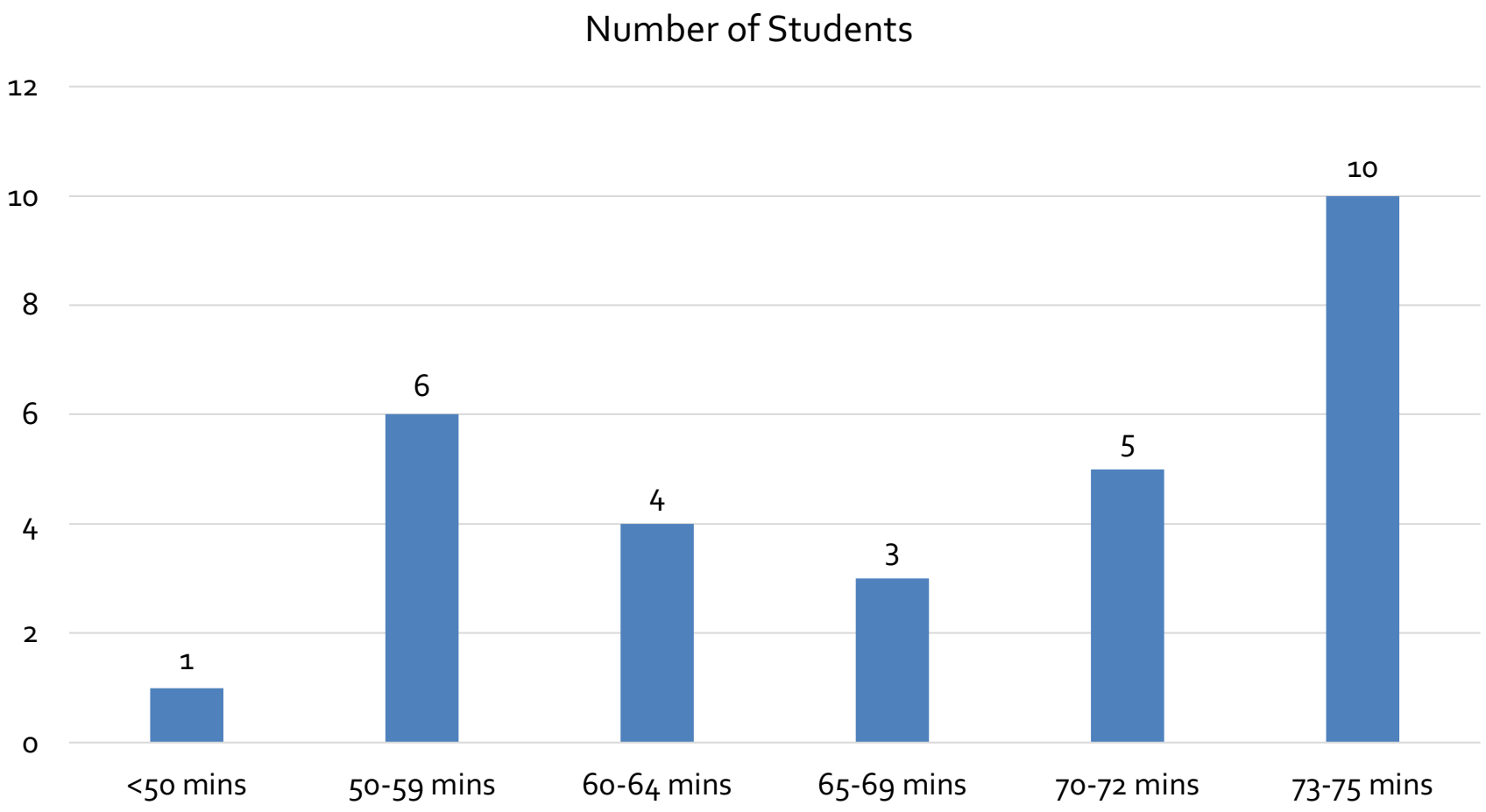
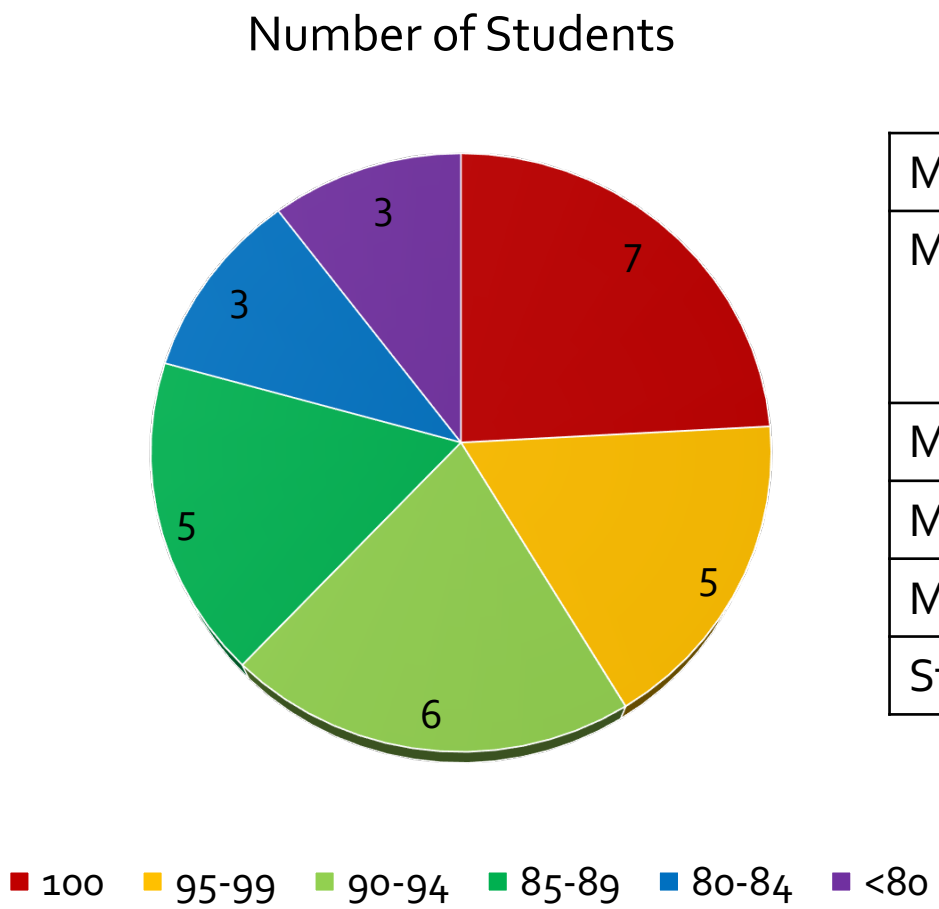
Mid-term Exam: Stats

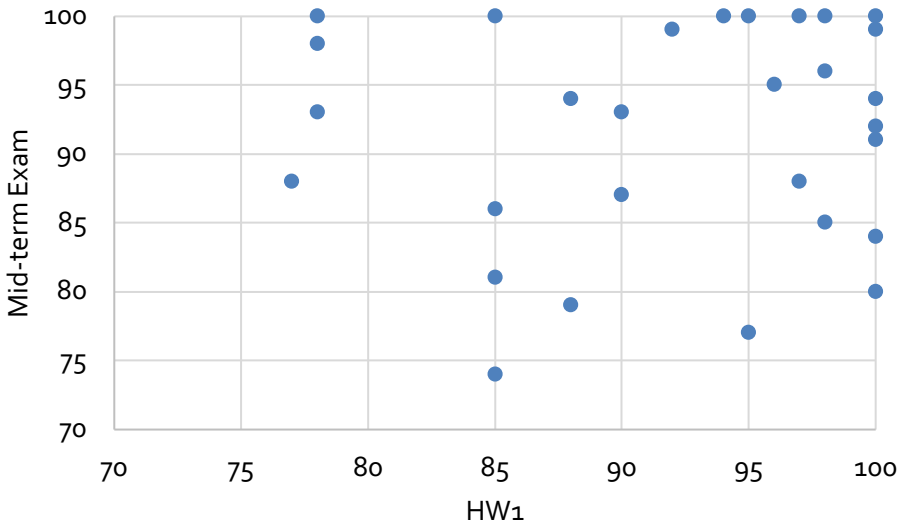


Mid-term Exam: Stats

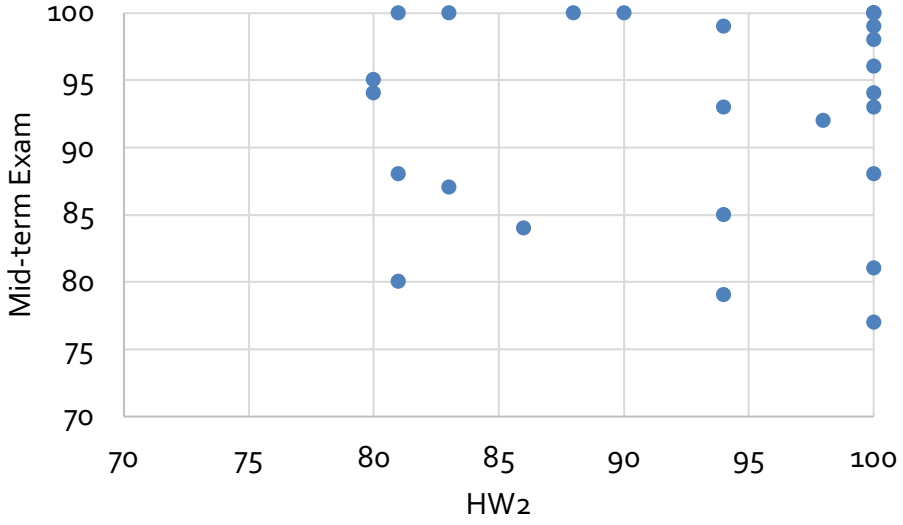


Min	74
Max	100 • 99+2 (3 students) • 98+2 (4 students)
Mean	91.5
Median	93
Mode	100
Std. Dev.	7.9

Correlation between Midterm and HW



Corr.(midterm, HW1) = 0.066



Corr.(midterm, HW2) = 0.183

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

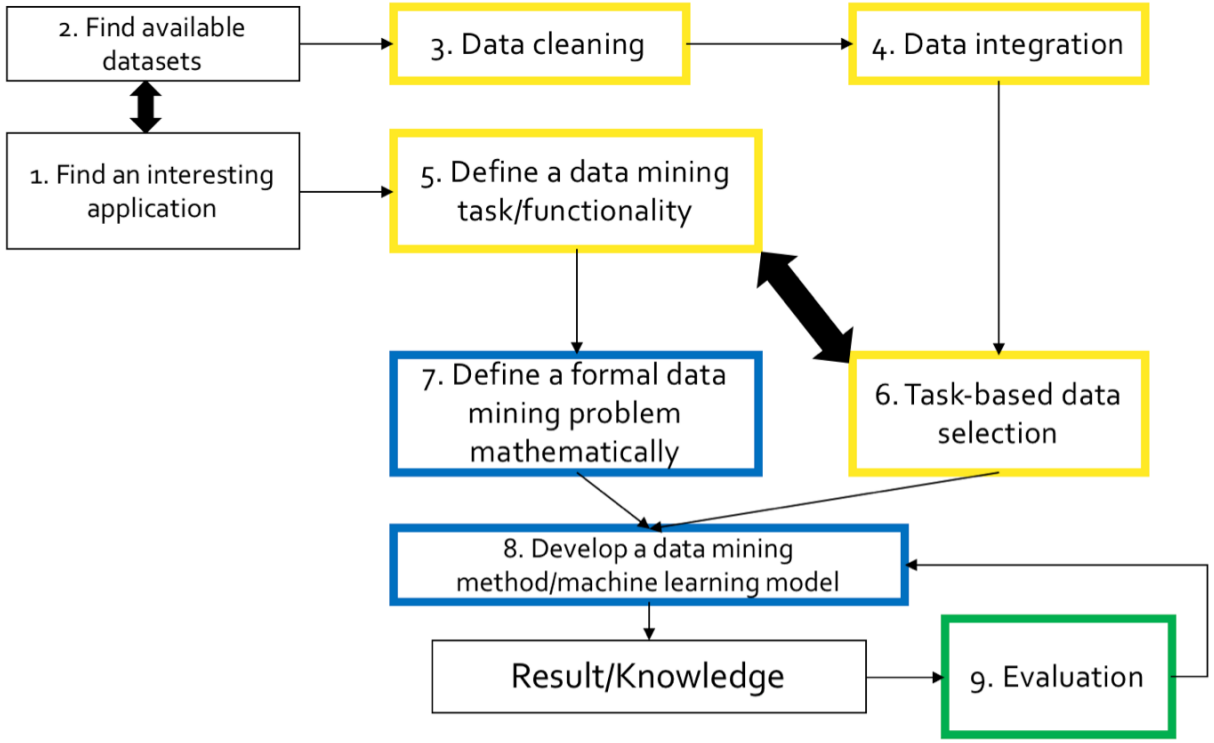
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q1: Name at least **five** steps in "Data Science Research" or called "Knowledge Discovery from Data" (KDD).

(Seven steps)



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q1: Name at least **five** steps in "Data Science Research" or called "Knowledge Discovery from Data" (KDD).

Any five of the following:

(1) Data cleaning
(2) Data integration
(3) Define data mining functionality/task
(4) Task-based data selection
(5) Define mathematical problem
(6) Data mining/machine learning (model construction)
(7) Evaluation

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

Q2: Given two data objects and four features, we have feature vectors of the two data objects as

(7, 4, −2, 1) and (4, 5, −1, 6).

What are the three specific Minkowski distance measures? Please give their names and calculate the distance measures between the two objects.

Chalk show!

Chapter 1:
Introduction (Jan. 16)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q3: Technique-to-Task

Technique	Data Preprocessing Task
1) Chi-square test	
2) Covariance analysis	a) Data cleaning: Handling missing data
3) Estimate/use parameters in linear regression	b) Data cleaning: Handling noisy data
4) Fill in with attribute mean	c) Data integration: Handling redundancy
5) Principle component analysis	d) Data reduction
6) Remove outliers using cluster analysis	e) Dimensionality reduction
7) Singular value decomposition	
8) Smooth feature values by bin means	
9) Stratified sampling	
10) Z-score normalization	

Chalk show!

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing

Mid-

(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q4: ID3, Naïve Bayes, and predictions.

Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

Chalk show!

Log Table

$\log_2(\frac{1}{5}) = -2.32$	$\log_2(\frac{2}{5}) = -1.32$	$\log_2(\frac{3}{5}) = -0.74$	$\log_2(\frac{4}{5}) = -0.32$
$\log_2(\frac{1}{4}) = -2$	$\log_2(\frac{2}{4}) = -1$	$\log_2(\frac{3}{4}) = -0.42$	
$\log_2(\frac{1}{3}) = -1.58$	$\log_2(\frac{2}{3}) = -0.58$		

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Artificial Neural Networks
(Feb. 22)

Q5: Compare the number of parameters:

(1) "4 (input) - 2 (hidden) - 2 (hidden) - 1 (output)" fully-connected feed-forward ANN;

(2) "4 (input) - 3 (hidden) - 1 (output)" fully-connected feed-forward ANN.

Chalk show!

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Evaluation (Feb. 15)

Q6: Confusion matrix, precision, recall, F1, and MAE.

ID	Ground truth label	Predicted label
1	Positive	Positive
2	Positive	Positive
3	Positive	Negative
4	Positive	Negative
5	Negative	Negative

Chalk show!

Reasoning in Project

- Suppose your project is making loans to customers.

Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

- Milestone: Try Naïve Bayes model.
- Preliminary results:

Test Customer ID	Prediction	Ground Truth
6	No	Yes
7	No	No
…	No	No
	No…	Yes…

Observation: So many "No" in prediction! Why?

Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

Analysis: If student="no", the prediction="no".

Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

Analysis: If income="high", the prediction="no".

Further Analysis

- Issue: Zero-probability
- Idea: Adding auxiliary instances fairly
- Concretely how? Having 5 + 16 = 21 training instances!

Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

1′	no	low	fair	no
2′	no	low	fair	yes
3′	no	low	excellent	no
…	…	…	…	…
16′	yes	high	excellent	yes

- Name it: "Laplace correction"
  - Because in statistics, additive smoothing, also called Laplace smoothing, is a technique used to smooth categorical data

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

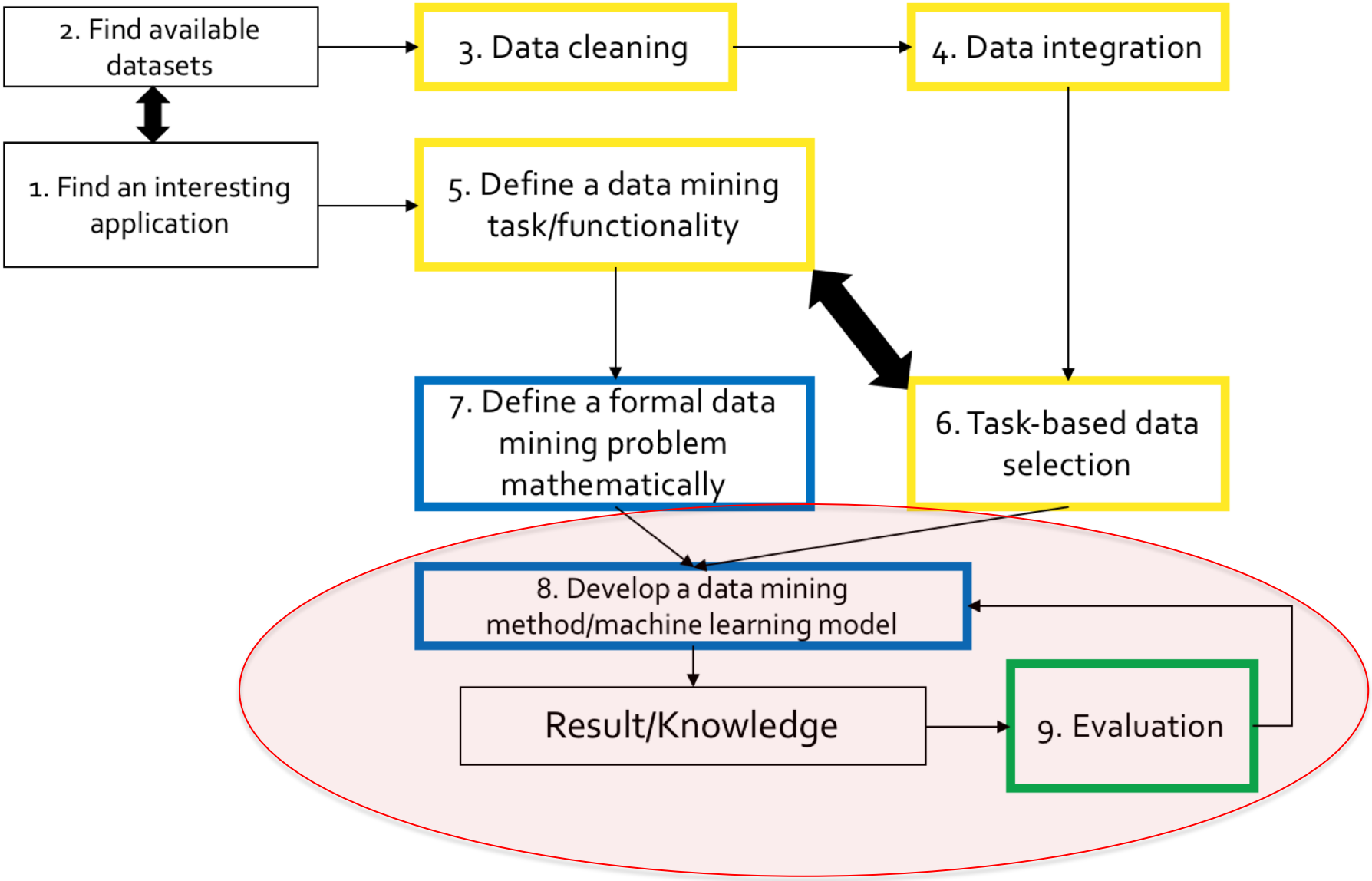
Borrow the idea! So… you'd better read more…

Implement Your Idea and Test It

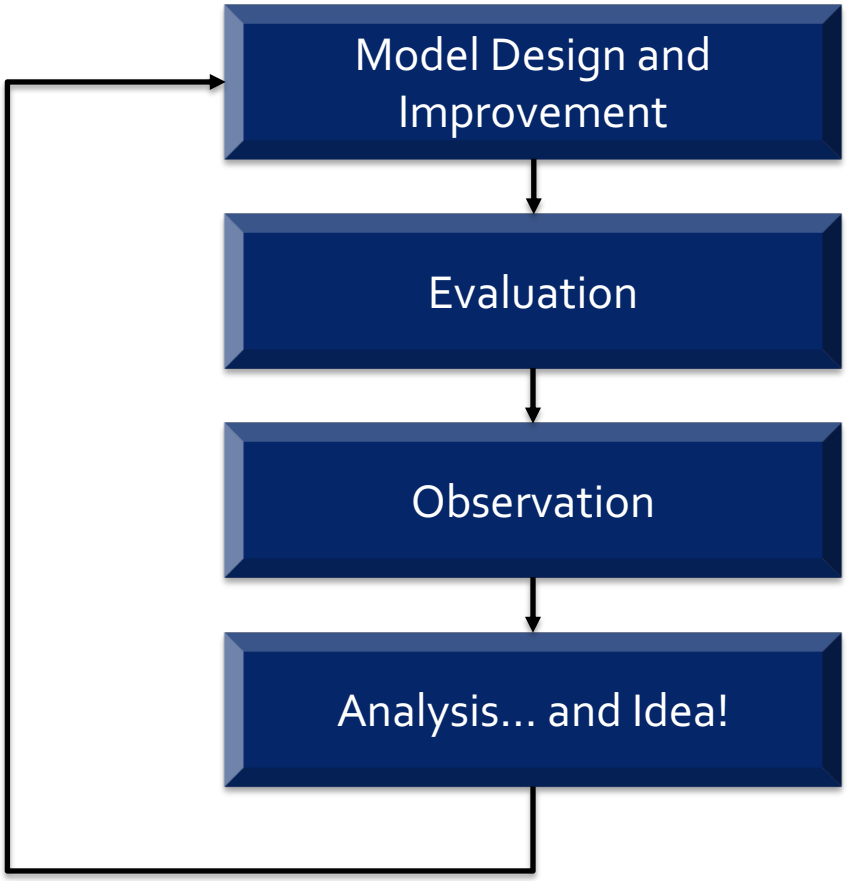
Customer ID	student	income	credit_rating	buys_computer
1	no	high	fair	no
2	no	low	excellent	no
3	yes	low	fair	yes
4	yes	low	excellent	yes
5	yes	low	fair	no
6	**yes**	**high**	**excellent**	**?**
7	**no**	**low**	**fair**	**?**

1′	no	low	fair	no
2′	no	low	fair	yes
3′	no	low	excellent	no
…	…	…	…	…
16′	yes	high	excellent	yes

Chalk show!

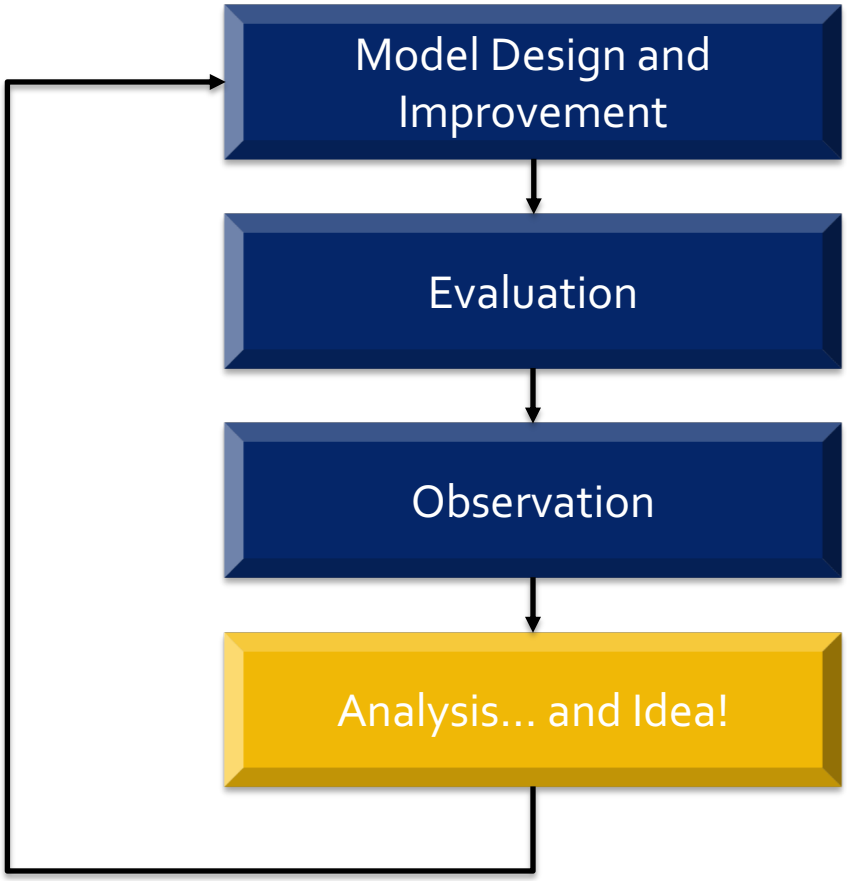

2. Find available datasets
3. Data cleaning
4. Data integration
1. Find an interesting application
5. Define a data mining task/functionality
7. Define a formal data mining problem mathematically
6. Task-based data selection
8. Develop a data mining method/machine learning model
Result/Knowledge
9. Evaluation

Basic Skills of Being a Data Scientist



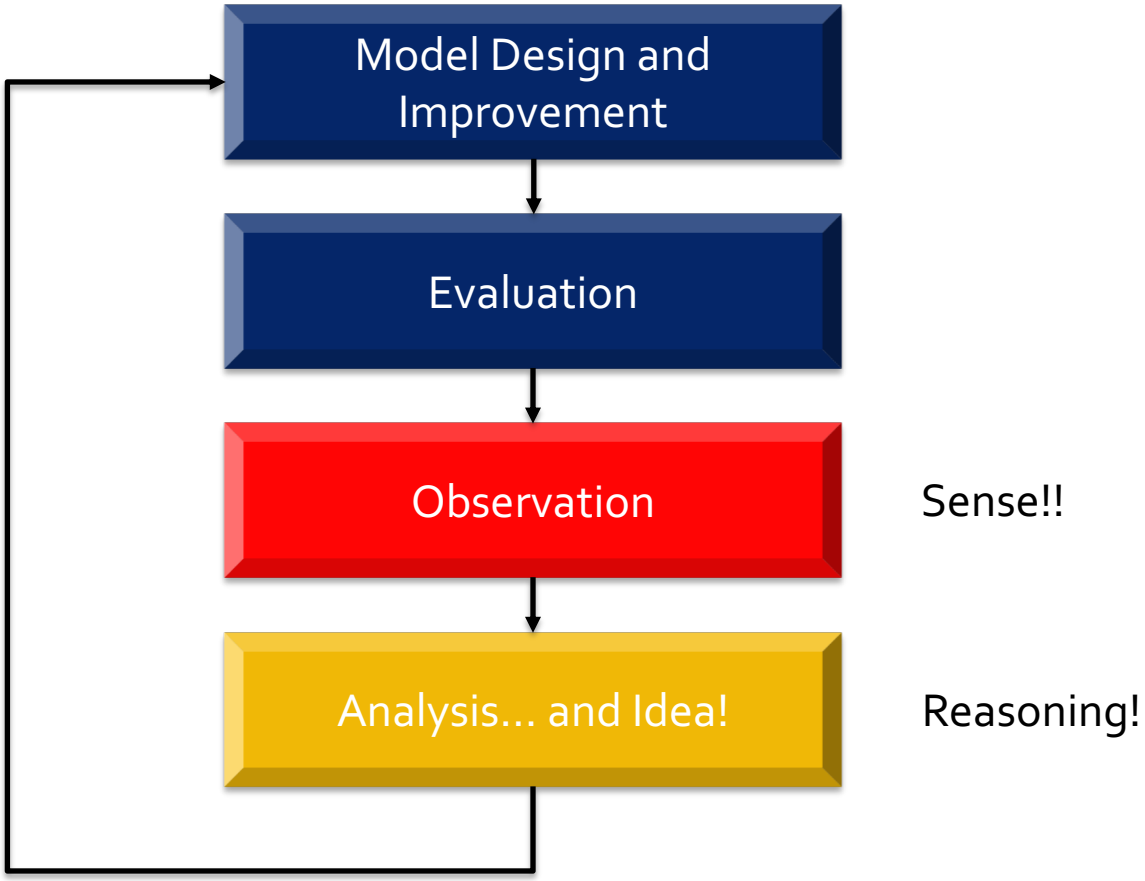
What is the most difficult?

Basic Skills of Being a Data Scientist

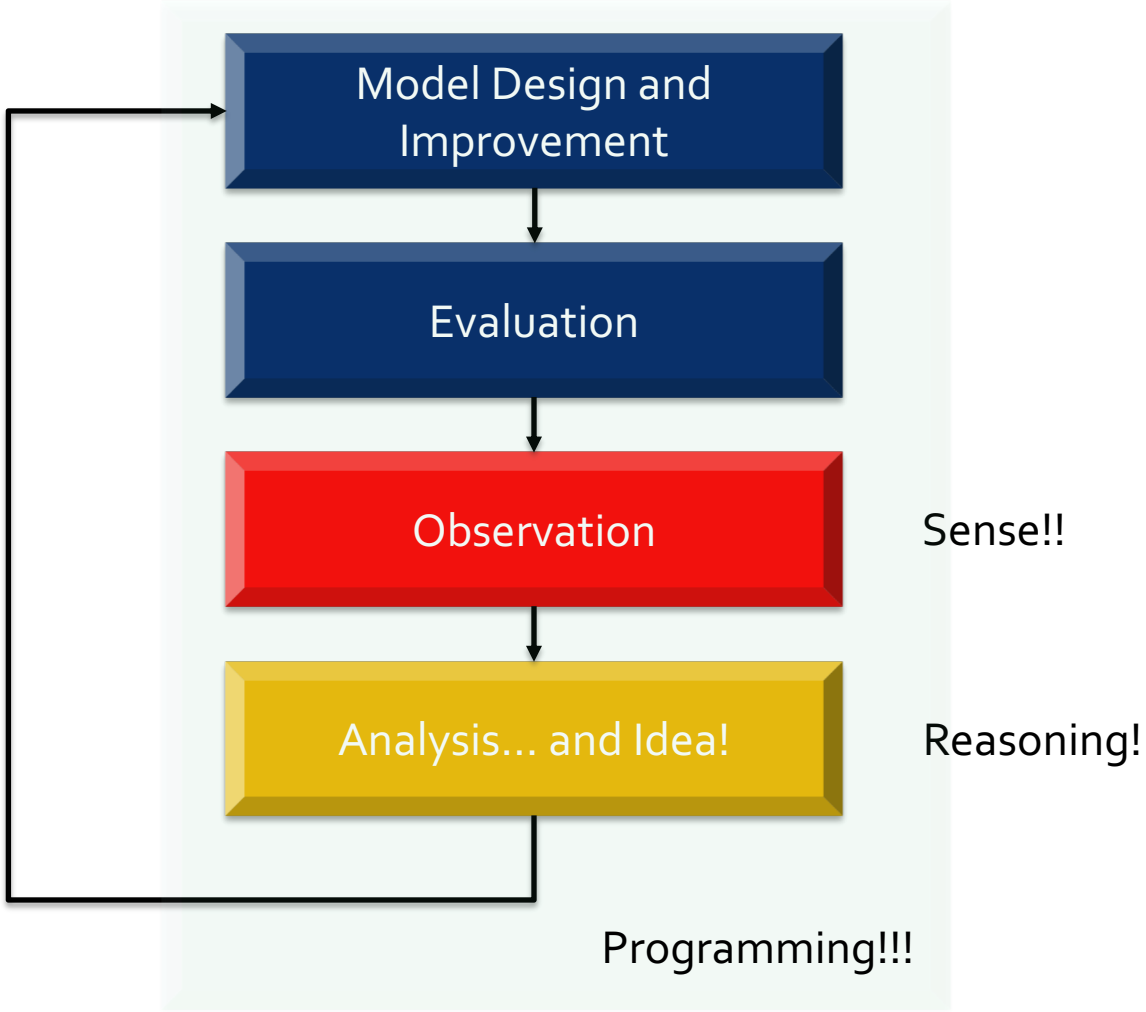


Reasoning!

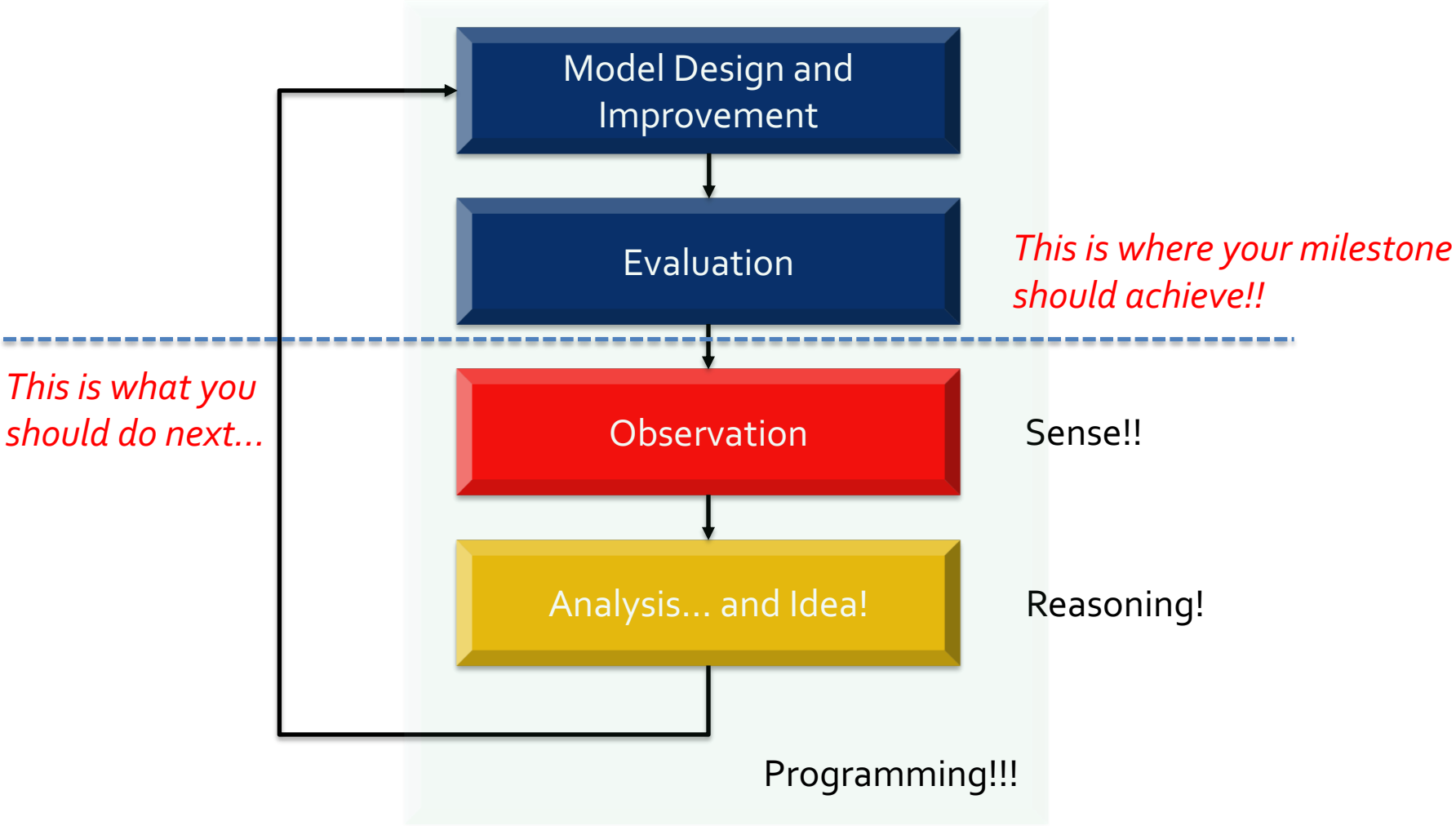
Basic Skills of Being a Data Scientist



Basic Skills of Being a Data Scientist



Basic Skills of Being a Data Scientist



Milestone Grading Policy

The **project paper** (milestone paper, final term paper) will be graded as follows:

Introduction:	15%	Provide context and motivation. What questions are being addressed? Why are these questions interesting or important?
Related Work:	10%	What other methods have addressed these or similar questions? How do these methods differ from your method?
Solution/Method:	25%	What did you do? What tools and techniques did you use? Was any innovation attempted?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable? What preprocessing was done the data?
Evaluation and Results:	25%	Did you properly evaluate your experiments? Did you test for statistical significance? Do your conclusions match your results?
Writing Quality:	15%	Clarity of writing (5%), organization (5%), and grammar (5%).

(5 points/100) * 30% = 1.5/100 in final grade = a 7.5 point-question in mid-term

The **project presentation** (milestone presentation, final poster, final oral presentation) will be graded as follows:

Introduction:	15%	Provide context. What questions are being addressed?
Solution/Method:	30%	What did you do? Why did you choose this method? What tools and techniques did you use?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable?
Evaluation and Results:	30%	What evaluation did you do? Do your conclusions match your results?
Presentation Quality:	15%	Clarity of speaking (5%), organization (5%), and visuals (5%).

(10 points/100) * 30% = 3/100 in final grade = a 15 point-question in mid-term

Milestone Paper Outline Example

- Introduction
  - Given customer's profile, predict if we should make a loan for him/her to buy a computer
- Related Work
  - Decision Trees (ID3, C4.5)
  - Naïve Bayes
- Solution/Method
  - Implemented ID3, C4.5, and Naïve Bayes
- Data/Experiments
  - …
- Evaluation/Results
  - …

Milestone Paper Outline Example (Better)

- Introduction
  - Given customer's profile, predict if we should make a loan for him/her to buy a computer
- Related Work
  - Decision Trees (ID3, C4.5)
  - Naïve Bayes
- Solution/Method
  - Reasoning on the Zero-Probability
  - Implemented ID3, C4.5, and Naïve Bayes
  - Designed Laplace correction on Naïve Bayes, and implemented
- Data/Experiments
  - …
- Evaluation/Results
  - …

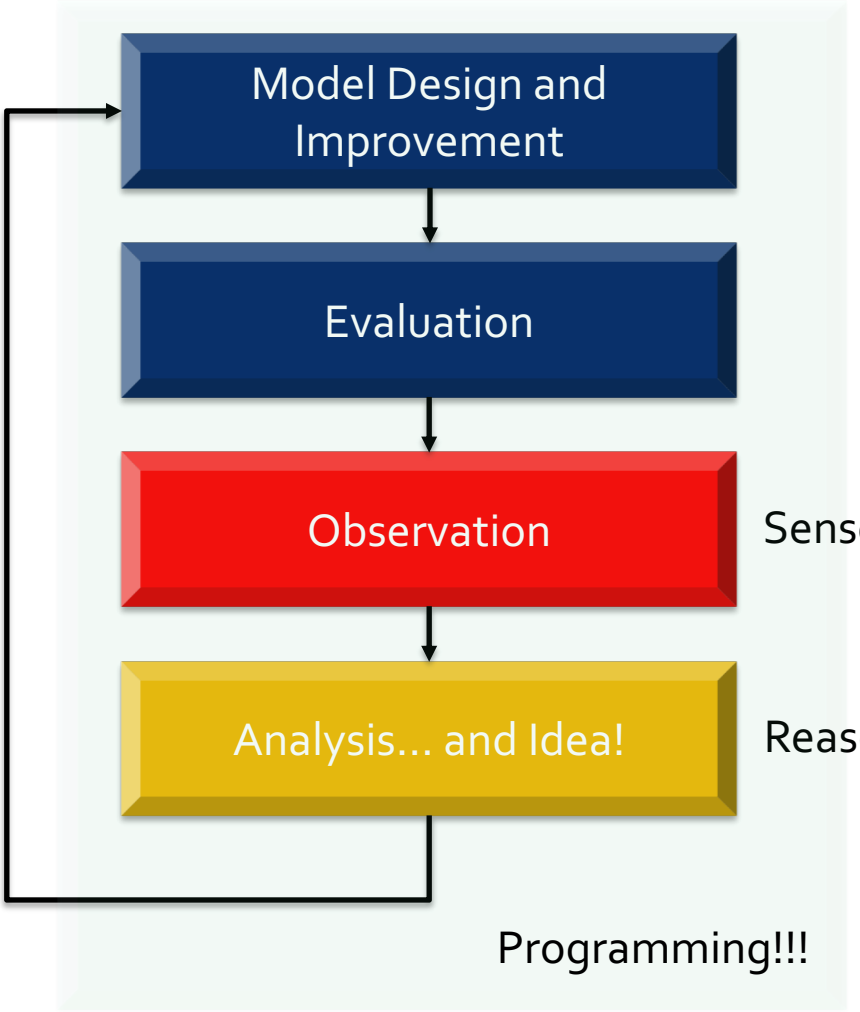
Commenting and Grading

- You will receive a hard copy of milestone papers and grading forms (like what we had for project proposal) on March 8.
- All members should present in class on March 8
  - Except two students who sent me their excuses
- All students should make constructive comments and grade the milestone papers during the Spring Break and hand in on March 20.
  - Student who miss the Milestone Presentation Day should download a print copy from the course website and print and comment and grade.
- Your team will lose some points if you do not show up in class with no excuse beforehand.
- Your team will lose points if you do not return grading forms to me.
- You may have extra points if you give constructive comments.

Milestone Presentation (5-8 mins)

- Introduction
  - Given customer's profile, predict if we should make a loan for him/her to buy a computer
- Solution/Method
  - Reasoning on the Zero-Probability
  - Implemented ID3, C4.5, and Naïve Bayes
  - Designed Laplace correction on Naïve Bayes, and implemented
- Data/Experiments
  - …
- Evaluation/Results
  - …
- Plan
- Team task distribution
  - Who did what?
  - Who will do what?

Questions to Each Team Now…



Q1: Which step have you achieved?
If all, how many iterations now?

Q2: What skill do you think
you need to enhance more?

- Programming
- Reasoning
- Sense

If you want to improve one of these,
which teammate do you want
to ask for help: discussion,
collaboration, etc.?

Two Reference Papers from ACM SIGKDD 2016 for Sports Session

Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights

Joel Brooks
Massachusetts Institute of Technology
Cambridge, MA
brooksjd@mit.edu

Matthew Kerr
Massachusetts Institute of Technology
Cambridge, MA
mattkerr@alum.mit.edu

John Guttag
Massachusetts Institute of Technology
Cambridge, MA
guttag@mit.edu

ABSTRACT

Quantitative evaluation of the ability of soccer players to contribute to team offensive performance is typically based on goals scored, assists made, and shots taken. In this paper, we describe a novel player ranking system based entirely on the value of passes completed. This value is derived based on the relationship of pass locations in a possession and shot opportunities generated. This relationship is learned by applying a supervised machine learning model to pass locations in event data from the 2012-2013 La Liga season. Interestingly, though this metric is based entirely on passes, the derived player rankings are largely consistent with general perceptions of offensive ability, e.g., Messi and Ronaldo are near the top. Additionally, when used to rank midfielders, it separates the more offensively-minded players from others.

Keywords

machine learning; sports analytics; soccer analytics

1. INTRODUCTION

Although soccer is by far the world's most popular sport [19], published work in soccer analytics has yet to achieve the same level of sophistication as analytics being performed in other professional sports. Crude summary statistics such as goals, shots, and assists are still the most common way to compare player performance analytically. More work is emerging [23] that leverages the rich datasets available to make discoveries about soccer, but there has not been much focus on quantitative metrics for evaluating player performance.

In this paper, we describe a novel way of quantitatively measuring a player's *passing* performance using existing data. We chose to focus on passing because it is one of the more strategic elements of soccer. Currently, players are often considered good passers if they accumulate many assists. Assists identify when a player makes a pass that directly leads to a goal, but this measure alone is quite limited. For example, assists do not capture passes that would have been assists except for an errant shot, or an excellent save by the opposing team's goalkeeper. Opta [17] extends the idea of assists to include all passes that lead to shots (whether or not they lead to a goal) in their "key passes" metric, but both this metric and assists are only applicable to passes immediately preceding a shot. There may be players who are excellent passers that create many opportunities for their team, but rarely make the last pass before a shot or goal.

Instead, we want to be able to quantitatively measure the importance of *any* pass. We accomplish this by first training a classifier that uses information about the locations of a set of passes to identify when that group of passes results in a shot. Since we use a linear classifier, we can directly utilize the model weights to understand the relative importance of pass origins and destinations for generating shot opportunities. These weights allow us to compute an estimated value of any pass for creating shots. We can then rank players by the value of the passes they complete over the course of a season.

In this paper, we use data from the 2012-2013 La Liga season to:

1. Construct a model relating pass origins and destinations during a possession with the probability of a shot. This model accurately identifies whether a possession ends in a shot from the pass locations alone.

2. Show how the resulting weights offer insights into the offensive utility of passes.

3. Utilize this model to rank players by the frequency with which their passes are highly valued by the model.

The rest of the paper is organized as follows. In Section 2, we outline some related work on using machine learning for knowledge discovery in soccer and other sports. In Section 3, we describe our event-based dataset. In Section 4, we present our methodology for building a model for predicting shots at the end of possessions. In Section 5, we demonstrate that this model can be used to rank players in an objective manner by how often they complete passes our model rates as valuable. Finally, in Section 6, we summarize the overall contributions of this paper and discuss possible future work.

2. RELATED WORK

Much of the published research in sports analytics, especially research that utilizes spatiotemporal data, has focused on sports that are easily discretized, such as baseball,

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2939695

Analyzing Volleyball Match Data from the 2014 World Championships Using Machine Learning Techniques

Jan Van Haaren
KU Leuven — Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
jan.vanhaaren@cs.kuleuven.be

Horesh Ben Shitrit
PlayfulVision
Ch. de la Raye 13, 1024 Ecublens, Switzerland
horeshb@playfulvision.com

Jesse Davis
KU Leuven — Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
jesse.davis@cs.kuleuven.be

Pascal Fua
EPFL — Computer Vision Laboratory
Station 14, 1015 Lausanne, Switzerland
pascal.fua@epfl.ch

ABSTRACT

This paper proposes a relational-learning based approach for discovering strategies in volleyball matches based on optical tracking data. In contrast to most existing methods, our approach permits discovering patterns that account for both **spatial** (that is, partial configurations of the players on the court) and **temporal** (that is, the order of events and positions) aspects of the game. We analyze both the men's and women's final match from the 2014 FIVB Volleyball World Championships, and are able to identify several interesting and relevant strategies from the matches.

Keywords

Sports analytics, Spatial data, Strategy detection

1. INTRODUCTION

Optical tracking enables analyses that go beyond descriptive statistics based on event data by providing sufficient context about the game state such that it is possible to gain insights into tactics and strategies. However, automatically detecting and understanding tactics in dynamic sports such as volleyball is a challenging problem as tactics comprise complex interactions among multiple players that evolve across both time and space. Hence, strategy detection must consider and evaluate a huge number of possible spatio-temporal movement patterns. Existing approaches, which largely look at basketball and soccer, tackle this problem by restricting the analysis in some way such as only considering a specific strategy [9], a short time window [4], or a small subset of players [5, 3].

This paper addresses automatically discovering attacking strategies in volleyball matches based on camera-tracking data of the players and the ball as a pattern mining task [1]. Specifically, we consider the following two tasks:

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08. . . $15.00

DOI: http://dx.doi.org/10.1145/2939672.2939725

Task 1: Identify a team's attacking patterns in volleyball matches that occur frequently in *won* rallies and infrequently in *lost* rallies.

Task 2: Identify attacking patterns in a volleyball match that are used by one team but not the opposing team.

In contrast to most existing approaches, we attempt to identify patterns that account for both **spatial** and **temporal** aspects of the game. That is, we want to model (partial) configurations of players' positions on the court as well as how play evolves over time. To illustrate this, consider the following simple pattern automatically discovered by our approach:

```
IF player #13 performs the dig  AND NEXT
   player #1 performs the set
      in the front center zone AND NEXT
   player #8 performs the spike
      in position 81 of the court
THEN the attack is likely to be successful.
```

This pattern is temporal as the dig occurs first, the set second, and the spike third. The pattern is spatial as it states the location on the court where the set and spike occur (see Figure 1 for a description of the locations).

In order to automatically identify patterns like the one just shown, we use a relational-learning based approach. As much as possible, we attempt to employ a data-driven approach that can automatically determine which players and characteristics of the game state are relevant to the strategy and should be included in the pattern. We analyze data from both the men's and women's final match from the 2014 FIVB Volleyball World Championships. Our top-ranked discovered patterns represent strategies that are both interesting and relevant from a volleyball perspective.

2. BACKGROUND ON VOLLEYBALL

Volleyball [10, 2] is a ball sport that is played by two teams of six players each. A volleyball court is 18 meters (59 feet) long and 9 meters (29.5 feet) wide. Each team occupies one half of the court, which is 9 meters by 9 meters. The halves are separated by a net whose top is 2.43 meters above the floor in men's competitions and 2.24 meters in women's competitions. The overall goal is to score points by grounding the ball on the opponent's court.

You Must Use
ACM LaTeX Template!
Two-columns paper

You can change the title a little bit different from the proposal if you want.

Submitting slides (not more than 5 pages) along with the paper is recommended!

Paper due: March 7 (Wednesday)

Presentation date: March 8 in class