
CSE 40647/60647
Data Science

The Instructor

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

B.S. and Ph.D.

Visiting Ph.D.

Postdoc Researcher

Assistant Professor

Visiting Researcher

Visiting Researcher

Chapter 1. Introduction

Meng Jiang
Data Science

Today

- Know general/concrete learning goals;

- Describe what is data science;
- Describe components of data science research;
- Describe data science functionalities.

- Know syllabus and class schedule;
- Know course project and project schedule;
- Know grading policy;
- Know time, location, and textbook.

General Learning Goals

- Learn *basic* data science concepts
- Learn *basic* methods of mining knowledge from data
- Prerequisites:
  - Programming with *Python*
  - Data structures and Algorithms
- As a prerequisite for:
  - CSE 40625/60625: Machine Learning
- You will see *Concrete Learning Goals* soon.

Expected and Not Expected

- 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. _____

What is Data Science?

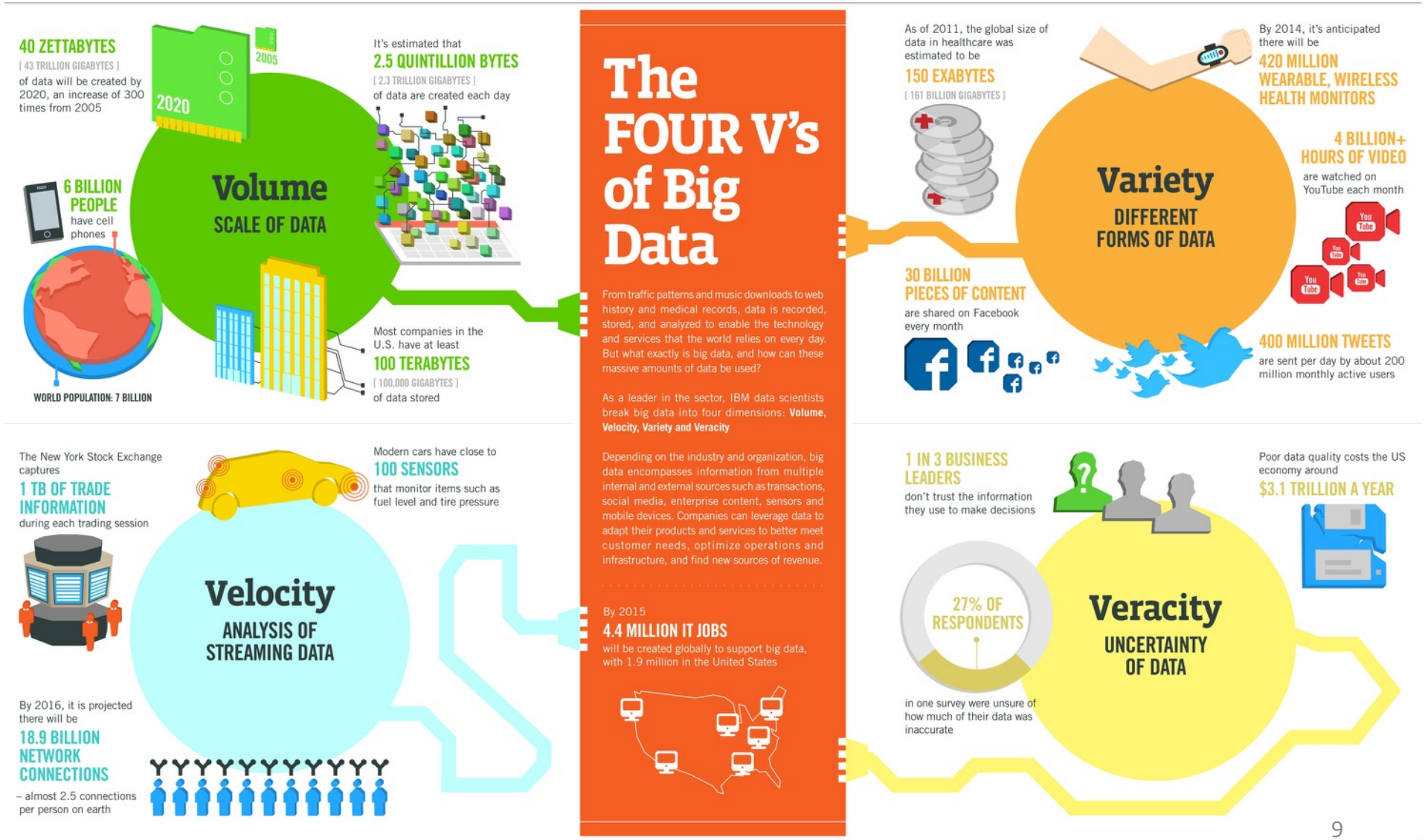
- "…the process of automatically discovering *useful information* in *large* repositories of data." — *Introduction to Data Mining* (Tan, Steinbach, & Kumar)
- "…the process of discovering *patterns* in data." — *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition* (Witten, Frank,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

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 is often so large that the process of extracting knowledge must be automated – or at least augmented – by computer.

Big Data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

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."

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.

[ ] Identifying strongly significant genes when doing multiple tests across many genes.

[ ] Finding the most popular hobby among us.

[ ] Inferring a student's hobby.

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.

The dataset is often not large.

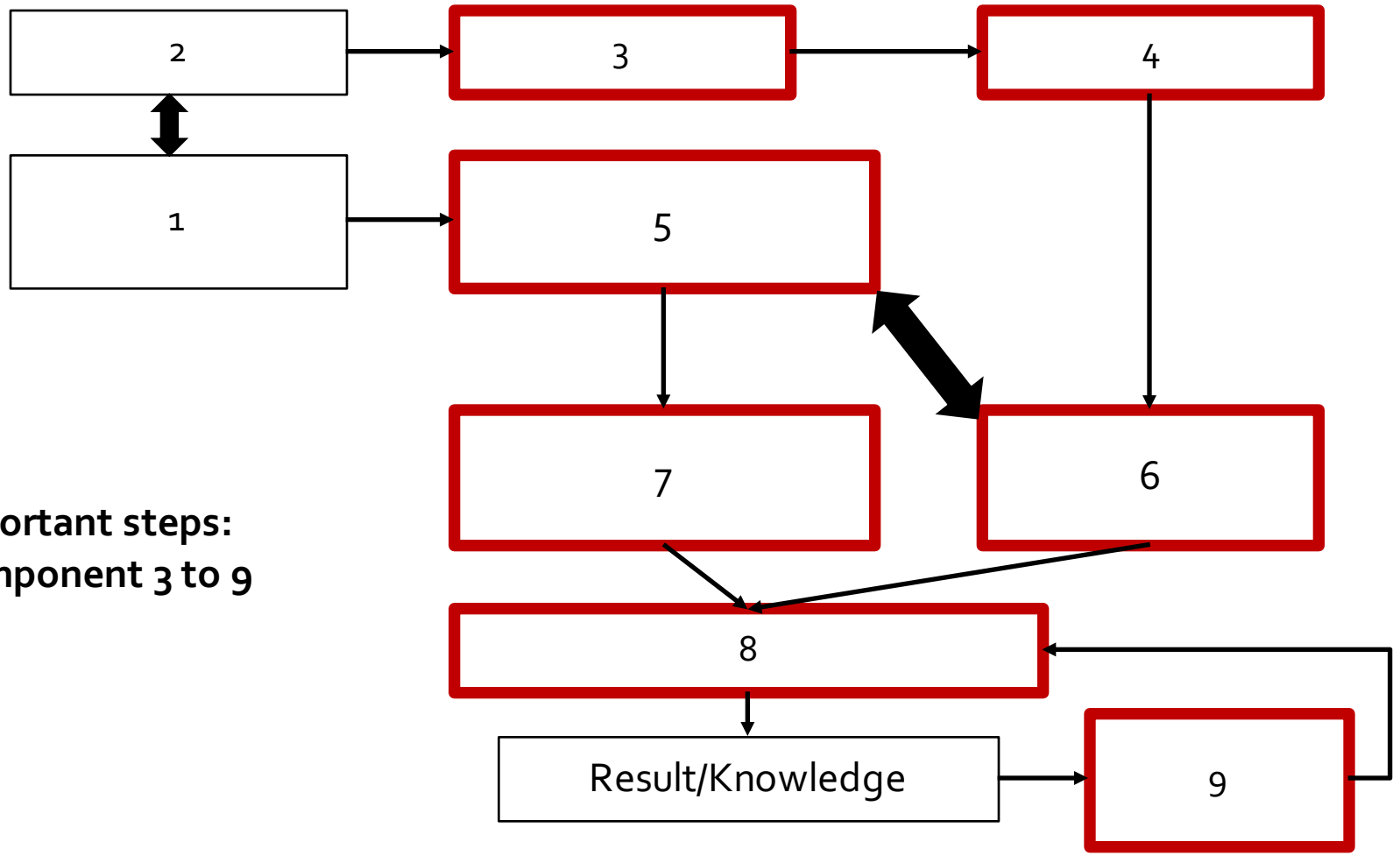
[√] Identifying strongly significant genes when doing multiple tests across many genes.

[×] Finding the most popular hobby among us.

[√] Inferring a student's hobby.

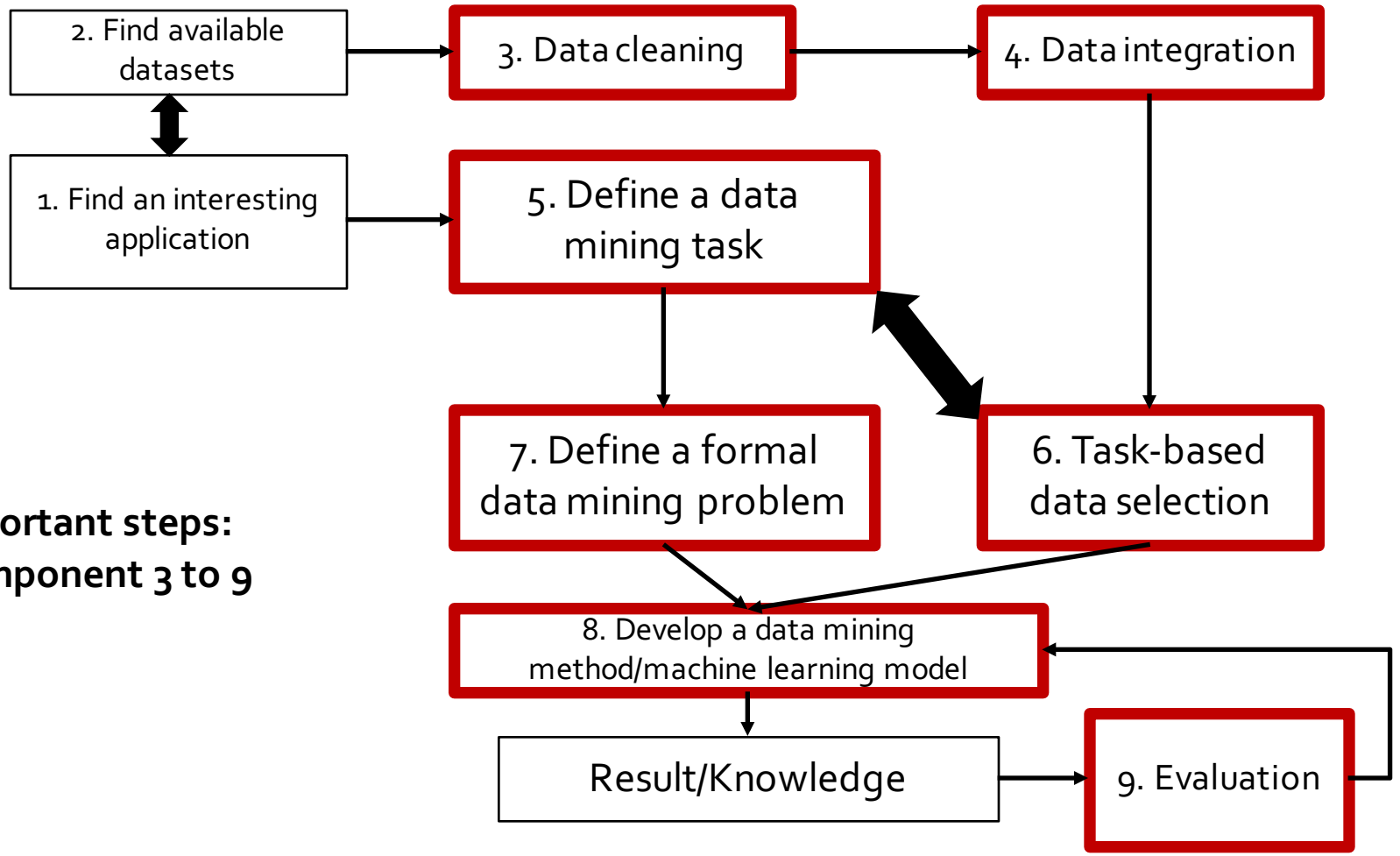
What is the first step of finding the most popular hobby?

Components of Data Science Research



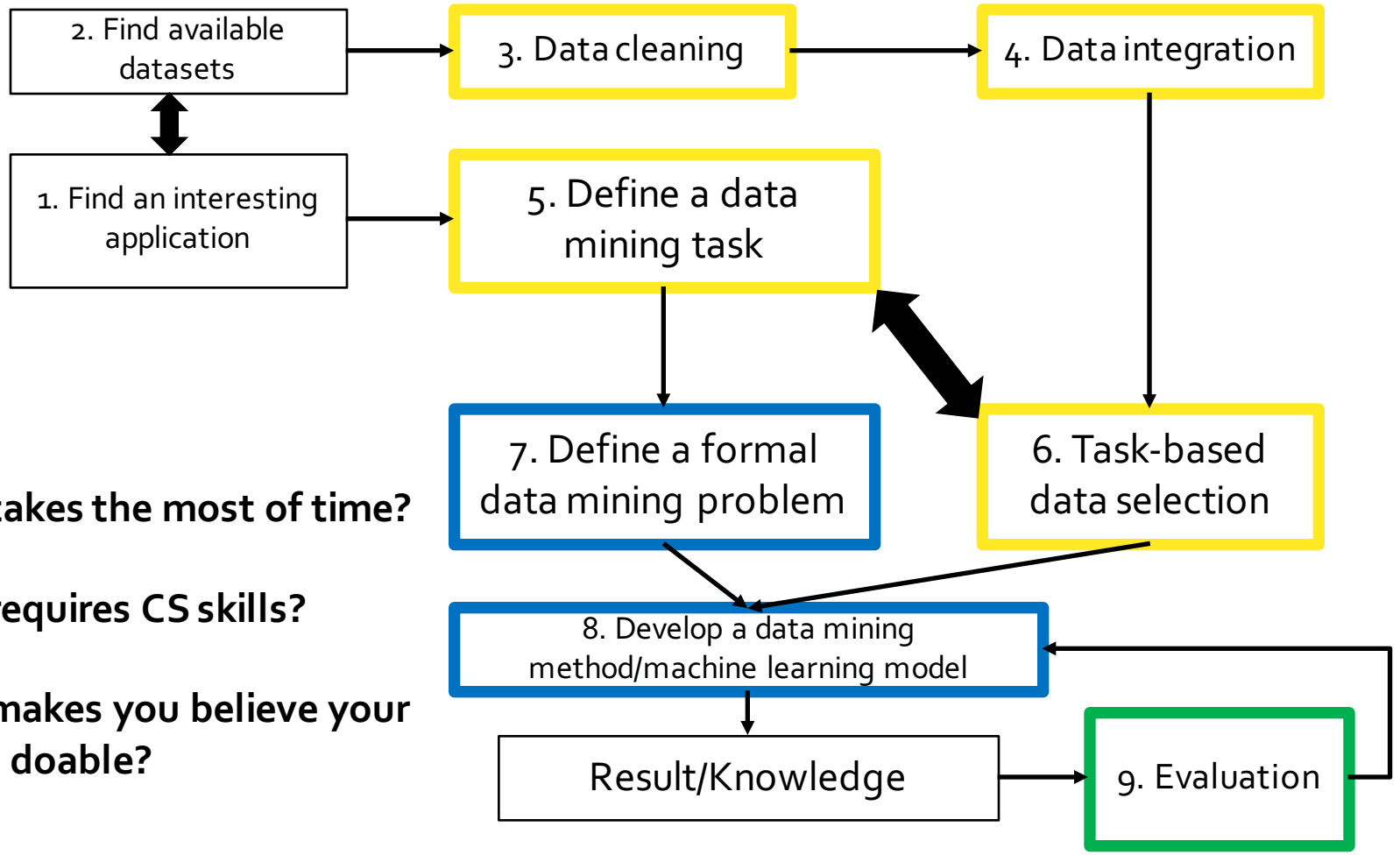
Important steps:
Component 3 to 9

Components of Data Science Research



Important steps:
Component 3 to 9

Components of Data Science Research



Which takes the most of time?

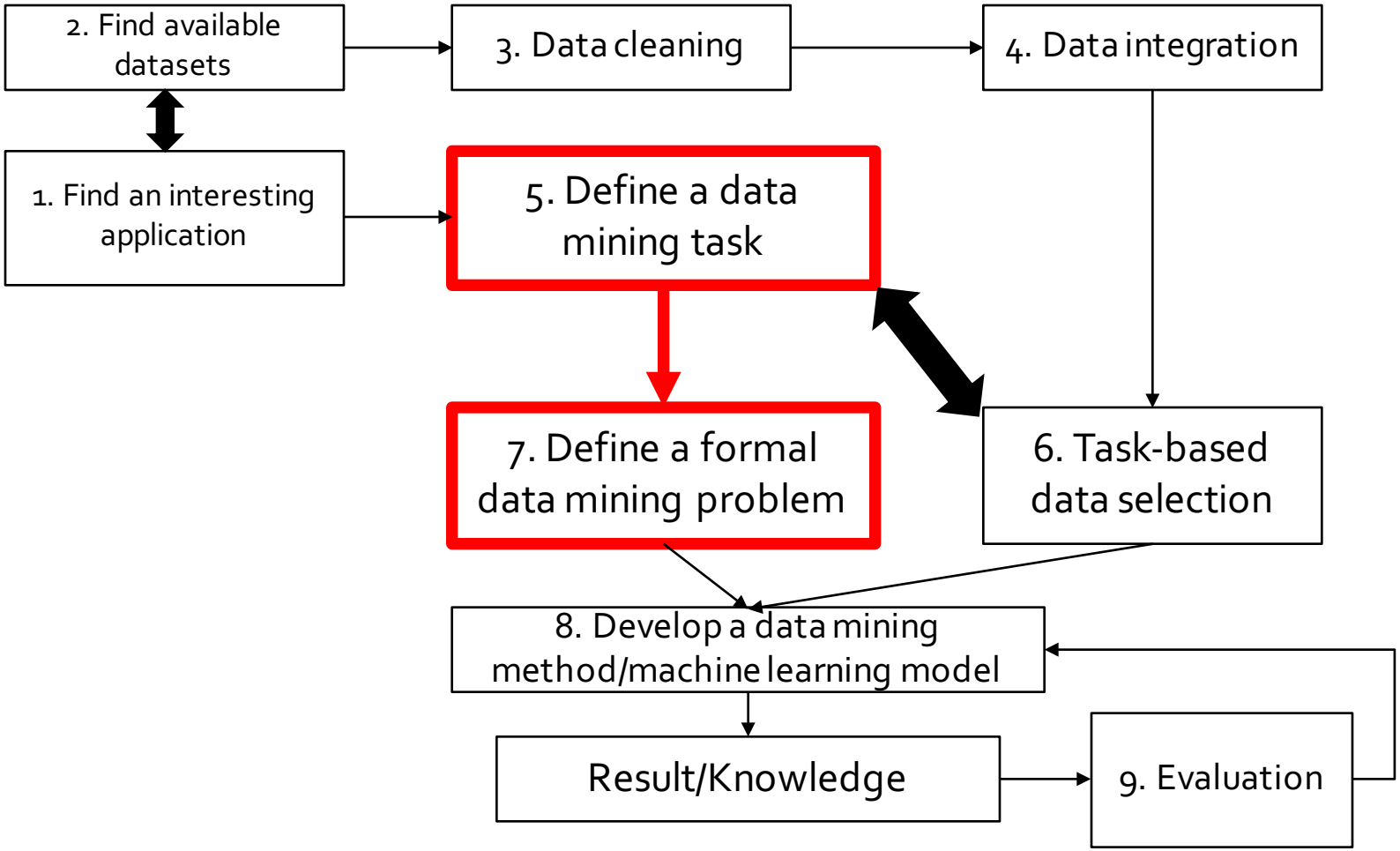
Which requires CS skills?

Which makes you believe your study is doable?

Machine Learning and Data Mining

- "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*

What are Data Science Functionalities?

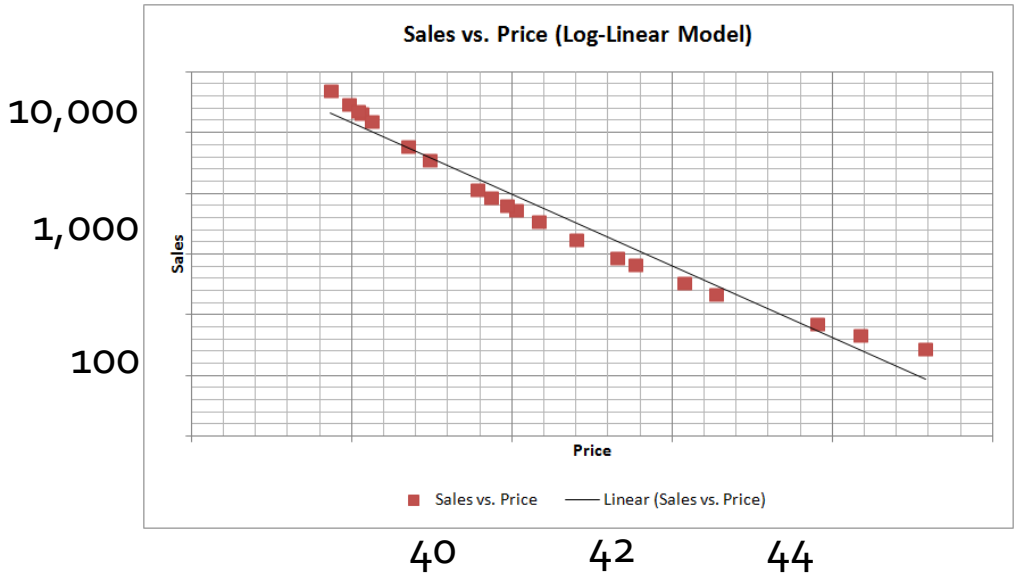


Classification

daynight = Classify[{

Regression


Happiness vs. Sleep
Happiness
Hours Slept
0
2
4
6
8
10
12
14


Sales vs. Price (Log-Linear Model)
10,000
1,000
100
Sales
Price
40
42
44
Sales vs. Price
Linear (Sales vs. Price)

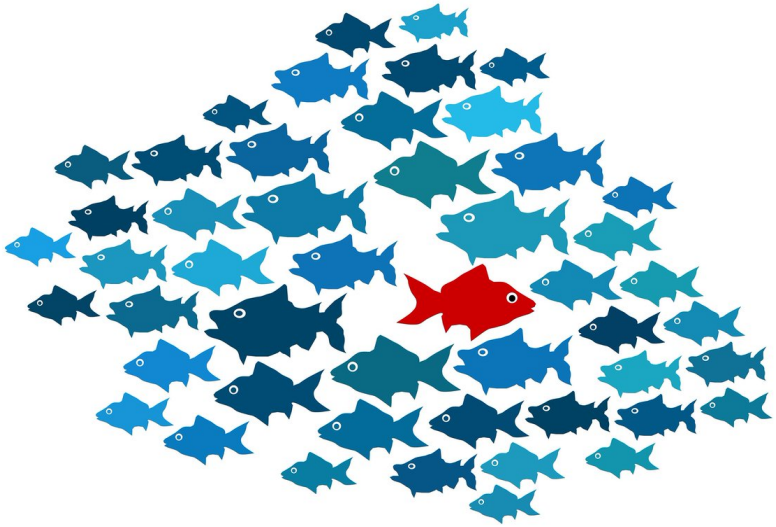
Clustering

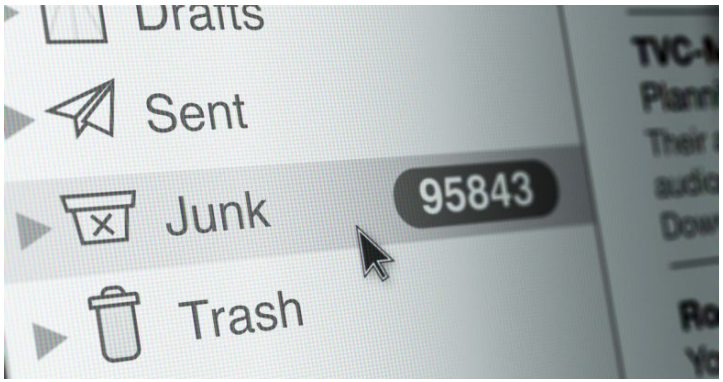
World News 2008

I have a Dream - Martin Luther King Jr.

Pattern/Association Mining

Outlier Detection





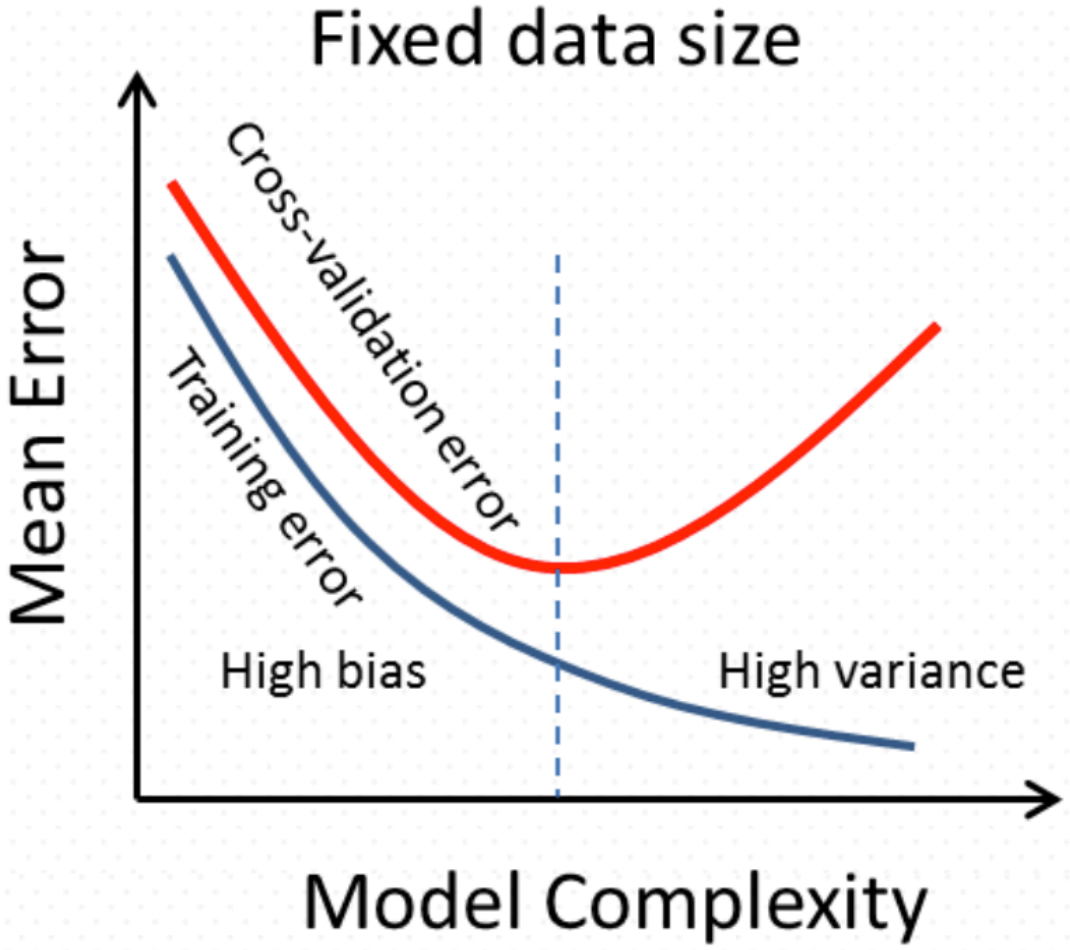
Data Mining Tasks

- Classification: the task of assigning (discrete) target variables to one of several predefined categories.
- Regression: the task of finding a function that models (continuous) target variables.
- Clustering: the task of discovering groups and structures.
- Outlier/Anomaly Detection: the task of detecting unusual deviations.
- Pattern/Association Analysis: the task of discovering patterns that describe relationships.

Overfitting and Generalization

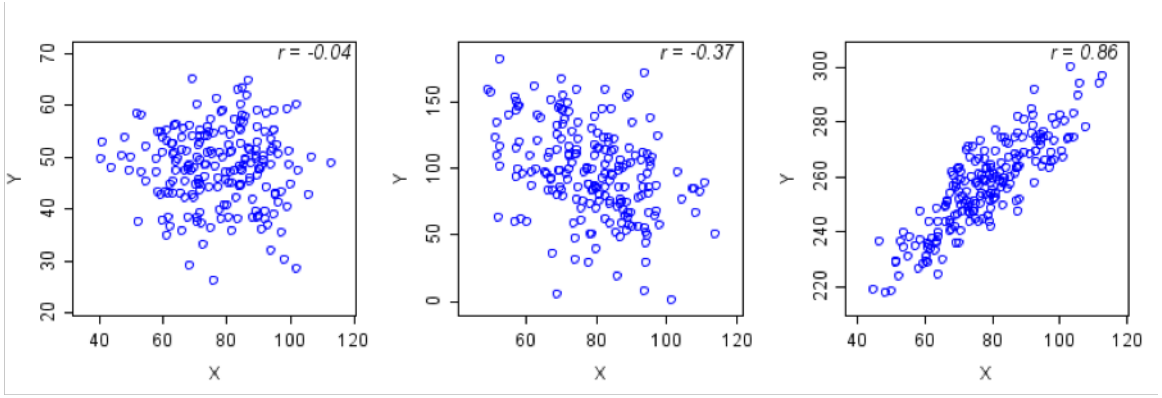
- Overfitting means doing very well on training data but poorly on the test data, lest test data exactly mimics the training data.
  - Counterintuitive?
  - Think about preparing for an exam. What is better: rote memorization or understanding a concept?
  - Learner could not induce a function that generalizes well.
- Generalization is often a tenet of machine learning.

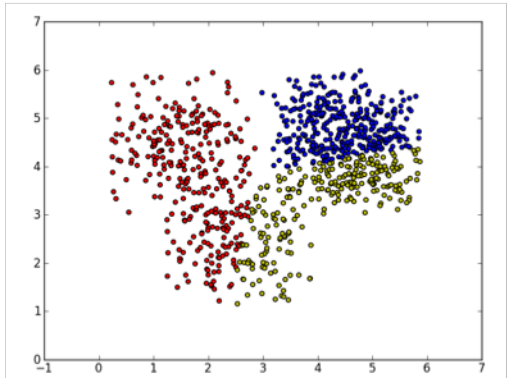
Generalization Behavior

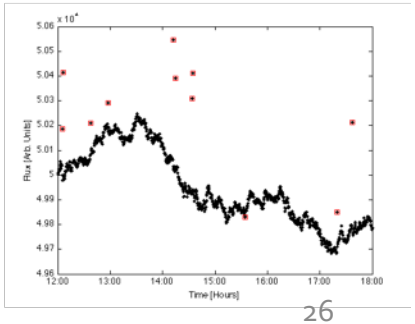


Data Science Functionalities

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







Defining a Data Mining Task

- Generate a problem statement.
- Utilize background knowledge.
- Posit the right question.
- Understand the data.
- Implement one or more modeling approaches.
- Identify performance measurement criteria.
- Interpret the model(s).
- Visualize and present the results.

Concrete Learning Goals

- Syllabus: http://www.meng-jiang.com/teaching/CSE647Spring18-Syllabus.pdf
- At the end of the course, students will be able to:
  - **Use** raw data processing techniques: data description, data visualization, data cleaning, data integration, data reduction, and dimension reduction
  - **Use** Decision Trees, Naïve Bayes, and SVMs for classification
  - **Describe** Ensembles and Neural Networks models for classification
  - **Use** K-Partitioning methods for clustering
  - **Describe** hierarchical clustering, kernel-based clustering and density-based clustering
  - **Use** Apriori and FP-Growth for frequent pattern mining and association mining
  - **Describe** diverse patterns, sequential patterns, graph patterns
  - **Use** appropriate measures to evaluate results of different functionalities (classification, clustering, and frequent pattern mining)

Grading

- **Individual HWs:** 20% = 5% * 4
  - Written + programming
- **Team project:** 30%
- **Mid-term exam:** 20%
  - Data processing + classification
- **Final exam:** 30%
  - Clustering + pattern/association mining

Project Instruction

- http://www.meng-jiang.com/teaching/CSE647Spring18-Project.pdf
  - Requirement
  - Grading
  - Schedule

Grading distribution: (100 points)

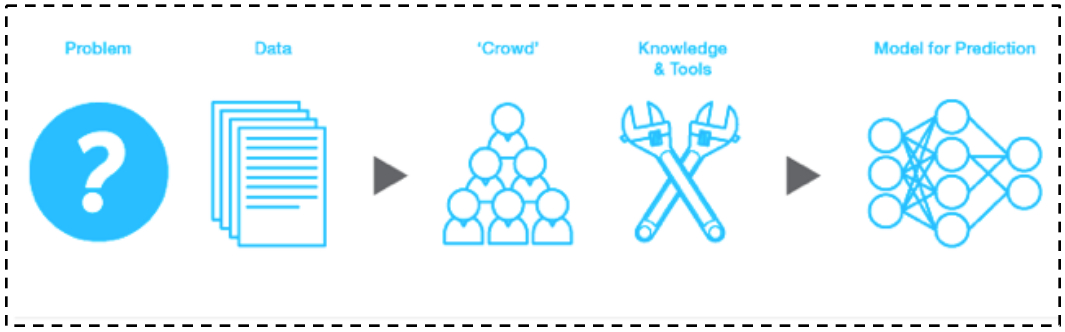
- **Proposal paper (10 points)**
- **Milestone presentation (10 points)**
- **Milestone paper (5 points)**
- **Poster (25 points)**
- **Final term paper (25 points)**
- **Code package and data (25 points)**
- **Oral presentation (+10 points)**

Schedule:

Date	Lecture#	Topic	Goals
01-16 (T)	1	Introduction	Understand what is data science research Know project grading policy and schedule Start looking for your **teammates** and find them ASAP Start looking for **interesting and doable** topics ASAP
02-06 (T)	7	Proposal: Teaming and proposal	Write down your teammate names in HW1 and proposal paper Submit your proposal paper: • **What is your project topic/research problem?** • **How will you find your dataset?** • **What is your proposed method?** You will listen to proposals from your classmates. This may help you if you still want to improve your idea.
03-06 (T)	14	QA	In case that you need to discuss about your project and you don't have time to come to office hours, we offer a great chance for you to briefly introduce your idea in class – everybody in the class will be happy to help you! **Keep in mind: In two days, you'll submit your milestone paper and give a presentation.**
03-08 (R)	15	Milestone	Submit your milestone paper: • **Your topic, dataset, and method** • **Milestone progress: Some preliminary results** • **Challenges and proposed solutions** • **Plan for the next two months** You will give milestone presentations in class. Believe me: Audience will help you, not argue with you.
04-26 (R)	26	Oral and QA	Volunteer to present in class on your full result. **Extra credit will be offered.** We have QA session.
05-01 (T)	27	Poster and project due	Every team makes a poster and shows in class. **Classmates will evaluate your poster.** You have to submit your code package, data, and term paper at 11:59PM this date.

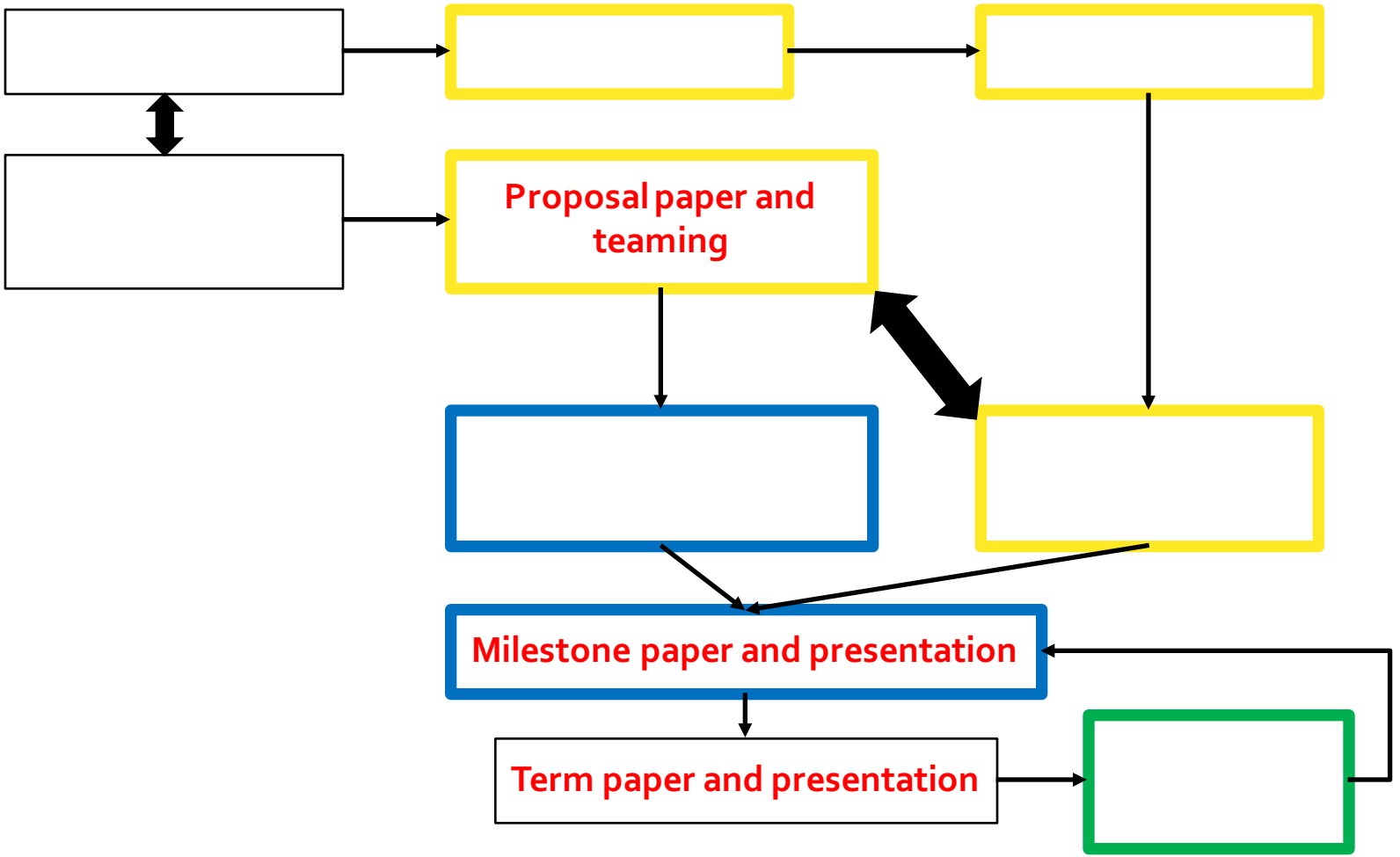
Some Data Portals

- Kaggle: https://www.kaggle.com/



- DATA.GOV: https://www.data.gov/
- City of Chicago Data Portal: https://data.cityofchicago.org/
- City of South Bend Open Data: http://data-southbend.opendata.arcgis.com/
- Index of Complex Networks: https://icon.colorado.edu/
- The Koblenz Network Collection: http://konect.uni-koblenz.de/
- Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/

Project


Proposal paper and teaming
Milestone paper and presentation
Term paper and presentation

Paper Template, No Page Limit

ACM Conference Proceedings Template: https://www.acm.org/publications/proceedings-template

Example paper below: http://www.meng-jiang.com/pubs/metapad-kdd17/metapad-kdd17-paper.pdf

Abstract

Introduction

Related Work

Methodology

Methodology

Introduction

Related Work

Problem Definition

Methodology

Experiments

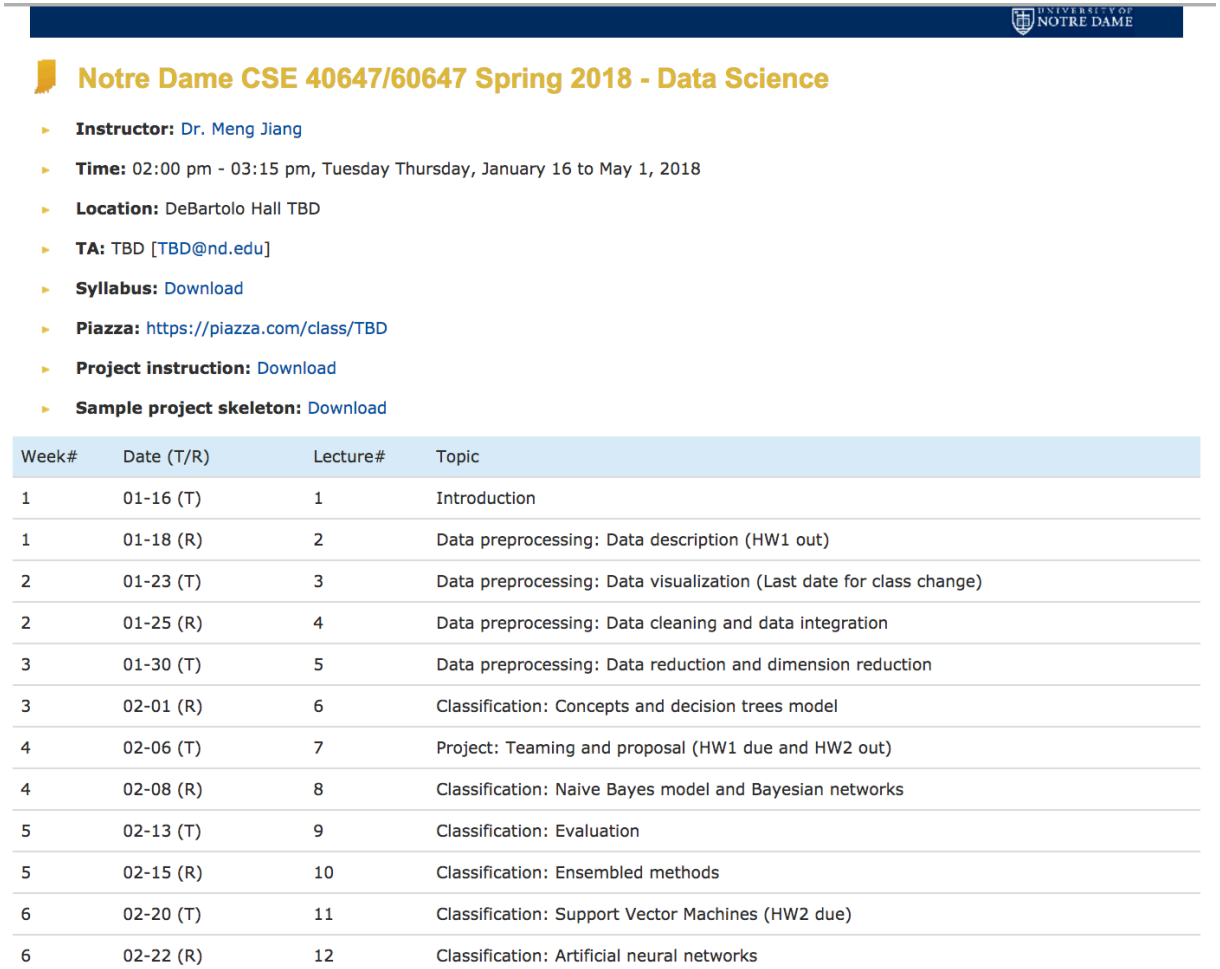
Experiments

Experiments

Acks and References

Course Website and Schedule

- http://www.meng-jiang.com/teaching-cse647-s18.html



UNIVERSITY OF NOTRE DAME

Notre Dame CSE 40647/60647 Spring 2018 - Data Science

- **Instructor:** Dr. Meng Jiang
- **Time:** 02:00 pm - 03:15 pm, Tuesday Thursday, January 16 to May 1, 2018
- **Location:** DeBartolo Hall TBD
- **TA:** TBD [TBD@nd.edu]
- **Syllabus:** Download
- **Piazza:** https://piazza.com/class/TBD
- **Project instruction:** Download
- **Sample project skeleton:** Download

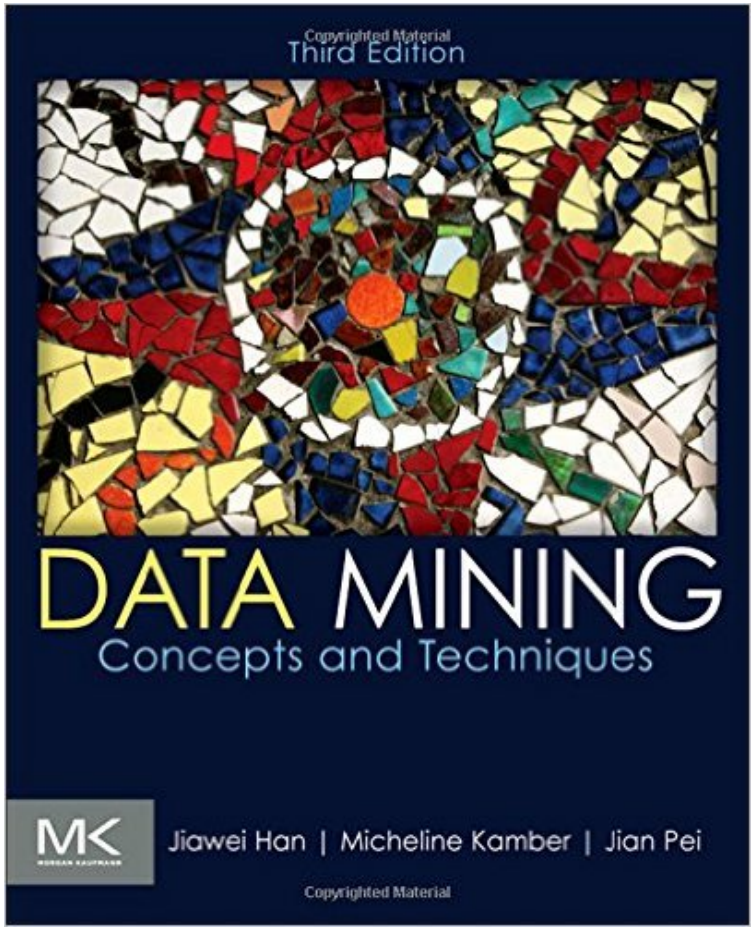
Week#	Date (T/R)	Lecture#	Topic
1	01-16 (T)	1	Introduction
1	01-18 (R)	2	Data preprocessing: Data description (HW1 out)
2	01-23 (T)	3	Data preprocessing: Data visualization (Last date for class change)
2	01-25 (R)	4	Data preprocessing: Data cleaning and data integration
3	01-30 (T)	5	Data preprocessing: Data reduction and dimension reduction
3	02-01 (R)	6	Classification: Concepts and decision trees model
4	02-06 (T)	7	Project: Teaming and proposal (HW1 due and HW2 out)
4	02-08 (R)	8	Classification: Naive Bayes model and Bayesian networks
5	02-13 (T)	9	Classification: Evaluation
5	02-15 (R)	10	Classification: Ensembled methods
6	02-20 (T)	11	Classification: Support Vector Machines (HW2 due)
6	02-22 (R)	12	Classification: Artificial neural networks

Time and Location

- Lecture: 2:00 pm – 3:15 pm (**Tuesday** and **Thursday**), DeBartolo Hall 117
- Office hour: 3:30 pm – 4:30 pm (**Friday**), Cushing Hall 326C
- Teaching Assistant: Qi Li (qli8@nd.edu)
- TA hour: 3:30 pm – 4:30 pm (**Monday**), Cushing Hall 212C
  - HW due is often on **Tuesday** ☺
- Website (slides): http://www.meng-jiang.com/teaching-cse647-s18.html
- Piazza: https://piazza.com/class/jcan36kl02c1e9

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.



Conclusion

- Know general/concrete learning goals;

- Describe what is data science;
- Describe components of data science research;
- Describe data science functionalities.

- Know syllabus and class schedule;
- Know course project and project schedule;
- Know grading policy;
- Know time, location, and textbook.

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