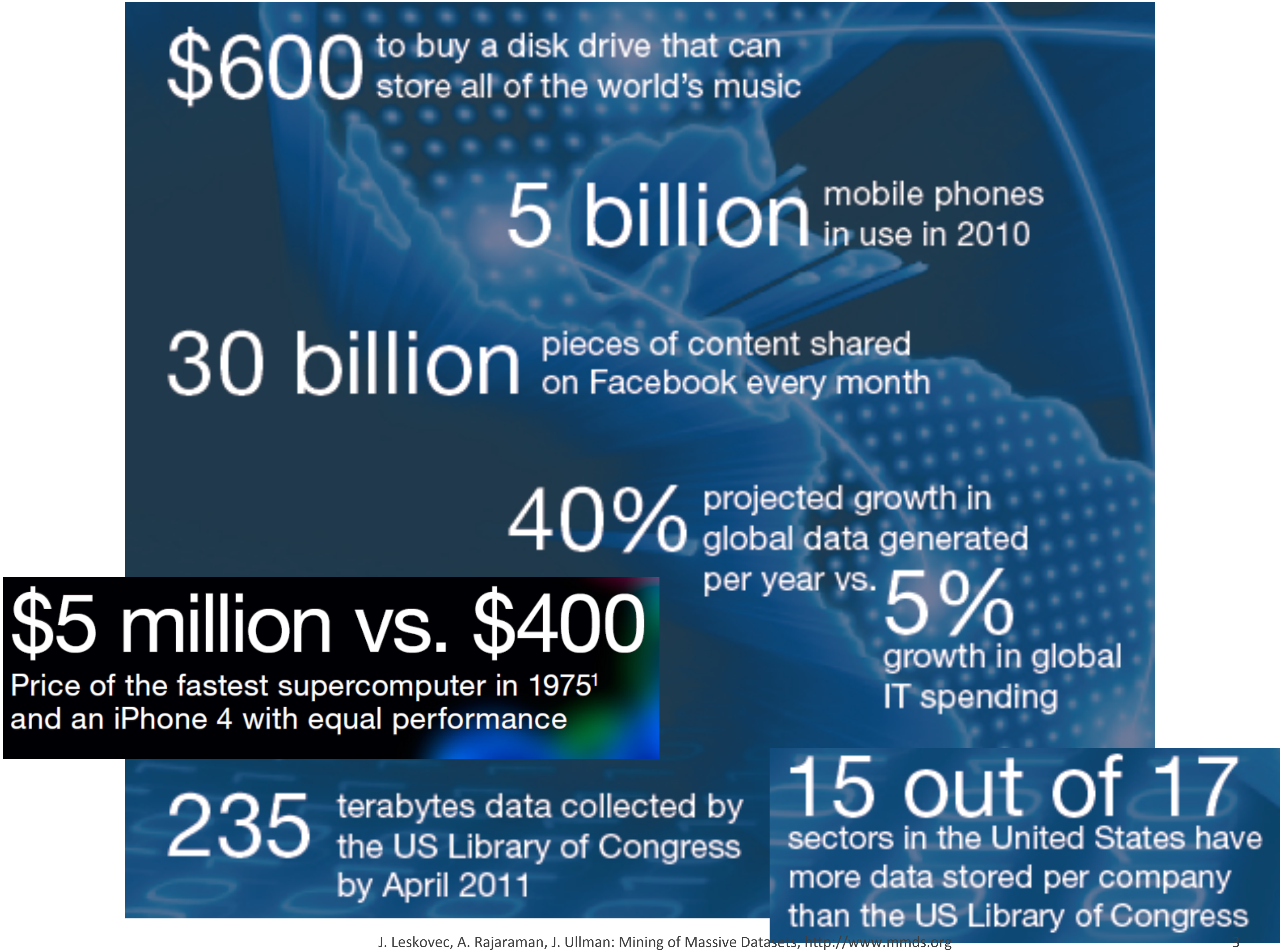
Note to other teachers and users of these slides: We would be delighted if you found this our material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. If you make use of a significant portion of these slides in your own lecture, please include this message, or a link to our web site: http://www.mmds.org

Mining of Massive Datasets: Course Introduction

Mining of Massive Datasets
Jure Leskovec, Anand Rajaraman, Jeff Ullman
Stanford University

http://www.mmds.org

What is Data Mining? Knowledge discovery from data



$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

$5 million vs. $400

Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Data contains value and knowledge

Data Mining

- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYZED ← this class**

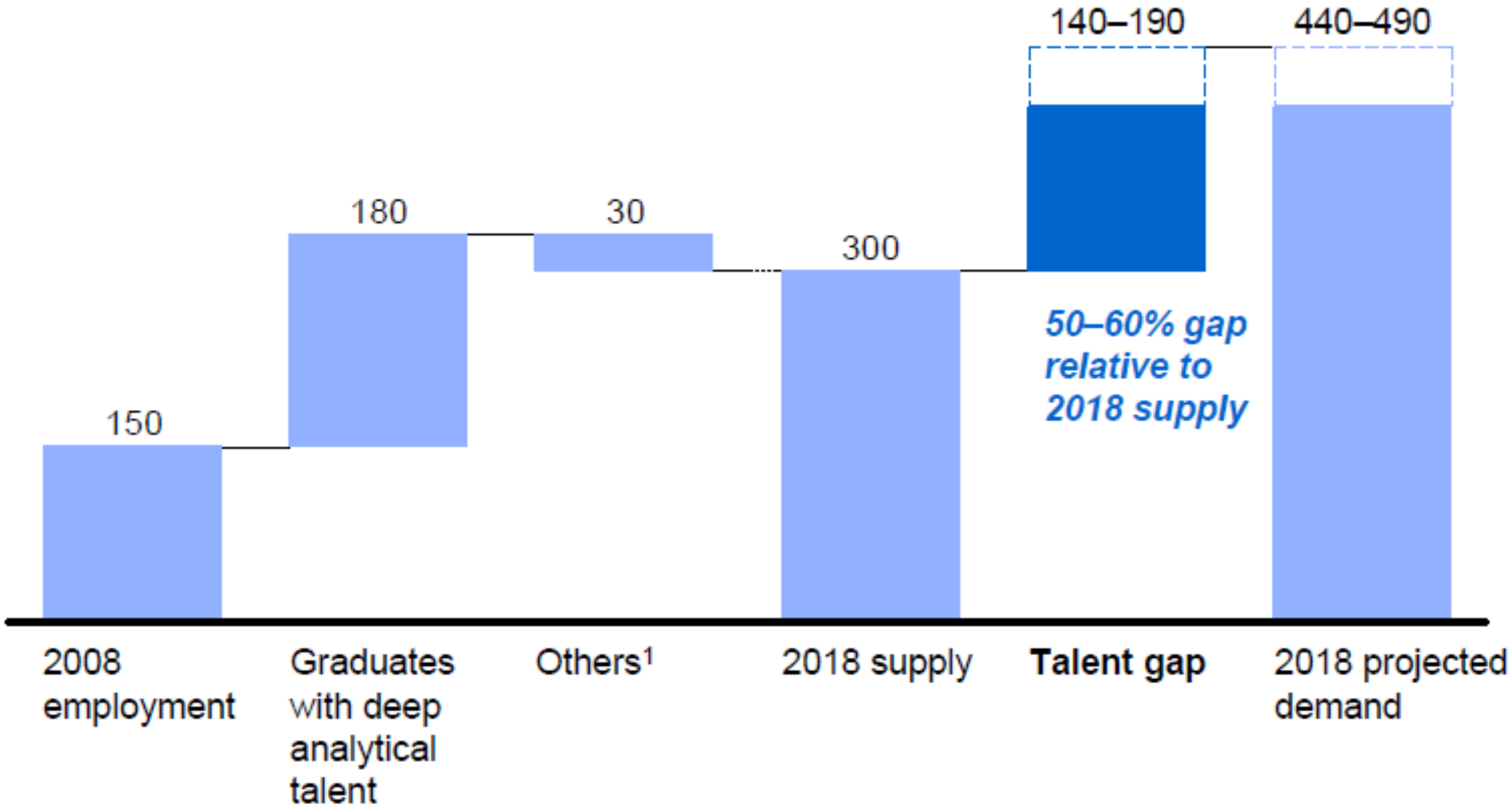
Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



[1] Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

What is Data Mining?

- **Given lots of data**
- **Discover 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

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- Statisticians call it **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 crap

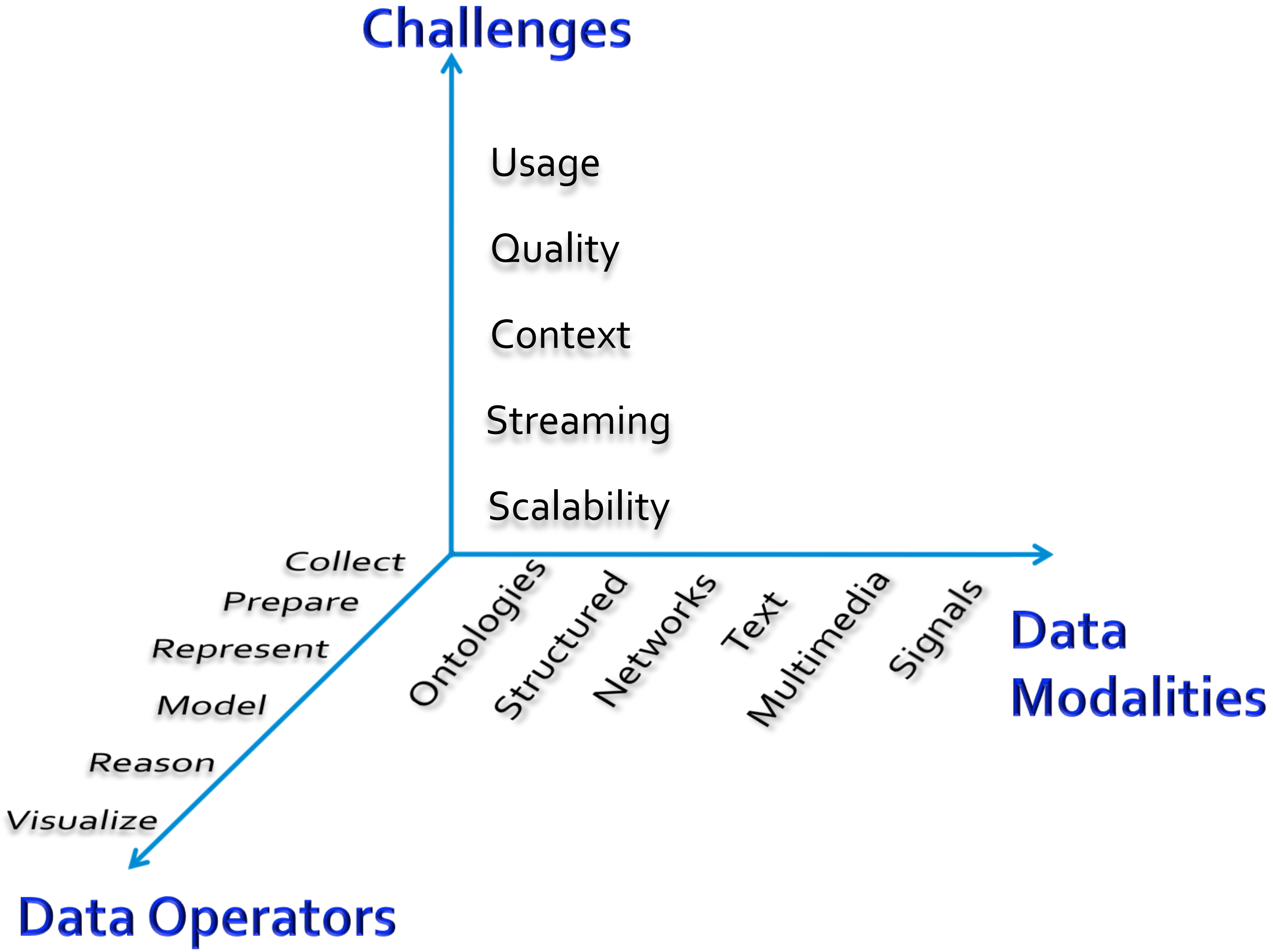
DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

Meaningfulness of Analytic Answers

Example:

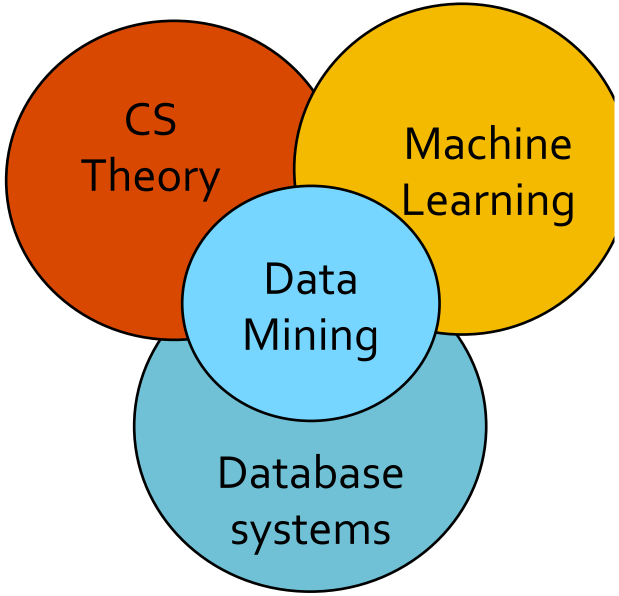
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **Expected number of "suspicious" pairs of people:**
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What matters when dealing with data?



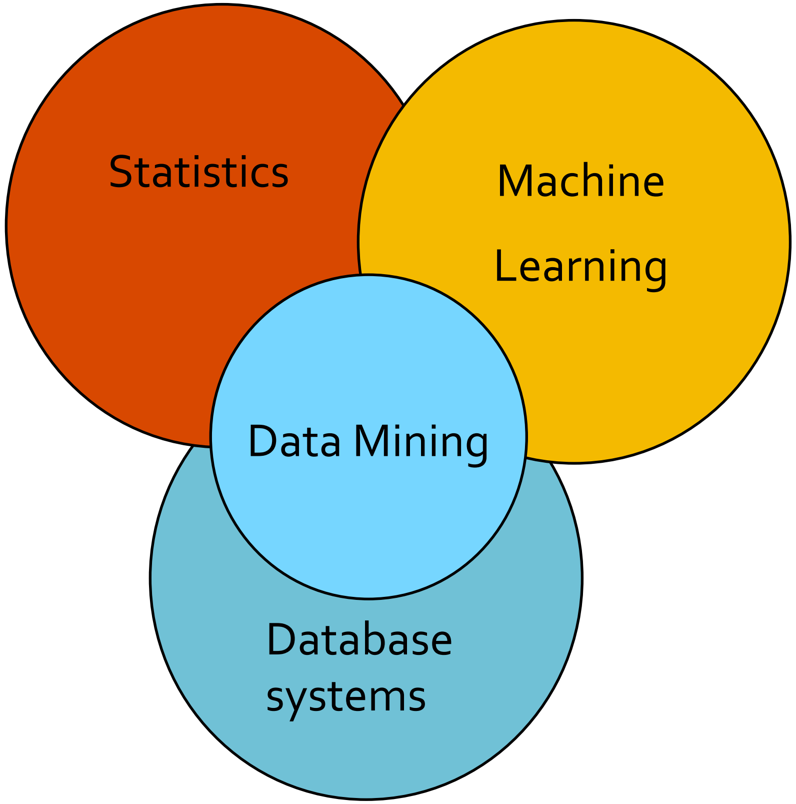
Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model
- **In this class we will do both!**



This Class: CS246

- **This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on**
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing architectures**
  - Automation for handling **large data**



What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
- **We will learn various “tools”:**
  - Linear algebra (SVD, Rec. Sys., Communities)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

How It All Fits Together

High dim. data	Graph data	Infinite data	Machine learning	Apps
Locality sensitive hashing	PageRank, SimRank	Filtering data streams	SVM	Recommender systems
Clustering	Community Detection	Web advertising	Decision Trees	Association Rules
Dimensionality reduction	Spam Detection	Queries on streams	Perceptron, kNN	Duplicate document detection

How do you want that data?

About the Course

2014 CS246 Course Staff

- TAs:
  - We have 9 great TAs!
    - **Sean** Choi (Head TA), **Sumit** Arrawatia, **Justin** Chen, **Dingyi** Li, **Anshul** Mittal, **Rose** Marie Philip, **Robi** Robaszkiewicz, **Le** Yu, **Tongda** Zhang

- Office hours:
  - **Jure:** Wednesdays 9-10am, Gates 418
  - See course website for TA office hours
  - For SCPD students we will use Google Hangout
    - We will post Google Hangout links on Piazza

Course Logistics

- **Course website:** http://cs246.stanford.edu
  - Lecture slides (at least 30min before the lecture)
  - Homeworks, solutions
  - Readings
- **Readings:** Book **Mining of Massive Datasets** with A. Rajaraman and J. Ullman

  Free online: http://www.mmds.org

Logistics: Communication

- **Piazza Q&A website:**
  - https://piazza.com/class#winter2013/cs246
    - Use Piazza for all questions and public communication with the course staff
    - If you don't have @stanford.edu email address, send us your email and we will manually register you to Piazza
- **For e-mailing us, always use:**
  - cs246-win1213-staff@lists.stanford.edu
- **We will post course announcements to Piazza (make sure you check it regularly)**

Auditors are welcome to sit-in & audit the class

Work for the Course

- **(1+)4 longer homeworks: 40%**
  - Theoretical and programming questions
  - **HW0 (Hadoop tutorial) has just been posted**
  - **Assignments take lots of time. Start early!!**
- **How to submit?**
  - **Homework write-up:**
    - **Stanford students:** In class or in Gates submission box
    - **SCPD students:** Submit write-ups via SCPD
    - **Attach the HW cover sheet** (and SCPD routing form)
  - **Upload code:**
    - Put the code for 1 question into 1 file and submit at: http://snap.stanford.edu/submit/

Work for the Course

- **Short weekly quizzes: 20%**
  - Short e-quizzes on Gradiance
  - You have exactly 7 days to complete it
    No late days!
  - First quiz is already online
- **Final exam: 40%**
  - **Friday, March 22** 12:15pm-3:15pm
- **It's going to be fun and hard work. ☺**

Course Calendar

- **Homework schedule:**

Date	Out	In
01/08, Tue	HW0	
01/10, Thu	HW1	
01/15, Tue		HW0
01/24, Thu	HW2	HW1
02/07, Thu	HW3	HW2
02/21, Thu	HW4	HW3
03/07, Thu		HW4

- **2 late “days” (late periods) for HWs for the quarter:**
  - 1 late day expires at the start of next class
  - **You can use max 1 late day per assignment**

Prerequisites

- Algorithms (CS161)
  - Dynamic programming, basic data structures
- Basic probability (CS109 or Stat116)
  - Moments, typical distributions, MLE, …
- Programming (CS107 or CS145)
  - Your choice, but C++/Java will be very useful

- **We provide some background, but the class will be fast paced**

Recitation Sessions

- **3 recitation sessions:**
  - **Hadoop:** Thurs. 1/10, 5:15-6:30pm
    - We prepared a virtual machine with Hadoop preinstalled
    - **HW0** helps you write your first Hadoop program
  - **Review of probability&stats:** 1/17, 5:15-6:30pm
  - **Review of linear algebra:** 1/18, 5:15-6:30pm
  - All sessions will be held in Thornton 102, Thornton Center (Terman Annex)
  - **Sessions will be video recorded!**

What’s after the class

- **InfoSeminar (CS545):**
  - http://i.stanford.edu/infoseminar
  - Great industrial & academic speakers
  - Topics include data mining and large scale data processing
- **CS341:** Project in Data Mining (Spring 2013)
  - Research project on big data
  - Groups of 3 students
  - We provide interesting data, computing resources (Amazon EC2) and mentoring
- **We have big-data RA positions open!**
  - **I will post details on Piazza**

3 To-do items

- **3 To-do items for you:**
  - **Register to Piazza**
  - **Complete HW0: Hadoop tutorial**
    - HW0 should take your about 1 hour to complete (Note this is a "toy" homework to get you started. Real homeworks will be much more challenging and longer)
  - **Register to Gradiance and complete the first quiz**
    - **Use your SUNet ID to register!** (so we can match grading records)
    - You have 7 days (sharp!) to do so
    - Quizzes typically take several hours
- **Additional details/instructions at** **http://cs246.stanford.edu**