Lecture 4
Machine learning: concepts

GEOL 4397: Data analytics and machine learning for geoscientists

Jiajia Sun, Ph.D.

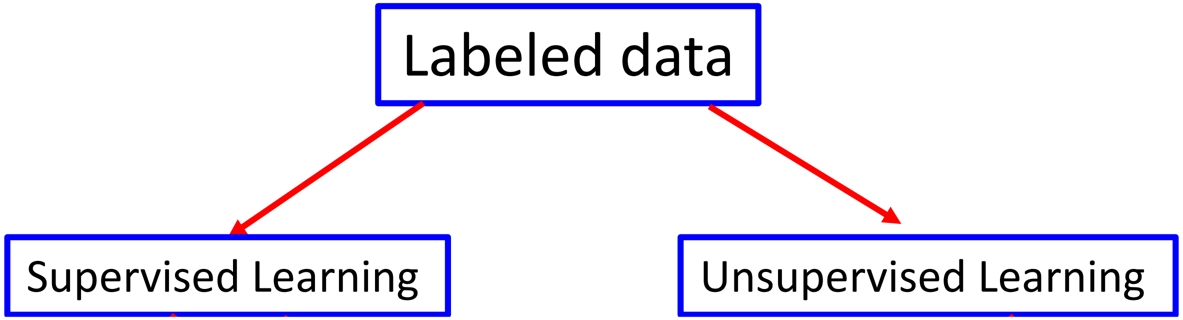
Jan. 31st, 2019

UNIVERSITY of HOUSTON
YOU ARE THE PRIDE
EARTH AND ATMOSPHERIC SCIENCES

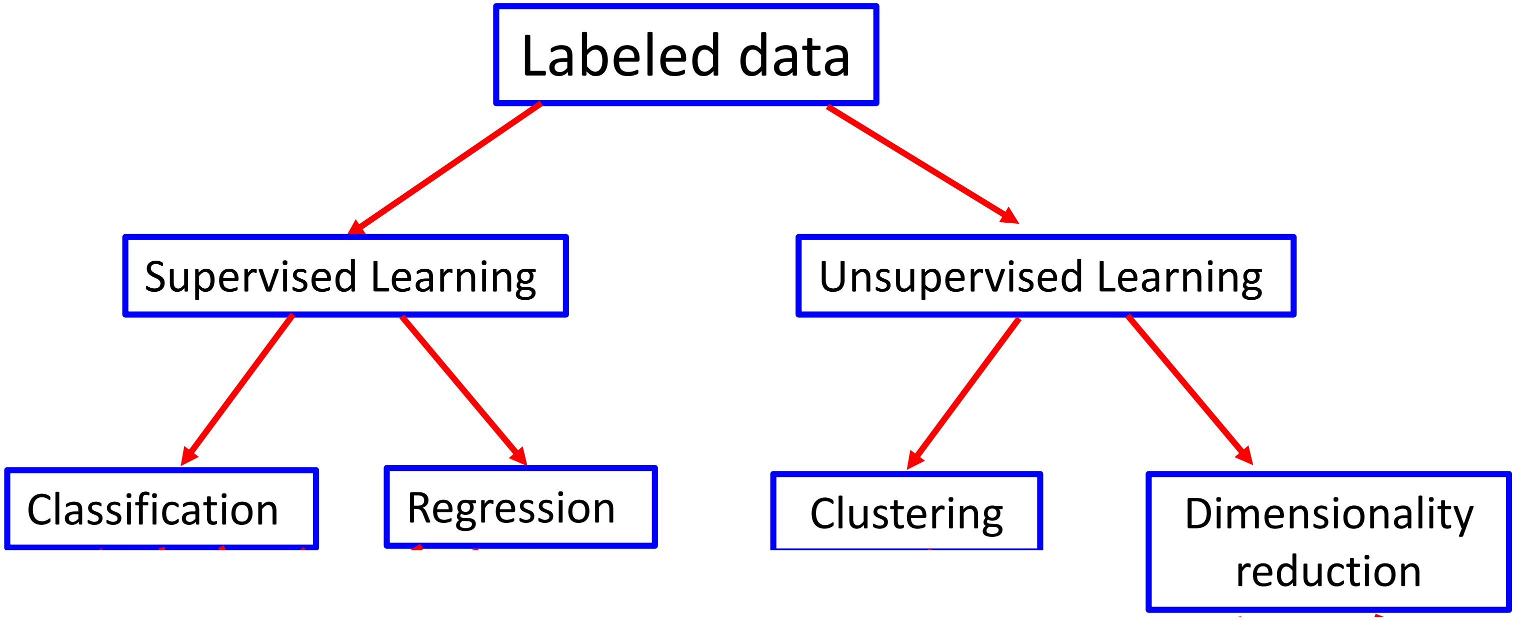
Today's agenda

- Supervised vs. unsupervised learning
- Regression vs. classification
- Overfit vs. underfit
- Bias vs. variance

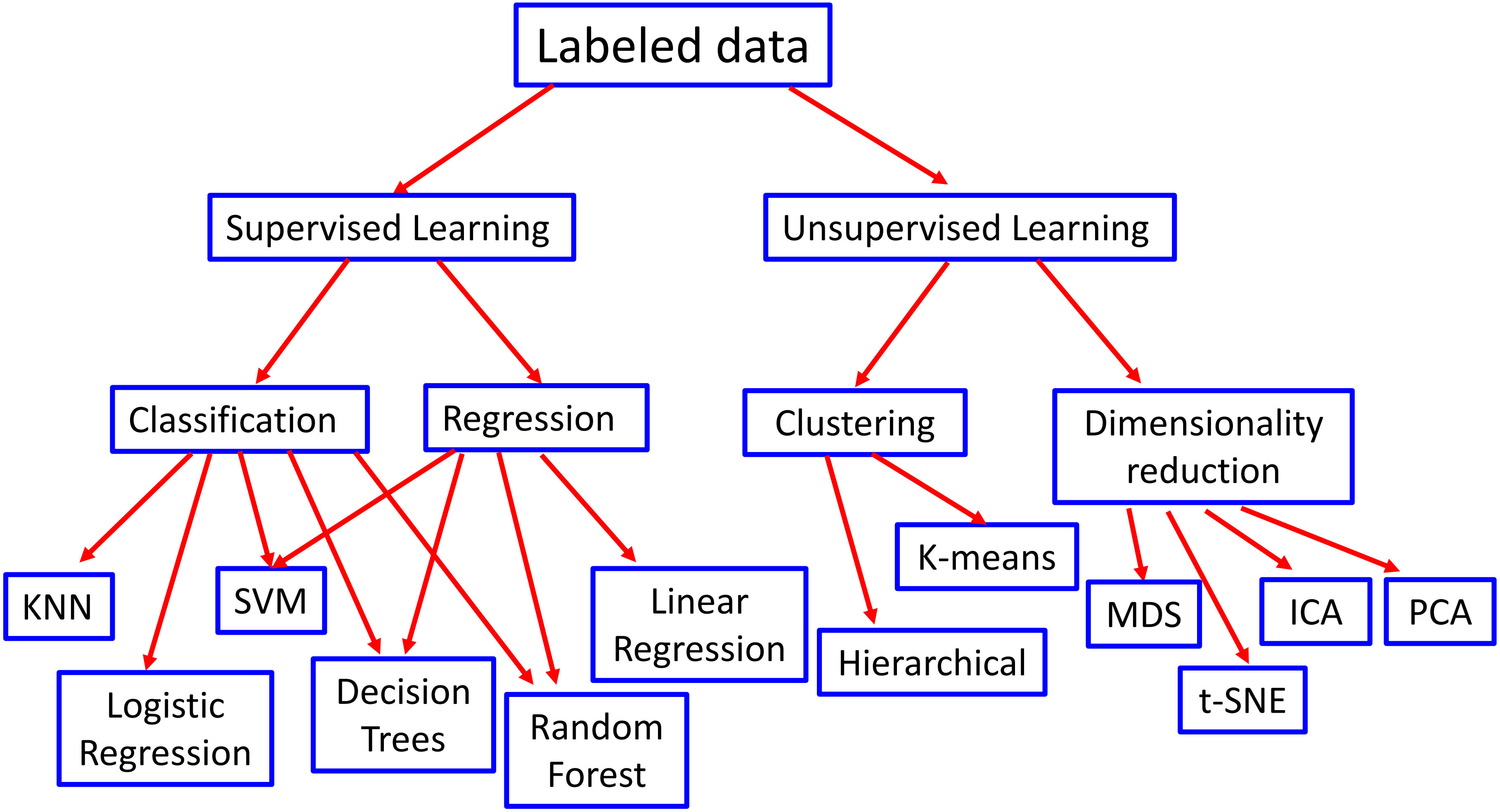
Machine learning algorithms



Machine learning algorithms



Machine learning algorithms



Supervised learning

- Training data come with labels (i.e., answers)
- Also termed labeled data
- E.g., cat classifier

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“cat”

http://animalsbirds.com/cats-hd-wallpapers-and-images-gallery/

“non-cat”

http://animalsbirds.com/animals-reindeer-photos-hd-wallpapers/

“non-cat”

http://animalsbirds.com/cheetah-photos-animals-hd-wallpapers-free-download/

“non-cat”

http://animalsbirds.com/animals-roe-deer-pictures/

“non-cat”

http://animalsbirds.com/discus-fish-images-gallery/symphysodon-discus-latest-hd-wallpapers-free-download/#main

Supervised learning: what is it?

- Let us consider each image as an input variable, x
- Also, consider each label as an output variable, y
- Supervised learning is all about learning a mapping function from the input to the output

$$y = f(x)$$

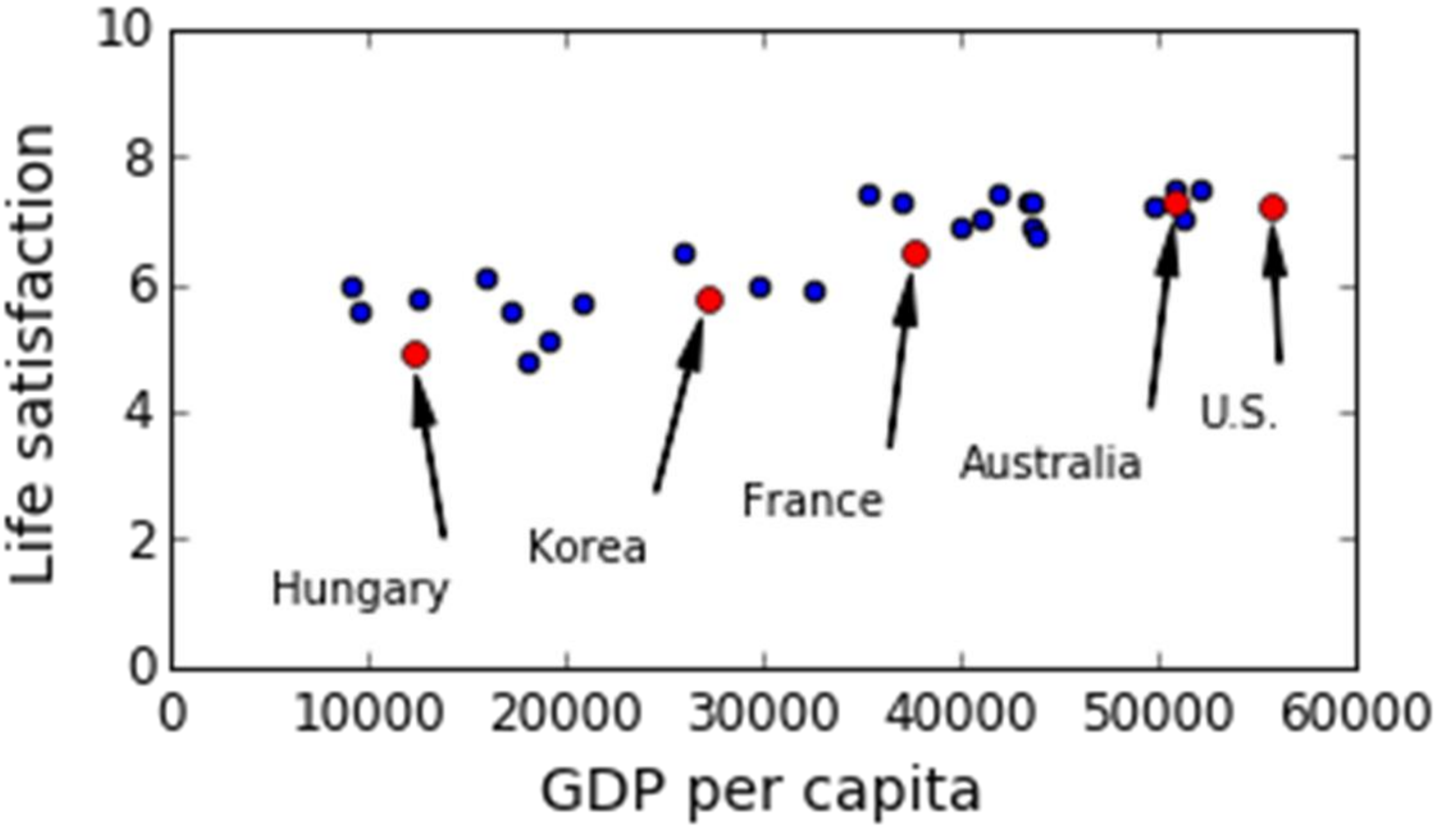
Supervised learning: what is it?

- Let us consider each image as an input variable, x
- Also, consider each label as an output variable, y
- Supervised learning is all about learning a mapping function from the input to the output

$$y = f(x)$$

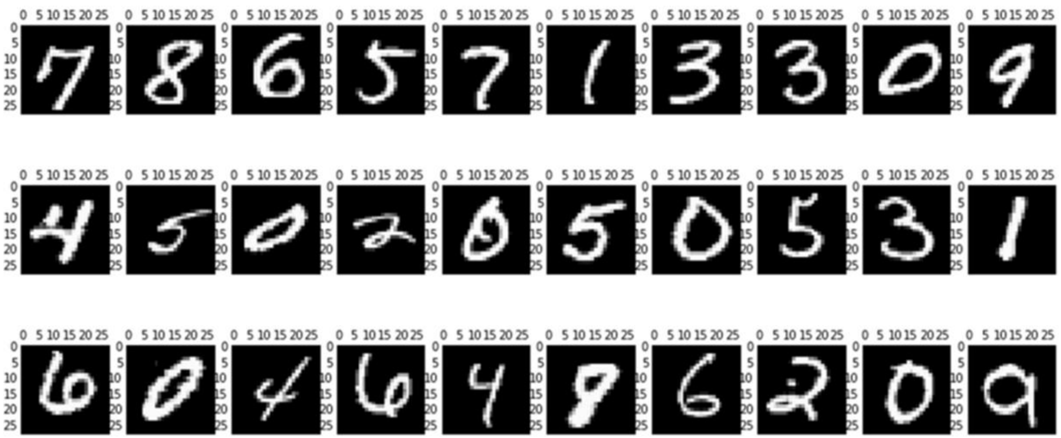
- So that, given a new image, x, your model learning model can predict y.

Supervised learning: applications



Supervised learning: applications

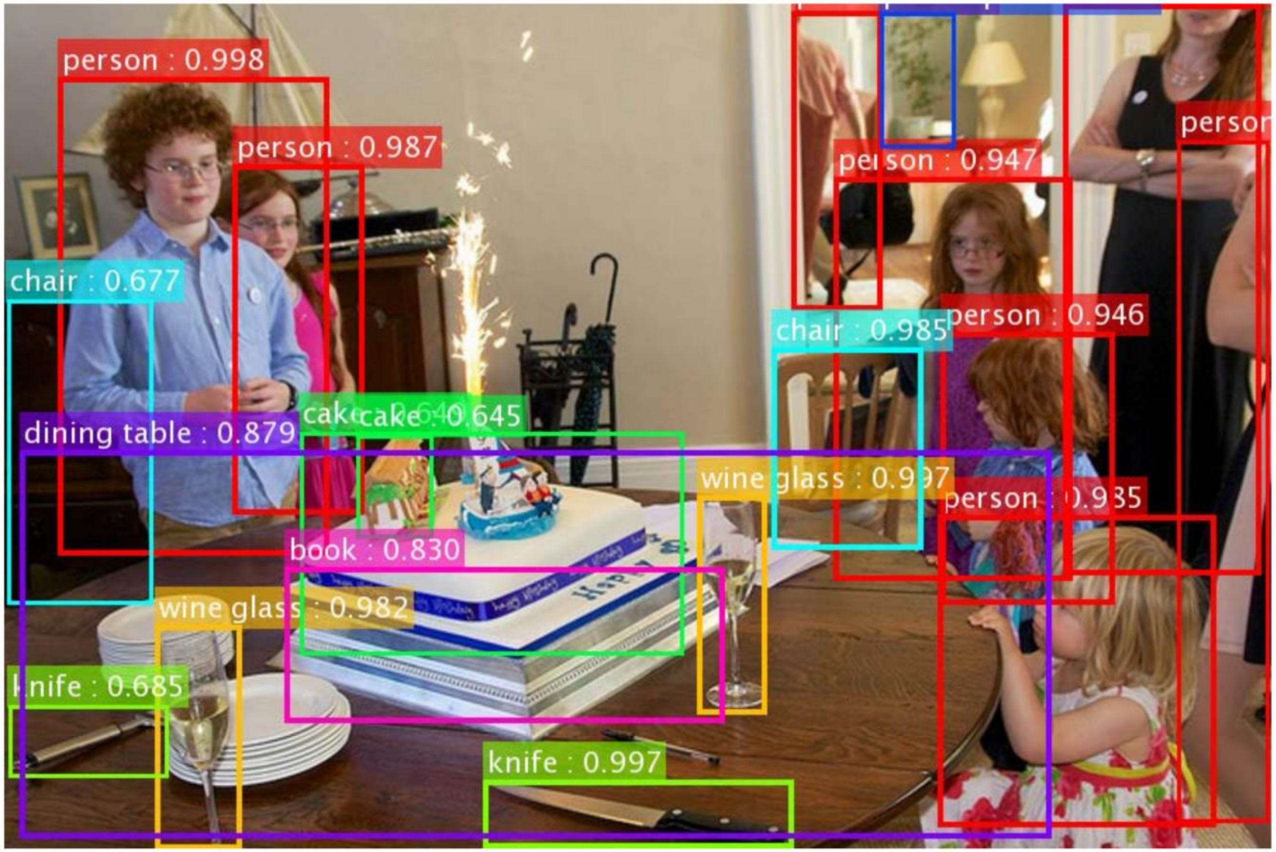
- Majority of practical machine learning products are based on supervised learning



Handwriting recognition

Speech recognition/Natural Language Processing (NLP)

Object dection



ResNet applied to COCO dataset.

Source: He et al., Deep residual learning for image recognition, CVPR, 2016 www.image-net.org/

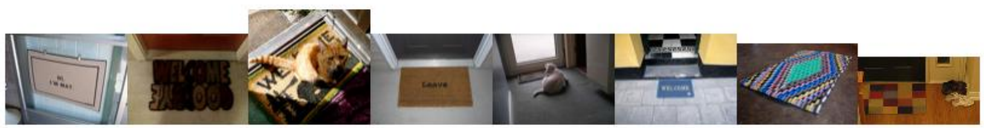
IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications CoolStuff About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.



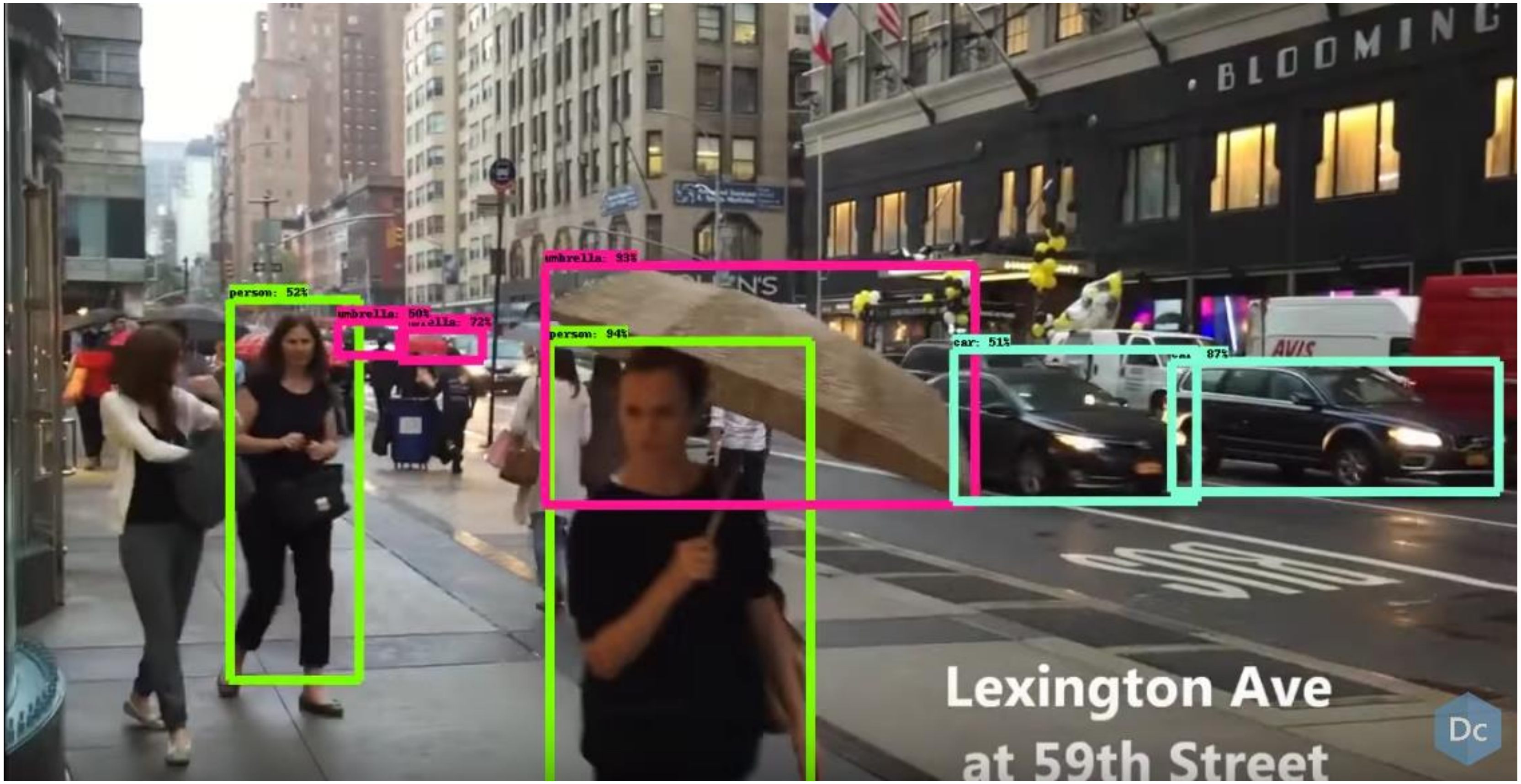
ImageNet

From Wikipedia, the free encyclopedia

The **ImageNet** project is a large visual database designed for use in visual object recognition software research. More than 14 million[1][2] images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided.[3] ImageNet contains more than 20,000 categories[2] with a typical category, such as "balloon" or "strawberry", consisting of several hundred images.[4] The database of annotations of third-party image URLs is freely available directly from ImageNet, though the actual images are not owned by ImageNet.[5] Since 2010, the ImageNet project runs an annual software contest, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where software programs compete to correctly classify and detect objects and scenes. The challenge uses a "trimmed" list of one thousand non-overlapping classes.[6]

The 2012 breakthrough in solving the ImageNet Challenge by AlexNet is often considered to be the beginning of the deep learning revolution of the 2010s. According to The Economist, "Suddenly people started to pay attention, not just within the AI community but across the technology industry as a whole."[4][7][8]

Real time object detection

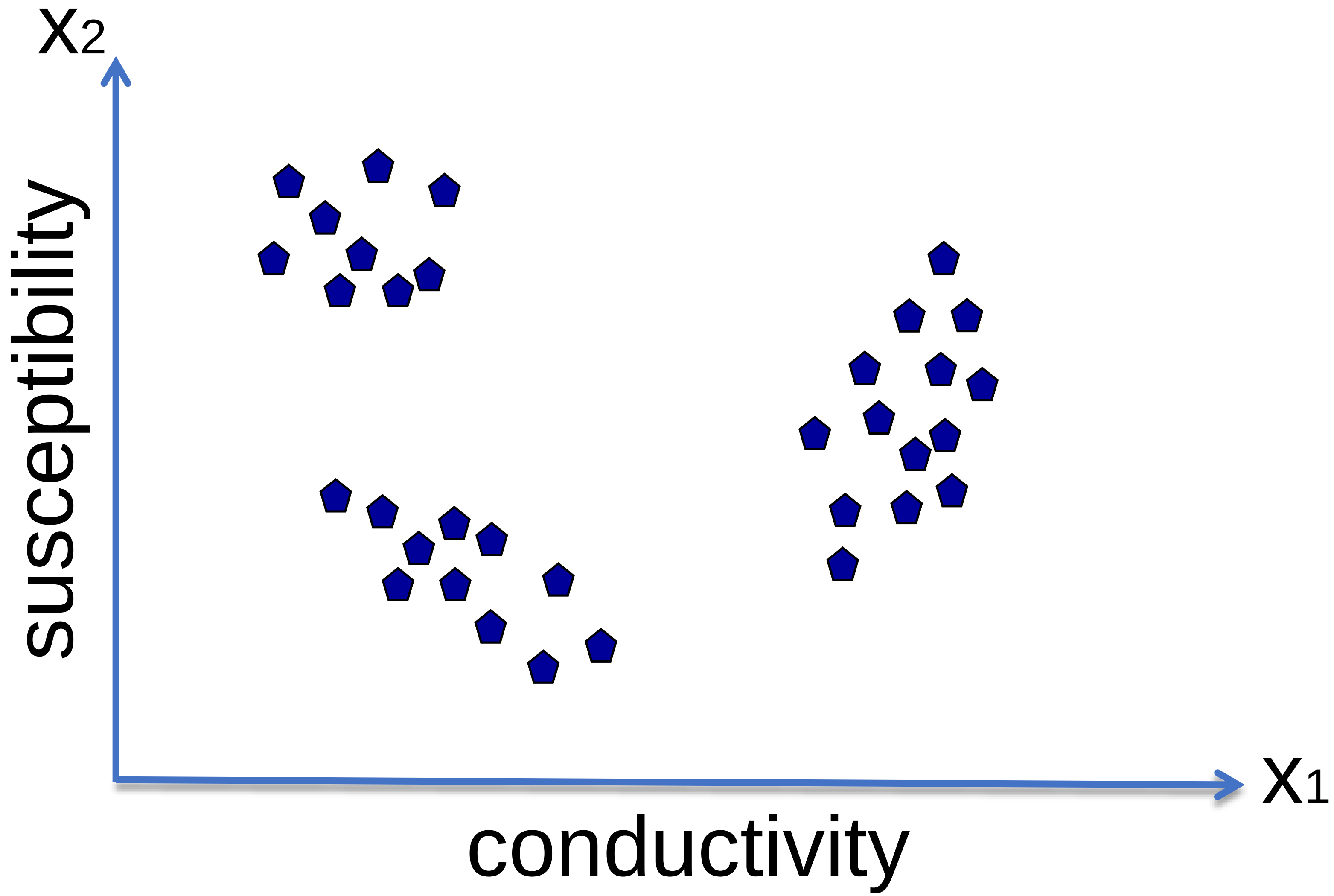


Video online: https://www.youtube.com/watch?v=_zZe27JYi8Y

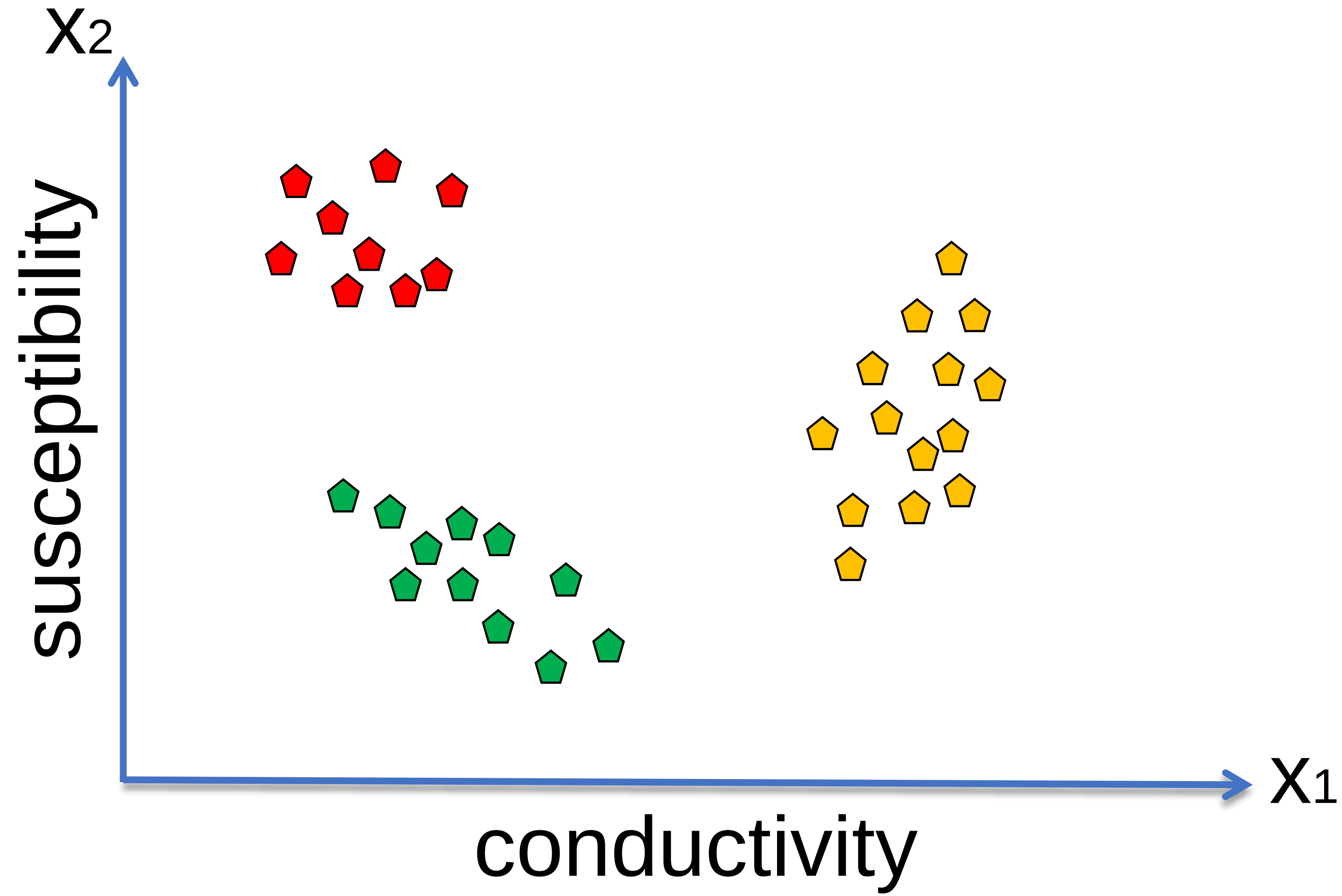
Unsupervised learning

- Training data does not have any label
- Unlabeled data
- The goal is to discover the intrinsic, and often complicated structures among data for better decision-making.
- These structures are not obvious to humans.
- No answer available. Algorithms are left to their own to discover the interesting structures in data.

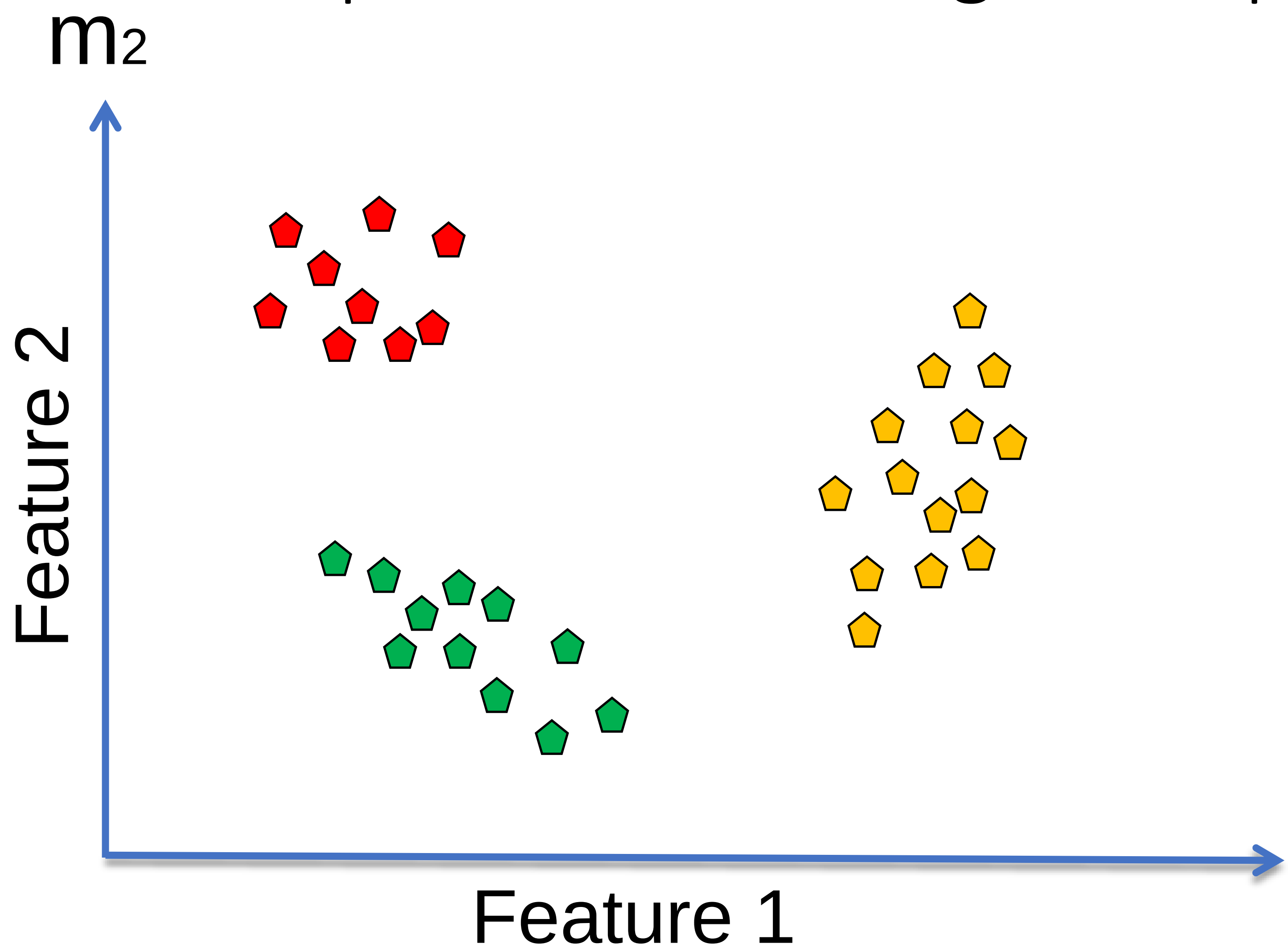
Unsupervised learning: example



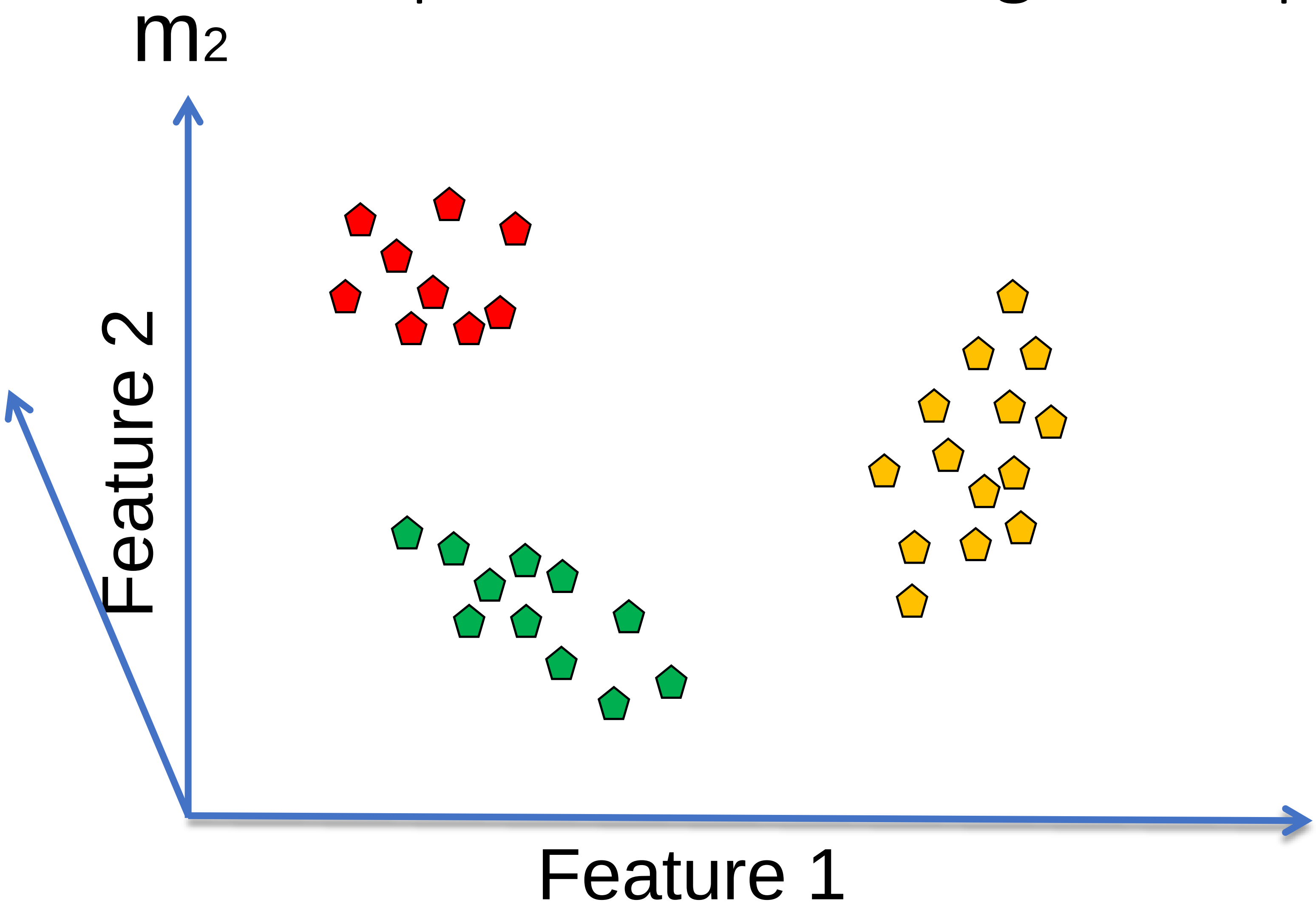
Unsupervised learning: example



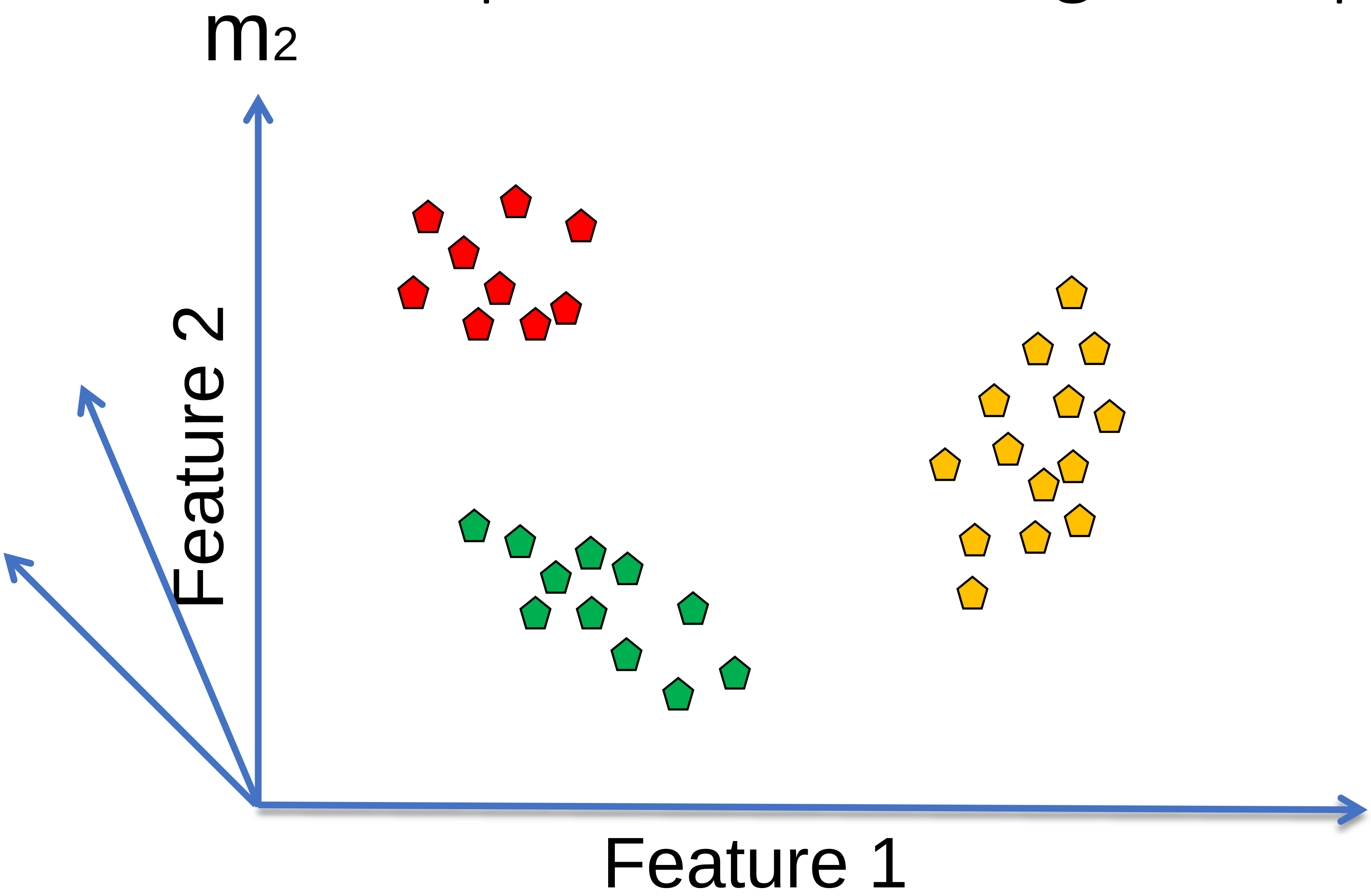
Unsupervised learning: example



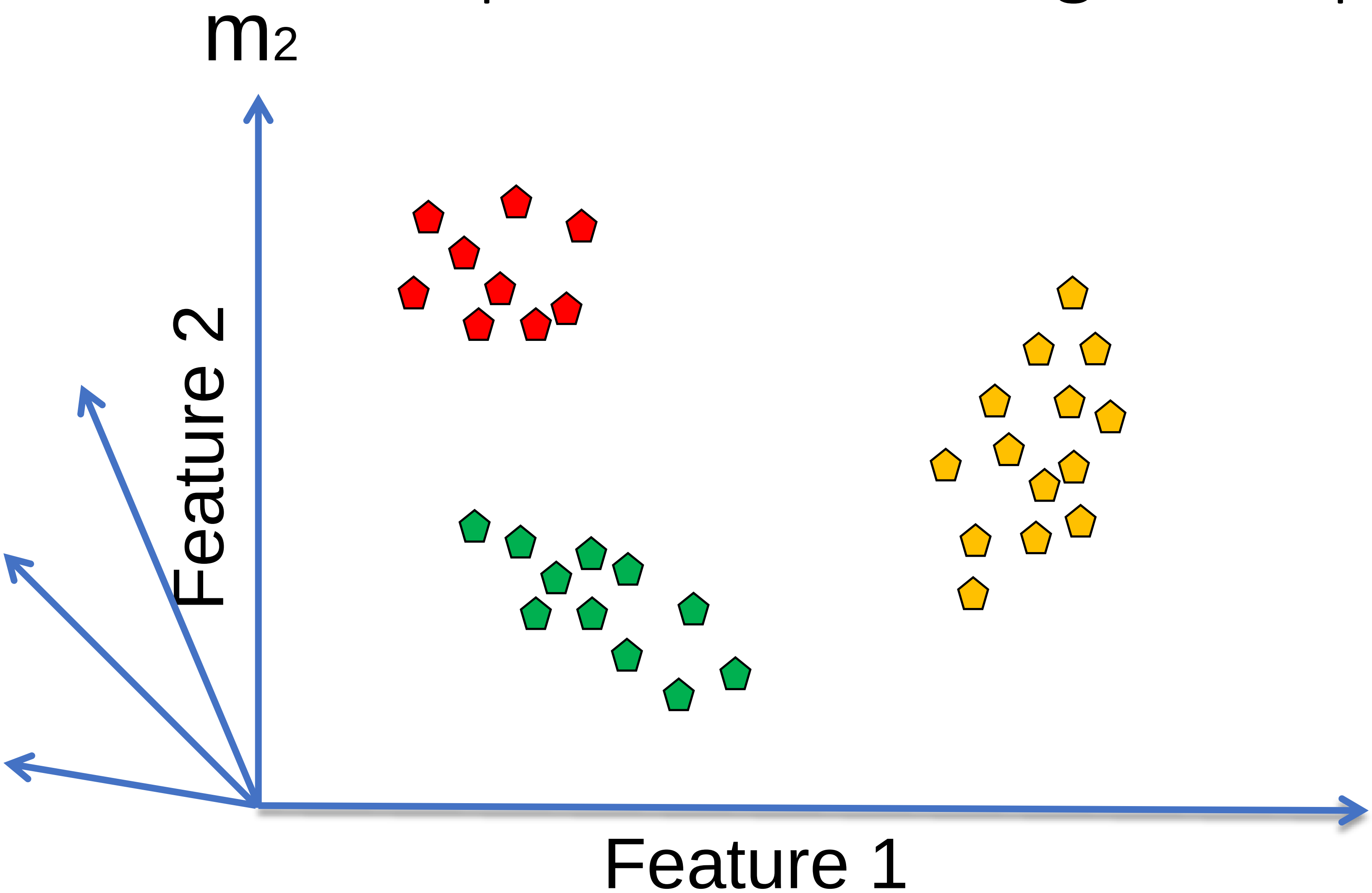
Unsupervised learning: example



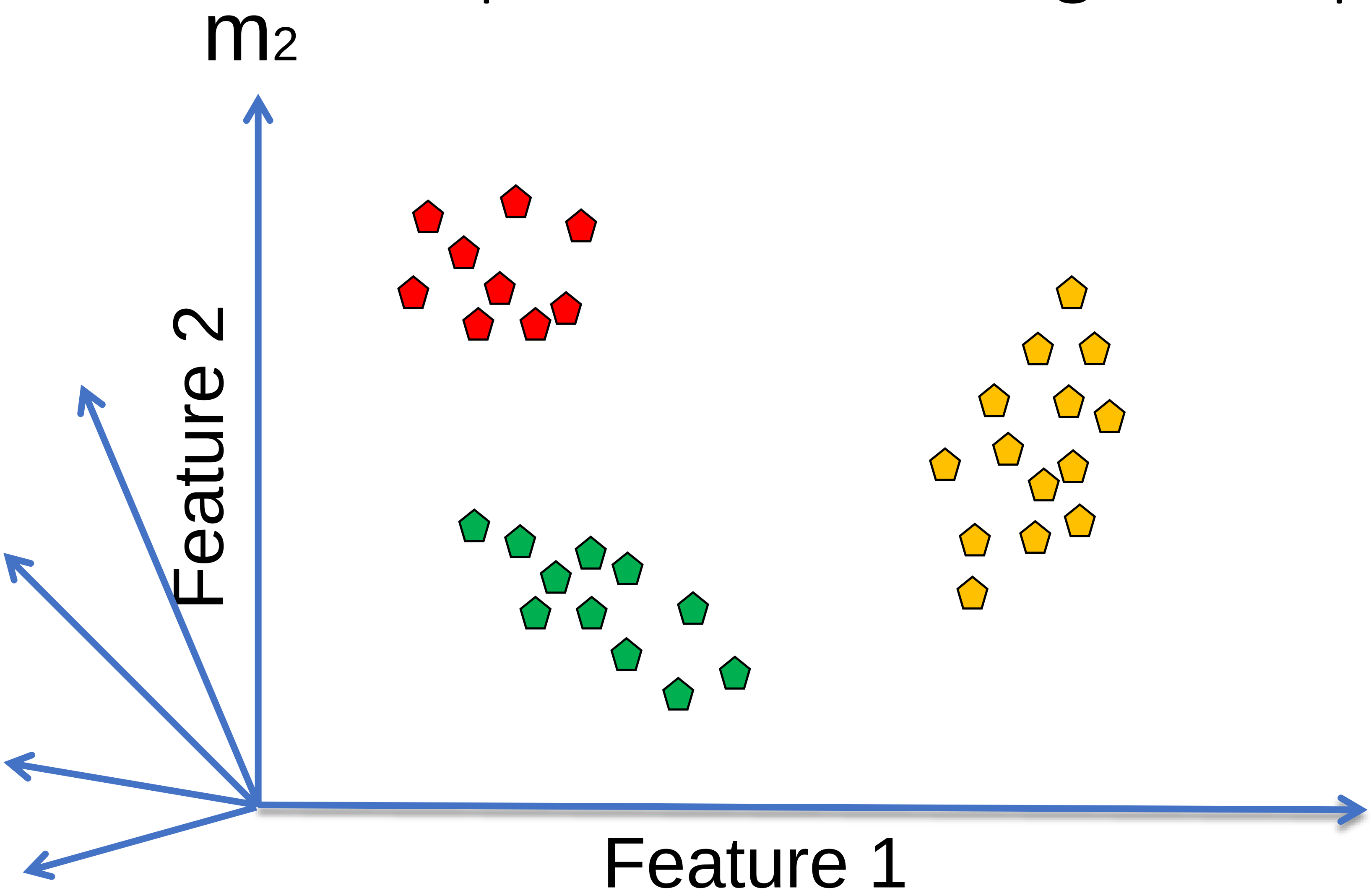
Unsupervised learning: example



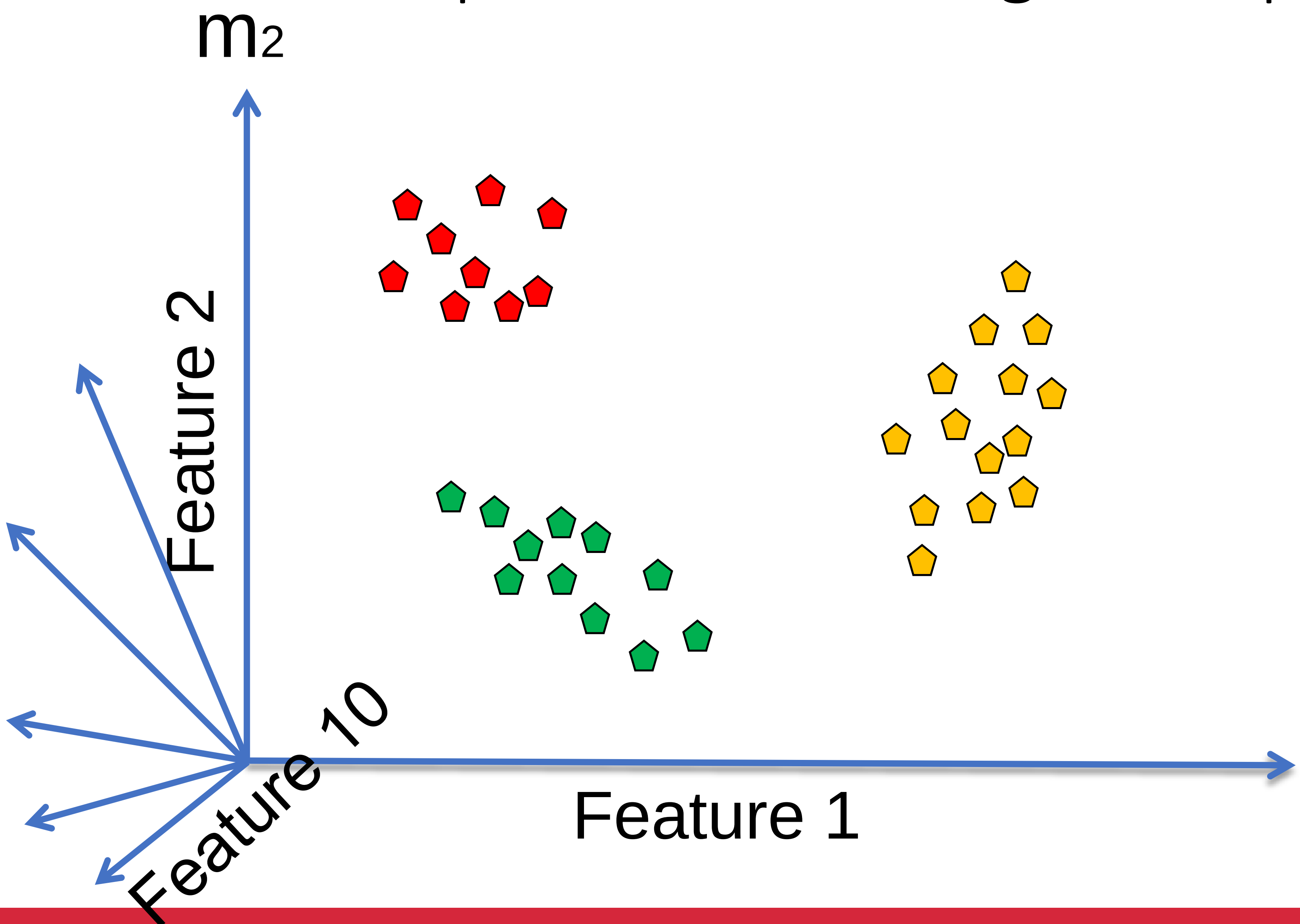
Unsupervised learning: example



Unsupervised learning: example



Unsupervised learning: example



Geochemical facies analysis

- Data: XRF (X-ray fluoresence) measurements of cutting from the lateral section of an unconventional well
- Measurements made at 10 m interval
- 22 measurements at each location
- Each measurement is the weight percentage of a chemical component
- 269 locations

Geochemical facies analysis

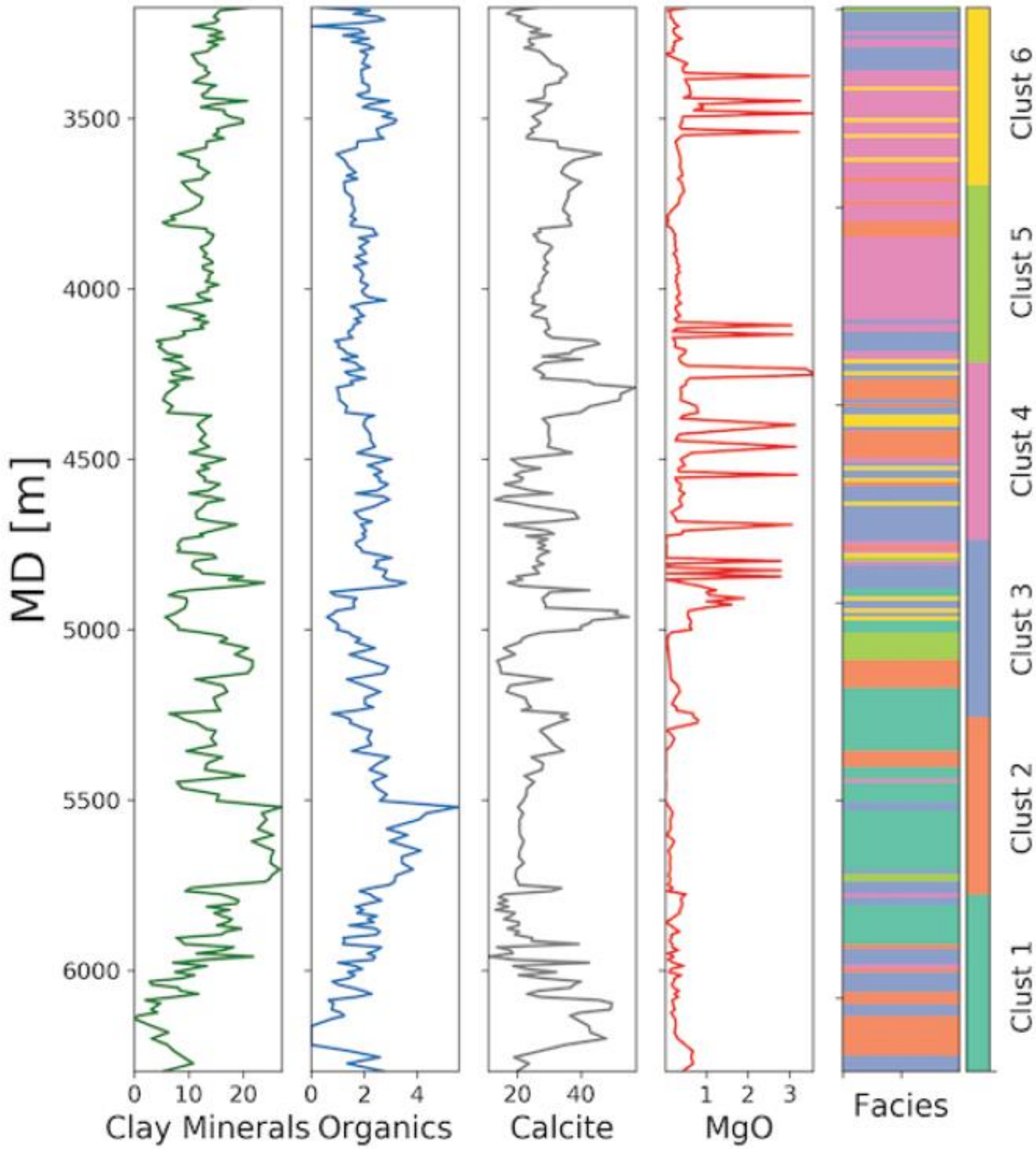
- Data: XRF (X-ray fluoresence) measurements of

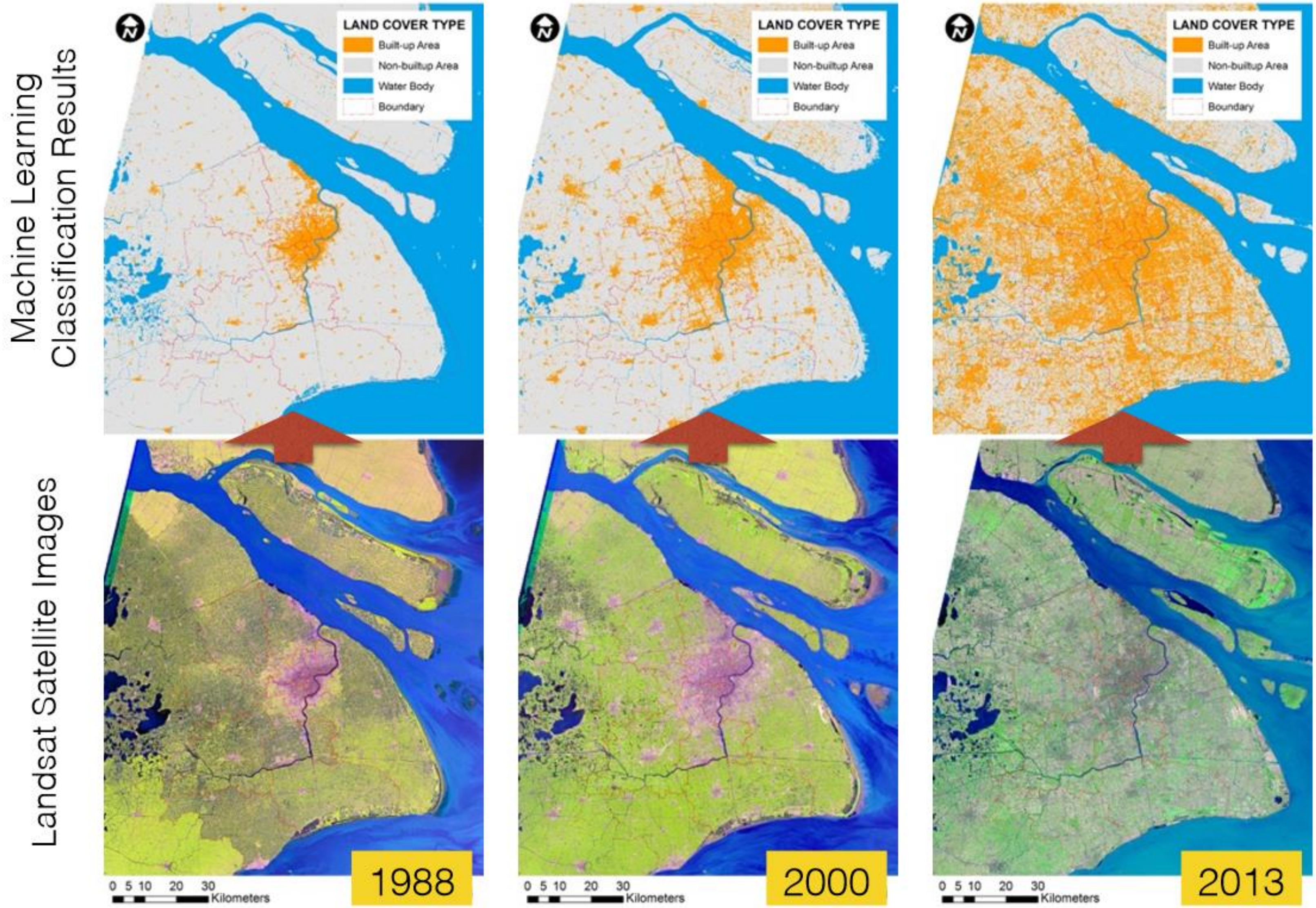
	Well Name	Depth	Quartz	...	SO3	Cl	Zr
0	Well 1	3173.97	27.56	...	1.20	0.28	201.70
1	Well 1	3183.11	42.92	...	0.81	0.26	395.35
2	Well 1	3192.26	44.55	...	0.76	0.23	362.70
...	...	...	...	...	...	...	...
266	Well 1	6255.50	45.04	...	0.97	0.02	337.37
267	Well 1	6273.78	41.21	...	1.05	0.02	356.98
268	Well 1	6296.64	46.72	...	0.77	0.02	360.96

Geo

- Data: X of

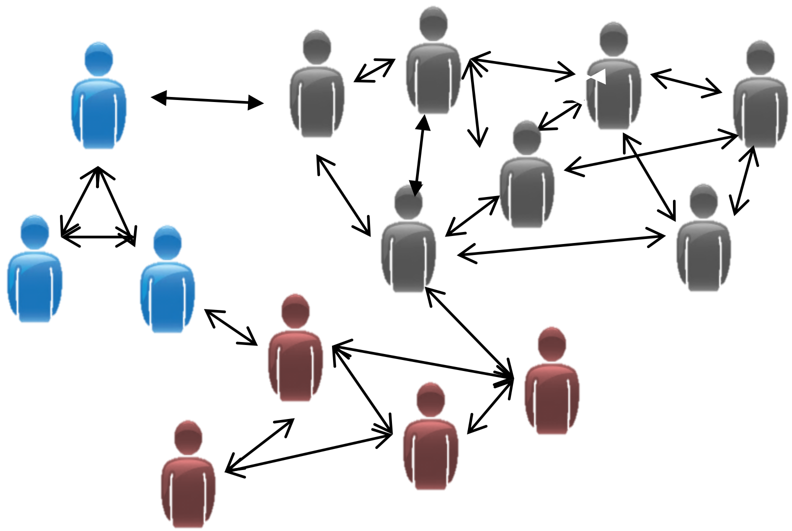
	Well	Cl	Zr
0	Well	0.28	201.70
1	Well	0.26	395.35
2	Well	0.23	362.70
...	...	...	...
266	Well	0.02	337.37
267	Well	0.02	356.98
268	Well	0.02	360.96



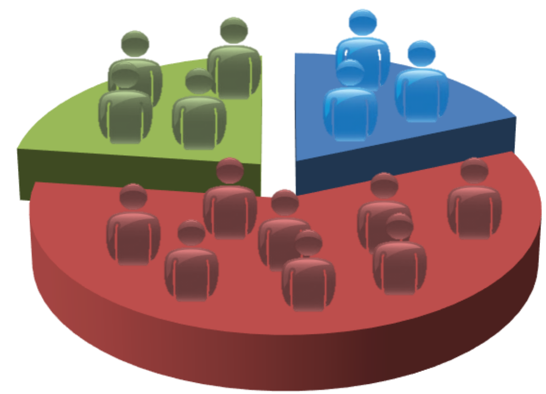


http://giswin.geo.tsukuba.ac.jp/sis/tutorial/Machine_learning%20_in_geoscience.pdf

Unsupervised learning: applications



Social network analysis



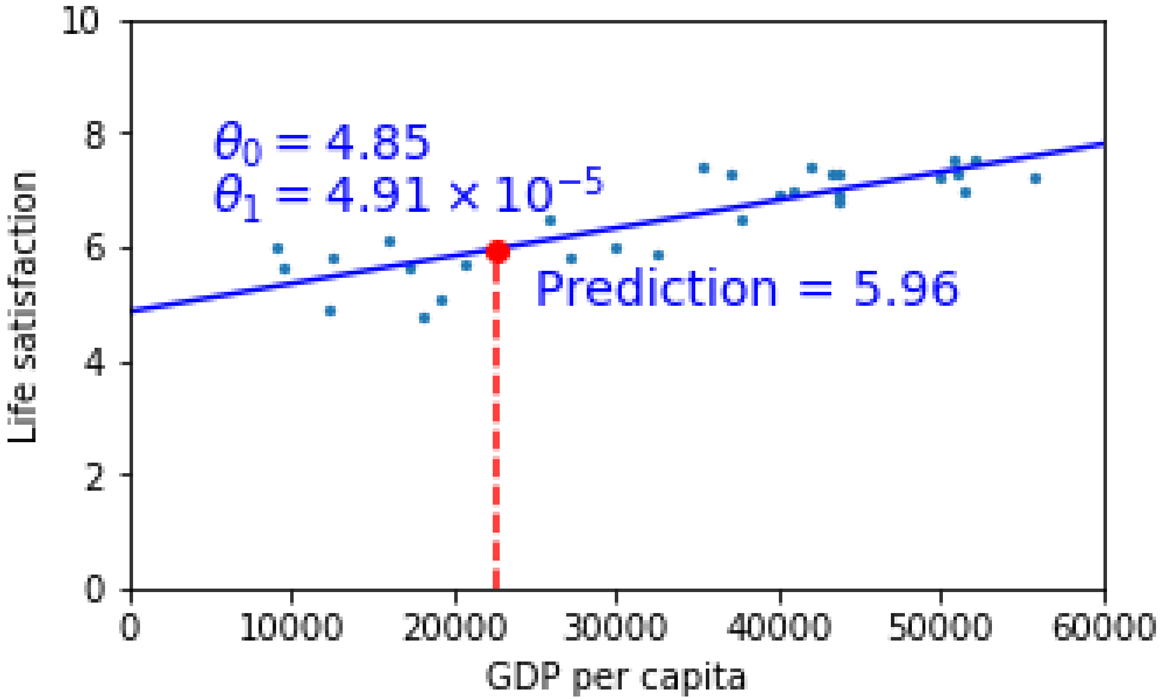
Market segmentation

Today’s agenda

- Supervised vs. unsupervised learning
- Regression vs. classification
- Overfit vs. underfit
- Bias vs. variance

Regression

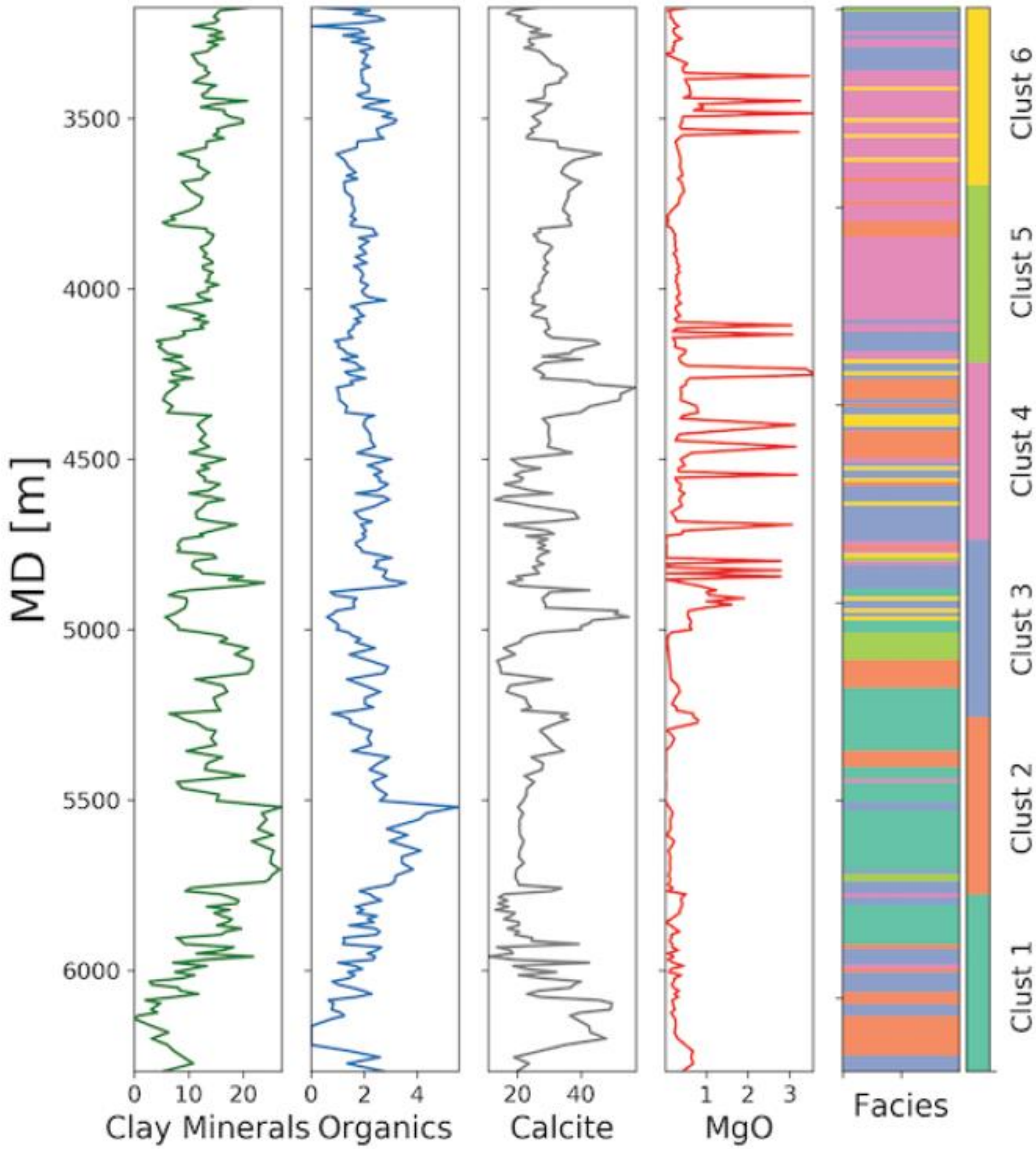
- Predict continuous numerical values, such as prices, temperatures, etc.



Classification

- Predict discrete categorical values, such as class 1, 2, 3, etc.

Clas



- Pred class 1, 2, 3,

Today’s agenda

- Supervised vs. unsupervised learning
- Regression vs. classification
- Overfit vs. underfit
- Bias vs. variance

Overfit: example

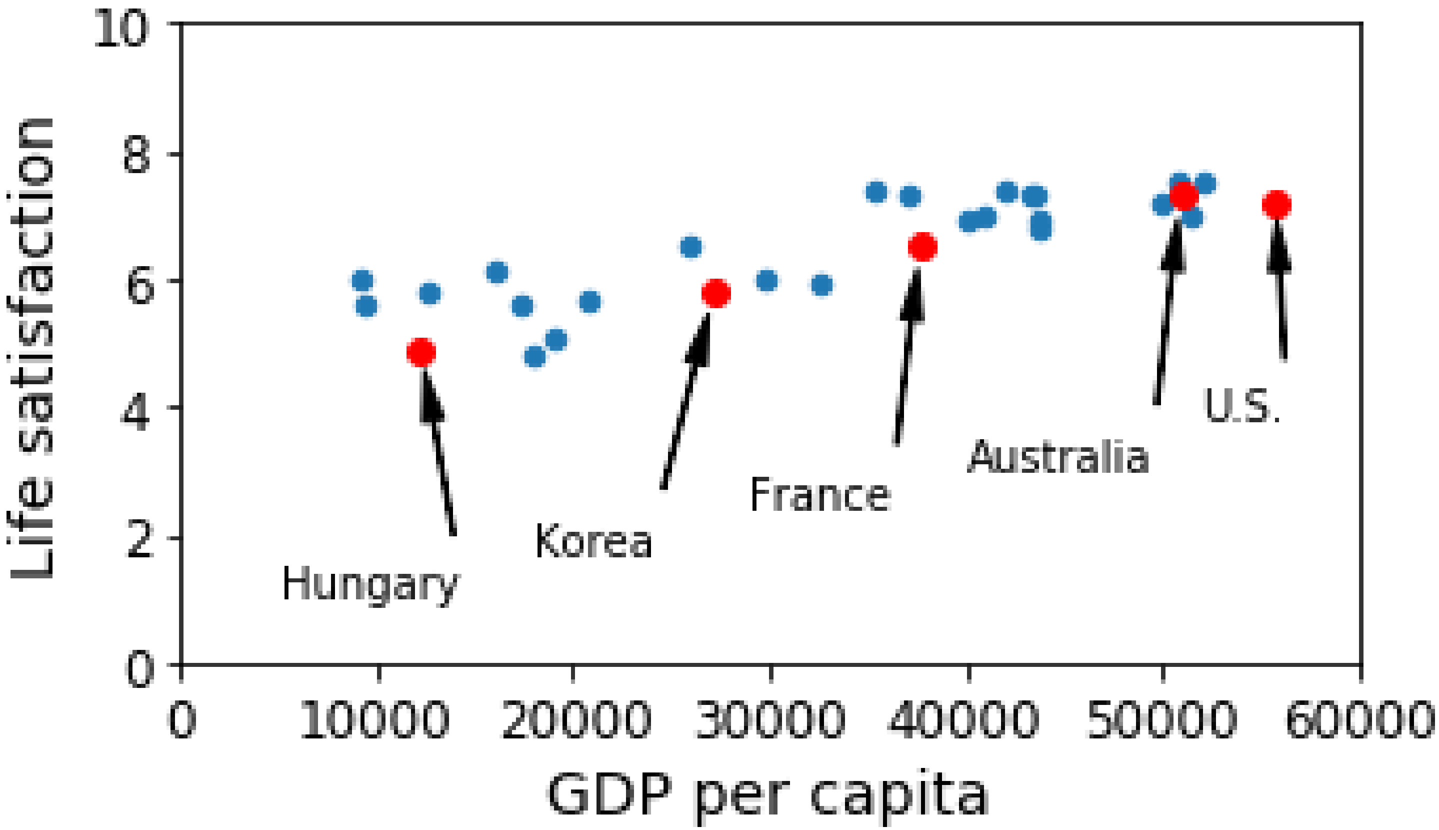


Figure from Aurelien Geron's ML book, page 19

Good fit

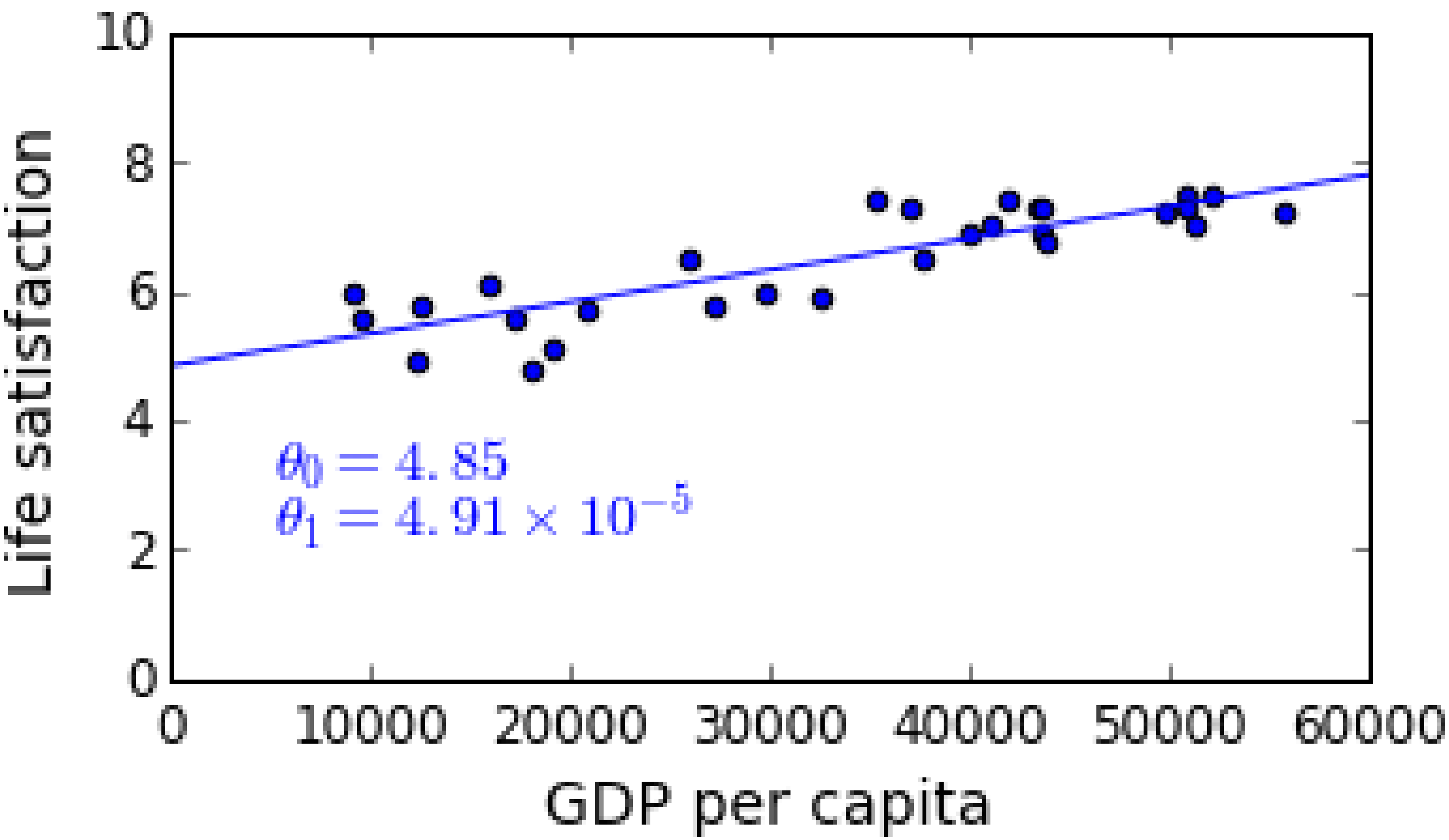


Figure from Aurelien Geron's ML book, page 20

Overfit (polynomial degree = 60)

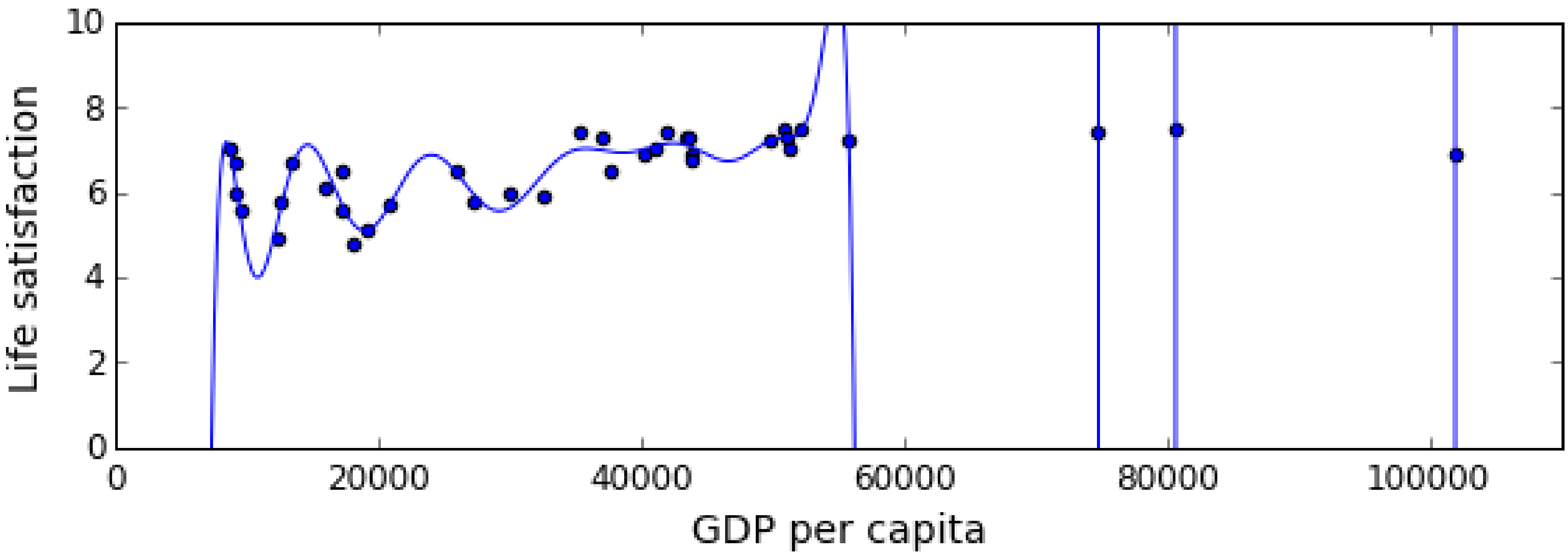


Figure from Aurelien Geron's ML book, page 26

Overfit

- Fit the training data very well (actually too well)
- But, does not generalize well to new data.
- That is, predictions on new data will be bad!

Overfit

- Fit the training data very well (actually too well)
- But, does not generalize well to new data.
- That is, predictions on new data will be bad!
- Remember, the whole purpose of machine learning is to make predications.
- If a machine learning model only works well on training data, but not on new (i.e., unseen) data, it is NOT a good model/product.

How to tell if you are overfitting?

- Split your training data into three parts:

1. Training set
2. Validation set
3. Test set

How to tell if you are overfitting?

- Split your training data into three parts:

1. Training set
2. Validation set
3. Test set

- Use only training set for training (put the other two sets of data aside), calculate the prediction error J_{train}
- After training, apply the learned model to cross-validation set, calculate the prediction error J_v
- If J_{train} is very small, J_v is large, you overfit your data!

$$J_{train} = \frac{1}{2m_{train}} \sum_{i=1}^{m_{train}} (h_\theta\,(x_{train}^{(i)}) - y_{train}^{(i)})^2$$

$$J^v = \frac{1}{2m^{cv}} \sum_{i=1}^{m^{cv}} (h_\theta\,(x^{(i)} - y^{(i)}))^2$$

Remedy for overfitting

- Overfitting happens when your ML model is overly complex
- Therefore, possible solutions are:
  1. Collect more training data
  2. Reduce data noise
  3. Simplify model
     - using linear model rather than a high-degree polynomial model
     - using regularization
     - …

Underfit

- The opposite of overfitting
- Your model is too simple to capture the important information/structures/relations in the data.

Underfit: example

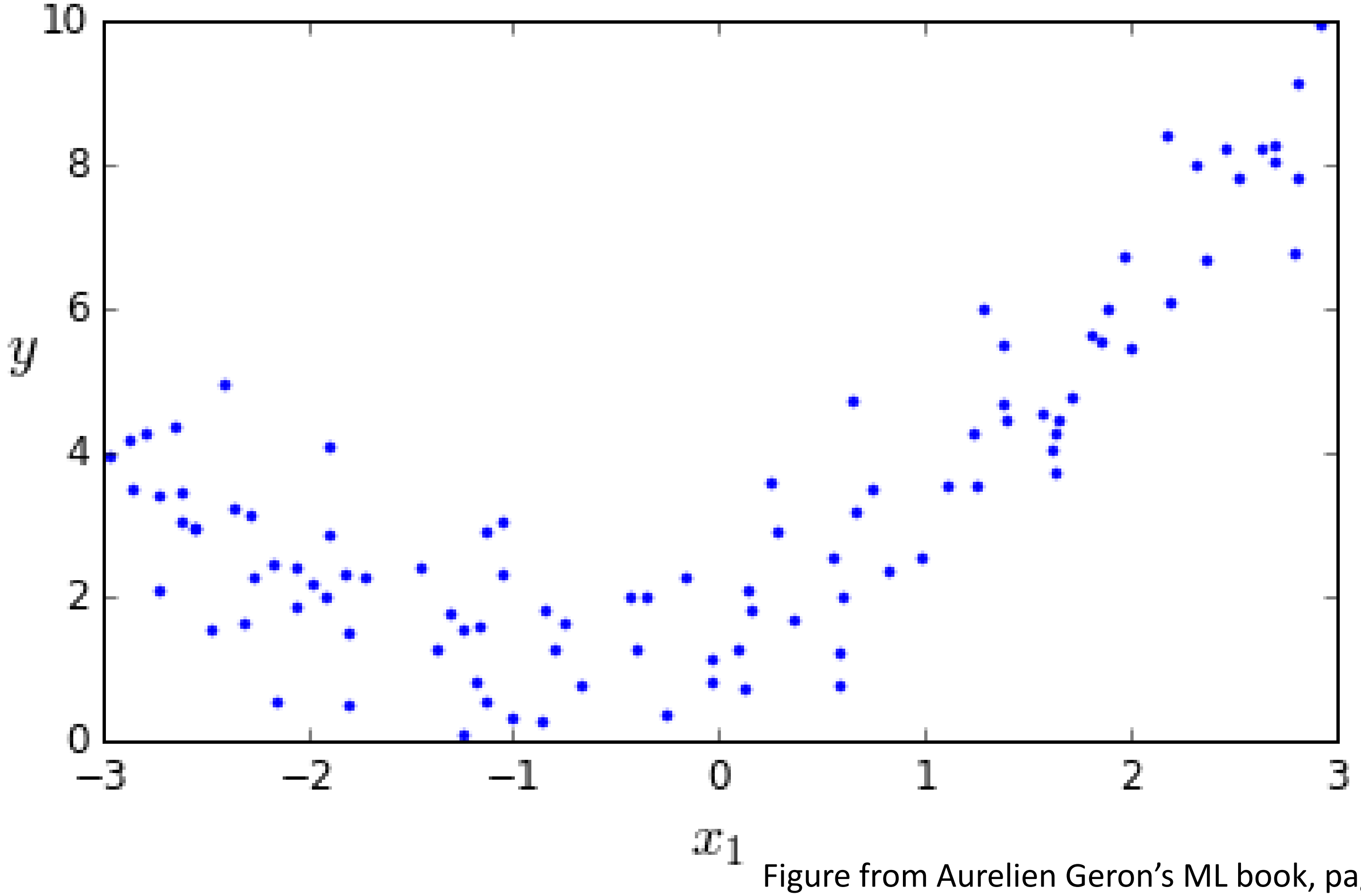
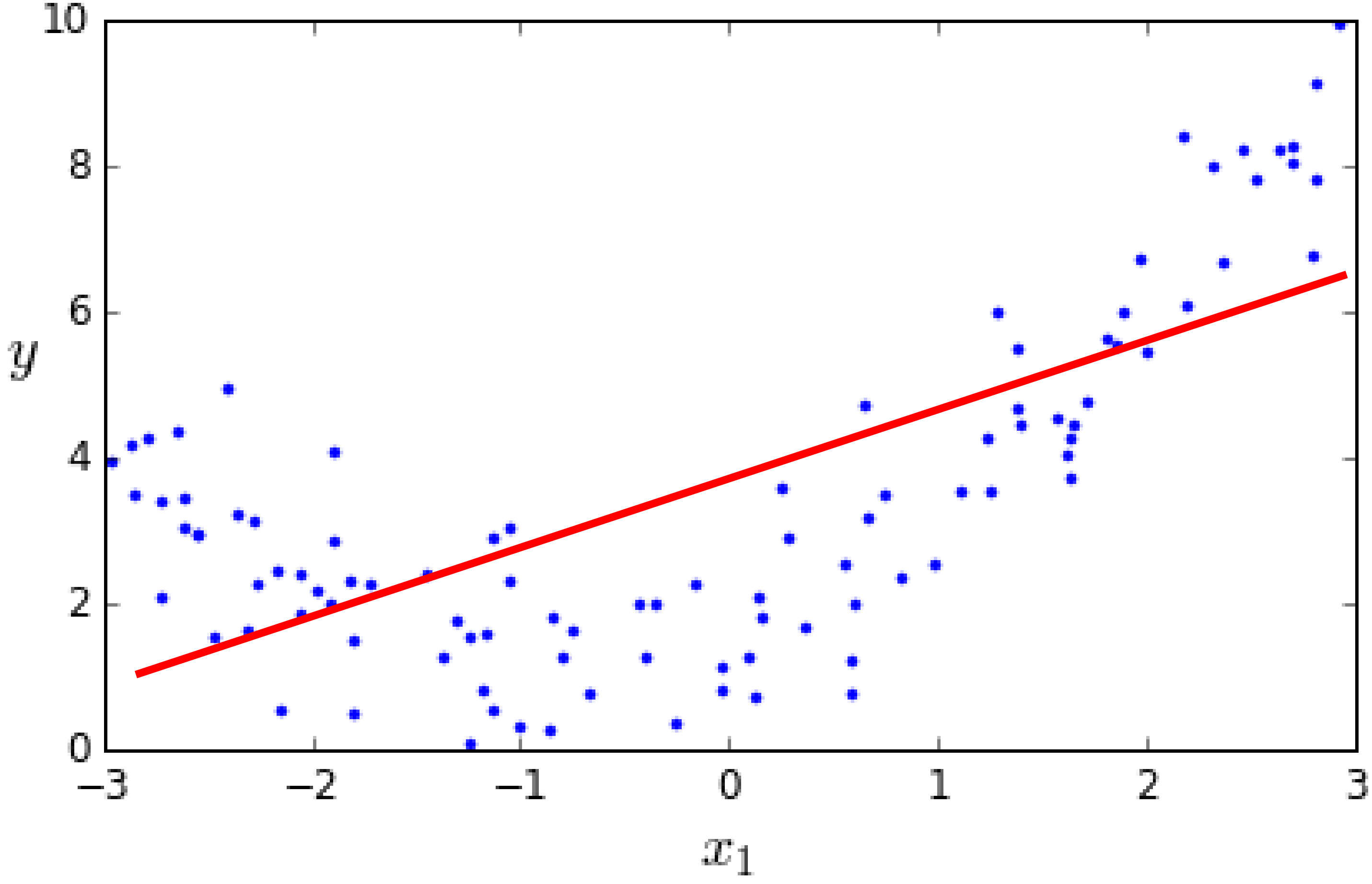


Figure from Aurelien Geron's ML book, page 121

Underfit: example



Overfit vs. underfit

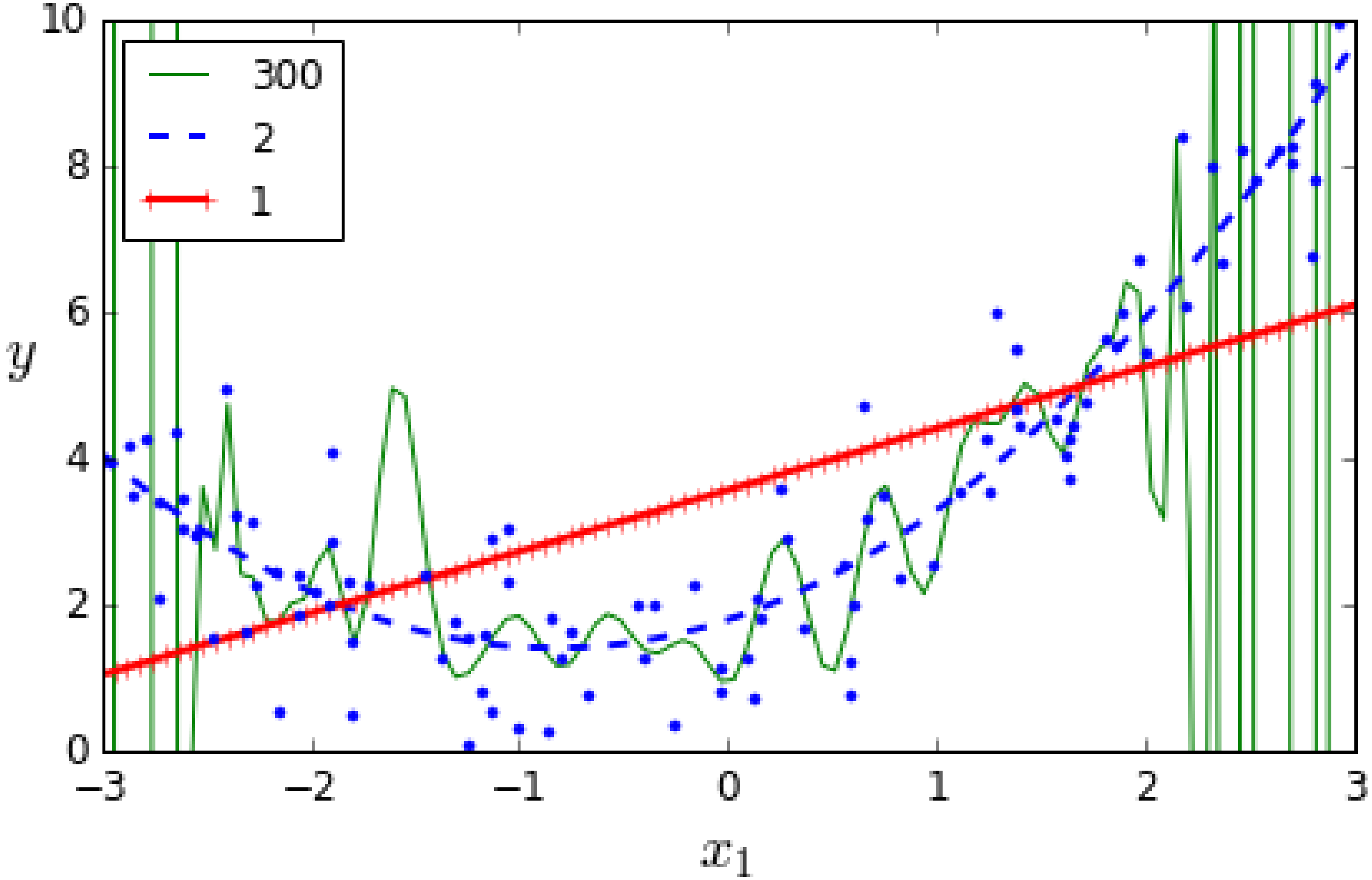


Figure from Aurelien Geron's ML book, page 123

How to tell if you are underfitting?

- Split your training data into three parts:

1. Training set
2. Validation set
3. Test set

How to tell if you are overfitting?

- Split your training data into three parts:

1. Training set
2. Validation set
3. Test set

- Use only training set for training (put the other two sets of data aside), calculate the prediction error J_{train}
- After training, apply the learned model to cross-validation set, calculate the prediction error J_v
- If J_{train} is large, J_v is large, you underfit your data!

Remedy for underfitting

- Underfitting happens when your ML model is overly simple
- Therefore, possible solutions are:
    1. ~~Collect more training data~~
    2. ~~Reduce data noise~~
    3. Make your model more complex
        - using a high-degree polynomial model rather than a linear model
        - using less regularization
        - Adding more features such as $(x_1^2, x_2^2, x_1 x_2)$ to the learning algorithm (feature engineering)

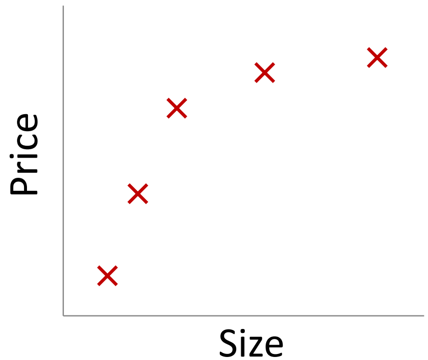
Remember,

- If you are underfitting your data, collecting more data won't help!

Today’s agenda

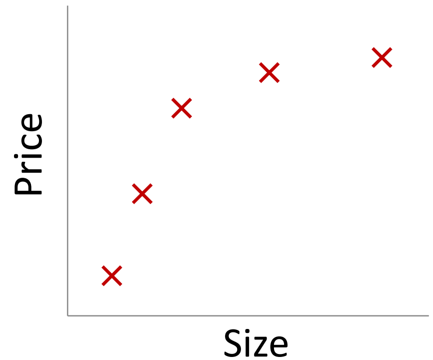
- Supervised vs. unsupervised learning
- Regression vs. classification
- Overfit vs. underfit
- Bias vs. variance

Bias/variance



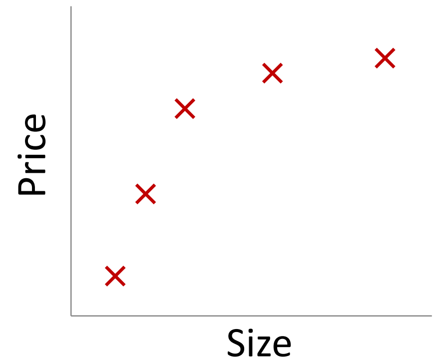
$$\theta_0 + \theta_1 x$$

High bias
(underfit)



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

"Just right"



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

High variance
(overfit)

This slide is taken from Andrew Ng's ML class on coursera

Bias vs. Variances

Bias

- Due to over-simplified assumptions
- E.g., assuming a linear model when the training data are actually from a non-linear model
- Lead to underfitting the training data

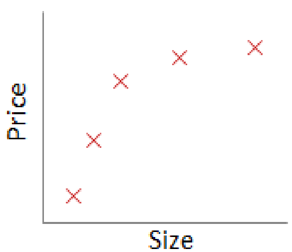
Variance

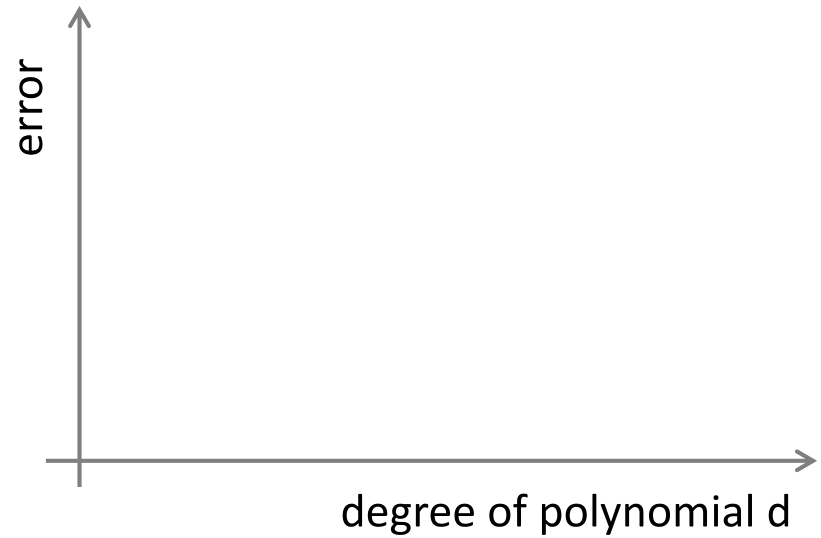
- Due to your model's excessive sensitivity to small variations in the training data
- E.g., assume a highly nonlinear model when the data are actually linear
- Lead to overfitting the data

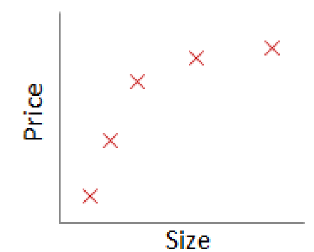
Bias/variance

Training error: $J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

Validation error: $J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$



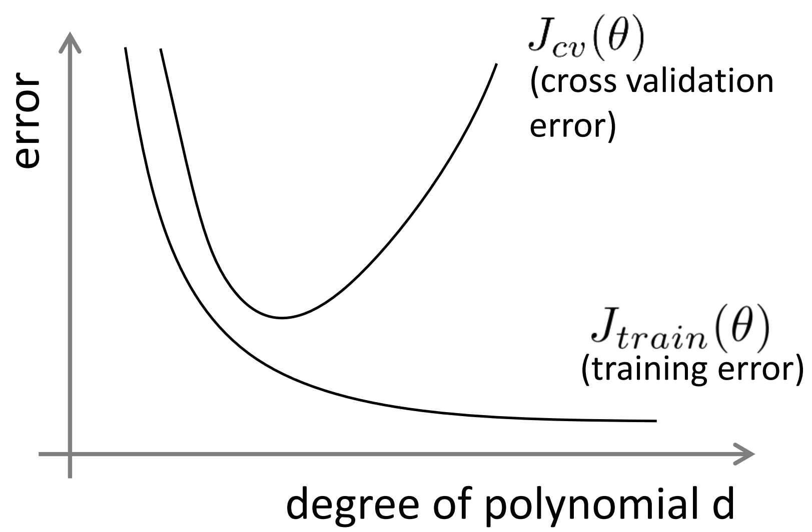




This slide is taken from Andrew Ng's ML class on coursera

Diagnosing bias vs. variance

Suppose your learning algorithm is performing less well than you were hoping. ($J_{cv}(\theta)$ or $J_{test}(\theta)$ is high.) Is it a bias problem or a variance problem?



Bias (underfit):

Variance (overfit):

This slide is taken from Andrew Ng's ML class on coursera