Berkeley
MSSE
MOLECULAR SCIENCE &
SOFTWARE ENGINEERING

Lecture 01:

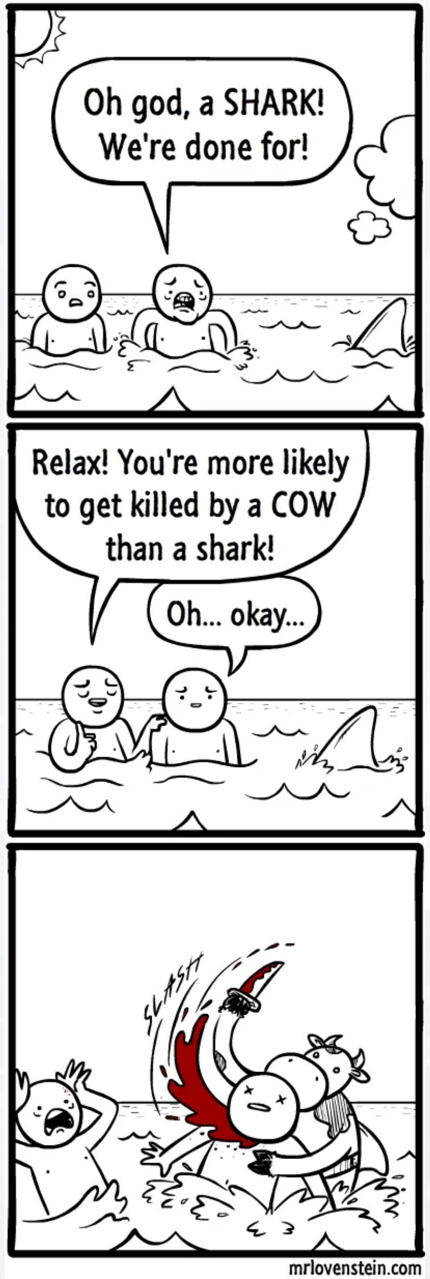
Introduction to Data Science

Markus Hohle

University California, Berkeley

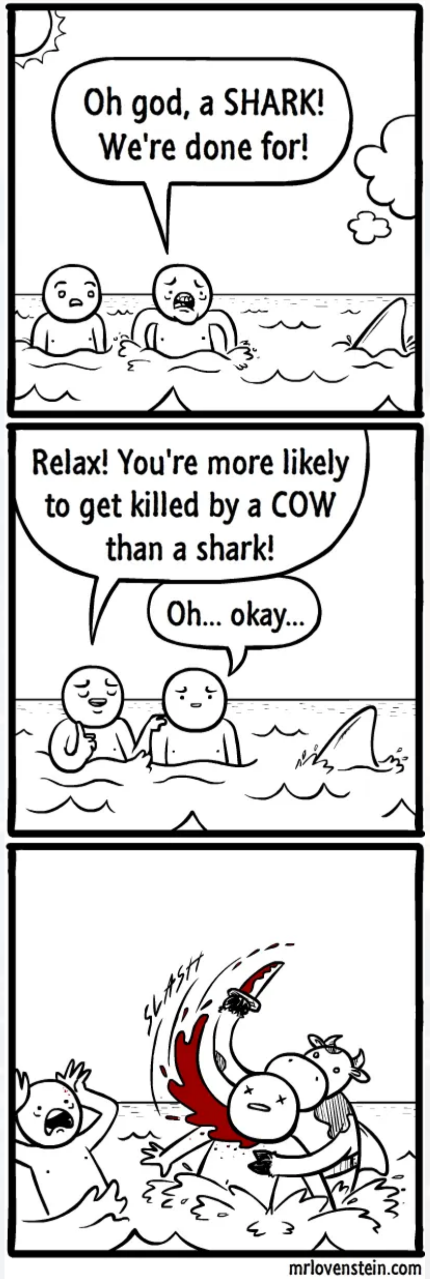
Data Science for Scientific Computing

MSSE 277A, 3 Units



<u>Outline</u>

- motivation for this course
- structure
- syllabus
- guide through the recordings



Outline

- **motivation for this course**
- structure
- syllabus
- guide through the recordings

Introduction

typical workflow:

start:

```
>GBDP50119-19|Aedes africanus|COI-5P|MF183656
TTAAATTTCGATCTGTTAATAATATAGTAATAGCTCCTGCTAAAAC
ATTGCTAAA
GAATCCTCC
```

```
>GBMIN56476-17|Aedes albopictus|COI-5P|KY378921
GTTTTAATTCGTATTGAACTTAGACATCCTGGTATATTTATTGGAAATGA
TACTAGGAGCCCCTGATATAGCTTTTCCTCGAATAAATAATATAAGTTTT
TCATGCTGGGGCTTCAGTTGATTTAGCAATTTTTTCTTTACATTTAGCGG
GTAATTACAGCTATTTTATTACTTCTTTCTCTACCCGTATTAGCCGGAGC
```

raw data

data exploration (EDA)
- How many files?
- How big are they?
- How many sets?
- consistency
- homogeneity
- gaps? Missing data?

data extraction
- sampling
- filtering

feature selection I
- Which information is relevant?
- Which information is redundant?
- Is there missing information/ biases?

feature engineering
- normalization
- encoding
- interpolation/extrapolation

feature selection II
- correlation analysis

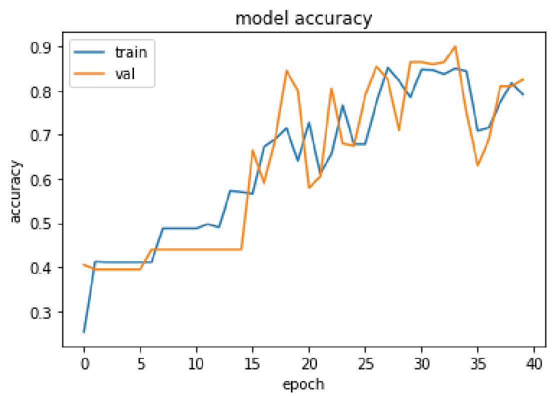
model selection

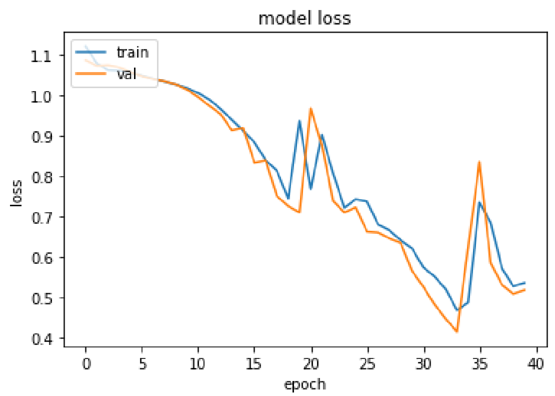
goal: classification/ regression/ prediction/ generation

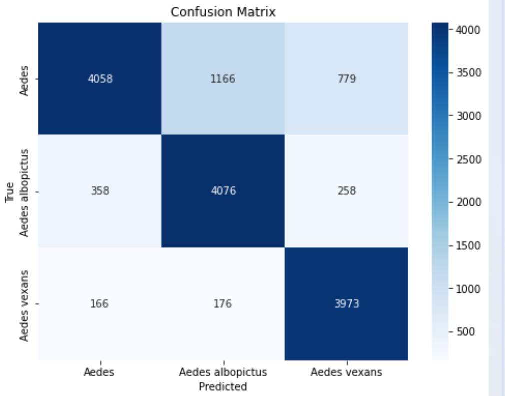
structure: timeseries/images/ sequences/perm invariance

feature selection III
- significance
- "leave one out"

goal: a **model** that:
- well **describes** the data
- covers **relevant information** of the data set
- can be used for **predictions**
- ...







typical tools:

- seaborn
- pandas
- python
- dask
- polars
- matplotlib
- NumPy
- SciPy
- scikit learn
- TensorFlow
- Keras
- PyTorch

start: raw data

data exploration (EDA)
- How many files?

data extraction
- sampling
- filtering

feature selection I
- Which information is relevant?
- Which information is redundant?
- Is there missing information/ ...ses?

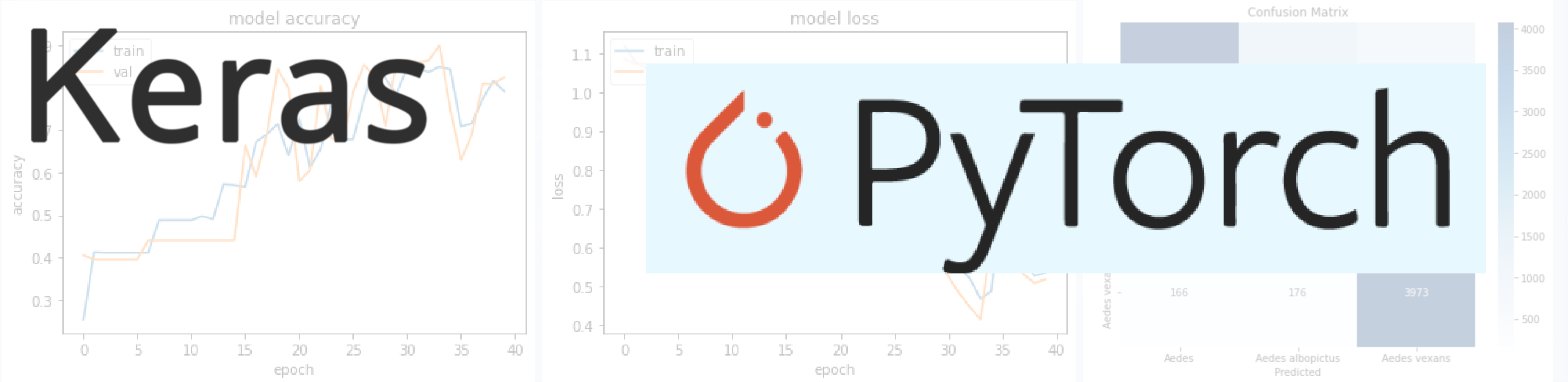
...ysis

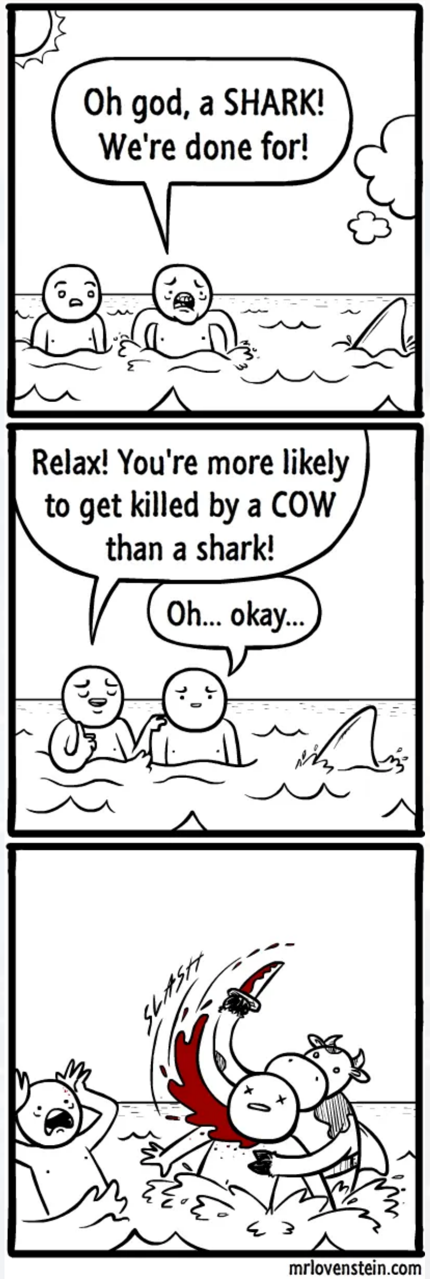
- interpolation/extrapo...

model selection
goal: classification/ regression/ prediction/ generation
structure: ... sequences/perm invariance

feature selection III
- significance
- "leave one out"

goal: a model that:
- well describes th...
- covers relevant information of the data set
- can be used for predictions
- ...





Outline

- motivation for this course
- **structure**
- syllabus
- guide through the recordings

GSI: Elizabeth (Lizzie) Gilson

Toxicology Data Scientist at EPA,

UC Berkeley Alumna (MSSE)

Lecturer: Markus Hohle

Lecturer at UC Berkeley &

Data Analysis Consultant

PhD Physics

Introduction

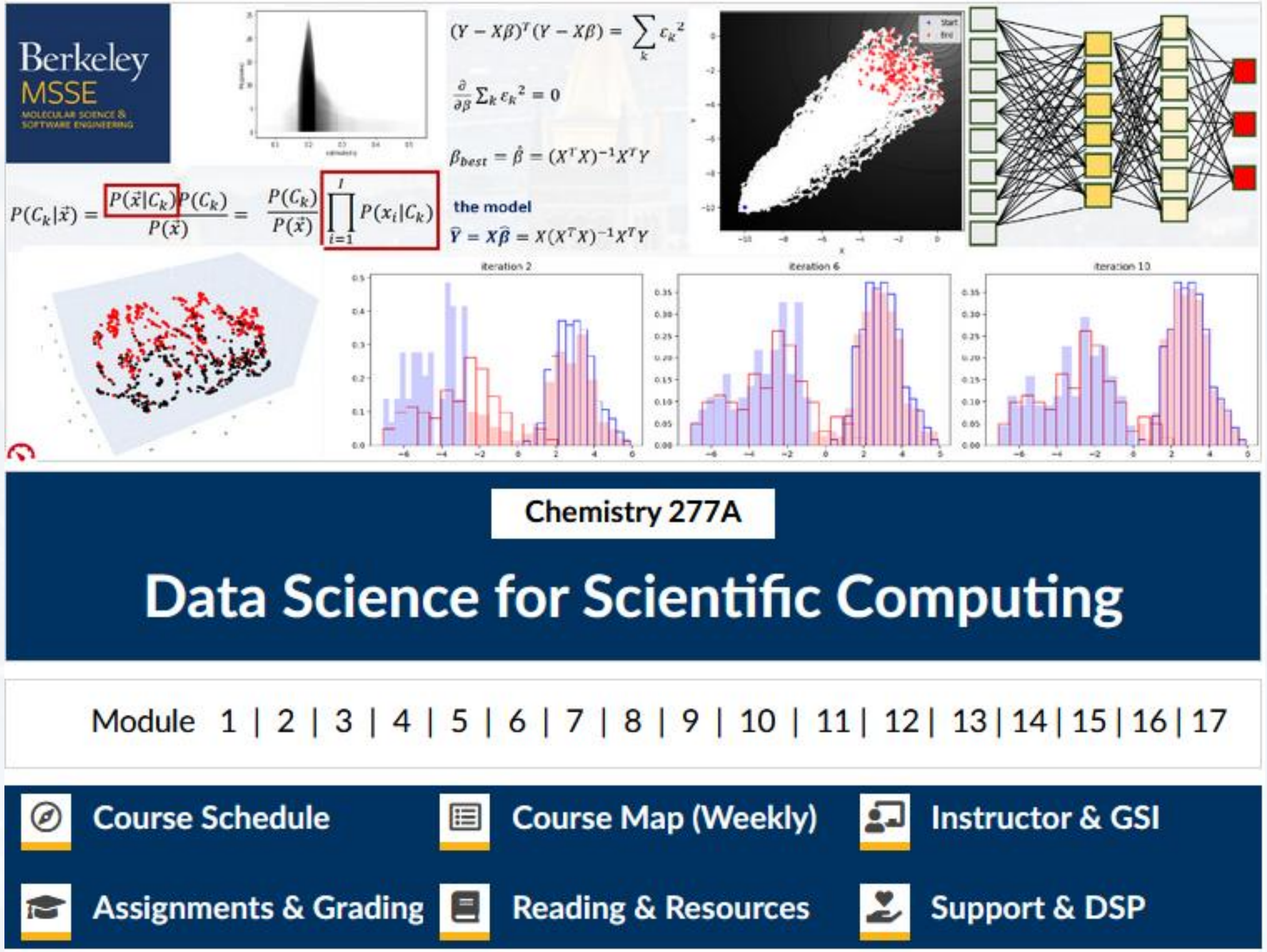
Lecture (Markus):	3 hours of **asynchronous recorded** lectures per week watch any **time, but prior** to **Discussions/Lab Sessions/Homework**	
Discussion (Lizzie/ Markus):	Tuesday,	5:30 – 6:30pm PT
Lab Session (Lizzie/ Markus):	every other Wednesday,	6:00 – 8:00pm PT
Office Hours (Markus):	Friday,	5:00 – 6:00pm PT

Grades:

assignment	**weight**
5 Problem Sets:	40%
2 Programming Projects (midterm & final project)	20%
Lecture Exercises	20%
Discussion & Lab Participation (be active! ask/answer questions!)	20%

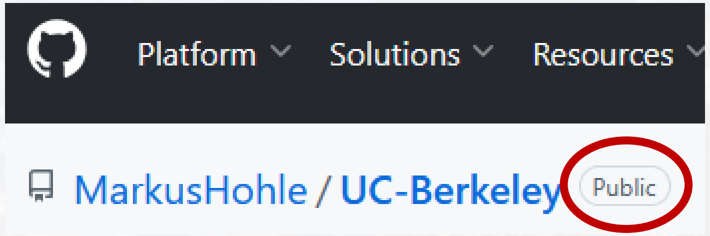
bcourses

- dates & times
- homework assignments and sample solutions
- videos/slides/notebooks/data/codes
- links to discussions/labs/office hours



GitHub

slides/notebooks/data/codes



bcourses

- **dates & times**
- homework assignments and sample solutions
- videos/slides/notebooks/data/codes
- links to discussions/labs/office hours

Course Map

Exact weeks and dates might differ slightly, depending on progress during class. Lectures which fall on a holiday will be recorded and made available on bcourses.

Course Map

Week	Dates	Topics	Reading/Quizzes	Material
1	Jan 20th	Introduction to Data Science		Module 1
2	Jan 26th	Data Sampling and Probability, Pandas		Module 2
3	Feb 2nd	Exploratory Data Analysis (EDA), and Regex Part I		Module 3
4	Feb 9th	Exploratory Data Analysis (EDA), and Regex Part II		Module 4
5	Feb 16th	Introduction to SQL		Module 5
6	Feb 23rd	Feature Analysis, Engineering and Encoding		Module 6

Introduction

bcourses

- **dates & times**
- homework assignments and sample solutions
- videos/slides/notebooks/data/codes
- **links to discussions/labs/office hours**

Course Start Date: Tuesday, January, 20th, 2026

Course End Date: Friday, May 8th, 2026

Spring Recess: Monday, March 23–Thursday, March 26, 2026

All times are PST.

Schedule

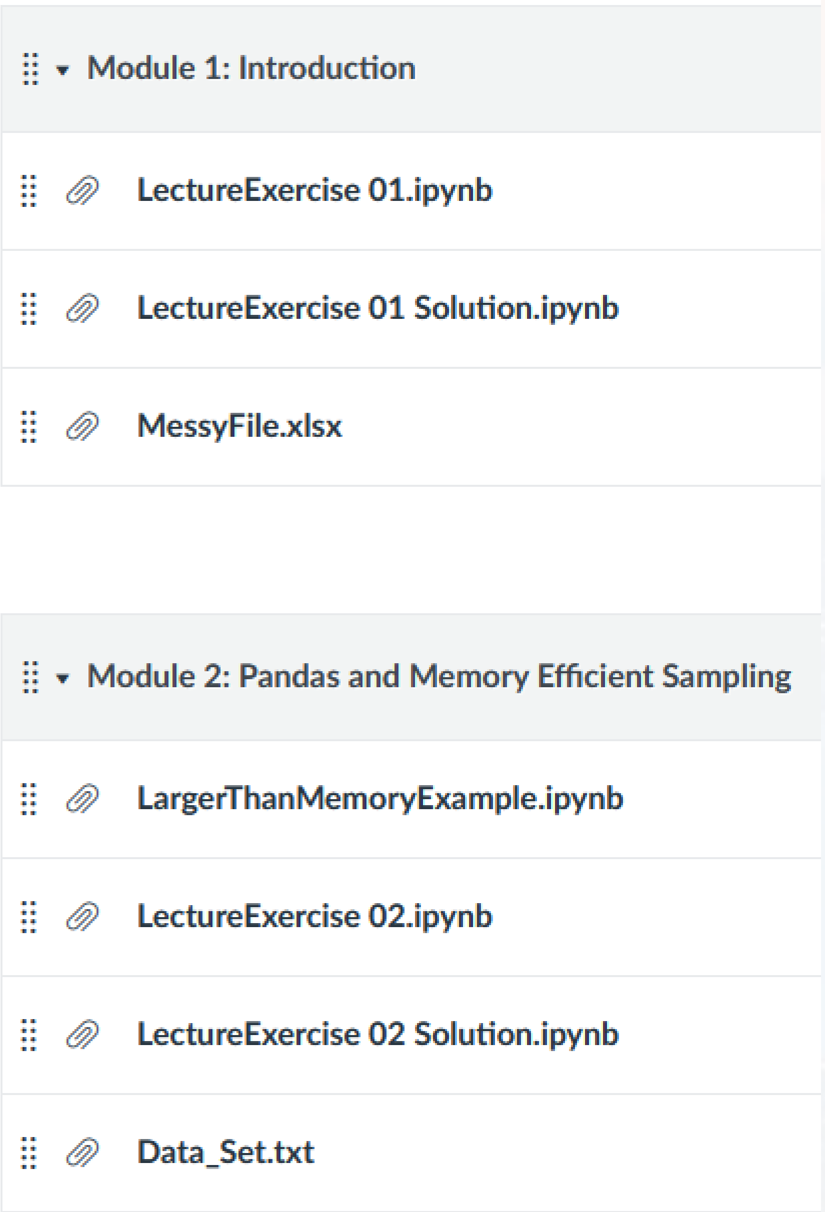
Event	Day	Time	Link
Lecture	asynchronous	3 hours prerecorded	
Discussions	Tuesday	5:30 pm - 6:30 pm	here
Lab (every other week)	Wednesday	6:00 pm - 8:00 pm	here
Office Hour GSI	TBD	TBD	here
Office Hour Lecturer	Friday	5:00 pm - 6:00 pm	here

Introduction

structure

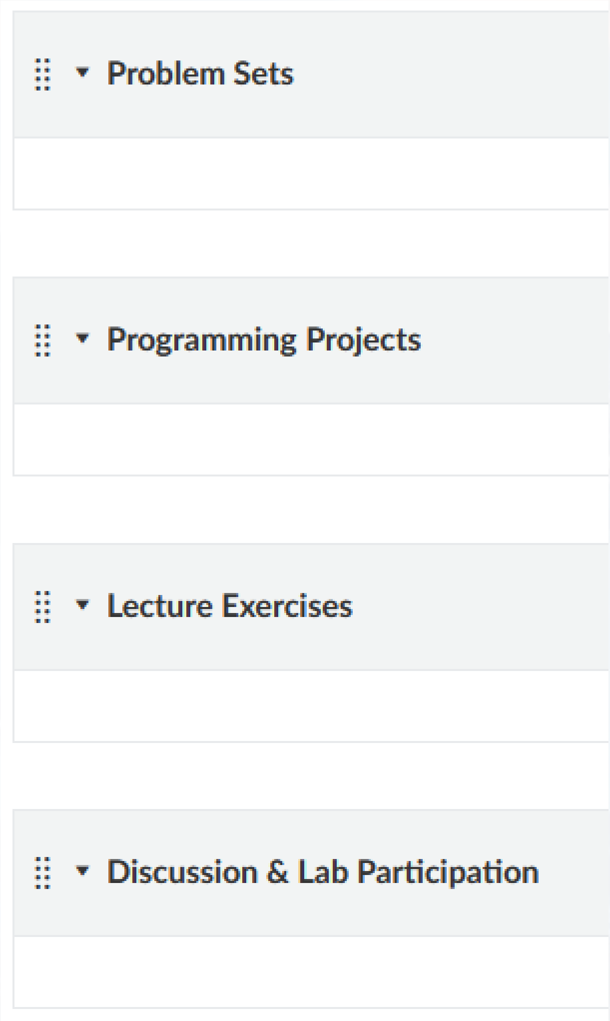
bcourses

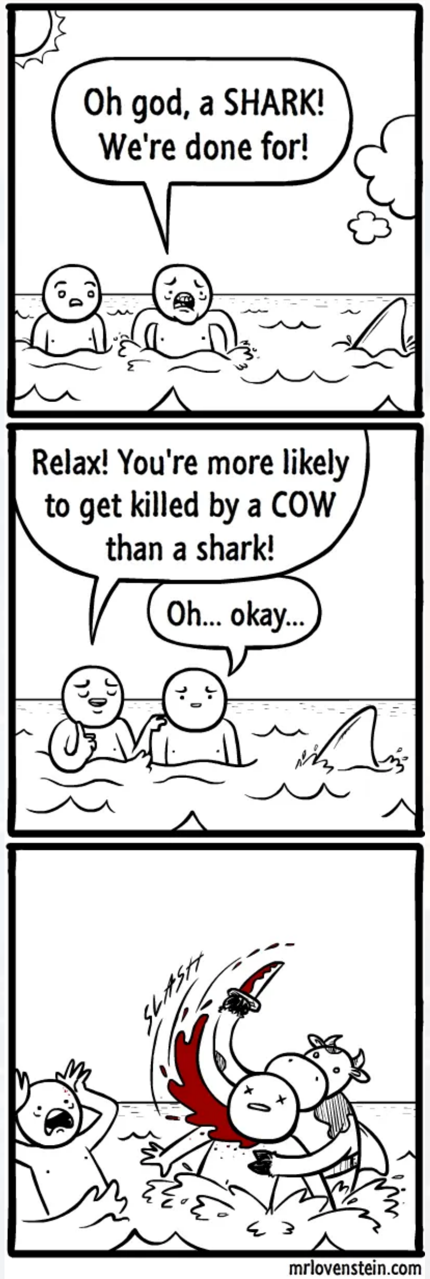
- dates & times
- **homework assignments and sample solutions**
- **videos/slides/notebooks/data/codes**
- links to discussions/labs/office hours



bcourses

- dates & times
- **homework assignments and sample solutions**
- videos/slides/notebooks/data/codes
- links to discussions/labs/office hours





Outline

- motivation for this course
- structure
- **syllabus**
- guide through the recordings

Introduction

syllabus

Lecture 1: Introduction to Data Science

data acquisition and analysis

Lecture 2: Data Sampling and Probability, Pandas

Lecture 3: Exploratory Data Analysis (EDA), and Regex Part I

Lecture 4: Exploratory Data Analysis (EDA), and Regex Part II

Lecture 5: Introduction to SQL

feature Selection and Analysis

Lecture 6: Feature Analysis, Engineering and Encoding

Lecture 7: PCA, LDA and Correlation, Dimension Reduction

Lecture 8: Advanced Visualization Tools (UMAP, T-SNE) in Python

project 1

Lecture 9: Feature Selection via Correlation Analysis

Spring Recess

modelling

Lecture 10: Introduction to linear and logistic regression; OLS

Lecture 11: Avoiding Overfitting and Regularization (L1 & L2)

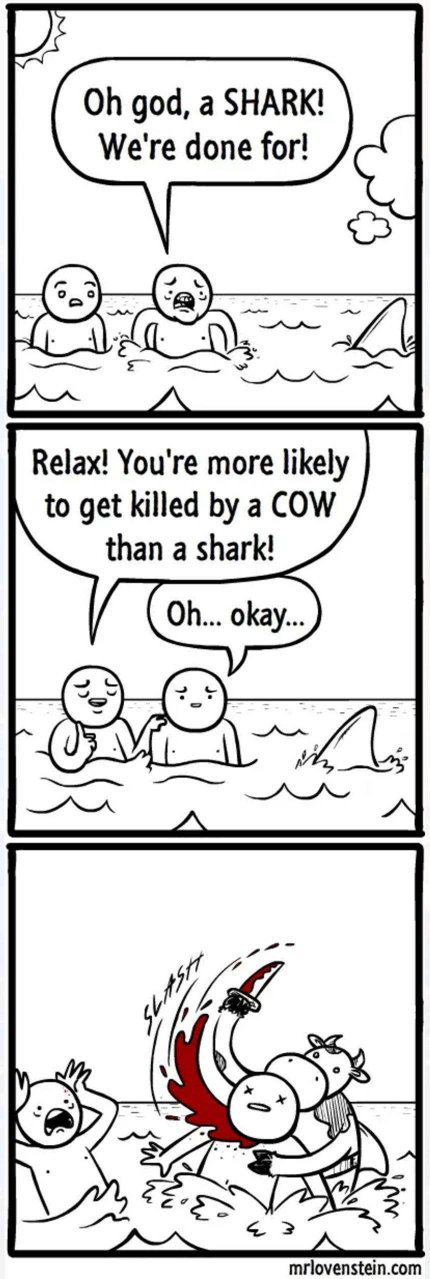
Lecture 12: Feature Selection via Regression Analysis

Lecture 13: Gradient Descent

Lecture 14: Clustering and Classification: Trees, KNN, K-Mean, GMM

project 2

Lecture 15: Feature Selection via ANNs ("leave-one-out", Accuracy Drop & Entropy)



Outline

- motivation for this course
- structure
- syllabus
- **guide through the recordings**

Introduction

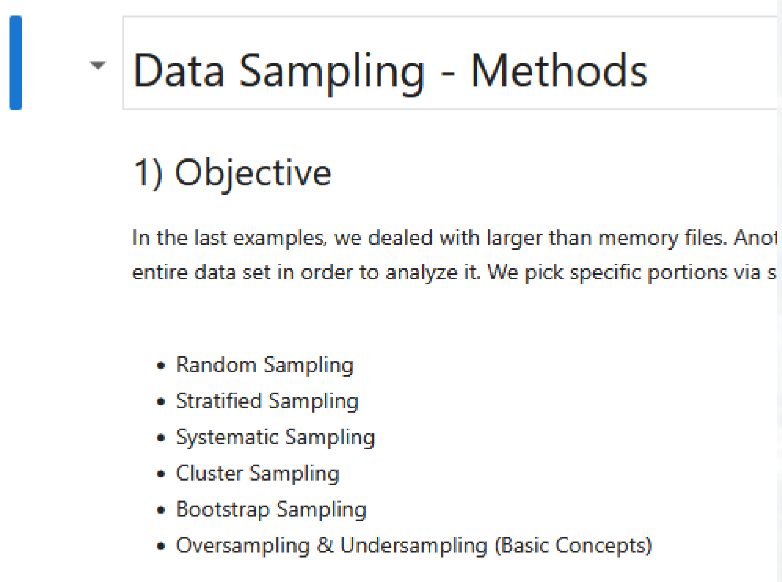
1st: watch the lecture recordings

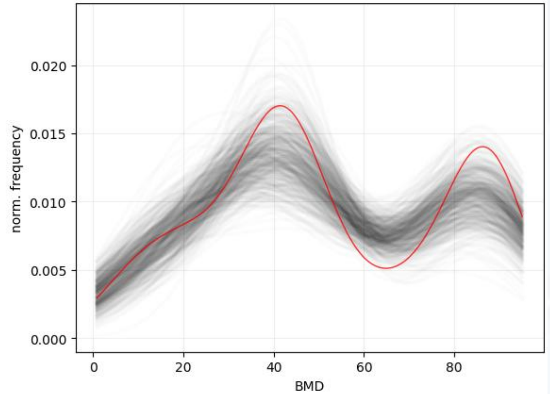
- Lect_5_1_Kmeans_Intro
- Lect_5_2_Kmeans_WalkThrough
- Lect_5_3_Kmeans_Summary
- Lect_5_4_GMM_Intro
- Lect_5_5_GMM_WalkThrough

lectures are ordered: *Lecture_Module_Order_Topic*

- try to understand as much as possible
- it is ok, if you don't understand everything!

2nd: explore the `jupyter notebooks and/or .py scripts`





- follow the instructions
- try to understand each step

Introduction

guide through the recordings

3rd: rewatch the lecture recordings if necessary

- Lect_5_1_Kmeans_Intro
- Lect_5_2_Kmeans_WalkThrough
- Lect_5_3_Kmeans_Summary
- Lect_5_4_GMM_Intro
- Lect_5_5_GMM_WalkThrough

lectures are ordered: *Lecture_Module_Order_Topic*

4th: write down questions for the lab and/or discussions

note: students have to actively attend the lab sessions and discussions (**asking/answering questions**) for grades!

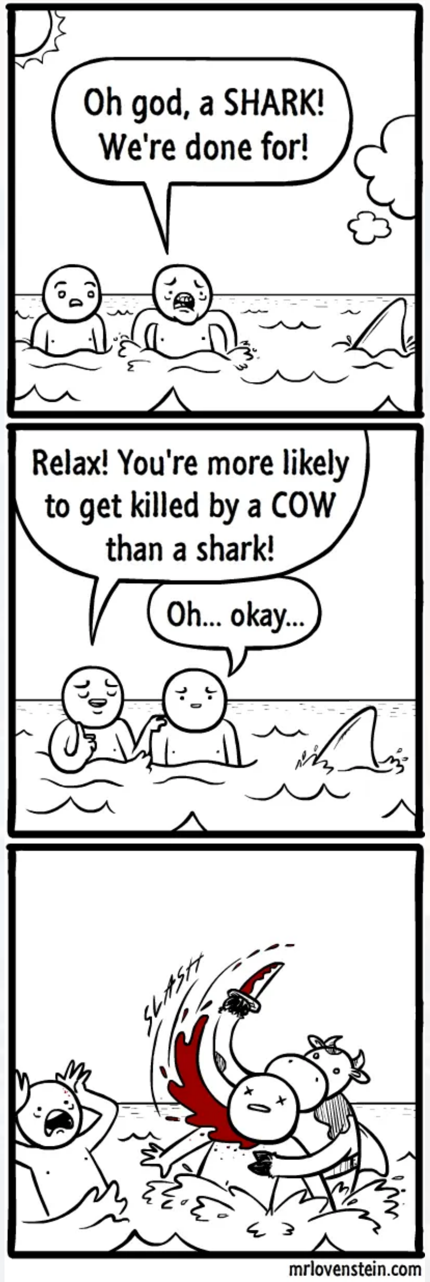
5th: solve the homework assignments

3) Task

Read the file *Data_Set.txt* to a standard pandas dataframe.

1. Write a function that is similar to "*def ReadWithAnyToolAnyMethod.py*"
   - .txt
   - .csv
   - .pkl
   - .parquet

   and monitor time and memory usage.
2. list the size of all these files
3. generate a table (as dataframe) that lists time, memory usage and size



Enjoy the Course ☺!