Lecture Notes
Week 12

INFS3200 Advanced Database Systems

Semester 1, 2021

Advanced Database Applications

Professor Xue Li

The University Of Queensland
AUSTRALIA

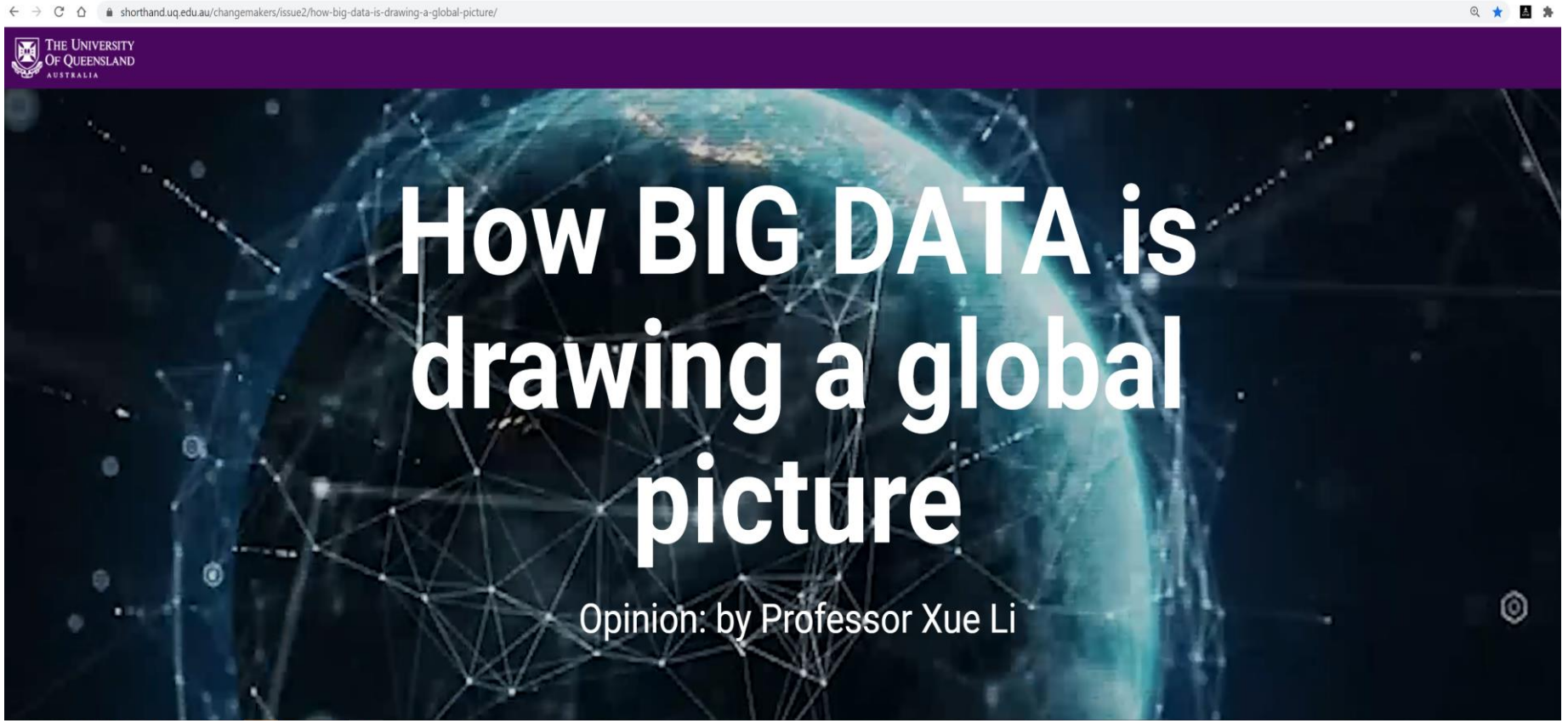
+ Teaching Activities

- Assessments due
  - All Practicals **Prac1, Prac2** and **Prac3** are due this week at **4pm, on Friday 21 May 2021**
  - **Assignment** is due next week **at 4pm, on Friday 28 May 2021**
  - Late submission will involve penalty (see course EPC)
- Next lecture (Week 13) will be **Course review (final exam related)**
- **SECats** course and teaching survey opened
  - Open from Monday 17 May 2021
  - Close at 11:59pm on Friday 04 June 2021
  - Link provided in email
  - Any comment is appreciated
- **SETutor** tutor survey is still running
  - Opened on Monday 03 May 2021
  - Ends on this weekend, at 11:59pm, on Sunday 23 May 2021

+ Outline

- Big Data Applications
  - Connecting Dots: From Small to Big
  - Discovering Specifics: From Big to Small
  - Inferencing: Knowing Unknown
  - Key Values
- Network Science: Data in Complex Graphs
  - The Curse of Dimensionality
  - Scale-Free Networks
  - Different Types of Networks
  - Computing Issues of Complex Networks
- Explaining Outliers in Aggregate Queries
- Effective Storage of Big Data
  - Colum-based vs Row-based Data database storages
  - Applicability of Row and Column Storages
  - Compare and Contrast of Row and Column Storages
  - SQL vs NOSQL

What do we do with Big Data?



https://shorthand.uq.edu.au/changemakers/issue2/how-big-data-is-drawing-a-global-picture/

Big Data is a new challenge to the society as

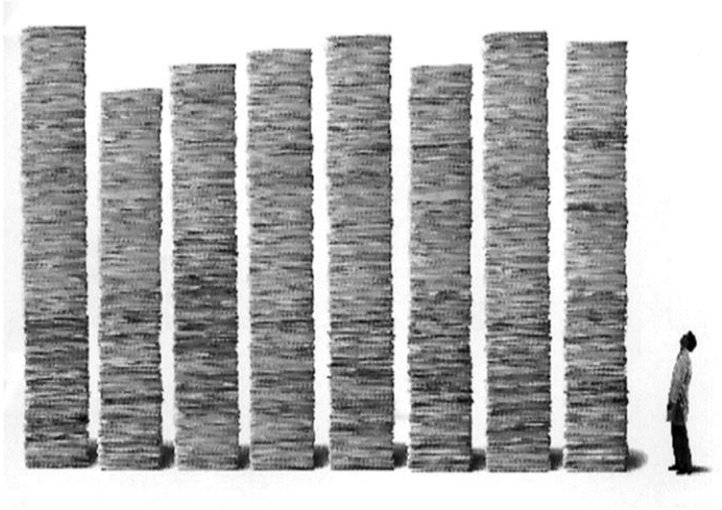
- New technology
- New type of software systems
- New opportunity for …

How do we use data?

				Round 1		Totals					Shooting			Per Game				Advanced			
Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	3P%	FT%	MP	PTS	TRB	AST	WS	WS/48	BPM	VORP
1	1	PHI	Ben Simmons	Louisiana State University																	
2	2	LAL	Brandon Ingram	Duke University	1	79	2279	740	317	166	.402	.294	.621	28.8	9.4	4.0	2.1	0.3	-.007	-3.8	-1.1
3	3	BOS	Jaylen Brown	University of California	1	78	1341	515	220	64	.454	.341	.685	17.2	6.6	2.8	0.8	1.5	.053	-4.0	-0.7
4	4	PHO	Dragan Bender		1	43	574	146	103	23	.354	.277	.364	13.3	3.4	2.4	0.5	0.3	-.029	-4.3	-0.3
5	5	MIN	Kris Dunn	Providence College	1	78	1333	293	166	188	.377	.288	.610	17.1	3.8	2.1	2.4	0.1	.004	-2.2	-0.1
6	6	NOP	Buddy Hield	University of Oklahoma	1	82	1888	866	269	121	.426	.391	.842	23.0	10.6	3.3	1.5	1.3	.032	-2.8	-0.4
7	7	DEN	Jamal Murray	University of Kentucky	1	82	1764	811	214	170	.404	.334	.883	21.5	9.9	2.6	2.1	1.4	.037	-2.6	-0.3
8	8	SAC	Marquese Chriss	University of Washington	1	82	1743	753	348	60	.449	.321	.624	21.3	9.2	4.2	0.7	1.8	.050	-1.6	0.2
9	9	TOR	Jakob Poeltl	University of Utah	1	54	626	165	165	12	.583		.544	11.6	3.1	3.1	0.2	1.6	.125	-0.7	0.2
10	10	MIL	Thon Maker		1	57	562	226	114	23	.459	.378	.653	9.9	4.0	2.0	0.4	1.3	.113	-1.5	0.1
11	11	ORL	Domantas Sabonis	Gonzaga University	1	81	1632	479	287	82	.399	.321	.657	20.1	5.9	3.5	1.0	0.7	.022	-4.9	-1.2
27	27	TOR	Pascal Siakam	New Mexico State University	1	55	859	229	185	17	.502	.143	.688	15.6	4.2	3.4	0.3				
28	28	PHO	Skal Labissiere	University of Kentucky	1	33	612	289	162	27	.537	.375	.703	18.5	8.8	4.9	0.8				
29	29	SAS	Dejounte Murray	University of Washington	1	38	322	130	42	48	.431	.391	.700	8.5	3.4	1.1	1.3				
30	30	GSW	Damian Jones	Vanderbilt University	1	10	85	19	23	0	.500		.300	8.5	1.9	2.3	0.0				
				Round 2		Totals					Shooting			Per Game							
Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	3P%	FT%	MP	PTS	TRB	AST				
31	31	BOS	Deyonta Davis	Michigan State University	1	36	238	58	60	2	.511		.556	6.6	1.6	1.7	0.1				
32	32	LAL	Ivica Zubac		1	38	609	284	159	30	.529	.000	.653	16.0	7.5	4.2	0.8				
33	33	LAC	Cheick Diallo	University of Kansas	1	17	199	87	73	4	.474		.714	11.7	5.1	4.3	0.2				
34	34	PHO	Tyler Ulis	University of Kentucky	1	61	1123	444	95	226	.421	.266	.775	18.4	7.3	1.6	3.7				
35	35	BOS	Rade Zagorac																		
36	36	MIL	Malcolm Brogdon	University of Virginia	1	75	1982	767	213	317	.457	.404	.865	26.4	10.2	2.8	4.2				

How do we make sense of Big Data?

- Big data is to solve three problems:
  - Problem of from-small-to-big (Connecting Dots)
  - Problem of from-big-to-small (Discovering Specifics)
  - Problem of "*knowing unknown*" – everything is related to everything else (Inferencing)



Connecting Dots: Big data is about the problem of "from small-to-big"

Given a piece of data, how do we find the ones that are relevant to it?

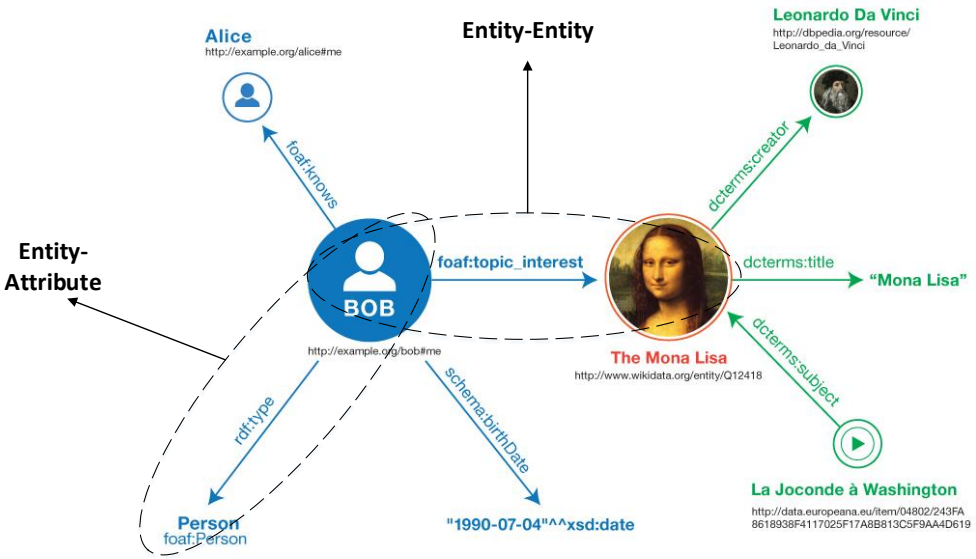
Aristotle: The whole is more than the sum of its parts.

Key Value – The smallest semantic unit of Big Data

- [key, value] is a minimum semantic unit representing an elementary fact.
- Every [key value] is unique with a timestamp.
- Key values are associated to each other.
- All key values can be drawn as a grah.

RDF triples:

Bob	type	Person.
bob	knows	alice.
bob	birthDate	"1990-07-04".
bob	topic_interest	The Mona Lisa.
The Mona Lisa	creator	Leonardo_da_Vinci .
La Joncode a Wa.	subject	The Mona Lisa.

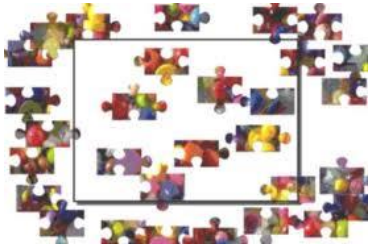


https://www.w3.org/TR/rdf11-primer/

+ Key-Value Stores

- A simple user interface
  - Data model: (key, value) pairs
  - Operations: Insert(key, value), Fetch(key),
  - Update(key), Delete(key)
  - Some allow (non-uniform) columns within value
  - Some allow Fetch on range of keys
- Example systems
  - Google BigTable, Amazon Dynamo, Cassandra, Voldemort, HBase, …

From-small-to-big: Key Values in Unstructured Data



If every piece of value is represented by a unique key, all values can be connected together.

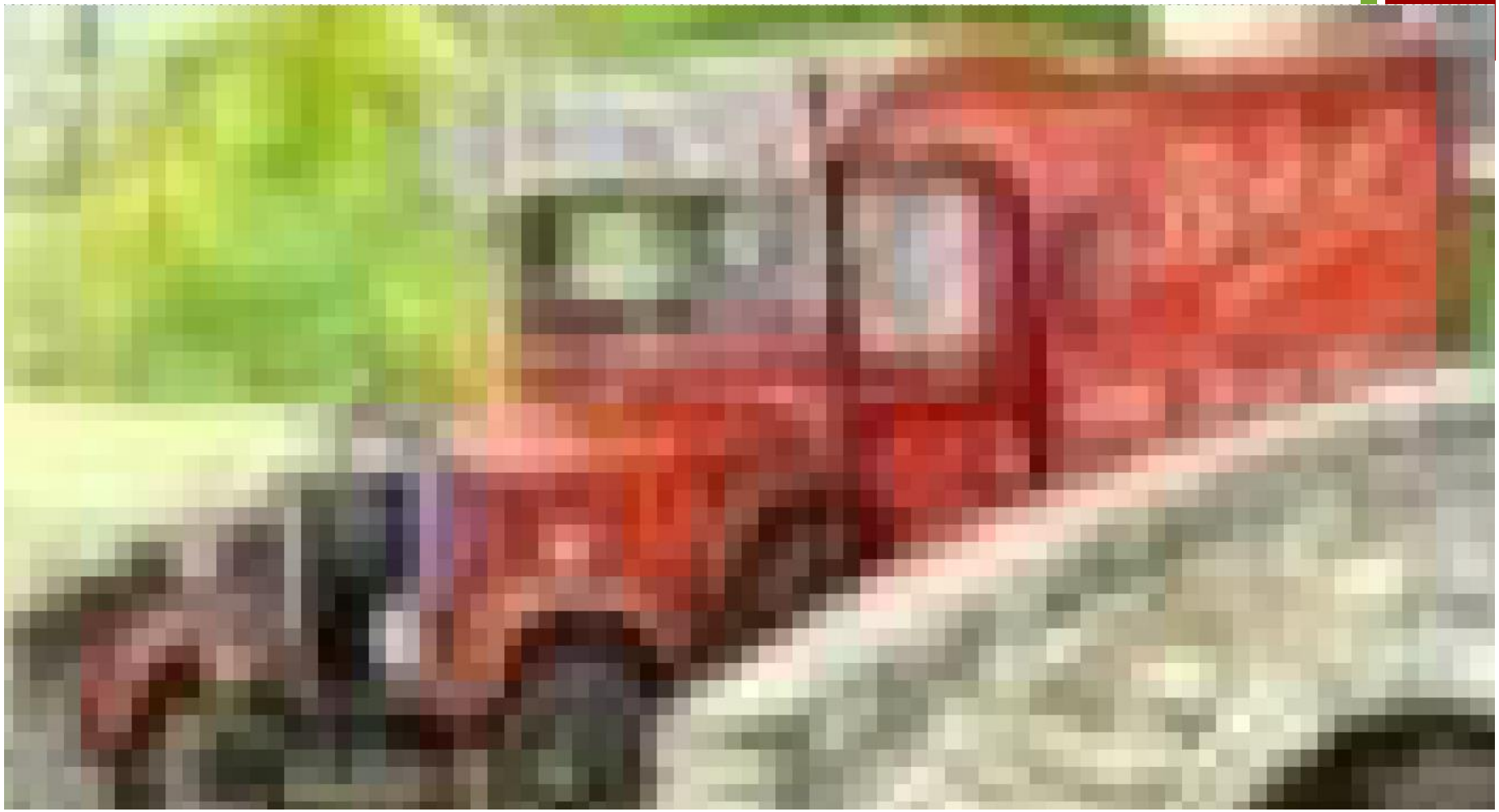
What do you get in pixels of a picture?

00101001	10101001	11100011	00011110	11100001	10001111	01010111	01010001	11100011	10101001
01010111	01010001	11100011	01010111	01010111	01010001	11100011	01010111	00010001	01010001
00001110	00101000	10010010	00001110	11100011	10100011	00010101	00001110	10110000	00101000
11001100	10100011	00110011	01010101	11100101	01010111	01010111	11100011	10100011	10100011
11100011	10100011	11100000	10101001	11010101	00001110	10101001	01010001	00010101	01010111
00110011	01001100	11100010	01010001	11100011	10100011	01010001	00101000	11000011	00001110
01010111	01010001	11100011	00101000	11100011	10100011	00101000	10100011	01010111	10101010
01010111	01010001	11100011	10100011	11100011	10100011	10100011	11011011	00001110	00101011

What can you see?

What can you really see?

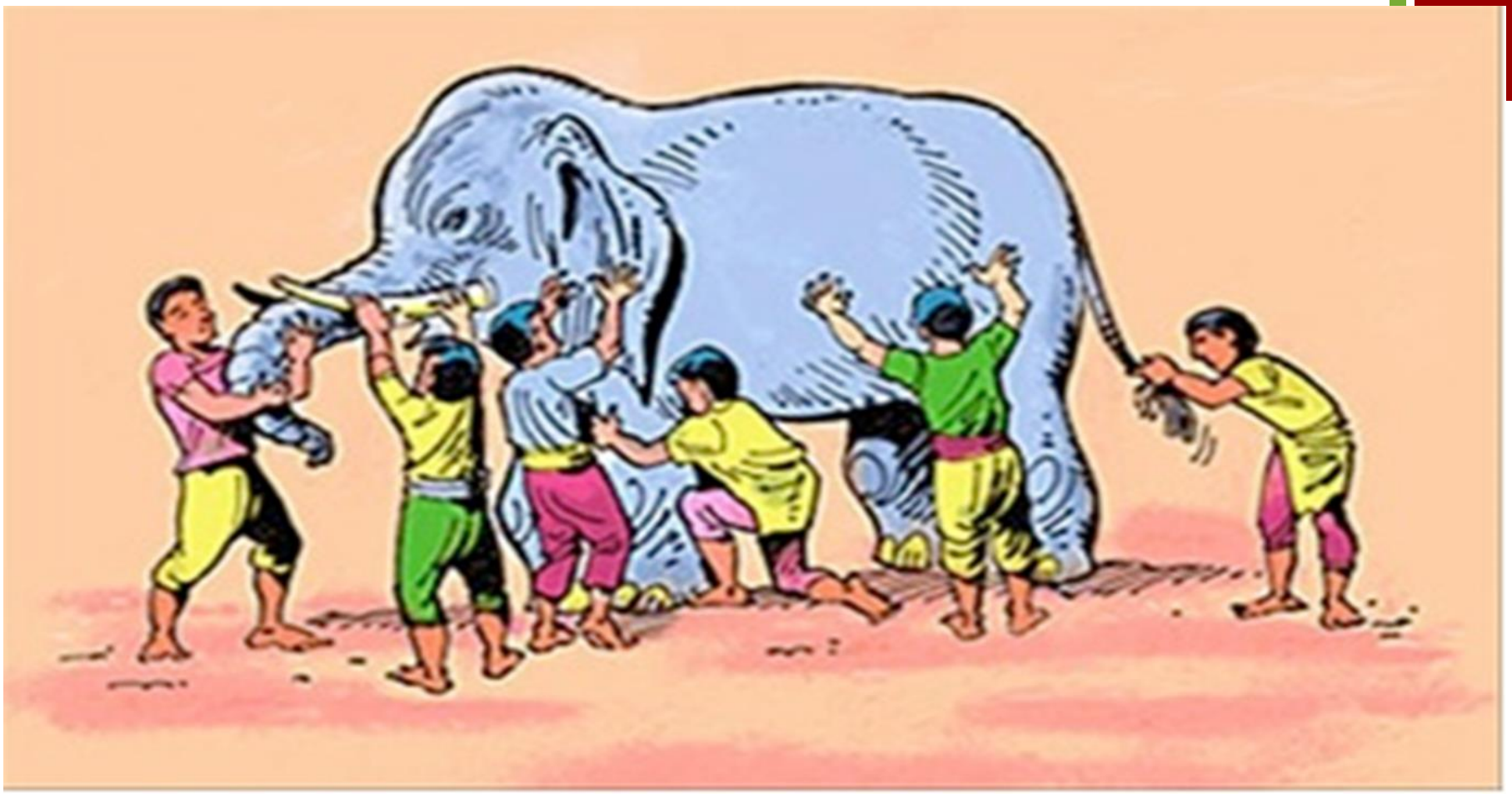
What can you really see?



From-small-to-big: Key Values in Unstructured Data

http://www.jigsawsuk.co.uk/product/shop-by-theme/house-of-puzzles-river-cottage-ray-cresswell-250piece/

People can only see what they want to see



Perspective changes opinions!

http://mytrailthoughts.blogspot.com.au/2010/06/six-blind-men-and-elephant.html

What is this?

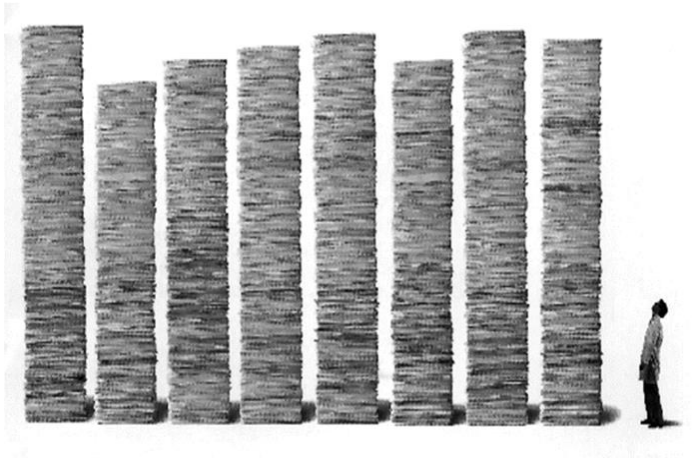
What is this?

What is this?

Data can only make sense if it is in a context.

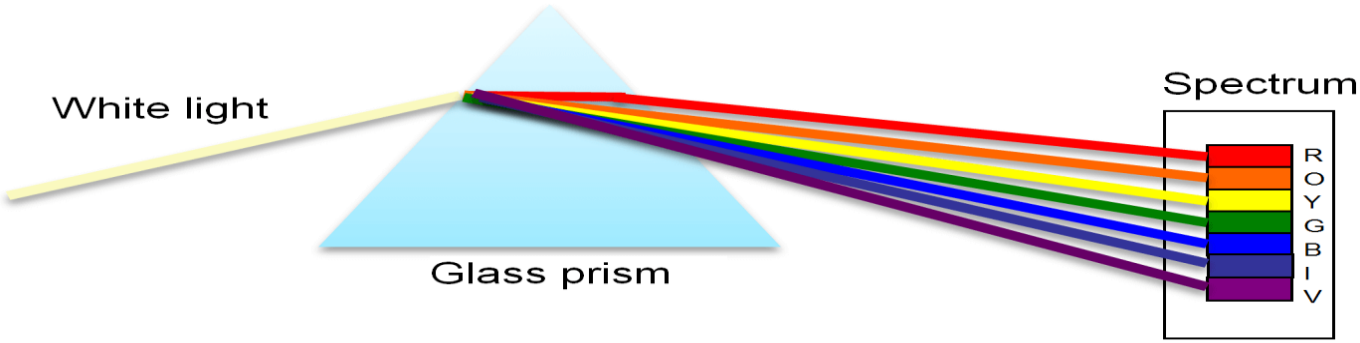
How do we make sense of Big Data?

- Big data is to solve three problems:
  - Problem of from-small-to-big (Connecting Dots)
  - **Problem of from-big-to-small (Discovering Specifics)**
  - Problem of "*knowing unknown*" – everything is related to everything else (Inferencing)



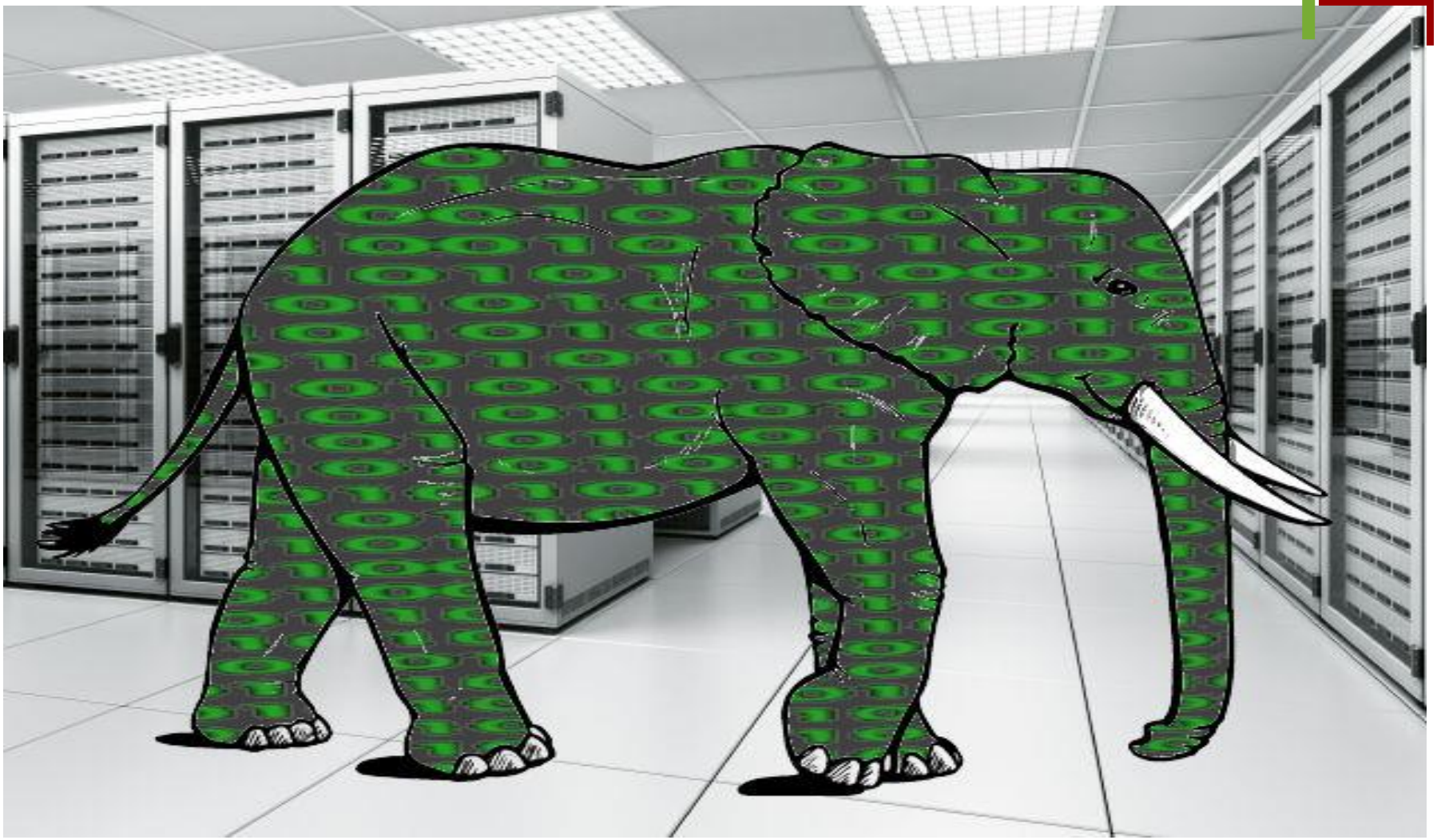
Discovering Specifics: Big data is about the problem of "from-big-to-small"

Given a data set, how do we find **outliers**, predict **trends**, provide **summaries**, and give **explanations**?



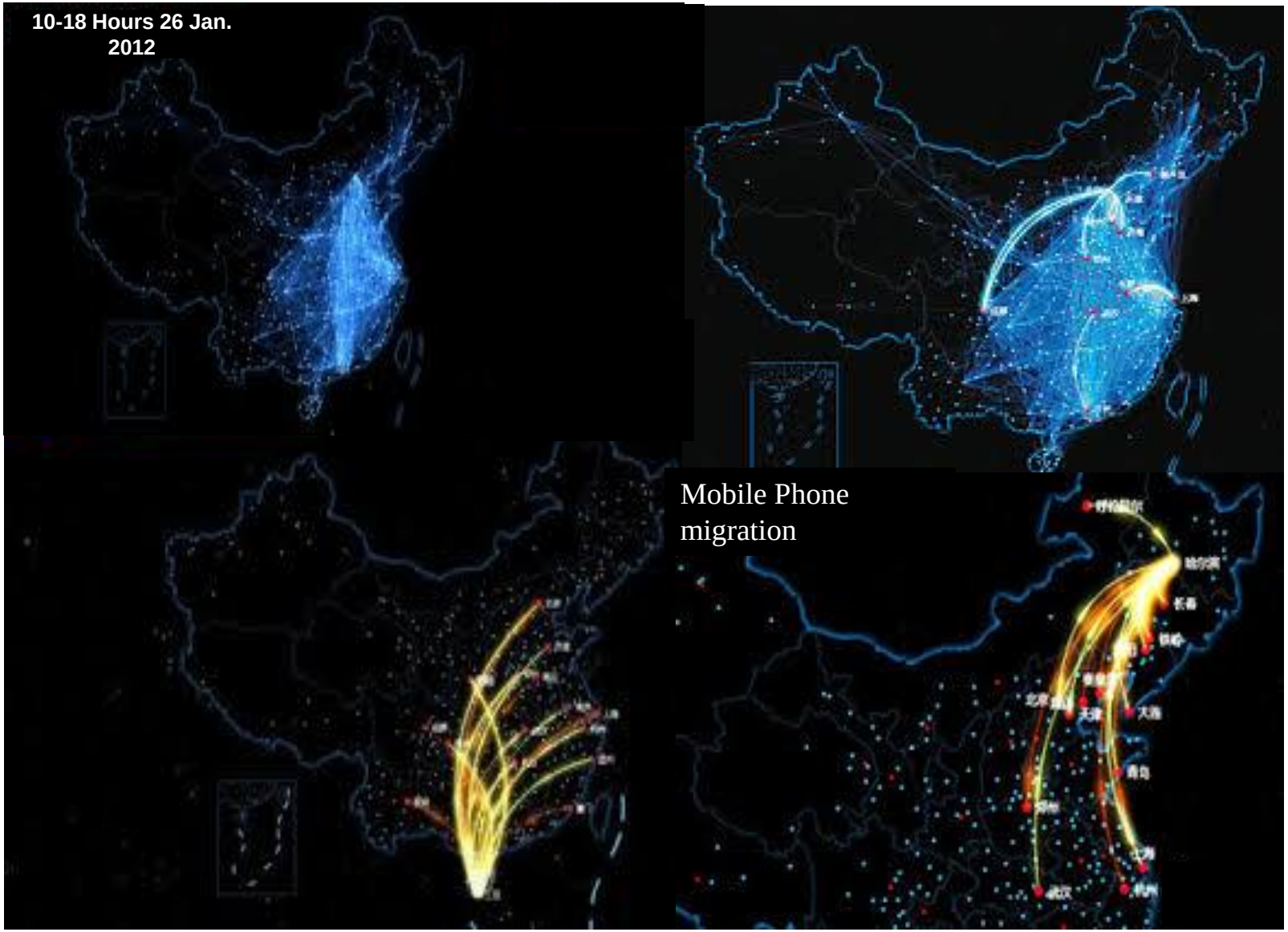
Picture from : http://www.funscience.in/study-zone/Physics/RefractionOfLight/DispersionOfLight.php

From Big-to-small: see details from a global picture



http://www.business2community.com/marketing/customer-data-management-with-big-data-part-1-0104110

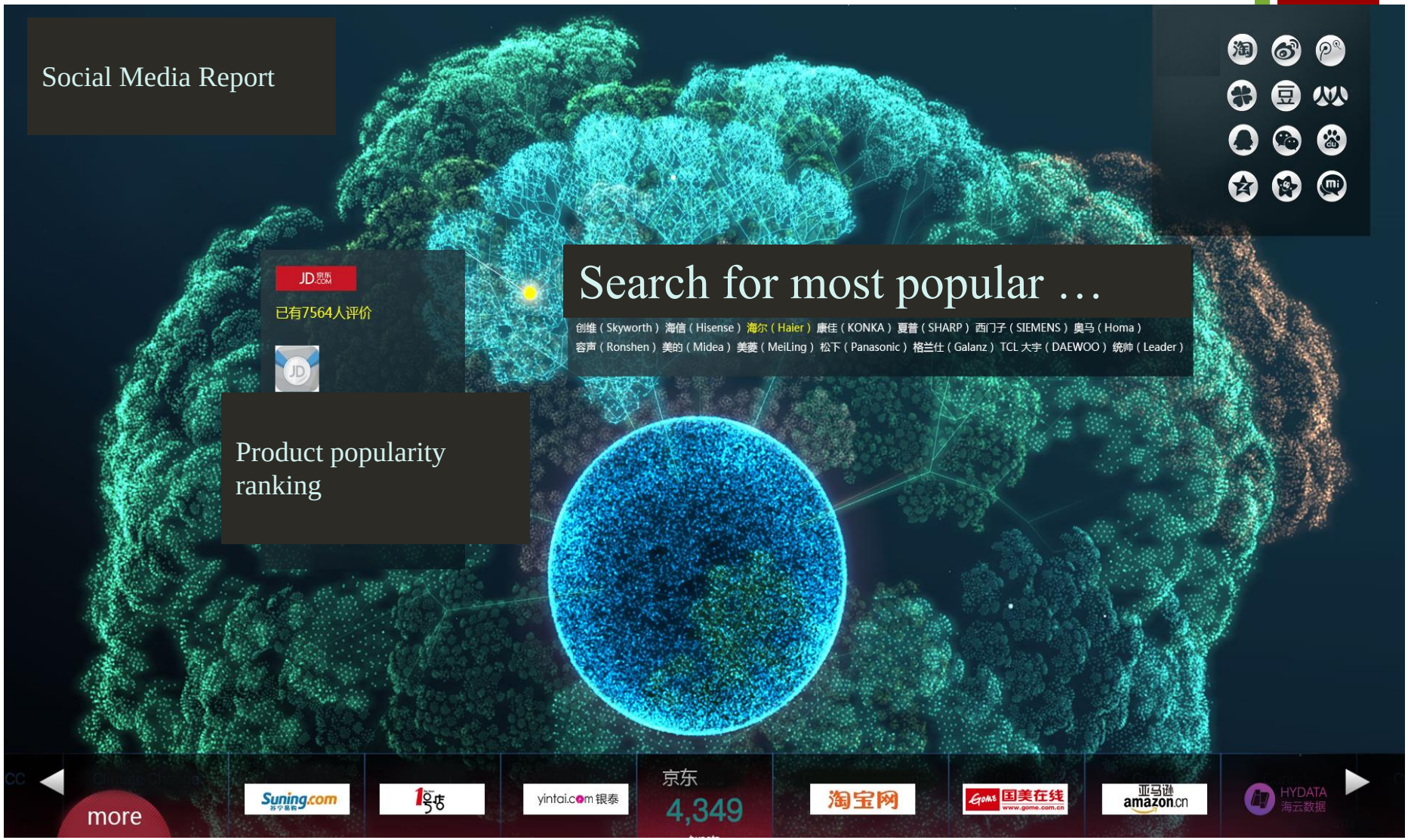
From Big-to-small: Big Data "People travelling"



From Big-to-small: Big Data of China Southern Airlines

Slides copied from HYDATA

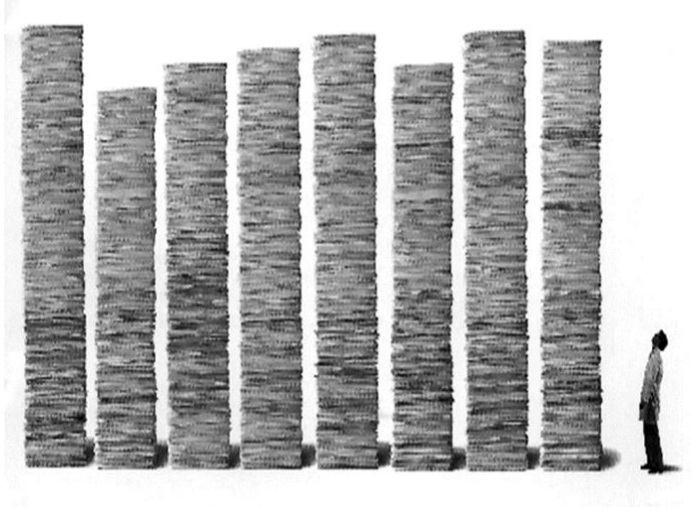
From Big-to-small: Big Data on Social Media Opinions



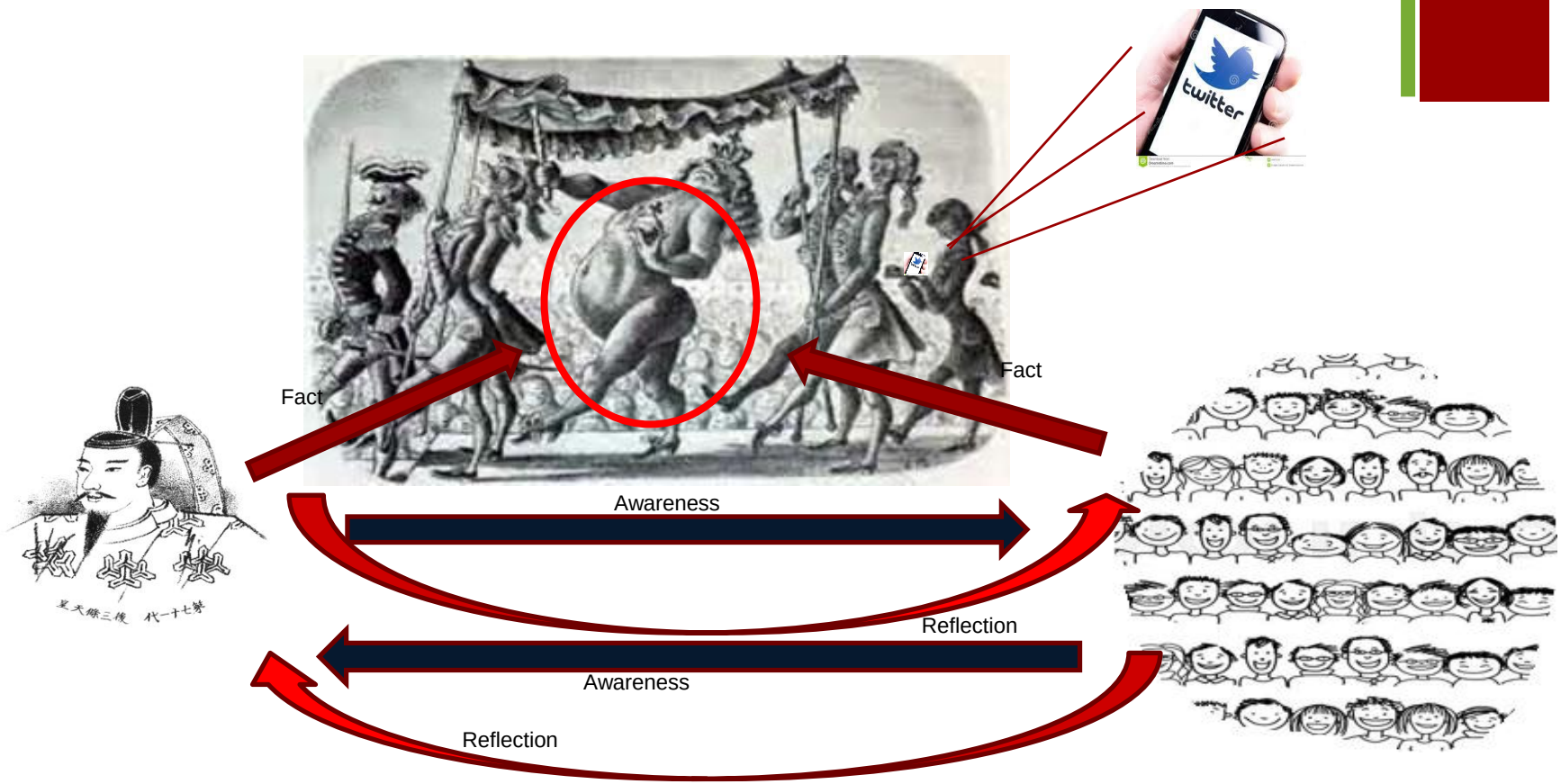
Slides copied from HYDATA

How do we make sense of Big Data?

- Big data is to solve three problems:
  - Problem of from-small-to-big (Connecting Dots)
  - Problem of from-big-to-small (Discovering Specifics)
  - Problem of "*knowing unknown*" – everything is related to everything else (Inferencing)



Big data enables us to "know unknowns"

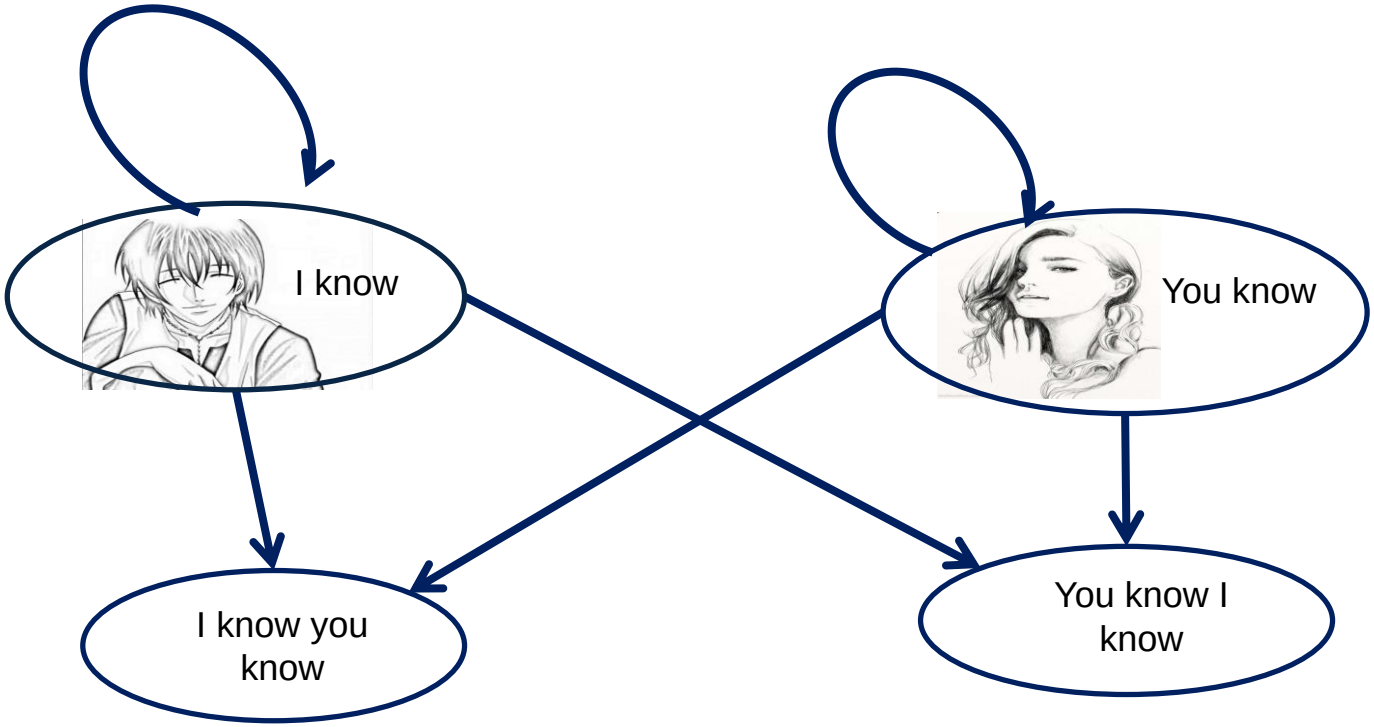


Information Island is caused by information blocking

https://www.colourbox.com/vector/group-of-people-sketch-for-your-design-vector-12724509
https://www.phrases.org.uk/meanings/the-emperors-new-clothes.html

https://commons.wikimedia.org/wiki/File:Emperor_Gosanj%C5%8D.jpg

Information Prosperity by Big Data



1. Fact: Does the Emperor know he wares nothing?
2. Fact: Do people know the Emperor wares nothing?
3. Awareness: Does the Emperor know people know he wares nothing?
4. Awareness: Do people know the Emperor knows he wares nothing?
5. Reflection: Does the Emperor know people know he knows he wares nothing?
6. Reflection: Do people know the emperor know people know he wares nothing?

Imbalanced information is the source of all problems!

Big Data can solve a big problem: Knowing unknowns

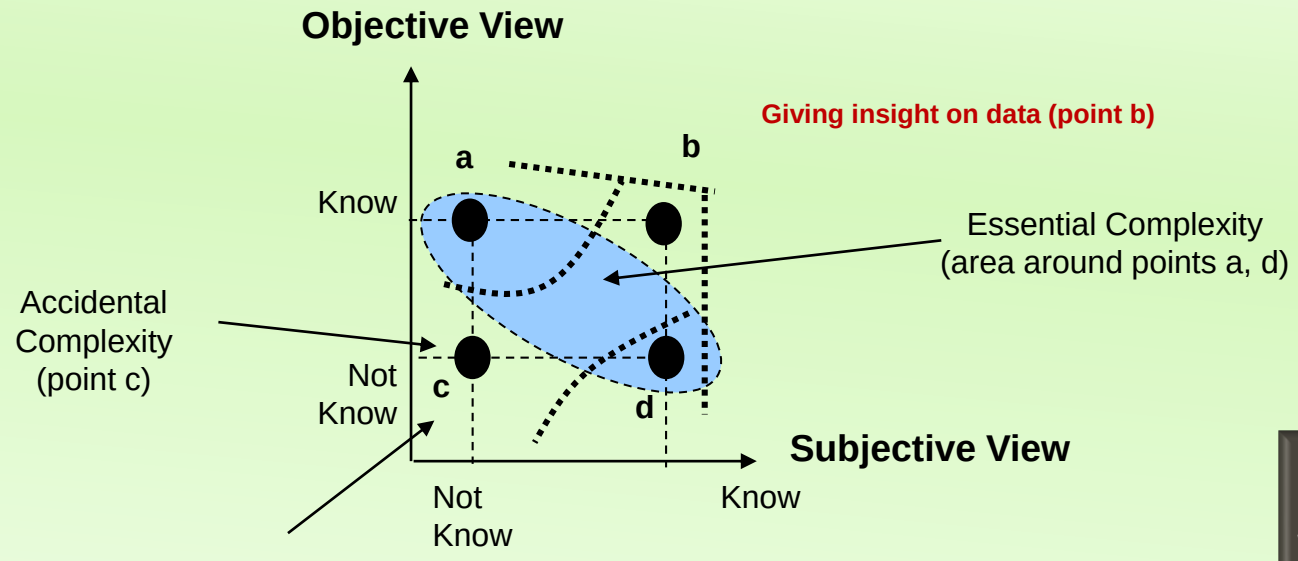
http://www.dianying.com/ft/title/xls1978

What is the problem of Xiangling ?

What is the problem of 911 event ?

https://sites.google.com/a/stonybrook.edu/lu-xun1/literary-works-1

Big data is to solve a big problem:
We don't know what we don't know

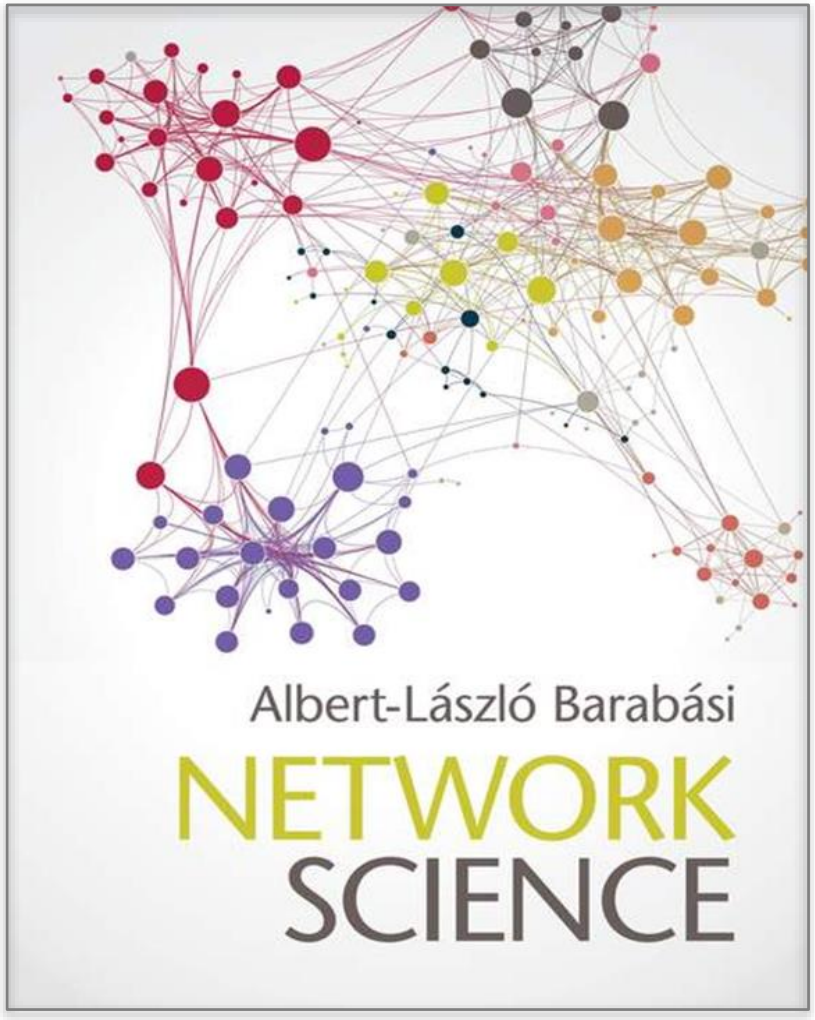


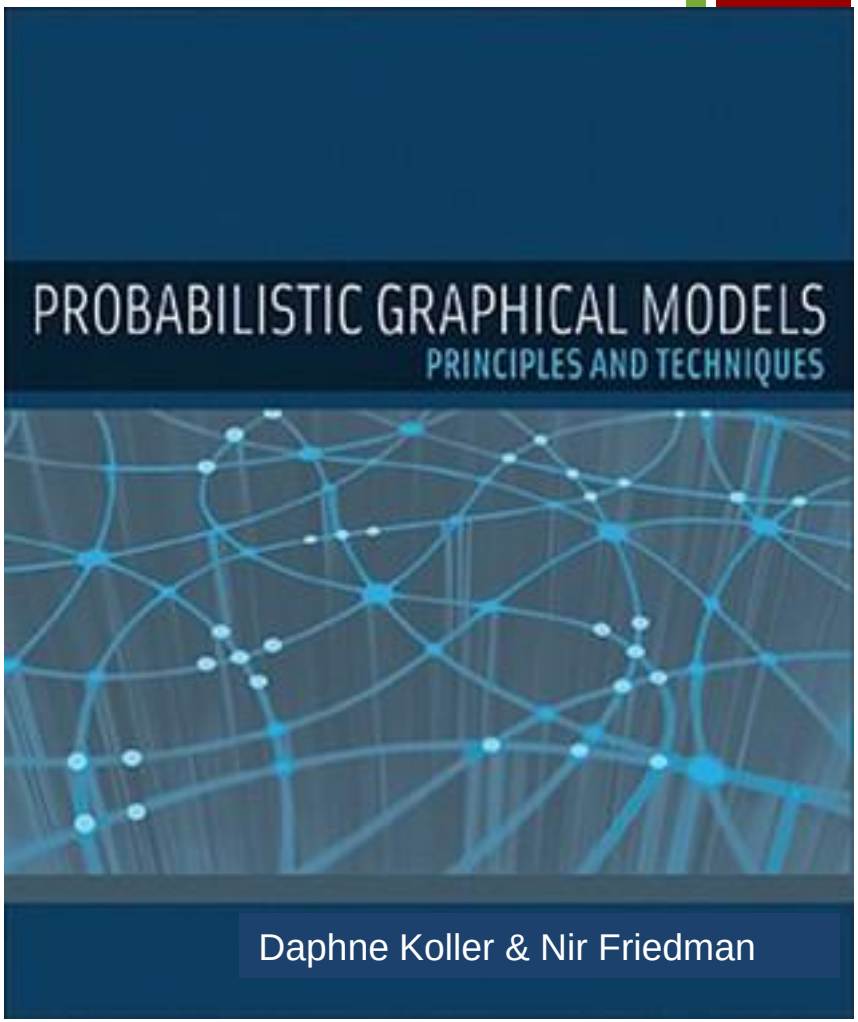
Difficulties of Data Processing (area around point c)

You know what you know and you know what you don't know – You know

Xue.Li, (2003) Learning Complexity. (http://community.flexiblelearning.net.au/TeachingTrainingLearners/content/article_3383.htm

What is the Science of Big Data?



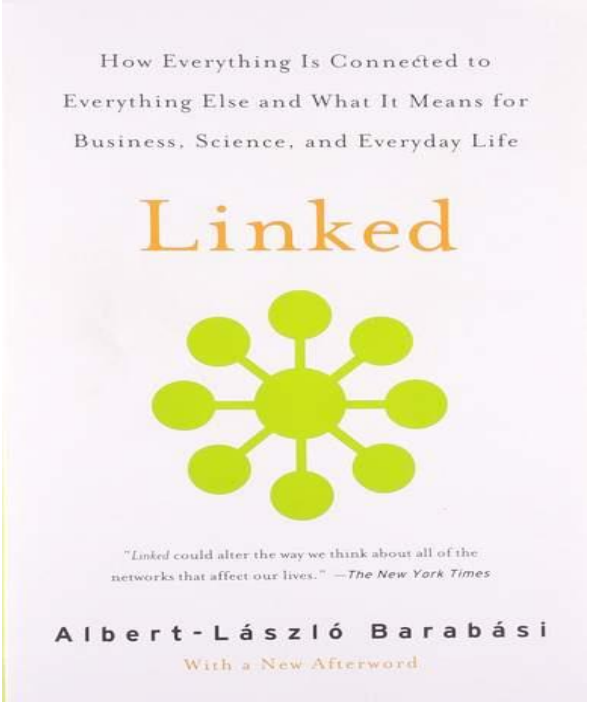


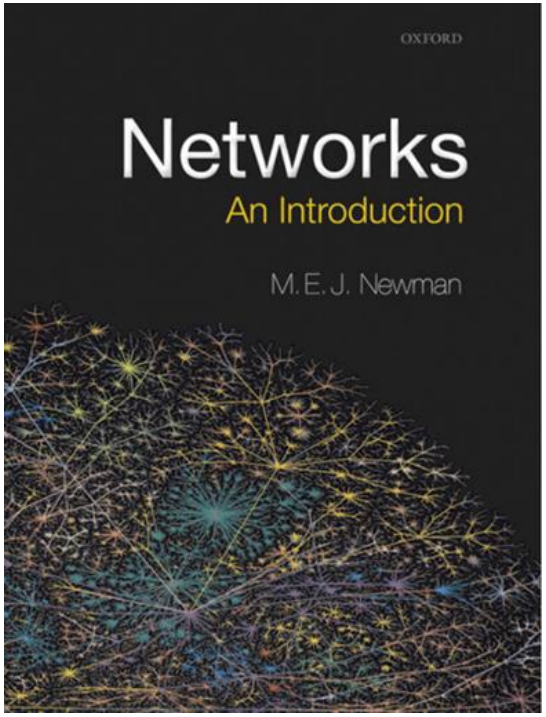
http://barabasi.com/book/network-science

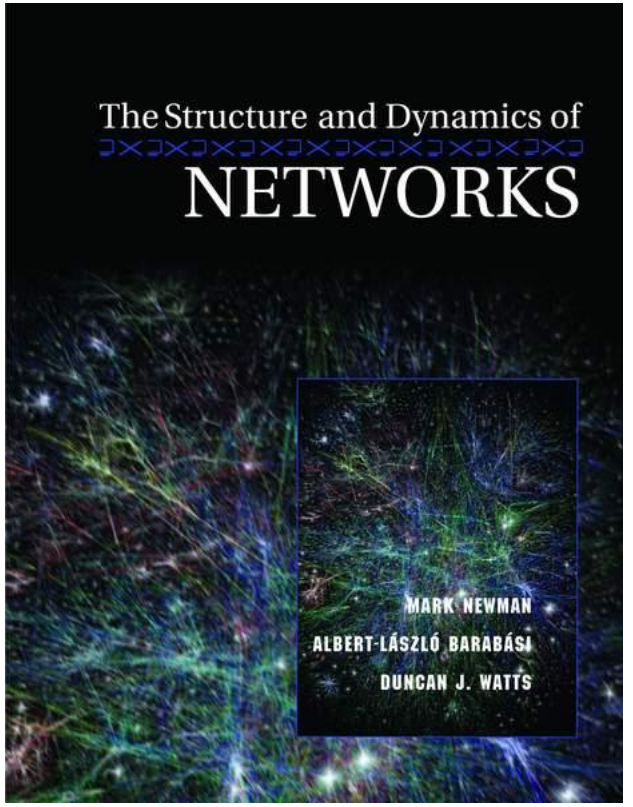
https://www.goodreads.com/book/show/6676555-probabilistic-graphical-models

Objectives of Network Science

- **Ideas** of Data Science are used in Network Science
- **Concepts** in Information Networks are in Network Science
- **Algorithms** on Computing Complex Networks
- **Applicability** of Network Algorithms







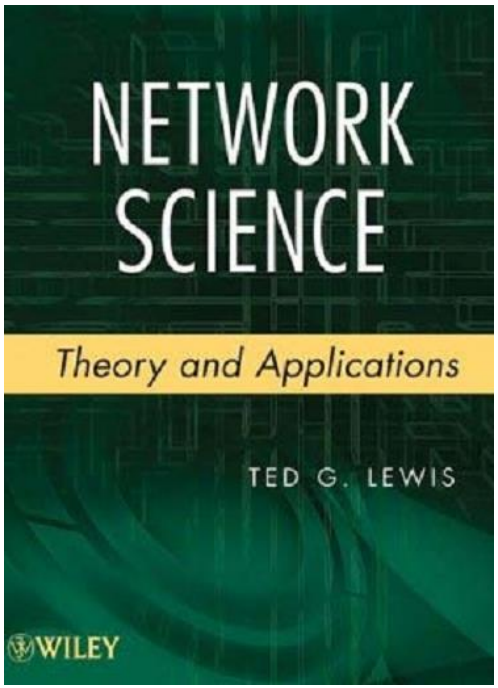
Managing Big Data: Information Networks in Network Science

- Information Network is a representation of physical, biological, and social phenomena leading to predictive models of these phenomena.
- Information Network is used to improve the predictability of the engineering design of **complex networks**.

Many scientifically important problems can be represented and empirically studied using networks. For example, biological and social patterns, the World Wide Web, metabolic networks, food webs, neural networks and pathological networks are real world problems that can be mathematically represented and topologically studied to reveal some unexpected structural features.

https://en.wikipedia.org/wiki/Modularity_(networks)

+ References in Network Science



SIAM REVIEW
Vol. 45, No. 2, pp. 167–256

The Structure and Function of Complex Networks*

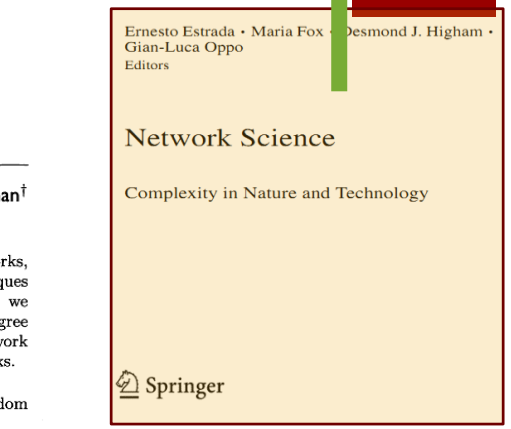
M. E. J. Newman†

Abstract. Inspired by empirical studies of networked systems such as the Internet, social networks, and biological networks, researchers have in recent years developed a variety of techniques and models to help us understand or predict the behavior of these systems. Here we review developments in this field, including such concepts as the small-world effect, degree distributions, clustering, network correlations, random graph models, models of network growth and preferential attachment, and dynamical processes taking place on networks.

Key words. networks, graph theory, complex systems, computer networks, social networks, random graphs, percolation theory

AMS subject classifications. 05C75, 05C90, 94C15

PII. S0036144503424804



Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physics Reports 424 (2006) 175–308

Networks, Crowds, and Markets:
Reasoning about a Highly Connected World

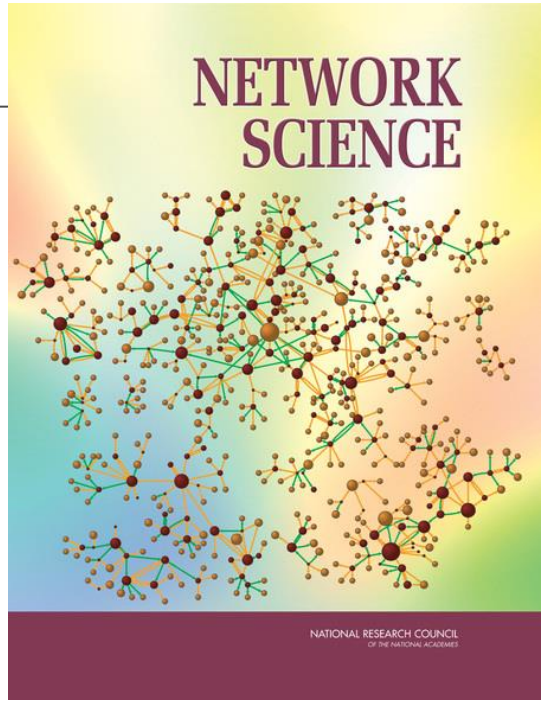
David Easley
Dept. of Economics
Cornell University

Jon Kleinberg
Dept. of Computer Science
Cornell University

Cambridge University Press, 2010
Draft version: June 10, 2010.

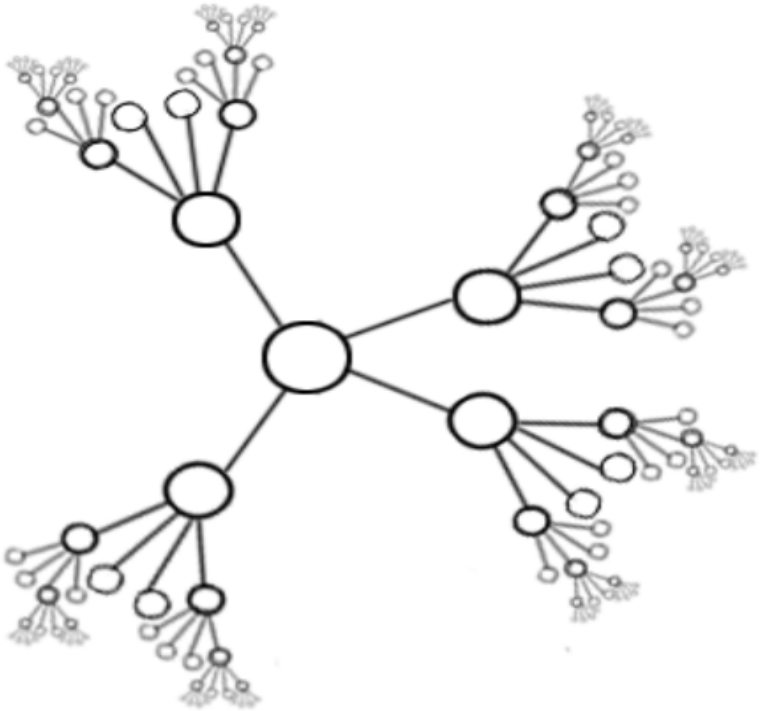
Complex networks: Structure and dynamics

S. Boccaletti[a,*], V. Latora[b,c], Y. Moreno[d,e], M. Chavez[f], D.-U. Hwang[a]



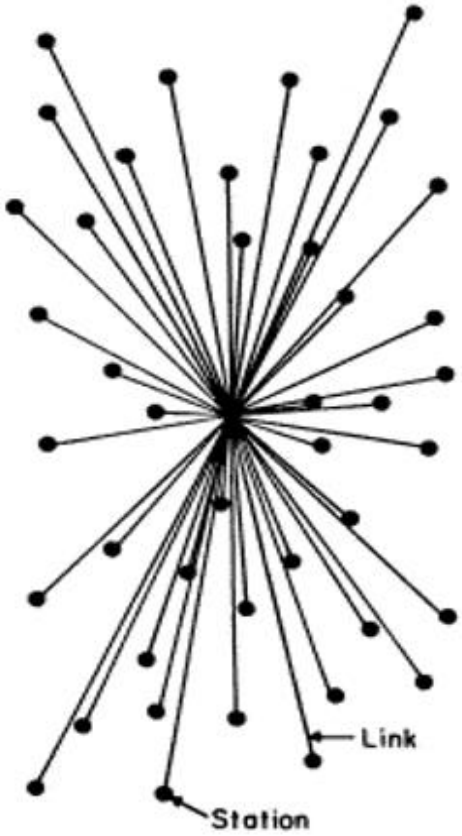
Scale-free Networks

- A network is called **scale-free** if the characteristics of the network are **independent of the size of the network**, i.e. the number of nodes.
- That means that when the network grows, the underlying **structure remains the same**.

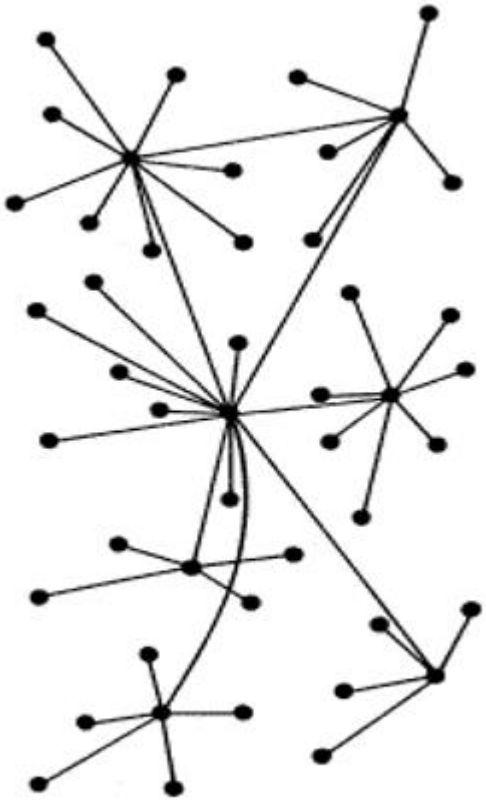


https://www.futurelearn.com/courses/complexity-and-uncertainty/0/steps/1855

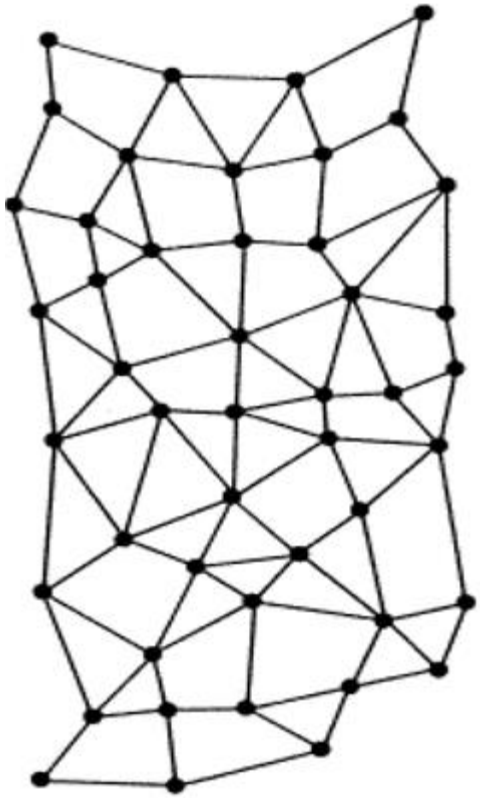
Types of Networks



CENTRALIZED
(A)



DECENTRALIZED
(B)

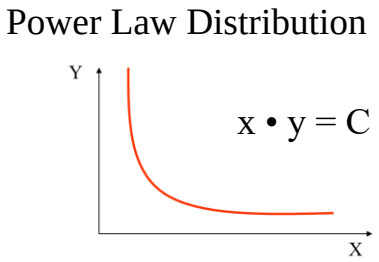


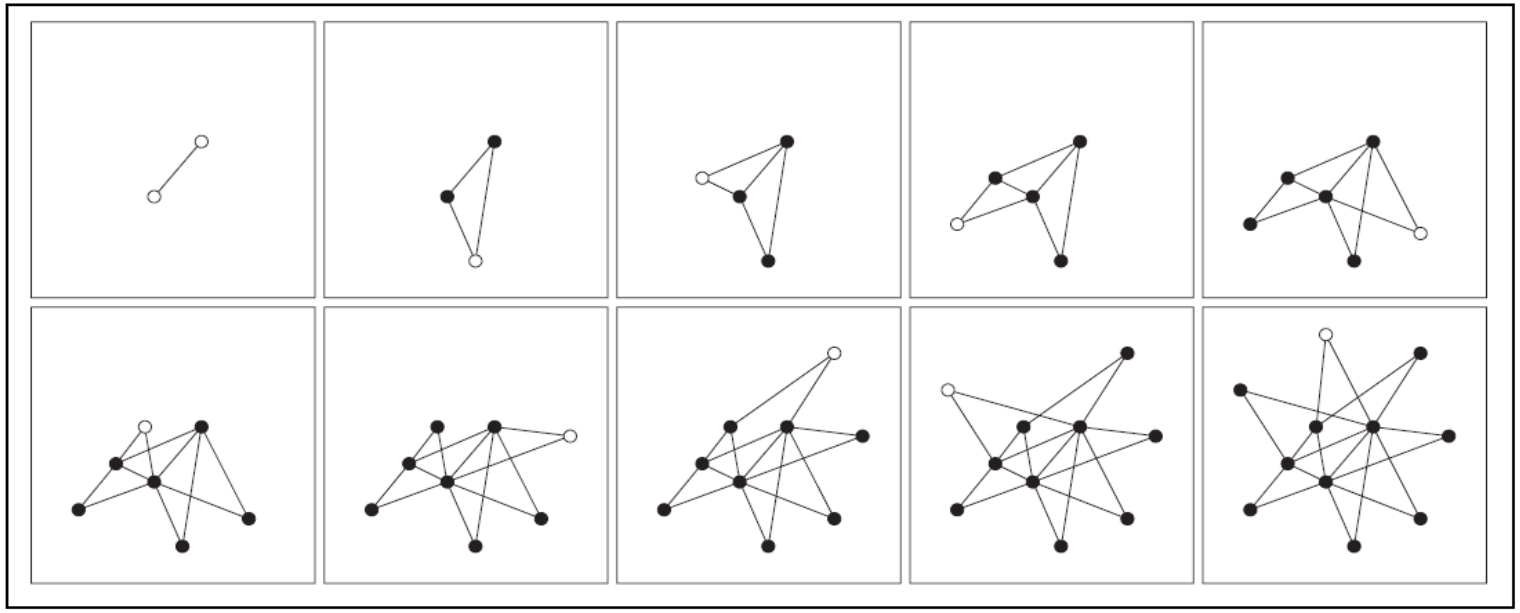
DISTRIBUTED
(C)

http://networkcultures.org/unlikeus/resources/articles/what-is-a-federated-network/

How Scale-Free Networks Generated?

- Why does rich get richer?
- How are the Hub Nodes generated?
- How is Power Law emerged?





Generation of a Scale-Free Network:

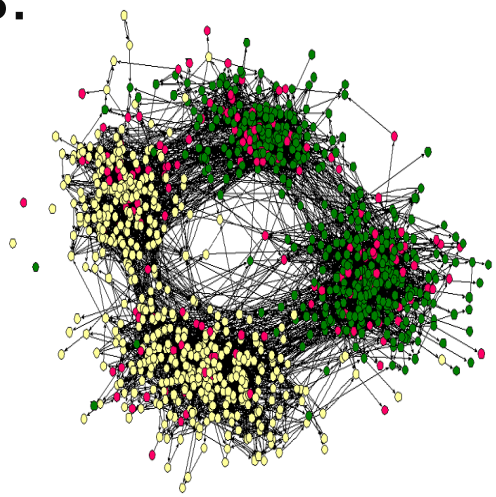
1. **Growth:** The new nodes keep coming
2. **Preferential Attachment**: A new coming point is free to join into the network with other nodes in the network but it inclines to **link with the highly connected** points.

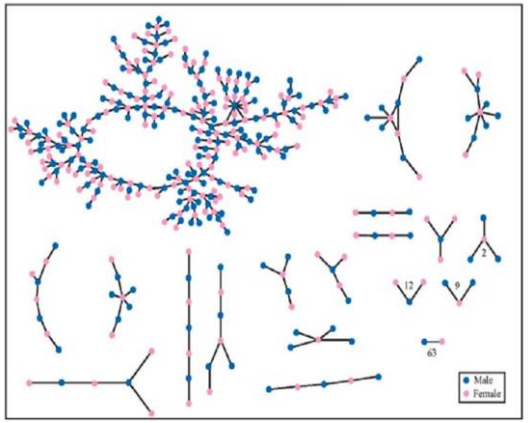
http://barabasi.com/book/the-structure-and-dynamics-of-networks

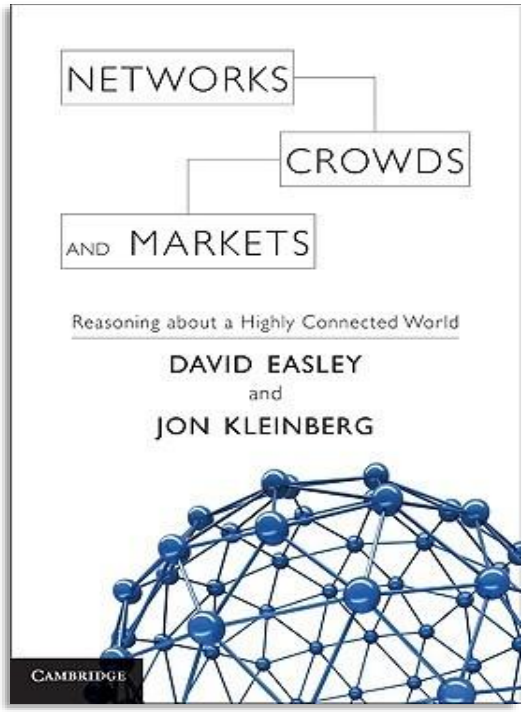
Examples of Data Networks

- Three questions to be asked for the following examples of networks:

1. What type is a given network example?
2. How can this network be created from scratch?
3. What kind of problems can be solved by using this network?

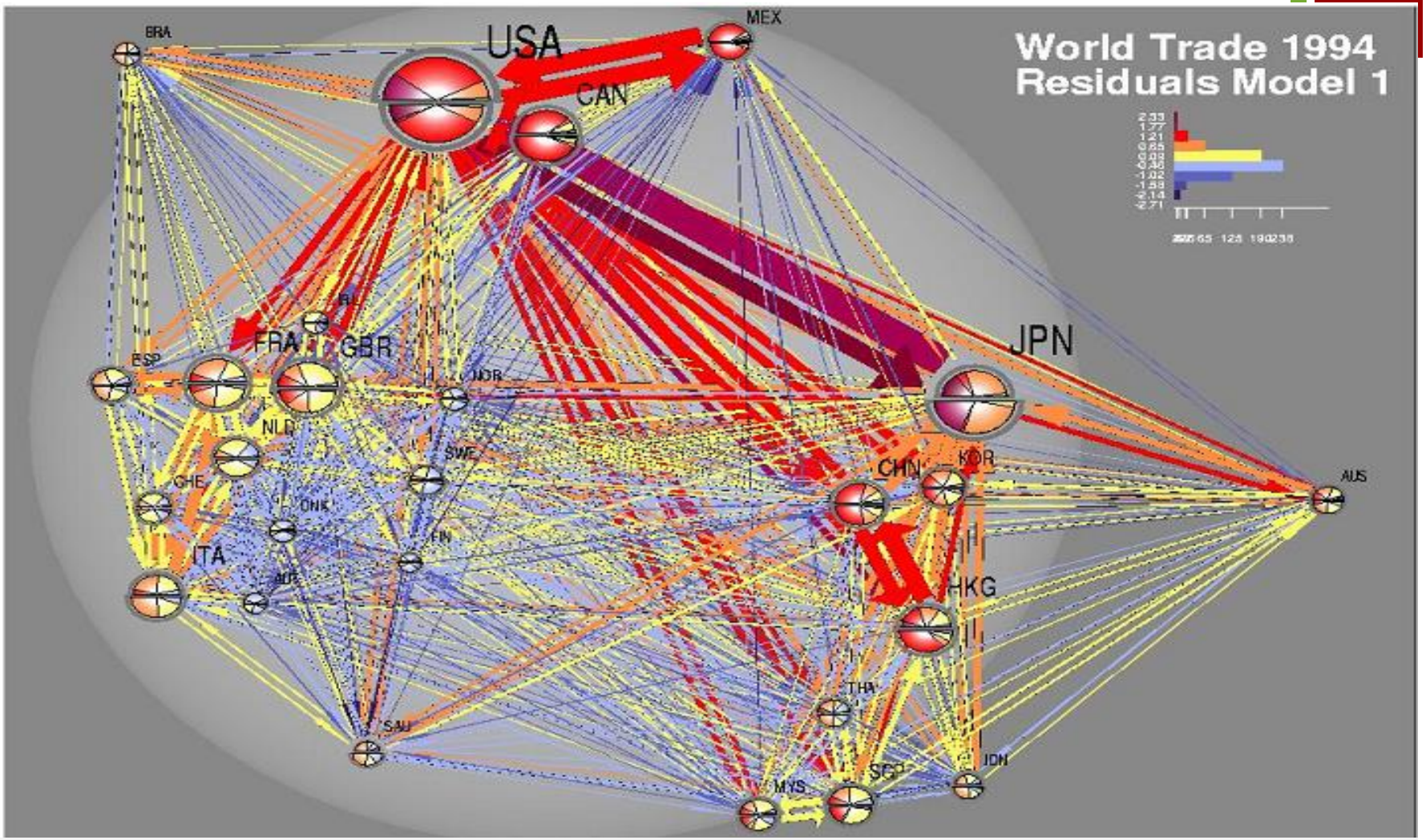






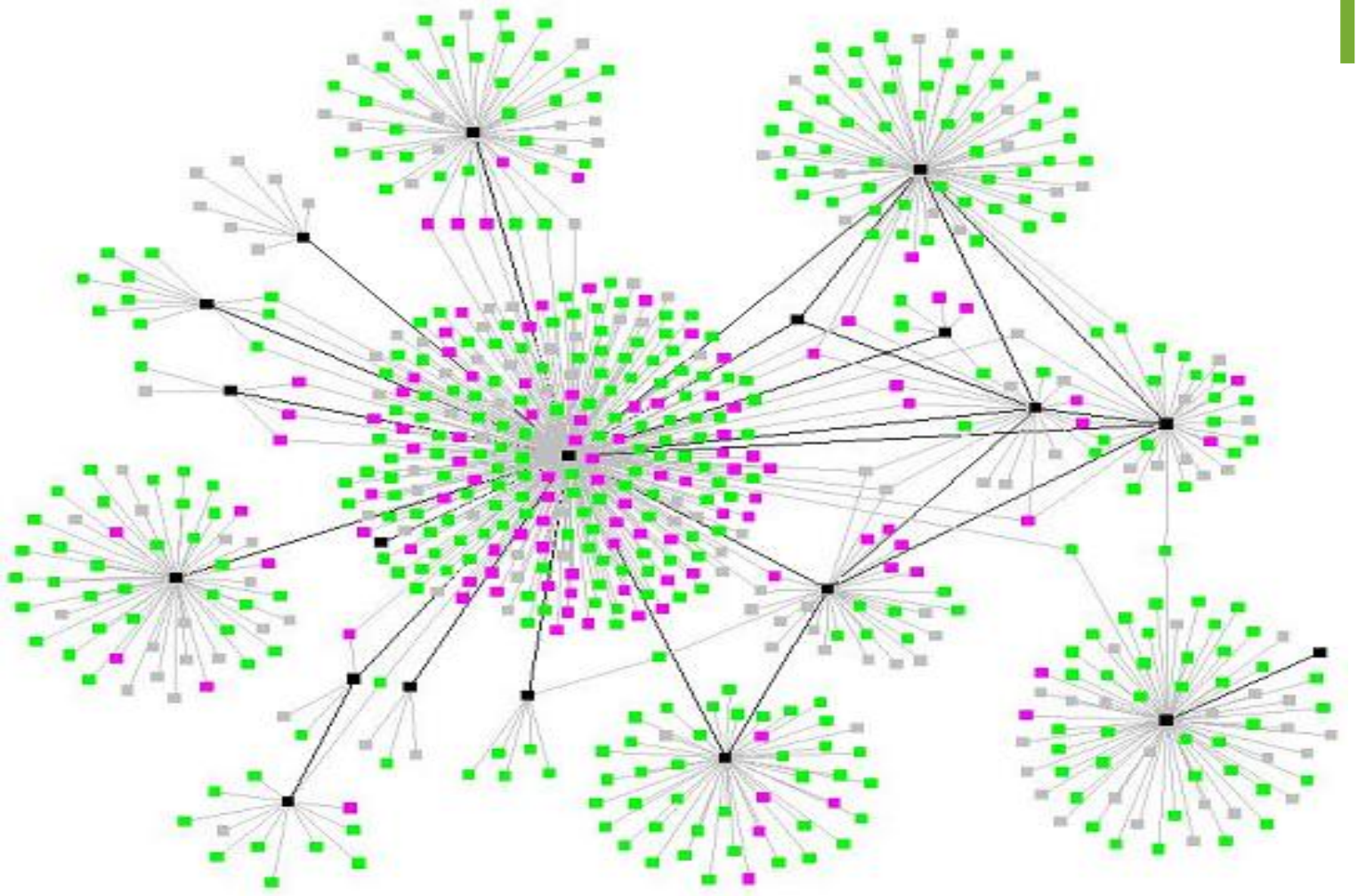
https://www.goodreads.com/book/show/8267287-networks-crowds-and-markets

+ Which country is trading with which?



From Book by David Easley and Jon Kleinberg

+ Who-infected-whom?



From Book by David Easley and Jon Kleinberg

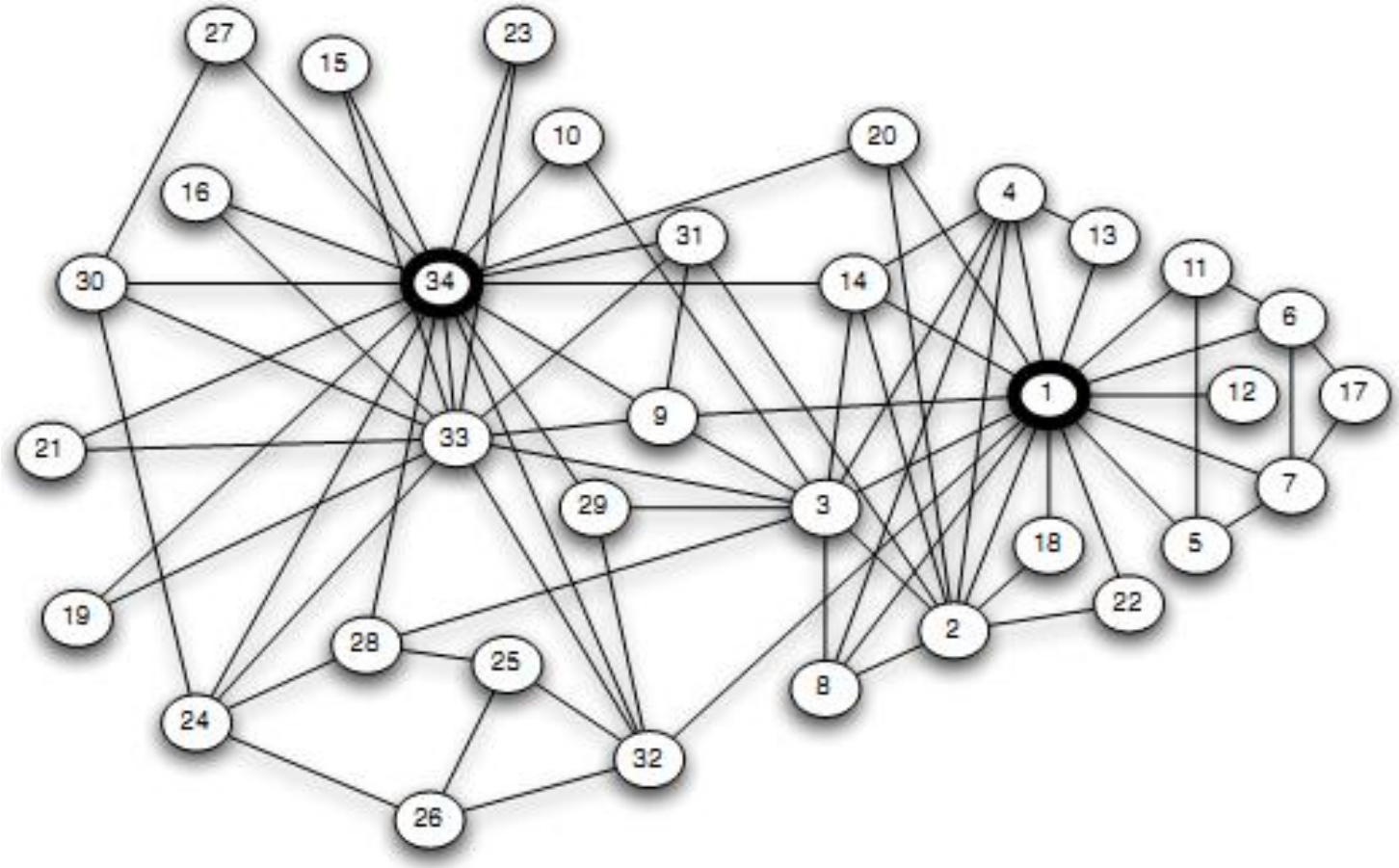
+ Who-email-Whom?

The pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the social organizational hierarchy .

http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg

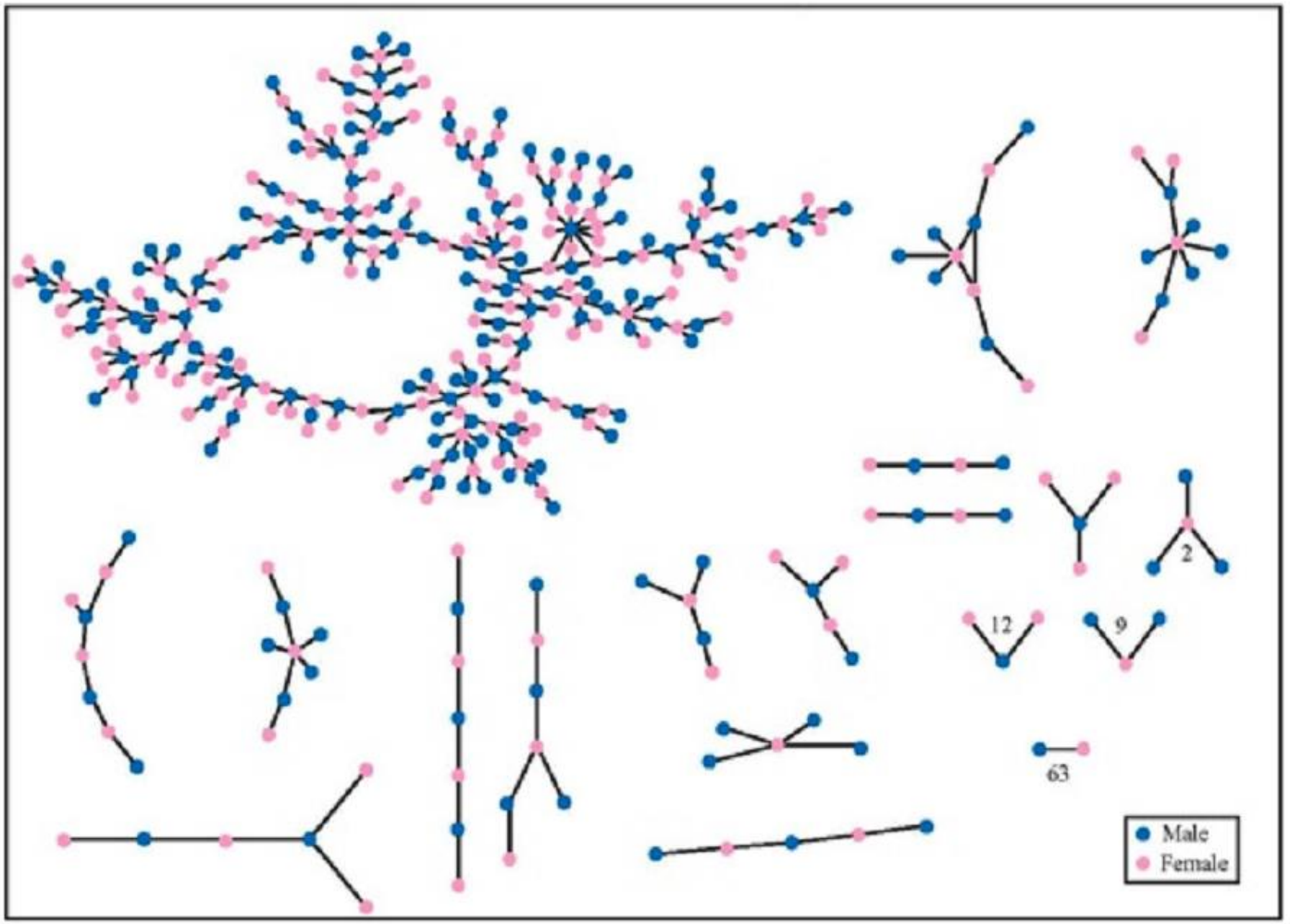
From Book by David Easley and Jon Kleinberg

+ Who-is-friend-of-whom?



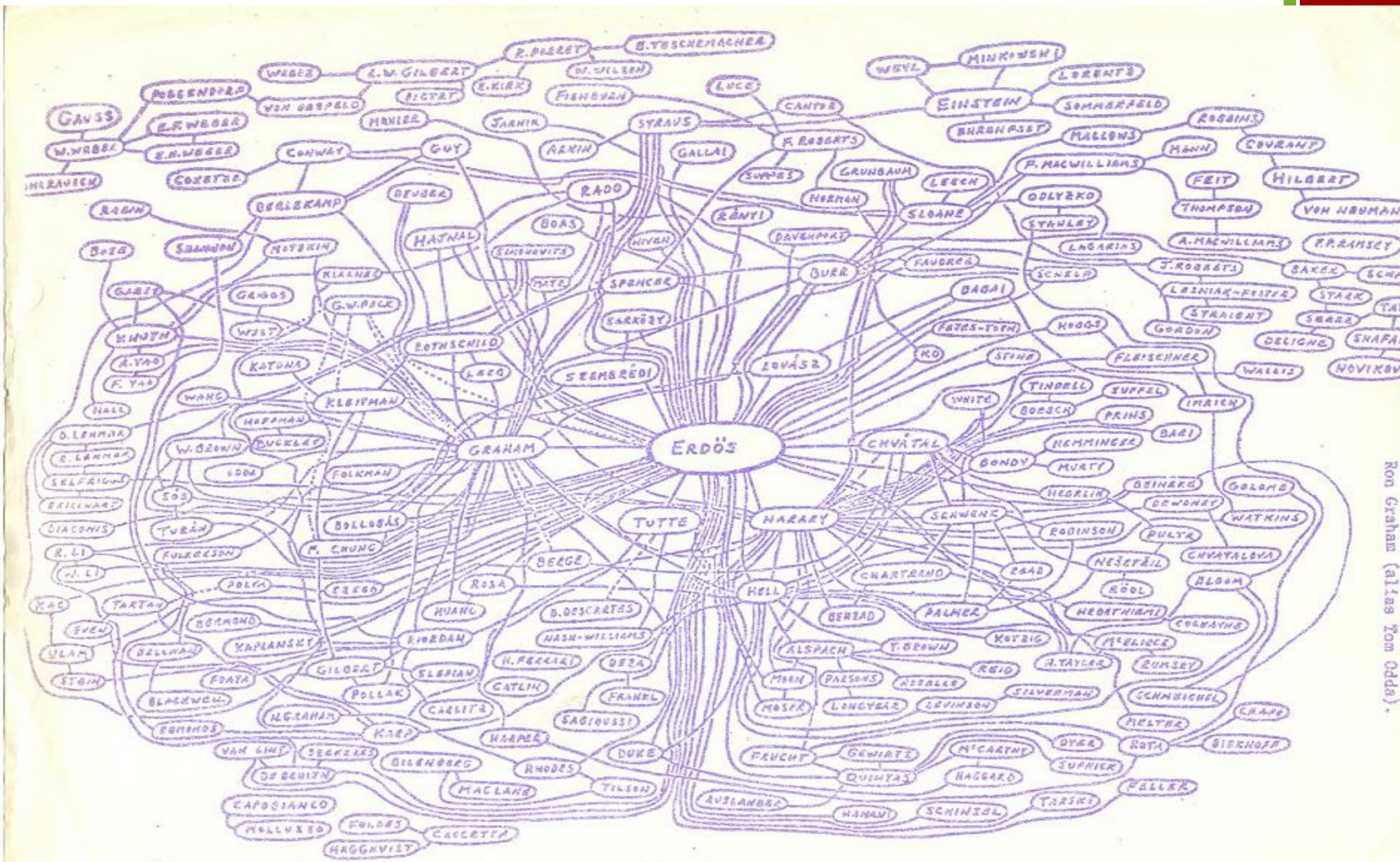
The social network of friendships within a 34-person karate club

High School Romance



From Book by David Easley and Jon Kleinberg

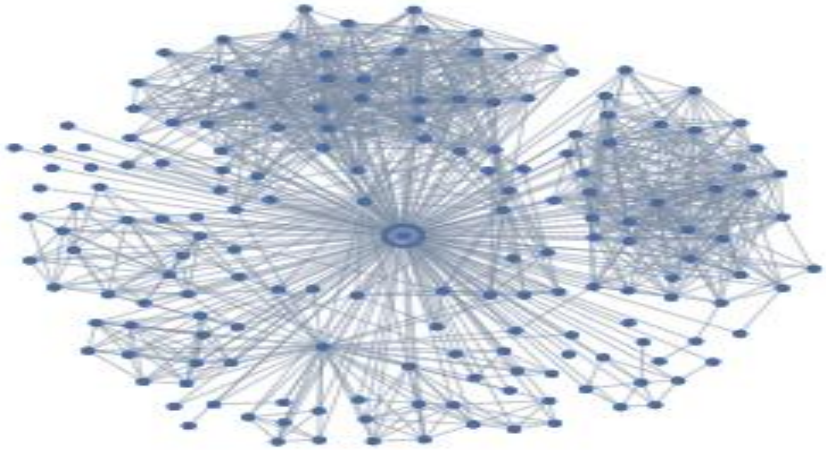
+ Mathematics Collaborations



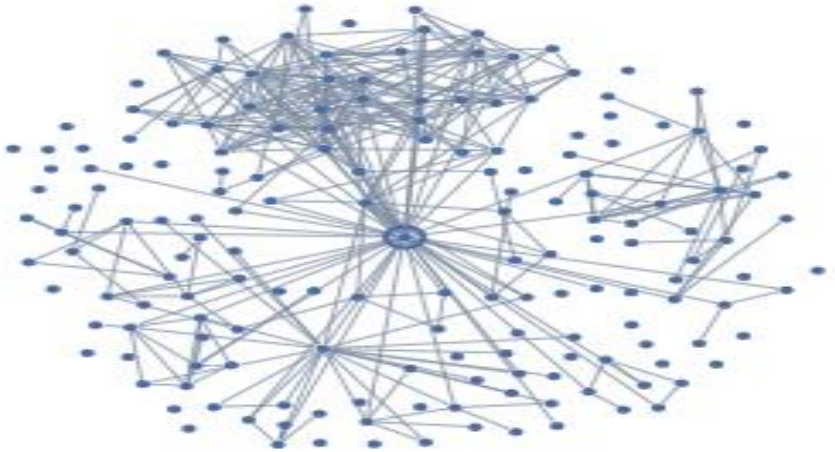
From Book by David Easley and Jon Kleinberg

Friends on Facebook

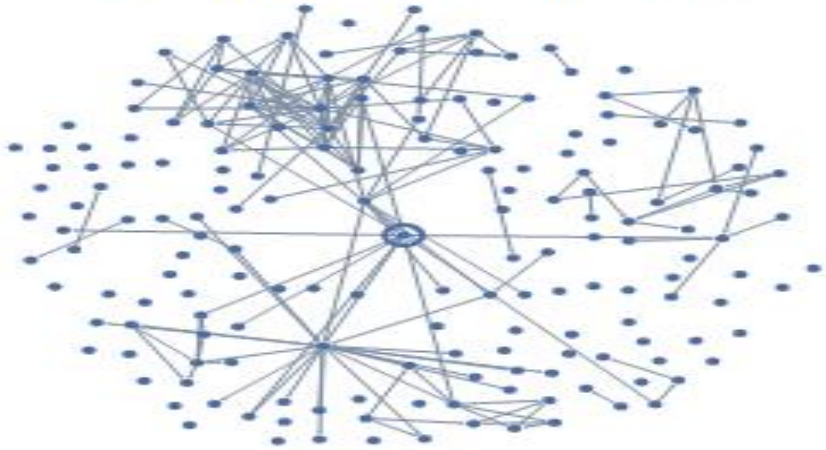
All Friends



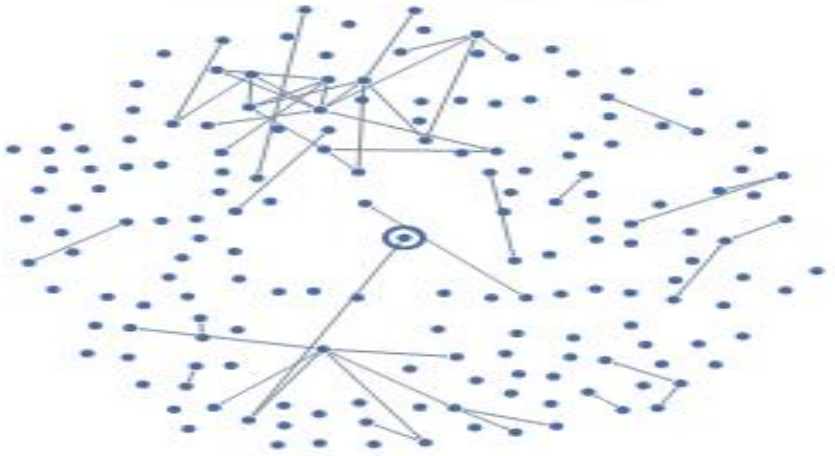
Maintained Relationships



One-way Communication

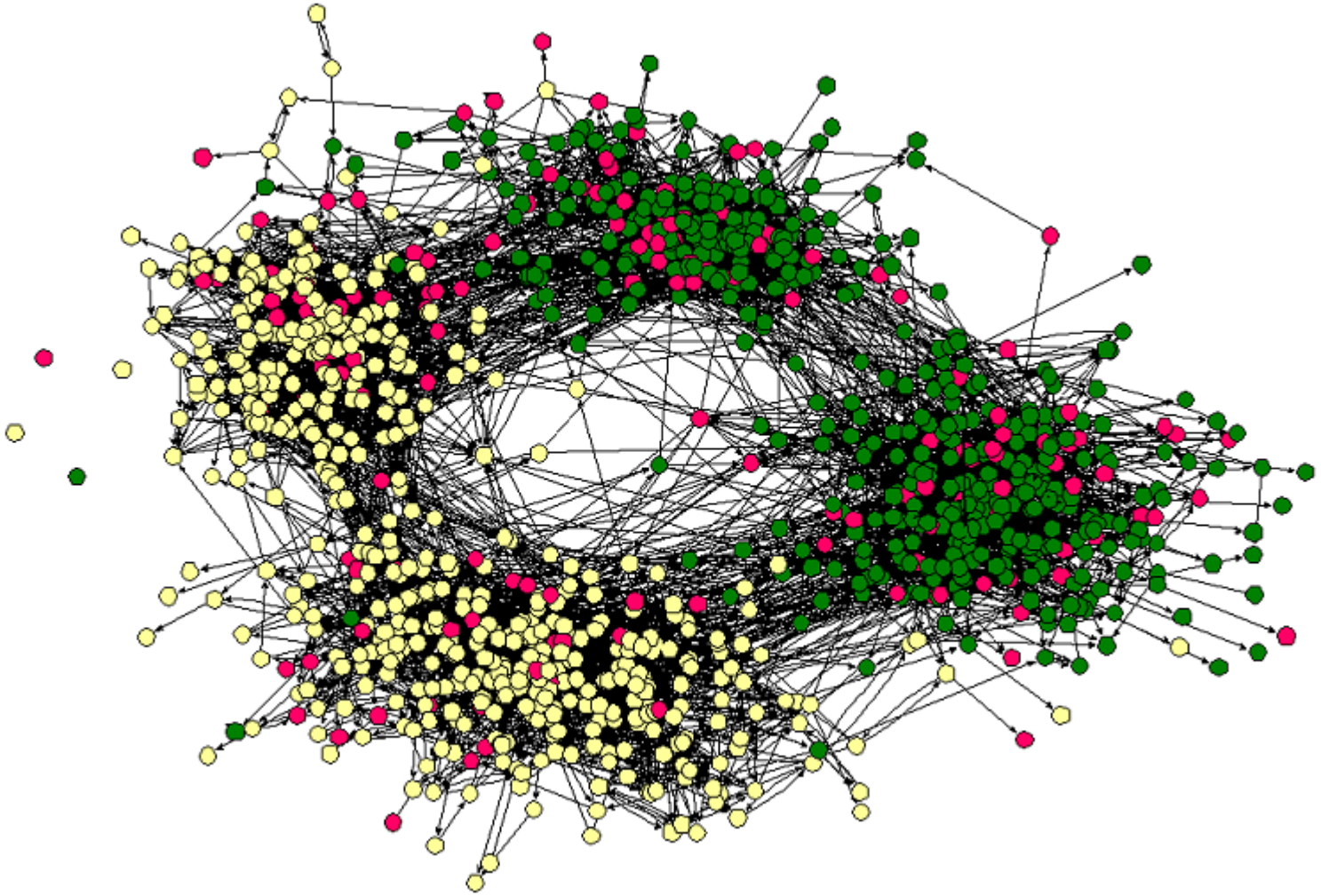


Mutual Communication



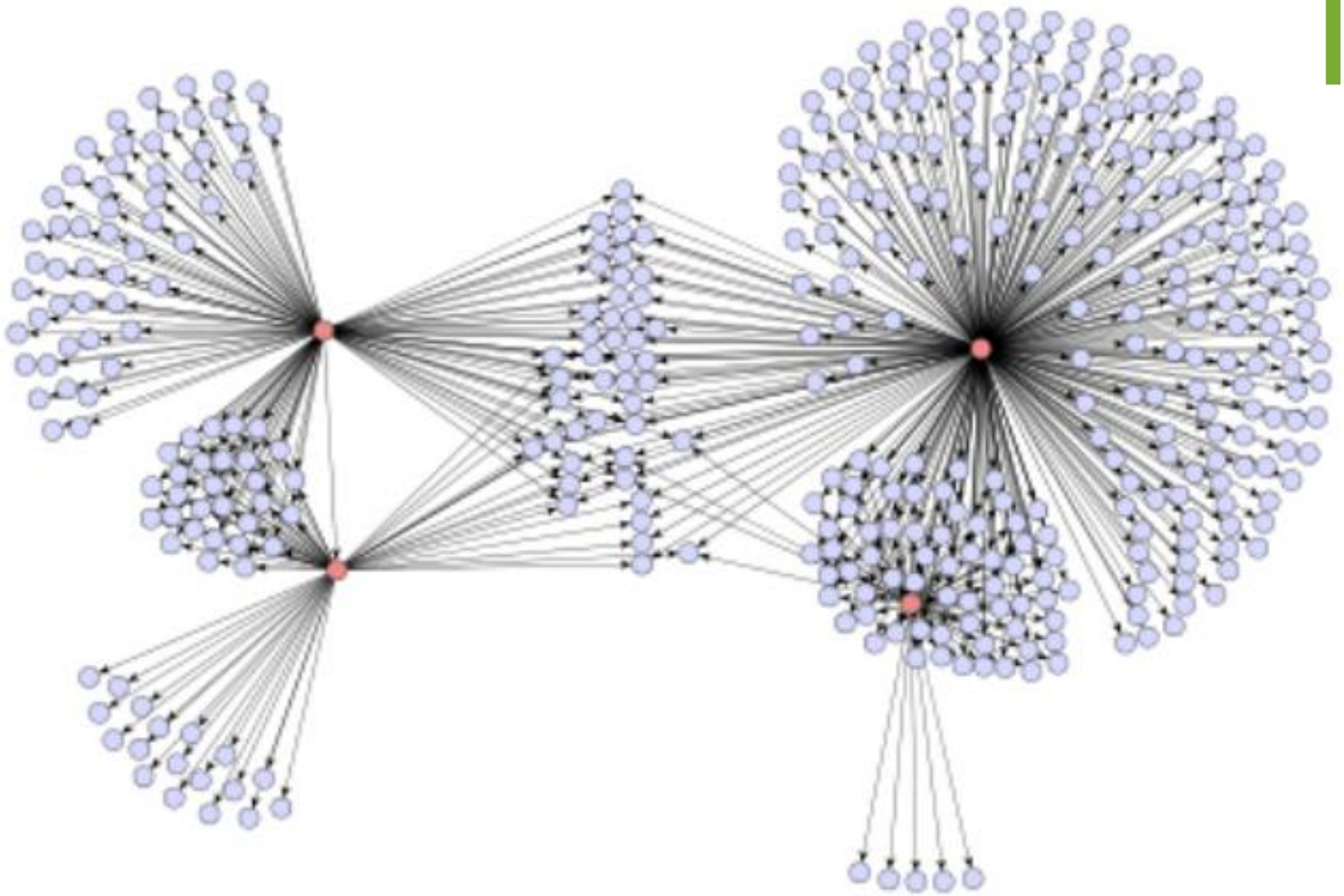
From Book by David Easley and Jon Kleinberg

Friends from middle school and high school grouped by races on Facebook



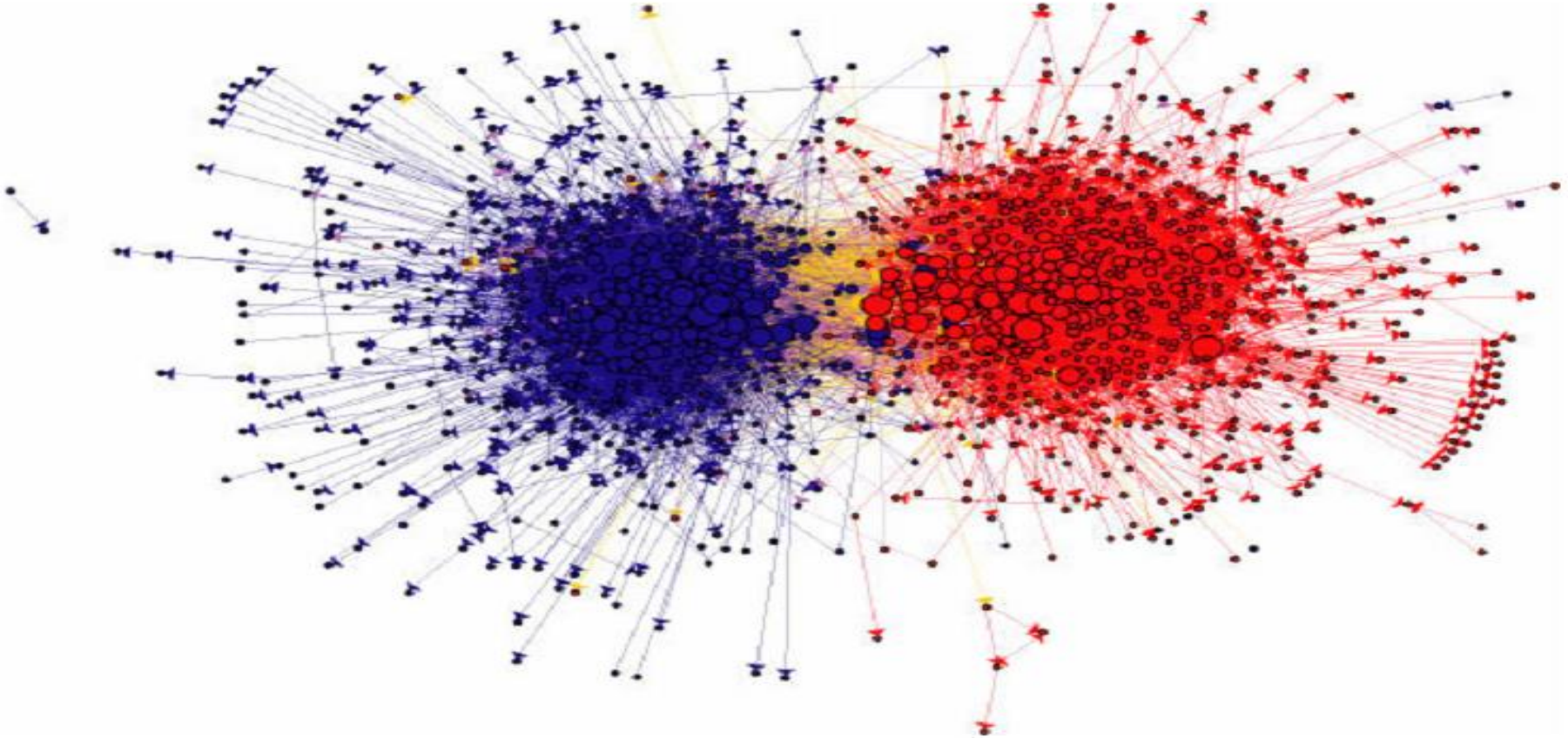
From Book by David Easley and Jon Kleinberg

Who-influence-whom?



e-mail recommendations for a Japanese graphic novel spread in a kind of informational (Book by David Easley and Jon Kleinberg)

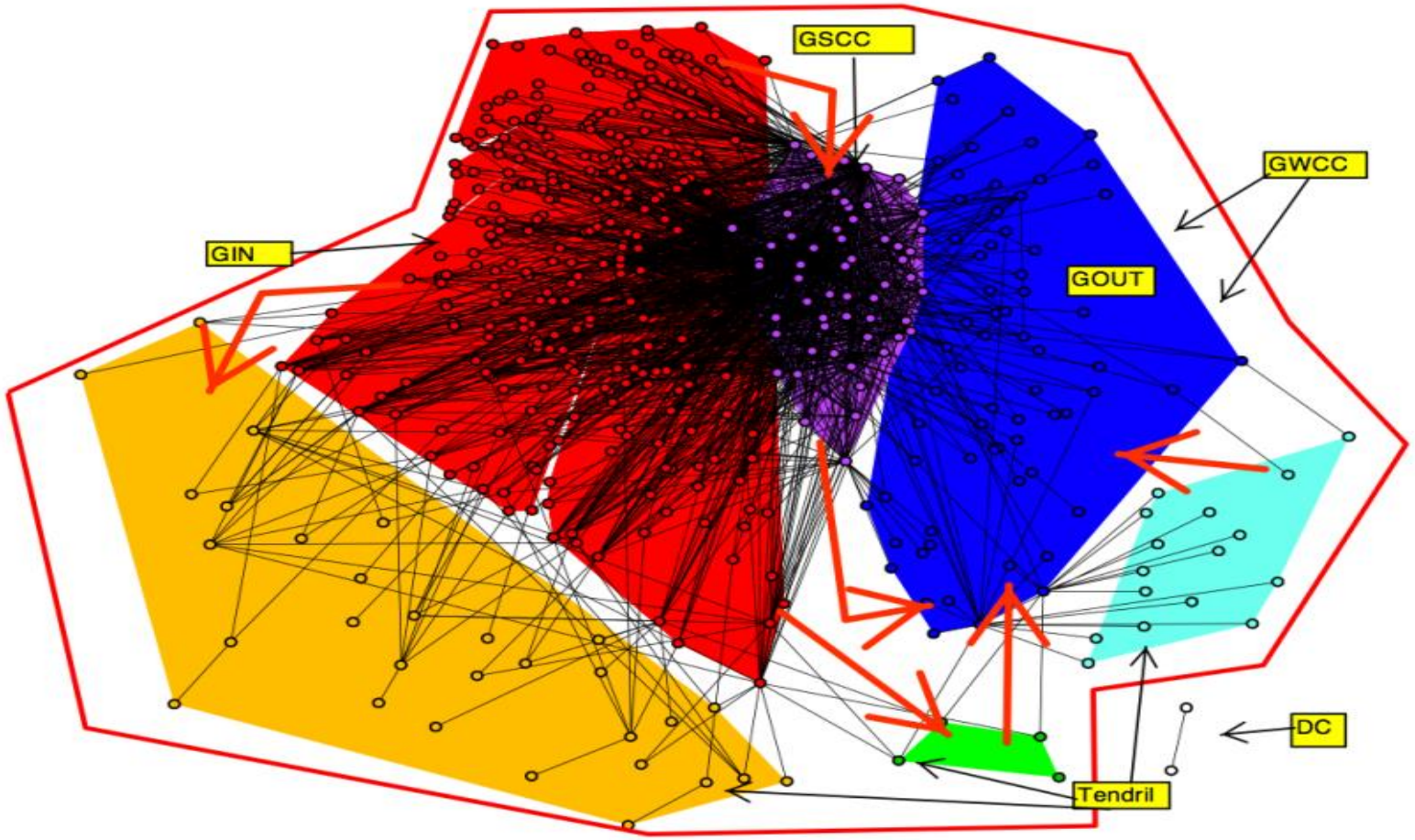
+ Who-support-which-party in Blogs?



- 2004 USA Presidential Election of Blogs' citations

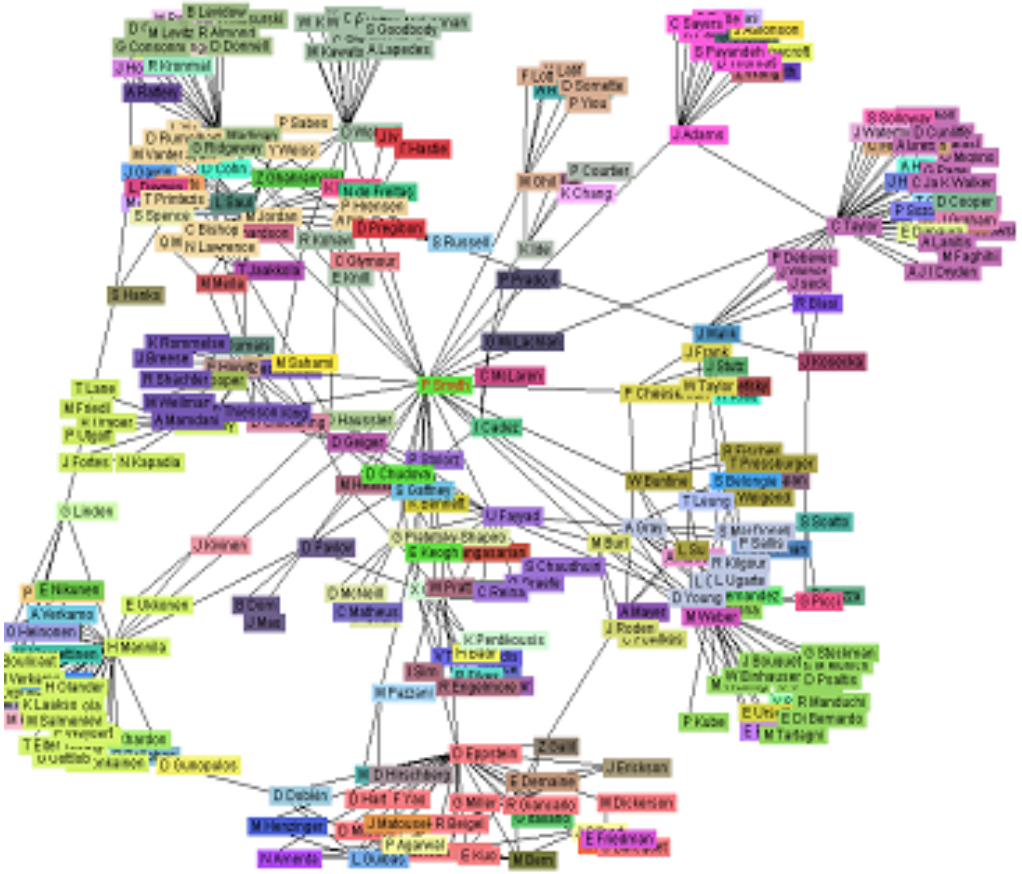
From Book by David Easley and Jon Kleinberg

+ Who-owe-whom?

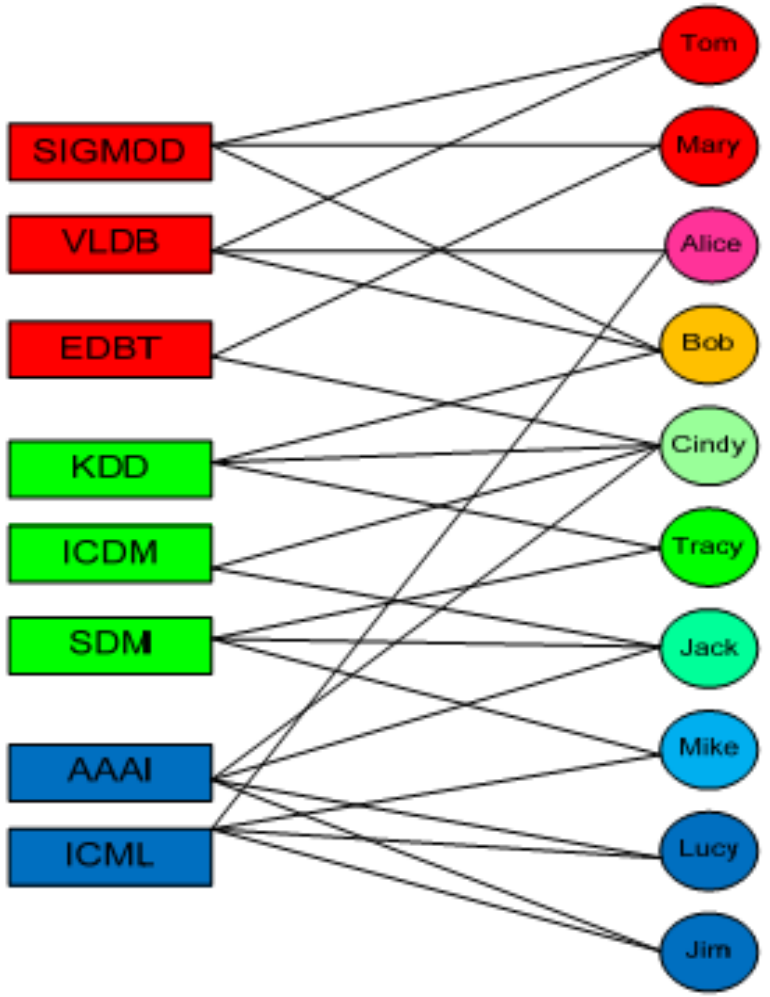


From Book by David Easley and Jon Kleinberg)

+ Homogenous vs Heterogeneous Networks

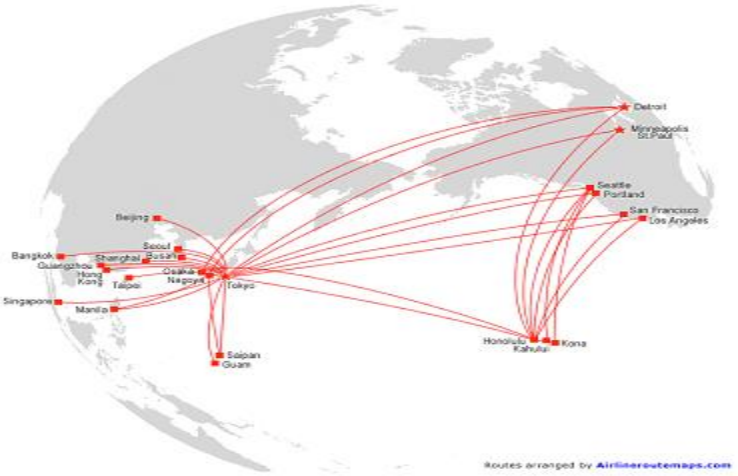


Co-author Network



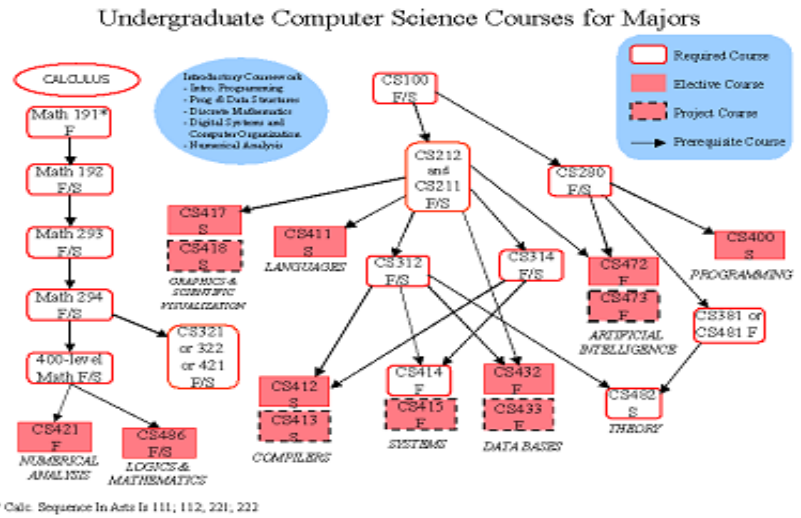
Conference-Author Network

Maps and Charts

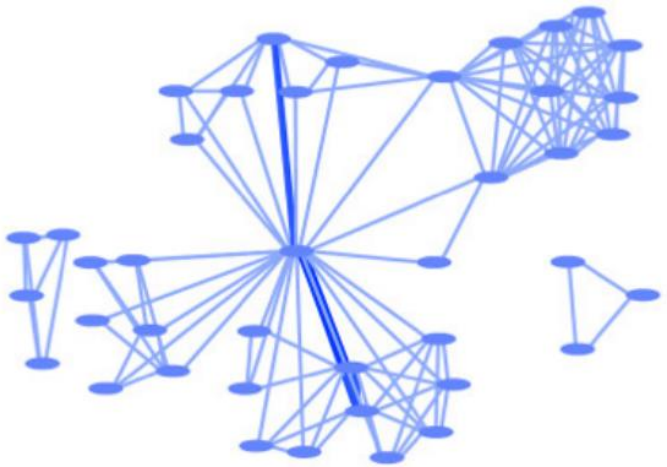


(a) *Airline routes*

(b) *Subway map*



(c) *Flowchart of college courses*

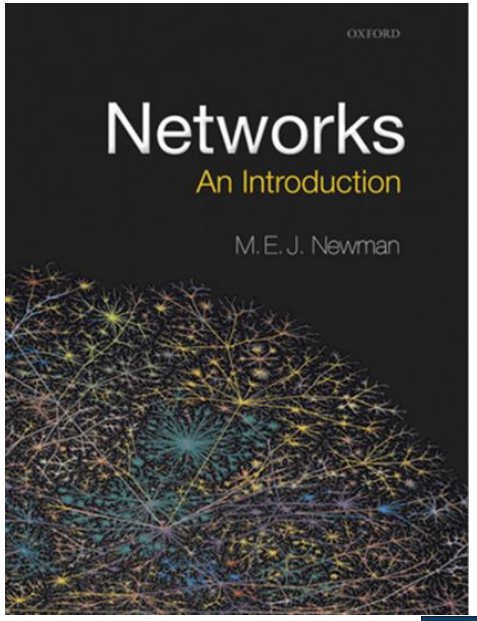


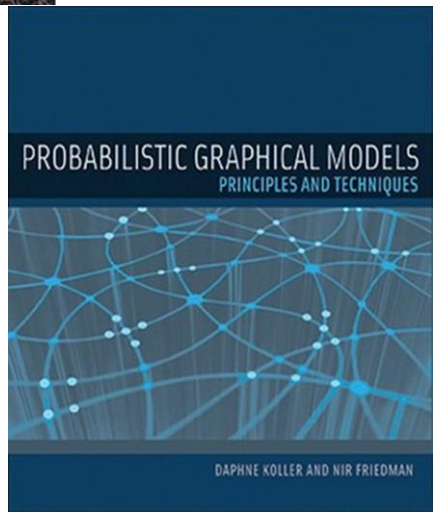
Research fields in Information networks

Research Area	Key Objective	Time Frame	Commercial Interest	Priority for Army Investment
Modeling, simulating, testing, and prototyping very large networks	Practical deployment tool sets	Mid term	High	High
Command and control of joint/combined networked forces	Networked properties of connected heterogeneous systems	Mid term	Medium	High
Impact of network structure on organizational behavior	Dynamics of networked organizational behavior	Mid term	Medium	High
Security and information assurance of networks	Properties of networks that enhance survival	Near term	High	High
Relationship of network structure to scalability and reliability	Characteristics of robust or dominant networks	Mid term	Medium	Medium
Managing network complexity	Properties of networks that promote simplicity and connectivity	Near term	High	High
Improving shared situational awareness of networked elements	Self-synchronization of networks	Mid term	Medium	High
Enhanced network-centric mission effectiveness	Individual and organizational training designs	Far term	Medium	Medium
Advanced network-based sensor fusion	Impact of control systems theory	Mid term	High	Medium
Hunter-prey relationships	Algorithms and models for adversary behaviors	Mid term	Low	High
Swarming behavior	Self-organizing UAV/UGV; self-healing	Mid term	Low	Medium
Metabolic and gene expression networks	Soldier performance enhancement	Near term	Medium	Medium

+ Data Networks: What are we computing?

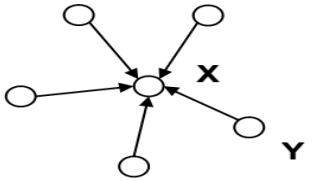
- Network Centrality
- Network Modularity
- Network Reachability
- Ranking
- Similarity
- Causality
- Correlation
- Outlier
- Small World (Six Degree Separation)
- Ontology
- Trajectory
- Graph Embedding



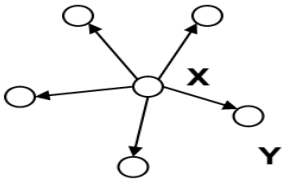


Network Centrality

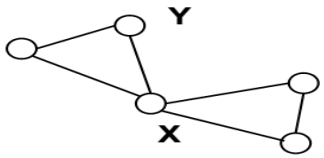
In each of the following networks, X has higher centrality than Y according to a particular measure



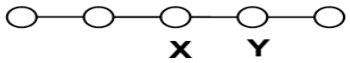
indegree



outdegree



betweenness



closeness

maximum value in the network

$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(i)\right]}{[(N-1)(N-2)]}$$

$$M = (\prod_{t=1}^{n-1} W_{X_t C} D^{-1}_{C X_{t+1}} W_{C X_{t+1}}) W_{X_n C} D^{-1}_{C X_1} W_{C X_1}$$

Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$

$$C_B(i) = \sum_{j<k} g_{jk}(i)/g_{jk}$$

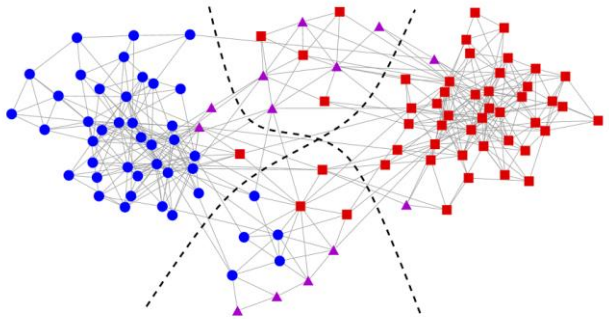
Where g_{jk} = the number of geodesics connecting *jk*, and $g_{jk}(i)$= the number of geodesics that actor *i* is on.

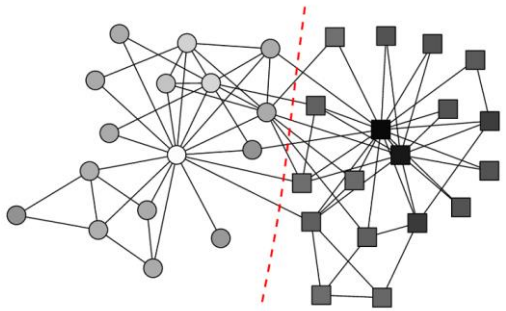
http://dl.acm.org/citation.cfm?id=2525417

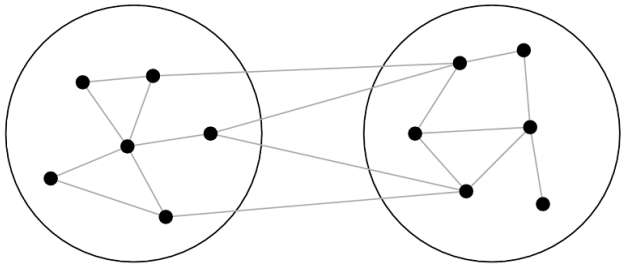
http://cs.brynmawr.edu/Courses/cs380/spring2013/section02/slides/05_Centrality.pdf

https://en.wikipedia.org/wiki/Centrality

Network Modularity







$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

where e_{ij} is the fraction of edges with one end vertices in community *i* and the other in community *j*:

$$e_{ij} = \sum_{vw} \frac{A_{vw}}{2m} 1_{v \in c_i} 1_{w \in c_j}$$

and a_i is the fraction of ends of edges that are attached to vertices in community *i*:

$$a_i = \frac{k_i}{2m} = \sum_j e_{ij}$$

https://en.wikipedia.org/wiki/Modularity_(networks)

http://vw.indiana.edu/netsci06/conf-slides/conf-mon/netsci-talk-mark-newman.pdf

Network Reachability

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\cdot\mathbf{x}(t) + \mathbf{B}\cdot\mathbf{u}(t)$$

$\mathbf{x}(t) \in R^{N\times 1}$: state vector.

$\mathbf{u}(t) \in R^{M\times 1}$: input vector $(M \le N)$.

$\mathbf{A} \in R^{N\times N}$: state matrix

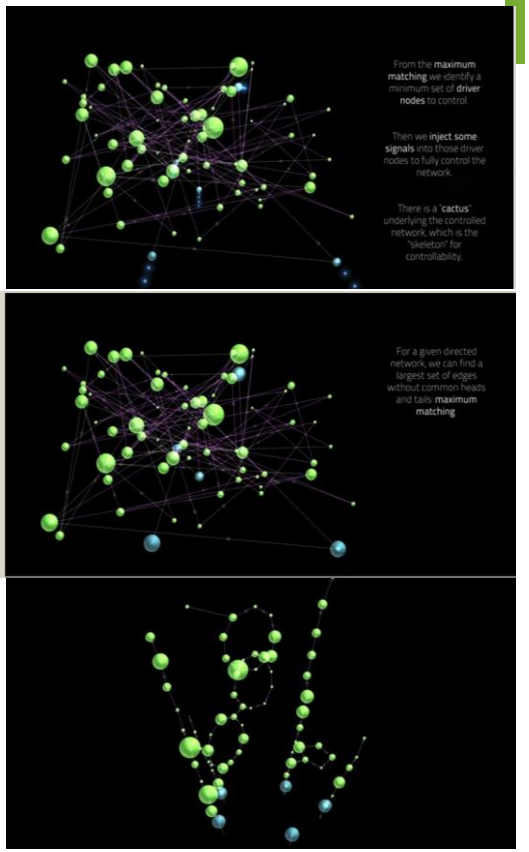
(weighted wiring diagram).

$\mathbf{B} \in R^{N\times M}$: input matrix

($\Rightarrow$ control configuration).

$$\text{rank } \mathbf{C} = N$$

$$\mathbf{C} = [\mathbf{B}, \mathbf{A}\cdot\mathbf{B}, \mathbf{A}^2\cdot\mathbf{B}, \cdots, \mathbf{A}^{N-1}\cdot\mathbf{B}]$$



$$N_D = \max\{1, N_{\text{unmatched}}\}$$

$$P(k_{\text{in}}, k_{\text{out}}) \rightarrow N_D$$

https://journals.aps.org/pre/pdf/10.1103/PhysRevE.71.046119

Research Project by: A.L. Barabasi, Yang-Yu Liu, Northeastern University, USA

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5541621

http://www.ns-cta.org/ns-cta-blog/?page_id=552

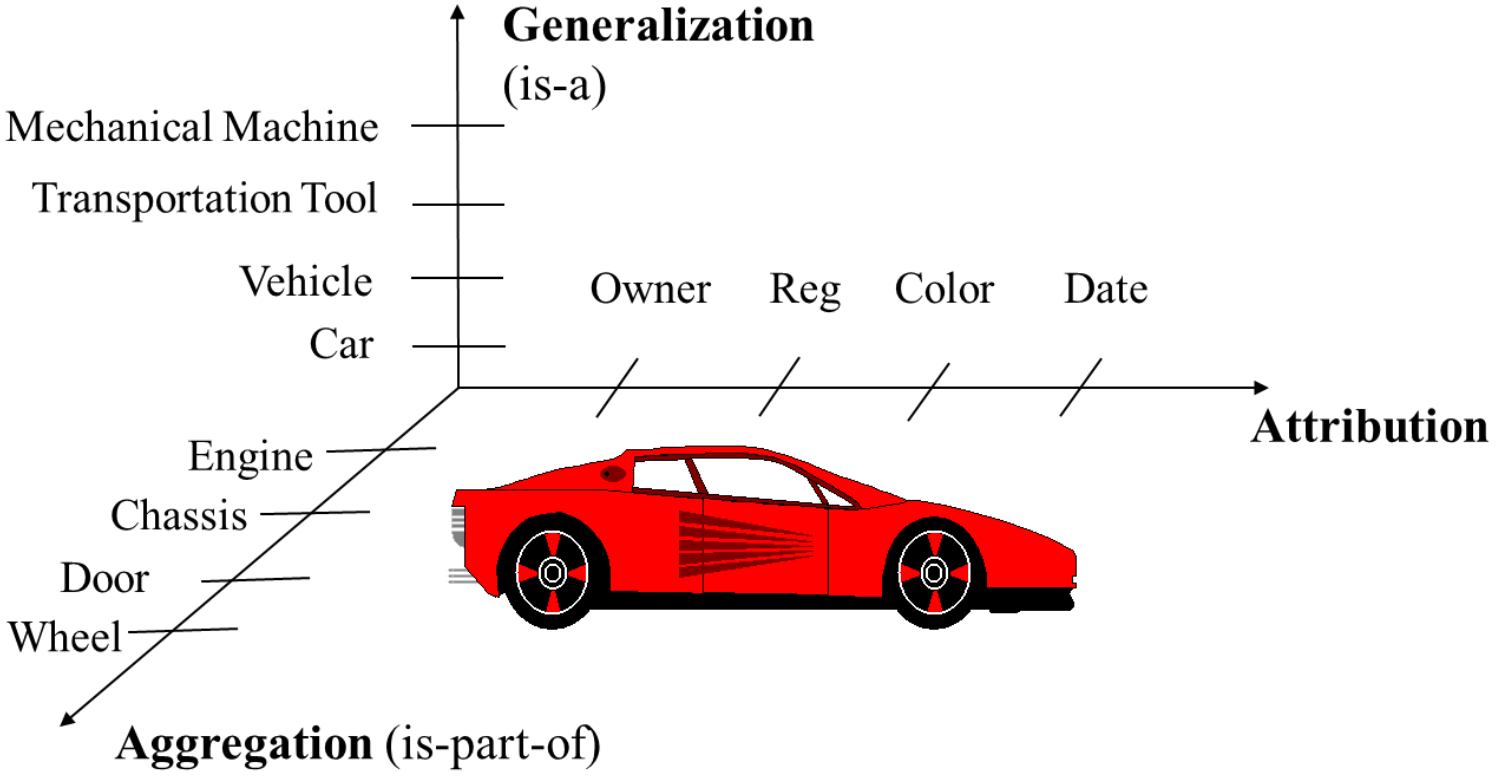
Dealing with Data: The **Curse of Dimensionality**

Adding extra dimensions to a data space will exponentially increase the volume of data space.

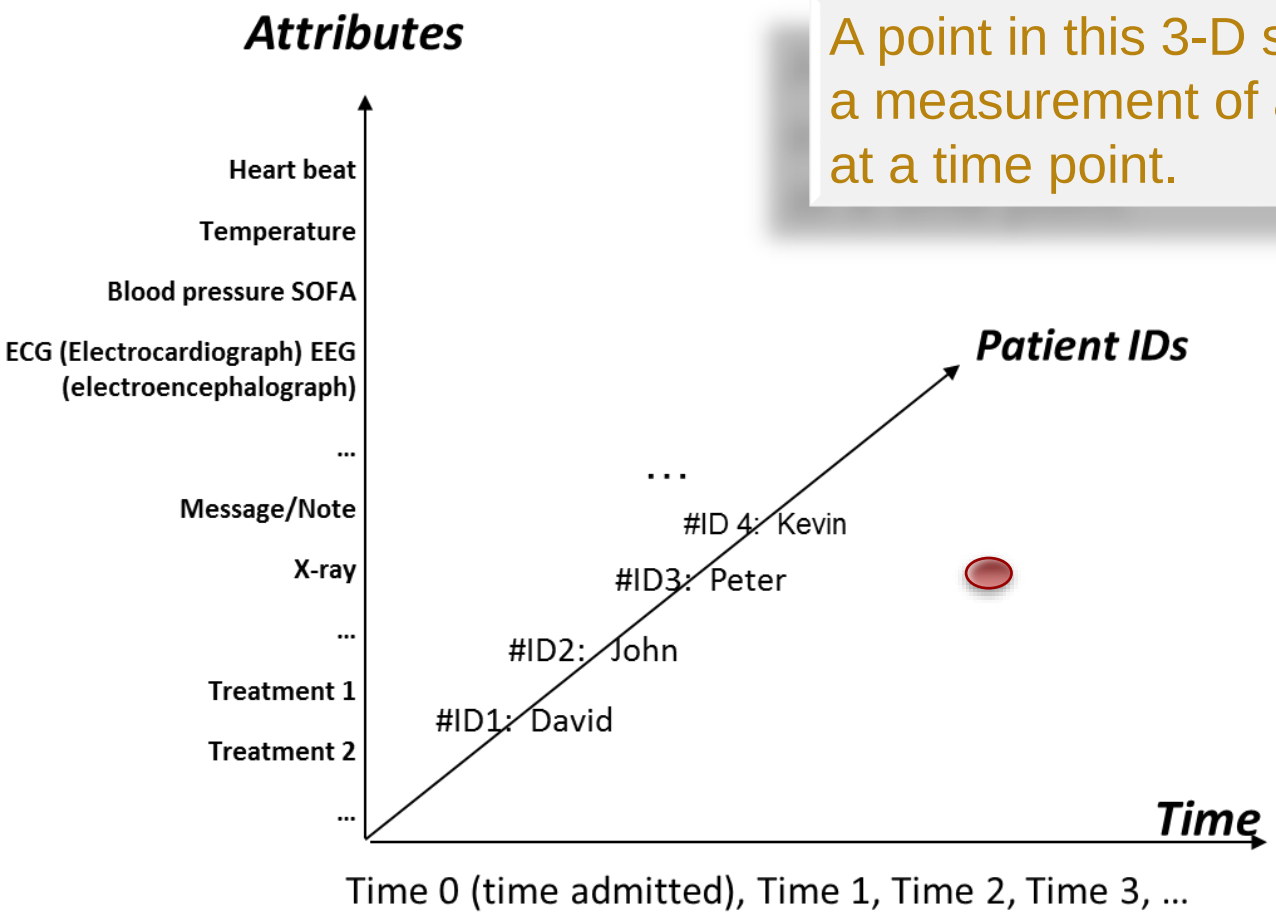
Curse of Dimensionality

http://www.liaotuo.org/remen/zhouyu.html

Example: Data for a multidimensional car from an Object-Oriented Perspective



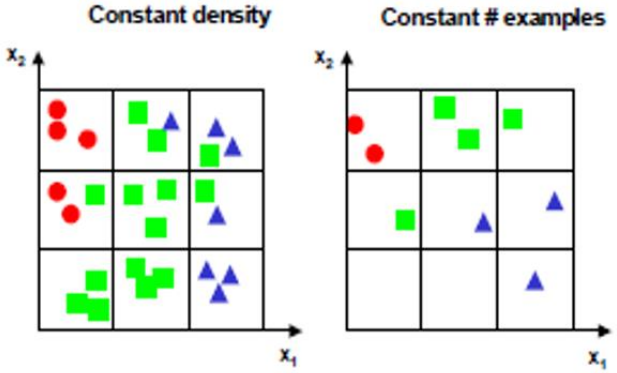
Example: Data for a multidimensional car from an Object-Oriented Perspective



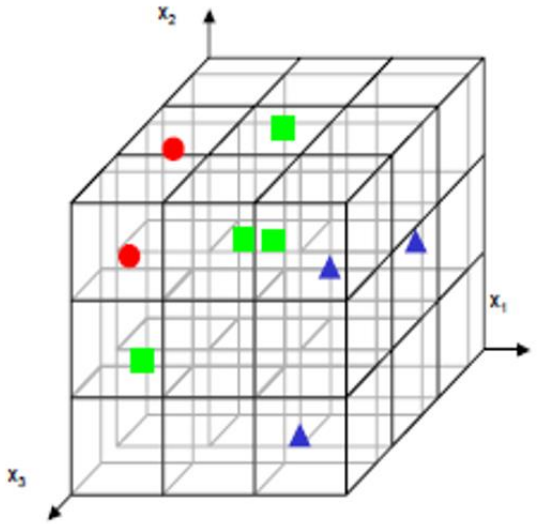
A point in this 3-D space is a measurement of a patient at a time point.

The Curse of Dimensionality

- **We decide to preserve the granularity of each axis, which raises the number of bins from 3 (in 1D) to 3^2=9 (in 2D)**
  - At this point we are faced with a decision: do we maintain the density of examples per bin or do we keep the number of examples we used for the one-dimensional case?
    - Choosing to maintain the density increases the number of examples from 9 (in 1D) to 27 (in 2D)
    - Choosing to maintain the number of examples results in a 2D scatter plot that is very sparse

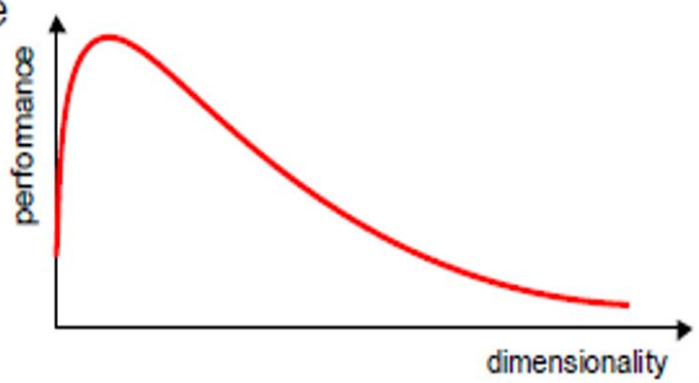


- **Moving to three features makes the problem worse:**
  - The number of bins grows to 3^3=27
  - For the same density of examples the number of needed examples becomes 81
  - For the same number of examples, well, the 3D scatter plot is almost empty



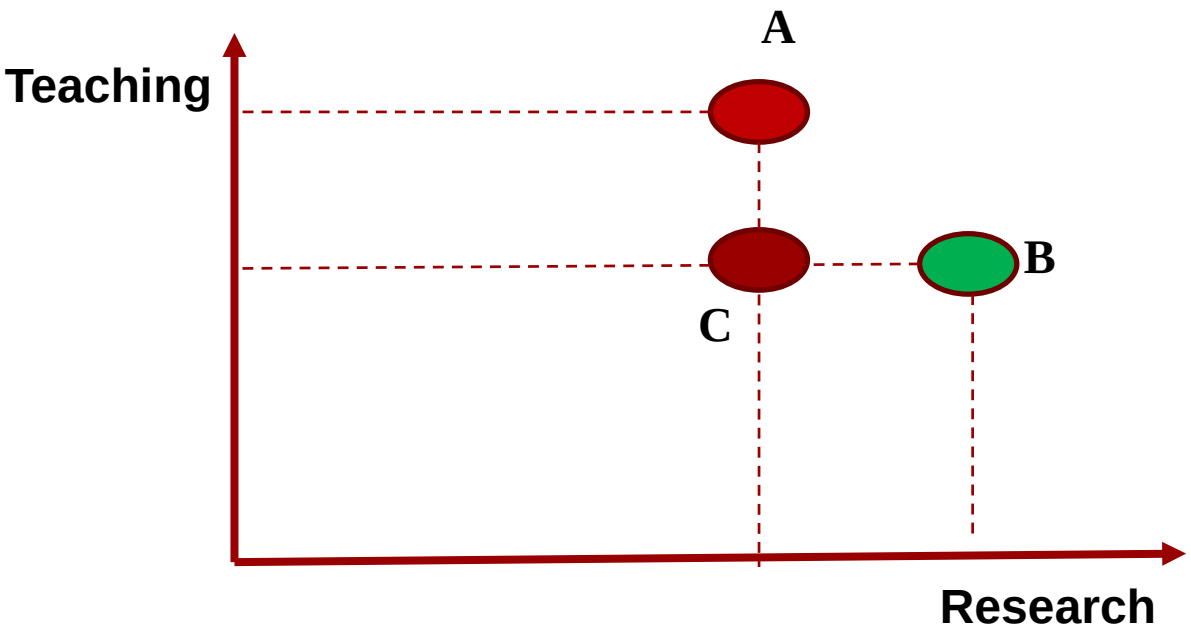
The Curse of Dimensionality

- **How do we beat the curse of dimensionality?**
  - By incorporating prior knowledge
  - By providing increasing smoothness of the target function
  - By reducing the dimensionality
- **In practice, the curse of dimensionality means that, for a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve**
  - In most cases, the additional information that is lost by discarding some features is (more than) compensated by a more accurate mapping in the lower-dimensional space



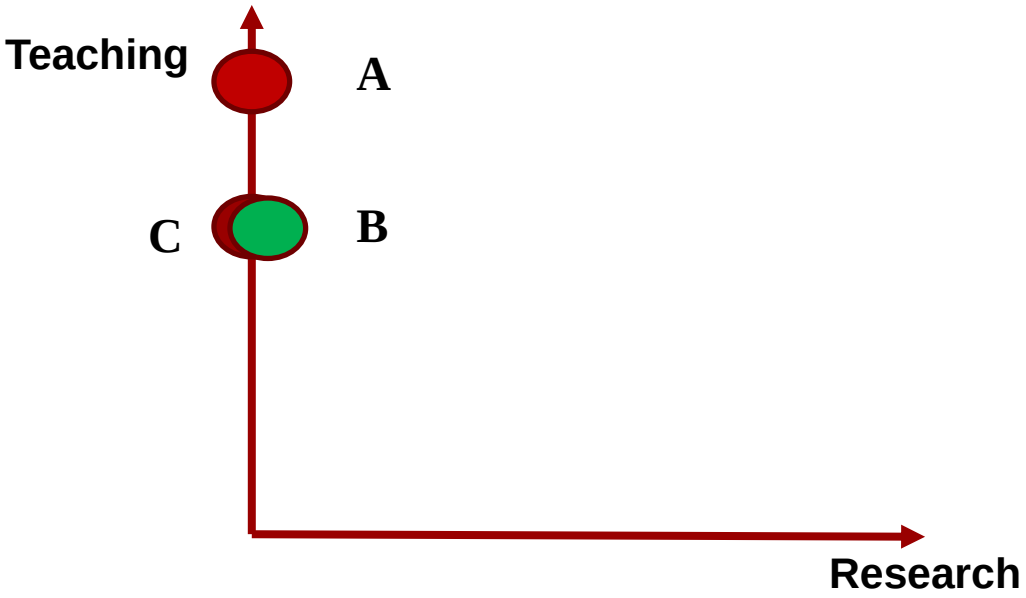
Should the dimensionality be high or **Low**?

- **Example:** Given the following scores for three candidates (A, B, C) for a position of a Faculty Position in a university, which one of them should be selected to offer a job?



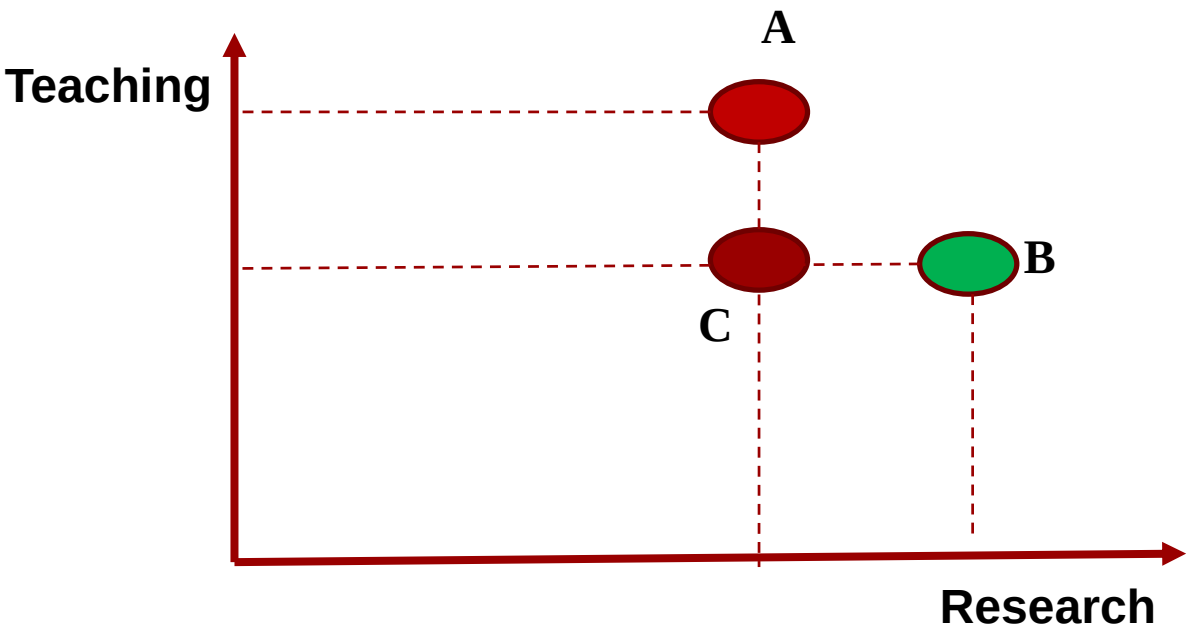
Should the dimensionality be high or **Low**?

- **Example:** Given the following scores for three candidates (A, B, C) for a position of a Faculty Position in a university, which one of them should be selected to offer a job?



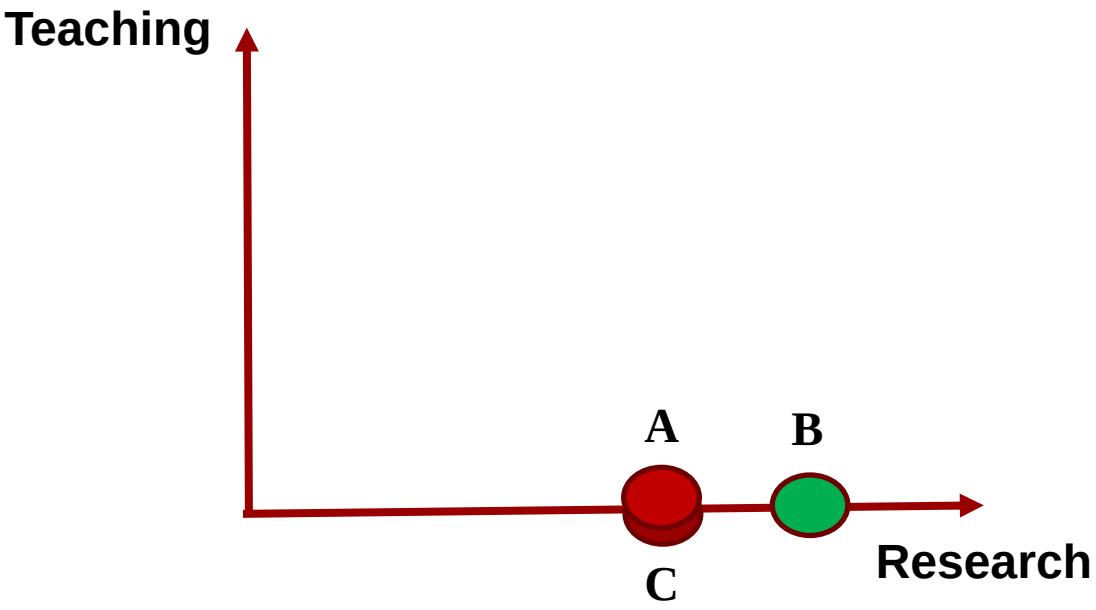
Should the dimensionality be high or **Low**?

- **Example:** Given the following scores for three candidates (A, B, C) for a position of a Faculty Position in a university, which one of them should be selected to offer a job?



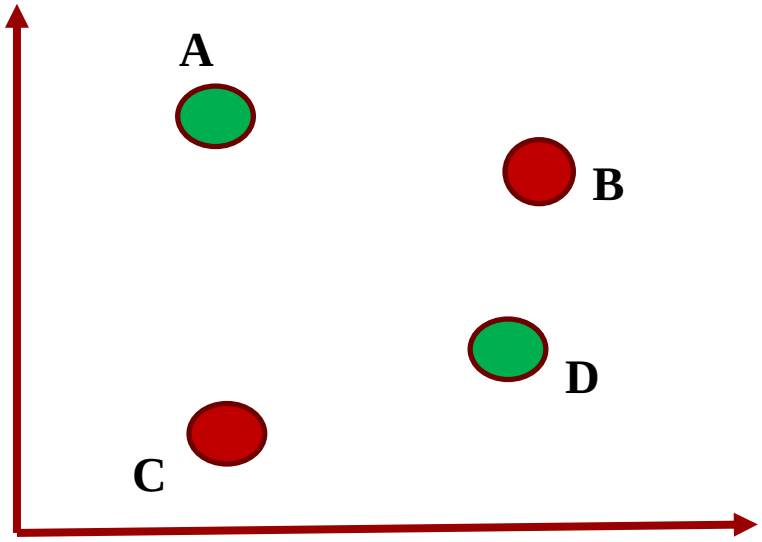
+ Should the dimensionality be high or **Low**?

- **Example:** Given the following scores for three candidates (A, B, C) for a position of a Faculty Position in a university, which one of them should be selected to offer a job?



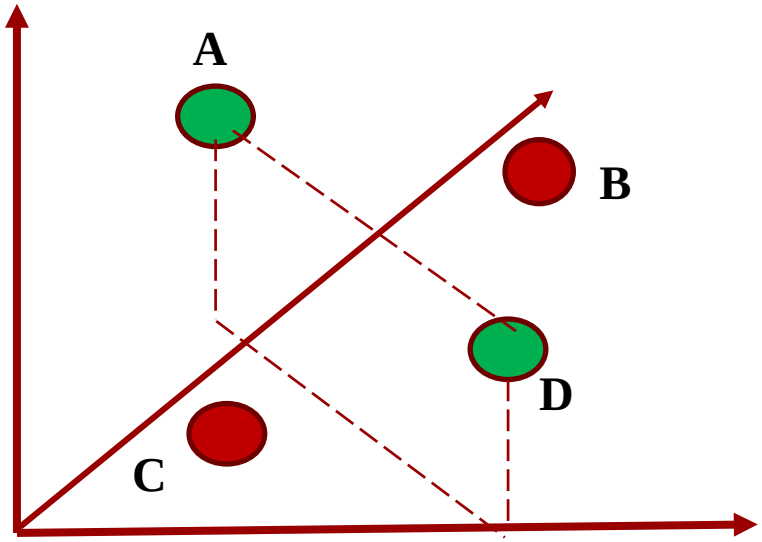
+ Should the dimensionality be **high** or Low?

- **Example:** Given two classes points, Red and Green in a 2D space, how do we classify these two classes of points?



Should the dimensionality be **high** or Low?

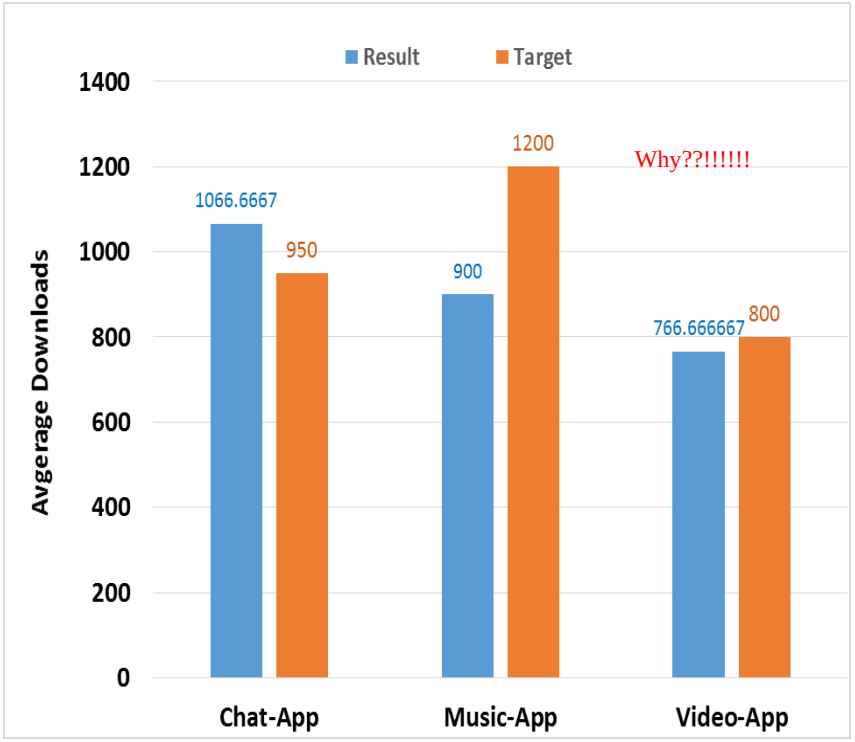
- **Example:** Given two classes points, Red and Green in a 2D space, how do we classify these two classes of points?



Data Exploration: How do we give Explanations Automatically?

Mobile App Example

Why the target were not met for these Apps?



id	App name	Carrier	Crash rate	Loading time	Session time	Downloads
T1	Chat-app	Vodafone	1.5	4	10	1000
T2	Video-app	Vodafone	2	3	15	1200
T3	Music-app	Vodafone	4	1	20	800
T4	Chat-app	Optus	3	2	20	1300
T5	Video-app	Optus	6	2	10	700
T6	Music-app	Optus	1	3	30	1200
T7	Chat-app	Telstra	1	4	11	900
T8	Video-app	Telstra	1.2	9	15	400
T9	Music-app	Telstra	4	6	25	700

Reasons:

Crash rate BETWEEN 3.0 AND 6.0 and
Session time BETWEEN 18.0 AND 26.0

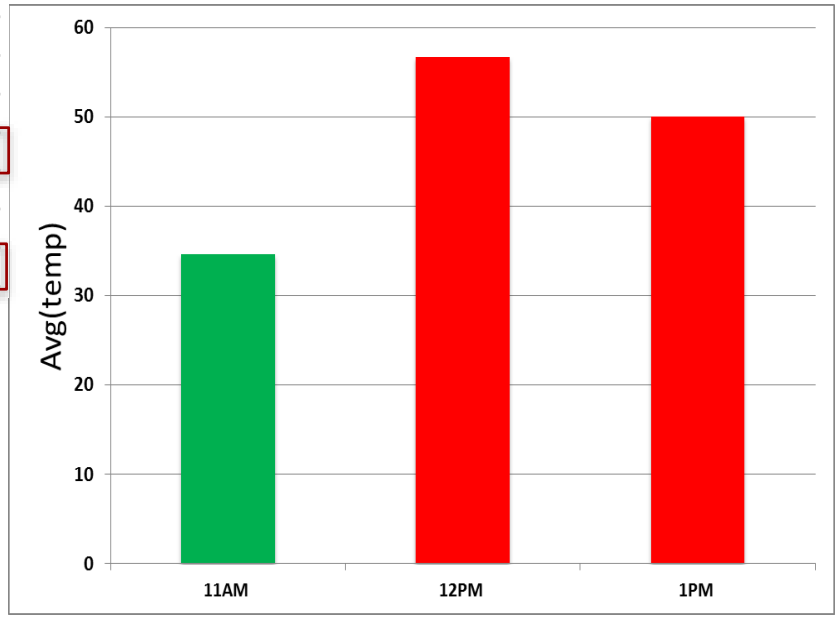
Q: SELECT app_name, Avg(downloads)
FROM app_data
GROUP BY app_name

[1] Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB Endow., 6(8):553–564, June 2013.

+ Another Example: Scorpion System

Sensors

Tuple id	Time	SensorID	Voltage	Humidity	Temp.
T1	11AM	1	2.64	0.4	34
T2	11AM	2	2.65	0.5	35
T3	11AM	3	2.63	0.4	35
T4	12PM	1	2.7	0.3	35
T5	12PM	2	2.7	0.5	35
T6	12PM	3	2.3	0.4	100
T7	1PM	1	2.7	0.3	35
T8	1PM	2	2.7	0.5	35
T9	1PM	3	2.3	0.5	80



SELECT avg(temp), time
FROM sensors GROUP BY time

- Scorpion[1] : a system that takes a set of user-specified outlier points in an aggregate query result and finds predicates that explain the outliers.

- It seeks to find a predicate over an input dataset that most influences a user selected set of query outputs.

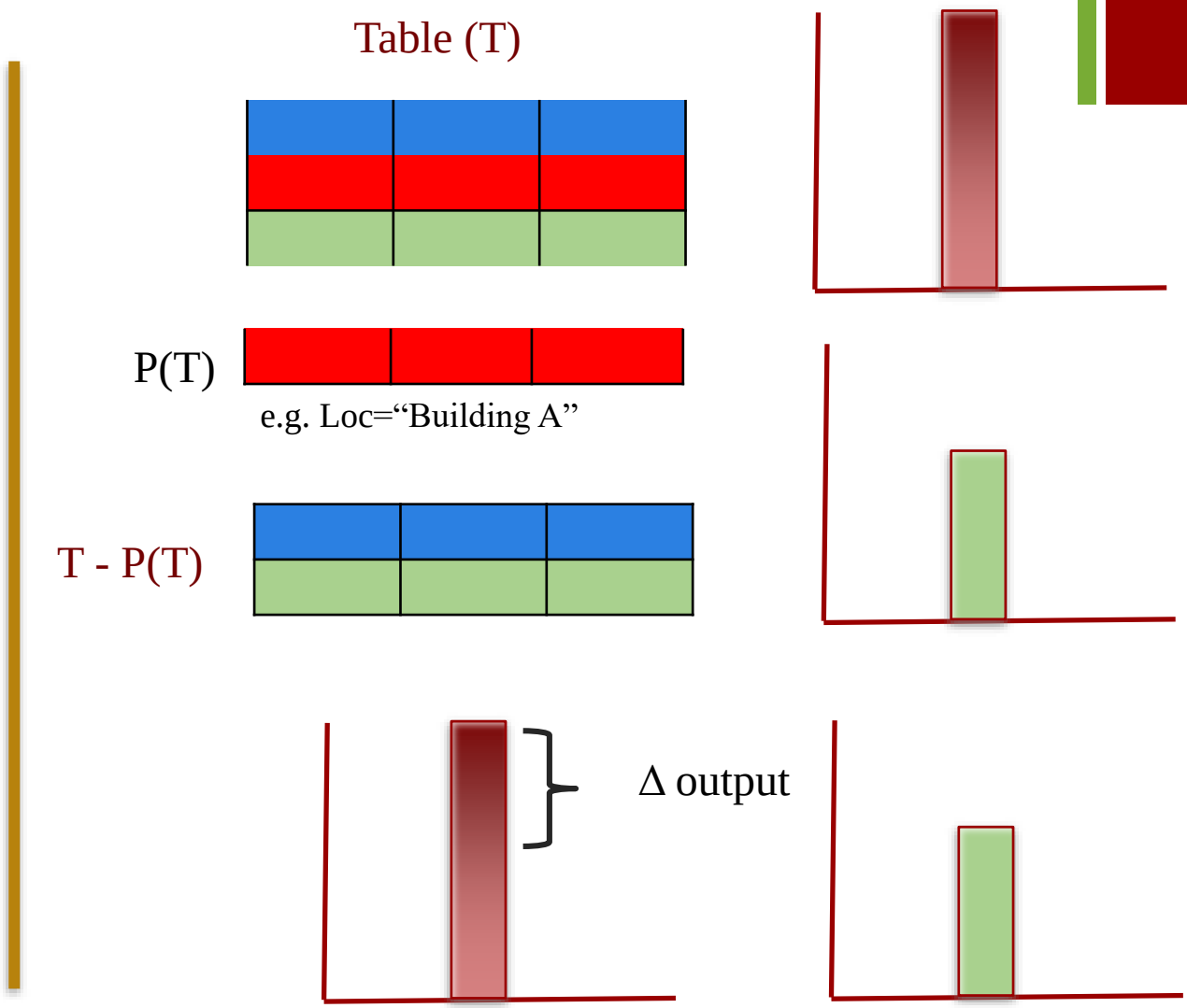
[1] Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB Endow., 6(8):553–564, June 2013.

Find tuples (predicates) that influence the Outlier

Predicate Influence=

$$\frac{\Delta\ output}{|P(T)|}$$

The ratio between the change in the output and the number of tuples that satisfy the predicate



[1] Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB Endow., 6(8):553–564, June 2013.

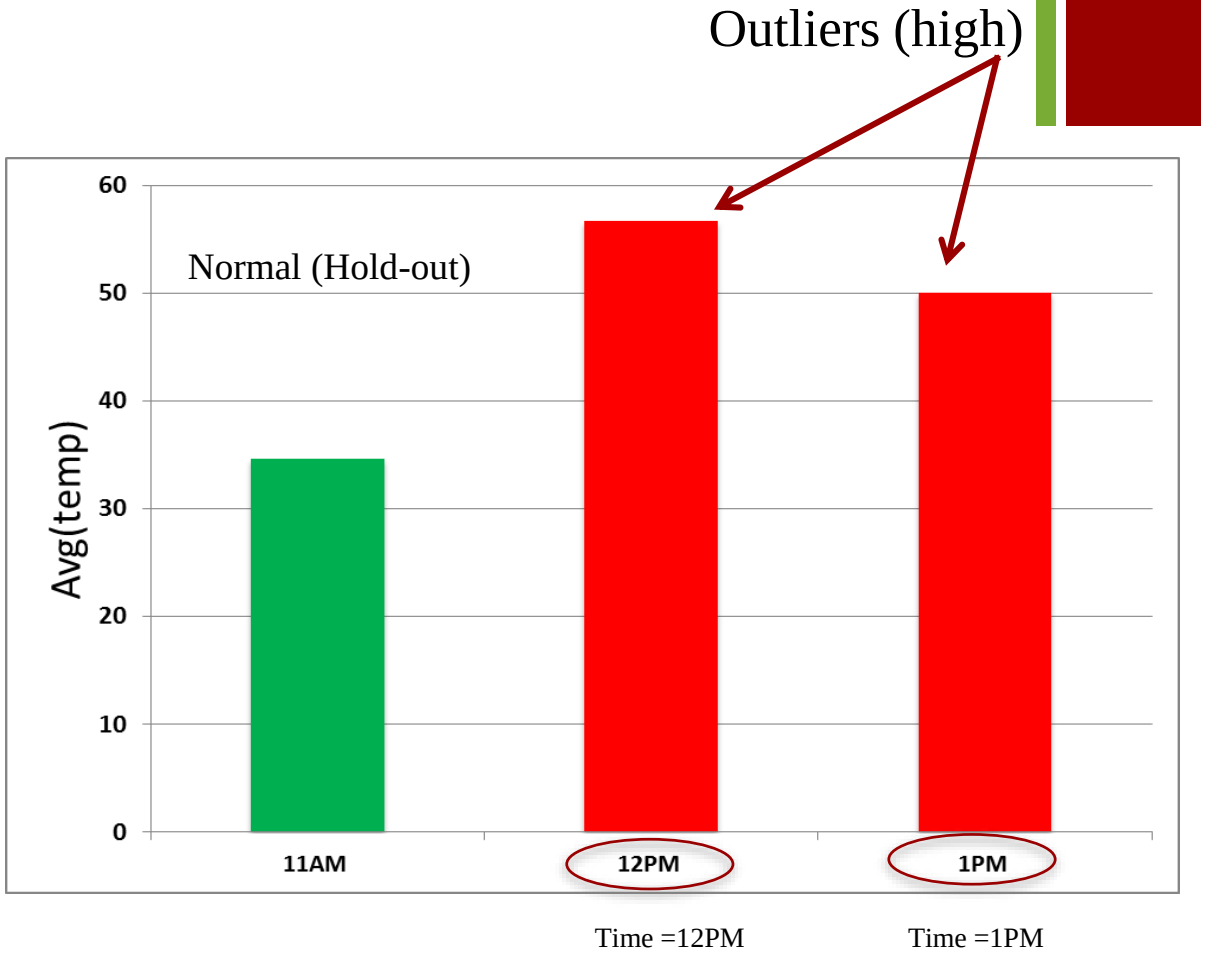
Scorpion: Which value in GROUP_BY output will cause the outlier disappear?

Inputs

Outliers and hold-out results

Finds

Predicates that responsible of outliers
i.e. adding predicate to original aggregate query "removes" outliers and maintains results look normal



SELECT avg(temp), time
FROM sensors GROUP BY time

[1] Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB Endow., 6(8):553–564, June 2013.

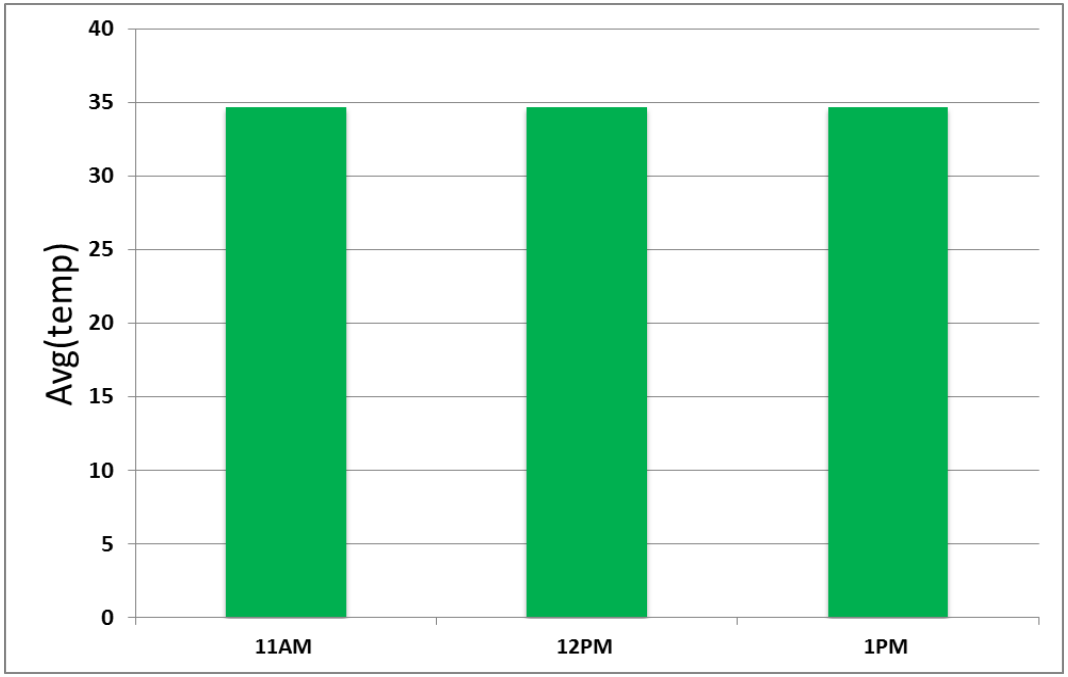
+ WHERE Clause to make the outlier disappear

Inputs

Outliers and hold-out results

Finds

Predicates **P** that are responsible for outliers
i.e. adding predicate to original aggregate query "removes" outliers and maintains results look normal



SELECT avg(temp), time
FROM sensors GROUP BY time
WHERE P

[1] Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB Endow., 6(8):553–564, June 2013.

NoSQL Technology

- Cloud Platform: - a viable alternative to relational databases operating on **cluster servers**
- No Schema: - **different types of data** is collected, stored, accessed without a Schema
- Data Fusion: - a data integration technique for **multi-source data**
- Flexible Access: - A **query model** accessing data without using traditional SQL

http://www.couchbase.com/why-nosql/nosql-database
http://smoothspan.wordpress.com/2011/07/22/nosql-is-a-premature-optimization/

NoSQL

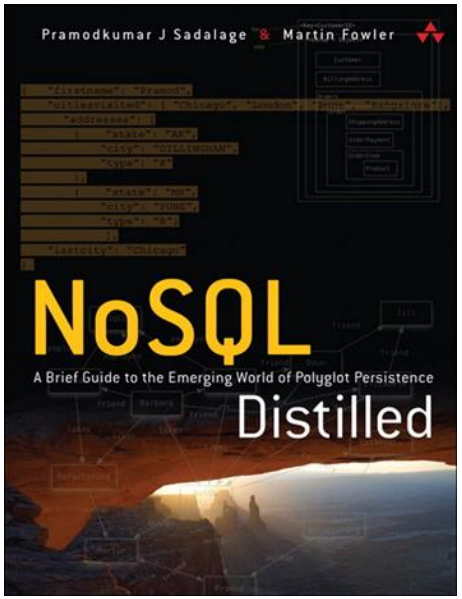
- To store objects using **key-values**.
- To add/delete/query massive arrays and **still allow for persistence** and fault tolerance.
- To implement large data query on **MapReduce** framework

  e.g an SQL code:

  SELECT * FROM users WHERE age > 10

  can be implemented by a Javascript function that runs against every item in the database:

```
function (doc)
{
    if (doc.objType == "users") {
        if (doc.age > 10) {
            emit(doc._id, null)
        }
    }
}
```



http://stackoverflow.com/questions/2328169/nosql-crash-course-tutorial

http://martinfowler.com/books/nosql.html

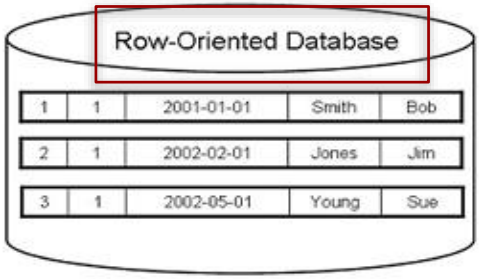
NoSQL Systems

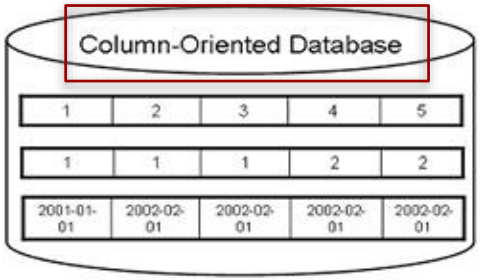
- **CouchDB** - a document-oriented database that can be queried and indexed in a MapReduce fashion using JavaScript. CouchDB also offers incremental replication with bi-directional conflict detection and resolution.
- **MongoDB** - a scalable, high-performance, open source, document-oriented database system.
- **Hadoop and HBase** - The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. HBase is the Hadoop database system. It supports random, real-time read/write access to “Big Data.”

http://www.thedatabasesite.com/page105.html

A change of traditional perception on data storage

Emp_no	Dept_id	Hire_date	Emp_ln	Emp_fn
1	1	2001-01-01	Smith	Bob
2	1	2002-02-01	Jones	Jim
3	1	2002-05-01	Young	Sue
4	2	2003-02-01	Stemle	Bill
5	2	1999-06-15	Aurora	Jack
6	3	2000-08-15	Jung	Laura





Row Store vs. Column Store

Record #	Name	Address	City	State
0003623	ABC	125 N Way	Cityville	PA
0003626	Newburg	1300 Forest Dr.	Troy	VT
0003647	Flotsam	5 Industrial Pkwy	Springfield	MT
0003705	Jolly	529 S 5th St.	Anywhere	NY

Record #	Name	Address	City	State
0003623	ABC	125 N Way	Cityville	PA
0003626	Newburg	1300 Forest Dr	Troy	VT
0003647	Flotsam	Industrial Pkwy	Springfield	MT
0003705	Jolly	529 S 5th St.	Anywhere	NY

www.enterpriseirregulars.com
http://www.enterpriseirregulars.com/39209/the-real-potential-impact-of-sap-hana/

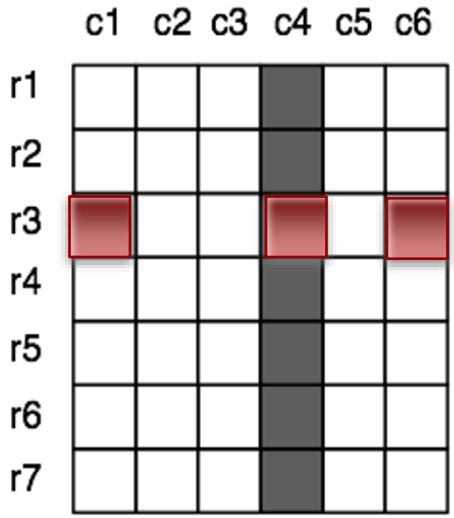
+ Row vs. Column Data Storages

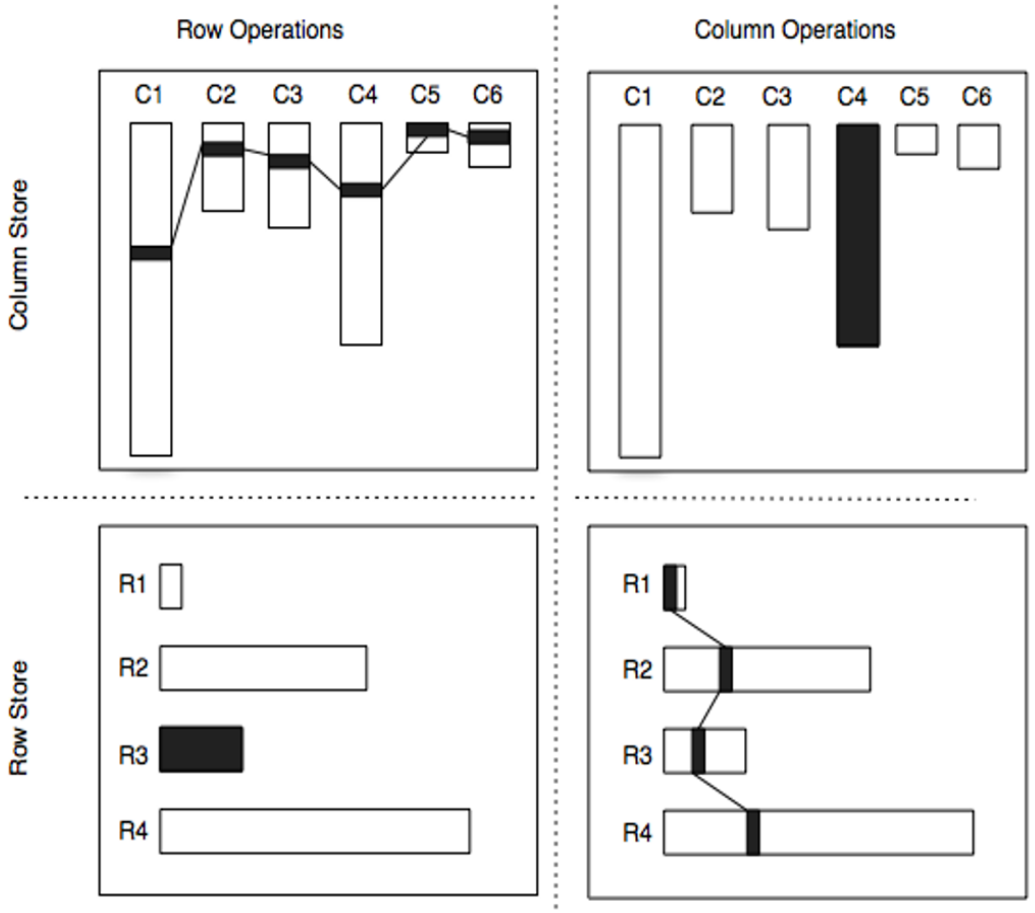
- **Why Row?**
  - Efficient when **many columns of a single row are required** at the same time, and when row-size is relatively small, as the entire row can be retrieved with a single disk seek.
  - Efficient when **writing a new row** if all of the row data is supplied at the same time, as the entire row can be written with a single disk seek in row-oriented storage.
  - Well-suited for **OLTP-like workloads** which are more heavily loaded with interactive transactions.
- **Why Column?**
  - Efficient on **aggregation operation** over many rows but only for a notably smaller subset of all columns of data, because reading that smaller subset of data can be faster than reading all data.
  - Efficient when **inserting new values of a column for all rows** at once, because that column data can be written efficiently and replace old column data without touching any other columns for the rows.
  - Well-suited for **OLAP-like workloads** (e.g., data warehouses) which typically involve a smaller number of highly complex queries over all data (possibly terabytes).

Query Processing in Row or Column Databases

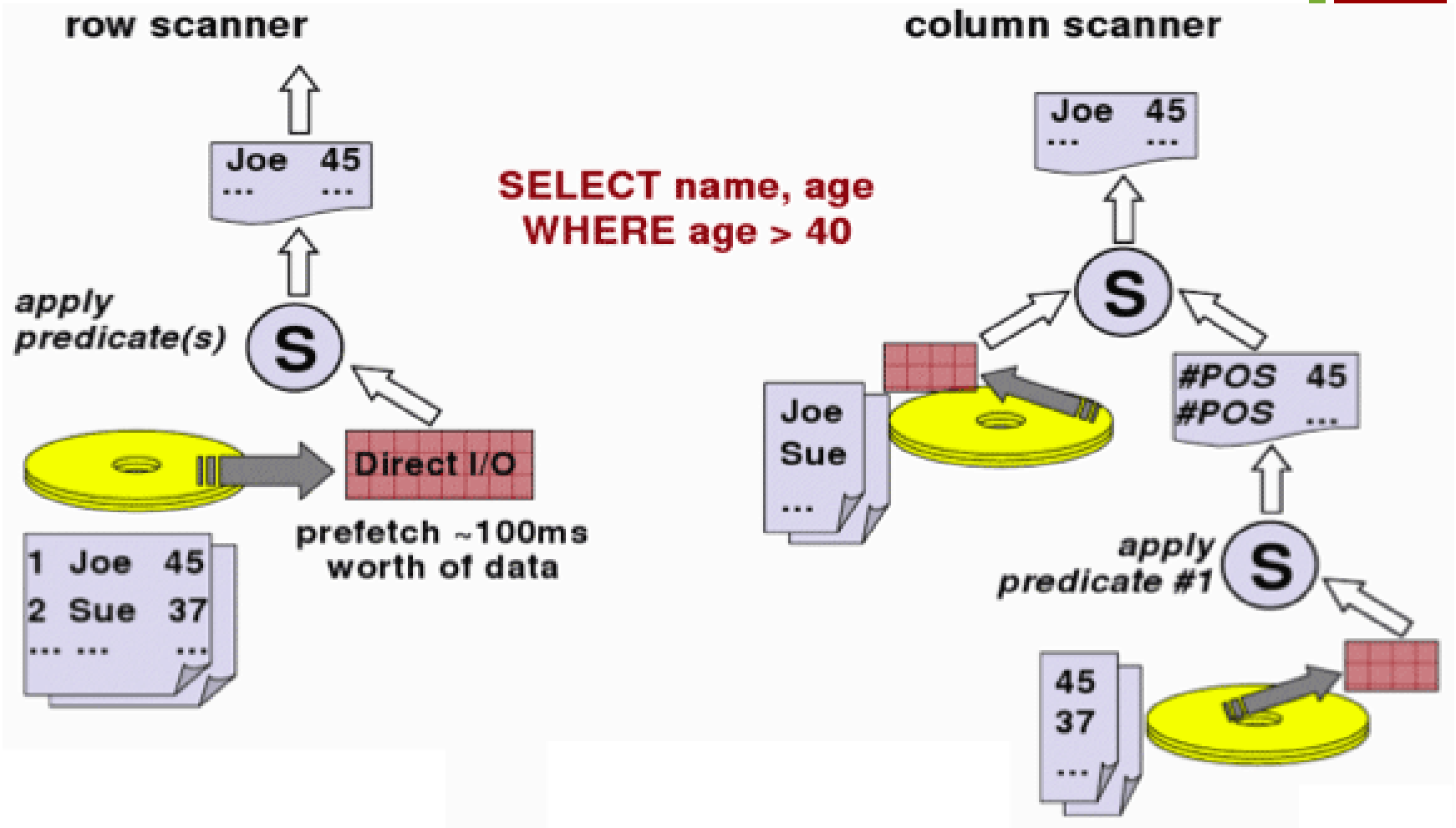
Row vs. Column Stored data

SELECT c1,c4, c6 FROM table WHERE c4 < ?





+ Query execution: Row vs. Column databases



Row Stored Data **Row Stored Data**

+ Summary

- Big Data is of three Utilities:
  - Connecting Dots – "*From small to big*"
  - Discovering Specifics – "*From big to small"*
  - Data Inferencing – "*Knowing unknown"*
- Network Science is a Foundation of Big Data Studies
  - Scale-Free Networks (*Structure is independent from the size*)
  - Different types of networks (*Centralised, Decentralised, Distributed)*
  - Curse of Dimensionality
  - Computations of Network *Centrality*, *Modularity*, *Reachability*, etc
- Row Storage vs Column Storage in Database Systems
  - SQL vs. NOSQL
  - OLTP with row storage and OLAP with column storage
  - Frequent update operations with row storage and read-only data with column storage

Readings

- "How Big Data is drawing a global picture?"
  https://shorthand.uq.edu.au/changemakers/issue2/how-big-data-is-drawing-a-global-picture/
- Book references listed in the slides in this lecture
- Eugene Wu and Samuel Madden. Scorpion: Explaining away outliers in aggregate queries. Proc. VLDB, Endow., 6(8):553–564, June 2013
- Row Store vs. Column Store Databases:
  https://dzone.com/articles/row-store-and-column-store-databases#:~:text=Row%20stores%20have%20the%20ability,hours%20are%20completed%20in%20seconds

Next Week: Course Revision & Preparation of the Final Exam