Data Mining and Analysis

Welcome and Administrivia

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

What is “Data Mining and Analysis?

What is “Data Mining and Analysis?

Collecting info and making use of it

Pattern extraction

AI

ML

Learn techniques to gather data and visualize it

Statistics

Web scraping … collect, extract, do something via interface

Make my data useful

Filter out crap

Science!!! Hypothesis — verify or reject it based on my data/model

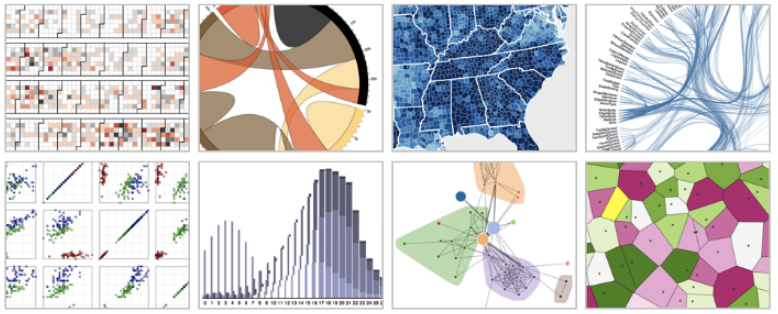
Anything using hadoop/spark

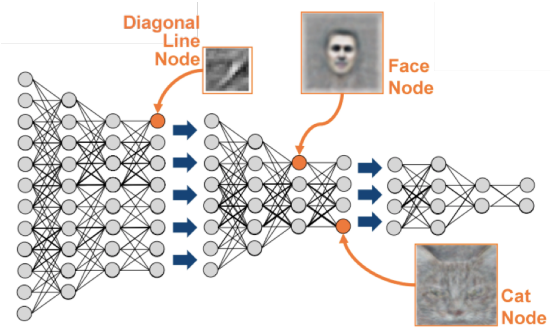
Data modeling and viz

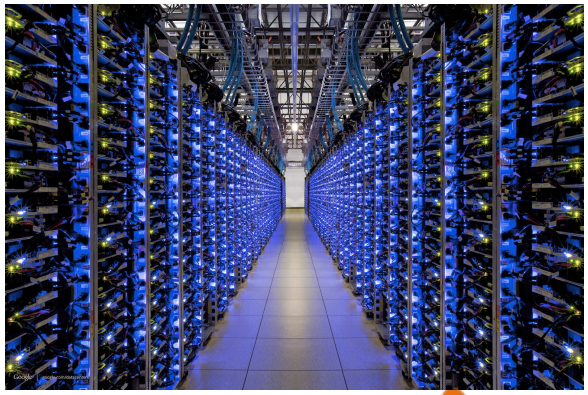
Data science?

…

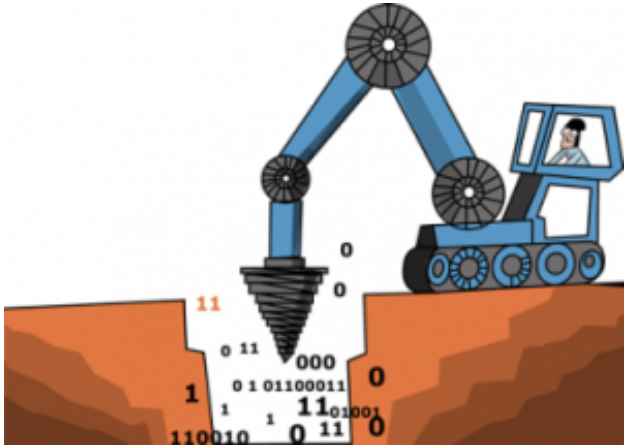
Data Mining and Analysis







Spark



Okay, but why?

Make decisions

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

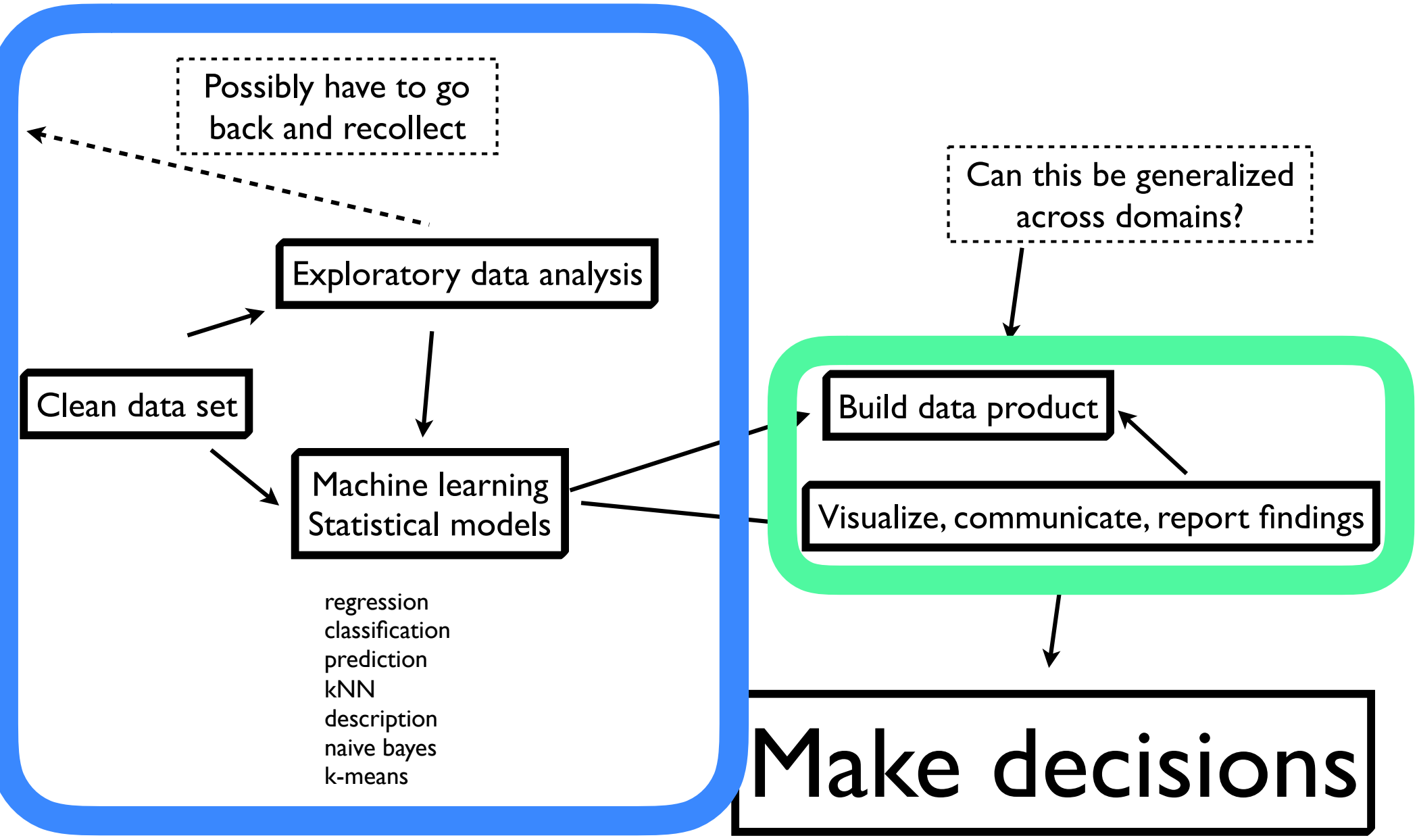
Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data set

"clean" table

http://columbiadatascience.com/2012/09/24/reflections-after-jakes-lecture/



Our main emphasis

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

email

medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data set

"clean" table

http://columbiadatascience.com/2012/09/24/reflections-after-jakes-lecture/

Goal of 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

Who are we?

Instructor: Prof. James Caverlee

TA: Ziwei Zhu

http://courses.cse.tamu.edu/caverlee/csce676/

Grading

5% Participation

10% Quiz

30% Final

25% Homework

30% Project

Participation (5%)

Roughly — come to class and participate on Piazza, and you will be fine

Requirements:

- Initiate at least one post on Piazza
- Participate in at least three threads on Piazza

Last day for your posts to count:

November 27

Quiz (10%)

In-class on October 7 (Monday)

50 minutes

You can bring one cheatsheet

Cheatsheet = 8.5” x 11” standard sheet of paper with anything on it, front and back

Final Exam (30%)

In our regular classroom

Monday, December 9 from 8-10am

120 minutes

Comprehensive

You can bring two cheatsheets

Homeworks (25%)

(1%) HW0: setup Python, get ready, submit via elearning

(6%) HW1: Data mining (single machine)

(6%) HW2: Spark + AWS + DM

(6%) HW3: Data Visualization++

(6%) HW4: Advanced data mining

Homework Late Days

Due by 11:59pm on the due date

You get FIVE late days total

But can only use up to THREE at a time

Late day = indivisible 24-hour unit

> e.g., if due date is 11:59pm on Monday, and you submit at 12:01am Tuesday = 1 late day

Once you are out of late days = 0

Homework Collaboration Policy

Homeworks are individual = you should write your own code, by yourself

But, we want you to talk amongst yourselves about approaches/methods

- Example: sit in a group with no laptops, just talking = totally fine
- Example: sit next to each other while you code = BAD NEWS

You must acknowledge all help in your homework

Concerned? Talk to me or the TA

Regrade Policy

Once you receive your graded assignment (e.g., hw, quiz), you have SEVEN days to request a regrade

After seven days = no regrades

You must give us a written explanation of what the issue is

We will re-grade the entire assignment

Project (30%)

Data Mining for Social Good

3-4 members per team

Proposal: October 13 on Piazza

Pitch: In-class, tell us about your project

Peer feedback: Post feedback to Piazza

Showcase: In-class December 2/4

Website

Communication

Piazza!!!

- Check often, post often

Email

- Use sparingly
- Put CSCE676 in the subject line
- We will respond, but may be slow

Office hours

- Mine: TBD
- Ziwei: 3-4 on Tuesday/Thursday in 408A

Readings

See schedule on the webpage

Class mechanics

Start at 9:10 sharp

End at 10:00 sharp

unless …

Open issues

Sign up for Piazza:

https://piazza.com/tamu/fall2019/csce676

Stay on top of readings

HW0 out later this week

Slides posted to Piazza

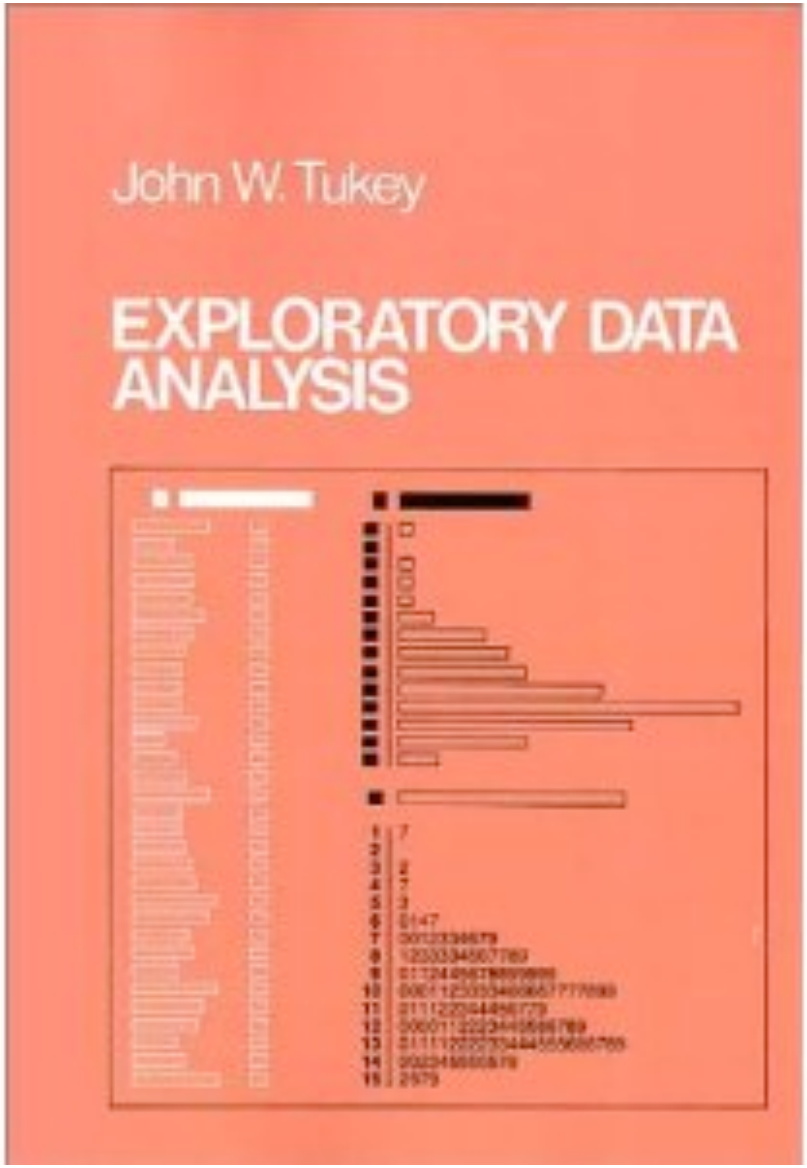
Quick history lesson …

- Before 1600: Empirical science
- 1600-1950s: Theoretical science
  - Each discipline has grown a theoretical component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s: Computational science
  - Over the last 50 years, most disciplines have grown a third, computational branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now: Data science
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
  - Data mining is a major new challenge!
- Jim Gray and Alex Szalay, The World Wide Telescope: An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002

“Business intelligence”

- 1958: “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal”
- 1989: “concepts and methods to improve business decision making by using fact-based support systems”
- http://en.wikipedia.org/wiki/Business_intelligence

1977



Fayyad (1996)

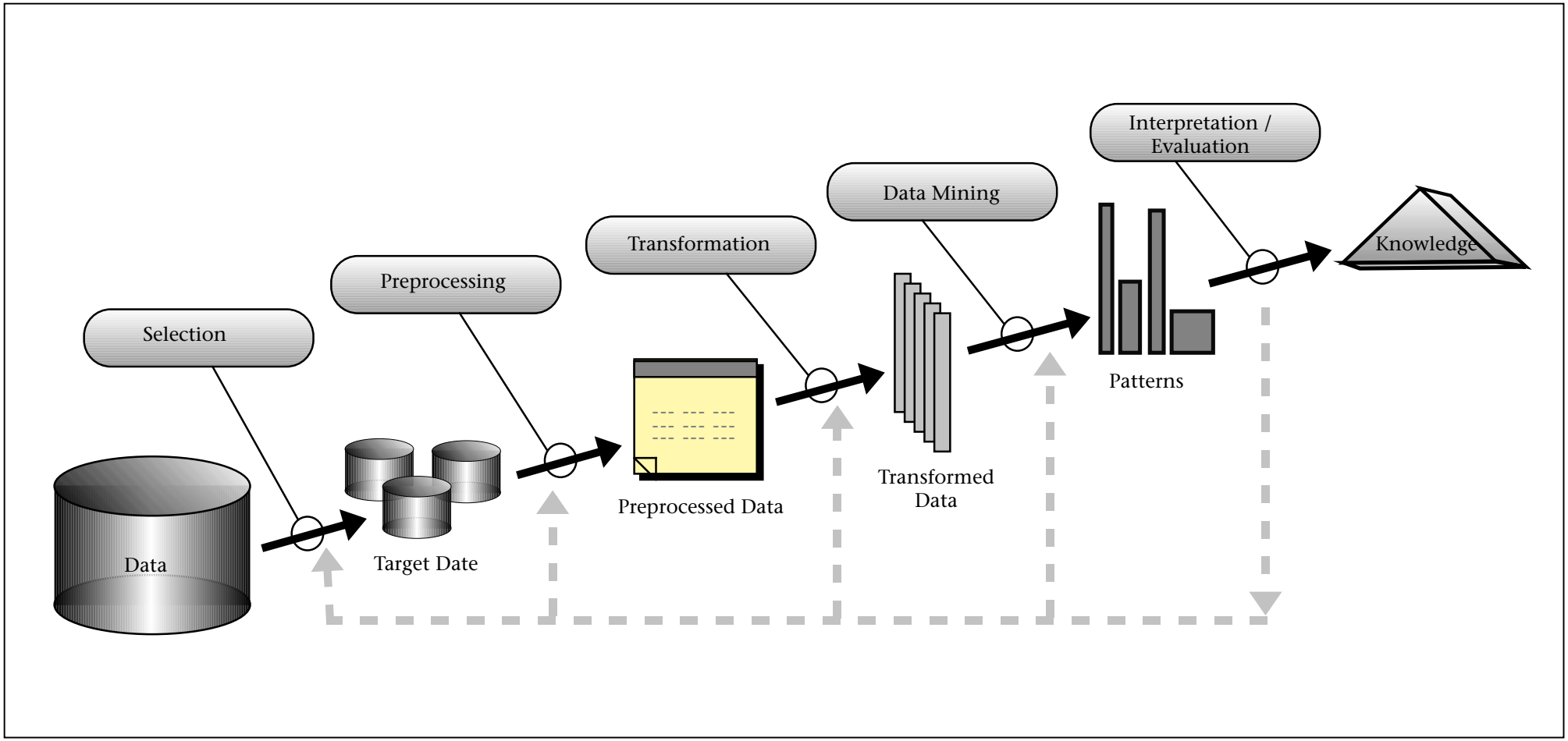


Figure 1. An Overview of the Steps That Compose the KDD Process.

The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
McKinsey Quarterly 2009
http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286

The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
McKinsey Quarterly 2009
http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286

Goal of Data Mining

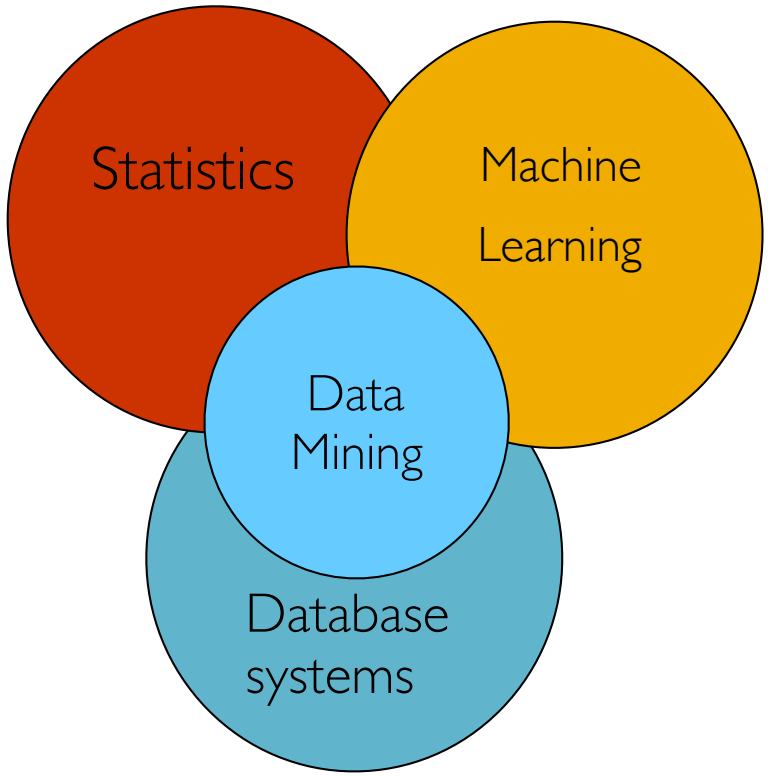
Given lots of data

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

CSCE 676

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



What we will 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/Spark
  - Streams and online algorithms
  - Single machine in-memory

What we will learn?

- We will learn to solve real-world problems:
  - Community detection
  - Market basket analysis
  - Outlier/anomaly detection
  - Duplicate document detection
- We will learn various “tools”:
  - Linear algebra (SVD, Communities)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH)

Be Careful

Meaningfulness of Answers

A big data mining risk 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 crap

Example: Rhine Paradox

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!

Example: Rhine Paradox

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?

He concluded that you shouldn't tell people they have ESP; it causes them to lose it!