Exploratory Analysis and mlib

This Lecture

Exploratory Data Analysis

Some Important Distributions

Spark mllib Machine Learning Library

Descriptive vs. Inferential Statistics

- Descriptive:
  - E.g., Median – describes data but can't be generalized beyond that
  - We will talk about Exploratory Data Analysis in this lecture

- Inferential:
  - E.g., t-test – enables inferences about population beyond our data
  - Techniques leveraged for Machine Learning and Prediction

Examples of Business Questions

- Simple (descriptive) Stats
  - » “Who are the most profitable customers?”
- Hypothesis Testing
  - » “Is there a difference in value to the company of these customers?”
- Segmentation/Classification
  - » What are the common characteristics of these customers?
- Prediction
  - » Will this new customer become a profitable customer?
  - » If so, how profitable?

adapted from Provost and Fawcett, “Data Science for Business”

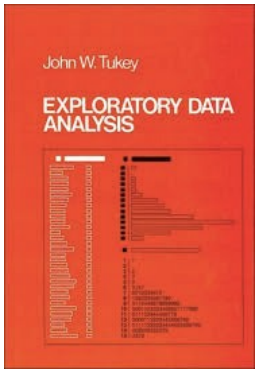
Applying Techniques

- Most business questions are causal
  - » What would happen if I show this ad?
- Easier to ask correlational questions
  - » What happened in this past when I showed this ad?
- Supervised Learning: Classification and Regression
- Unsupervised Learning: Clustering and Dimension reduction
- Note: UL often used inside a larger SL problem
  - » E.g., auto-encoders for image recognition neural nets

Learning Techniques

- Supervised Learning:
  - kNN (k Nearest Neighbors)
  - Naive Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- Unsupervised Learning:
  - Clustering
  - Factor Analysis
  - Latent Dirichlet Allocation

Exploratory Data Analysis (1977)



- Based on insights developed at Bell Labs in 1960's
- Techniques for visualizing and summarizing data
- What can the data tell us? (vs "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5-Number summary:
  - Extremes (min and max)
  - Median & Quartiles q1, q3
  - More robust to skewed and long-tailed distributions

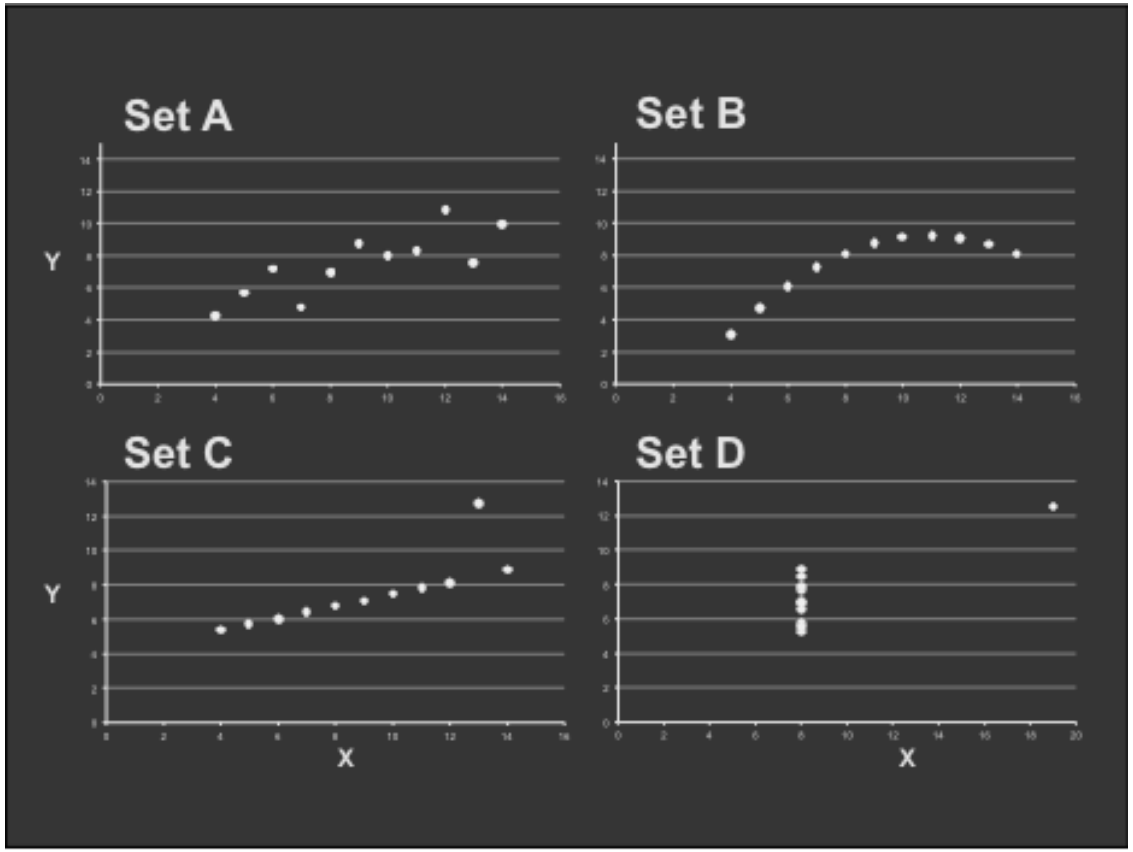
The Trouble with Summary Stats

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

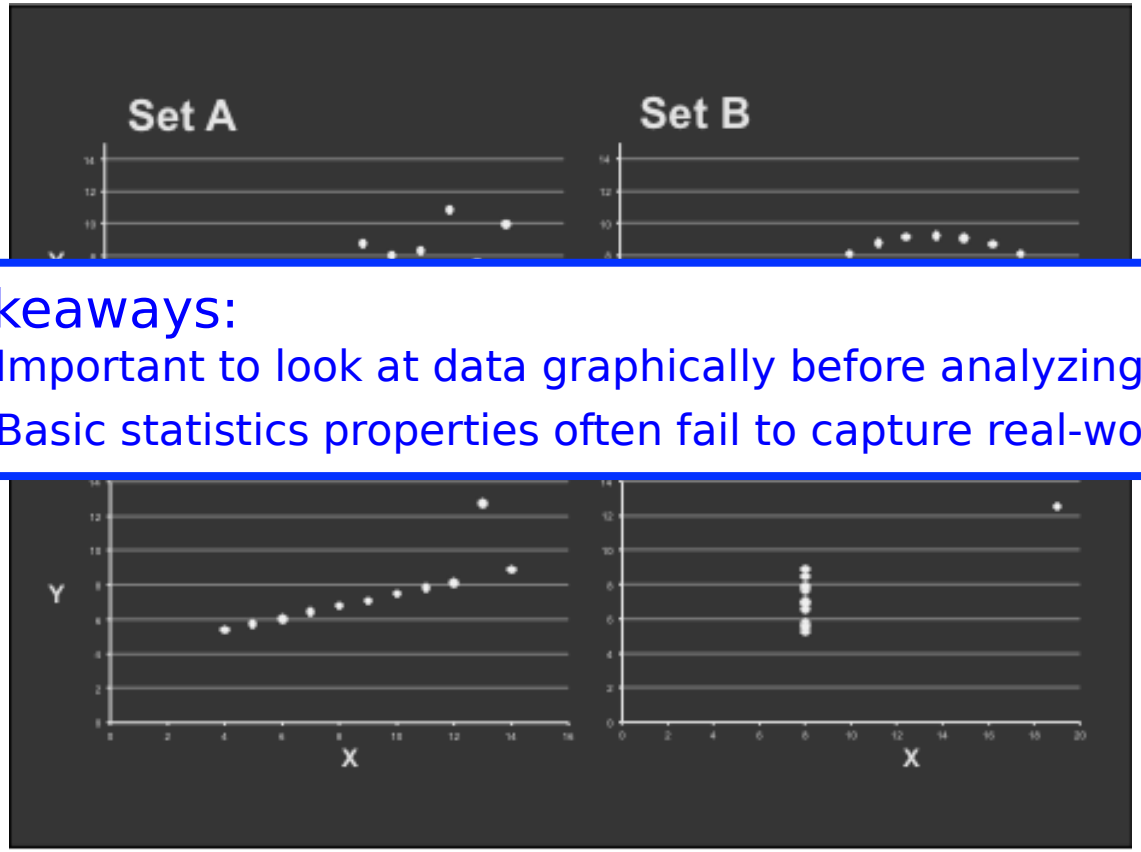
Anscombe's Quartet 1973

Property in each set	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.122
Linear Regression	$y = 3 + 0.5x$

Looking at The Data



Looking at The Data



Takeaways:

- Important to look at data graphically before analyzing it
- Basic statistics properties often fail to capture real-world complexities

Data Presentation

- Data Art – Visualizing Friendships

https://www.facebook.com/note.php?note_id=469716398919

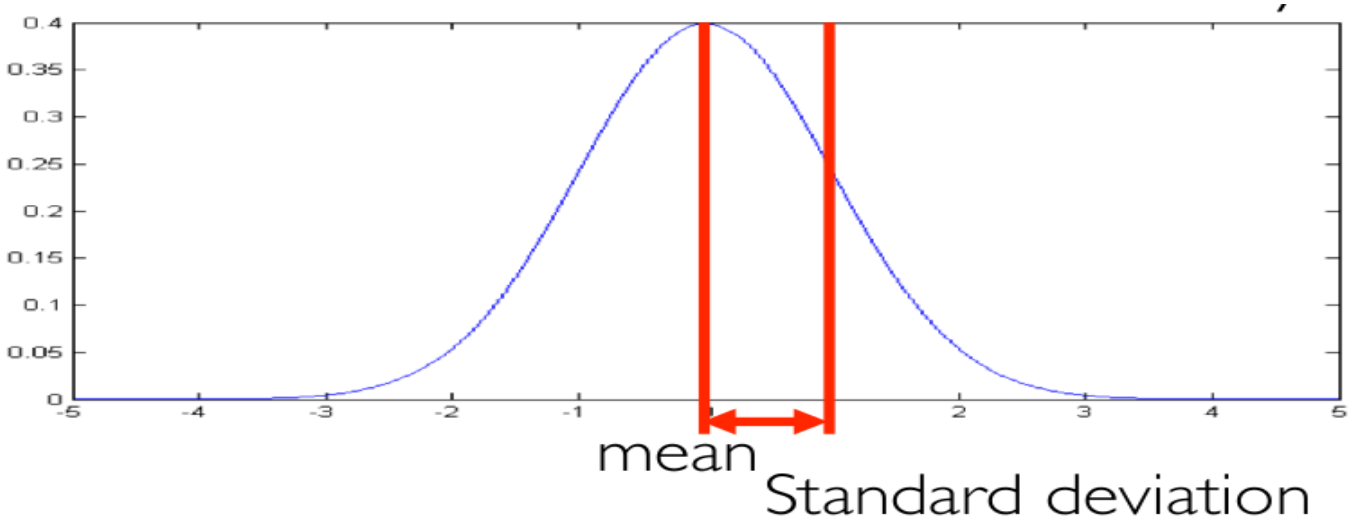
The “R” Language

- Evolution of the “S” language developed at Bell labs for EDA
- Idea: allow interactive exploration and visualization of data
- Preferred language for statisticians, used by many data scientists
- Features:
  - » The most comprehensive collection of statistical models and distributions
  - » CRAN: large resource of open source statistical models

Jeff Hammerbacher 2012 course at UC Berkeley

Normal Distributions, Mean, Variance

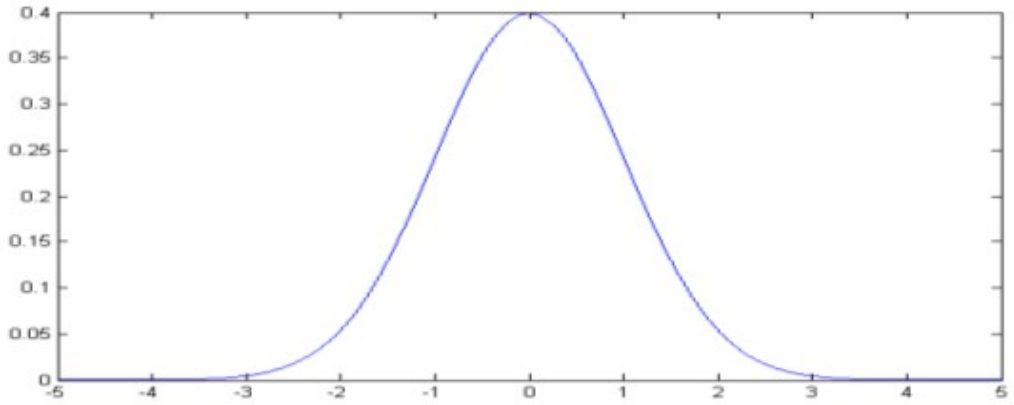
- The mean of a set of values is the average of the values
- Variance is a measure of the width of a distribution
- The standard deviation is the square root of variance
- A normal distribution is characterized by mean and variance



Central Limit Theorem

- The distribution of sum (or mean) of n identically-distributed random variables X i approaches a normal distribution as $n \to \infty$

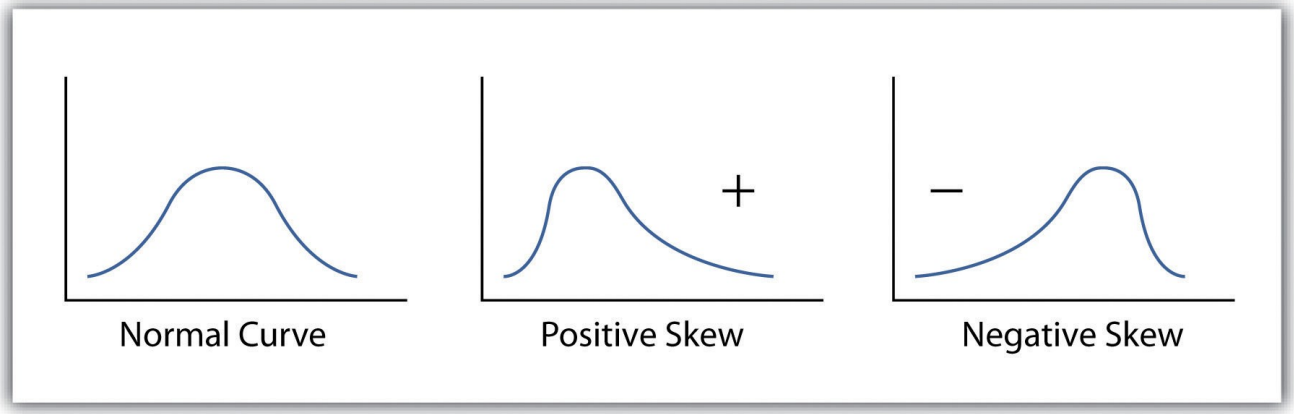
- Permite averiguar la probabilidad de que la media de una muestra concreta esté en un cierto intervalo.
- Permite calcular la probabilidad de que la suma de los elementos de una muestra esté, a priori, en un cierto intervalo.
- Inferir la media de la población a partir de una muestra.



Correcting Distributions

- Many statistical tools (mean, variance, t-test, ANOVA) assume data are normally distributed

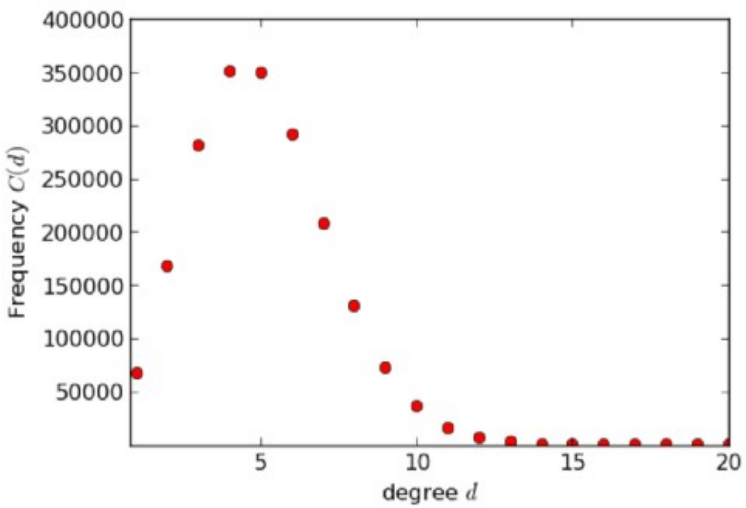
- Very often this is not true – examine the histogram

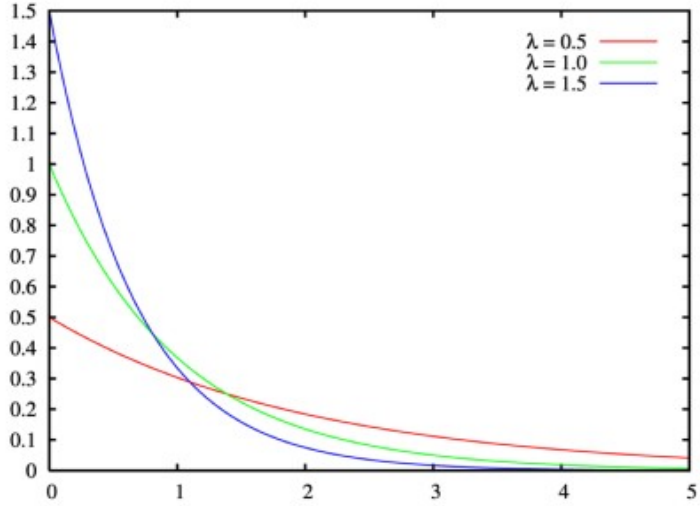


Other Important Distributions

- Poisson: distribution of counts that occur at a certain “rate”
  - » Observed frequency of a given term in a corpus
  - » Number of visits to web site in a fixed time interval
  - » Number of web site clicks in an hour

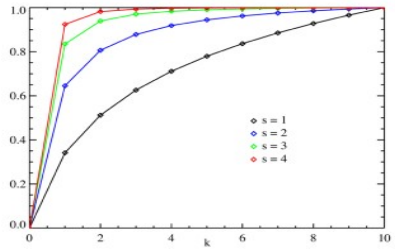
- Exponential: interval between two such events

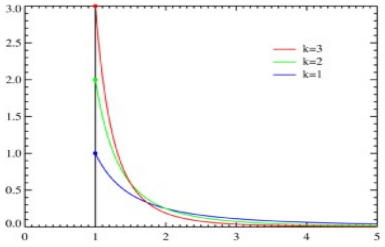


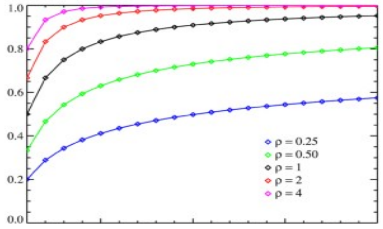


Other Important Distributions

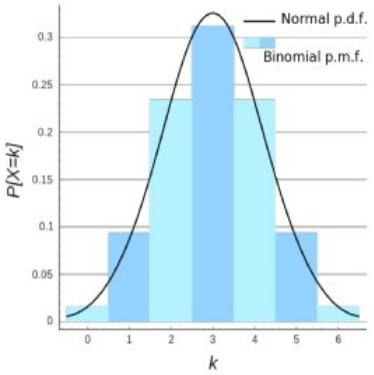
- Zipf/Pareto/Yule distributions:
  - » Govern frequencies of different terms in a document, or web site visits







- Binomial/Multinomial:
  - » Number of counts of events
  - » Example: 6 die tosses out of n trials



- Understand your data's distribution before applying any model

Rhine Paradox*

- Joseph Rhine was a parapsychologist in the 1950's
  - » Experiment: subjects guess whether 10 hidden cards were red or blue
- He found that about 1 person in 1,000 had *Perception Extra Sensory*!
  - » They could correctly guess the color of all 10 cards

*Example from Jeff Ullman/Anand Rajaraman

Rhine Paradox

- Called back “psychic” subjects and had them rep
  - » They all failed

- Concluded that act of telling psychics that they h psychic abilities causes them to lose it…(!)

- *Q: What’s wrong with his conclusion?*

Rhine’s Error

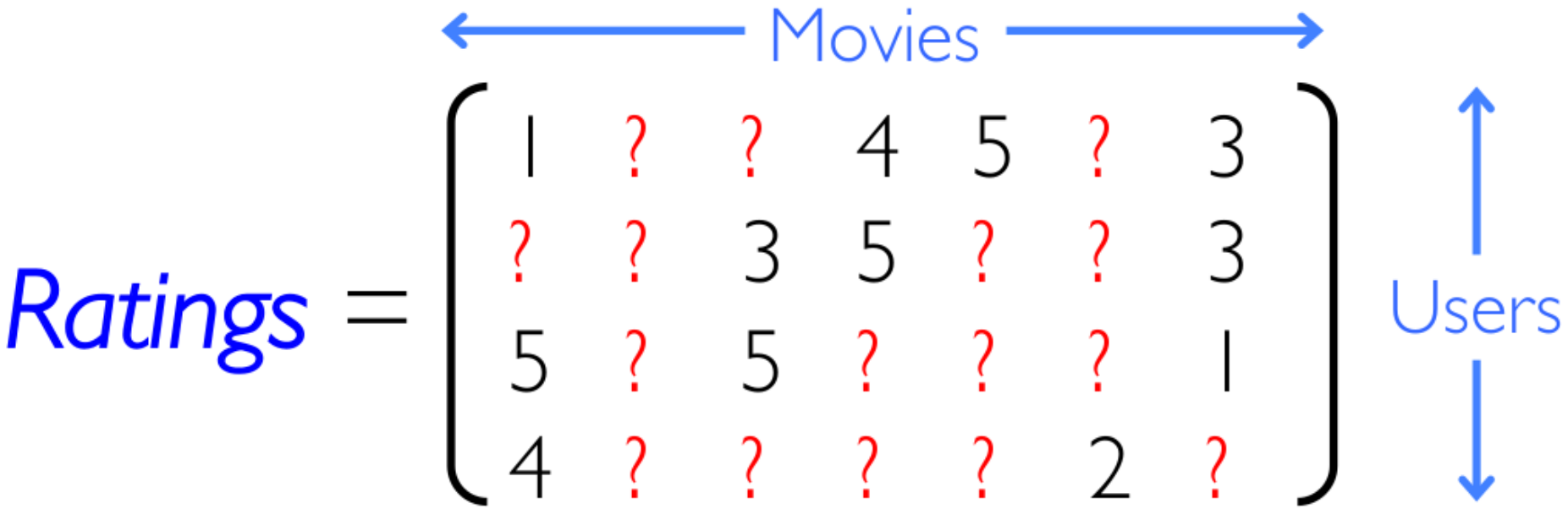
- What’s wrong with his conclusion?
- 2^{10} = 1,024 combinations of red and blue of length 10
- 0.98 probability at least 1subject in 1,000 will guess correctly

Spark's Machine Learning Toolkit

- mllib: scalable, distributed machine learning library
  - » Scikit-learn like ML toolkit, Interoperates with NumPy
- Classification:
  - » SVM, Logistic Regression, Decision Trees, Naive Bayes, …
- Regression: Linear, Lasso, Ridge, …
- Miscellaneous:
  - » Alternating Least Squares, K-Means, SVD
  - » Optimization primitives (SGD, L-BGFS)
  - » …

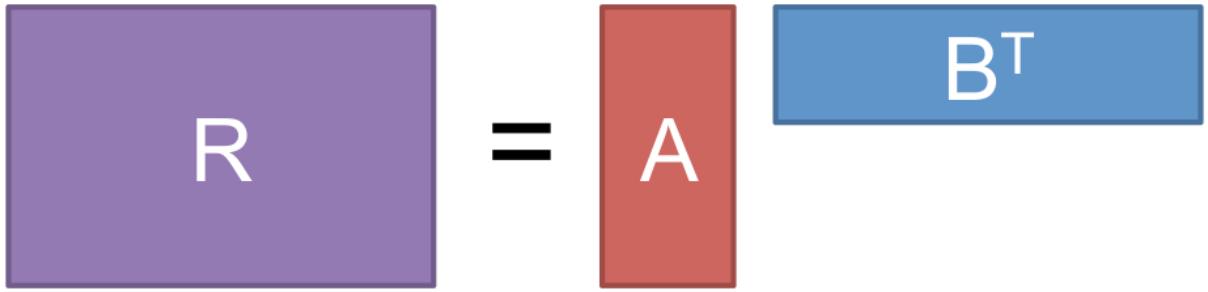
Lab: Collaborative Filtering

Goal: predict users' movie ratings based on past ratings of other movies



Model and Algorithm

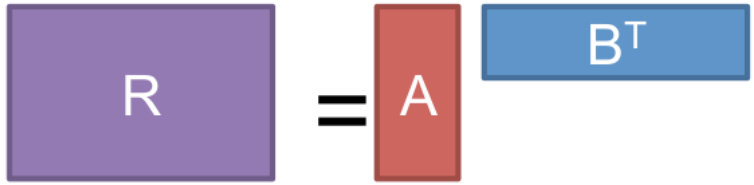
- Model *Ratings* as product of *User (A)* and *Movie Feature (B)* matrices of size U×K and M×K



- *K*: rank
- Learn *K* factors for each user
- Learn *K* factors for each movie

Model and Algorithm

- Model *Ratings* as product of *User (A)* and *Movie Feature (B)* matrices of size U×K and M×K



- Alternating Least Squares (ALS)
  - » Start with random A and B vectors
  - » Optimize user vectors (A) based on movies
  - » Optimize movie vectors (B) based on users
  - » Repeat until converged