CSE217
INTRODUCTION TO DATA SCIENCE

LECTURE 1: DS & ML

Spring 2019

Marion Neumann

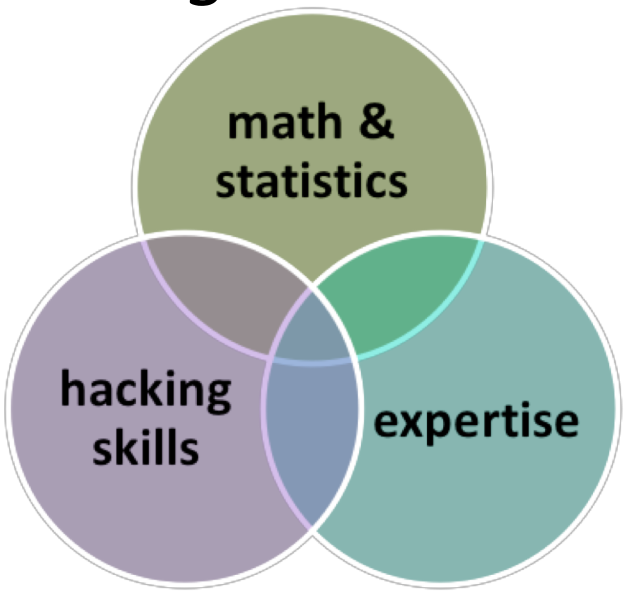
Washington University in St. Louis

WHAT IS DATA SCIENCE?

…solving problems with data…

…sounds cool!

What makes a good data scientist?

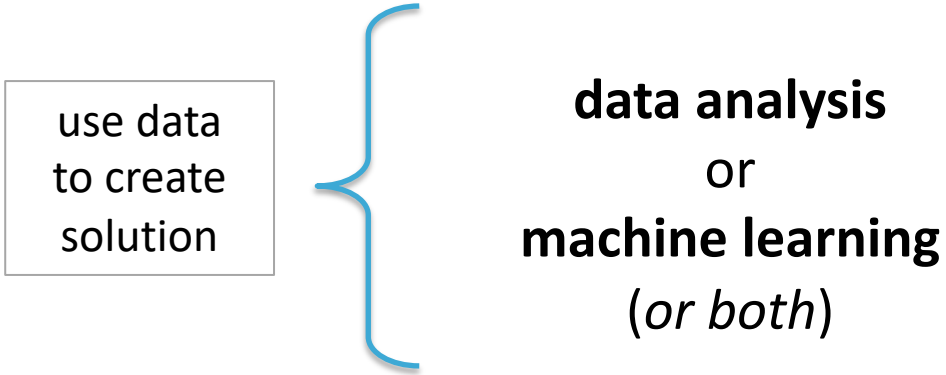


[PDSH] p xi
[DSFS] p 1-2

WHAT IS DATA SCIENCE?

…solving problems with data…

…which step is most challenging?

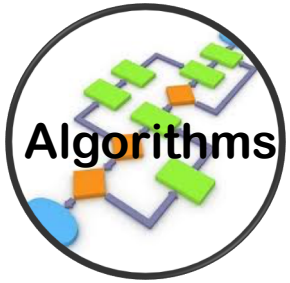


WHAT IS DATA ANALYSIS?

…using data to discover useful information…

- **data**: anything you can *measure* or *record*

- **statistics**: summarize (and visualize) *main characteristics* of the data

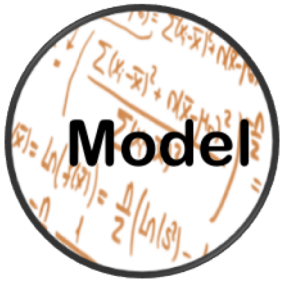


- **algorithms**: apply algorithms to find *patterns* in the data

WHAT IS MACHINE LEARNING?

…creating and using models that learn from data…

- **data**: anything you can *measure* or *record*



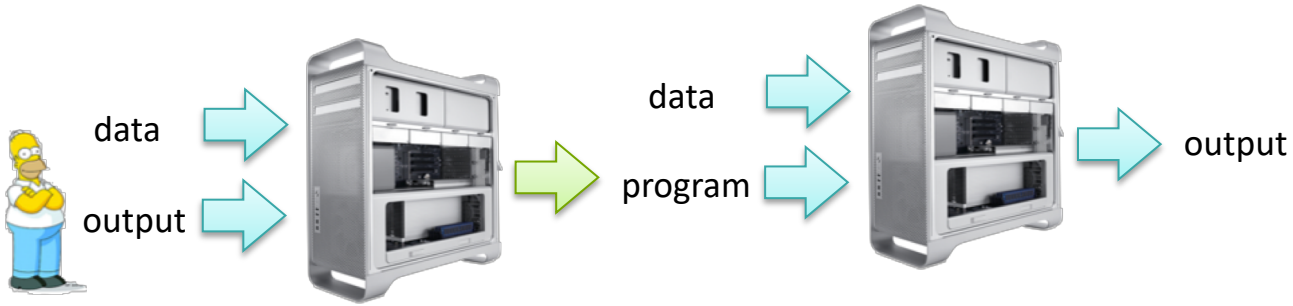
- **model**: specification of a (mathematical) *relationship* between different variables

- **evaluation**: how well does the model *work*?

WHAT IS MACHINE LEARNING?

- Traditional CS

- Machine Learning

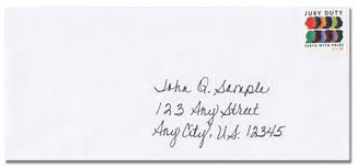


WHAT IS MACHINE LEARNING?

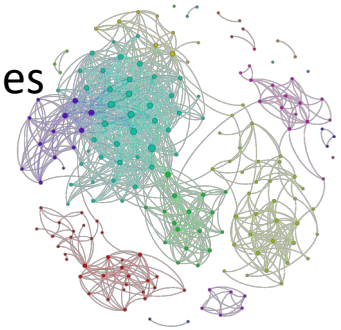
…creating and using models that learn from data…

Examples

Identifying zip code from handwritten digits



Detecting communities in social networks

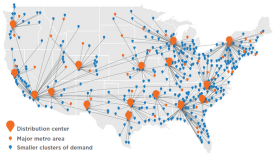


Predicting the traffic volume at rush hour

2.000 bis 50.000
1.500 bis 2.000
1.000 bis 1.250
800 bis 1.000
600 bis 800
400 bis 600
200 bis 400
0 bis 200

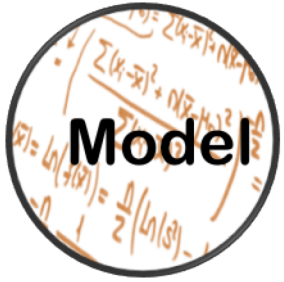
Detecting fraudulent credit card transactions

Determining the location of distribution centers based on customers' residence

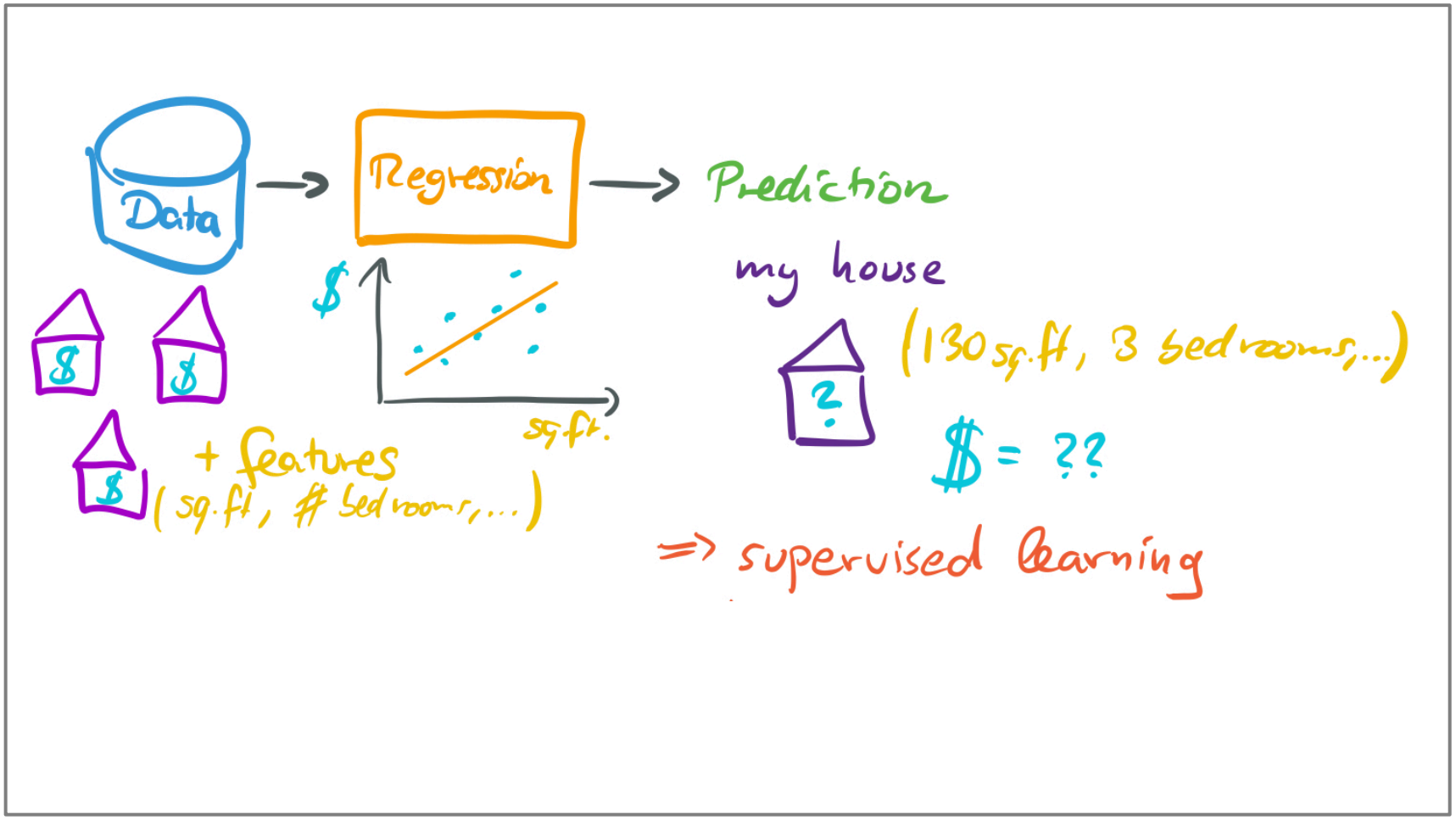


[DSFS] p 3-13

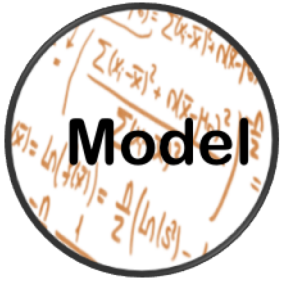
LEARNING FROM DATA



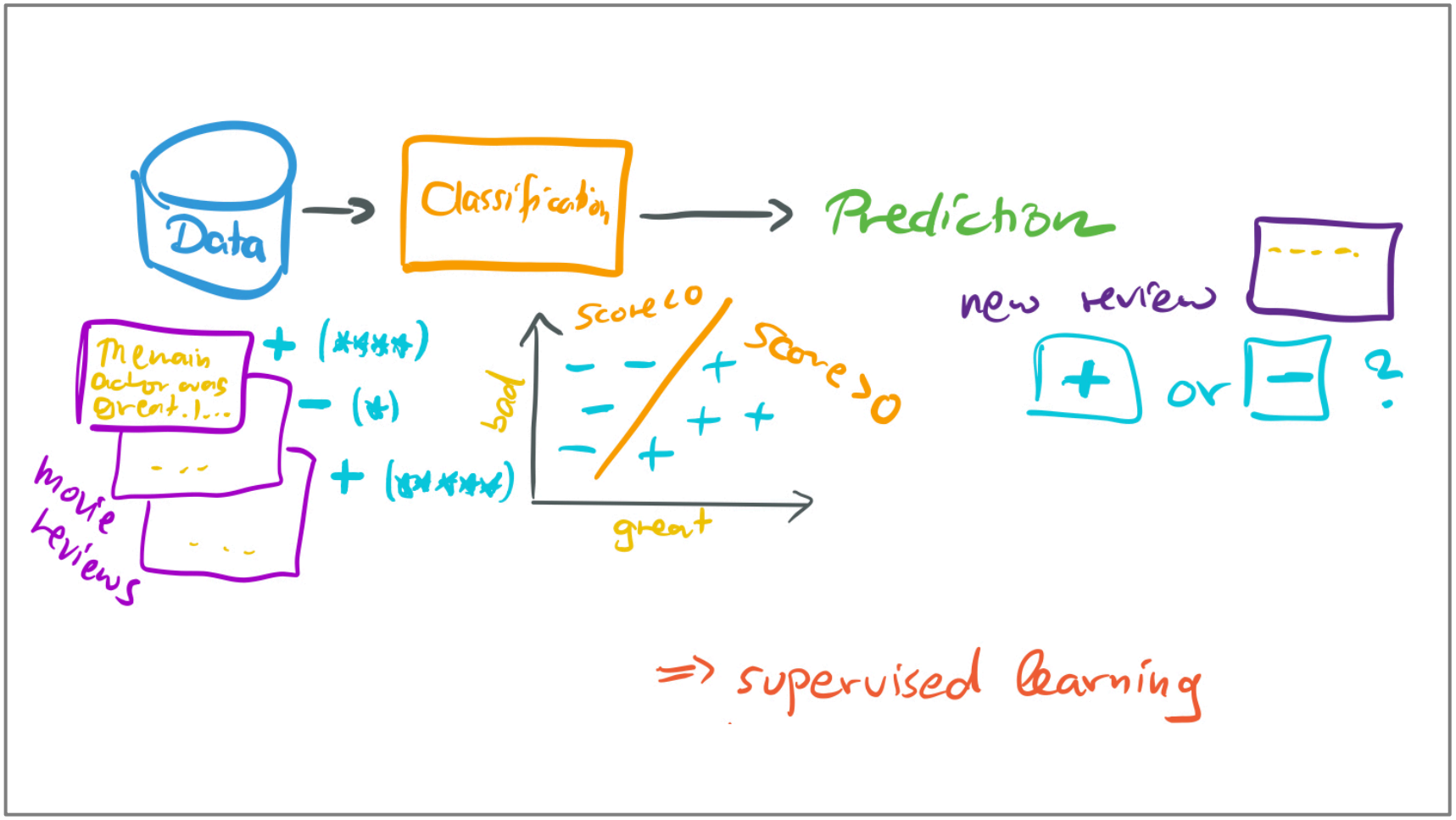
- Regression



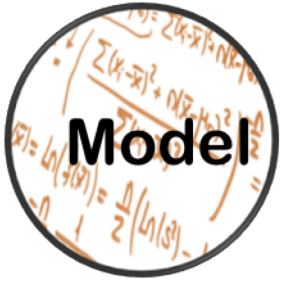
LEARNING FROM DATA



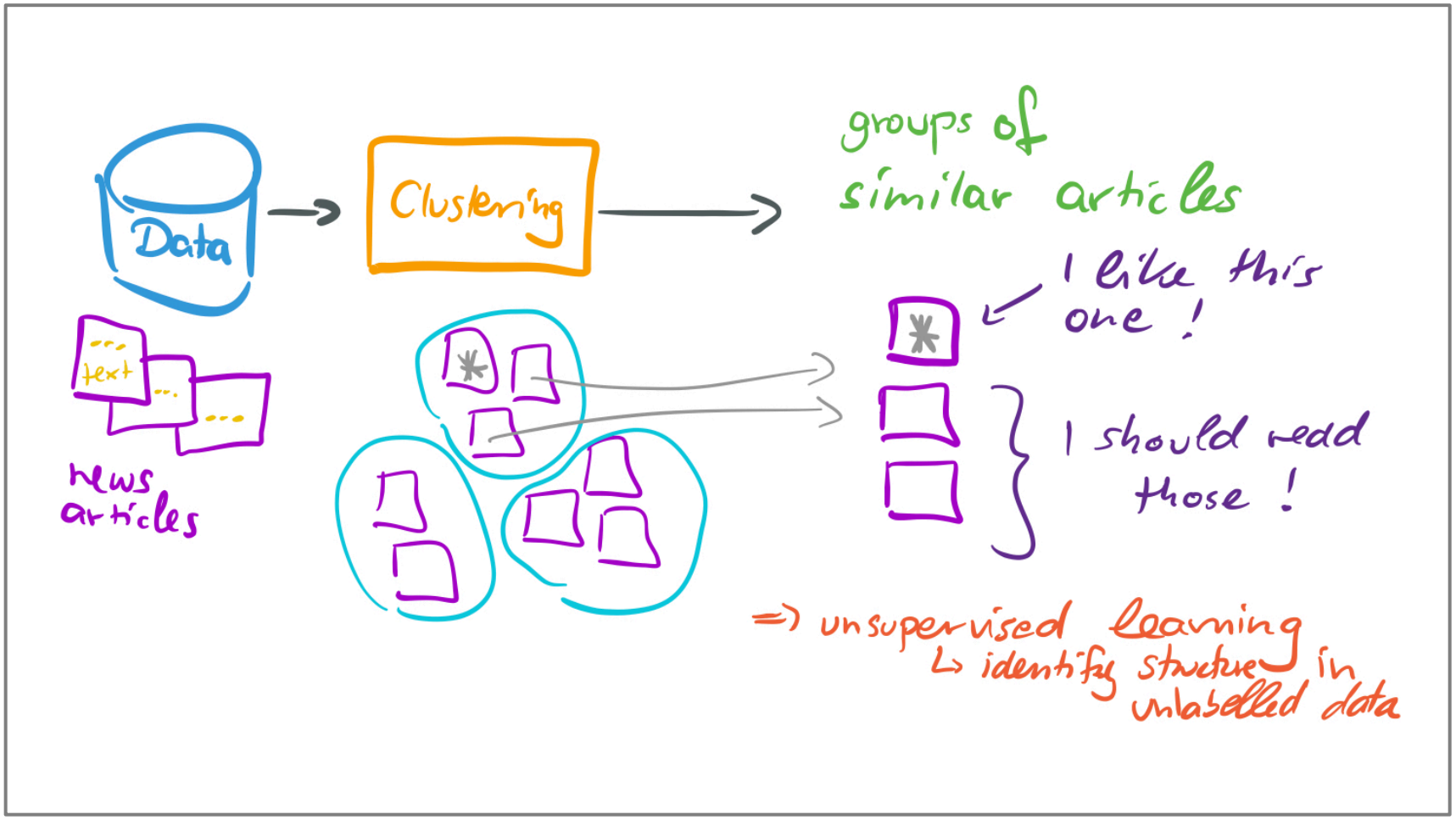
- Classification



LEARNING FROM DATA



- Clustering



WHAT IS MACHINE LEARNING?

regression, classification, clustering

…creating and using models that learn from data…

→ **supervised learning**/predictive modelling

- come up with predictions
- extract knowledge/insights

→ **unsupervised learning**/data mining

[PDSH] p 332-342
[DSFS] p 141-142

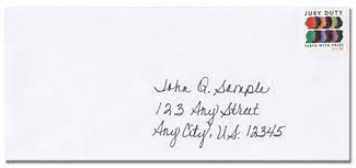
ACTIVITY 1

regression, classification, clustering

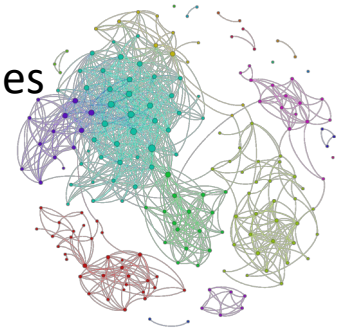
…creating and using models that learn from data…

Categorize these Examples

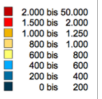
Identifying zip code from handwritten digits



Detecting communities in social networks

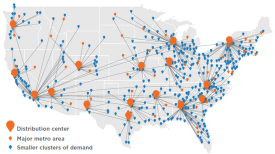


Predicting the traffic volume at rush hour



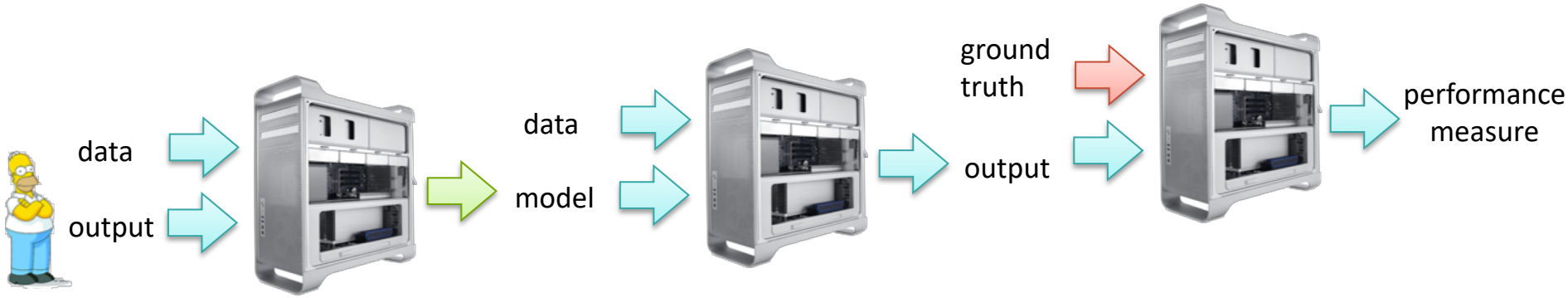
Detecting fraudulent credit card transactions

Determining the location of distribution centers based on customers' residence



MACHINE LEARNING WORKFLOW

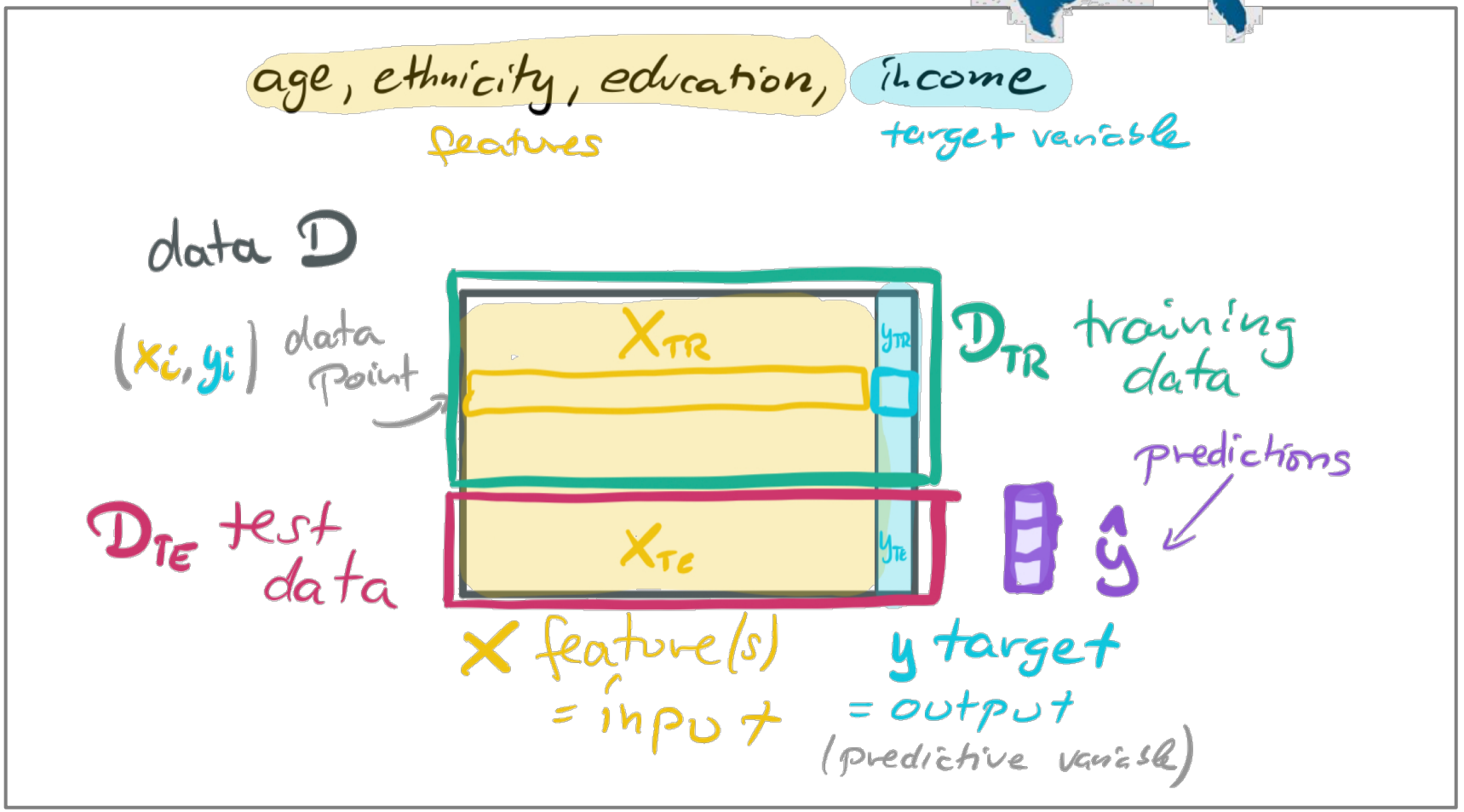
- *training phase*, *test phase*, *evaluation phase*



→ let's have a closer look at the *data* we are using

ACTIVITY 2

- Example: Census Data



- *training data* and *test data*

DATA

- Notation:
  - D all observed data
  - X all features
  - y observations
  - $\square_{TE}$ test
  - $\square_{TR}$ training
  - $\hat{y}$ predictions

Helper Notation:

n number of data points

d number of features

m number of training points

$\square_{1,\ldots,i,\ldots,n}$: indices for data points

$\square_{1,\ldots,j,\ldots,d}$: indices for features

- What data structure to use?
  - *set*, *list*, or *array*?

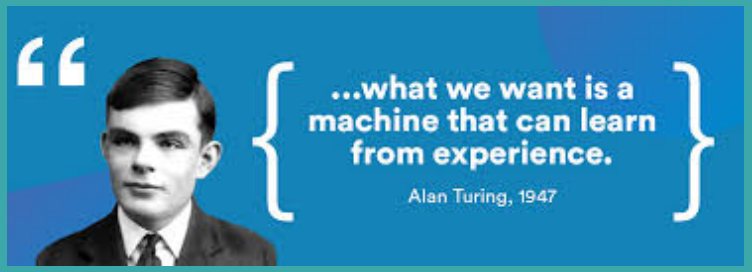
SUMMARY & READING

- *Data Science* is about

  data, **models**, and **evaluation**

- *Data Science* can solve a wide **variety of problems** – once we have the *right* data *and* model!

- 9 DS problems https://hackernoon.com/9-unusual-problems-that-can-be-solved-using-data-science-e7dbb89aa0c4
- **DSFS**
  - Ch1: Introduction
  - Ch11: Machine Learning (p141-142)
- **PDSH**
  - Preface: xi-xii
  - Ch5: Machine Learning (p331-342)



- History of ML https://www.forbes.com/sites/insights-intelai/2018/07/17/from-imitation-games-to-the-real-thing-a-brief-history-of-machine-learning/ - 240977b22056