Discovering Knowledge from Massive Networks and Science Data – Next Frontier for HPC

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

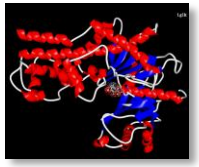
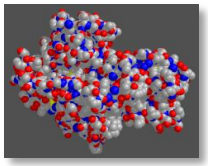
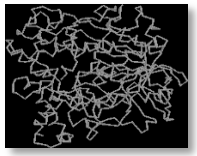
DOE – CSGF July 2012

National Science Foundation
WHERE DISCOVERIES BEGIN

ACKNOWLEDGEMENTS

U.S. DEPARTMENT OF ENERGY

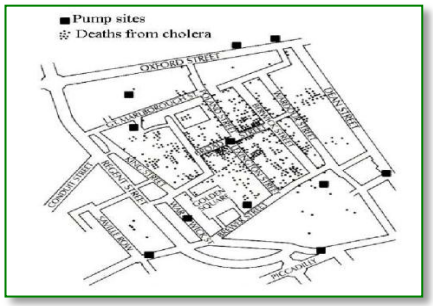
Data-Data Everywhere

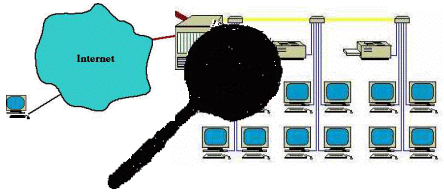




Biomedical Data

- Supermarket scanners
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Web server logs
- Customer web site trails
- Podcasts
- Blogs

- Scientific experiments
- Sensors
- Cameras
- Interactions in social networks
- Newswires
- Speech-to-text translation
- Email
- Closed caption





Information Assurance
Network Intrusion Detection

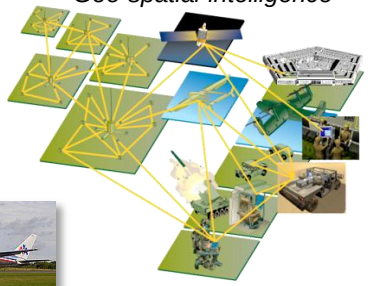
Homeland Security

•Print, film, optical, and magnetic storage: 5 Exabytes (EB) of new information in 2002, doubled in the last three years [How much Information 2003, UC Berkeley]

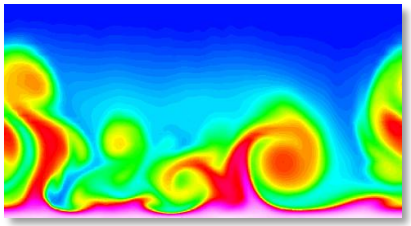
Google

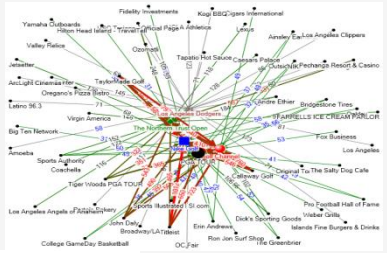
at&t

Geo-spatial intelligence



amazon.com





Business

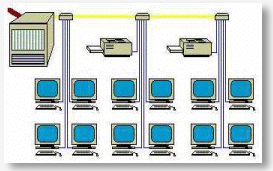
Google

at&t

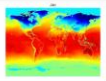
facebook

amazon.com

BIG DATA, High-End Analytics,



Engineering



Knowledge Discovery

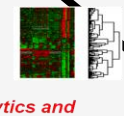
Visualization

Analytics and Mining

Massive datasets

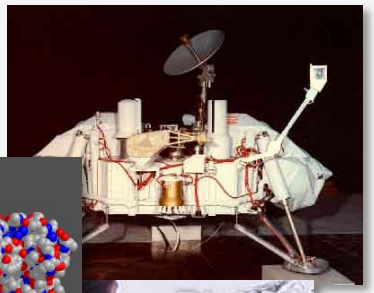
Observations Instruments Experiments

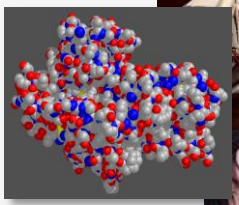
Large-Scale Scientific Simulation



Jaguar - Cray XT4/XT3 - Oak Ridge National Laboratory

Science





"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
  - Massive transactions
  - BIG analytics
    - Top-down query (well-defined operations)
    - Bottom up discovery (unpredictable time-to-result)
  - BIG data processing
  - Predictive computing
- 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

The Data Driven Discovery Ecosystem

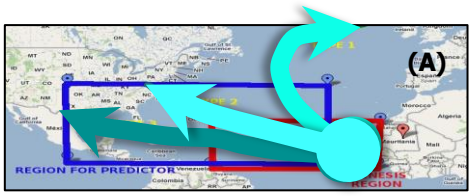
Transactional: Data Generation

Historical : Data Processing/Organization

Relational: Discovery/Predictive Modeling

Act, Refine Feedback

- Data reduction
- Data Management
- Data Query



Historical data → Learning Models → Trigger/questions → Predict

Supercomputers (Current): Illustration of Simulation Dataset Sizes

Application	On-Line Data	Off-Line Data
FLASH: Buoyancy-Driven Turbulent Nuclear Burning	75TB	300TB
Reactor Core Hydrodynamics	2TB	5TB
Computational Nuclear Structure	4TB	40TB
Computational Protein Structure	1TB	2TB
Performance Evaluation and Analysis	1TB	1TB
Kinetics and Thermodynamics of Metal and Complex Hydride Nanoparticles	5TB	100TB
Climate Science	10TB	345TB
Parkinson's Disease	2.5TB	50TB
Plasma Microturbulence	2TB	10TB
Lattice QCD	1TB	44TB
Thermal Striping in Sodium Cooled Reactors	4TB	8TB
Gating Mechanisms of Membrane Proteins	10TB	10TB

Outline

Climate Science

- Data Driven Approach
- Beyond Broad Trends
- Identifying Patterns and Predictions

Scalable Text, Network Analysis

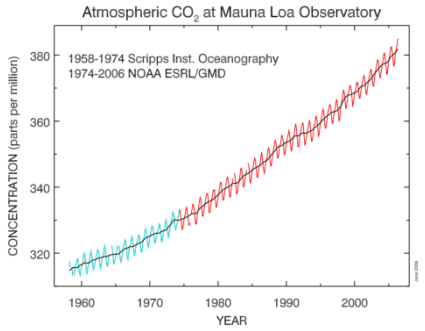
- Sentiment
- Influence
- Networks

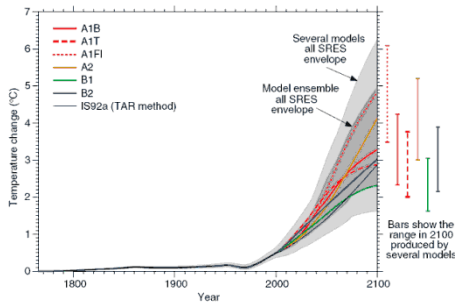
Social, Political and Business Applications

- Measuring the Pulse in real-time, Understanding Egyptian Revolution
- Networks and Action Based Connections

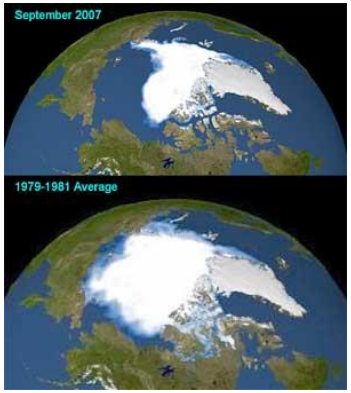
Climate Change: The defining issue of our era

- **The planet is warming**
  - Multiple lines of evidence
  - Credible link to human GHG (green house gas) emissions
- **Consequences can be dire**
  - Extreme weather events, regional climate and ecosystem shifts, abrupt climate change, stress on key resources and critical infrastructures
- **There is an urgency to act**
  - Adaptation: "Manage the unavoidable"
  - Mitigation: "Avoid the unmanageable"
- **The societal cost of both action and inaction is large**





Russia Burns, Moscow Chokes
NATIONAL GEOGRAPHIC, 2010



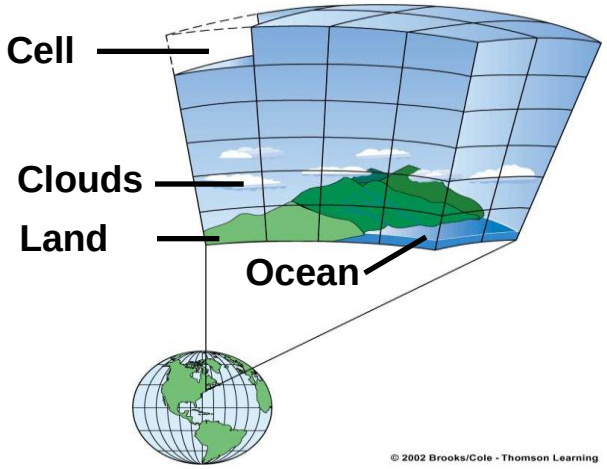
The Vanishing of the Arctic Ice cap
ecology.com, 2008

Key outstanding science challenge:
Actionable predictive insights to credibly inform policy

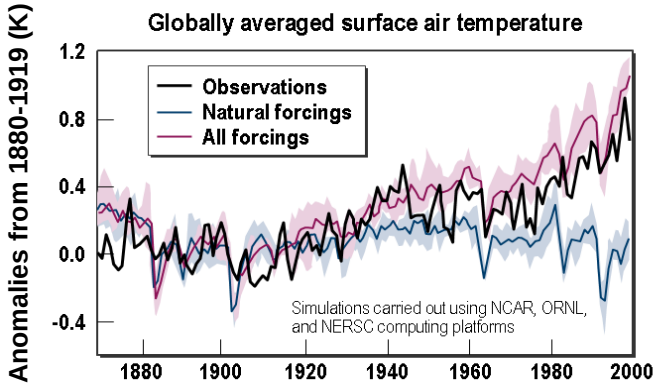
Understanding Climate Change - Physics based Approach

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

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



Temperature increases are human-induced
The anthropogenic climate change "fingerprint"



In the absence of human-induced changes to the atmosphere, the earth would be in a cooling trend

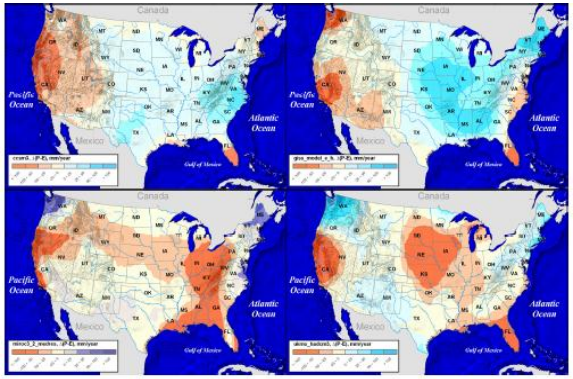
Figure Courtesy: ORNL

Physics-based models are essential but not adequate

- 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)

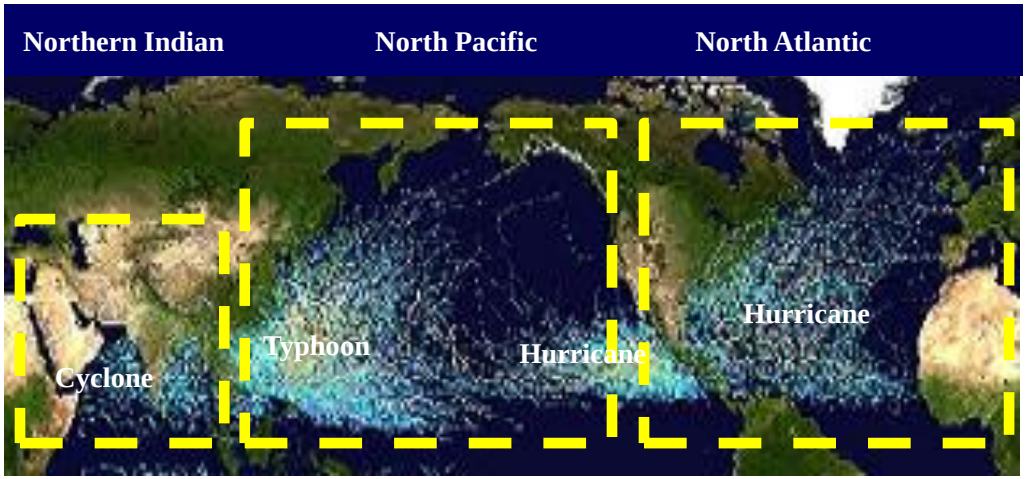
Disagreement between IPCC models



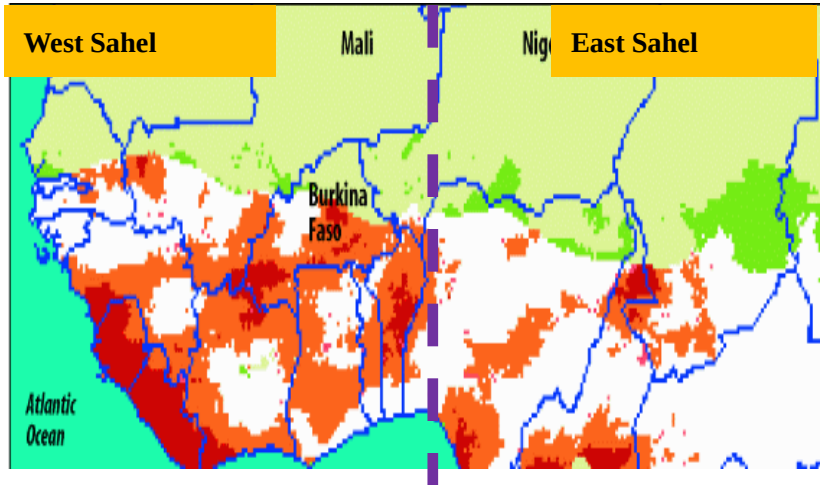
Regional hydrology exhibits large variations among major IPCC model projections

Example Use Cases: Extreme Events Prediction

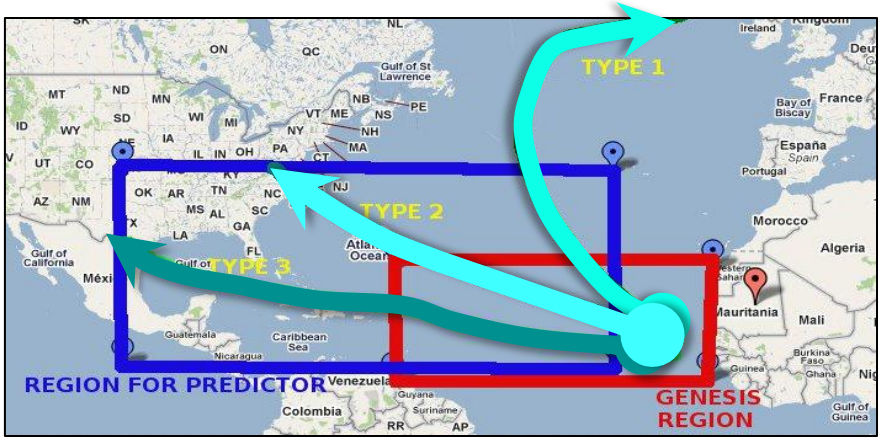
NH Tropical Cyclone (TC) Activity



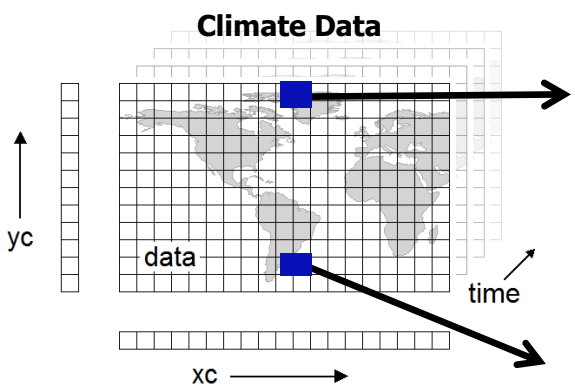
Climate-Meningitis Outlook

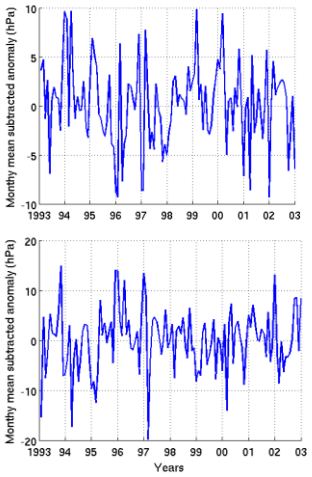


Forecasting NA Hurricane Tracks

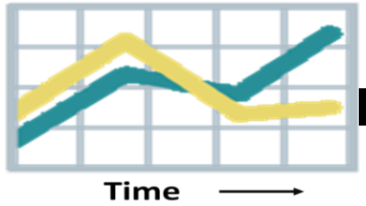


Modeling a Climate System as a Network

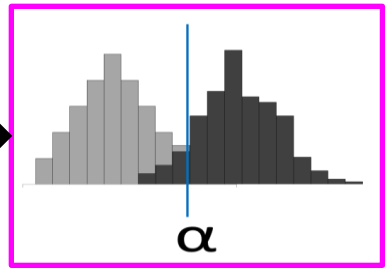




Anomaly time series at each node

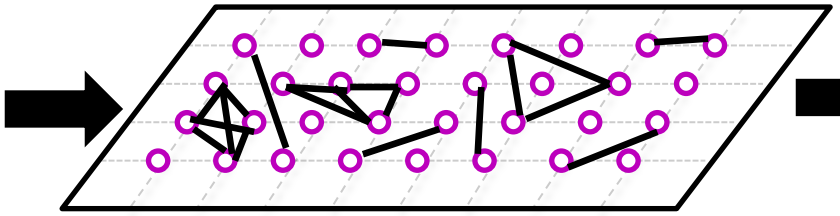


Correlation between two anomaly time series



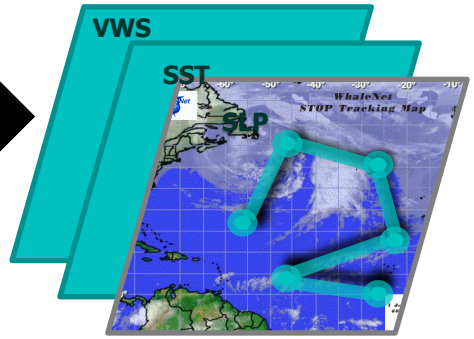
Stat. significant correlations

Climate Network

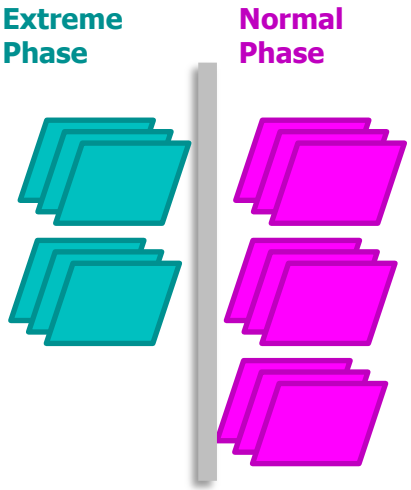


Edge weights: significant correlations

Nodes in the graph: grid points on the globe



Multivariate Networks

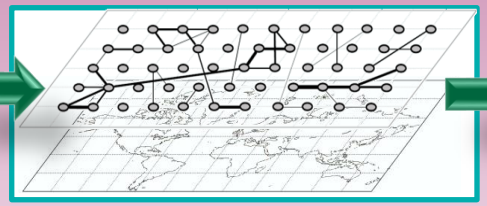


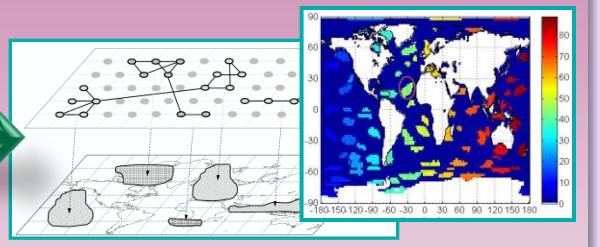
Multiphase Networks

Towards Predictive Insights: Hurricane Frequency

Steps for Discovery of Multivariate Non-linear Interactions

IPCC AR4 Models: CMIP3 datasets
Monthly mean sea surface temperature
Monthly mean atmospheric temperature
Daily horizontal wind at 250/850 hPa



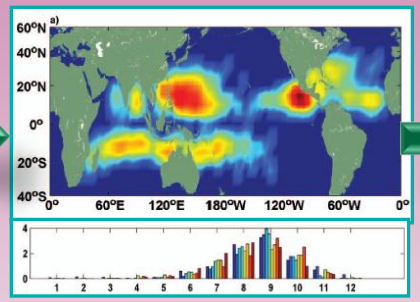


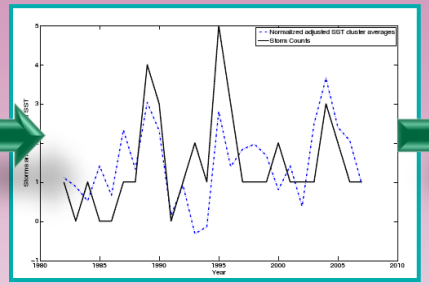
1. Pre-process ancillary climate model outputs

2. Construct multivariate nonlinear climate network

3. Detect & track communities

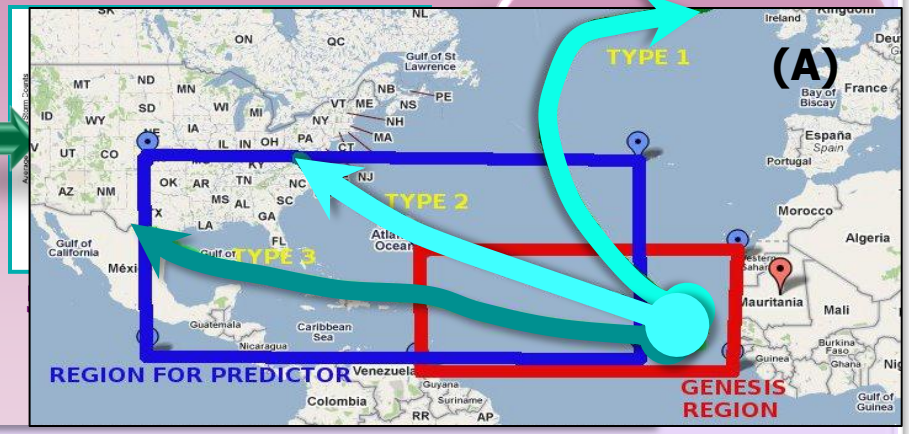
Steps for Predictive Modeling of Hurricanes





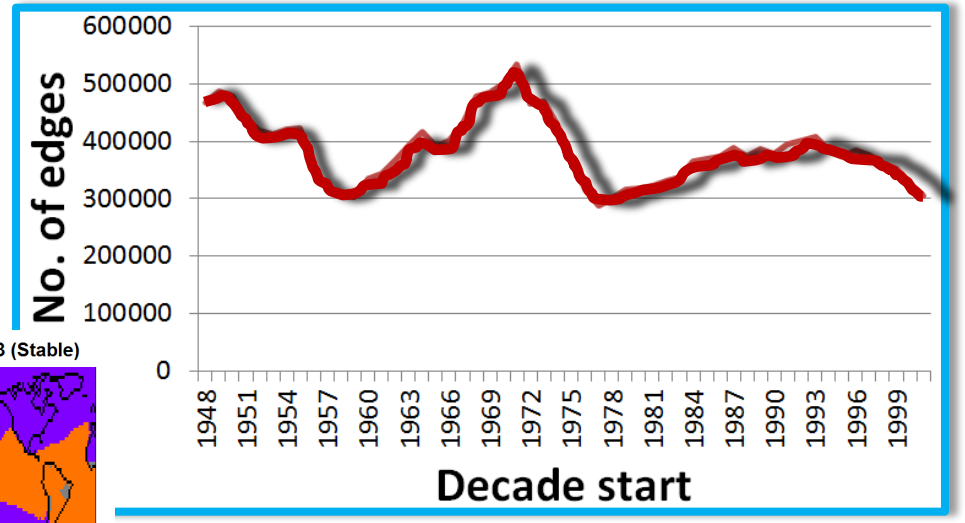
3. Find non-linear relationships

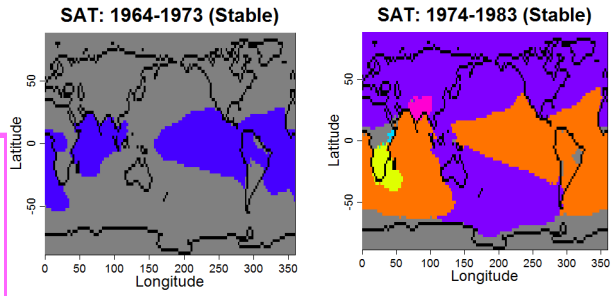
4. Validate w/ hindcasts

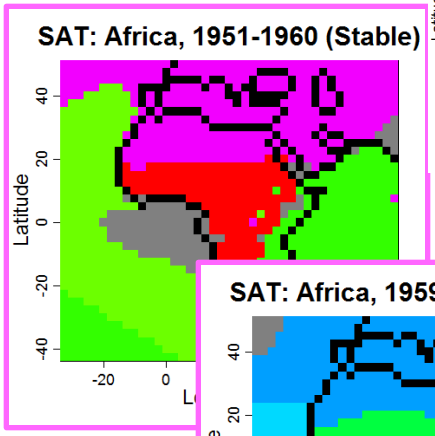


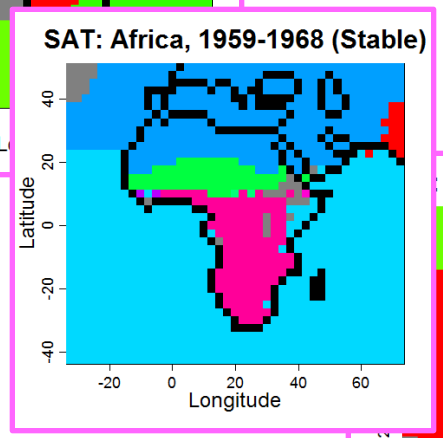
Decadal Trends Discovery via Community Dynamics

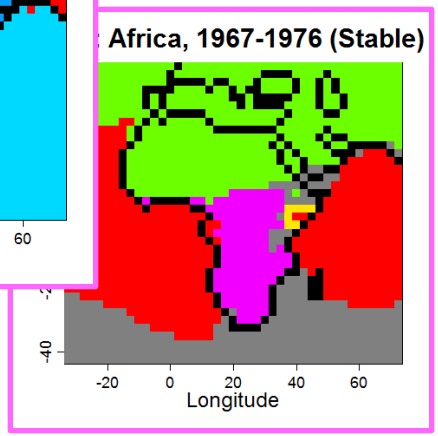
- Characterize the evolution of network communities that are:
  - Recurring/stable or
  - Exhibiting significant shifts
  - With long distance connections







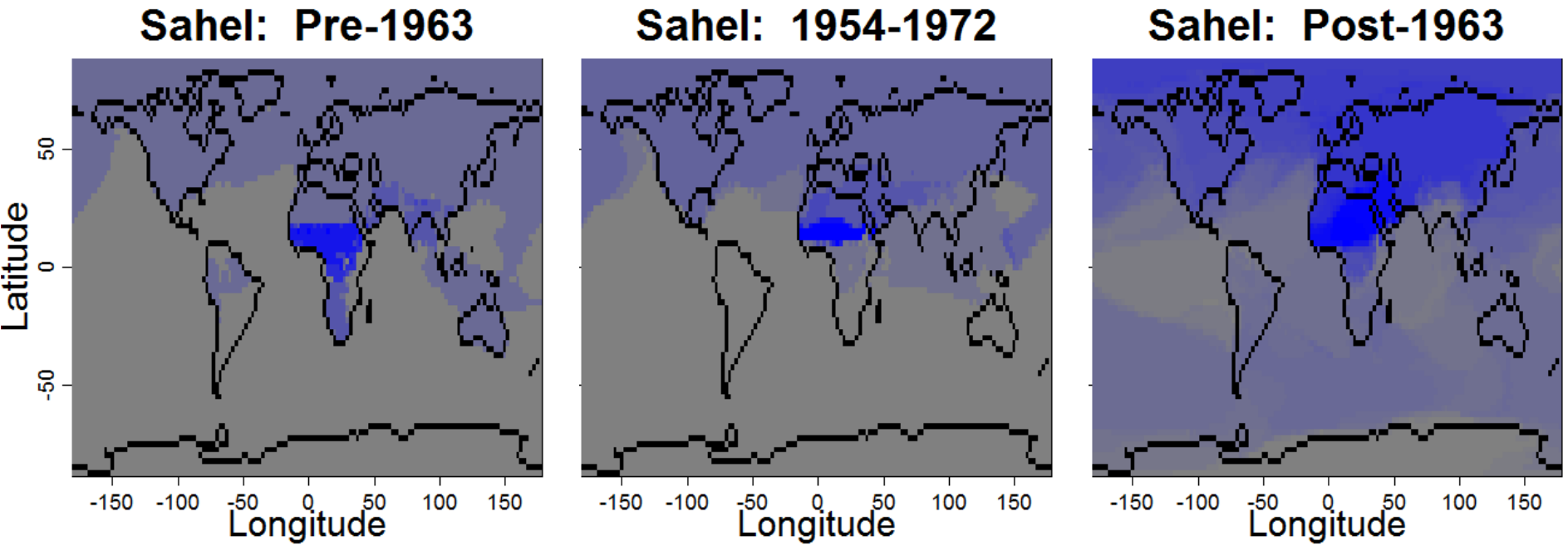




Surface Air Temperature Community Dynamics

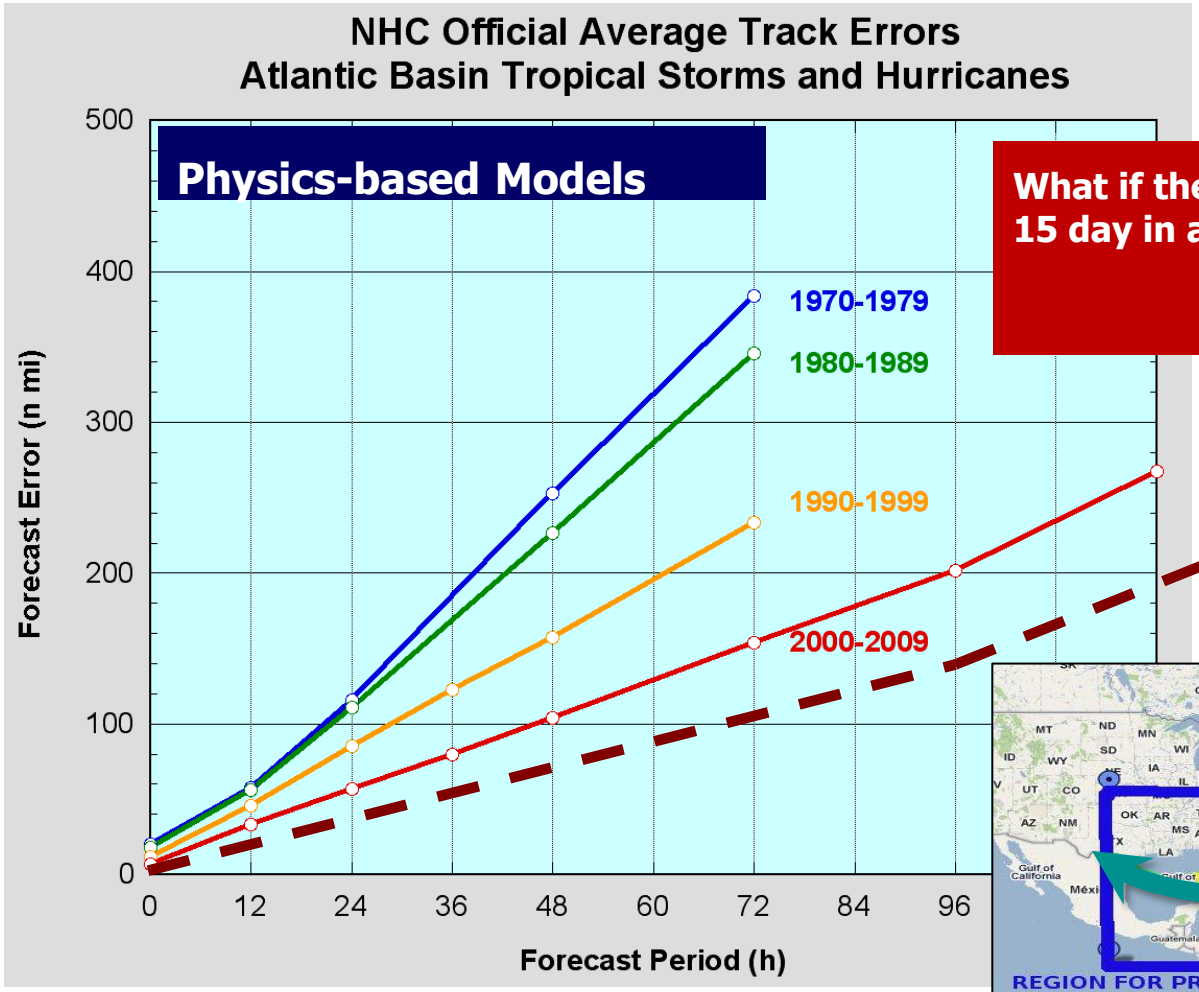
- **Evidence of modulation in planetary-scale climatic pattern**
- **Stable teleconnections between Nino-3 region and Indian Ocean**
- **Realignment of Sahel region to northern Africa**
  - Indirect evidence of desertification

A Closer Look: Desertification in the Sahel



Forecasting Hurricane Tracks

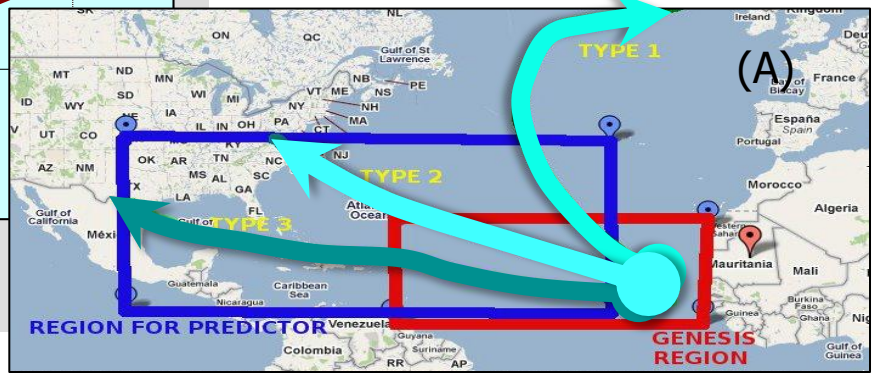
Improving but have mean error (>185km) beyond 48 h



What if the error gets interpolated to 10-15 day in advance forecast?

~500 km

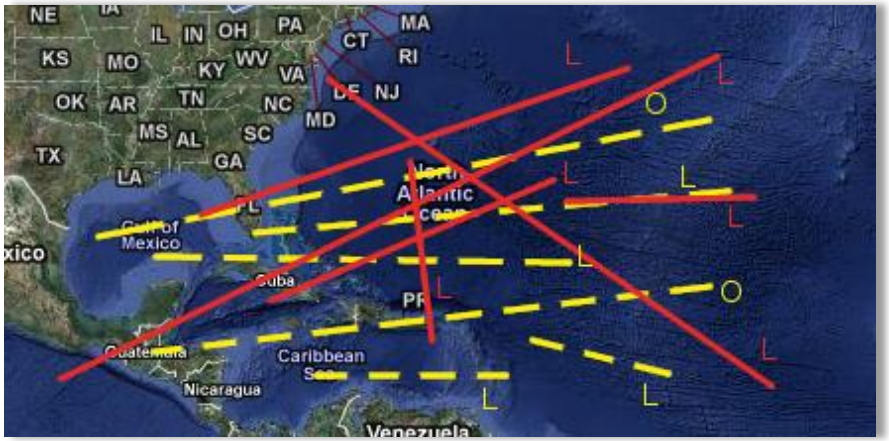
HURDAT Historic Data



Src: NOAA/NHC: http://www.nhc.noaa.gov/verification/index.shtml

Hurricane End-game Track Forecast

Forecast **10-15 days in advance** the **end-game** of a North Atlantic since hurricane embryonic formation in Western Africa.



SLP (yellow/dashed) and SST (red/solid) (+)correlated teleconnections;
L—biased toward land-hitting tracks;
O—biased toward offshore tracks.

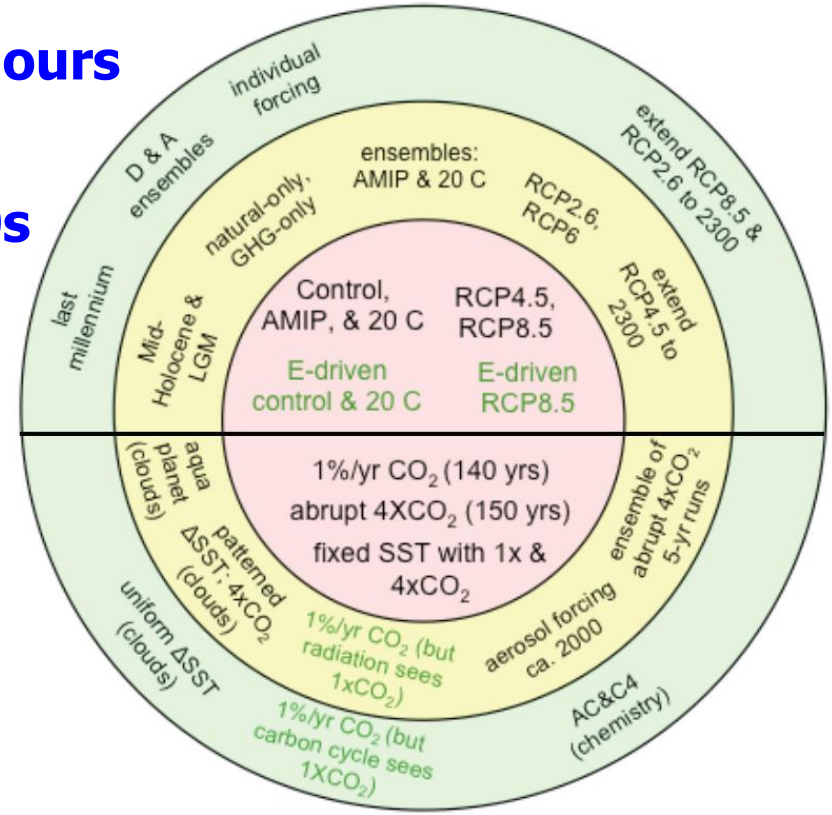
- Nearly **east-oriented SLP** edges suggest horizontal pressure gradient configuration in the same direction.
- Based on Buys Ballot's law, this pressure gradient would be associated with **wind flow in the north-south direction**.
- Onshore wind anomaly flow would promote favorable conditions for landfall; opposite flow anomaly would be more favorable for hurricanes tracks in no-landfall.

Performance of Land-hitting vs. Offshore

	LOO			10-FOLD	
	SLP	**SST**	**SLP+SST**	**SLP**	**SST**
Accuracy	0.88	0.90	0.92	0.90	0.90
Sensitivity	0.91	0.96	0.97	0.95	0.97
Specificity	0.77	0.76	0.81	0.80	0.74
Precision	0.90	0.90	0.92	0.92	0.90
F1-meas.	0.90	0.93	0.94	0.93	0.93

CMIP3 → CMIP5 => BIG DATA and Computing

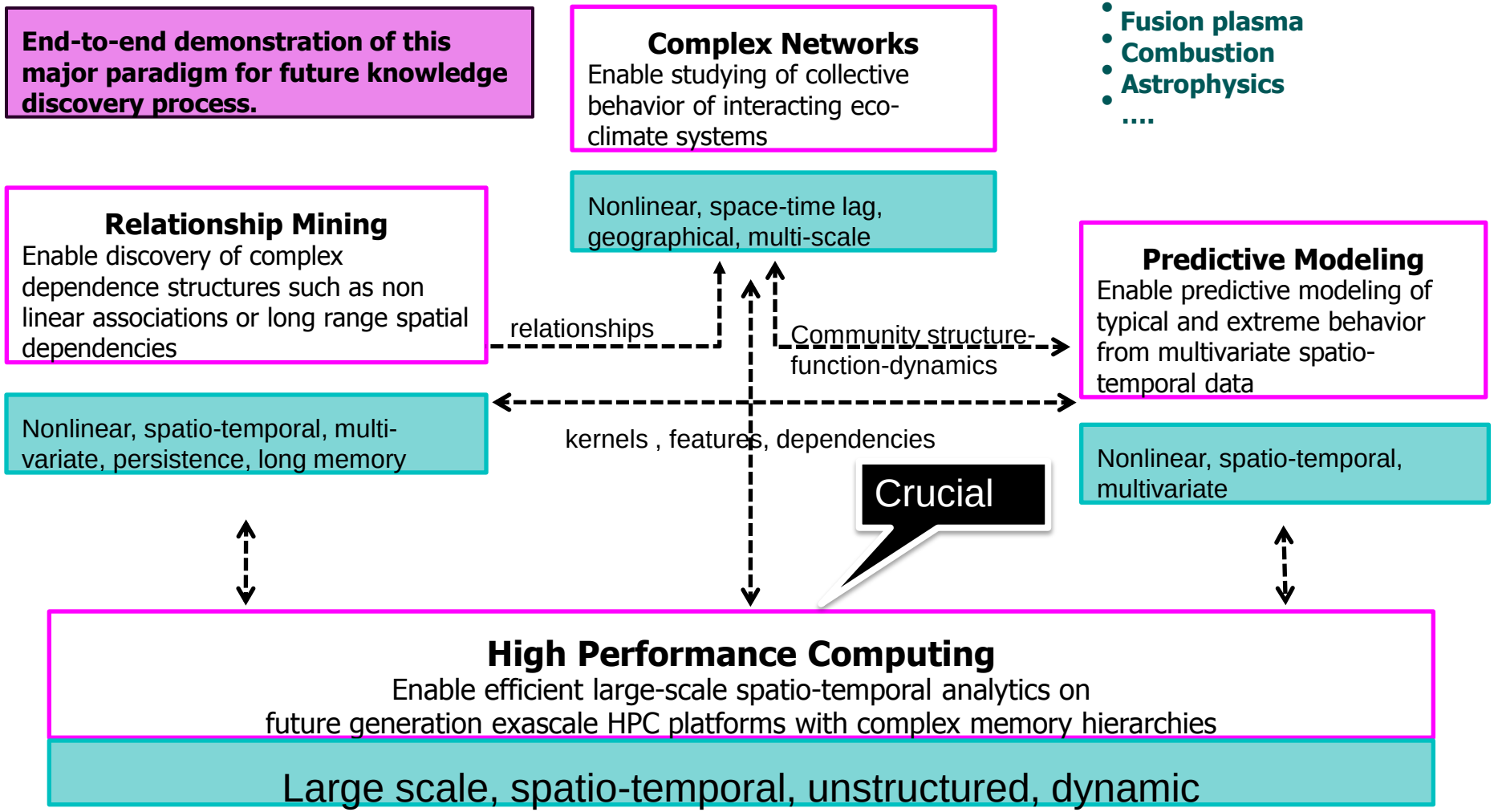
- **Coupled Model Inter comparison Project**
- **Spatial resolution: 1 – 0.25 degrees**
- **Temporal resolution: 6 hours – 3 hours**
- **Models: 24 - 37**
- **Simulation experiments: 10s - 100s**
  - Control runs & hindcast
  - Decadal & centennial-scale forecasts
- **Covers 1000s of simulation years**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



Summary of CMIP5 model experiments, grouped into three tiers

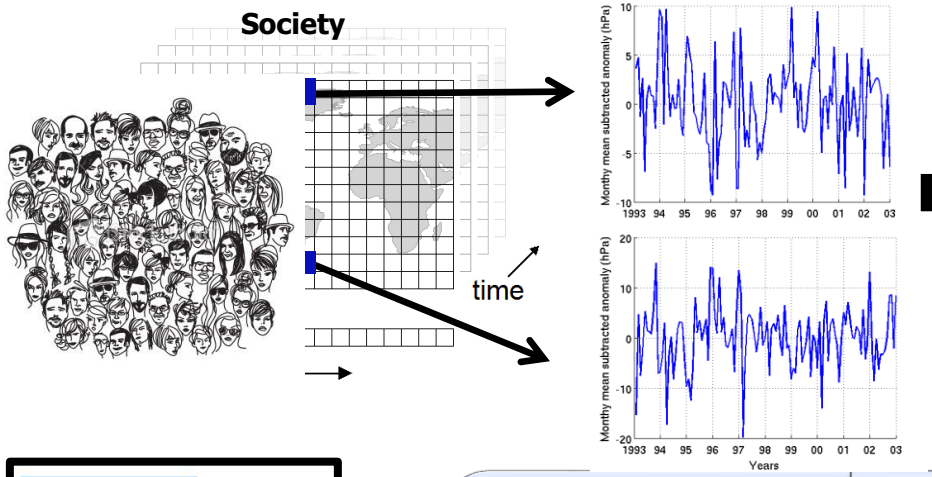
Transformative Computer Science Research

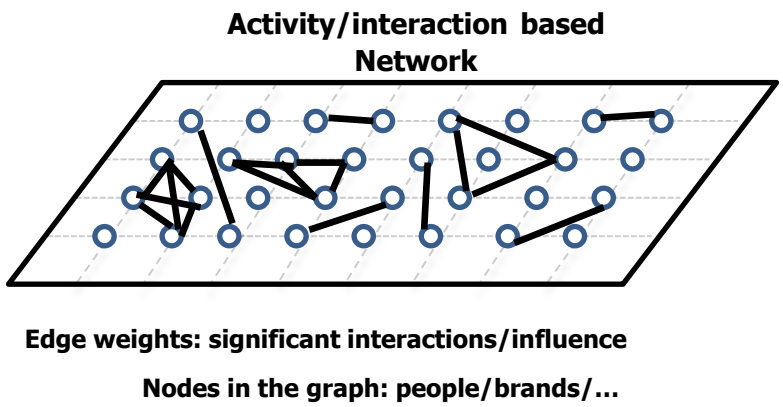
Enabling large-scale data-driven science for complex, multivariate, spatio-temporal, non-linear, and dynamic systems:

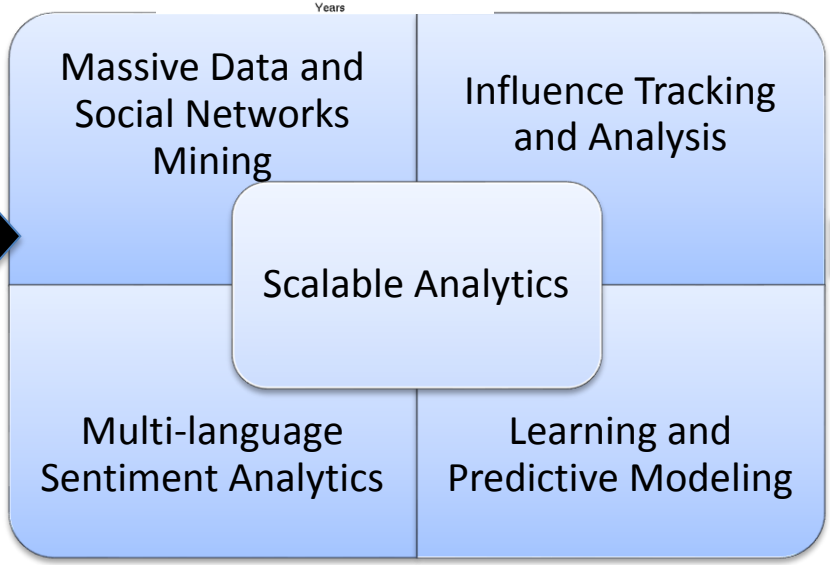


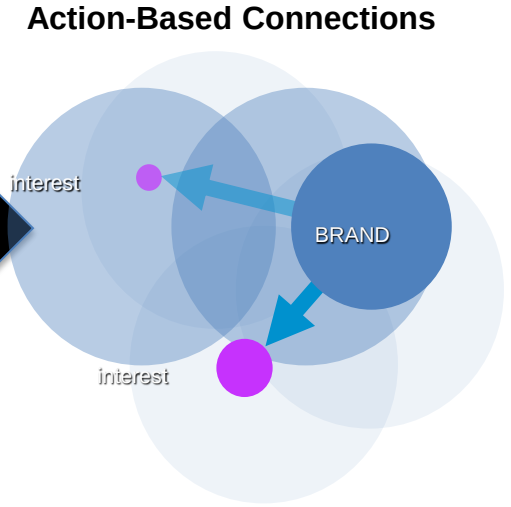
From Science to Social

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





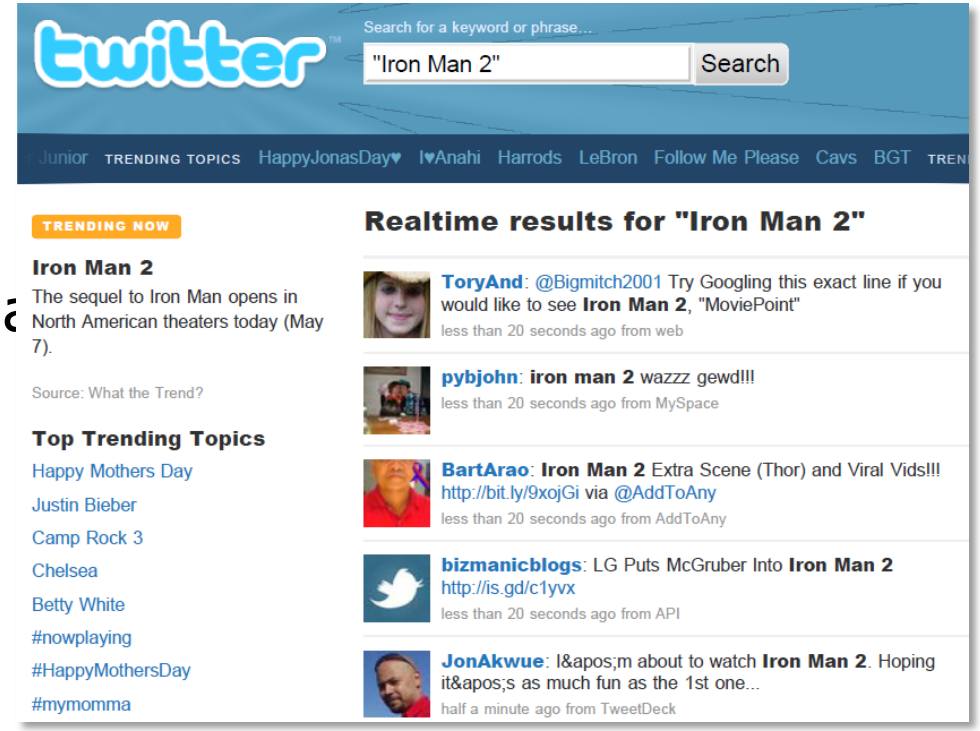




http://www.pulseofthetweeters.com

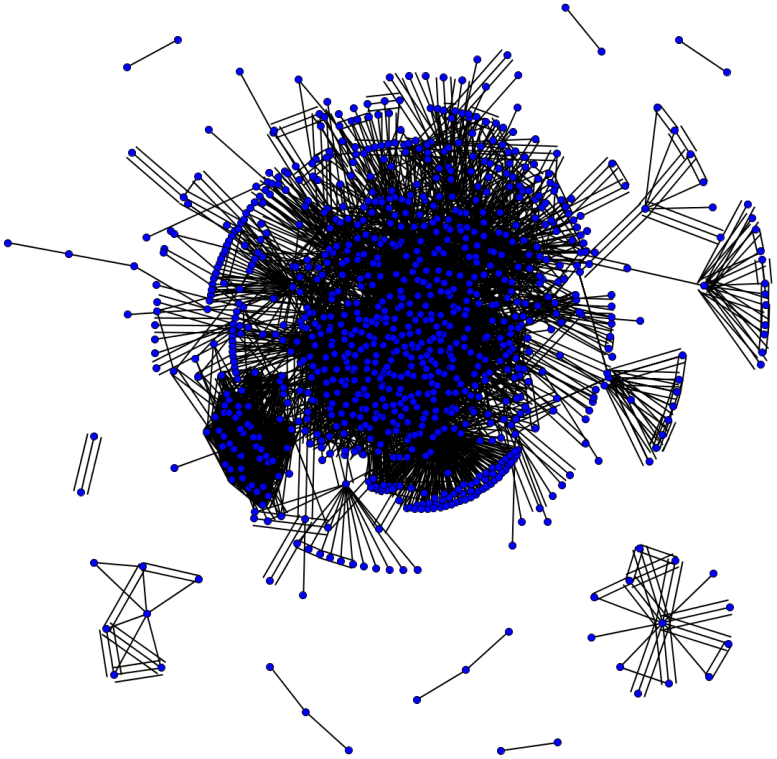
About Twitter

- ✓ Twitter: a micro blogging social network
- ✓ Millions of users
- ✓ Short messages of up to 140 characters called `tweets'
- ✓ ~100 million tweets per da
- ✓ News, Politics, Sports, Entertainment …..
- ✓ Trending topics on Twitter

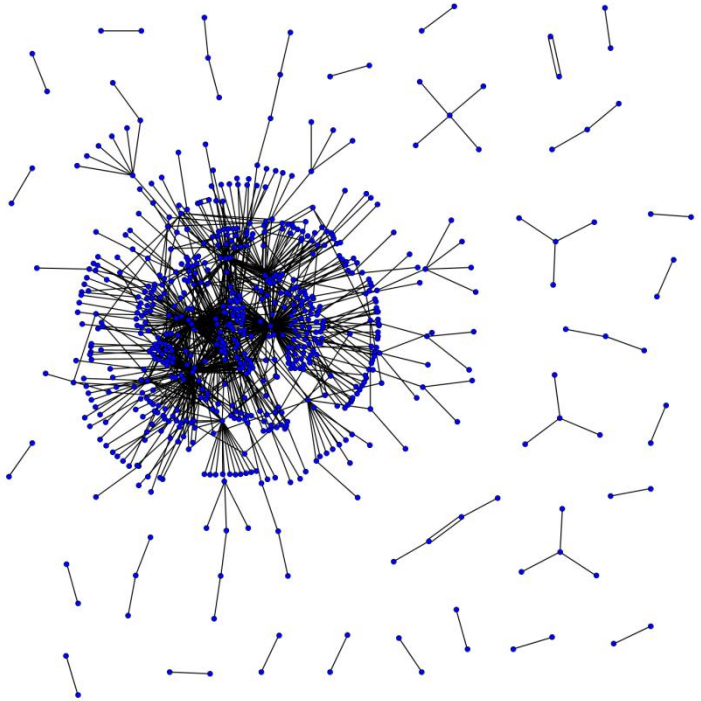


Social Network Pulse and (Soft) Real-time Sentiment Analysis

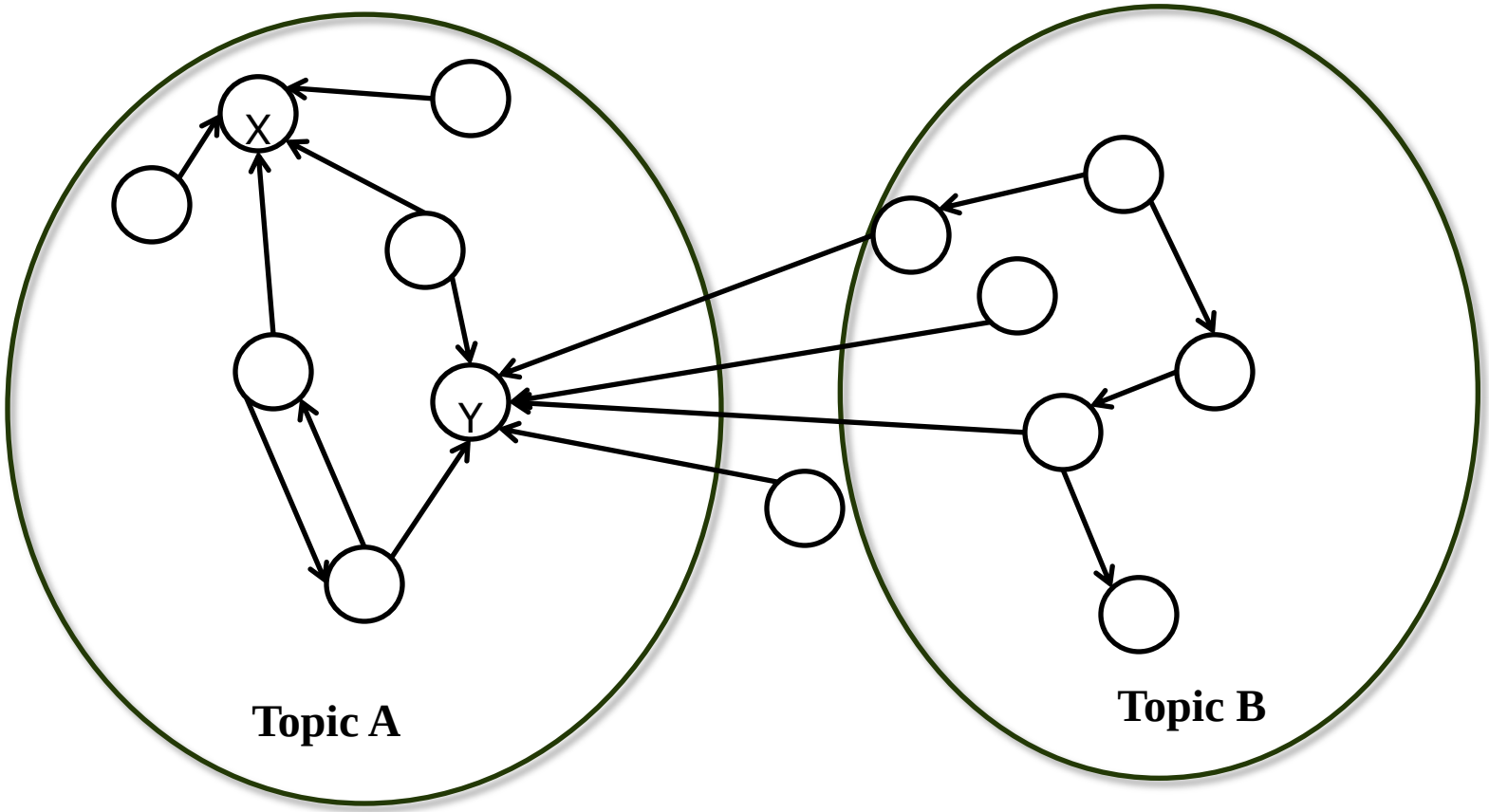
Static (Slowly Changing) Social Network(1000 users)



Dynamic Topic based response network



Identifying Influencers



$$R_x = (1-d) + d * \sum_{i=1}^{n} R_i \frac{W_e(U_i, U_x)}{\sum_{j=1}^{m} W_e(U_i, U_j)}$$

Social Media Evolution of the Egyptian Revolution

BY ALOK CHOUDHARY, WILLIAM HENDRIX, KATHY LEE, DIANA PALSETIA, AND WEI-KENG LIAO

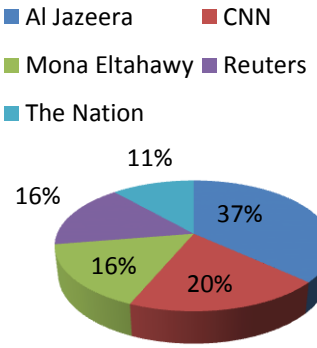
74 COMMUNICATIONS OF THE ACM | MAY 2013 | VOL. 56 | NO. 5

- The 2011 Egyptian revolution resulted in the removal of longtime leader Hosni Mubarak
- By most accounts, social media played an integral role in organizing and building support for the revolution

Protests in Cairo's Tahrir Square

Ousted President Hosni Mubarak

The Twitter Revolution



Over **800,000 tweets** in six trending topics on **Egyptian revolution**

Topics: *egypt*, *cairo*, *tahrir*, *egyptians*, *hosni_mubarak*, and *omar_suleiman*

Leaders

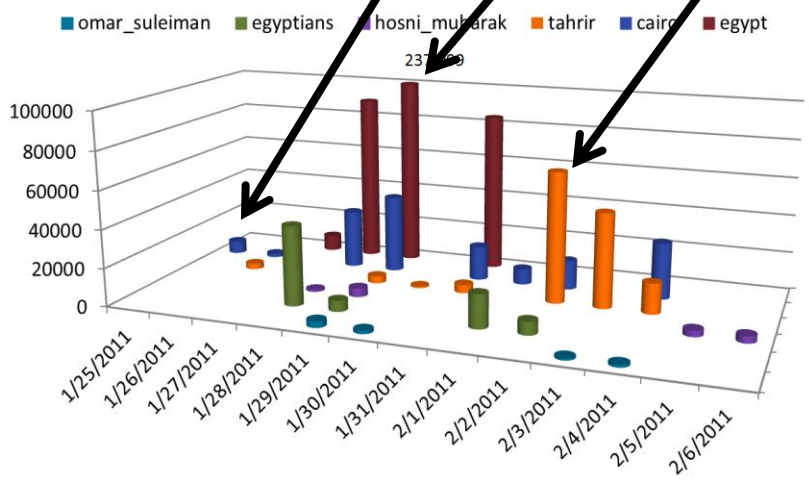
Influencer	Description	Twitter user names
Al Jazeera	Arabic news channel based in Doha, Qatar; name translates to "the island"	AJEnglish, Dima_AlJazeera, AJGaza, AlanFisher, FatimaNaib, mohamed, SherineT
CNN	American news network founded by Ted Turner	bencnn, cnnbrk, CNNLive, natlsecuritycnn, vhernandezcnn
Mona Eltahawy	Egyptian journalist and public speaker	monaeltahawy
Reuters	News agency headquartered in London	Reuters, JimPathokoukis
The Nation	Left-leaning weekly magazine headquartered in New York City	jeremyscahill, thenation

- Measured **influence** based on how often followers started discussing topics the influencer tweeted about
- Most influential organizations: **Al Jazeera** and **CNN**
- Most influential individual: independent journalist **Mona Eltahawy**

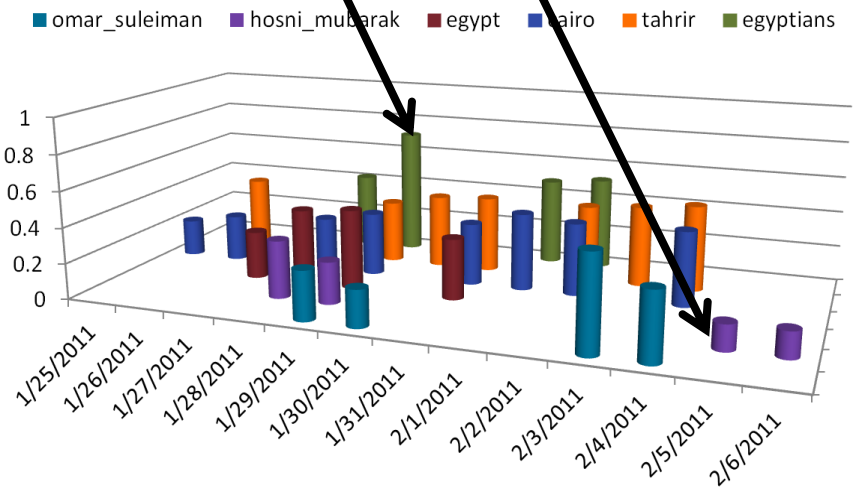
The Evolution of the Revolution

Date	Major events	Twitter activity
Jan. 25	"Day of Rage" protests in Cairo signal the start of major changes in Egypt.	Topic *cairo* is only trending topic related to Egypt Negative tweets outnumber positive tweets four to one
Jan. 27	The Egyptian government begins limiting internet access in Egypt.	Little activity overall Topic *egypt* begins trending Topic *egyptians* (the most positive topic) begins trending Jan. 28
Jan. 29	President Hosni Mubarak dismisses his cabinet and appoints Omar Suleiman as Vice President of Egypt.	All Egypt-related topics trend Largest single-day volume of tweets on revolution Topic *egyptians* peaks in sentiment, with almost 3 positive tweets for every negative tweet
Feb. 2	Blockage of internet access by Egyptian government ends.	Huge increase in tweets on *tahrir* (about 1500% as many tweets) Topic *omar_suleiman* trends on Feb. 3 with more positive tweets than negative
Feb. 6	Egypt-related topics stop trending on Twitter. Mubarak resigns Feb. 11.	Sentiment on topic *hosni_mubarak* decreased progressively throughout the revolution, with about six times as many negative tweets as positive by Feb. 6

Tweet volume



Sentiment



Activity Signature: Not Just Another #superbowl

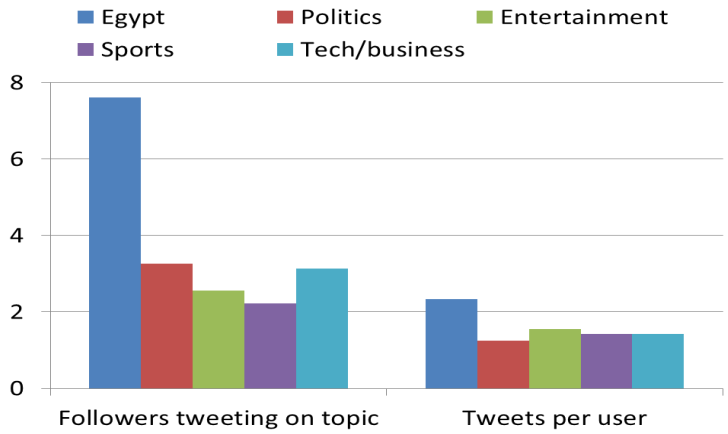
Compared with **twenty** other trending topics from the same time period

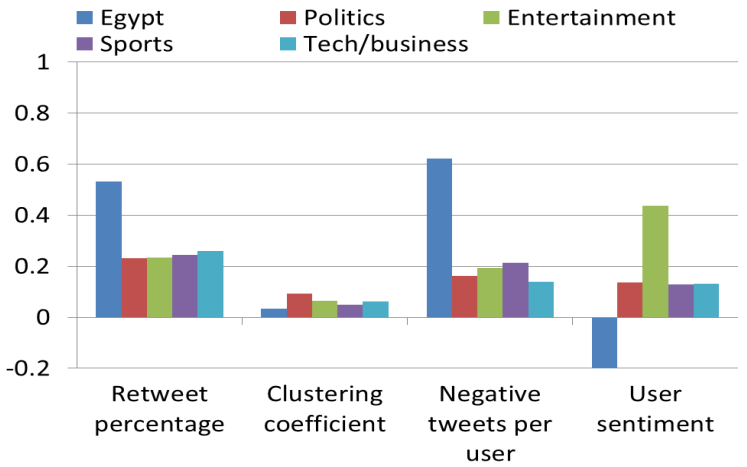
Four categories of topics
- Politics
- Entertainment
- Sports
- Tech/business

Comparison based on **tweets**, **sentiment**, and **network structure** of Twitter followers

Significant differences

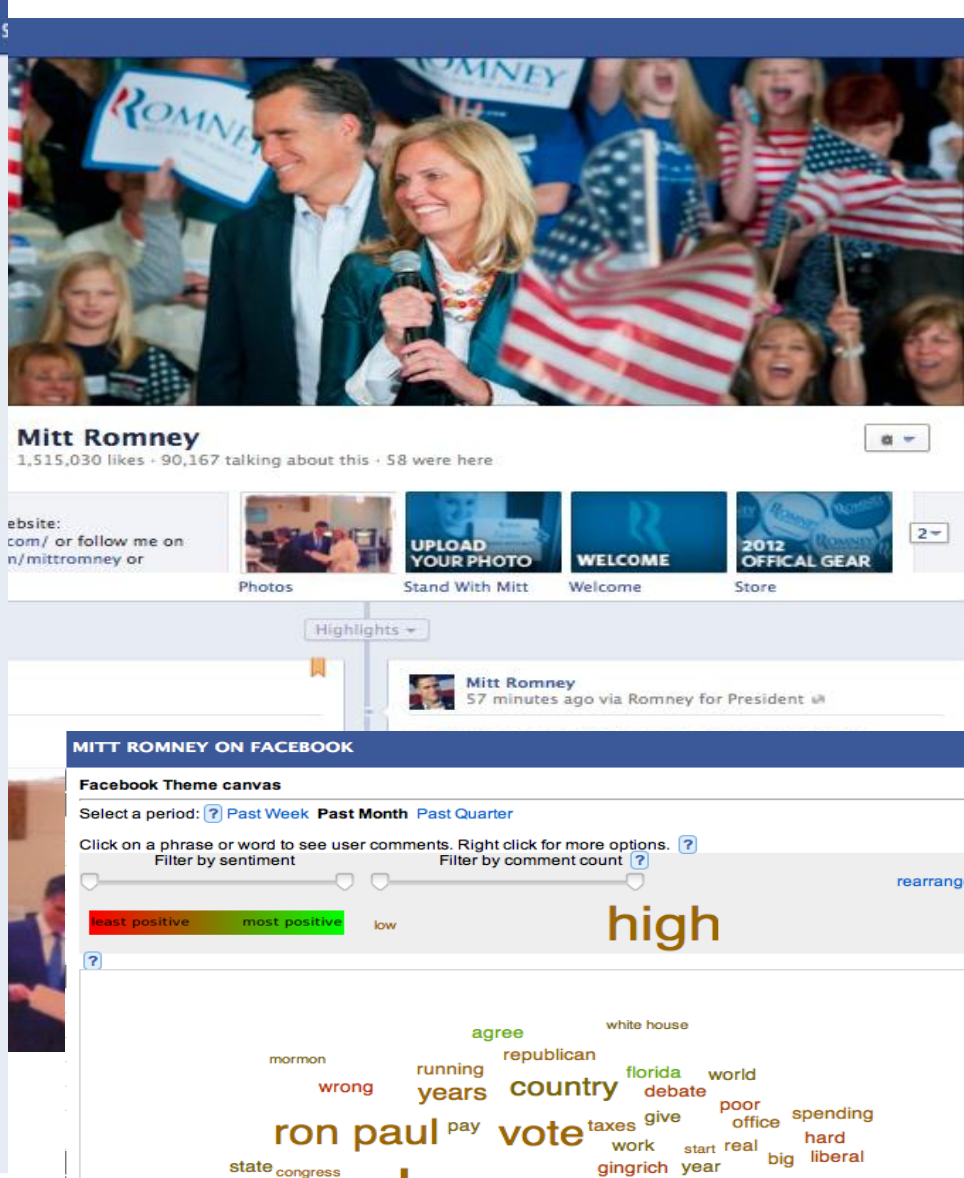
- Large percentage of **retweets**, **more tweets** overall
- Tweets significantly **more negative**
- Followers more likely to be **engaged**
- Follower network **less tightly knit**





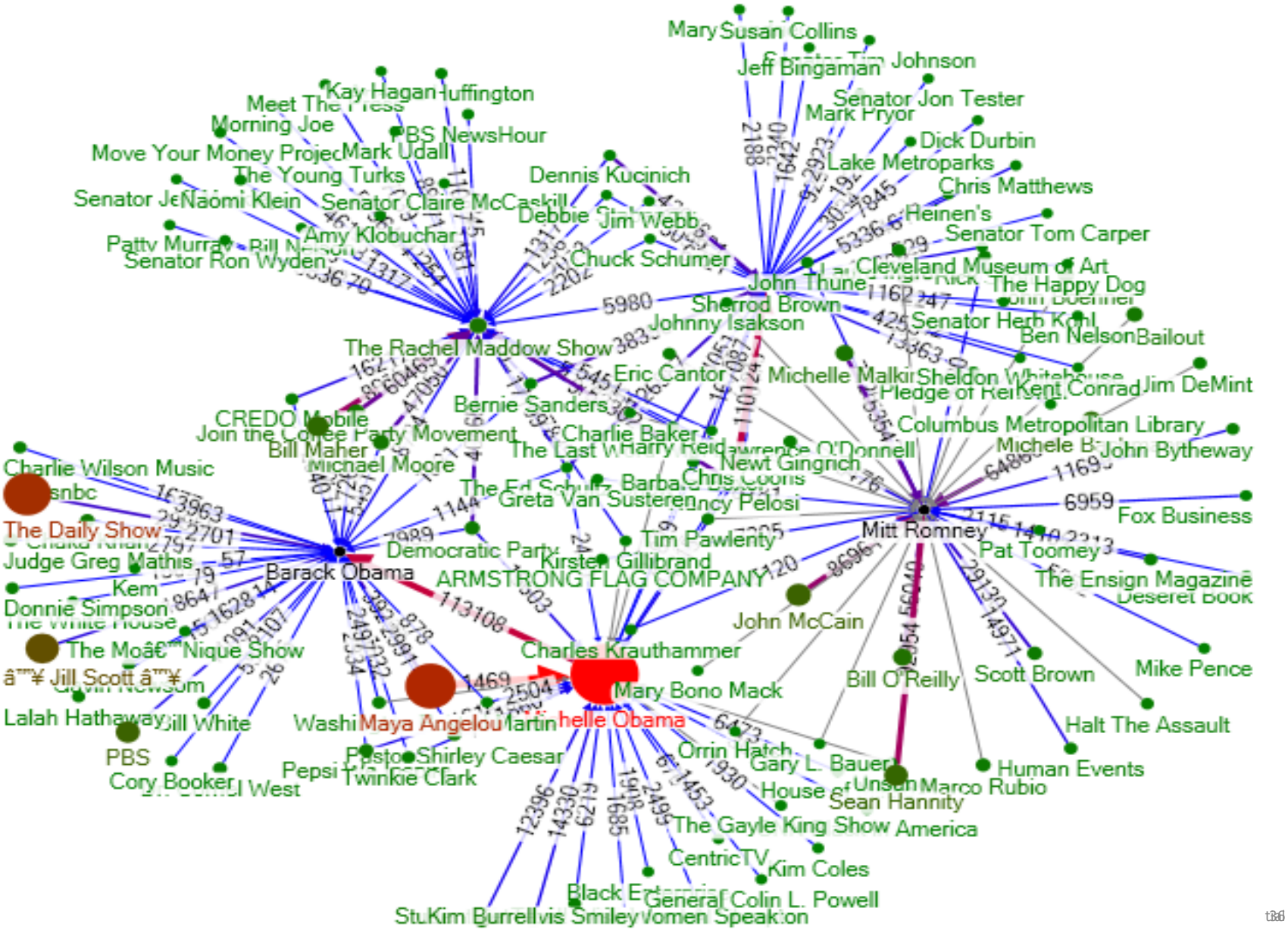
Interest based Community Extraction in Social Networks: Facebook and Twitter Data with 150M+ users


facebook
Barack Obama
25,469,380 likes · 339,731 talking about this
Politician
This page is run by Obama for America, President Obama's 2012 campaign. To visit the White House Facebook page, go to
About
Photos
Donate
Videos
Store
DONATE
ONLINE STORE
Highlights
Barack Obama
February 26
This photo has been making the rounds. What's your #1 reason for supporting President Obama?
Why I Support Obama—
1. For 30 years I've heard politicians talking about health
Likes
See All
Veterans for Obama
Politician
Obama for America - Ohio
Political Organization
Generation Forty Four (Gen44)
Political Organization
Democratic Party

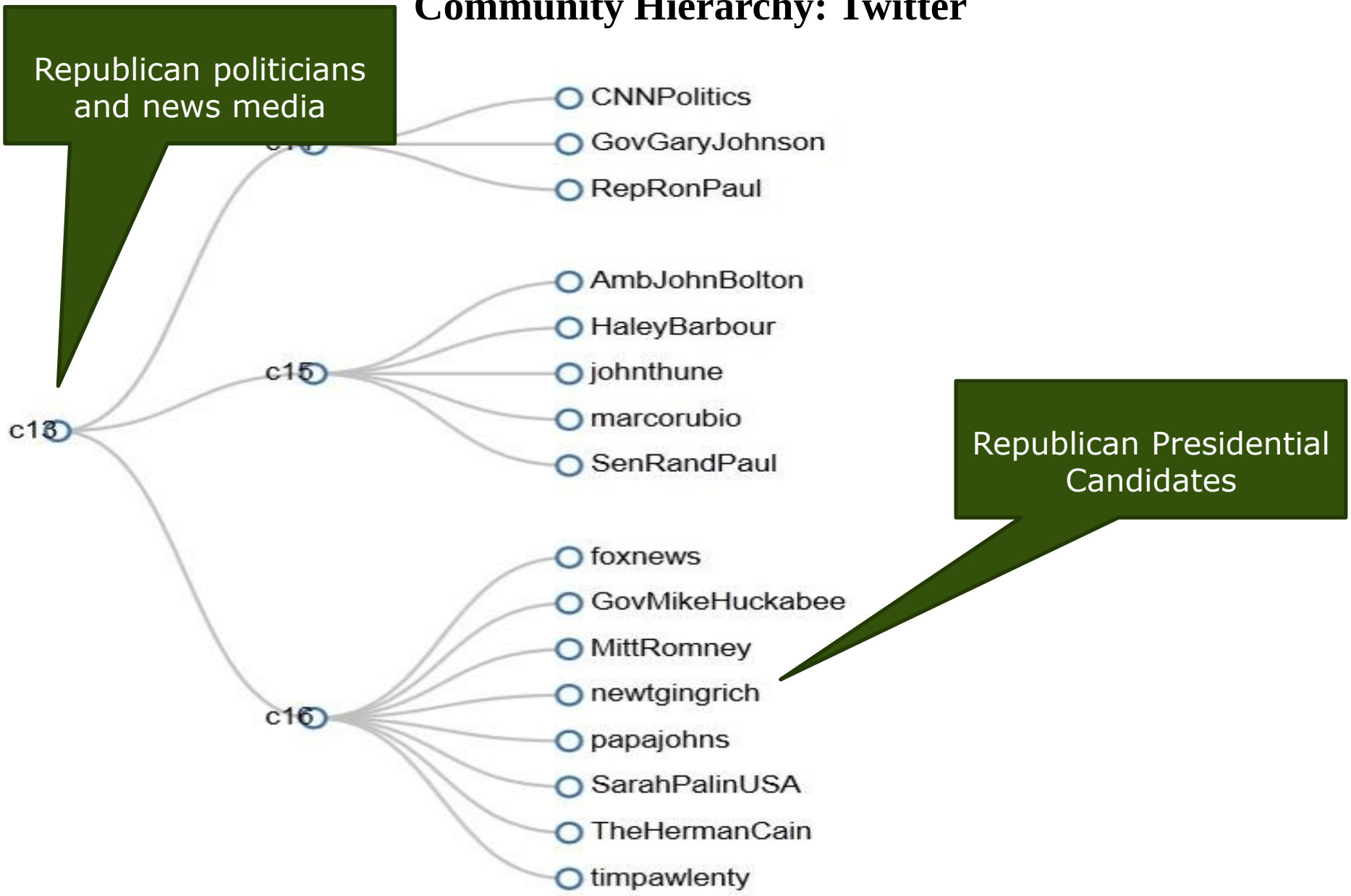

Mitt Romney
1,515,030 likes · 90,167 talking about this · 58 were here
Photos
Stand With Mitt
Welcome
Store
UPLOAD YOUR PHOTO
WELCOME
2012 OFFICAL GEAR
Highlights
Mitt Romney
57 minutes ago via Romney for President
MITT ROMNEY ON FACEBOOK
Facebook Theme canvas
Select a period: Past Week Past Month Past Quarter
Click on a phrase or word to see user comments. Right click for more options.
Filter by sentiment
Filter by comment count
least positive
most positive
low
high
agree
white house
mormon
republican
running
florida
world
wrong
years
country
debate
poor
ron paul
pay
vote
taxes
give
office
spending
work
start
real
hard
state
congress
gingrich
year
big
liberal
party
obama
guy
stop
talk
media
money
gop
win
change
rich
job
supporters
americans
tax
care
newt
page
government
jobs
support
economy
american
hope
true
conservative
time
paul
life
candidates
run
back
business
santorum
constitution
stand
voting
person
candidate
election
campaign
governor
show
plan
republicans

7/27/2012

Network Effect and Precise Interest Targeting

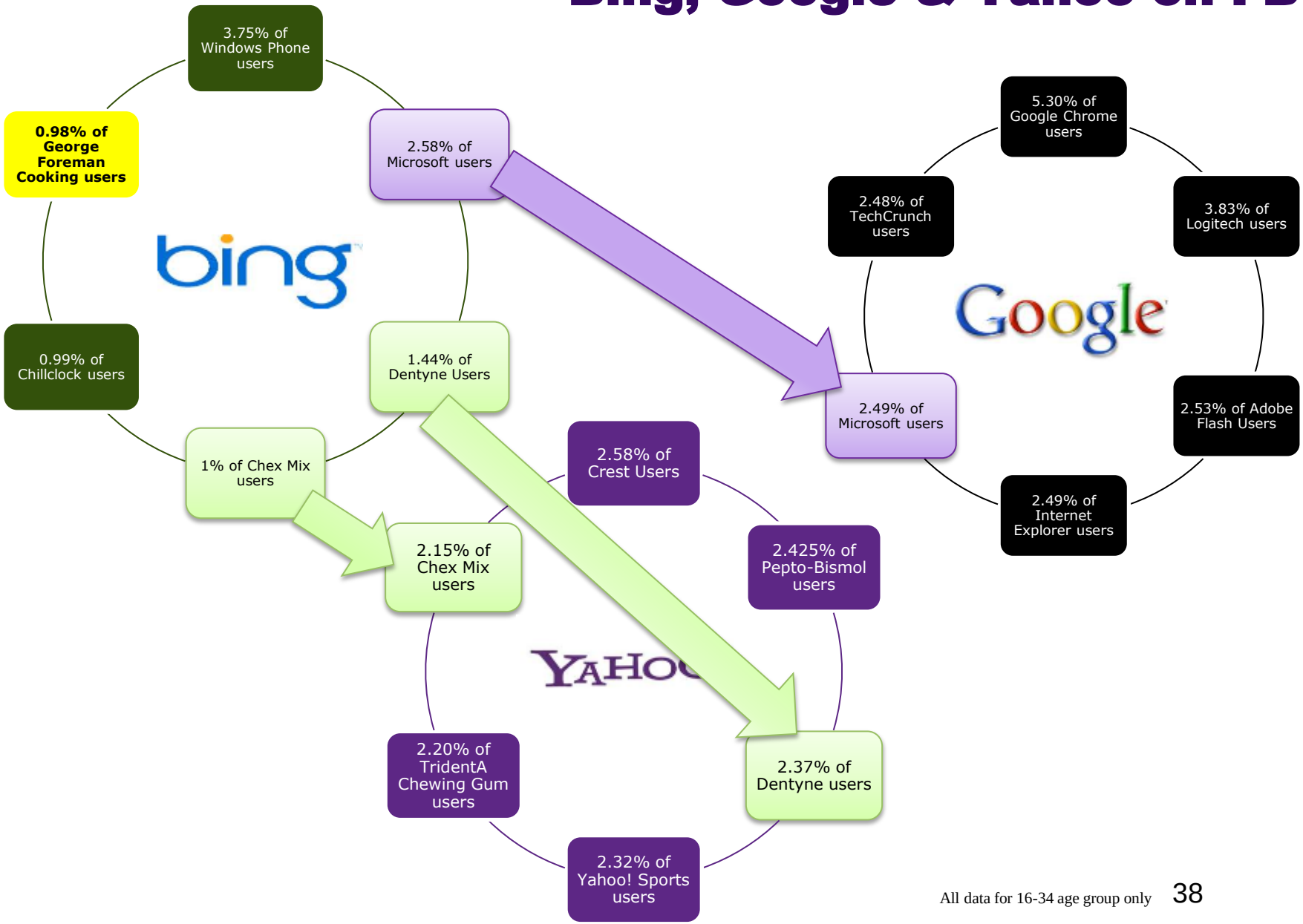


Community Hierarchy: Twitter



E.g. U.S. Politics

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

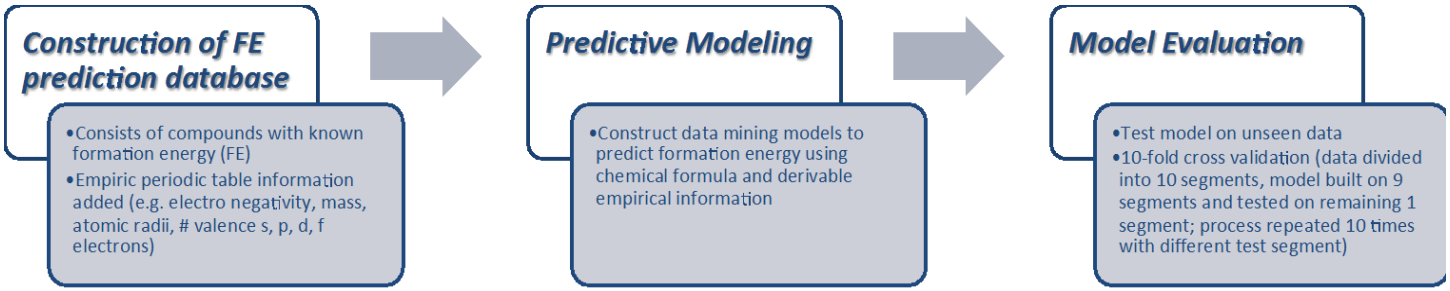


All data for 16-34 age group only

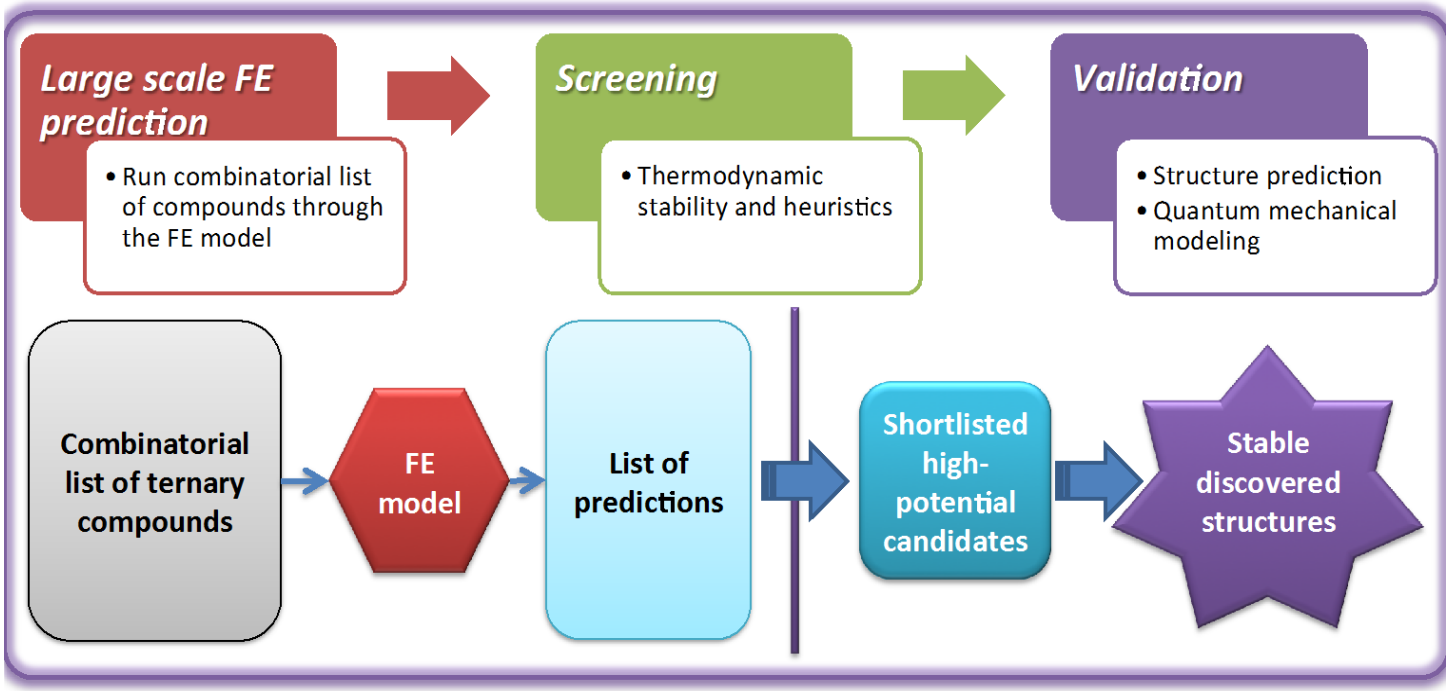
A different way of thinking?

A "DATA DRIVEN DISCOVERY" WORTH A THOUSAND SIMULATIONS?

Discovering Materials : Simulations → Analytics

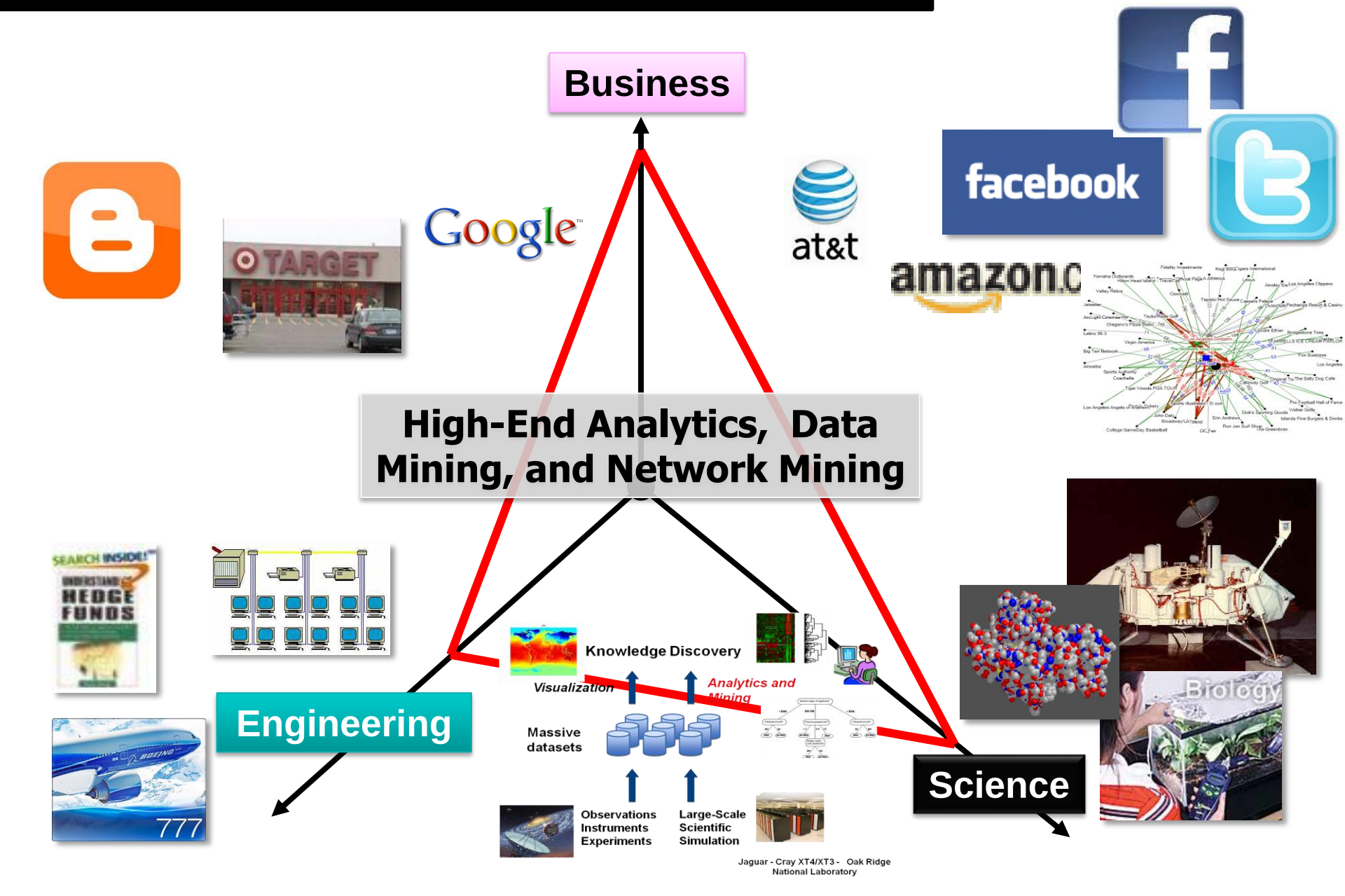


(a)



(b)

Summary: Discovering Knowledge from Massive Data



Thank You.

Alok Choudhary, John G. Searle Professor
Dept. of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Director of the Center for Ultra-Scale Computing and Security
Northwestern University
choudhar@eecs.northwestern.edu