INFSCI 2725:
Data Analytics
(an introduction)

Philip J. Cwynar MSIS, MBA

Chirayu Wongchokprasitti PhD

University of Pittsburgh
School of Information Sciences
and Intelligent Systems Program

INFSCI 2725 DATA ANALYTICS

Introduction to fundamental technologies dealing with distributed **storage** and efficient **analysis** of **very large amounts of data.**

It is an **overview** of approaches for extracting **information** and **knowledge** from data, verification, testing, and presentation of results.

It is a required course in the "Big Data Analytics" Track of Study

Outline

- GIST “Big Data Analytics” Track
- Introducing each other
- Organization of the course
- Some useful advice
- What is data analytics?
- Contents of the course
- Course outline

GIST "Big Data Analytics" Track

- Faculty
- Prerequisites
- Plan of Study

GIST "Big Data Analytics" track

- The "Big Data Analytics" specialization aims at preparing SIS graduates for careers in the field of "Big Data."
- Provide the essential in-depth knowledge of technologies relevant to big data management.
- Coursework will cover the design and maintenance of infrastructure to efficiently store, easily access, and transfer extremely large amounts of data.
- Education to design, develop and deploy complex information systems and applications that deal with multi-terabyte data sets.

GIST "Big Data Analytics" lead faculty

Lead faculty for the "Big Data Analytics" specialization (listed alphabetically):

Marek J. Druzdzel (decision support, data analytics)
Hassan Karimi (Geographic Information Systems)
Prashant Krishnamurthy (telecommunications)
Vladimir Zadorozhny (databases, wireless sensor networks)




Other key GIST faculty with related interests

Rosta Farzan (social computing)
Stephen C. Hirtle (information visualization, cluster analysis, data mining)
James Joshi (security)
Michael Lewis (human Factors; human-computer interaction; visualization)
Yu-Ru Lin (social and political networks, computational and visualization methods for understanding network data)
Paul Munro (neural information processing, modeling and simulation, models of learning, visualization)
Balaji Palanisamy (Distributed Systems, Location and Data Privacy, Big Data and Cloud Computing)







GIST "Big Data Analytics" prerequisites

Students must have taken **IS 2500 Data Structures** or an equivalent as well as a course in Java programing language prior to entering the "Big Data Analytics" specialization.

This is in addition to the other pre-requisites for the MSIS program (listed on http://www.ischool.pitt.edu/ist/degrees/msis-admissions.php), i.e., one three-credit college course in each of the following:

- **A structured programming language** (Java, Python, C# or C++)
- **Statistics** (data collection, descriptive and inferential statistics)
- **Mathematics** (discrete mathematics or calculus)

http://www.ischool.pitt.edu/ist/degrees/specializations/big-data.php

GIST "Big Data Analytics" plan of study

6 credits in the Mathematical and Formal Foundations area:

Required courses:

- INFSCI 2160: Data Mining
- INFSCI 2591: Algorithm Design

6 credits in the Cognitive Science area:

Recommended courses:

- INFSCI 2410 Introduction to Neural Networks
- INFSCI 2415 Information Visualization
- INFSCI 2430 Social Computing
- INFSCI 2480 Adaptive Information Systems
- INFSCI 2130: Decision Analysis and Decision Support Systems

GIST "Big Data Analytics" plan of study

18 credits in the Systems and Technology area:

Required courses:

- INFSCI 2710: Database Management
- INFSCI 2711: Advanced Topics in Database Management or
  INFSCI 2750 Cloud Computing
- INFSCI 2725: Data Analytics

Recommended courses:

- INFSCI 2150 Security and Privacy
- INFSCI 2711 Advanced Topics in Database Management
- INFSCI 2750 Cloud Computing
- TELCOM 2120 Network Performance
- TELCOM 2321 Computer Networking

GIST "Big Data Analytics" plan of study

6 credits of Electives:

Recommended courses:

- **INFSCI 2000 Introduction to Information Science**
- **INFSCI 2801 Geospatial Information Systems**
- **INFSCI 2802 Mobile GIS and Location-Based Services**
- **INFSCI 2809 Spatial Data Analytics**

The electives can be chosen to meet the individual needs of the student and may include classes in Machine Learning, Advanced Statistics, and domain-specific areas.

GIST "Big Data Analytics" plan of study

Departures from the distribution above are possible (especially if they make sense) but must be requested in advance through a petition to the GIST faculty.

Do not treat this lightly but if your case does not fit the standard requirements, ask for a special treatment.

Introductions

The instructors

Philip J. Cwynar MSIS, MBA

Office: IS 708

Email: TBA

WWW: www.linkedin.com/in/philip-cwynar-msis-mba-ba393b4

- Introducing each other
- Organization of the course
- Some useful advice
- What is decision analysis?
- Contents of the course
- Course outline

The Instructors

Chirayu Wongchokprasitti, PhD

Center for Causal Discovery

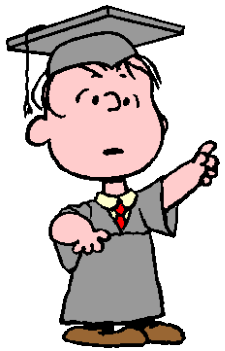
Department of Biomedical Informatics, PITT

Office : IS 708 or BAUM 435F

Email : chw20@pitt.edu

WWW : www.pitt.edu/~chw20/

Introductions -- Who are you?



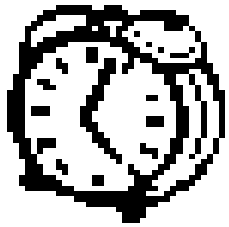
Tell us about yourself:

- What is your name, what do you want to be called?
- What is your educational background (prior studies, current program)?
- What is your professional background (prior and current work experience)?
- What can you do? What are your strengths?

A word of advice: Listen carefully and look for partners for your assignments and term project ☺!

Organization of the Course

Meeting times



Classes (404 IS Building):

Mondays, 6:00-8:50pm

Philip's office hours:

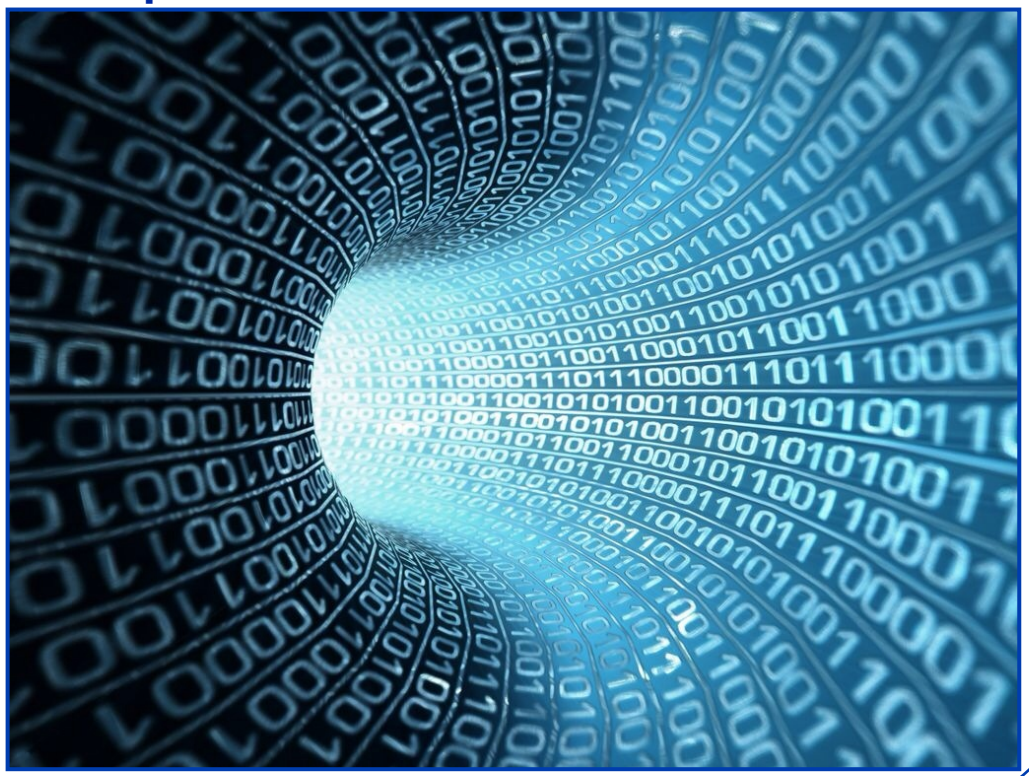
IS 708 by appointment

Chirayu's office hours:

IS 708 or BAUM 437F by appointment

Objective of the course

The primary objective of this course is to make you acquainted with analytical procedures that are useful in processing very large amounts of data. This should make you better prepared for the deluge of data that you will encounter in practical environments.



The textbook

Readings for this course will be taken from several sources, listed in the syllabus.

Additional readings may be assigned in the course of the semester.

Assignments

Nine assignments planned over the course of the semester.

Group work (at most 3 students in each group).

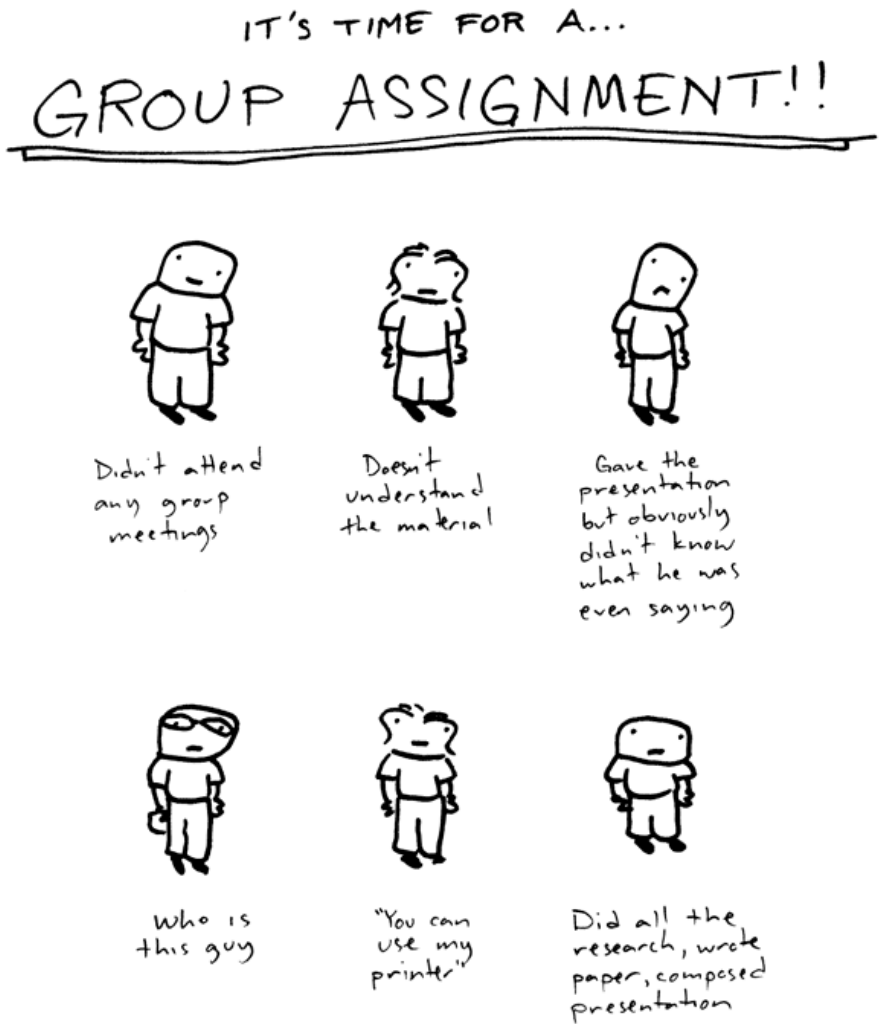
Deadlines are marked on the syllabus.

Will be “recycled” but please do not feel tempted to use past solutions!

This is bad for you and is also explicitly forbidden by the University anti-plagiarism policies.

Group work (assignments and project)

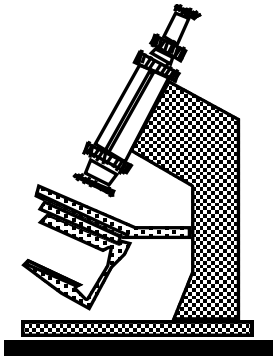
- Group work means generally learning more with a smaller effort.
- Some communication overhead but it is generally worth it.
- Make sure that the groups that you form are not like in this cartoon!
- Small groups (2-3 students).



Term project

- Play with a fairly large (2GB+) data file.
- Team work (2-3 people, do not necessarily have to be the same as for the assignments).
- Develop ways of efficiently storing the data and processing it over the course of the semester.
- Important ultimate performance/accuracy but also computational efficiency.

Exams



There will be one midterm exam and one comprehensive final exam, both closed book.

You can bring with you to the exam one double-sided letter-size sheet of paper with notes.

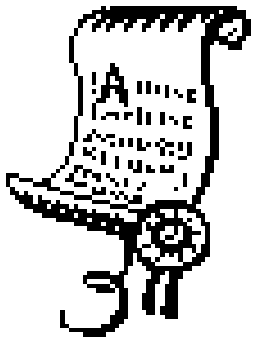
There are no limits on the font size – you can cram as much information on these two pages as you wish – but the notes have to be handwritten personally by you and this is a strict requirement.

Copied or computer-printed sheets are not allowed.

Expected effort (time load)

- Expect to spend about six hours (preferably nine) quality time outside of class for every class meeting.
  - Four hours (and two more) for readings
  - Two hours (and one more) to do the assignments.
- The term project should normally demand between twenty and thirty hours of your time.
- The actual load will vary, of course, depending on your background and preparation.

Grading



Your final grade for the course will be determined as follows:

Assignments : 30%

Term project : 30%

Midterm exam : 20%

Final exam : 20%

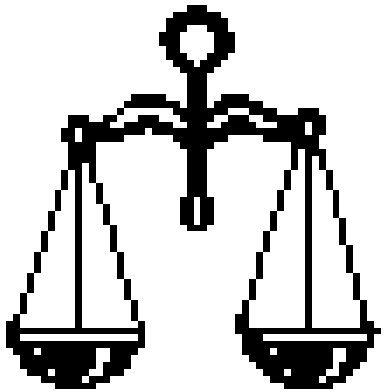
On the top of this all, you can obtain up to 10% of the total score for in-class quizzes and participation.

Useful Advice (Hopefully)

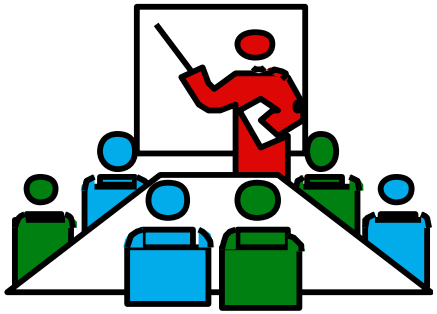
Do you really want to take this course?

Please ask yourself the following questions:

- Do I really want to take this course?
- Is this the right time for me to take this course?
- Do I have enough time to take this course?
- Do I want to take this course with this teacher?



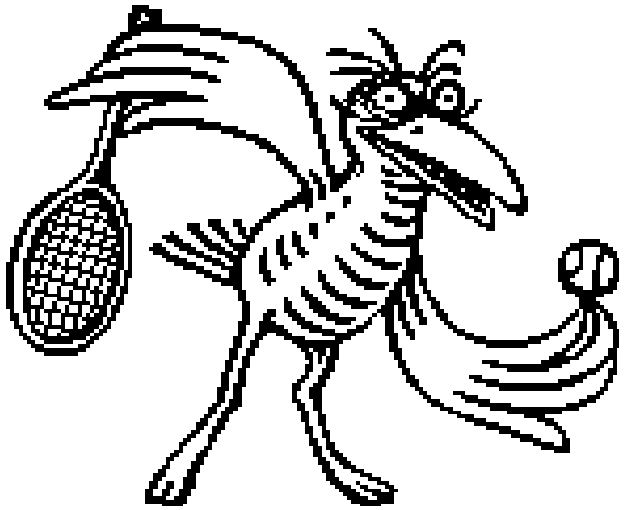
Come to classes …



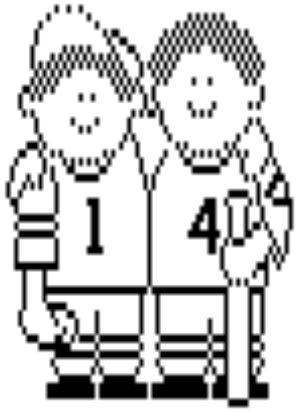
- Class attendance is important in learning.
- Coming to class stimulates timely reading of the material and helps you to be up to date on what is happening in the course.
- Our in-class discussions and exercises will be an important factor in your learning.
- Understanding difficult parts of the material on your own may often cost you a multiple of what it takes in class.

... and be their active participant

- This is the best way to learn
- Do not hesitate to ask questions
- We'll reward your participation



Be good to your classmates



As somebody in a biology lab has once put it:

"if you are a good colleague, you will not need to be afraid that somebody pisses in your cultures when you are not in the lab."

All work in this course is collaborative.

Do the readings before the class

You will be amazed how efficient you will be in your studies!

What is Data Analytics?

From data to knowledge

Analytics

1. **Data**: symbols
2. **Information**: data that is processed to be useful; provides answers to "who", "what", "where", and "when" questions
3. **Knowledge**: application of data and information; answers "how" questions
4. **Understanding**: appreciation of "why"
5. **Wisdom**: evaluated understanding

Ackoff, R. L., *"From Data to Wisdom", Journal of Applied Systems Analysis*, 16:3-9, 1989

From data to knowledge

From wisdom to … ?

1. **Data**: symbols
2. **Information**: data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions
3. **Knowledge**: application of data and information; answers "how" questions
4. **Understanding**: appreciation of "why"
5. **Wisdom**: evaluated understanding

"Wisdom does not make you a good man" – *Confucius*?

"Data is not information, Information is not knowledge, Knowledge is not understanding, Understanding is not wisdom"

– *Cliff Stoll & Gary Schubert*

"Science is organized knowledge. Wisdom is organized life."

– *Immanuel Kant*

What is “Big Data?”

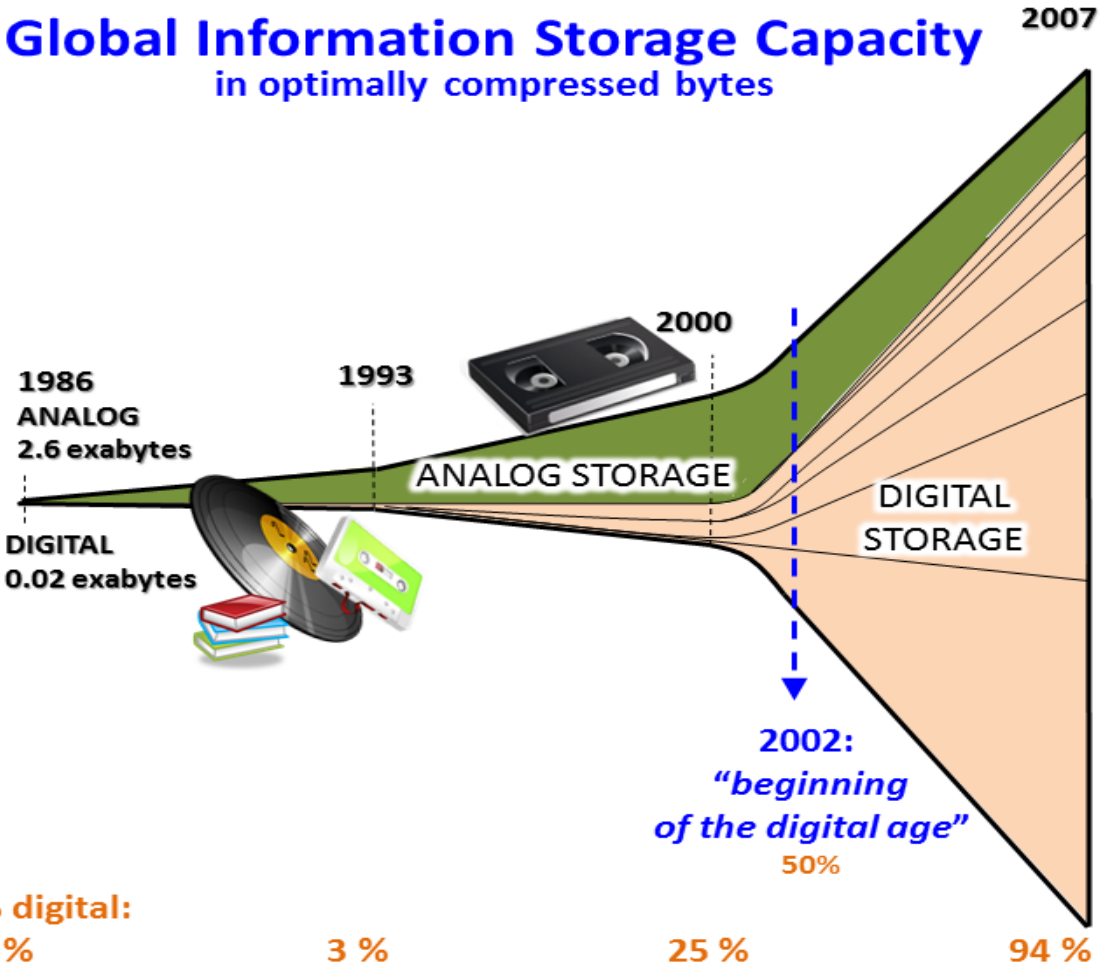
What is "Big Data"?

"**Big data** is a broad term for **data** sets so **LARGE** or **COMPLEX** that traditional **data** processing applications are inadequate. Challenges include analysis, capture, **data** curation, search, sharing, storage, transfer, visualization, querying and information privacy."

http://en.wikipedia.org/wiki/Big_data

Global Information Storage Capacity

in optimally compressed bytes



2007 **ANALOG**
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

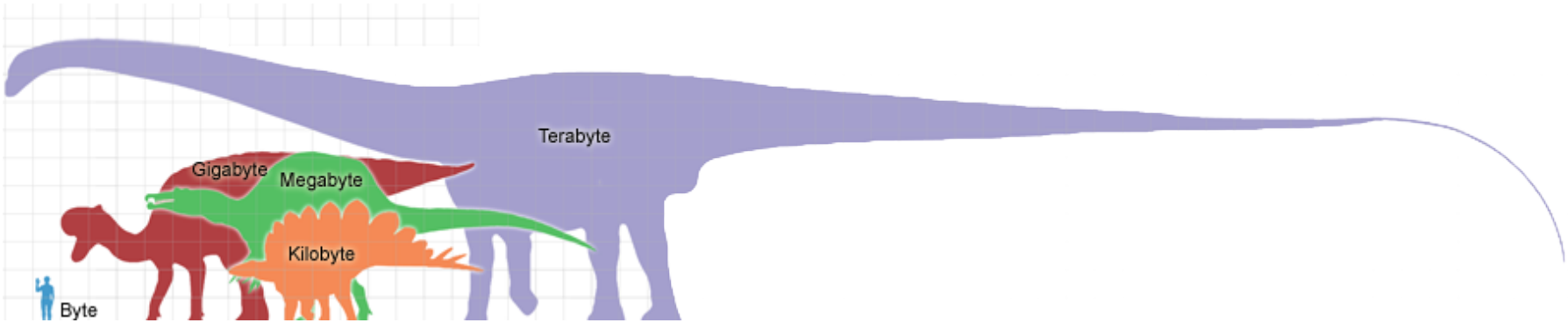
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes

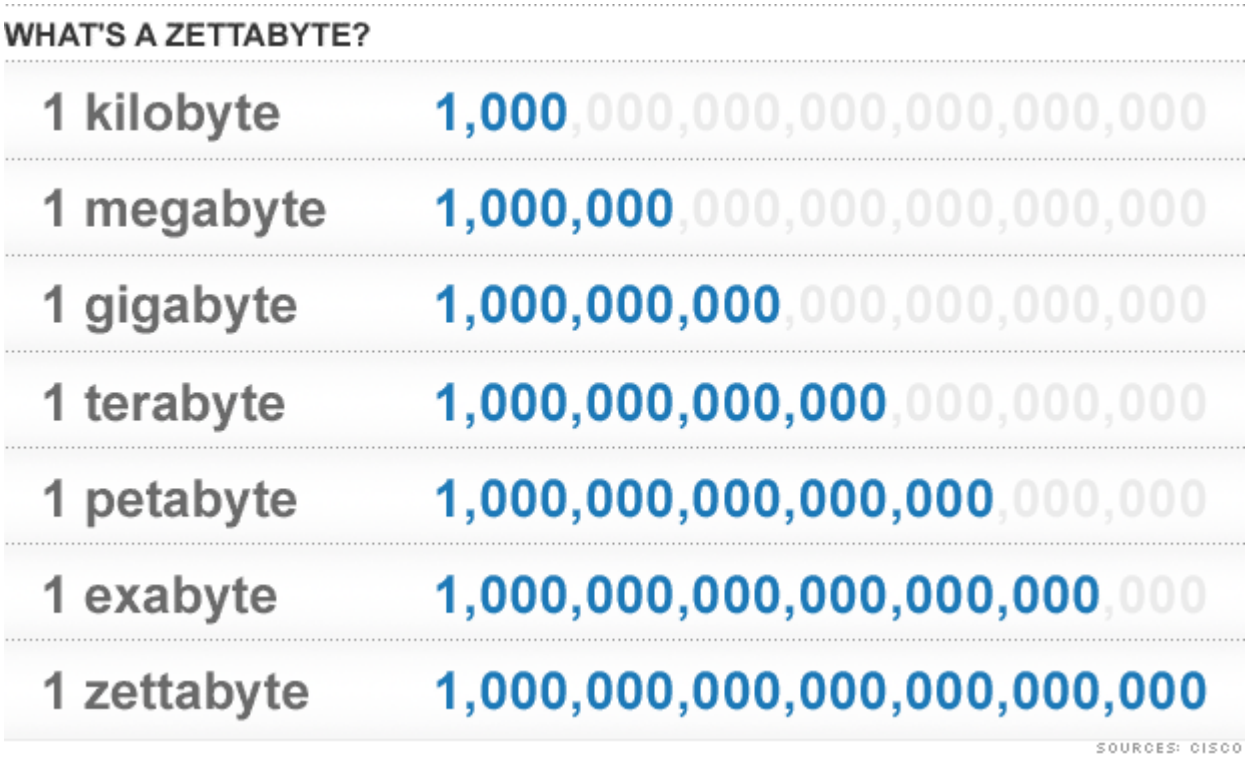
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

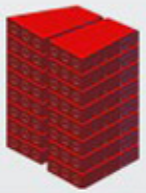

Terabyte
Gigabyte
Megabyte
Kilobyte
Byte

A **zettabyte** is a measure of storage capacity and is 2 to the 70th power bytes, also expressed as 10^{21} or **1 sextillion bytes**. One **zettabyte** is approximately equal to a thousand *exabytes* or a billion terabytes.



Anything Bigger?????????

How big is a Yottabyte?

TERABYTE	Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.	
PETABYTE	Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.	
EXABYTE	Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.	
ZETTABYTE	Will fill 1,000 datacenters or about 20% of Manhattan, New York.	
YOTTABYTE	Will fill the states of Delaware and Rhode Island with a million datacenters.	

Definition: Starbucks analogy

Some examples of "Big Data"

- Walmart handles more than 1 million customer transactions every hour, estimated to contain more than 2.5 petabytes of data
  - Equivalent of 167 times the information contained in all the books in the US Library of Congress
- Facebook handles 40 billion photos from its user base.
- Four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes of data in 2010 (13,000 terabytes).
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.

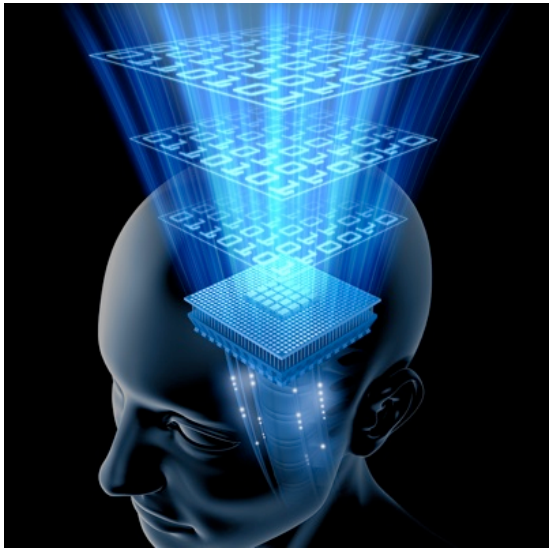
http://en.wikipedia.org/wiki/Big_data

Components of “Big Data”

Technical components of “Big Data”

Storage

Analytics



Presentation of results

This is not the whole story!

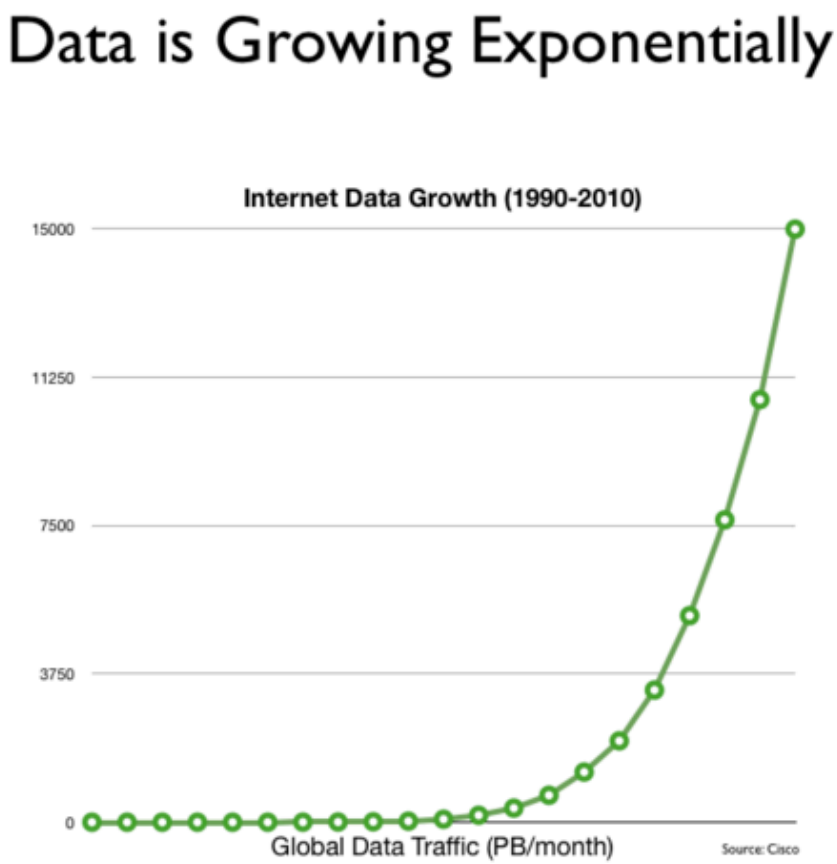
Important non-technical components of “Big Data”: **Legal** and **ethical** issues

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Why is collecting, storing, and analyzing “Big Data” hard?

Unprecedented size (that makes some of the techniques that you have learned unusable)

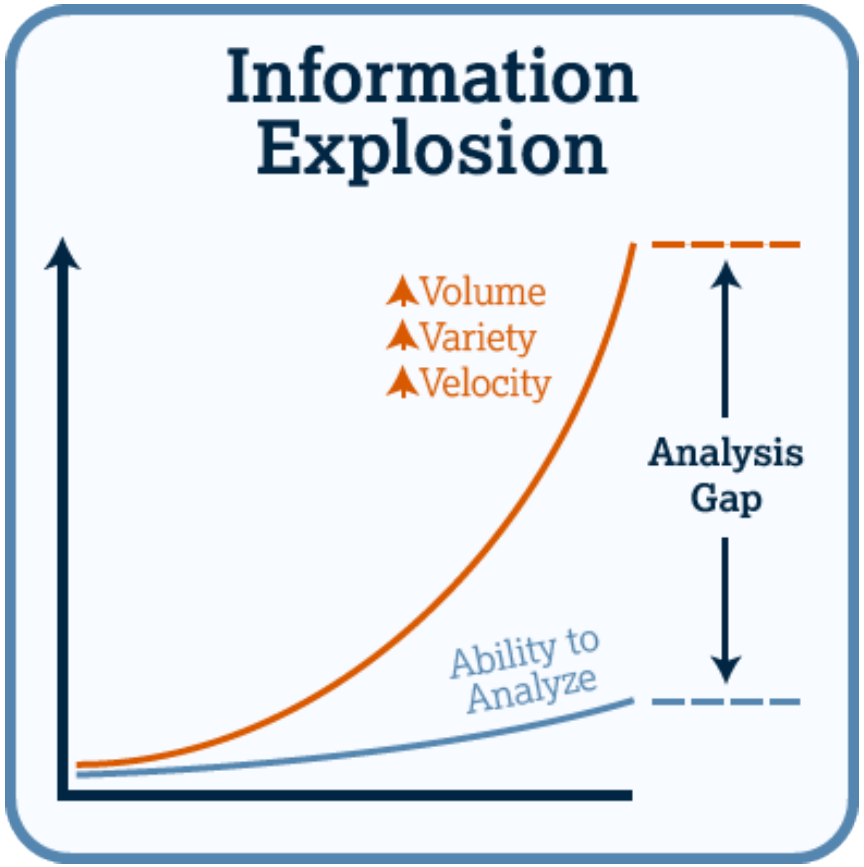
Why is collecting, storing, and analyzing "Big Data" hard?



http://trendspottr.tumblr.com/post/12525895145/real-time-trends-and-the-paradox-of-big-data

The amount of data collected grows exponentially with time

Why is collecting, storing, and analyzing "Big Data" hard?



http://www.jisc.ac.uk/publications/reports/2012/activity-data-delivering-benefits.aspx

Conventional techniques for analyzing data have a hard time catching up

Why is collecting, storing, and analyzing “Big Data” hard?

What is really important in “Big Data?”

What is really important in "Big Data?"

"The purpose of computing is insight, not numbers"

Richard Hamming
(preface to his 1962 book on numerical methods)
http://en.wikipedia.org/wiki/Richard_Hamming

What is “Big Data?”

"If you aren’t taking advantage of big data, then you don’t have big data, you have just a pile of data."

— Jay Parikh, VP of infrastructure at Facebook

Analytics (+ presentation of results, i.e., the user interface) seem to be the critical thing

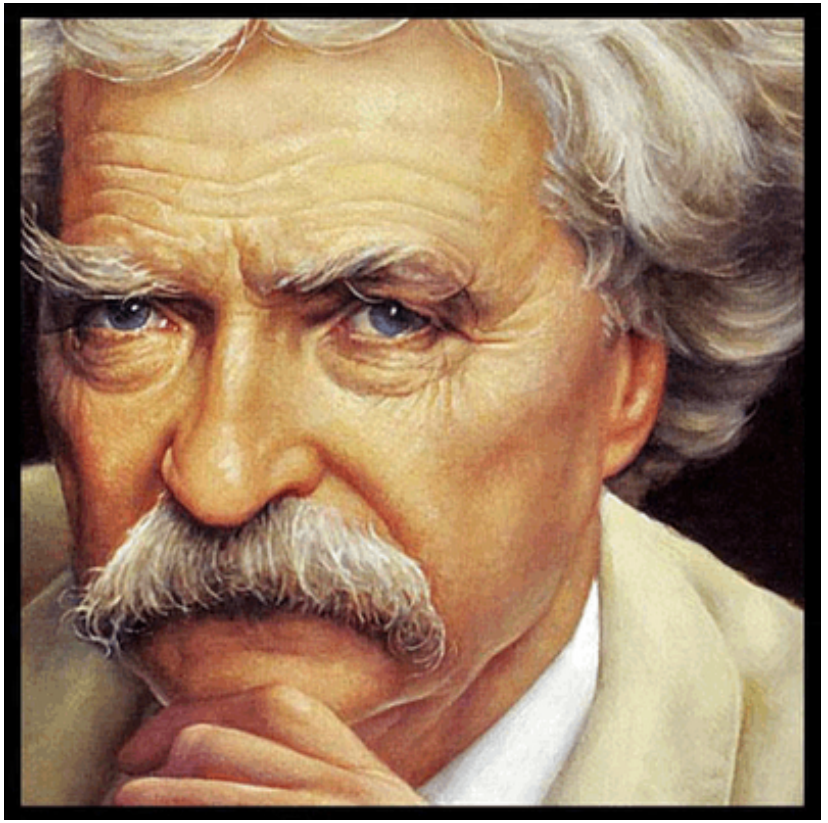
The goal of “Big Data”

Analytics!

Why would you even think of collecting and storing data without wanting to analyze them?

What is "Big Data?"

"A man who does not read has no advantage over a man who cannot read" — Mark Twain



"A man who does not analyze his data has no advantage over a man who has no data"

— Mar(e)k Druzdzel ☺

What is "Big Data?"

"Big data" – a personal view

"Big Data" does not seem to be more (above data analytics) than a sound use of old computer science techniques, such as distributed storage and distributed processing

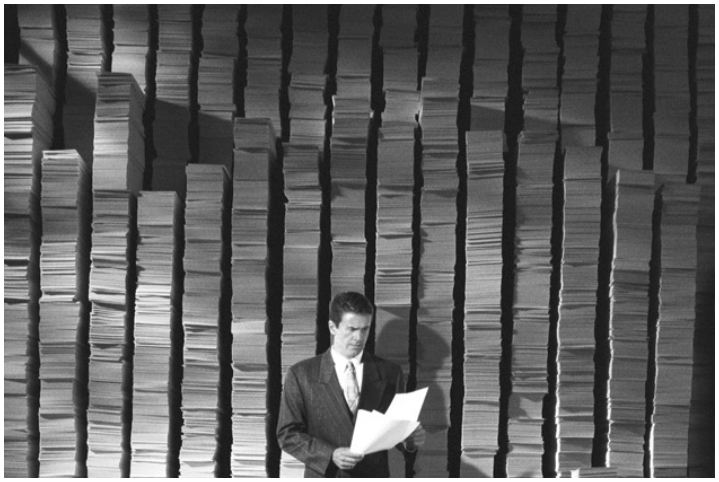
These techniques are simply a necessity when the amount of data and the complexity of computing becomes too large

The term "Big Data" will disappear, although the problems of efficient storage and retrieval, analysis, and presentation of results will stay

Foundations of data analytics

Base the analysis on procedures that are well grounded in statistics

What we will do in this course?



- In this course, you will go through the principles of collecting, storing and analyzing very large amounts of data.
- All this is amenable to automation.
- Storing and distributed processing of data will take just one block of classes

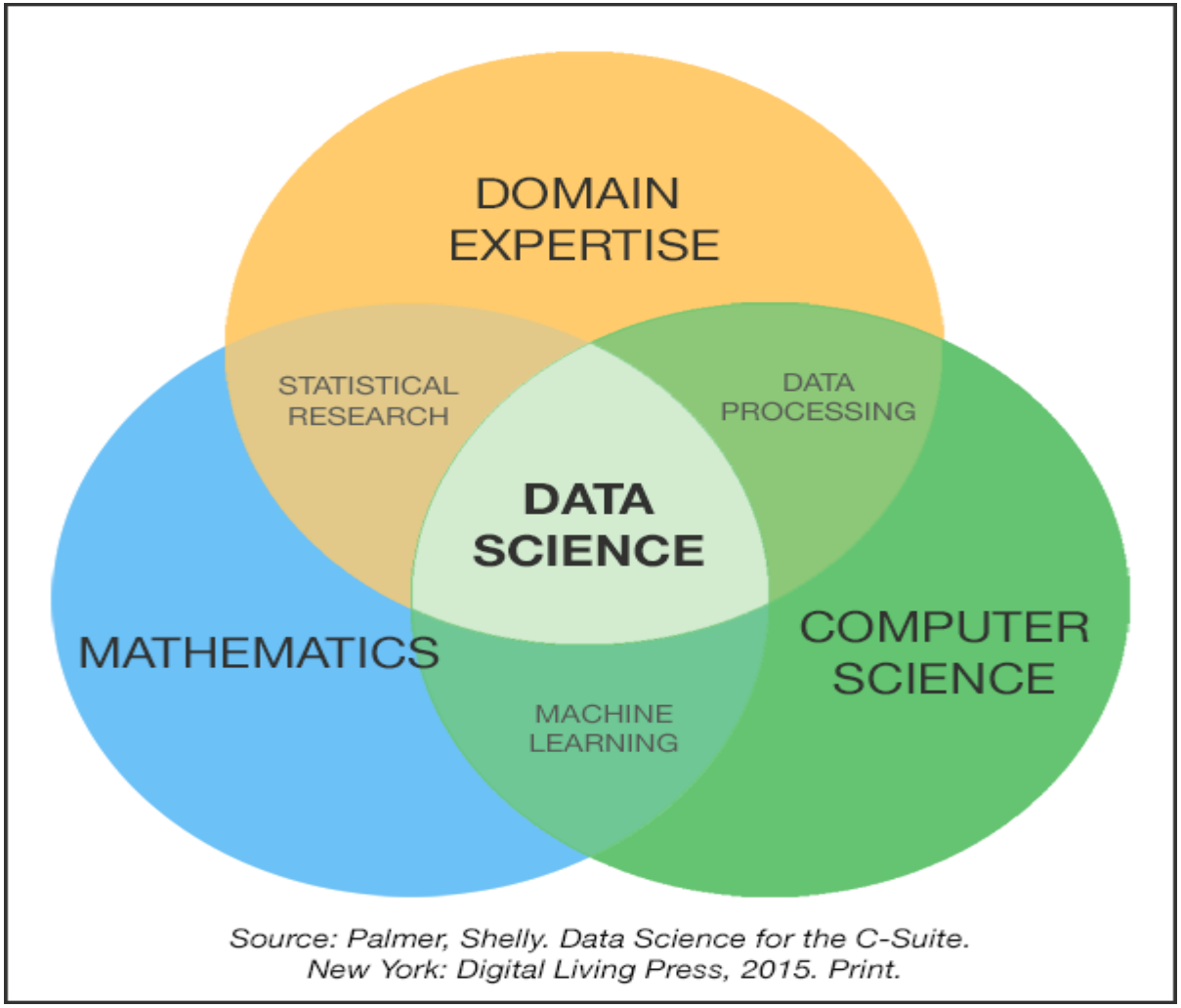
What we will do in this course?

We will overview the following

Data Storage/Processing → Data Analysis → Insights → Decision /Actions

Course relevance diagram

Data Analytics Skills



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Term project

From the following page:

http://www.kaggle.com/competitions/

Competition:

TBA

Your task: Win the competition

While winning will be rewarding (literary and in terms of your further career in information science), getting close will be sufficient for an excellent grade in this course.

Course outline

See the syllabus!

Please read for next class

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation