

# Analytics and Visualization of Big Data

*Fadel M. Megahed*

## ***Lecture 02: Course Intro (Cont.) + Data Mining***



**AUBURN UNIVERSITY**

SAMUEL GINN  
COLLEGE OF ENGINEERING

***Department of Industrial and Systems Engineering***

*Spring 13*

- “(...) you will select what work you complete and thus, have a strong control over your learning outcomes from the class. I will also promote the concept of a **learning community**, where the participants will be encouraged to share/develop resources that are relevant to the material covered in class. *After all an industrial engineering data-mining class that does not efficiently and effectively utilize its resources in generating knowledge should not exist!!*”



# What are the Assignments that you will Choose from?

3

- In this course, you select what work you complete bearing these 3 simple rules in mind:
  - At least 70 percent of the total points possible for each individual assignment must be earned; otherwise, no points will be recorded for the assignment.
  - Once the due date for an assignment has passed, that assignment cannot be completed.
  - With the exceptions of exams, you are allowed/encouraged to ask your colleagues questions as long as you use Piazza for this purpose. To access our Piazza page, please click [here](#).



# What are the Assignments that you will Choose from?

4

- Homework:
  - Help you gauge your understanding of the material
  - ~ One HW Set every 2 chapters (6 total)
  - Recommendation: Read the text and review class slides before exploring the homework
  - Each homework is worth 10 points (i.e. you can get up to 60 total)






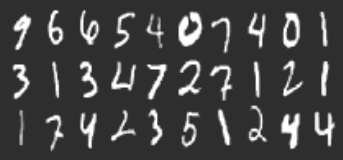

Disclaimer: The information in this slide is aimed at providing you with an overview. Please read the syllabus for more details.



# What are the Assignments that you will Choose from?

5

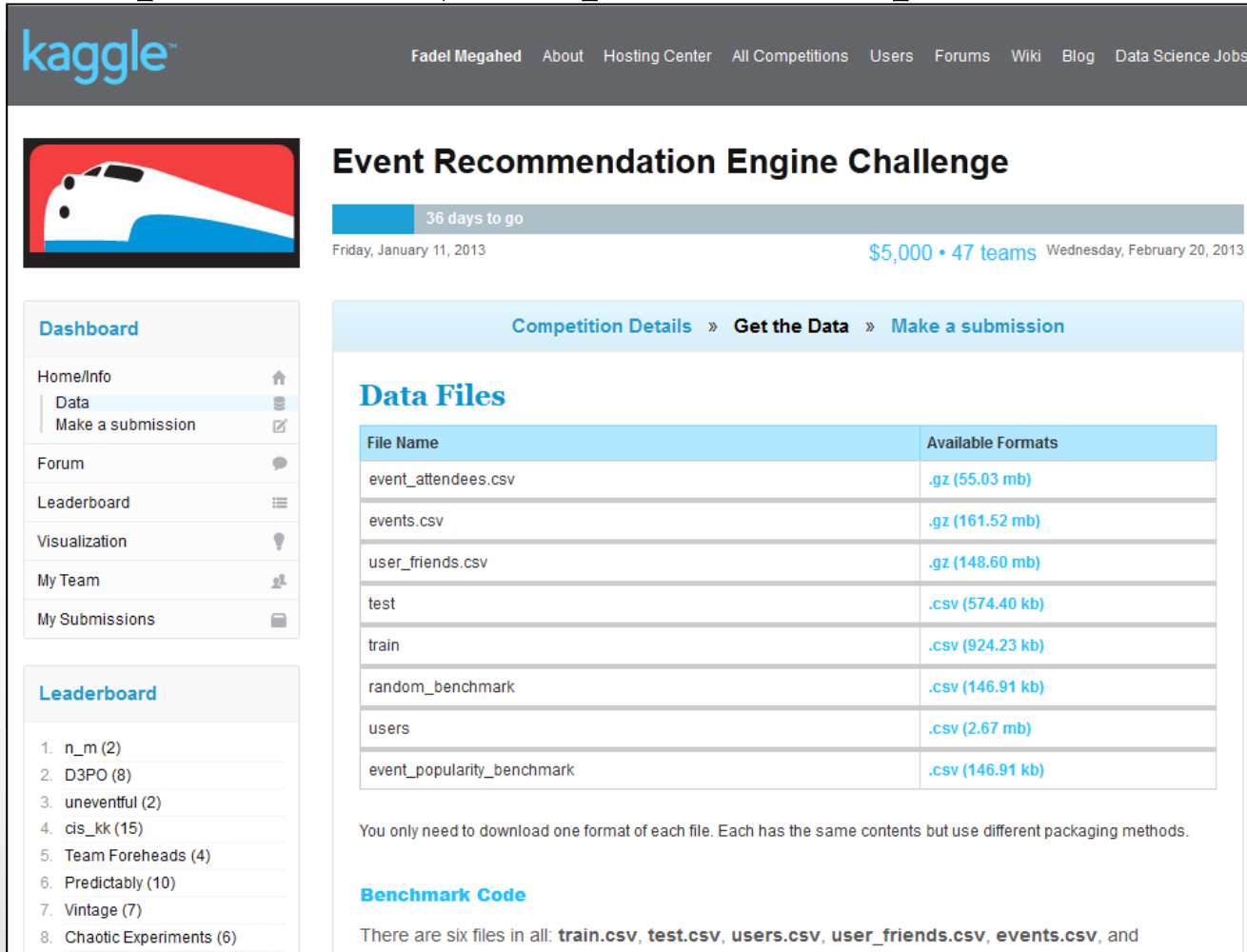
- DM Competitions\*: ( $\leq 2$ , possible 50 pts each, click [here](#))

 Featured Competitions		
	<b>Traveling Santa Problem</b> Solve ye olde traveling salesman problem to help Santa Claus deliver his presents	3.4 days next deadline 325 teams \$3,000
	<b>Event Recommendation Engine Challenge</b> Predict what events our users will be interested in based on user actions, event metadata, and demographic information.	36 days next deadline \$5,000
	<b>Heritage Health Prize</b> Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)	2 months next deadline 1501 teams \$3 million
 Getting Started Competitions		
	<b>Digit Recognizer</b> Classify handwritten digits in this "Getting Started" competition.	6 months next deadline 813 teams Knowledge
	<b>Titanic: Machine Learning from Disaster</b> Getting Started Competition, with tutorials in Excel, Python and introduction to Random Forests.	8 months next deadline 1433 teams Knowledge

# What are the Assignments that you will Choose from?

6

- DM Competitions\*: ( $\leq 2$ , possible 50 pts each, click [here](#))



The screenshot shows the Kaggle website interface for the "Event Recommendation Engine Challenge". The top navigation bar includes the Kaggle logo and links to Fadel Megahed, About, Hosting Center, All Competitions, Users, Forums, Wiki, Blog, and Data Science Jobs. The main header features a train icon and the challenge title "Event Recommendation Engine Challenge". A progress bar indicates "36 days to go" with the start date "Friday, January 11, 2013" and the end date "Wednesday, February 20, 2013". The prize pool is "\$5,000 • 47 teams". The left sidebar contains a "Dashboard" with links to Home/Info, Data, Make a submission, Forum, Leaderboard, Visualization, My Team, and My Submissions. Below the dashboard is a "Leaderboard" showing the top 8 teams: n\_m (2), D3PO (8), uneventful (2), cis\_kk (15), Team Foreheads (4), Predictably (10), Vintage (7), and Chaotic Experiments (6). The main content area has a "Competition Details" section with links to "Get the Data" and "Make a submission". Below this is a "Data Files" table listing the available data files and their formats. The table has two columns: "File Name" and "Available Formats". The files listed are event\_attendees.csv, events.csv, user\_friends.csv, test, train, random\_benchmark, users, and event\_popularity\_benchmark. The formats are .gz (55.03 mb), .gz (161.52 mb), .gz (148.60 mb), .csv (574.40 kb), .csv (924.23 kb), .csv (146.91 kb), .csv (2.67 mb), and .csv (146.91 kb) respectively. Below the table, a note states: "You only need to download one format of each file. Each has the same contents but use different packaging methods." There is also a "Benchmark Code" section with the text: "There are six files in all: train.csv, test.csv, users.csv, user\_friends.csv, events.csv, and".

**Event Recommendation Engine Challenge**

36 days to go

Friday, January 11, 2013 \$5,000 • 47 teams Wednesday, February 20, 2013

[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

### Data Files

File Name	Available Formats
event_attendees.csv	.gz (55.03 mb)
events.csv	.gz (161.52 mb)
user_friends.csv	.gz (148.60 mb)
test	.csv (574.40 kb)
train	.csv (924.23 kb)
random_benchmark	.csv (146.91 kb)
users	.csv (2.67 mb)
event_popularity_benchmark	.csv (146.91 kb)

You only need to download one format of each file. Each has the same contents but use different packaging methods.

### Benchmark Code

There are six files in all: **train.csv**, **test.csv**, **users.csv**, **user\_friends.csv**, **events.csv**, and

# What are the Assignments that you will Choose from?

7

- Project\*: (You can only do one of the two options)
  - A data mining project\* (100 points possible)
  - A big data project (200 points possible)
- These two projects should be thought about in 3 Phases:
  - Project Identification
  - Data Manipulation
  - Presenting your results (Grad Students must document their work in a conference/journal-style paper, see example in the Syllabus)




# What are the Assignments that you will Choose from?

8


**Teammate Search** *(Posted by Piazza)*

Need to form teams? Create a post below to initiate a search and we'll notify you via email when others respond.

**add new post:**

☒

I'm **one student** looking for more people to work with.

☐

I'm **from a group** looking for more students.

**\*Name**

Fadel Megahed

**\*Email**

fmegahed@auburn.edu

**\*About Me**

Introduce yourself. What kind of teammate(s) are you looking for?  
...


*(Things you could include: your location, grad/undergrad, when you're available... help people get to know you!)*

Submit

**1 open search:**

# students

about

  
one student

**Robin Gautham Muthukumar** (rzm0024@auburn.edu)  
I'm from India currently residing in W Glenn Avenue. Did my undergrad in Aeronautical Engineering. Developed interest in Big Data from Dec 2012. Interested to work in manufacturing, health care and other interesting topics.  

hide responses

**Fadel Megahed** (fmegahed@auburn.edu) 1/14/13 

delete

Hi Robin, do you have access to any datasets? Can you choose a dataset of interest from the last two slides in Thursday's class? I will go over that material tomorrow in class, remind me to publicize your post!! Please feel free to revisit your post with specific examples of datasets that you may want to look at. Any Kaggle competitions that you want to tackle? You can also highlight that to provide your fellow colleagues with more details on your interests :)

Write your response... 

respond





# What are the Assignments that you will Choose from?

9

- Animated Visualization: (50 points possible)
  - We provided an example of visual analytics in sports
  - Another example is highlighted in this video

See Hans Rosling's Talk at the State Dept. (not this one) at [http://www.ted.com/talks/hans\\_rosling\\_at\\_state.html](http://www.ted.com/talks/hans_rosling_at_state.html)



# What are the Assignments that you will Choose from?

10

- Class Blog: (260 points possible, click [here](#))
  - Bloggers:
    - At least 16 blogs in 8 weeks (zero credit if less than 16)
    - At least four of them should be tutorials
    - Purpose of the non-tutorials: is to bring to the attention of class (and the world) interesting readings, news stories, and visual-analytic tools relevant to DM/Big Data.
  - Participants:
    - Utilize the information gained from the blog and apply it to a different dataset.
    - Provide ~250 word comment (with screenshots, figures, and/or tables) showing what you have learned.
    - At least 8 posts (zero credit if less than 8)



# What are the Assignments that you will Choose from?

11

JAN

14



Dear Big Data Students,

First...Thanks to everyone for helping establish a collegial, inviting and thoughtful classroom environment. Your willingness to critically engage with and talk about topics from the presentations suggests that we'll have a great opportunity to learn about big data, while having fun in the process.

With a large section of 55+ intellectually adept INSY students, we will rarely (if ever) have enough time to adequately address all topics and questions. Thankfully, you now have a digital space for such intellectual endeavors. The guidelines for using the blog are highlighted in both the [Syllabus](#) and in the [lecture materials](#) for the first two classes. In addition, please feel free to use the blog to:

1. elaborate upon an idea from in-class discussion; or
2. engage in another mode of critical reflection.

From the Dashboard, all you need to do is click "New Post".

If questions or suggestions arise, please don't hesitate to contact [me](#). Don't forget, this blog is made available on the web to educate not only your INSY 4970 colleagues, but also to share your thoughts, reflections, tutorials and Big Data-related news to the rest of the world. Essentially, this is our space for showcasing the skills of ENG students at Auburn University in making sense of Big Data and using the knowledge captured to tackle large-scale engineering problems.

For the time being, please familiarize yourself with the Big Data Discussion blog. Whenever you are ready, feel free to compose an engaging and critically reflective post pertaining to big data analytics.

Let the digital critical engagement begin...

War Eagle!!

Fadel Megahed

[www.fadelmegahed.com](http://www.fadelmegahed.com)

Note: If you have not sent me your Gmail yet, please add your Gmail (you will need to create one if you do not have one) to the following spreadsheet, click [here](#) to access.



# What are the Assignments that you will Choose from?

12

- Exams ( $\times 2$ ); each worth 80 points
  - Somewhat similar to the homework
  - There will be questions based on the blog
  - The exams are 75 mins each



The points total for each assignment group is provided below:

Homework, 6 total at	
10 points per homework	60 points
DM competitions, 2 total at	
50 points per competition	100 points
DM Project	100 points
Big Data Project	200 points
Conference-Style Paper	100 points [Mandatory for Grads doing a Project]
Visualization Assignment	50 points
Blog – Blogger, 20 max (16 min)	
10 points per post	200 points
Blog – Participant, 8 max (8 min)	
7.5 points per post	60 points
Exam I	80 points
Exam II	80 points
<hr/>	
930 points TOTAL [Assuming Big Data Project]	

Letter grades will be assigned using the following scale:

Letter Grade	Points
A (UG)	515 and above
A (G)	535 and above
B (UG)	465-514
B (G)	485-534
C (UG)	400-464
C (G)	420-484
F (UG)	399 and below
F (G)	419 and above

\*Please note that I will round your grades to the *nearest* integer at the *end* of the course.



- Course calendar is the last page in the syllabus
- I expect that each student in this class will develop an expertise in one or multiple areas related to the topics that we discuss in class.
- You totally control your learning in this class.
  - Zig Ziglar, a motivational speaker/author once said *“people often say that motivation doesn't last. Well, neither does bathing - that's why we recommend it daily.”*
  - There are a ton of supplementary data that is available online
- Start learning about big data (or DM) today ☺



# A Sample of the Available Datasets (and Info) Online

15

- UC Berkley class on Big Data and Twitter,  
<http://www.youtube.com/playlist?list=PLE8C1256A28C1487F>
- The Linguistic Data consortium, <http://ldc.upenn.edu/>
- U Michigan compilation of several datasets,  
<http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/subject.jsp>
- Wikipedia dumps, <http://dumps.wikimedia.org/enwiki/>
- Stanford collection of social media datasets,  
<http://snap.stanford.edu/news.html>
- The global terrorism database, <http://www.start.umd.edu/gtd/about/> (Need to request data using a contact form)



# A Sample of the Available Datasets (and Info) Online

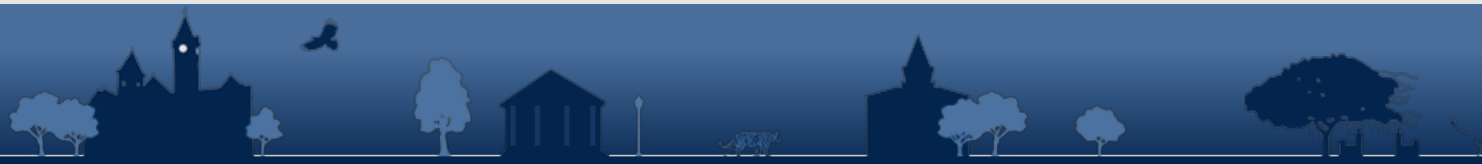
16

- Public datasets on AWS,  
[http://aws.amazon.com/datasets?\\_encoding=UTF8&jiveRedirect=1](http://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1)
- Some suggested data mining projects with their datasets, <http://cms.uhd.edu/faculty/chenp/class/4319/project/index.html>
- Large health data sets, <http://www.ehdp.com/vitalnet/datasets.htm>
- Google Ngrams, <http://books.google.com/ngrams/>
- Analyzing twitter data webtool, <http://www.tweetcharts.com/>
- A great list of ted talks on using large datasets,  
[http://www.ted.com/playlists/56/making\\_sense\\_of\\_too\\_much\\_data.html](http://www.ted.com/playlists/56/making_sense_of_too_much_data.html).





## Chapter 01: Data Mining



## What is Data Mining (DM)?

- **Statistical Model**
- **Machine Learning**
- **Summarization**
- **Feature Extraction**

## Statistical Limits on DM

- **Total Information Awareness Act**
- **Bonferroni's Principle**
- **Examples on Bonferroni's Principle**

## Things Useful To Know

- **Importance of Words in Documents**
- **Hash Functions**
- **Indexes**
- **Power Laws**



- The most common definition of data mining is the discovery of models from data.
- Discovery of **patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
- Subsidiary issues:
  - **Data cleansing:** detection of bogus data
  - **Visualization:** something better than MBs of output
  - **Warehousing** of data (for retrieval)

Source: The slide is adapted from Jure Leskovic, Stanford CS246, Lecture Notes, see <http://cs246.stanford.edu>



- These models can be one of the following:
  - Statistical models
  - Machine learning
  - Summarization
  - Feature Extraction

}

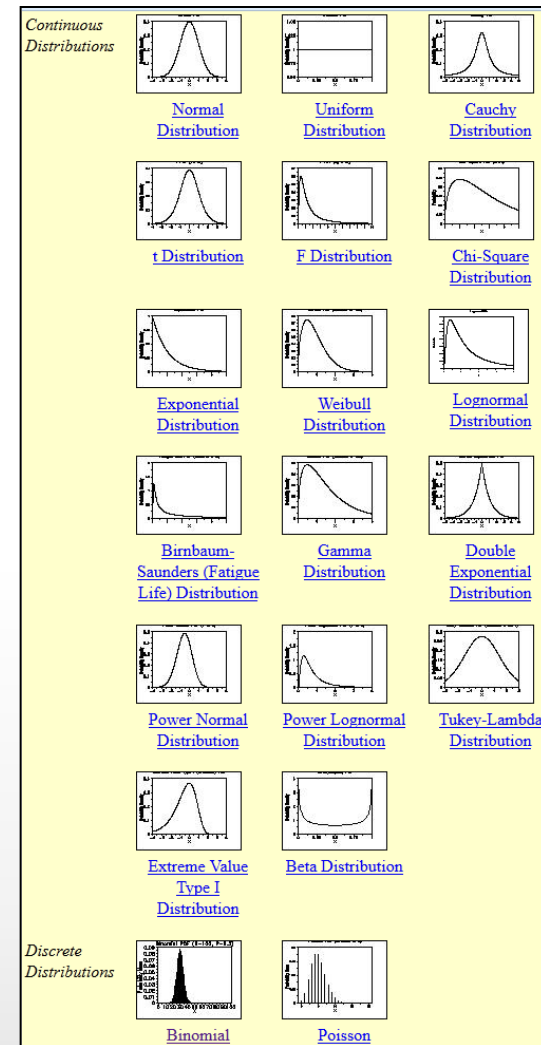
Predictive methods

Descriptive methods
  
- Data Mining is hard since it has the following issues:<sup>1</sup>
  - Scalability
  - Dimensionality
  - Complex and Heterogeneous Data
  - Data Quality
  - Data Ownership and Distribution
  - Privacy Preservation

1. Adapted from Jure Leskovic, Stanford CS246, Lecture Notes, see <http://cs246.stanford.edu>



- Fitting a distribution to the data and figuring out the *most likely parameters* for this distribution.
- This involves the following two steps:
  - Determination of the "best-fitting" distribution.
  - Estimation of the parameters (shape, location, and scale parameters) for that distribution.



Important Note: For more information on this important topic, please check the NIST Engineering Statistics Handbook, click [here](#).

- ML practitioners use available data to train an algorithm that is used in discovering new knowledge from the data.
- Typically, used when we have little idea of what we are looking for in the data;
  - Good example: Netflix challenge (see a brief description [here](#))
  - Bad example: locating resumes/CVs on the web
- In this class, we will discuss ML in Chapter 12, click [here](#) for obtaining the chapter.

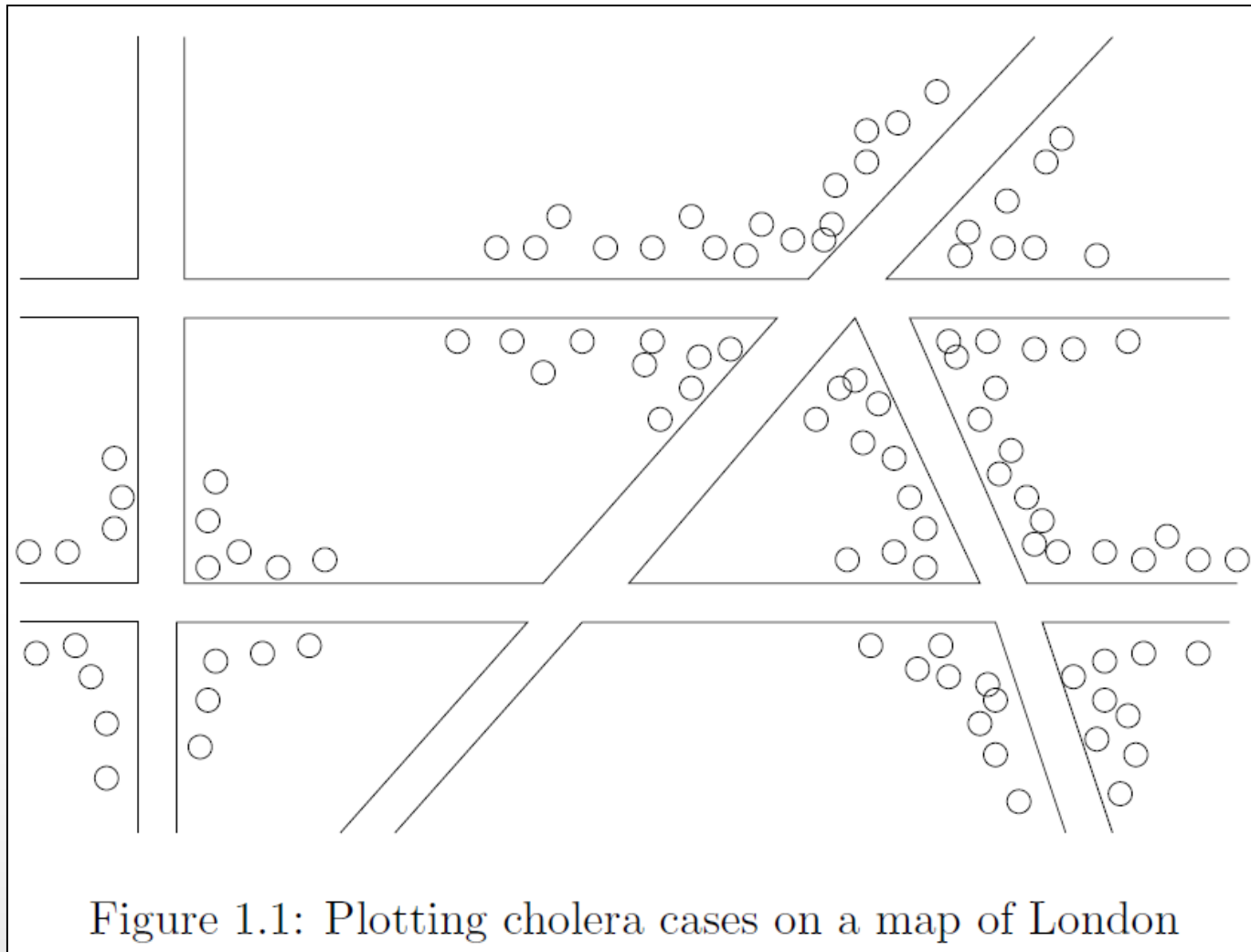


- There are several ways to summarize very complex data. Two of them are of particular interest:
  - **Page Rank:**
    - This is what made Google successful.
    - Essentially, the entire web is summarized by a single number, “Page Rank”.
  - **Clustering:**
    - Data can be viewed as points in a multidimensional space.
    - Points that are close are assigned to the same cluster.
    - Cluster summaries (e.g. centroids) become the summary of the entire dataset



# Summarization – The 1854 Cholera Outbreak

24



Source: A. Rajaraman, J. Leskovec, J.D. Ullman. (2012). "Mining of Massive Datasets". <http://i.stanford.edu/~ullman/mmds.html>



- In these models, we look for the most extreme examples of a phenomenon.
- Some of the important feature extraction models that we will discuss in this course include:
  - **Frequent Itemsets:**
    - We look for items that appear together in many “baskets”
    - Original application was the true market basket
  - **Similar Items:**
    - Used for recommendations
    - E.g. recommender systems in Ch. 9 and LSH in Ch.3



## What is Data Mining (DM)?

- Statistical Model
- Machine Learning
- Summarization
- Feature Extraction

## Statistical Limits on DM

- **Total Information Awareness Act**
- **Bonferroni's Principle**
- **Examples on Bonferroni's Principle**

## Things Useful To Know

- Importance of Words in Documents
- Hash Functions
- Indexes
- Power Laws



- A big risk when data mining is that you will discover patterns that are meaningless.
- Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find.

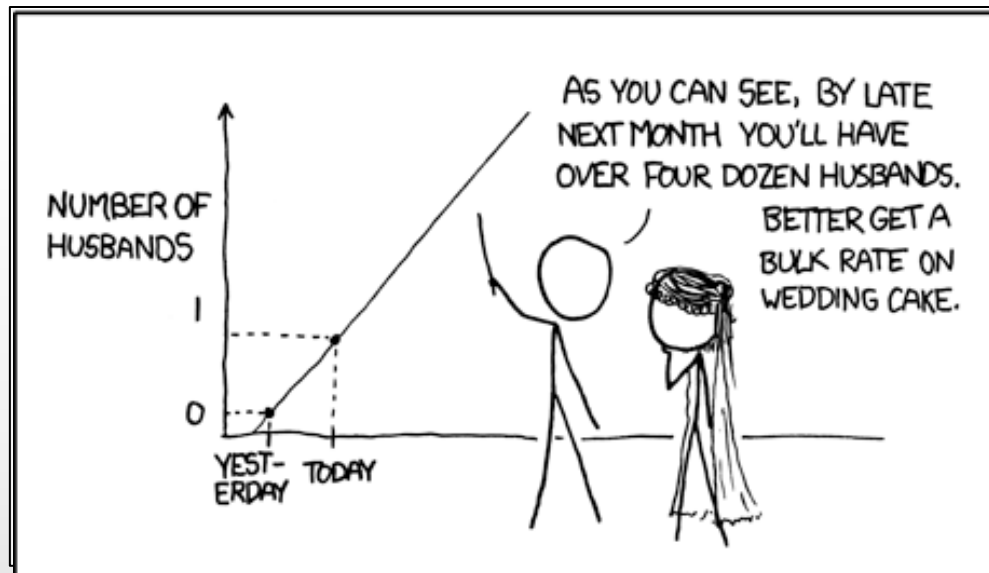


Figure Source: <http://stats.stackexchange.com/questions/423/what-is-your-favorite-data-analysis-cartoon>

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception .
- He devised an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

1. Source: Jure Leskovic, Stanford CS246, Lecture Notes, see <http://cs246.stanford.edu>



## Rhine's Paradox: An Example of Overzealous DM?<sup>1</sup>

29

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP ☹
- What did he conclude?

1. Source: Jure Leskovic, Stanford CS246, Lecture Notes, see <http://cs246.stanford.edu>



- In 2002, the Bush administration put forward a plan to mine all the data it can find, including credit-card receipts, hotel records, and travel data to track terrorist activity.
- This project was coined **Total Information Awareness** and was (officially) killed by the Congress.
- Neglecting privacy concerns, it does raise technical questions about its feasibility and realism of its assumptions!!



# A Simplified Example to Illustrate Potential Issues with TIA<sup>31</sup>

- Problem Definition:
  - Suppose that “evil-doers” periodically gather at a hotel to plot their evil. We want to detect them based on the following assumptions:
    - There are one billion people who might be evil-doers.
    - Everyone goes to a hotel one day in 100.
    - A hotel holds 100 people.
    - We shall examine hotel records for 1000 days.
  - To find “evil-doers”, we shall look for people who, on two different days, were both at the same hotel.
- Question: If there are no “evil-doers” will the data mining detect anything suspicious? (i.e. What are the # of pairs that will look suspicious based on this?)



## What is Data Mining (DM)?

- Statistical Model
- Machine Learning
- Summarization
- Feature Extraction

## Statistical Limits on DM

- Total Information Awareness Act
- Bonferroni's Principle
- Examples on Bonferroni's Principle

## Things Useful To Know

- Importance of Words in Documents
- Hash Functions
- Indexes
- Power Laws





- In several applications, we will want to categorize documents by their topic.
- Typically, topics are identified by finding special words that characterize that topic!!
- Words of Caution:
  - The most frequent words in a document are typically of little value
  - Not all rare words are equally useful
    - “Notwithstanding” versus “chukker”



# TF.IDF – Term Frequency × Inverse Document Frequency 34

- A formal way to measure how concentrated a given word in relatively few documents.

- Calculation:

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

is the no. of occurrences of term  $i$  in document  $j$

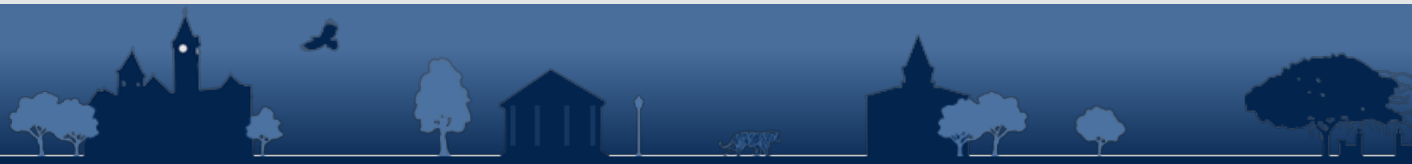
is the no. of occurrences of term  $k$  in document  $j$

$$IDF_i = \log_2 \left( \frac{N}{n_i} \right)$$

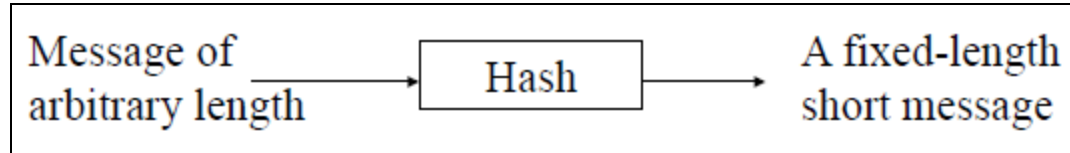
is the total number of documents examined

is the total # documents that term  $i$  appeared in

- Score:  $TF.IDF = TF_{ij} \times IDF_i$
- The terms with the highest score are often the best characterize the topic of the document.



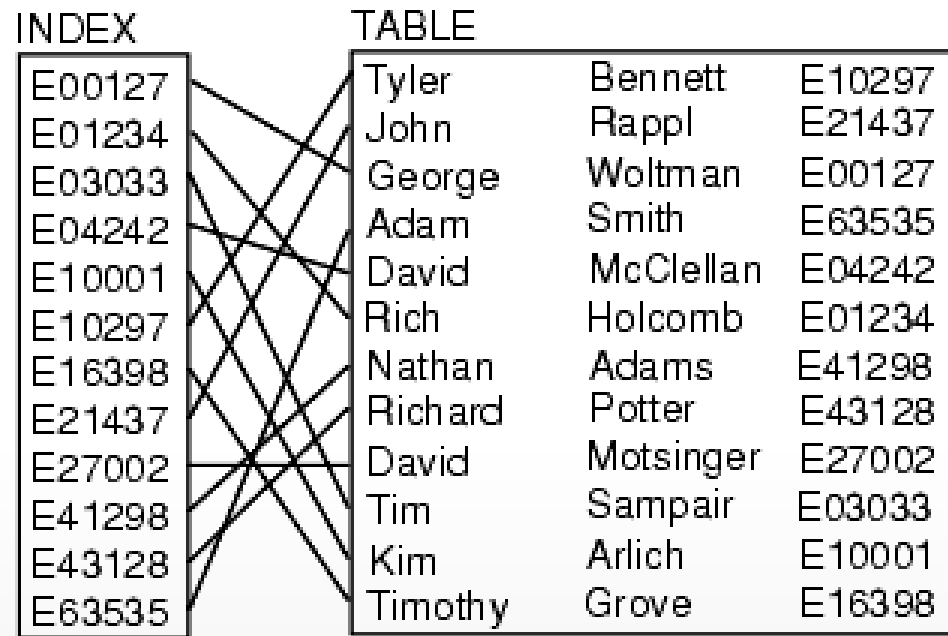
- Essential components of many DM algorithms



- Also known as:
  - Message digest
  - One-way function
  - Hash
- Length of  $H(m)$  much shorter than length of  $m$
- Usually fixed lengths: 128 or 160 bits



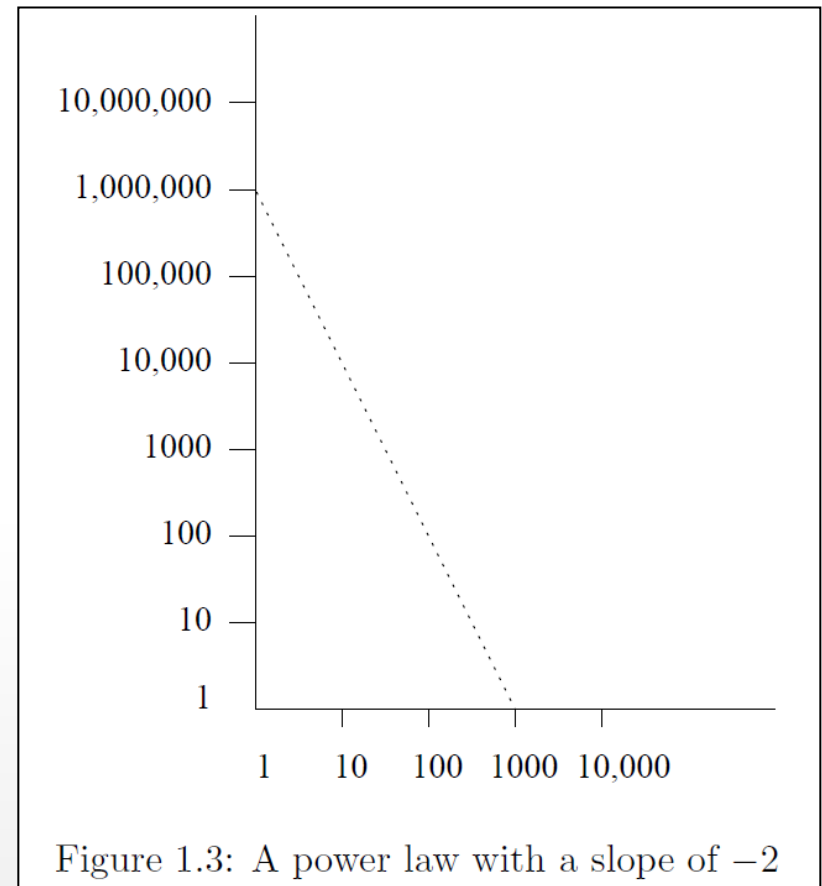
- A data structure that makes it **efficient to retrieve objects** given one or more value of the object.
- The most common situation is one where the objects are records and the index is **one of the fields of that record**.



Source: <http://www.datadirect.com/resources/odbc/using-indexes/index.html>

- Many phenomena can be represented by a linear relationship between the logarithms of the variables.
- General form:

$$\log y = b + a \log x$$
$$y = cx^a$$
- Power law is handy in many applications, see P.14-15 in book



Source: A. Rajaraman, J. Leskovec, J.D. Ullman. (2012). "Mining of Massive Datasets". <http://i.stanford.edu/~ullman/mmds.html>

## HW 0 – No Submission Needed (Except for Gmail address) <sup>38</sup>

- Develop your own gameplan for the course, i.e.,
  - Which assignments are you targeting?
  - Why are you targeting these assignments?
  - What skill sets are you trying to achieve from this class?
  - Do these assignments prepare you to achieve these goals?
- Provide your Gmail to this spreadsheet so we can add to you the class's blog 😊
- Get access to Piazza by signing up here so you can have the PDFs for the class lectures 😊
- Explore the datasets provided in this presentation.
  - Other datasets that you want to share? Highlight in Piazza 😊



# Analytics and Visualization of Big Data

*Fadel M. Megahed*

## ***Lecture 02: Course Intro (Cont.) + Data Mining***



**AUBURN UNIVERSITY**

SAMUEL GINN  
COLLEGE OF ENGINEERING

***Department of Industrial and Systems Engineering***

*Spring 13*