CSE 40647/60647 Data Science

Chapter 1. Introduction

Meng Jiang

Data Science

The Instructor

- Dr. Meng Jiang (www.meng-jiang.com)

B.S. and Ph.D.

Tsinghua University

Visiting Ph.D.

Carnegie Mellon University

Postdoc Researcher

Assistant Professor

Visiting Researcher

University of Maryland

Visiting Researcher

General Learning Goals

- Learn *basic* data science concepts
- Learn *basic* methods for mining datasets
- Prerequisites:
  - Programming with **Python**
  - Data structures and Algorithms
- As a prerequisite for:
  - CSE 40625/60625: Machine Learning

What is Data Science?

- "…the process of automatically discovering *useful information* in *large* repositories of data." — *Introduction to Data Mining* (Tan, Steinbach, & Kumar)
- "…the process of discovering *patterns* in data." — *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition* (Witten, Frank, & Hall)
- "…the process of discovering *interesting patterns and knowledge* from *large* amounts of data." — *Data Mining: Concepts and Techniques, 3rd Edition* (Han, Kambler, & Pei)

More Definitions

- "…the analysis of (often large) observational data sets to find unsuspected *relationships* and to summarize the data in novel ways that are both *understandable and useful to the data owner*." (Hand, Mannila, & Smyth)

- "…the set of methods and techniques for exploring and analysing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data *certain unknown or hidden rules, associations or tendencies*, special systems output the essentials of the useful information while reducing the quantity of data." (Tuffery)

Our Definition of the Course

- "…the art and craft of extracting *knowledge* from *large* bodies of *structured and unstructured* data using methods from many disciplines, including (but not limited to) machine learning, databases, probability and statistics, information theory, and data visualization."

Two Key Features

- Extracting Knowledge
  – Previously unknown patterns, descriptions, or relations—potentially useful information are being extracted from data. Discovering this knowledge often requires some form of learning (modeling).

- Large Bodies of Data
  – Datasets are structured, often as a database. The data the contain is often so large that the process of extracting knowledge must be automated—or at least augmented—by computer.

Big Data

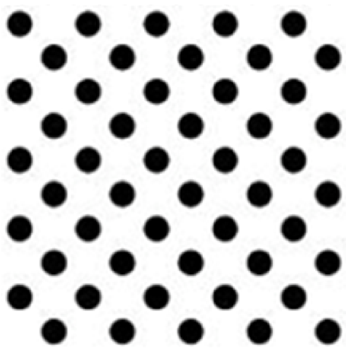
Volume

Data at Rest

Terabytes to Exabytes of existing data to process

Big Data

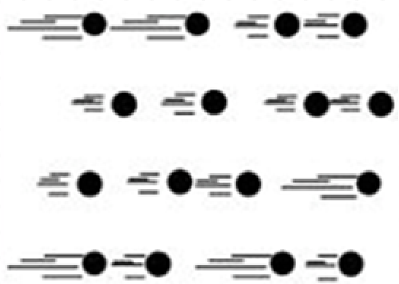
Volume



Data at Rest

Terabytes to Exabytes of existing data to process

Velocity



Data in Motion

Streaming data, requiring milliseconds to seconds to respond

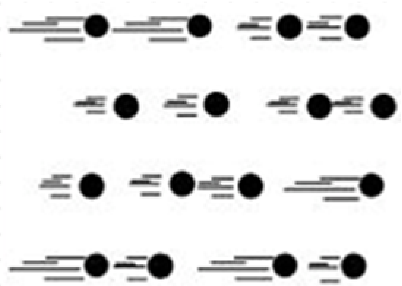
Big Data

Volume

Data at Rest

Terabytes to Exabytes of existing data to process

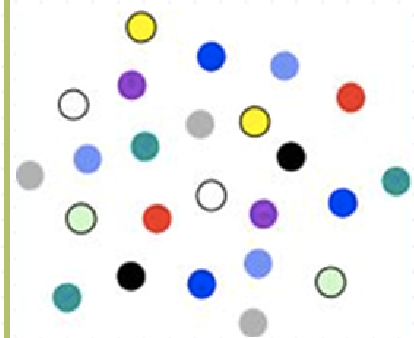
Velocity



Data in Motion

Streaming data, requiring milliseconds to seconds to respond

Variety



Data in Many Forms

Structured, unstructured, text, multimedia,...

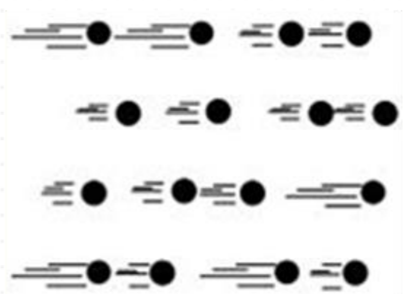
Big Data

Volume

Data at Rest

Terabytes to Exabytes of existing data to process

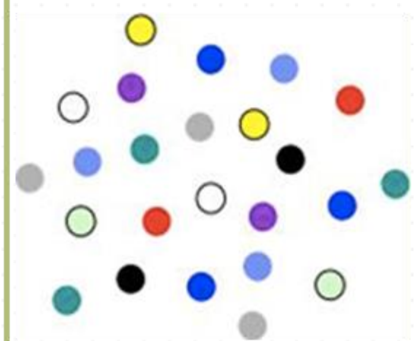
Velocity



Data in Motion

Streaming data, requiring milliseconds to seconds to respond

Variety



Data in Many Forms

Structured, unstructured, text, multimedia,…

Veracity

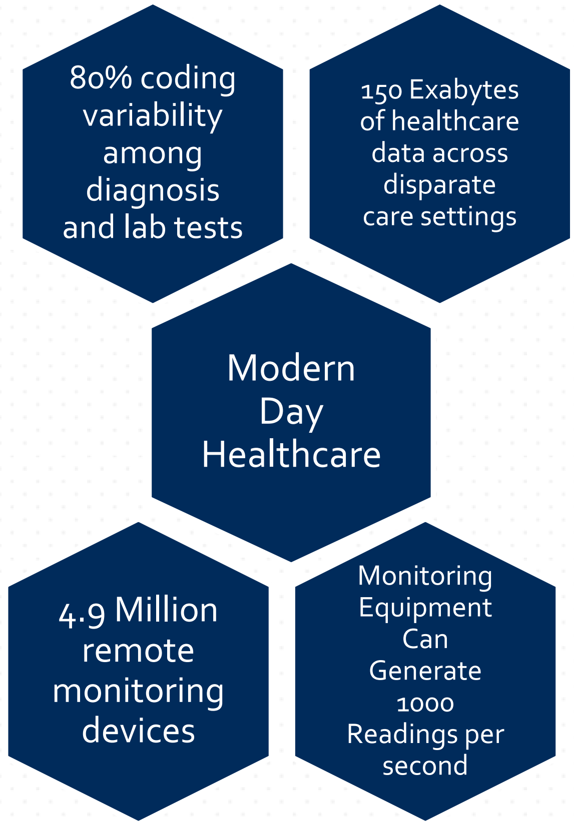
Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data in Healthcare

Variety (structured, unstructured, multimedia)

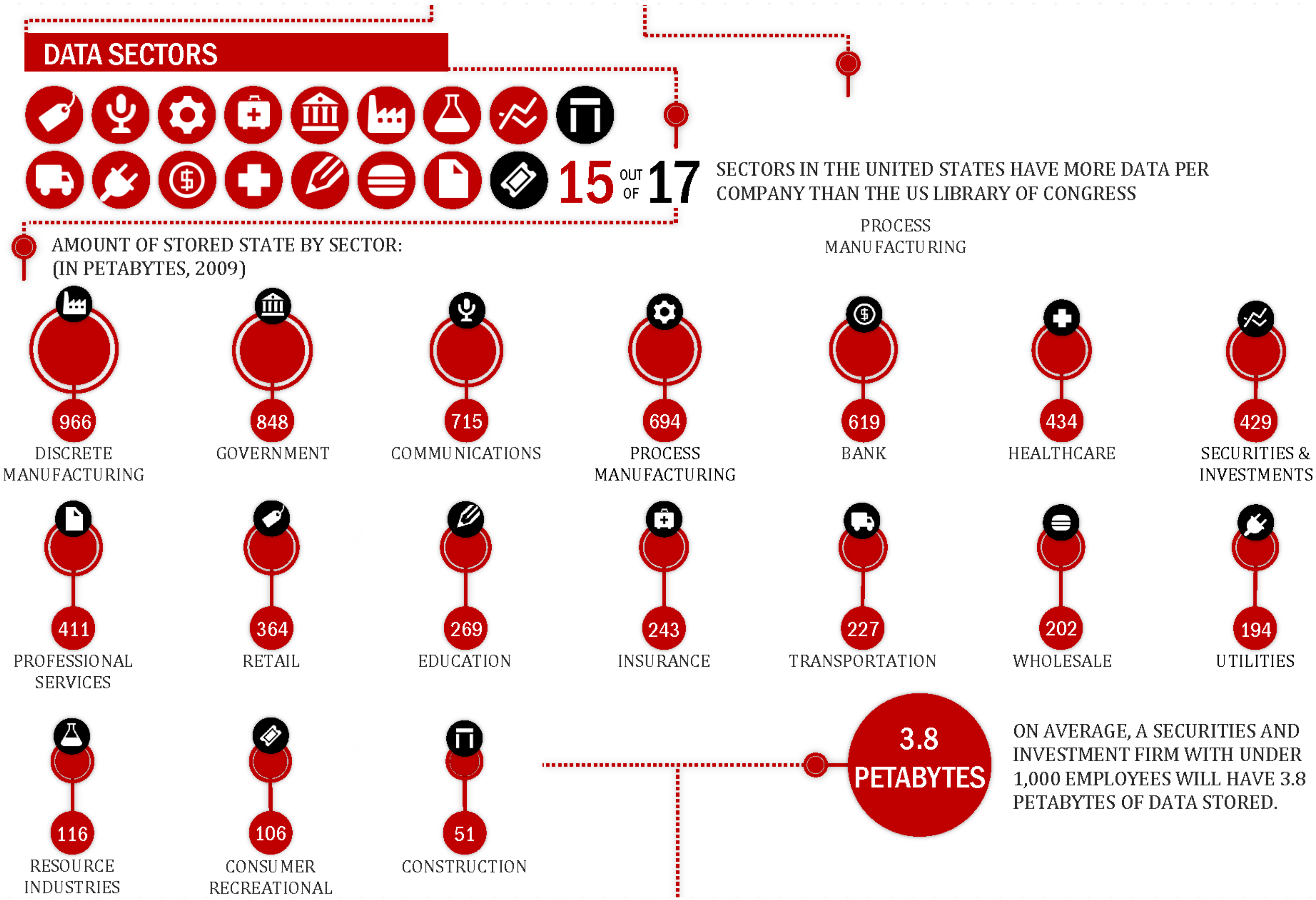
Volume (massive data)



Veracity (reliability and predictability of data)

Velocity (streaming data)

Why did you enroll in this course?



What is/isn't Data Science?

[ ] Looking up a record in a database.

[ ] Noting that some last names occur in certain geographical areas.

[ ] Searching for a term on Google.

[ ] Taking all query results from Google and discovering that they can be grouped or categorized.

[ ] Testing a two-sample hypothesis in a clinical trial.

[ ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[   ] Noting that some last names occur in certain geographical areas.

[   ] Searching for a term on Google.

[   ] Taking all query results from Google and discovering that they can be grouped or categorized.

[   ] Testing a two-sample hypothesis in a clinical trial.

[   ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[√] Noting that some last names occur in certain geographical areas.

[   ] Searching for a term on Google.

[   ] Taking all query results from Google and discovering that they can be grouped or categorized.

[   ] Testing a two-sample hypothesis in a clinical trial.

[   ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[√] Noting that some last names occur in certain geographical areas.

[×] Searching for a term on Google.

This is simply a "match" or "non-match".

[ ] Taking all query results from Google and discovering that they can be grouped or categorized.

[ ] Testing a two-sample hypothesis in a clinical trial.

[ ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[√] Noting that some last names occur in certain geographical areas.

[×] Searching for a term on Google.

This is simply a "match" or "non-match".

[√] Taking all query results from Google and discovering that they can be grouped or categorized.

[ ] Testing a two-sample hypothesis in a clinical trial.

[ ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[√] Noting that some last names occur in certain geographical areas.

[×] Searching for a term on Google.

This is simply a "match" or "non-match".

[√] Taking all query results from Google and discovering that they can be grouped or categorized.

[×] Testing a two-sample hypothesis in a clinical trial.

This falls more to pure statistics, does this generalize?

[ ] When doing multiple tests across many different genes, identifying strongly associated genes.

What is/isn't Data Science?

[×] Looking up a record in a database.

No pattern is revealed by this lookup.

[√] Noting that some last names occur in certain geographical areas.

[×] Searching for a term on Google.

This is simply a "match" or "non-match".

[√] Taking all query results from Google and discovering that they can be grouped or categorized.

[×] Testing a two-sample hypothesis in a clinical trial.

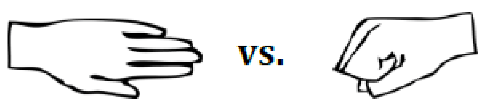
This falls more to pure statistics, does this generalize?

[√] When doing multiple tests across many different genes, identifying strongly associated genes.

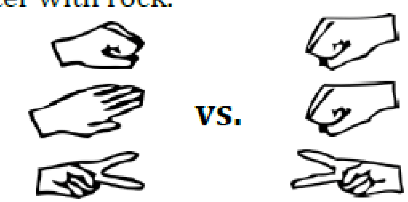
An Example

Rock, Paper, Scissors

1 Rock-it:
Males have the tendency to produce rock on their first throw. If you are playing against one, try using paper.



2 Double on the Rocks:
When you see a two-Rock run, it is highly likely that your opponent's next move will be Scissors or Paper. People dislike being predicable. Counter with rock.



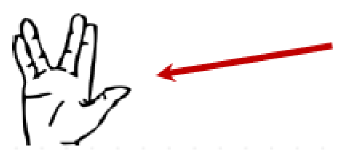
3 Paper Please:
Paper is thrown the least in a match. Use it as an unexpected options.

Paper is thrown 29.6% of the time.

Rock is thrown 35.4% of the time.

Scissors is thrown 35% of the time.

4 Spock & Roll:
When in doubt, and all seems lost, go for the Spock. It is unexpected and highly illegal, but also impossible to counter.



Okay, this part is not really data mining.

Is This Data Science?

[ ] Find the most popular hobby among us.

Is This Data Science?

[?] Find the most popular hobby among us.

Is This Data Science?

[?] Find the most popular hobby among us.

[ ] Infer the hobby of a student.

Is This Data Science?

[?] Find the most popular hobby among us.

[ ] Infer the hobby of a student.

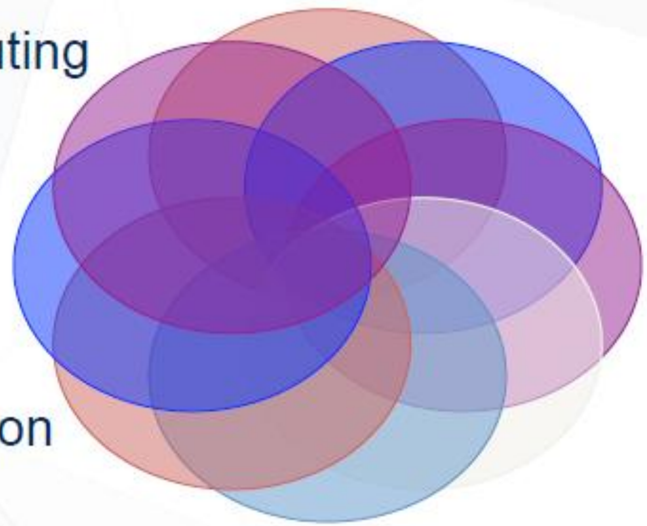
If I ask you to do the "research", what's the first step?

Defining a Data Science Task

- Problem statement & background knowledge
  - Posit the right question
- Aggregate, integrate, and understand data
- Design the machine learning or modeling approach
- Identify performance measurement criteria
- Communicate results and actionable insights

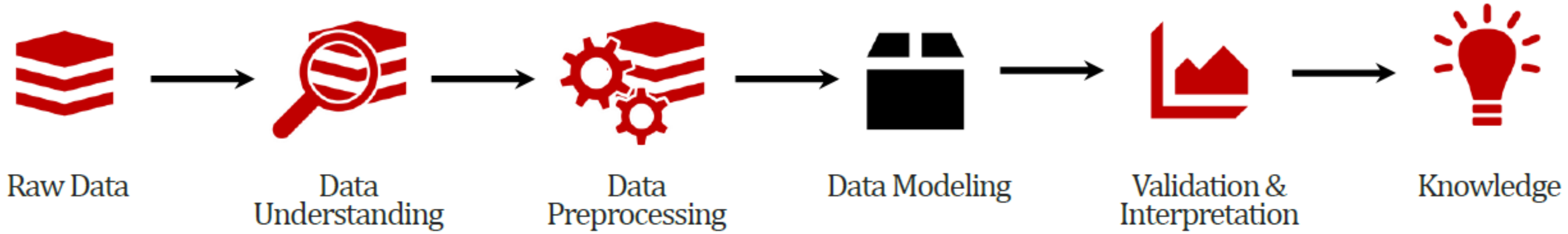
Data Science a Confluence of Fields

- Parallel and distributed computing
- Databases
- Visualization
- Pattern Recognition
- Algorithms

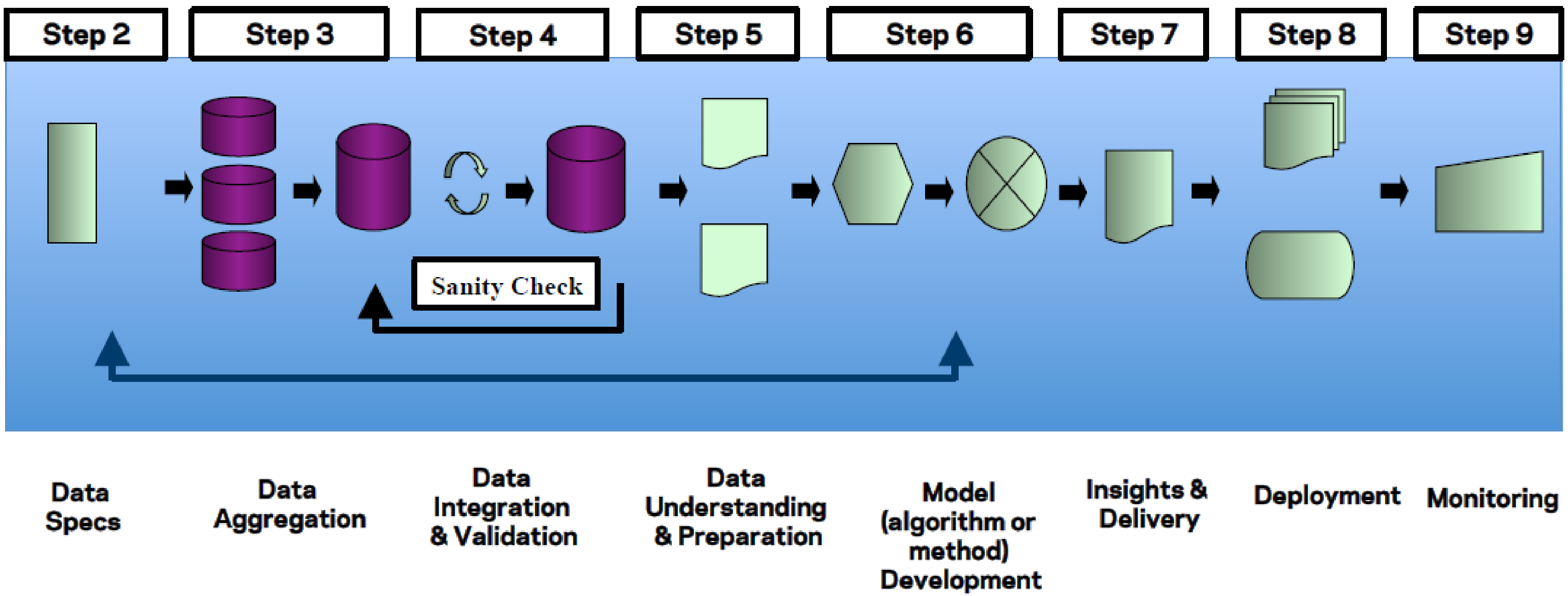


- Machine Learning
- Statistics
- Information Theory
- Probability

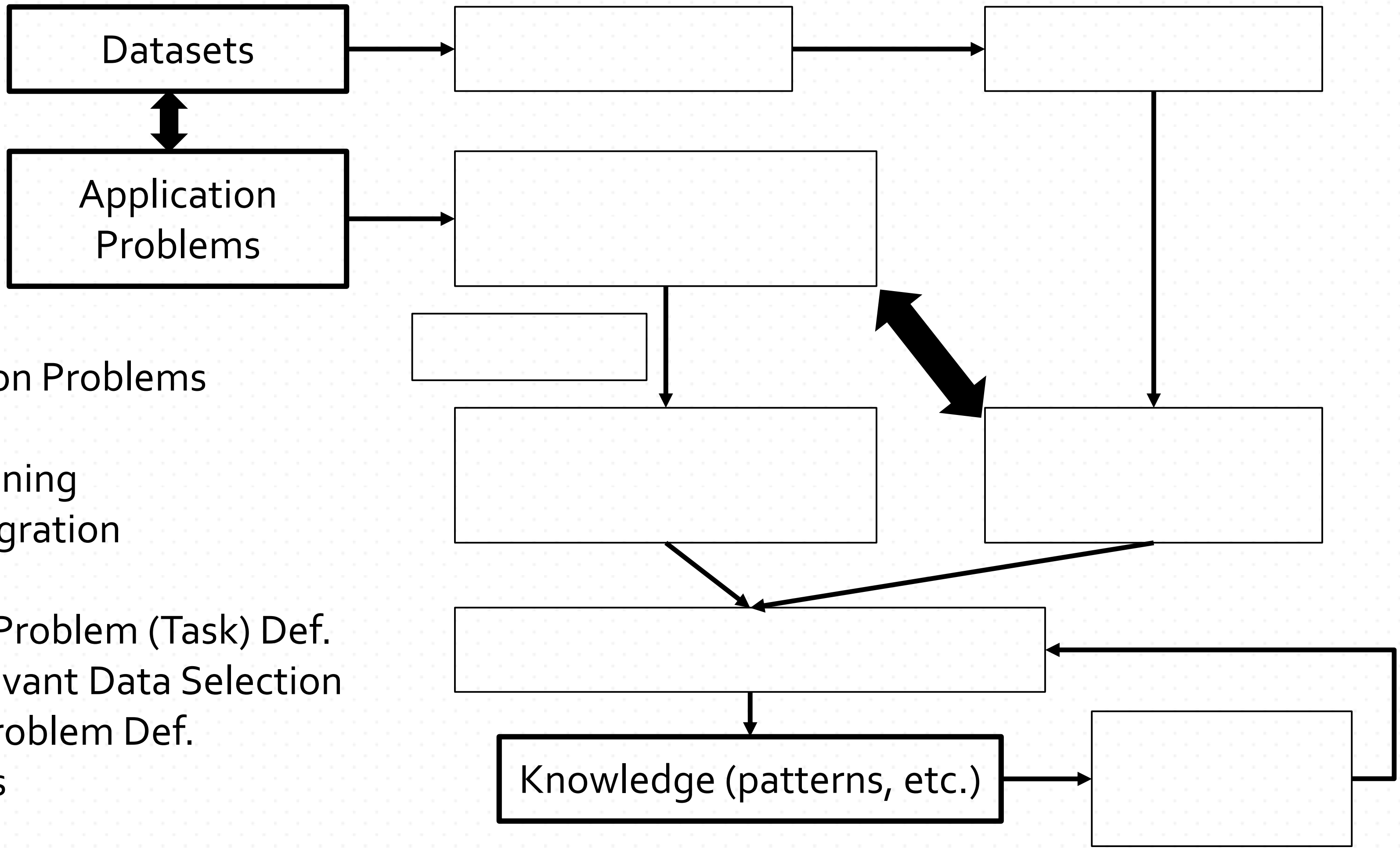
Data Science Pipeline



Data Science Pipelines



Data Science Research



1. Datasets
2. Application Problems

3. Data Cleaning
4. Data Integration

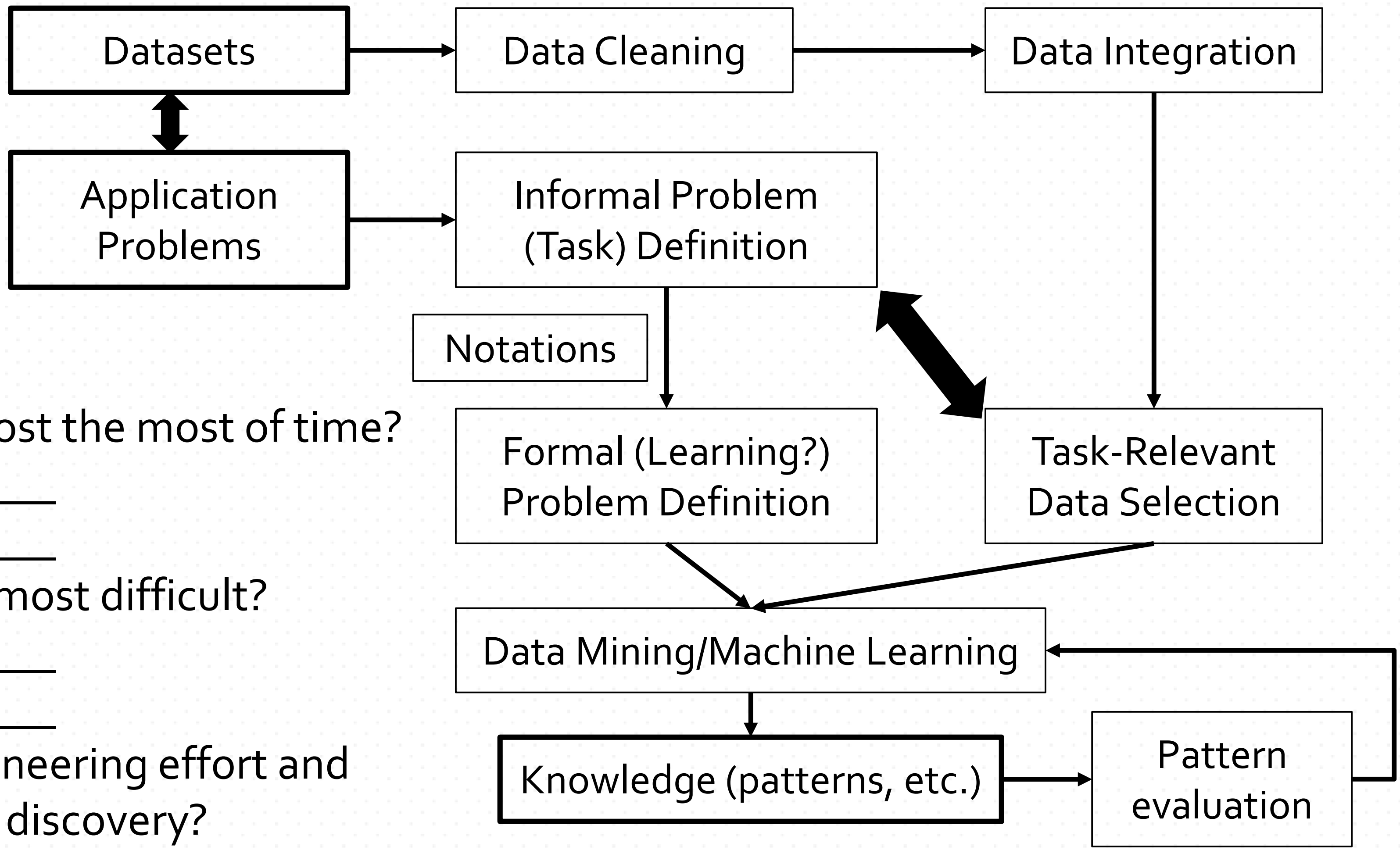
5. Informal Problem (Task) Def.
6. Task-Relevant Data Selection
7. Formal Problem Def.
8. Notations

9. Data Mining
10. Knowledge (patterns, descriptions, relations, etc.)
11. Pattern evaluation

Example

1. **Datasets:** Walmart transaction data
2. **Application Problems:** Optimize products placement for more sales
3. **Data Cleaning:** Incomplete data, noisy data, etc.
4. **Data Integration:** Multiple operational databases (markets)
5. **Informal Problem (Task) Def.:** Given transactions, which two items are often purchased together?
6. **Task-Relevant Data Selection:** Input and validation data for a task
7. **Formal Problem Def.:** Given $T = \{T_1, \ldots\}$ and $T_i \subseteq X$, find *associations* $X_j \rightarrow X_k$ that have high *support* and *confidence*.
8. **Notations:** Transaction set T, itemset/transaction T_i, the set of all the items X, items X_j
9. **Data Mining:** Propose an approach for association mining
10. **Knowledge (patterns, etc.):** The associations
11. **Pattern evaluation:** Sales increase?

Data Science Research



What may cost the most of time?

1. ______________
2. ______________

What is the most difficult?

1. ______________
2. ______________

What is engineering effort and what makes discovery?

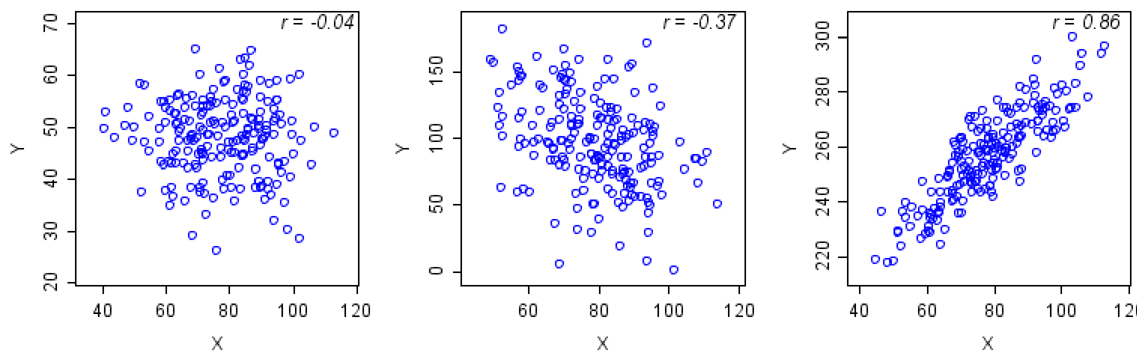
1. ______________
2. ______________

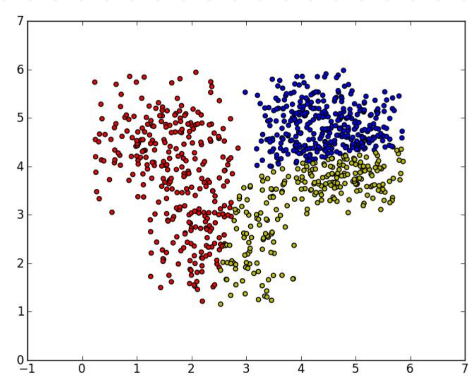
Machine Learning

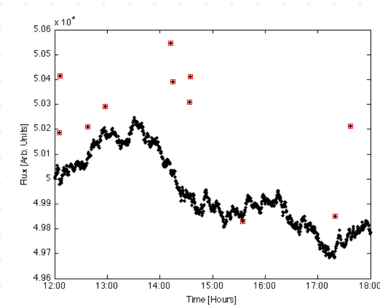
- "A computer program is said to *learn* from experience, E, with respect to some class of tasks, T, and performance measure, P, if its performance at tasks in T, as measured by P, improves with experience, E." — Tom Mitchell, *Machine Learning*
- "*Machine learning* algorithms have proven to be of great practical value in a variety of application domains. They are especially useful in *data mining problems*…" — Tom Mitchell, *Machine Learning*

Data Science Functionalities

- Generalization
- Visualization
- Frequent pattern mining and association mining
- Classification
- Clustering
- Outlier analysis



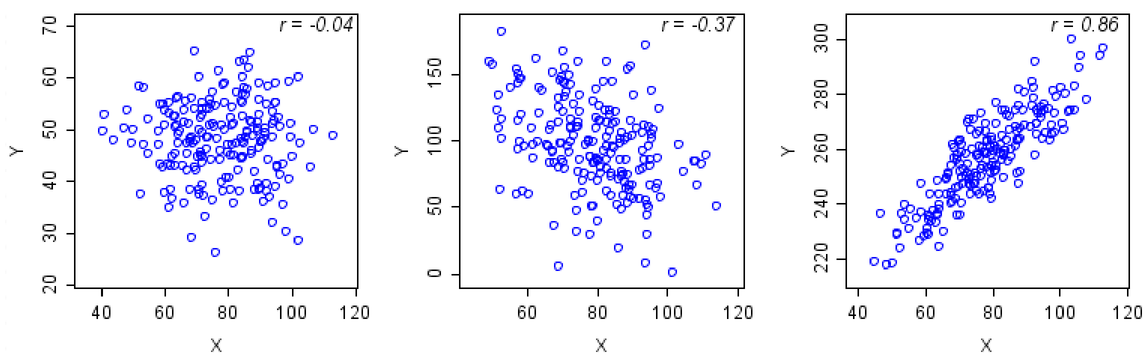




Pattern Discovery

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?

- A typical association rule
  - Diaper and Beer [0.5%, 75%] (support, confidence)
  - **Support:** the proportion of transactions in the dataset which contains the itemset (Diaper, Beer).
  - **Confidence:** the proportion of the transactions that contains Diaper which also contains Beer.

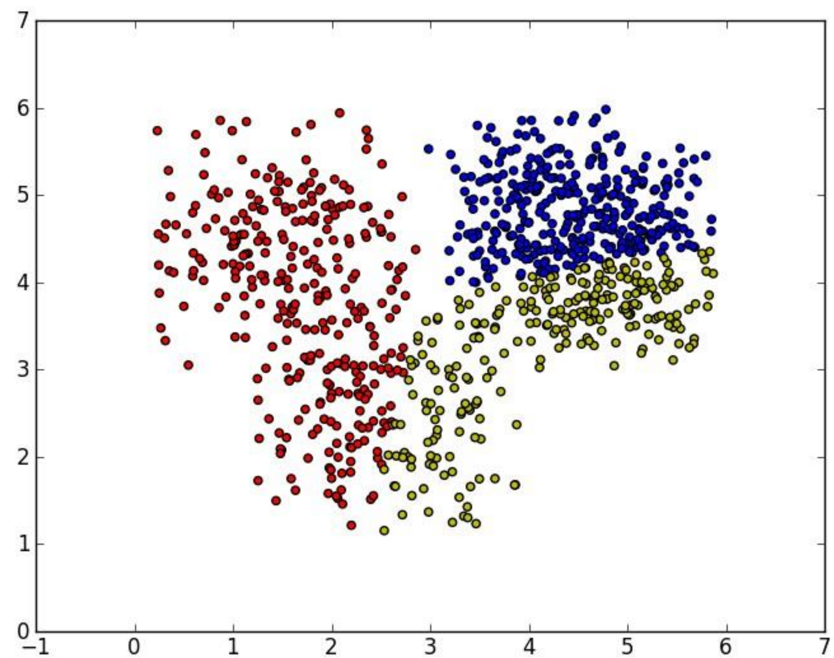


Classification

- Classification and label prediction
  - Construct models (functions) based on some **training** examples
  - Describe and distinguish classes or concepts for future prediction
    - Ex. 1. Classify countries based on (climate)
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown **class labels**
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

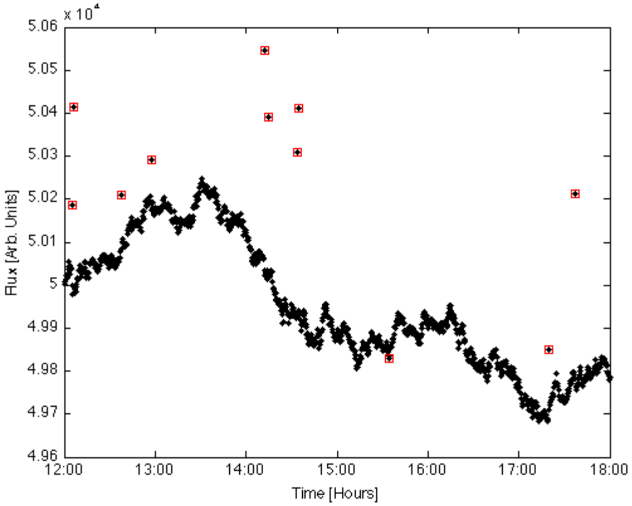
Clustering

- Unsupervised learning (i.e., class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-classsimilarity & minimizing interclass similarity



Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behaviorof the data
  - Methods: by product of clustering or regression analysis…
  - Useful in fraud detection, rare events analysis



Data Science Tasks

- **Descriptive Tasks**
  - Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that are able to summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.
- **Predictive Tasks**
  - The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Data Science Challenges

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in **multi-dimensional** space
  - Data mining: An **interdisciplinary** effort
  - Boosting the power of discovery in a networked environment
  - Handling **noise, uncertainty, and incompleteness** of data
  - **Pattern evaluation** and pattern- or constraint-guided mining
- User Interaction
  - **Interactive** mining
  - Incorporation of **background knowledge**
  - **Presentation and visualization** of data mining results

Data Science Challenges

- Efficiency and Scalability
  - **Efficiency and scalability** of data mining algorithms
  - **Parallel, distributed, stream, and incremental** mining methods
- Diversity of data types
  - Handling **complex** types of data
  - Mining **dynamic, networked, and global** data repositories
- Data mining and society
  - **Social** impacts of data mining
  - **Privacy-preserving** data mining
  - **Invisible** data mining

Expect and Not Expect

- Expect to have:
  - The *first tiny* step of being a "data scientist"
- Don't expect to have:
  - *State-of-the-art* machine learning/AI models
    1. ______________
    2. ______________
  - *All* skills that your start-up idea requires
    1. ______________
    2. ______________
    3. ______________

Concrete Learning Goals

- **Can process raw data: data cleaning, data integration, data reduction, dimension reduction**
- Can describe data warehouse, OLAP, data cube concepts and technology that work on multi-dimensional datasets
- **Can use Apriori and FP-Growth for frequent pattern mining**
- Can describe diverse patterns, sequential patterns, graph patterns
- **Can use Decision Tree, Naïve Bayes, Ensembles for classification**
- Can describe SVMs and Neural Networks for classification
- **Can use K-Partitioning Methods (K-Means, etc.) for clustering**
- Can describe Kernel-based Clustering and Density-based Clustering
- **Can use appropriate measures to evaluate results of different functionalities**

Syllabus and Schedule

08-22T	**Introduction**	10-12R	Classification: Naïve Bayes
08-24R	Data description	10-24T	Classification: Evaluation
08-29T	Data visualization	10-26R	Classification: Ensembles
08-31R	**Project introduction**	10-31T	Classification: SVMs
09-05T	Data cleaning and data integration	11-02R	Classification: Neural networks
09-07R	Data reduction and dimension reduction	11-07T	Clustering: Concepts
09-12T	Data cube: Concepts and operations	11-09R	Clustering: Partitioning methods
09-14R	Data cube: Data warehouse and OLAP	11-14T	Clustering: Kernel-based
09-19T	Frequent pattern mining: Apriori	11-16R	Clustering: Density-based
09-21R	Frequent pattern mining: FP-Growth	11-21T	Clustering: Evaluation
09-26T	Frequent pattern mining: Evaluation	11-28T	**Course review 2**
09-28R	Frequent pattern mining: Beyond itemset	11-30R	**Course review 3**
10-03T	**Course review 1**	12-05T	**Project presentation 1**
10-05R	**Mid-term**	12-07R	**Project presentation 2**
10-10T	Classification: Decision tree induction	12-12T	**Final**

Five Written Assignments and One Project

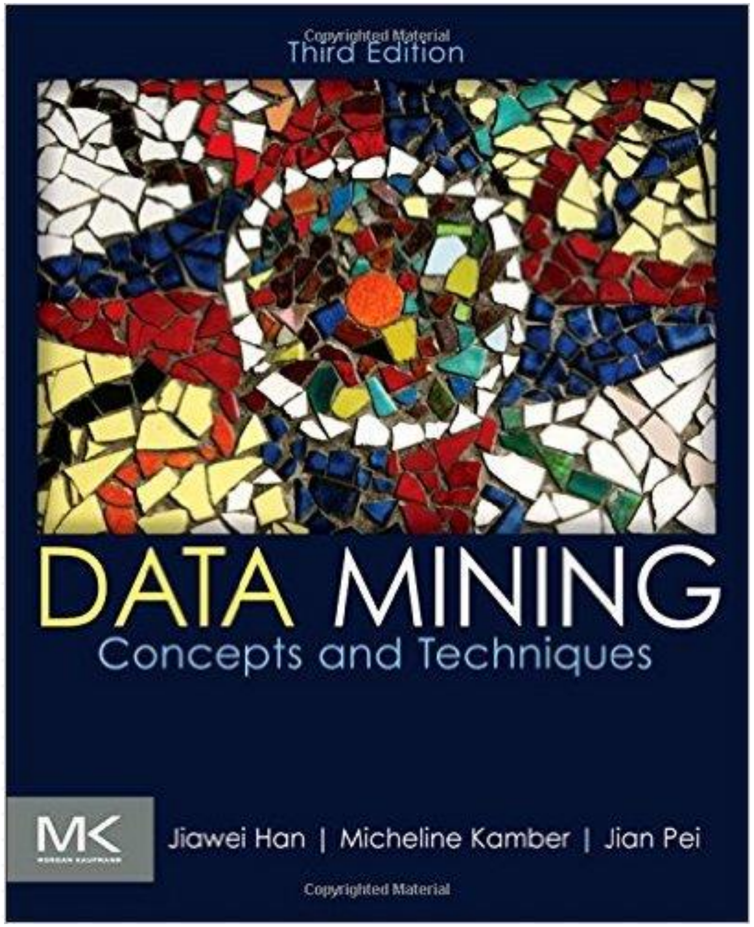
08-22T	Introduction	10-12R	HW4 out
08-24R	Data processing	10-24T	
08-29T	HW1 out	10-26R	
08-31R	Project introduction Project out	10-31T	
09-05T		11-02R	
09-07R		11-07T	Clustering
09-12T	Data cube HW1 due, HW2 out	11-09R	HW4 due, HW5 out
09-14R		11-14T	
09-19T	Frequent pattern mining	11-16R	
09-21R	HW2 due, HW3 out	11-21T	
09-26T		11-28T	Course review 2 HW5 due
09-28R		11-30R	Course review 3 Project due
10-03T	Course review 1 HW3 due	12-05T	Project presentation 1
10-05R	Mid-term	12-07R	Project presentation 2
10-10T	Classification	12-12T	Final

Grading

- **Uniform grading policy for undergraduates**
- **Individual HWs: 25%** = 5% * 5
- **Individual project: 25%** (Graduates are graded separately)
  - "Data science research bot"
  - Fed with *thousands* of data science publications
  - QA with *discovered knowledge*: Help data scientists on their research
  - Techs
    - Data cube: Paper/expert recommendation
    - Frequent pattern mining and classification: Entity recognition
    - Classification: Entity typing ($Problem, $Method, $Dataset, $Metric, $Digit…)
    - Clustering: Entity clustering
    - Evaluations
    - *Inference and prediction
  - Monitored in HWs (cube stat., ten most freq. patterns, etc.); volunteer to present and be graded by classmates and instructor; others graded by the instructor
- **Mid-term: 20%**
- **Final: 30%**
- **No quiz.**

Textbook

- Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques (3rd ed.), Morgan Kaufmann, 2011
- Our lecture does *not cover all* the content of the book.
- We provide lecture notes from the 2nd ed. of the text book.



Time and Location

- Lecture: 2:00 pm – 3:15 pm (Tuesday and Thursday), DeBartolo Hall 140
- Office hour: 3:30 pm – 4:30 pm (**Thursday**), Cushing Hall 326C
- Teaching Assistant: Qi Li (qli8)
- TA hour: 3:30 pm – 4:30 pm (**Tuesday**), Fitzpatrick Hall 247A
- Website (slides): http://www.meng-jiang.com/teaching-csex0647.html
- Forum: (Piazza) https://piazza.com/class/j6dmfs52c6d50v

References

- Charu C. Aggarwal, Data Mining: The Textbook, Springer, 2015
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005 (2nd ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014