香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

DDA2001: Introduction to Data Science

Introduction

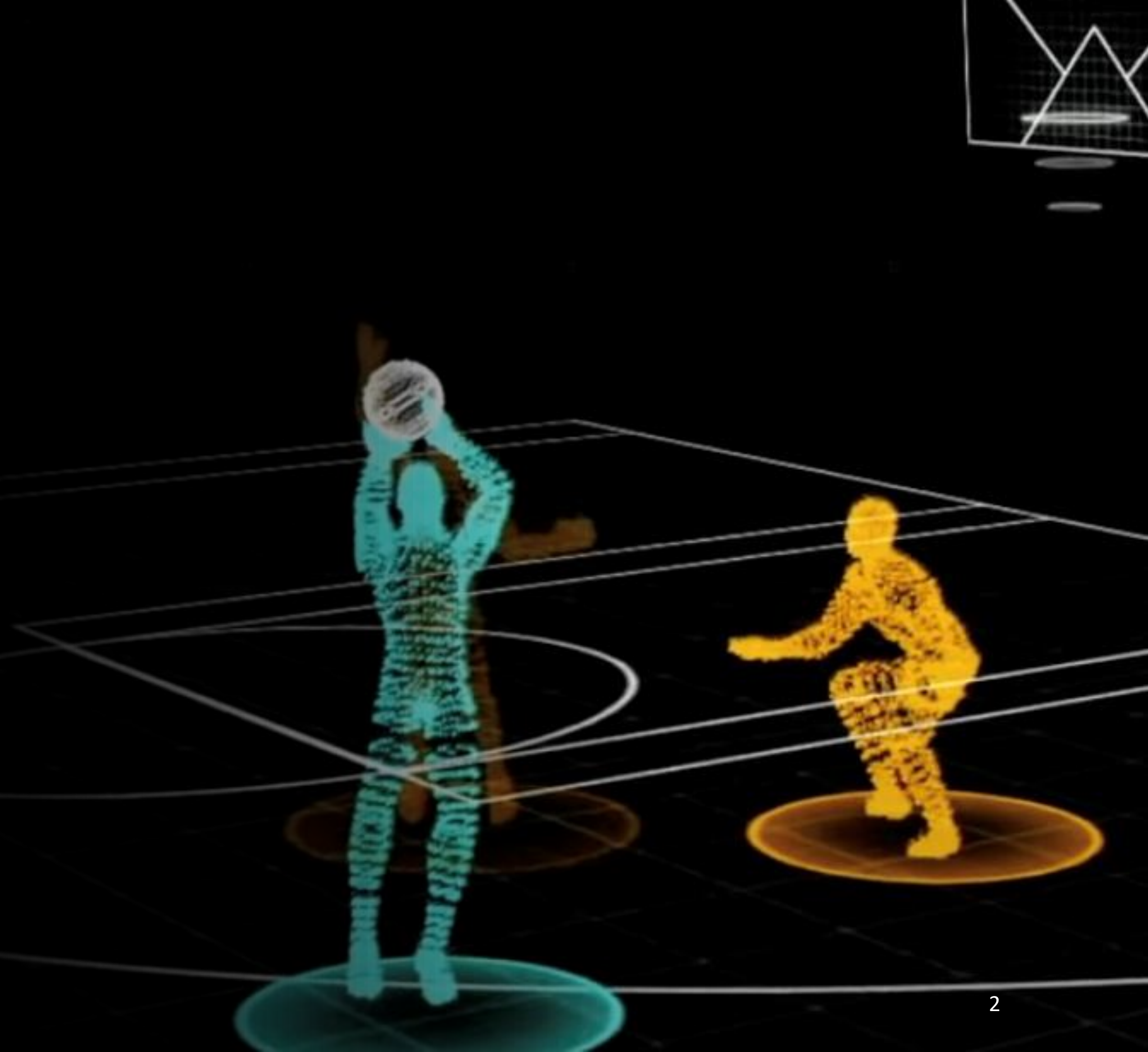
2023/01/08

Zicheng Wang

An appetizer of data science application

The Math Behind Basketball's Wildest Moves | Rajiv Maheswaran | TED Talks

https://www.youtube.com/watch?v=66ko_cWSHBU&ab_channel=TED (6:40-8:25)



Data science aims to understand the world by collecting and analyzing data, and makes better decisions for the future.

Outline

Course organization

Course introduction

Course organization

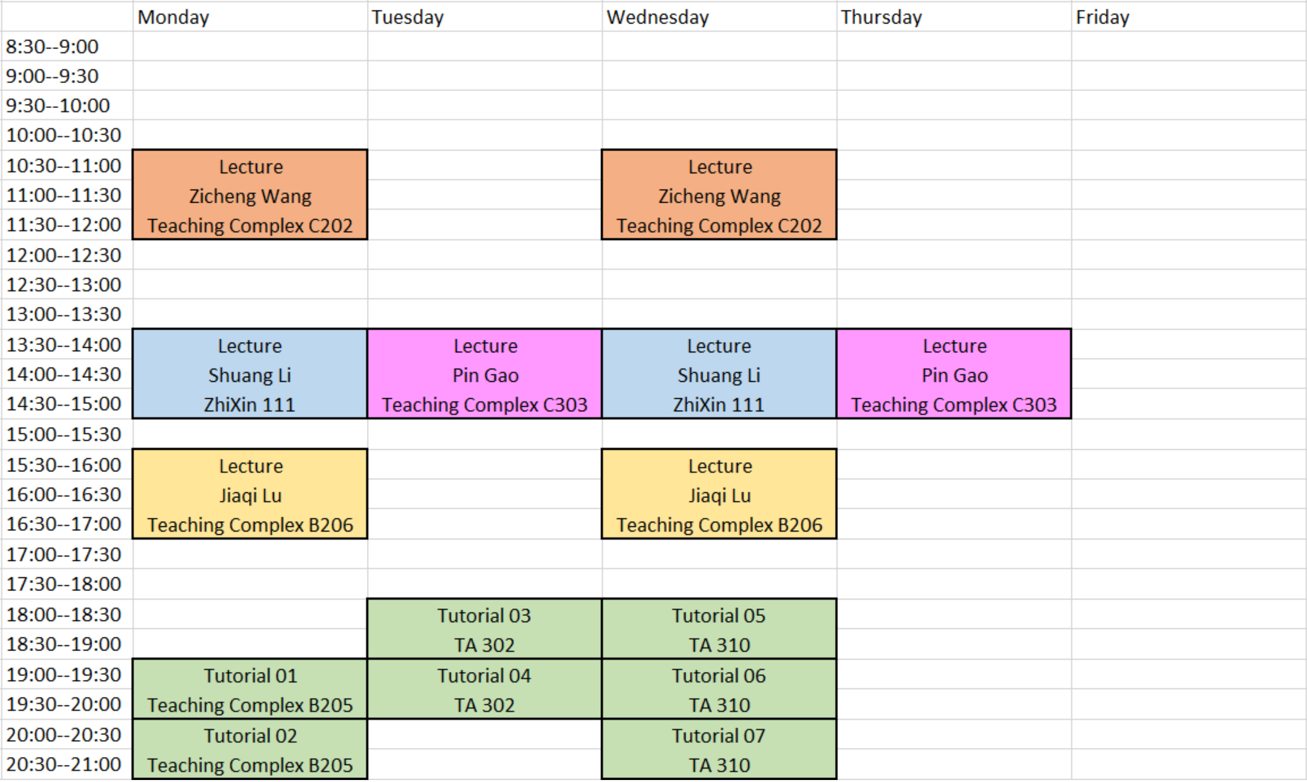
Course staff

- Instructor:
  - *Zicheng Wang*
    - Office: DY 322B, TXC711
    - Office Hours: Monday 9am – 10am
    - Email: wangzicheng@cuhk.edu.cn
- Lectures:
  - Room: Teaching Complex C202
  - Times: Mon Wed 10:30 – 11:50am

Course staff (TAs)

•Chenguang Wang	王晨光
Email:	222043009@link.cuhk.edu.cn
•Chao Yang	杨超
Email:	222043011@link.cuhk.edu.cn
•Weihuang Wen	温伟煌
Email:	223040256@link.cuhk.edu.cn
•Sheng Xu	徐圣
Email:	223040246@link.cuhk.edu.cn
•Wenjun Zheng	郑文军
Email:	118010446@link.cuhk.edu.cn

Schedule



	Monday	Tuesday	Wednesday	Thursday	Friday
8:30--9:00					
9:00--9:30					
9:30--10:00					
10:00--10:30					
10:30--11:00	Lecture Zicheng Wang Teaching Complex C202		Lecture Zicheng Wang Teaching Complex C202		
11:00--11:30					
11:30--12:00					
12:00--12:30					
12:30--13:00					
13:00--13:30					
13:30--14:00	Lecture Shuang Li ZhiXin 111	Lecture Pin Gao Teaching Complex C303	Lecture Shuang Li ZhiXin 111	Lecture Pin Gao Teaching Complex C303	
14:00--14:30					
14:30--15:00					
15:00--15:30					
15:30--16:00	Lecture Jiaqi Lu Teaching Complex B206		Lecture Jiaqi Lu Teaching Complex B206		
16:00--16:30					
16:30--17:00					
17:00--17:30					
17:30--18:00					
18:00--18:30		Tutorial 03 TA 302	Tutorial 05 TA 310		
18:30--19:00					
19:00--19:30	Tutorial 01 Teaching Complex B205	Tutorial 04 TA 302	Tutorial 06 TA 310		
19:30--20:00					
20:00--20:30	Tutorial 02 Teaching Complex B205		Tutorial 07 TA 310		
20:30--21:00					

Tutorials

- **Tutorials start on the second week**

Probability	Tutorial Session	TA	Location	Time
	T01	Chenguang Wang	TCB 205	Mon:19:00-19:50
	T02	Chenguang Wang	TCB 205	Mon:20:00-20:50
	T03	Chenguang Wang	TA302	Tues: 18:00-18.50
	T04	Chenguang Wang	TA302	Tues: 19:00-19.50
	T05	Chenguang Wang	TA310	Wed: 18:00-18:50
	T06	Chenguang Wang	TA310	Wed: 19:00-19:50
	T07	Chenguang Wang	TA310	Wed: 20:00-20:50
Statistics	Tutorial Session	TA	Location	Time
	T01	Sheng Xu	TCB 205	Mon:19:00-19:50
	T02	Sheng Xu	TCB 205	Mon:20:00-20:50
	T03	Sheng Xu	TA302	Tues: 18:00-18.50
	T04	Sheng Xu	TA302	Tues: 19:00-19.50
	T05	Sheng Xu	TA310	Wed: 18:00-18:50
	T06	Sheng Xu	TA310	Wed: 19:00-19:50
	T07	Sheng Xu	TA310	Wed: 20:00-20:50
Optimization	Tutorial Session	TA	Location	Time
	T01	Wenjun Zheng	TCB 205	Mon:19:00-19:50
	T02	Wenjun Zheng	TCB 205	Mon:20:00-20:50
	T03	Wenjun Zheng	TA302	Tues: 18:00-18.50
	T04	Wenjun Zheng	TA302	Tues: 19:00-19.50
	T05	Wenjun Zheng	TA310	Wed: 18:00-18:50
	T06	Wenjun Zheng	TA310	Wed: 19:00-19:50
	T07	Wenjun Zheng	TA310	Wed: 20:00-20:50
ML	Tutorial Session	TA	Location	Time
	T01	Weihuang Wen	TCB 205	Mon:19:00-19:50
	T02	Weihuang Wen	TCB 205	Mon:20:00-20:50
	T03	Weihuang Wen	TA302	Tues: 18:00-18.50
	T04	Weihuang Wen	TA302	Tues: 19:00-19.50
	T05	Weihuang Wen	TA310	Wed: 18:00-18:50
	T06	Weihuang Wen	TA310	Wed: 19:00-19:50
	T07	Weihuang Wen	TA310	Wed: 20:00-20:50

Tentative Calendar

Weeks	**Content/ topic/ activity**
	Course Introduction
1	Course plan, grading, overview, examples, and applications, etc.
	Probability
3	1 Terminologies: sample spaces, probability function, event, etc.. 2. Main Properties: mean, variance, CDF. 3. Common Discrete RV: Bernoulli, Binomial, Poisson, Geometric. 4. Common Continuous RV: uniform, normal, exponential. 5*. Advanced Concepts: correlation and conditional probability (Bayes rule). 6#. Additional examples and exercises.
	Statistics
3	1. Sampling: Monte Carlo (integration, MDP, etc.). 2. Sampling: generating uncommon RV. 3. Point estimator: MLE. 4. MLE Application: Regression Analysis. 5*. Confidence interval 6#. Additional examples and exercises.
0.5	**Midterm Review**

Tentative Calendar

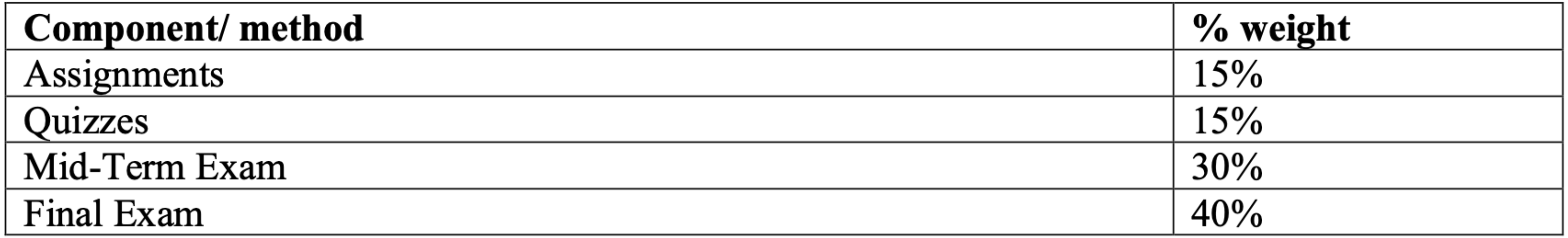
	Optimization
3.5	1. Problem formulation (objective function, constraints, decision variables, etc.) Application examples in revenue management, online advertising, etc. 2. Optimization: convex set. 3. Convex Optimization: convex function. 4. Convex Optimization: convex function Property. 5. Convex Optimization: quasi-convex function. 6*. Gradient Descent and Stochastic Gradient Descent 7#. Additional examples and exercises.
	Machine Learning
2.5	1*. Supervised learning: K-Nearest Neighbors (K-NN). 2*. Supervised learning: Logistic regression, Neural Nets. 3*. Unsupervised learning: Clustering (K means). 4**. Cross Validation. 5**: Deep Learning and Reinforcement Learning.
0.5	**Final Review**

# Problem exercise.

* Only concepts will be examined.

** Contents that will be covered if time permits. If covered, only concepts will be examined.

Learning Assessment Strategy



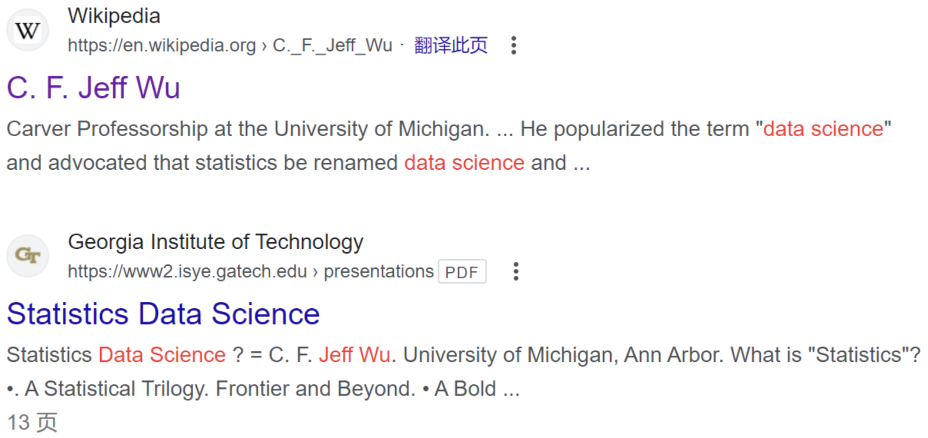
Component/ method	% weight
Assignments	15%
Quizzes	15%
Mid-Term Exam	30%
Final Exam	40%

- **3 to 5 Guest Lectures on Fridays (Dates TBD). Quizzes on Guest Lectures.**
- **No late homework will be accepted.**
- **Midterm (tentative): TBD**
- **There will be no make-up midterm exams**
- **Zero tolerance policy when it comes to cheating so, DON’T CHEAT**
- **Missing the midterm or final exam without prior notification to and approval of the instructors will automatically result in the "0" grade for the exam.**

About me

- **My academic pathway**
  - Undergraduate: Mathematics, University of Minnesota
  - Graduate: PhD in Industrial and Systems Engineering, University of Minnesota (advisor: Kevin Leder)
  - Join SDS of CUHK(SZ) since July 2023, and currently an assistant professor
- **My Research Interest related to “Data Science”: Operations Research**
  - Applied probability and its application in mathematical modeling of cancer and operations management
  - Delineate tumor dynamics
  - Analyze complex spatial systems

Course introduction



Wikipedia
https://en.wikipedia.org › C._F._Jeff_Wu · 翻译此页

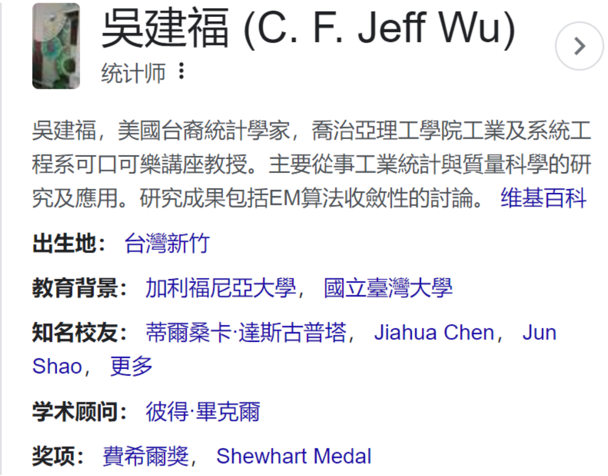
C. F. Jeff Wu

Carver Professorship at the University of Michigan. ... He popularized the term "data science" and advocated that statistics be renamed data science and ...

Georgia Institute of Technology
https://www2.isye.gatech.edu › presentations PDF

Statistics Data Science

Statistics Data Science ? = C. F. Jeff Wu. University of Michigan, Ann Arbor. What is "Statistics"? •. A Statistical Trilogy. Frontier and Beyond. • A Bold ...
13 页



吴建福 (C. F. Jeff Wu)
统计师

吳建福，美國台裔統計學家，喬治亞理工學院工業及系統工程系可口可樂講座教授。主要從事工業統計與質量科學的研究及應用。研究成果包括EM算法收斂性的討論。 维基百科

出生地： 台灣新竹

教育背景： 加利福尼亞大學， 國立臺灣大學

知名校友： 蒂爾桑卡·達斯古普塔， Jiahua Chen， Jun Shao， 更多

学术顾问： 彼得·畢克爾

奖项： 費希爾獎， Shewhart Medal

What is data science

- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.
- Data science is not one single subject! It incorporates probability, statistics, operations research, machine learning, and business applications…
- You need a solid mathematical/technical foundation, together with some industrial expertise in order to be a data scientist.
- Analytical skills + business understanding.

How are they related?

- Probability
- Statistics
- Optimization
- Machine learning

11.11
11.11
天猫全球狂欢季
2021
11.11
天猫全球狂欢季
2021
美好生活 共同向往
2021天猫双11全球狂欢季消费者发布会

GLOBAL SHOPPING FESTIVAL 2021

The festival has just passed.
Did you buy a lot of things?

剁手人、吃土人、尾款人、晚八人

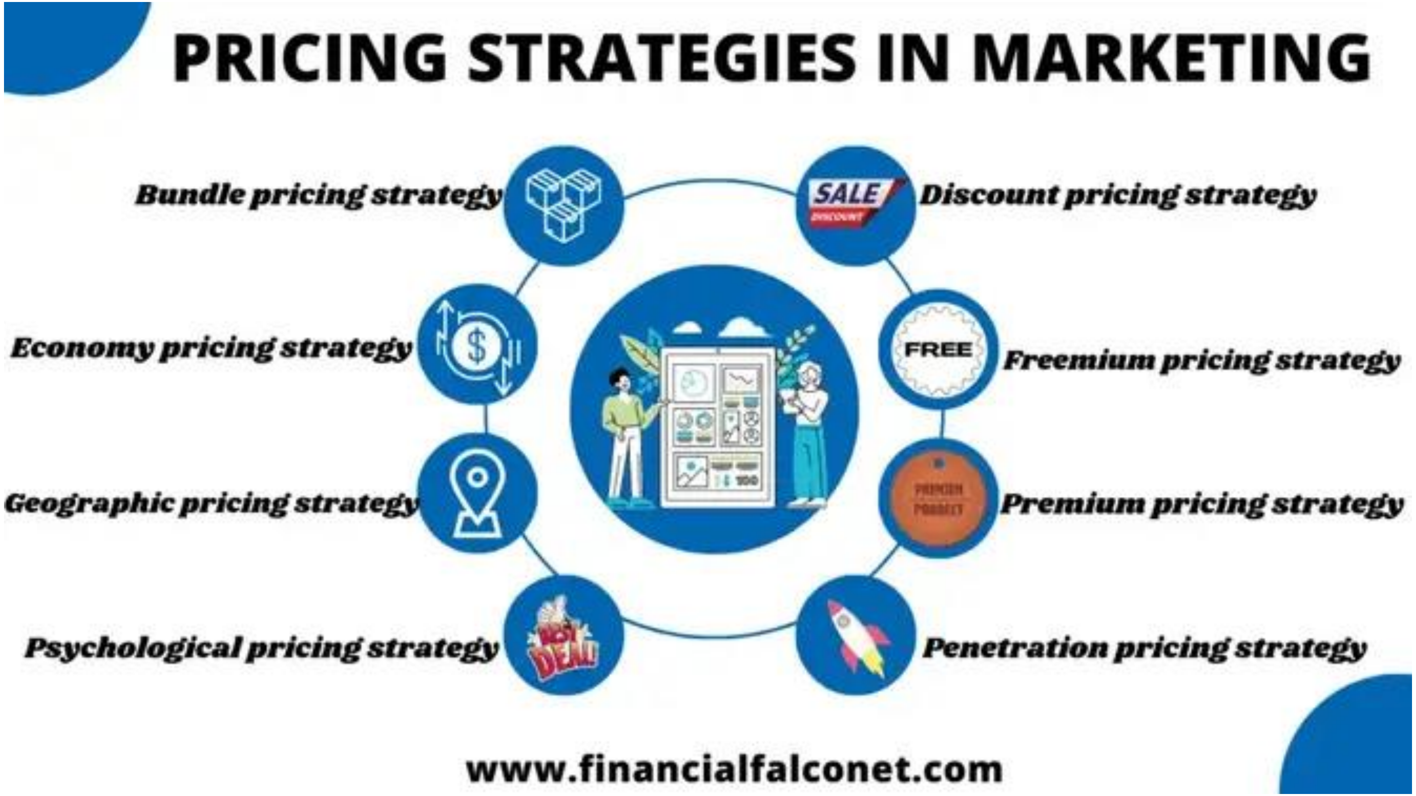
1.剁手人 (Duò shǒu rén): Literally translates to "hand-chopping people". It's a humorous way to describe someone who can't resist buying things online, implying they might as well chop off their hands to stop spending.

2.吃土人 (Chī tǔ rén): Literally means "eating dirt people". This phrase is used to describe people who have spent so much money, particularly on shopping, that they are left with no money for essentials, metaphorically left to 'eat dirt’.

3.尾款人 (Wěi kuǎn rén): This translates to "tail payment people". It's used to describe people who make a down payment for a product (often during a sale like Singles' Day in China) and are waiting to pay the balance amount.

4.晚八人 (Wǎn bā rén): This phrase is a bit more context-dependent. Literally, it means "late-eight people". This could refer to people who work late (until 8 PM), or it could be a specific cultural reference that needs more context to accurately translate.

- To stand out from the event, various promotional rules were designed


PRICING STRATEGIES IN MARKETING
Bundle pricing strategy
Discount pricing strategy
SALE
DISCOUNT
Economy pricing strategy
Freemium pricing strategy
FREE
Geographic pricing strategy
Premium pricing strategy
PREMIUM PRODUCT
Psychological pricing strategy
BEST DEAL
Penetration pricing strategy
www.financialfalconet.com

Things to consider when choosing a pricing strategy

Cost of Products

Customers

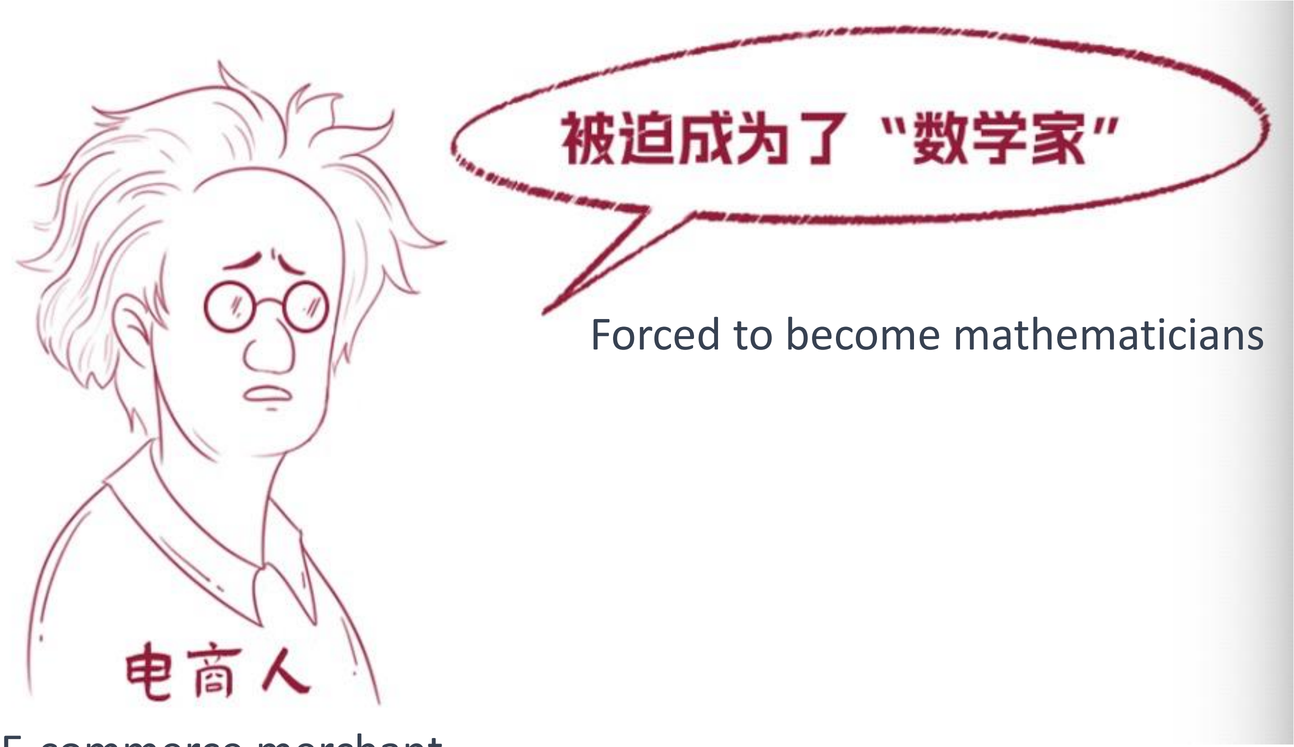
Market Positioning

Competitors

Profit Margins

Geography

Product Offerings



Forced to become mathematicians

E-commerce merchant

https://shanshu.ai/news/detail?id=243

- Even if you are already good at these pricing techniques, it is still very difficult to implement them.

Do I have to set the prices for all these?

How complicated are the price rules?

Has the price not changed?

1, 2, 3, click the link, one yuan for three bottles of craft beer

Sorry!

https://shanshu.ai/news/detail?id=243

How complicated are the price rules?

Sorry, setting this price was too troublesome. Full refunds for all purchases and free giveaways! Let's try another round!

What? Still the original price? Crashed again?

Sorry, I am a new broadcaster.

对不起，这个价格设置太麻烦了
已购买全部退款免费送！再抢一轮
啥？还是原价，
又翻车了？
对不起，
我是新人主播

https://shanshu.ai/news/detail?id=243

How complicated are the price rules?

There may be hundreds of SKUs (stock-keeping units) in one live broadcast

For e-commerce platforms, what they are facing can be more than millions of SKUs

https://shanshu.ai/news/detail?id=243

Say NO to manual pricing!

https://shanshu.ai/news/detail?id=243

How to design prices more
scientifically and efficiently?

1. Let a machine think like you

2. Let the machine automatically design prices

Data science

- Let a machine think like you
  - Extract knowledge from **data** sets

- Let the machine automatically design prices
  - Apply the knowledge and actionable insights from data to solve problems in a wide range of application domains

To be a good data scientist,

- Programming languages + analytical skills + business understanding.

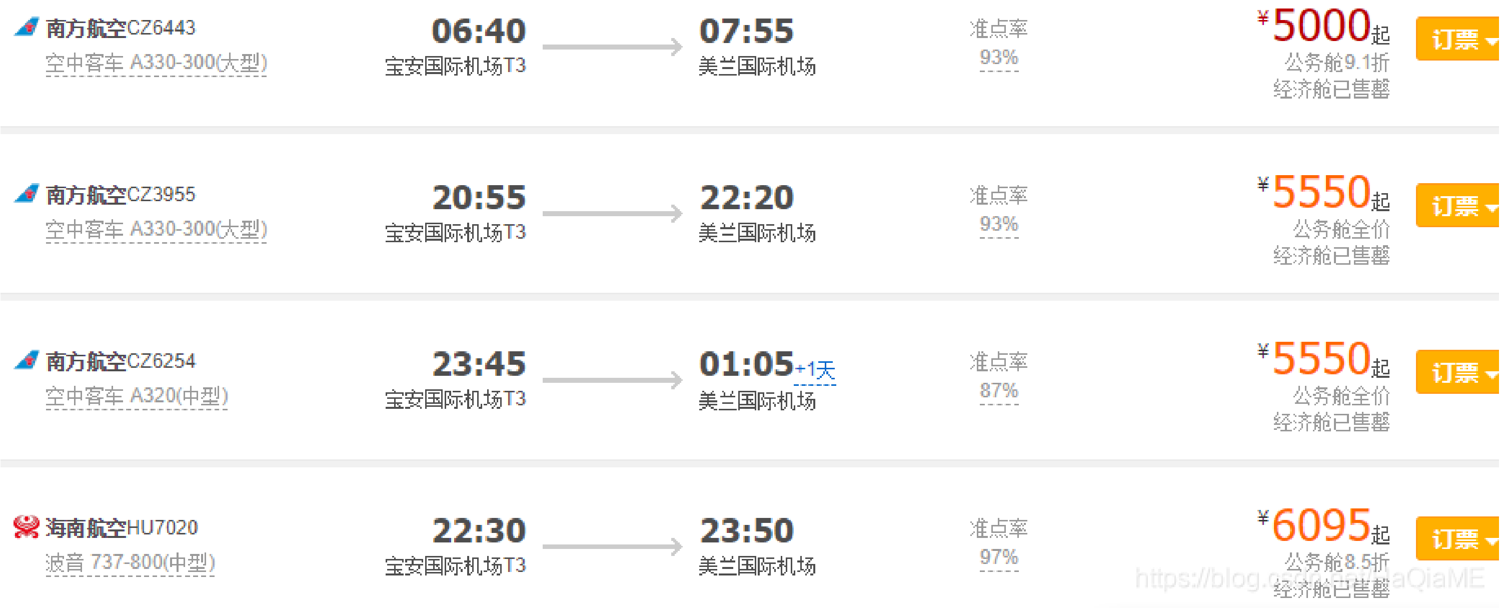
- It incorporates probability, statistics, operations research, machine learning, and business applications…

A concrete example

How to price a flight ticket?

1. Data collection? **Programming**.

- Obtain historical data, obtain current competitor prices, etc.



南方航空CZ6443 空中客车 A330-300(大型)	06:40 宝安国际机场T3	07:55 美兰国际机场	准点率 93%	¥5000起 公务舱9.1折 经济舱已售罄	订票
南方航空CZ3955 空中客车 A330-300(大型)	20:55 宝安国际机场T3	22:20 美兰国际机场	准点率 93%	¥5550起 公务舱全价 经济舱已售罄	订票
南方航空CZ6254 空中客车 A320(中型)	23:45 宝安国际机场T3	01:05+1天 美兰国际机场	准点率 87%	¥5550起 公务舱全价 经济舱已售罄	订票
海南航空HU7020 波音 737-800(中型)	22:30 宝安国际机场T3	23:50 美兰国际机场	准点率 97%	¥6095起 公务舱8.5折 经济舱已售罄	订票

1. Data collection? **Programming**.

- Obtain historical data, obtain current competitor prices, etc.

- Programming is the process of creating a set of instructions that tell a computer how to perform a task

- Learning to program is to free your hands

- Programming is a fundamental part of data science.

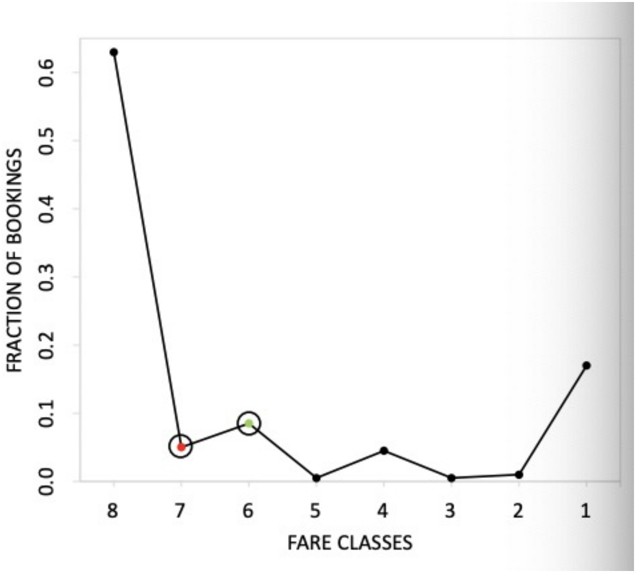
- Programming Languages for Data Science: python, R, C, C++, …

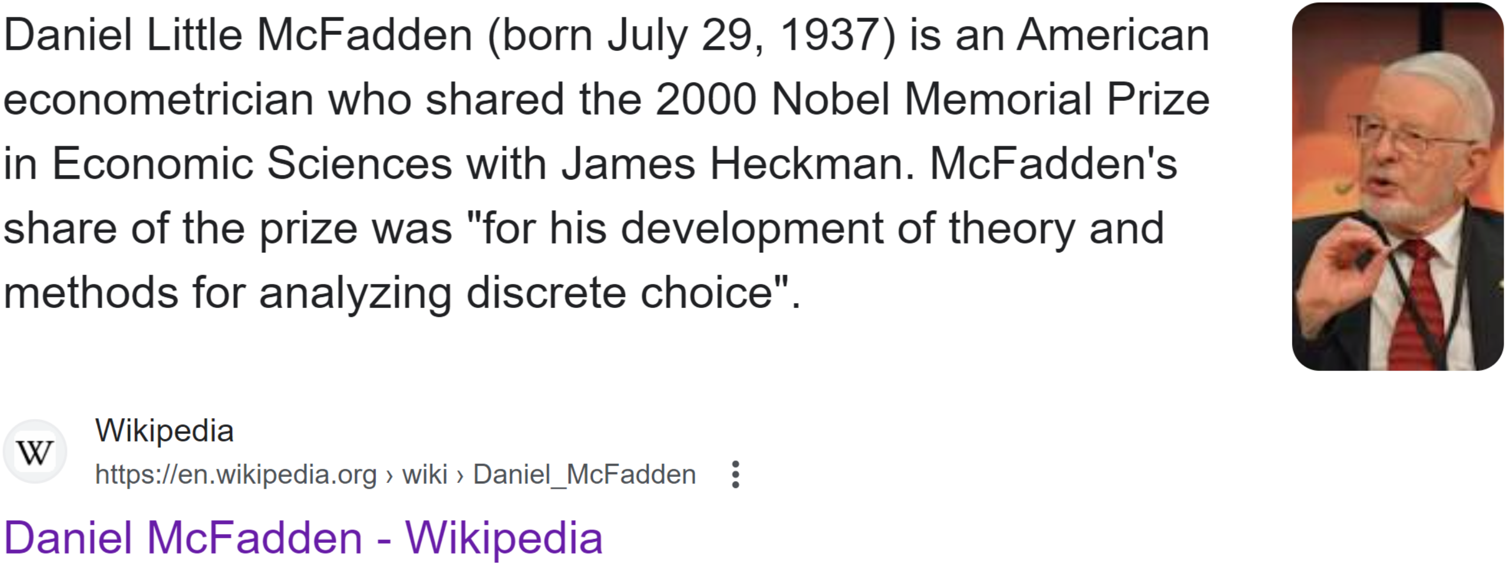
- In this course,
  - We will give some coding demos
  - There will be no coding questions in the homework and the exams.

1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.

You observe:
The product with the lowest price always sells significantly higher than other products.

How to describe the data generation process?





Daniel Little McFadden (born July 29, 1937) is an American econometrician who shared the 2000 Nobel Memorial Prize in Economic Sciences with James Heckman. McFadden's share of the prize was "for his development of theory and methods for analyzing discrete choice".

Wikipedia
https://en.wikipedia.org › wiki › Daniel_McFadden

Daniel McFadden - Wikipedia

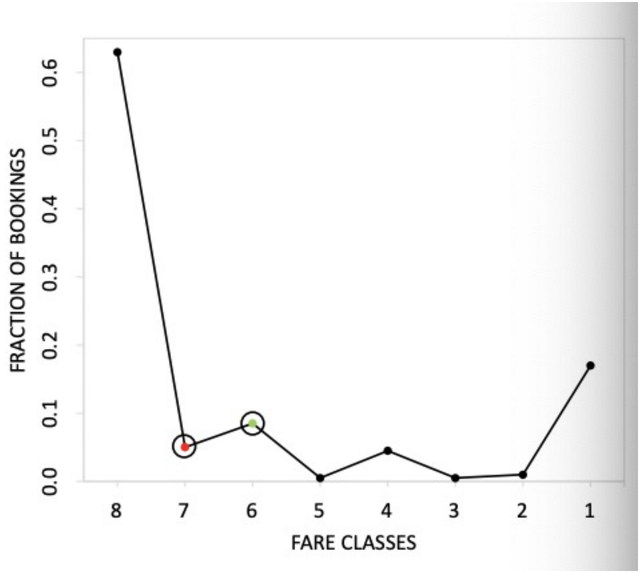
1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.

Consider a toy model:

With a set of products S, assume that the purchase probability of product i

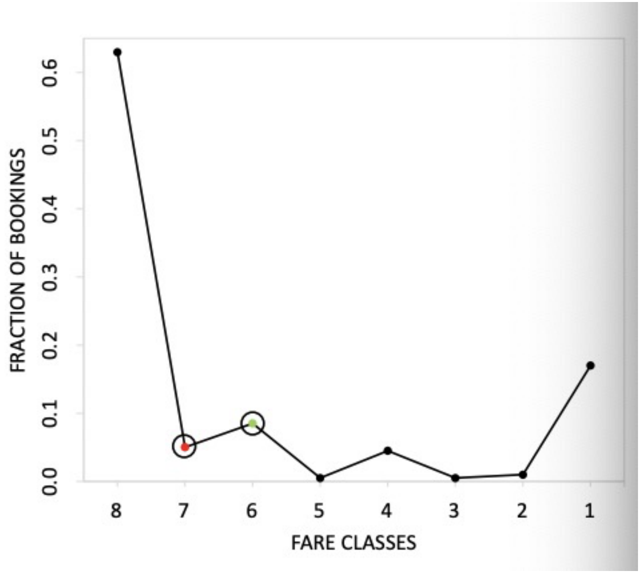
$$P_i(S) = \frac{a_i + x_i v_i}{1 + \sum_{j \in S}(a_j + x_j v_j)}$$

- a_i and v_i are known functions of prices
- $x_i \in \{0,1\}$ and x_i=1 if and only if this product has the lowest price



1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.
3. How to choose a reliable data pattern? Statistics.

$$P_i(S) = \frac{a_i + x_i v_i}{1 + \sum_{j \in S}(a_j + x_j v_j)}$$

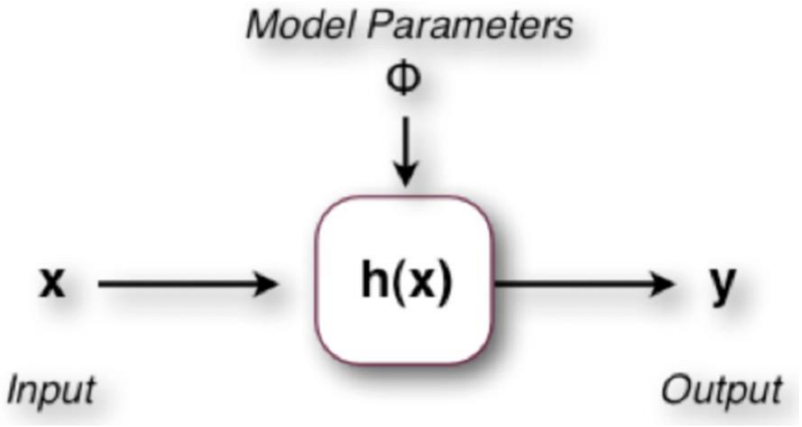


- How to determine the values of a_i and v_i?
- Statistics can be used to test the quality of the model.

1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.
3. How to choose a reliable data pattern? Statistics.
4. How to learn a complicated function? Machine Learning.

$$P_i(S) = \frac{a_i + x_i v_i}{1 + \sum_{j \in S}(a_j + x_j v_j)}$$

- a_i and v_i may depend on product attributes and consumer profiles.



1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.
3. How to choose a reliable data pattern? Statistics.
4. How to learn a complicated function? Machine Learning.
5. What decision should be made to maximize objective? Optimization

$$\max_r \sum_{j \in S} r_i P_i(S|r)$$

- Choose a price vector to maximize the above objective.

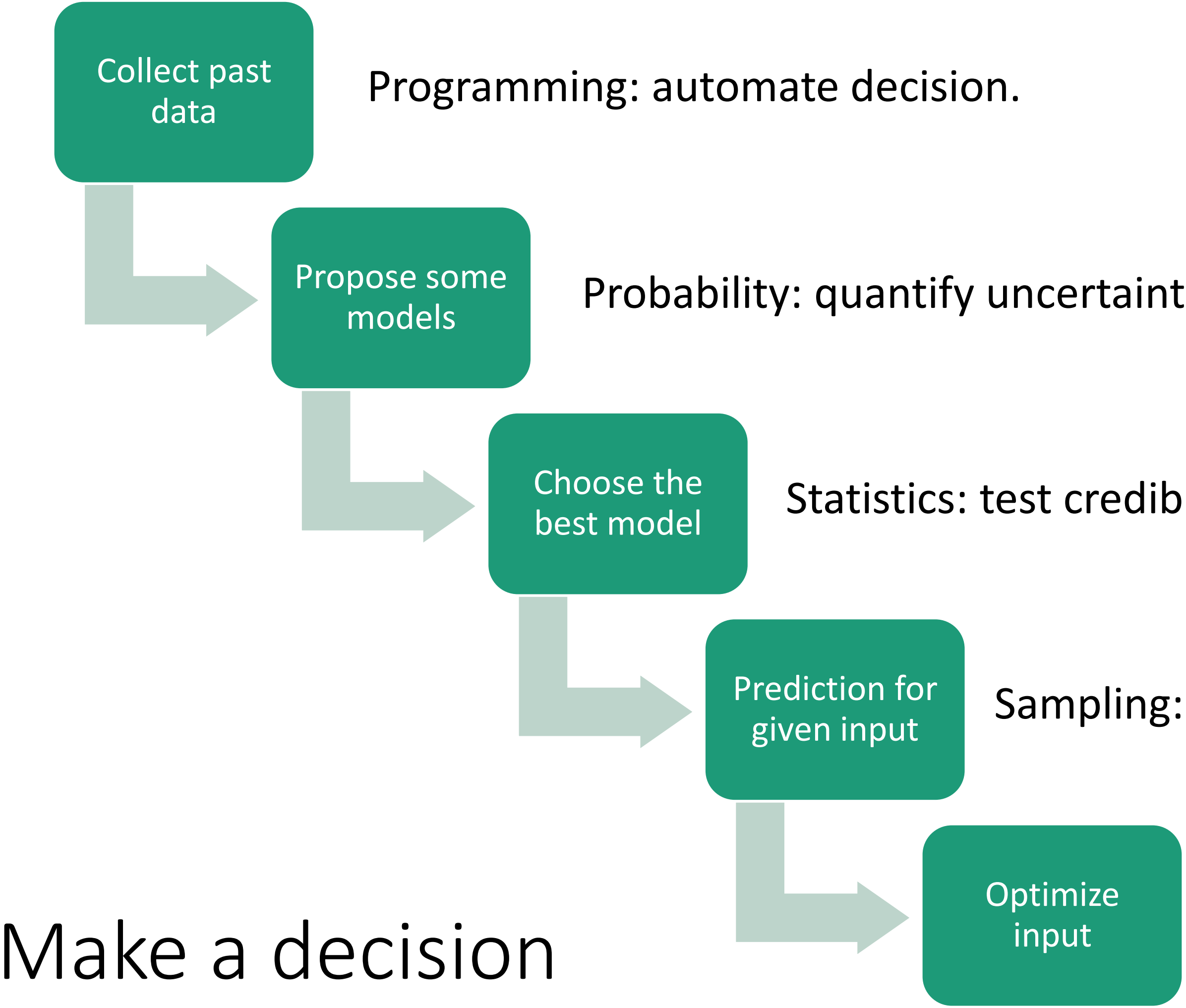
- The selection of a best element, with regard to some criterion, from some set of available alternatives

1. Data collection? **Programming**.
2. How to describe the data pattern? Probability theory.
3. How to choose a reliable data pattern? Statistics.
4. How to learn a complicated function? Machine Learning.
5. What decision should be made to maximize objective? Optimization
6. Calculate a complicated objective? Simulations

$$\sum_{j\in S} r_i P_i(S|r)$$

A computerized mathematical technique that allows people to generate random samples and aid quantitative analysis and decision making.

Make a decision



Collect past data

Programming: automate decision.

Propose some models

Probability: quantify uncertainty.

Choose the best model

Statistics: test credibility.

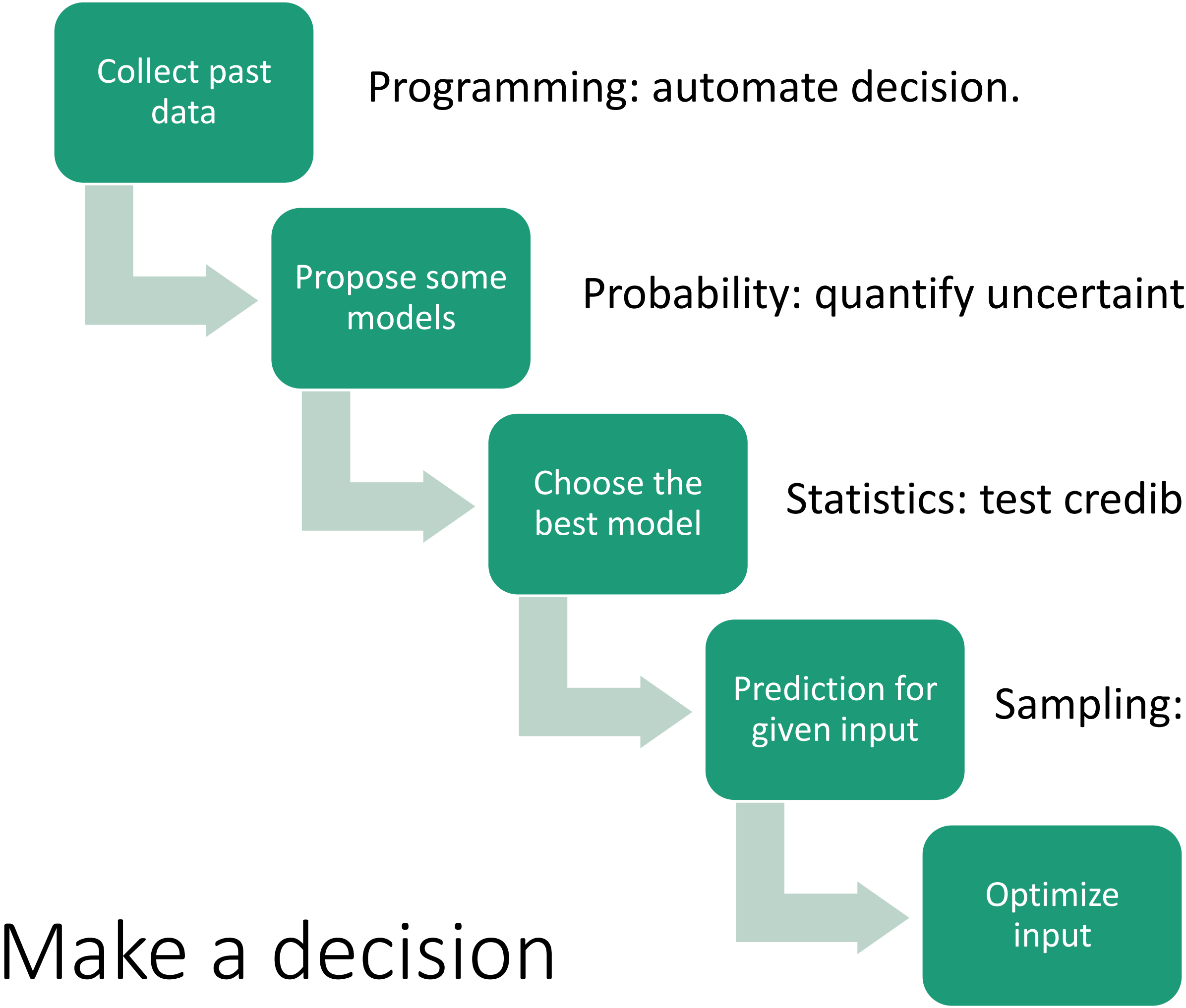
Prediction for given input

Sampling: calculate complicated objective.

Optimize input

Optimization: optimize objectives.

Make a decision



Programming: automate decision.

Probability: quantify uncertainty.

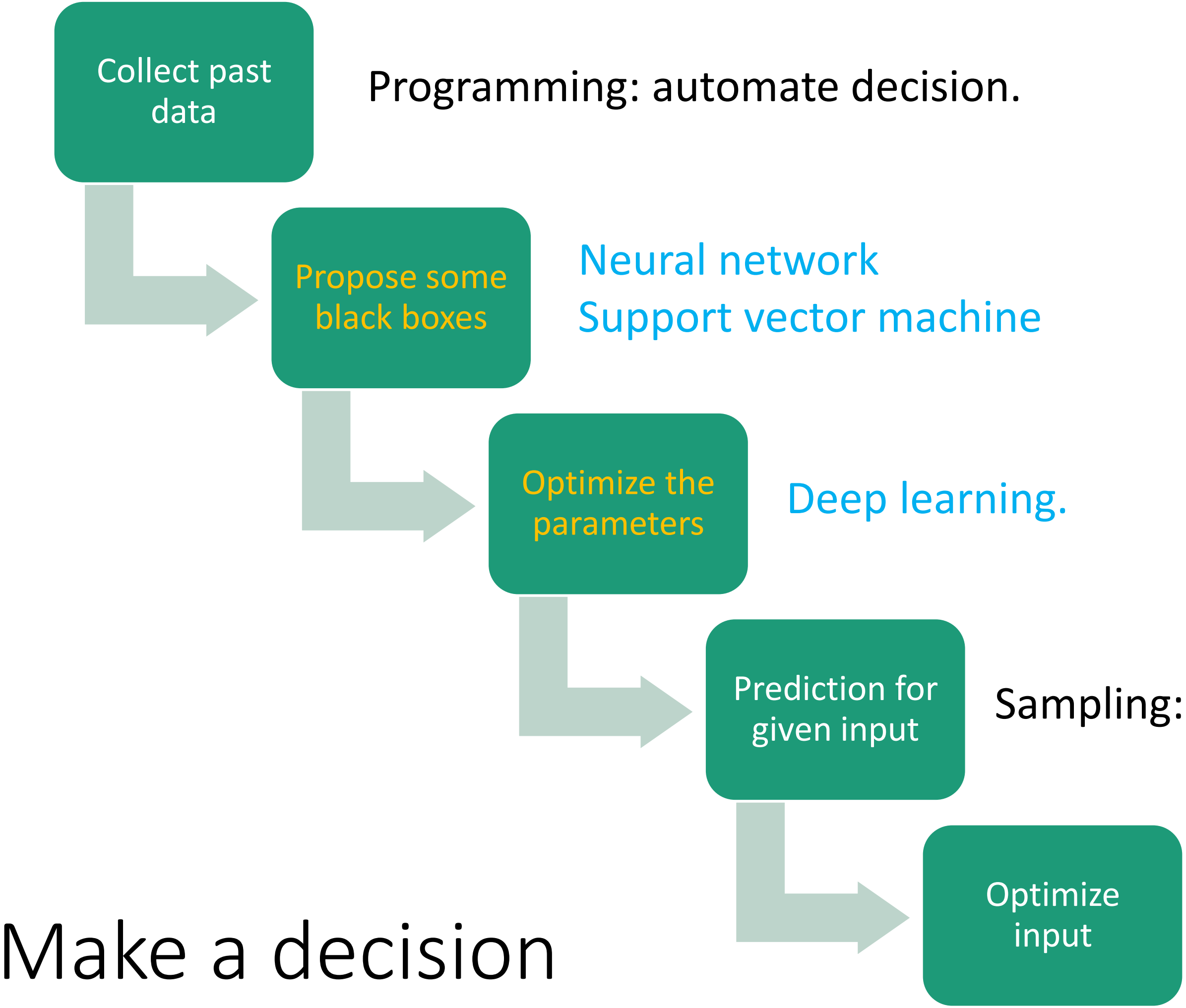
Statistics: test credibility.

Sampling: calculate complicated objective.

Optimization: optimize objectives.

No models to propose?

Make a decision



Collect past data

Propose some black boxes

Optimize the parameters

Prediction for given input

Optimize input

Programming: automate decision.

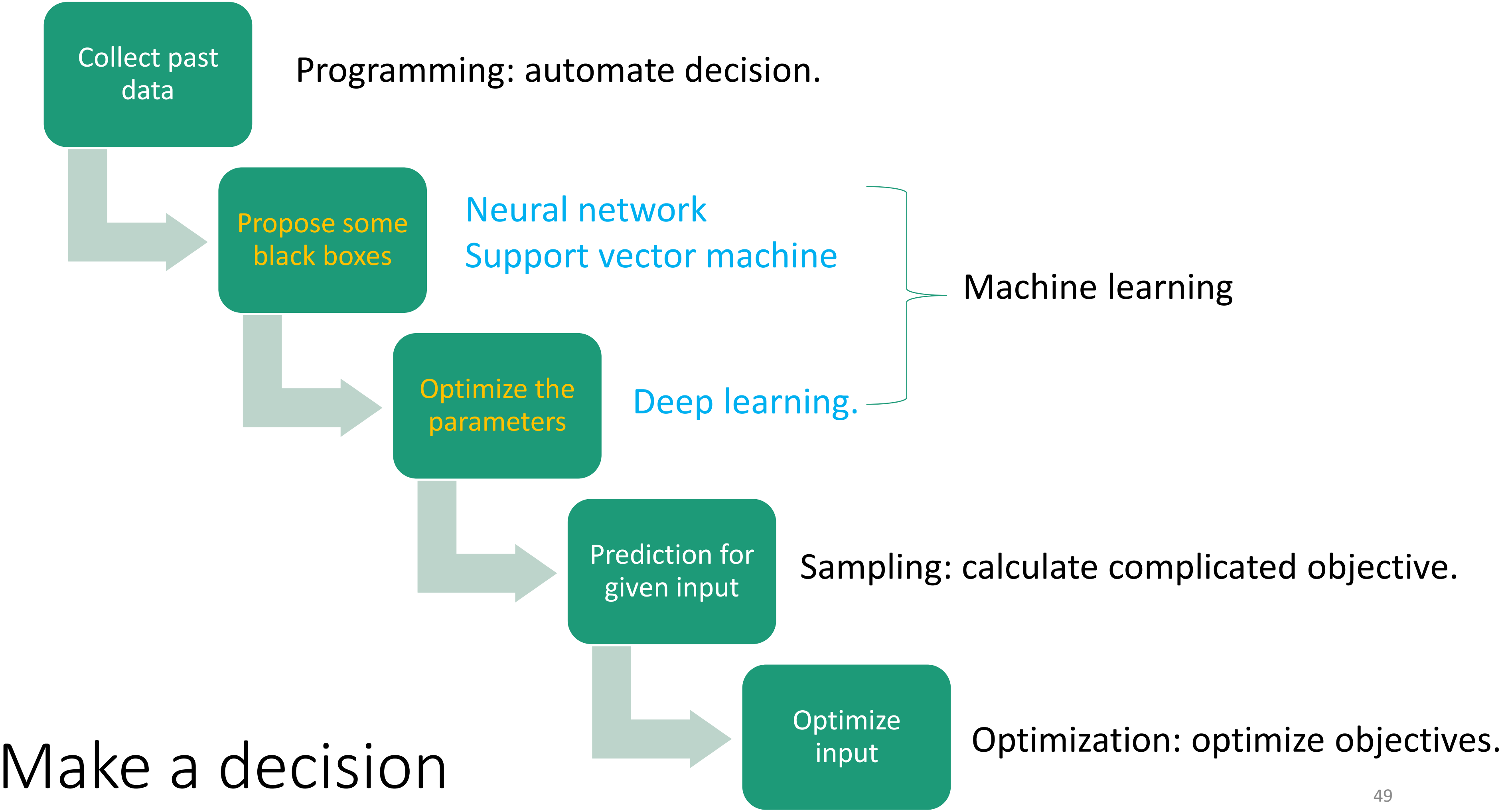
Neural network
Support vector machine

Deep learning.

Sampling: calculate complicated objective.

Optimization: optimize objectives.

Make a decision



Collect past data

Programming: automate decision.

Propose some black boxes

Neural network

Support vector machine

Optimize the parameters

Deep learning.

Machine learning

Prediction for given input

Sampling: calculate complicated objective.

Optimize input

Optimization: optimize objectives.

Distinction with other courses

- This is an interdisciplinary course and will cover broadly
- More focus on the big picture and the connection between different concepts
- Probability: STA2001/STA2003
- Statistics: STA2002/STA2004
- Optimization: MAT3007
- Machine Learning: DDA3020