OPEN MACHINE LEARNING COURSE

LECTURE 0: INTRODUCTION

19TH, FEBRUARY

11 lectures • Each Wednesday at 7pm • Practical assignment • Competitions

https://www.meetup.com/Dubai-Data-Science-Meetup/

Organized by:

ods.ai

In partnership with:

HULT
INTERNATIONAL
BUSINESS SCHOOL

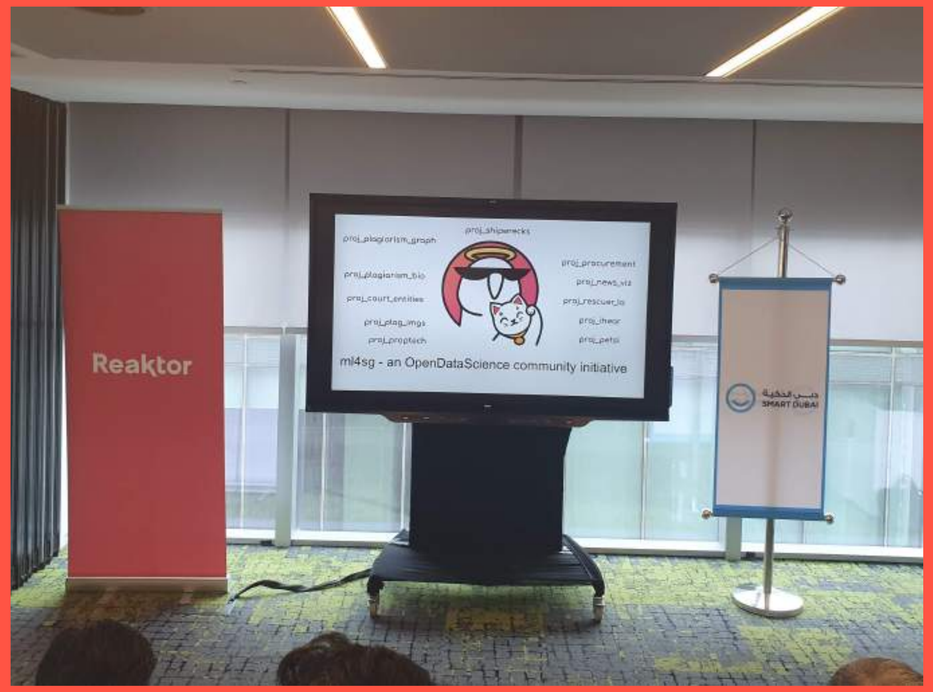
Lecturers from top companies:

R FICO EXOR DIUM VISA D.


Industry context
Conservative to any innovations (negotiable)
Won't tune anything if it works
Inexcusably high prices for errors and mistakes
(e.g. plugging the well with rapid change of pump parameters)

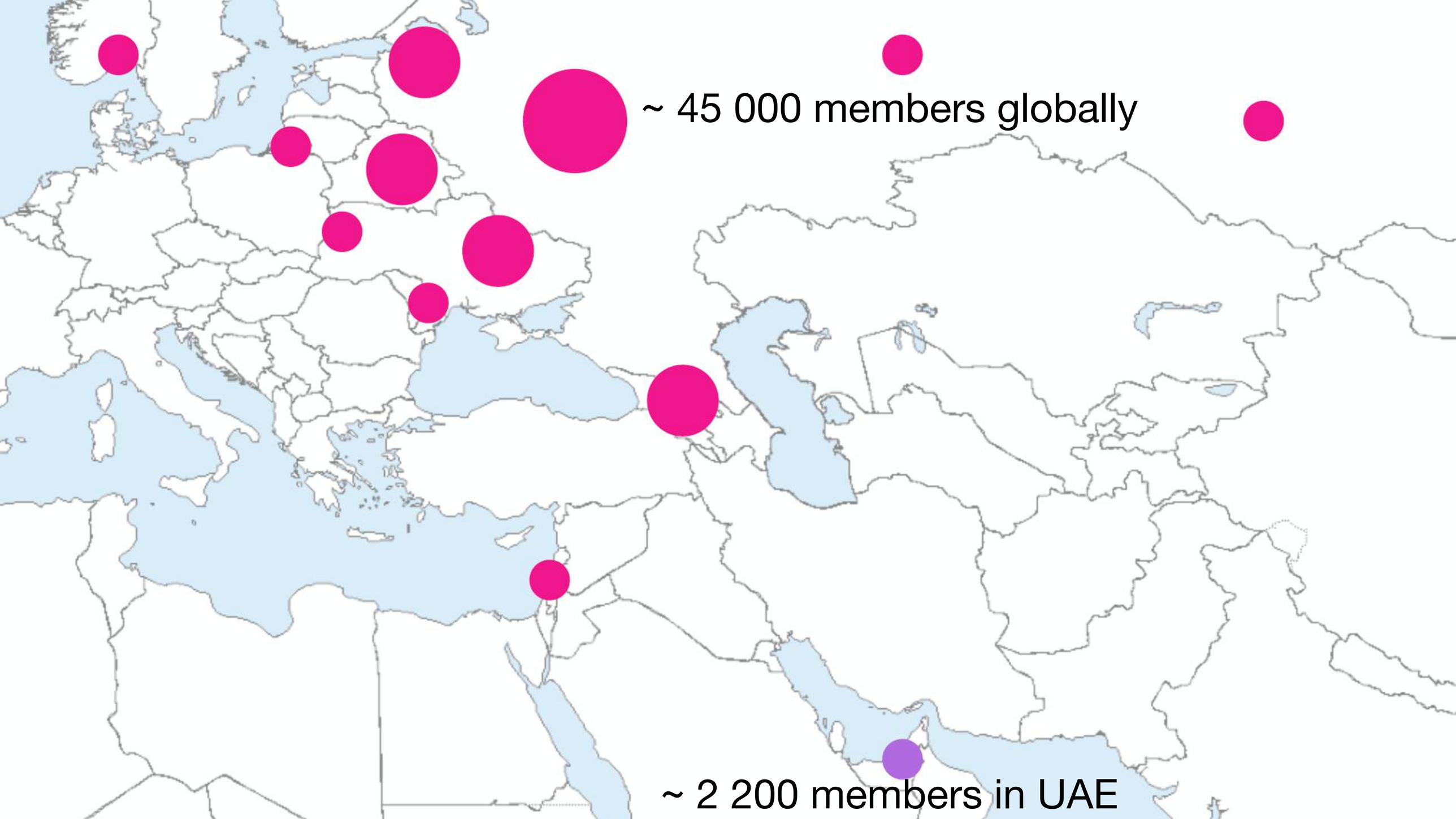

SONY

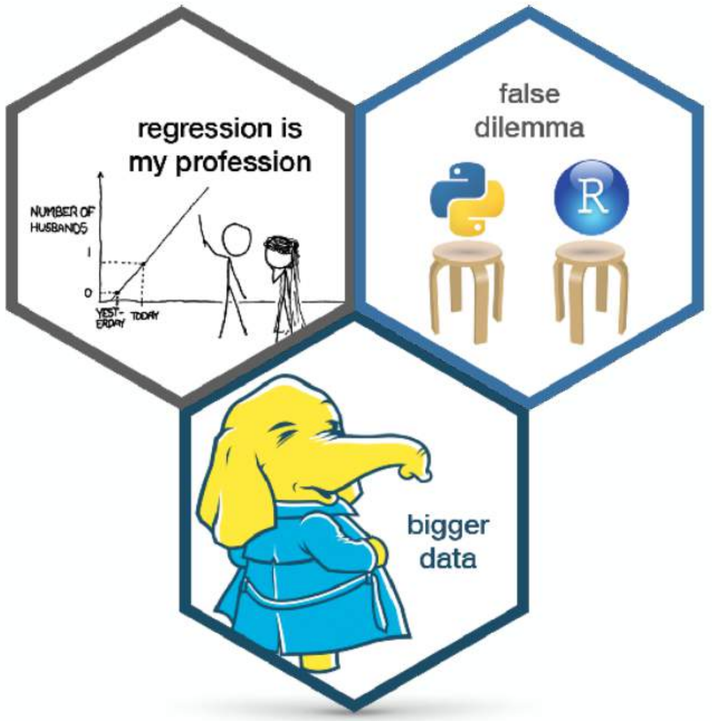

Impact
Data Science solutions are in infancy and not visible by business yet
Level 1: kickstart
Level 2 narrow focus
Level 3


Reaktor
proj_plagiarism_graph
proj_shipwrecks
proj_plagiarism_bio
proj_procurement
proj_court_entities
proj_news_viz
proj_plag_imgs
proj_rescuer_bi
proj_proptech
proj_theor
proj_petsi
ml4sg - an OpenDataScience community initiative
SMART DUBAI


Reaktor
SMART DUBAI

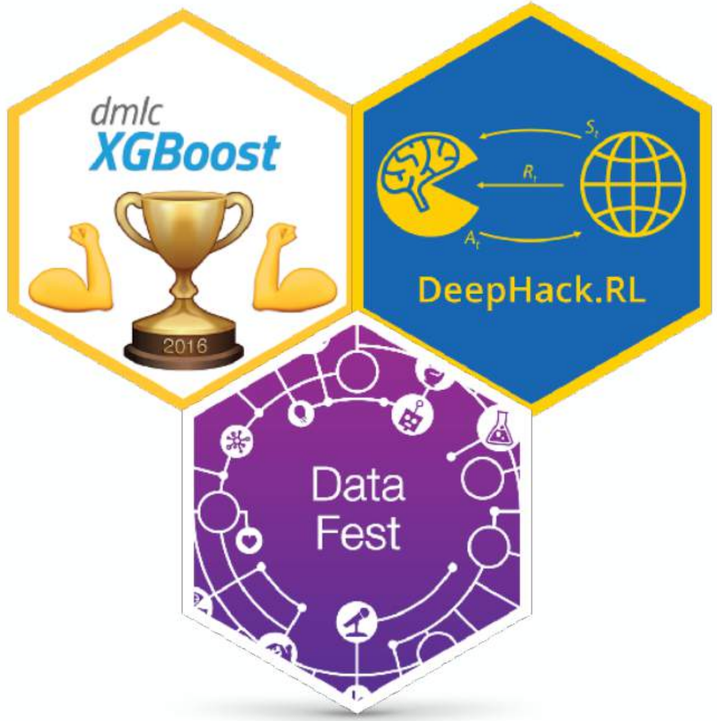

SMART DUBAI


~ 45 000 members globally
~ 2 200 members in UAE



DS\ML community development

Events to grow and facilitate

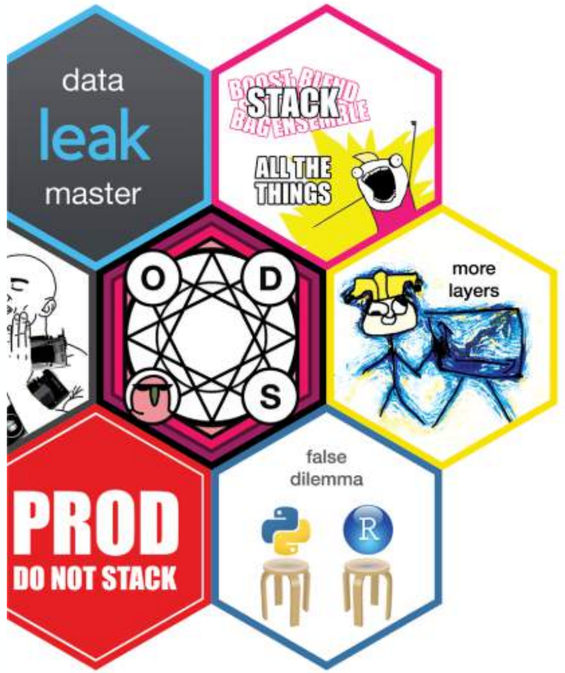


Research and discovery hub

Overview Data Kernels Discussion Leaderboard Rules

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	[ods.ai] STAMP			0.989642	109	8mo
2	▲6	[ods.ai] GPU_muscles_SPcu...			0.987976	124	8mo
3	▼2	FIIGO_SPcup_eligible			0.987857	93	8mo
4	▲5	Guanshuo Xu			0.987023	133	8mo
5	▲2	[ods.ai] 10011000			0.986547	77	8mo
6	▼3	[ods.ai] Evgeny Nizhibitsky			0.986190	31	8mo
7	▲4	blzr_SPcup_eligible			0.985595	65	8mo
8	▲4	Master			0.985595	17	8mo
9	▼5	[ods.ai] SVM punks			0.985357	228	8mo
10	▼4	Make Ensemble Great Again!			0.984761	173	8mo
11	▼1	[ods.ai] Nokia3310			0.984404	100	8mo
12	▲2	Yusaku | Branden | KazAnova			0.982738	117	8mo

ods.ai



Open Data Science

Global community that unites all researchers, engineers and developers around Data Science and related areas.

- Create awesome projects, events and educational courses
- Share experience, developing each other's skills
- Promote open Data Science and push the field forward

JOIN #SLACK

https://ods.ai/en/

THE TEAM

- Pavel Nesterov, Principal Data Scientist @ Reaktor
  - http://linkedin.com/in/nesterovpavel
- Yuri Chekalin, Data Scientist & Pre-Sales Solution Architect @ FICO
  - https://www.linkedin.com/in/ychekalin/
- Dmitriy Dovgan, Ph. D., Data Science Manager @ VISA
  - https://www.linkedin.com/in/dmitriy-dovgan-15b22585/
- Shams Shapsough, CEO @ Exordium
  - https://www.linkedin.com/in/sh-shapsough/
- Dmitry Denisov, Data Scientist @ Deloitte
  - https://www.linkedin.com/in/dmitry-denisov-022102103/
- Pavel Golubev, Principal Data Scientist @ Reaktor, CEO & Founder @ MaritimeAI
  - https://www.linkedin.com/in/pavel-golubev/

https://mlcourse.ai/

PRIORITIZE

Yury Kashnitsky

Ph.D., Data Scientist at KPN

Head of mlcourse.ai

https://www.linkedin.com/in/kashnitskiy/

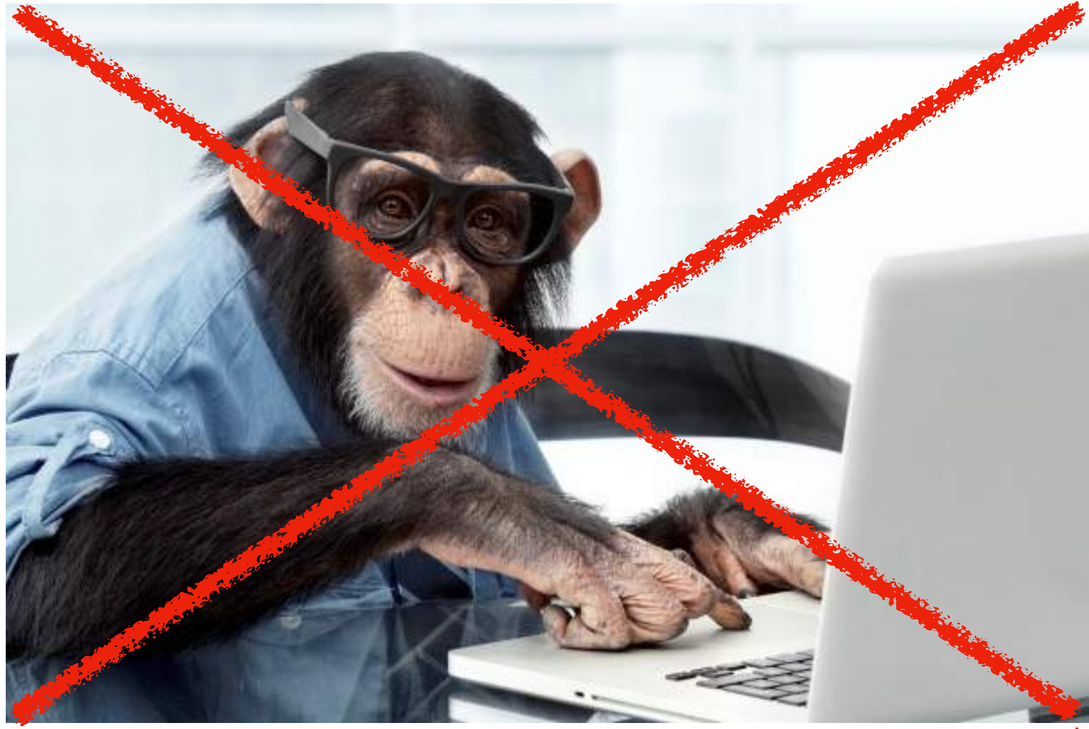
SYLLABUS

- Each Wednesday at 7pm in Hult
- 10 more lectures
- Basic ML algorithms and their applications
- Assignments and in-class practice
- Competitions
- Individual projects
- Tutorials

https://github.com/DmitriiDenisov/mlcourse_dubai

WHAT MAKES IT DIFFERENT

- Lots and lots of practice
- Theoretical understanding of applied techniques
- Delving into competitions
- Your own projects
- Really vibrant community!

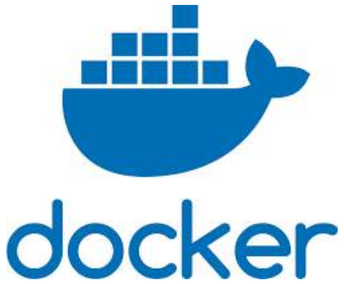


ROADMAP AND LOGISTICS

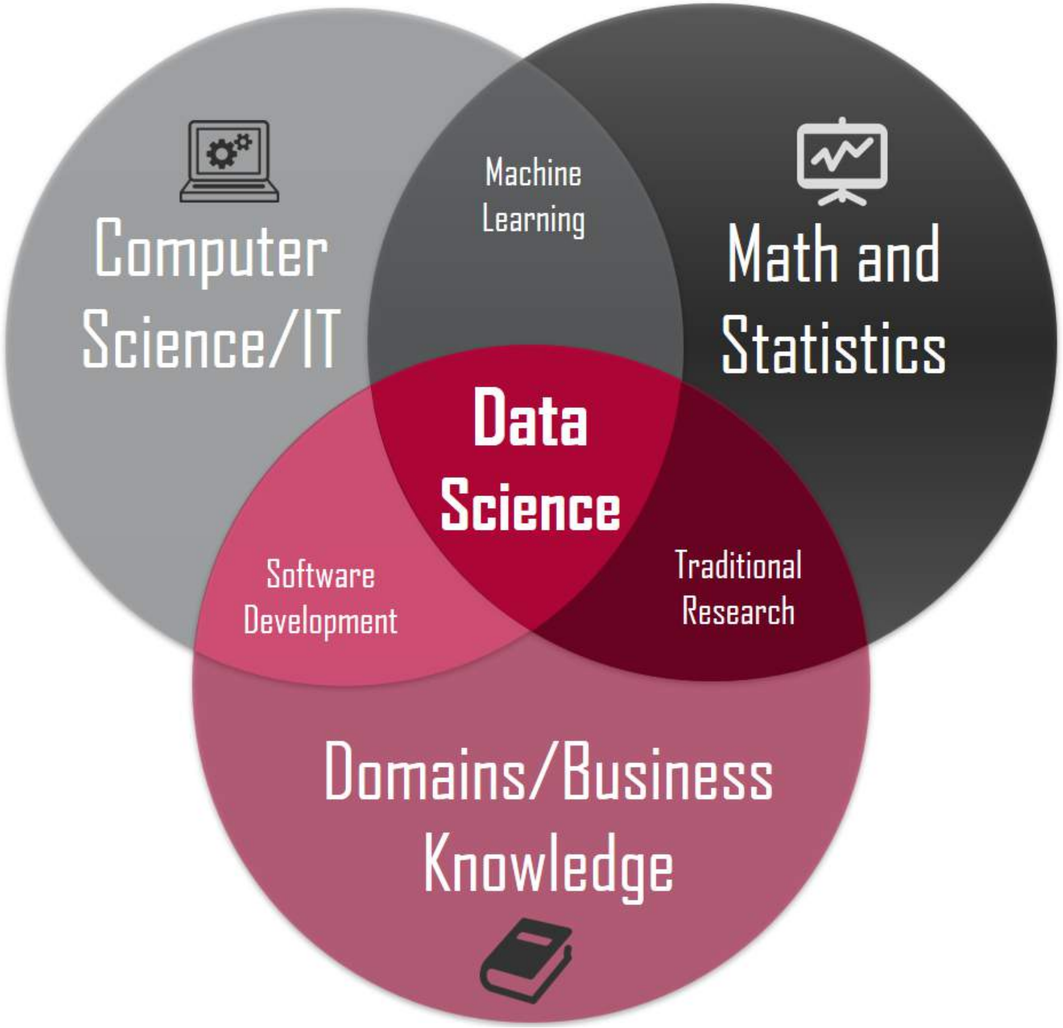
- All communication in ODS Slack, **#mlcourse_dubai**
- Roadmap: https://github.com/DmitriiDenisov/mlcourse_dubai
- 10 assignments – ~10 credits each
- Projects, competitions, tutorials – up to 40 crd. each
- Current rating: https://docs.google.com/spreadsheets/d/1pbfhIyDURLiLWVHZGWAPC_g9l7DHSmQb2v7mrvKzmPc/edit?usp=sharing
- All materials are stored on GitHub
  - https://github.com/DmitriiDenisov/mlcourse_dubai
- mlcourse.ai and https://mlcourse.ai
- Top-100 participants will be mentioned on a special Wiki page

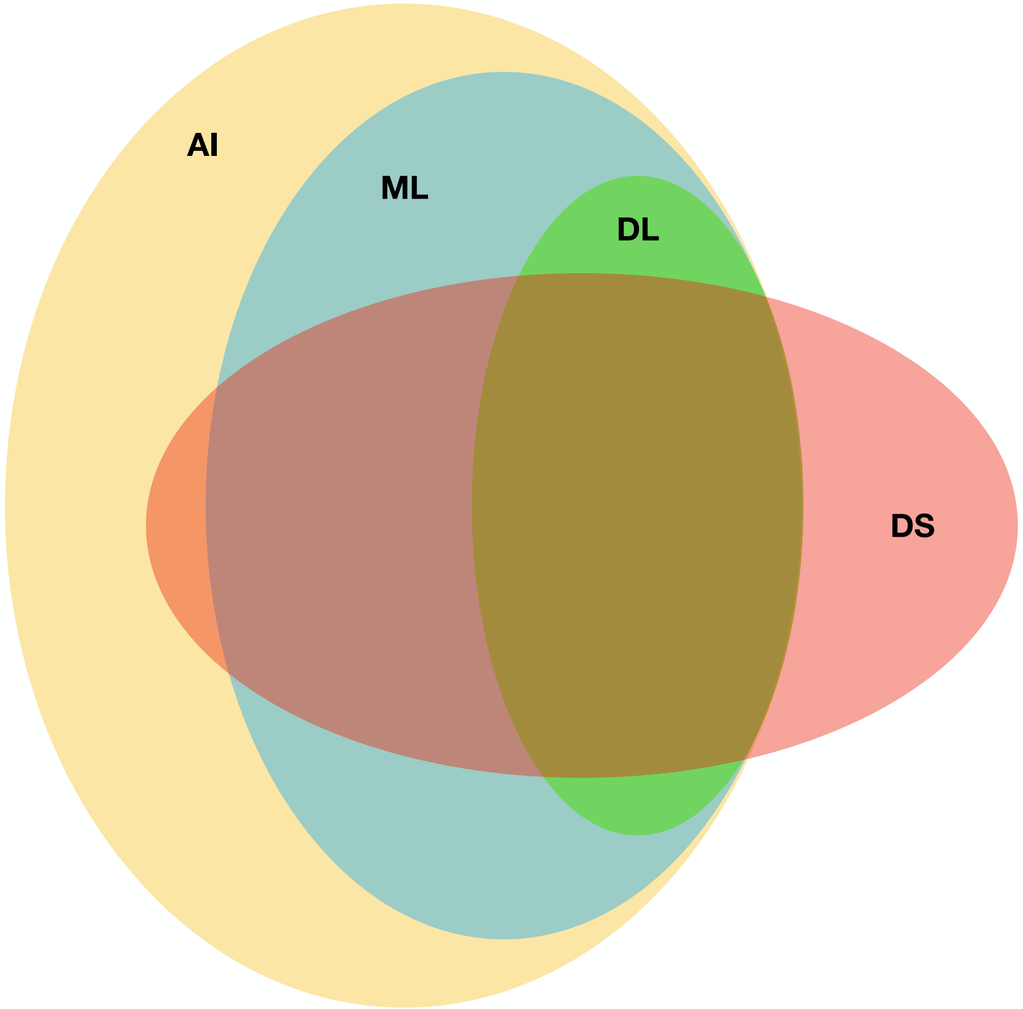
TOOLBOX

- Python
- Jupyter notebooks
- GitHub
- Docker (optional)
- Other libs like Vowpal Wabbit & Xgboost
- Instructions https://mlcourse.ai/prerequisites



DS/AI/ML/DL


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge


AI
ML
DL
DS

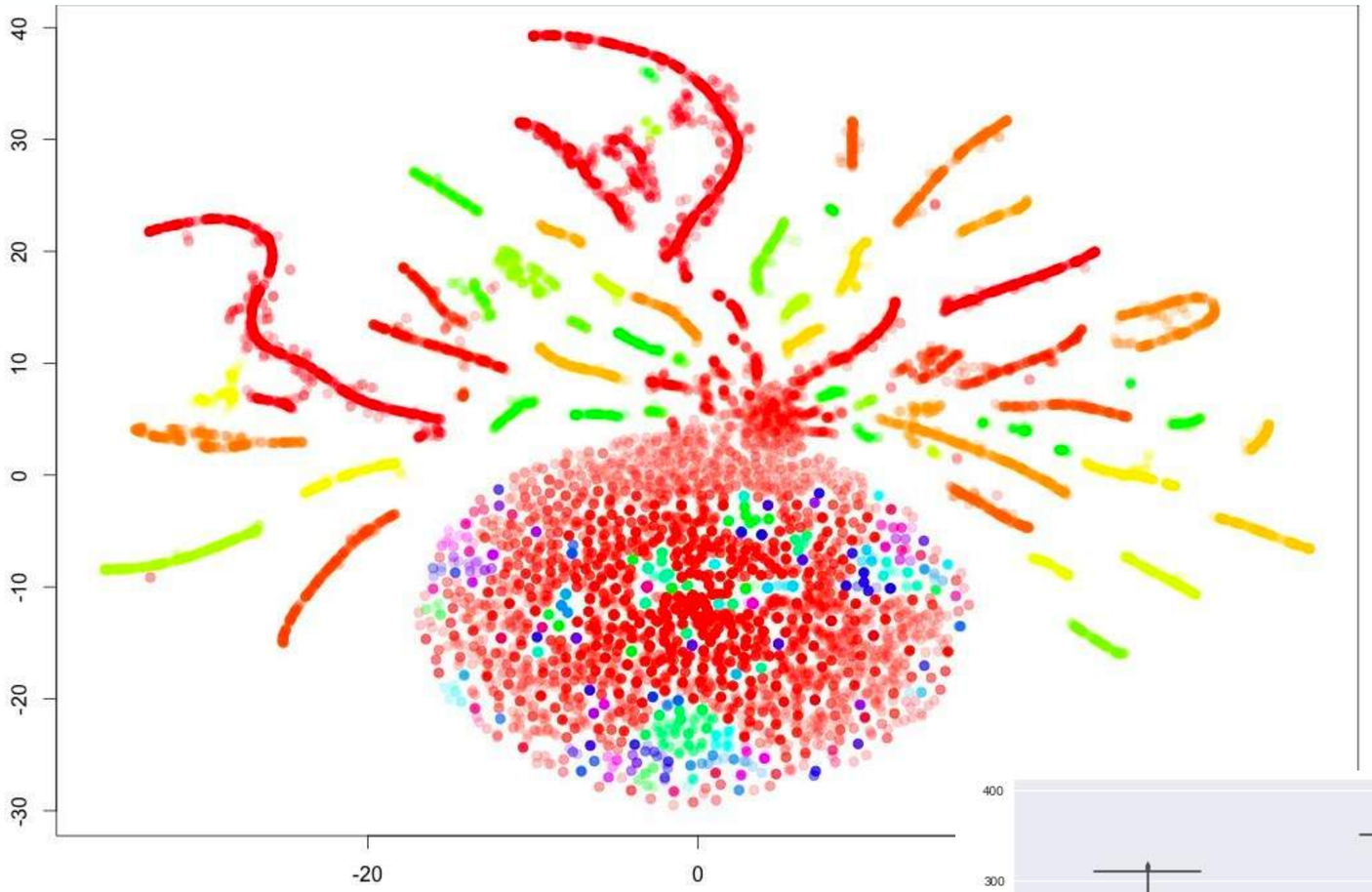
LECTURE 1

```
df.head(4)
```

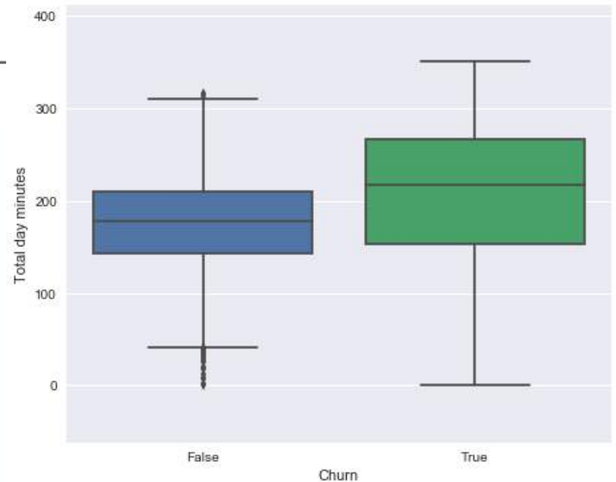
	wage	exper	union	goodhlth	black	female	married
0	5.73	30	0	1	0	1	1
1	4.28	28	0	1	0	1	1
2	7.96	35	0	1	0	1	0
3	11.57	38	0	1	0	0	1

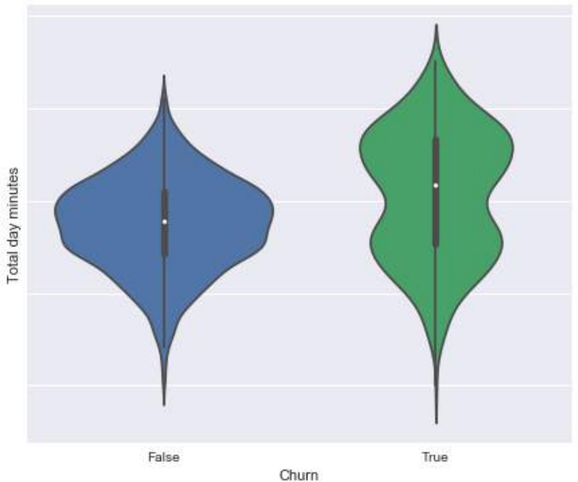
- Data analysis with Pandas
- Practice on first steps after getting data
- EDA

LECTURE 2

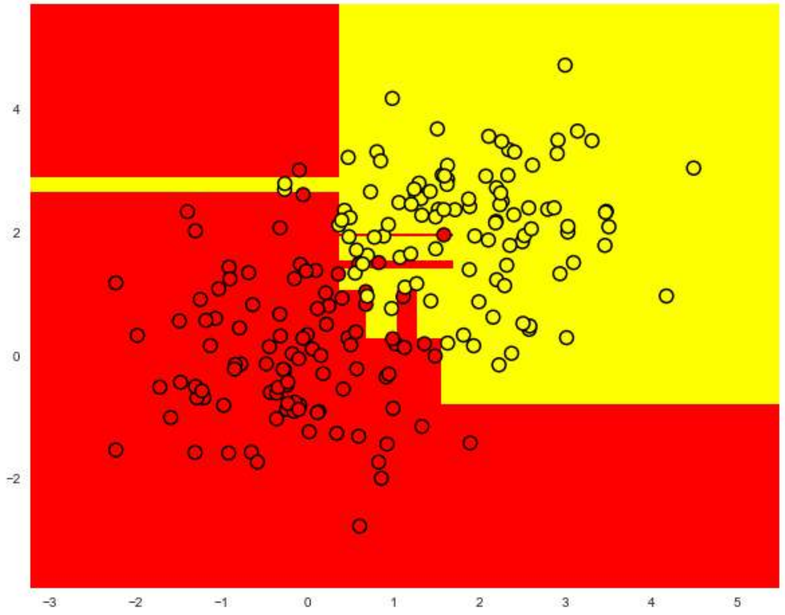


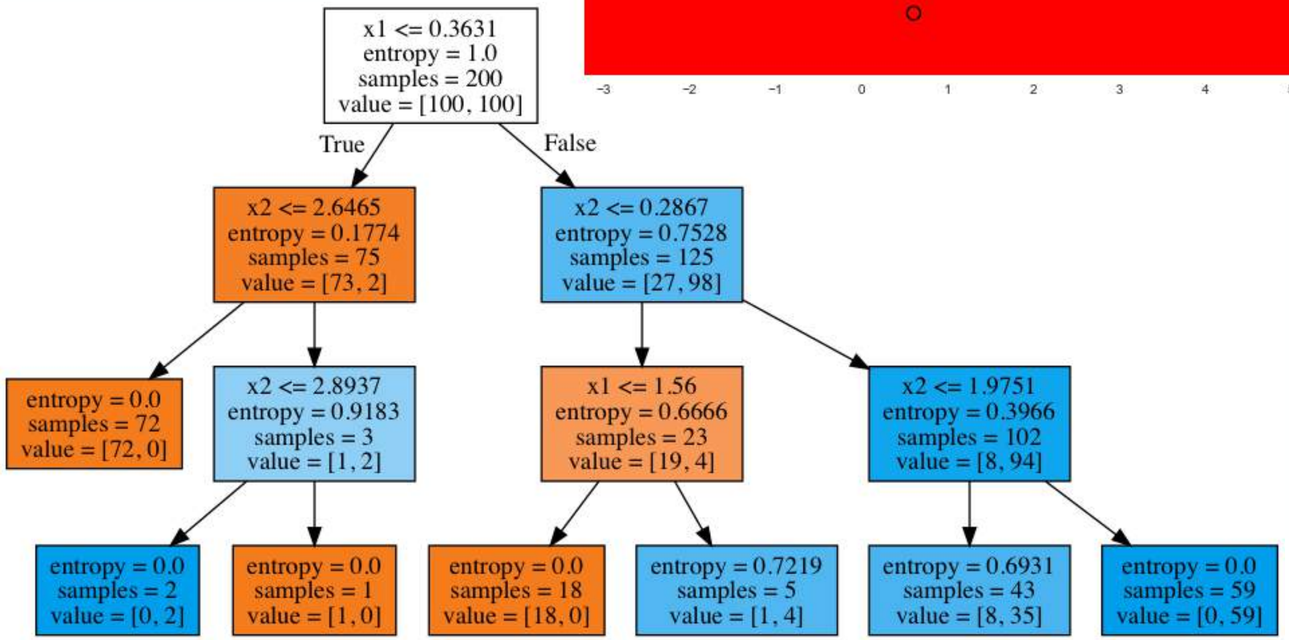
- Visual data analysis with
- Pandas, Matplotlib and Seaborn
- Crucial plots for feature exploration
- Practice on «drawing»





LECTURE 3

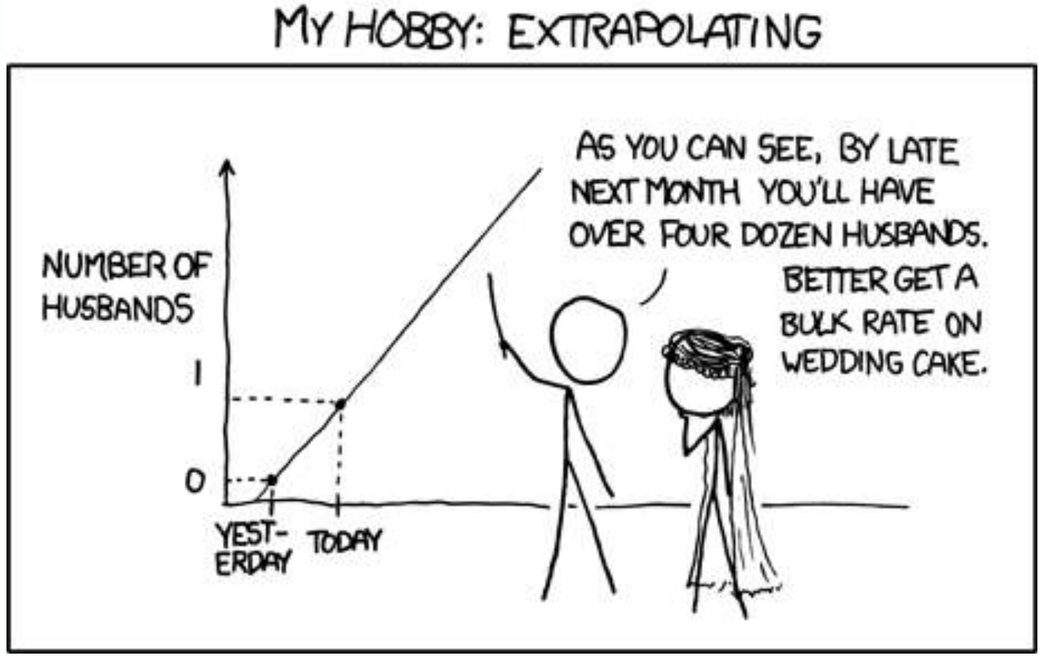


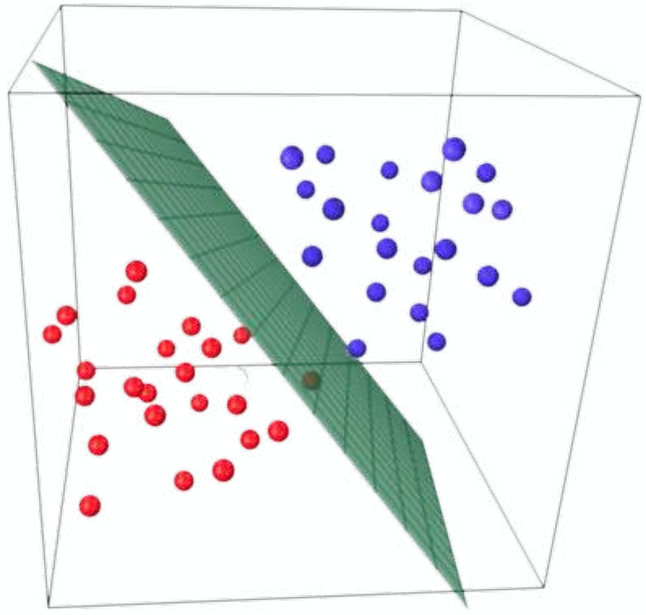


- Foundations of Machine Learning
- Supervised learning
- Decision trees
- k Nearest Neighbours
- Practice: first steps with
- Scikit-learn

LECTURE 4

- Linear classification models
- Regularization
- Cross-validation
- Practice on logistic regression for a "real-world" task

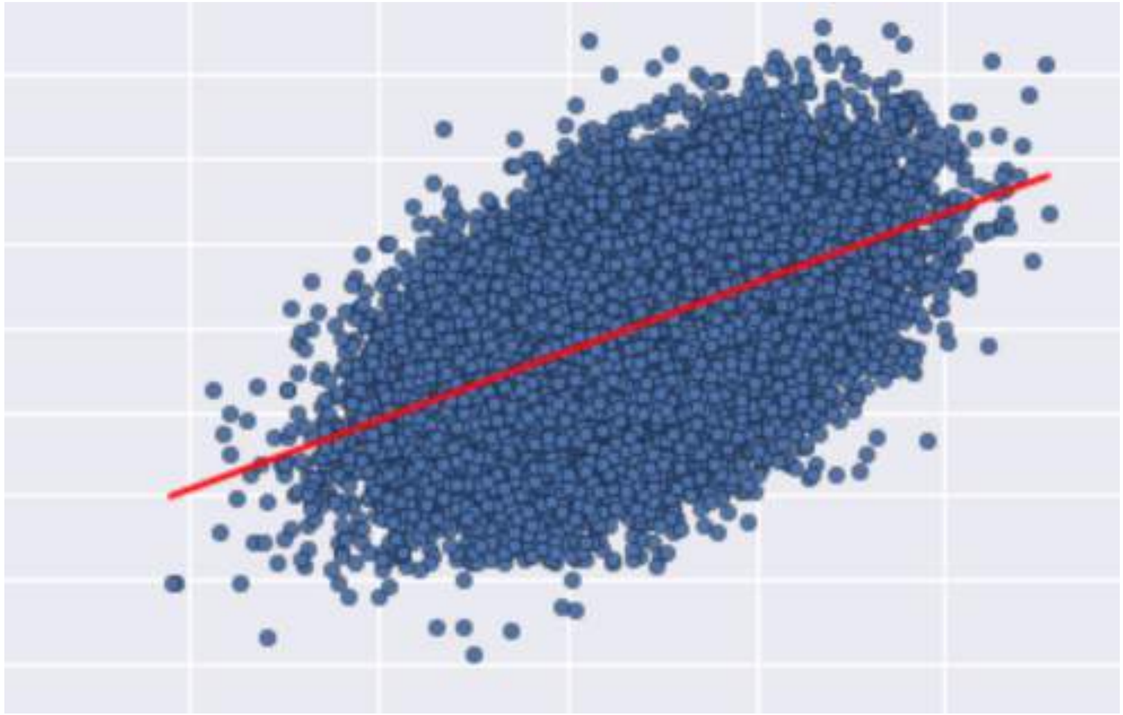




LECTURE 5

- Ensembles, random forest
- Feature importance
- Practice on random forest and assessing feature importance

LECTURE 6



The Gauss-Markov theorem ensures

$$\mathsf{Var}[\widehat{\beta} \mid X] \preceq \mathsf{Var}[\widetilde{\beta} \mid X],$$

where $\widehat{\beta}$ is the OLS estimator and $\widetilde{\beta}$ is any other unbiased linear estimator of β^*.

(a) Does the theorem ensure $\mathsf{var}[\widehat{\beta}_j \mid X] \leq \mathsf{var}[\widetilde{\beta}_j \mid X]$ for any $j \in [d]$?

(b) Show that the theorem ensures $\mathsf{Var}[\widehat{\beta}] \preceq \mathsf{Var}[\widetilde{\beta}]$.
Hint: Recall the decomposition of variance:

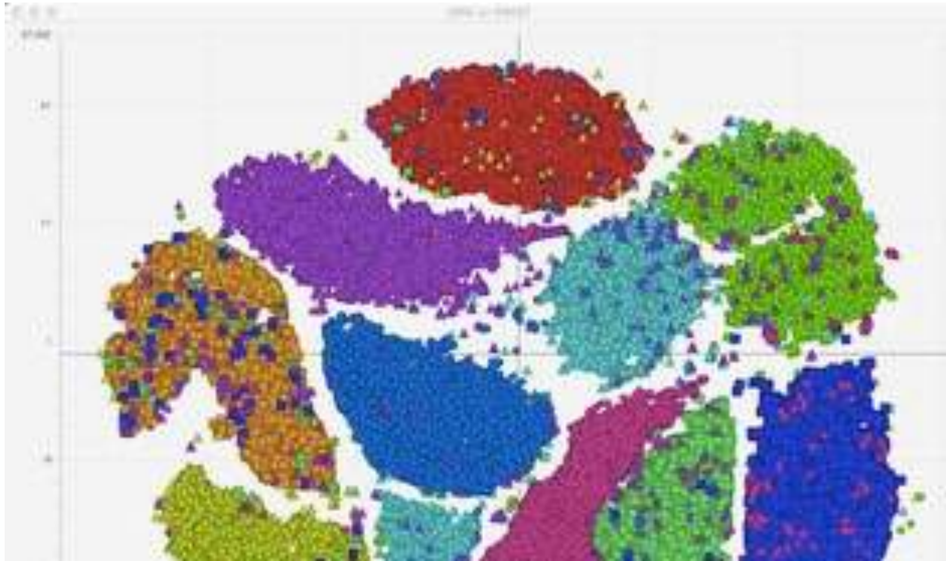
$$\mathsf{var}[\mathrm{x}] = \mathbb{E}[\mathsf{var}[\mathrm{x} \mid \mathrm{y}]] + \mathsf{var}[\mathbb{E}[\mathrm{x} \mid \mathrm{y}]].$$

(c) Is there a linear, but not necessarily unbiased estimator of β^* that has smaller variance than the OLS estimator? If yes, give an example of such an estimator. Otherwise, explain why there is no such estimator.

- Regression task
- Linear and non-linear regression models
- Practice on grasping core ideas behind linear regression

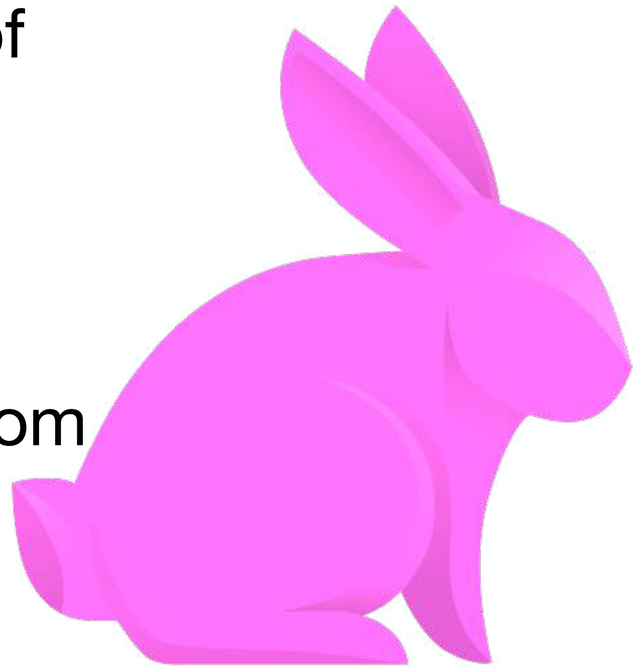
LECTURE 7

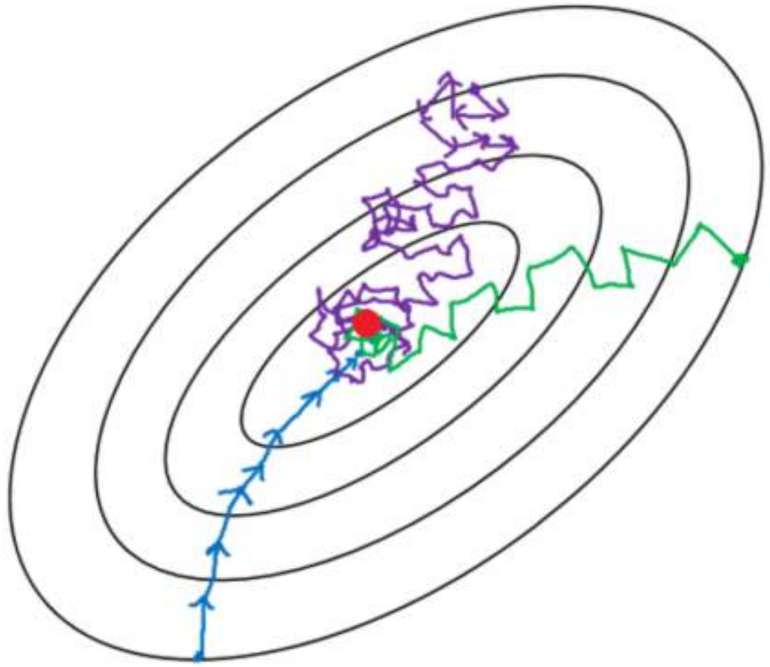
- Unsupervised Learning
- Principal Component Analysis
- Clustering
- Practice: clustering Samsung
- Galaxy S3 sensor data into types of human activity



LECTURE 8

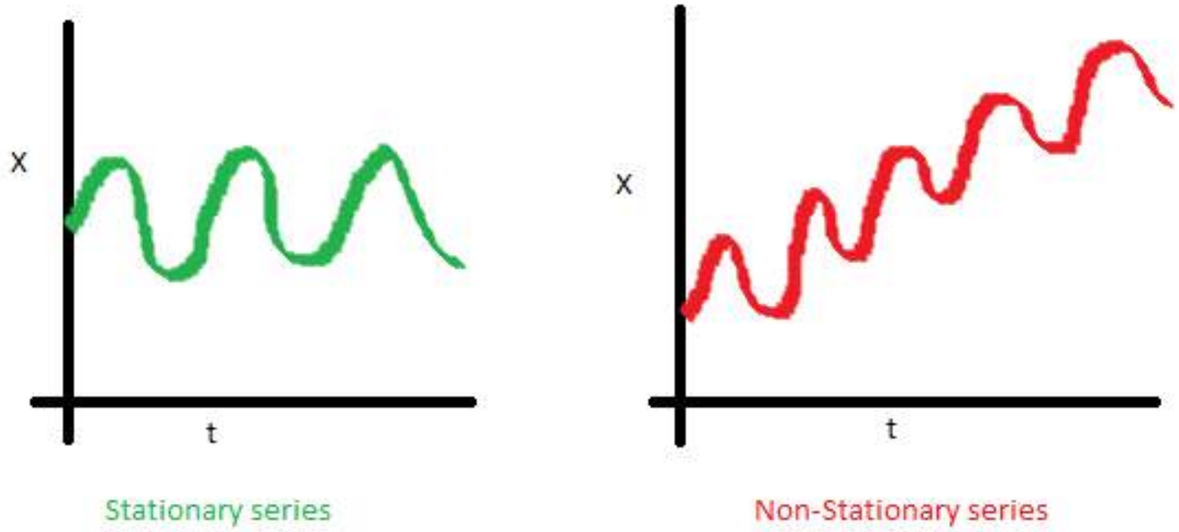
- Stochastic Gradient Descent & Online learning
- Learning with a couple GB of data
- Vowpal Wabbit
- Extracting simple features from texts
- Practice: text classification

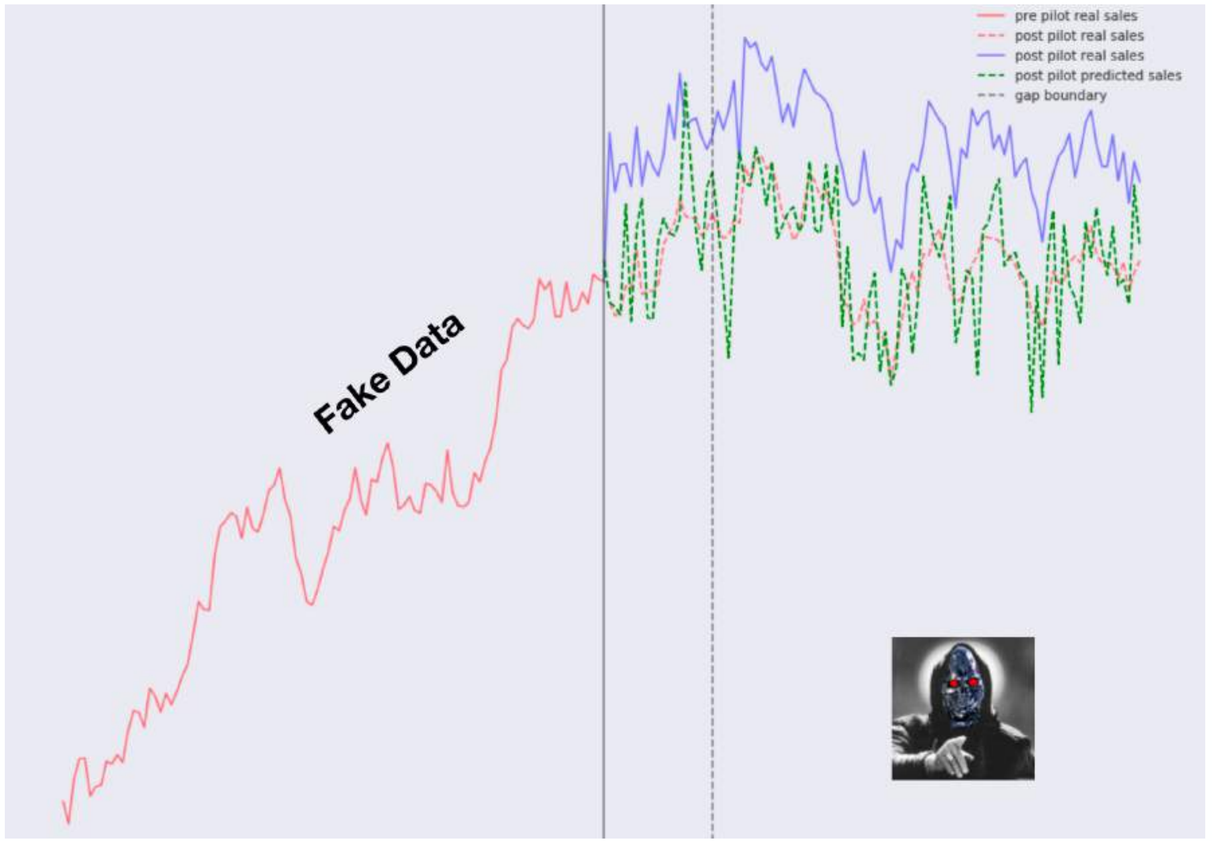




VOWPAL WABBIT

LECTURE 9





- Time series
- Classical and modern approaches
- Practice: ARIMA model, Facebook Prophet

LECTURE 10

- Gradient boosting: a modern view
- Theoretical basis for gradient boosting
- Best implementations
- Practice: beating a baseline in a Kaggle Inclass competition

HOMEWORKS

- Assignments
- Kaggle competitions
- Individual projects / Tutorials (peer review)

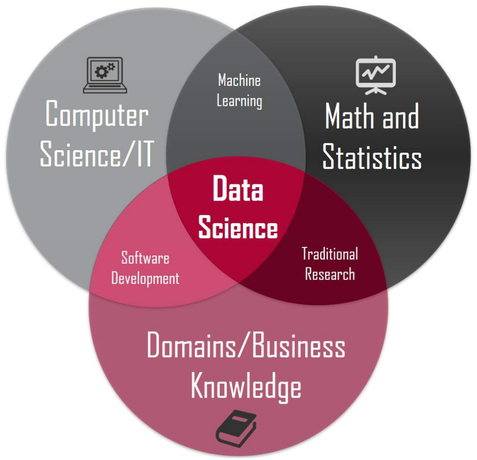
Project “Alice" Tutorials

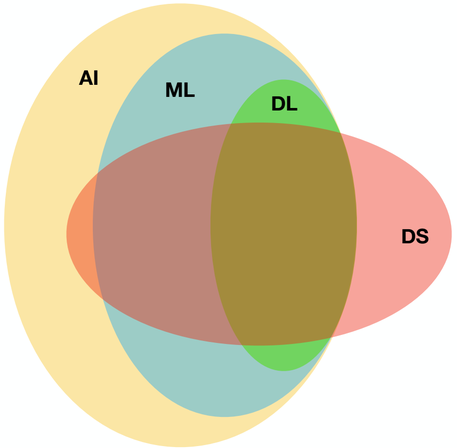
- A substitute for an individual project if you don't have cool ideas for one
- Clear instructions
- 6 weeks, 6 notebooks to complete
- Solutions are not shared

BONUS

Three parts of the course cover all aspects of AI

- Machine Learning Course (now)
- Deep Learning Course (Summer)
- Business Side of Data Science (Autumn)





PREPARATION

- http://leetcode.com/problems
- https://www.youtube.com/watch?v=_uQrJ0TkZlc (Python)
- http://cs231n.github.io/python-numpy-tutorial/
- https://mml-book.github.io/
- http://www.deeplearningbook.org/
- https://ml-cheatsheet.readthedocs.io/en/latest/
- https://www.kaggle.com/c/titanic
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques

STAY TUNED

More info in Slack **#mlcourse_dubai**

(pinned items)

https://github.com/DmitriiDenisov/mlcourse_dubai

https://forms.gle/XTvhyNhuevV1QV3F8