CPT-S 415

Big Data

Yinghui Wu

EME B45

Welcome!

Instructor: Yinghui Wu

Office: EME B45

Email: yinghui@eecs.wsu.edu

http://eecs.wsu.edu/~yinghui/

Office hour: Tu/Thu (3PM to 4PM) or by appointment

TA: Sheng Guan

Email: sheng.guan@wsu.edu

Initial survey

Complete the survey in Lecture 1 (Blackboard)

“Big Data Era”

- ✓ 90% of worlds’ data generated over last two years

- ✓ A single jet engine produces *20TB* (10^{12}B) *of data per hour*
- ✓ Facebook has 1.2 billion users, 140 billion links, about 300 PB of data (2015)
- ✓ *Genome of human:* sampling, biochemistry, immunology, imaging, genetic, phenotypic data
  - 1 person: 1PB (10^{15}B)
  - 1000 people: 1EB (10^{18}B)
  - 1 billion people: 1ZB *(10^{24}B)*

Big data is a relative notion: 1TB is already too big for your laptop

But What is big data anyway?

Big data: What is it anyway?

No standard definition!

- **Big data** is the **term** for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

- The **trend** due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

- The **challenges** include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

- This course:
  - Study Big Data models and management,
  - Introduce algorithm design and analysis methodologies, and
  - Tackle Big Data challenges

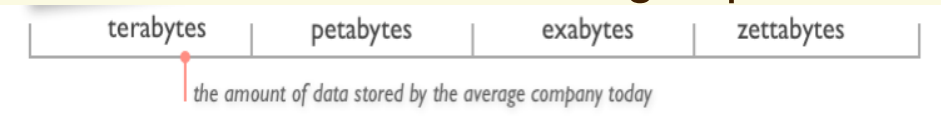
Big data: the 4 V's

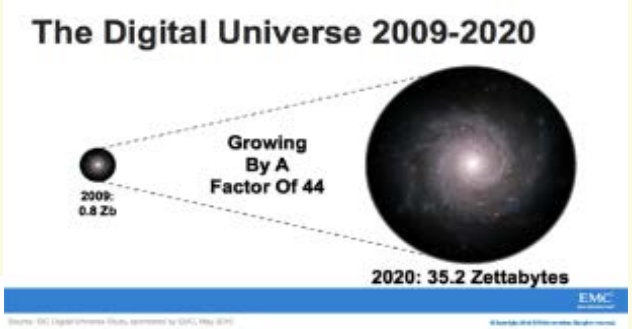
- ✓ *Volume: horrendously large*
  - PB (10^{15}B)
  - *EB (10^{18}B)*
- ✓ *Variety: heterogeneous, semi-structured or unstructured*
  - *9:1* ratio of unstructured data vs. structured data
  - collecting 95% restaurants requires at least 5000 sources
- ✓ *Velocity: dynamic, streams*
  - think of the Web and Facebook, …
- ✓ *Veracity: trust in its quality*
  - *real-life data is typically dirty!*

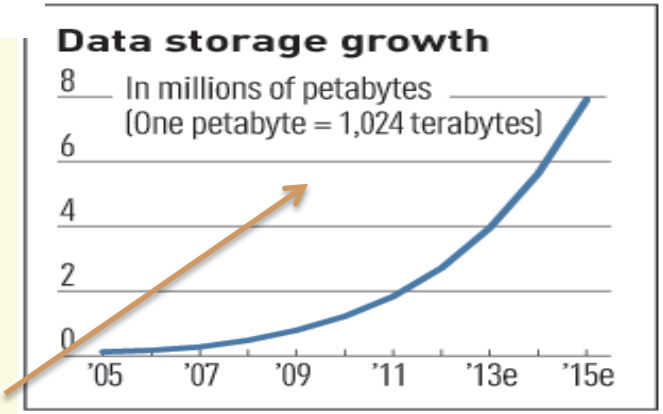
A departure from our familiar data management!

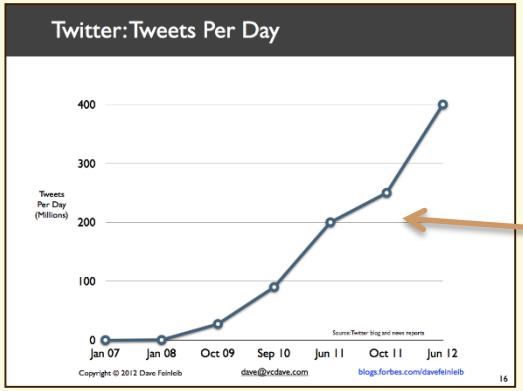
Volume (Scale)

- ✓ **Data Volume**
  - 44x increase from 2009 to 2020
  - From 0.8 zettabytes to 35zb
- ✓ Data volume is increasing exponential







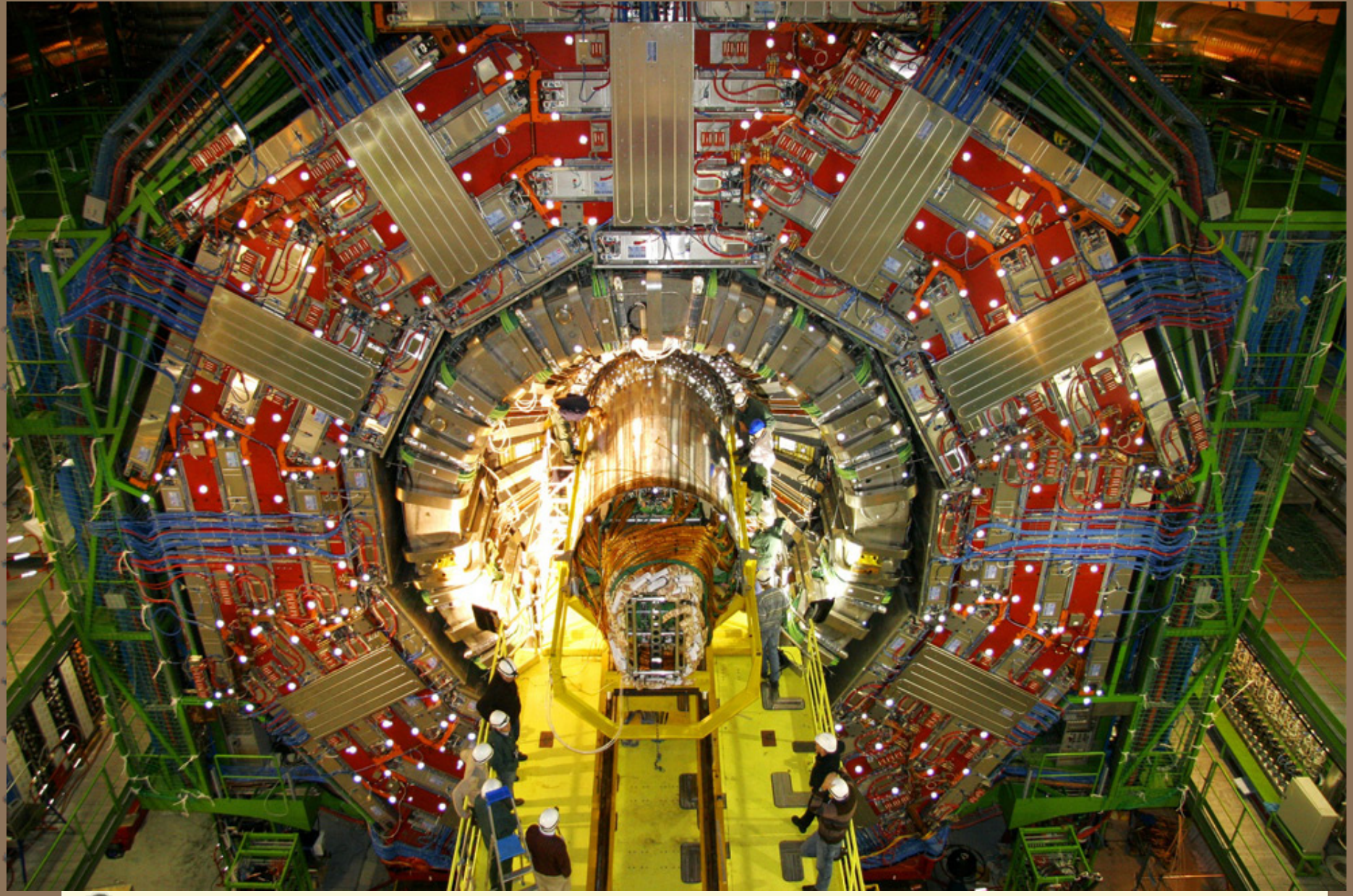


Exponential increase in collected/generated data

The Earthscope

- “The Earthscope is designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.”

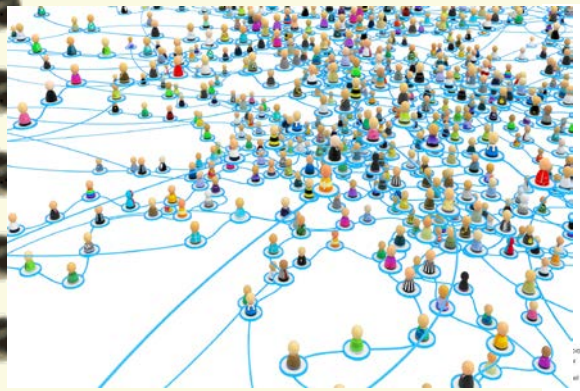
(http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)



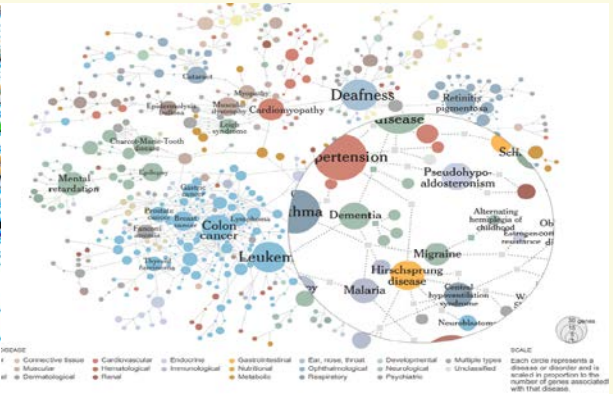
CERN’s Large Hydron Collider (LHC) generates 15 PB a year

Maximilien Brice, © CERN

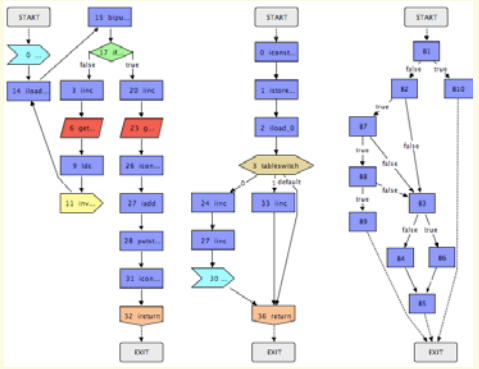
… and no data is an island



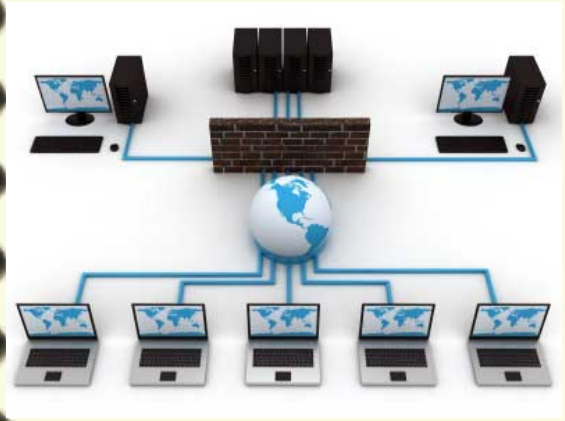
social networks



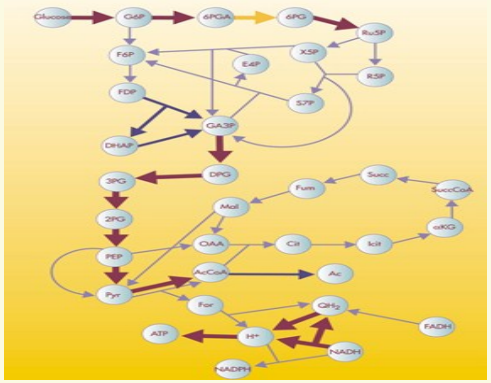
knowledge graph



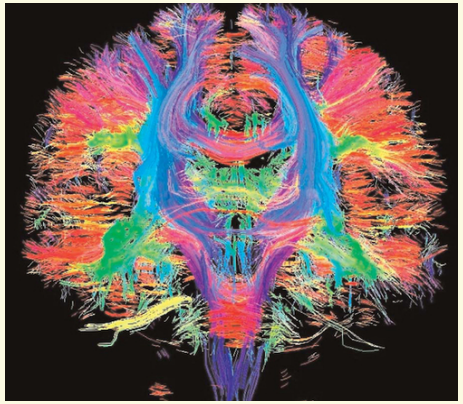
control flow graph



cyber networks



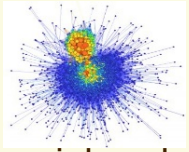
metabolic networks



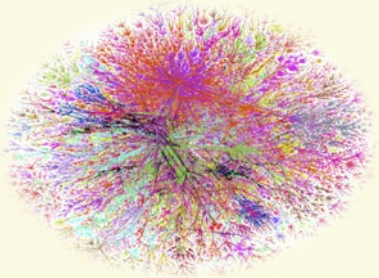
brain network

Real-life scope

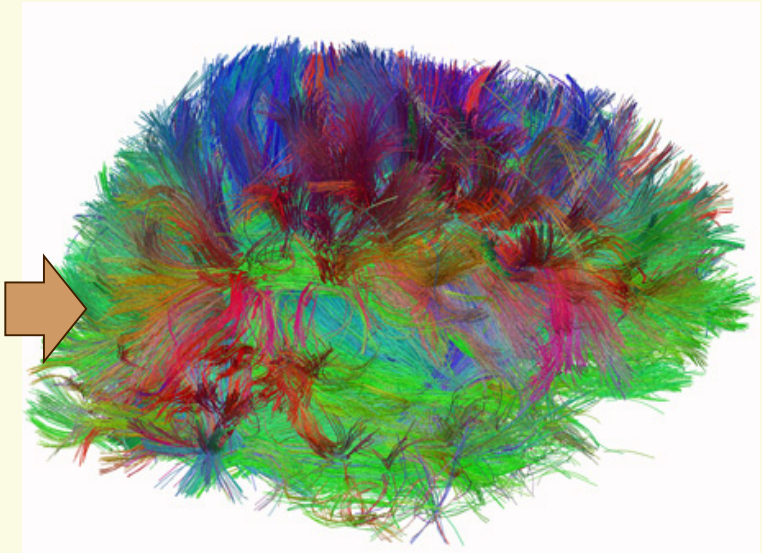
100M(10^8)



social scale
100B (10^{11})



Web scale
1T (10^{12})



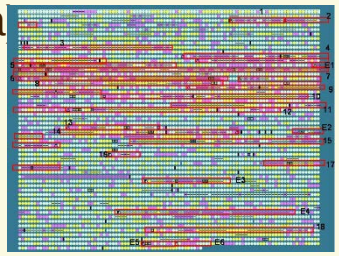
brain scale, 100T (10^{14})
(human brain project)

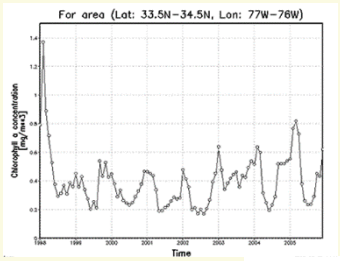
Real-life scope

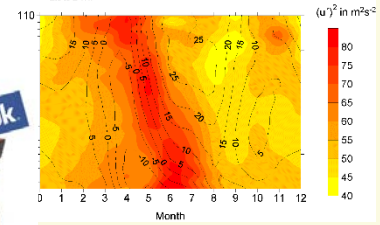
Challenge: Find needle in the haystack?

Variety (Complexity)

- ✓ Relational Data (Tables/Transaction/Legacy Data)
- ✓ Text Data (Web)
- ✓ Semi-structured Data (XML)
- ✓ Graph Data
  - Social Network, Semantic Web (RDF), …
- ✓ Streaming Data
  - You can only scan the data once
- ✓ A single application can be generating/collecting many types of data
- ✓ Big Public Data (online, weather, finance, etc)

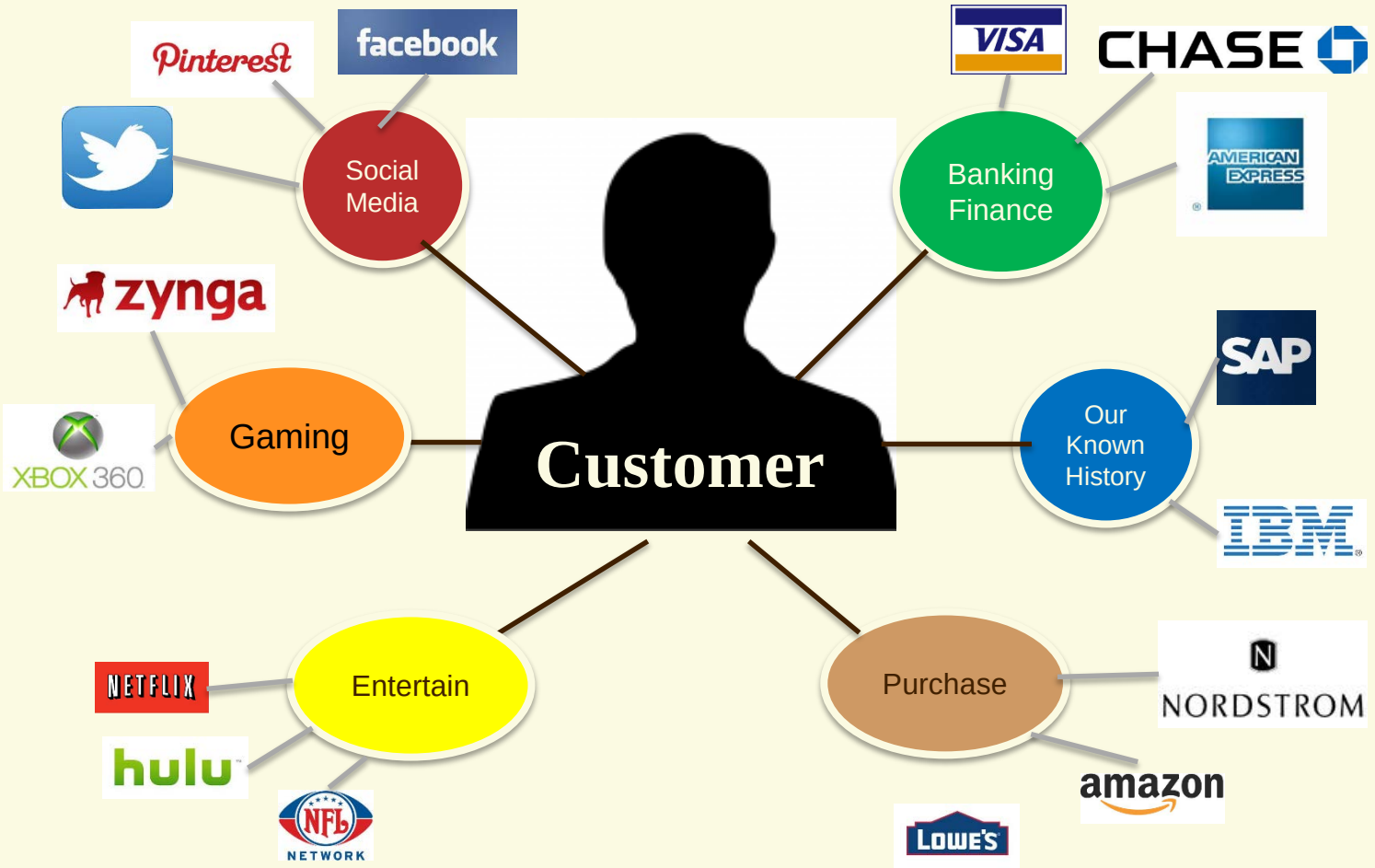




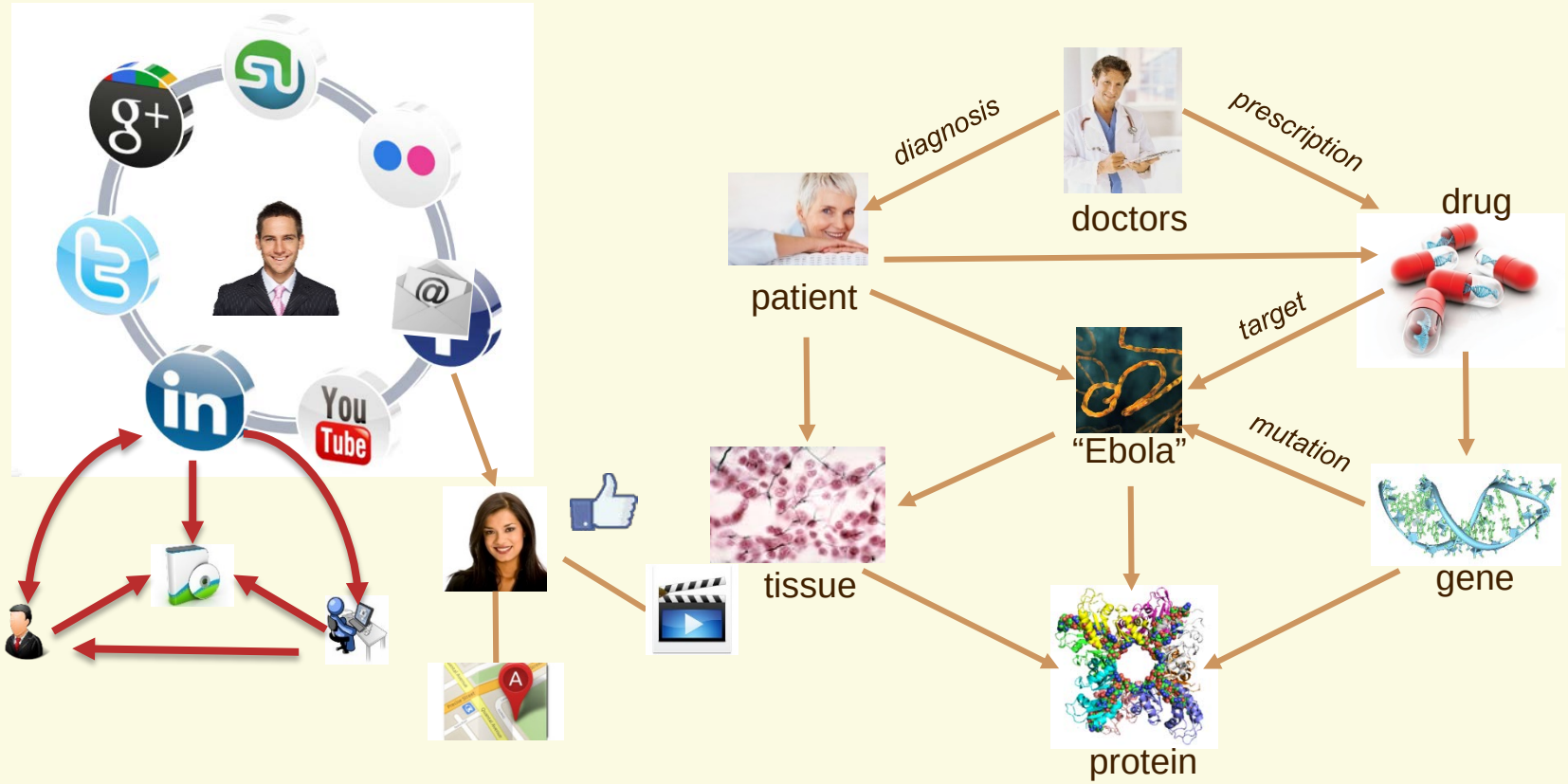


To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



A Global View of Linked Big Data



Challenge: "data wrangling", model and fusion?


IN 60 SECONDS..
1 NEW DEFINITION IS ADDED ON URBAN DICTIONARY
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
Y! THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
QUESTIONS ASKED ON THE INTERNET...
100+
40+
Answers.com
Yahoo!
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
1,500+ BLOG POSTS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
Google
6,600+ NEW flickr
50+ WordPress DOWNLOADS
695,000+ facebook STATUS UPDATES
1,700+ Firefox DOWNLOADS
125+ PLUGIN DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
GO-Globe.com

Velocity (Speed)

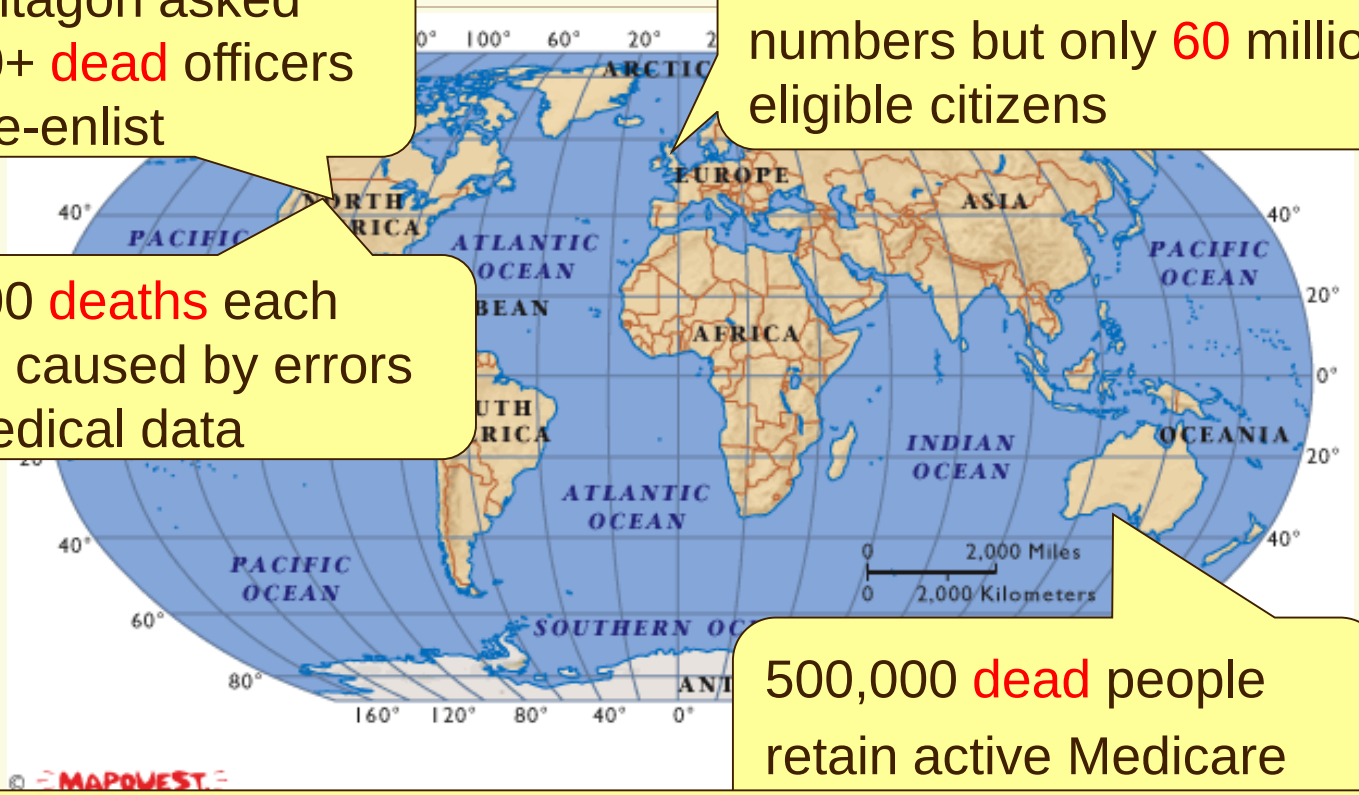
- ✓ Data is begin generated fast and need to be processed fast
- ✓ Online Data Analytics
- ✓ Late decisions ➔ missing opportunities
- ✓ The progress and innovation is no longer hindered by the ability to collect data
- ✓ But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion
- ✓ **Challenge: "Drinking from a firehose"**

Data in real-life is often dirty

Pentagon asked 200+ dead officers to re-enlist

81 million National Insurance numbers but only 60 million eligible citizens

98000 deaths each year, caused by errors in medical data



500,000 dead people retain active Medicare

Data error rates in industry: 30% (Redman, 1998)

Challenge: Dirty data: inconsistent, inaccurate, incomplete, stale

Veracity (quality & trust)

Data = quantity + quality

When we talk about big data, we typically mean its quantity:

- ✓ What capacity of a system provides to cope with the sheer size of the data?
- ✓ Is a query feasible on big data within our available resources?
- ✓ How can we make our queries tractable on big data?
- ✓ . . .

Can we trust the answers to our queries?

- ✓ Dirty data routinely lead to misleading financial reports, strategic business planning decision ⇒ loss of revenue, credibility and customers, disastrous consequences

The study of data quality is as important as data quantity

Dirty data are costly

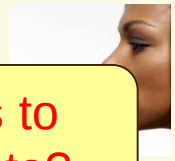
- ✓ Poor data cost US businesses $611 billion annually
- ✓ Erroneously priced data in retail databases cost US customers $2.5 billion each year
- ✓ 1/3 of system development projects were forced to delay or cancel due to poor data quality
- ✓ 30%-80% of the development time and budget for data warehousing are for data cleaning
- ✓ CIA dirty data about WMD in Iraq!

THE DATA WAREHOUSING INSTITUTE

DMReview 2000

PRICEWATERHOUSECOOPERS 2001

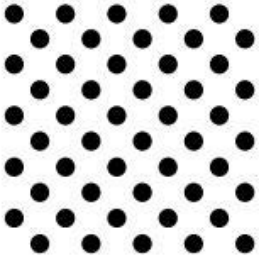
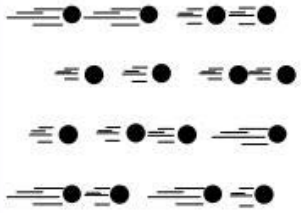
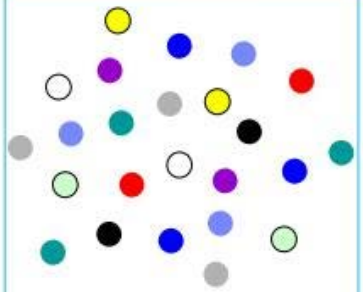
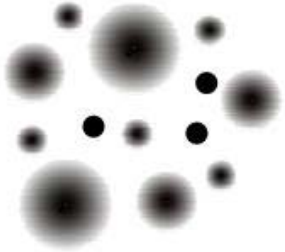
Merrill Lynch 1998



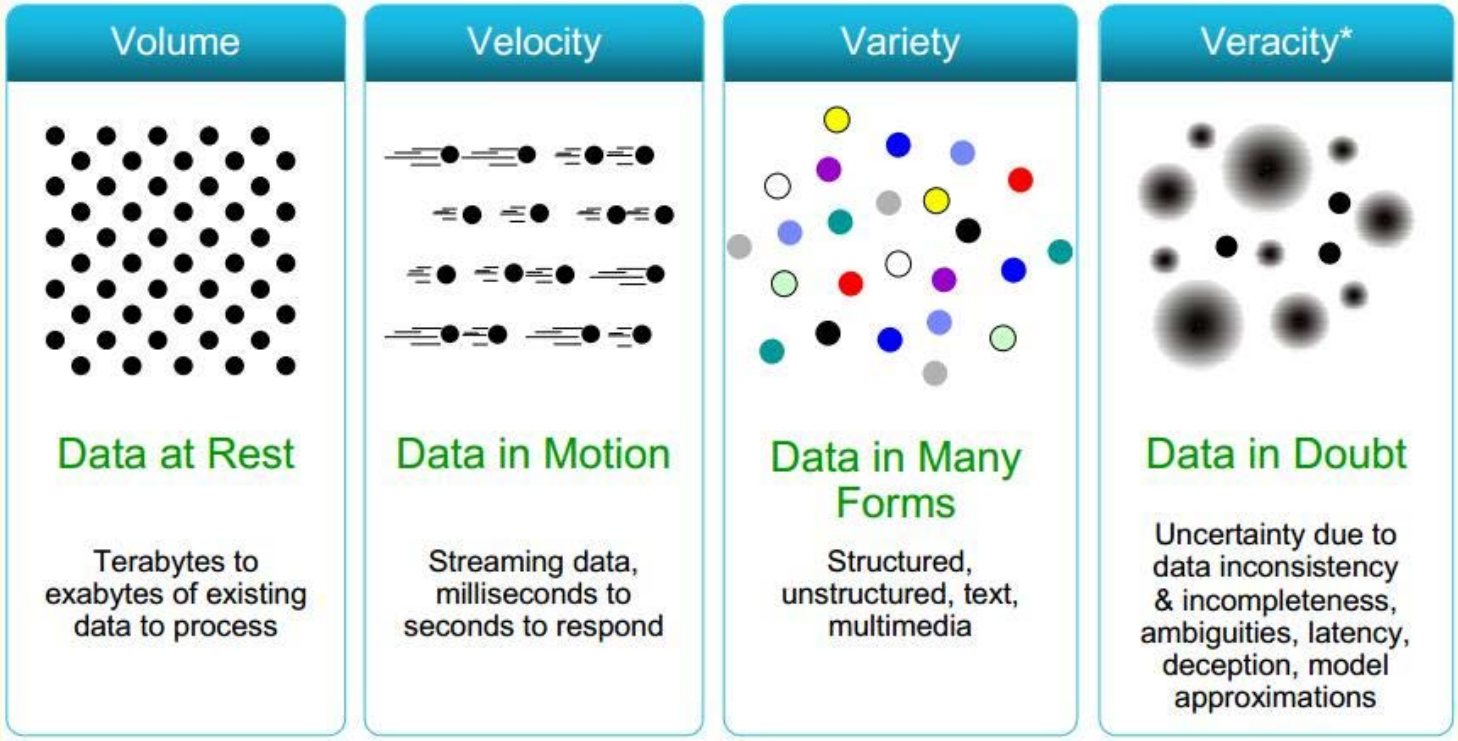
Can we trust answers to our queries in dirty data?

The scale of the data quality problem is far worse on big data!

The 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

The 4V's + n Vs…



Venue (location)
Vocabulary (semantics)
Value

Why do we care about big data?

Big data is needed everywhere

- ✓ Social media marketing:
  - 78% of consumers trust peer (friend, colleague and family member) recommendations – only 14% trust ad
  - if three close friends of person X like items P and W, and if X also likes P, then the chances are that X likes W too
- ✓ Social event monitoring:
  - Prevent terrorist attack
  - transient population
- Scientific research:
  - A new yet more effective way to develop theory, by exploring and discovering correlations of seemingly disconnected factors

The world is becoming data-driven, like it or not…

The big data market is BIG

- ✓ US HEALTH CARE $300 B
  Increase industry value per year by $300 B
- ✓ US RETAIL 60+%
  Increase net margin by 60+%
- ✓ MANUFACTURING –50%
  Decrease development and assembly costs by 50%
- ✓ GLOBAL PERSONAL LOCATION DATA $100 B
  Increase service provider revenue by $100 B
- ✓ EUROPE PUBLIC SECTOR ADMIN 250 B Euro
  Increase industry value per year by 250 B Euro

McKinsey Global Institute

Big Data: The frontier for innovation, competition and productivity

Why study big data?

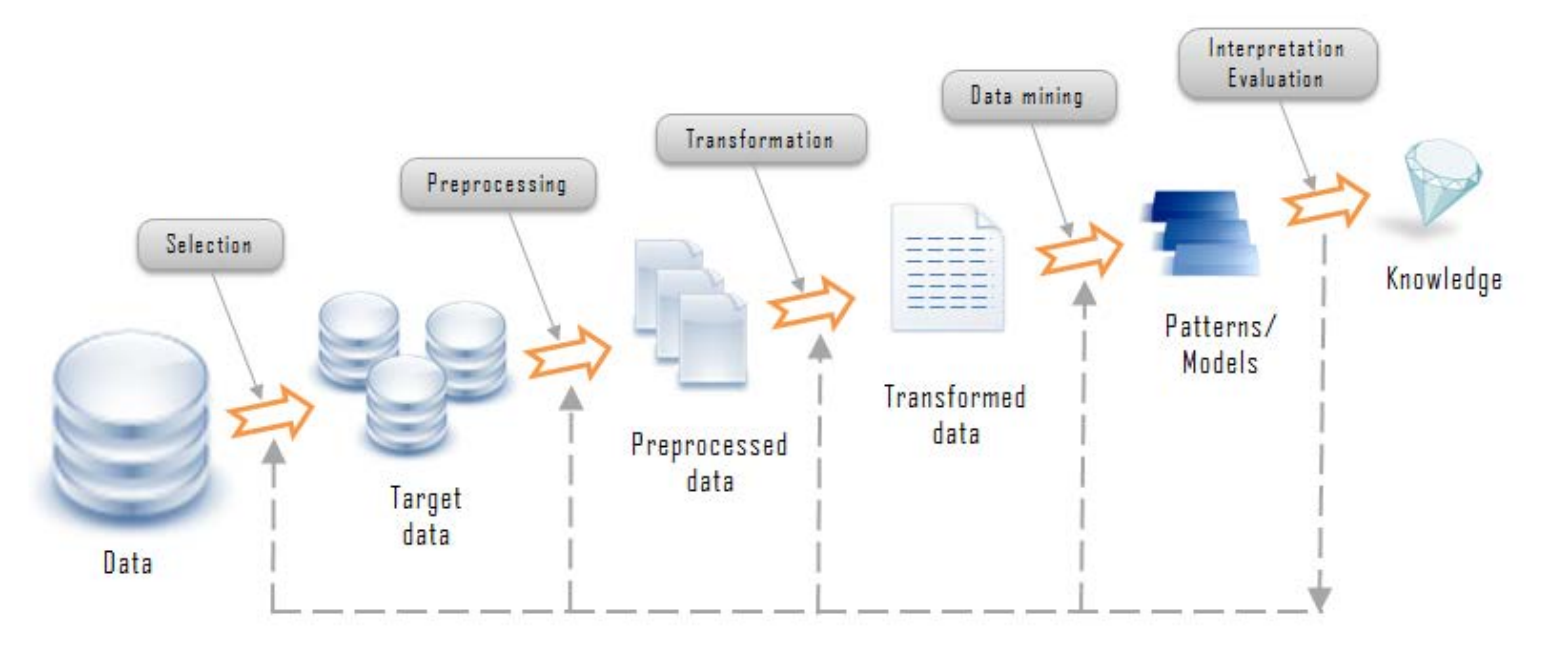
- ✓ Want to find a job?
  - Research and development of big data systems:
    ETL, distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:
    social marketing, healthcare, …
  - Data analysis: to get values out of big data
    discovering and applying patterns, predictive analysis, business inte
- ✓ Prepare you for
  - graduate study: current research and practical issues;
  - the job market: skills/knowledge in need

complexity theory, distributed databases, query answering, algorithms, data quality

Big data = Big $$$

What does this course cover?

A process of knowledge discovery



Data models, storage and management

Topic 1:
Data models, storage and management

Relational data models and DBMS:

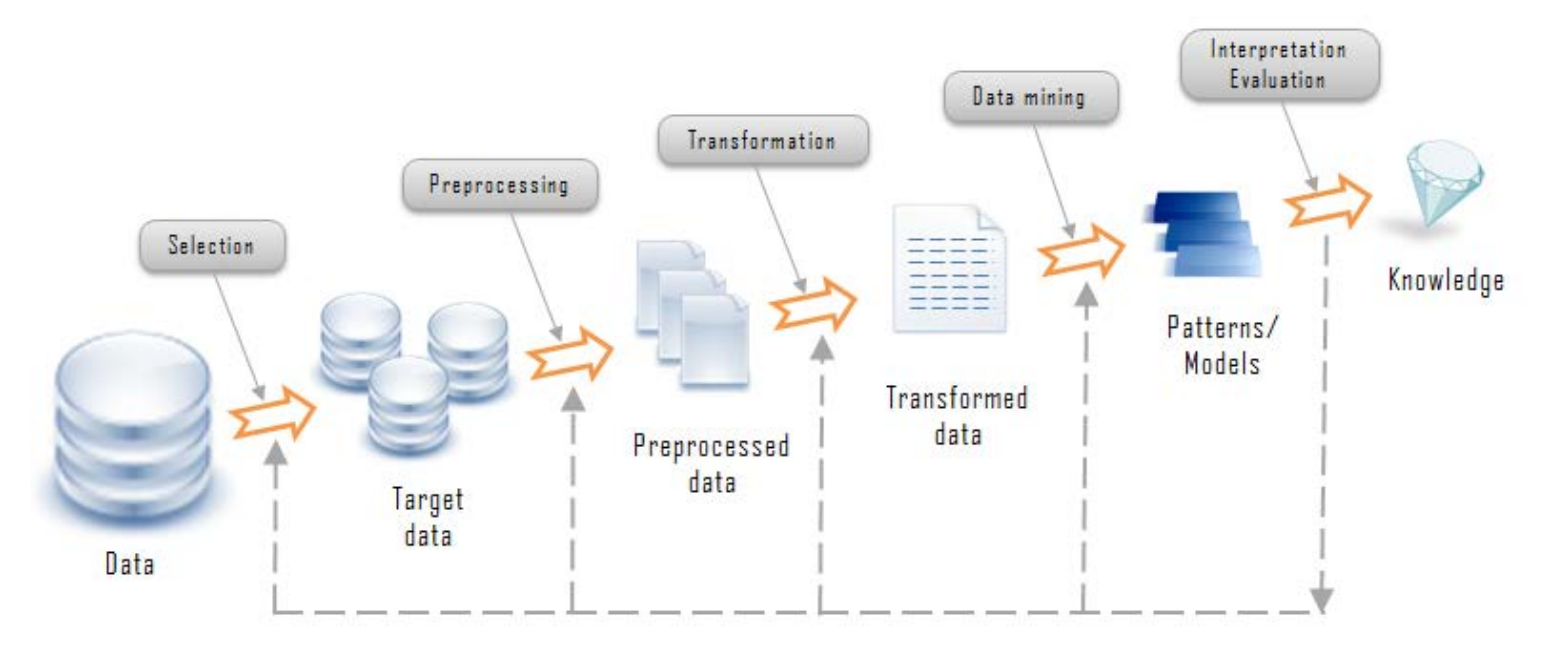
- ✓ Relational data and relation algebra
- ✓ DBMS: centralized; single processor (CPU)
- ✓ Relational databases

Challenge 1: How to store and represent Big data?

Beyond Relational databases

- ✓ Non-relational data, semi-structured data
- ✓ noSQLs, newSQLs, Key-value stores, column-stores, document stores…
- ✓ Graph data and graph databases

A process of knowledge discovery



Data models, storage and Management

Data analytics (search, mining and learning)

Topic 2:
Search Big Data

- ✓ Popular query languages
  - SQL fundamentals
  - XML, XQuery and SPARQL

Challenge 2: How to find needle in the Big Data haystack?

- ✓ Big data search algorithms: Design principles and Case study
  - Indexing and Views
  - Exact vs. Approximate search
  - Compression and summarization
  - Resource bounded search
  - Cope with data streams

Topic 3:
Parallel/Distributed systems

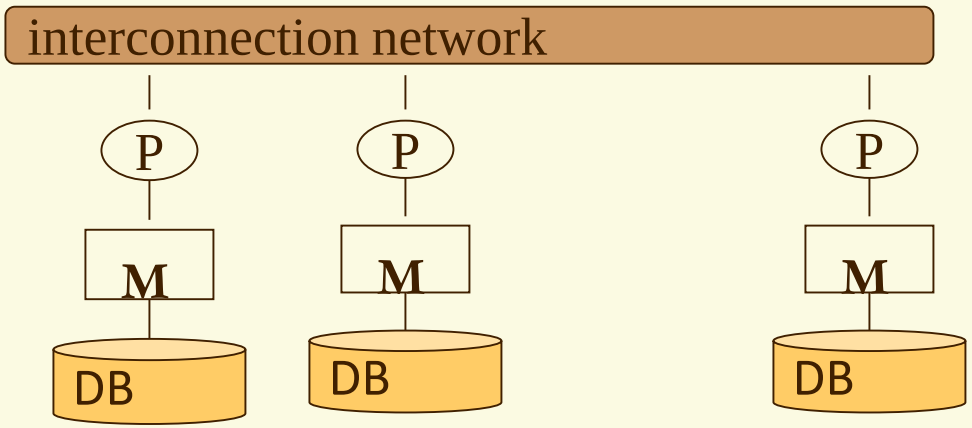
Recall traditional DBMS:

- ✓ Database: “single” memory, disk
- ✓ DBMS: centralized; single processor (CPU);

Can we do better provided with multiple processors?

Parallel DBMS: exploring parallelism

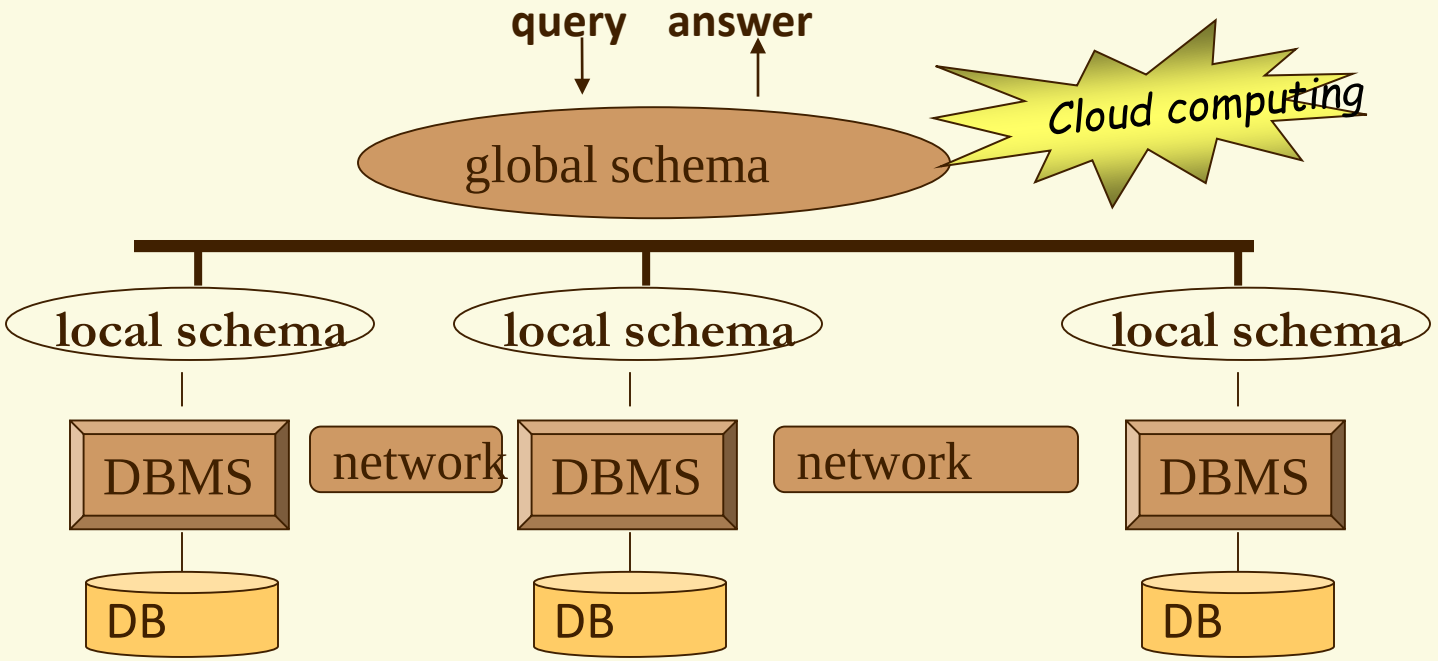
- ✓ Improve performance
- ✓ Reliability and availability



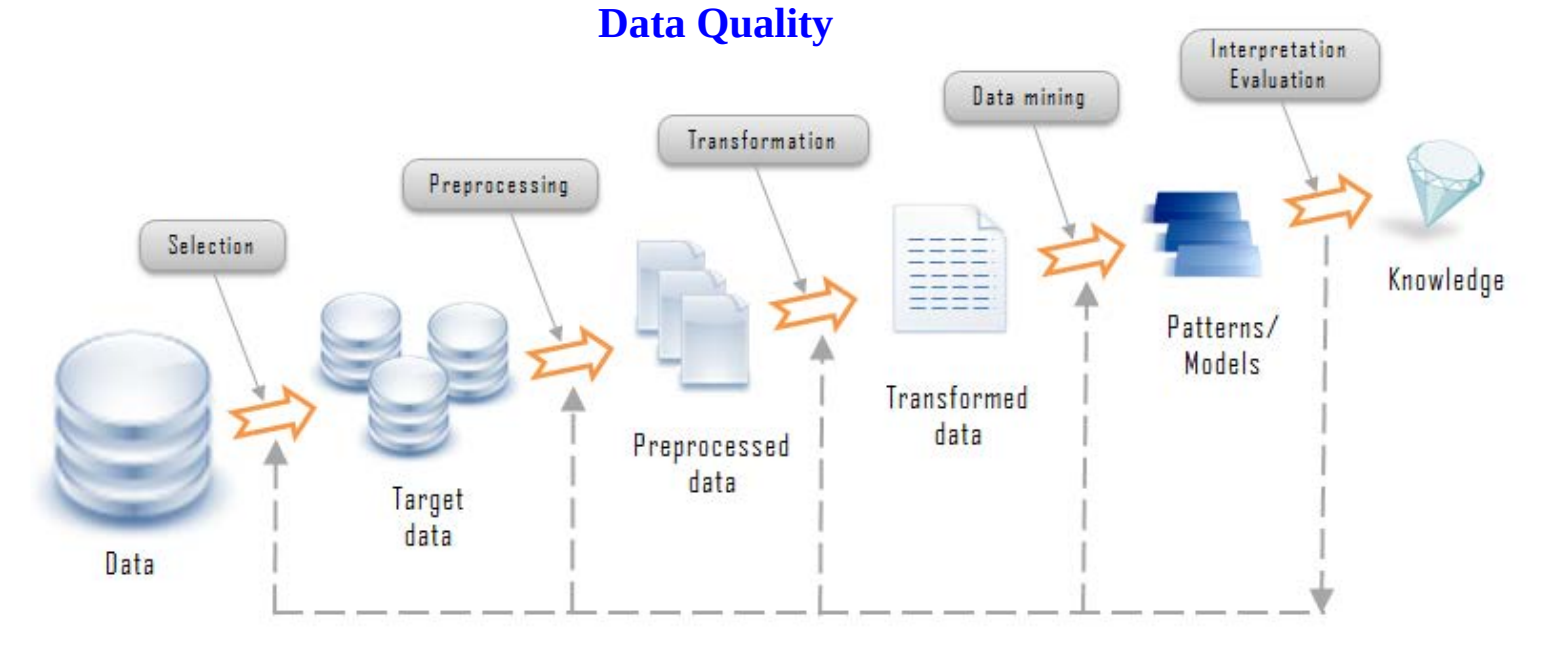
Distributed databases

Data is stored in several sites, each with an independent DBMS

- ✓ Local ownership: physically stored across different sites
- ✓ Increased availability and reliability
- ✓ Performance



A process of knowledge discovery



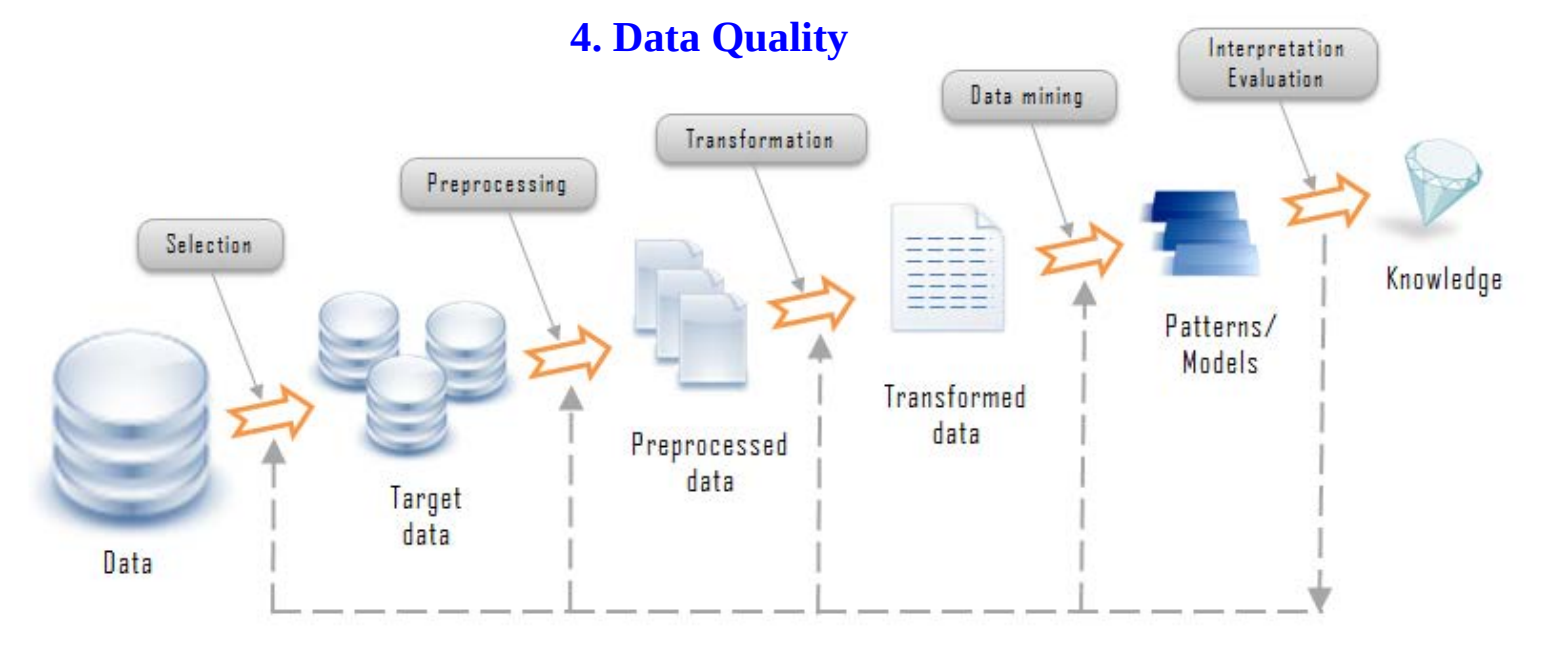
Data models, storage and Management

Data analytics (search, mining and learning)

Topic 4 & 5: data quality, security and ethics

- ✓ Data quality: Cleaning big data: error detection, data repairing, certain fixes (veracity)
- ✓ Privacy and Security (veracity)
- ✓ Data visualization

Putting together (CPTS 415)



1. Data models, storage and Management --- CPTS 451

2. Data analytics (search, mining and learning) - CPTS 315, 570, 575

3. Distributed/Parallel data analysis – CPTS 411

5. Privacy & ethics

Course format

Course information

- This course is not
  - a programming tool or programming language course
  - an independent database or data mining course
  - Plenty of online tutorial for Big data tools!

✓ This course is
  - to provide design principles for Big data challenge
  - an overview of state-of-the-art big data techniques, tools, and principles of Big data solutions
  - provides pointers to Big Data research projects, papers, source code, commercial, open source projects

✓ This course is unique in
  - a complete overview of major big data techniques
  - algorithm design techniques for Big Data
  - academic & industrial practice

Course format

- ✓ A Seminar-style course: *there will be no exam*!
  - Lectures: background.
  - 6 Homework
  - 1 Final course project
- Suggested Textbook:

  Database Systems: The Complete Book (2nd Edition)

  https://www.amazon.com/Database-Systems-Complete-Book-2nd/dp/0131873253
- References:
  - Hadoop: The Definitive Guide, Tom White, O'Reilly
  - Hadoop In Action, Chuck Lam, Manning
  - Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer (*www.umiacs.umd.edu/~jimmylin/**MapReduce**-book-final.pdf*)
  - Data Mining: Concepts and Techniques, Third Edition, by Jiawei Han et al.
- ✓ Online Tutorials and Papers
  - Research papers or chapters related to the topics (3-4 each)
    - At the end of lecture notes from Ln3
  - Check out the "resource" on the course homepage (keep updating)

Grading

- ✓ Class participation: 10%
- ✓ Homework: 40%
- ✓ Project: 40%
- ✓ Final Project report and presentation: 10%

Homework: Six sets of homework, starting from week 3; deadlines:

- 11:59 pm, Thursday, Sep 13, week 4
- 11:59 pm, Thursday, Sep 27, week 6
- 11:59 pm, Thursday, Oct 11, week 8
- 11:59 pm, Thursday, Oct 25, week 10
- 11:59 pm, Thursday, Nov 15, week 12
- 11:59 pm, Thursday, Dec 16, week 14

– See course website for grading policy

Project – Research and development

- Research and development:
  - Topic: pick one from the recommended list, or come up with your own proposal

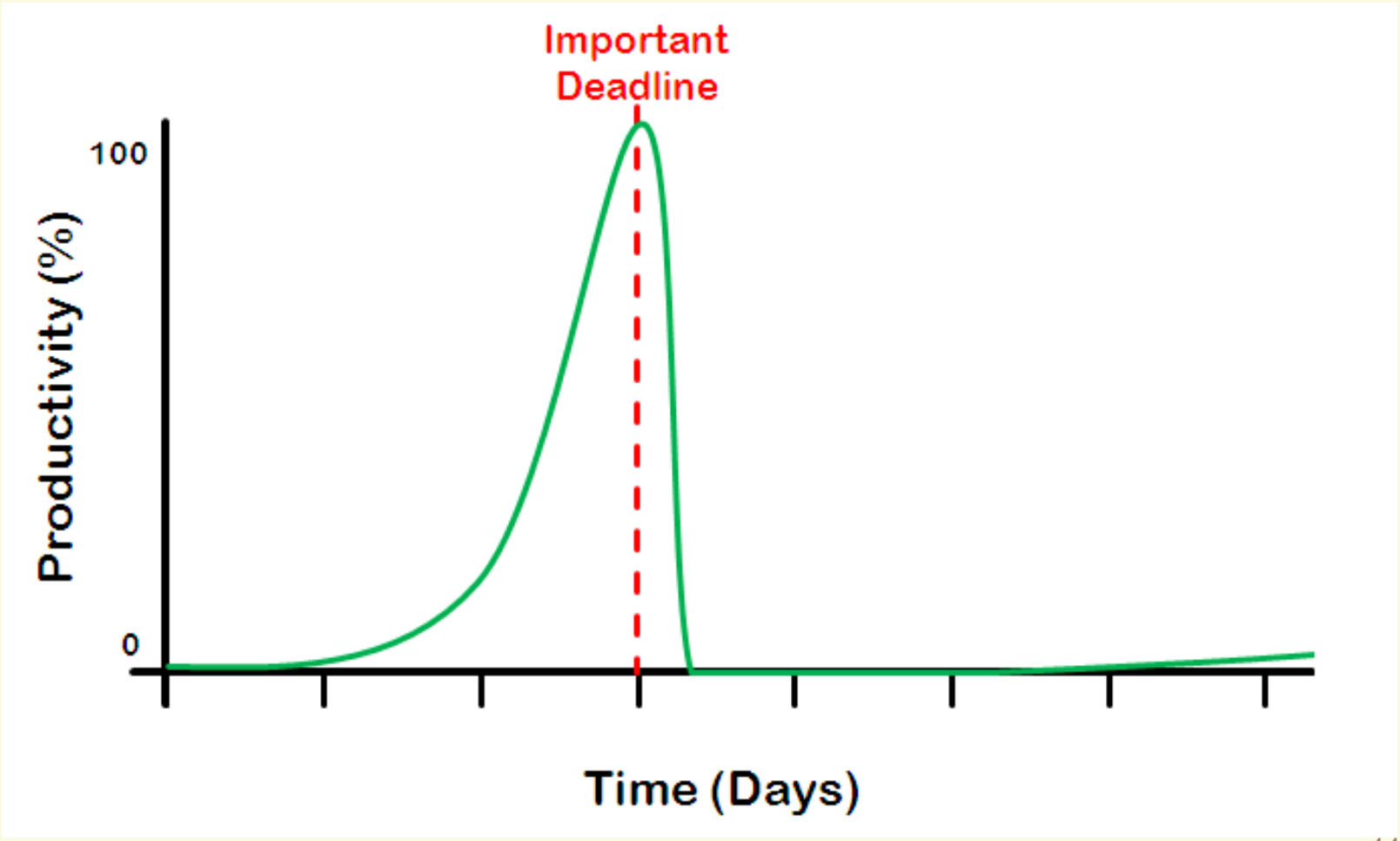
  Example: Airport search engine supported by Hadoop and effective querying algorithms

  You are encouraged to come up with your own project – talk to me first

  (recommended for graduate students) from the reading list given in the lecture notes. Implement main algorithms
  - Conduct experimental study

  Multiple people may work on the same project independently

Start early!


Important
Deadline
100
0
Productivity (%)
Time (Days)

Grading – Project

- Distribution:
  - Algorithms: technical depth, performance guarantees 30%
  - Prove the correctness, complexity analysis and performance guarantees of your algorithms 30%
  - Justification (experimental evaluation or demo) 20%
  - Writing report: 20%

- Report: in the form of technical report/research paper
  - Introduction: problem statement, motivation
  - Related work: survey
  - Techniques; algorithms, illustration via intuitive examples
  - Correctness/complexity/property/proofs
  - Experimental evaluation
  - Possible extensions

Grading - presentation

- ✓ A clear problem statement
- ✓ Motivation and challenges
- ✓ Key ideas, techniques/approaches
- ✓ Key results – what you have got, intuitive examples
- ✓ Findings/recommendations for different applications
- ✓ Demonstration
- ✓ Presentation: question handling (show that you have developed a good understanding of the line of work)

Learn how to present your work

Summary and Review

- ✓ How do we characterize big data?
- ✓ What is the volume of big data? Variety? Velocity? Veracity?
- ✓ Why do we care about big data?
- ✓ Why study Big Data?
- ✓ Fundamental challenges introduced by big data? (topics/projects overview; next lecture)