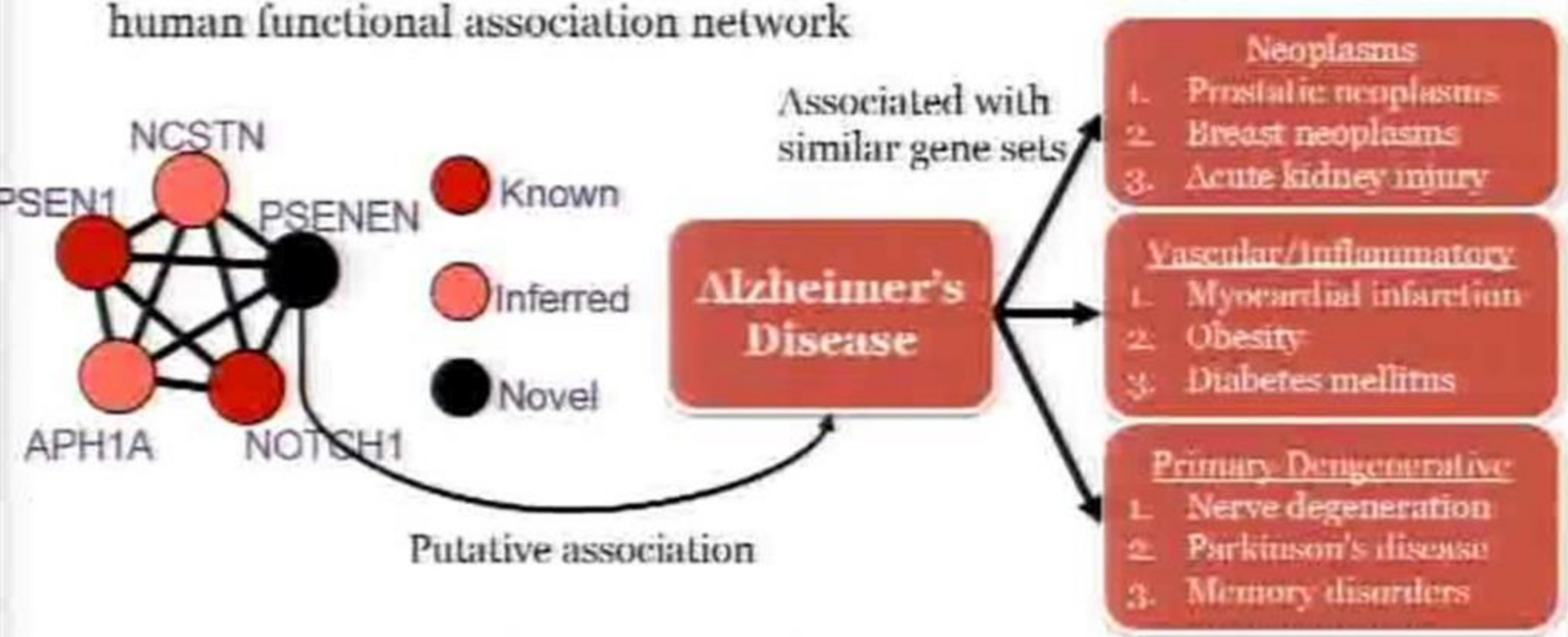
As high as 100x speedups over state-of-the-art when utilizing pre-computed index.

Significantly reduce peak memory usage compared to state-of-the-art, up to 1000x (regularly >40x).

Steve Harenberg, et al: Memory-efficient Query-driven Community Detection with Application to Complex Disease Associations. SIAM Data Mining 2014: 1010-1018

Community Detection with *Knowledge Priors*: Complex Disease Associations

- **Guilt-by-association:** genes functionally associated to many other genes related to Alzheimer's are also likely associated
- Community detection used to find groups of functionally related genes in human functional association network



Steve Harenberg, et al: Memory-efficient Query-driven Community Detection with Application to Complex Disease Associations. SIAM Data Mining 2014: 1010-1018

Response-associated Community Detection

- **Problem**: Identify communities that can be used to analyze or predict a response variable of interest (e.g., rainfall).
- **State-of-the-art**: Community detection techniques are traditionally *unsupervised* learning methods→communities identified may *not* be associated with the response variable.
- ***Supervised* community detection**: *identify communities associated with a response variable of interest* by explicitly incorporating information of this variable during the community detection process.

Gonzalo Bello, et al: Supervised Community Detection: Application to Climate Index Discovery. 2015 (Under Review)

Supervised Community Detection

- Goal: partition a given graph into a set of disjoint communities.
- Joint optimization criterion: maximize both the "goodness" of the partition and the association of the communities with the response variable:

$$\alpha \cdot q(C) + (1 - \alpha) \cdot \bar{\phi}_C$$

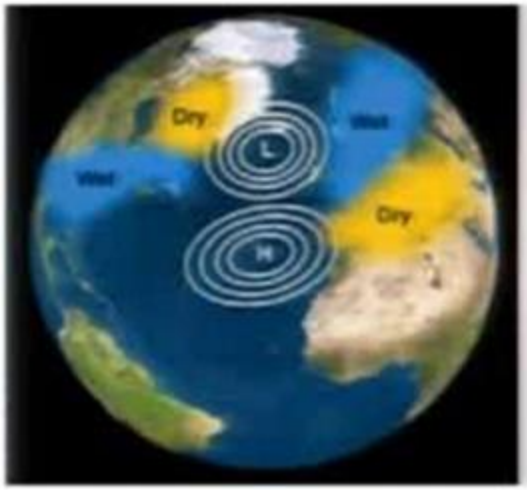
where C is a set of communities, q is a function of the "goodness" of C, $\bar{\phi}_C$ is the average association of the communities with the response variable, and α is a tuning parameter to balance the trade-off between $q(C)$ and ϕ_C.

- Use modularity as the "goodness" function q.
- Use correlation to measure the association between a community and the response variable.
- Modify heuristic algorithm for modularity maximization, such as the Louvain method (Blondel et al., 2008), to maximize the joint optimization criterion.

Communities in Climate Networks: Climate Indices

Climate indices are defined to quantify climatic phenomena

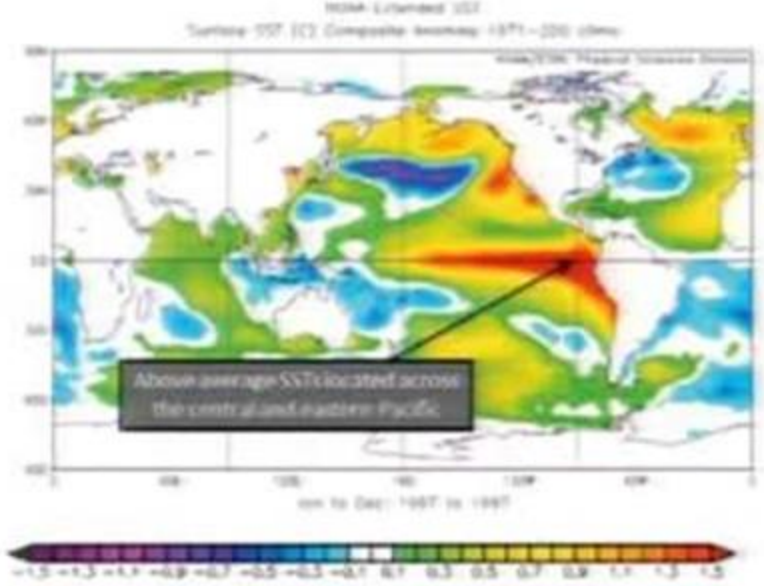
Many of them are defined in terms of teleconnection patterns or dipoles



North Atlantic Oscillation

Dipole - difference in sea level pressure between the azores and a region near Iceland

NAO



El Niño (Warm Phase)

Teleconnection pattern - above average Sea Surface Temperature across the tropical Pacific

leads to drought like conditions in the Sahel region

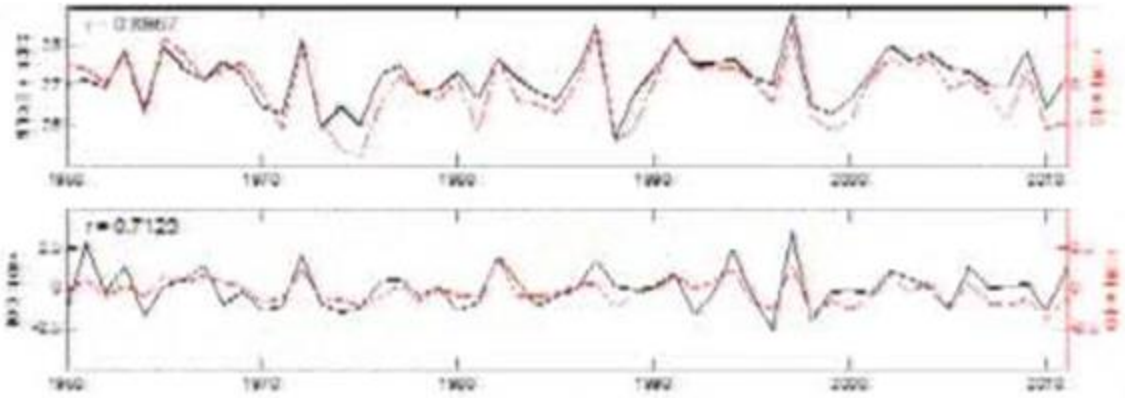
Nino1+2 | Nino3 | Nino4 | Nino3.4

ENSO index family

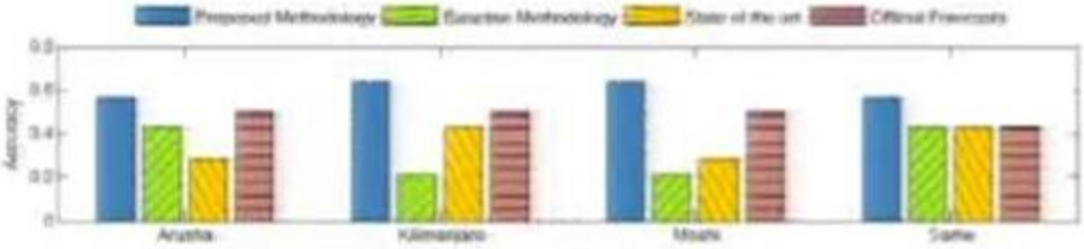
Evaluation: Climate Indices Discovered

- Climatological relevance of climate indices discovered is supported by their association with traditional climate indices known to be related to seasonal rainfall in the region.

- Improvement of forecast skill for seasonal rainfall in the region with respect to existing methodologies for climate index discovery and official forecasts.



Time series and correlation of traditional climate indices (Nino 3.4 and IOD, respectively) with climate indices discovered



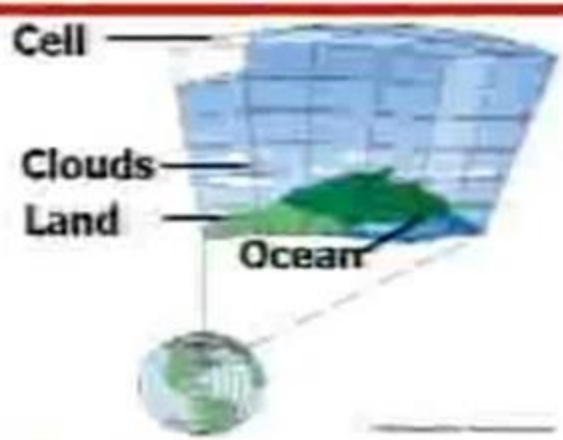
Accuracy of the prediction of seasonal rainfall at 4 stations in the Greater Horn of Africa from 1998 to 2011

Take-away Message

Incorporating information of the response variable of interest during the community detection process allows for the discovery of communities with a higher association with this response variable and better predictive skill.

Understanding Climate Change – A Systems Challenge

Climate systems are complex because of non-linear coupling of its subsystems (e.g., the ocean and the atmosphere).



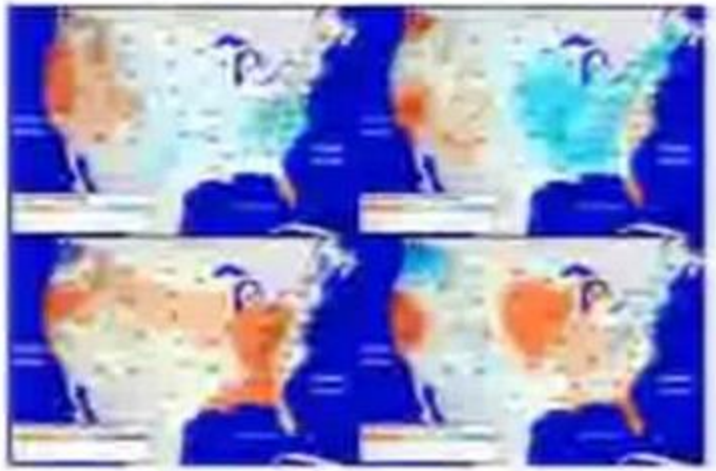
Parameterization and non-linearity of differential equations are sources for uncertainty!

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

Physics-based models are essential but not adequate

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

Disagreement between IPCC models

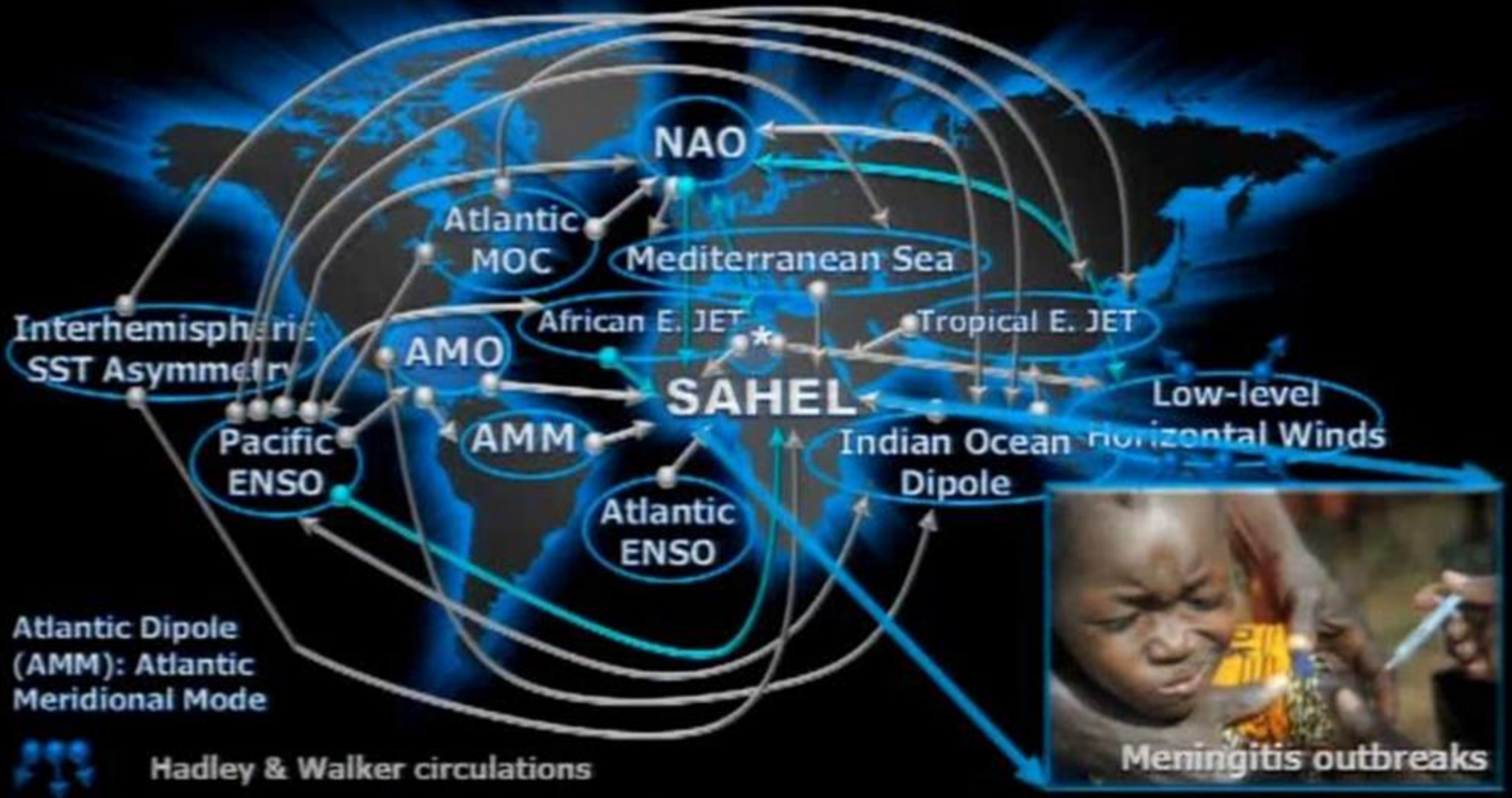


Regional hydrology variations among major IPCC model projections

1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

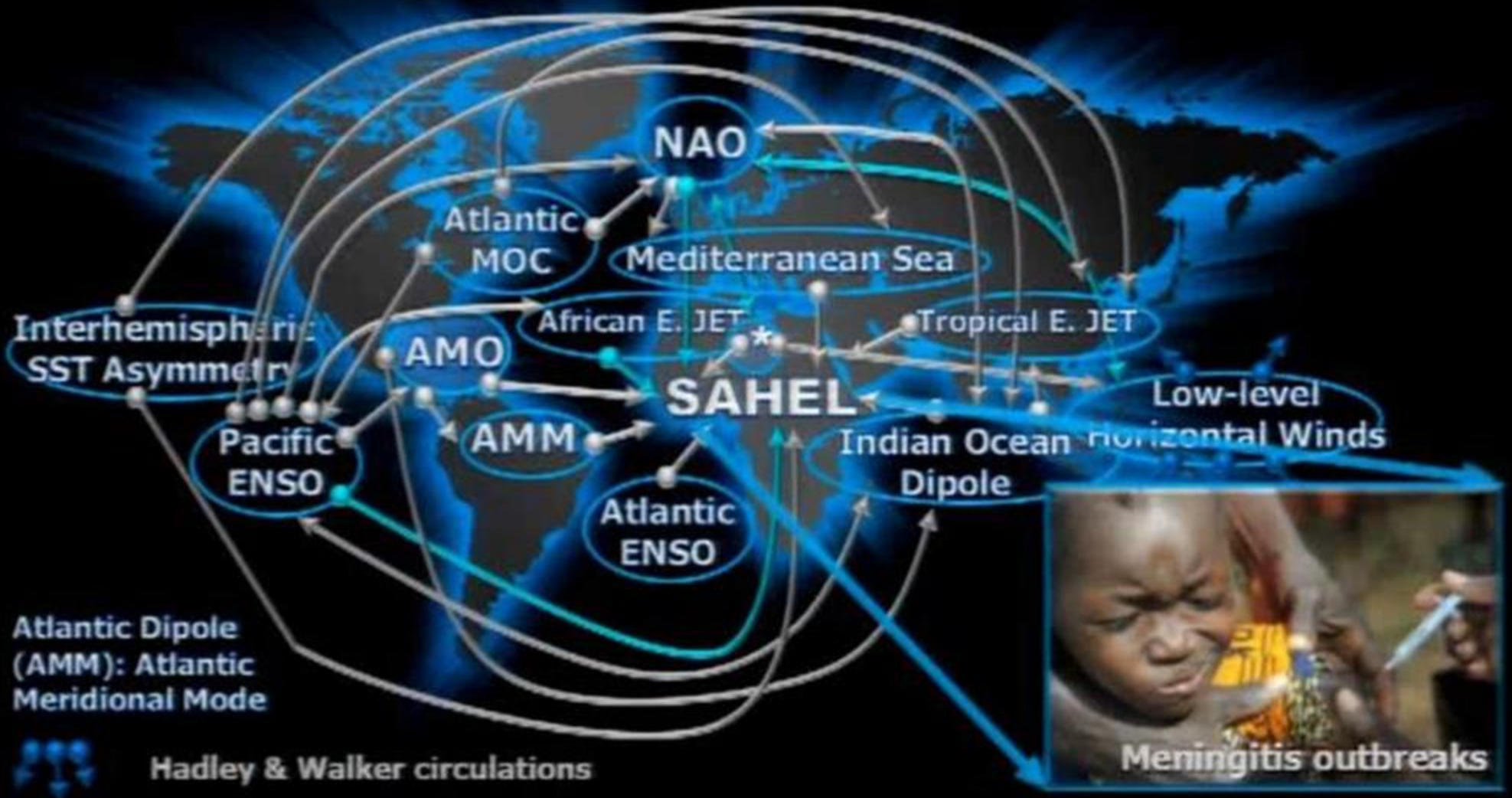
1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

Can data-driven approaches expedite such discoveries?



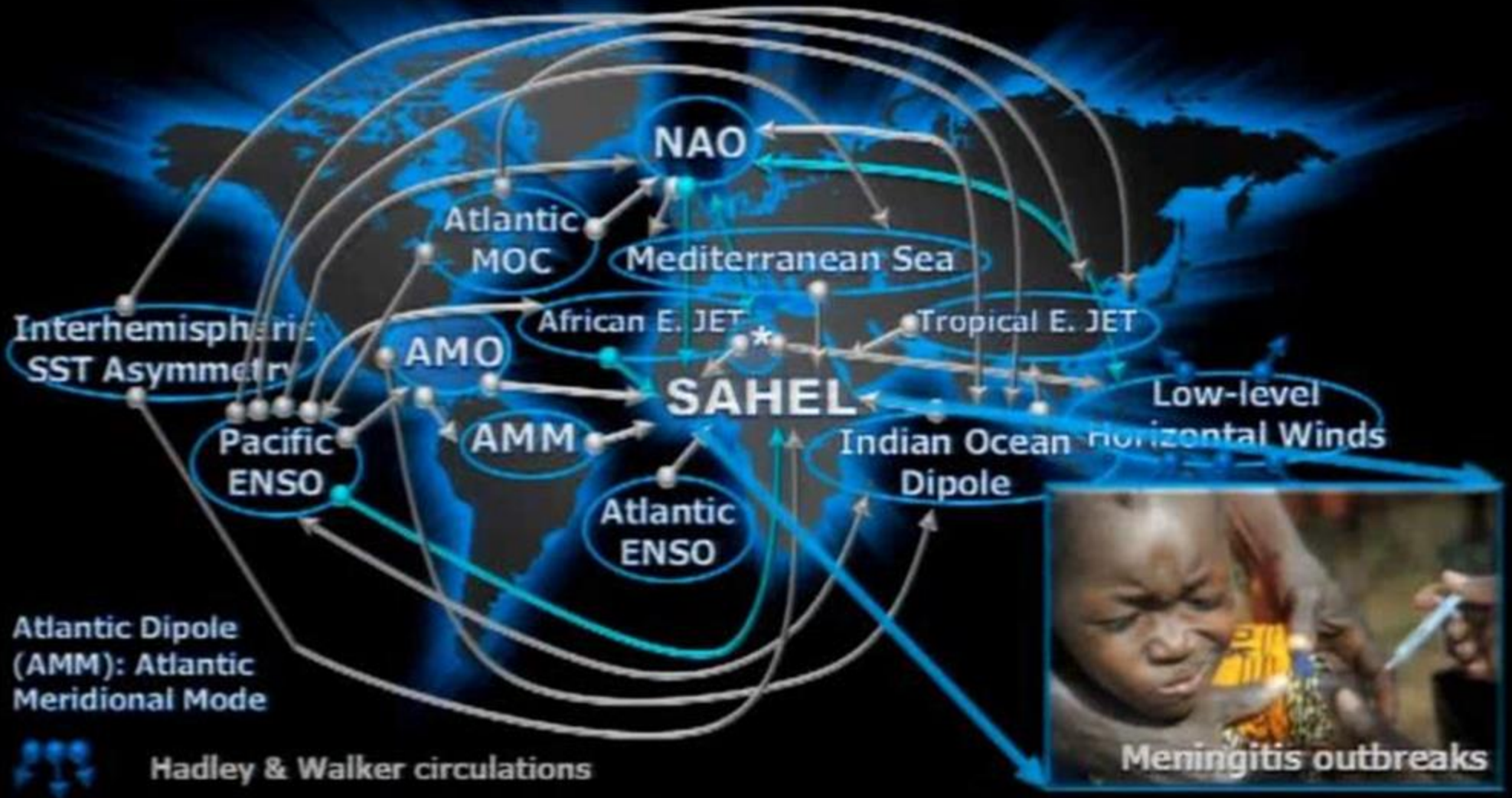
1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

Can data-driven approaches expedite such discoveries?

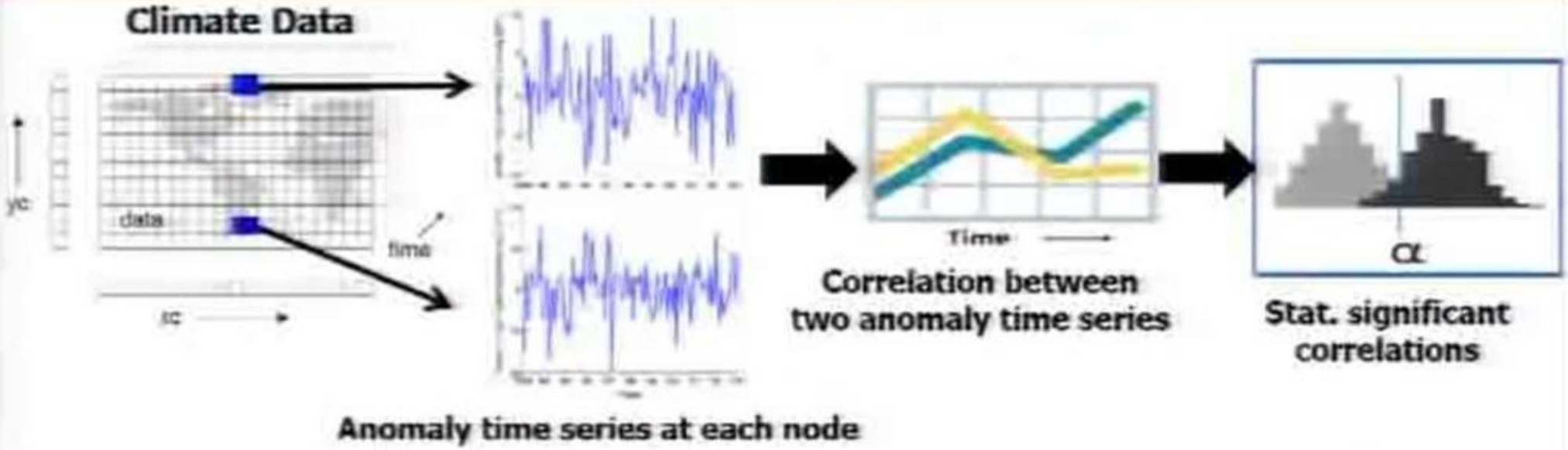


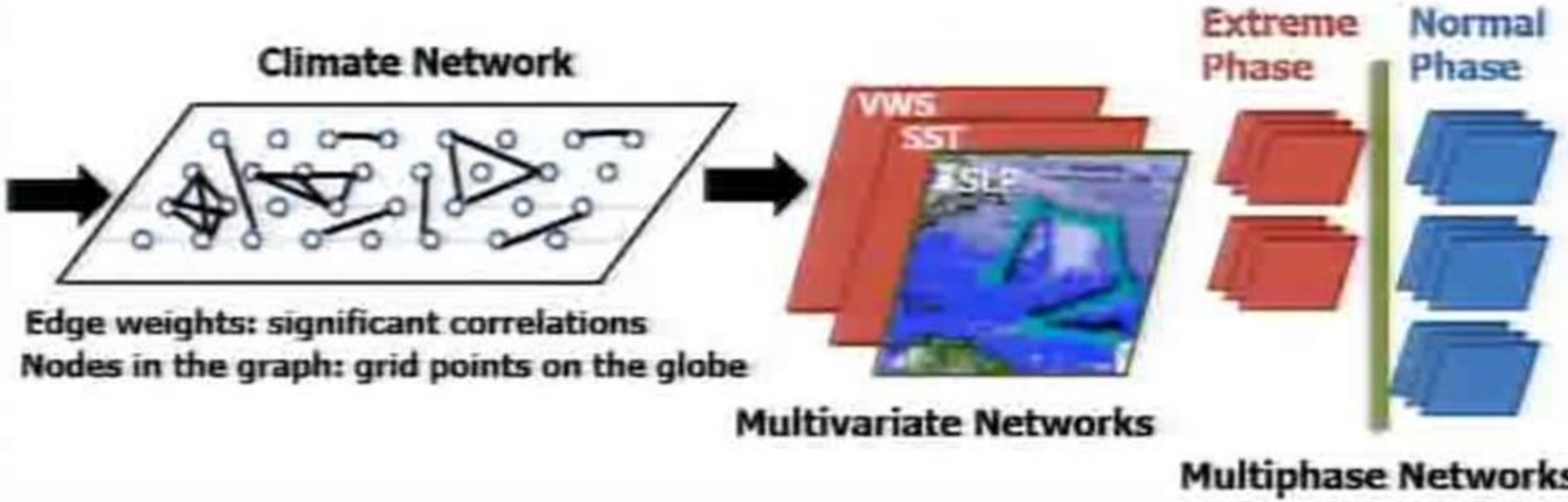
1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

Can data-driven approaches expedite such discoveries?



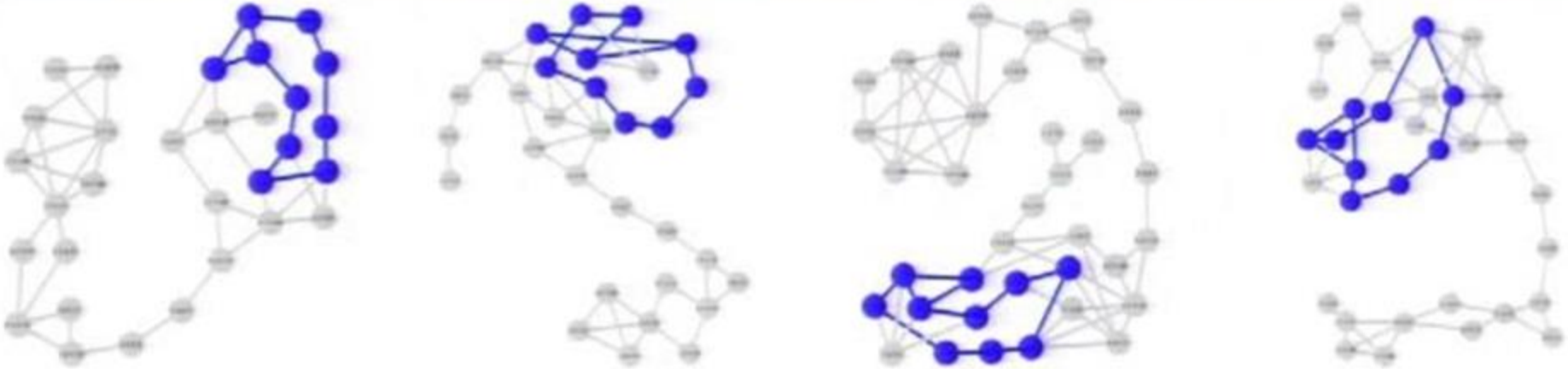
Data-driven Modeling of a Climate System as a Complex Network





Subgraphs Common to Extreme Event Climate Networks

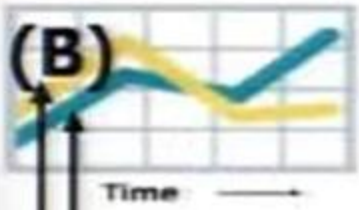
Networks for Climate Systems during Extreme Events

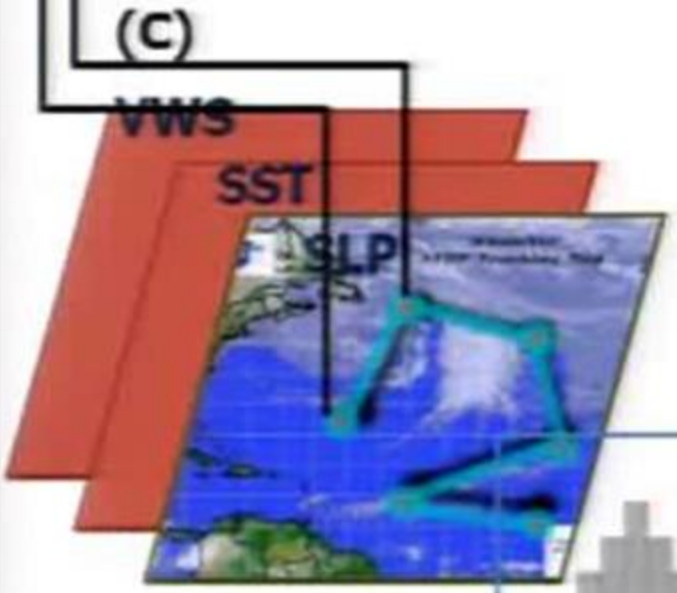


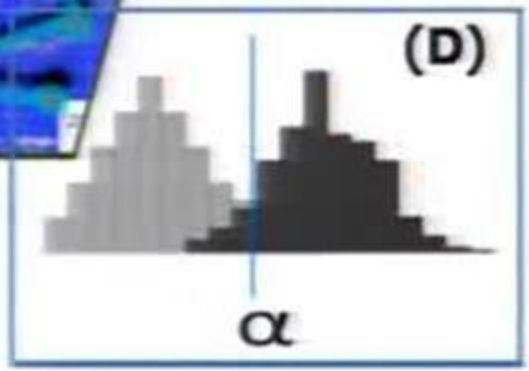
Networks for Climate Systems during Normal Events

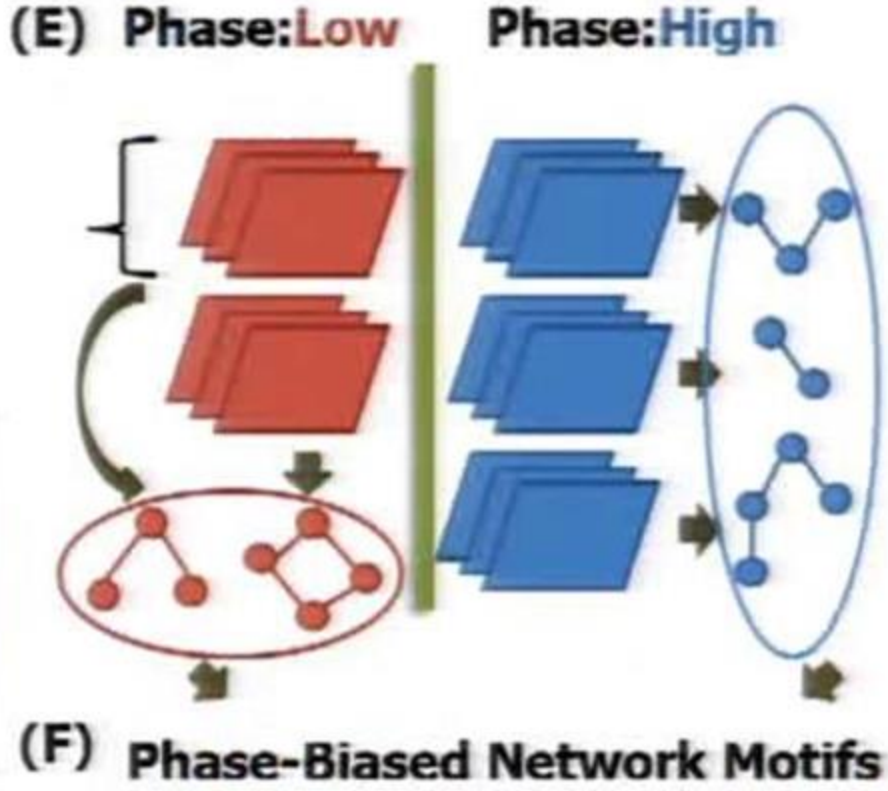
Contrast-based Network Motif Discovery for Extreme Event Forecasting

Intuition: If an extreme event (e.g. hurricanes) is in one of its key phases (e.g. high activity season), then there exist network motifs (recurrent patterns in climate networks) that are specific to that phase.









Hurricane Activity Class Forecast vs. State-of-art

FORECASTER Performance on North Atlantic Hurricane

Metric	FORECASTER NC State	[1], 2009 Colorado	[2], 2010 GA Tech	Random Forest	Bagging	Boosting
Accuracy (%)	**93.3**	64.0	65.5	76.7	73.3	75.0
HSS	**0.90**	0.45	0.49	0.66	0.60	0.62
PSS	**0.92**	0.44	0.50	0.65	0.63	0.63
GSS	**0.96**	0.50	0.68	0.65	0.67	0.66

ML-based **Regression** **Hybrid**

[1] P. J. Klotzbach and W. M. Gray, "Twenty-five years of Atlantic basin seasonal hurricane forecasts (1984-2008)," Geophys. Res. Lett., vol. 36, pp. L09 711, 5pp, May 2009.
[2] H. M. Kim and P. J. Webster. Extended-range seasonal hurricane forecasts for the North Atlantic with a **hybrid dynamical-statistical model**. Geophys. Res. Lett., 37(21):L21705, 2010.

HSS: Heidke score, measures how well relative to a randomly selected forecast;
PSS: Peirce score, difference between the hit rate and the false alarm rate;
GS: Gerrity score, occurrences substantially less frequent.

Graph Data Analytics

PART III: SEQUENCE OF GRAPHS