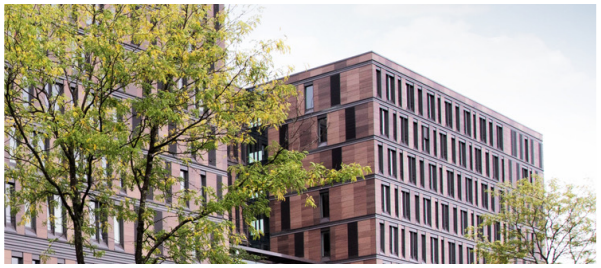
Frankfurt School
of Finance & Management

German Excellence. Global Relevance.

Class 01

Introduction to the Module



Computational Statistics

Prof. Dr. Jens Witkowski

Agenda

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

- Session 1
  - Catching up
  - Module overview and Organization
  - Forecasting Game

- Session 2
  - Artificial Intelligence vs Machine Learning vs Classical Statistics
  - Linear Regression re-visited

Organization

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

- **Attendance and Participation**
  - Mutual respect: risk-free learning environment.
  - No explicit participation grading
  - Bring to class:
    - Your name tags.
    - Fully charged laptops.
- **All In-Class, on Campus**
- **Office hours: by appointment**
  - Please reach out beforehand via email with your specific questions.
- **Communication**
  - Check Canvas regularly for course materials, announcements, and other updates.
  - Email is the preferred way to get in touch with me: j.witkowski@fs.de

Respectful Conduct

- Arrive on time. This includes coming back from the break.
- If you will arrive late: write me an email beforehand.
- If you will have to leave early: write me an email or let me know at beginning of class.
- In interest of learning experience: turn off your cell phones and use laptops only as instructed.
- If you have to use the bathroom, leave and enter the room quietly.

General guidance: Think of each lecture as a business meeting.

High-Level Course Outline

- Part 1: Introduction to Supervised Machine Learning
- Part 2: Crowd Forecasting and Decision Making

Both parts are about predictions!

First Part: Supervised Machine Learning

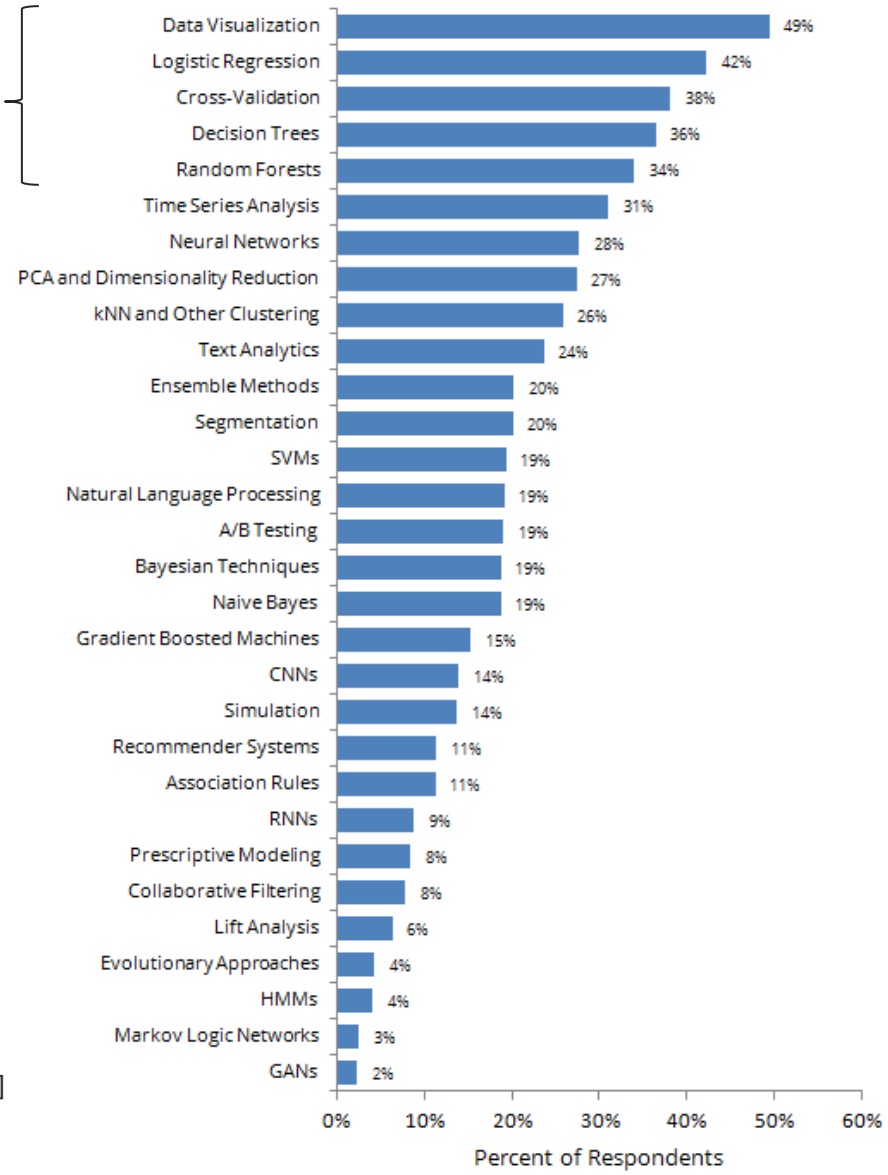
Example: Should a bank give a loan to person X?

- Depends on probability of default by X and bank's profit in cases of "default" and "no default"
- Machine learning approaches use large datasets of previous loans and their outcomes to estimate probability (make prediction) that X will default.

Data Science Methods Used at Work

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

Will get to know some of these



[Kaggle, 2018]

Second Part: Crowd Predictions

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

What is probability of armed conflict between Mainland China and Taiwan within the next 6 months?

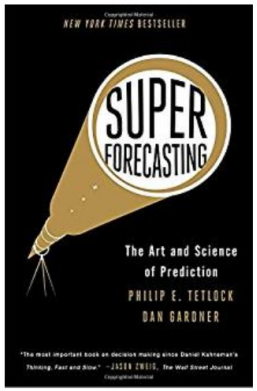
- Not enough relevant historical data for pure data-driven approach
- Crowd forecasting methods aggregate forecasts of many individuals into single probabilistic forecast
- Used by governmental agencies, financial institutions, and retailers (e.g., demand predictions for particular style or color of clothing)

Most large-scale, strategic decisions are *supported* by data but there are too few similar situations with known outcomes ("labels") that pure machine learning approaches are practical!

Crowd Forecasting

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

Wisdom of the Crowd: Eliciting and aggregating many people's beliefs (expert or non-expert).

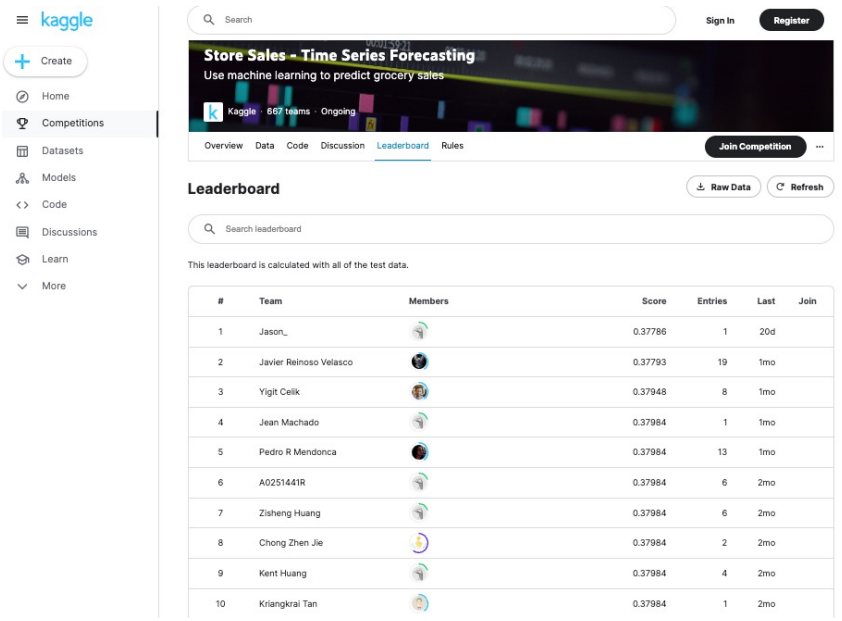


Superforecasting: The Art and Science of Prediction
by Philip E. Tetlock
Paperback
$11.55 ~~$17.00~~ Prime
You save $5.45 (32%)
Get it by **Tomorrow, Apr 4**
66 offers from $7.00
★★★★½ (256)
4 Other Formats: Kindle Edition, Hardcover, +2 more

Some questions we will address:

- How to identify good forecasters in a crowd?
- How to aggregate forecasts from different forecasters?
- How to incentivize forecasters to invest effort and report truthfully.

Data Science and Forecasting Competitions

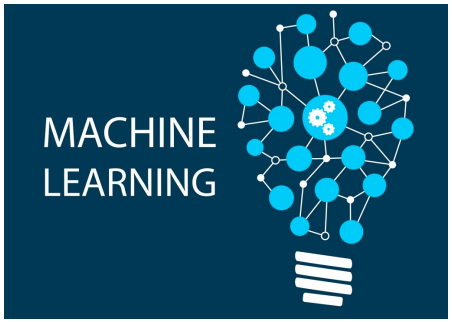


- Crowd-sourcing the Analytics: let the crowd come up with best machine learning model for you!
- Will discuss incentive issues with these tournaments and how to address those issues.

Human versus Machine Predictions

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

- With historical data for similar* situations:



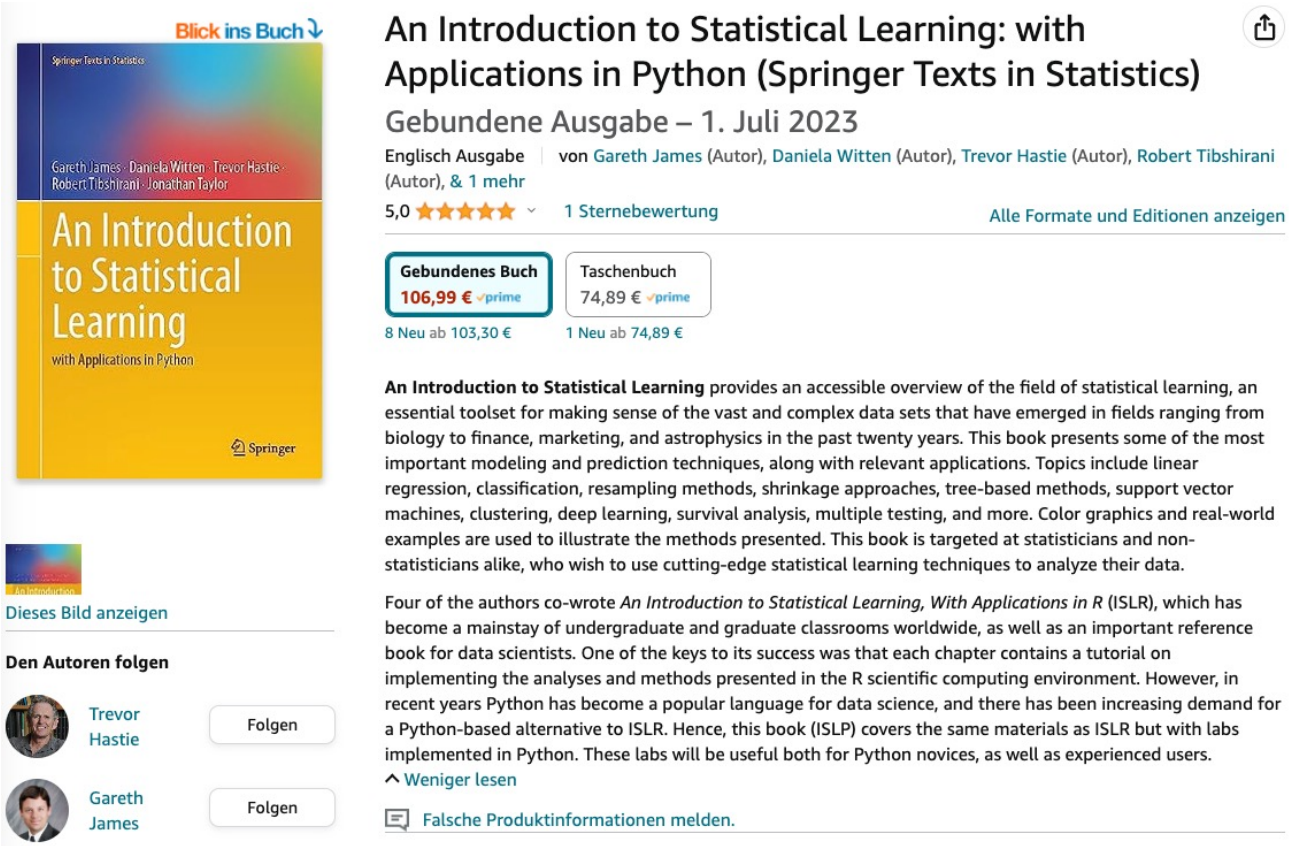
*will make this more clear

- Without appropriate historical data:

- Or, potentially better:

Crowd

Literature Pointer for First Part of this Class



Excellent text book and available as free download: https://hastie.su.domains/ISLP/ISLP_website.pdf

- Recommended for Class 01 (today): Section 2.1
- Recommended for Class 02: Sections 3.1, 3.2, 5.1, 7.1

What is Artificial Intelligence (AI)?

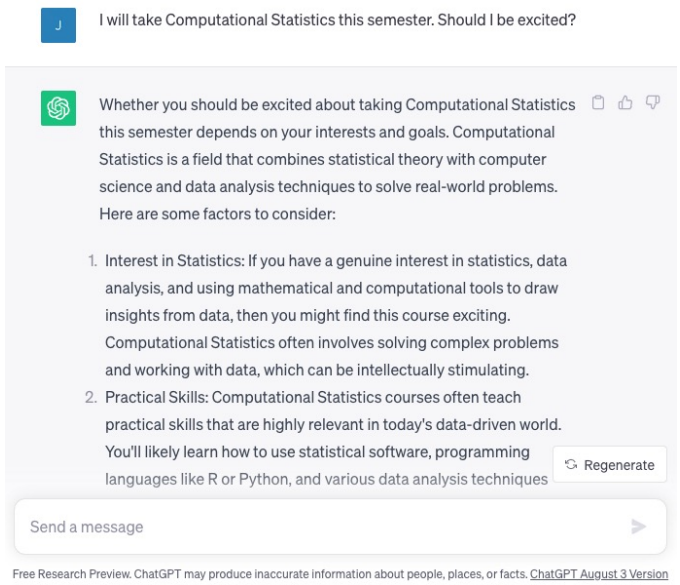
"The science and engineering of making intelligent machines"
(John McCarthy, '56)

What does this mean in reality?

What is Artificial Intelligence: Examples

AI more than just learning; in many real-world systems combined, e.g.:

- Planning and Search
- Game Theory
- Machine Learning

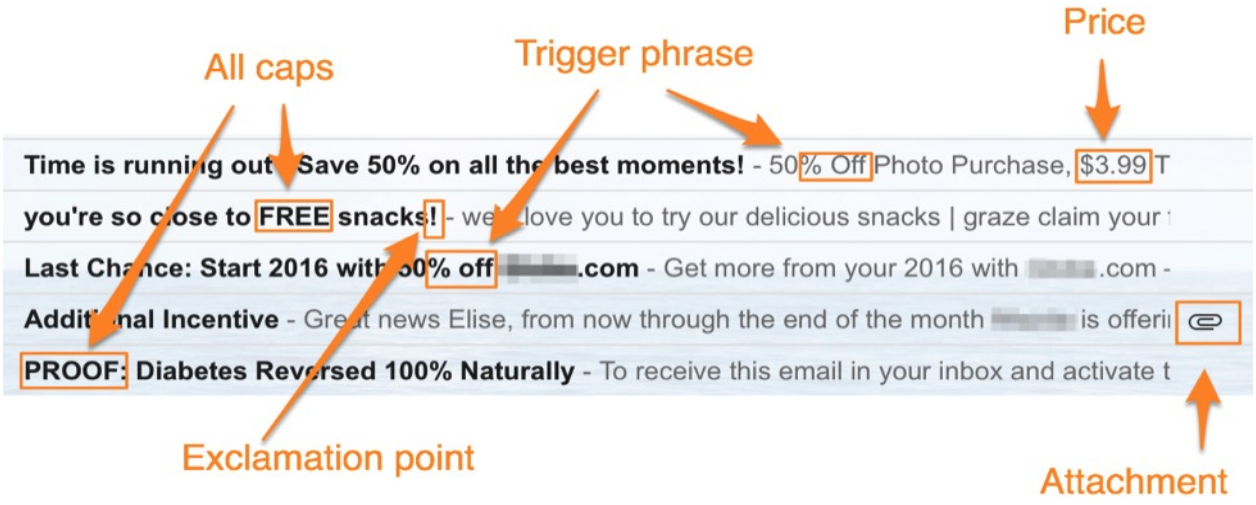


What is Machine Learning (ML)?

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed." (Arthur Samuel, 1959)

What is Machine Learning: Example

- Classify email messages as "spam" or "no spam"

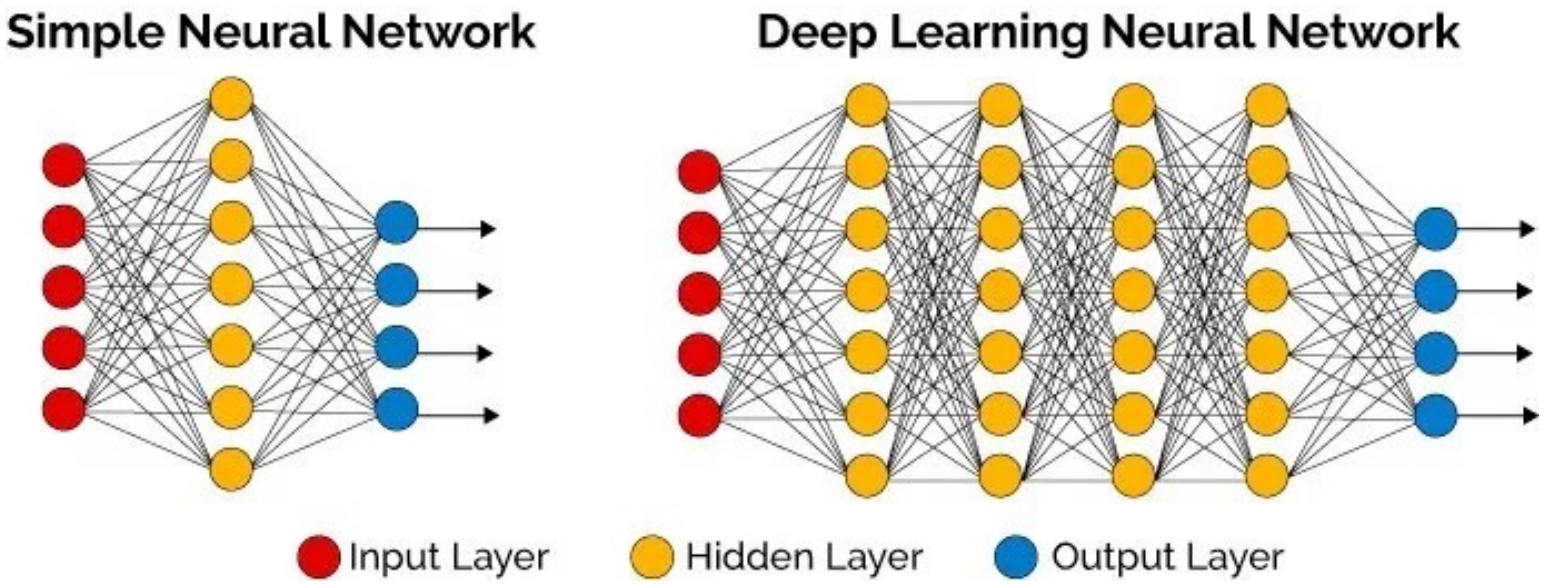


- Classical approach: manually write down rules:

  IF text contains "50% off" AND email has attachment
  THEN classify as "spam" ELSE classify as "no spam"

- Machine Learning: automatic identification of rules from training data (examples)

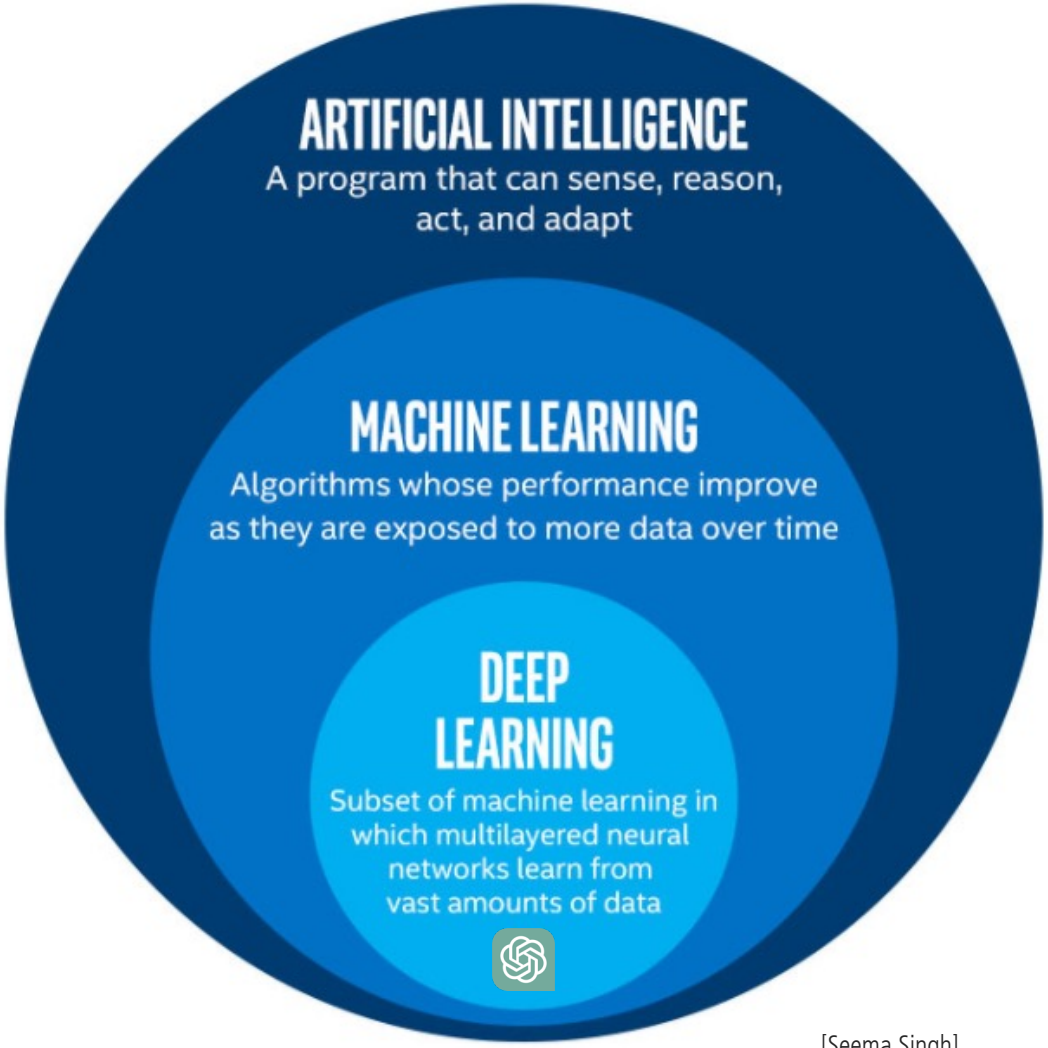
What is Deep Learning?



[Picture: https://thedatascientist.com/what-deep-learning-is-and-isnt]

- Sub-area of machine learning
- Employs neural network models, which are loosely inspired by the human brain
- "Deep" derives from using several hidden layers (until ~2010 believed to be unnecessary as single hidden layer proven to be "general" function approximator)
- Tremendous success in applications in recent years

Artificial Intelligence and Machine Learning



[Seema Singh]

Main Areas within Machine Learning

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

Supervised Learning: training data given to us contains "correct" target.

Regression: real-valued target, i.e. $y \in \mathbb{R}$

- Amount of coffee sold in a particular area
- Growth of the US economy in 2021
- A person's height

Classification: categorial target, i.e. $y \in \{1, \ldots, m\}$

- Emails: "spam" / "no spam"
- Credit card transactions: "regular" / "fraud"
- Image recognition: "dog" / "person" / "house"

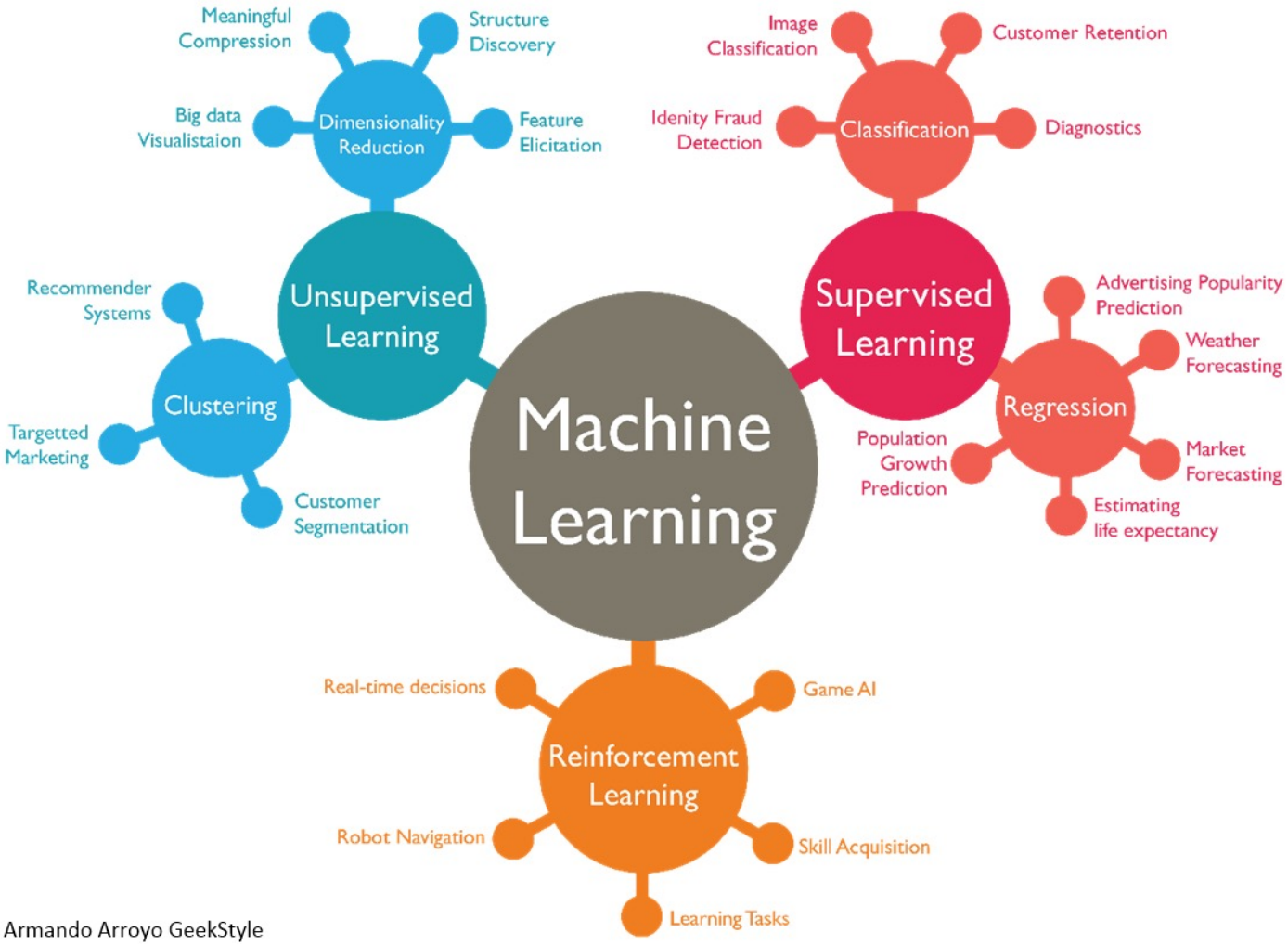
Unsupervised Learning: training data contains no targets.

Dimensionality reduction

Clustering: group "similar" data points, e.g., google photos grouping similar pictures

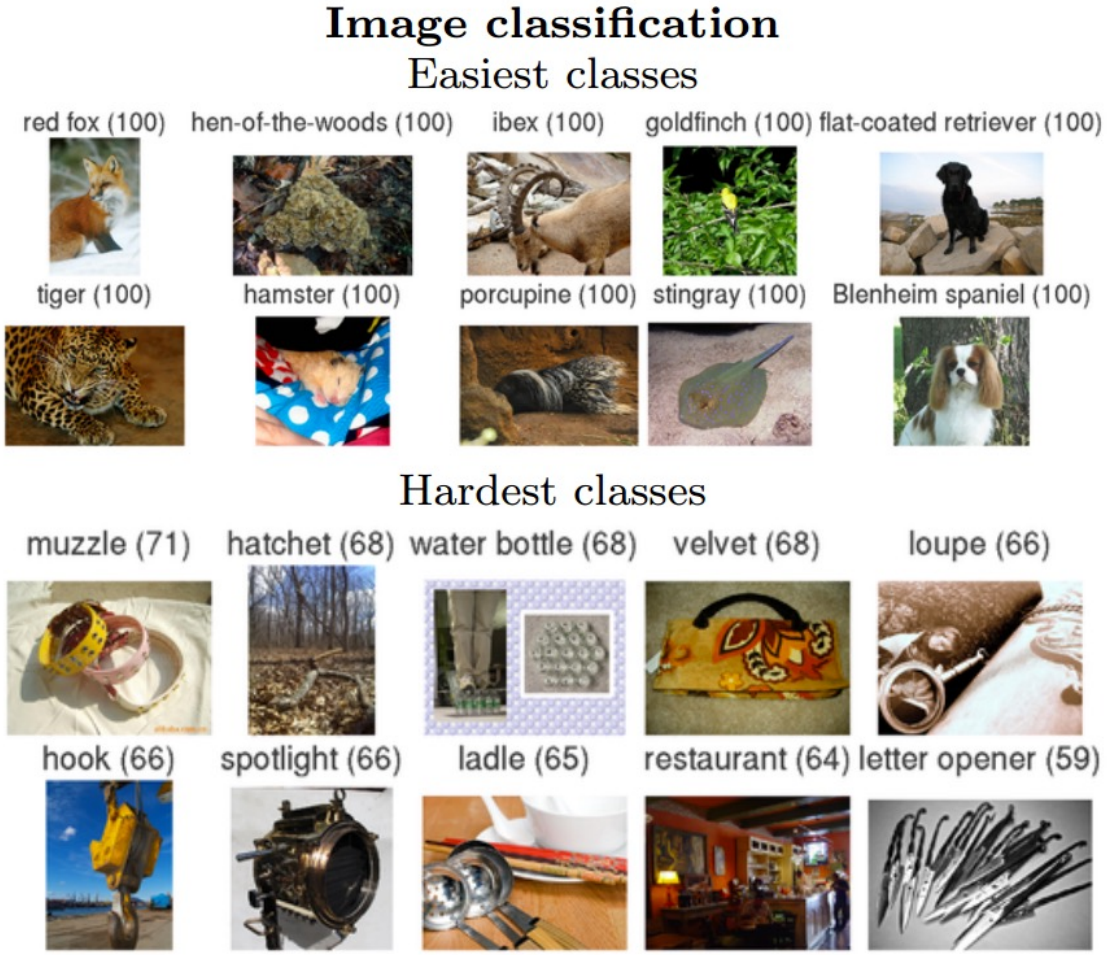
Reinforcement Learning: learning from trial and error with feedback from environment.

Main Areas within Machine Learning (cont.)



Armando Arroyo GeekStyle

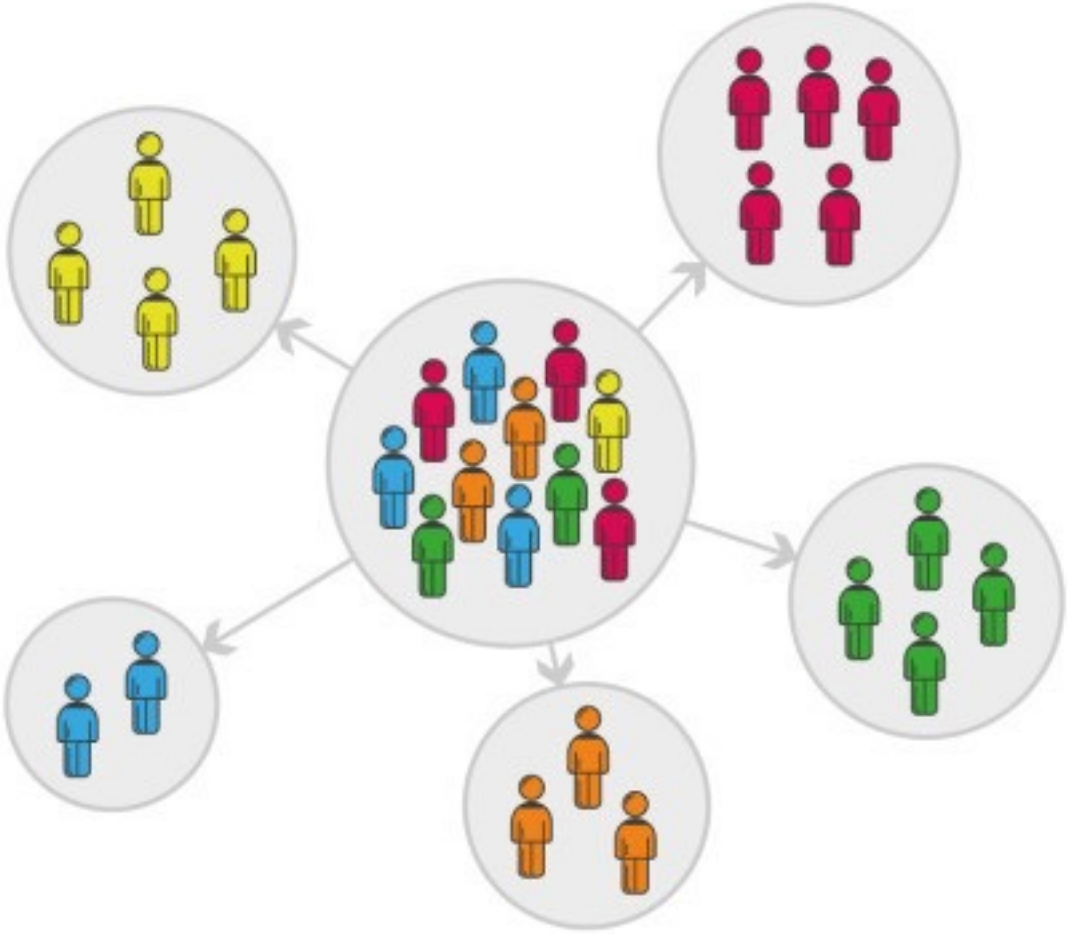
Supervised Machine Learning: Example



Unsupervised Machine Learning: Example

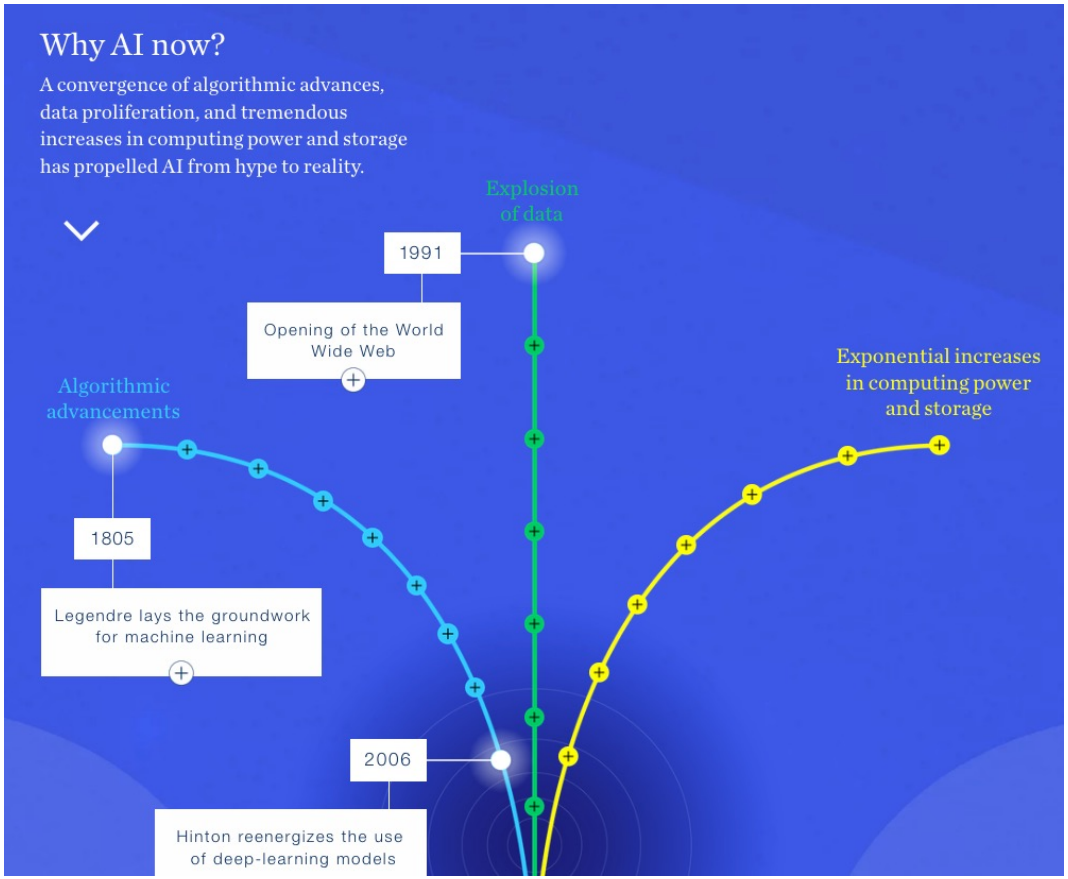
Frankfurt School of Finance & Management
German Excellence. Global Relevance.

Customer segmentation



[Project A, 2015]

Why now?



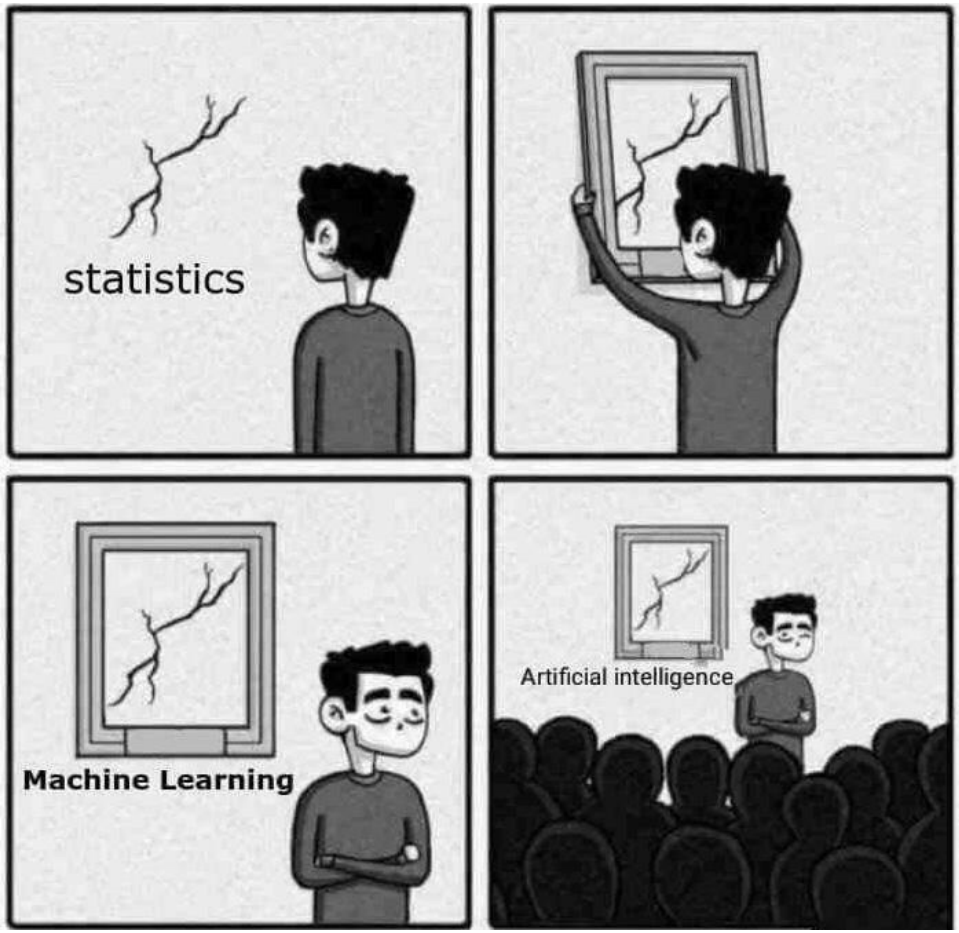
[McKinsey, 2018]

Main drivers of recent ML successes:

- More (labeled!) data
- More computing power
- New algorithms (note though: basic neural network structure has been around at least since the 1990s)

Machine Learning vs Statistics

The cynical view



[https://i.imgur.com/pWiyz4Z.jpg]

(not quite right as we have seen but still contains some truth)

Machine Learning vs Classical Statistics

- The two fields are converging: machine learning and statistical learning theory sometimes used interchangeably

- However, (supervised) **ML focuses on prediction**
  - Historically cares less about causality / explanation / interpretation / statistical significance
  - ML lets algorithm figure out which variables (and their interactions, e.g, $x_1 * x_2$) should be included
  - Often very much non-linear (don't need to be able to have explanation for why learned function has the form it has as long as it's accurate)
  - Evaluation on unseen data / cross validation

- ML developed out of computer science; hence: more computational
  - often: "bigger" data
  - often more "engineer-y" (and, say, less hypothesis testing)

- ML arguably more concerned with performance / accuracy than elegant theory (still a lot of research needed as to why deep learning works as well as it does)

Dictionary: Statistics ←→ Machine Learning

These terms are used inter-changeably:

- Regression coefficients = weights = parameters
- Residual variance = noise variance
- Inputs = predictors = features = explanatory variables = regressors = IVs = covariates
- Outputs = outcomes = targets = response variables = DVs = labels
- Training = learning = fitting (kind of; depends on context)

Supervised Machine Learning: Regression

Want to learn function $f(\boldsymbol{x})$ using samples (x_i, y_i) coming from

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

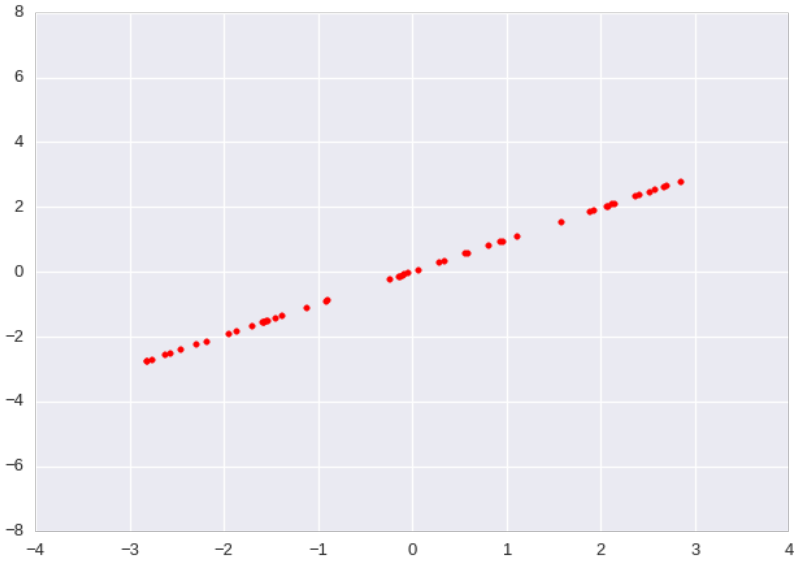
where $\boldsymbol{\epsilon}$ is a noise term with $\mathbf{E}[\boldsymbol{\epsilon}] = 0$.

Final objective: **accurate predictions for unseen points**!

Generative Model of Linear Regression

Frankfurt School of Finance & Management
German Excellence. Global Relevance.

Data is generated from a linear model with unknown parameters but corrupted with Gaussian noise:

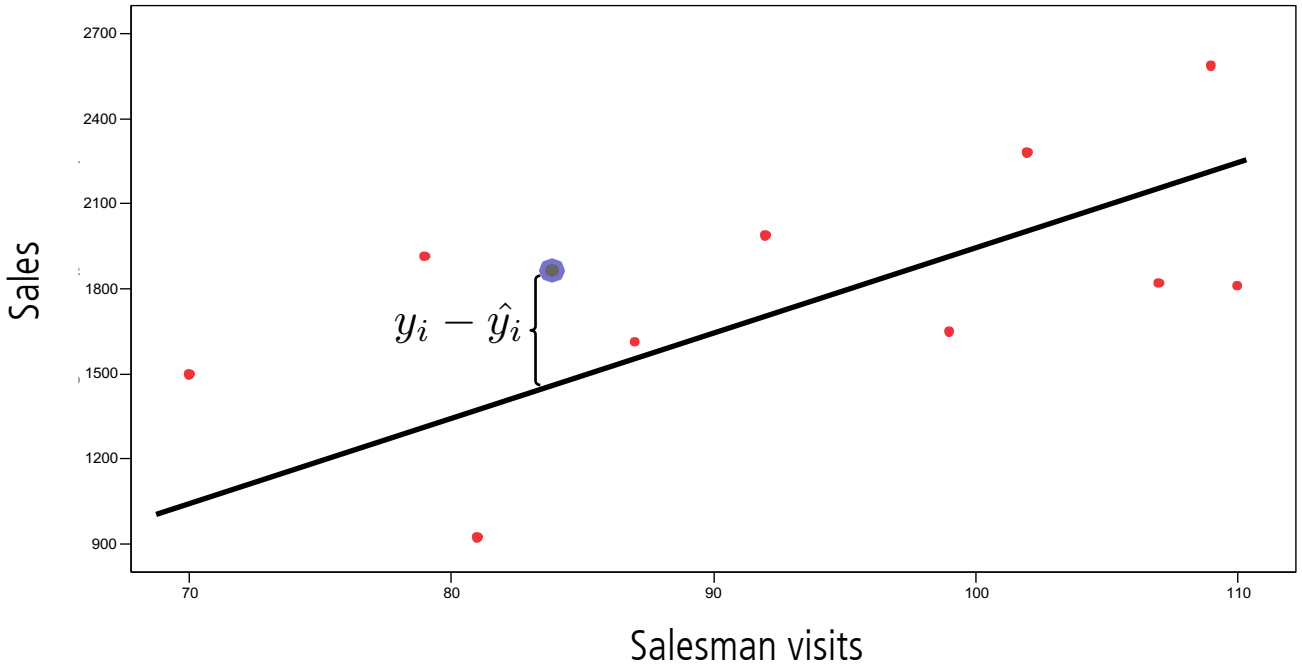


[Parkes and Rush, 2017]

Assume data comes from linear model, but we get to see only noisy version.

Linear Regression for Prediction

- After training a linear regression model, can use it to make predictions:



- How will we measure accuracy of the prediction (goodness of fit)?

  Typically: Mean Squared Error (MSE), i.e. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Python Practice

Two standard libraries:

```
import numpy as np
import matplotlib.pyplot as plt
```

Let's create two simple Numpy arrays:

```
x = np.array([1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 2.5])
y = np.array([7.6, 5.0, 4.0, 3.5, 4.0, 5.0, 7.0, 4.0])
```

Let's plot this:

```
plt.scatter(x,y)
plt.show()
```

What would be a good fit for this data?

Fitting a Linear Regression Model in Python

First, we need a new library:

```
from sklearn.linear_model import LinearRegression
```

Create a linear regression object using

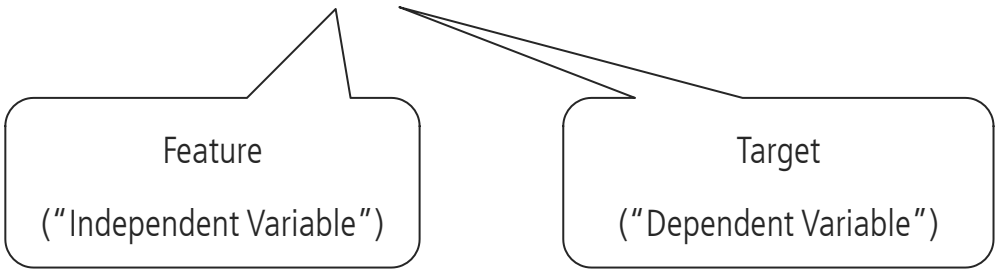
```
polyreg = LinearRegression()
```

Fitting a model requires a 2D feature array and reshape(-1,1) creates a 2D array from a 1D array:

```
X = x.reshape(-1, 1)
```

Fit the linear regression model using

```
polyreg.fit(X,y)
```



Predictions in Python

Use `predict` on your trained model to make predictions on unseen data, e.g., if you fitted your model using

```
polyreg.fit(X,y)
```

you can predict (unknown) value `Y` for `X = 10` running

```
new_X = np.array([[10]])
polyreg.predict(new_X)
```

Remember: 2D array

To predict for several values:

```
new_X = np.array([[-10],[15],[50]])
polyreg.predict(new_X)
```

For a whole range of values, use `arange` or `linspace`:

```
X_seq = np.linspace(0,10,101).reshape(-1,1)
polyreg.predict(X_seq)
```

Plotting the fitted model:

```
plt.plot(X_seq, polyreg.predict(X_seq), color="black")
```

Prof. Dr. Jens Witkowski
Associate Professor of Computer Science and Management

Frankfurt School of Finance & Management
Adickesallee 32-34
60322 Frankfurt am Main
Germany

Email: j.witkowski@fs.de
Tel.: +49 (0) 69-154008-875
Fax: +49 (0) 69-154008-4875