RICE | D2K
Data to Knowledge Lab

Welcome to DSCI 101

Introduction to Data Science

Week 11 Recap

- Linear regression models
  - correlation and simple linear regression
  - multiple linear regression
  - categorical predictors
  - interpret regression coefficients

- Logistic regression with binary response
  - logit and logistic function
  - interpret logistic coefficients as log odds ratio

Week 12 Preview

- Introduction to Machine Learning
  - Supervised, unsupervised and semi-supervised learning
  - Real world examples

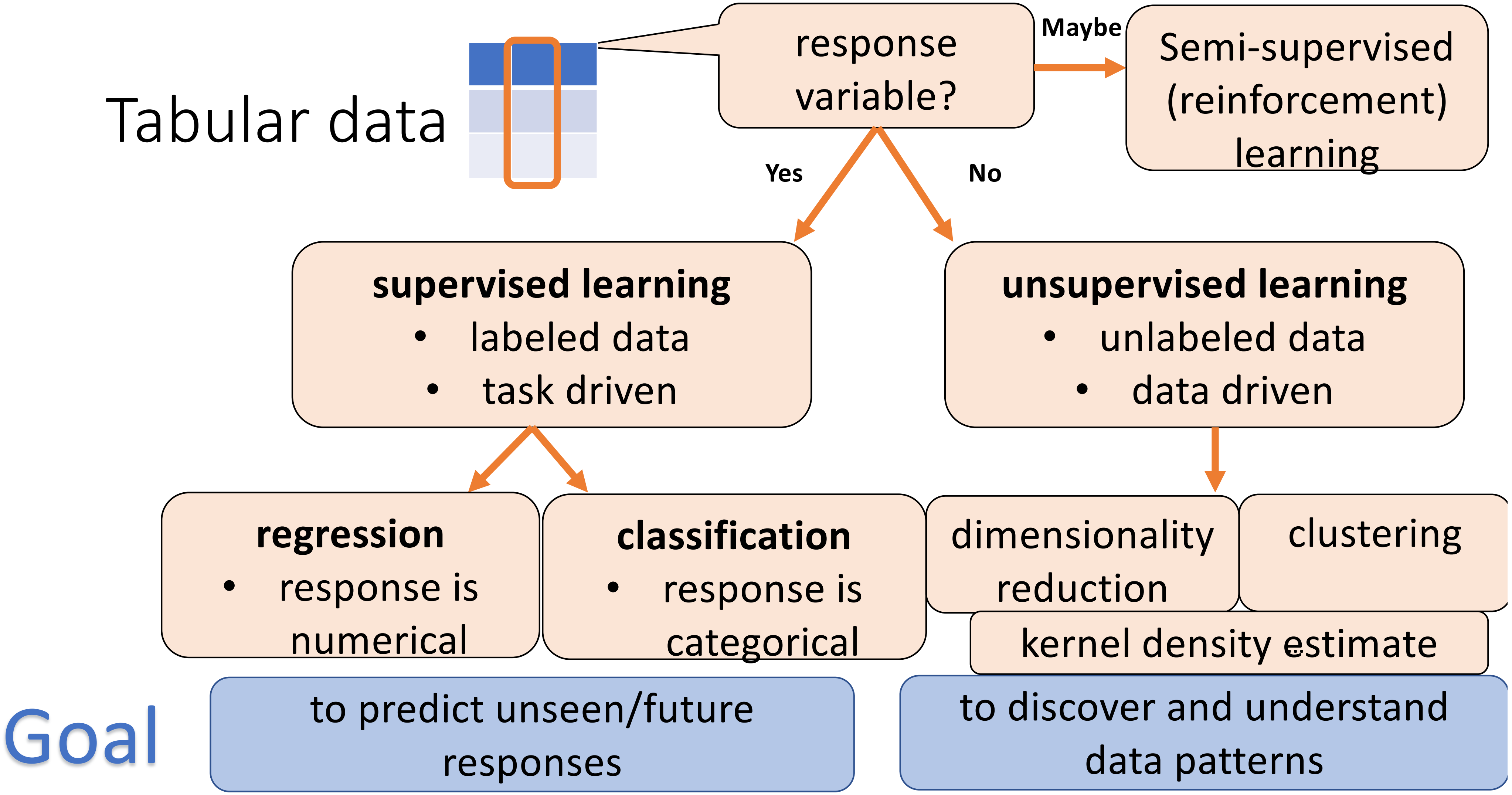
- Supervised learning
  - Regression vs. classification
  - Deep learning and AI

- Your first supervised learning model
  - K-nearest neighbor

Machine Learning

- Definition according to Wiki:
  - study of **computer algorithms** that can **improve automatically** through experience and by **the use of data**
  - when computational skills meets statistical thinking

- Learning goals:
  - vocabulary, concepts, intuitions
  - some basic but very useful models
  - DSCI303 – Machine Learning in DSCI minor
    - ELEC378, ELEC478, COMP341, STAT413…



Tabular data

response variable?

Maybe

Semi-supervised (reinforcement) learning

Yes

No

supervised learning

- labeled data
- task driven

unsupervised learning

- unlabeled data
- data driven

regression

- response is numerical

classification

- response is categorical

dimensionality reduction

clustering

kernel density estimate

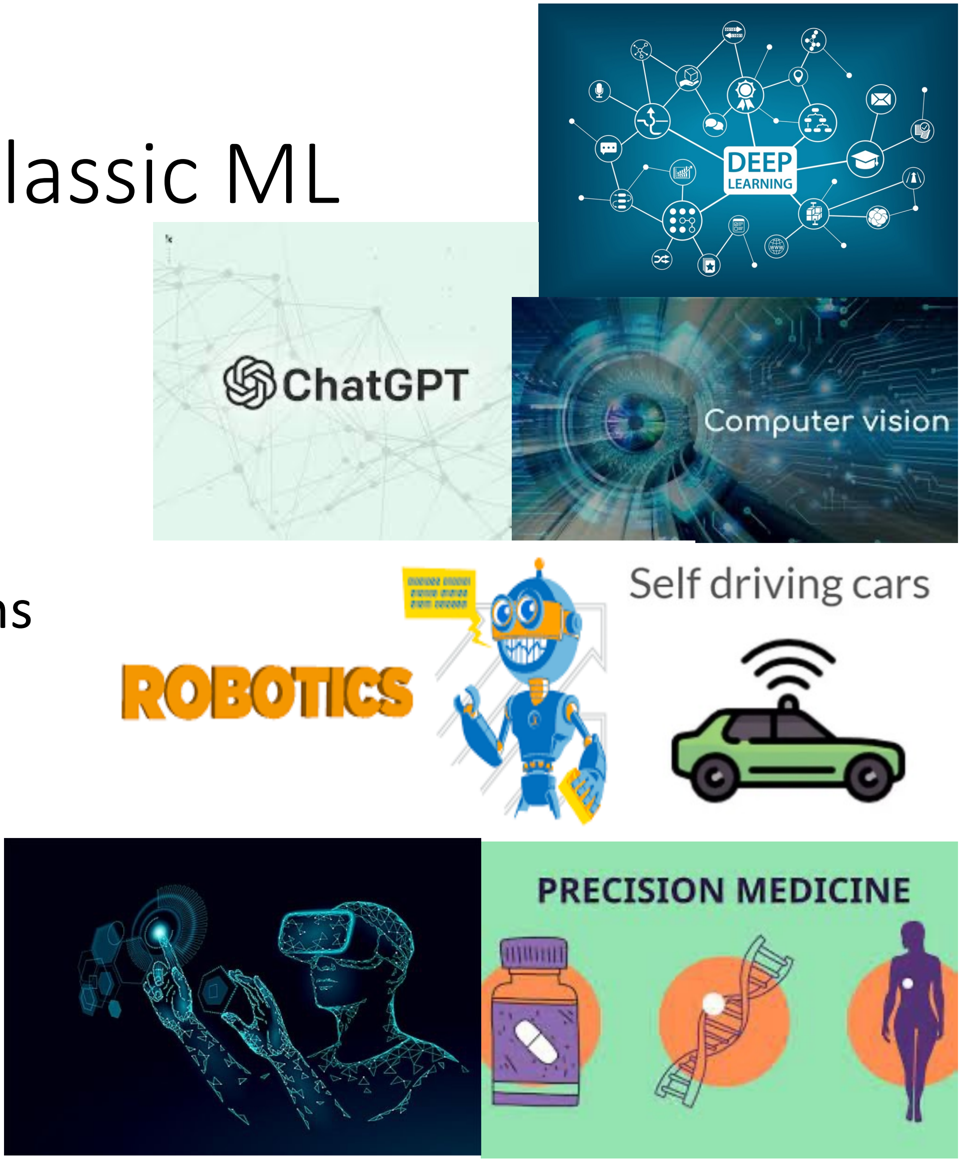
Goal

to predict unseen/future responses

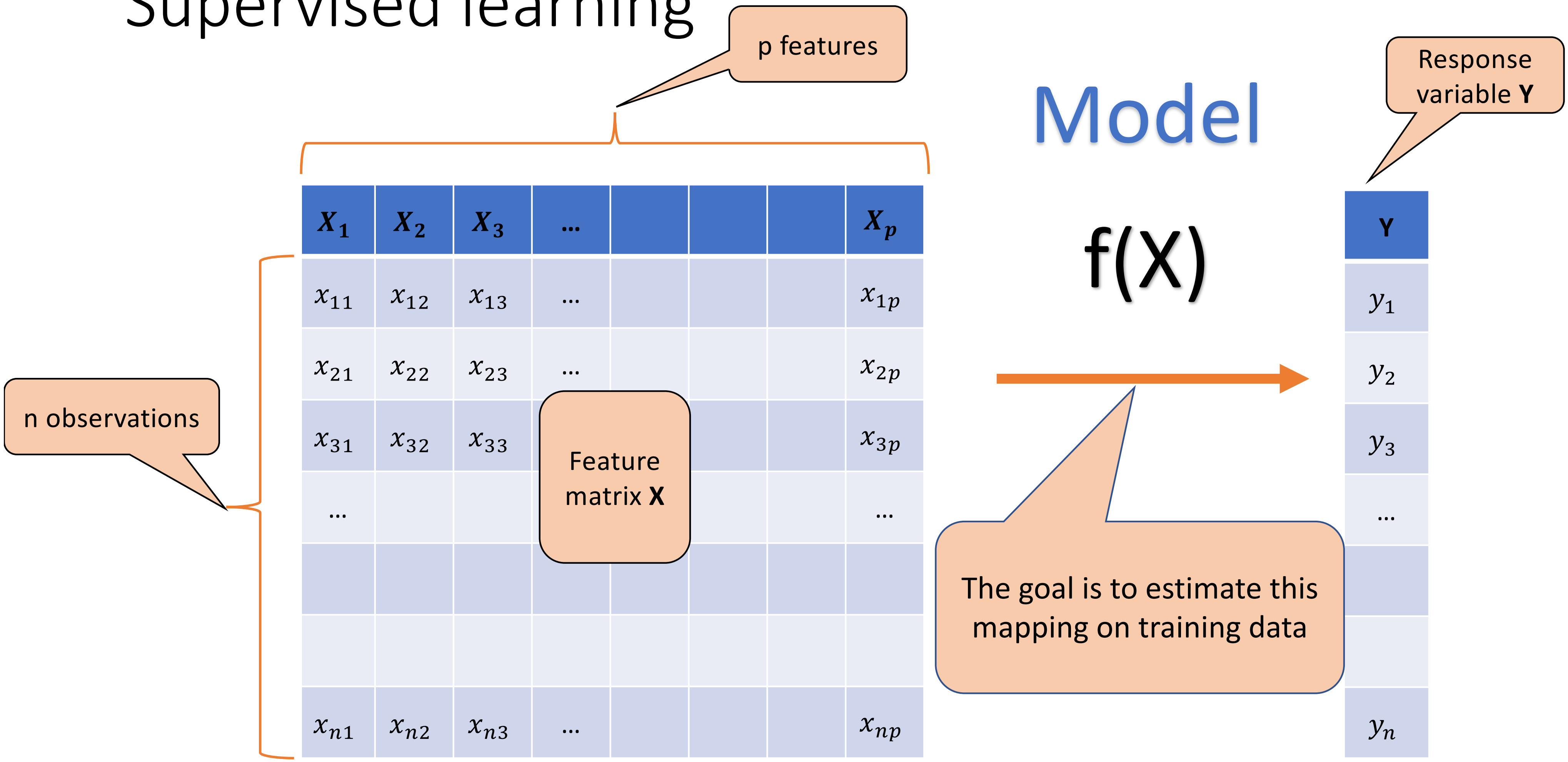
to discover and understand data patterns

Real-world examples of classic ML

- Spam email filter
- Credit card fraud detection
- Pricing Airbnb properties
- Forecast number of Covid19 infections
- Identify the risk factors for wildfire
- Discover breast cancer subtypes
- Recommendation system
- …



Supervised learning



p features

Model

$f(X)$

Response variable **Y**

n observations

Feature matrix **X**

The goal is to estimate this mapping on training data

X_1	X_2	X_3	…				X_p
x_{11}	x_{12}	x_{13}	…				x_{1p}
x_{21}	x_{22}	x_{23}	…				x_{2p}
x_{31}	x_{32}	x_{33}					x_{3p}
…							…
x_{n1}	x_{n2}	x_{n3}	…				x_{np}

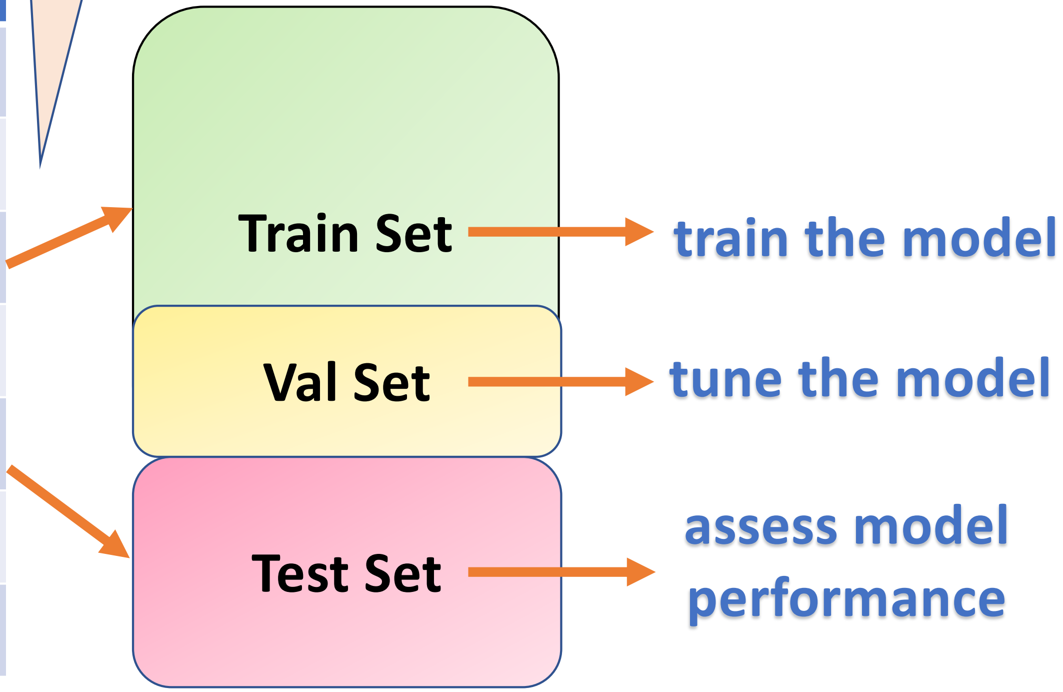
Y
y_1
y_2
y_3
…
y_n

Data

Randomly split the rows into training vs. testing

Model

X_1	X_2	X_3	**…**				X_p	Y
x_{11}	x_{12}	x_{13}	…				x_{1p}	y_1
x_{21}	x_{22}	x_{23}	…				x_{2p}	y_2
x_{31}	x_{32}	x_{33}					x_{3p}	y_3
…							…	…
x_{n1}	x_{n2}	x_{n3}	…				x_{np}	y_n



Train Set → **train the model**

Val Set → **tune the model**

Test Set → **assess model performance**

Regression model

- Model input: features, predictors, independent variables, covariates…
  - **X = $(\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots \boldsymbol{X_p})$** – always multidimensional
- Model output: response, dependent variable…
  - **Y** – a numerical variable
- Data: feature matrix **X** and response vector **Y**.
- Regression model
  - $\boldsymbol{Y = f(X_1, X_2, \cdots X_p) + \epsilon}$ $\left\{ \begin{array}{l} y_1 = f\left(x_{11}, x_{22}, \cdots x_{1p}\right) + \epsilon_1 \\ y_2 = f\left(x_{21}, x_{22}, \cdots x_{2p}\right) + \epsilon_2 \\ \cdots \\ \cdots \\ y_n = f\left(x_{n1}, x_{n2}, \cdots x_{np}\right) + \epsilon_n \end{array} \right.$

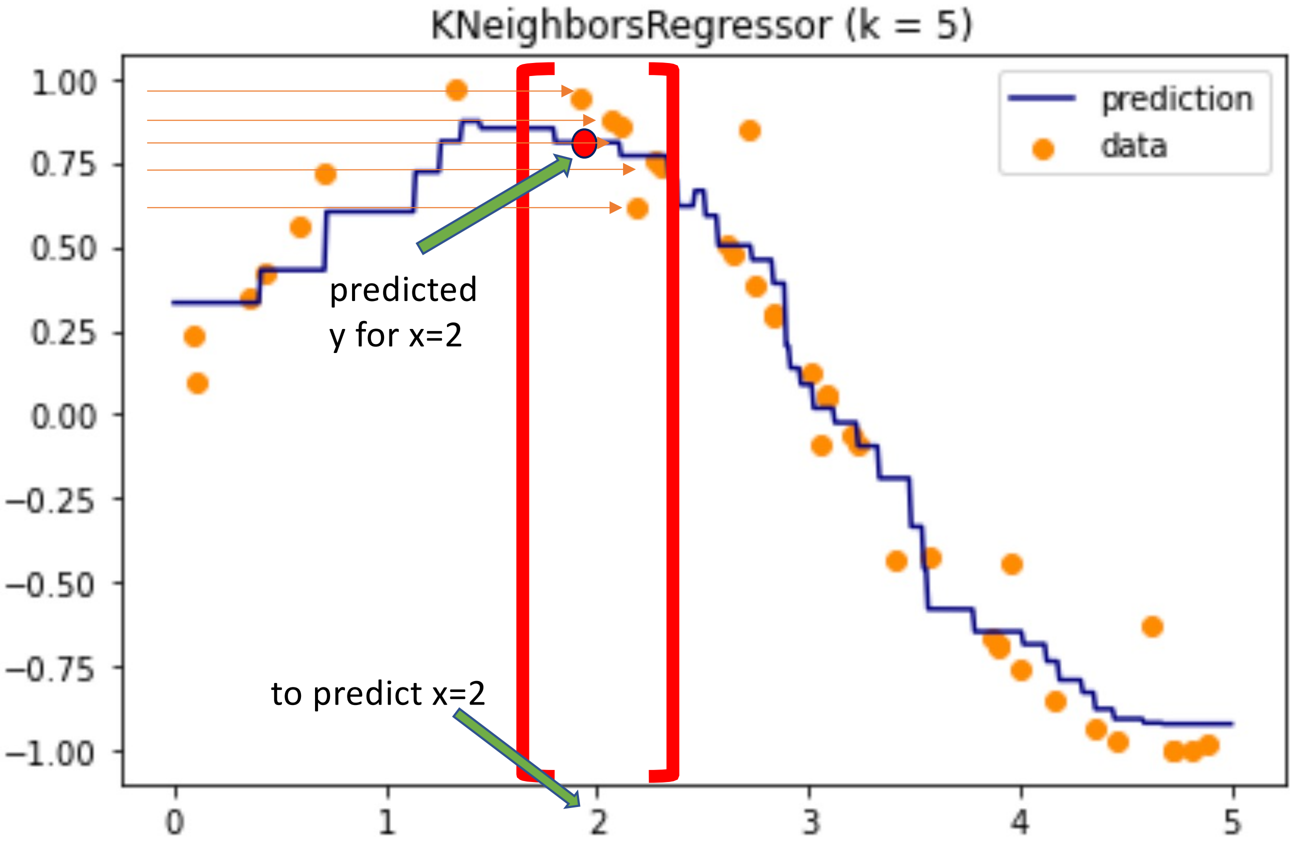
Model training using optimization

- How can we train a regression model?
  - With some estimate $\hat{f}$ we can have $\hat{y}$=$\hat{f}(\mathrm{x})$
  - How can we find **"the best"** $\hat{\boldsymbol{f}}$???

- **Loss function / cost function / error measure:**
  - $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ → **Square Loss**
  - Or equivalently **Mean Square Error** (MSE)
- The goal is to find **the** $\hat{\boldsymbol{f}}$ that can **minimize mean square loss**
  - other loss: mean absolute loss, Huber loss…

K-nearest neighbor regression

- Example of K=5
  - to predict x=2, find a neighborhood of x=2 that contains 5 training data
  - average those 5 corresponding y value as prediction for x=2
  - **avg to min MSE**
- What values can K take?



Classification model

- Most concepts in regression also apply in classification models
- Now this model $\boldsymbol{f}(\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots \boldsymbol{X_p})$ needs to output a class label

- **Loss function / cost function / error measure:**
  - proportion of labels predicted wrong → Misclassification Rate
  - also known as 0-1 loss

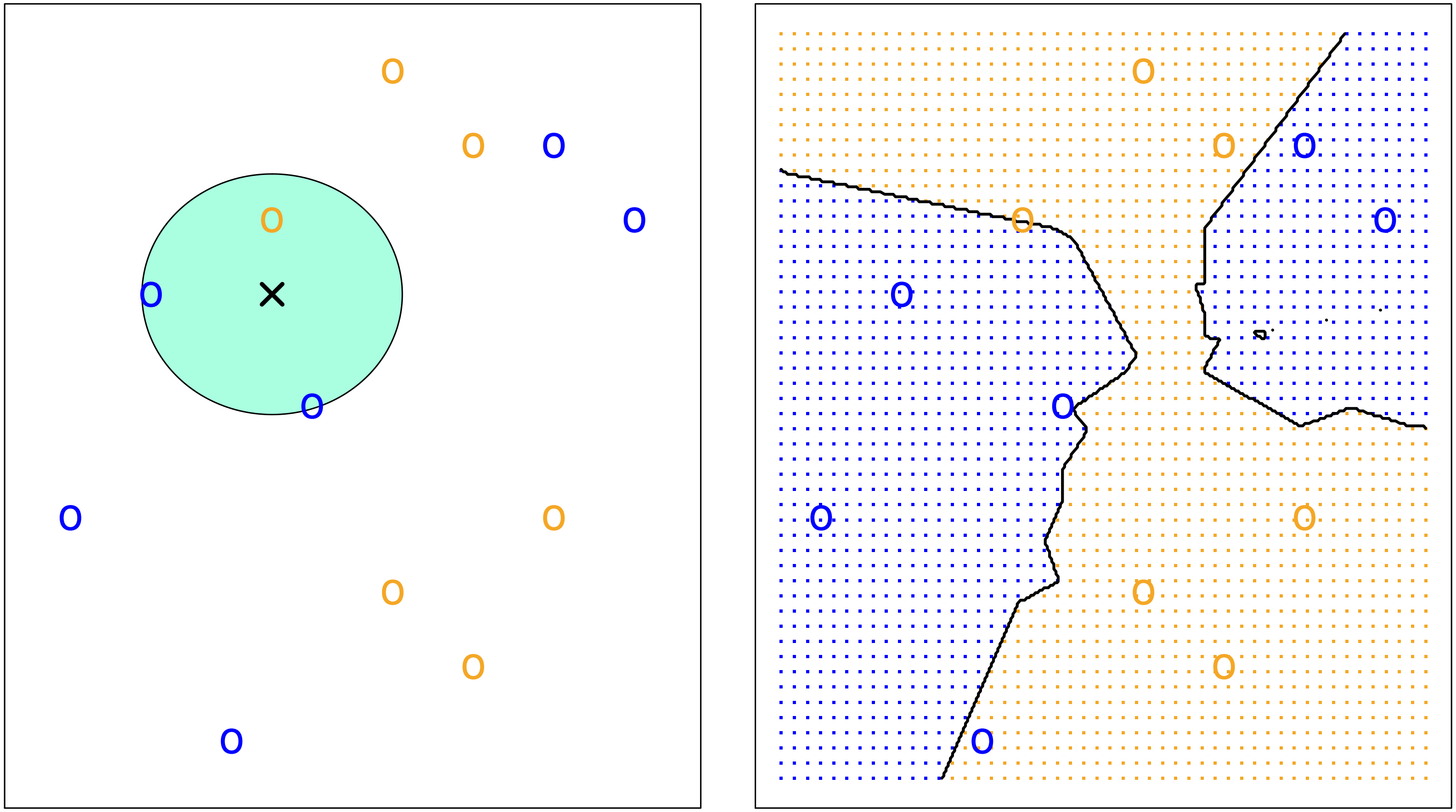
- The goal is to find **the** $\hat{\boldsymbol{f}}$ that can minimize the chosen loss
  - other loss: cross entropy loss, hinge loss

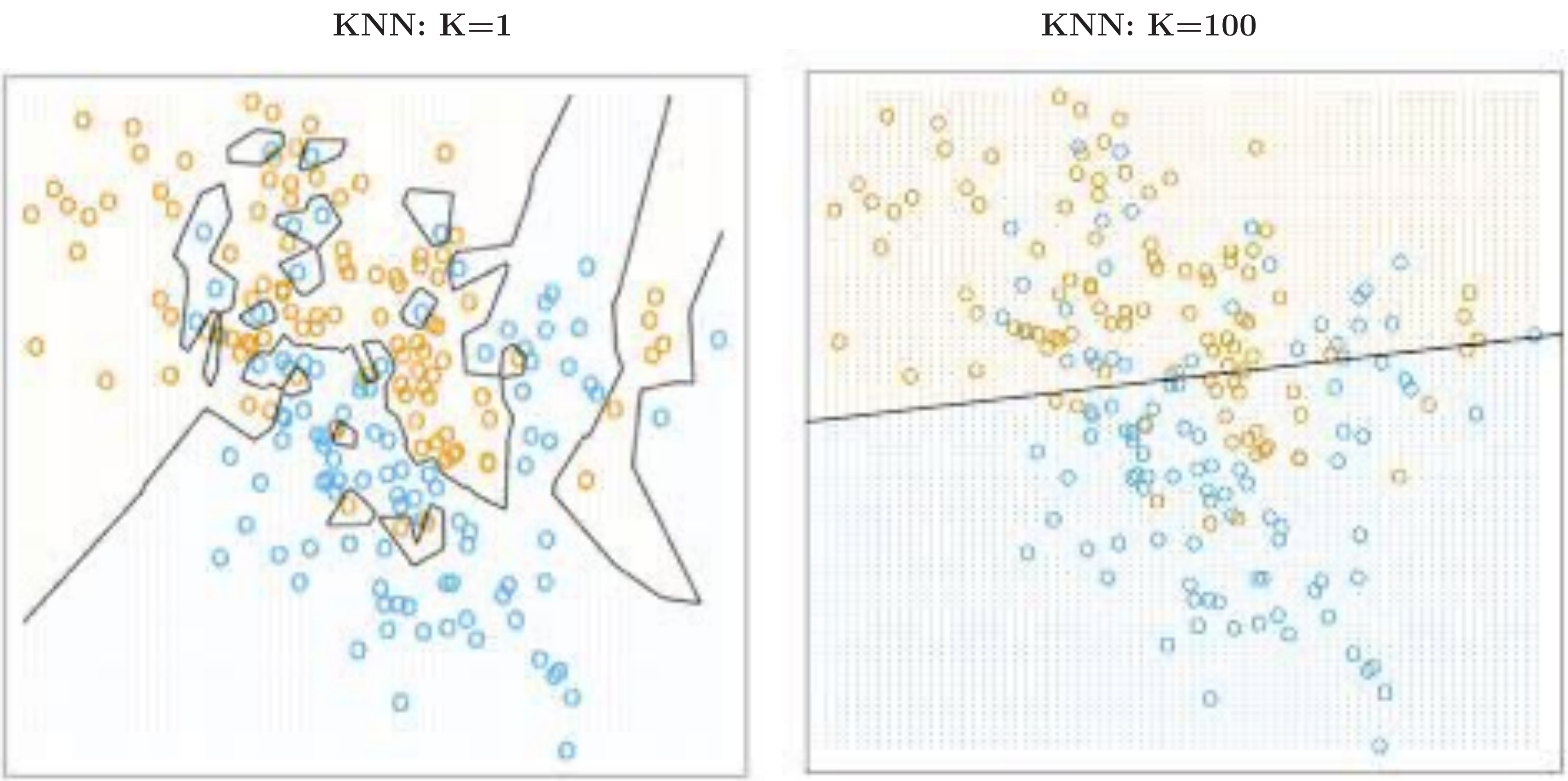
K-nearest neighbor classifier

- Can this idea apply to classification?

- Yes!
  - instead of average over y value in the neighborhood
  - take a **majority of vote** for class labels in the neighborhood

- KNN classifier works surprisingly well if
  - K is properly tuned!
  - number of features is not too large





KNN: K=1

KNN: K=100

Model choice and trade-off

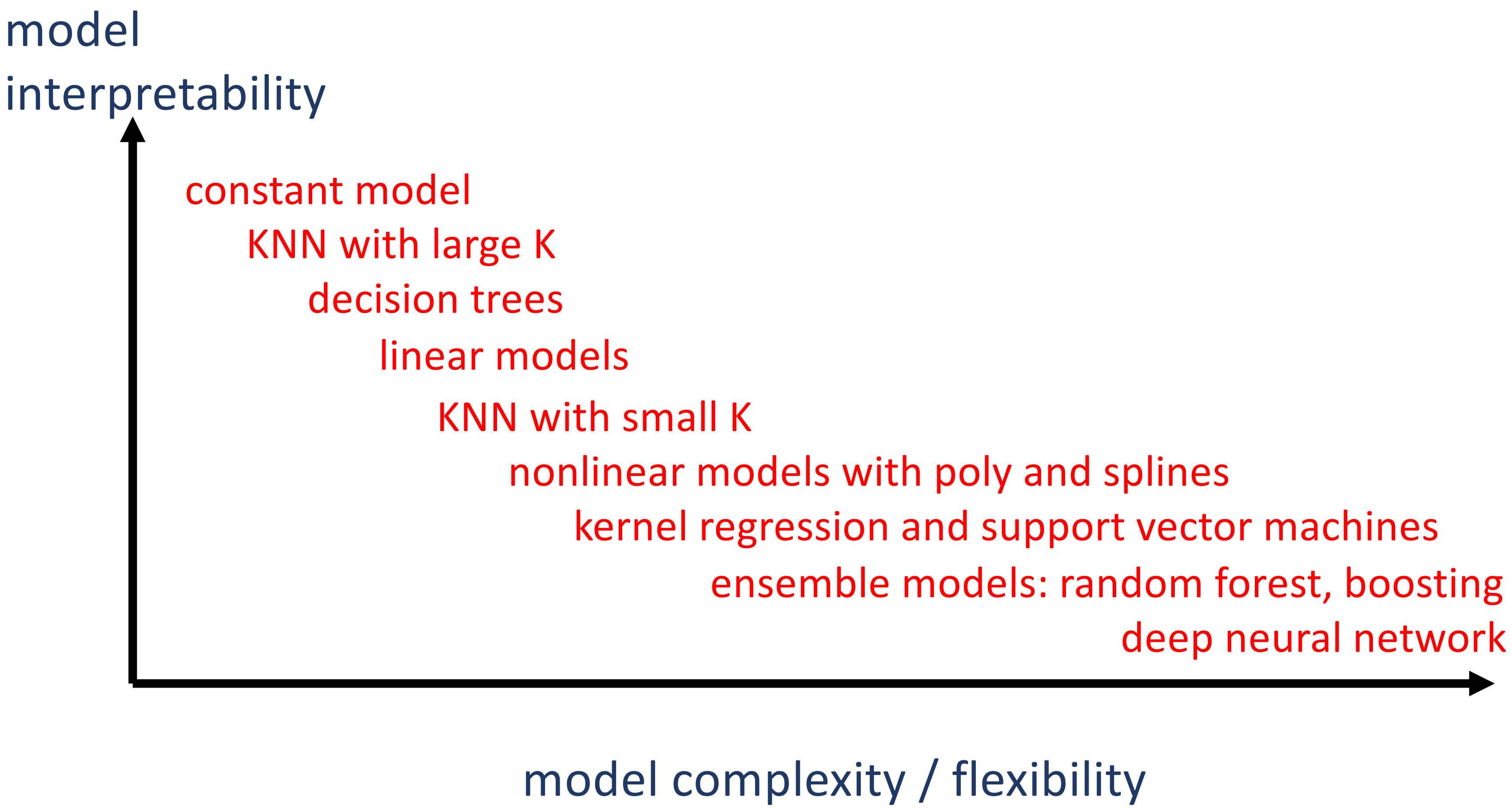
- Trade-off between model complexity and interpretability

- Intuitions about high complex models:
    - more flexible/wiggly/expressive regression function & decision boundary
    - more parameters in the model
    - more training data and computation

- Intuitions about high interpretable models:
    - can explain why the model makes such prediction
    - can explain relation between features and response

- Occam’s razor (aka principle of parsimony)

Model complexity and interpretability trade-off



model interpretability

constant model

KNN with large K

decision trees

linear models

KNN with small K

nonlinear models with poly and splines

kernel regression and support vector machines

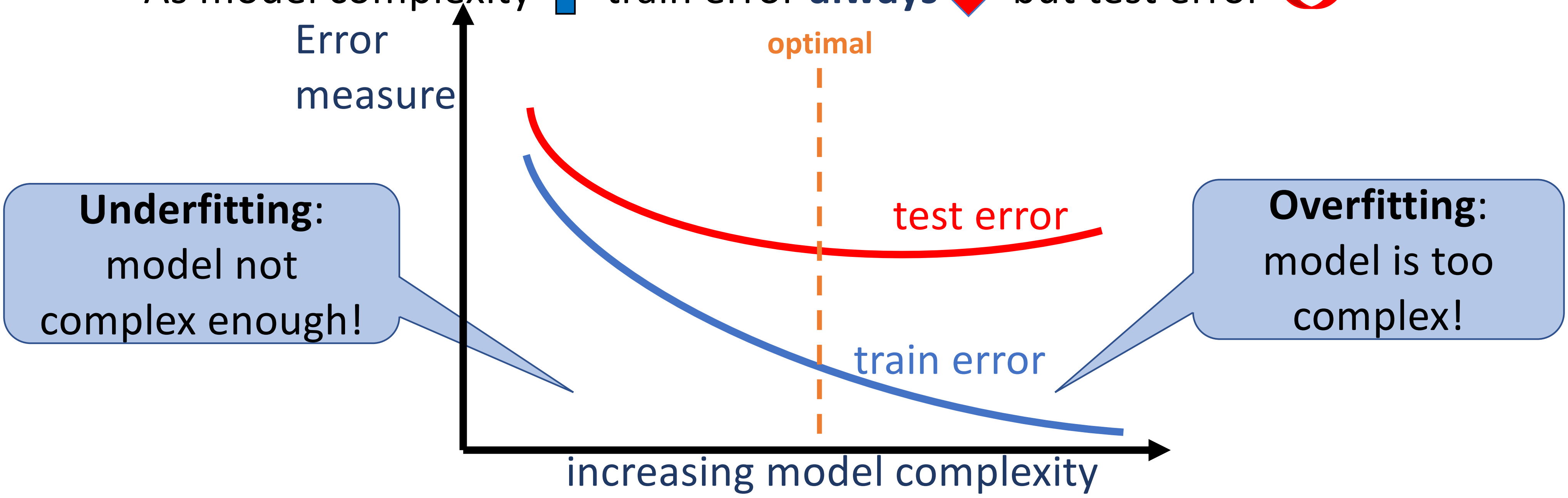
ensemble models: random forest, boosting

deep neural network

model complexity / flexibility

Train vs. test error

- Testing error: evaluate your model on a fresh testing set
    - training error < test error (on average)
- As model complexity ⬆ train error **always** ⬇ but test error ⤴



Underfitting: model not complex enough!

Overfitting: model is too complex!

Underfitting:

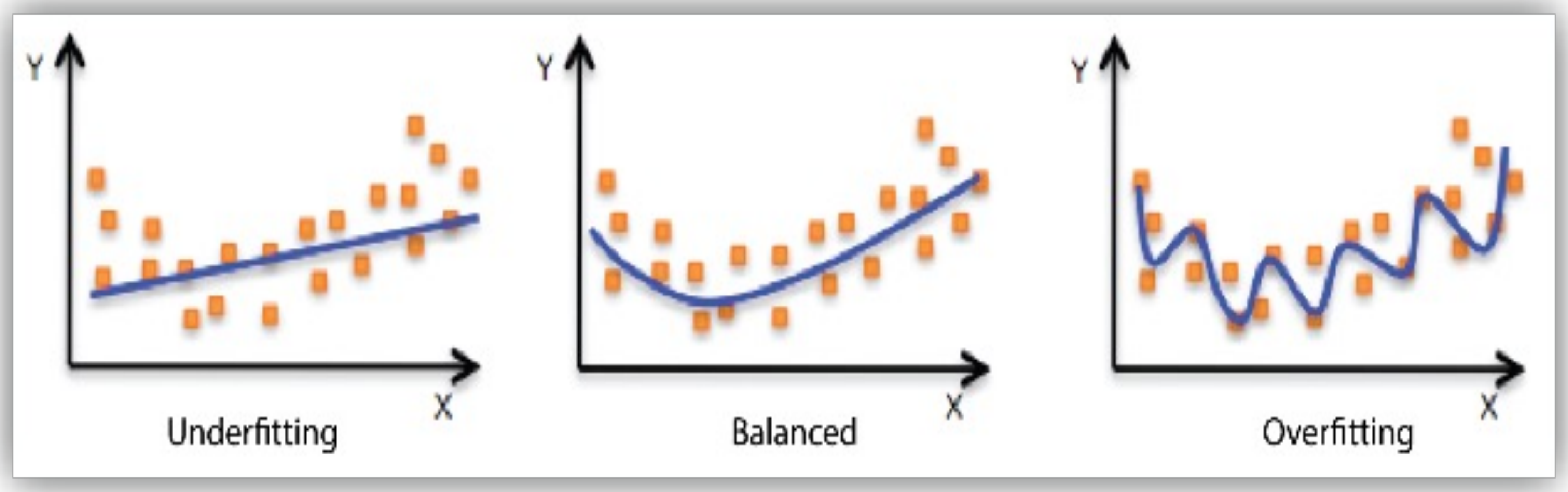
- training error: ☹
- testing error: ☹

Balanced:

- training error: ☺
- testing error: ☺

Overfitting:

- training error: ☺
- testing error: ☹



Underfitting:

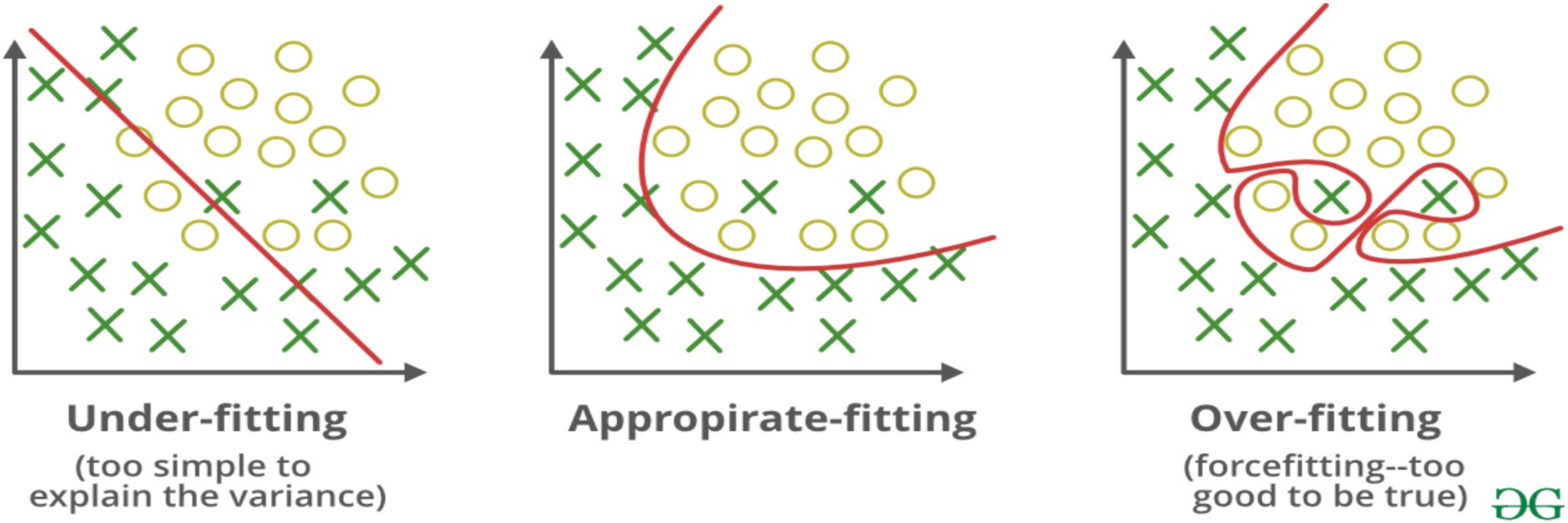
- training error: ☹
- testing error: ☹

Balanced:

- training error: ☺
- testing error: ☺

Overfitting:

- training error: ☺
- testing error: ☹


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)
ƏG

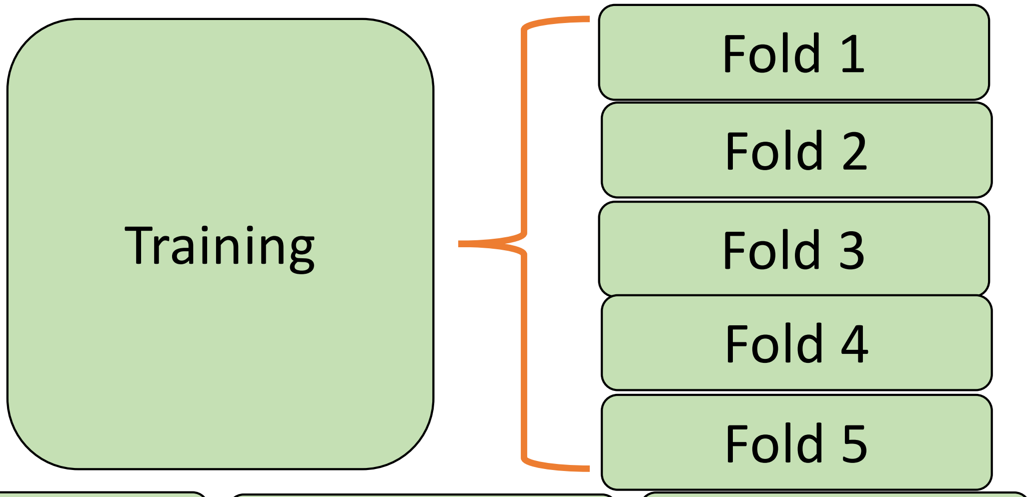
How to find the best K?

- Validation set!
  - For each value of K you want to try (for example 3, 5, 13, 25, 55…)
  - Fit KNN on training set and predict on validation set
  - Pick the K with the best validation set performance
  - Predict final model performance on test set

- Why do we need the validation set?
  - If we use the same training set to make prediction…
  - What is the training error for K=1?

K fold Cross-validation

- No validation set: how to tune hyperparameters?



Training

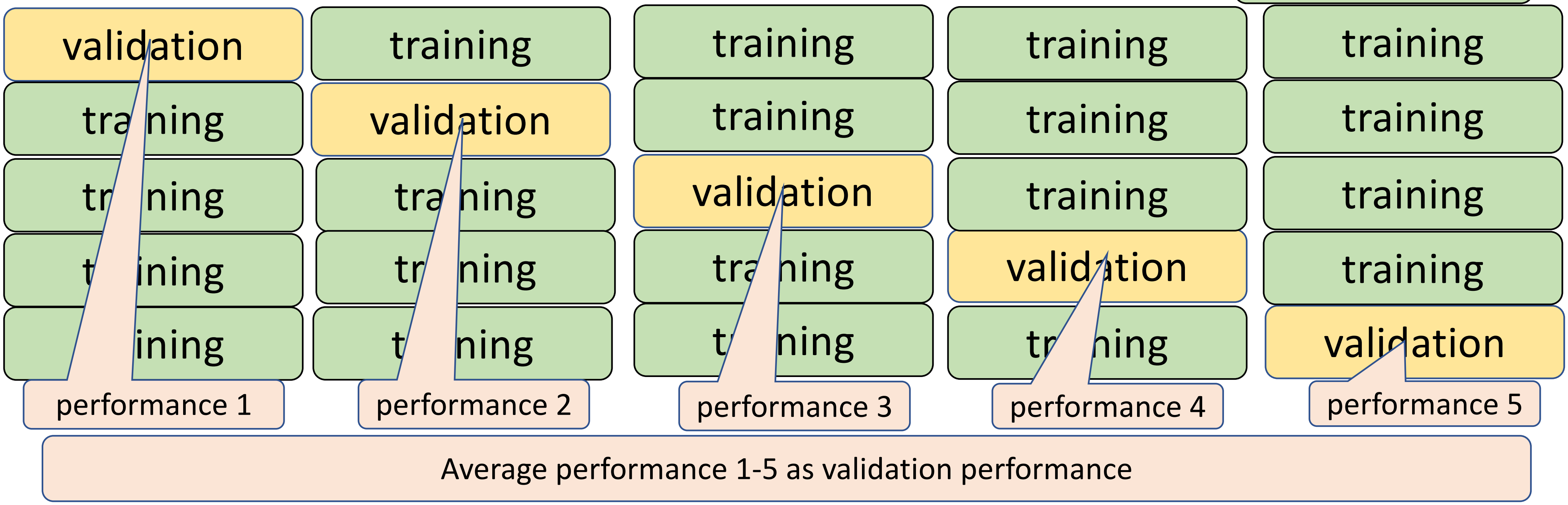
Fold 1

Fold 2

Fold 3

Fold 4

Fold 5



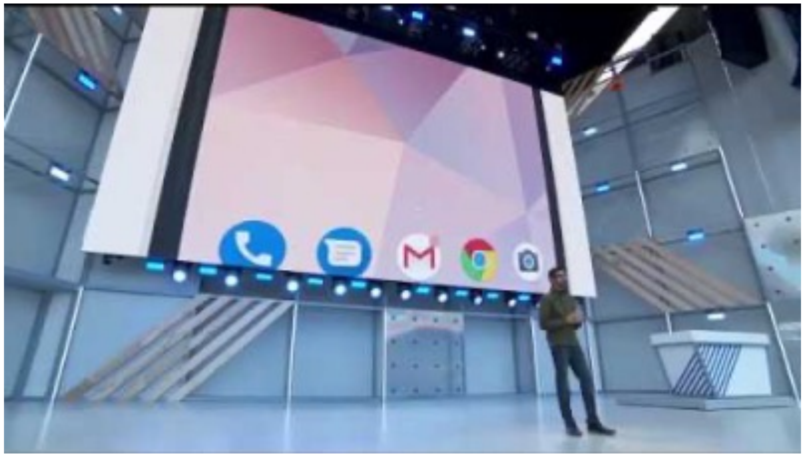
validation	training	training	training	training
training	validation	training	training	training
training	training	validation	training	training
training	training	training	validation	training
training	training	training	training	validation
performance 1	performance 2	performance 3	performance 4	performance 5

Average performance 1-5 as validation performance

Artificial Intelligence

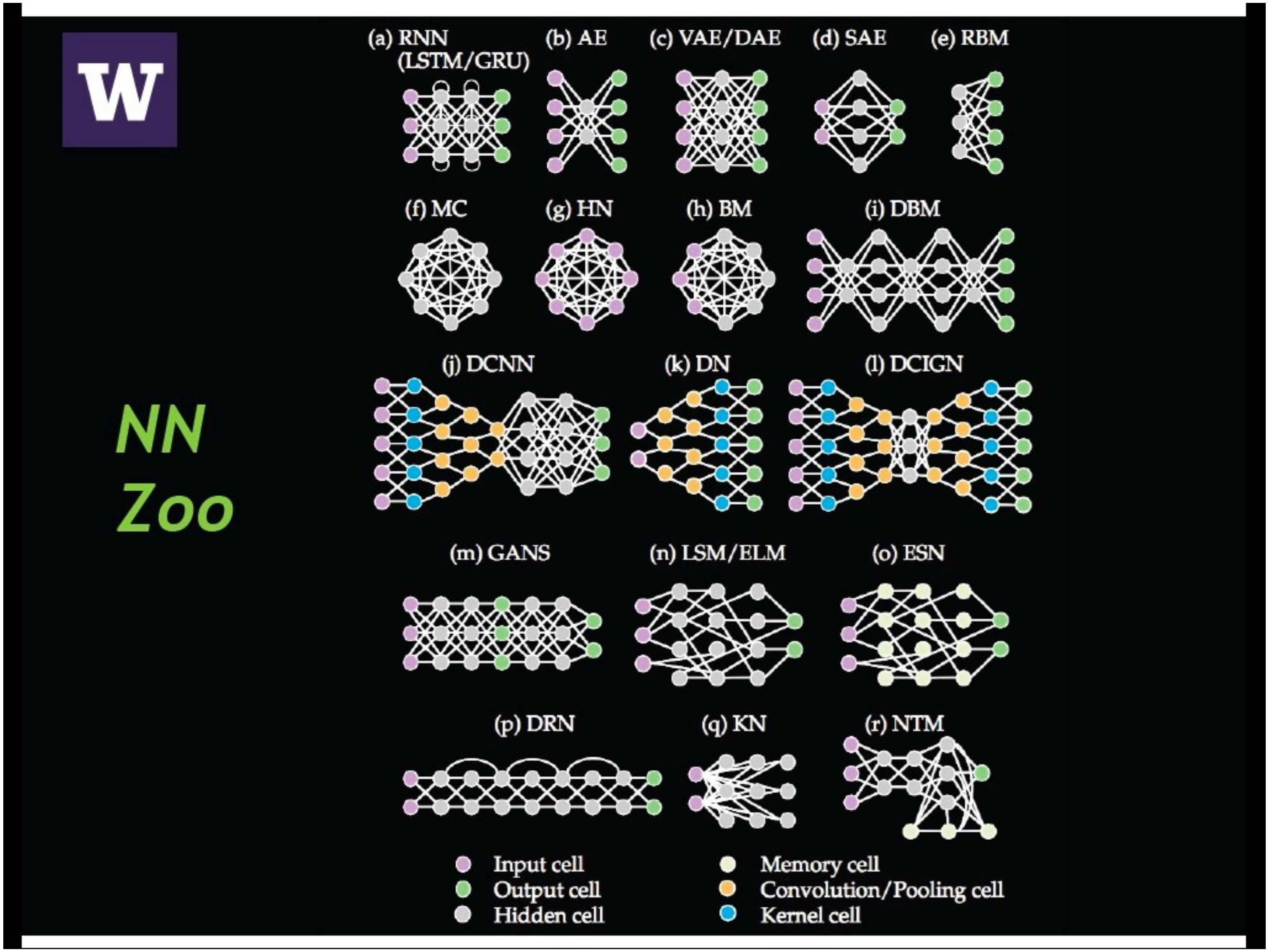
- AI is changing the way we work and live!
- DL is a subset of ML, and ML is a subset of AI, AI is a subset of CS
- Demystifying AI:
    - ANI: current AI can do tasks that only take human brains seconds to do, recognize an image, listening and responding in a conversation, read a sentence of text and translate, drive a car
    - AGI: do anything human can do!

- Alexa, Siri, Google Assistant
    - Movie Her (2013)

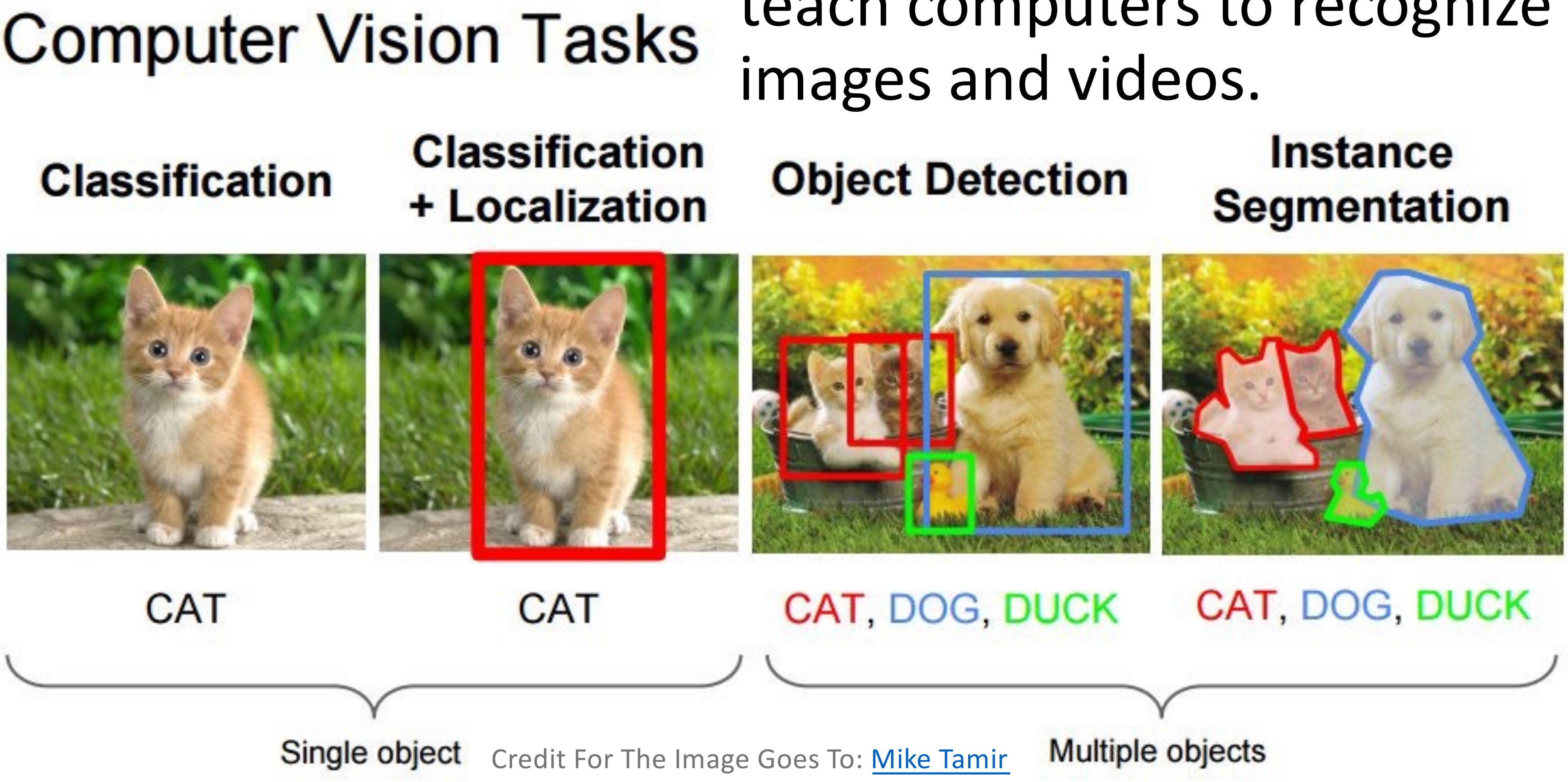


Deep Learning

- Apply Artificial Neural Network (NN) models to **supervised learning tasks**.
- NN has almost nothing to do with the brain.



teach computers to recognize images and videos.



Credit For The Image Goes To: [Mike Tamir]

Computer Vision Model: CNN

Donald J. Trump
@realDonaldTrump

You cannot trust CNN! They are FAKE!!!

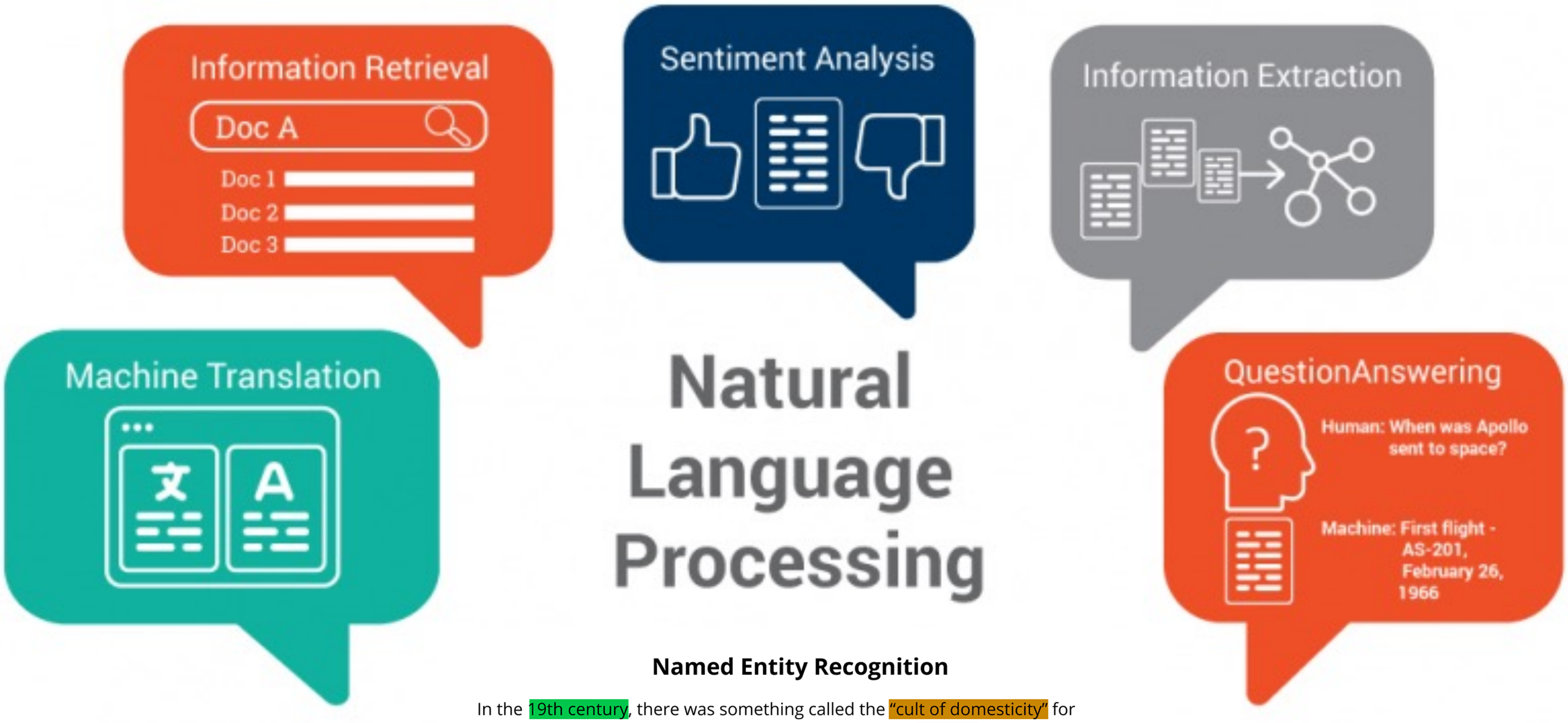
RETWEETS 7,771 LIKES 2,094

11:07 AM - 21 Nov 2017

364
8K
2K
what people think he's referring to
what he's actually referring to
Convolution
Fully co
L1
256x256
L2
128x128
L3
64x64
L4
32x32

Try it yourself:

- Google Vision AI
  - detect objects, emotion, landmark and location


Information Retrieval

Sentiment Analysis

Information Extraction

Machine Translation

Natural Language Processing

QuestionAnswering

Human: When was Apollo sent to space?

Machine: First flight - AS-201, February 26, 1966

Named Entity Recognition

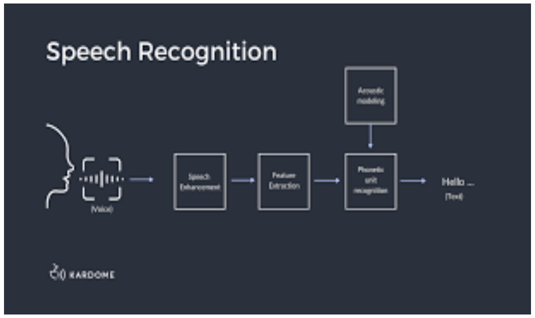
In the 19th century, there was something called the "cult of domesticity" for many American women. This meant that most married women were expected to stay in the home and raise children. As in other countries, American wives were very much under the control of their husband, and had almost no rights. Women who were not married had only a few jobs open to them, such as working in clothing factories and serving as maids. By the 19th century, women such as Lucretia Mott and Elizabeth Cady Stanton thought that women should have more rights. In 1848, many of these women met and agreed to fight for more rights for women, including voting. Many of the women involved in the movement for women's rights were also involved in the movement to end slavery.

Wikipedia

Tag colors:

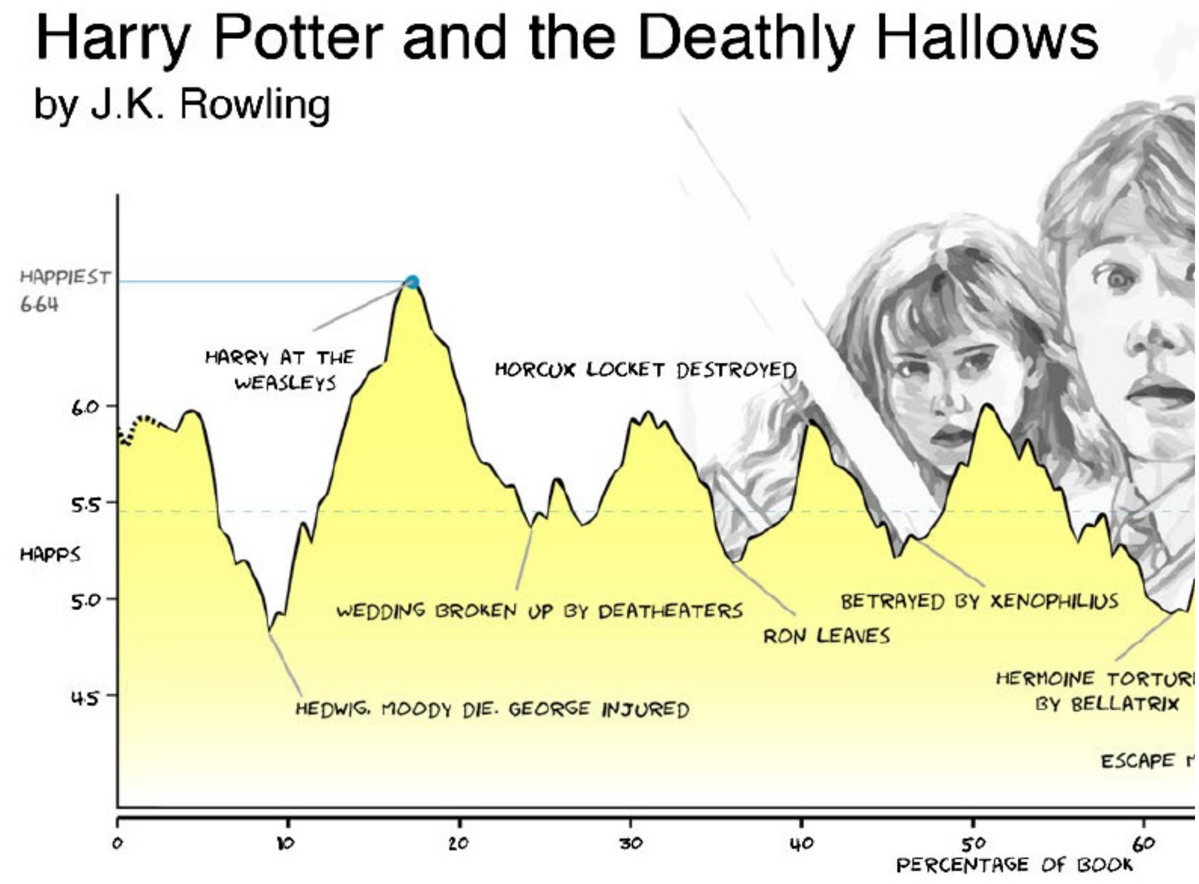
LOCATION PERSON TERM DATE CONDITION PROCESS PEOPLE

mobidev


Speech Recognition

The emotional arcs of stories are dominated by six basic shapes

Andrew J Reagan[1*], Lewis Mitchell[2], Dilan Kiley[1], Christopher M Danforth[1] and Peter Sheridan Dodds[1]



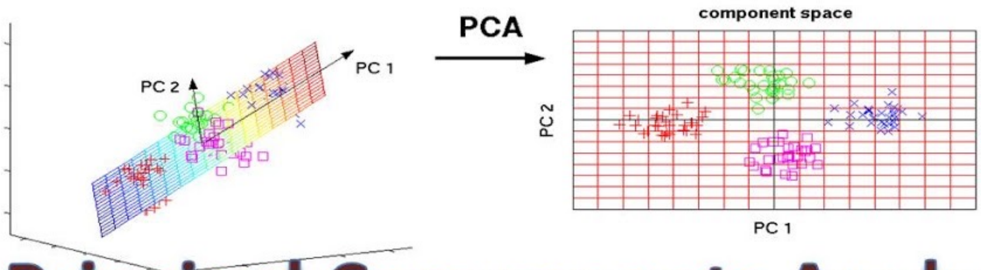
Try it yourself:

- Google Natural Language AI
  - sentiment analysis
  - syntax and grammar

Unsupervised learning

visualize big data

finding subgroups



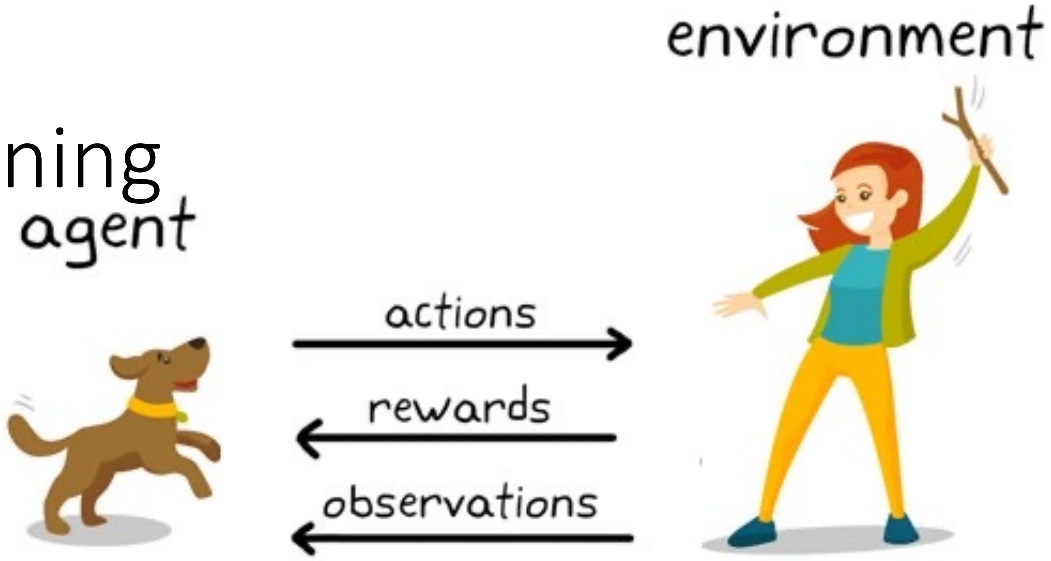
Principal Components Analysis (PCA)

Clustering

many others MDS, NMF, t-SNE…

K-means and so many more…

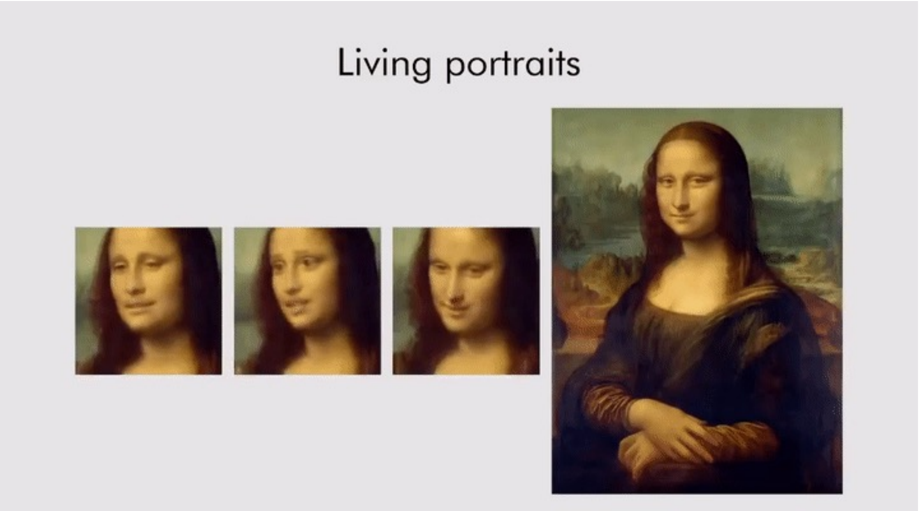
Semi-supervised learning



- In general when data is partially labeled

- Reinforcement Learning:
  - AlphaGo
  - Player of Games

- Generative Adversarial Networks (GAN)
  - generate data to fool the model



Take away message

- Most important concepts:
  - Supervised, unsupervised and semi-supervised learning
  - Regression vs. classification in supervised learning
  - Deep learning and AI

- Trade-offs: to find the “best” fitting model to your data
  - Training error vs. testing error
  - Under fitting vs. over fitting

- Next week: supervised learning models
  - Decision trees for regression and classification
  - Ensemble model Random Forest