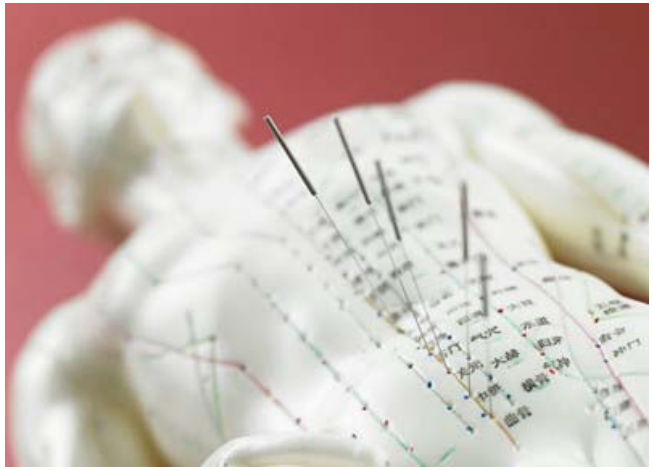
Big Data Analytics: Foundations and Applications

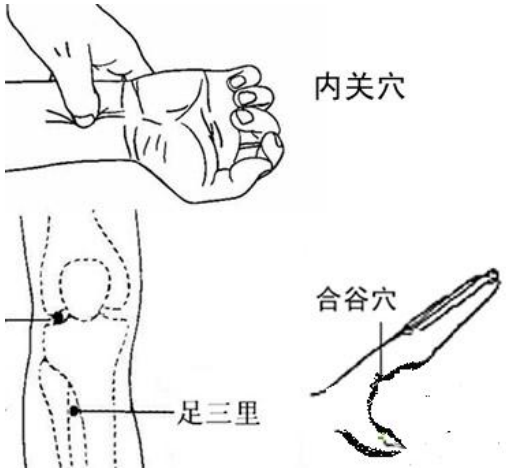
http://www.comp.nus.edu.sg/~atung/northcluster7_talk_english.pdf

Anthony K. H. Tung
atung@comp.nus.edu.sg
School of Computing
National University of Singapore

What is Big Data Analytics?

- **You leave data trails wherever you go**
  凡走过必留下数据
- Perform reverse engineering on data to construct models that represent the scenario on which the data is generated
- "Essentially, all models are wrong, but some are useful.“ --- Box, George E. P.; Norman R. Draper (1987).





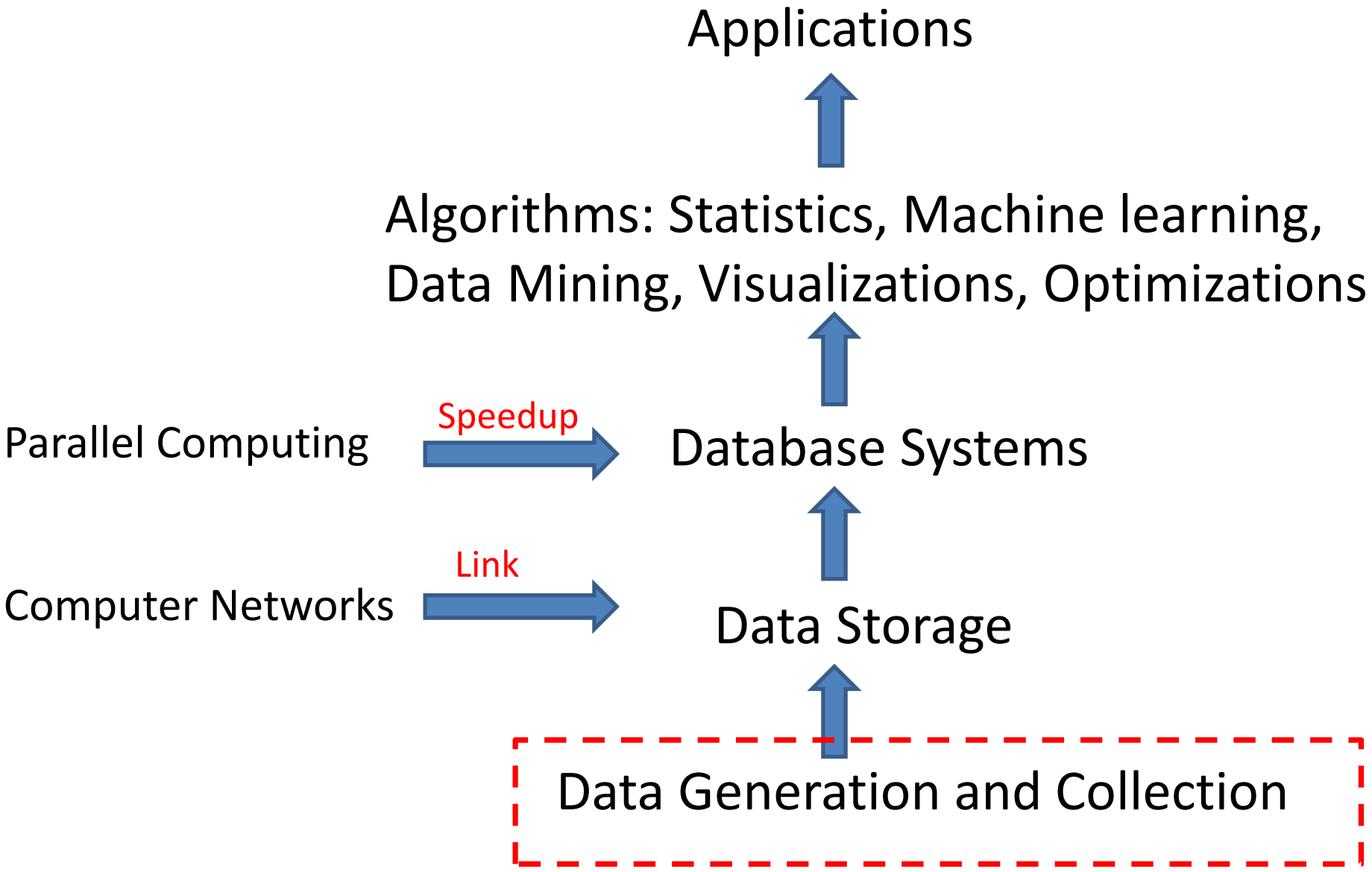
Outline

- Big Data: Characteristics and Components 
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

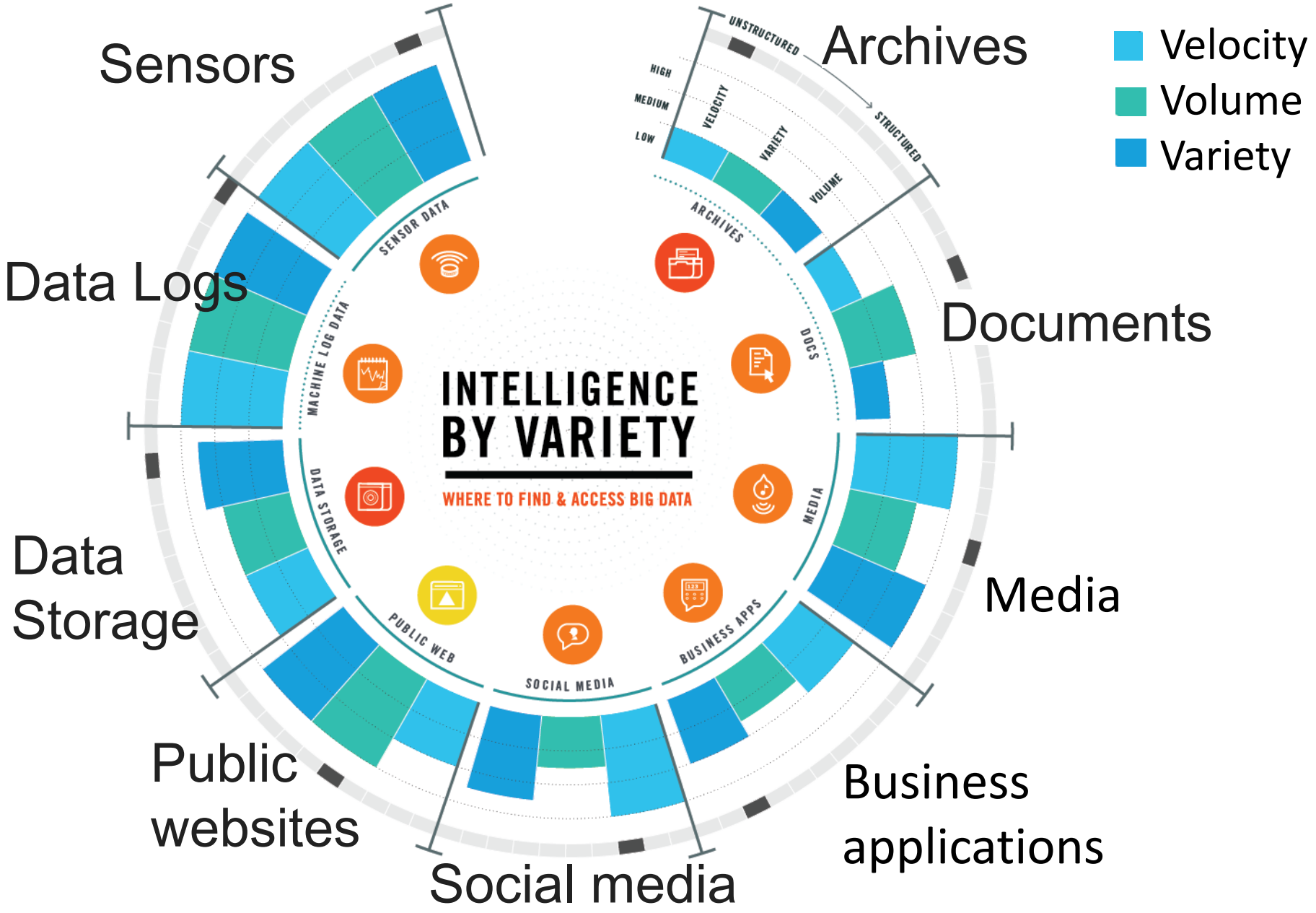
3Vs of Big Data

- Volume: (Solvable by putting more computing hardware resources)
  - huge amount of data
- Velocity: (Solvable by putting more computing hardware resources)
  - High throughput
  - Timeliness
  - Require fast processing speed
- Variety: (**NOT** solvable by putting in more computing hardware resources)
  - Structured data: documents, media
  - Unstructured data: websites, logs, streams
  - Mixed data types

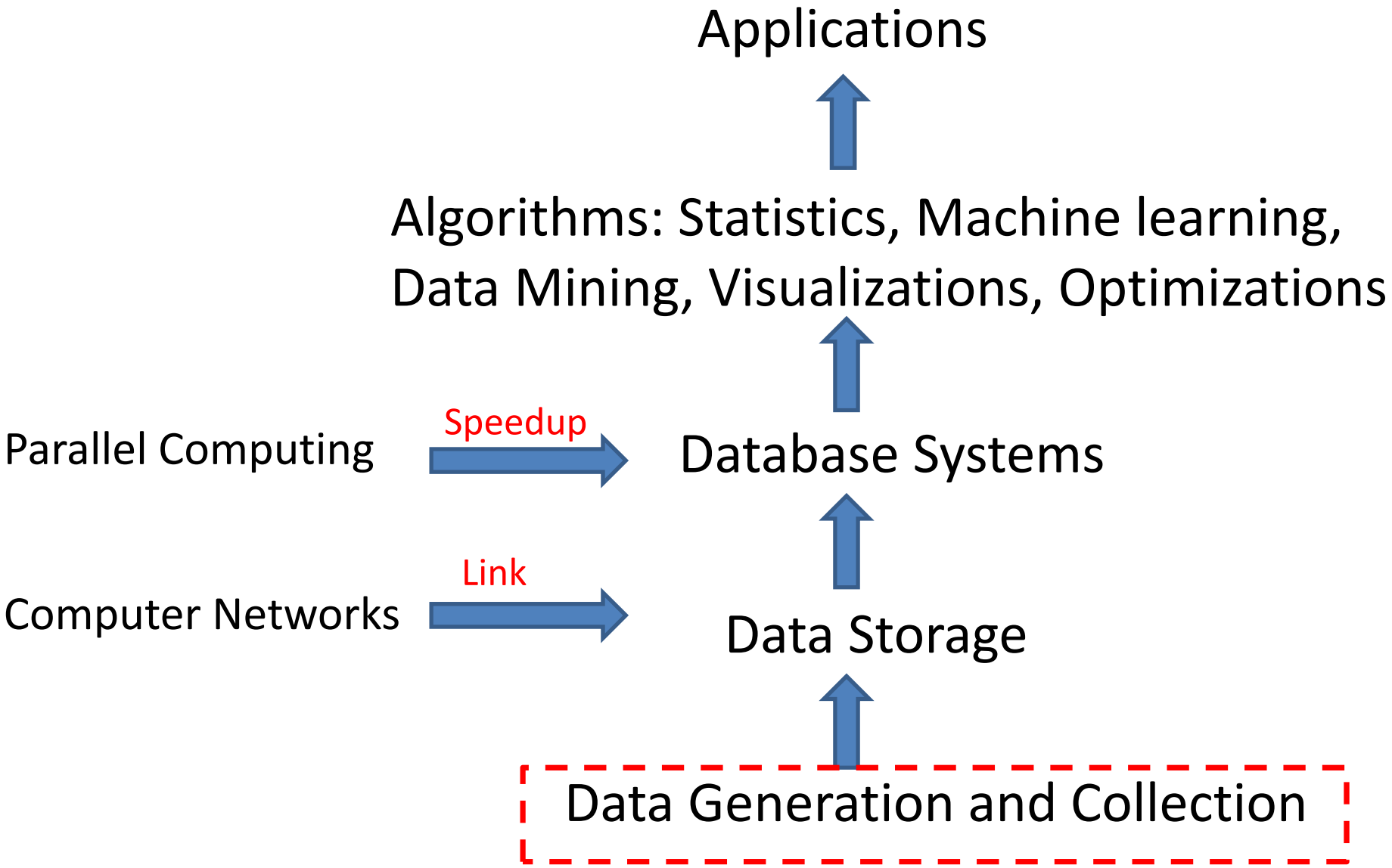
Overview of Big Data Components



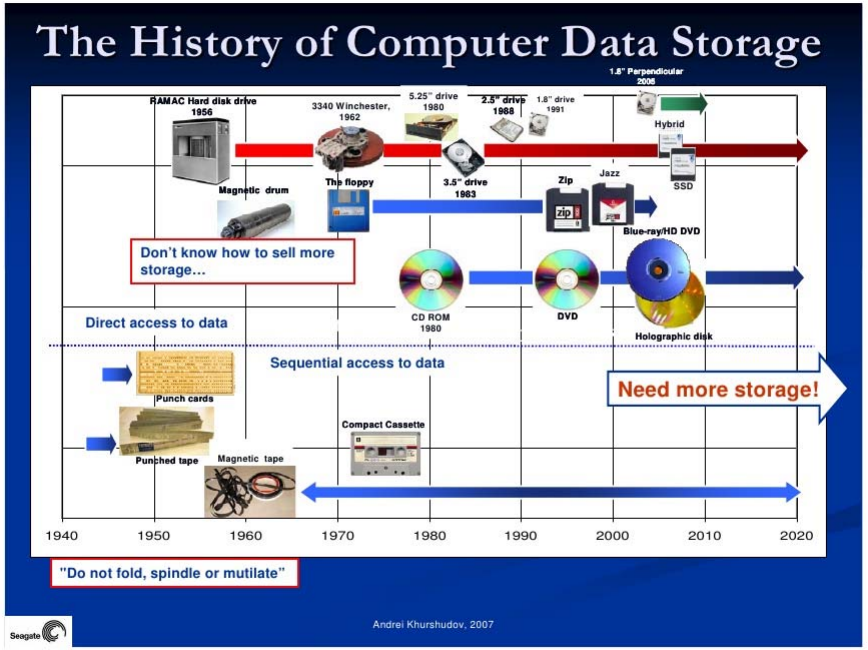
Big Data Components(I):
Data generations and collections

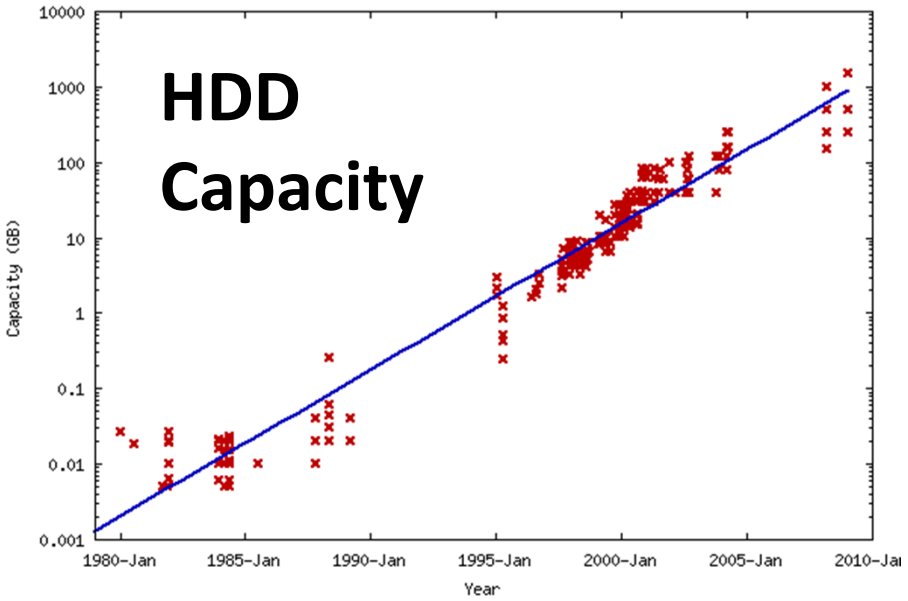


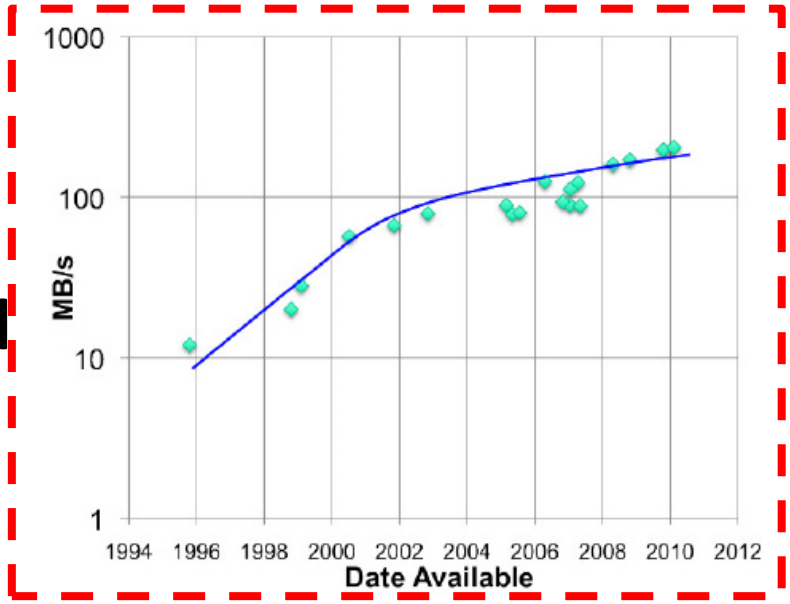
Overview of Big Data Components

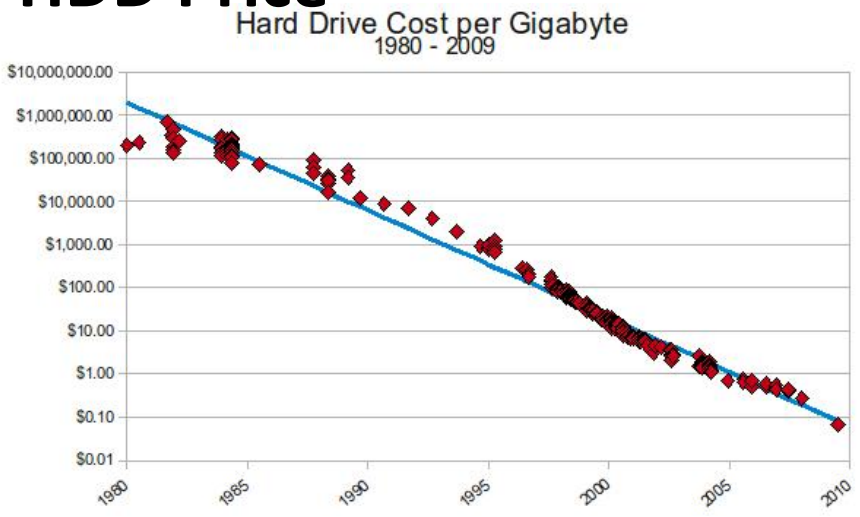


Big Data Components(II): Data storage

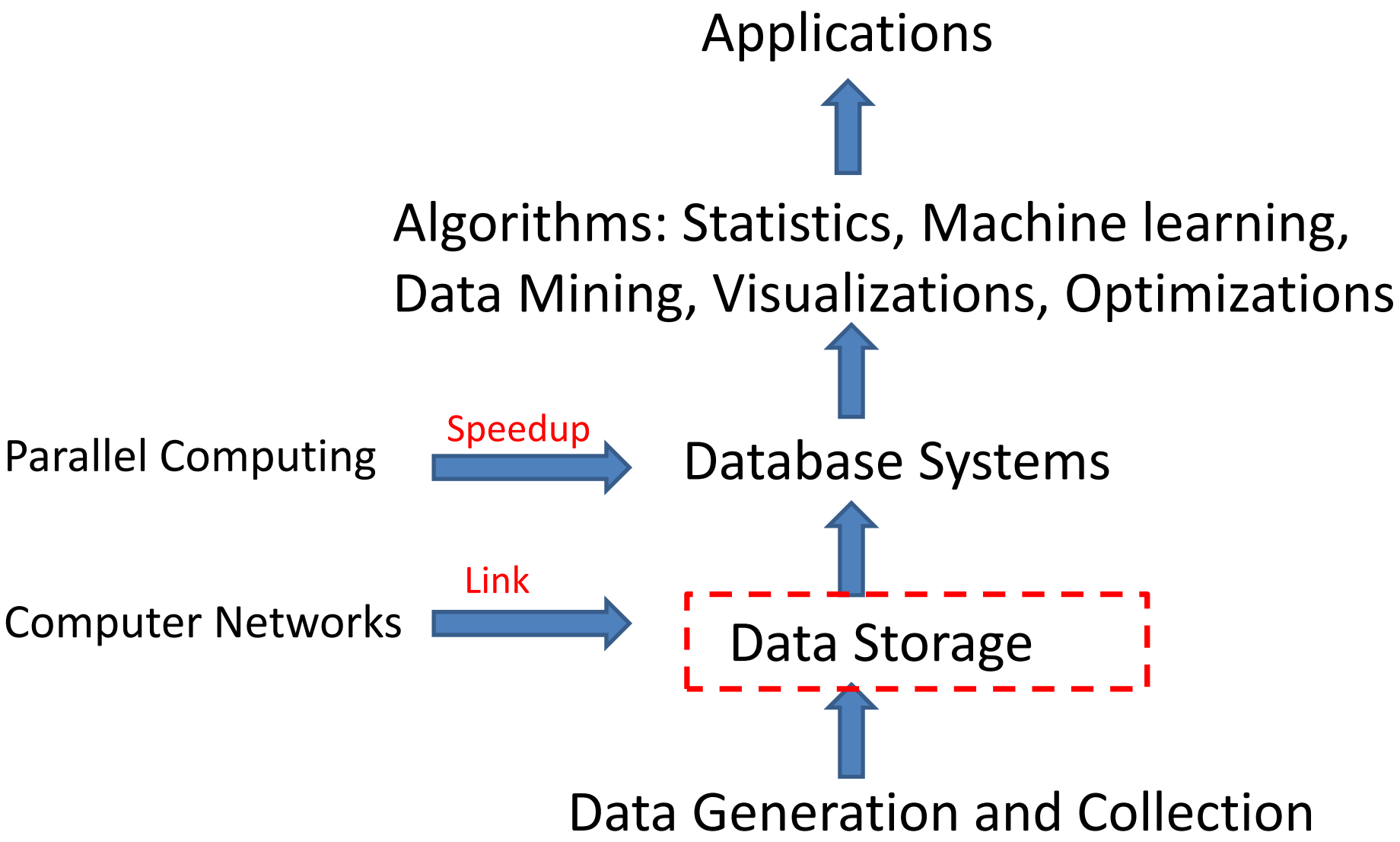




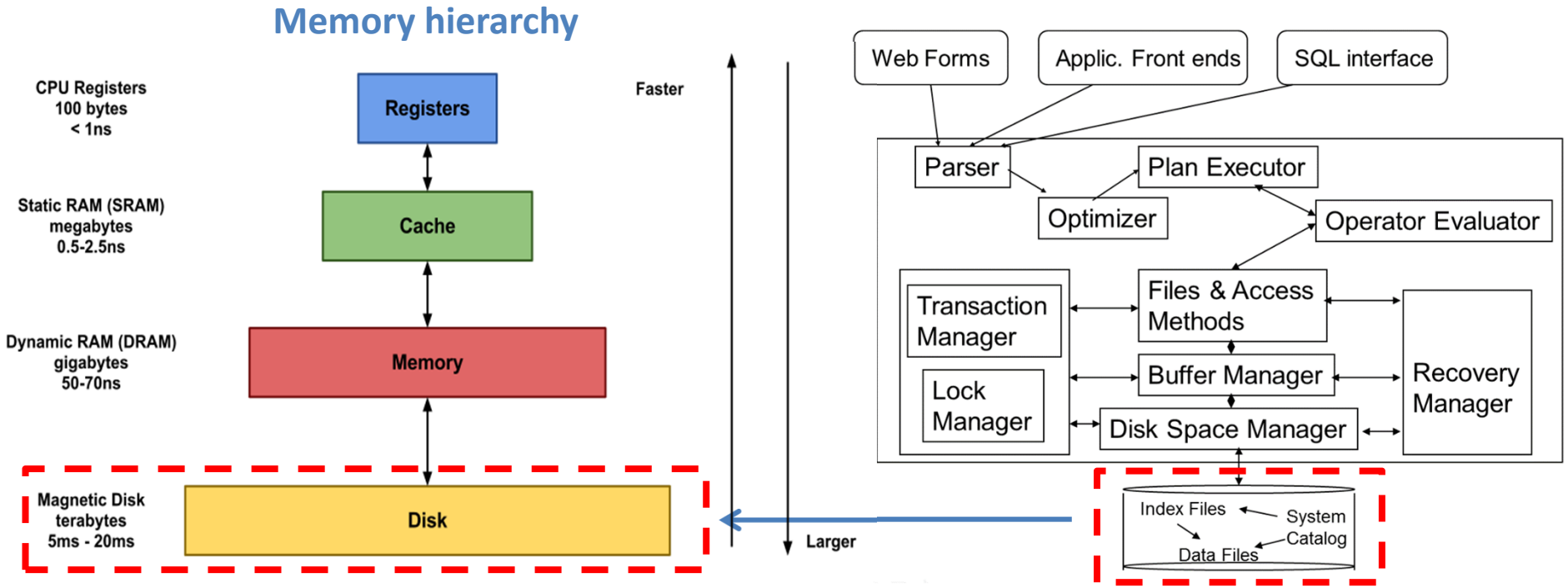




Overview of Big Data Components



Big Data Components(III): Database systems and techniques



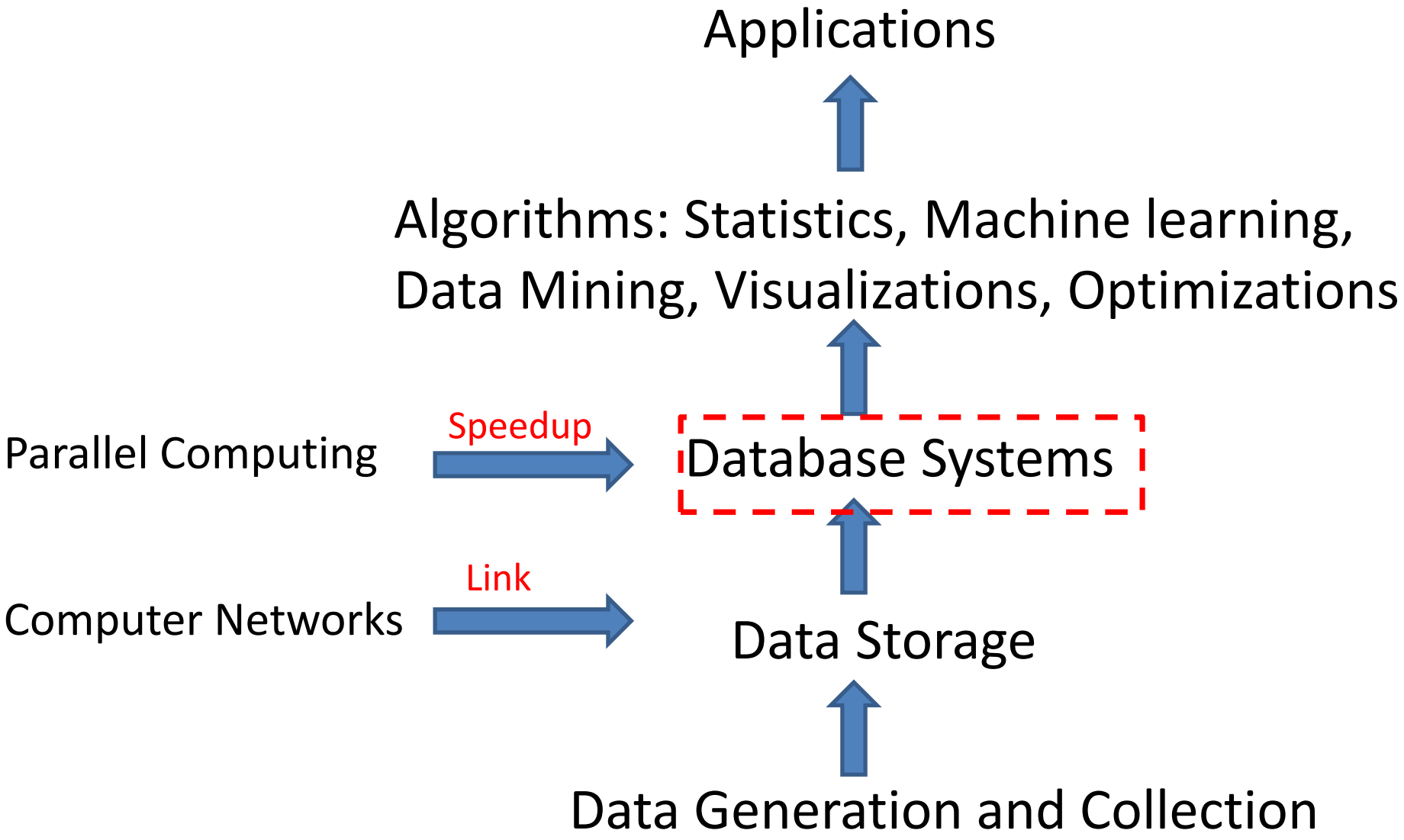
Concept: store the frequently-used data in the memory, to avoid accessing disk.

Problem: How to manage the data in the memory/hard disk?

Solution: Use database systems

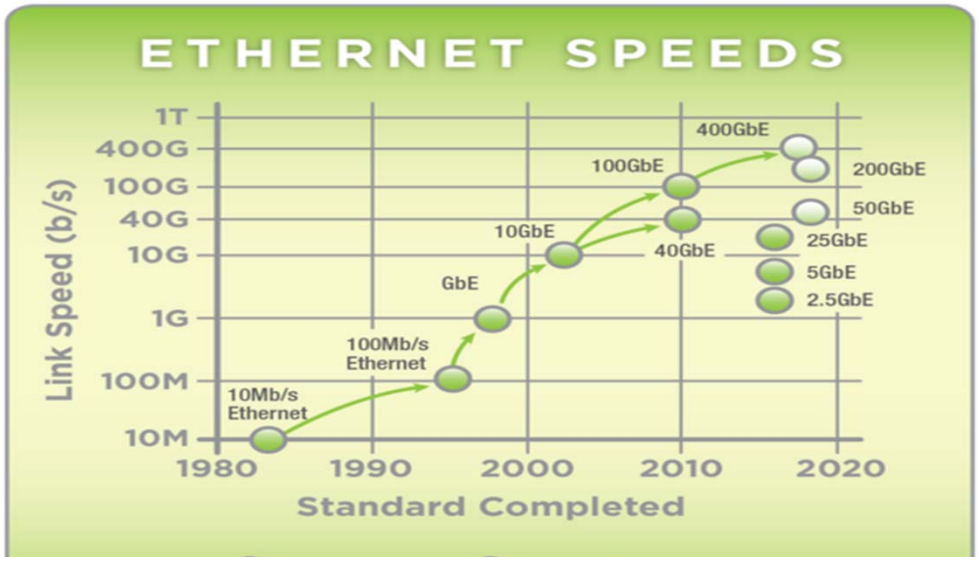
1. **Manage dataflow in memory/hard disk**
2. **Build and maintain index, speedup query processing**
3. **Use locks to guarantee reliability.**
4. **Equip query languages (SQL or NoSQL) to support high-level applications**

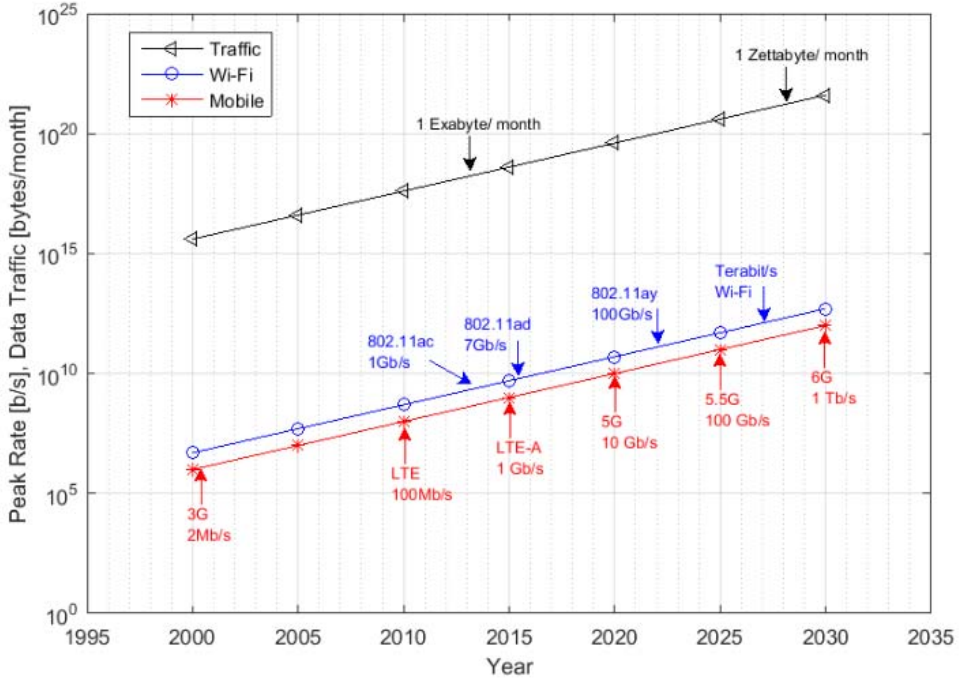
Overview of Big Data Components



Big Data Components (IV): Computer Networks

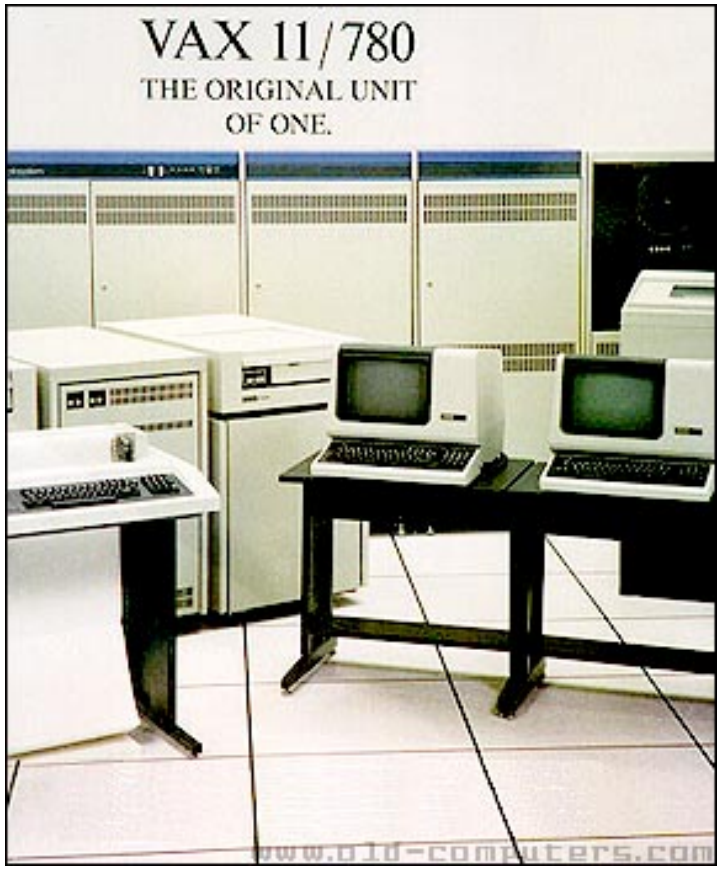
- Data highway
- Wired network
  - Data communication between servers in data centers
- Wireless network
  - Data communication between devices and data centers
  - Wi-fi： short distance, fast
  - Mobile： long distance, slow
- That which is long divided must unify; that which is long unified must divide. 合久必分，分久必合
  - Your future cell phone might have the same computational power as today's high-end servers.
  - Data communication is for merging and analyzing data.





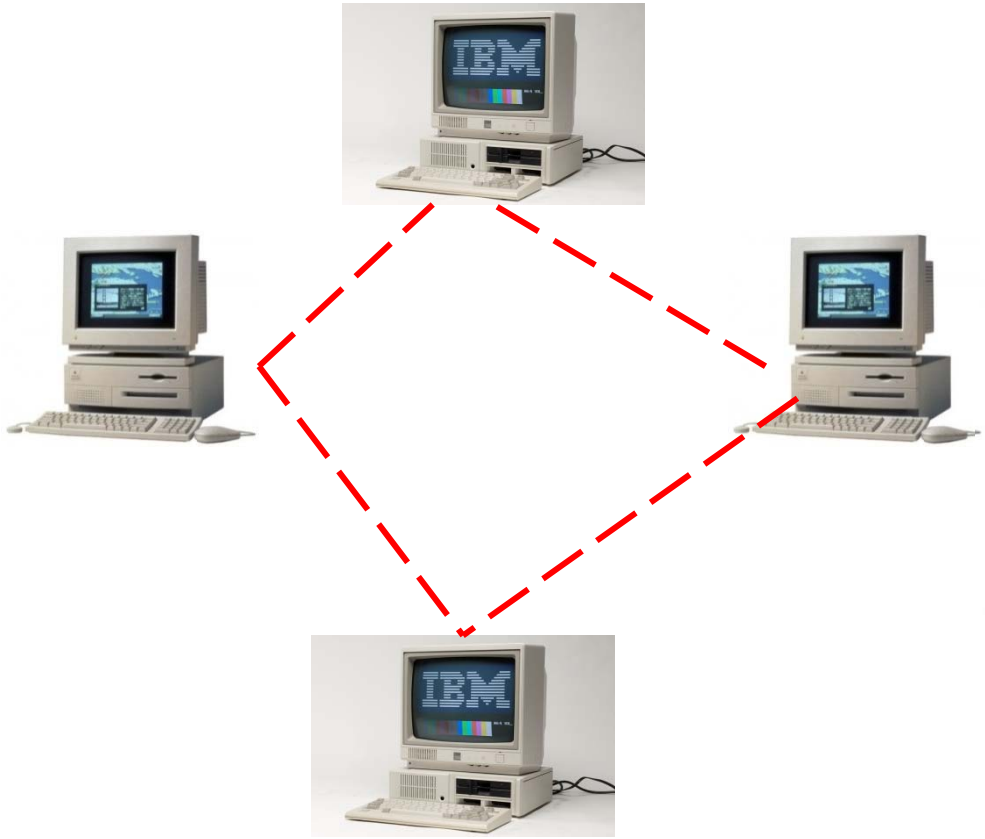
That which is long divided must unify; that which is long unified must divide

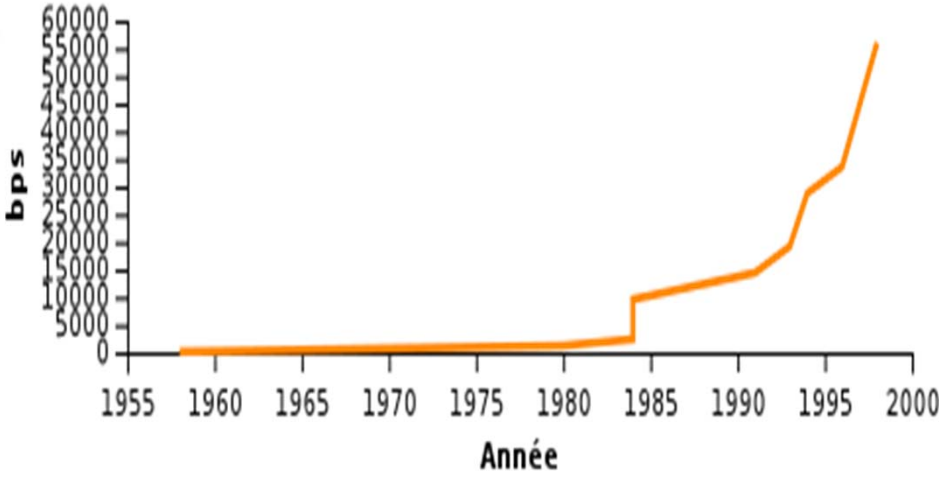
- Initial cost of new hardware is expensive, so everyone share the hardware through simple terminals



That which is long divided must unify; that which is long unified must divide

- Then PCs become cheap enough for everyone to own one. Low network bandwidth however ensure isolated processing





That which is long divided must unify; that which is long unified must divide

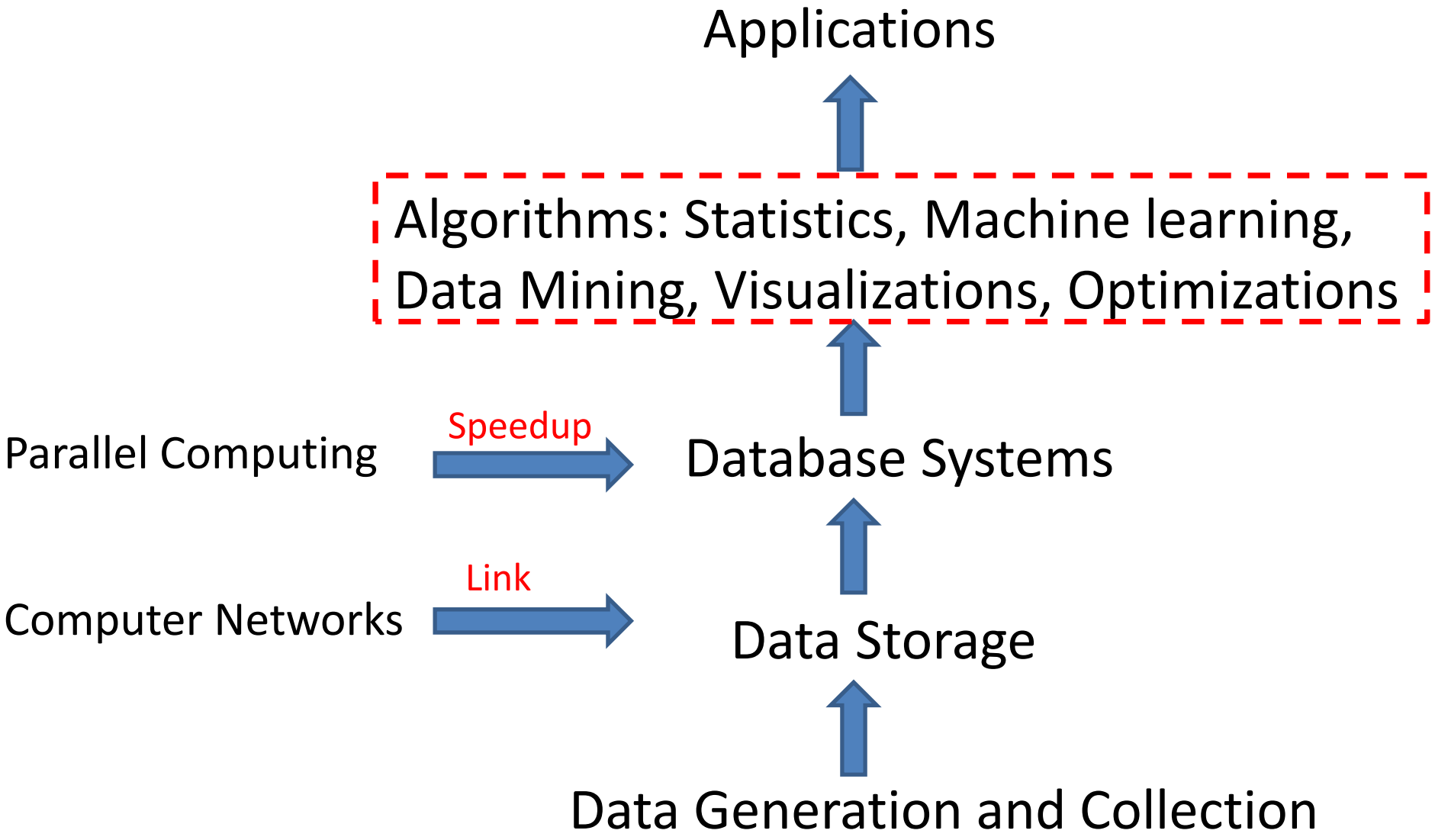
- Further drop in price of computing hardware and growth in both wired and wireless network then result in the current trend in cloud-based processing

Cloud(clustered PCs)

High Speed Data Transfer

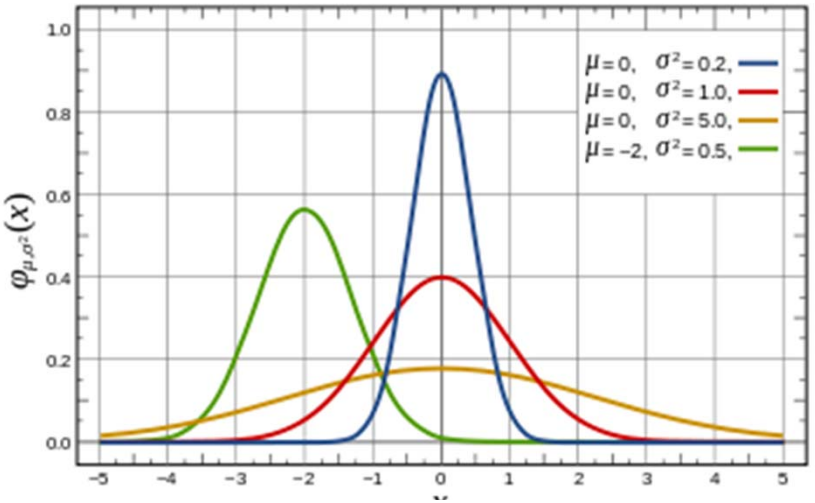
What if these devices become much more powerful and energy efficient while maintaining low cost?

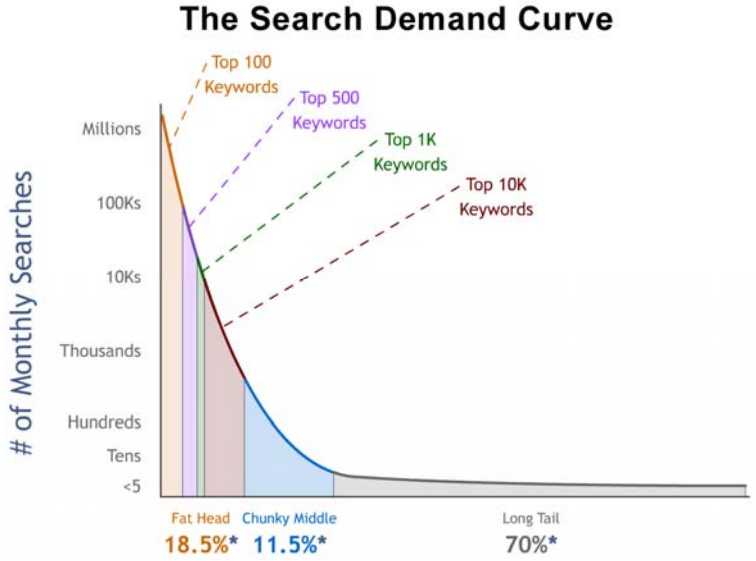
Overview of Big Data Components

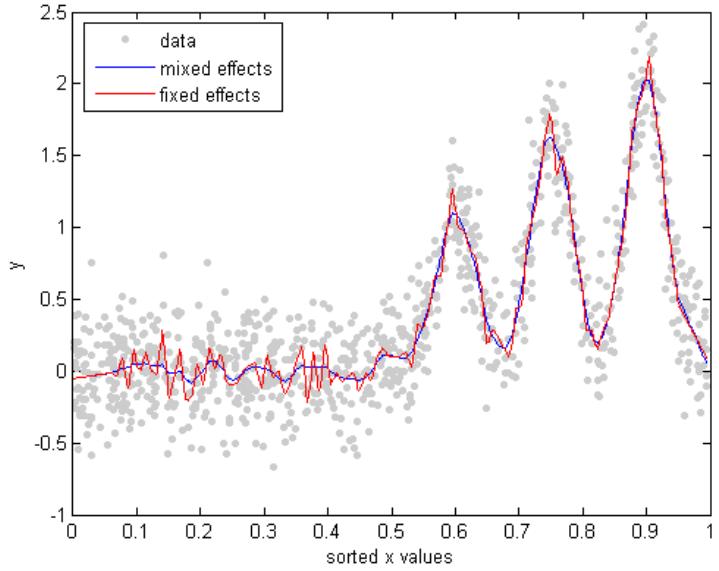


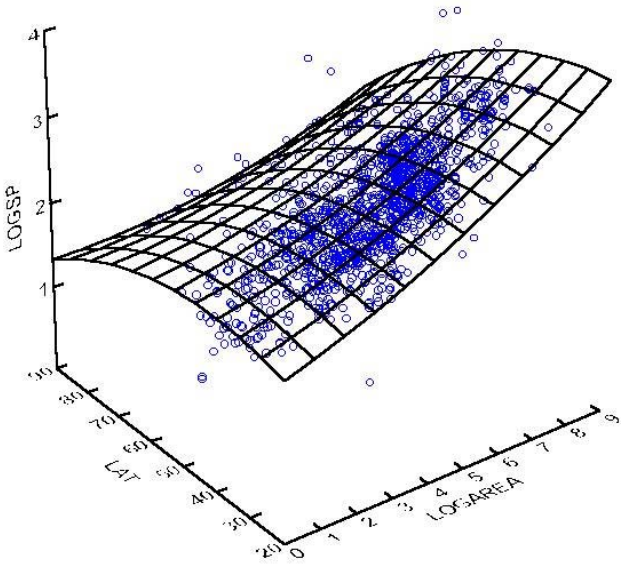
Big Data Components (V): Algorithms (I)

- Statistics: understand data distribution using statistical models



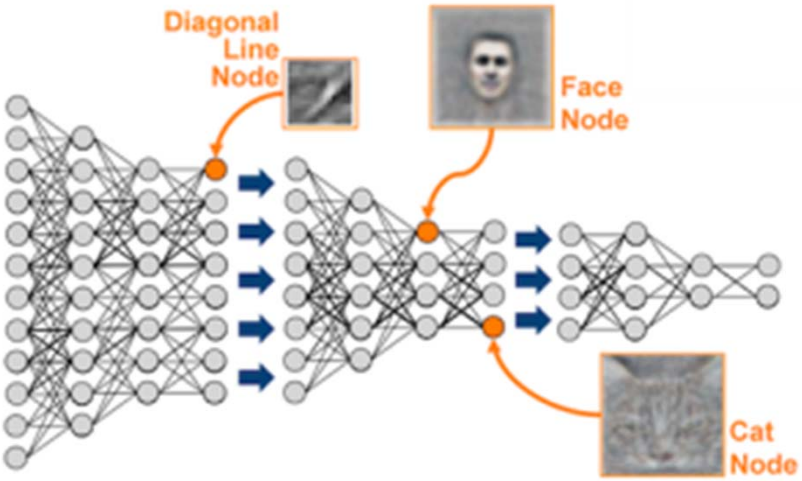


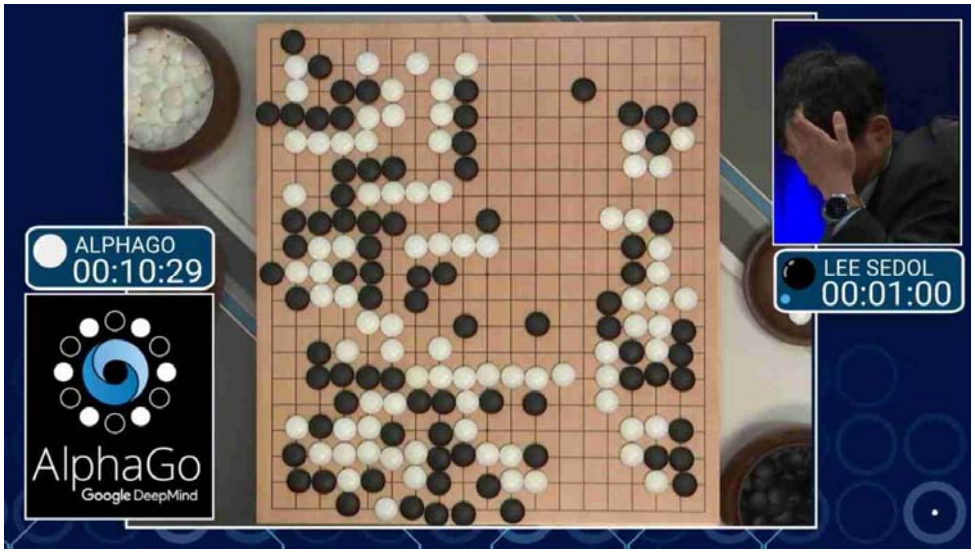




Big Data Components (V): Algorithms(II)

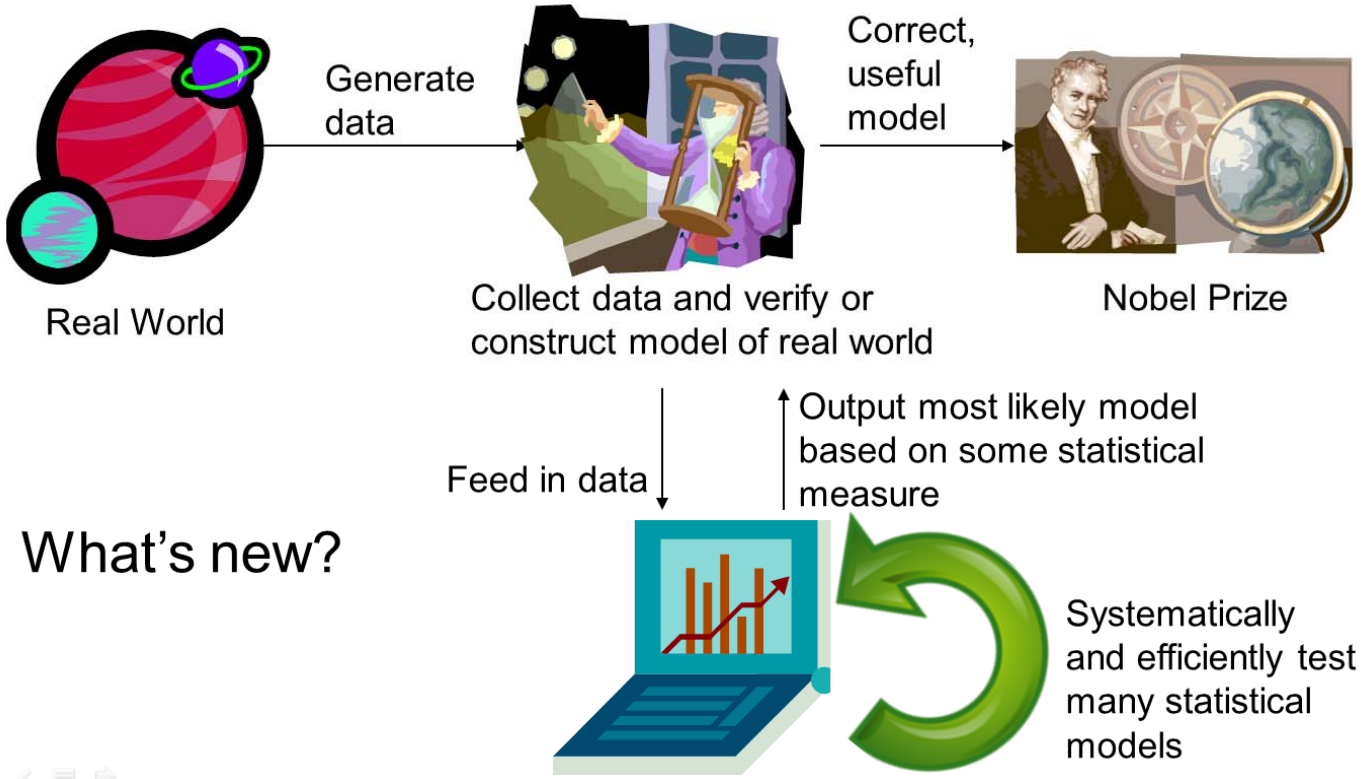
- Machine learning: borrows models from statistics, trains models efficiently using modern computers, and focuses on prediction tasks.
  - Can be operated as a **black box**





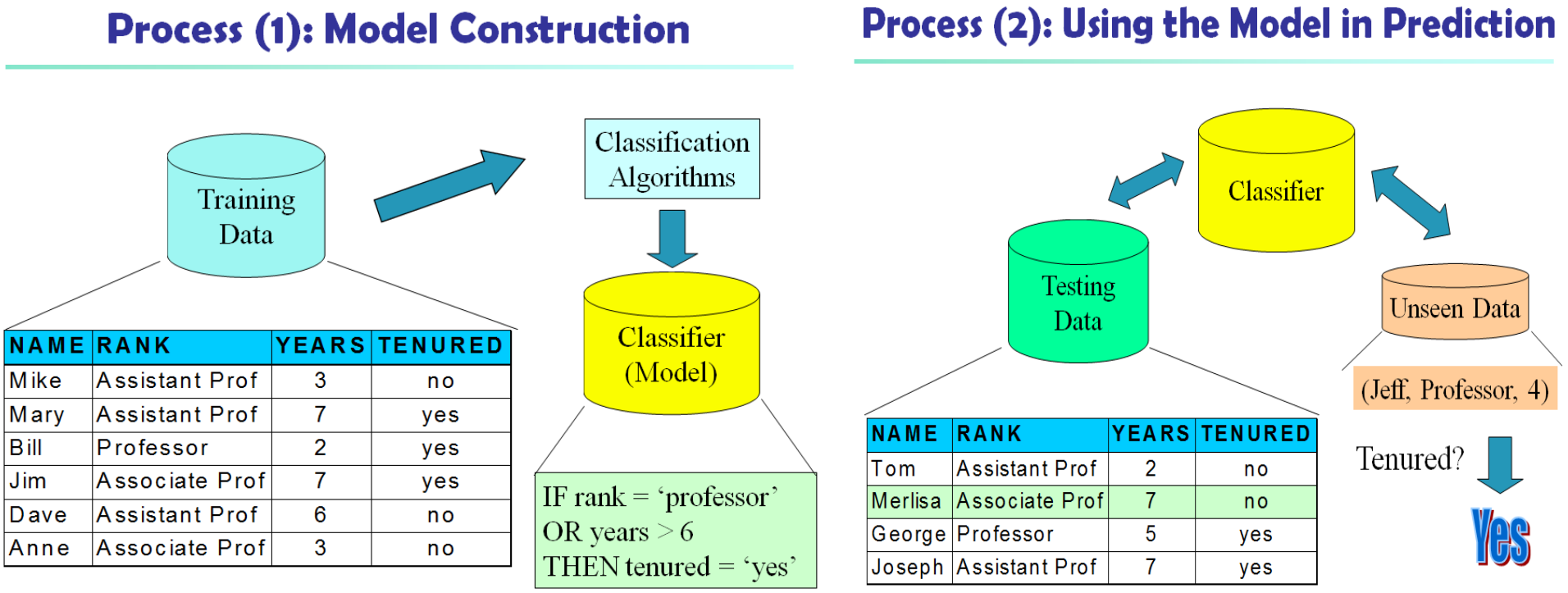
Big Data Components (V): Algorithms(III)

- Data mining:
  - Based on methods from statistics and machine learning, involves database and data management aspects to improve efficiency
  - Goal: extract knowledge



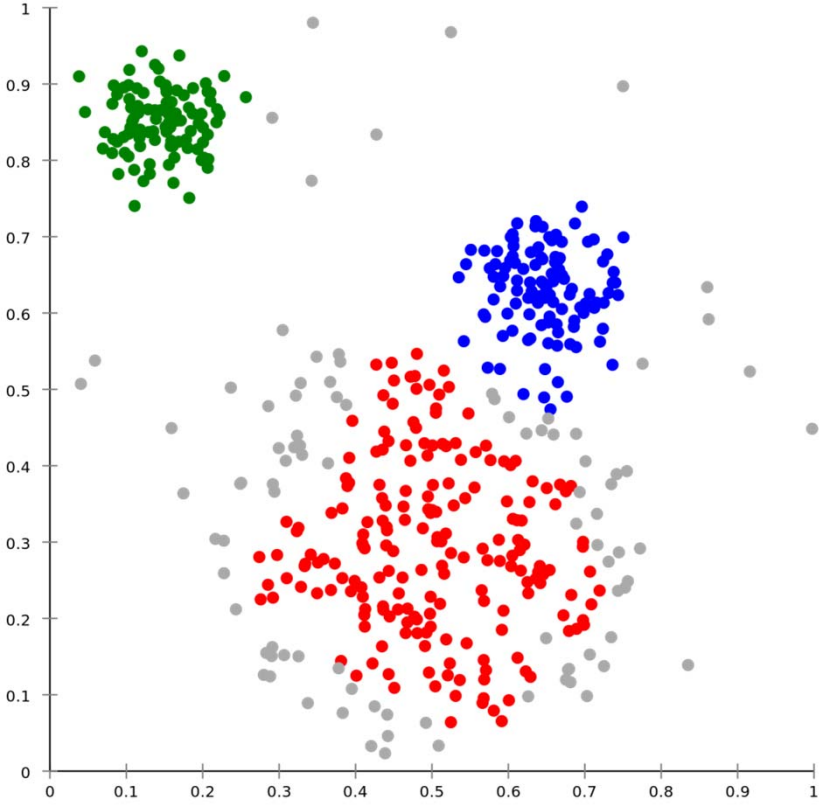
Big Data Components (V): Algorithms(IV)

- Data mining techniques(I): **Classification**



Big Data Components (V): Algorithms(V)

- Data mining techniques(II): **Clustering**



Separate into k=2 clusters

age	salary	credit	sex
20	30k	poor	M
25	76k	good	F
30	90k	good	F
40	100k	poor	M
50	110k	good	F
60	50k	good	M
70	35k	poor	F
75	15k	poor	M

Error Tolerance（容错性）:

age	salary	credit	sex
5	25k	0	0

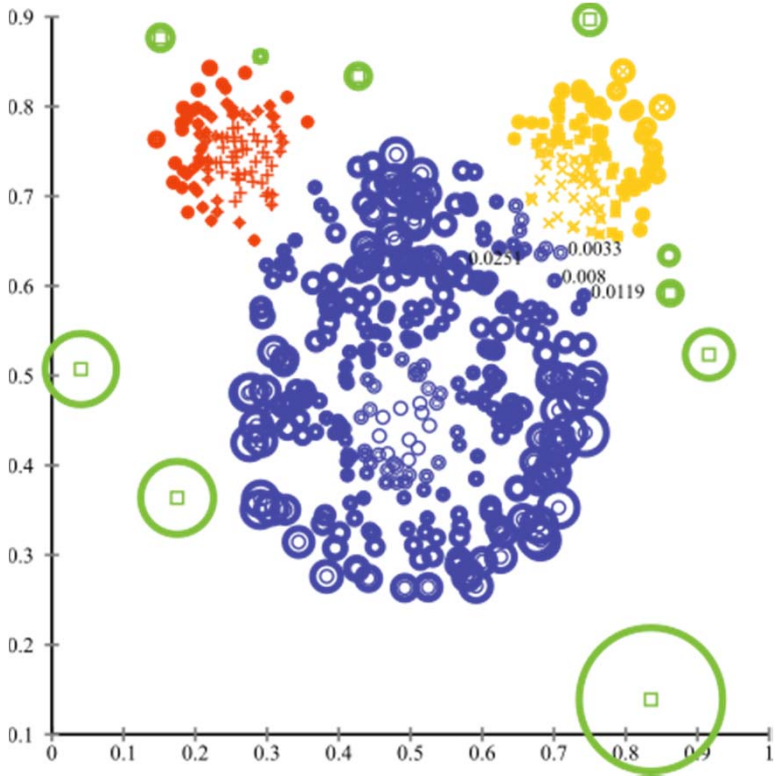
	RRid	age	salary	credit	sex
代表1	1	70	30k	poor	M
代表2	2	25	90k	good	F

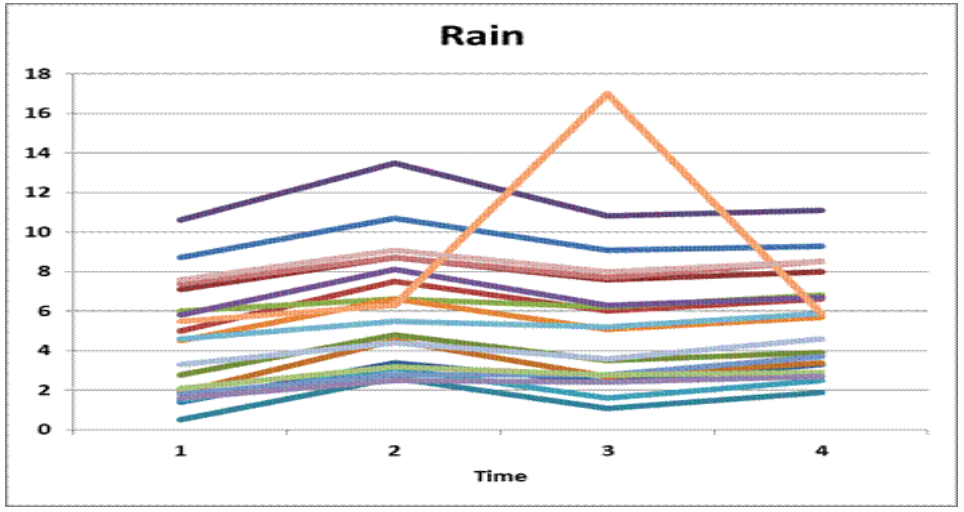
age	salary	credit	sex
20	30k	poor	M
40	100k	poor	M
60	50k	good	M
70	35k	poor	F
75	15k	poor	M

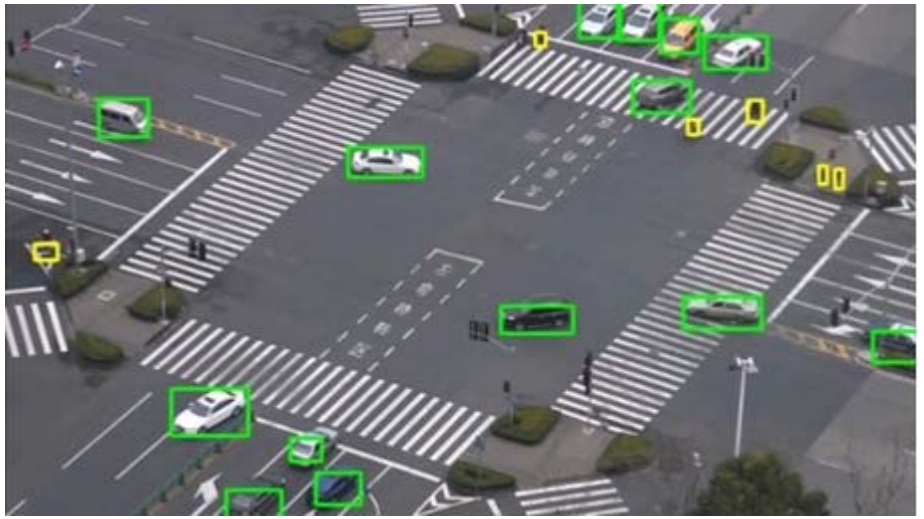
age	salary	credit	sex
25	76k	good	F
30	90k	good	F
50	110k	good	F

Big Data Components (V): Algorithms(VI)

- Data mining techniques(III): **Anomaly detection**







Big Data Components (V): Algorithms(VII)

- Data mining techniques(IV):

  Association rule mining

Transactions	Items
1	Milk, ice cream, jam, bread
2	ice cream, jam, bread, coffee
3	Milk, jam, bread
4	Milk, coffee
5	Milk, bread, chocolate
6	Ice cream, bread, coffee
7	Milk, jam, bread, banana
8	Coffee, bread, grape

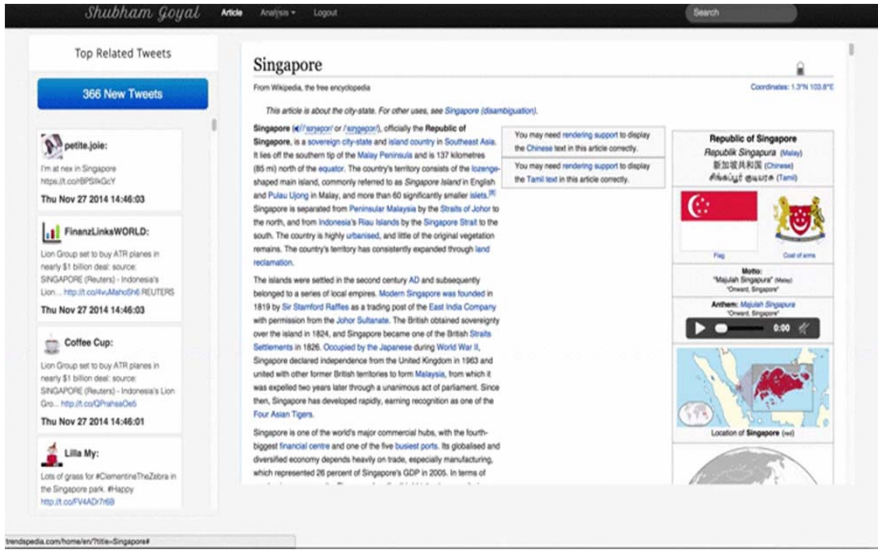
Milk, jam → bread

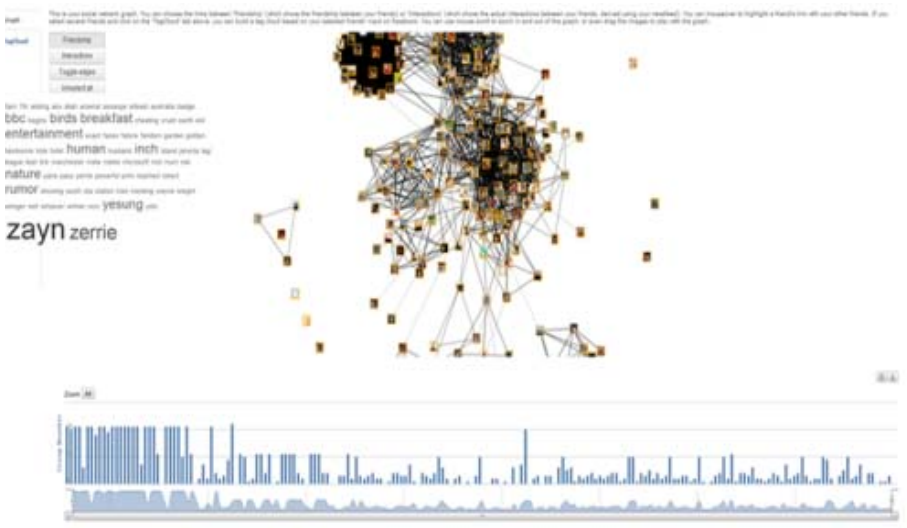
Support = 3/8

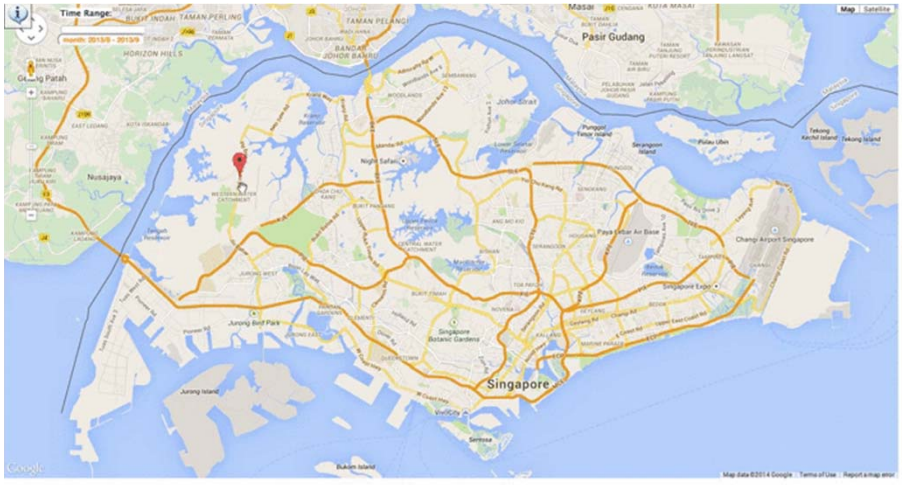
Confidence = 3/3

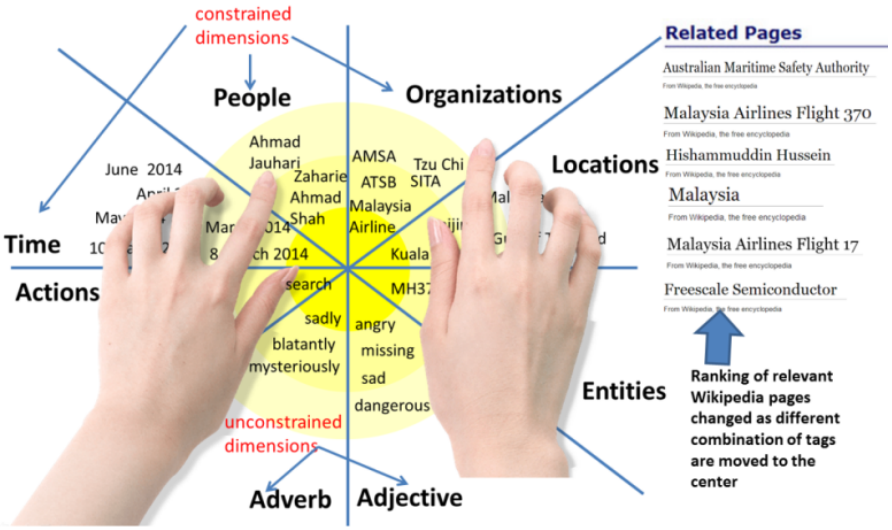
Big Data Components (V): Algorithms(VIII)

- Data mining techniques(V): **Visualization**









Big Data Components (V): Algorithms(VIII)

- Data mining techniques(V): **OLAP Operations**

Table 1: Mobile Game Activity Table

	player	time	action	role	country	gold
t_1	001	2013/05/19:1000	launch	dwarf	Australia	0
t_2	001	2013/05/20:0800	shop	dwarf	Australia	50
t_3	001	2013/05/20:1400	shop	dwarf	Australia	100
t_4	001	2013/05/21:1400	shop	assassin	Australia	50
t_5	001	2013/05/22:0900	fight	assassin	Australia	0
t_6	002	2013/05/20:0900	launch	wizard	USA	0
t_7	002	2013/05/21:1500	shop	wizard	USA	30
t_8	002	2013/05/22:1700	shop	wizard	USA	40
t_9	003	2013/05/20:1000	launch	bandit	China	0
t_{10}	003	2013/05/21:1000	fight	bandit	China	0

Table 2: Results of Q_s

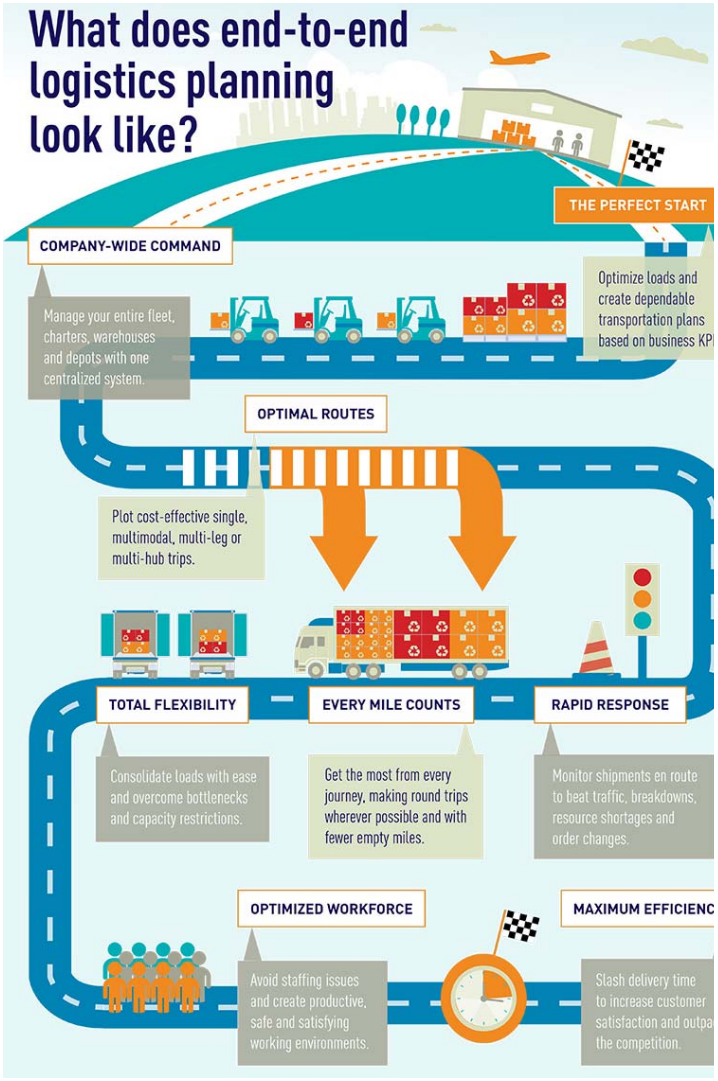
week	avgSpent
2013-05-19	50
2013-05-26	45
2013-06-02	43
2013-06-09	42
2013-06-16	45

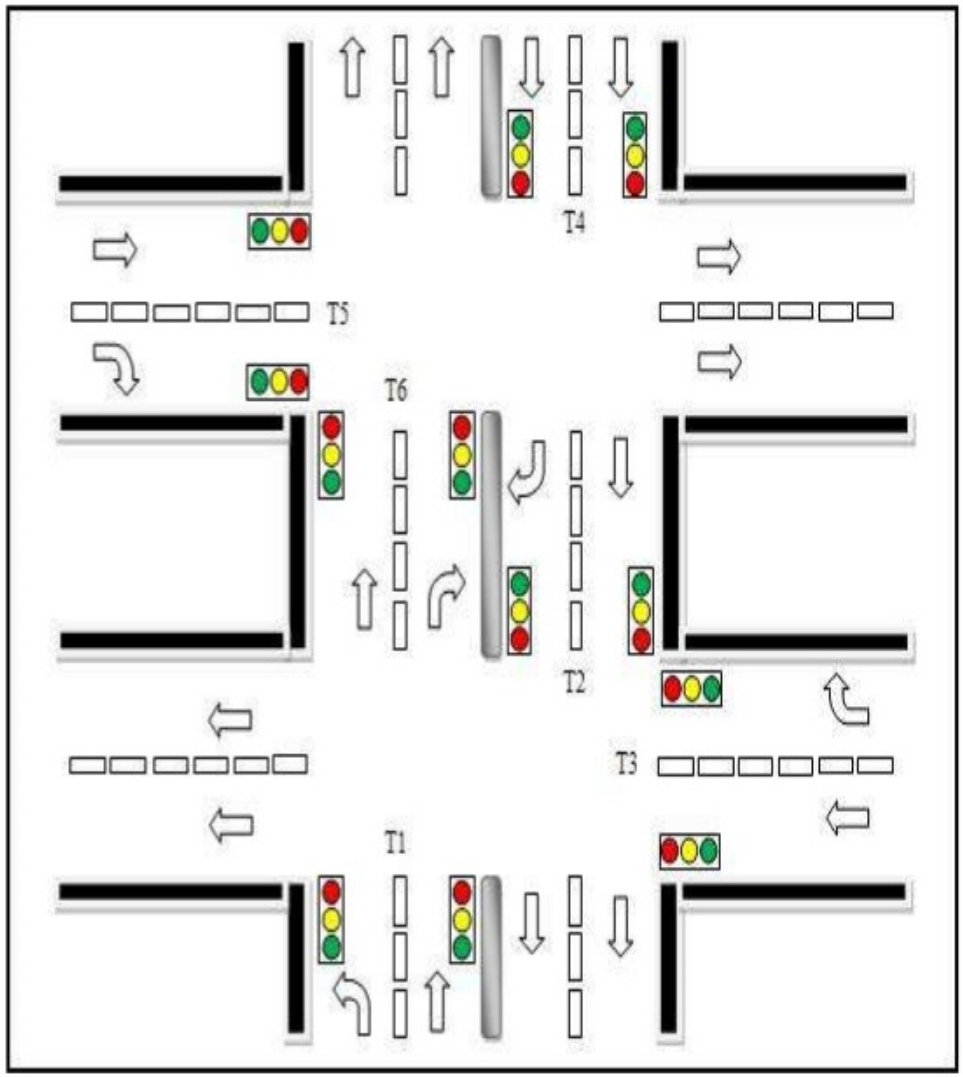
Table 3: Cohort Report for Shopping Trend

cohort	age (weeks)				
	1	2	3	4	5
2013-05-19 (145)	52	31	18	12	5
2013-05-26 (130)	58	43	31	21	
2013-06-02 (135)	68	58	50		
2013-06-09 (140)	80	73			
2013-06-16 (126)	86				

Big Data Components (V): Algorithms(IX)

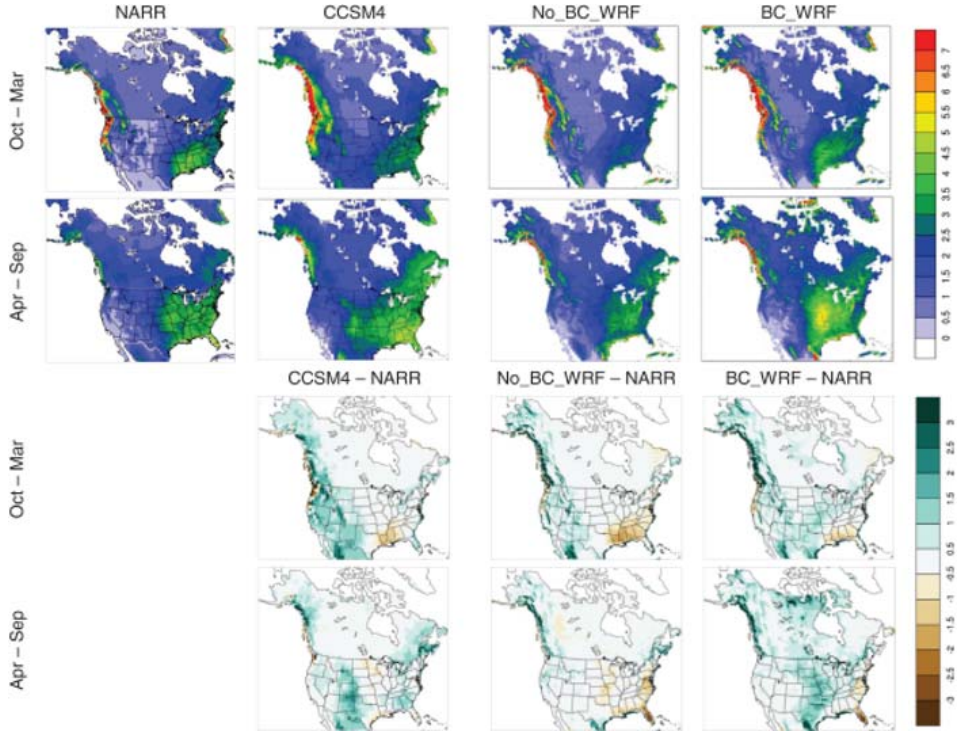
- Optimizations

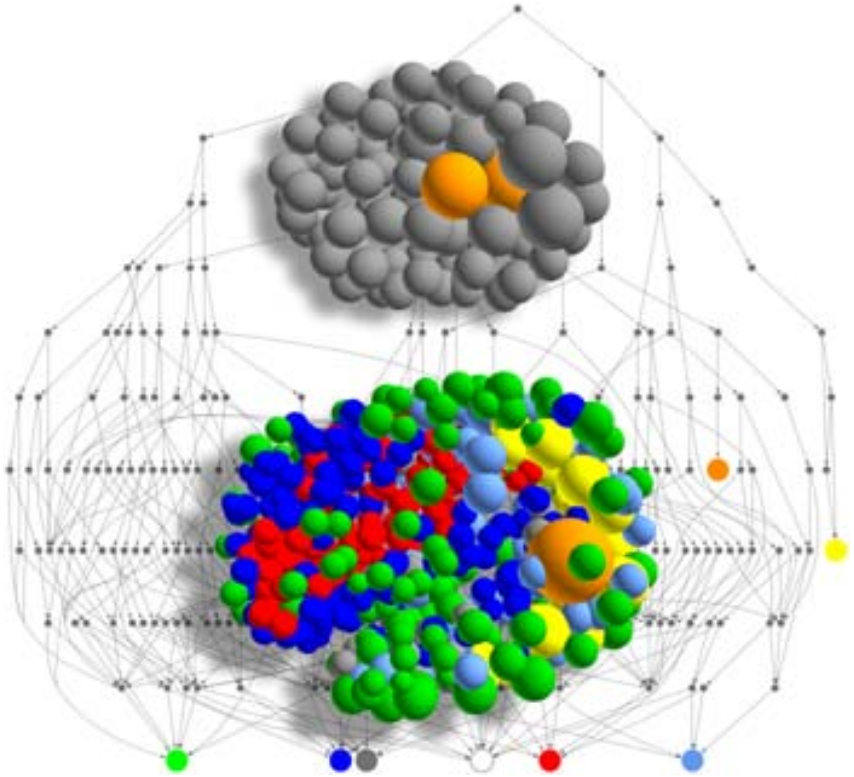




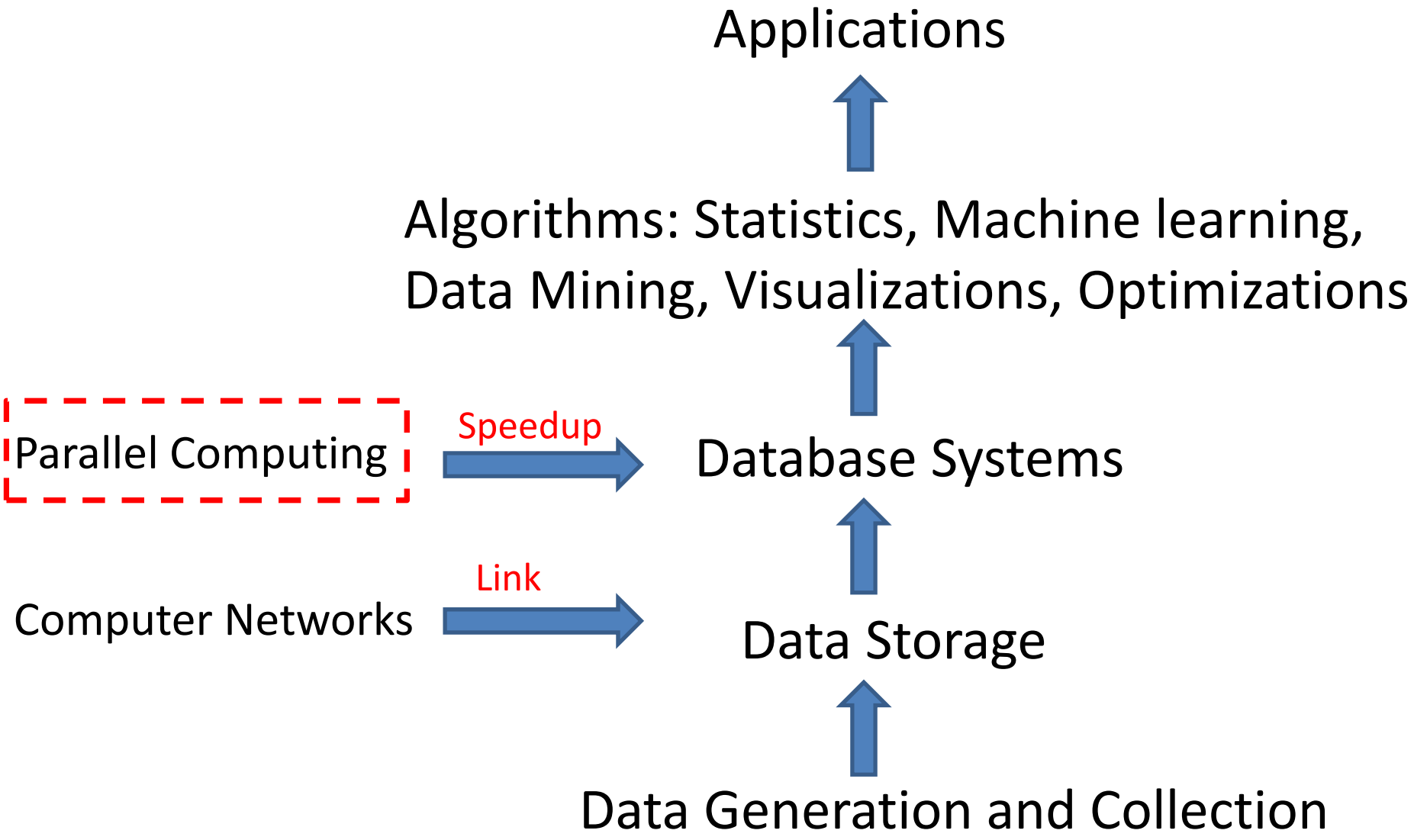
Big Data Components (V): Algorithms(X)

- Simulation:
  - Learn model parameters from Big Data
  - Compare the simulated data with the real data



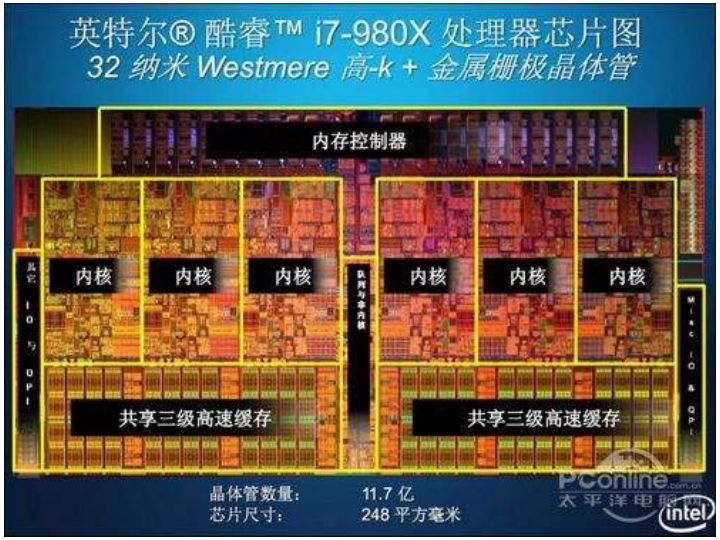


Overview of Big Data Components

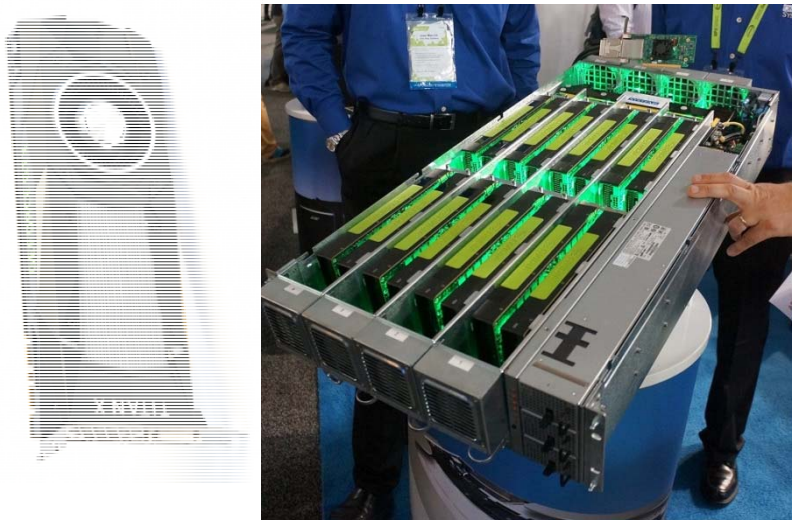


Big Data Components(VI):Parallel Computing(I)

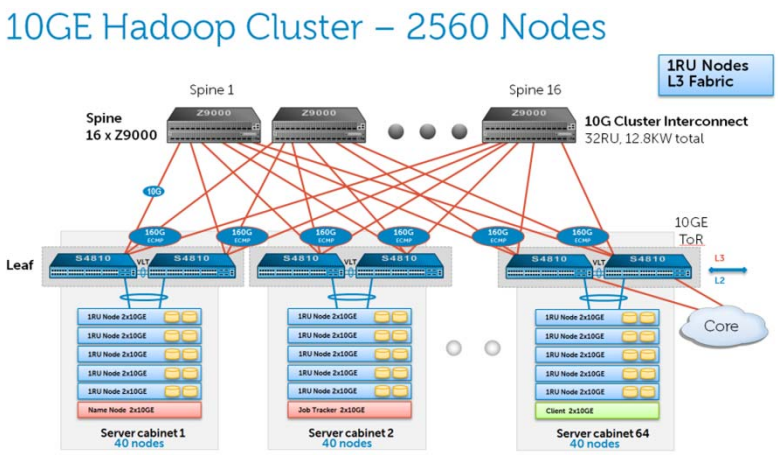
Multi-core processors



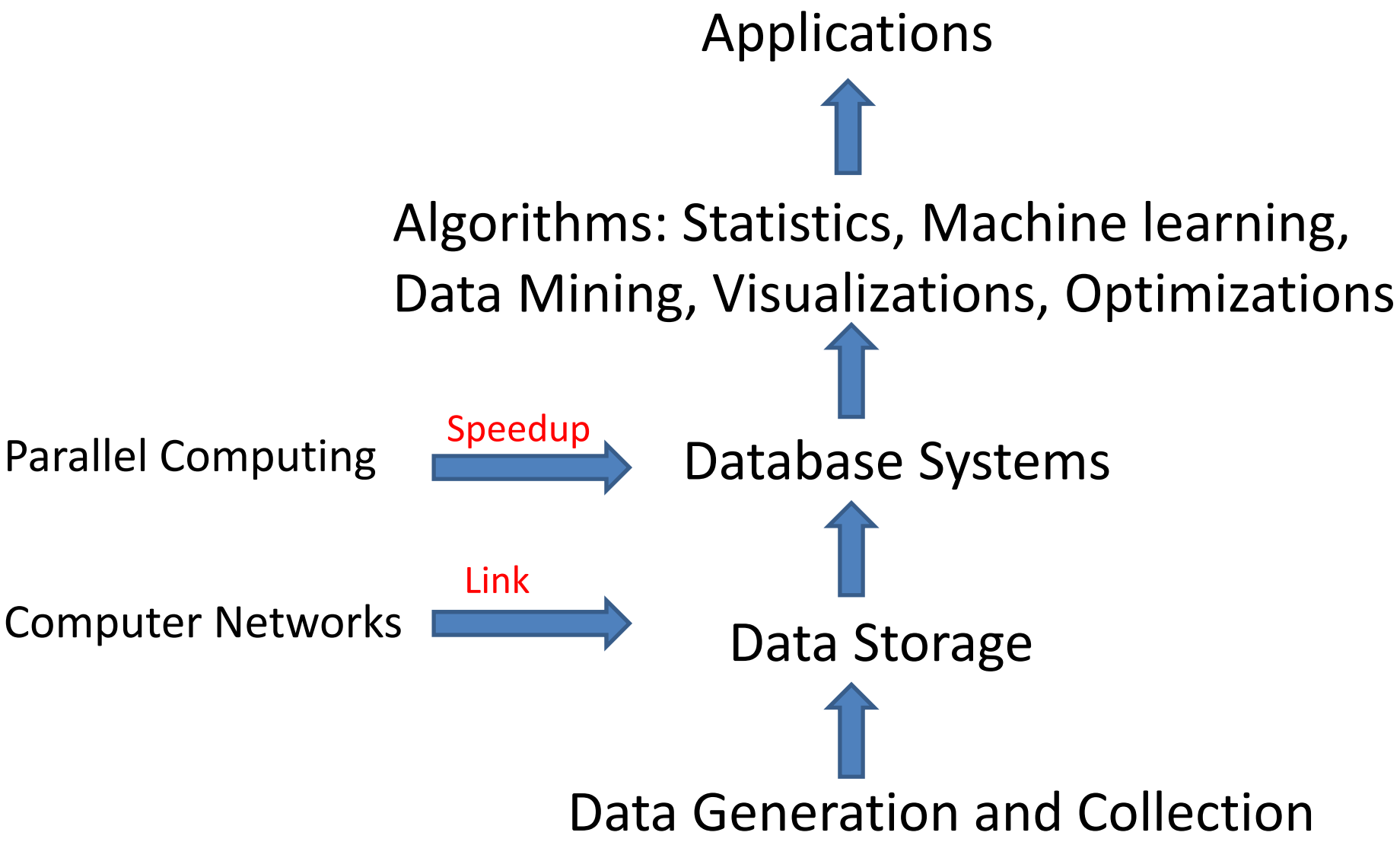
GPU servers



Computer clusters



Overview of Big Data Components



Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications 
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

Types of Big Data

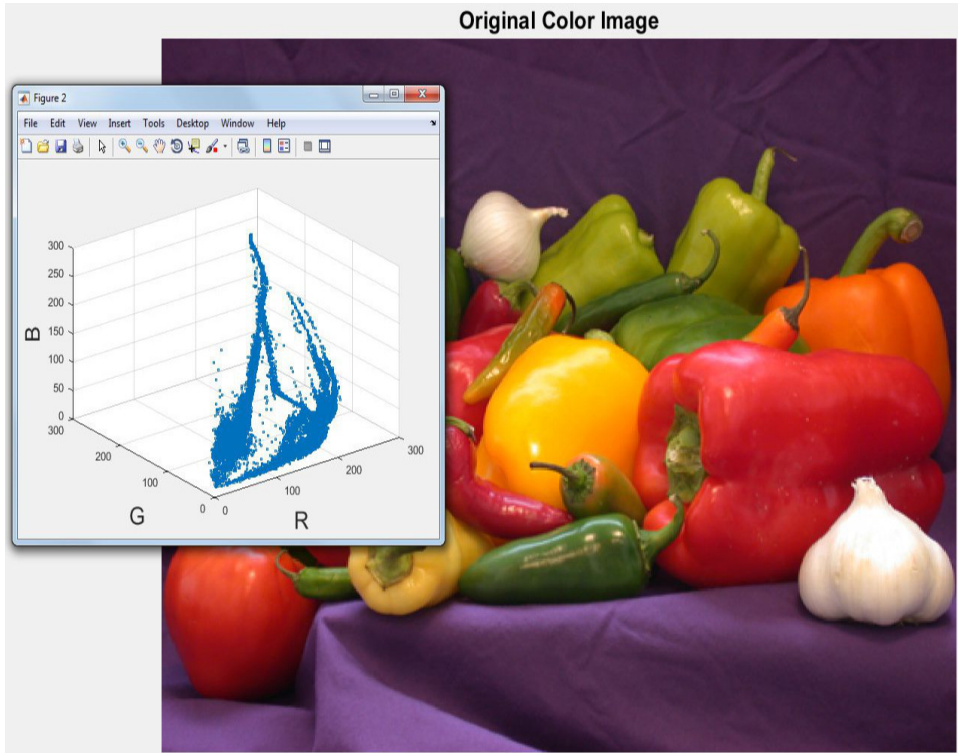
- Relational data
- High-dimensional data
- Sequence
- Tree
- Graph
- Mixed types
  - Sequences in a graph (social network)
  - Spatial-temporal data
  - Spatial-textual data
  - High-dimensional time series

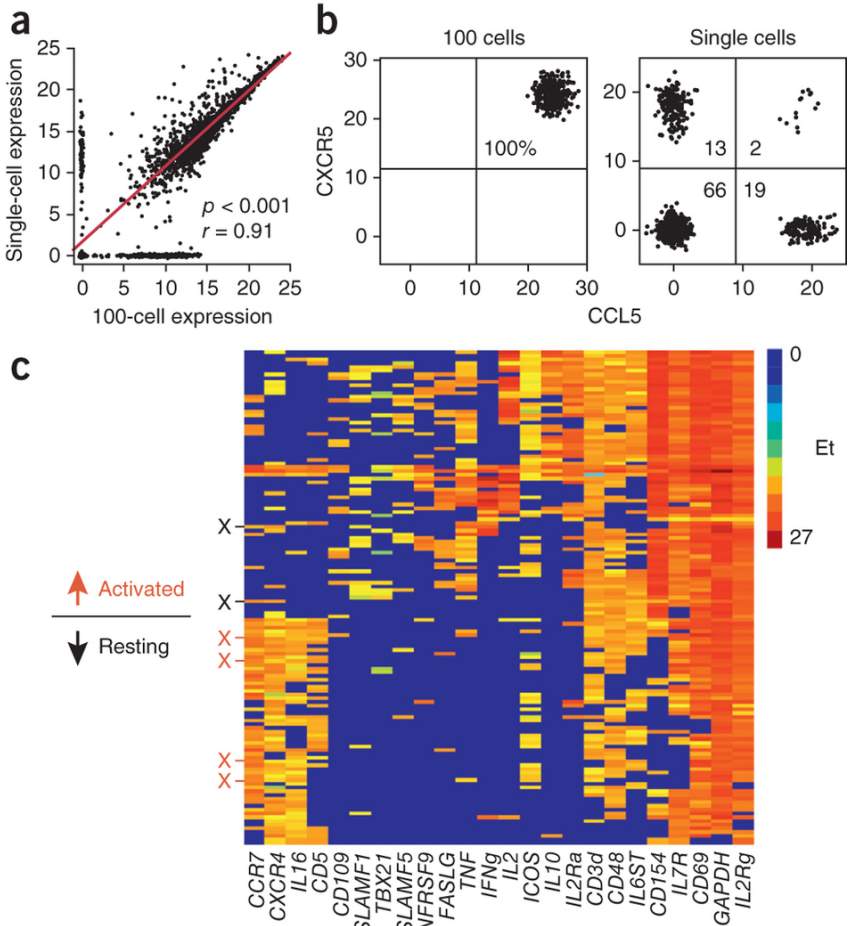
Types of Big Data(I): Relational Data

- Numeric(integer, float), categorical, binary, textual

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1

Types of Big Data(II): High-dimensional Data





Types of Big Data(III): Sequence

Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA

GTCGACCGGTGACCGTGCGTACACAGTGCTCCGGATAGCTGATAGCTCCGGTG

CAGCTGGCCACTGGCACGCATGTGTCACGAGGCCTATCGACTATCGAGGCCAC

Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA

GTCCCAGCCGGACCGTACCGGTAGATCAGCCGGTAGATTGATAGCGTGATGTG

CAGGGTCGGCCTGGCATGGCCATCTAGTCGGCCATCTAACTATCGCACTACAC

Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA

GTCTACGTAATCGTAGCCATCCGGACAGTGTGCACGATCGTACATGCTACGTG

CAGATGCATTAGCATCGGTAGGCCTGTCACACGTGCTAGCATGTACGATGCAC

Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA

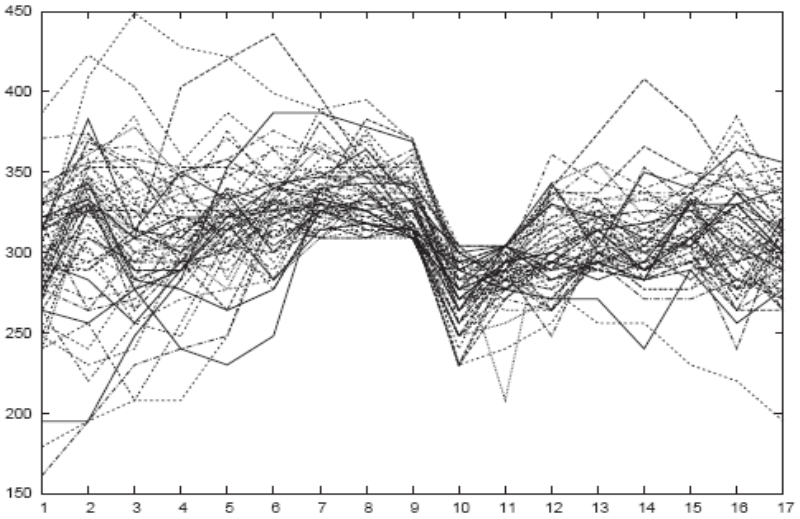
GTCGACCGGTGACCGTGCGTACACAGTGCTCCGGATAGCTGATAGCTCCGGTG

CAGCTGGCCACTGGCACGCATGTGTCACGAGGCCTATCGACTATCGAGGCCAC

Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA

GTCTCCATCCGGACTACCATACATCTGGTGTACCCGGTGATATCGTCCGGGTG

CAGAGGTAGGCCTGATGGTATGTAGACCACATGGGCCACTATAGCAGGCCCAC



重庆 官方旅游攻略2014版 CHONQING

欣欣旅游 www.cncn.com

必做的事

重庆市简称渝，位于我国西南地区东部，长江上游。1997 年以原四川省重庆、万县、涪陵三地级市区域设中央直辖市重庆,是我国面积最大、行政辖区最广、人口最多的中央直辖市。在重庆，你一定要：

1、吃一次秦妈老火锅，要去老巷子里

在重庆最出名就属火锅了，就连不吃辣的外国人都要来重庆品尝一下。重庆秦妈老火锅是重庆老字号的火锅酒楼，其锅底汤鲜味美，用料考究，油而不腻且易于消化，三百多种菜品任你选用。

图片来源：欣欣会员—dalianwantai

2、租一辆自行车，沿着南滨路，骑到洋人街。

这里西洋建筑扎成堆，不仅可以在这里体会到异域风情，同时还可以体验到由老外提供的端茶送水的服务，慢慢这里成为了重庆最大的“英语角”，很多老外在这里开店，顾客则是外国人、重庆人。最出名的厕所文化，来了才能感受，强烈建议来这里多拍几张照片，回去定会回味无穷。

3、爬南山，顺道品尝一下泉水鸡，再走到一棵树看重庆夜景，夏夜最好！

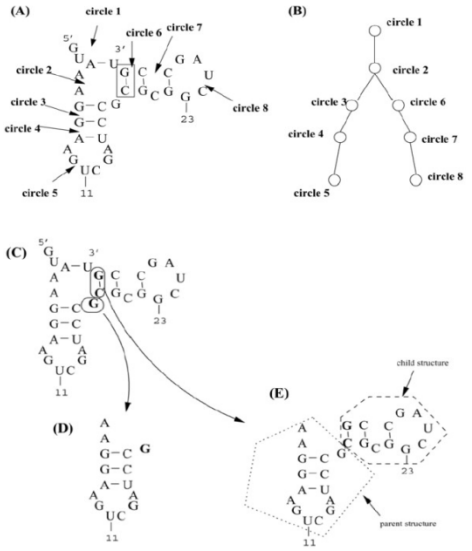
山城夜景是重庆的一张城市名片，不览“价值百万美元的夜景”，就不算到过重庆。夜晚的重庆让人沉迷，熙熙的人群，悦耳的重庆话，伴随着养眼的重庆美女，店里的茶女们打着小麻将，喝着重庆味道的茶水。空气有点湿热。去朝天门码头看夜景时，上空经常有风筝飘，台阶下的男女老幼，有的乘凉，有的索性泡在江中游泳。

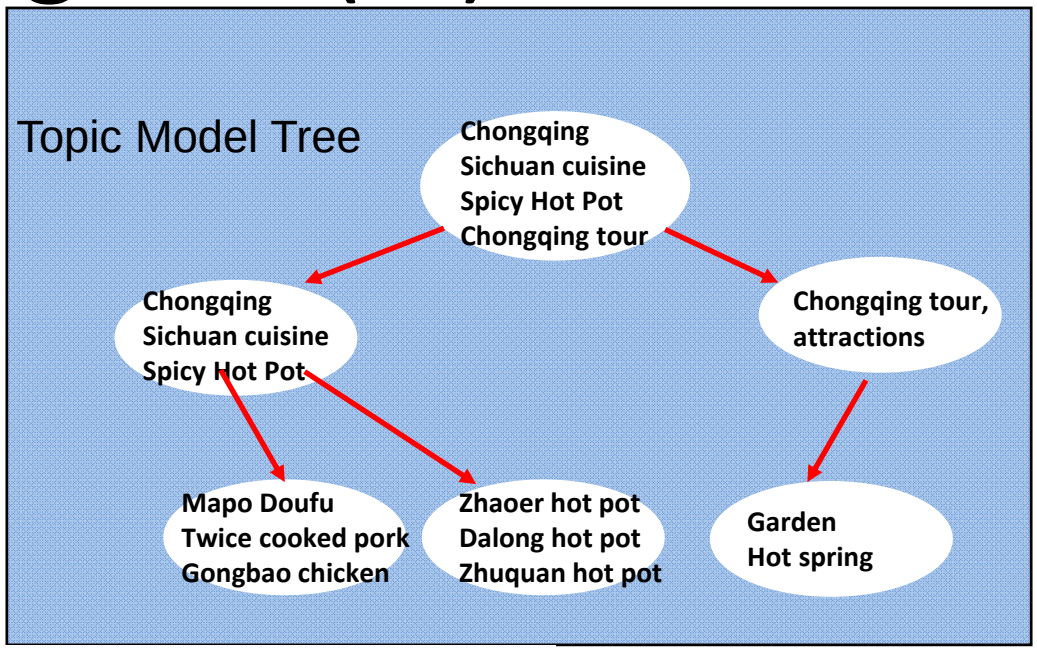
4、徒步老城区，十八梯，下半城去感受老重庆市井生活。

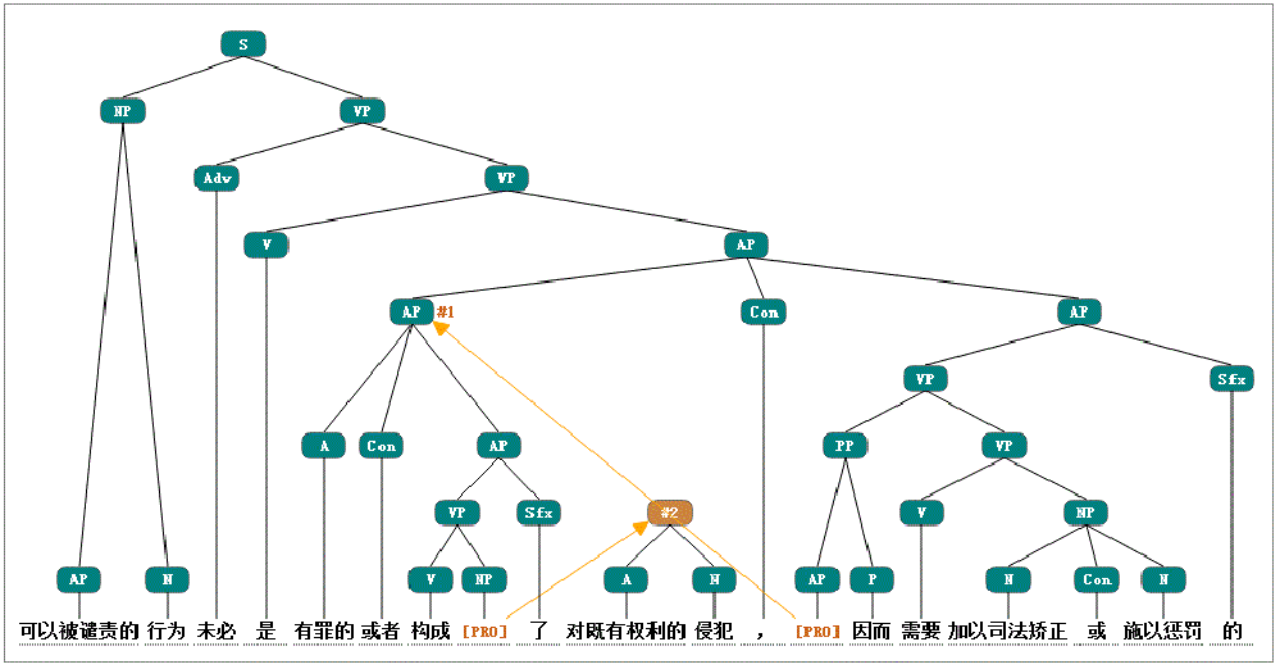
作为重庆老城里真正的老街，一走进这里，就进入了真正的重庆民间。十八梯最具爬坡上坎的山城特点，一色的青石板阶梯和有各个年代标志的老房子。最古老的吊脚楼、捆绑房、上世纪五六十年代的竹蔑抹灰墙房子和六七十年代的前苏联风格的砖瓦房，仿佛重庆的历史建筑展览。

5、坐一次长江索道，俯视脚下滚滚向东流去的江水。

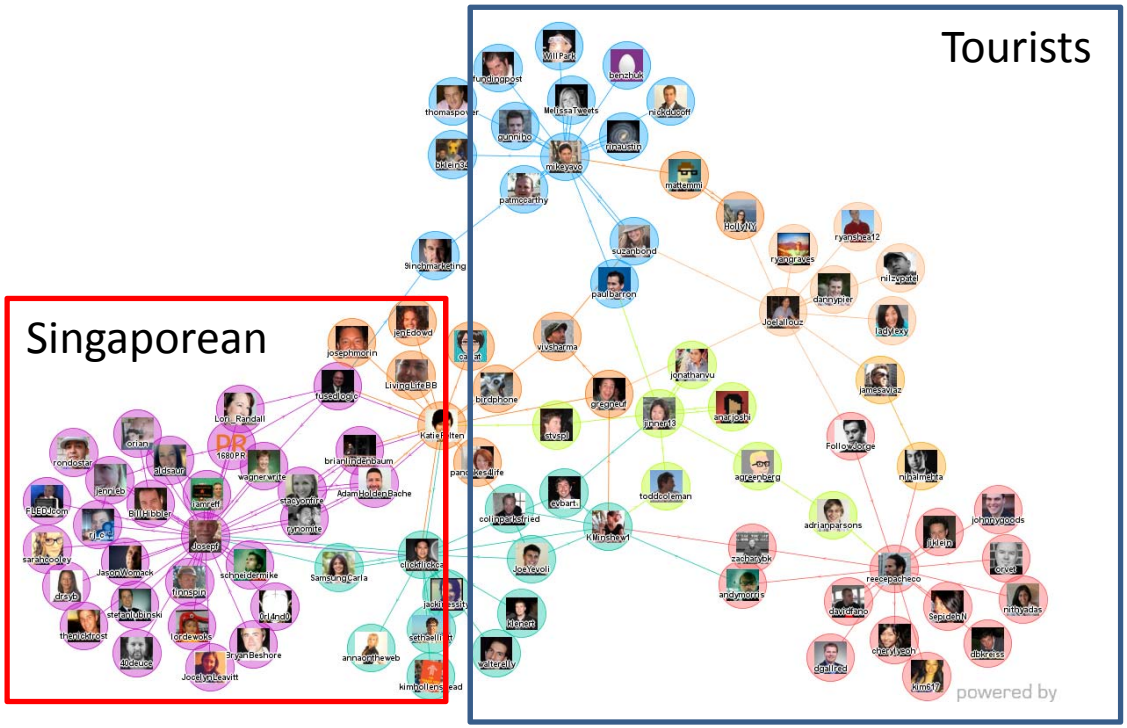
Types of Big Data(IV): Tree

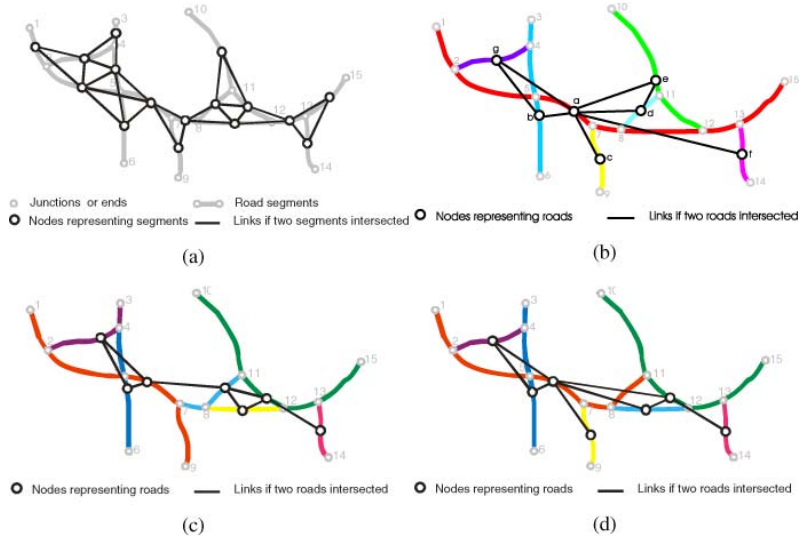


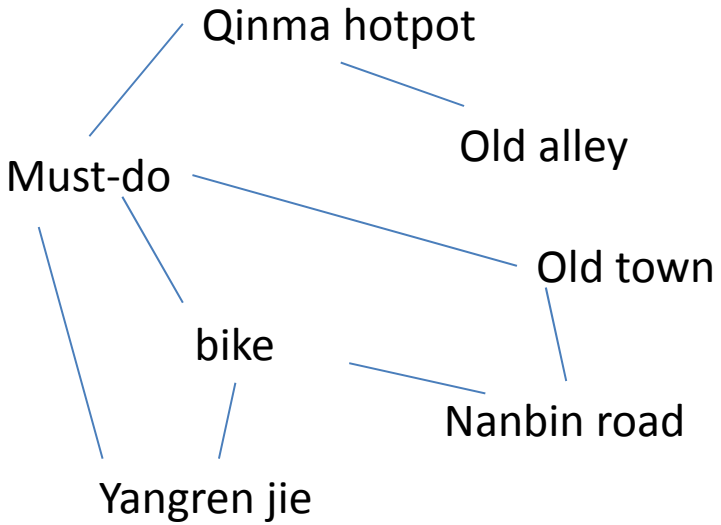




Types of Big Data(V): Graph







Recap

- Book can be represented as a sequence
- Book can be represented as a tree
- Book can be represented as a graph
- Which representation is correct?


重庆 CHONGQING 官方旅游攻略2014版

欣欣旅游 www.cncn.com

必做的事

重庆市简称渝，位于我国西南地区东部，长江上游。1997 年以原四川省重庆、万县、涪陵三地级市区域设中央直辖市重庆,是我国面积最大、行政辖区最广、人口最多的中央直辖市。在重庆，你一定要：

1、吃一次秦妈老火锅，要去老巷子里

在重庆最出名就属火锅了，就连不吃辣的外国人都要来重庆品尝一下。重庆秦妈老火锅是重庆老字号的火锅酒楼，其锅底汤鲜味美，用料考究，油而不腻且易于消化，三百多种菜品任你选用。

2、租一辆自行车，沿着南滨路，骑到洋人街。

这里西洋建筑扎成堆，不仅可以在这里体会到异域风情，同时还可以体验到由老外提供的端茶送水的服务，慢慢这里成为了重庆最大的“英语角”，很多老外在这里开店，顾客则是外国人、重庆人。最出名的厕所文化，来了才能感受，强烈建议来这里多拍几张照片，回去定会回味无穷。

图片来源：欣欣会员—dallanwantai

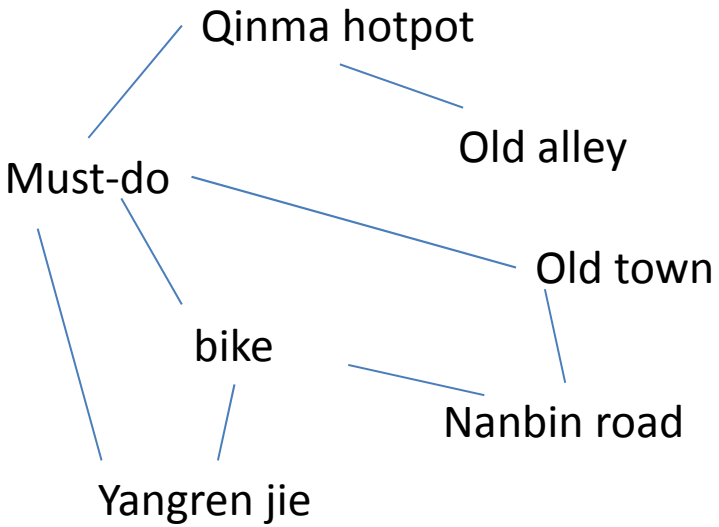
3、爬南山，顺道品尝一下泉水鸡，再走到一棵树看重庆夜景，夏夜最好！

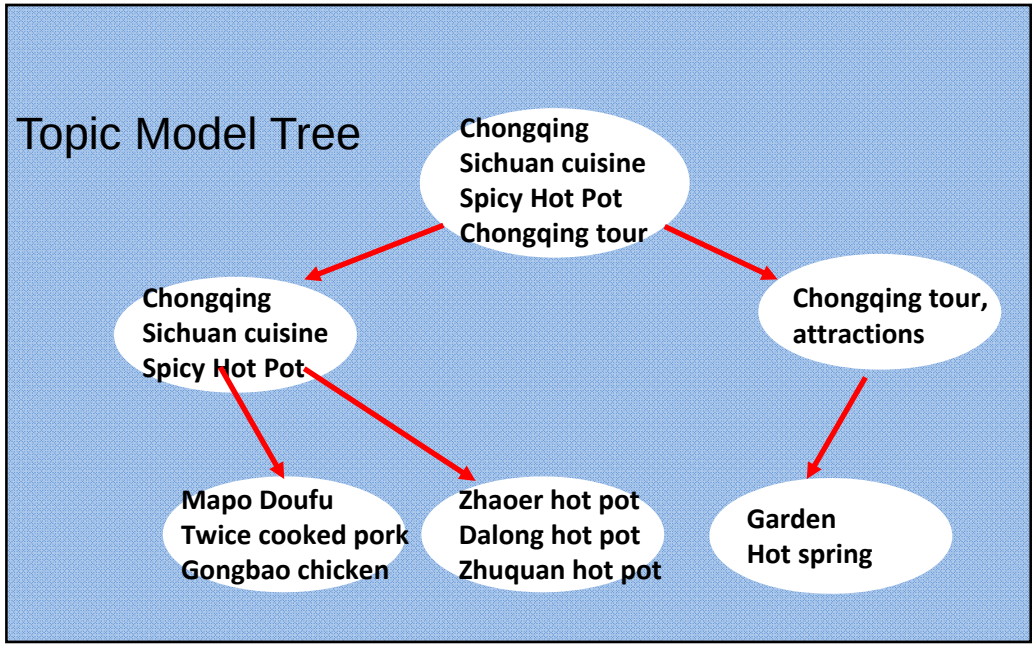
山城夜景是重庆的一张城市名片，不览“价值百万美元的夜景”，就不算到过重庆。夜晚的重庆让人沉迷，熙熙的人群，悦耳的重庆话，伴随着养眼的重庆美女，店里的茶女们打着小麻将，喝着重庆味道的茶水。空气有点湿热。去朝天门码头看夜景时，上空经常有风筝飘，台阶下的男女老幼，有的乘凉，有的索性泡在江中游泳。

4、徒步老城区，十八梯，下半城去感受老重庆市井生活。

作为重庆老城里真正的老街，一走进这里，就进入了真正的重庆同问。十八梯最具爬坡上坎的山城特点，一色的青石板阶梯和有各个年代标志的老房子。最古老的吊脚楼、捆绑房、上世纪五六十年代的竹篾抹灰墙房子和六七十年代的前苏联风格的砖瓦房，仿佛重庆的历史建筑展览。

5、坐一次长江索道，俯视脚下滚滚向东流去的江水。





Types of Big Data

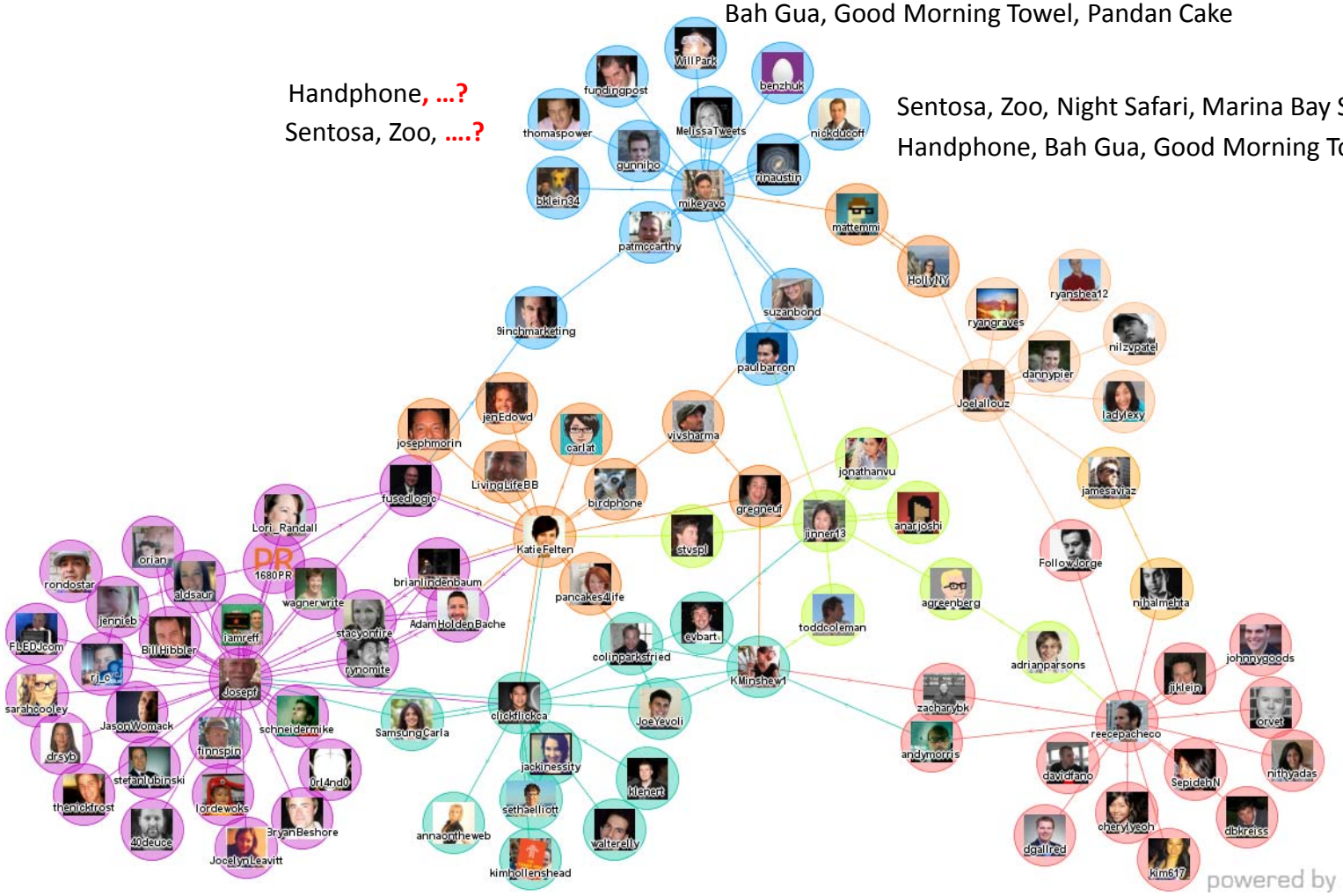
- Relational data
- High-dimensional data
- Sequence
- Tree
- Graph
- Mixed types
  - Sequences in a graph (social network)
  - Spatial-temporal data
  - Spatial-textual data
  - High-dimensional time series

Types of Big Data(IV): Sequences in a Graph

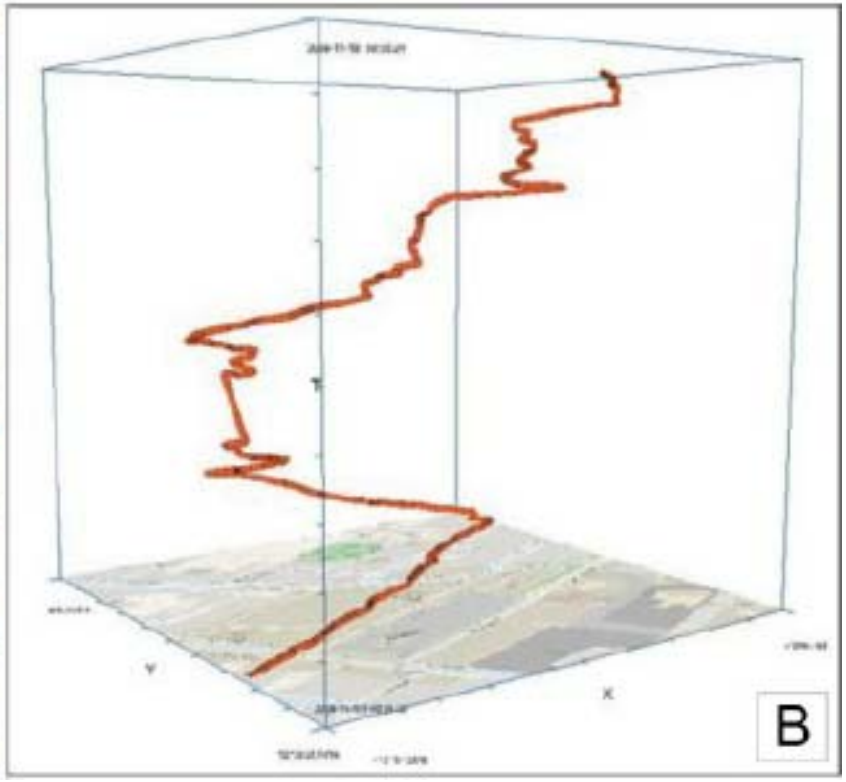
Sentosa, Night Safari, Marina Bay Sand
Bah Gua, Good Morning Towel, Pandan Cake

Handphone, …?
Sentosa, Zoo, ….?

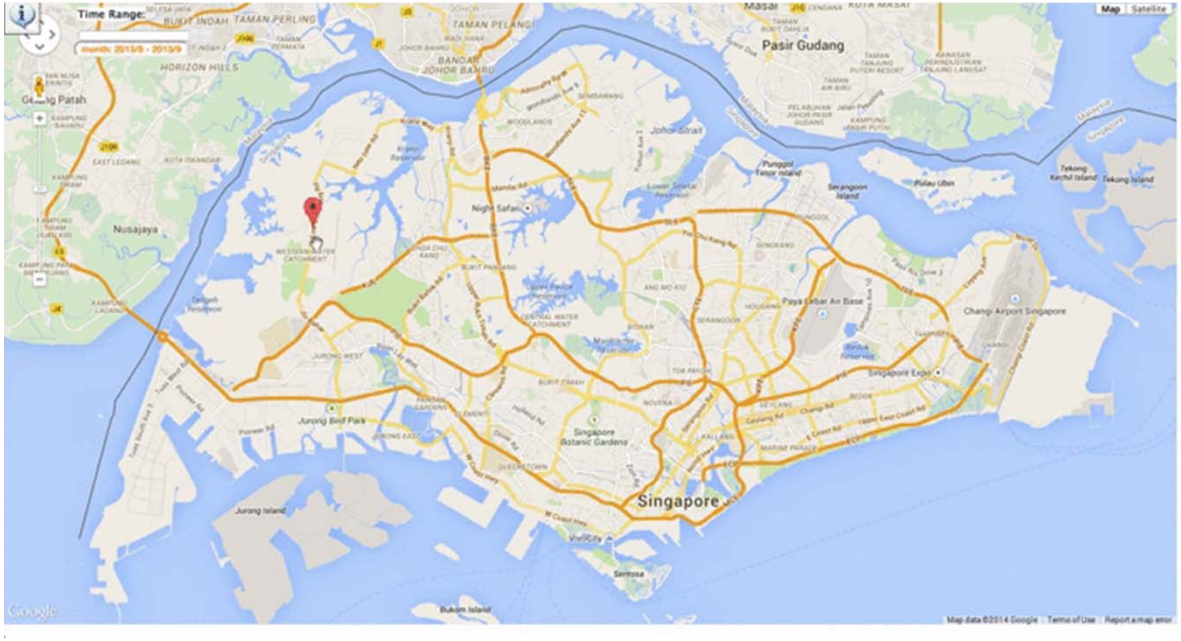
Sentosa, Zoo, Night Safari, Marina Bay Sand
Handphone, Bah Gua, Good Morning Towel

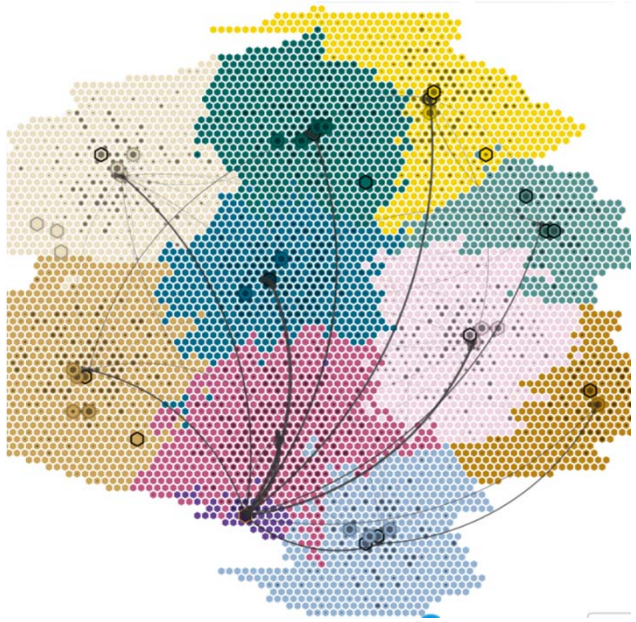


Types of Big Data(V):Spatial-temporal Data

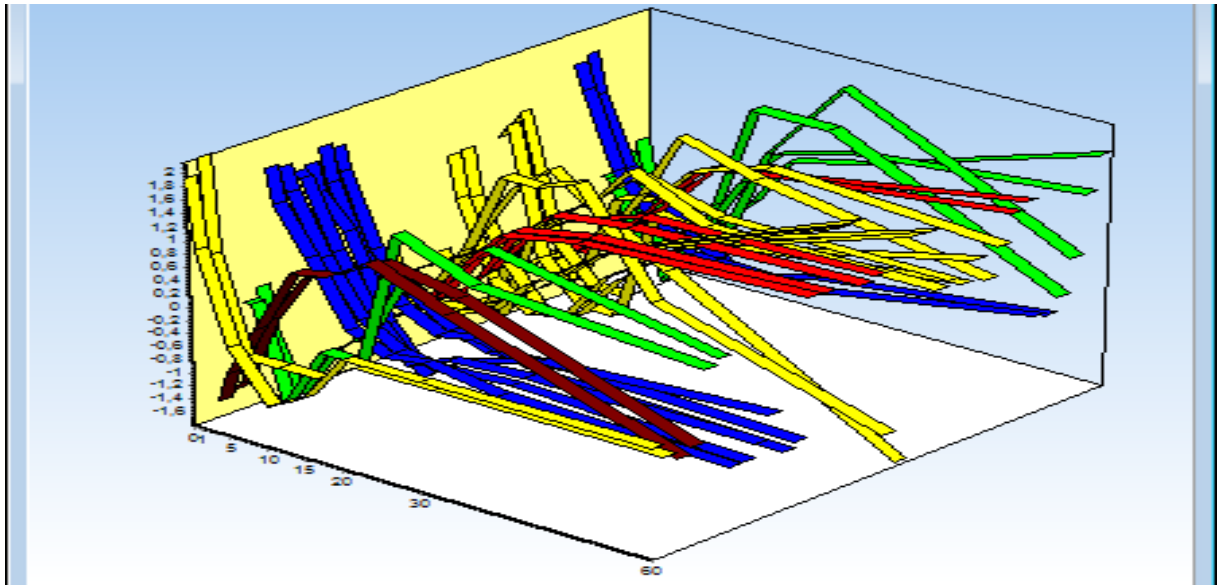


Types of Big Data(VI): Spatial-textual Data





Types of Big Data(VII): High-dimensional Time Series



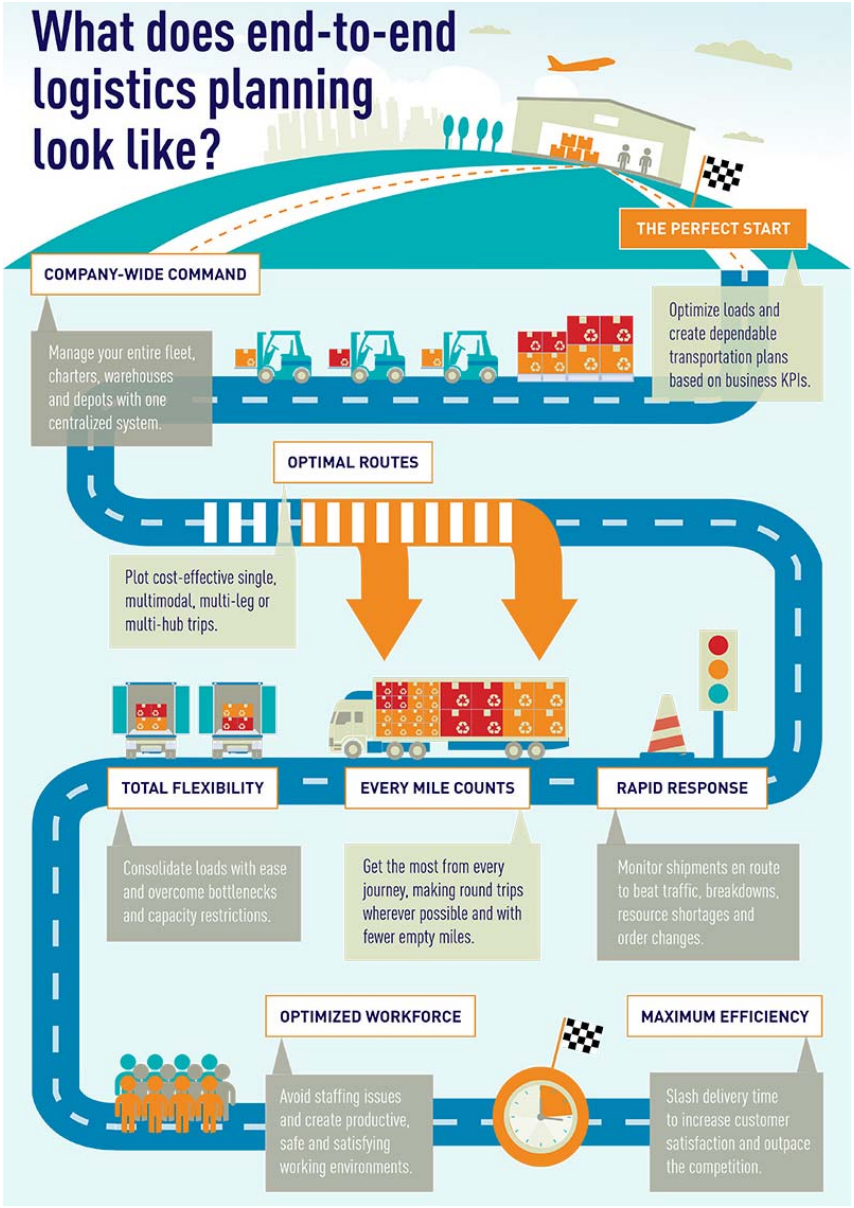
Types of Big Data: Summary

- The type of big data somehow depends on how data is generated or collected.
- In many cases, the types of data depends on the problem and application themselves.
- Big data applications often need to process multiple types of data
- Distilling raw data into appropriate data is the most fundamental problem.

Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications 
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

Big Data Applications(I): Logistic



- Supplier → Manufacture →Distributor→Customer
- Transport capacity
- The delivery time is affected by traffic and weather conditions.
- Storage capacity and price
- Accuracy of the plan

Big Data Applications(II): Transportation

Early Warning of Human Crowds Based on Query Data from Baidu Map: Analysis Based on Shanghai Stampede

Jingbo Zhou, Hongbin Pei and Haishan Wu*
Baidu Research – Big Data Lab, Beijing, China

[Media Report: MIT Technology Review, Wall Street Journal, South China Morning Post]

Figure 2: Human population density between 23:00-24:00 on Dec. 31th 2014.

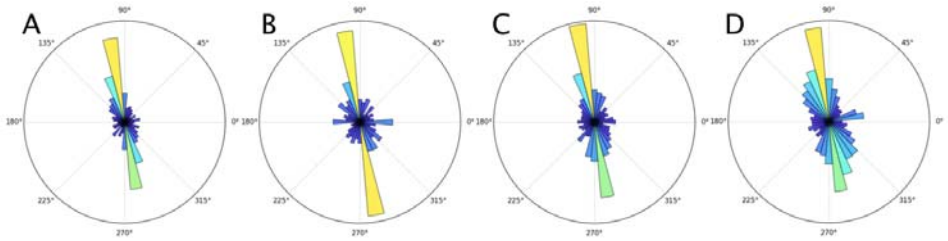
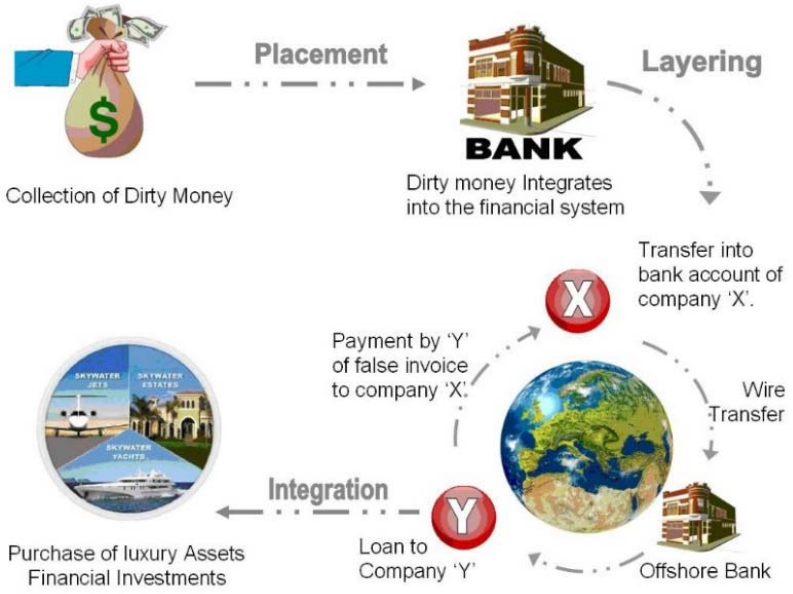
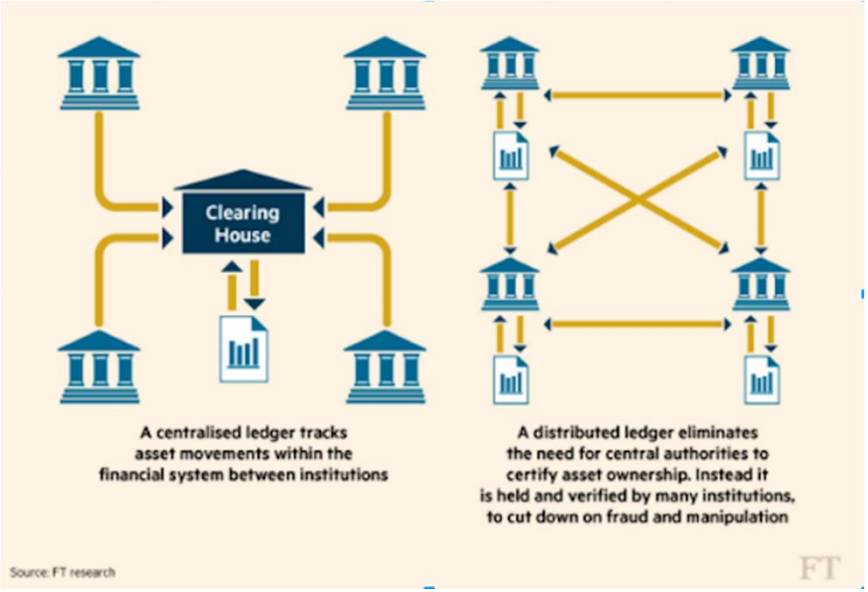


Figure 5: Human flow direction distribution in Chenyi Square (the specific disaster area of 2014 Shanghai Stampede) from 22:00 to 24:00 in: A – a common weekend (Aug. 23th 2014); B – the eve of the Mid-Autumn Festival (Sept. 7th 2014); C – the China's National Day (Oct. 1st 2014) and D – New Year's Eve of 2014

- Integration of transportation data
  - Multiple sources: car, taxi, bus, pedestrian, sensor
  - Multiple organizations: telecom corporation, taxi company, bus company, government
  - Data sharing and integrating
- Transportation planning
  - Construction of new roads
  - Location of transport junction
  - Answer "what-if" questions
- Transportation management
  - Prevent traffic jam
  - Optimize traffic lights
  - Direct human crowds
  - Identify bottlenecks

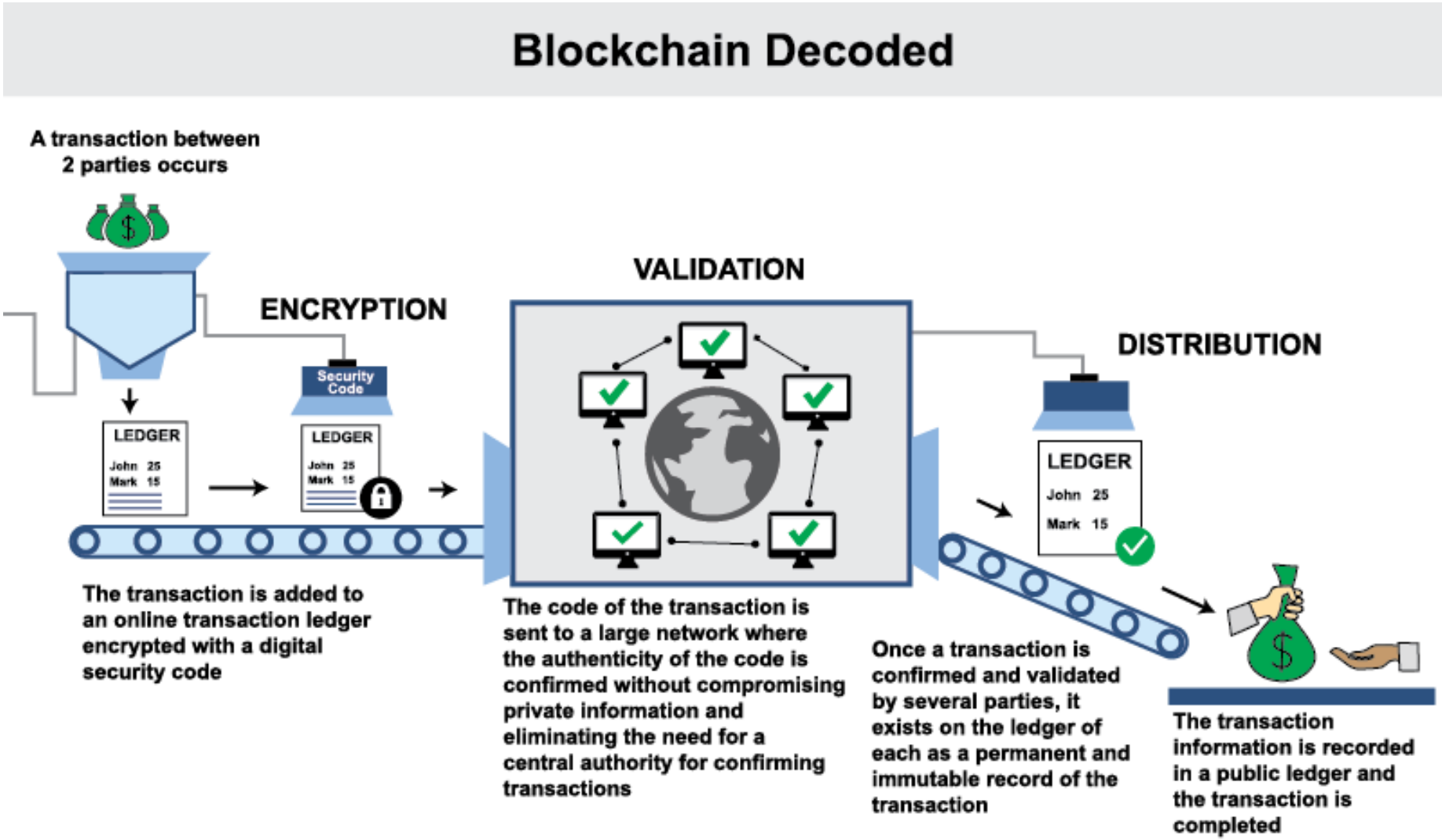
Big Data Applications(III): Finance



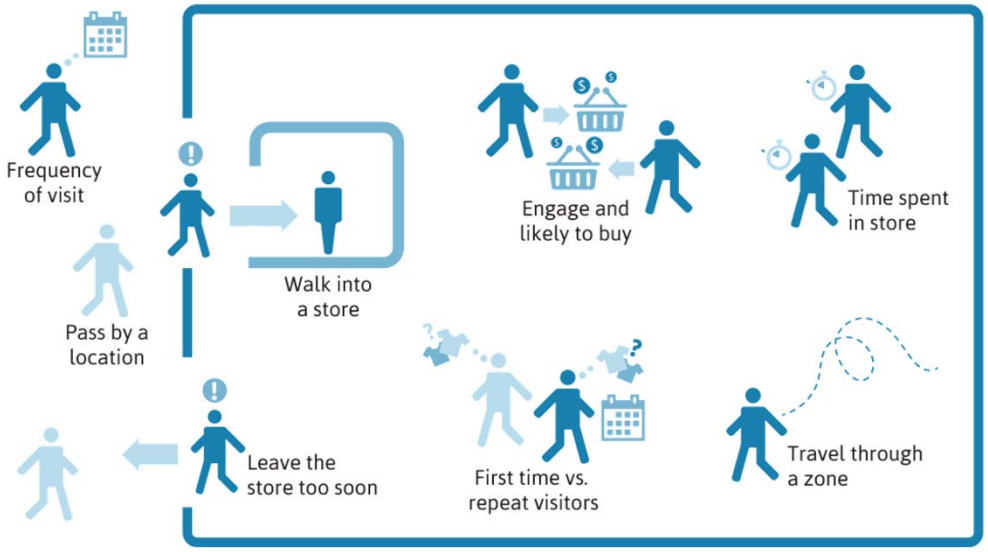


- Finance Prediction/Policy
  - Cash flow
- Abnormity Detection
  - Fraud
  - Money laundering
  - Tax evasion
- Fintech (financial technology)
  - Blockchain
  - P2P loan

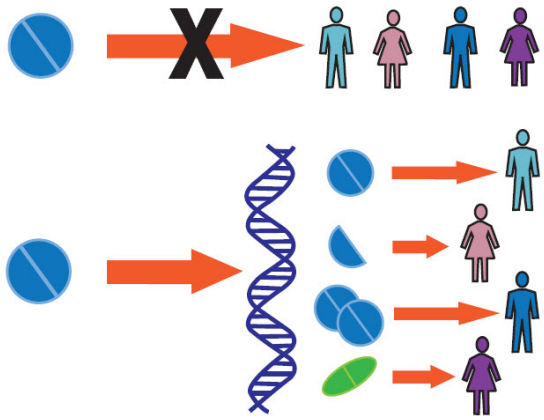
Big Data Applications(III): Finance(II)

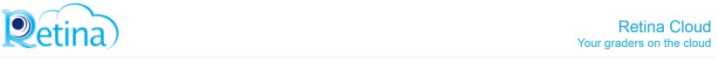


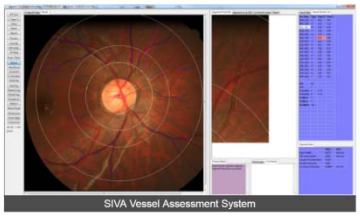
Big Data Applications(IV): Retail Analytics



Big Data Applications(V): Medical

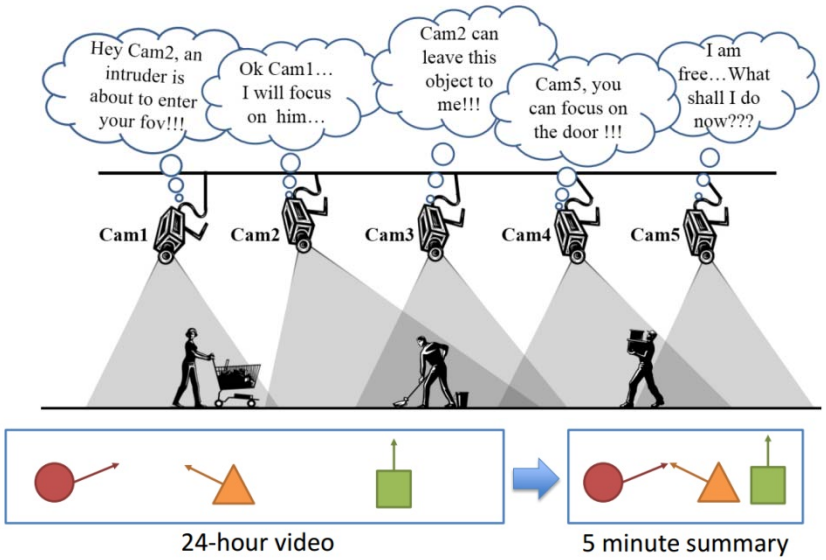


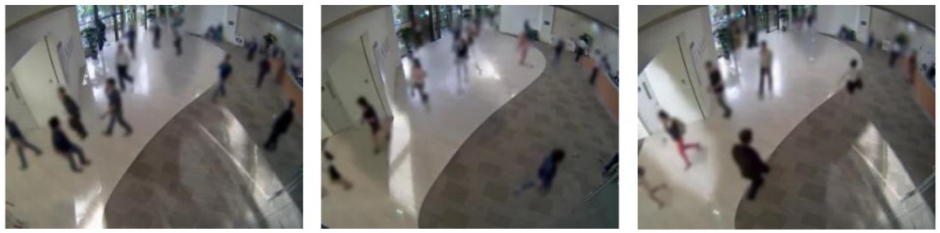




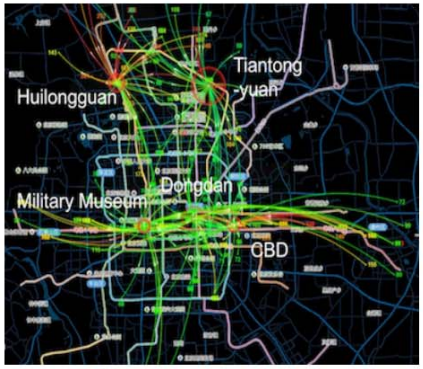
- Medicine control
  - Drug allergy
  - Drug and Poison Analysis
- Personal Medicine
- Hospital/Clinic management
  - Medical record
  - Probability of re-hospitalization
- Doctor on the Cloud: Retinal-scan analysis
  - https://retinacloud.d1.comp.nus.edu.sg/users/sign_in

Big Data Applications(VI): Security

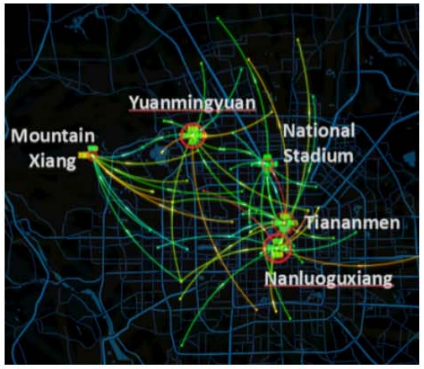




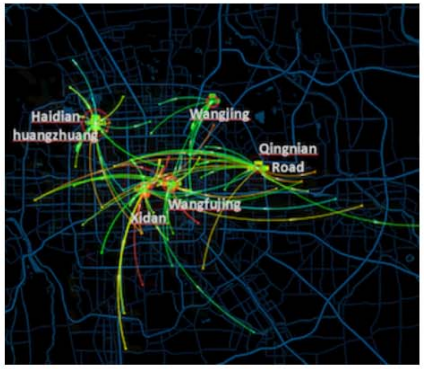
- Monitoring
  - CCTV
  - IC cards
- Facial recognition to detect strangers
- Exit passageway monitoring
- Crime analysis



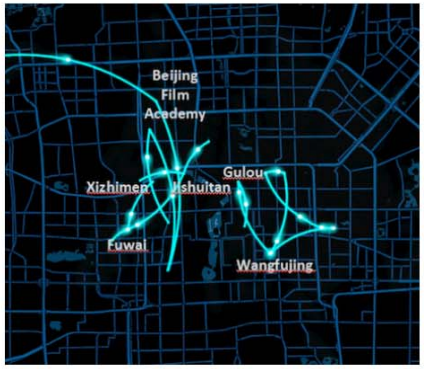
(a) all passengers



(b) visitors

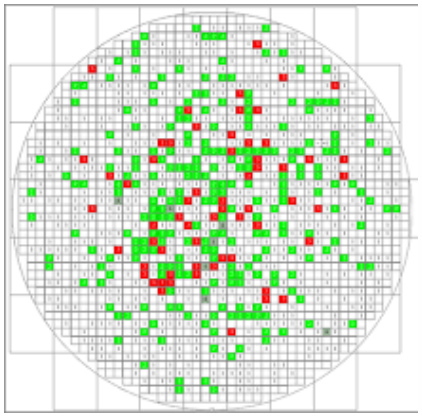


(c) shoppers



(d) thieves

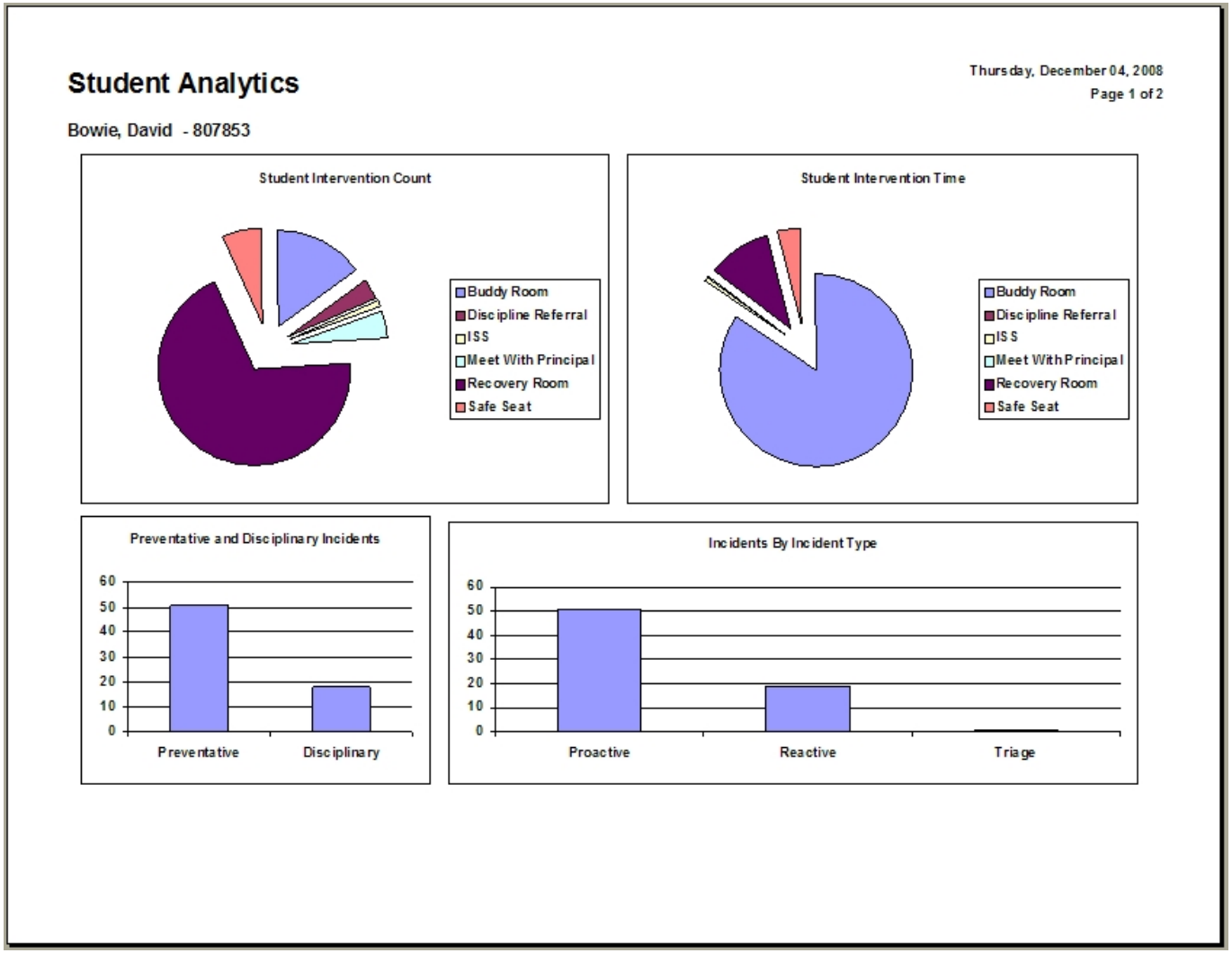
Big Data Applications(VII): Manufacturer



- High returns products
  - Wafer
  - Petroleum
- Manufacture data
  - Product imaging
  - Machine sensors
  - Machine repaired logs
- Usage/Applications
  - Predictive Maintenance
    - Which machine affected the quality?
    - Which part of the machine needs to be repaired?
    - How to fully utilize the machines?
  - Product quality control

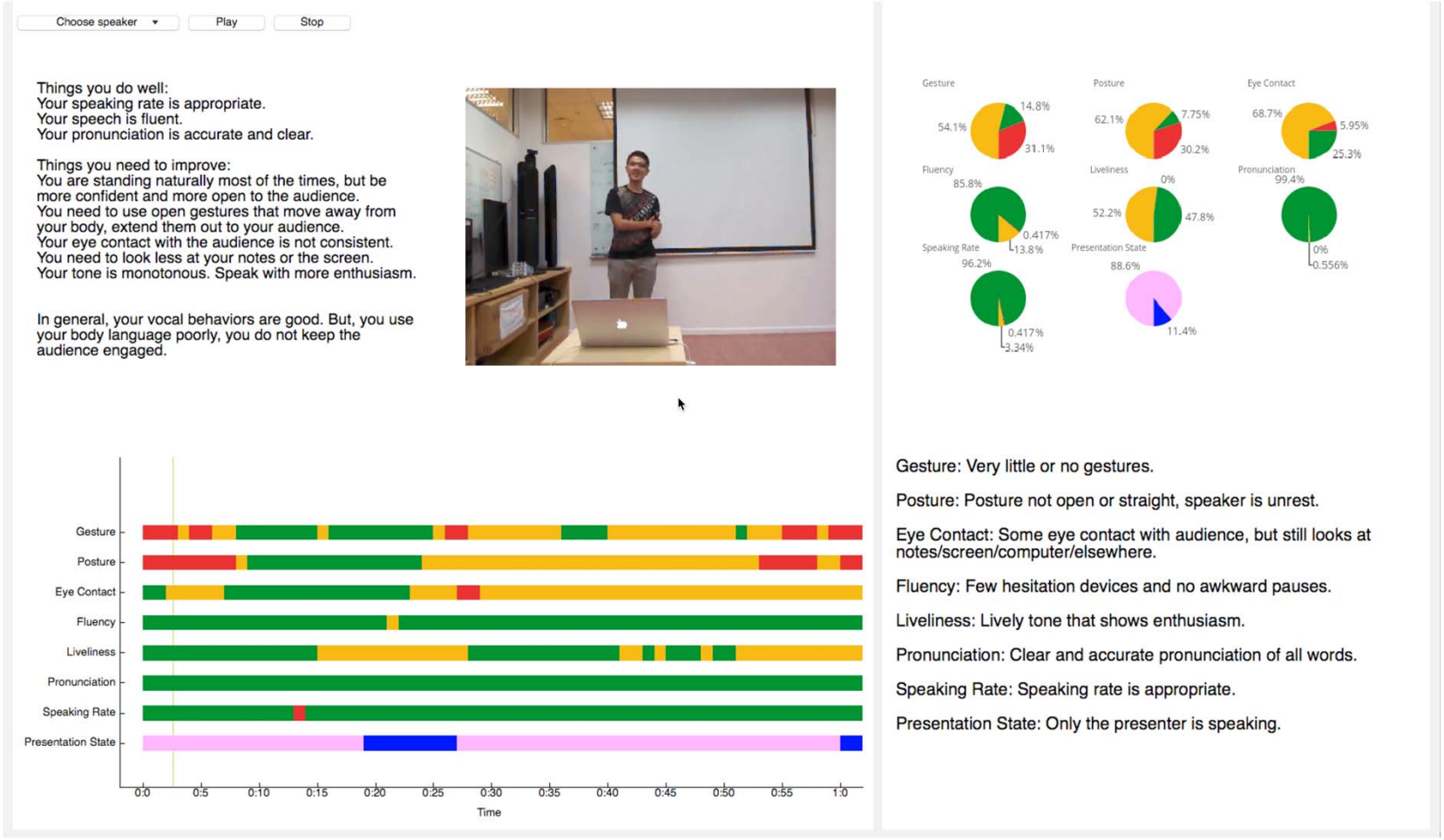
Big Data Application(VIII): Education(I)

- Student Analytics



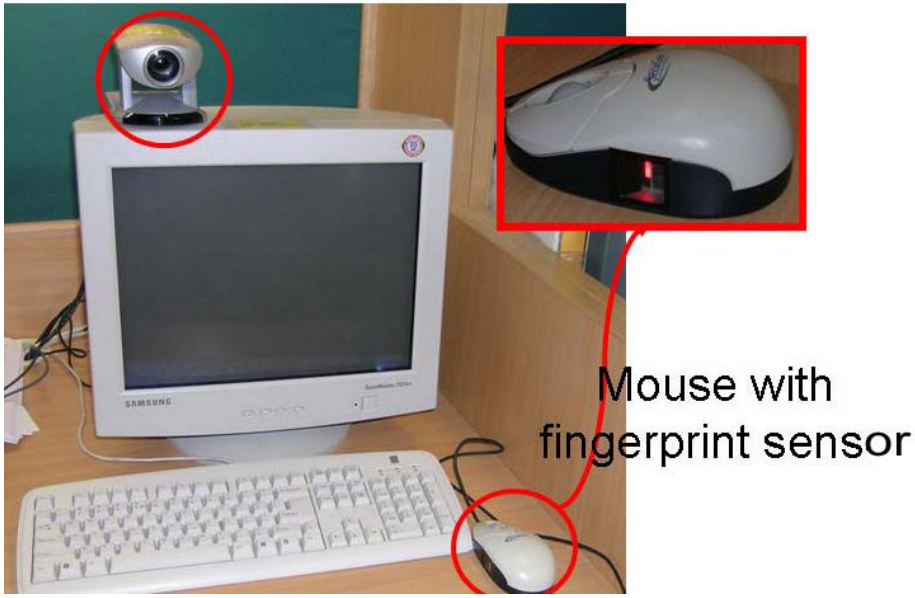
Big Data Application(VIII): Education(II)

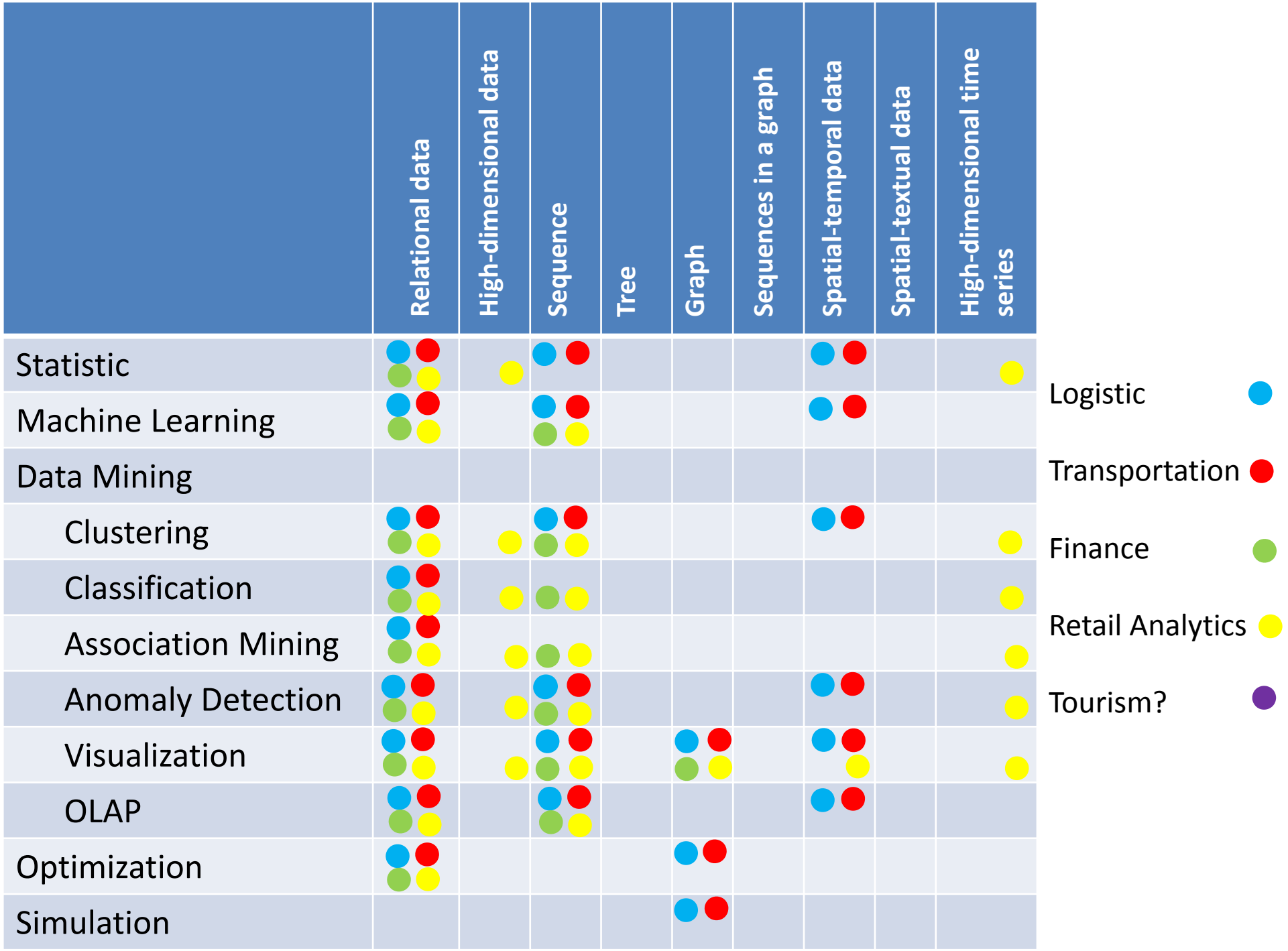
- PreSENSE:First-Person-View sensor to perform self-quantification of a presentation



Big Data Application(VIII): Education(III)

- Continuous Authentication using Multimodal Biometrics
- Use face + fingerprint images to continuously verify presence of legitimate user
- Remote but authentic examination



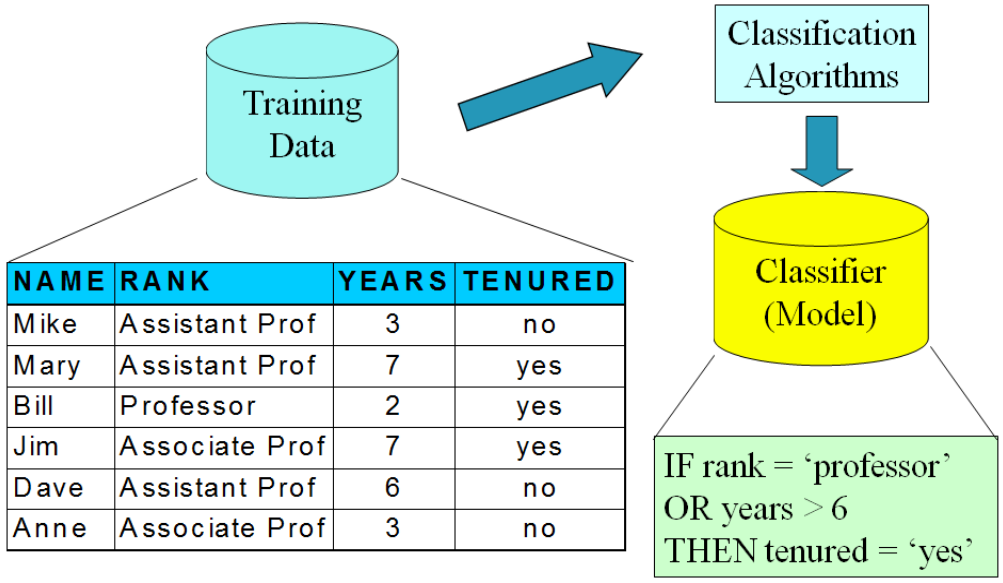

Relational data
High-dimensional data
Sequence
Tree
Graph
Sequences in a graph
Spatial-temporal data
Spatial-textual data
High-dimensional time series
Statistic
Machine Learning
Data Mining
Clustering
Classification
Association Mining
Anomaly Detection
Visualization
OLAP
Optimization
Simulation
Logistic
Transportation
Finance
Retail Analytics
Tourism?

Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data 
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

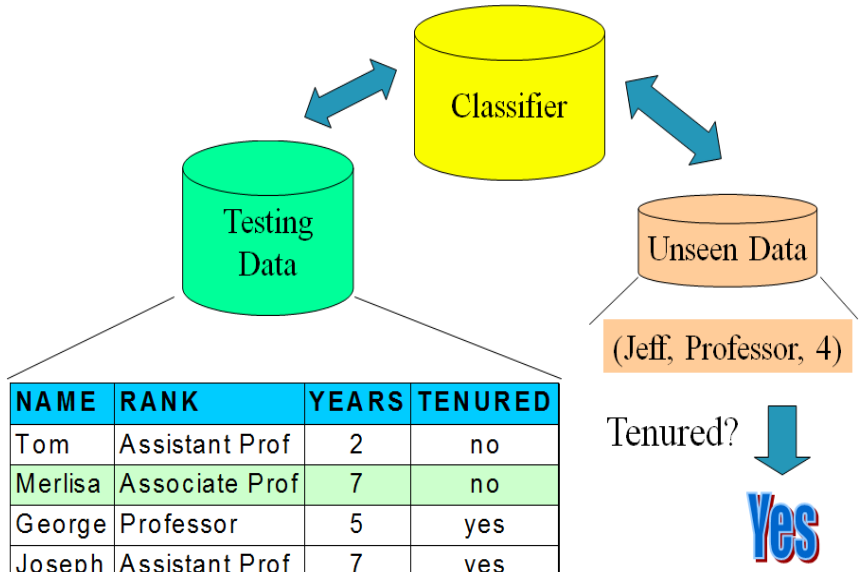
Classification

Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

LAMP: Semi-Lazy Mining Paradigm

- Semi-lazy: Do not pre-construct any model. Given a query for prediction, find a set of nearest neighbors(KNN) , construct model (eg. SVM, Decision tree, HMM) on KNN and then use the model for prediction
  - Like lazy learning
    - We do not commit to a global model but keep the whole historical time series data intact
  - Like lazy learning
    - When making prediction, we invoke a kNN search process
  - Like eager learning
    - We apply complex models on the search results to achieve predictive analysis.
- Goal:

No “concept drifting”, no “information loss”

With “powerful predictive functions”

Closed Book vs Open Book Exam

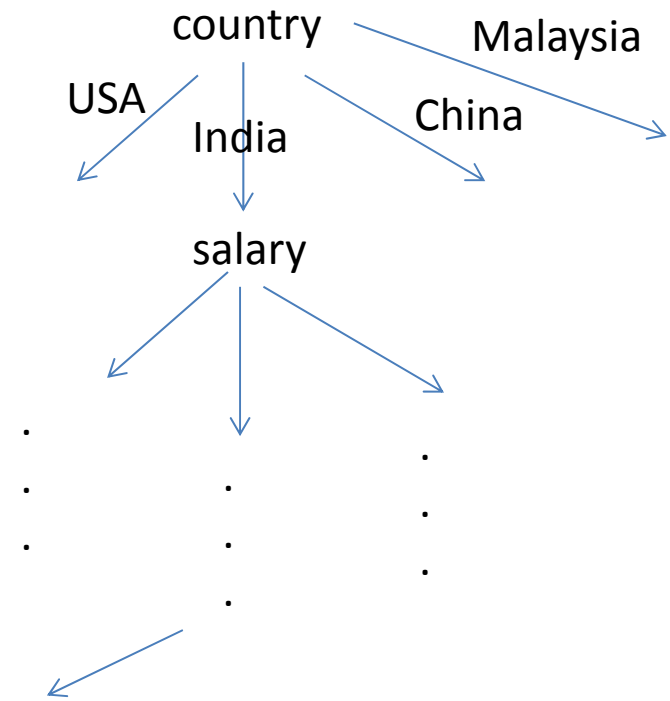
- Eager Learning=Closed Book Exam
  - Build model before the questions arrive
  - Throw away the book(data) but retain the model
  - Must cater to all possible questions with the model
- Lazy Learning=Open Book Exam
  - Bring the whole book(data)
  - After questions are given, search for relevant materials and read
  - If insufficient time, build weak model and give weak answer
  - If sufficient time?

Illustrating Example

- Predict where a new visitor will go and how much he/she will spend in Singapore

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1

Eager Learning



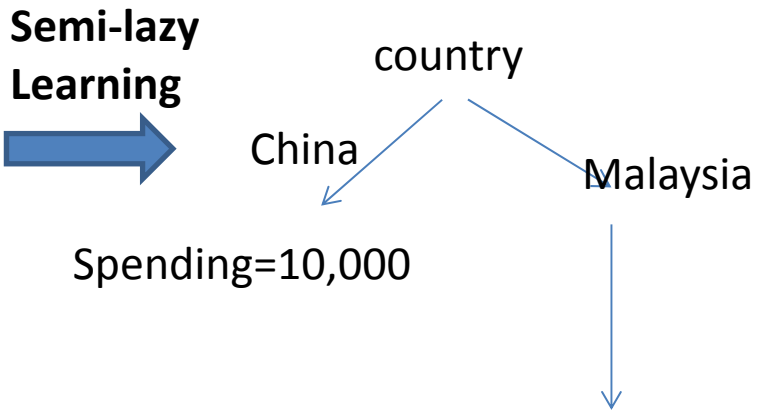
Match again global modal

New visitor, (35, 76k, good, M, Malaysia, ?,?,?,?,?)

Illustrating Example(II)

- Predict where a new visitor will go and how much he/she will spend in Singapore

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1



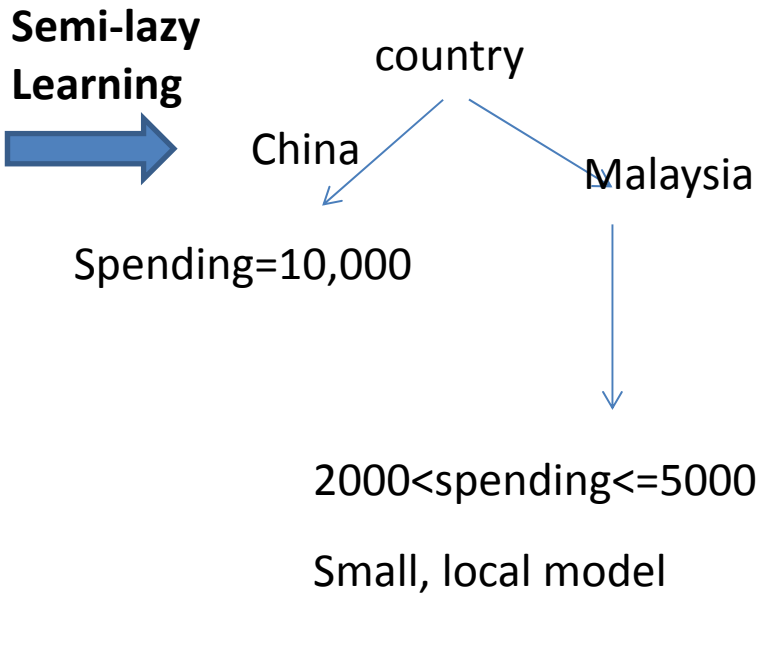
Match against database

New visitor, (35, 76k, good, M, Malaysia, ?,?,?,?,?)

Illustrating Example(II)

- Predict where a new visitor will go and how much he/she will spend in Singapore

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1



Match against database

New visitor, (35, 76k, good, M, Malaysia, 0,1,?,?,?)

Dynamic matching and change model

Illustrating Example(III)

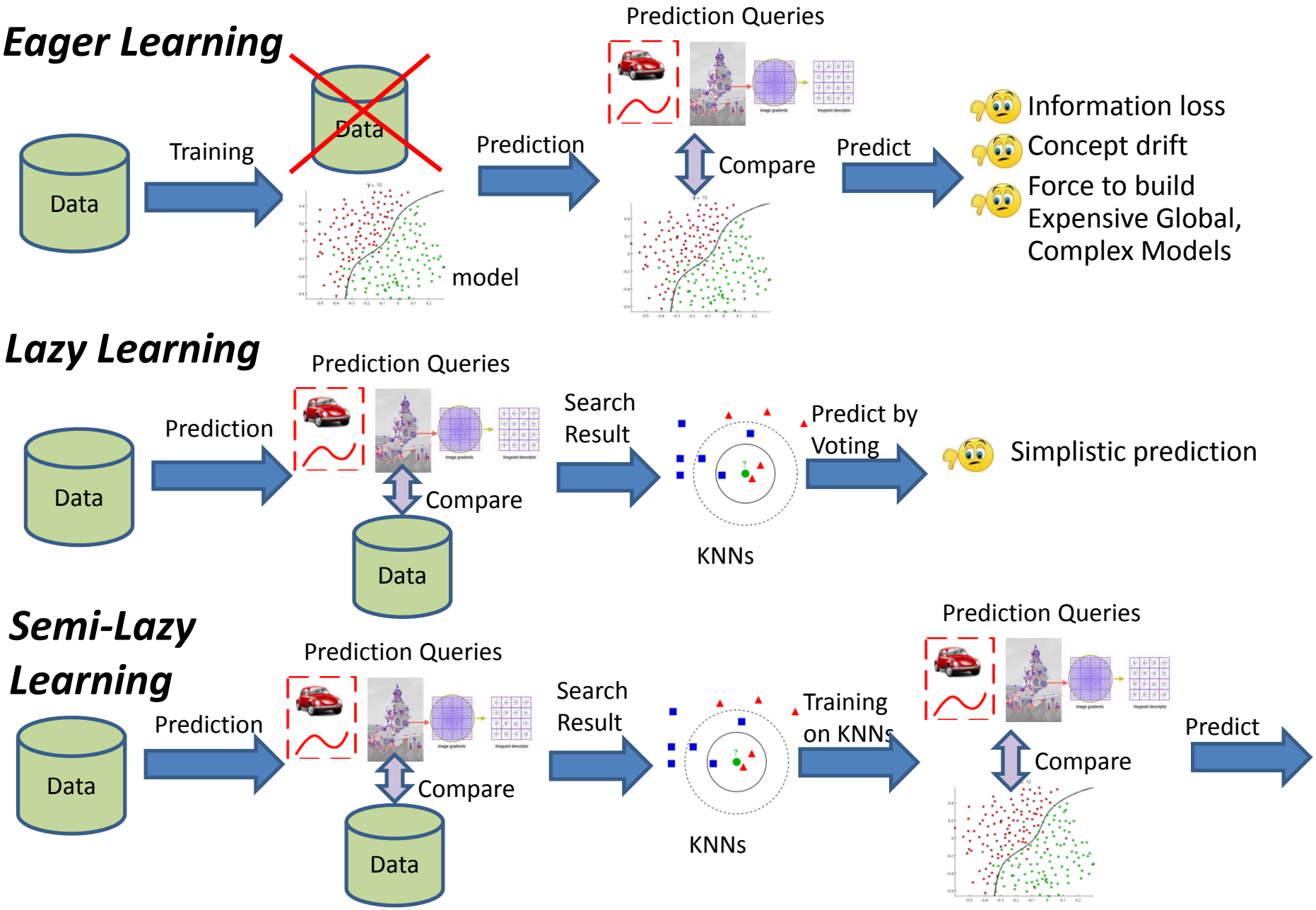
- Predict where a new visitor will go and how much he/she will spend in Singapore

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1
35	76K	good	M	Malaysia	4,000	0	1	0	1

New visitor, (35, 76k, good, M, Malaysia, 0,1,?,?,?)

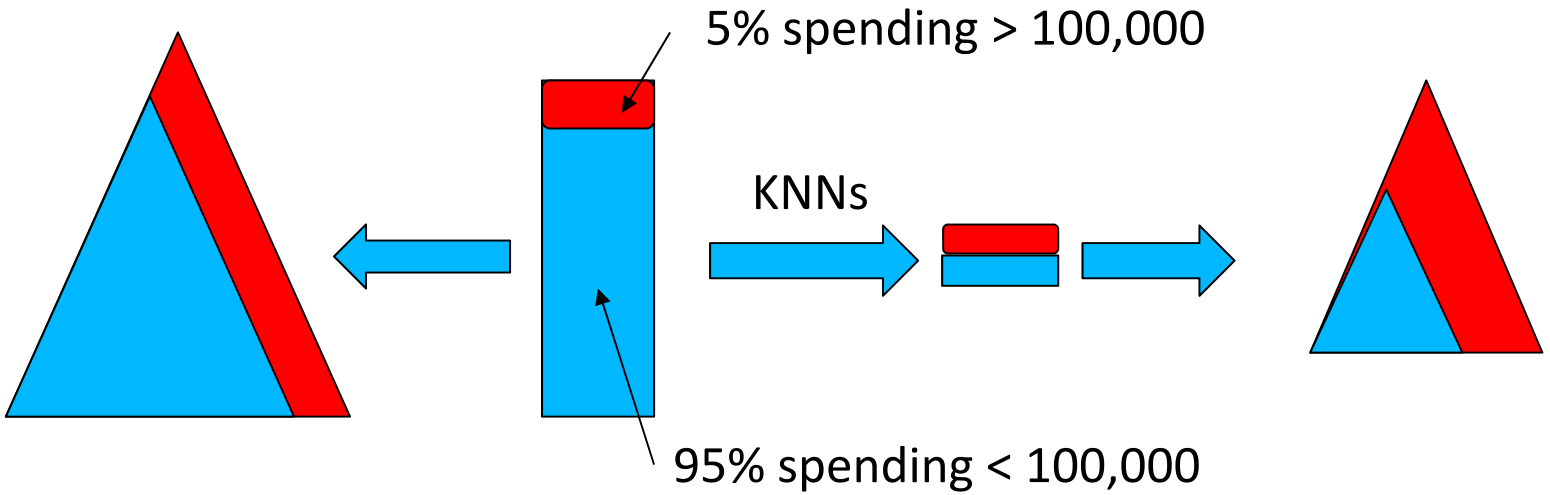
Dynamic Insertion

Motivation for LAMP



Motivation for LAMP(III)

- **Global model vs local model**



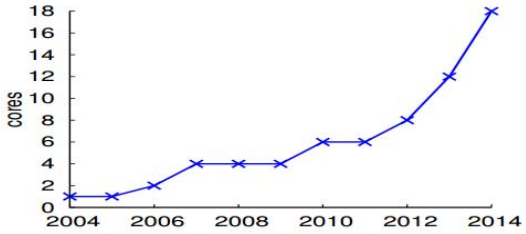
- Existing mining algorithm can be applied without modification
- Grow in hardware will increasing make just-in-time model construction feasible
- Support micro to macro prediction

Global model dominated by majority class: ***Big data build big, complex model but not necessary good model***

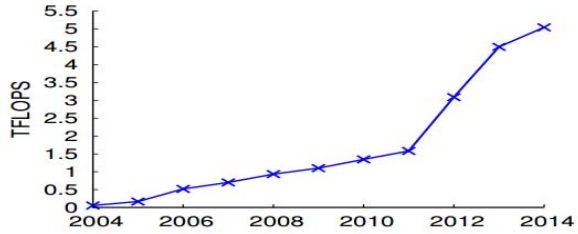
Local model catering to specific query: **Right data is more important**

Will LAMP be too slow?

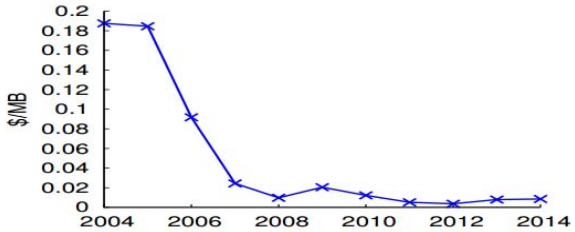
- Growth in modern hardware will substantially reduce time to construct local model on say 1000 **well-selected** training examples



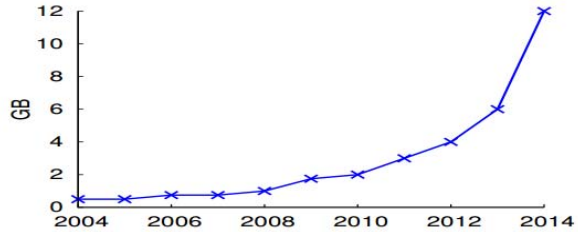
(a) CPU cores



(b) GPU single precision FLOPS



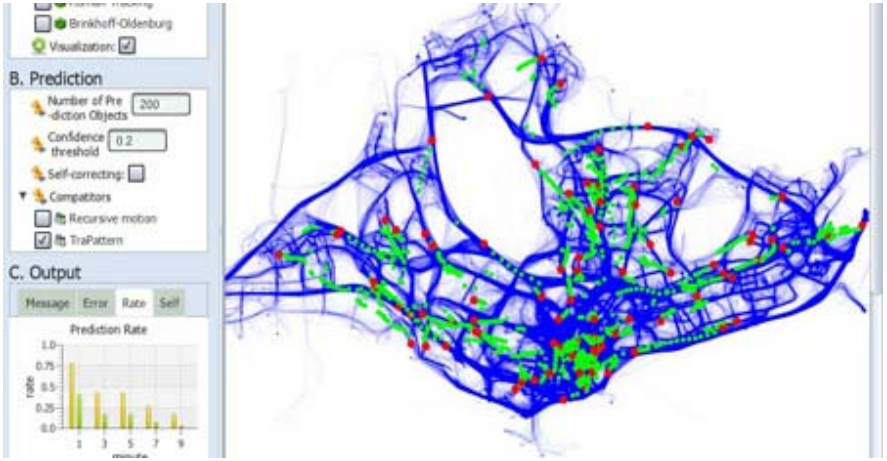
(c) CPU memory price

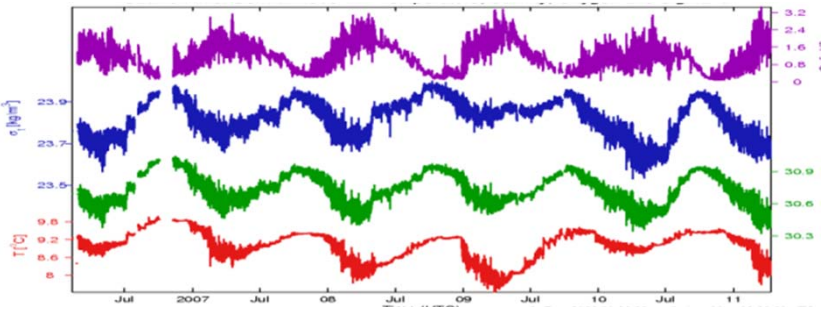


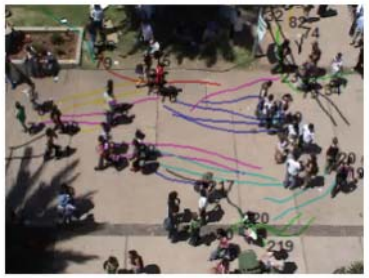
(d) GPU memory size

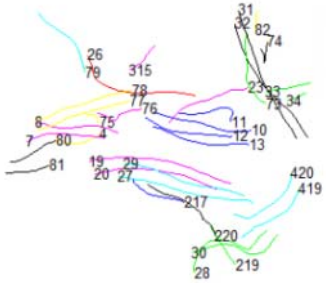
- Physical properties in real world don't change or grow too fast
  - Number of roads and speed of cars in most places
  - Number of sensors one put on the body and how fast the heart beat
  - How fast you type and number of photographs you take with your camera
- **Eventually, processing power overtake amount of useful data that can be generated especially in the context of Information-on-the-Go**

Published Result of LAMP







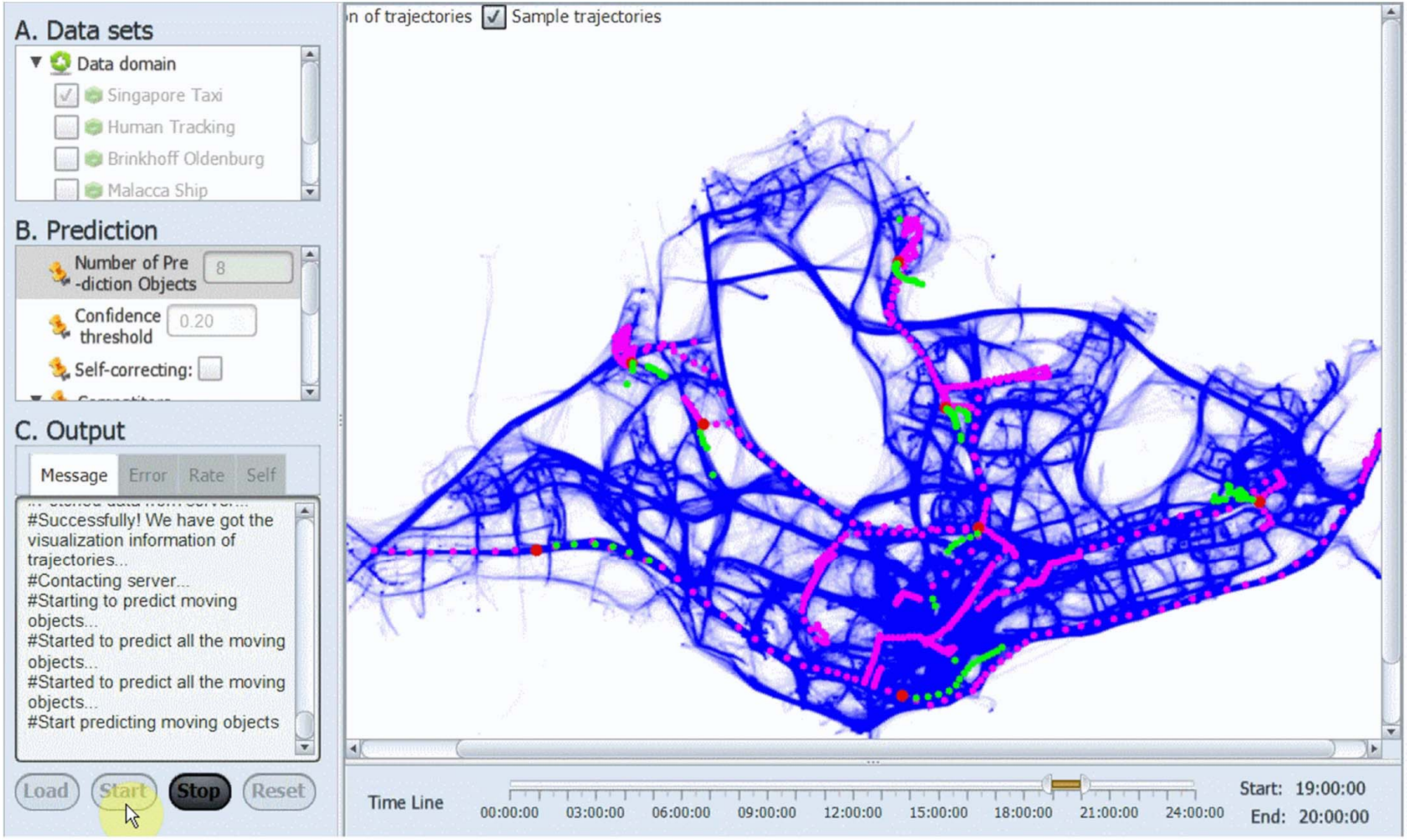


Applied in trajectory prediction(KDD'2014). Given a vehicle's present movement, find k similar historical trajectories and construct model dynamically to predict how the vehicle will move next.

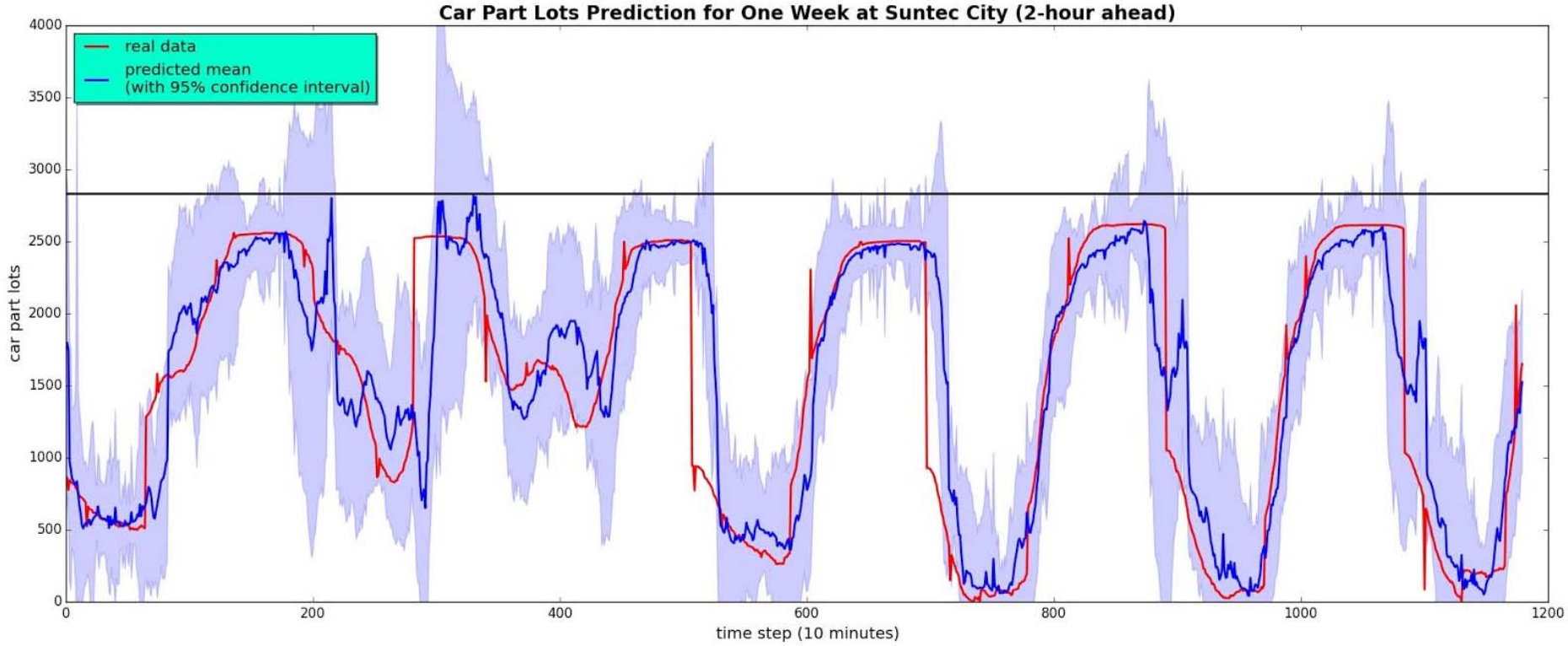
Time series prediction on sensors (SIGMOD'2015)

http://db128gb-b.ddns.comp.nus.edu.sg/jzhou/R2-D2/

LAMP and GENIE Example:“Semi-Lazy” Path Prediction(II)

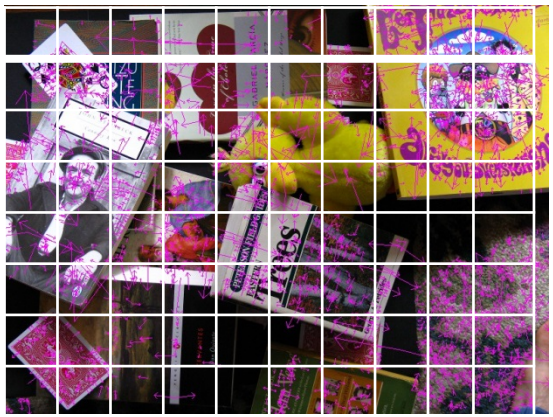


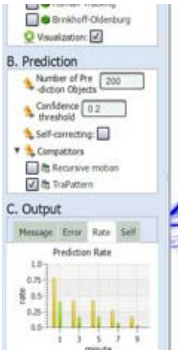
Car Park Lot Availability Prediction

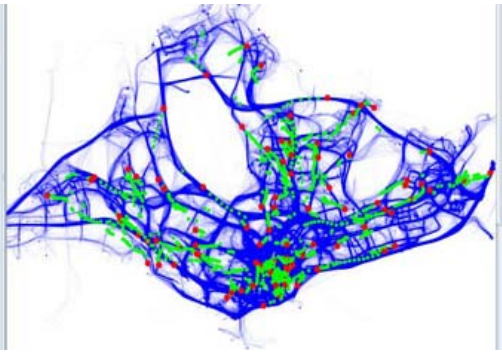


Many Little LAMP: Motivation

- Many applications for applying LAMP to large number of KNN sets







Query Document

W1	W2	W3	W4	W5	W6	W7
0	2	3	2	7	1	9

W1	W3	W3	W5	W7	W6
0	3	3	7	9	1

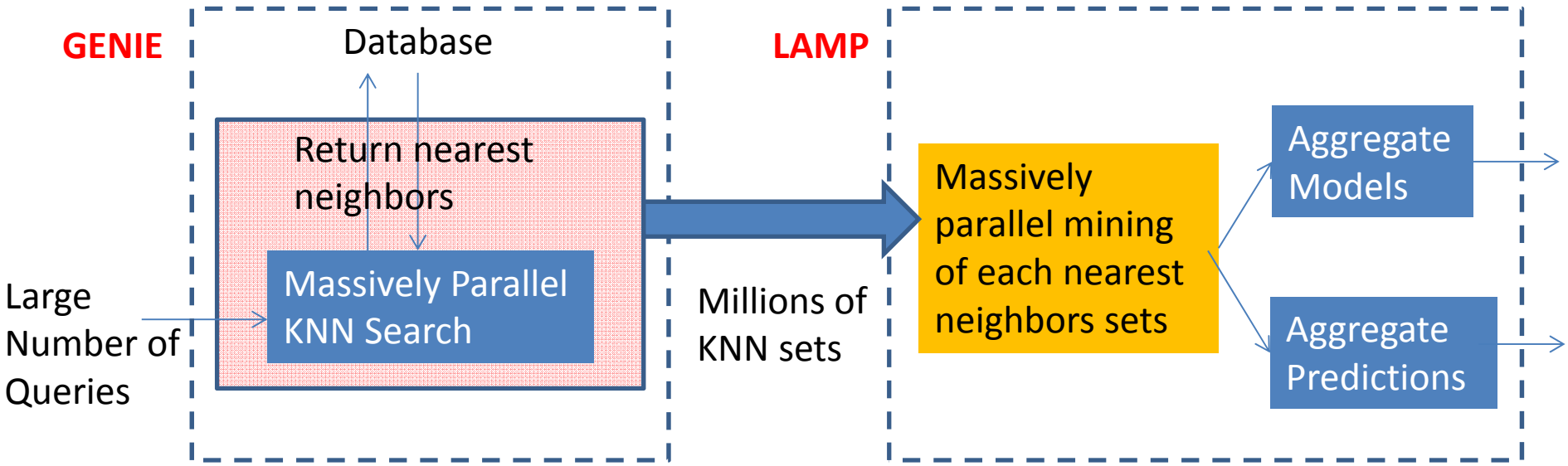
.....

Apply KNNs on each grid, classify each grid by constructing a classifier and then aggregate the result of all the grids

Perform prediction for every vehicles and aggregate them to predict traffic conditions

Randomly project prediction query on different subsets of attribute values find KNNs for different subsets, construct classifier for each subset and aggregate prediction

LAMP and GENIE



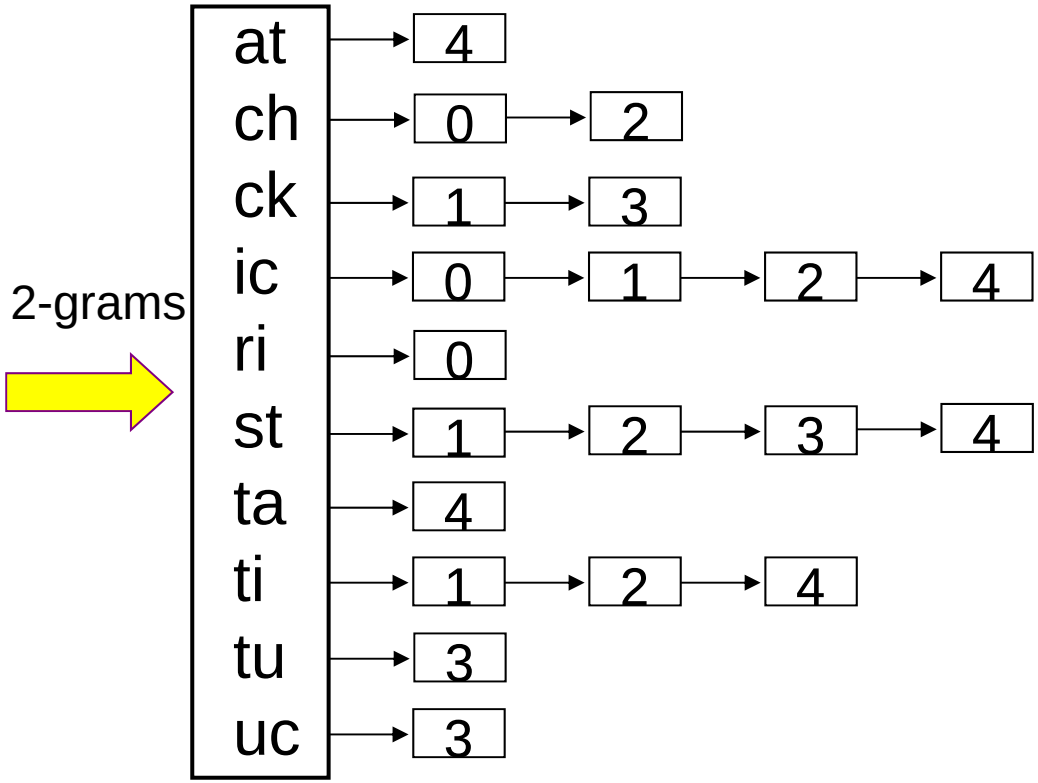
GENIE: A generic inverted index system to supports large number of k-Nearest Neighbors (kNN) search on large variety of complex data including sequences, trees, graphs and high dimensional data

LAMP:**Semi-lazy mining paradigm** which adopt a **Just-In-Time** construction of prediction model only AFTER a prediction request arrive

Rely on GENIE to find sets of kNN and construct many local models over these kNN sets in parallel

Similarity Search for Complex Data(GENIE) : Shotgun, Search and Assemble using Inverted Index

id	strings
0	rich
1	stick
2	stich
3	stuck
4	static



Searching using inverted lists

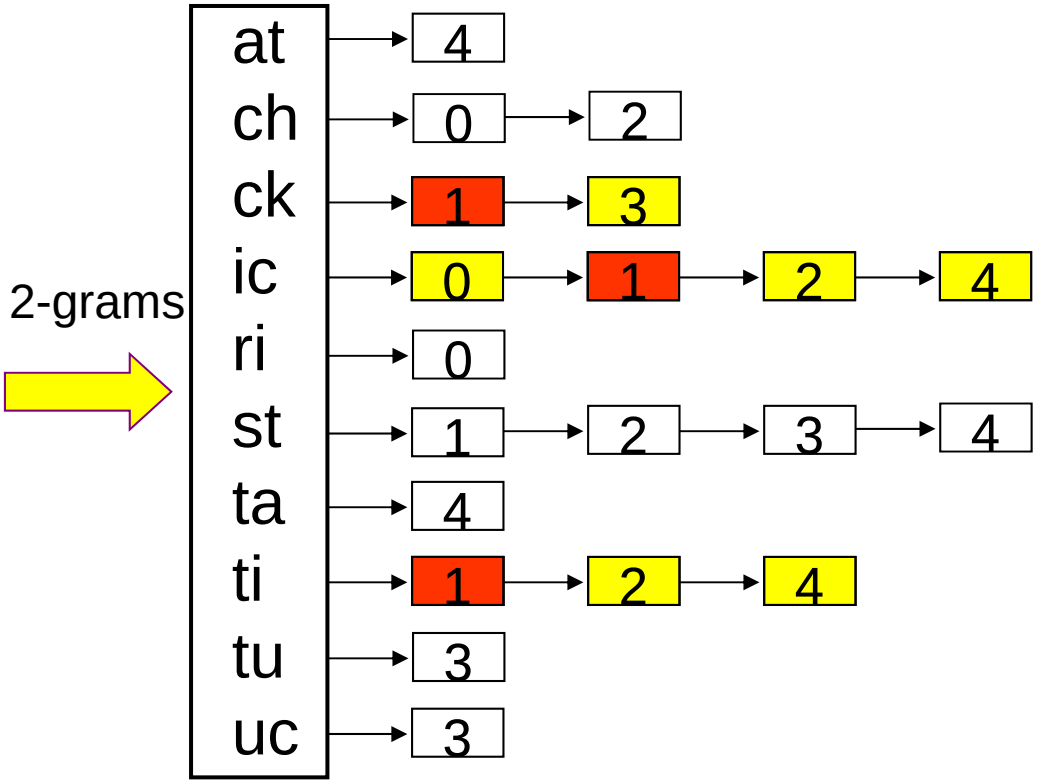
- Query: “shtick”, ED(shtick, ?)⩽1

sh ht ti ic ck

# of common grams >= 3

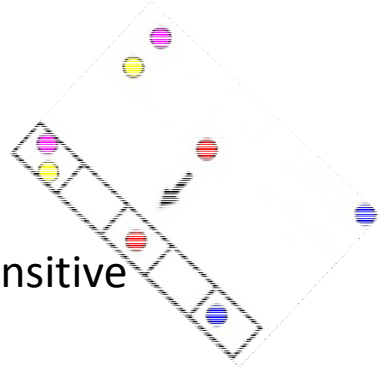
id	strings
0	rich
1	stick
2	stich
3	stuck
4	static

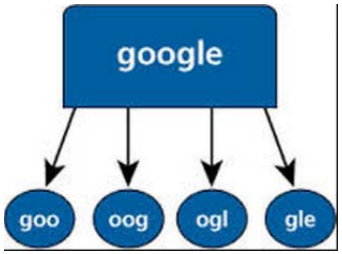
2-grams



Similarity Search for Complex Data(GENIE) : Shotgun, Search and Assemble using Inverted Index

Local Sensitive Hashing



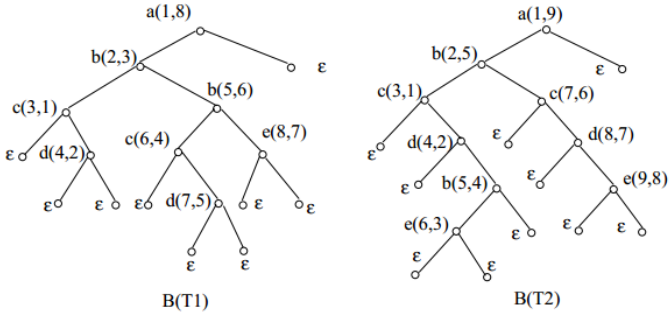


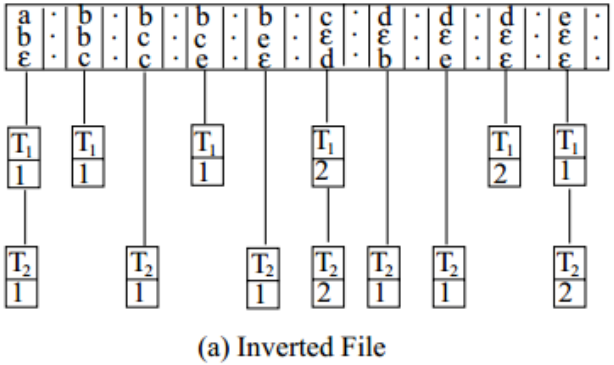
ID	String
10	Introduction
20	introductive

5-grams

age	salary	credit	sex
20	30k	poor	M
40	100k	poor	M
50	110k	good	F
60	50k	good	M
75	15k	poor	M

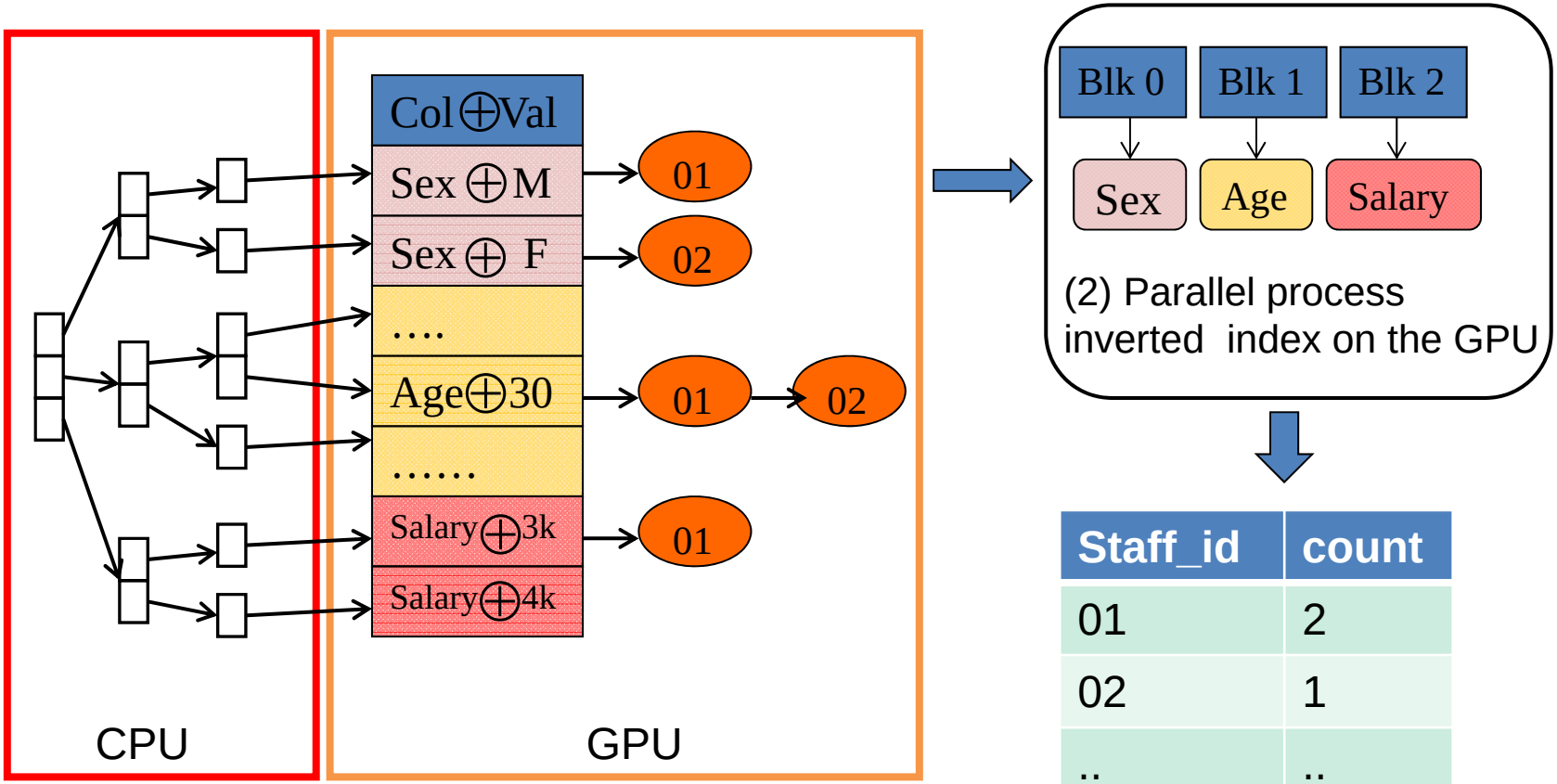
trodu	uctio	uctiv	ntrod	oduct	roduc	ction	intro	ctive	ducti
10 20	10	20	10 20	10 20	10 20	10	10 20	20	10 20





(a) Inverted File

GPU-GENIE:
Parallel kNN Searching on the GPU



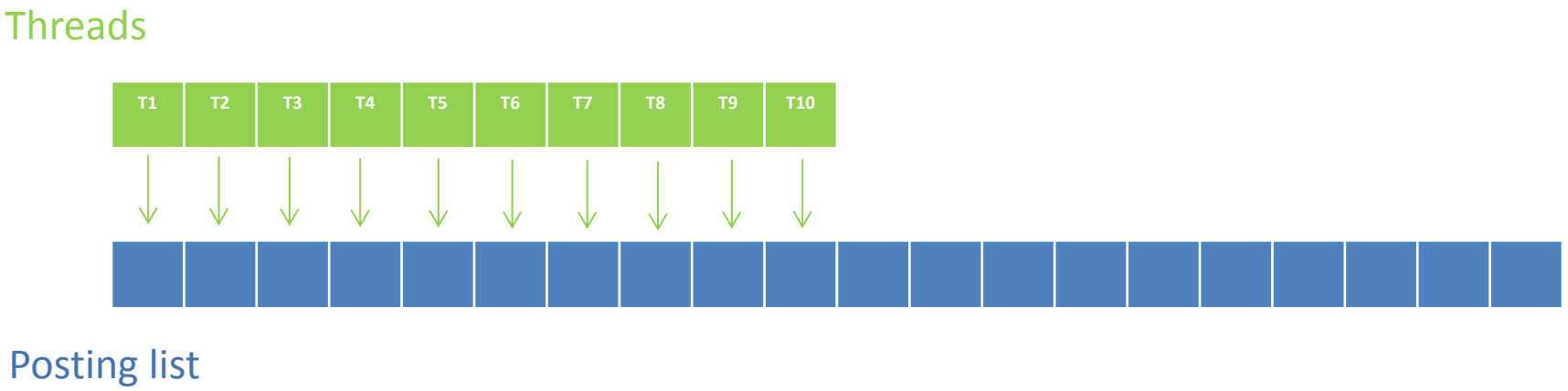
Staff_id	count
01	2
02	1
..	..

(1) Inverted Index on the GPU.
(A tree index on the CPU to organize the inverted index on the GPU)

(3) GPU k selection to get results

GPU Parallel Processing

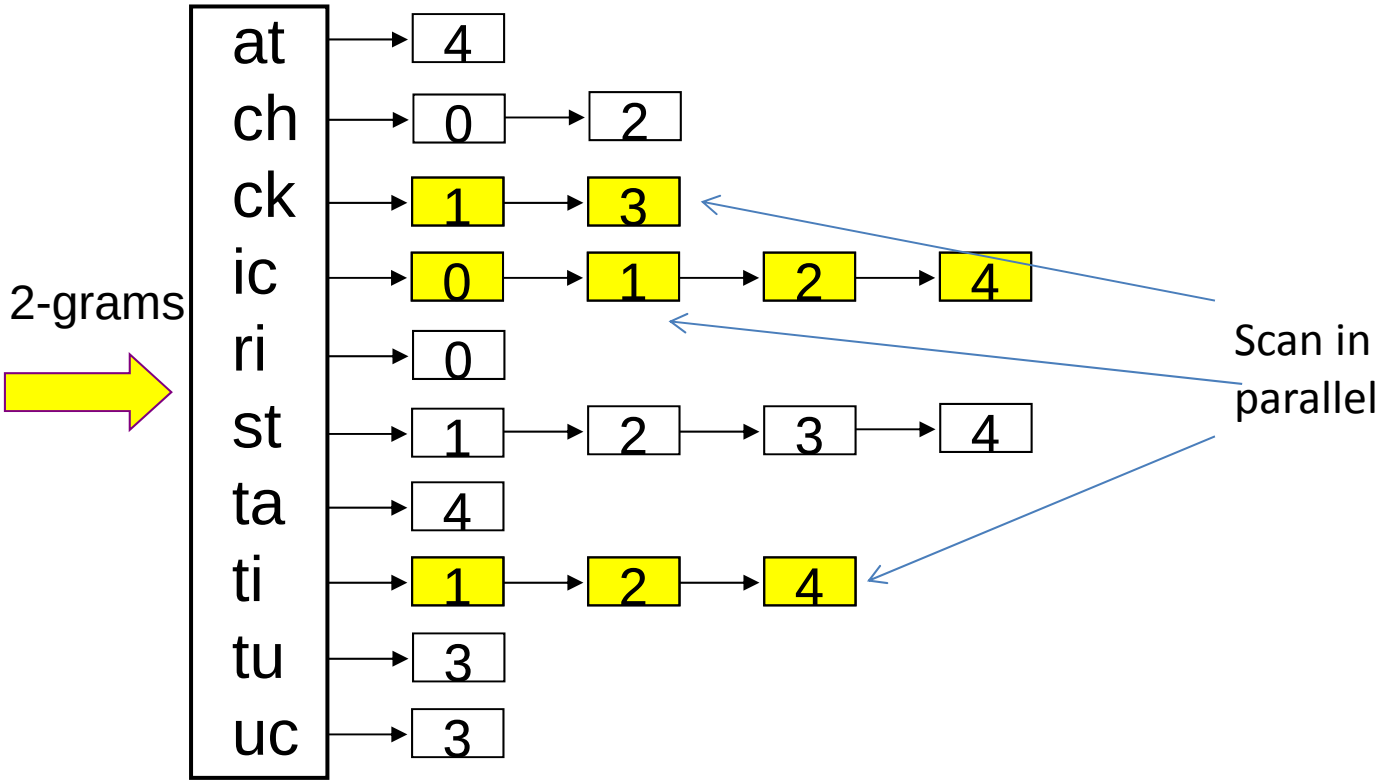
- Each thread in a thread block accesses an element in the inverted list and performs aggregation in parallel.



- A thread block has 1024 threads for modern GPUs.

Searching using inverted lists

id	strings
0	rich
1	stick
2	stich
3	stuck
4	static

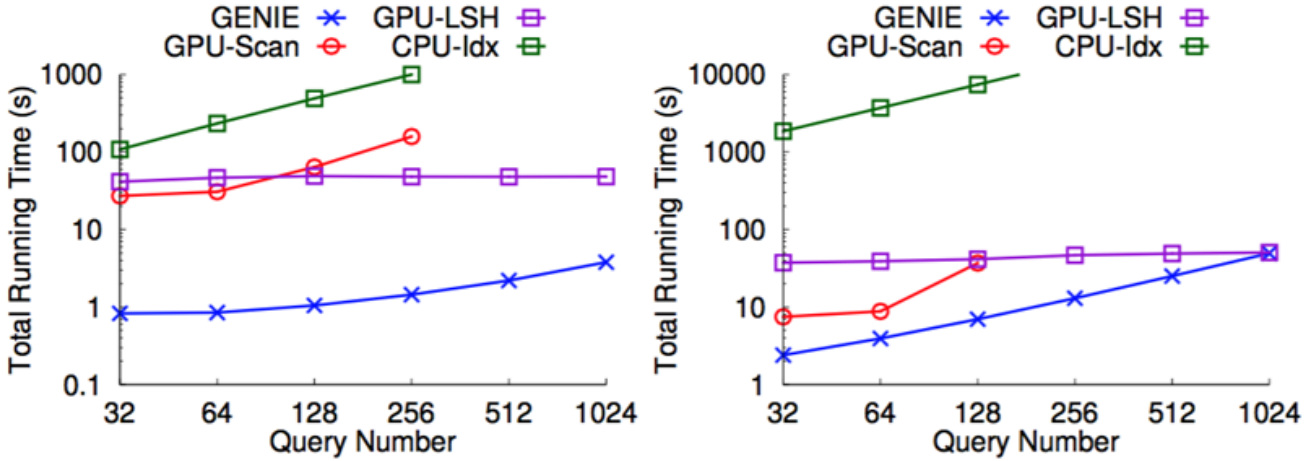


Multiple queries in parallel: shtick, pich, skich, puck, tatics,

Optimizations

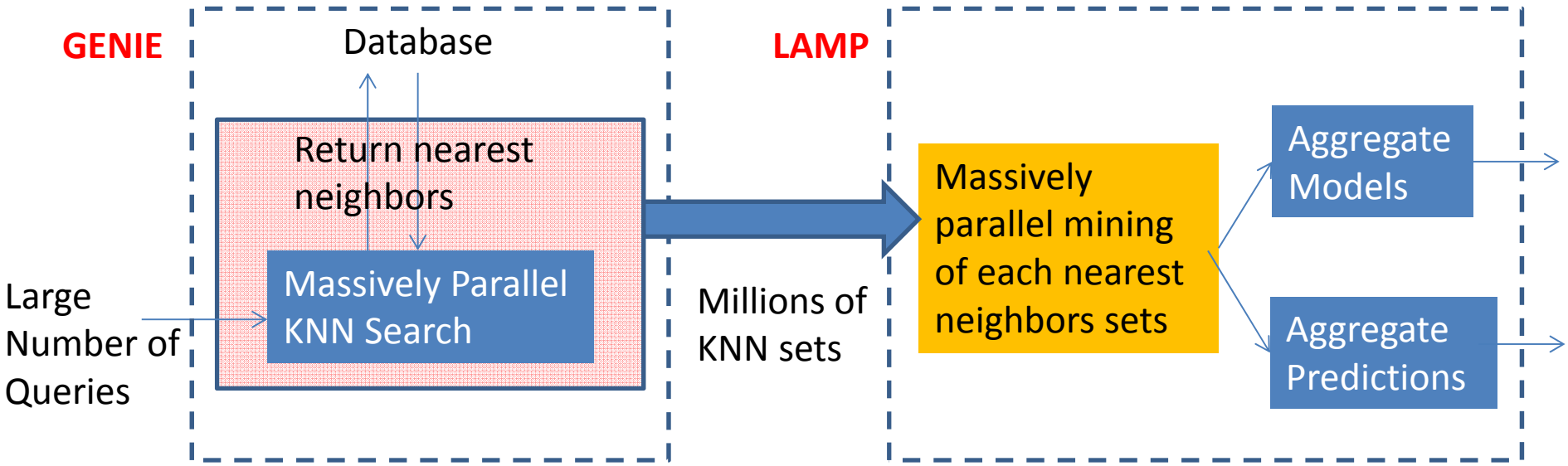
1. Bitmap Counter
   Compress the whole Count Table in a bitmap structure.
2. Hash Table
   Insert a small number of candidate objects into a hash table during the update of the bitmap counter. Later k-selection on this hash table only.
3. Load Balancing
   There may be some extreme long posting lists which can become the bottleneck of the system. Therefore the length of posting lists are limited and long lists touching the limit are divided into a set of sub-lists.

Performance



Total running time for multiple queries

LAMP and GENIE



GENIE: A generic inverted index system to supports large number of k-Nearest Neighbors (kNN) search on large variety of complex data including sequences, trees, graphs and high dimensional data

LAMP:**Semi-lazy mining paradigm** which adopt a **Just-In-Time** construction of prediction model only AFTER a prediction request arrive

Rely on GENIE to find sets of kNN and construct many local models over these kNN sets in parallel

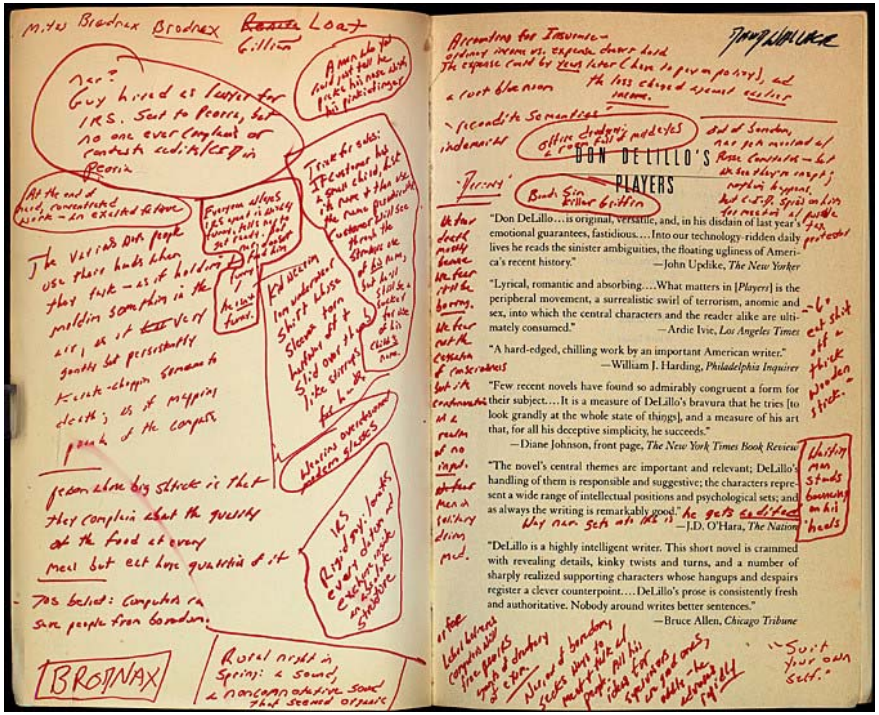
Outline

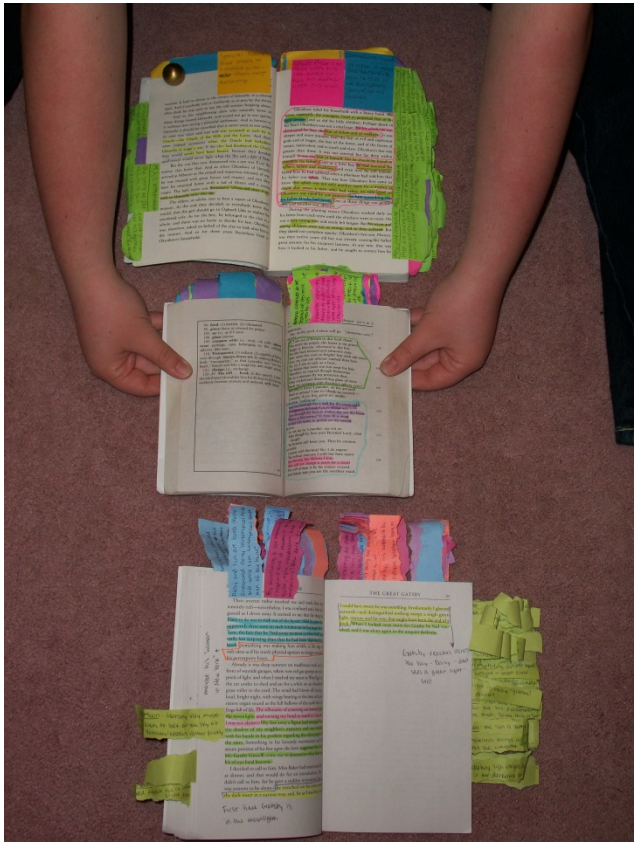
- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

Motivation: Collaborative Social Network System

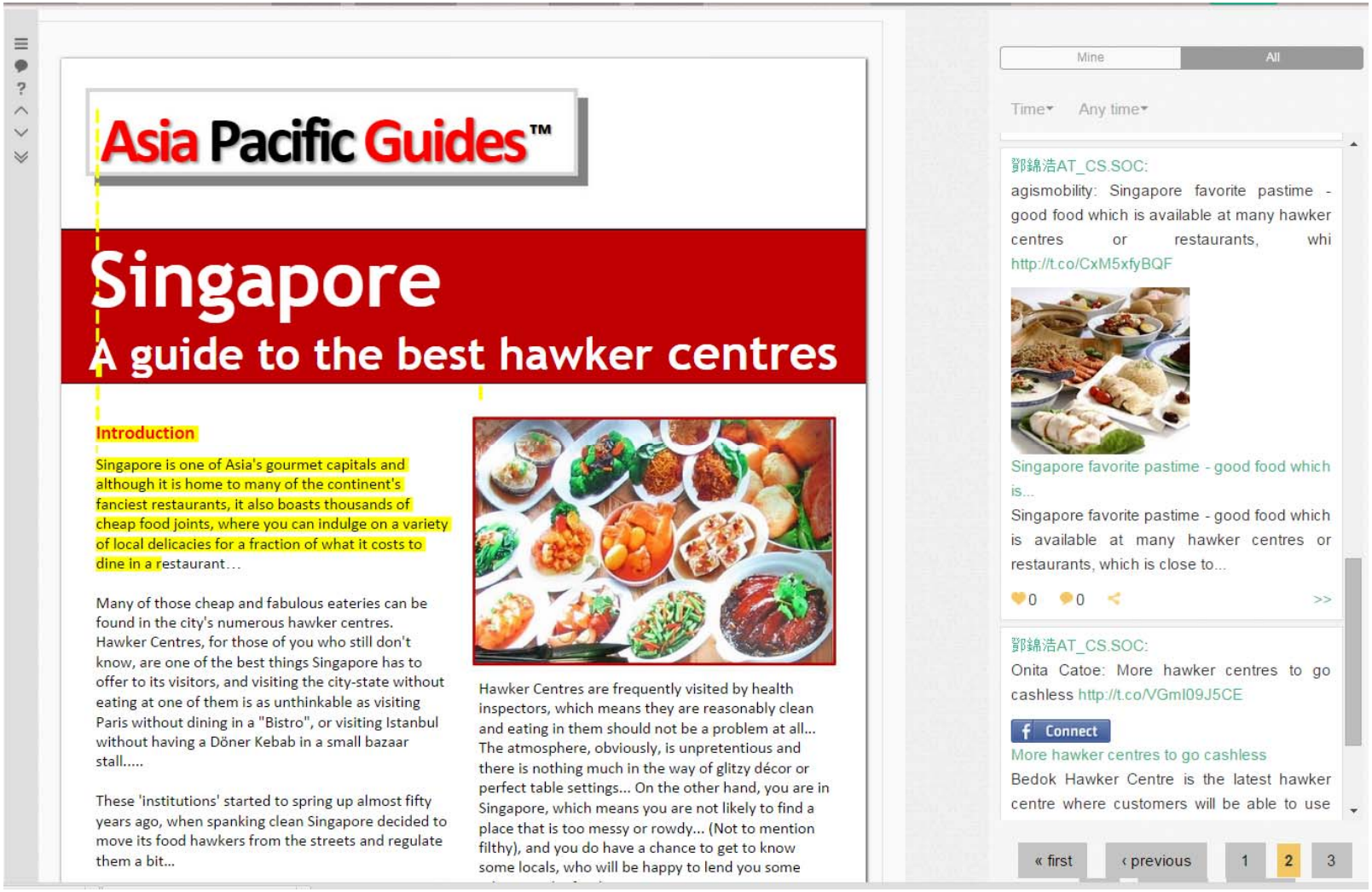
- "Things fall apart; the centre cannot hold; Mere anarchy is loosed upon the world"---- -- The Second Coming
- **Problem:** People, information are extremely fragmented and getting the right information from the right people at the right time is extremely difficult
- **Solution:** Build social communities and index information around certain "centers"

Manual Book Annotations





Readpeer: Building Social Communities Around Documents and Books



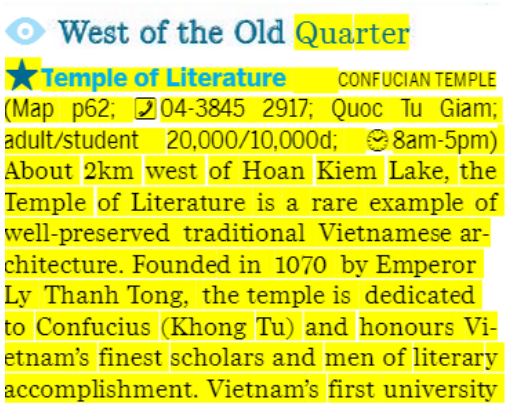
https://www.facebook.com/akhtung/videos/pcb.10155214811333972/10155214806188972/?type=3&theater

ReadPeer: Annotated Text are Recorded instead of Location in the Text

Users	Annotated Text in Food Guide	Annotations
Anthony	***Hainanese Chicken Rice***, one of "Singapore's national dishes", is named after the island of Hainan, in South China, where it originates from.	
Bernard	prepared in traditional Hainanese methods and served with aromatic rice that has been prepared with pandan leaves. The dish is then served with a variety of sauces, including chili, ginger and dark soy sauce.	
Anthony	Tiong Bahru Market is a newly reopened complex, relatively more spacious and pleasant than the ordinary Hawker Centre, and with a wider selection of excellent stalls (almost one hundred...) and some nice 'alfreso' and in doors seats.	
Chris	Lau pa sat Festival Market: Originally built in 1894 as a fish market, Telok Ayer Market (as it is better known) is one of the best and most popular Hawker/Food centres around this side of Chinatown. The imposing castiron structure of the market was prefabbed in Glasgow, Scotland, more than a century ago and shipped to Singapore in pieces, before being erected on si	

Massively Parallel Approximate Text Matching Index and Analytics

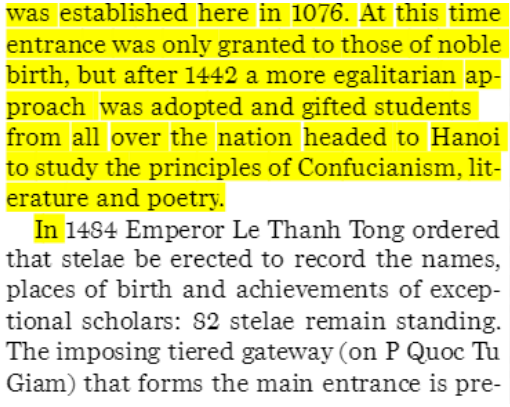
Readpeer: Annotations are Free to Flow to Any place using Approximate Sequence Matching



West of the Old Quarter

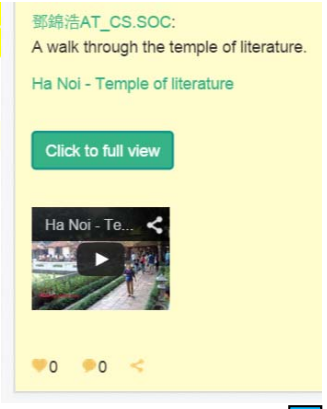
Temple of Literature CONFUCIAN TEMPLE

(Map p62; 04-3845 2917; Quoc Tu Giam; adult/student 20,000/10,000d; 8am-5pm) About 2km west of Hoan Kiem Lake, the Temple of Literature is a rare example of well-preserved traditional Vietnamese architecture. Founded in 1070 by Emperor Ly Thanh Tong, the temple is dedicated to Confucius (Khong Tu) and honours Vietnam's finest scholars and men of literary accomplishment. Vietnam's first university



was established here in 1076. At this time entrance was only granted to those of noble birth, but after 1442 a more egalitarian approach was adopted and gifted students from all over the nation headed to Hanoi to study the principles of Confucianism, literature and poetry.

In 1484 Emperor Le Thanh Tong ordered that stelae be erected to record the names, places of birth and achievements of exceptional scholars: 82 stelae remain standing. The imposing tiered gateway (on P Quoc Tu Giam) that forms the main entrance is pre-



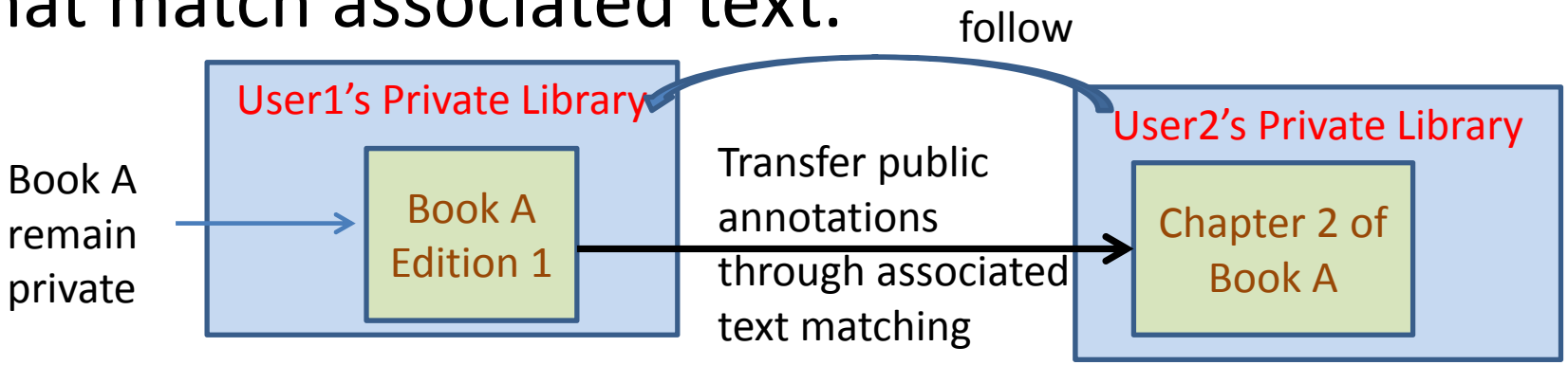
Different copy and version of the same ebook

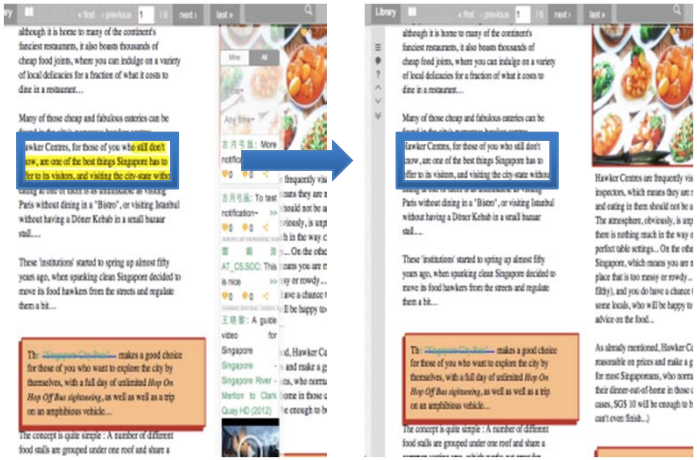
Different Format

Different Platform

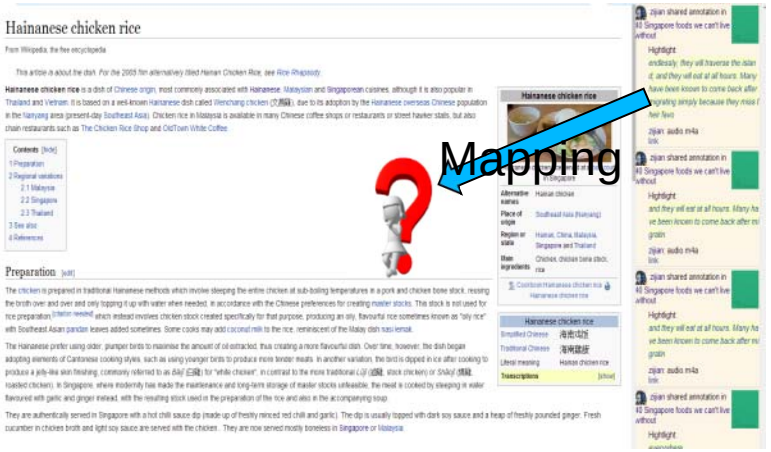
Readpeer: TakeLeaf

Transfer annotations without sharing the same copy of document or book. Adopt fast indexing method that match associated text.



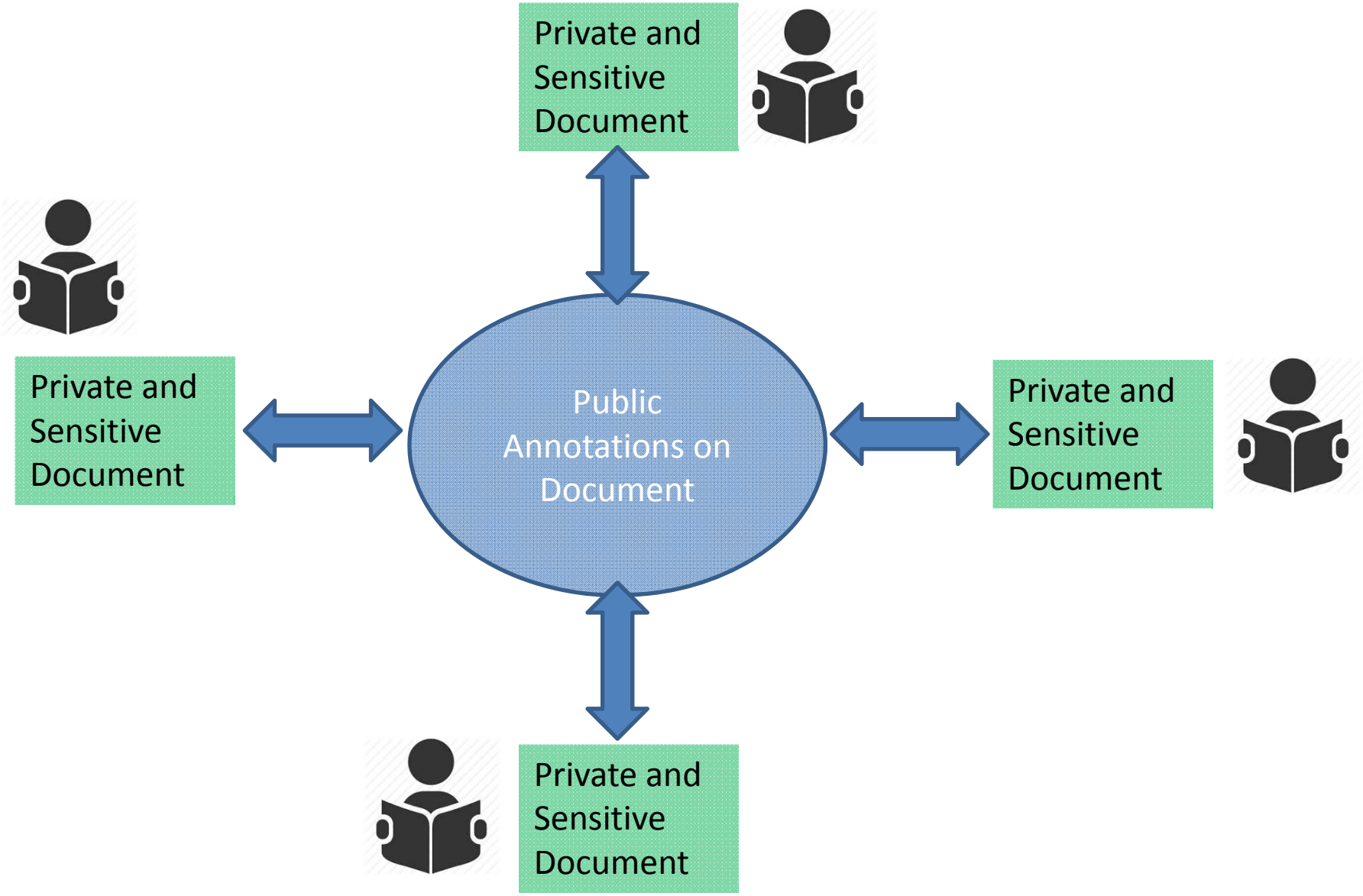


Book annotation migration problem across different edition of books



Annotations retrieval during web browsing using matching paragraph

Readpeer: Copyright and Content Protection



TakeLeaf(III)

DocRicher: An automatic annotation tool

Transfer public annotations through text matching

TakeLeaf: Migrate and synchronize annotations for various book versions

Technical support

Book index

Text of A

Book A

Annotation index

Version graph

Integrated graph

Search result

Migrated annotations to Book A	
Annotation1	From Book B
Annotation2	From Book C

Offline processing Book B, C, and F

User interfaces

Add annotations to Book A

The updated node

The referenced node

After processing:

1. New annotations from A will be synchronized into B, C, and F
2. New referenced adages will be added to reference graph if other results are found

Demo video

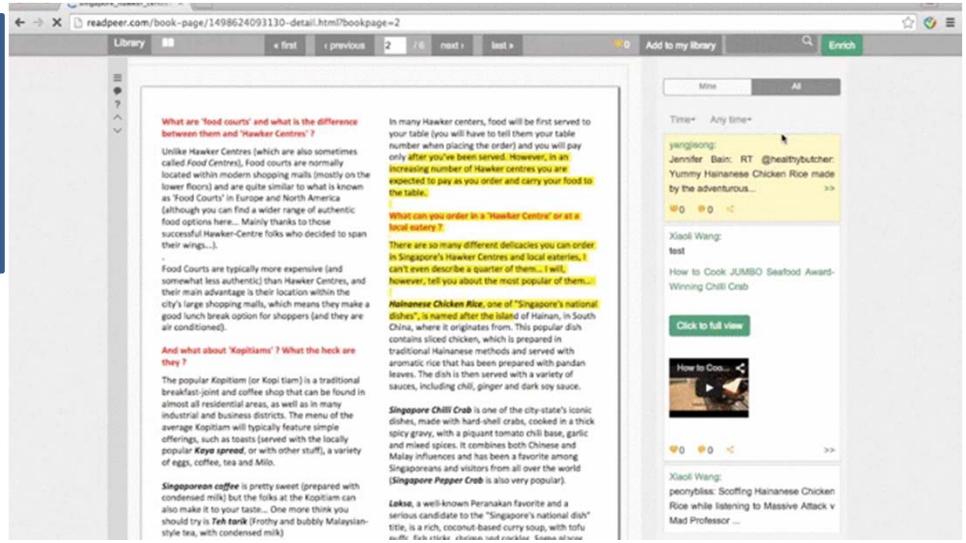
DocRicher: Enrich a textbook using social media

DocRicher: An automatic annotation tool

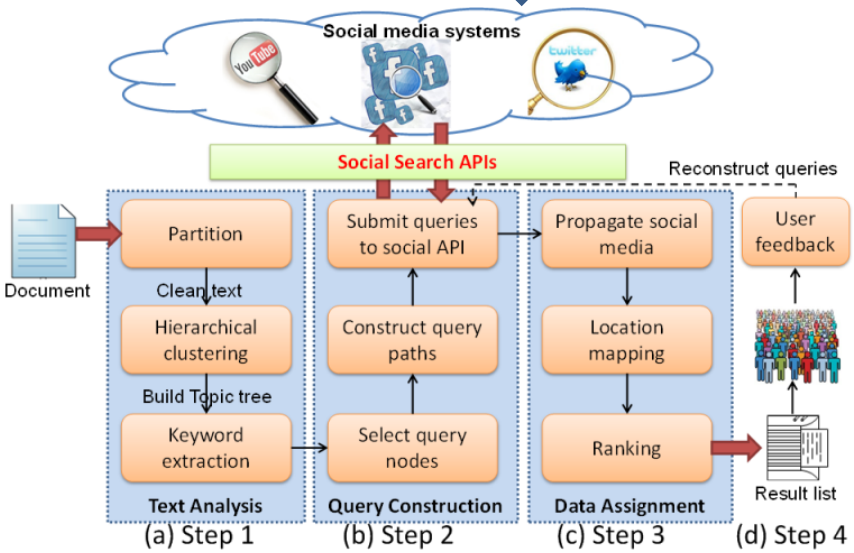
Analyze texts, query social media from social systems

LiveBook: Suggest annotations using social media like tweets

Demo video



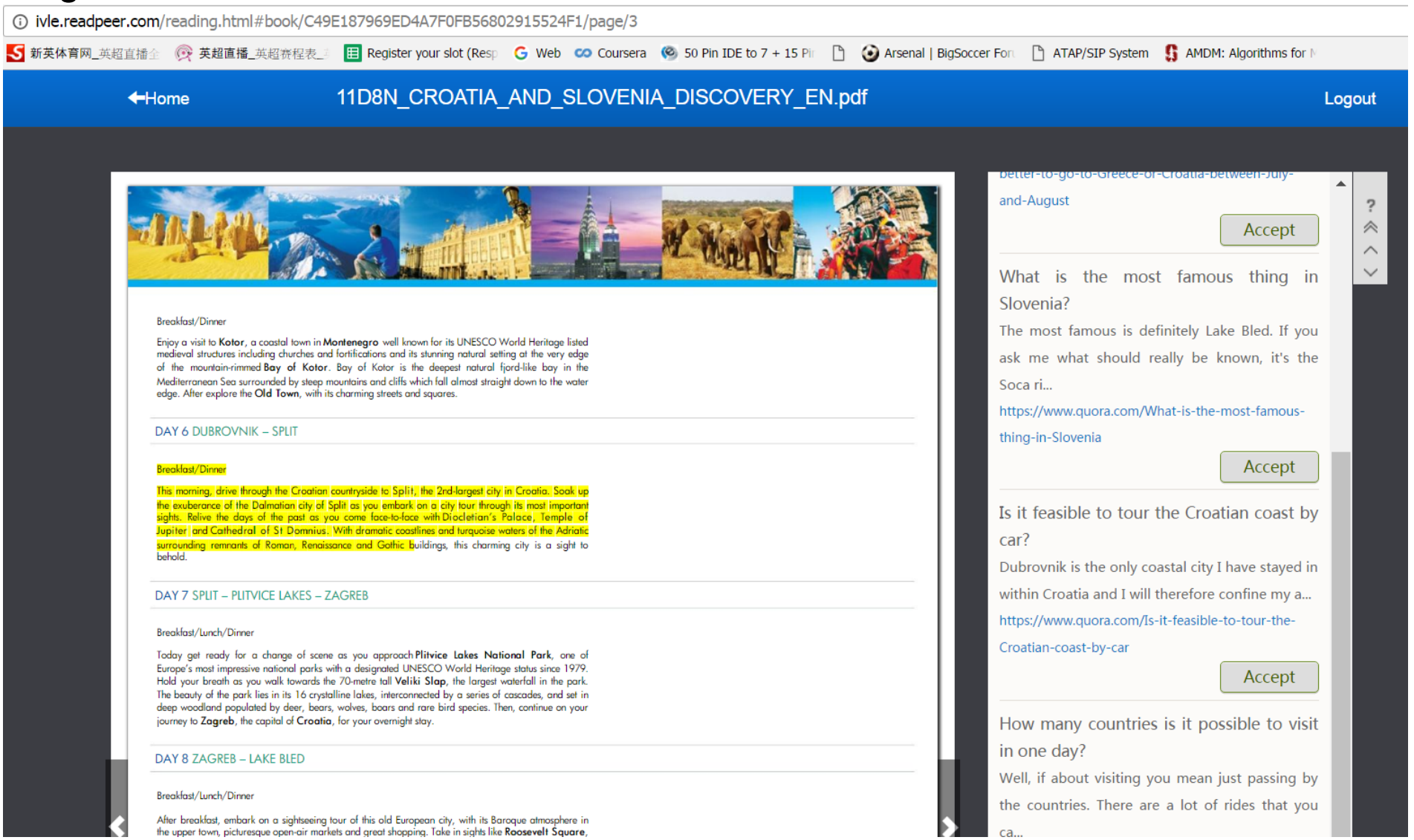
Technical support



The implementation of user feedback

Readpeer: Q&A Linking

Link relevant contents from Quora, Stackoverflow and Zhihu to document segments



Readpeer for Education

- Problem with current MOOC: Basically still a lecture-based approach with only one way of explanation by a single, smart lecturer(生而知之）
- Allows multiple ways of explanation for the same materials by different student mentors many of which went through the difficulty of learning the material(学而知之，困而知之)
- Student mentors learn through teaching
- Encourage students to read the book and then search for help when needed instead of relying on lectures right from start (i.e. self learning)

Readpeer for Education(II)

- Can we be the “Uber” of online education?

Students

Readpeer

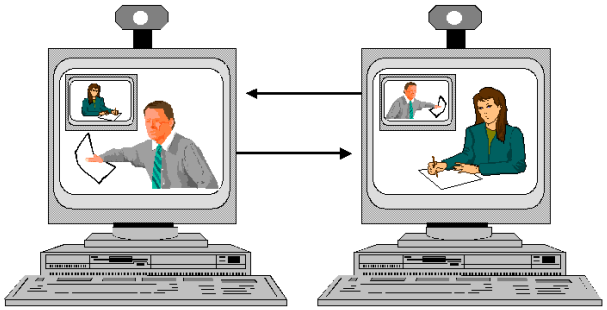
Mentors with Diversified View of the Textbook

Students

Book

Follows

Mentors



Online Video Question and Answering

Year 4 U. Students

teach

Year 3 U. Students

teach

Year 2 U. Students

teach

Year 1 U. Students

teach

Year 3 SM Students

teach

Year 2 SM Students

teach

Primary 1 Students

Everyone is a teacher to his/her juniors

Readpeer for Education(III)

- How do we motivate student mentors to make annotations on books?
- Answer: A proper ranking system based on
  - People following them
  - Number of online video questions and answer sessions
  - Provide annotations that are liked, with lots of discussions
- Becoming a respected expert in a book must be very desirable (fame, CV and earnings)

Readpeer: Other Target Domains

- Book Publishing
  - Books are free
  - Monthly subscription for author's updates and participation in the book community
- DIY/Sport Guide
  - Books are free
  - Online annotation services that provide training and guidance are chargeable
- Personal Social Feeds Indexing
  - Tie social network accounts to Readpeer
  - Attached all shared social messages to relevant part of books in Readpeer's personal library

Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

Cloud-based Augmented Reality for Shops

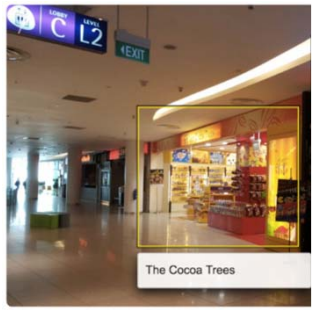
https://www.facebook.com/akhtung/videos/pcb.10155214827118972/10155214820483972/?type=3&theater

- Provide mall and shop owners cloud-based tools to use augmented reality in their malls and shops
- Provide shoppers to search for items in their shopping list and provide comments using AR

Shop owner

Take images →

Add annotations →



← Retrieve similar Images & transfer annotations

Shopper

Take an image →

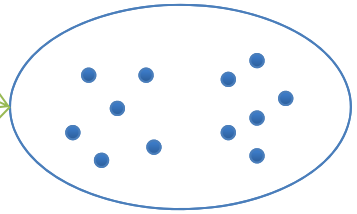
Two-level clustering

WIFI signal

WIFI Clusters

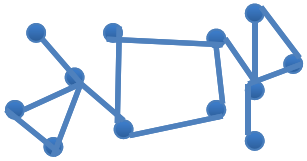
Take a cluster

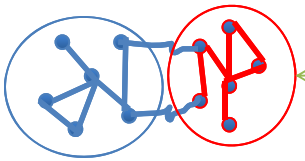
Cluster of Spatial Features



One cluster

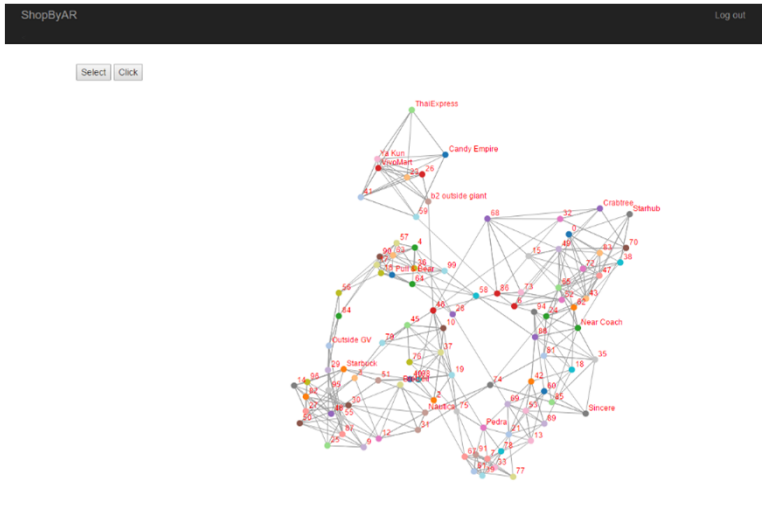
Connect visually similar images



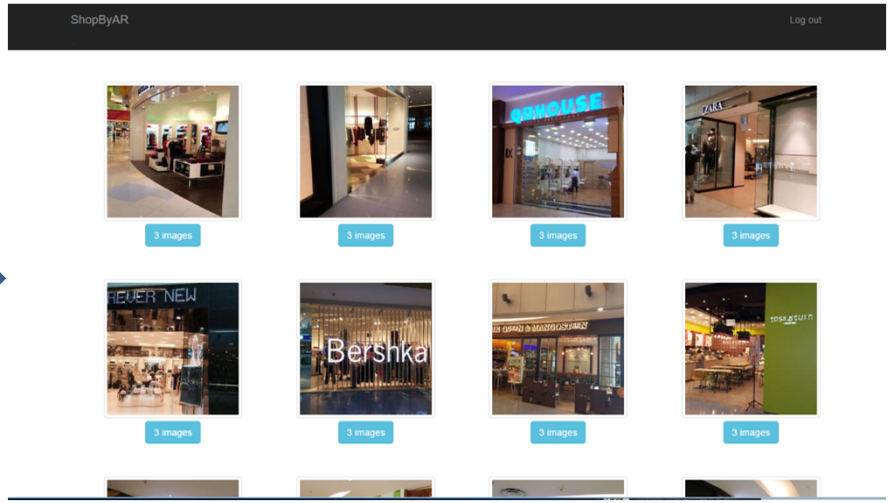


Spectral clustering

Cluster of Visual Features

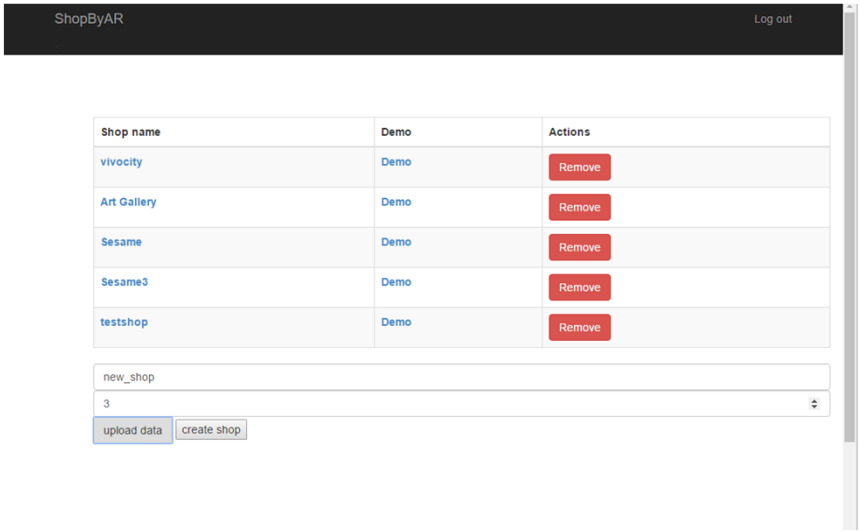


WIFI Clustering

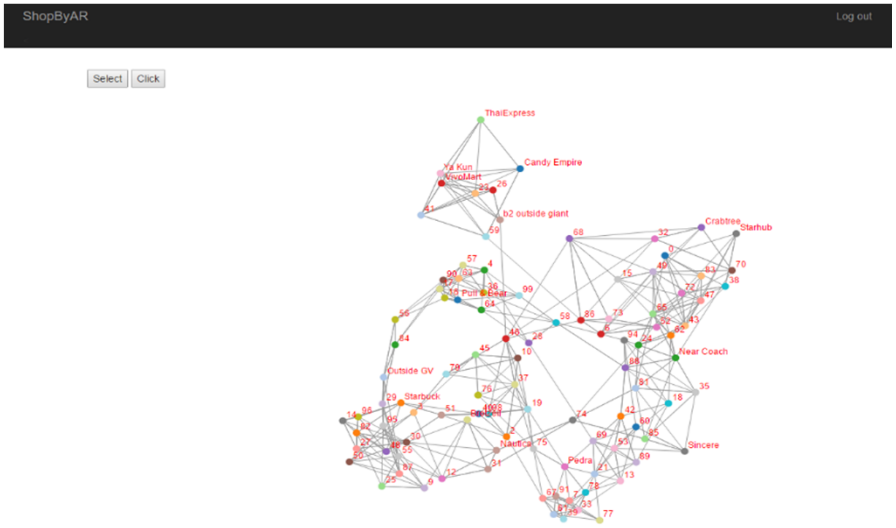


Visual Feature Clustering

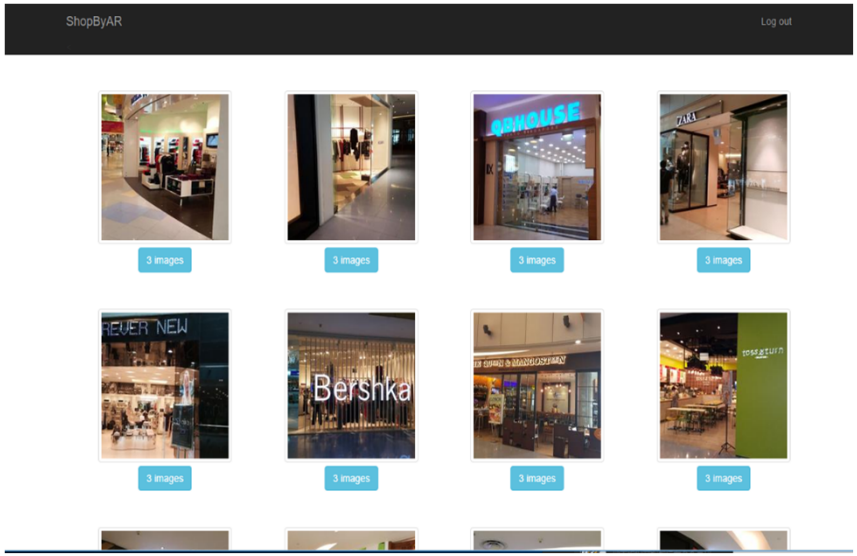
Website interface



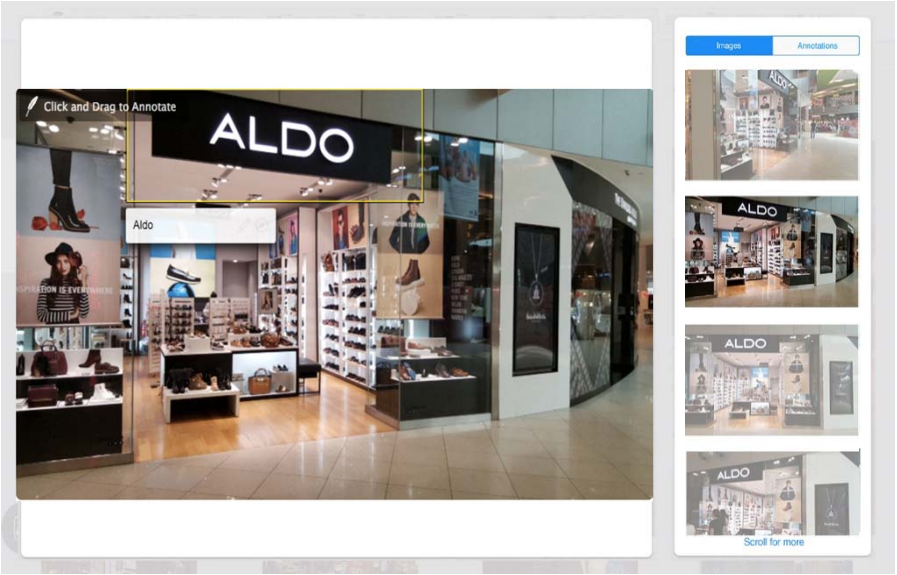
Create a shop by uploading a zip file



WIFI clusters view

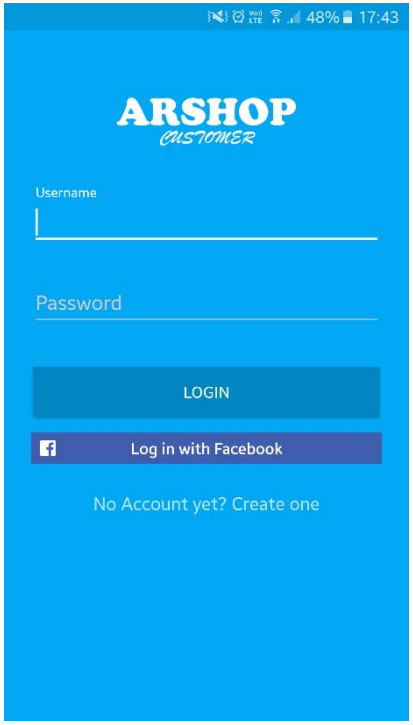


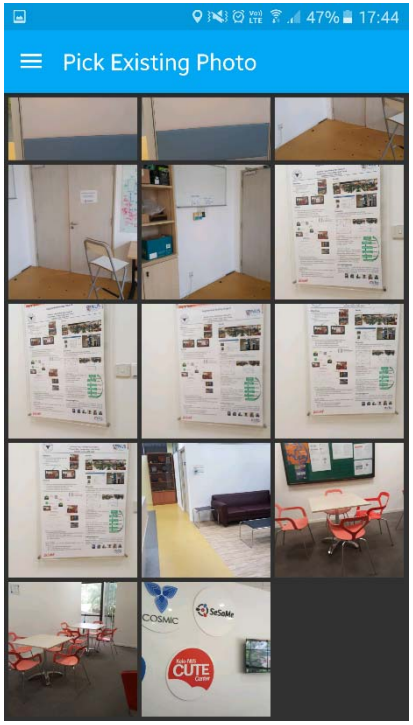
Visual feature clusters view

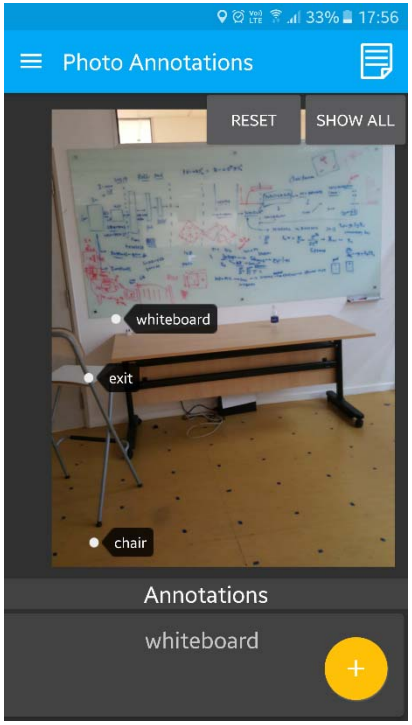


Annotation propagation

Application interface







- Log in by registering or using Facebook account
- Take a new photo or upload existing photos
- Perform a query and retrieve annotations
- Add a new annotation and share with Facebook friends

Experimental Datasets

- Vivocity (1500 images)

- National Gallery of Singapore (1600 images)

Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications

Kampong (village)

- How to inherit the amiable atmosphere from the old kampong?
- How to make the residence in modern HDB feel happy?

Three components of a Kampong system

- Cyber-physical Chatrooms
- Service Chatbots
- Sensor-based Event Detection and Prediction

Unverified Users

Verified Users

Verified, Power Users

A Hierarchical Organization of Cyber-Physical Chat Rooms

HDB Estate

Precinct 1

Blk 1

Void Deck

Level 2

Level 3

Study Corner

Blk 2

Level 2

Level 3

Car Park 1

Precinct 2

Food Center

Basketball Court

Health Corner

Blk 3

Level 2

Level 3

Level 4

Precinct 3

Senior's corner

Community Club

Park

Blk 4

Level 2

Event detection from distributed public/private sensors

Food Center Chat Room

Environment Chatbot

Security Chatbot

Trading Chatbot

Health/Sport Chatbot

Kindness Chatbot

Traffic Chatbot

Moderation and Posting

Virtual and Physical Bulletin for **Group Interaction**

One to One interaction

Cyber Presence (Single location)

Physical Presence

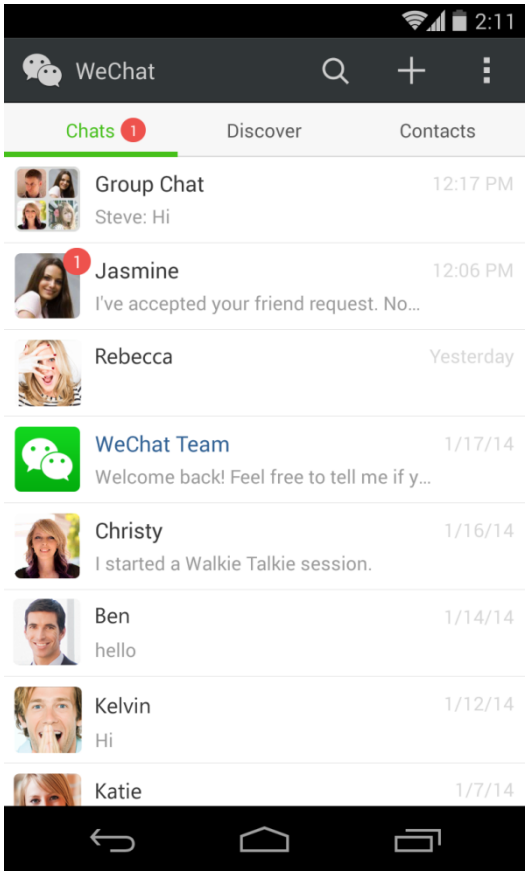
Cyber-physical Chatrooms

Each **Location of Interest (LOI)** in the HDB estate is represented as a chatroom (eg. a food center).

Residents who are physically in a LOI (detected through GPS and wireless signals) is said to have a **physical presence** in the LOI while **cyber presence** can be achieved by logging into a chatroom.

Residents who have cyber/physical presence in a LOI can see each other in the associated chatroom. They can choose to chat with each other on a one to one basis or even add each other as friends so that they can keep in contact after leaving the LOI.

Group interaction will be supported by sending messages to chatbots in the chatroom (to be discussed later) which will perform the role of spam/scam filtering before posting the message in a public bulletin board within the chatroom. Such a bulletin board can be virtual, observable on computing devices or physical, observable from community smart screens that are installed in selected LOIs.



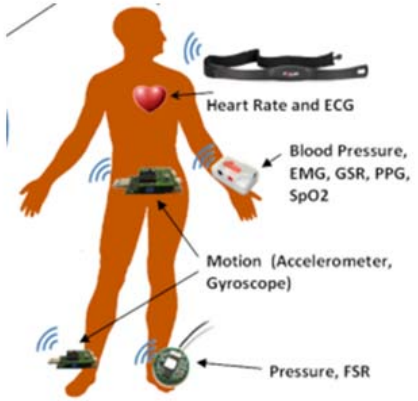
Chatbots

For each chatroom, we will deploy AI chatbots which will interact with residents based on the context of the LOI that is associated with the chatroom. These chatbots play multiple roles in the chatroom including guide, middle man for service providers/consumers and moderators for bulletin board. Different chatbots will be in charge of different services in the chatroom. Example of these chatbots includes:

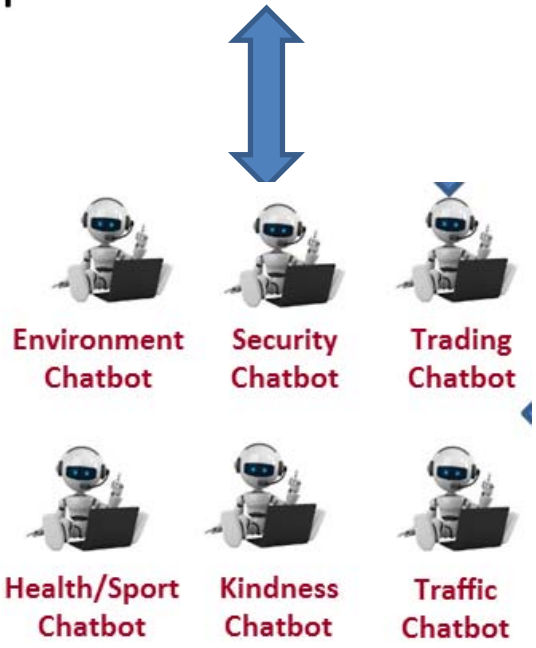
- *Environment chatbot,*
- *Security chatbot,*
- *Trading chatbot*
- *Health/sport chatbot*
- *Kindness chatbot*
- *Traffic chatbot*

Residents can also be friends with these chatbots who will send relevant information about the LOI without the need for the residents to have presence in the LOI.

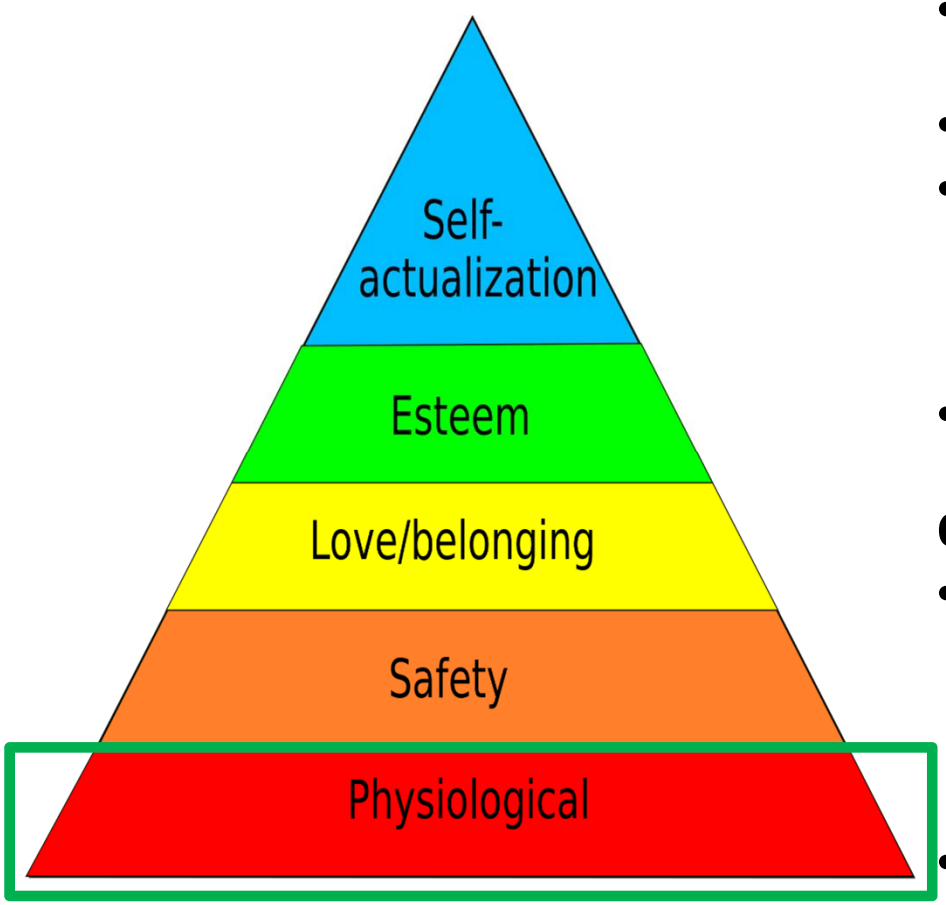
Sensor-based Event Detection and Prediction



Event detection from distributed public/private sensors



Happiness in Maslow's hierarchy of needs



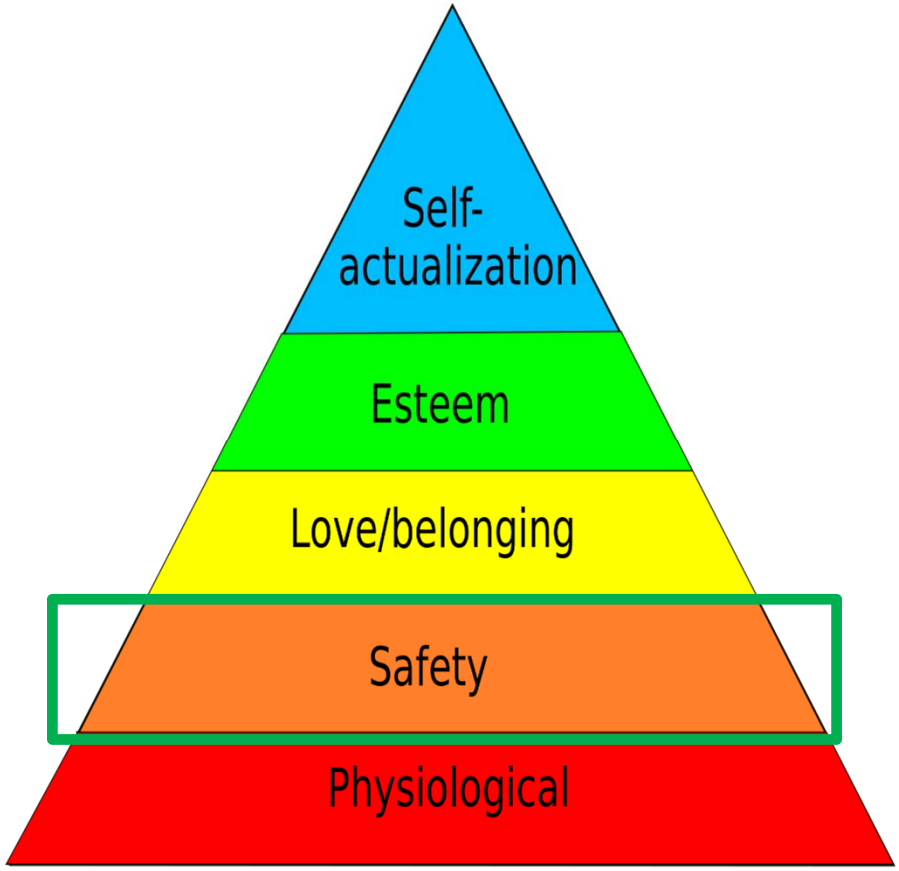
Time

- Happiness=Free Time to do what you want
- Free Time= Total time-Time Wasted
- Time Wasted on
  - Searching (for items, services, experts)
  - Queuing (traffic, grocery store, hawker center, shopping malls, clinics)
- Urban Kampong: Ensure minimum time spend on searching and queuing

Opportunities

- Happiness= Opportunities to provide/obtain chargeable/free services(eg. Tuition, babysitting) or goods (eg. Garage sales) to/from others. New jobs creation.
- Urban Kampong: Facilitate efficient, effective and reliable interactions.

Happiness in Maslow's hierarchy of needs



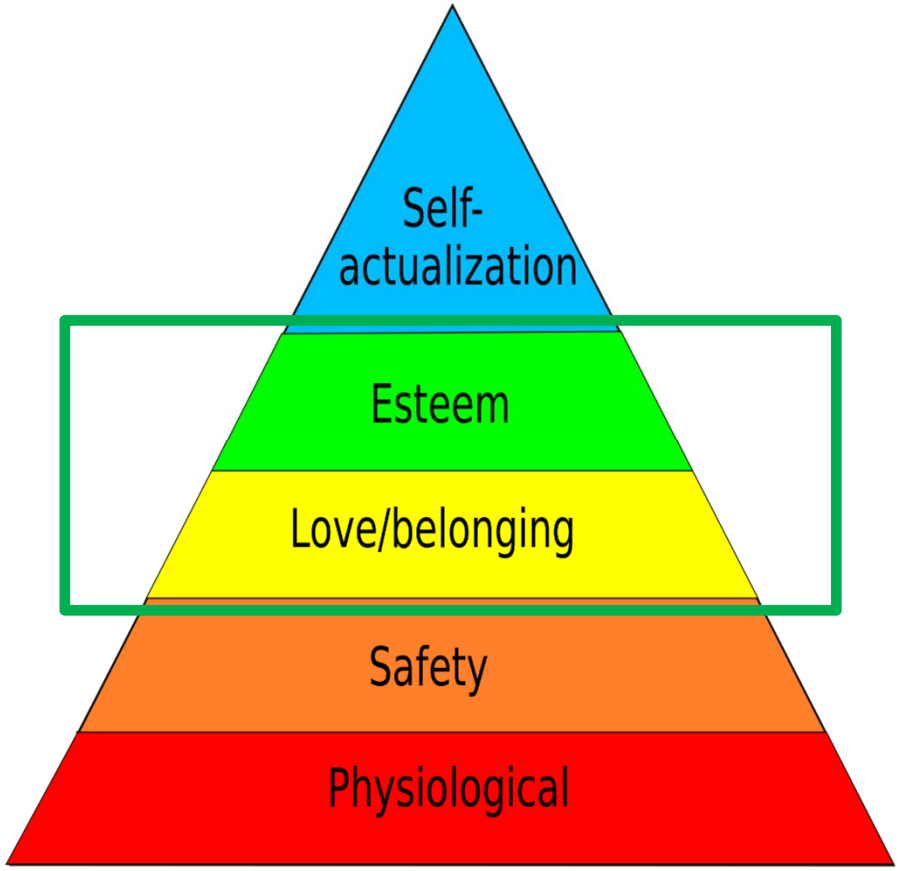
Security

- Happiness=Feeling secure
- Urban Kampong:
  - Cyber-physical neighborhood watch group
  - Intelligent safe path proposal and auto lighting control
  - Intelligent detection of anonymous faces in neighborhood(aggregated)

Health

- Happiness= Minimum health hazards
- Urban Kampong
  - Monitor and report pollution
  - Monitor and report infectious diseases
  - Promote sports etc.

Happiness in Maslow's hierarchy of needs



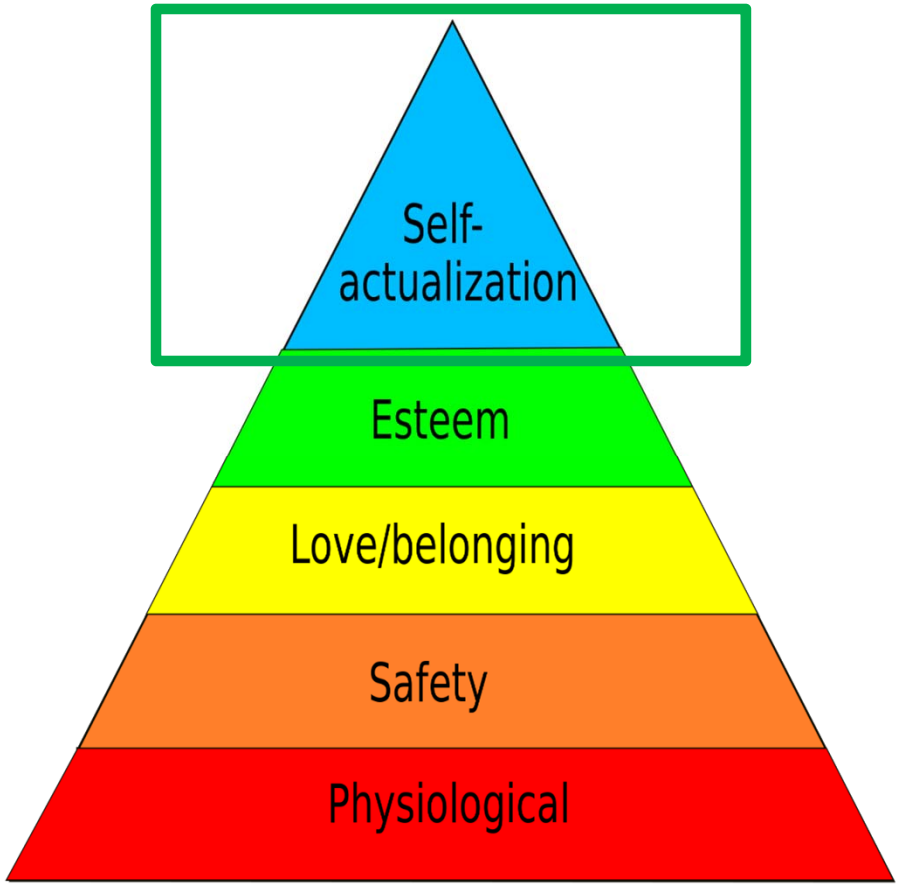
Social

- Happiness= Ability to connect and socialize with people in the neighborhood
- Urban Kampong:
  - Augmented Reality interaction
  - Spatial-Temporal Aware Messaging

Customized Environment

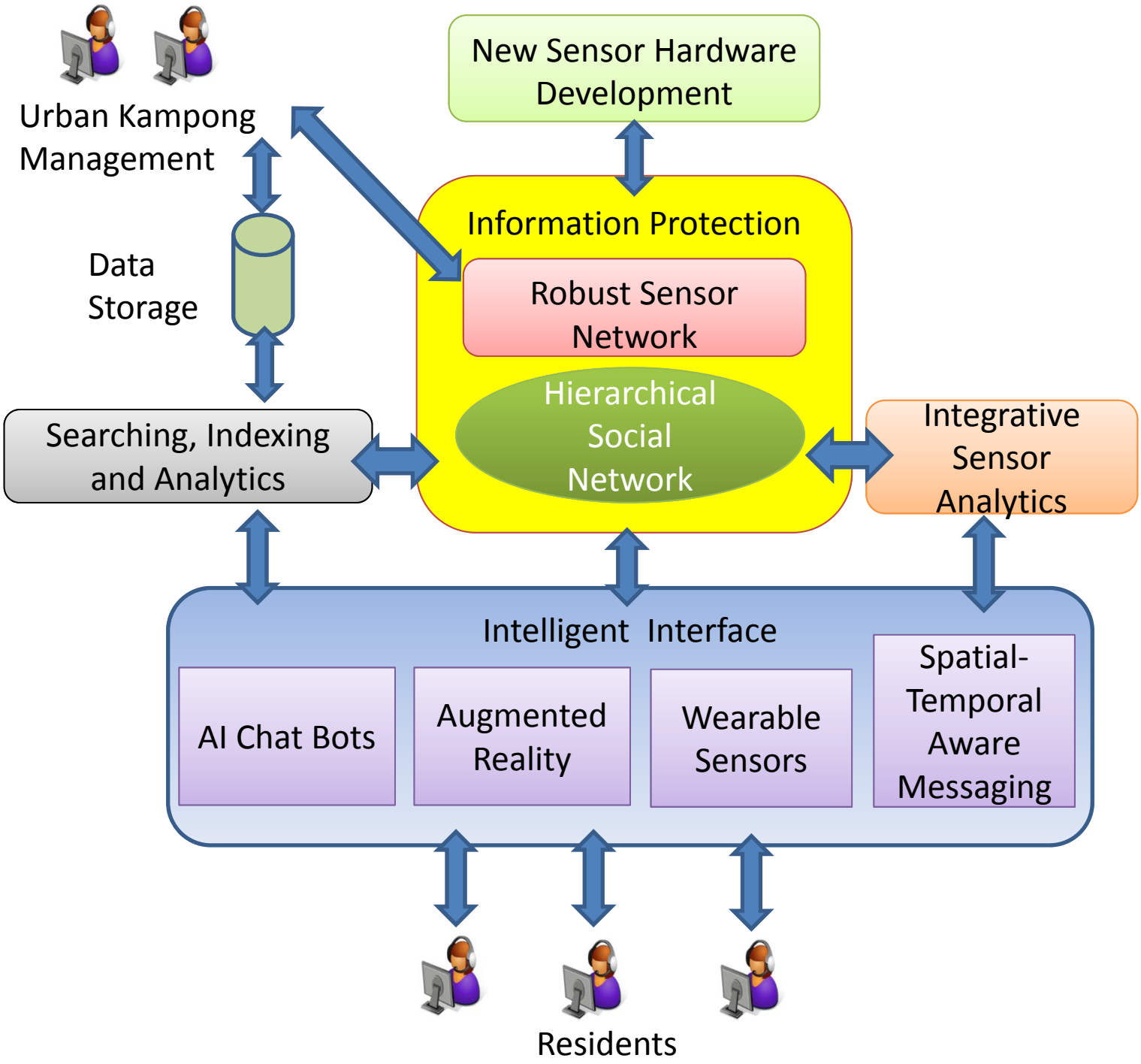
- Happiness= Feeling special and that people care
- Urban Kampong
  - Places with memory and intelligence: Playground, fitness corners, sport facilities, car parks that remember your preference and performance

Happiness in Maslow's hierarchy of needs



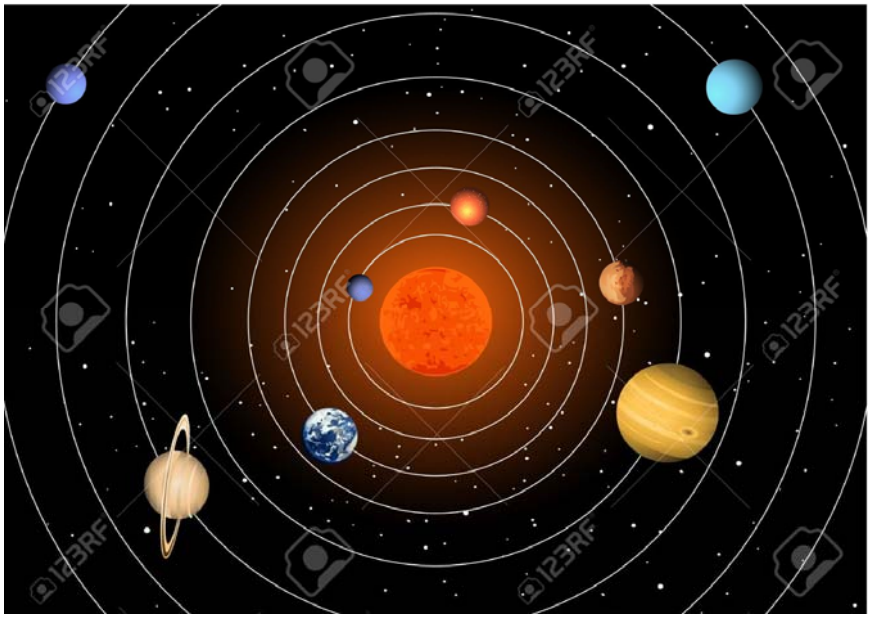
Quantified Self-Improvement

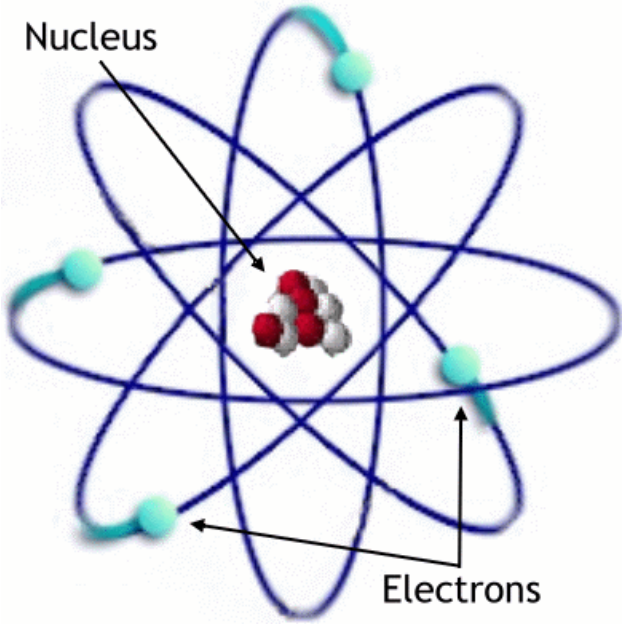
- Happiness= Having targets and seeing progress towards those targets plus helping other towards these targets
- Urban Kampong
  - Enhanced facilities in community clubs, common areas that can measure and quantify improvement in common targets like sports, performing arts, making presentation etc.
  - Support both long term cloud-based coaching and short term ad-hoc coaching


Urban Kampong Management
New Sensor Hardware Development
Information Protection
Data Storage
Robust Sensor Network
Hierarchical Social Network
Searching, Indexing and Analytics
Integrative Sensor Analytics
Intelligent Interface
AI Chat Bots
Augmented Reality
Wearable Sensors
Spatial-Temporal Aware Messaging
Residents

Summary: Collaborative Social Network System

- If you want to get people to integrate and collaborate on Big data, find a center(中) that is stable(庸) and revolve everything around it
- 中庸：不偏之谓中 不易之谓庸
- Model after nature when in doubt!





New trends

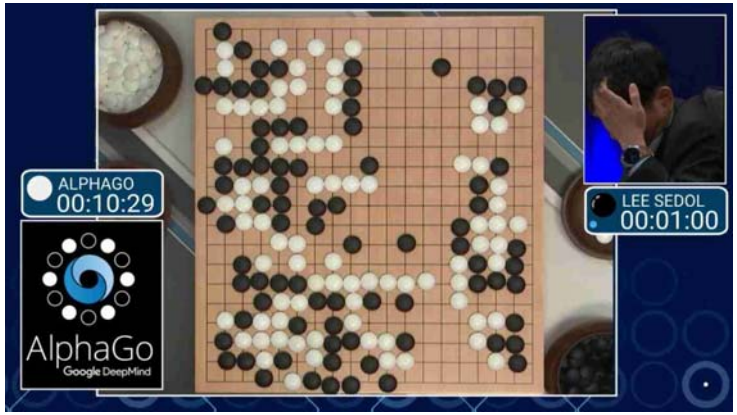
- New hardware = new algorithms
- The right data is better than big data.
- Three phases of data analytics
  - First Phase: Second class citizens, data provided are by-product of other process
  - Second Phase: First class citizens, collect and process data to solve the existing problems
  - Future: To identify and solve previous problems that we have conformed to and forgotten about

Vision of Big Data Analytics and its Impact on Education

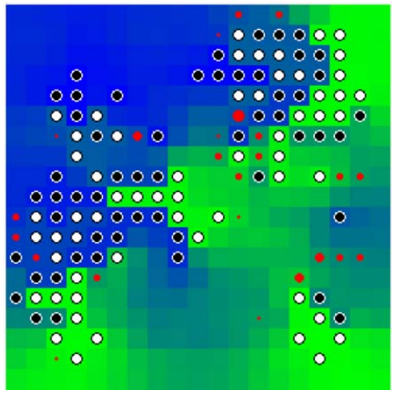
- AI-Assisted Training and Education
- Quantified Self Improvement
- Educational Credential over Distributed Ledger

AI-Assisted Training and Education

- Use AI to identify mistakes in students and propose remedy actions
- Provide analytics tools for human to understand why AI programs behave in a certain way and improve human's performance as a result



Surprising Moves



Analytics Tools

Train

Quantified Self Improvement

"There was once a scholar, Zhao, who put an empty bottle and two plates of beans on his table; one is a plate of white beans and one is a plate of black beans. Whenever Zhao had a good thought, he will put a white bean in the bottle and whenever he had an evil thought, he will put a black bean in the bottle. At the end of the day, Zhao will count the beans in the bottle to determine whether he had more good or evil thoughts. At first, the number of black beans is more than the number of white beans. But as days progress, Zhao became more aware of his thoughts and the number of white beans became more than the number of black beans. Finally, Zhao went on to forget about what are good and evil thoughts but remain a morally upright man even without counting beans to remind himself."----<<The Words of Zhu Zhi Chapter 129>>

赵概投豆

Being able to quantify your improvements can provide very good motivation for consistent effort

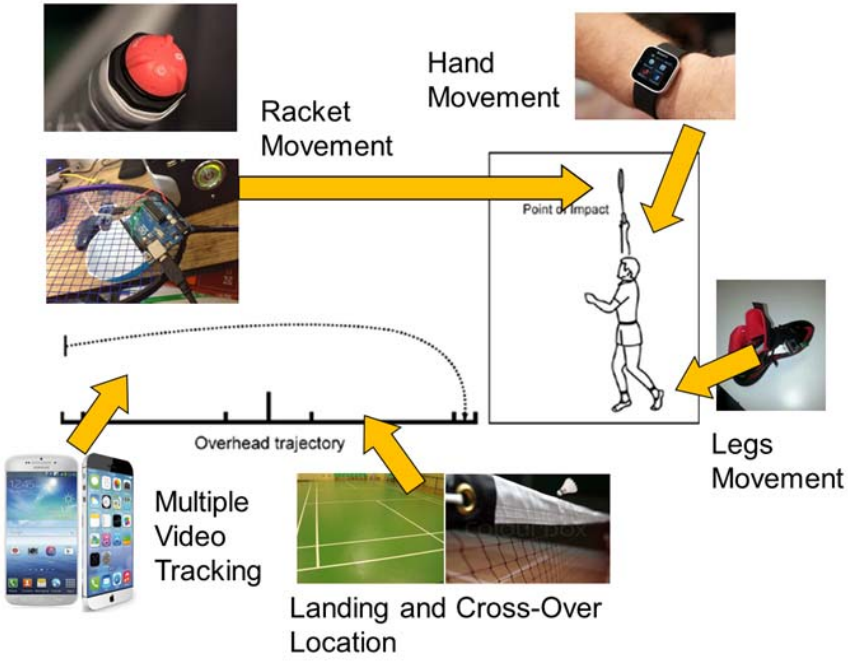
Quantified Self-Improvement in Sports and Performance Art

	Human Body	Equipment	Output(Eg. Ball Movement, Music)
Precision			
Strength			
Speed			
Consistency			
Rhyme			
Coordination			

effect

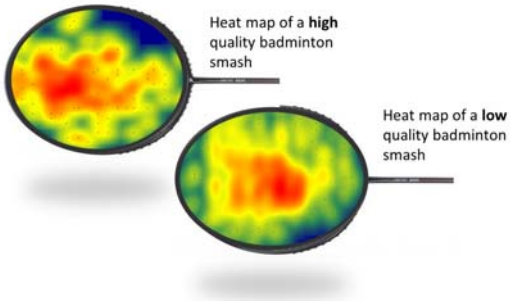
effect

A General Framework for Measurement of Skills. The precision, strength, speed, consistency, rhyme and coordination of human body affect the movement of equipment which in term affect the movement of balls, music generated etc.



Capturing Causality Effect of Body, Equipment and Shuttlecock using Multiple Private/Public Sensors

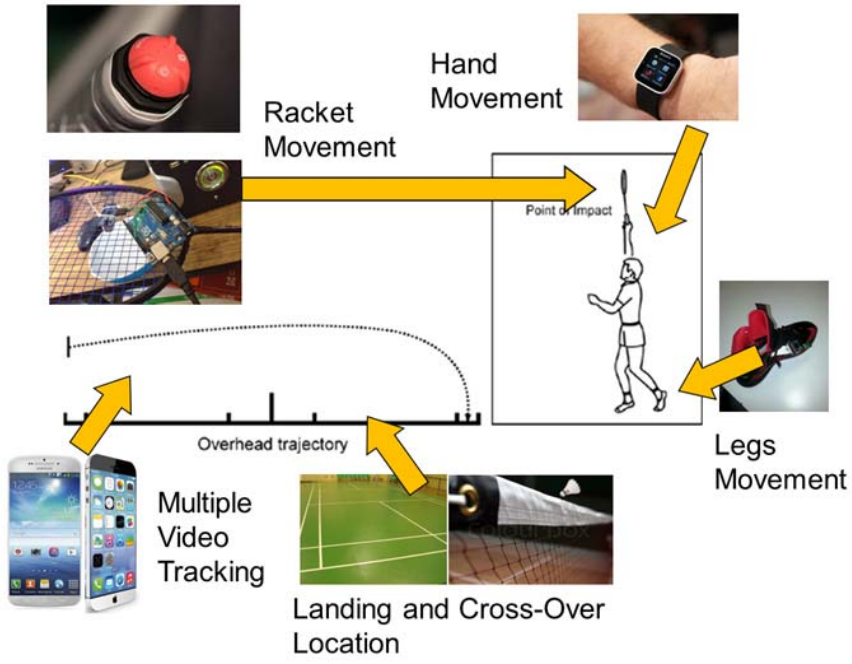
Heat map analysis of high and low quality badminton smashes



Quantified Self-Improvement in Sports and Performance Art(I)

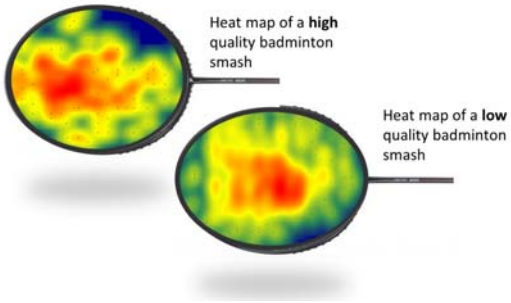
	Human Body	Equipment	Output(Eg. Ball Movement, Music)
Precision			
Strength			
Speed			
Consistency			
Rhyme	effect	effect	
Coordination			

A General Framework for Measurement of Skills. The precision, strength, speed, consistency, rhyme and coordination of human body affect the movement of equipment which in term affect the movement of balls, music generated etc.



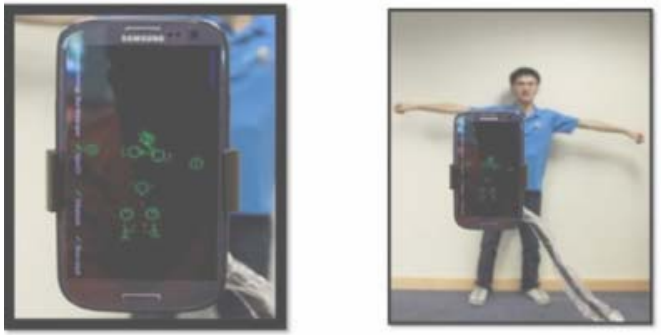
Capturing Causality Effect of Body, Equipment and Shuttlecock using Multiple Private/Public Sensors

Heat map analysis of high and low quality badminton smashes



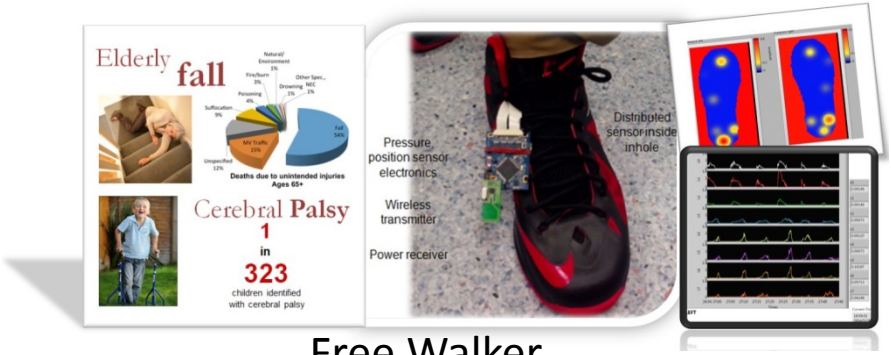
Quantified Self-Improvement in Sports and Performance Art(II)

Multi-view Video Analysis for a Basketball Throw



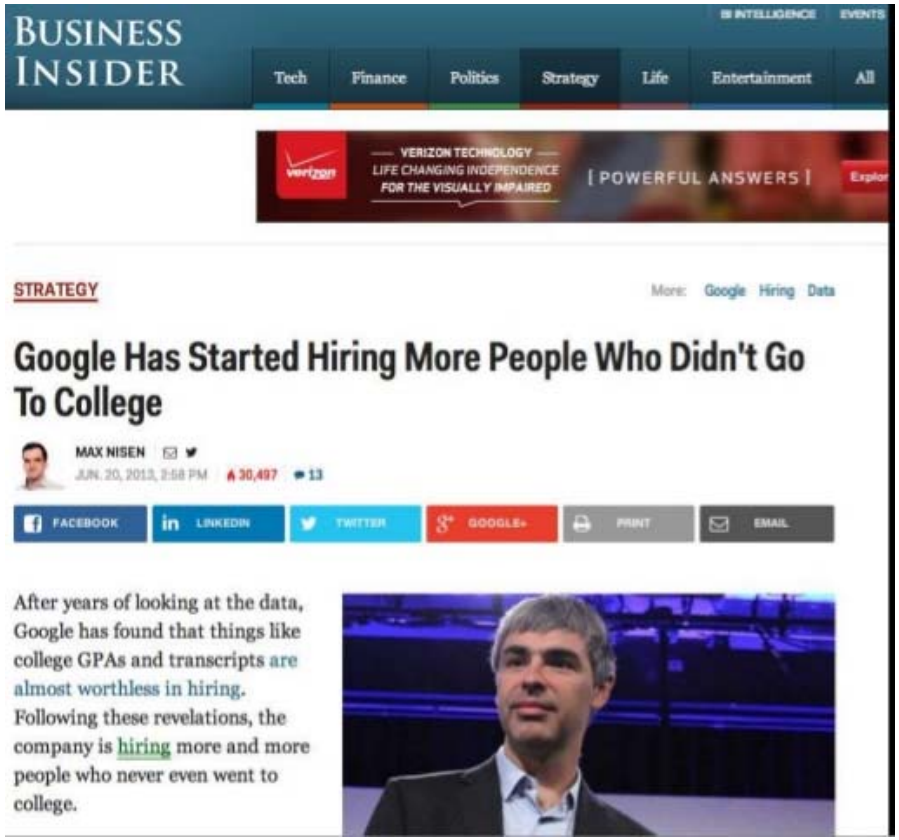
Social Gaming using Multiple Handphones

- Solve fundamental problem of not having enough coaches
- Potential creation of a new industry: *Sport and Performance Art Analytics on the Cloud*. Job creation.



Free Walker

Educational Credential over Distributed Ledger



- Can educational credentials like online book discussion, internships, courses, projects be recorded in a reliable distributed ledger that can be taken in consideration aside for jobs aside from college degree?
- Can recommendation credentials be passed around like bitcoins between students, employers, schools and certification institutes?

Outline

- Big Data: Characteristics and Components
  - Data Generation and Collection
  - Data Storage
  - Database System and Technology
  - Computer Networks
  - Algorithms: Statistics, Machine learning, Data Mining, Visualizations, Optimizations, Simulation
  - Parallel Computing
- Big Data: Types and Applications
  - Relational data, High-dimensional data, Sequences, Trees, Graphs, Mixed data types
  - Logistics, Transportations, Finance, Retail Analytics, Medical, Security, Manufacturing
- New Trends in Big Data
  - Building models on the Fly: Principles and applications
  - Collaborative Social Network System: Collective intelligence over Big Data
    - Readpeer: Building social communities around documents and books
    - ARShop: Augmented reality for shopping
    - Digital Kampong: Building amiable neighborhood community using big data
- Consideration when building big data applications 

Consideration for Building Big Data Applications

1. The Moral Law （一曰道）
   - Need advanced methods to handle big data, or just simpler methods?
2. The Heaven （二曰天）
3. The Earth （三曰地）
   - The same big data application might have different effects under different circumstances.
   - Big data applications must be localized.
4. The Commander （四曰将）
   - Do we have the right people to build big data applications?
5. Method and discipline （五曰法）
   - Who own the data?
   - Who maintain the system after building the applications?
   - What guidelines must be followed when using the data?

Acknowledgement

- Much help provided by my students, RAs and postdocs