Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 1: Overview

Jan-Willem van de Meent

Who are we?

Instructor

Jan-Willem van de Meent

Email: j.vandemeent@northeastern.edu
Phone: +1 617 373-7696
Office Hours: 478 WVH, Wed 1.30pm - 2.30pm

Teaching Assistants

Yuan Zhong

E-mail: yzhong@ccs.neu.edu
Office Hours: WVH 462, Wed 3pm - 5pm

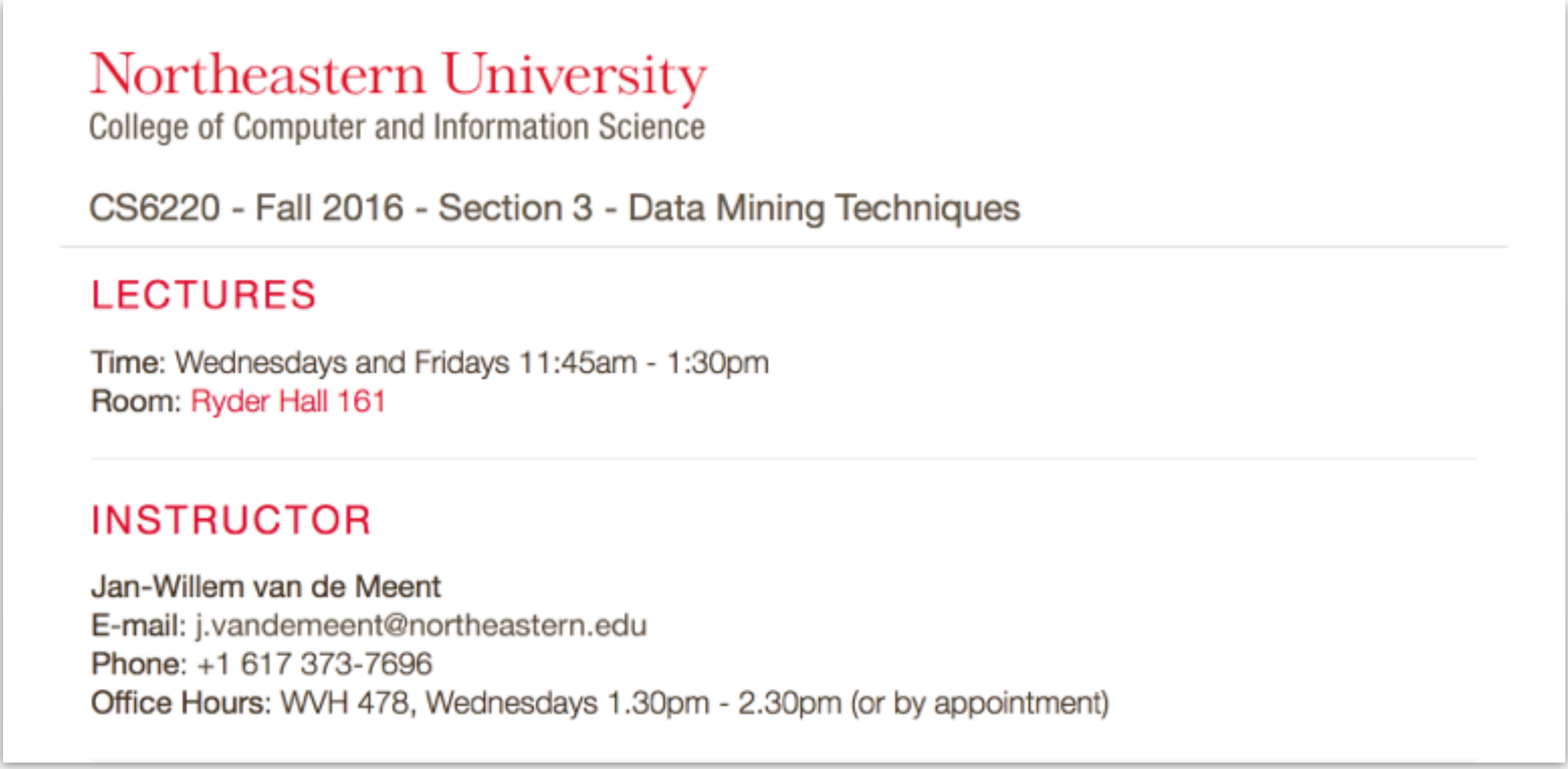
Kamlendra Kumar

E-mail: kumark@zimbra.ccs.neu.edu
Office Hours: WVH 462, Fri 3pm - 5pm

Who are you?

Syllabus

http://www.ccs.neu.edu/course/cs6220f16/sec3/



Northeastern University
College of Computer and Information Science

CS6220 - Fall 2016 - Section 3 - Data Mining Techniques

LECTURES

Time: Wednesdays and Fridays 11:45am - 1:30pm
Room: Ryder Hall 161

INSTRUCTOR

Jan-Willem van de Meent
E-mail: j.vandemeent@northeastern.edu
Phone: +1 617 373-7696
Office Hours: WVH 478, Wednesdays 1.30pm - 2.30pm (or by appointment)

Course Objectives

1. Lectures: <u>Understand</u> data mining methods

- Mathematical/algorithmic definitions
- When should each method be used?
- What are some limitations of each method?

2. Homework Problems: <u>Use</u> data mining methods

- Implement methods
- Use methods in existing libraries
- Visualize results, evaluate effectiveness

Homework Problems

- 4 or (more likely) 5 problem sets
- 30% - 40% of grade (depends on type of project)
- Can use any language (within reason)
- **Discussion is encouraged, but submissions must be completed individually**
  (absolutely **no** sharing of code)
- Submission via *zip* file by **11.59pm** on day of deadline
  (no late submissions)
- Please follow *submission guidelines* on website
  (TA's have authority to deduct points)

Project

Vote next week

1. *Freeform*: Develop your own project proposals
   - 30% of grade (homework 30%)
   - Present proposals after midterm
   - Peer-review reports
2. *Predefined*: Same project for whole class
   - 20% of grade (homework 40%)
   - More like a “super-homework”
   - Teaching assistants and instructors

Participation

1. Attend the Lectures
2. Ask questions!
3. Help Others

Self-evaluation

For Homework Problems

- Indicate time spent
- What was easy / hard?
- What did you learn?

After Midterm and Final Exams

- What was your favorite topic?
- What parts were easier / more difficult to follow?
- List 3 students that contributed to your understanding

Grading

Freeform Project

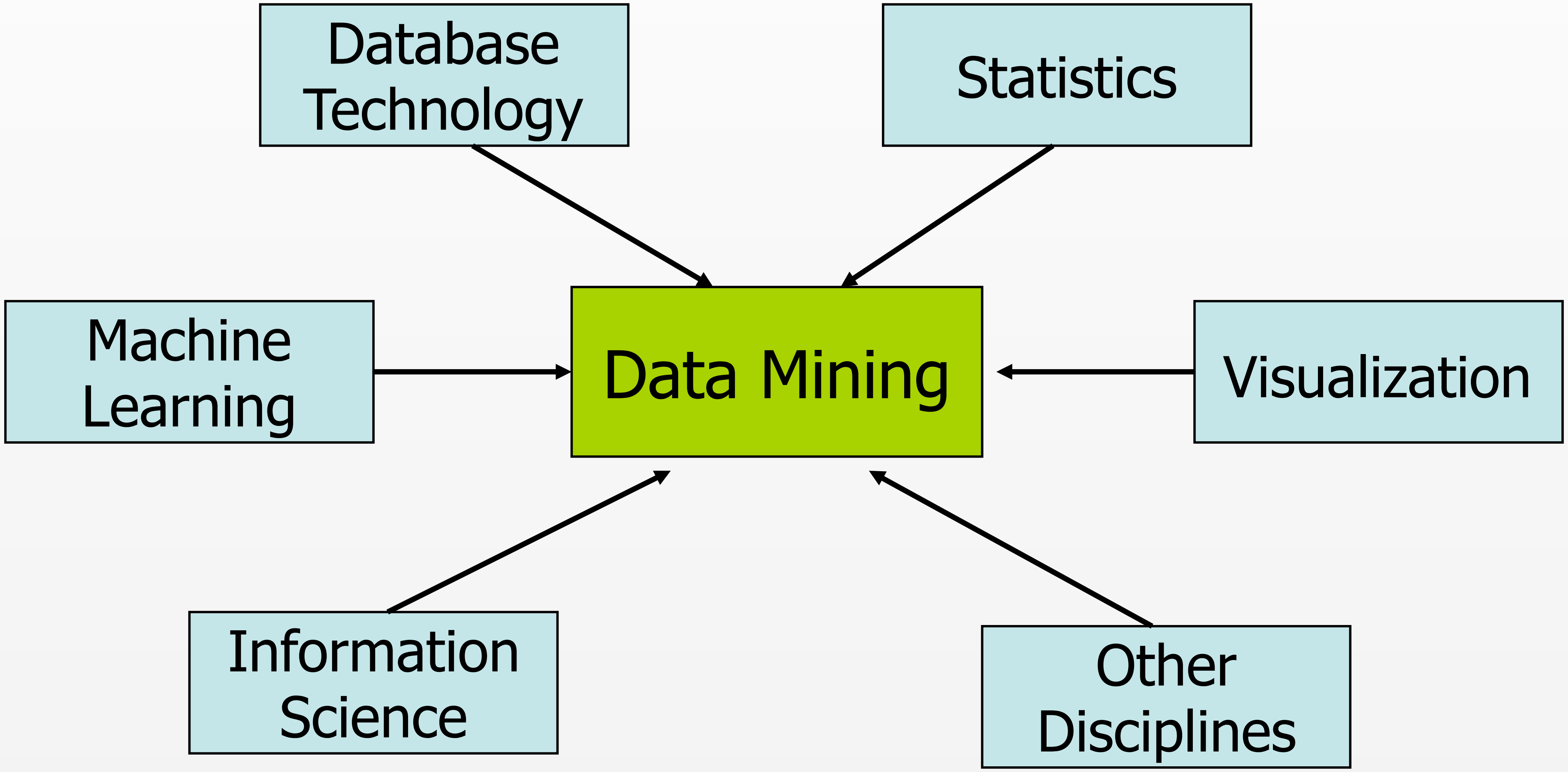
- Homework: **30%**
- Midterm: 20%
- Final: 20%
- Project: **30%**
- Participation (bonus): 10%

Predefined Project

- Homework: **40%**
- Midterm: 20%
- Final: 20%
- Project: **20%**
- Participation (bonus): 10%

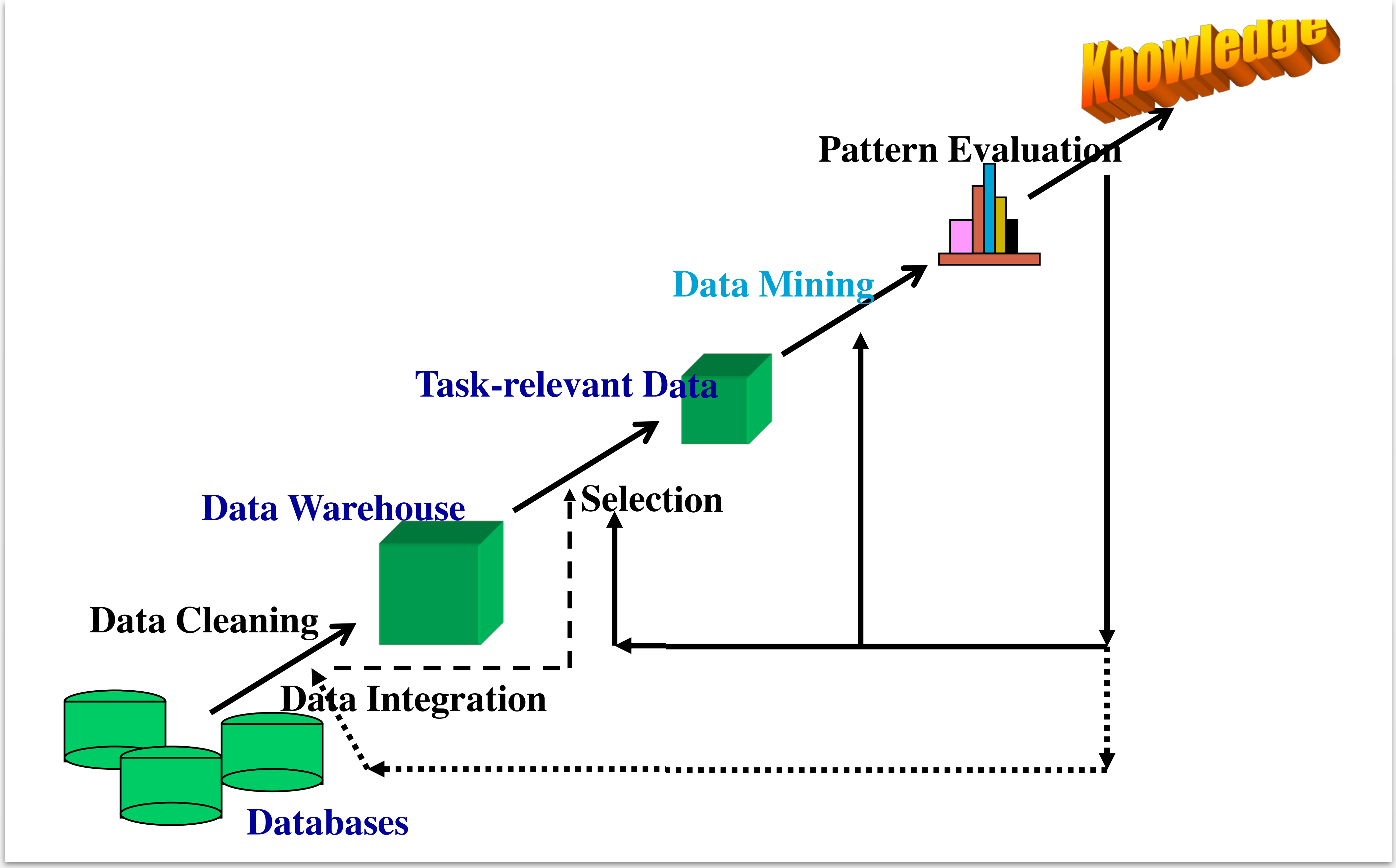
What is Data Mining?

Intersection of Disciplines



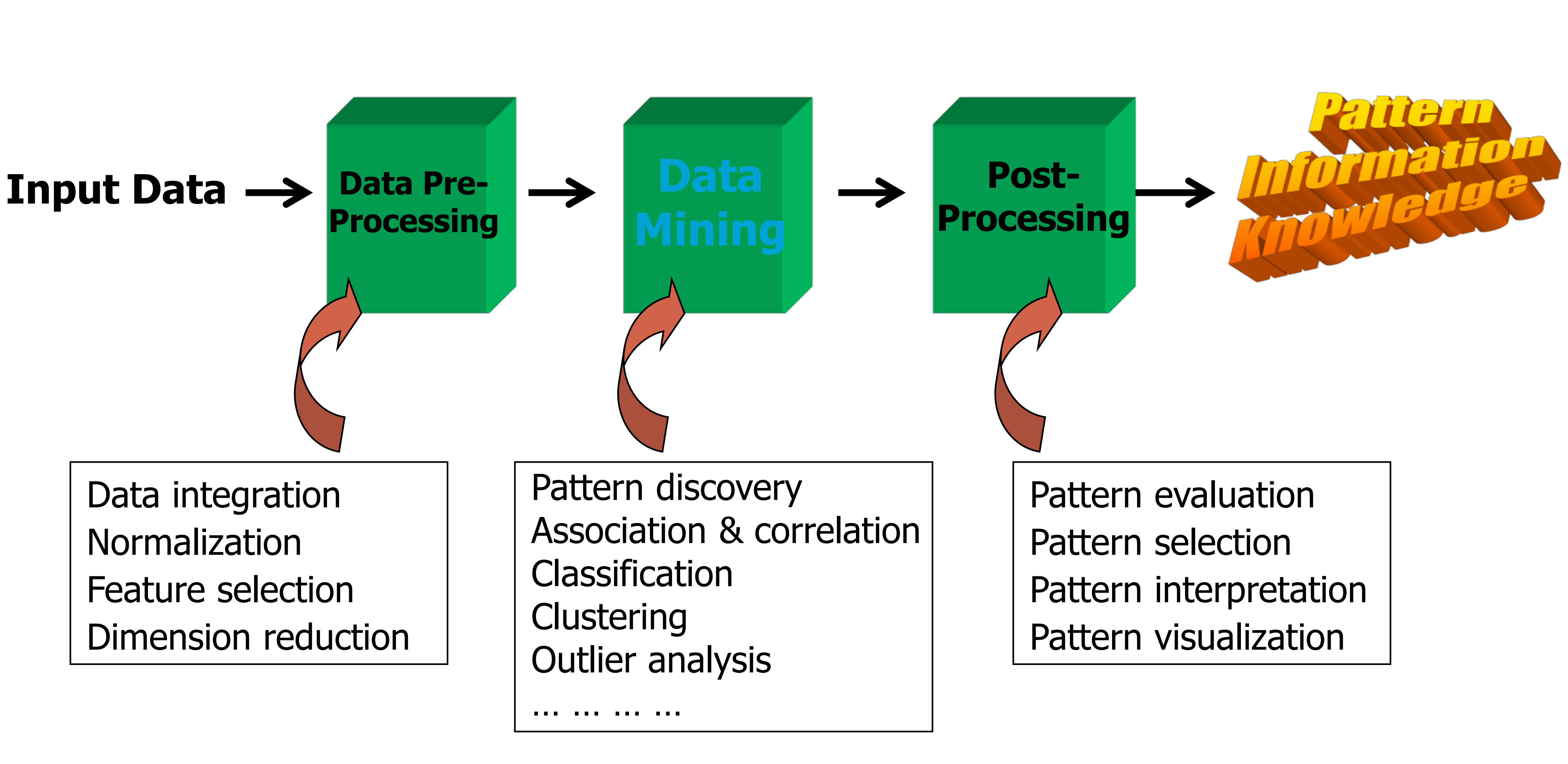
Knowledge Discovery in Databases

(a.k.a. database system / data warehouse perspective)



Data Mining ≃ Data Science

(a.k.a. machine learning and statistics perspective)



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

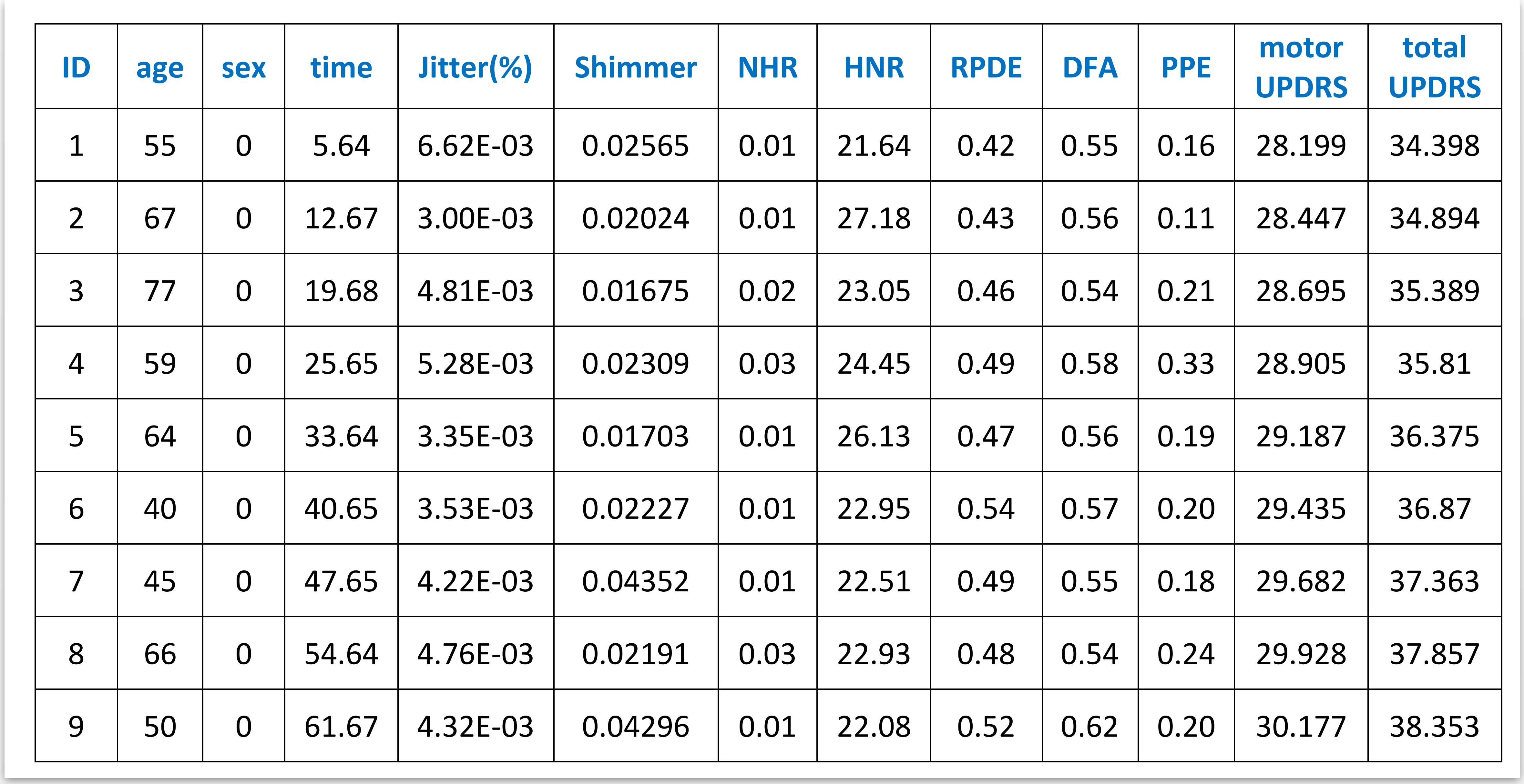
Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

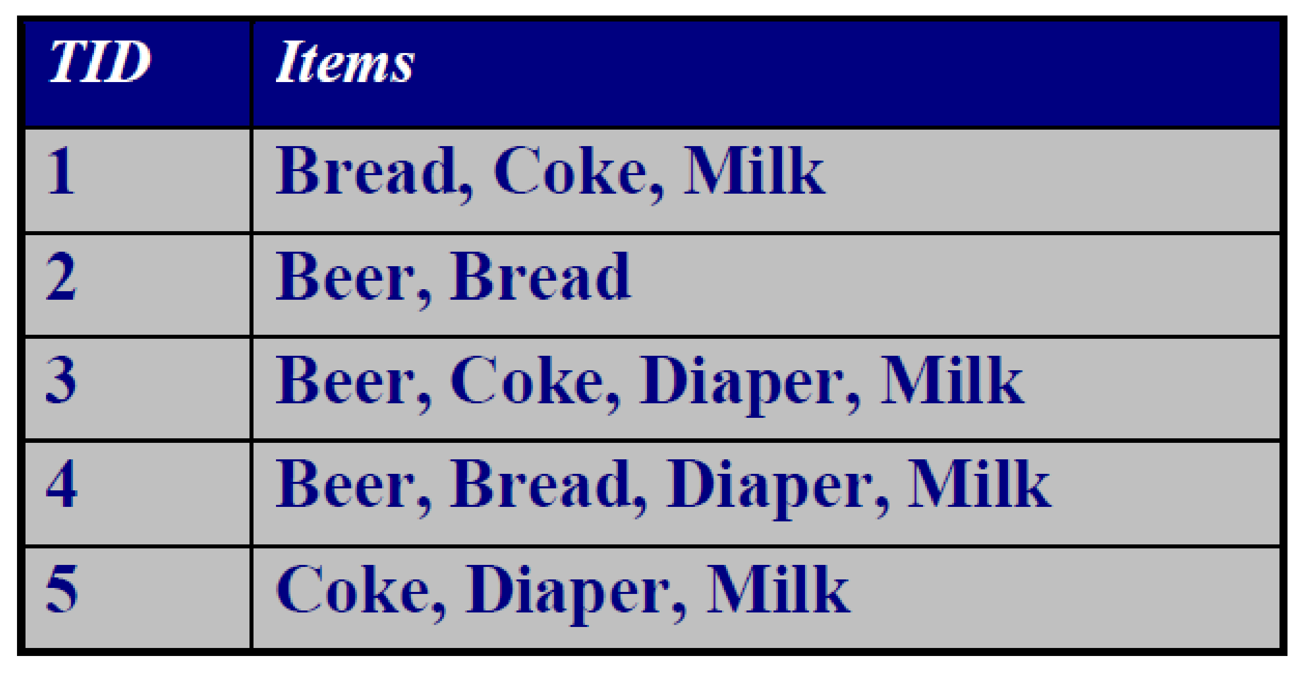
1. Types of Data

Matrix Data

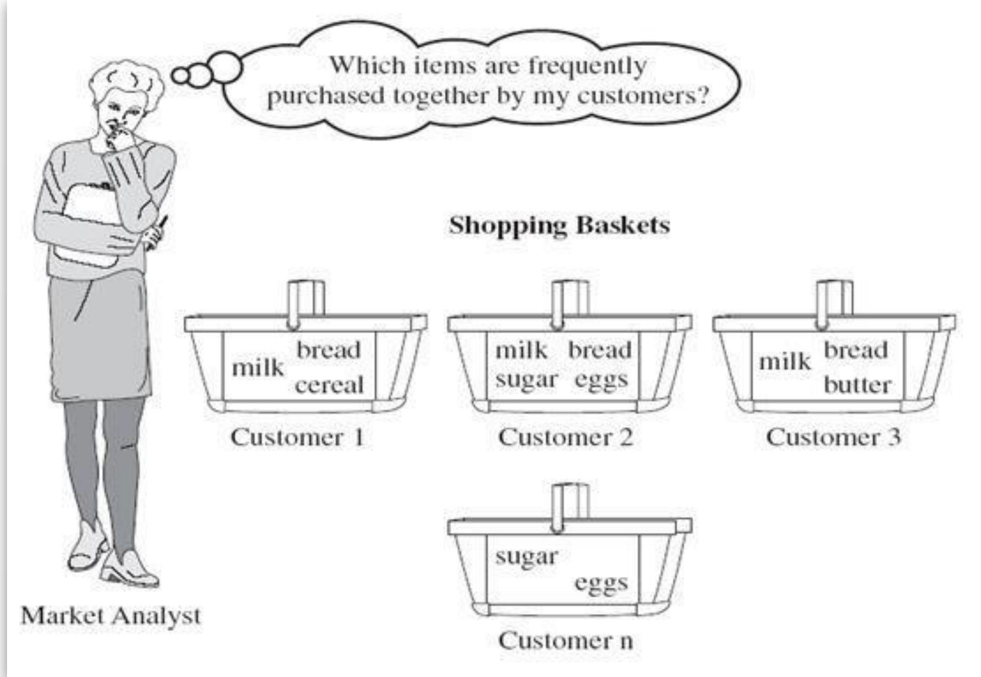


ID	age	sex	time	Jitter(%)	Shimmer	NHR	HNR	RPDE	DFA	PPE	motor UPDRS	total UPDRS
1	55	0	5.64	6.62E-03	0.02565	0.01	21.64	0.42	0.55	0.16	28.199	34.398
2	67	0	12.67	3.00E-03	0.02024	0.01	27.18	0.43	0.56	0.11	28.447	34.894
3	77	0	19.68	4.81E-03	0.01675	0.02	23.05	0.46	0.54	0.21	28.695	35.389
4	59	0	25.65	5.28E-03	0.02309	0.03	24.45	0.49	0.58	0.33	28.905	35.81
5	64	0	33.64	3.35E-03	0.01703	0.01	26.13	0.47	0.56	0.19	29.187	36.375
6	40	0	40.65	3.53E-03	0.02227	0.01	22.95	0.54	0.57	0.20	29.435	36.87
7	45	0	47.65	4.22E-03	0.04352	0.01	22.51	0.49	0.55	0.18	29.682	37.363
8	66	0	54.64	4.76E-03	0.02191	0.03	22.93	0.48	0.54	0.24	29.928	37.857
9	50	0	61.67	4.32E-03	0.04296	0.01	22.08	0.52	0.62	0.20	30.177	38.353

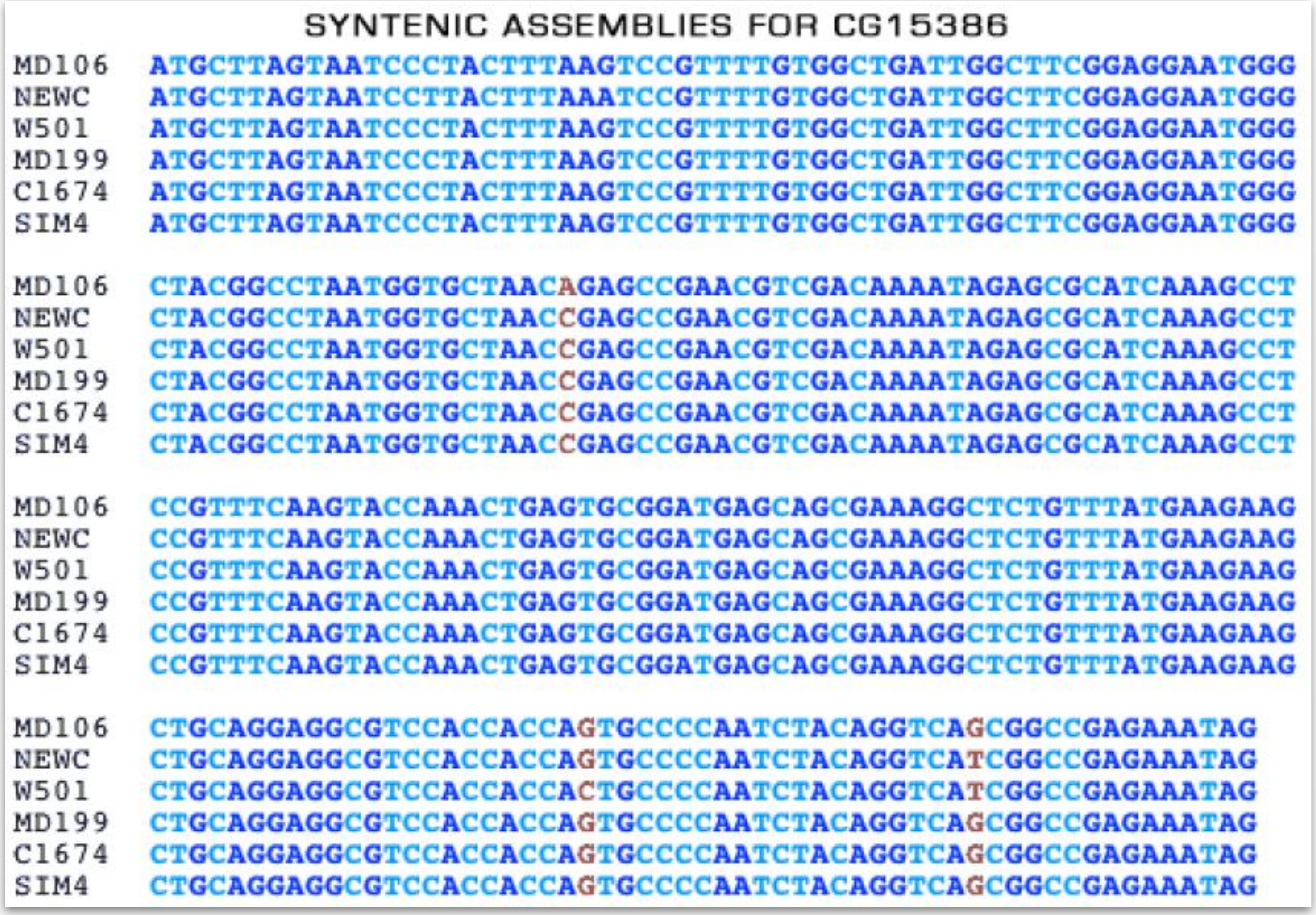
Set Data



TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



Sequence Data



SYNTENIC ASSEMBLIES FOR CG15386

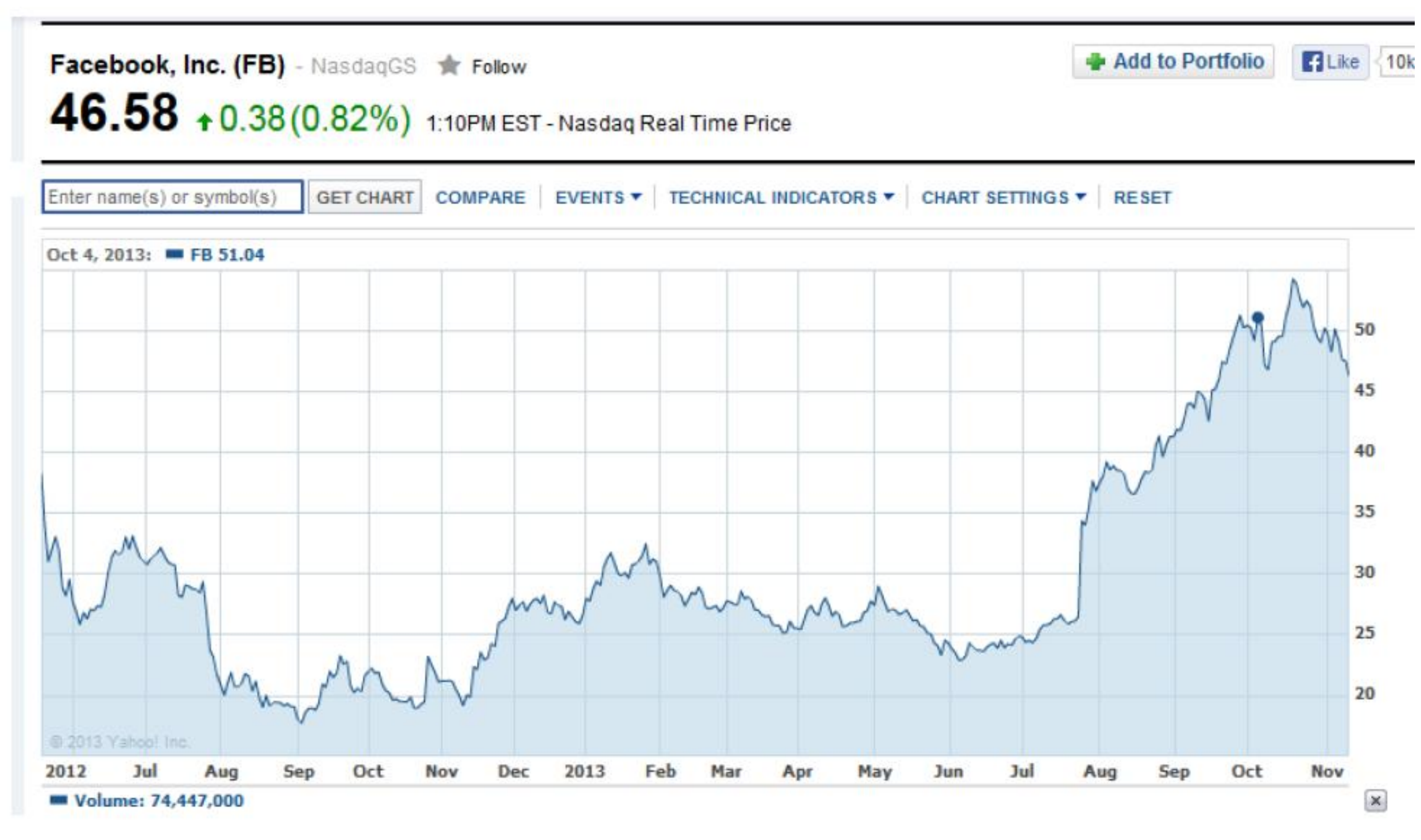
```
MD106   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC    ATGCTTAGTAATCCTTACTTTAAATCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
SIM4    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG

MD106   CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
SIM4    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT

MD106   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
NEWC    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
W501    CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
MD199   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
C1674   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
SIM4    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
```

Time Series Data



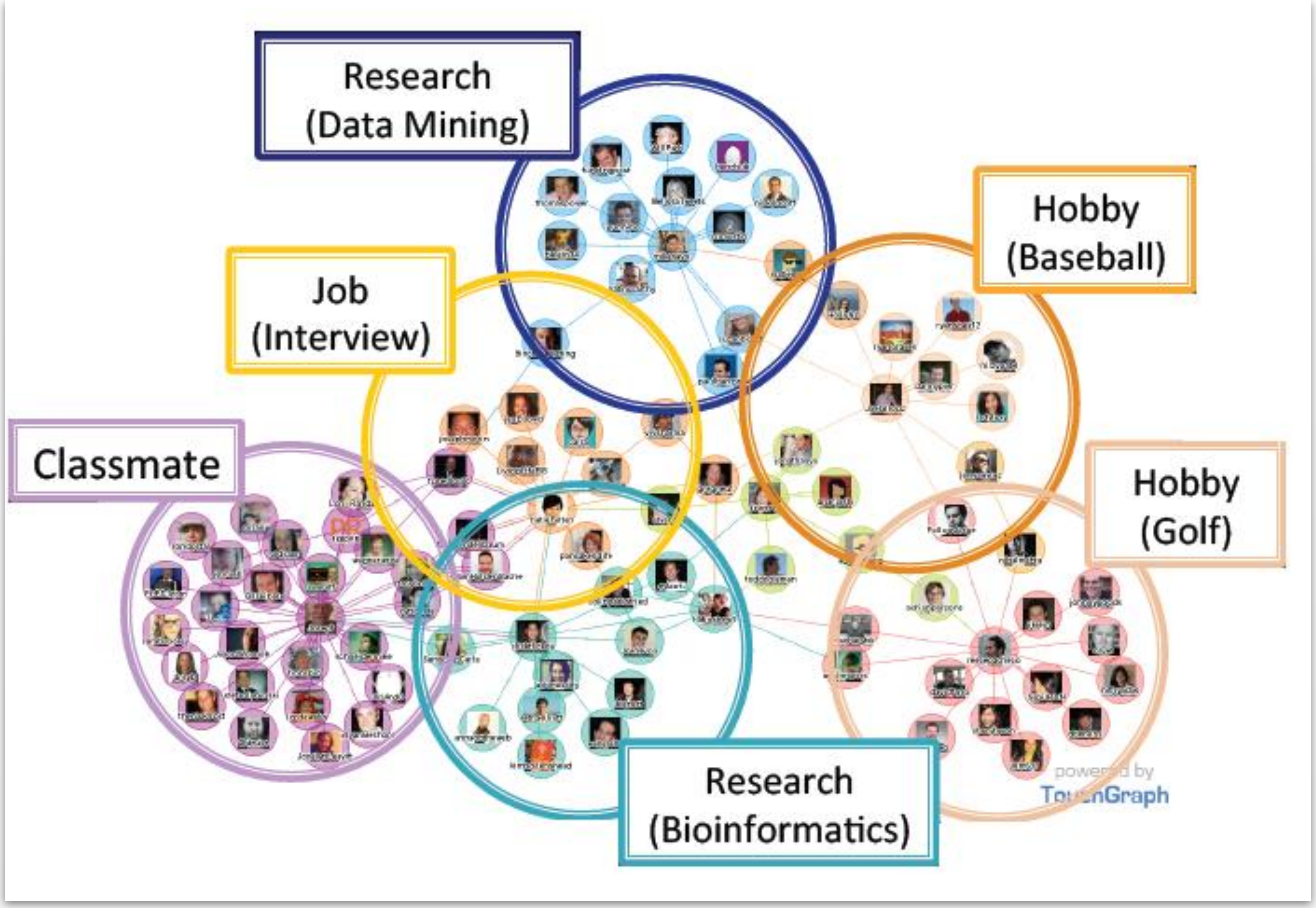
Facebook, Inc. (FB) - NasdaqGS

46.58 ↑ 0.38 (0.82%) 1:10PM EST - Nasdaq Real Time Price

Oct 4, 2013: FB 51.04

Volume: 74,447,000

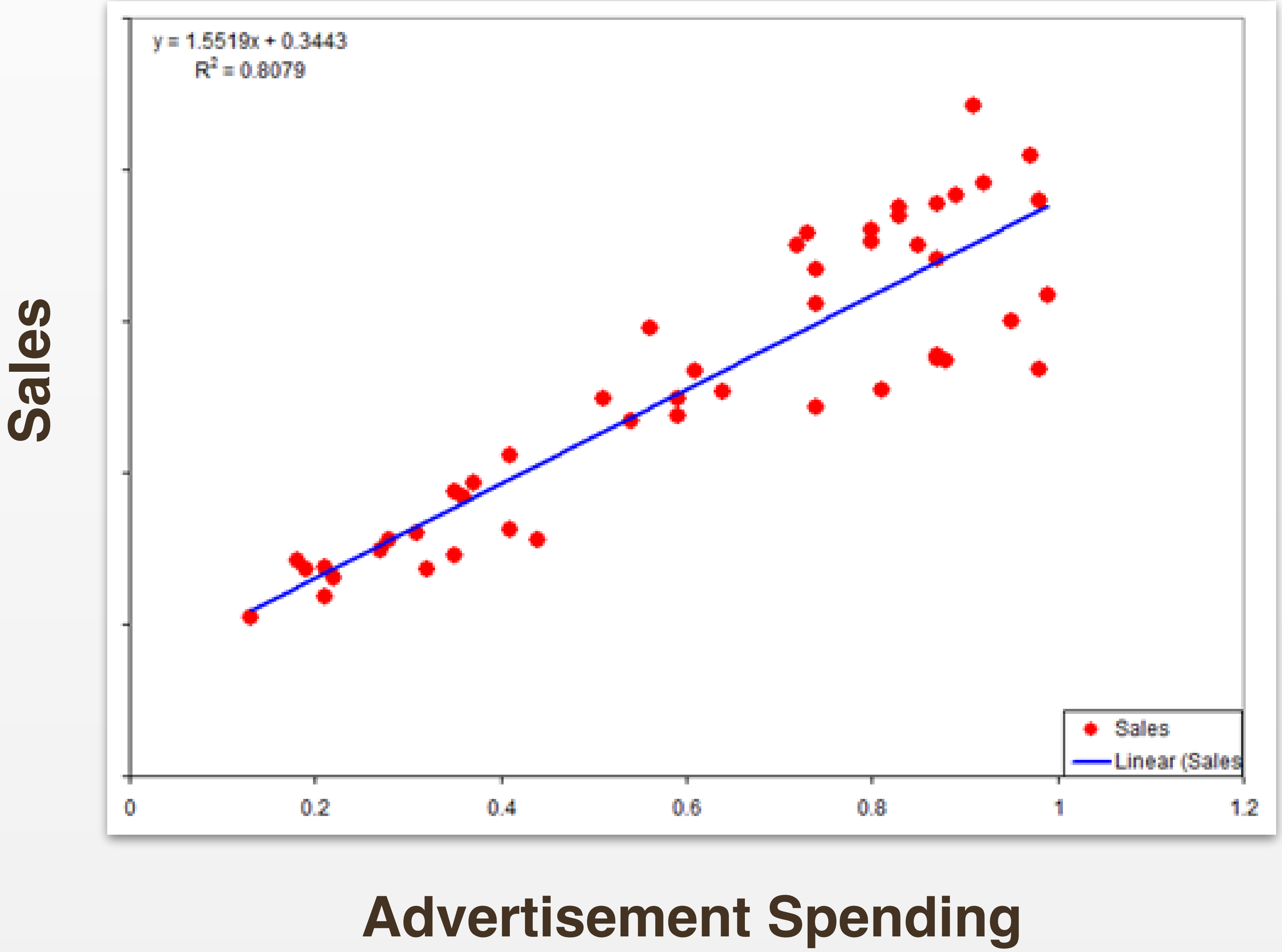
Graph / Network Data



2. Types of Methods

Regression

(a.k.a. predicting continuous things)

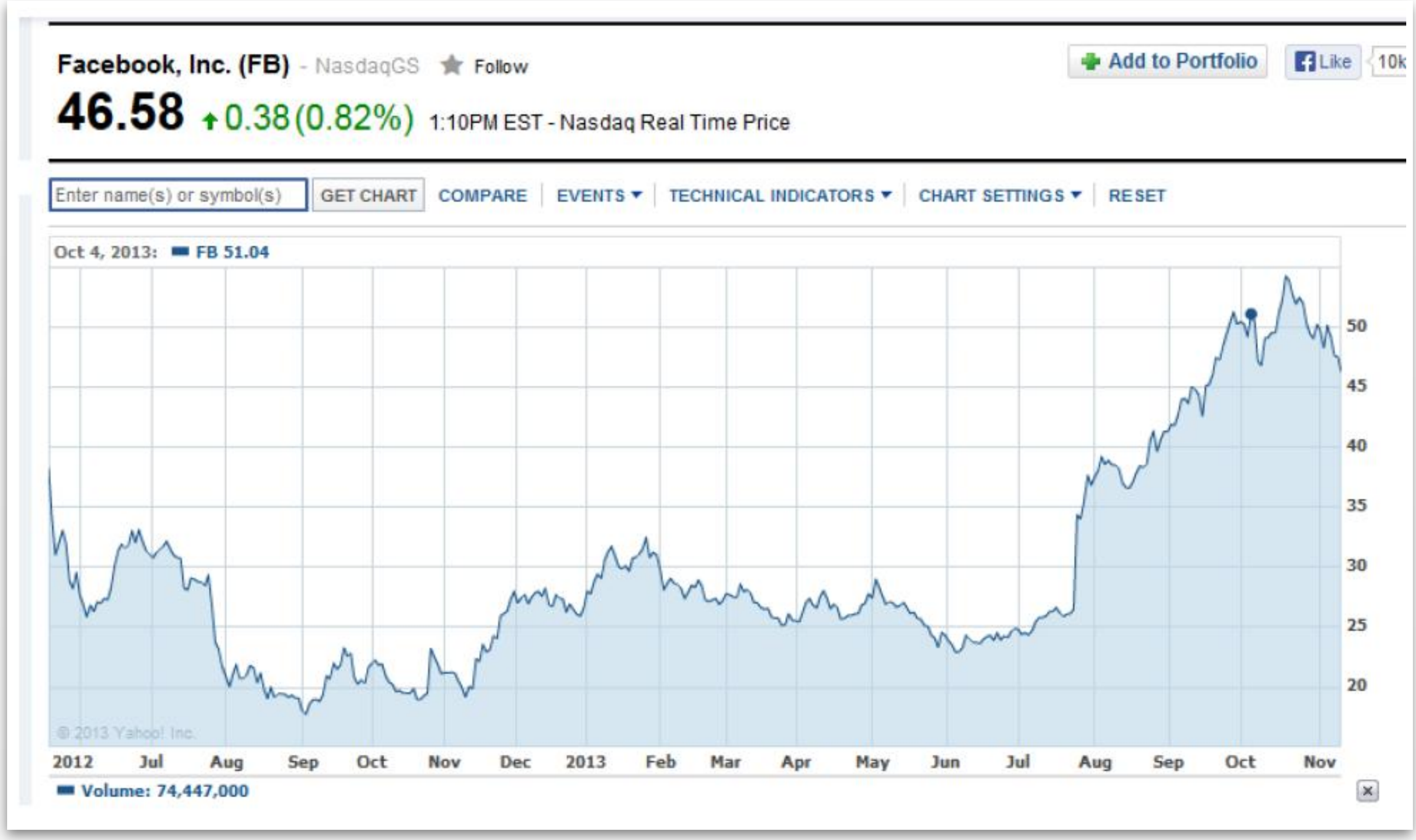


Methods

- Linear Regression
- Gaussian Processes
- Autoregressive Models

Regression

(a.k.a. predicting continuous things)



Facebook, Inc. (FB) - NasdaqGS

46.58 ↑0.38(0.82%) 1:10PM EST - Nasdaq Real Time Price

Oct 4, 2013: ▬ FB 51.04

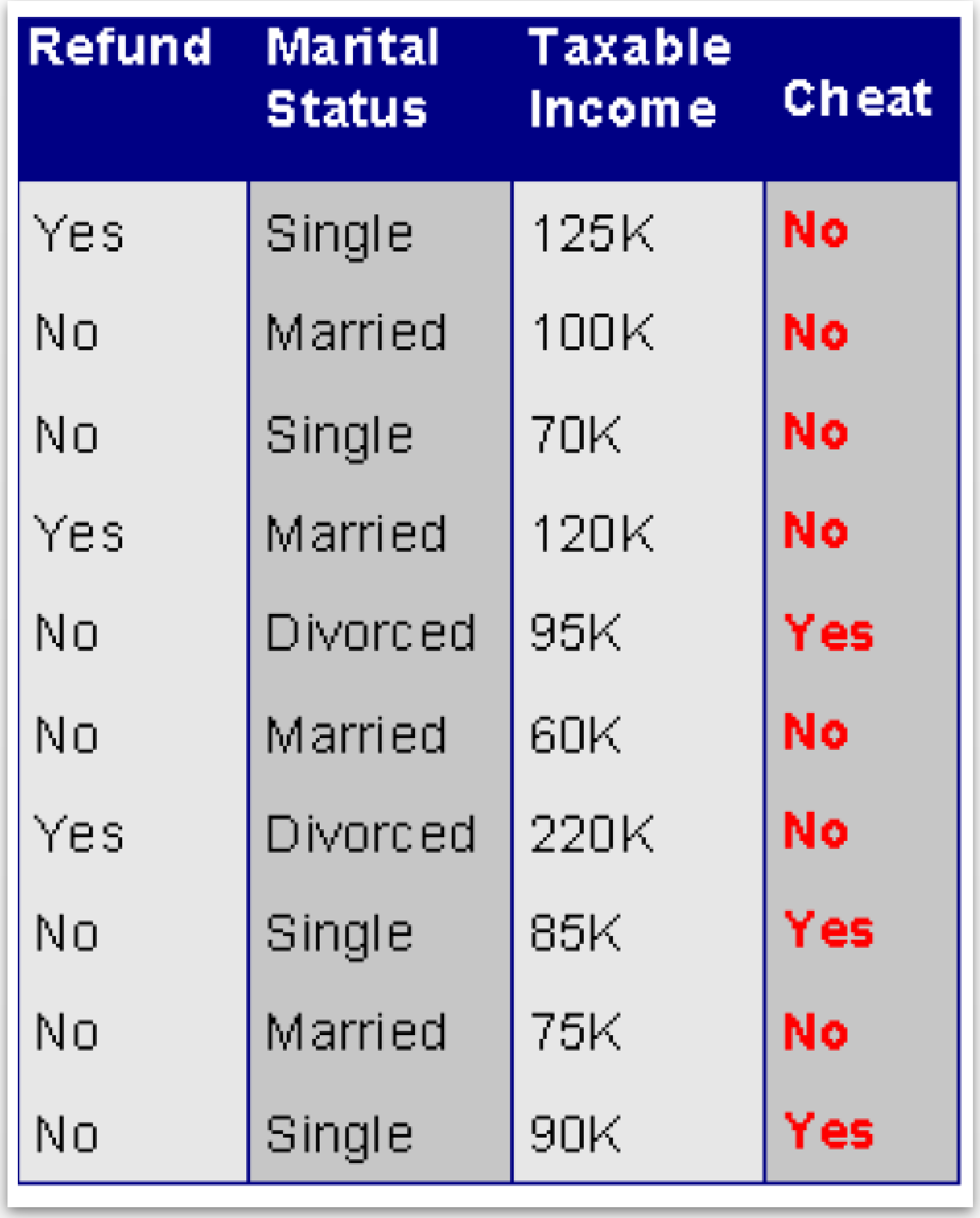
▬ Volume: 74,447,000

Methods

- Linear Regression
- Gaussian Processes
- Autoregressive Models

Classification

(a.k.a. predicting discrete things)



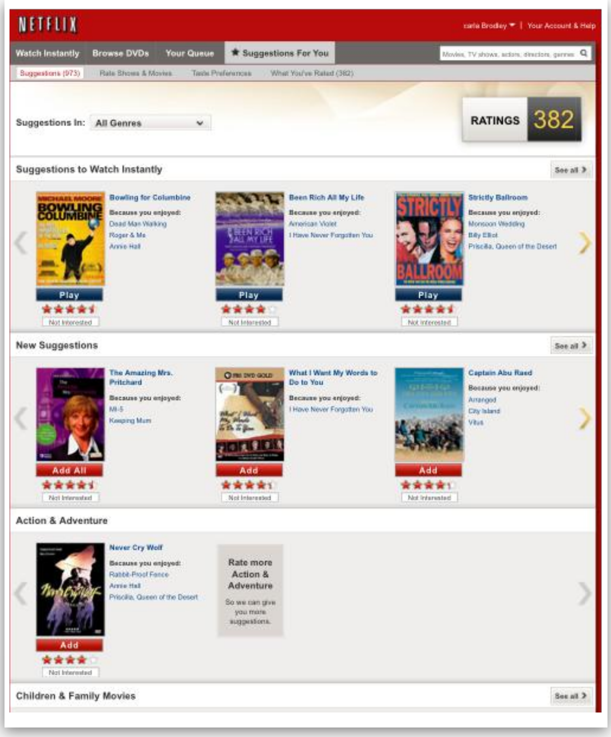
Refund	Marital Status	Taxable Income	Cheat
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

Methods

- Naive Bayes
- Decision Trees
  - Boosting
  - Random Forests
- Support Vector Machines
- Logistic Regression
- k-Nearest Neighbors

Regression/Classification Applications

Recommender Systems



Suggestions to Watch Instantly

Bowling for Columbine
Because you enjoyed:
Dead Man Walking
Roger & Me
Annie Hall

Been Rich All My Life
Because you enjoyed:
American Violet
I Have Never Forgotten You

Strictly Ballroom
Because you enjoyed:
Monsoon Wedding
Billy Elliot
Priscilla, Queen of the Desert

New Suggestions

The Amazing Mrs. Pritchard
Because you enjoyed:
MI-5
Keeping Mum

What I Want My Words to Do to You
Because you enjoyed:
I Have Never Forgotten You

Captain Abu Raed
Because you enjoyed:
Arranged
City Island
Vitus

Action & Adventure

Never Cry Wolf
Because you enjoyed:
Rabbit-Proof Fence
Annie Hall
Priscilla, Queen of the Desert

Rate more Action & Adventure
So we can give you more suggestions.

Children & Family Movies

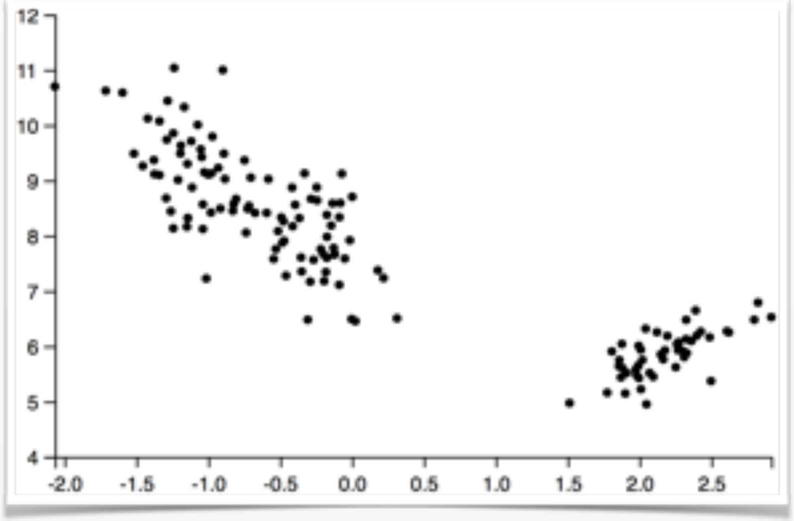
Character Recognition

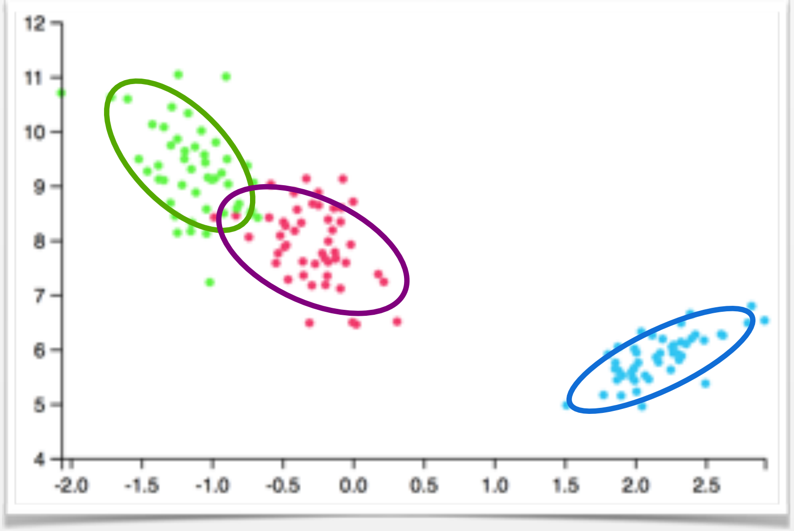
Healthcare

"I'm sorry, the doctor no longer makes diagnoses."

Clustering

(a.k.a. grouping things)



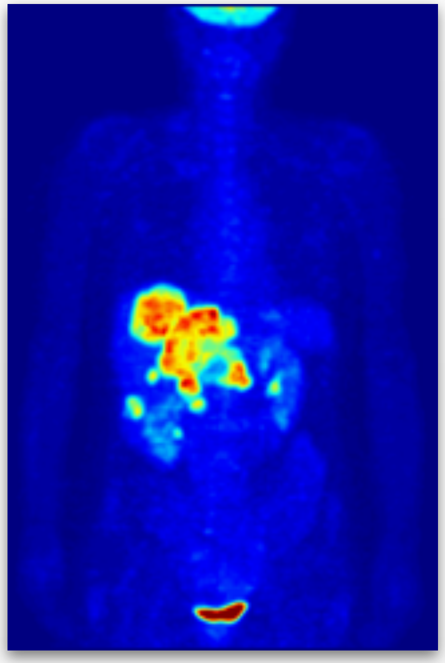


Methods

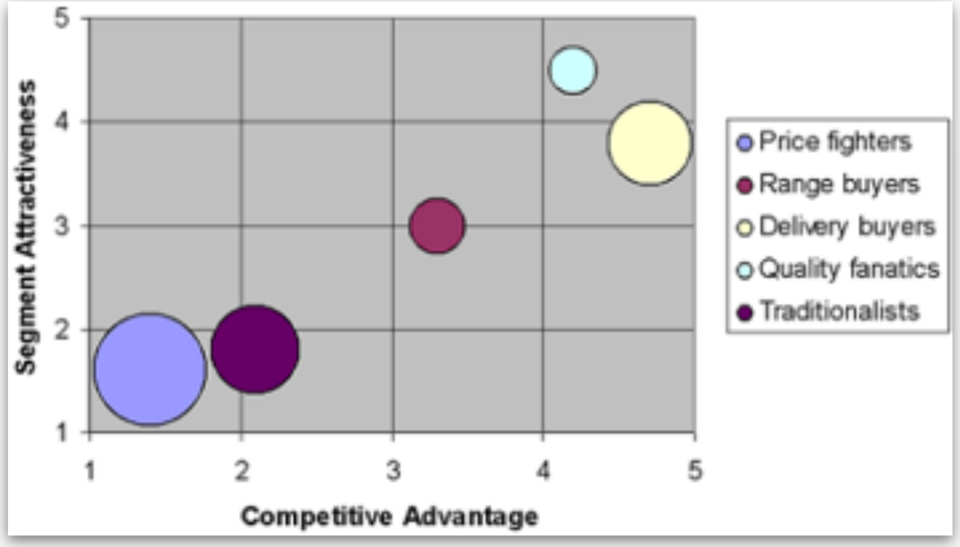
- K-means, K-medioids
- DBSCAN
- Gaussian Mixture Models (expectation maximization)

Clustering Applications

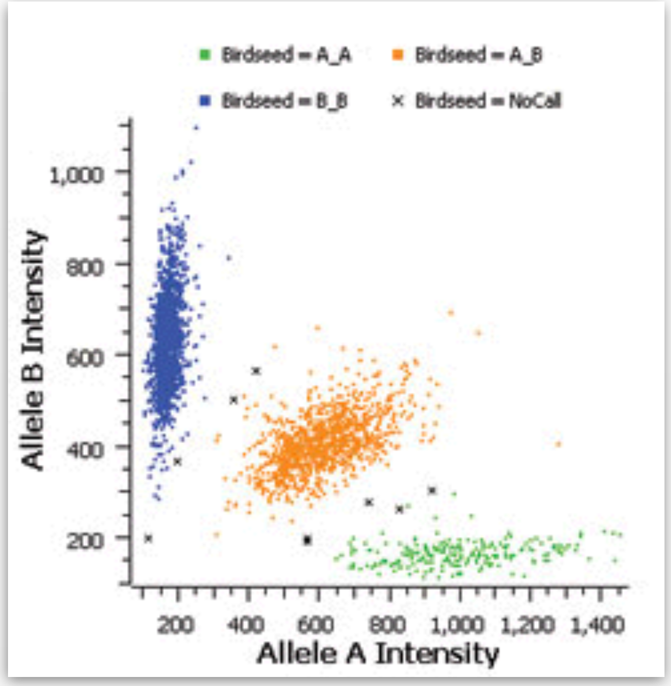
Medical Imaging



Market Research

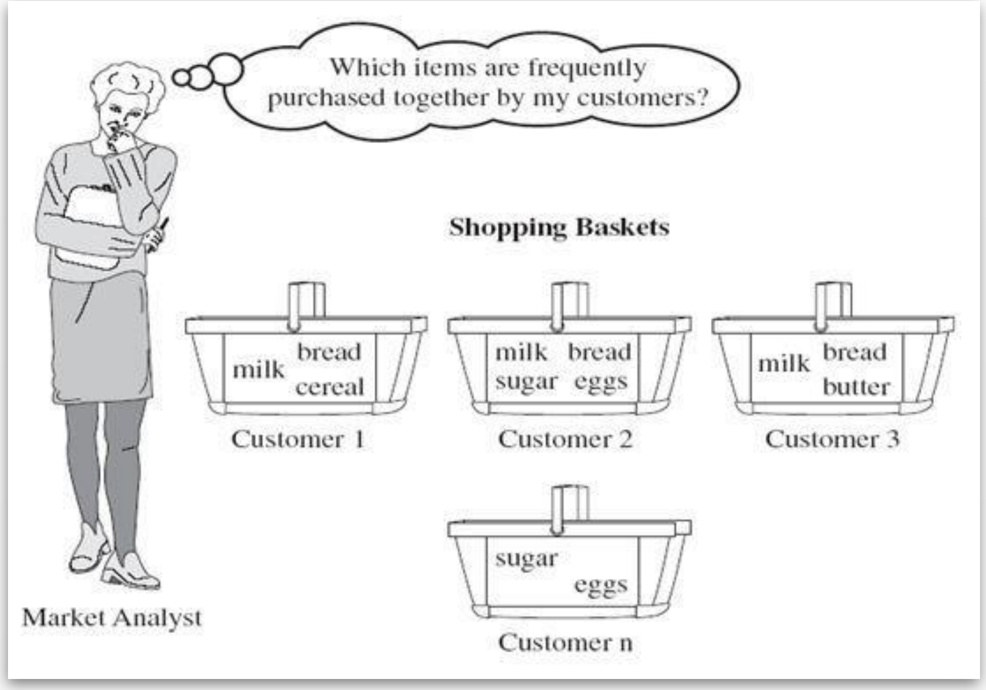


Genotyping



Association Rules Mining

(a.k.a. predicting sets of things)



Frequent Itemsets
What items are purchased together?

Association, correlation vs causality
Diaper -> Beer
[0.5% support, 75% confidence]

Methods

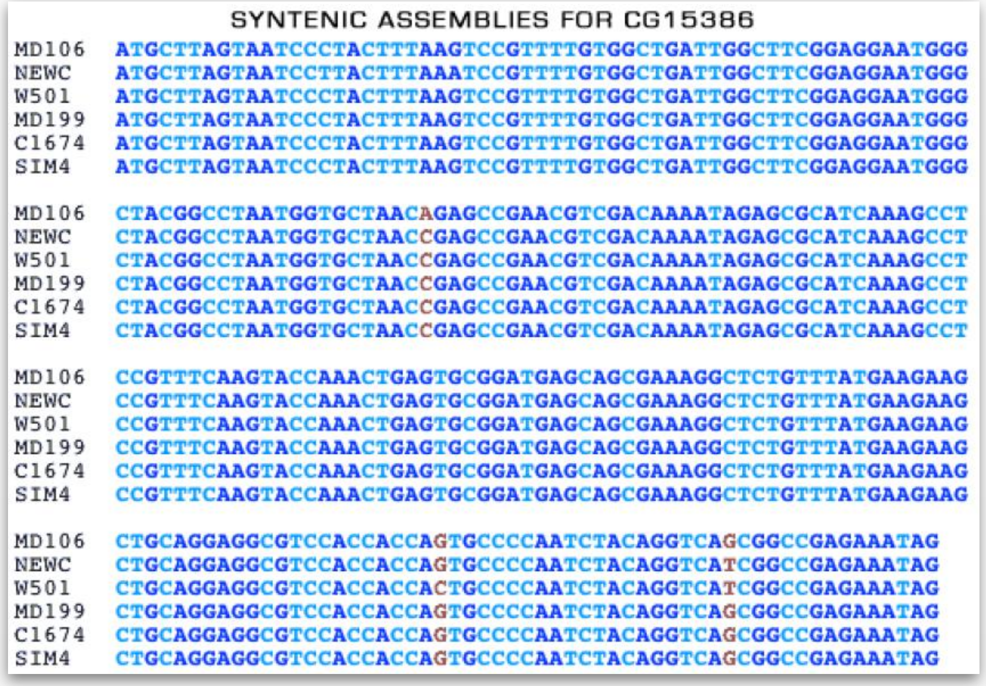
- Apriori
- FP-Growth

Association Rules Applications

- ***Market Basket Analysis***
  - Cross-selling
  - Promotions
  - Catalog design
- ***Customer Relationship Management***
  - Identify customer preference
  - Identify new product tailored to customer's liking (e.g. credit card)
- ***Census Data Analysis***
  - Plan public services (education, health, transportation, etc.)
  - Create new public business (banks, shopping malls, etc.)

Sequence Mining

(a.k.a. predicting *ordered* sets of things)



SYNTENIC ASSEMBLIES FOR CG15386

```
MD106   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC    ATGCTTAGTAATCCTTACTTTAAATCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
SIM4    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG

MD106   CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
SIM4    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT

MD106   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
NEWC    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
W501    CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
MD199   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
C1674   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
SIM4    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
```

Methods

- Generalized Sequential Patterns
- PrefixSpan
- Hidden Markov Models

Sequence Mining Applications

- **Telephone calling/webpage click patterns**
- **Speech Recognition / Speech synthesis**
- **Natural Language Processing**
  (part of speech tagging)
- **Computational biology**
  - ***Profile comparison***: identifying similarities between proteins
  - ***Gene prediction***: identifying the regions of genomic DNA that encode genes.
  - ***Sequence alignment***: identify homologous DNA sequences in a database.

Course Outline

- **Regression**
  Bias-variance tradeoff, overfitting, cross-validation
- **Classification**
  Naive Bayes, Logistic Regression, SVMs, Random Forests
- **Clustering**
  K-means, K-medioids, DBSCAN, EM for Mixture Models
- **Dimensionality Reduction**
  PCA, ICA, Random Projections
- **Time Series**
  ARIMA, HMMs
- **Recommender systems**
- **Frequent Pattern Mining**
  Apriori, FP-Growth
- **Networks**
  Page-rank, Spectral Clustering

Course Outline

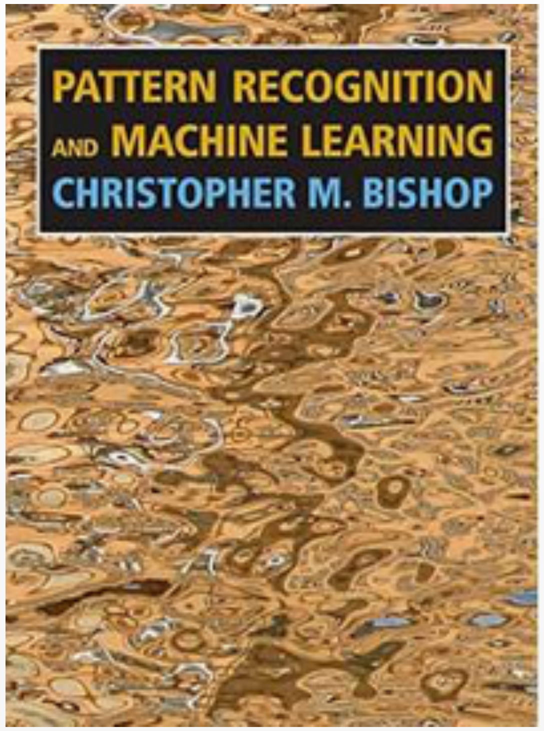
- **Regression**
  Bias-variance tradeoff, overfitting, cross-validation
- **Classification**
  Naive Bayes, Logistic Regression, SVMs, Random Forests
- **Clustering**
  K-means, K-medioids, DBSCAN, EM for Mixture Models
- **Dimensionality Reduction**
  PCA, ICA, Random Projections
- **Time Series**
  ARIMA, HMMs
- **Recommender systems**
- **Frequent Pattern Mining**
  Apriori, FP-Growth
- **Networks**
  Page-rank, Spectral Clustering

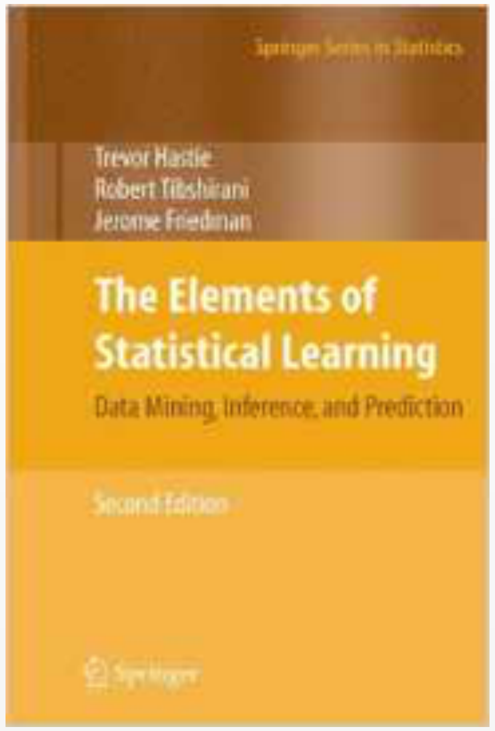
Supervised Learning

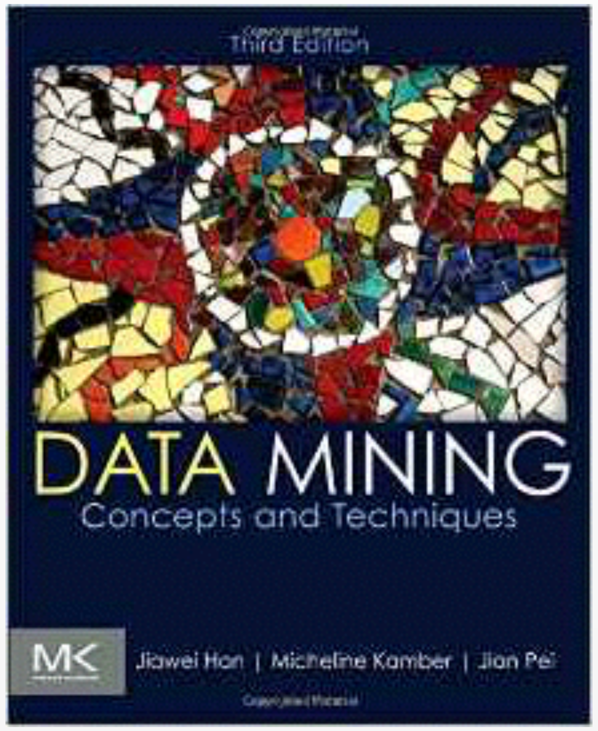
Unsupervised Learning

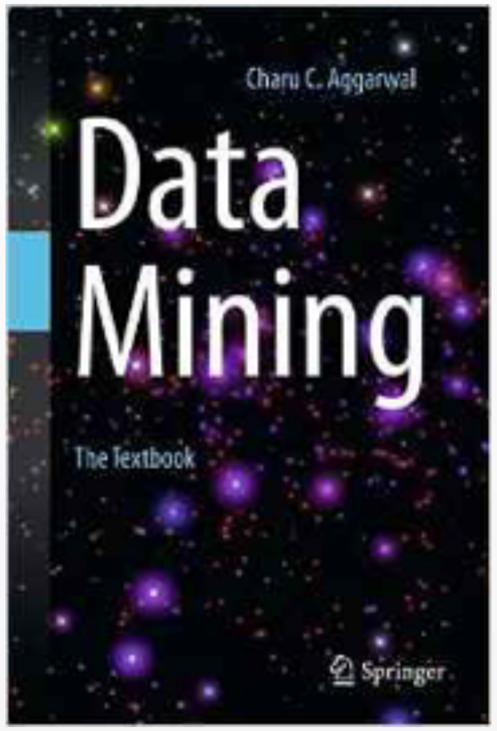
Data Mining

Textbooks









Bishop	Hastie	Han	Aggarwal
Machine Learning	Statistics	Data Mining	
On reserve at Snell	*PDF freely available*	*Ebook available through library*	*PDF available on campus network*

Question

What would ***you*** like to get out of this course?