INFSCI 2725:
Data Analytics
(an introduction)

Marek J. Druzdzel

University of Pittsburgh
School of Information Sciences
and Intelligent Systems Program

marek@sis.pitt.edu
http://www.pitt.edu/~druzdzel

Outline

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

The instructor

Marek J. Druzdzel

associate professor, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh

Office : 2B10 IS Building (Decision Systems Laboratory)

Email : *marek@sis.pitt.edu*

Phone : +1 (412) 624-9432 (office)

WWW : http://www.pitt.edu/~druzdzel

The Teaching Assistant

Marcin Kozniewski

doctoral student, School of Information Sciences

Office : B-212 IS Building (Decision Systems Laboratory)

Email : *mak295@pitt.edu*

Phone : (412) 624-7378 (Decision Systems Laboratory)

Who are you?

Say in no longer than 25 seconds:

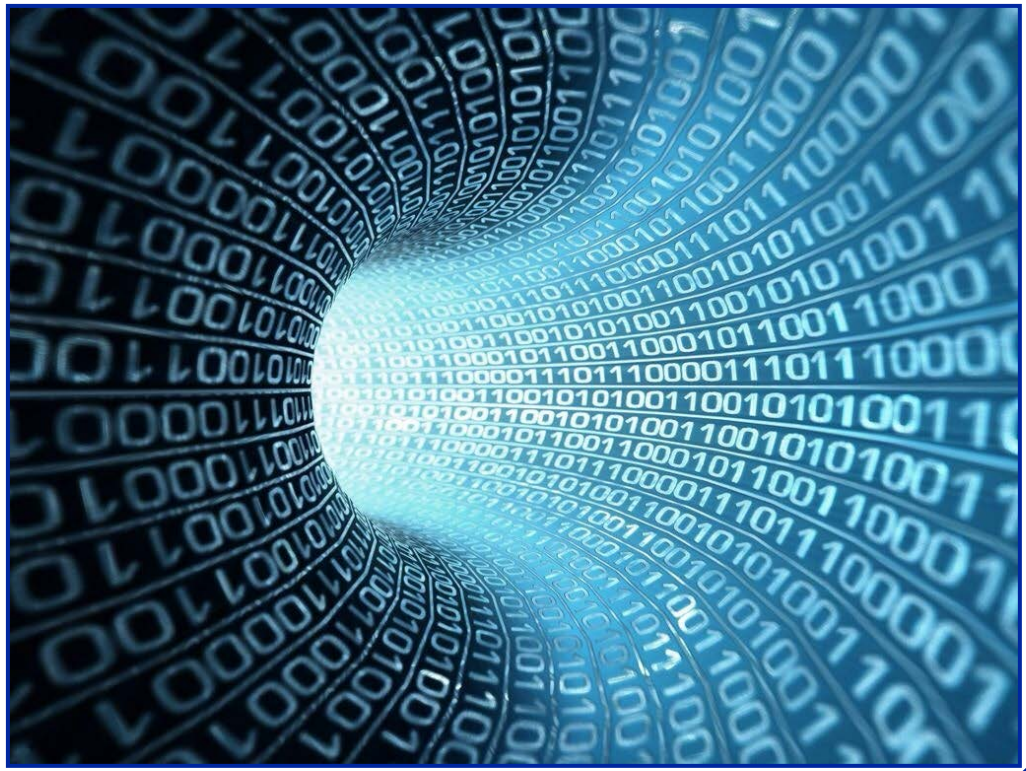
- **What is your name, what do you want to be called?**
- **What is your educational background (prior studies, current program)?**
- **What is your professional background (prior and current work experience)?**
- **What can you do? What are your strengths?**

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

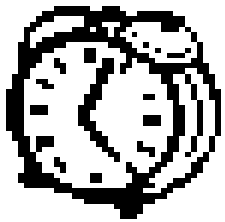
Organization of the Course

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.



Meeting times



Classes (403 IS Building):

Tuesdays, 12:00-2:50pm (break 1:20pm-1:35pm)

Marek's office hours (2B13 SIS Building):

Tuesdays, 3:00-4:00pm or by appointment

Marcin's office hours (2B12):

Mondays, 6:00-8:00pm or by appointment

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).

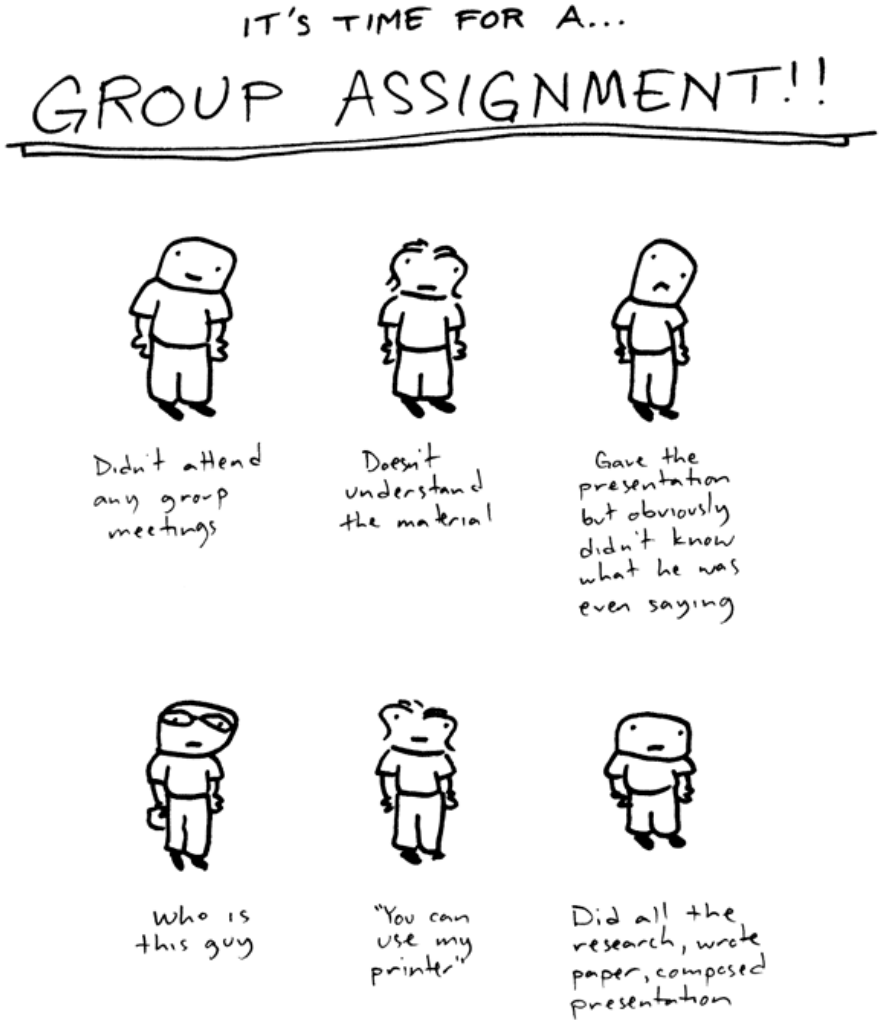
Deadlines are marked on the syllabus.

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

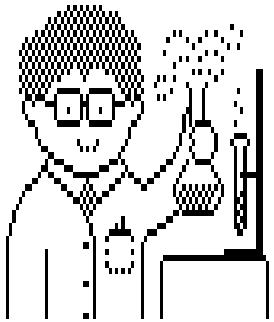
This is bad for you and is also explicitly forbidden by the University anti-plagiarism policies.

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).

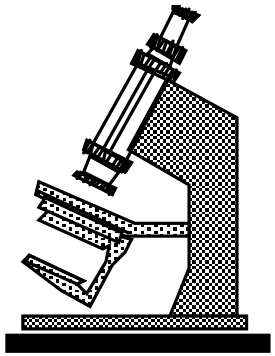


Term project



- 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.

Exam



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.

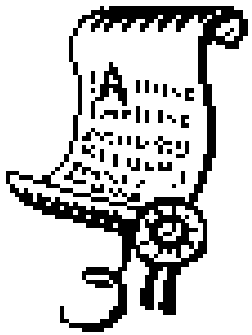
There are no limits on the font size – you can cram as much information on these two pages as you wish – but the notes have to be handwritten personally by you and this is a strict requirement.

Copied or computer-printed sheets are not allowed.

Expected effort (time load)

- Expect to spend about six hours quality time outside of class for every class meeting. I estimate that you will need about four hours to do the readings and two hours (on the average) 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

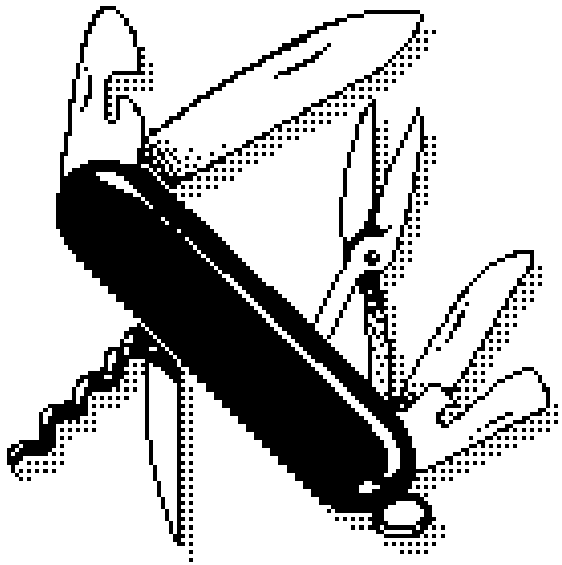


Your final grade for the course will be determined as follows:

- 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 participation.

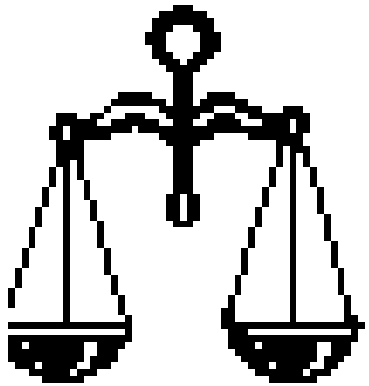
Useful Advice (Hopefully)



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?

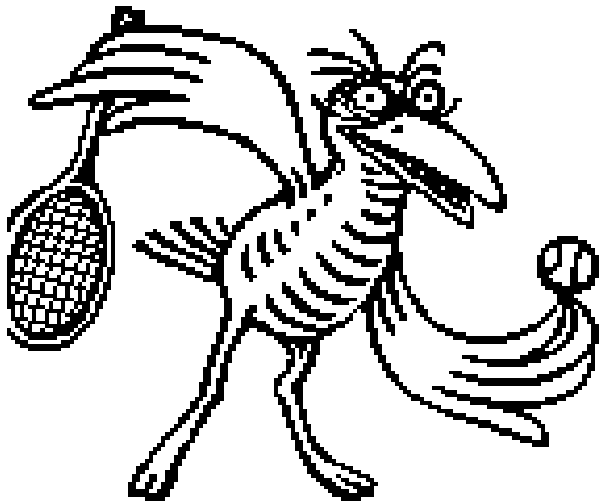


Come to classes …

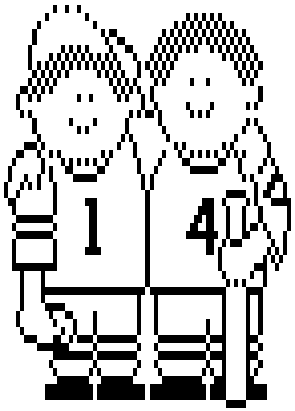
- Class attendance is important in learning.
- Coming to class stimulates timely reading of the material and helps you to be up to date on what is happening in the course.
- Our in-class discussions and exercises will be an important factor in your learning.
- Understanding difficult parts of the material 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, interrupt me if needed
- I reward your participation



Be good to your classmates

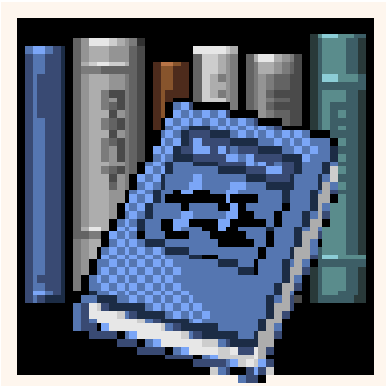


As somebody in a biology lab has once put it:

"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 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

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?”

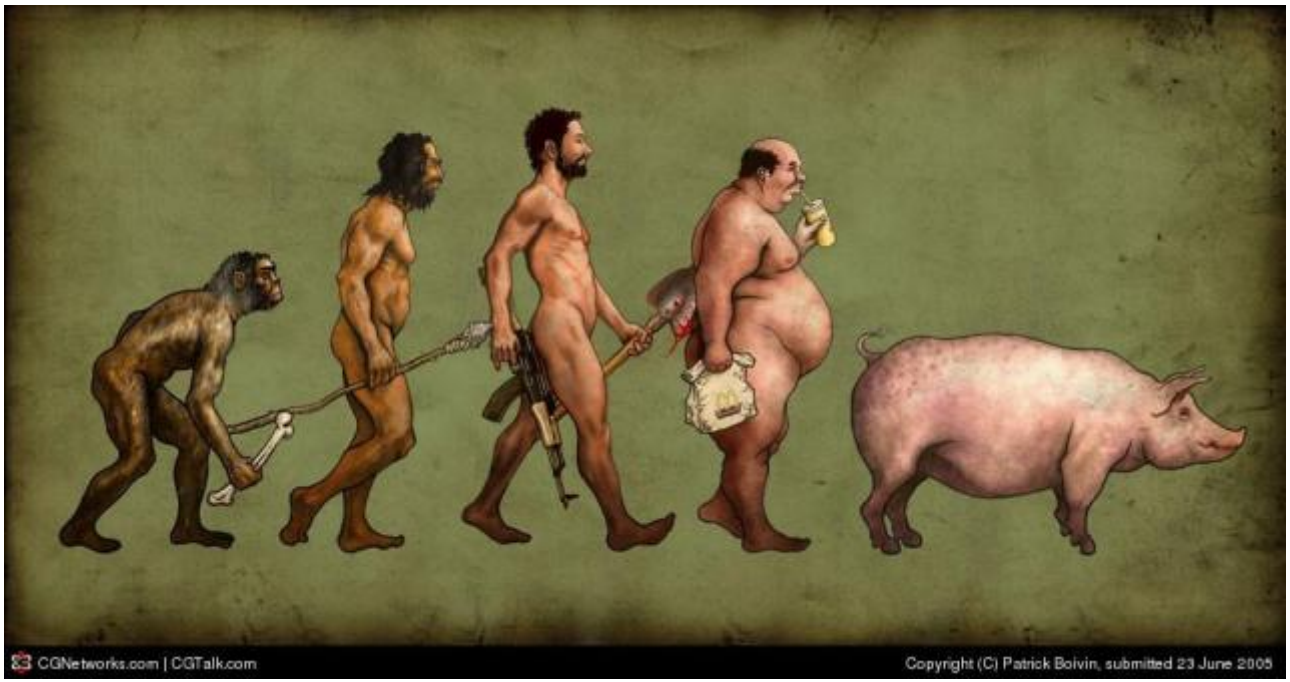
What is “Big Data”?

“Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, manage, and process the data within a tolerable elapsed time.”

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

Definition: Starbucks analogy

Definition: Obesity analogy



http://www.cdc.gov/obesity/data/adult.html

Some examples of “Big Data”

- When the Sloan Digital Sky Survey (SDSS) began collecting data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days.
- In total, the four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes of data in 2010 (13,000 terabytes).
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data - the 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.
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.

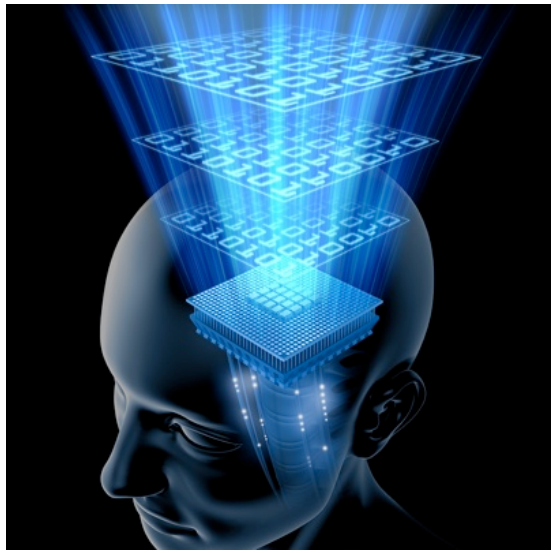
http://en.wikipedia.org/wiki/Big_data

Components of “Big Data”

Technical components of “Big Data”

Storage

Analytics



Presentation of results

This is not the whole story!

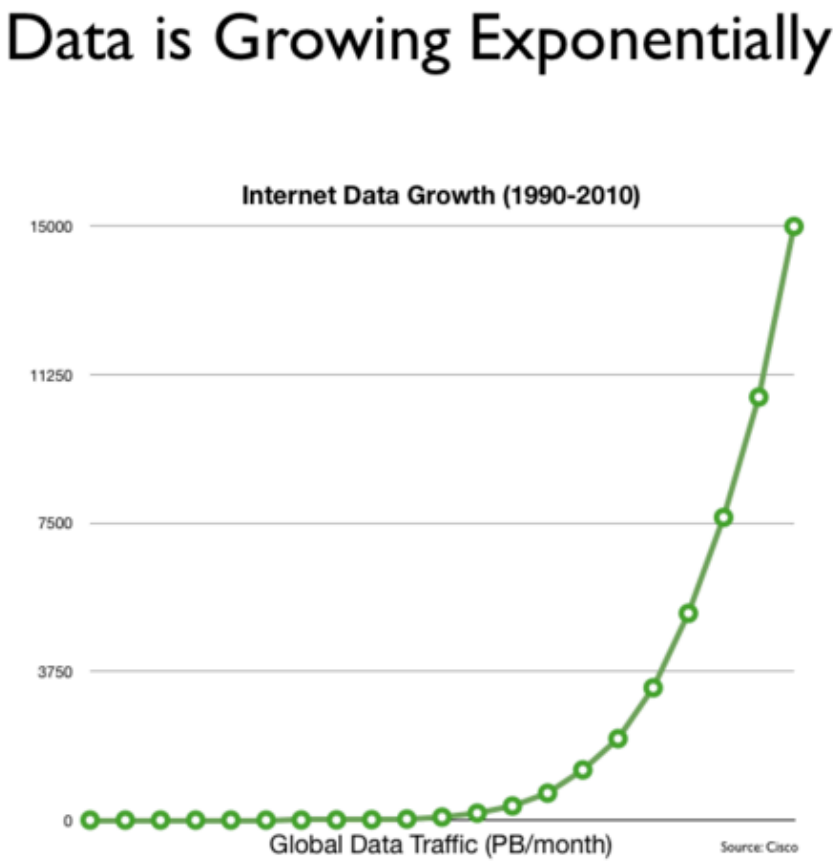
Important non-technical components of “Big Data”: Legal and ethical issues

"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)

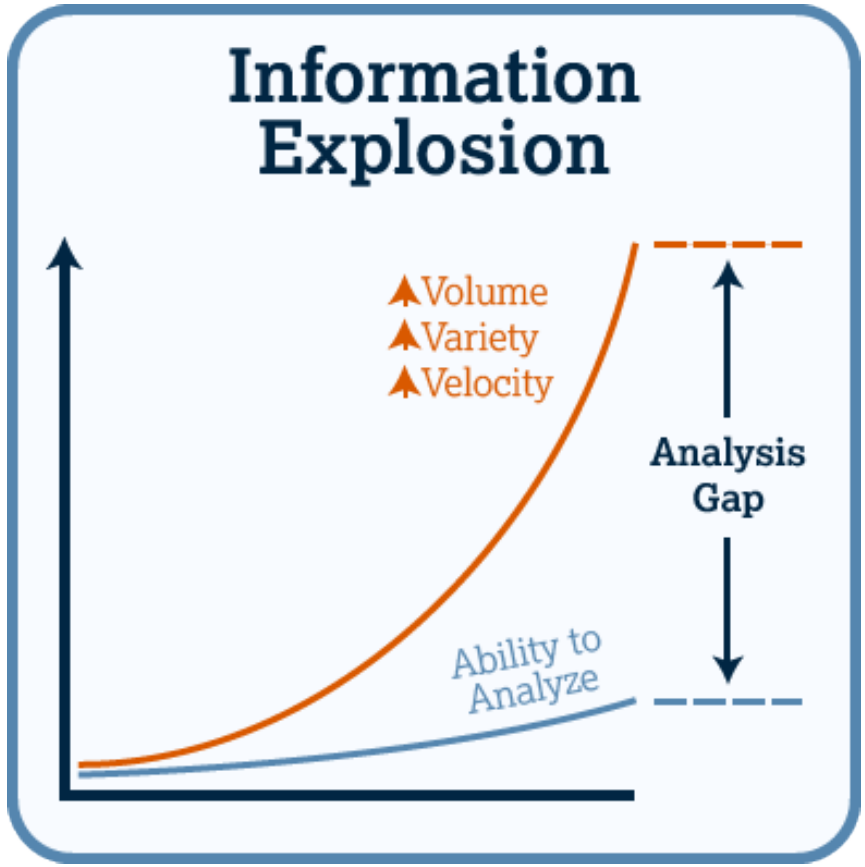
Why is collecting, storing, and analyzing "Big Data" hard?



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?”

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

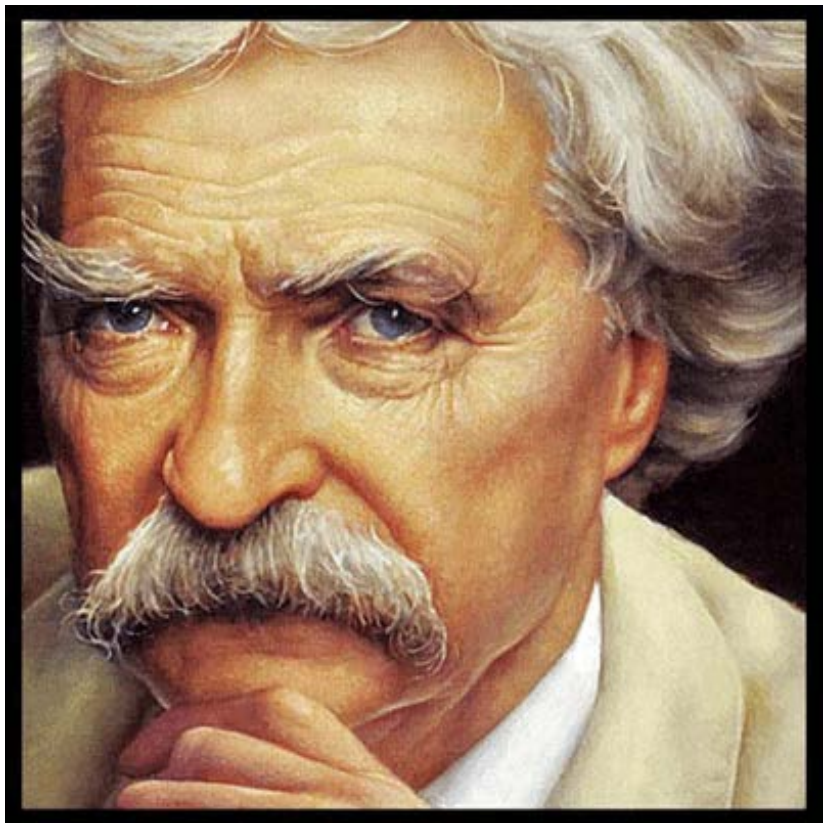
The goal of “Big Data”

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

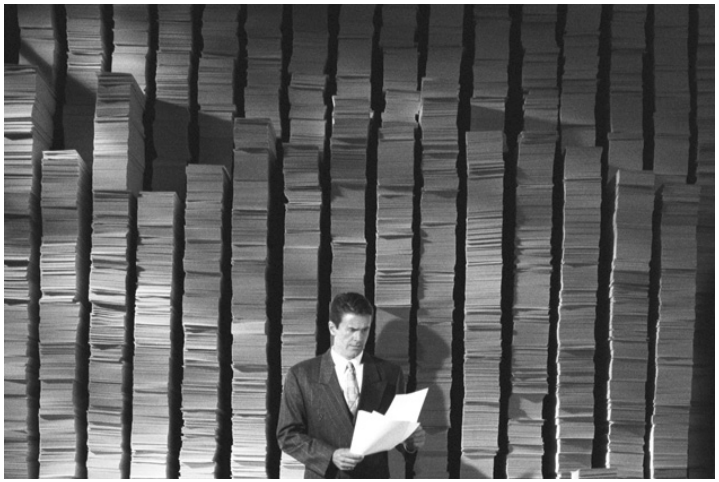
These techniques are simply a necessity when the amount of data and the complexity of computing becomes too large

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

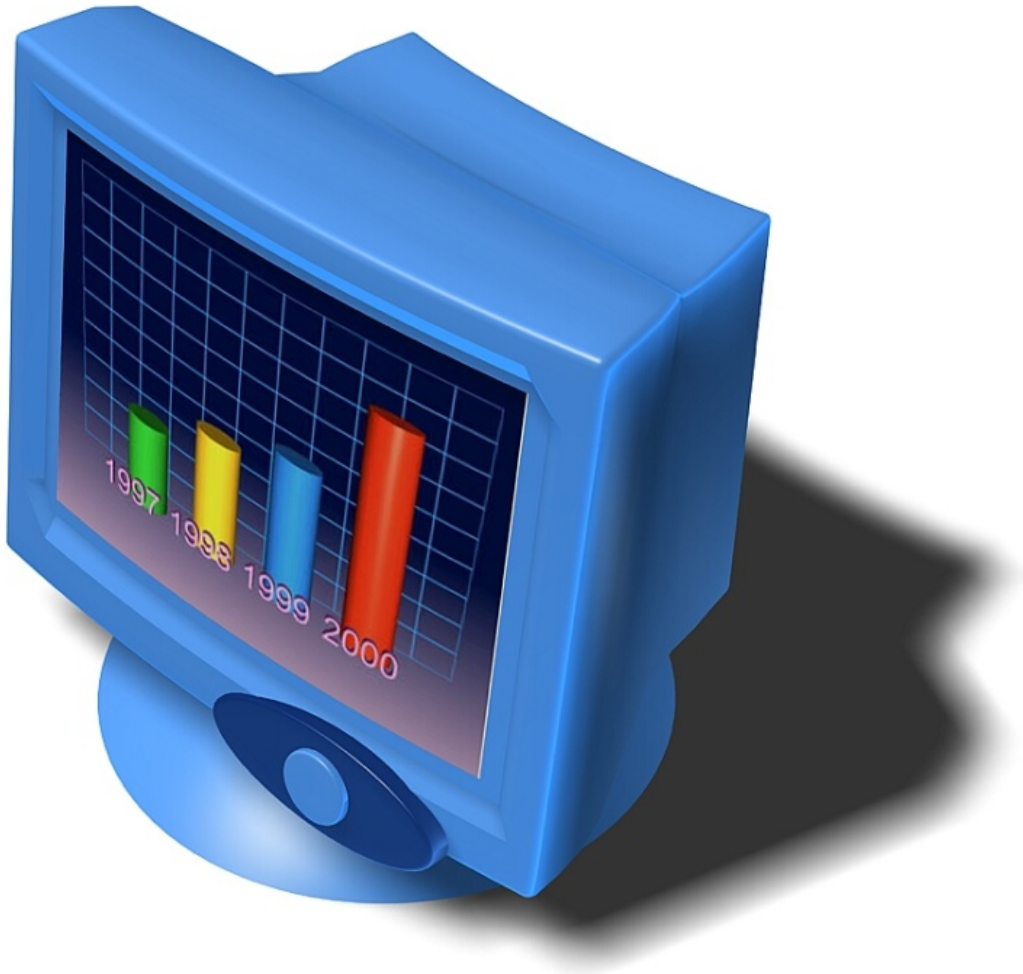
Base the analysis on procedures that are well grounded in statistics

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 (two meetings).
- A much harder thing is analytics!
- Much harder are issues that pertain to human users ☺.

What kind of things can you do when performing analytics?



Proposed Contents

(subject to slight changes ☺)

Course relevance diagram

Course Outline

(subject to changes ☺)

Course outline

See the syllabus!

Term project

From the following page:

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

Competition:

TBA (will be coordinated with the other section of this course)

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.