Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

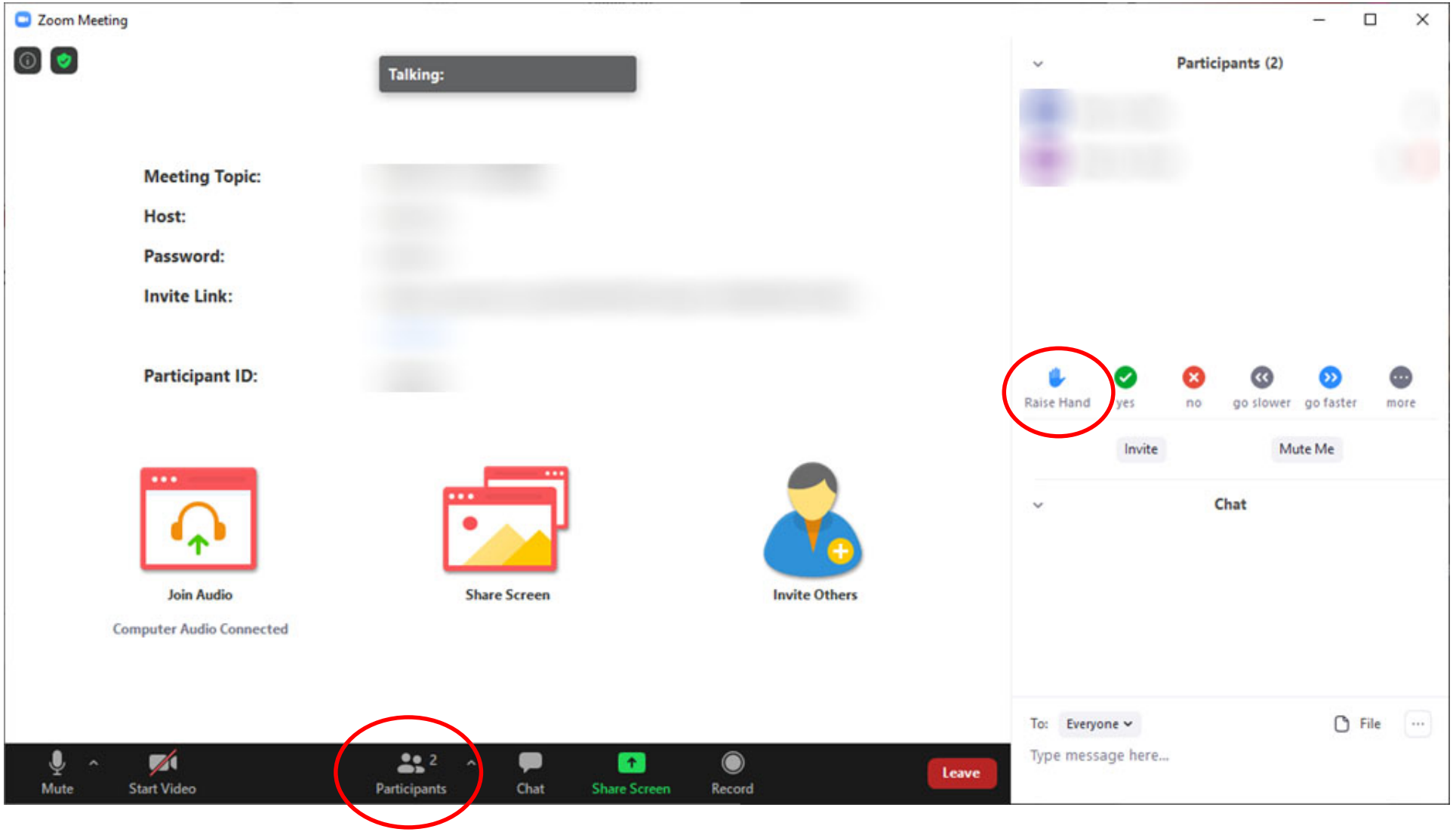

WELCOME
BACK TO
the
ClassroOM

A mixture of external mode and flexible mode

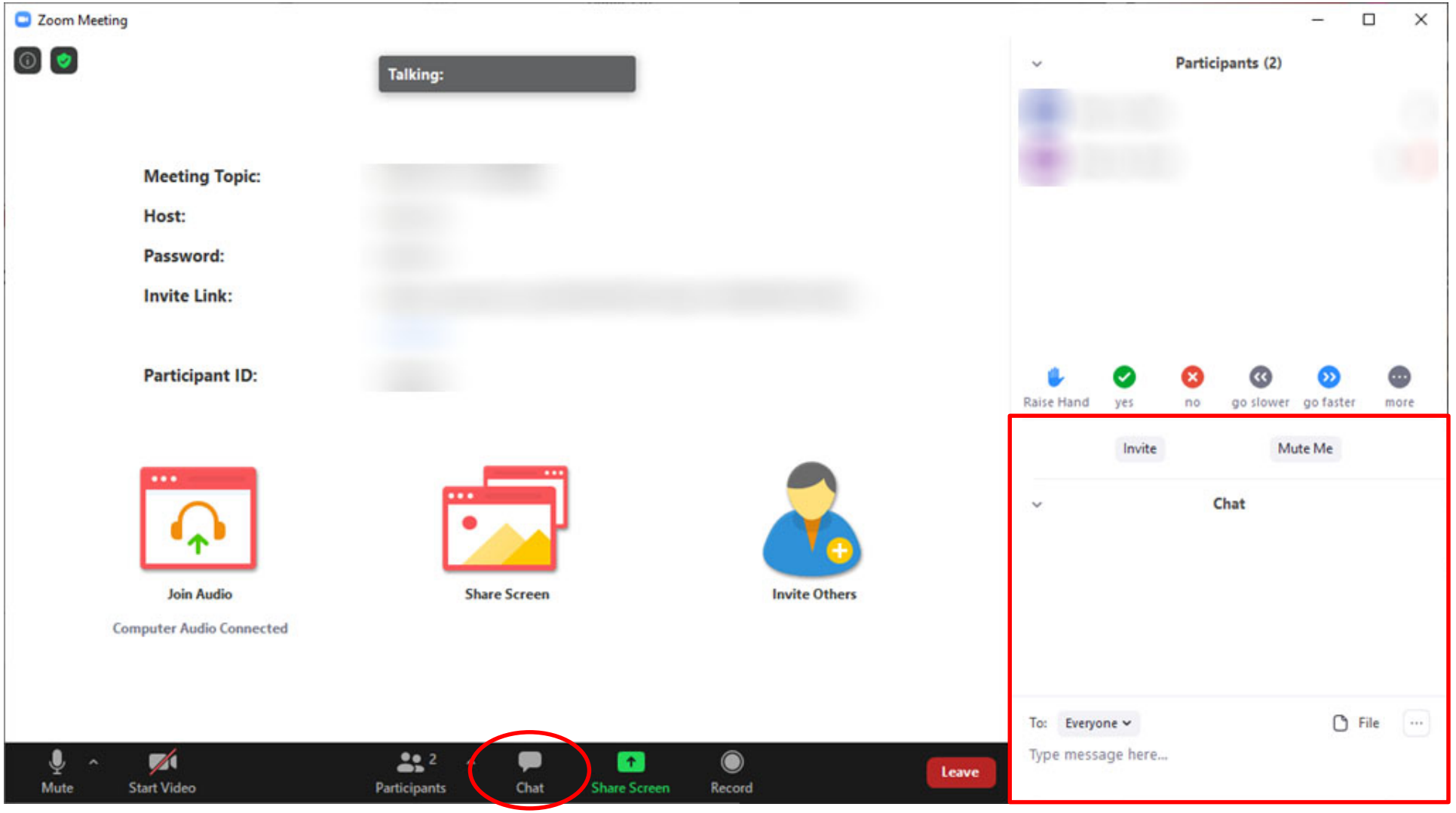
- Lectures:
  - Interactively online
- Contact:
  - Interactively online
- Tutorials:
  - Online and on campus

Stay healthy and enjoy the course!

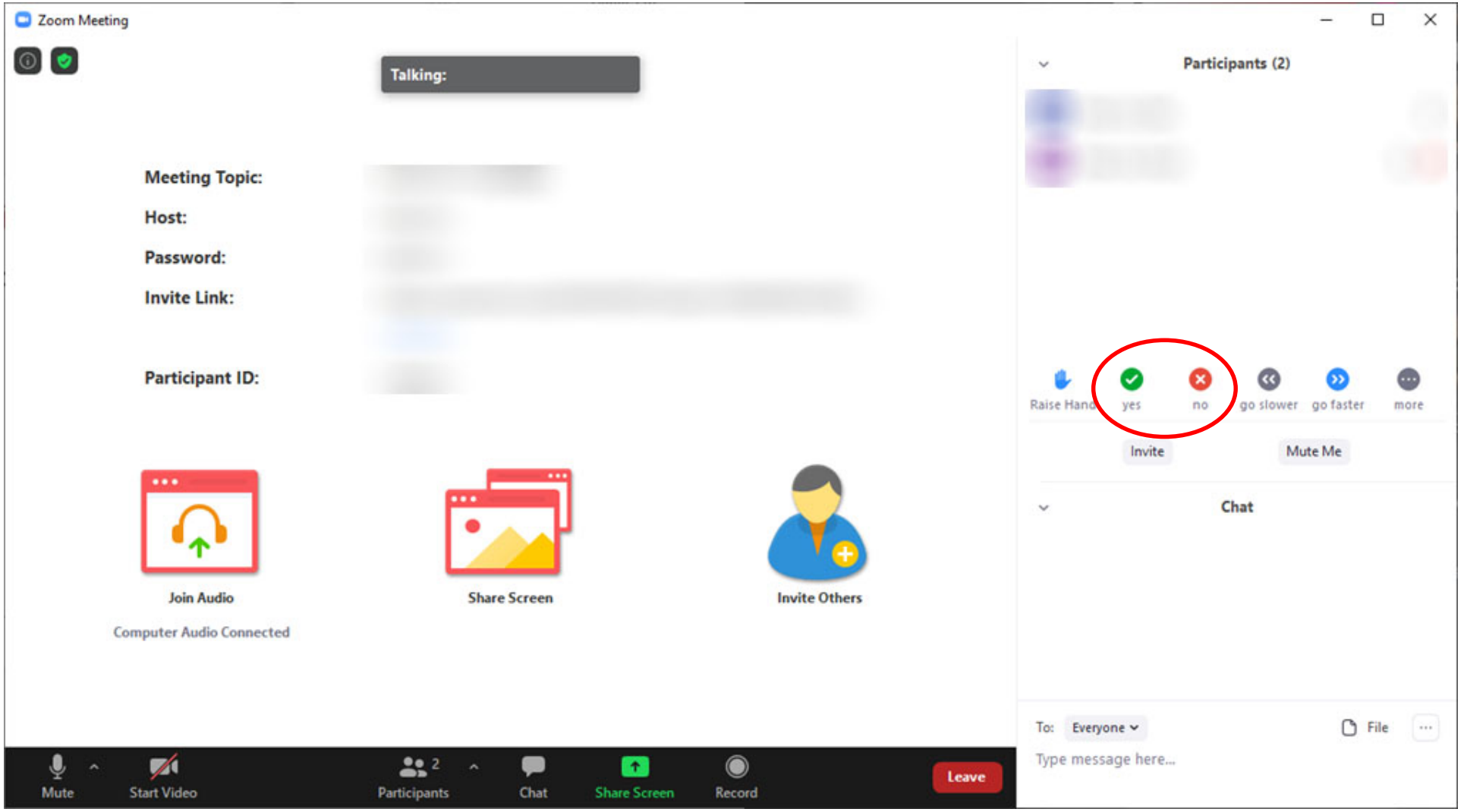
Zoom tips: asking a question by “Raise Hand”



Zoom tips: asking a question by “Chat”



Zoom tips: answering a yes/no question



This lecture:

- Basic information for INFS4203/7203

- Introduction to Data Mining

Basic Information for INFS4203/7203

Outline

- Who are your fellows?
- What you are supposed to learn?
- What you are supposed to do?
- Where you can ask for help?

Outline

- Who are your fellows?
- What you are supposed to learn?
- What you are supposed to do?
- Where you can ask for help?

Quick facts

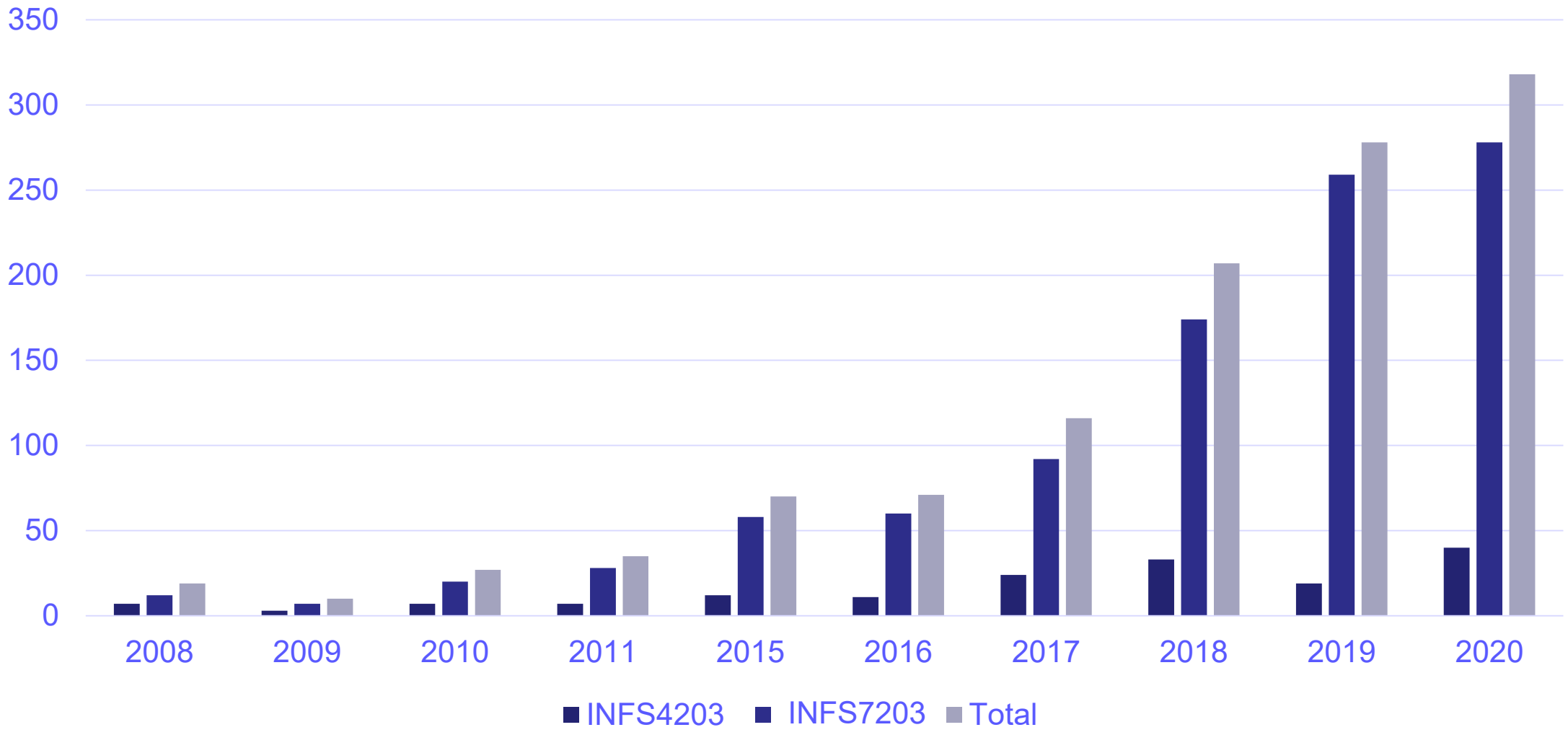
- Bachelor/Master/Grad Cert./Grad Dip./Study Abroad in
  - Data Science/ Computer Science/ Information Technology/ Bioinformatics/ Business/ Commerce/ Engineering Science/ Engineering/Geographic Information Science/ Applied Econometrics/ Science/ Mathematics…

Course	Mode	Enrolment
INFS4203 (Undergraduate)	Flexible	29
	External	13
INFS7203 (Postgraduate)	Flexible	193
	External	119

Activity: what is your major

apps.elearning.uq.edu.au/wordcloud/75753

Enrolment


350
300
250
200
150
100
50
0
2008
2009
2010
2011
2015
2016
2017
2018
2019
2020
INFS4203
INFS7203
Total

Activity: group discussion

- Task 1: introduce yourself
- Task 2: discuss why you take the course
- Task 3: what do you expect from this course
- Task 4: one representative records **Task 2** and **3** in https://padletuq.padlet.org/miaoxu/m7lfg4880nf26jis

Outline

- Who are your fellows?
- What you are supposed to learn?
- What you are supposed to do?
- Where you can ask for help?

INFS courses

- INFS1200/7900 Intro to Information Systems
- **INFS2200/7903 Relational Database Systems**
- INFS3200/7907 Advanced Database Systems
- INFS3202/7202 Web Information Systems
- INFS3208/7208 Cloud Computing
- **INFS4203/7203 Data Mining**
- INFS4205 Advanced Technology for High Dimensional Data
- INFS7410 Information Retrieval
- INFS7450 Social Media Analytics
- INFS7901 Database Principles

Pre-Requisites

INFS1200/7900 vs. INFS2200/7903

- Recall what you have learned?
- What are the difference between these two courses?

INFS1200/7900
Intro to Information Systems

INFS2200/7903
Relational Database Systems

INFS1200/7900 vs. INFS2200/7903

INFS1200/7900
Intro to Information Systems

INFS2200/7903
Relational Database Systems

INFS4200/7203 Data mining

"Data mining is 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."

[D. Hand et al. , Principles of Data Mining]

Data mining vs. machine learning

Data mining vs. machine learning

- Database originated
- Data-centred

- AI originated
- Learning-centred

Yes/No questions

- Is data mining the same as machine learning?
- Is data mining totally different from machine learning?
- Is data mining the result of the evolution of database technologies?
- Is data mining the result of the evolution of machine learning research?

Outline

- Who are your fellows?
- What you are supposed to learn?
- What you are supposed to do?
- Where you can ask for help?

Learning activities

	Lectures	Tutorials	Contact
When	Fri. 14:00-16:00	8 sessions on campus ~~4~~5 sessions online	Mon. 17:00-20:00
Where	Zoom	14-115/Zoom	Zoom

- Tutorials:
  - Start in Week 3
  - Problem Solving
  - Python Language

  https://www.learnpython.org/
- Contact:
  - Start in Week 2
  - Consultation
  - Not Compulsory
  - Join/Leave Freely

Staying Healthy @ UQ

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Stay home if you are unwell

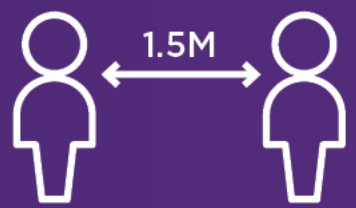
Cover your mouth and nose when you sneeze or cough

Avoid touching your face

Wash your hands thoroughly

Don't share personal items

Clean surfaces



Maintain space between each other

Put used tissues in the bin

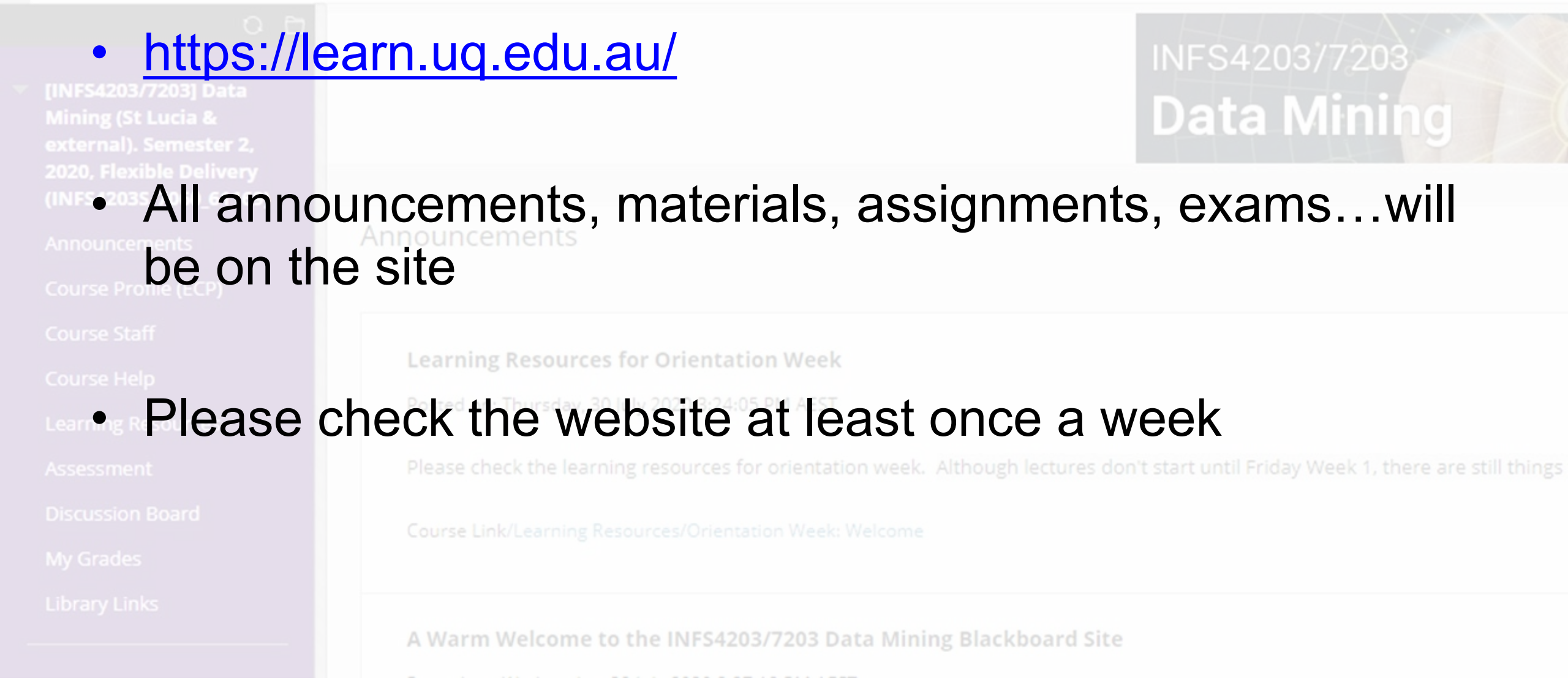
Call your General Practitioner (doctor) or UQ Health Care and explain your symptoms

CRICOS 00025B

Need the facts? **about.uq.edu.au/coronavirus-advice-uq-community**

Course website

- https://learn.uq.edu.au/



- All announcements, materials, assignments, exams…will be on the site
- Please check the website at least once a week

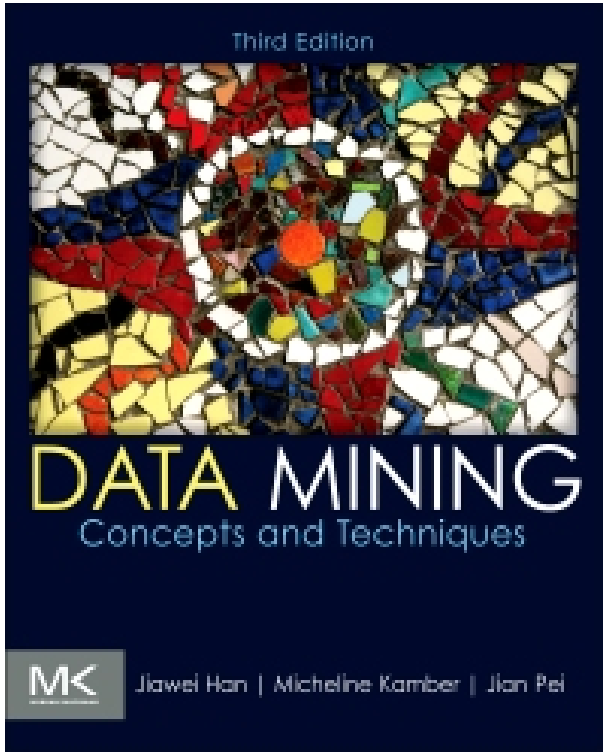
Learning materials

- Lecture notes/records
- Tutorial notes/records
- Weekly summarized frequently asked questions

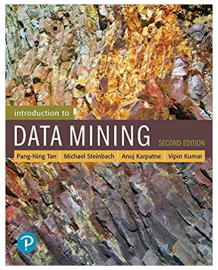
Recommended reference book

- Jiawei Han, Micheline Kamber, Jian Pei. *Data mining concepts and techniques 3rd ed*. Morgan Kaufmann, 2012.

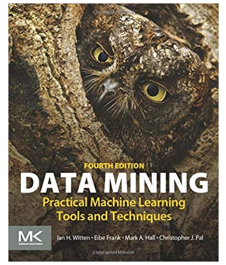
- Additional chapters from *2nd ed*
  - Mining stream and time-series data
  - Mining graph data
  - Mining text and web data



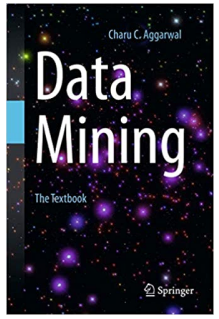
Other data mining books (for your interest)



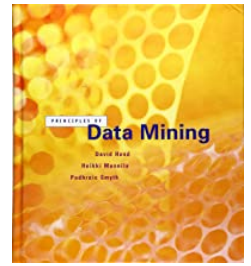
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar. *Introduction to Data Mining*, 2nd ed. Pearson, 2018. (the library link is for a previous version)



- Ian H. Witten, Eibe Frank, Mark A. Hall and Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4th edition, 2016.



- Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.



- David Hand, Heikki Mannila and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

Individual assessment

Assessment	When	Percentage
Assignment1	Due by Aug. 27th	4%
Assignment2	Due by Sep. 10th	6%
Assignment3	Due by Oct. 8th	6%
Assignment4	Due by Oct. 22nd	6%
Assignment5	Due by Oct. 29th	3%
Mid-term	Sep. 18th in class	25% Online Close-Book
Final-term		50% Invigilated Online Close-Book

Uh, Mom?
Remember that paper
you wrote for me about nuclear
reactors? What does it mean
when the big, red “meltdown”
light is flashing?

AI Awareness Poster, Brigham Young University
https://live-academicintegrity.pantheonsite.io/wp-content/uploads/2017/12/462c19_1dc7602ba4ec4f4586d803e7c43904cc_mv2-1.gif

Outline

- Who are your fellows?
- What you are supposed to learn?
- What you are supposed to do?
- Where you can ask for help?

Course staff

- Course coordinator: Dr. Miao Xu
  - Email: miao.xu@uq.edu.au
  - Office: General Purpose South Building (78) 633, St Lucia
- Tutors (see Blackboard for their contacts)
  - Mr. Rocky Chen
  - Ms. Sivangi Mund
  - Ms. Skye Sun
  - Mr. Zijian Wang
  - Mr. Ziwei Wang

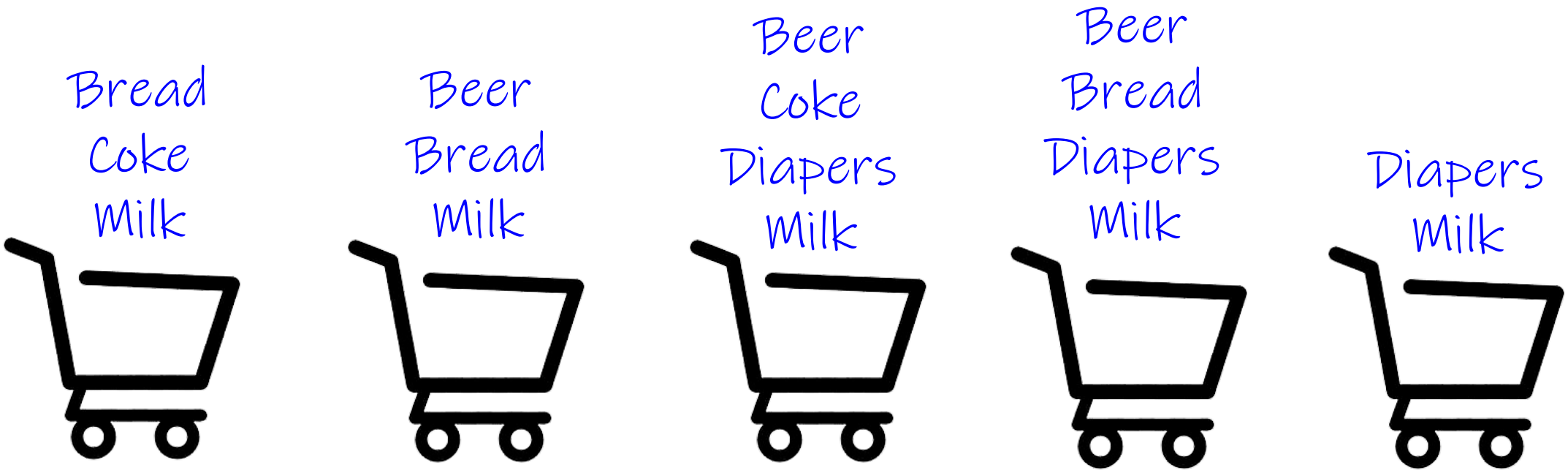
Ask for help

- Ask in class or immediately before/after class
- Post it on ~~Blackboard discussion board~~ Piazza
- Bring it to "Contact" time (Mon. 5pm-8pm)
- Email tutors
- Email course coordinator
- Visit the office: by appointment

You are encouraged to visit the **Piazza** to share your thoughts and help your fellows!

Lecture 1: Introduction to Data Mining

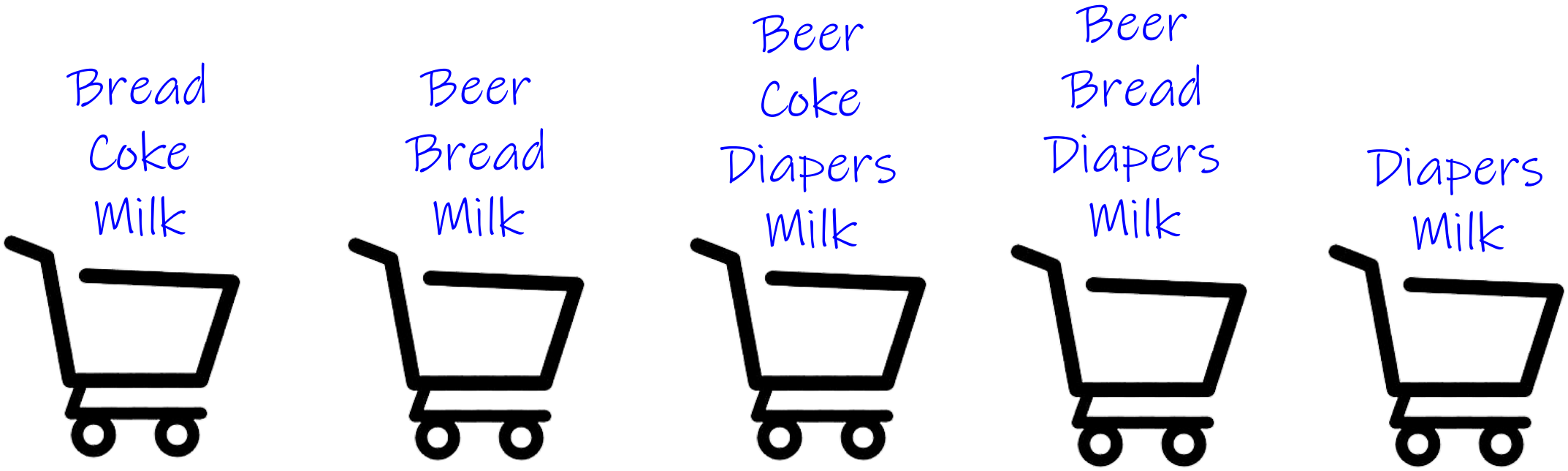
Example 1: Mining transaction data



Anything interesting?

- How many items are there?
- Do they appear frequently?
- Do they appear together?
- Any patterns?
- What can be the implicit reasons?

Example 1: Mining transaction data

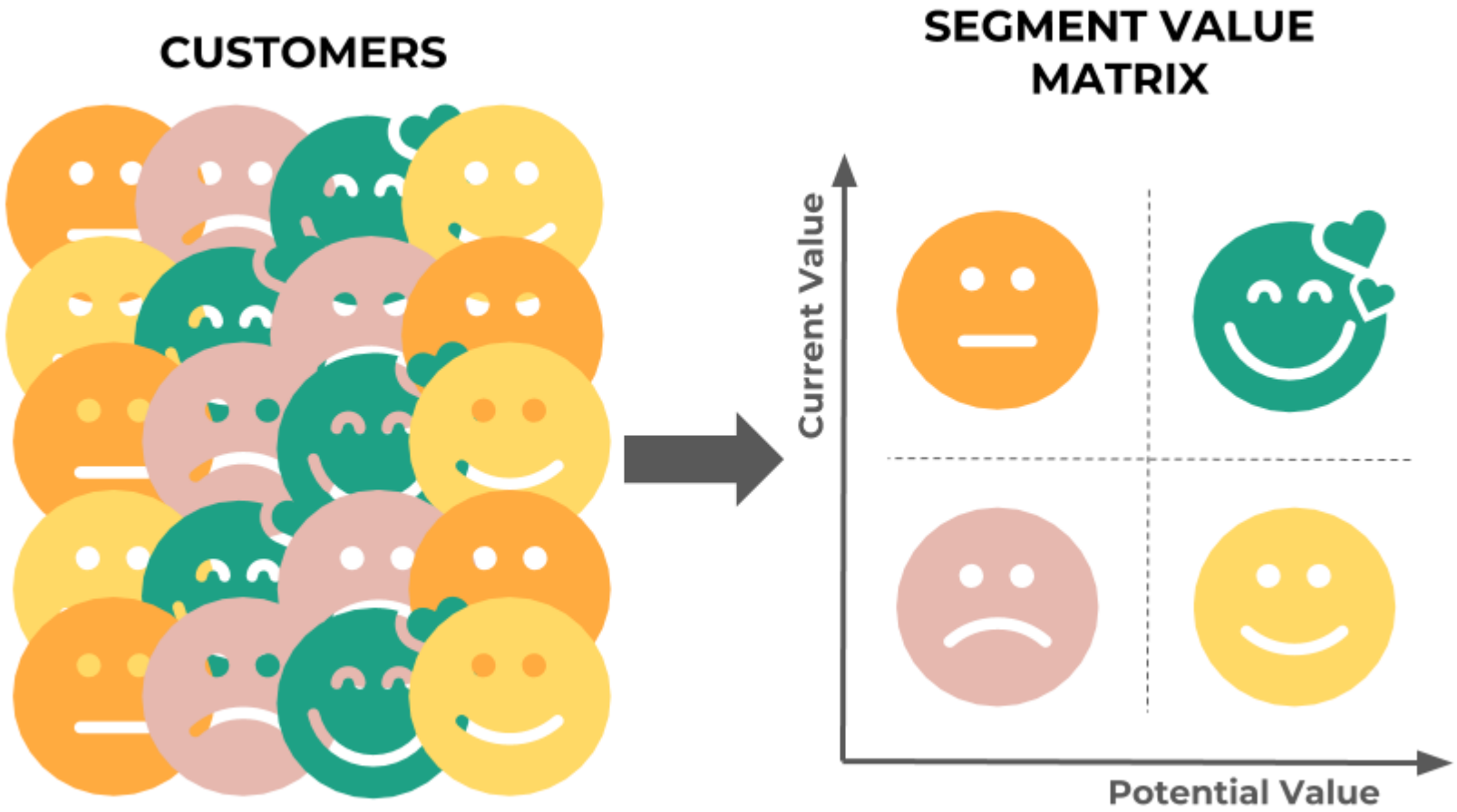


Anything interesting?

- ✓ Bread→ Milk 100%
- ✓ Diapers →Milk 100%
- ✓ Diapers→ Beer 66%

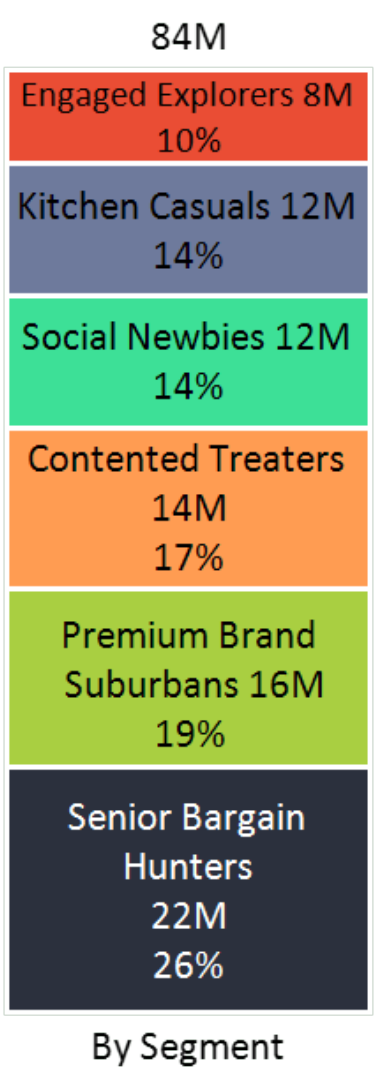
https://medium.com/@ghumare64/market-basket-analysis-using-association-rule-mining-2f87e4064ee7

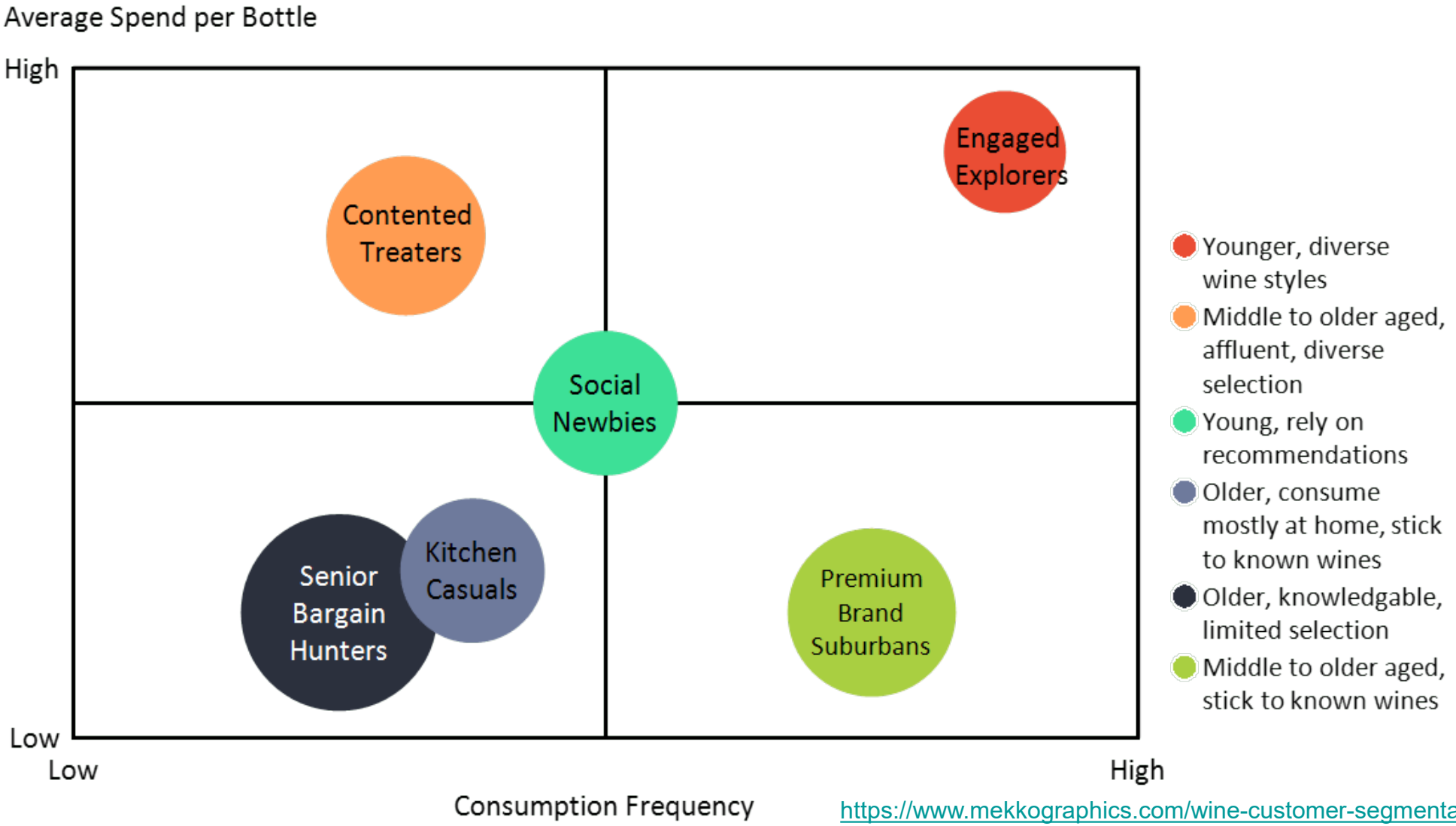
Example 2: Mining customer data



https://towardsdatascience.com/how-to-automatically-segment-customers-using-purchase-data-and-a-few-lines-of-python-36939fb587a4

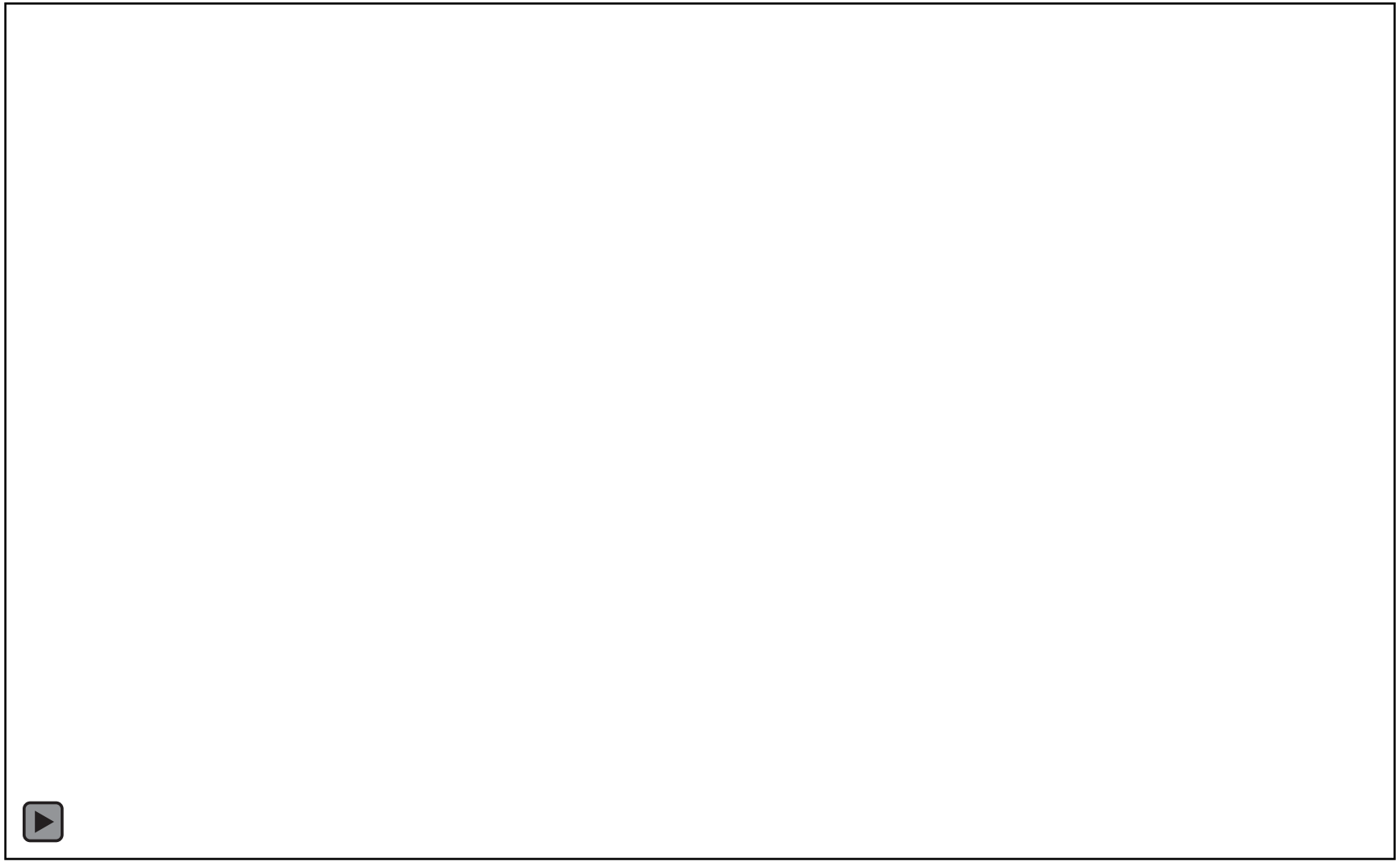
Example 2: Mining customer data





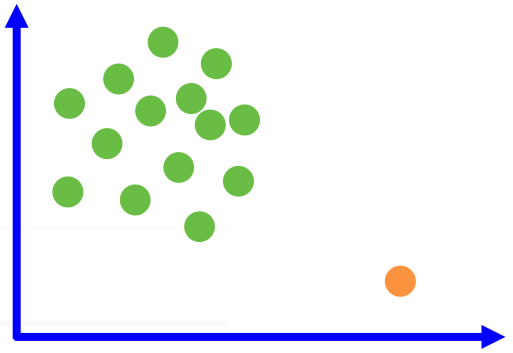
https://www.mekkographics.com/wine-customer-segmentation/

Example 3.1: Mining video surveillance data



https://cdn.vox-cdn.com/thumbor/9Hirq_5MsKwoDhz1biJWMJ1BbQY=/0x0:601x371/620x413/filters:focal(253x138:349x234):gifv():no_upscale()/cdn.vox-cdn.com/uploads/chorus_image/image/60183483/stop_thief_ai_cctv_automated_surveillance.0.gif

Example 3.2: Credit card fraud detection

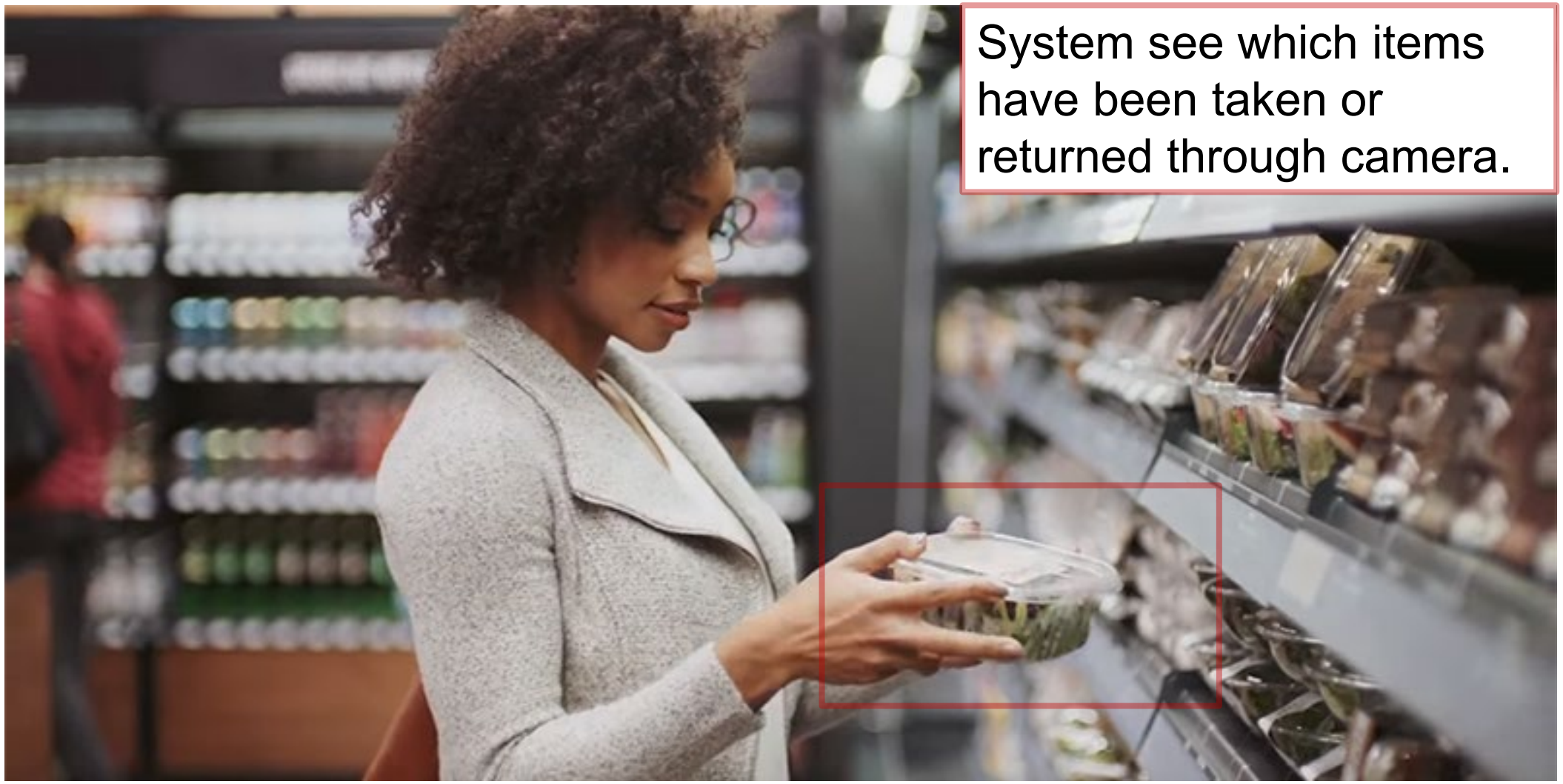


https://logentries.com/product/anomaly-detection/

Example 4: Amazon Go

https://tamebay.com/2020/04/amazon-just-walk-out-technology-now-available-to-retailers.html

Example 4: Amazon Go

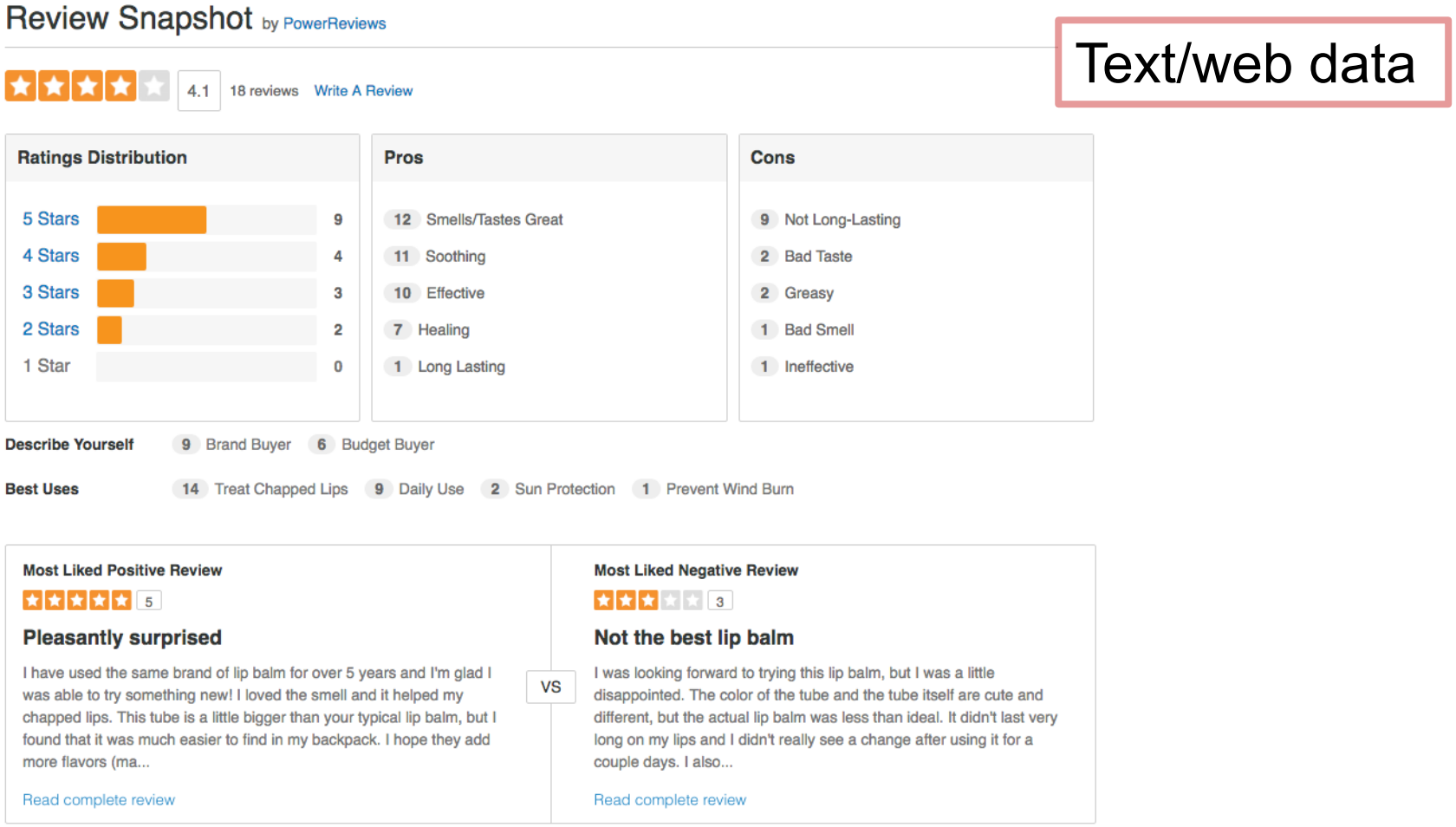


https://www.slideshare.net/AmazonWebServices/artificial-intelligence-machine-learning-on-aws

Example 5.1: sales prediction

Example 5.1: sales prediction

Example 5.2: an online store



Text/web data

https://www.powerreviews.com/blog/how-to-get-more-product-reviews/

*After learning the data mining course, you will have a **fundamental** knowledge of how these are achieved.*

Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

Why data mining

- How will you describe the haystack?
- How will you describe the golden needle?
- What is the meaning of fire on the top?
- Is this task hard?
- Why is it hard?

“Search for a golden needle in a haystack. So much hay and so little time”

https://www.siam.org/meetings/sdm08/TS2.ppt

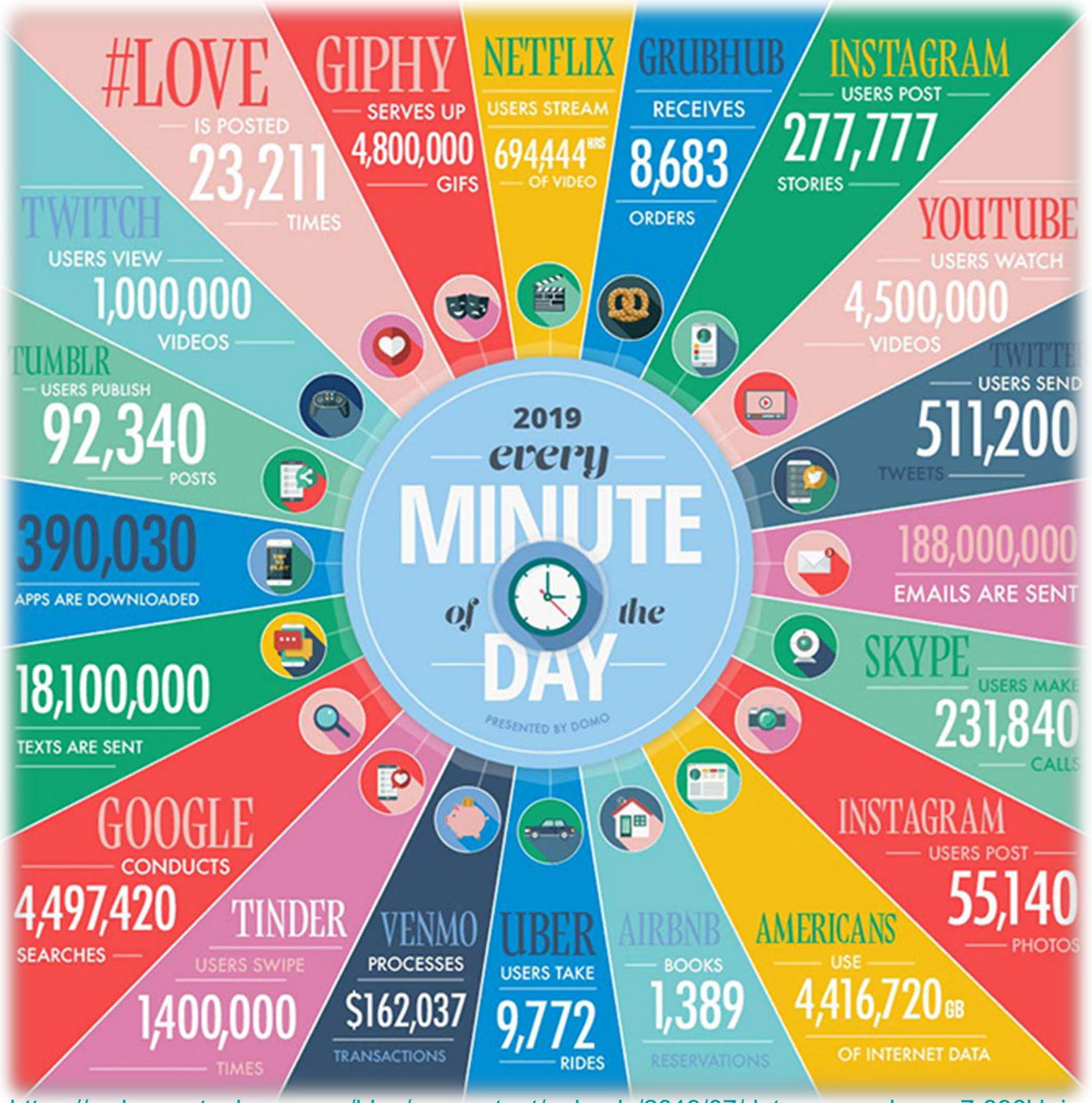
Why data mining

- ✓ Volume
- ✓ Value
- ✓ Velocity
- ✓ Variety

“Mining golden needle in a haystack. So much hay and so little time”

Data on the web

- ✓ Volume
- ✓ Value
- ✓ Velocity
- ✓ Variety



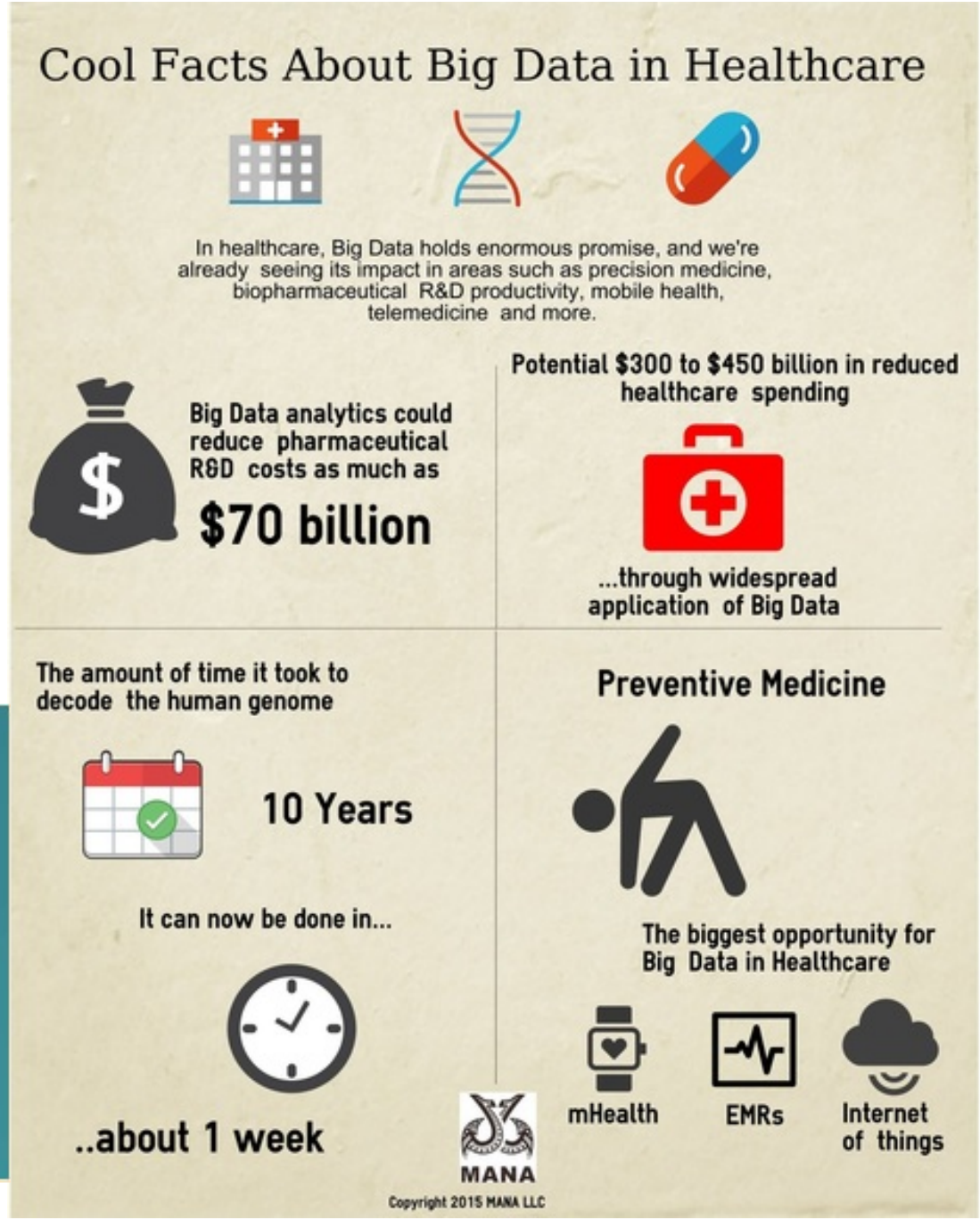
https://web-assets.domo.com/blog/wp-content/uploads/2019/07/data-never-sleeps-7-896kb.jpg

Data in healthcare

Data gives us the potential to improve quality of health care meanwhile reducing cost.

1PB = 1024 TB

https://www.scoop.it/topic/mobile-health-how-mobile-phones-support-health-care



Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

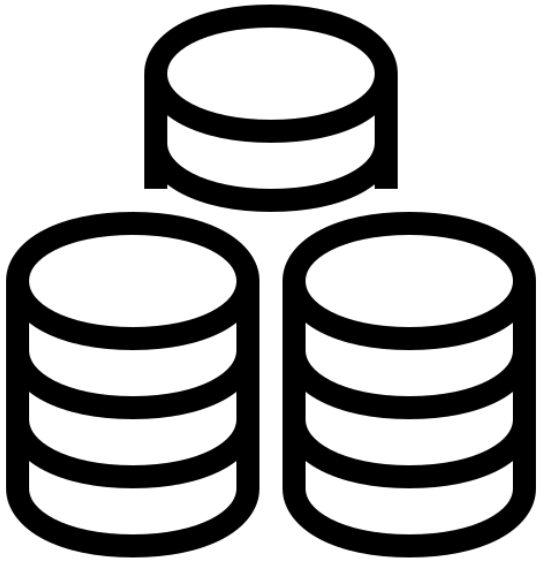
What is data mining

"Data mining is 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."

[D. Hand et al. , Principles of Data Mining]

Associates the words with the objects

Large | Unsuspected | Novel | Understandable | Useful



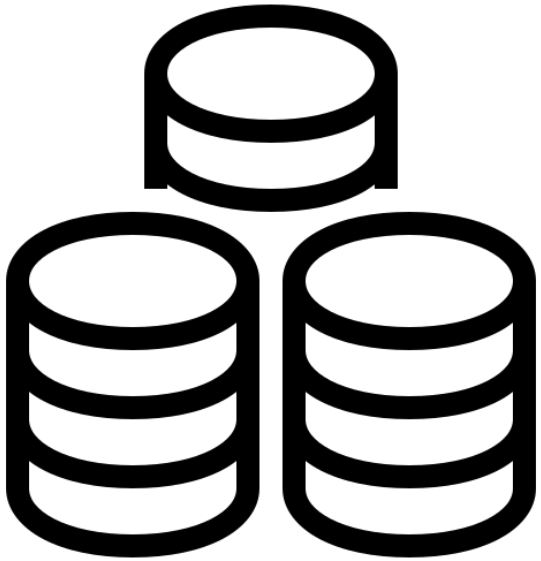
Data

Mining Technique

Knowledge

Associates the words with the objects

Large



Data

Mining Technique

Efficient in both time and space

Novel

Useful

Understandable

Unsuspected

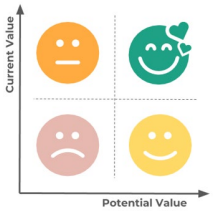
Knowledge

Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

What can data mining do

- Association rule mining (Week 3)
  - How to find highly correlated items out
- Clustering (Week 4, 5)
  - How to segment data with similar traits

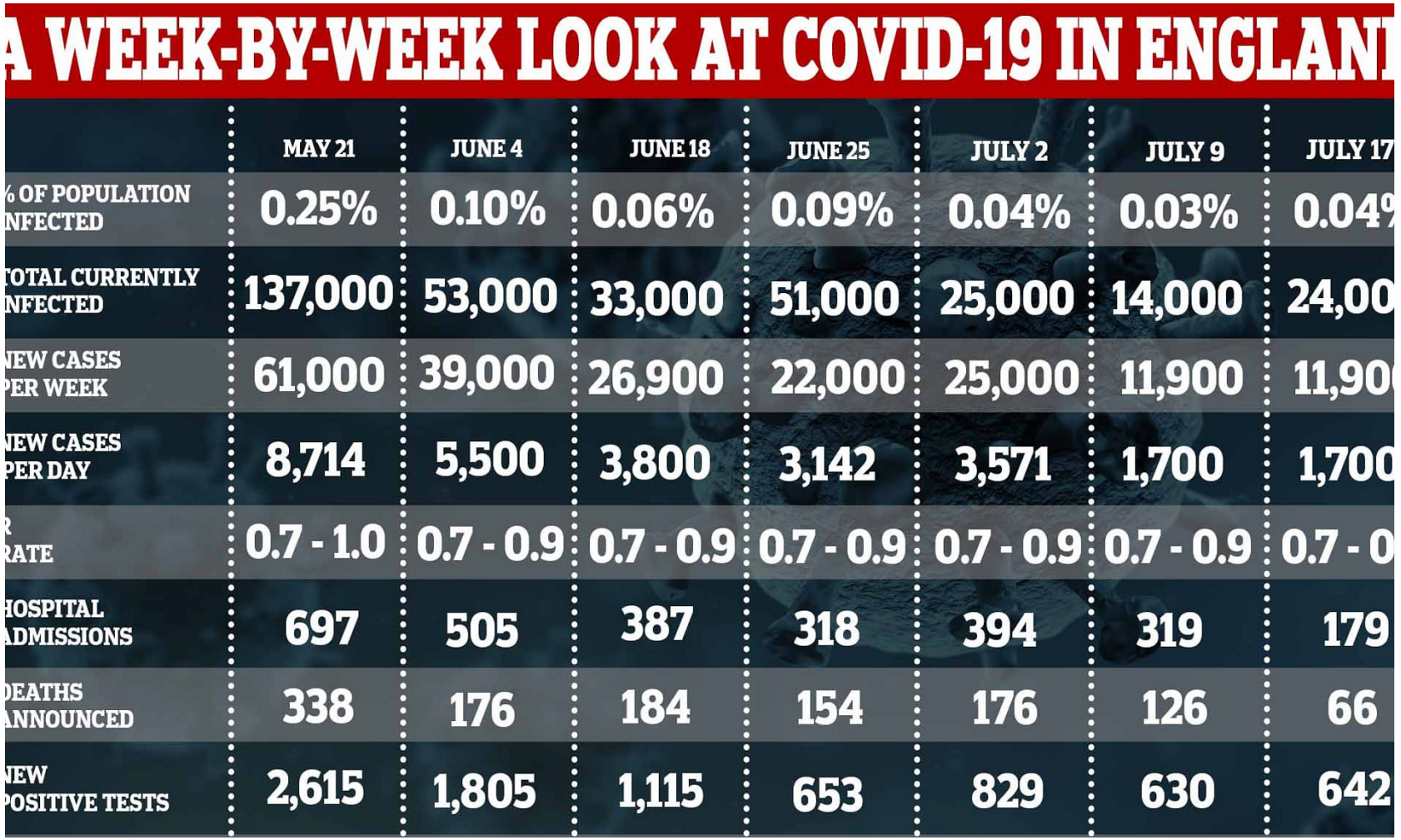


- Anomaly detection (Week 6)
  - How to detect outliers in a dataset

- Classification (Week 8, 9, 10)
  - How to construct the mapping from the input space to the nominal output space

Flat data

“Flat” data: vectors and matrix



A WEEK-BY-WEEK LOOK AT COVID-19 IN ENGLAN

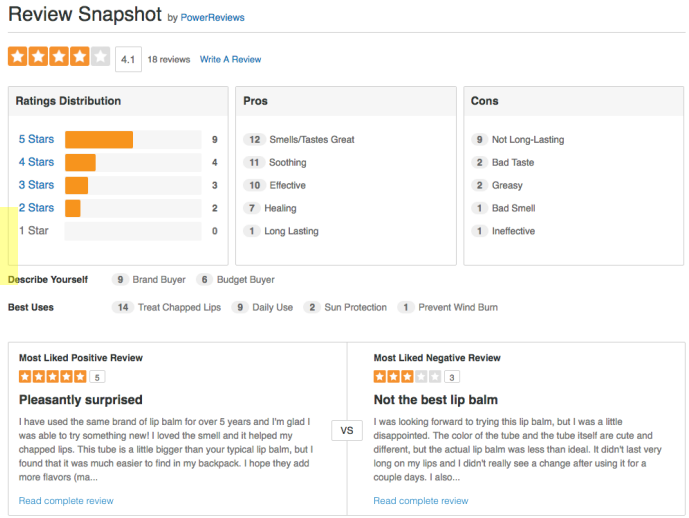
	MAY 21	JUNE 4	JUNE 18	JUNE 25	JULY 2	JULY 9	JULY 17
% OF POPULATION NFECTED	0.25%	0.10%	0.06%	0.09%	0.04%	0.03%	0.04%
TOTAL CURRENTLY NFECTED	137,000	53,000	33,000	51,000	25,000	14,000	24,00
NEW CASES PER WEEK	61,000	39,000	26,900	22,000	25,000	11,900	11,90
NEW CASES PER DAY	8,714	5,500	3,800	3,142	3,571	1,700	1,700
R RATE	0.7 - 1.0	0.7 - 0.9	0.7 - 0.9	0.7 - 0.9	0.7 - 0.9	0.7 - 0.9	0.7 - 0
HOSPITAL ADMISSIONS	697	505	387	318	394	319	179
DEATHS ANNOUNCED	338	176	184	154	176	126	66
NEW POSITIVE TESTS	2,615	1,805	1,115	653	829	630	642

What can data mining do – continue

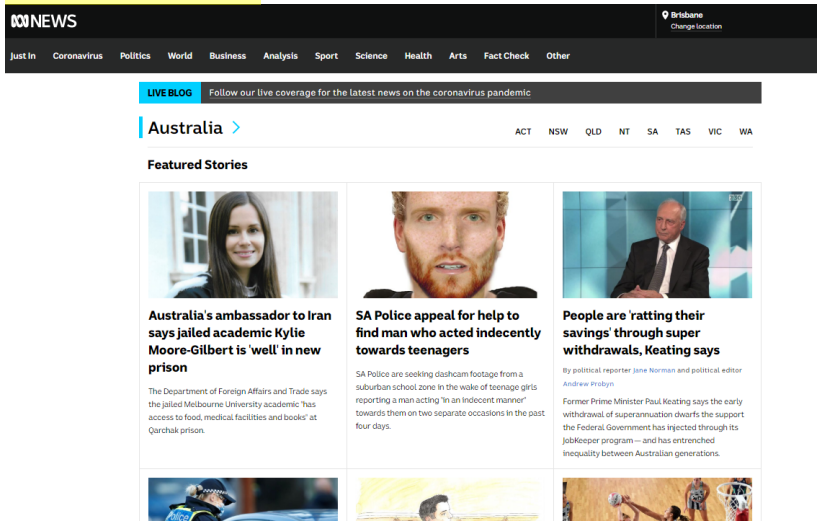
- Mining complex data (Week 11, 12)

Time-series data

Text data



Web data



Graph data

Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

Who need data mining

- Previous Sponsors and Supporters for KDD

Annual salary of data miners

Region	Employer Type	Salary or Income
US/Canada (221)	Company/Self	$131K
	Academic, Government	$103K
Australia/NZ (7)	Company/Self	$172K
	Academic, Government	$60K
W. Europe (78)	Company/Self	$82K
	Academic, Government	$54K
E. Europe (23)	Company/Self	$34K
	Academic, Government	$24K
Africa/Midlle East (7)	Company/Self	$62K
	Academic, Government	$105K
Latin America (12)	Company/Self	$59K
	Academic, Government	$58K
Asia (22)	Company/Self	$47K
	Academic, Government	$59K

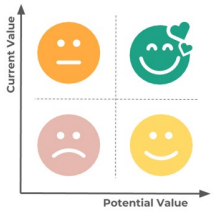
https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html

Outline

- Why data mining?
- What is data mining?
- What can data mining do?
- Who are using data mining?

Next lecture

- Association rule mining (Week 3)
  - How to find highly correlated items out
- Clustering (Week 4, 5)
  - How to segment data with similar traits
- Anomaly detection (Week 6)
  - How to detect outliers in a dataset
- Classification (Week 8, 9, 10)
  - How to construct the mapping from the input space to the nominal output space



Q&A Time

- In Week 2:

No lecture/tutorial due to public holiday.

Contact is open for consultation.