

# Computer science theory to support research in the information age

John Hopcroft  
Cornell University  
Ithaca, New York



# Time of change

- The information age is a revolution that is changing all aspects of our lives.
- Those individuals, institutions, and nations who recognize this change and position themselves for the future will benefit enormously.



# Computer Science is changing

Early years

- Programming languages
- Compilers
- Operating systems
- Algorithms
- Data bases

Emphasis on making computers useful



# Computer Science is changing

## The future years

- Tracking the flow of ideas in scientific literature
- Tracking evolution of communities in social networks
- Extracting information from unstructured data sources
- Processing massive data sets and streams
- Extracting signals from noise
- Dealing with high dimensional data and dimension reduction



# Drivers of change

- Merging of computing and communication
- Data available in digital form
- Networked devices and sensors



# Implications for TCS

- Need to develop theory to support the new directions
- Update computer science education

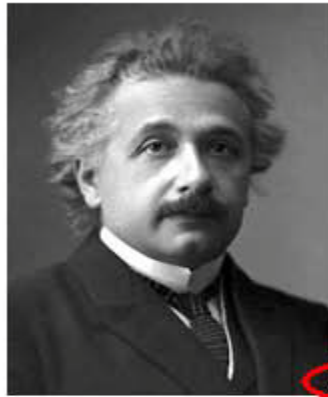


# Internet search engines are changing

## ■ When was Einstein born?

Einstein was born at Ulm, in Wurttemberg, Germany, on March 14, 1879.

List of relevant web pages



When was Albert Einstein  
born?

**Albert Einstein** was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where he later on began his schooling at the Luitpold Gymnasium. Later, they moved to Italy and Albert continued his education at Aarau, Switzerland and in 1896 he entered the Swiss Federal Polytechnic School in Zurich to be trained as a teacher in physics and mathematics. In 1901, the year he gained his diploma, he acquired Swiss citizenship and, as he was unable to find a teaching post, he accepted a position as technical assistant in the Swiss Patent Office. In 1905 he obtained his doctor's degree.

During his stay at the Patent Office, and in his spare time, he produced much of his remarkable work and in 1908 he was appointed Privatdozent in Berne. In 1909 he became Professor Extraordinary at Zurich, in 1911 Professor of





# Internet queries will be different

- Which car should I buy?
- What are the key papers in Theoretical Computer Science?
- Construct an annotated bibliography on graph theory.
- Where should I go to college?
- How did the field of computer science develop?



# Which car should I buy?

- Search engine response: Which criteria below are important to you?
  - Fuel economy
  - Crash safety
  - Reliability
  - Performance
  - Etc.



Make	Cost	Reliability	Fuel economy	Crash safety	Links to photos/articles
Toyota Prius	23,780	Excellent	44 mpg	Fair	photo article
Honda Accord	28,695	Better	26 mpg	Excellent	photo article
Toyota Camry	29,839	Average	24 mpg	Good	photo article
Lexus 350	38,615	Excellent	23 mpg	Good	photo article
Infiniti M35	47,650	Excellent	19 mpg	Good	photo article



# 2010 Toyota Camry - Auto Shows

## Toyota sneaks the new Camry into the Detroit Auto Show.

Usually, redesigns and facelifts of cars as significant as the hot-selling [Toyota Camry](#) are accompanied by a commensurate amount of fanfare. So we were surprised when, right about the time that we were walking by the Toyota booth, a chirp of our Blackberries brought us the press release announcing that the facelifted 2010 Toyota Camry and [Camry Hybrid](#) mid-sized sedans were appearing at the [2009 NAIAS](#) in Detroit.

We'd have hardly noticed if they hadn't told us—the headlamps are slightly larger, the grilles on the gas and hybrid models go their own way, and taillamps become primarily LED. Wheels are also new, but overall, the resemblance to the Corolla is downright uncanny. Let's hear it for brand consistency!

Four-cylinder Camrys get Toyota's new 2.5-liter four-cylinder with a boost in horsepower to 169 for LE and XLE grades, 179 for the Camry SE, all of which are available with six-speed manual or automatic transmissions. Camry V-6 and Hybrid models are relatively unchanged under the skin.

Inside, changes are likewise minimal: the options list has been shaken up a bit, but the only visible change on any Camry model is the Hybrid's new gauge cluster and softer seat fabrics. Pricing will be announced closer to the time it goes on sale this March.



### [Toyota Camry](#)

- › [Overview](#)
- › [Specifications](#)
- › [Price with Options](#)
- › [Get a Free Quote](#)

### News & Reviews

- [2010 Toyota Camry - Auto Shows](#)

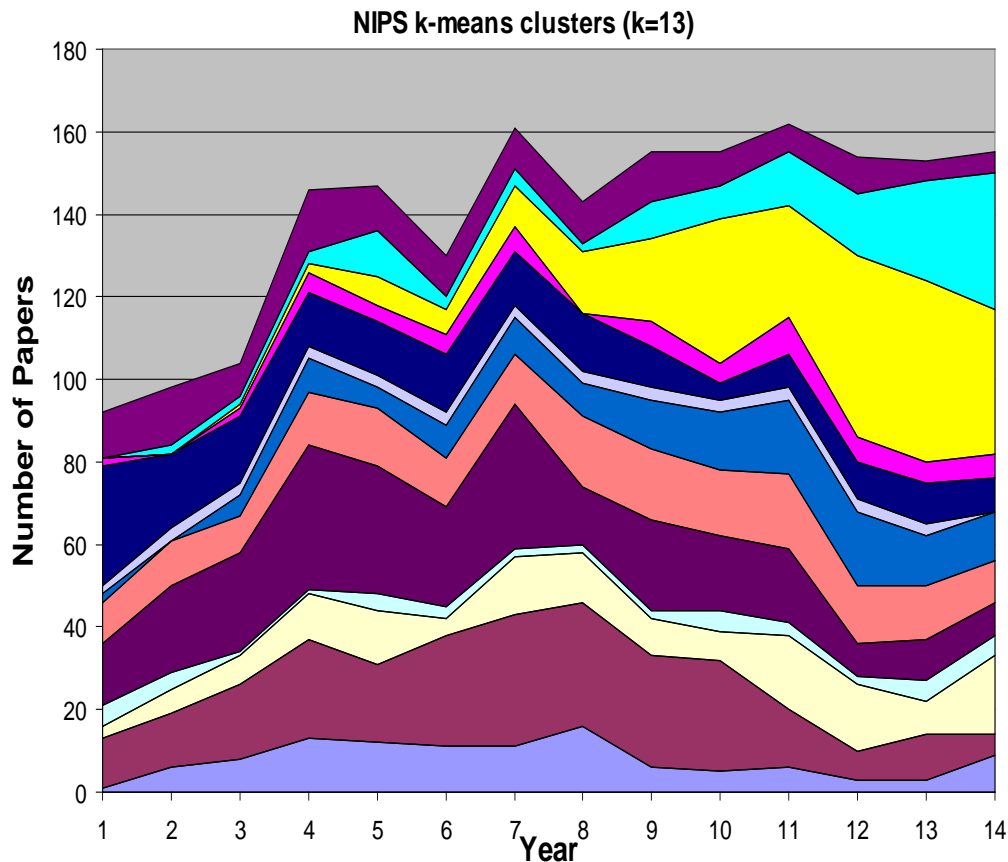
### Top Competitors

- [Chevrolet Malibu](#)
- [Ford Fusion](#)
- [Honda Accord sedan](#)

# Which are the key papers in Theoretical Computer Science?

- Hartmanis and Stearns, “On the computational complexity of algorithms”
- Blum, “A machine-independent theory of the complexity of recursive functions”
- Cook, “The complexity of theorem proving procedures”
- Karp, “Reducibility among combinatorial problems”
- Garey and Johnson, “Computers and Intractability: A Guide to the Theory of NP-Completeness”
- Yao, “Theory and Applications of Trapdoor Functions”
- Shafi Goldwasser, Silvio Micali, Charles Rackoff, “The Knowledge Complexity of Interactive Proof Systems”
- Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy, “Proof Verification and the Hardness of Approximation Problems”

# Temporal Cluster Histograms: NIPS Results



- 12: chip, circuit, analog, voltage, vlsi
- 11: kernel, margin, svm, vc, xi
- 10: bayesian, mixture, posterior, likelihood, em
- 9: spike, spikes, firing, neuron, neurons
- 8: neurons, neuron, synaptic, memory, firing
- 7: david, michael, john, richard, chair
- 6: policy, reinforcement, action, state, agent
- 5: visual, eye, cells, motion, orientation
- 4: units, node, training, nodes, tree
- 3: code, codes, decoding, message, hints
- 2: image, images, object, face, video
- 1: recurrent, hidden, training, units, error
- 0: speech, word, hmm, recognition, mlp

# Search

- When I query “coffee shop” I want one in Brazil, not Ithaca, NY.
- If someone queries “Michael Jordan”, they may want the basketball star. If I, as a computer scientist, query “Michael Jordan” I probably want the computer scientist at Berkeley.





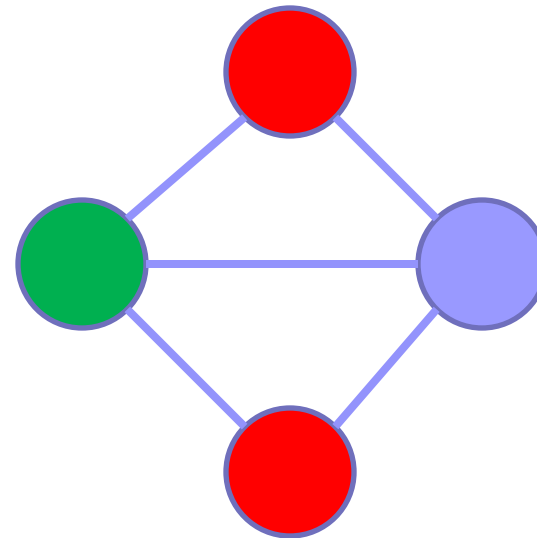
# Digitization of medical records

- Doctor – needs my entire medical record
- Insurance company – needs my last doctor visit, not my entire medical record
- Researcher – needs statistical information but no identifiable individual information

Relevant research – zero knowledge proofs, differential privacy

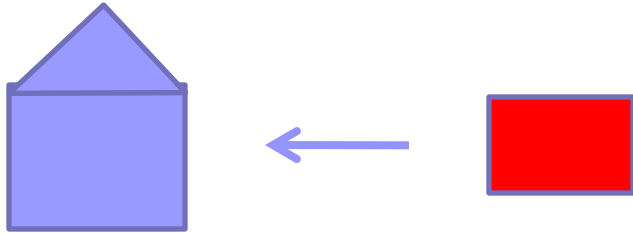
# Zero knowledge proof

- Graph 3-colorability

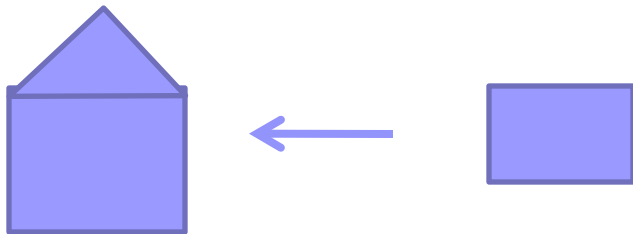


- Problem is NP-hard - No polynomial time algorithm unless  $P=NP$

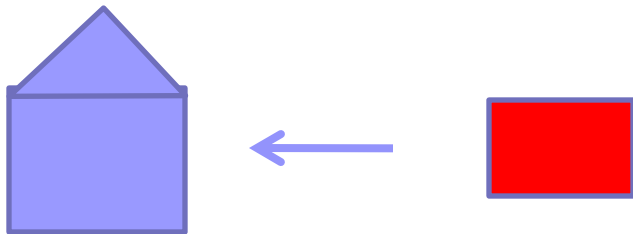
# Zero knowledge proof



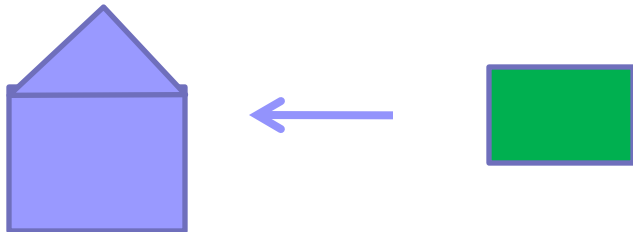
**I send the sealed envelopes.**




**You select an edge and open the two envelopes corresponding to the end points.**



**Then we destroy all envelopes and start over, but I permute the colors and then resend the envelopes.**





# Digitization of medical records is not the only system

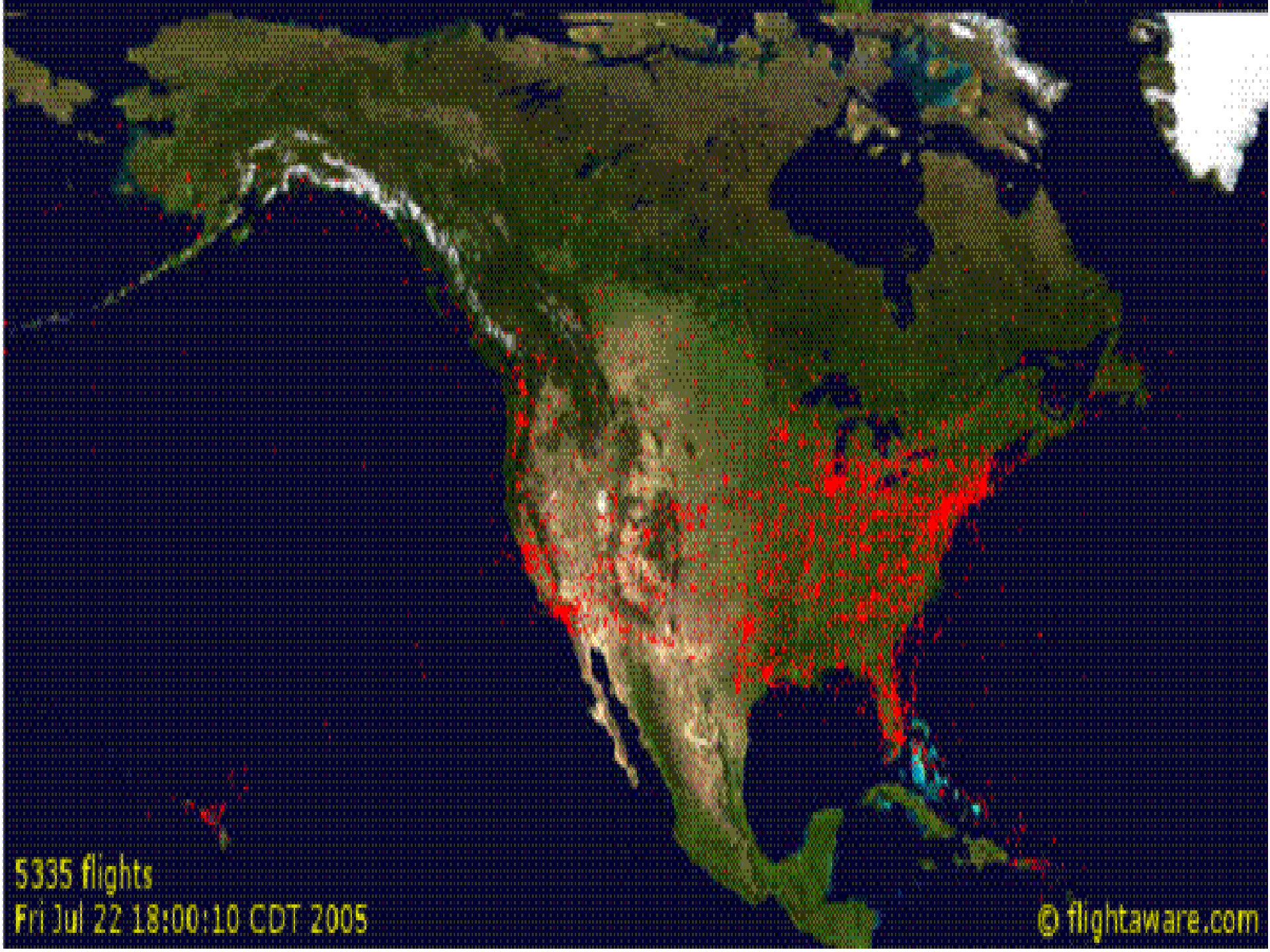
- Car and road – gps – privacy
- Supply chains
- Transportation systems



# Fed Ex package tracking

Tracking number	XXXXXXXXXXXXXXXX
Ship date	Dec 16, 2005
Delivered to	Receptionist/Front Desk
Destination	Ithaca, NY
Delivery date	Dec 19, 2005 9:28 AM
Signed for by	J. SMITH
Service type	Priority Pak

Date/Time	Location/Activity
Dec 19, 2005 9:28 AM	Ithaca, NY/Delivered
8:00 AM	ITHACA, NY/On FedEx vehicle for delivery
Dec 17, 2005 12:17 PM	ITHACA, NY/At local FedEx facility
9:26 AM	ITHACA, NY/At local FedEx facility
8:13 AM	SYRACUSE, NY/At dest sort facility
4:12 AM	MEMPHIS, TN/Departed FedEx location
12:03 AM	MEMPHIS, TN/Arrived at FedEx location
Dec 16, 2005 9:26 PM	WASHINGTON, DC/Left origin
6:58 PM	WASHINGTON, DC/Picked up
1:26 PM	/Package data transmitted to FedEx



5335 flights  
Fri Jul 22 18:00:10 CDT 2005

© flightaware.com



KSFO

COA41  
B738  
340 409

KEWR

0 420 840 1260 mi

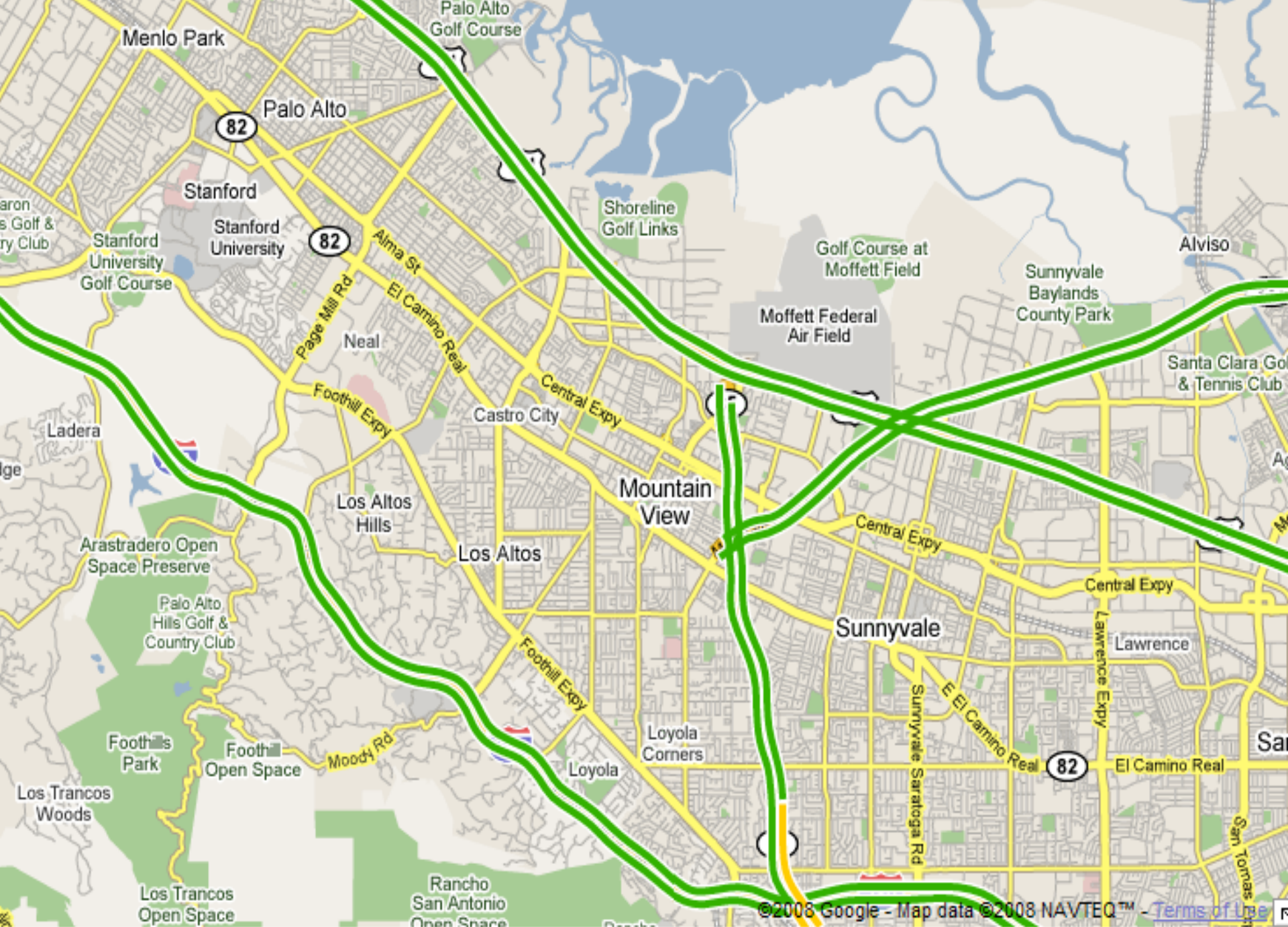
© 2005 FlightAware.com

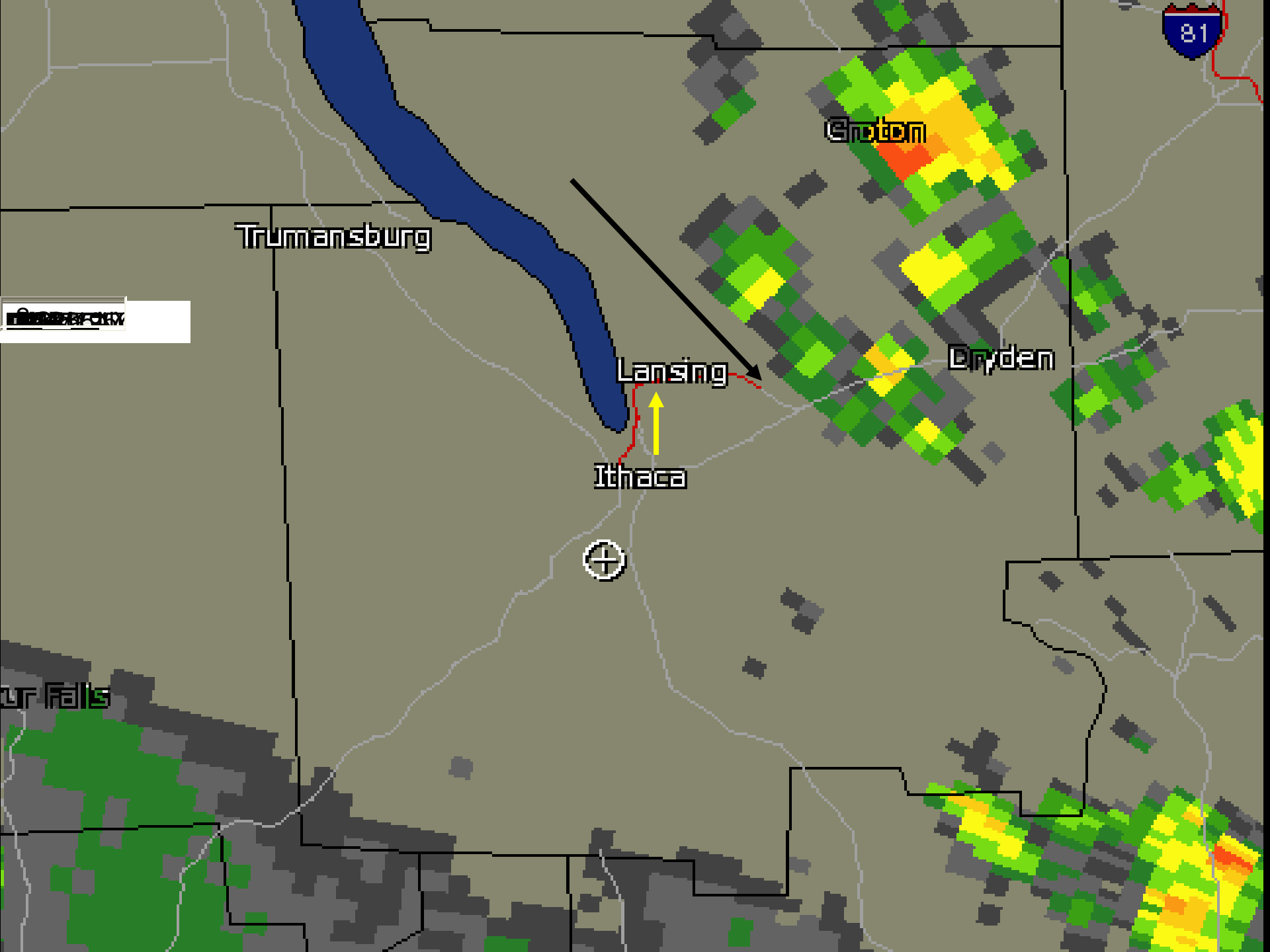




**Fri, 06 Jan 2006 11:58:00 PST**







81

Groton

Trumansburg

WFO

Lansing

Dryden

Ithaca

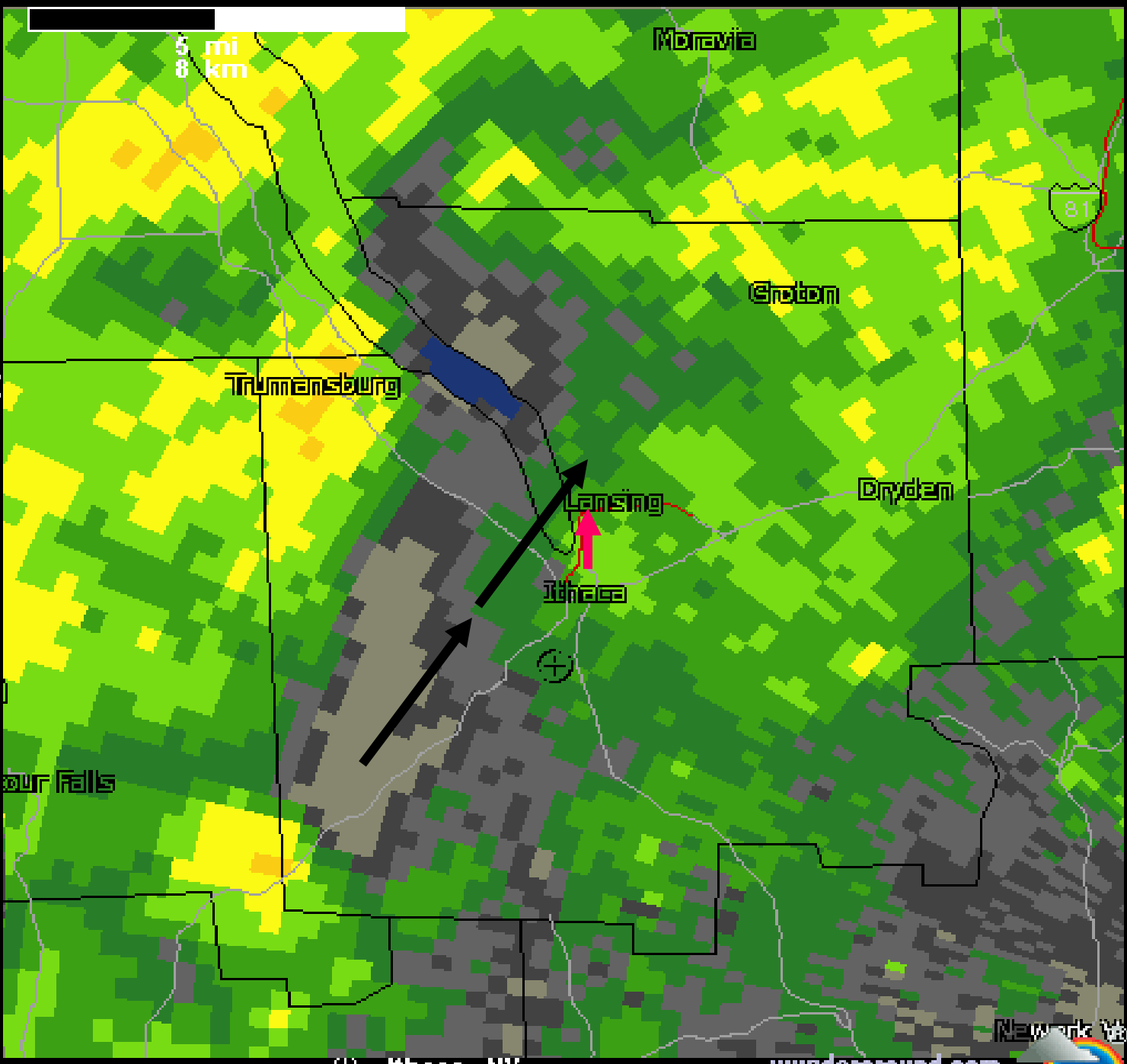
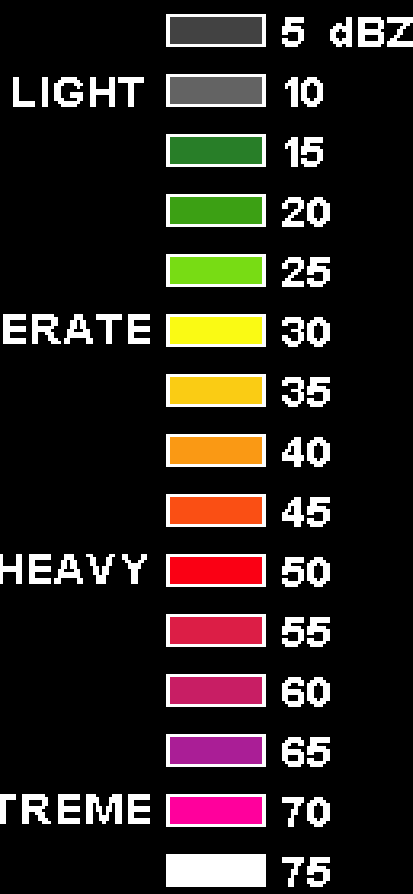
Thur Falls

13:37 EDT  
09/27/09

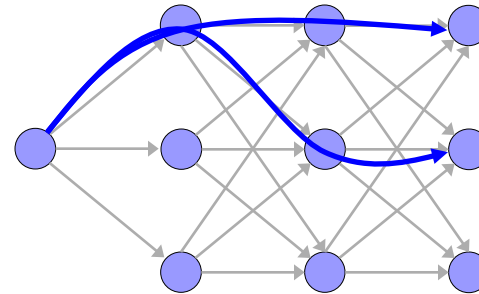
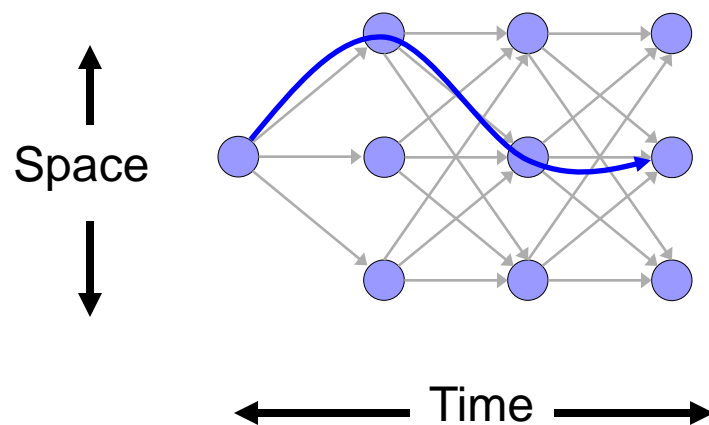


17:37 UTC  
09/27/09

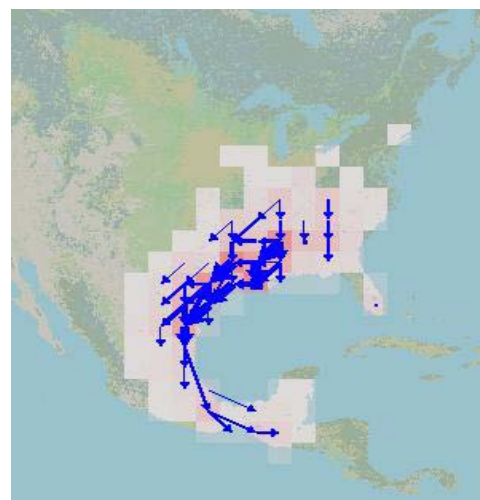
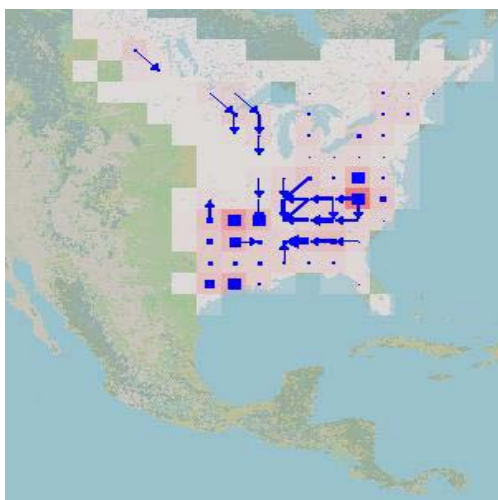
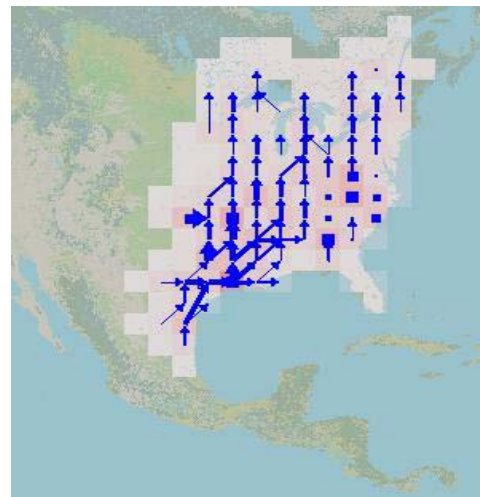
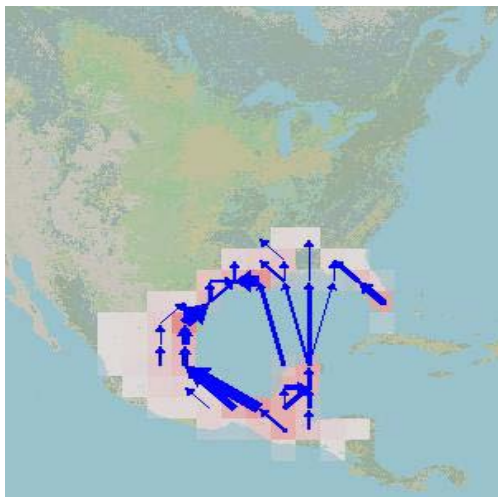
reflectivity 48 dbZ  
cov. pattern 21



# Collective Inference on Markov Models for Modeling Bird Migration







Daniel Sheldon, M. A. Saleh Elmohamed, Dexter Kozen



# Science base to support activities

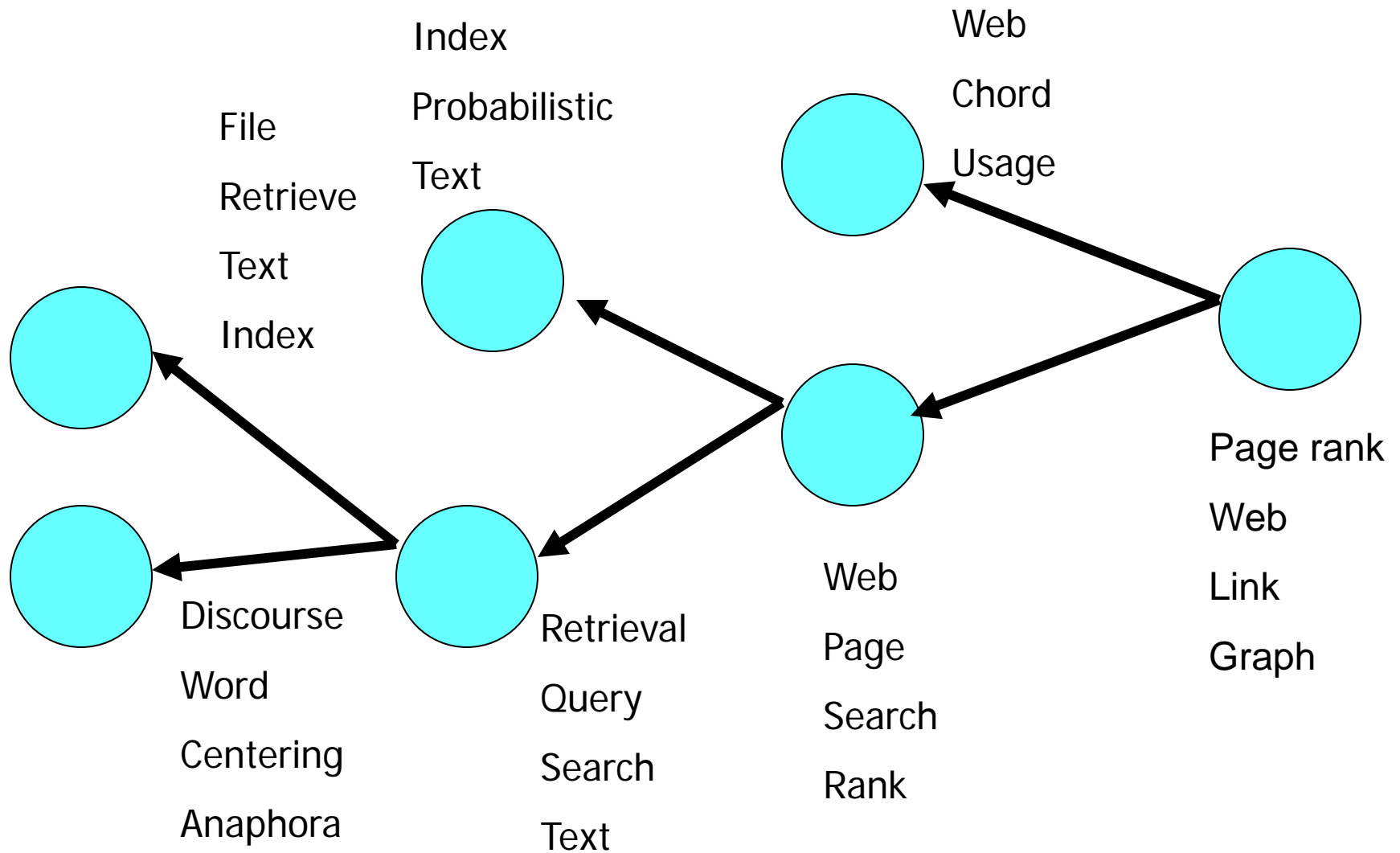
- Track flow of ideas in scientific literature
- Track evolution of communities in social networks
- Extract information from unstructured data sources.





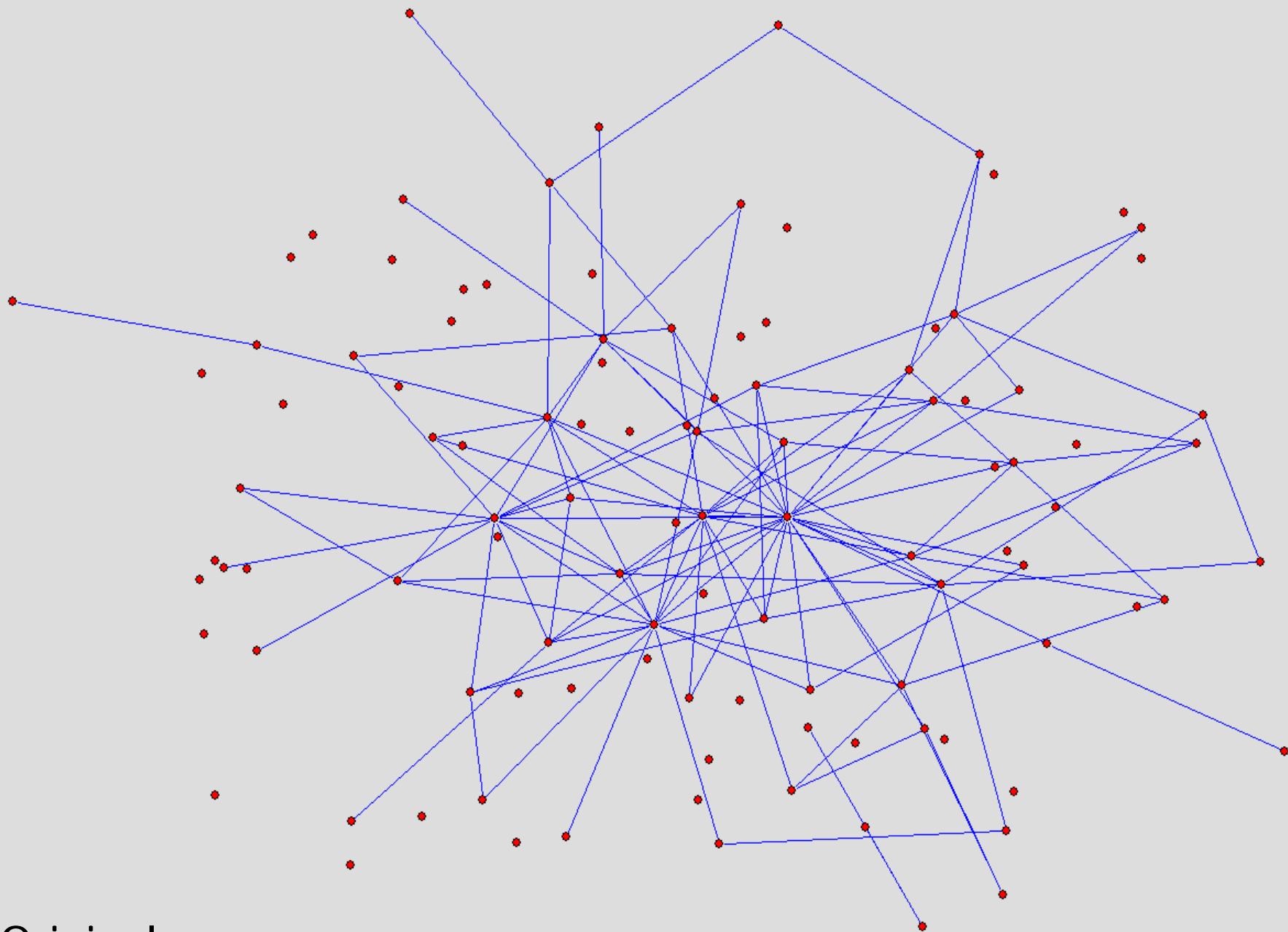
# Tracking the flow of ideas in scientific literature

Yookyung Jo

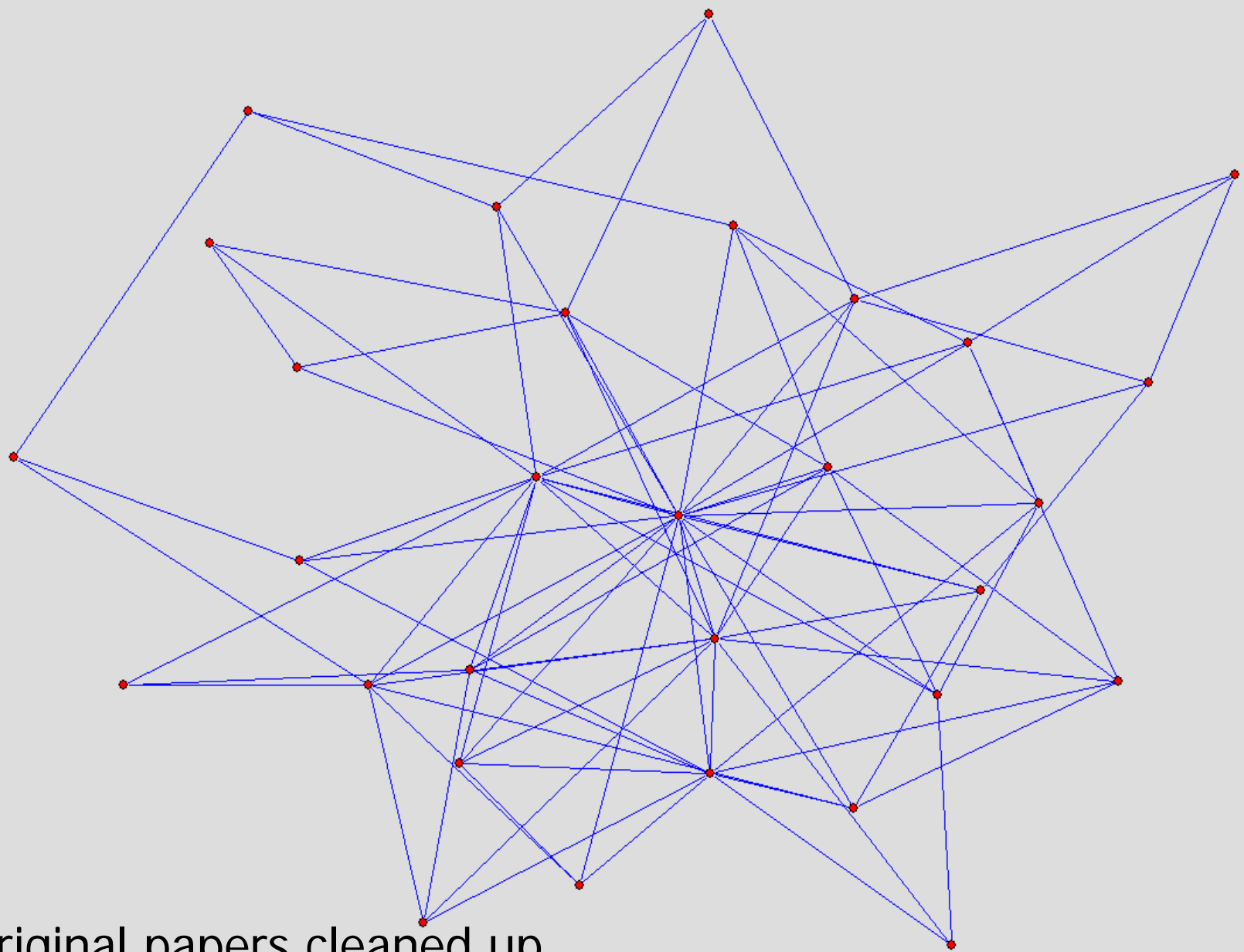


# Tracking the flow of ideas in scientific literature

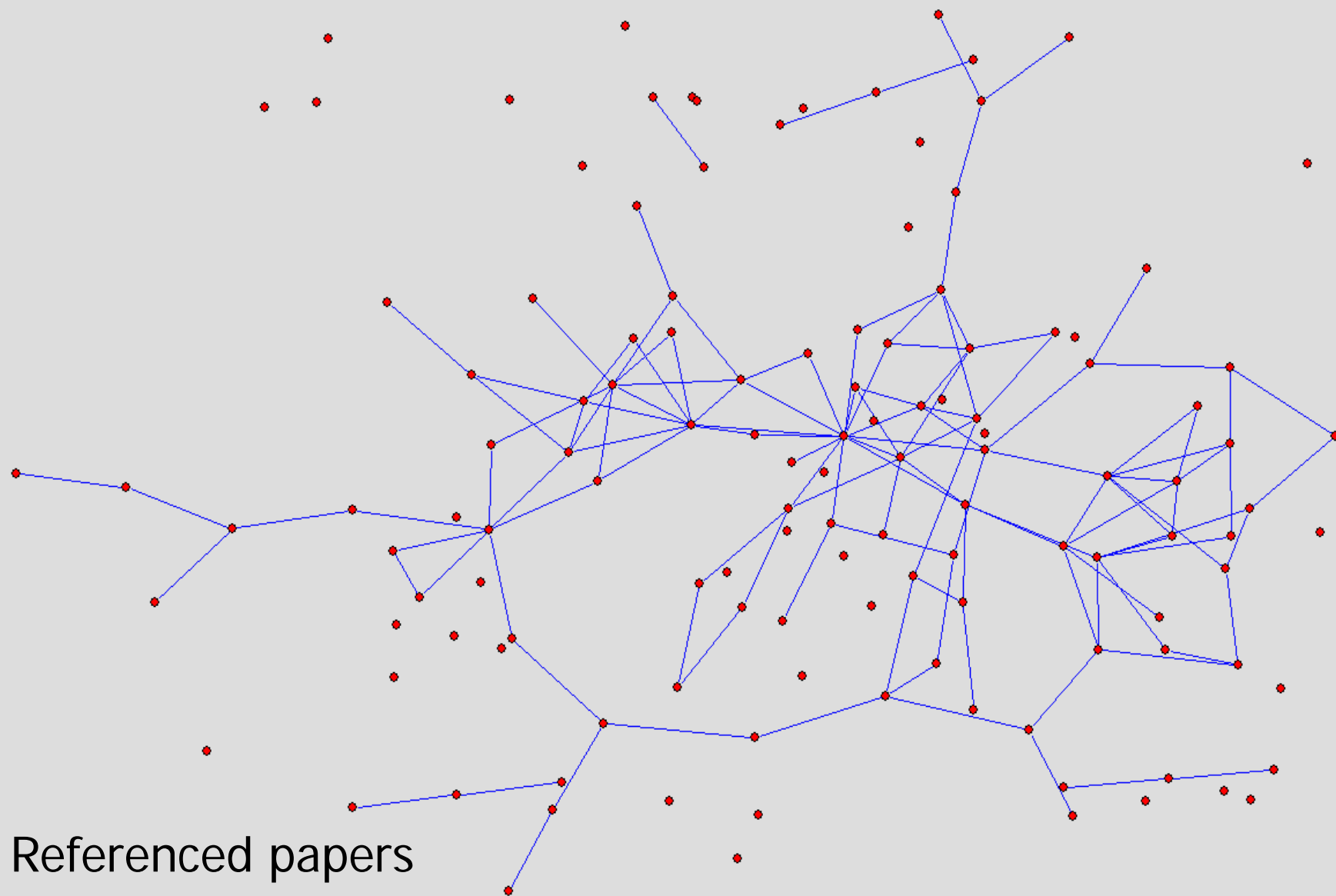
Yookyung Jo

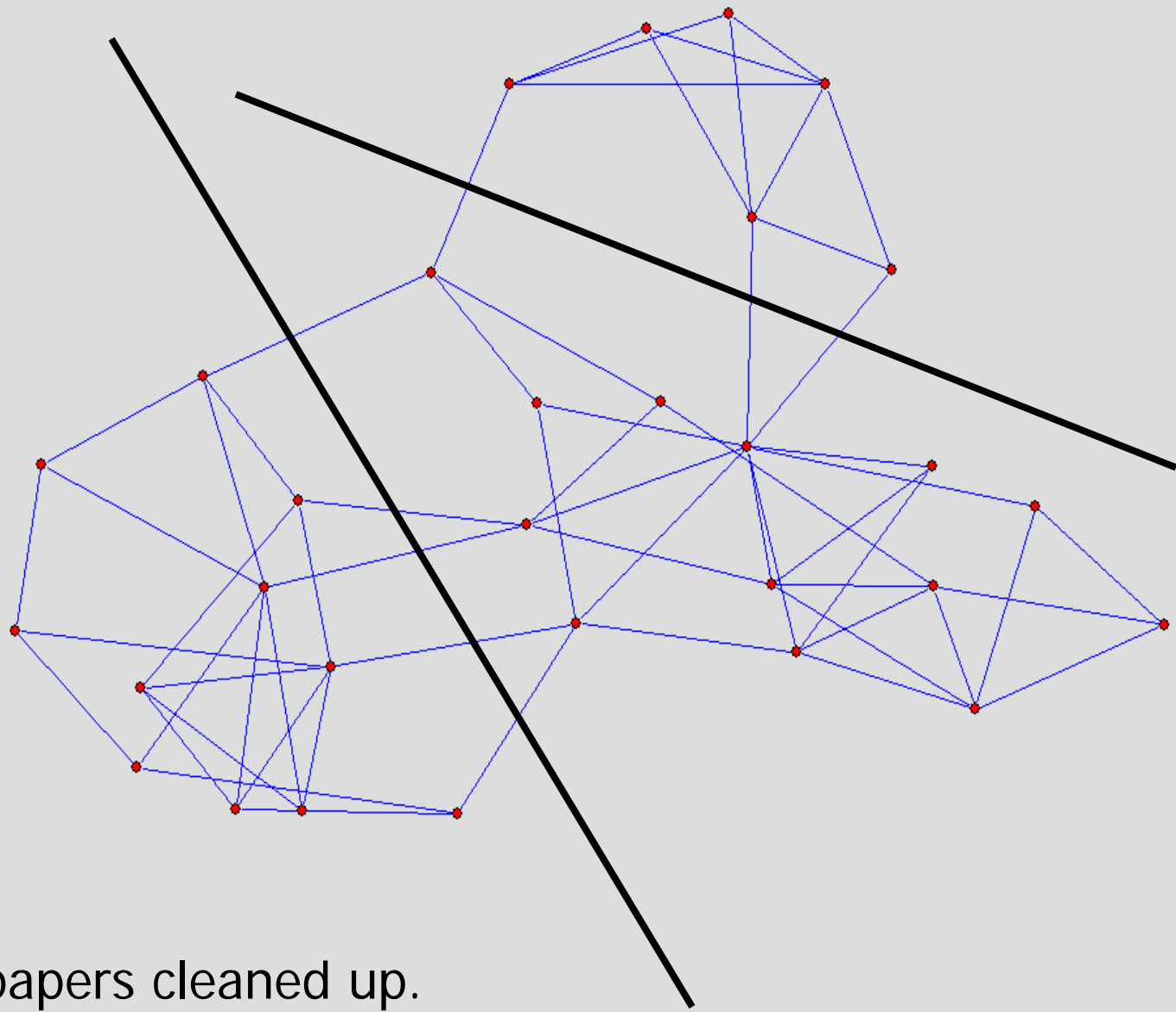


Original papers



Original papers cleaned up

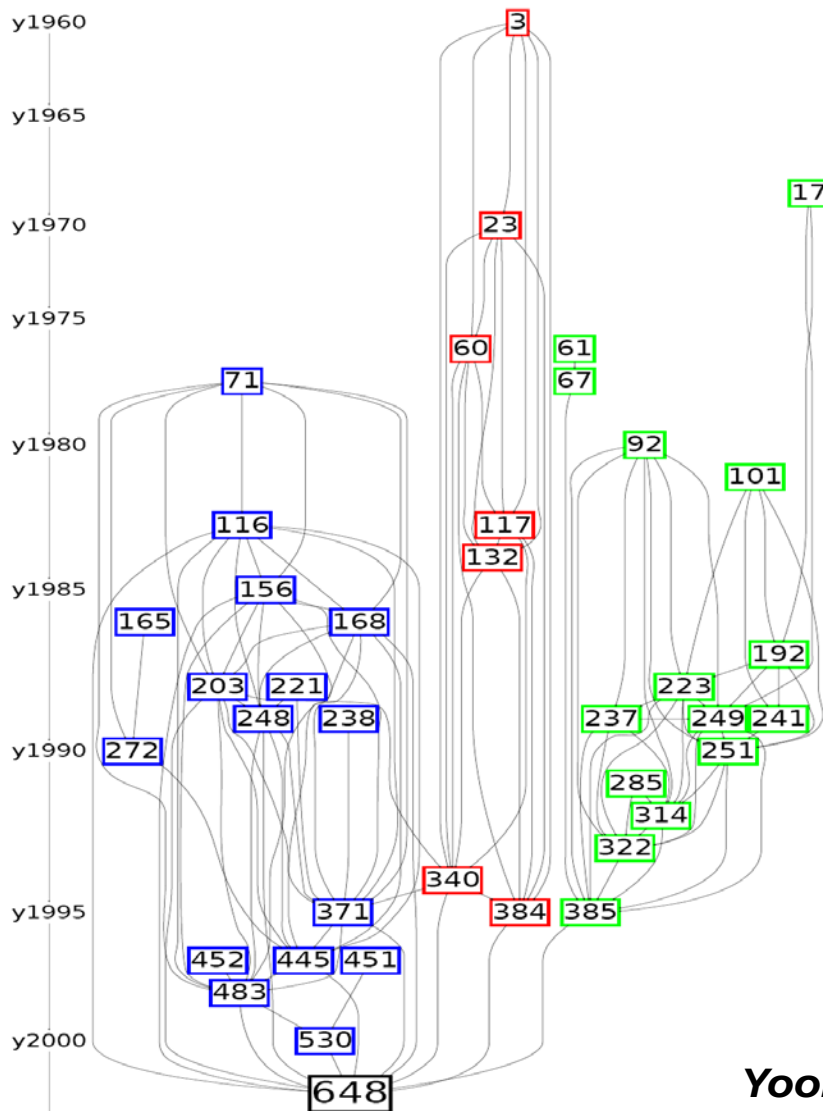




Referenced papers cleaned up.  
Three distinct categories of papers

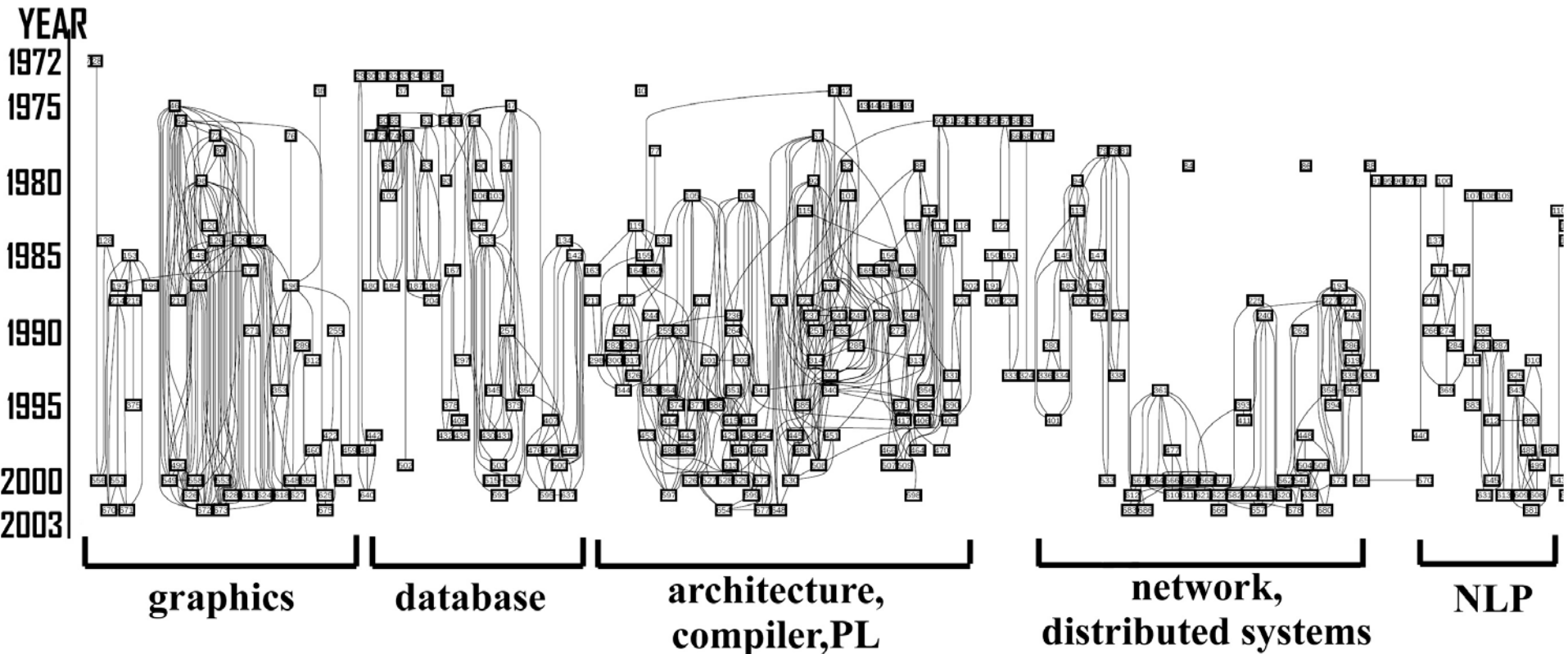
# Topic evolution thread

- Seed topic :
  - 648 : making c program type-safe by separating pointer types by their usage to prevent memory errors
- 3 subthreads :
  - **Type :**
  - **Garbage collection :**
  - **Pointer analysis :**



*Yookyung Jo, 2010*

# Topic Evolution Map of the ACM corpus




*Yookyung Jo, 2010*





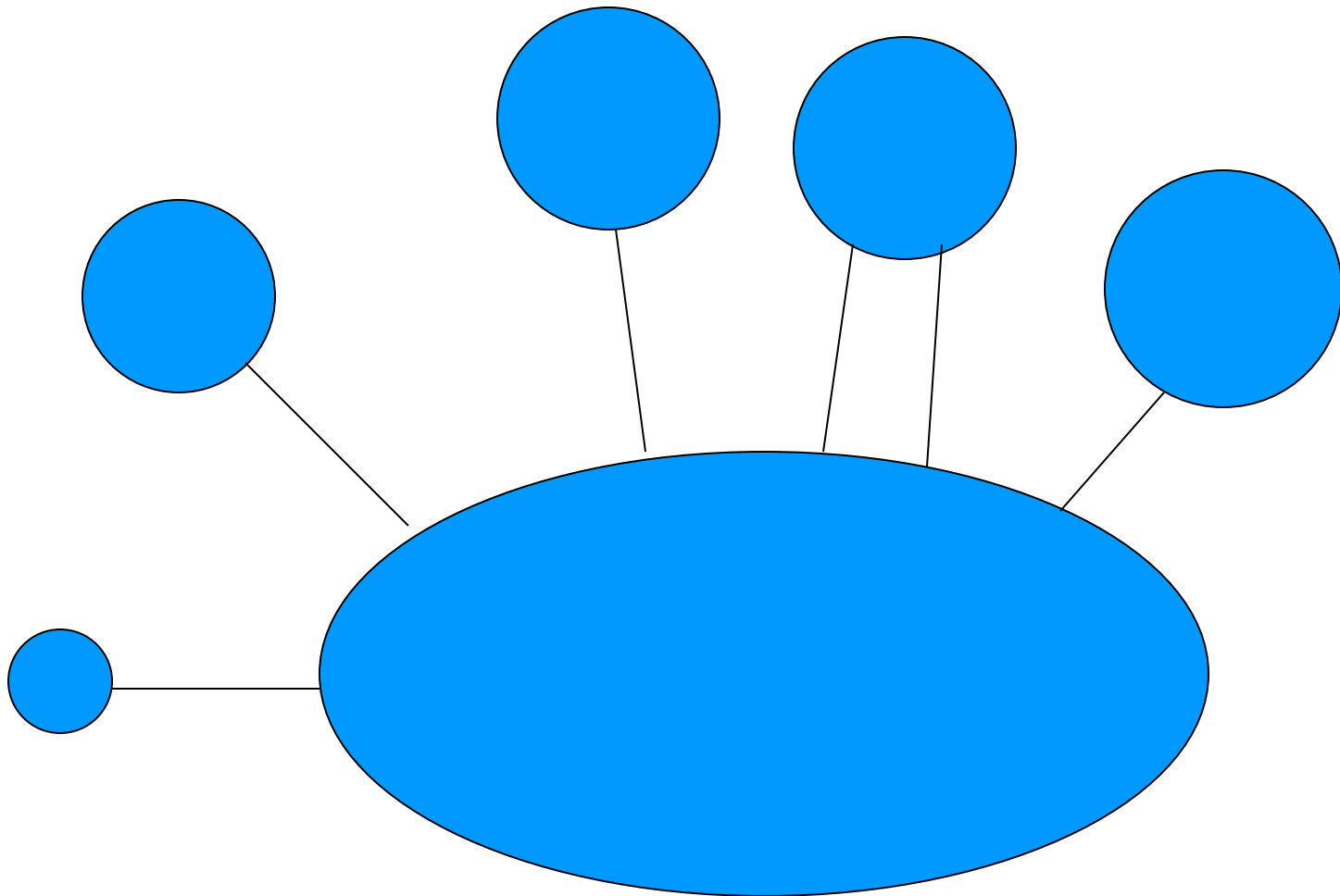
# Tracking communities in social networks

Liaoruo Wang



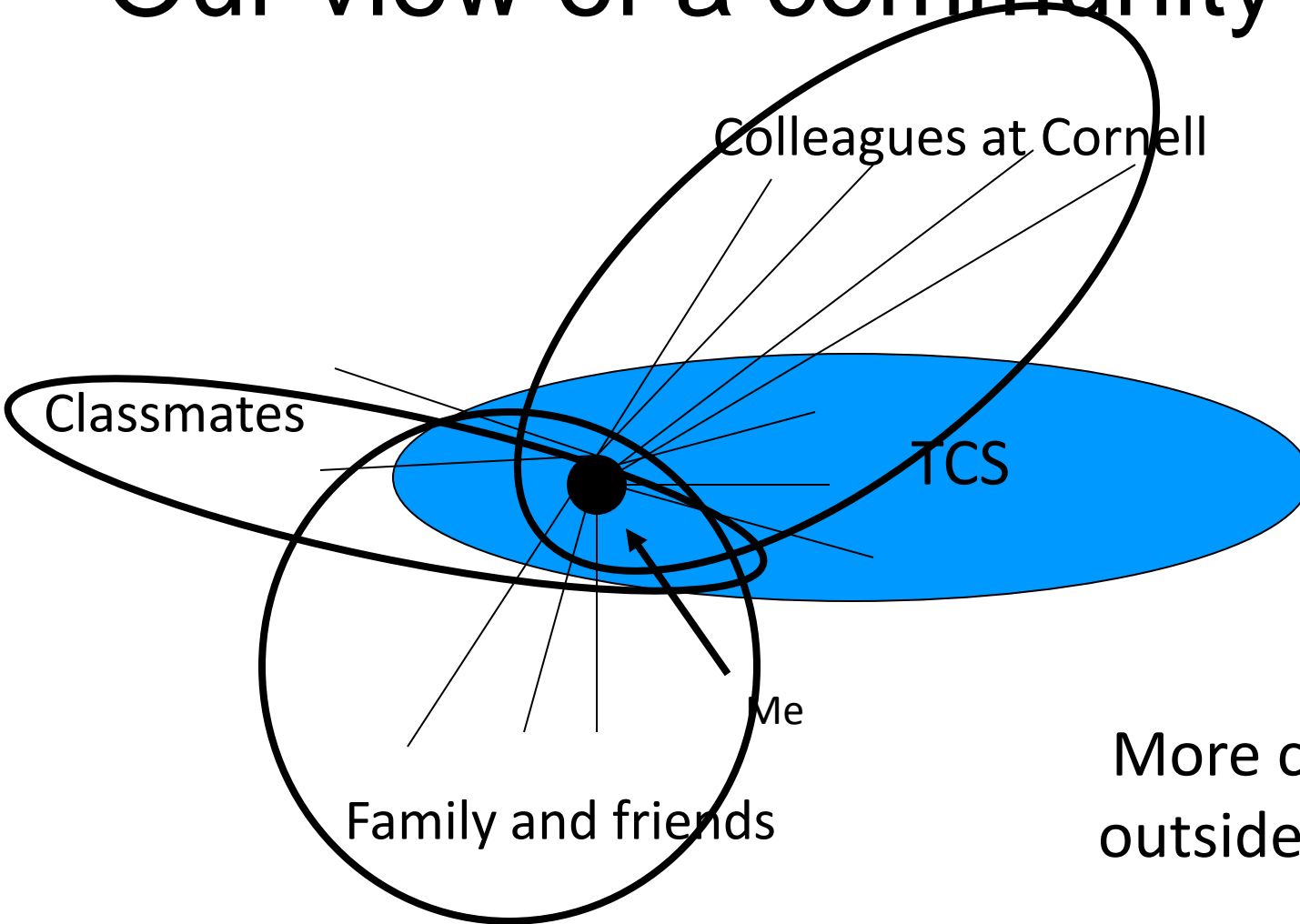
## “Statistical Properties of Community Structure in Large Social and Information Networks”, Jure Leskovec; Kevin Lang; Anirban Dasgupta; Michael Mahoney

- Studied over 70 large sparse real-world networks.
- Best communities are of approximate size 100 to 150.



Whisker: A component with  $v$  vertices connected by  $e \leq v$  edges

# Our view of a community



More connections  
outside than inside



## ■ Early work

- Min cut – two equal size communities
- Conductance – minimizes cross edges

## ■ Future work

- Consider communities with more external edges than internal edges
- Find small communities
- Track communities over time
- Develop appropriate definitions for communities
- Understand the structure of different types of social networks



# “Clustering Social networks”

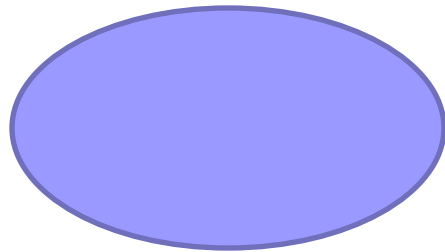
Mishra, Schreiber, Stanton, and Tarjan

- Each member of community is connected to a beta fraction of community
- No member outside the community is connected to more than an alpha fraction of the community
- Some connectivity constraint



# In sparse graphs

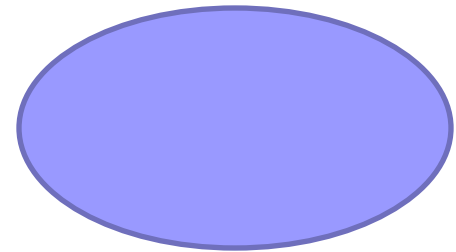
- Does every graph have an alpha beta community?
- How do you find alpha-beta communities?
- How many communities of a given size are there in a social network?



200 randomly  
chosen nodes



Apply alpha-  
beta algorithm



Resulting alpha-  
beta community

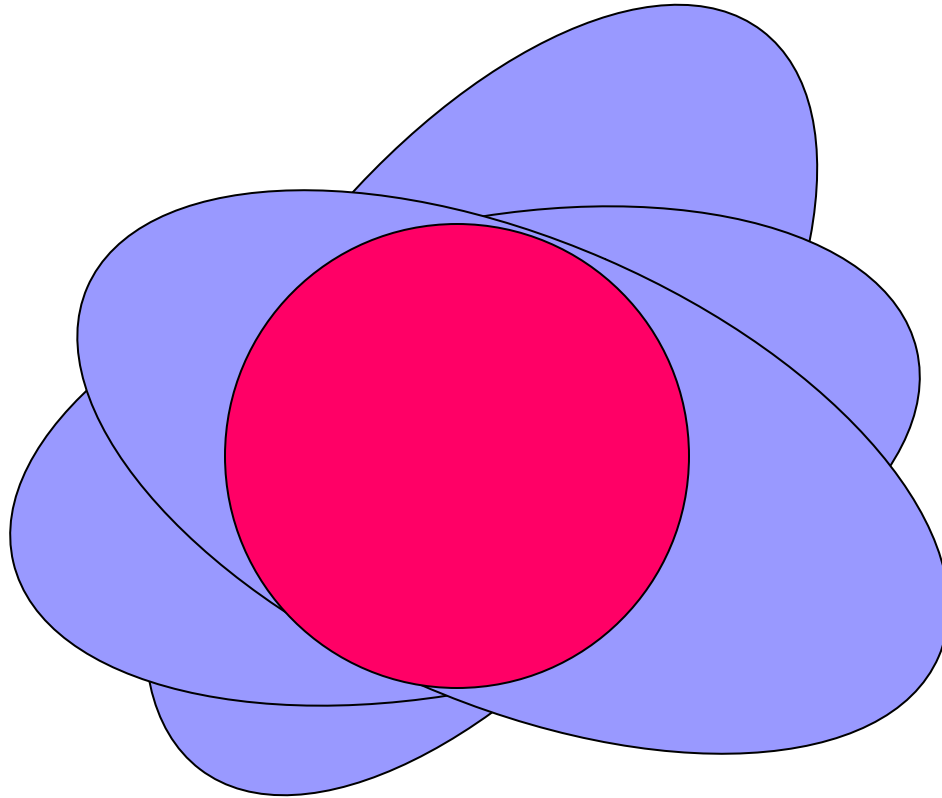
How many alpha-beta communities?







# Massively overlapping communities

- Are there a small number of massively overlapping communities that share a common core?
- Are there massively overlapping communities in which one can move from one community to a totally disjoint community?

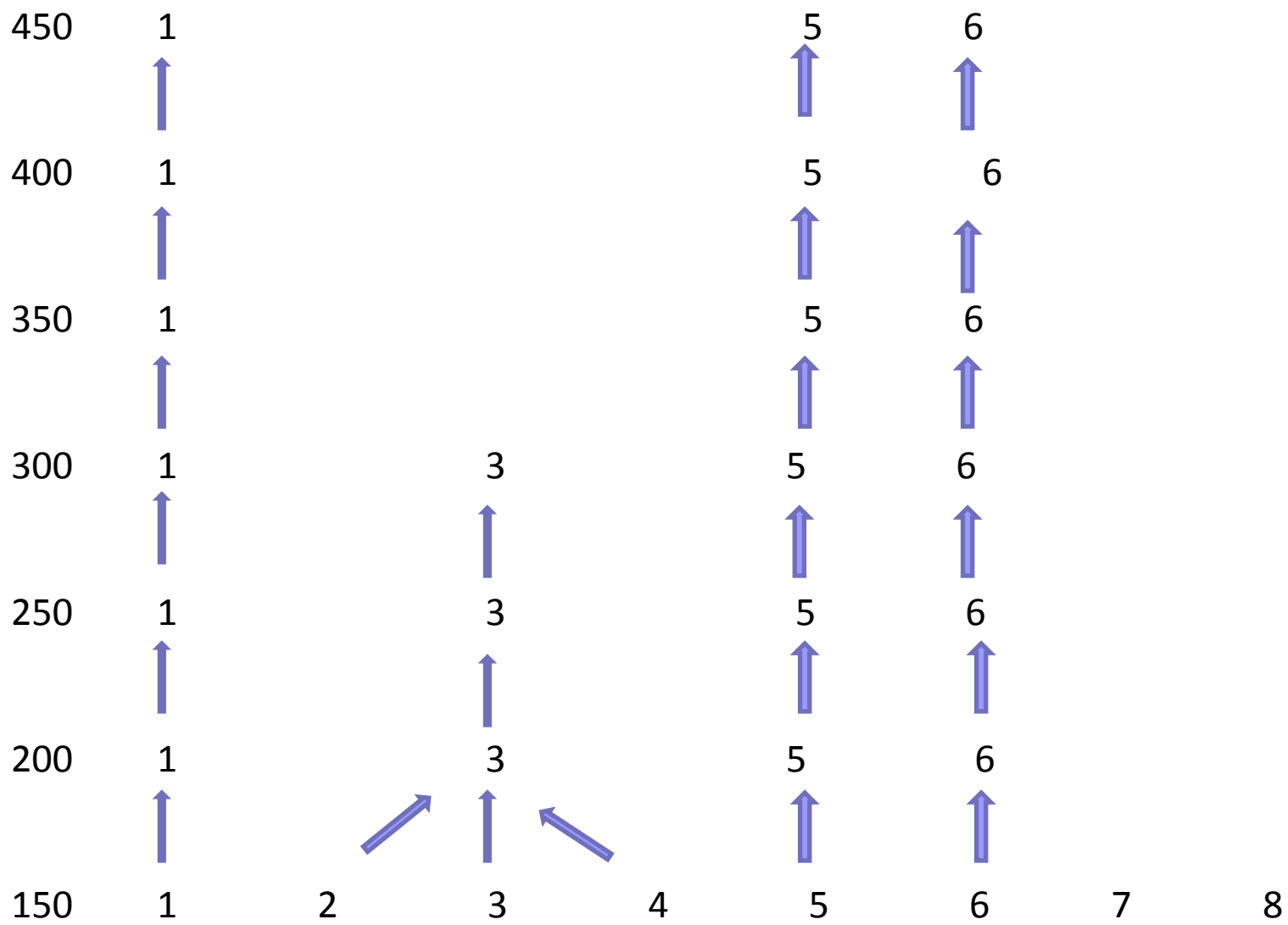



Massively overlapping communities with a common core

- 
- Define the core of a set of overlapping communities to be the intersection of the communities.
  - There are a small number of cores in the Tweeter data set.



Size of initial set	Number of cores
25	221
50	94
100	19
150	8
200	4
250	4
300	4
350	3
400	3
450	3



- 
- What is the graph structure that causes certain cores to merge and others to simply vanish?
  - What is the structure of cores as they get larger? Do they consist of concentric layers that are less dense at the outside?
  - Do different social networks have quite different structures?



# Sparse vectors

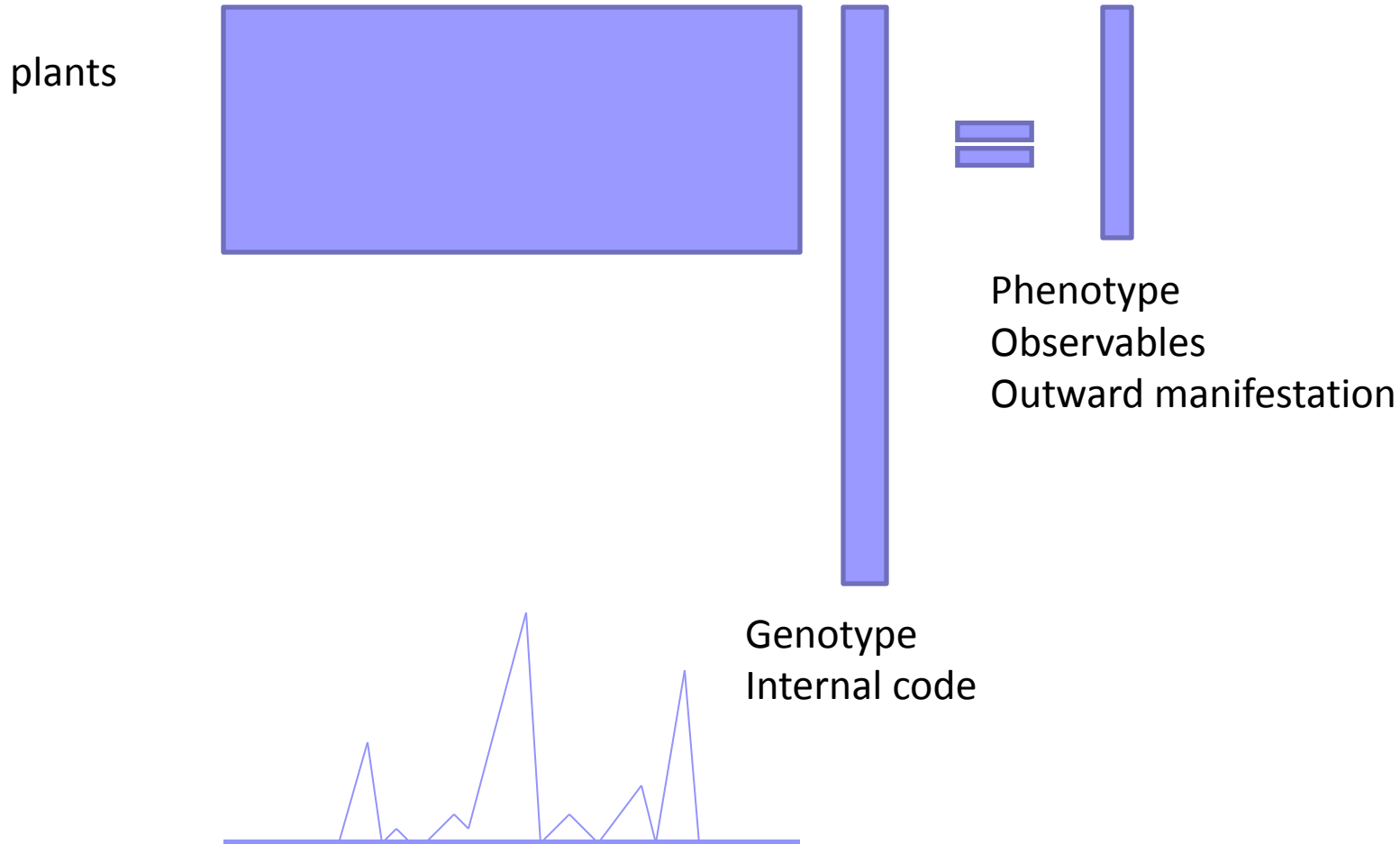
There are a number of situations where sparse vectors are important

Tracking the flow of ideas in scientific literature

Biological applications

Signal processing

# Sparse vectors in biology







# Theory to support new directions

- Large graphs
- Spectral analysis
- High dimensions and dimension reduction
- Clustering
- Collaborative filtering
- Extracting signal from noise

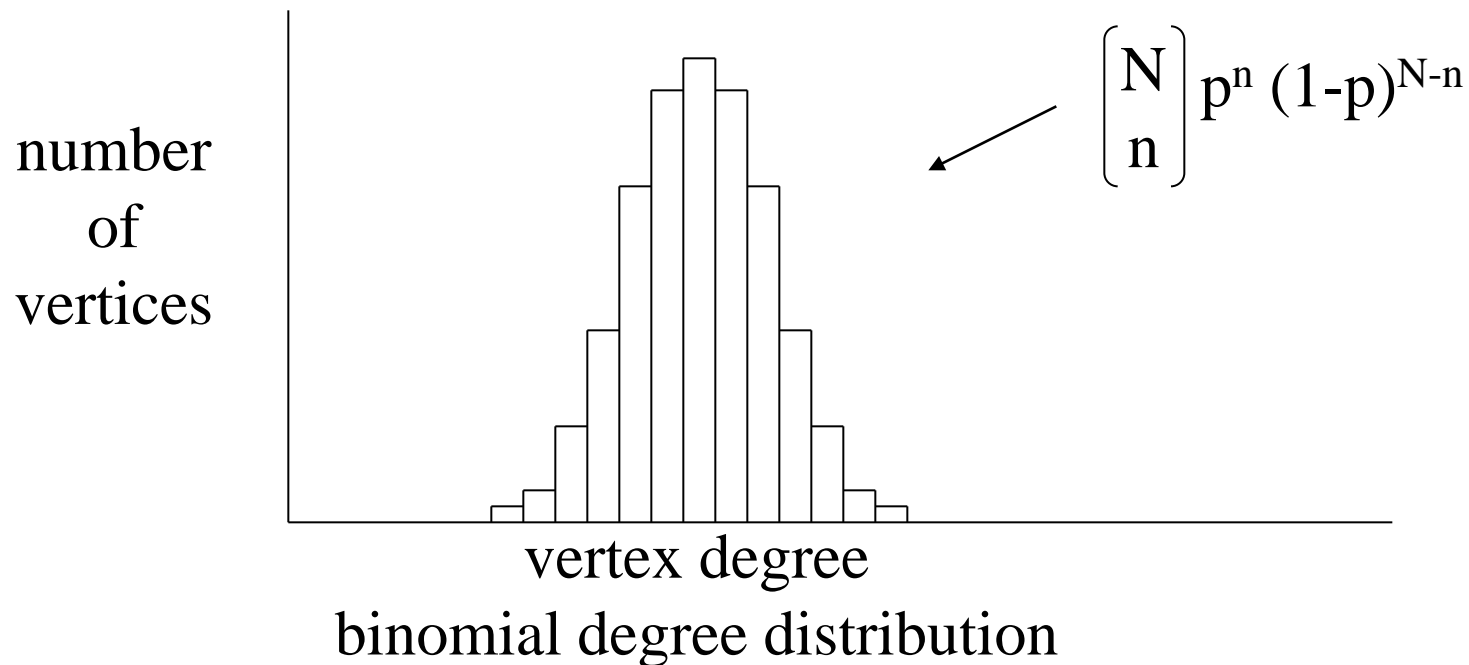


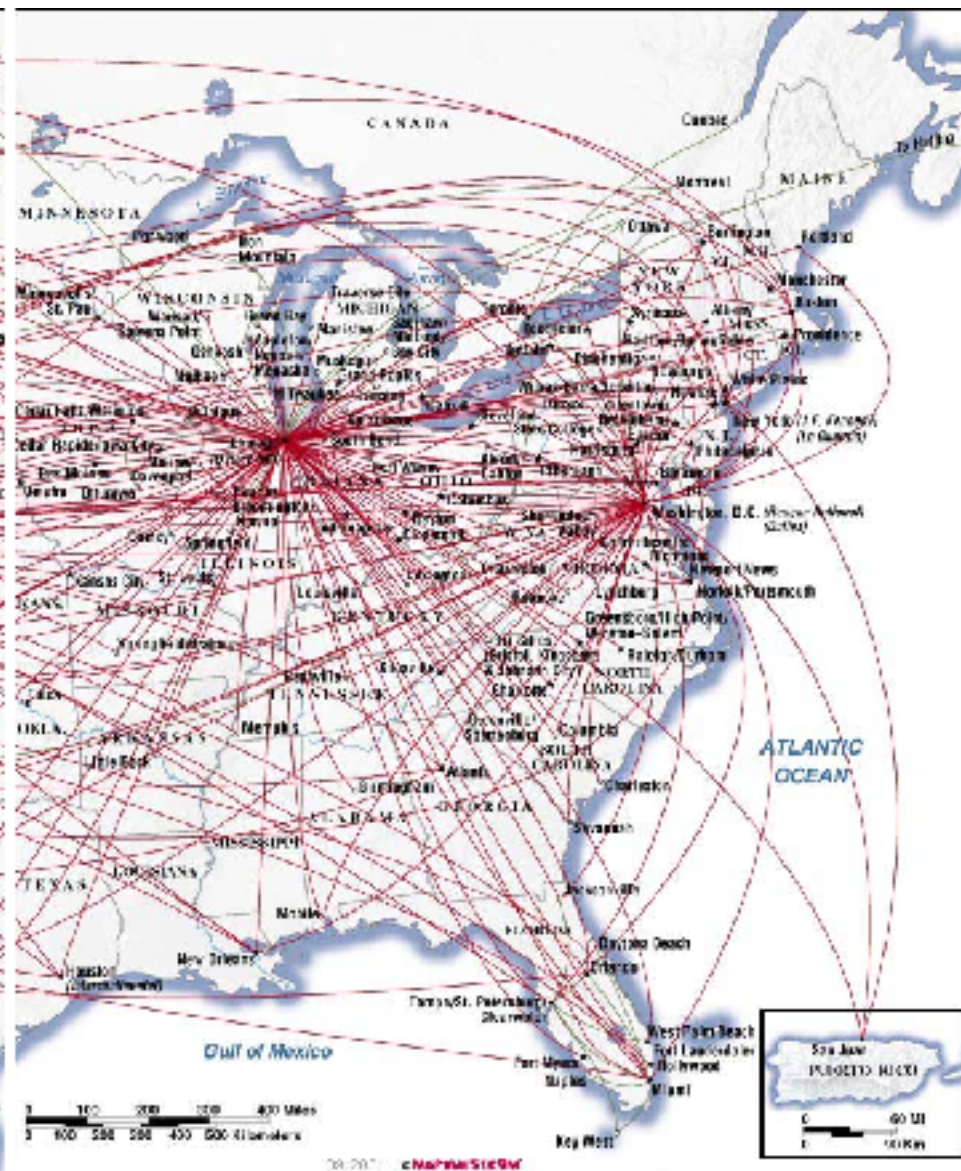
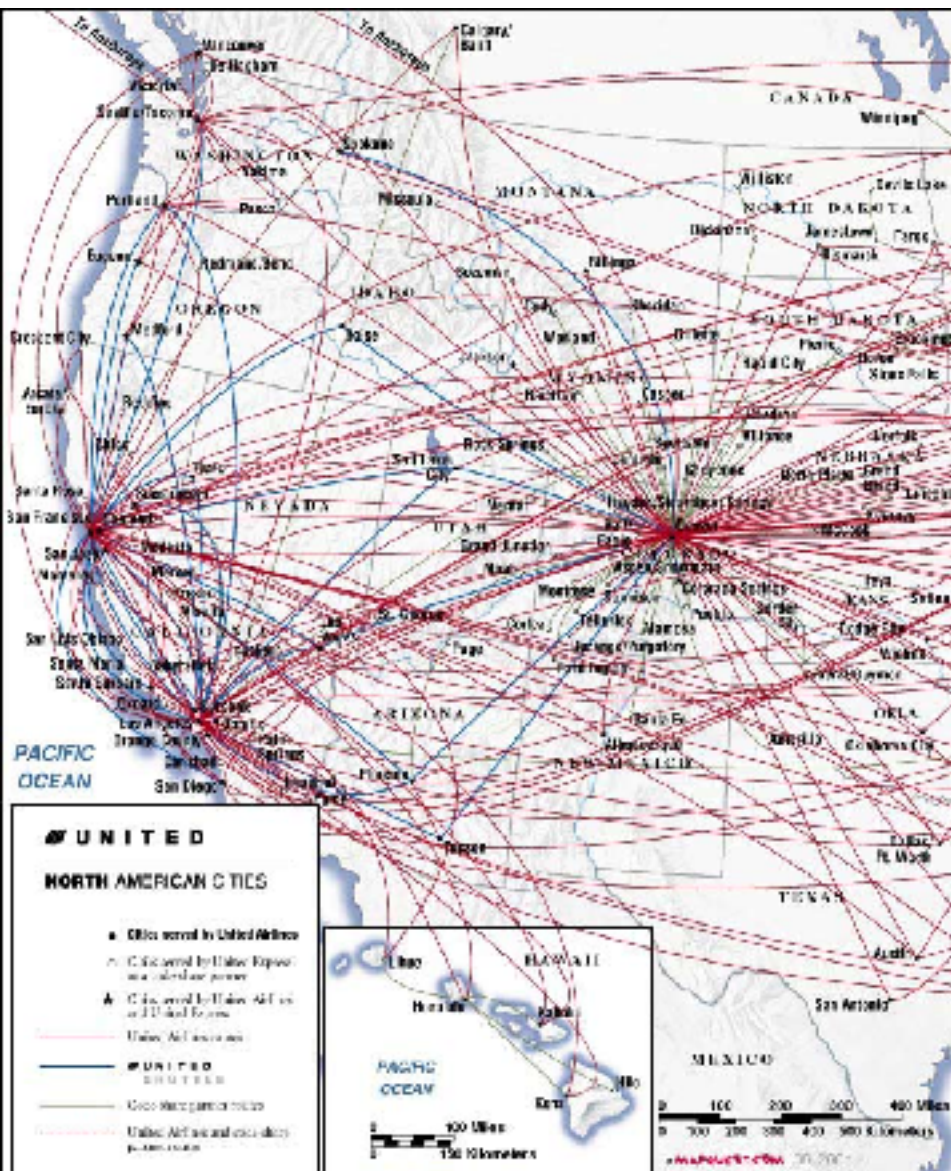
# Theory of Large Graphs

- Large graphs with billions of vertices
- Exact edges present not critical
- Invariant to small changes in definition
- Must be able to prove basic theorems

# Erdős-Renyi

- $n$  vertices
- each of  $n^2$  potential edges is present with independent probability





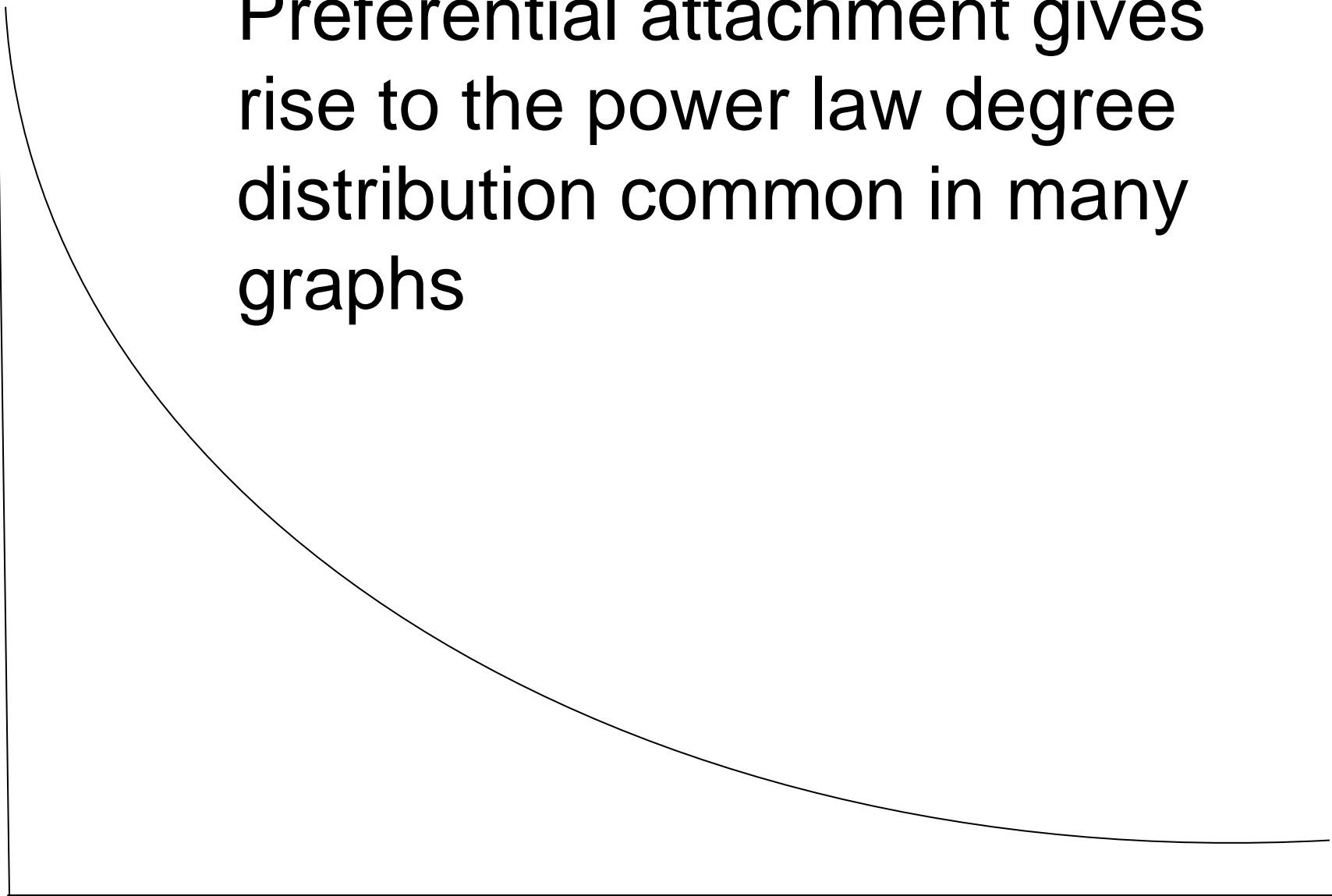


# Generative models for graphs

- Vertices and edges added at each unit of time
- Rule to determine where to place edges
  - Uniform probability
  - Preferential attachment - gives rise to power law degree distributions

Preferential attachment gives rise to the power law degree distribution common in many graphs

Number  
of  
vertices



Vertex degree

# Protein interactions

2730 proteins in data base

3602 interactions between proteins

SIZE OF COMPONENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	1000
NUMBER OF COMPONENTS	48	179	50	25	14	6	4	6	1	1	1	0	0	0	0	1		0

Only 899 proteins in components. Where are the 1851 missing proteins?

Science 1999 July 30; 285:751-753



# Protein interactions

2730 proteins in data base

3602 interactions between proteins

SIZE OF COMPONENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	1851
NUMBER OF COMPONENTS	48	179	50	25	14	6	4	6	1	1	1	0	0	0	0	1		1

Science 1999 July 30; 285:751-753

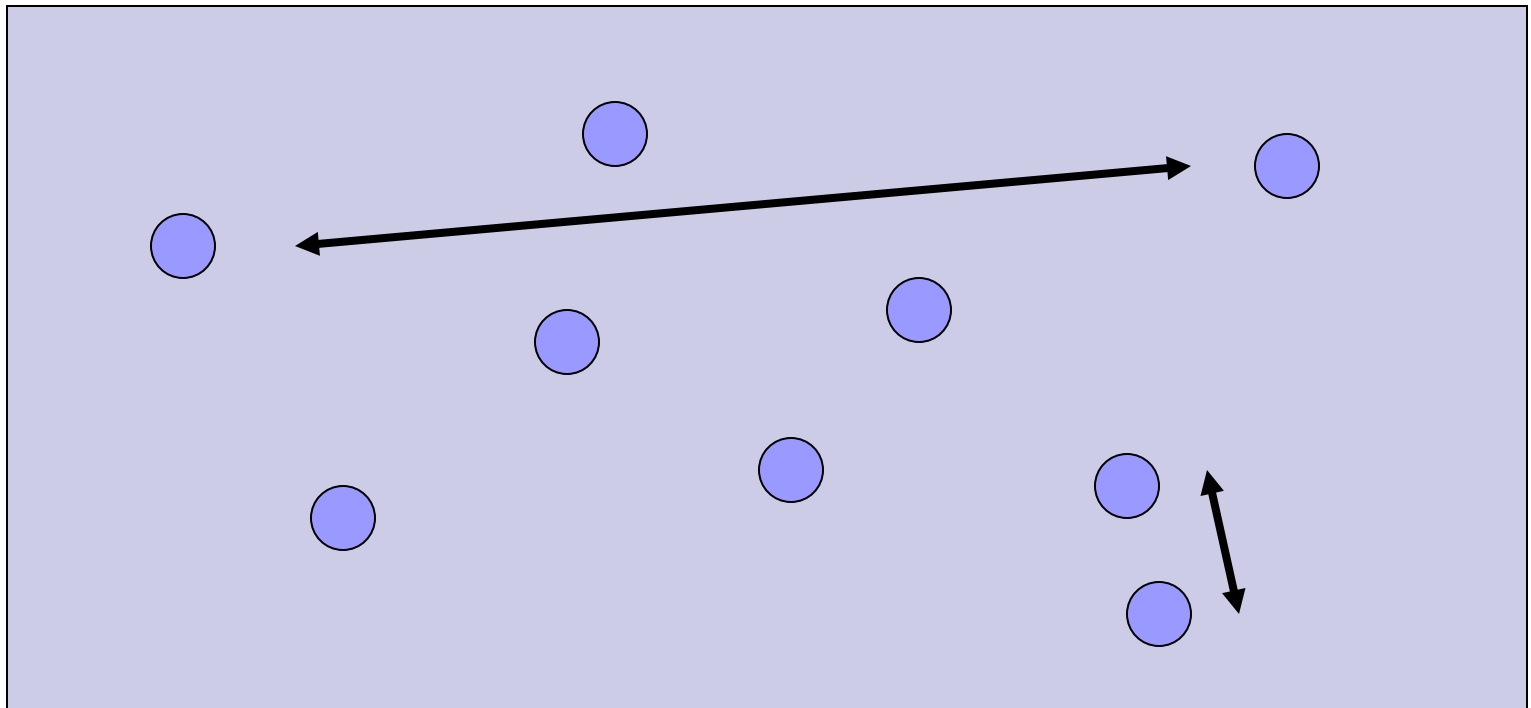




# Science base

- What do we mean by science base?
  - Example: High dimensions

High dimension is  
fundamentally different from 2  
or 3 dimensional space



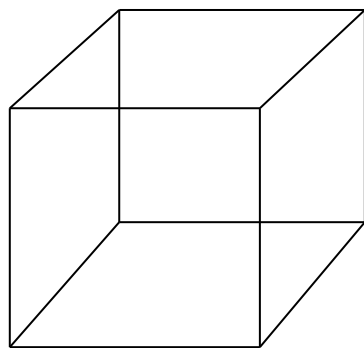
# High dimensional data is inherently unstable

- Given  $n$  random points in  $d$  dimensional space, essentially all  $n^2$  distances are equal.

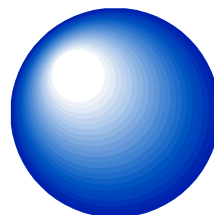
- $$|x - y|^2 = \sum_{i=1}^d (x_i - y_i)^2$$

# High Dimensions

Intuition from two and three dimensions not valid for high dimension

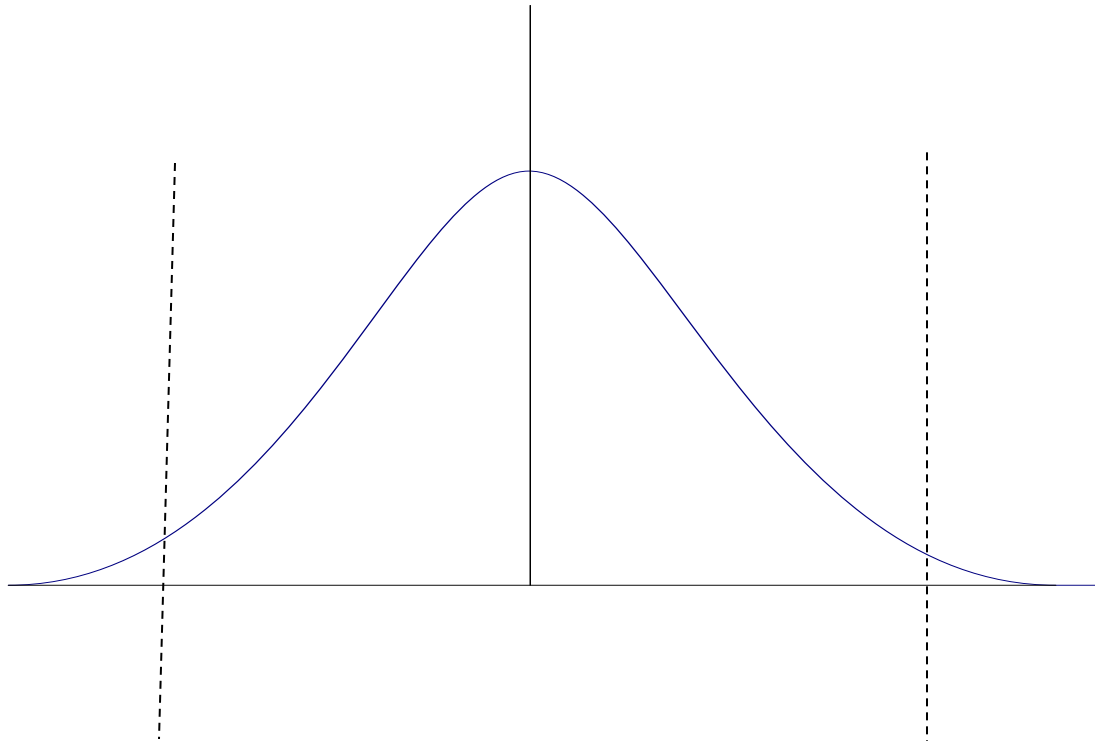


Volume of cube is  
one in all  
dimensions



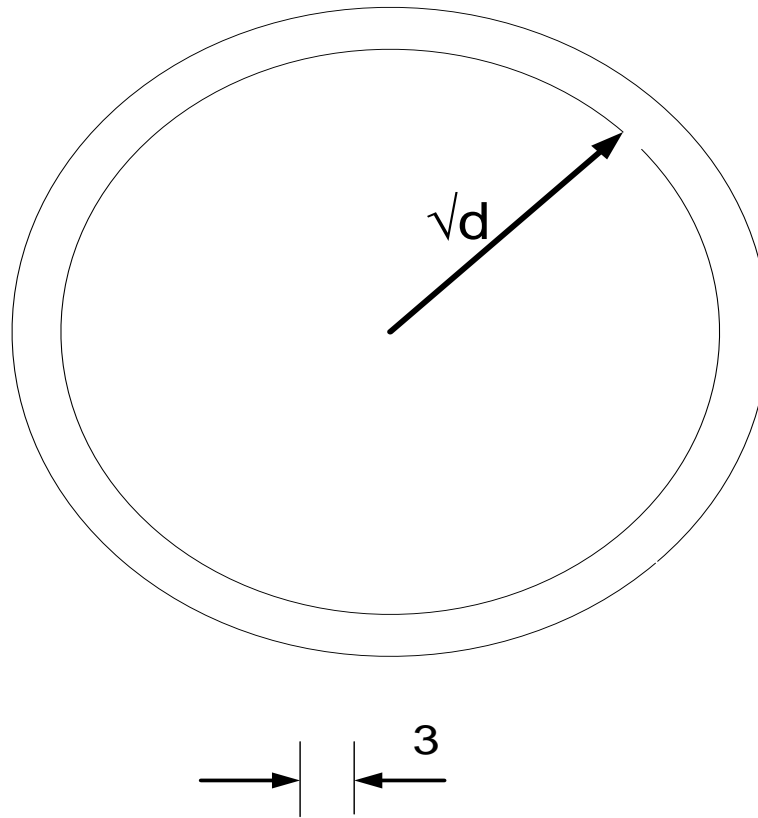
Volume of  
sphere goes to  
zero

# Gaussian distribution

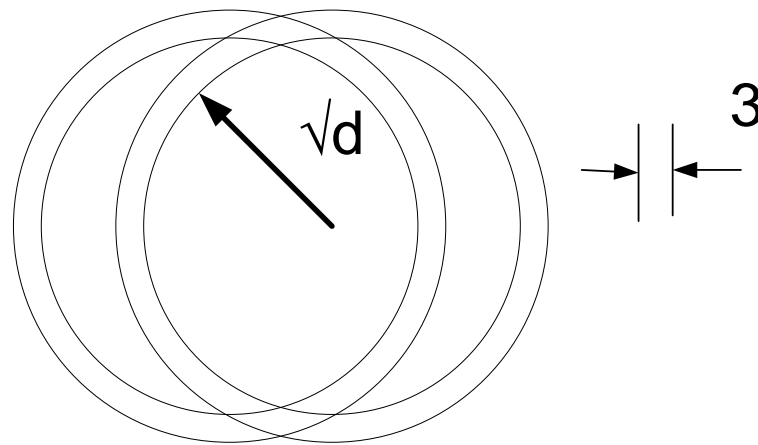


Probability mass concentrated  
between dotted lines

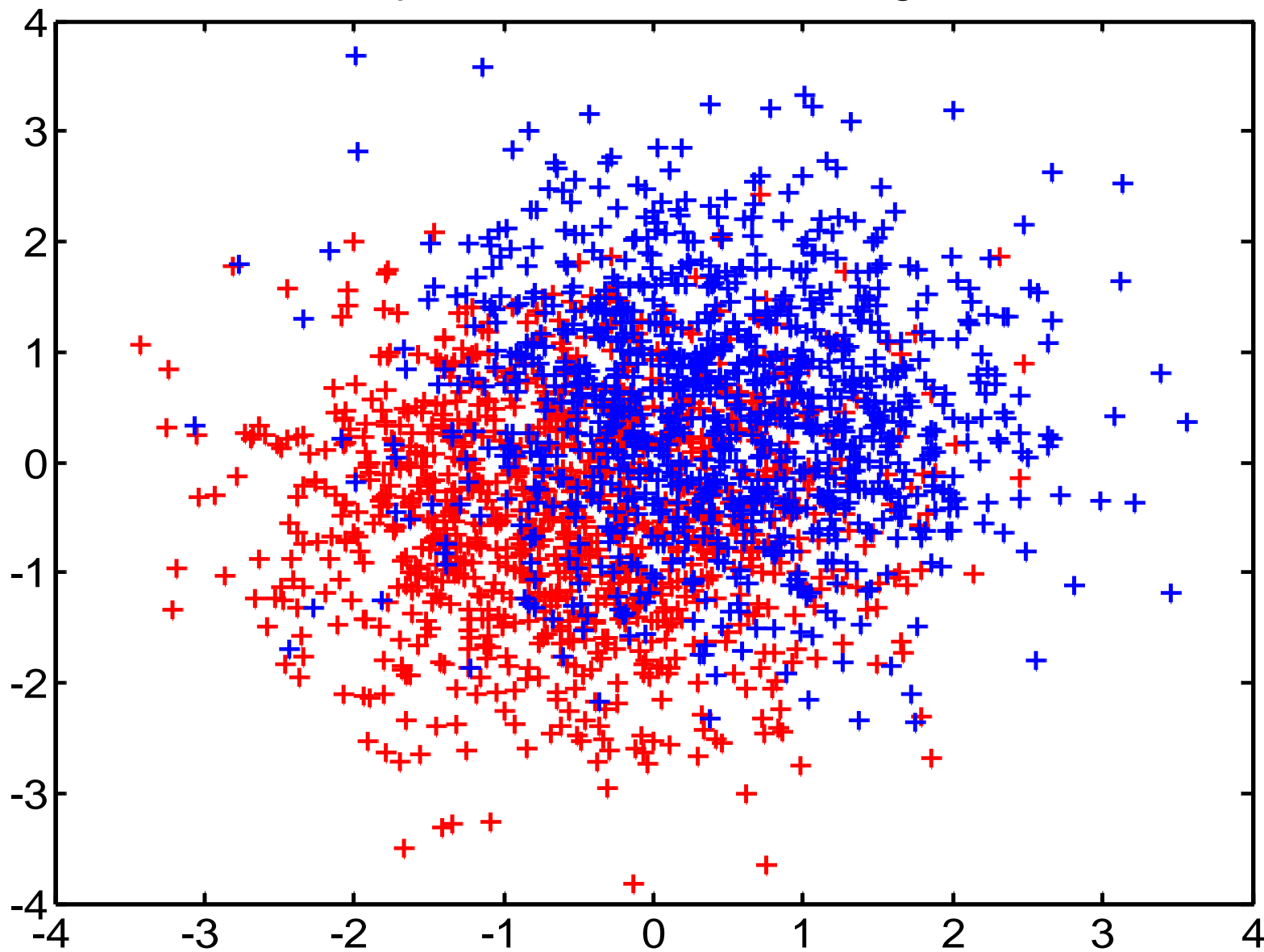
# Gaussian in high dimensions



# Two Gaussians

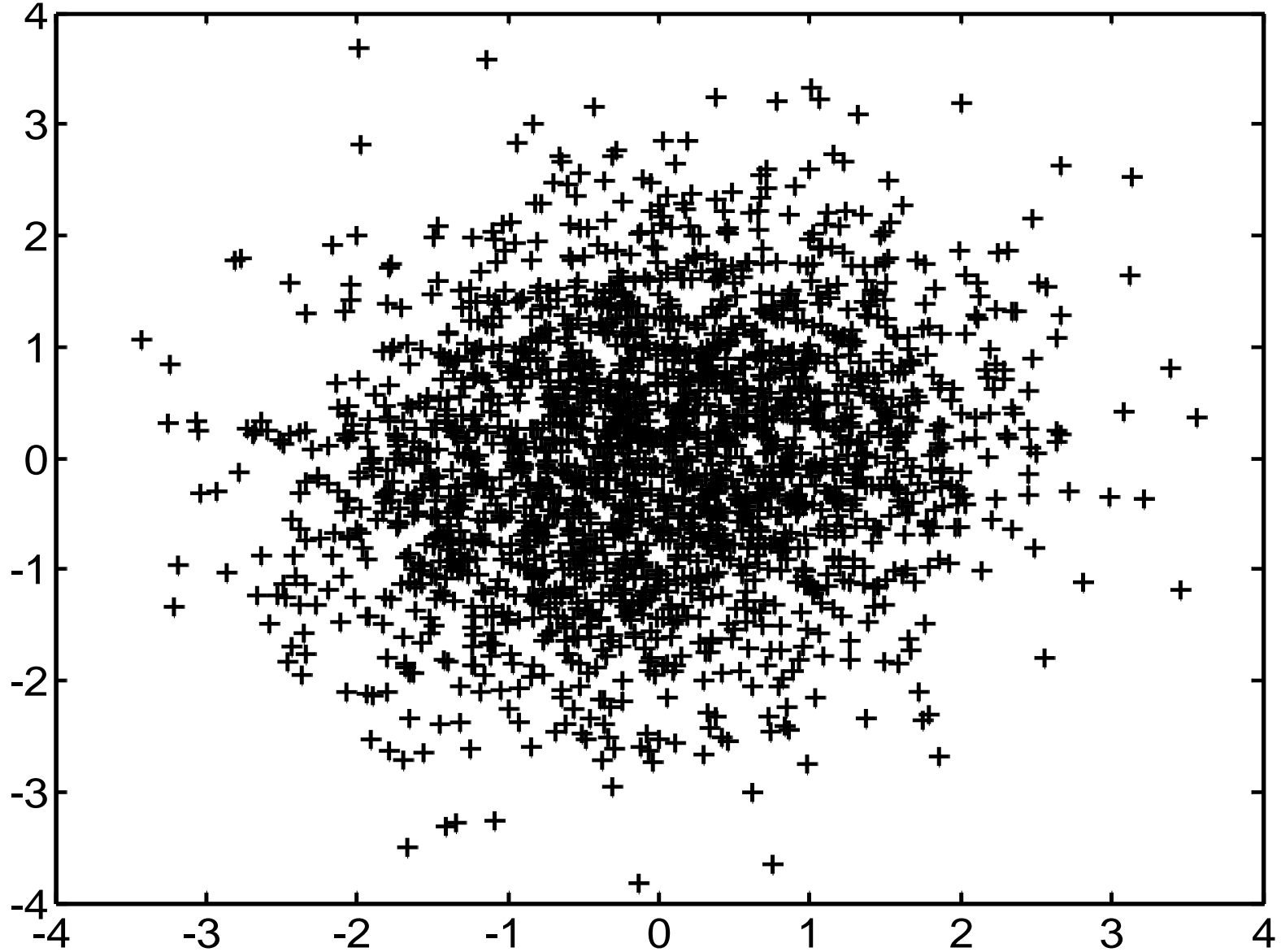


+ 2 Gaussians with 1000 points each:  $\mu=1.000$ ,  $\sigma=2.000$ ,  $\text{dim}=500$



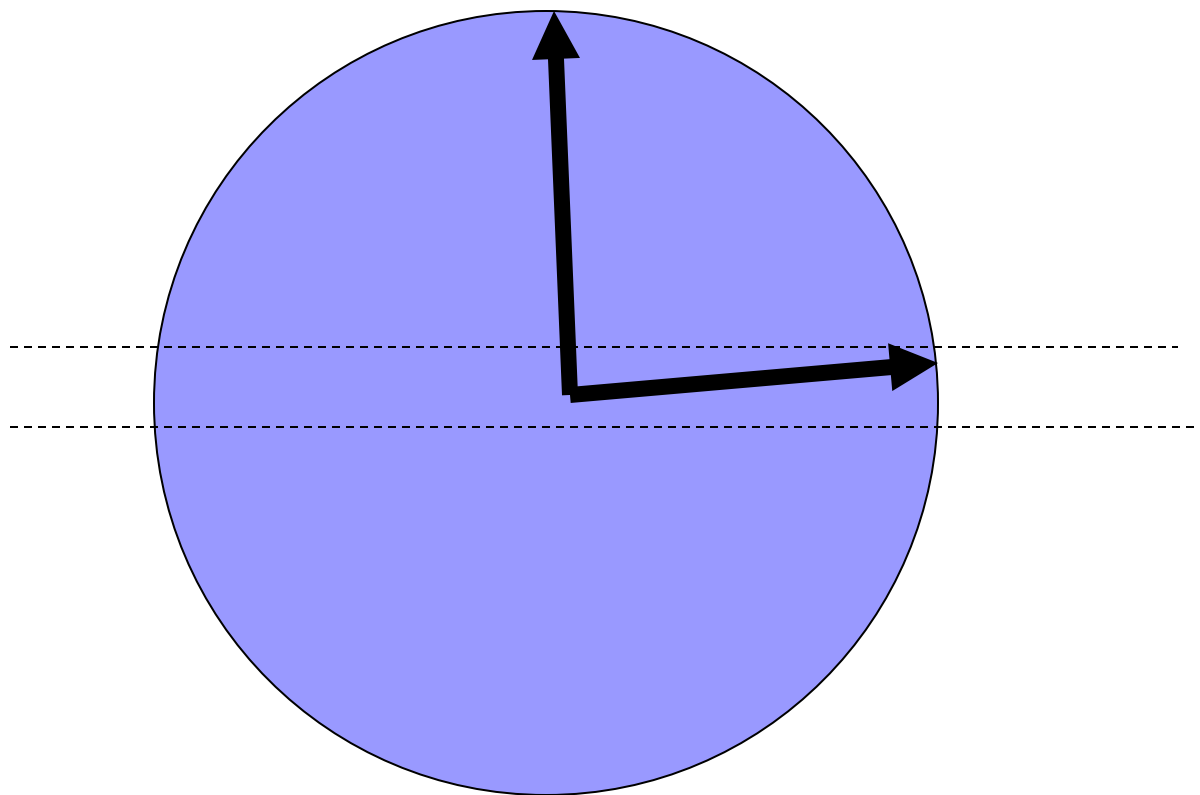


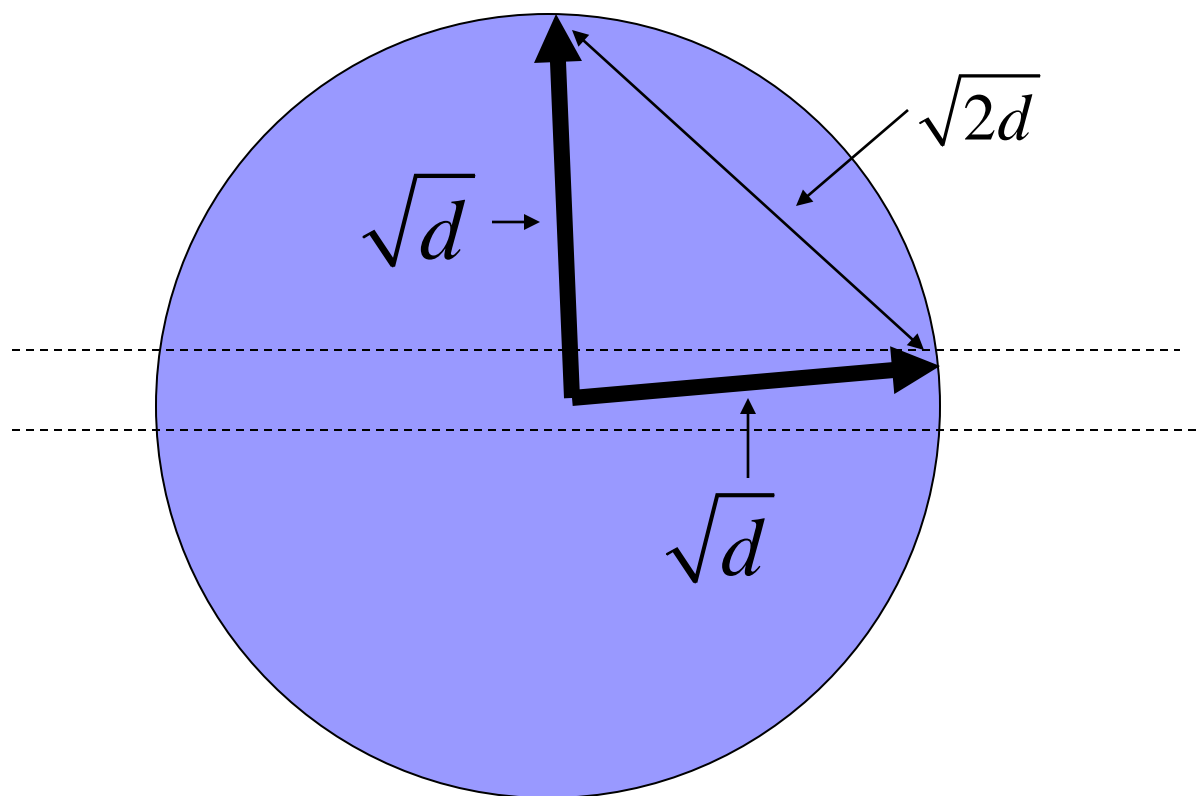
- + 2 Gaussians with 1000 points each:  $\mu=1.000$ ,  $\sigma=2.000$ ,  $\text{dim}=500$



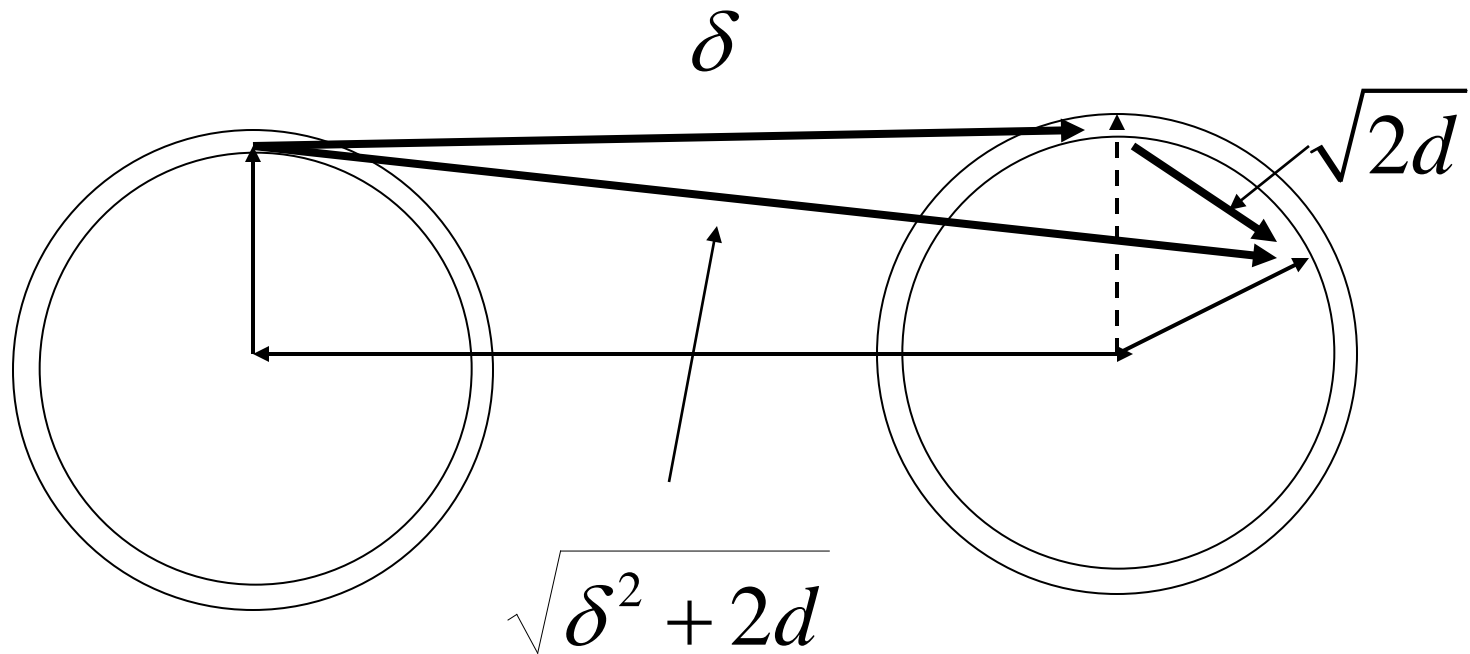
# Distance between two random points from same Gaussian

- Points on thin annulus of radius  $\sqrt{d}$
- Approximate by a sphere of radius  $\sqrt{d}$
- Average distance between two points is  $\sqrt{2d}$   
(Place one point at N. Pole, the other point at random.  
Almost surely, the second point is near the equator.)





# Expected distance between points from two Gaussians separated by $\delta$



# Can separate points from two Gaussians if

$$\sqrt{\delta^2 + 2d} > \sqrt{2d} + \gamma$$

$$\sqrt{2d} \left(1 + \frac{1}{2} \frac{\delta^2}{2d} + \dots\right) > \sqrt{2d} + \gamma$$

$$\frac{1}{2} \frac{\delta^2}{\sqrt{2d}} > \gamma$$

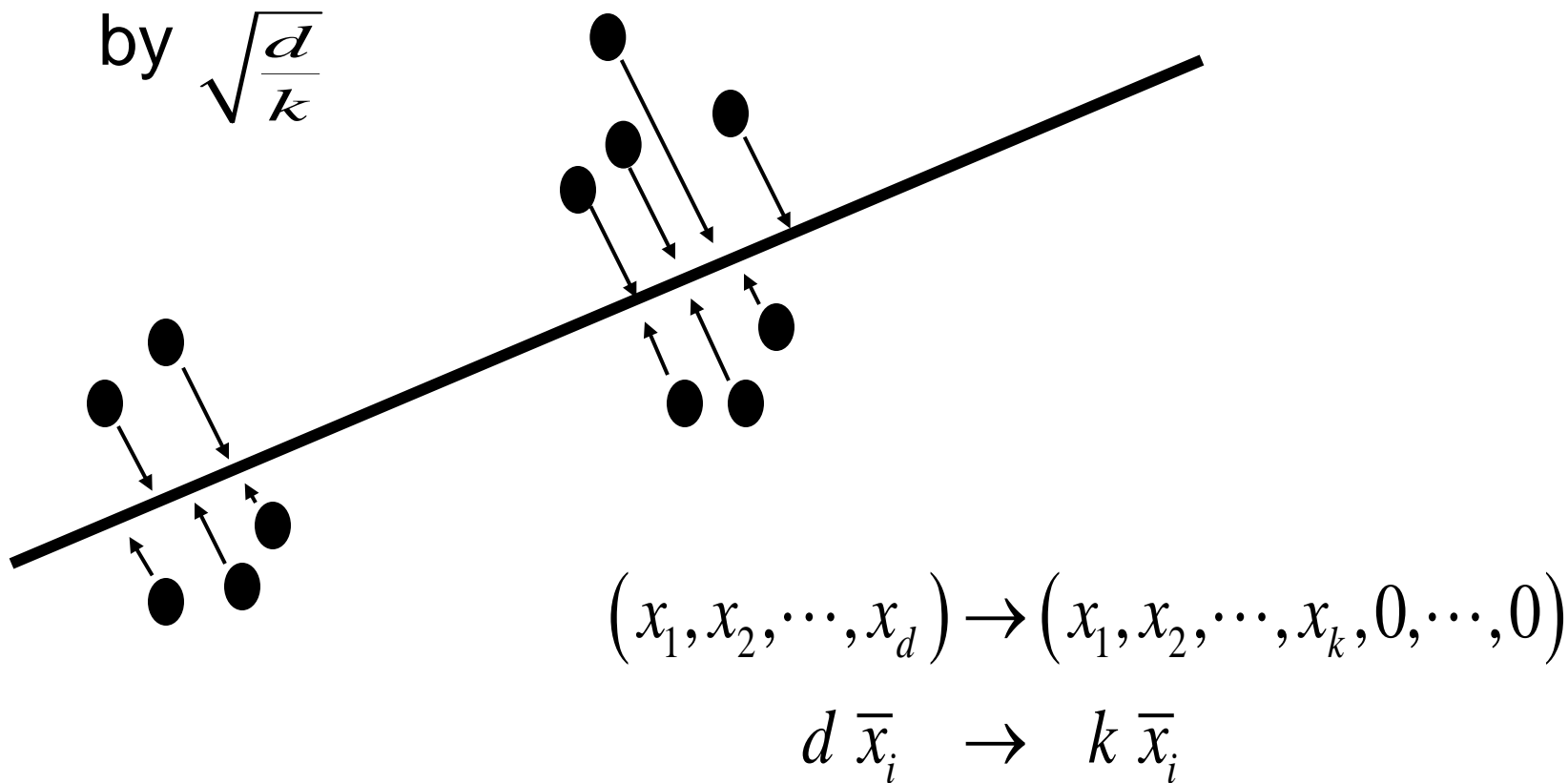
$$\delta > \sqrt{2\gamma} (2d)^{\frac{1}{4}}$$



# Dimension reduction

- Project points onto subspace containing centers of Gaussians
- Reduce dimension from  $d$  to  $k$ , the number of Gaussians

- Centers retain separation
- Average distance between points reduced







# Can separate Gaussians provided

$$\sqrt{\delta^2 + 2k} > \sqrt{2k} + \gamma$$

$\delta$  > some constant involving  $k$  and  $\gamma$   
independent of the dimension

- 
- We have just seen what a science base for high dimensional data might look like.
  - What other areas do we need to develop a science base for?

- 
- Ranking is important
    - Restaurants, movies, books, web pages
    - Multi-billion dollar industry
  - Collaborative filtering
    - When a customer buys a product, what else is he or she likely to buy?
  - Dimension reduction
  - Extracting information from large data sources
  - Social networks



# Time of change

- The information age is a fundamental revolution that is changing all aspects of our lives.
- Those individuals, institutions and nations who recognize this change and position themselves for the future will benefit enormously.



# Conclusions

- We are in an exciting time of change.
- Information technology is a big driver of that change.
- Computer science theory needs to be developed to support this information age.