TU Delft

Spatial Data Science

Machine Learning for Everyone

(EPA122A)

Lecture 9

Trivik Verma

Aims of the Final Project

- The final project is a **group work** with an aim to apply spatial data science methods to study a relevant spatial phenomenon.
- You should report your findings in a **concise bulleted document** – template provided on the course webpage.
  - Fairly common way of reporting when purpose is **not to test writing skills**
  - Emphasis is on presenting nice figures and maps, <u>less but adequate text</u>
- You will also conduct a **peer-assessment** for your group members.

Topic of the Final Project

- You may choose freely,
  - However, the topic should have a strong link to course contents. If unclear, check with your assigned TA.
  - I provide you with three ideas on Mobility, Vulnerability, and COVID-19. You may also use these to do another case-study by collecting your own data from a city/area of your choice.
  - The main criteria for a suitable topic is that you should relate it to a). processes unfolding in a city like examples above, and b). use spatial data science to understand the chosen phenomena.
- The objective of the group project is,
  - Work as a team and teach each other what you have learned about the chosen topic, and what kind of methods, visual and data-related challenges appear in your analysis.
  - Deepen your understanding about specific spatial issues that interest you.
  - Observe how a phenomena can unfold in space, highlighting the importance of looking at space and geography.

Contents of the Final Project

- **Abstract**: Summarise the entire report in < 250 words
- **Introduction**: introduce the reader/listener to the **key questions** that you aim to answer with the analysis
- **Related Work**: describe relevant background to the topic
- **EDA / Data & methods**: Introduce **key datasets** and methods that you used to derive your results
- **Analysis / Results**: Represent your results with nice (and big) visualizations: maps, graphs and other possible infographics
- **Discussion**: List/represent shortly the key findings of your study, i.e. provide the "take home messages”. Discuss about the potential data/methodology related challenges and other issues that might relate to your topic
- **References**: Add a short list of references relevant to your analysis and questions

Full template on website

Miscellaneous leftovers..

Groups formed? 4 members only!

Week 5 suggestions…

Scope of Work and Preliminary EDA

<u>Address the two most important things first:</u>

- Project statement. The project goal in the posted project description is not fully formulated or tuned. Based on the project description and references, state a well-defined question that you'll address in the project.
- Preliminary EDA. Explain your plans for preliminary data exploration. Please take care when planning, so that team members can work individually on these tasks if need be. Stay in touch with your team and communicate regularly.

<u>Consult with TAs next week. The order of groups and their TAs will be posted on Brightspace by Friday.</u>

Miscellaneous leftovers..

- Contact your group members to discuss about the project work and decide the final topic for your analysis: meet on campus, or use e.g. Zoom or Teams (whatever works the best for your group)
- Agree about the division of labour: Who take's the lead on the analytics side? Who focuses on finding good/relevant data? Who finds e.g. relevant literature and gathers relevant contextual information from the literature related to your topic? Who does visualisation based on principles of graphical excellence? Etc.
- Take advantage of the lecture slides and the recommended literature which is provided for you for each week. In addition, you search for more literature e.g. using Google Scholar.
- You can ask your assigned TAs for tips and hints about relevant data and literature if you don't know where to get started.
- Do not hesitate to contact me if there are any issues with the group work!

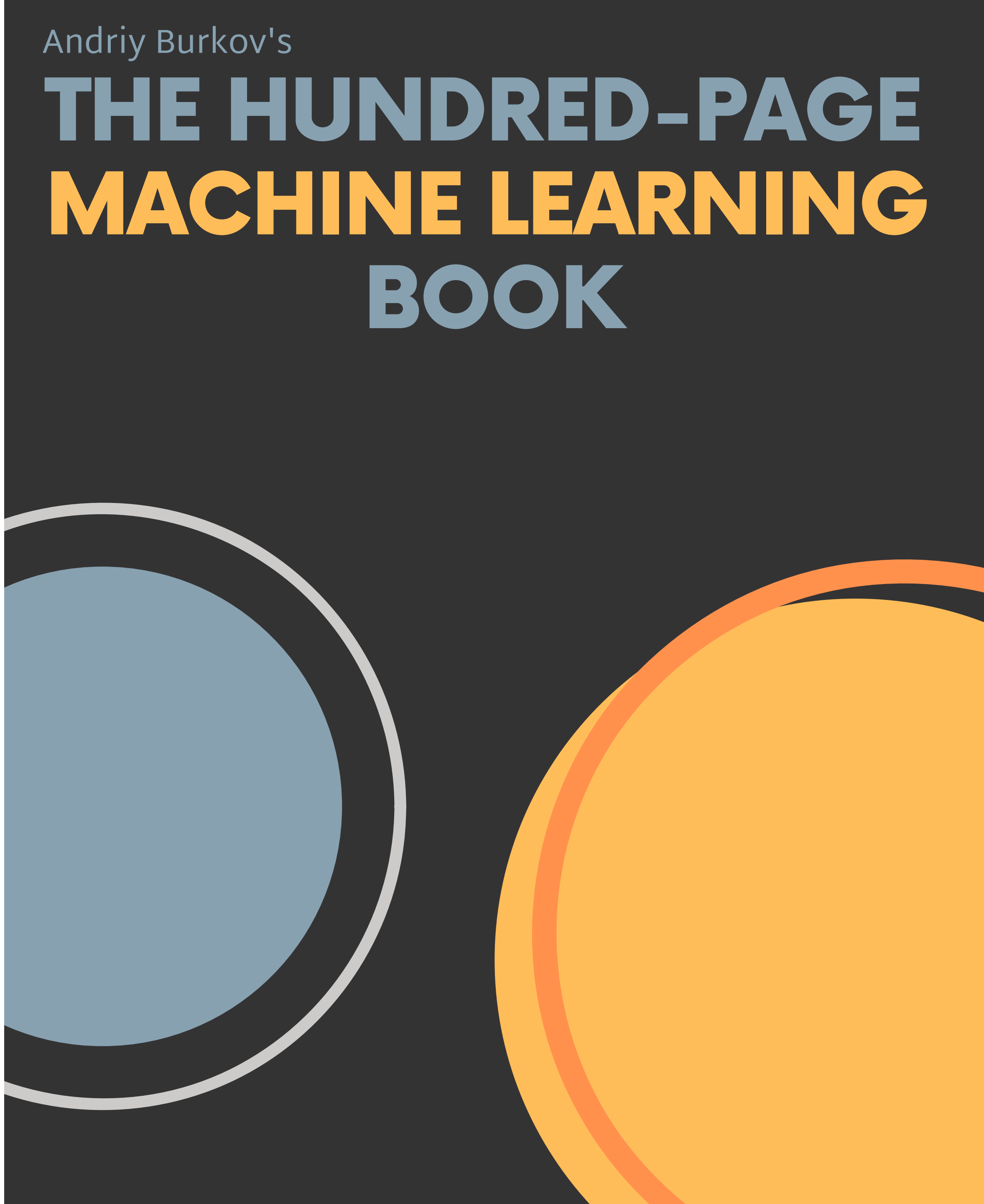
Last Time

- Exploratory Spatial Data Analysis (ESDA)
- Spatial Autocorrelation Measures
  - Global
  - Local

Today

- Machine Learning
- Predicting a Variable
- Error evaluation
- Model comparison
- Fitness of models

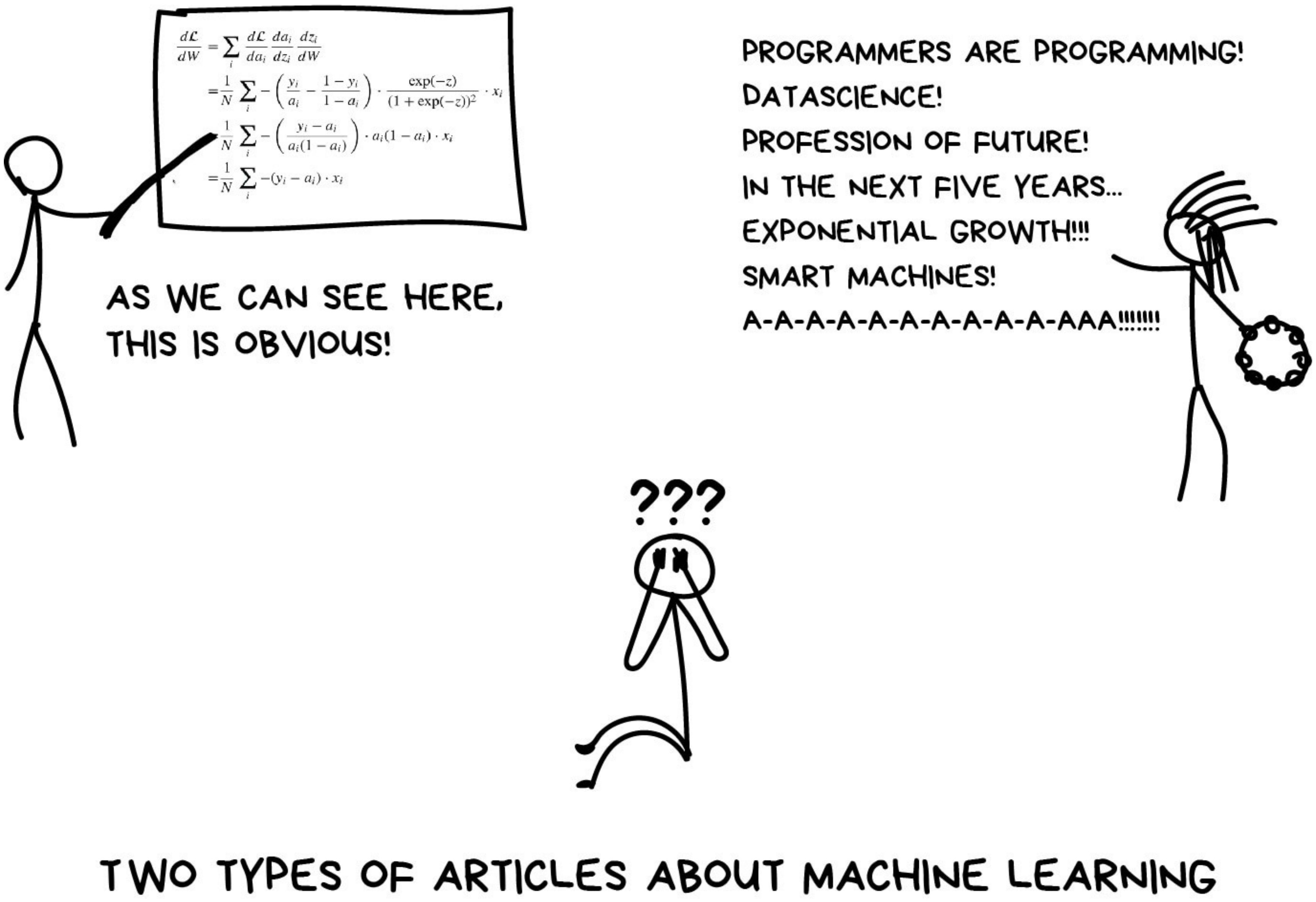
A beginner in machine learning will find in this book just enough details to get a comfortable level of understanding of the field and start asking the right questions. **[Optional]**



Burkov, A. (2019). The Hundred-Page Machine Learning Book by Andriy Burkov.

Let’s start with the Truth

Machines don’t learn



Images taken from the vas3k blog on Machine Learning

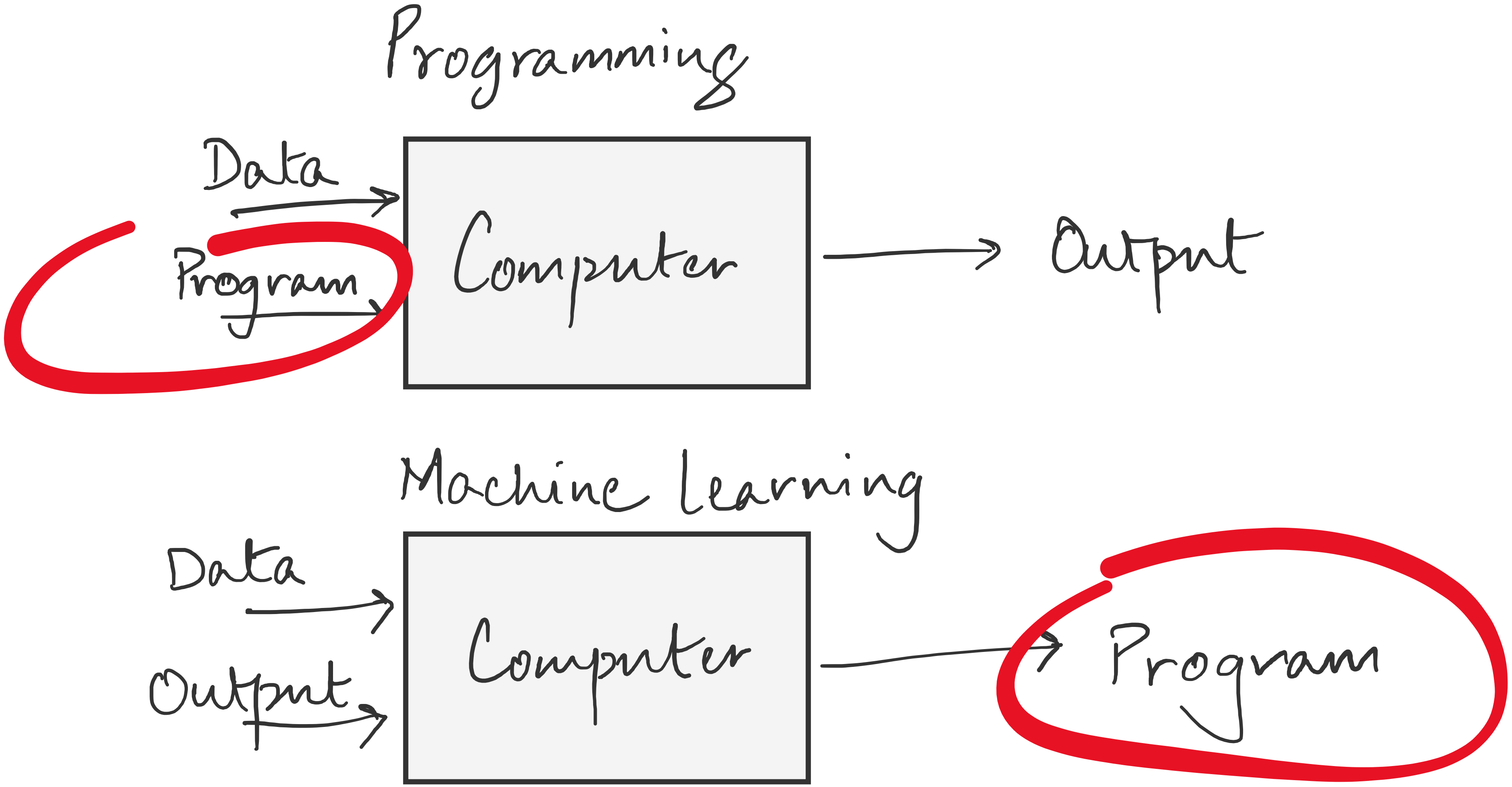
In this course: Only real-world problems, practical solutions, simple language, and no high-level theorems

TUDelft

Arthur Samuel (1959)

Source: Wikimedia Commons

"Field of study that gives computers the ability to learn without being explicitly programmed."


Programming
Data
Program
Computer
Output
Machine learning
Data
Output
Computer
Program

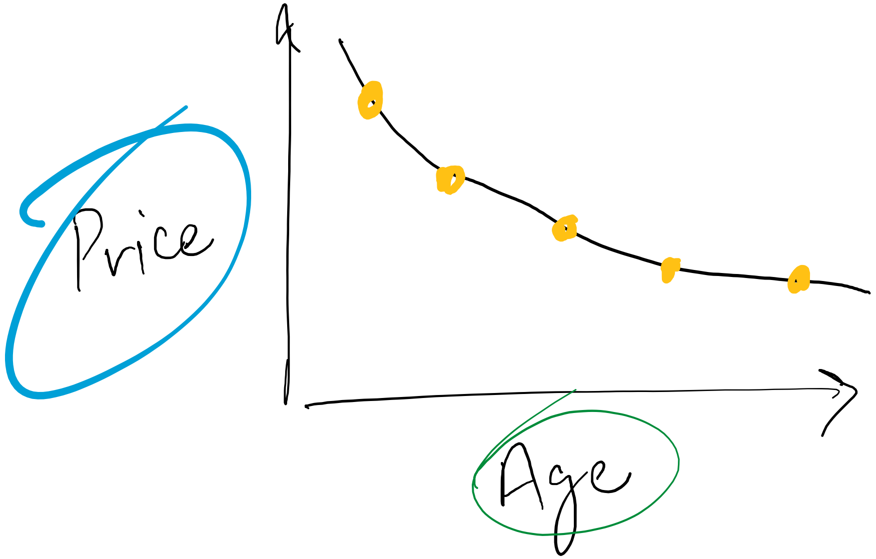
Why do we want machines to learn?

Year

1 €20,000
2 €19,000
3 €18,000
↓
N €10,000

Price ∝ Age

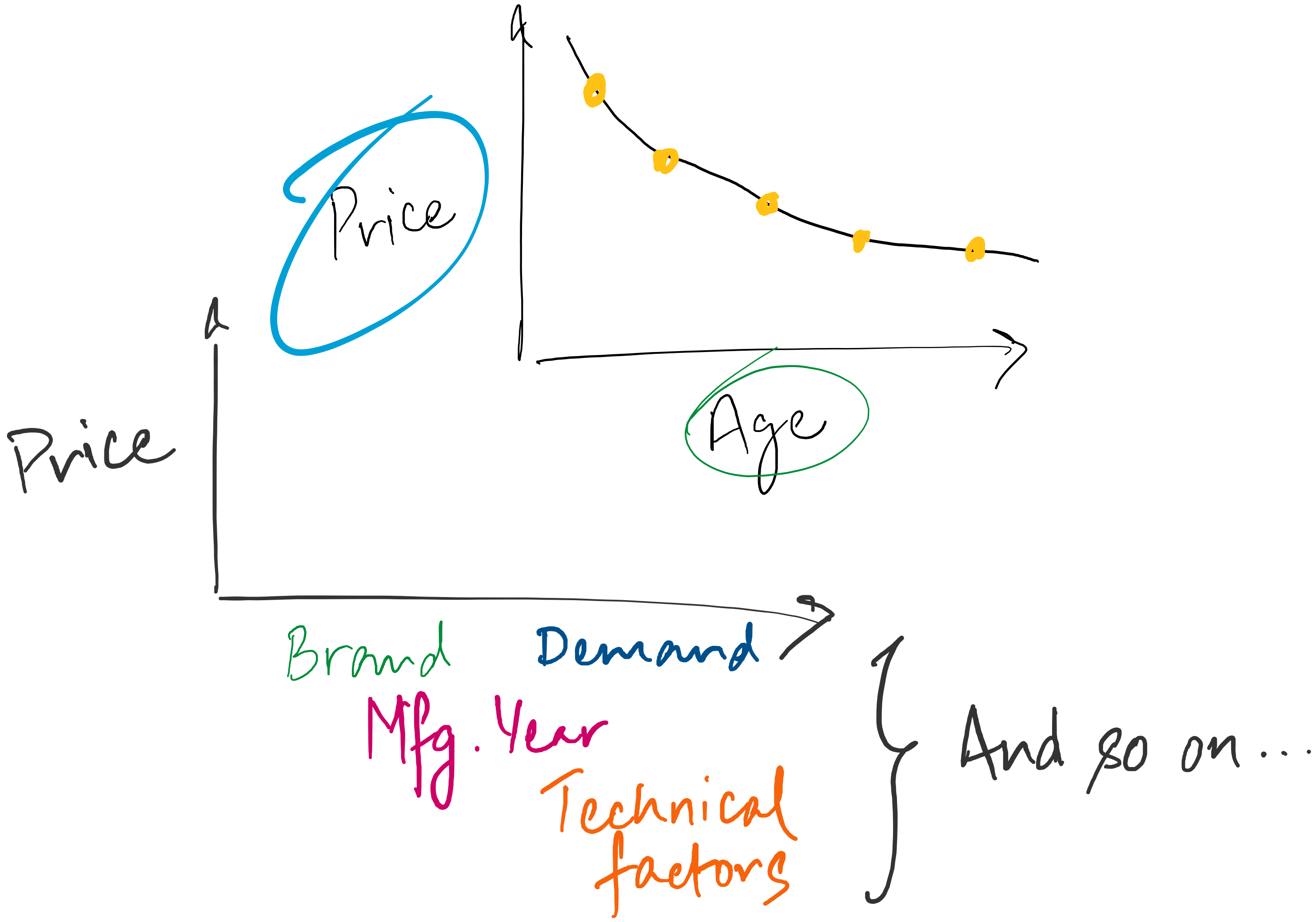
Why do we want machines to learn?



Year
(1) € 20,000
(2) € 19,000
(3) € 18,000
↓
(N) € 10,000

Price ∝ Age

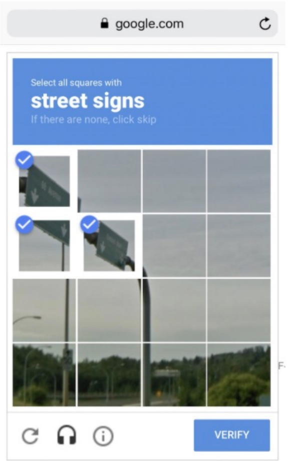
Images taken from the vas3k blog on Machine Learning



Price
Price
Age
Brand
Demand
Mfg. Year
Technical factors
And so on...

The goal of ML algorithms is to predict results based on incoming data. That's it!

What do we need to teach machines?



DATA

Crime
Poverty
Weather
Stocks
Socio-economic condition

Features

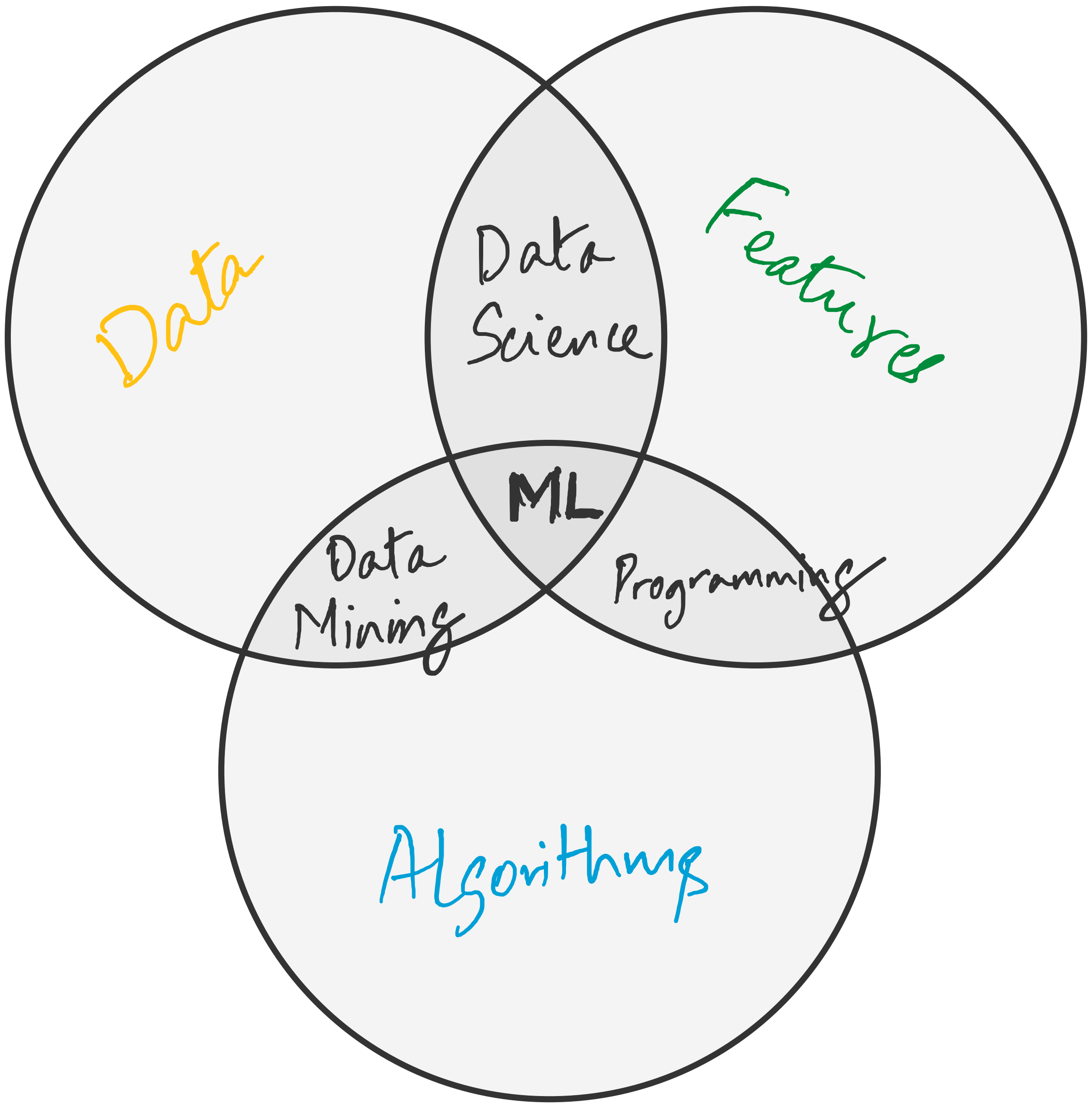
Income
Amenities
Gender
Stock price
Travel info
Family info

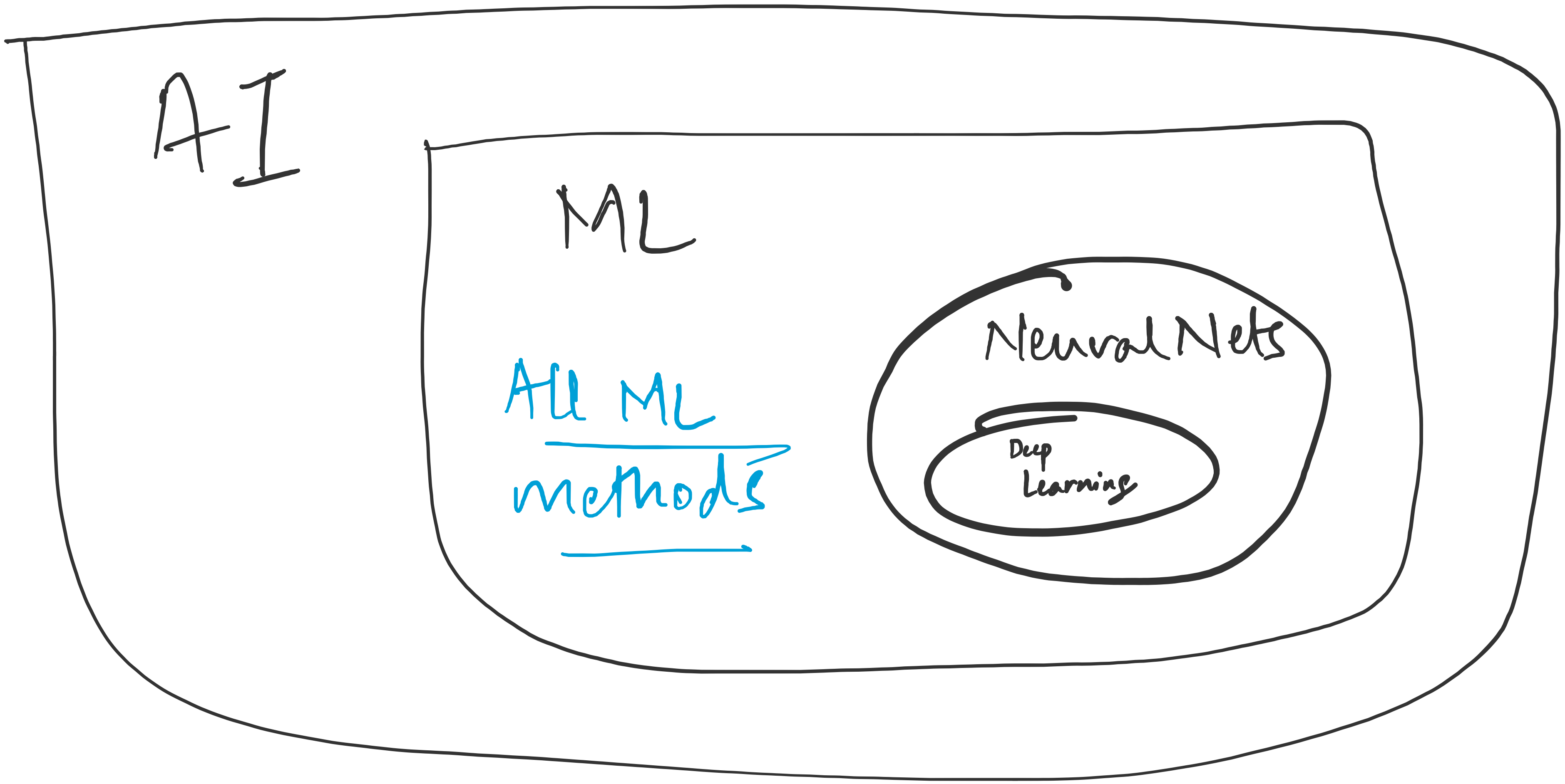
Algorithms

- precision
- performance
- size

[garbage in - garbage out]

Images taken from the vas3k blog on Machine Learning


Data
Features
Algorithms
Data Science
Data Mining
Programming
ML



AI
ML
All ML methods
Neural Nets
Deep Learning

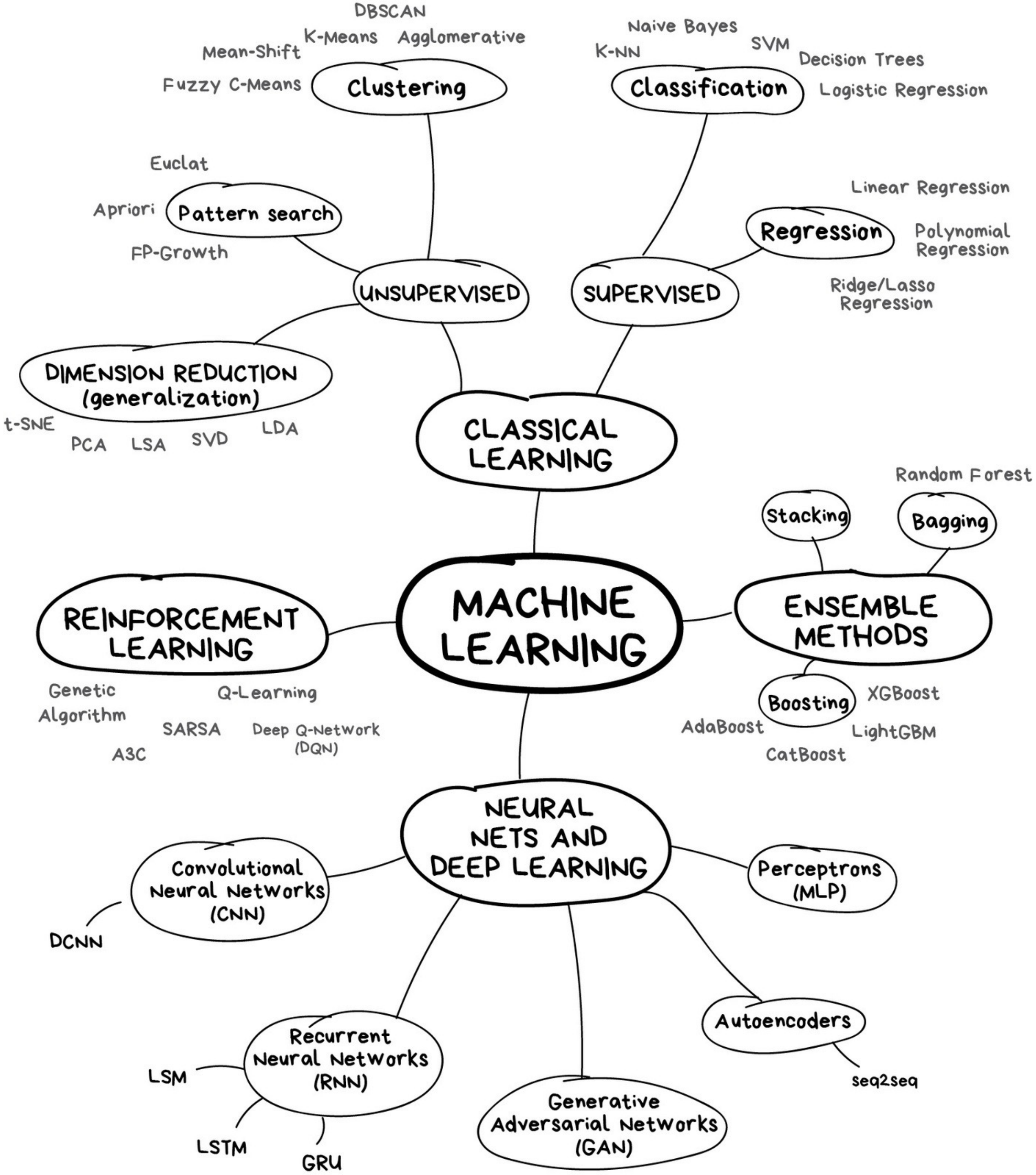
AI : Field like → physics
→ Chemistry
→ Math

ML : Part of AI / Not the only one

NN : Type of ML

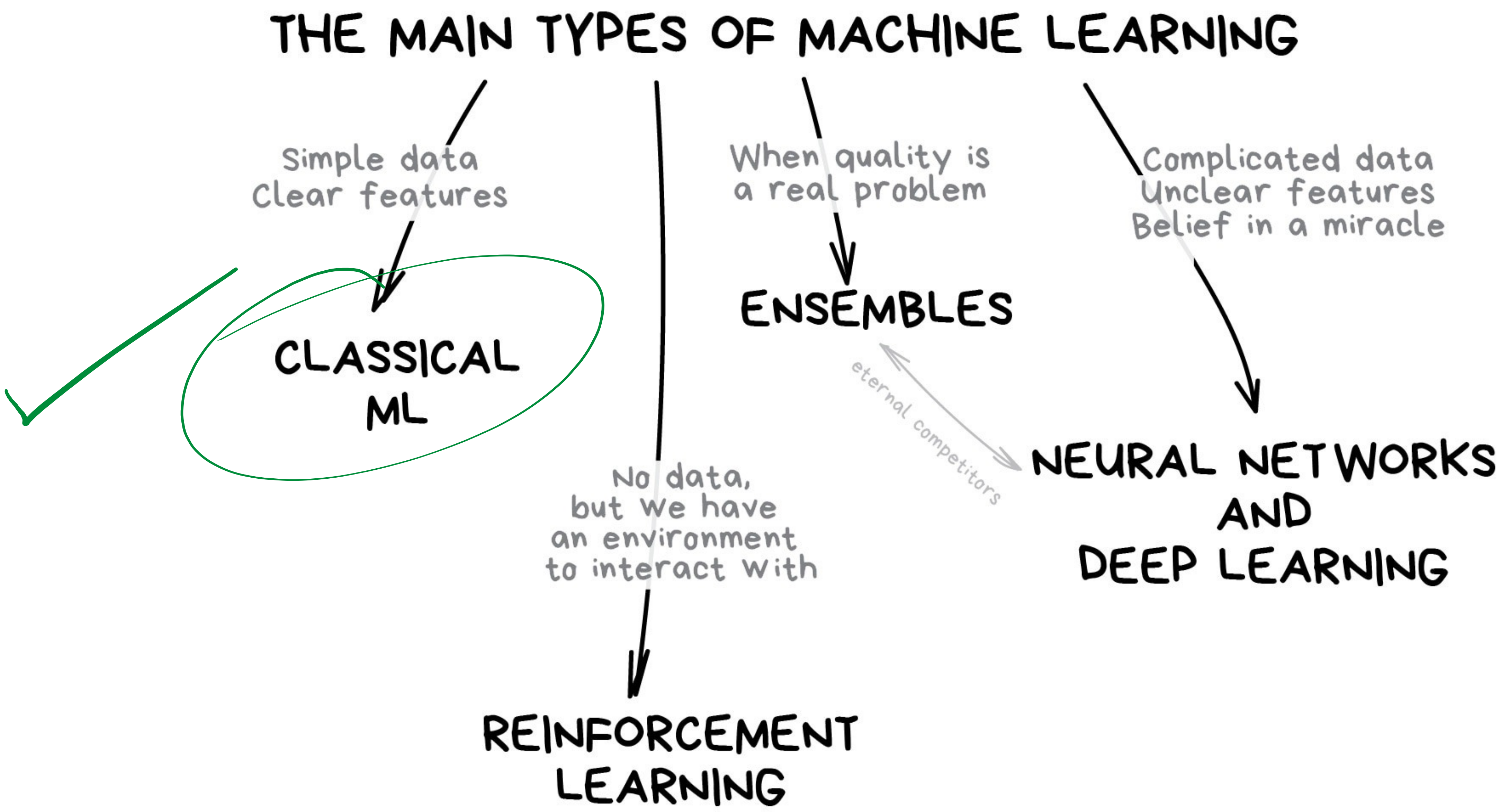
Deep Learning : Basically a new architecture for NN.

Machines **CAN**	Machines **CANNOT**
- Forecast	- Create / Innovate
- Memorise	- Get Smart
- Reproduce	- Go beyond the task
- Choose best object	- Take over the world

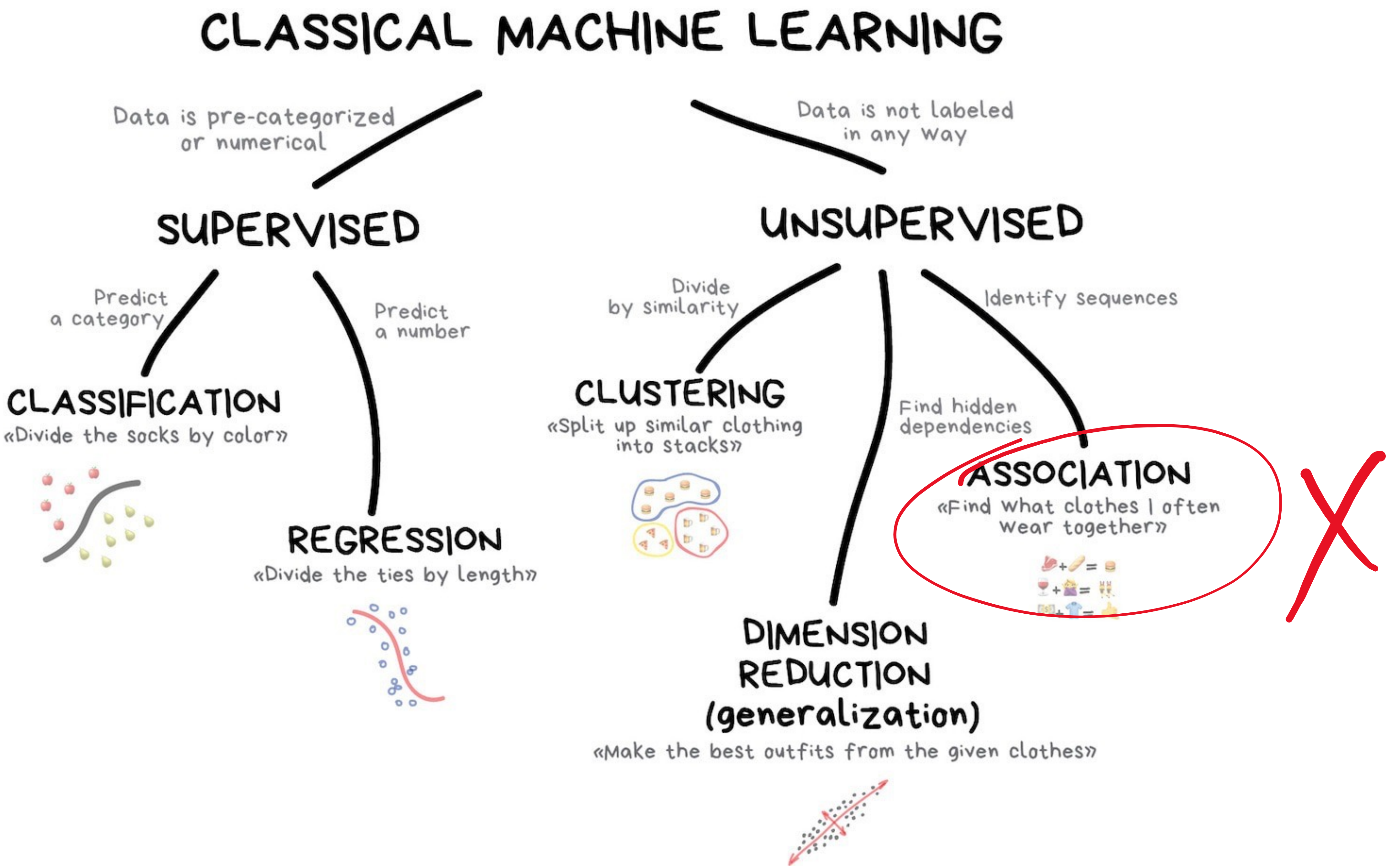


Images taken from the vas3k blog on Machine Learning

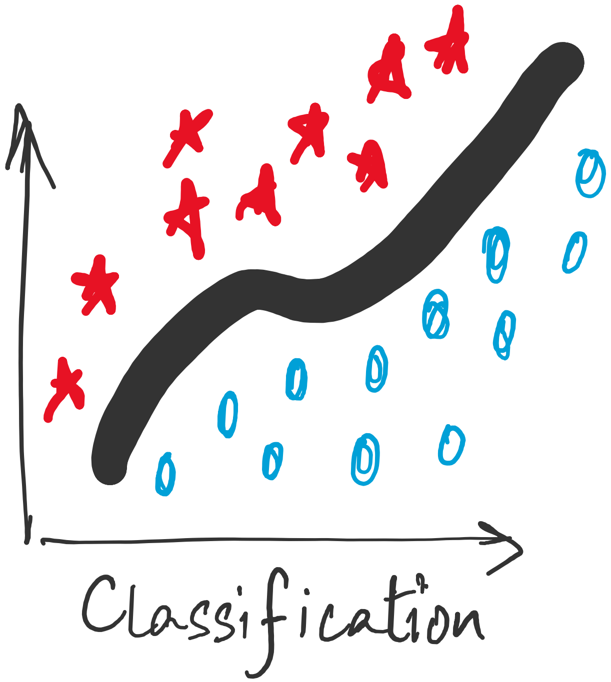
TU Delft



Images taken from the vas3k blog on Machine Learning

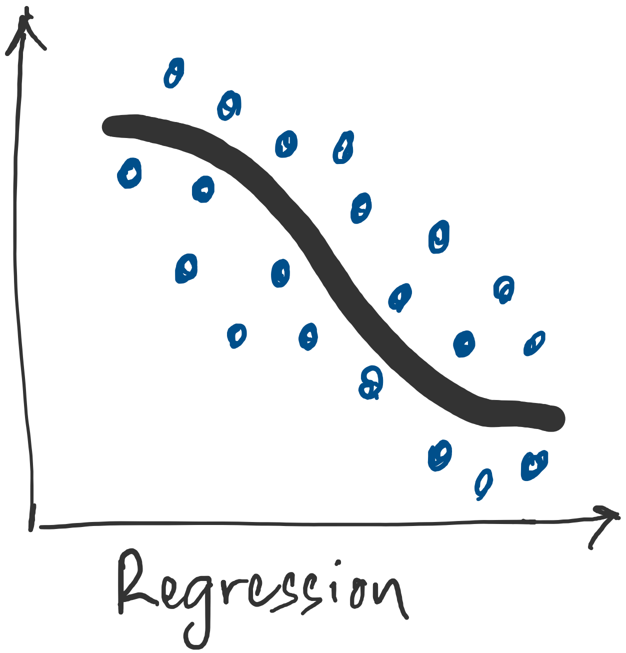


Images taken from the vas3k blog on Machine Learning



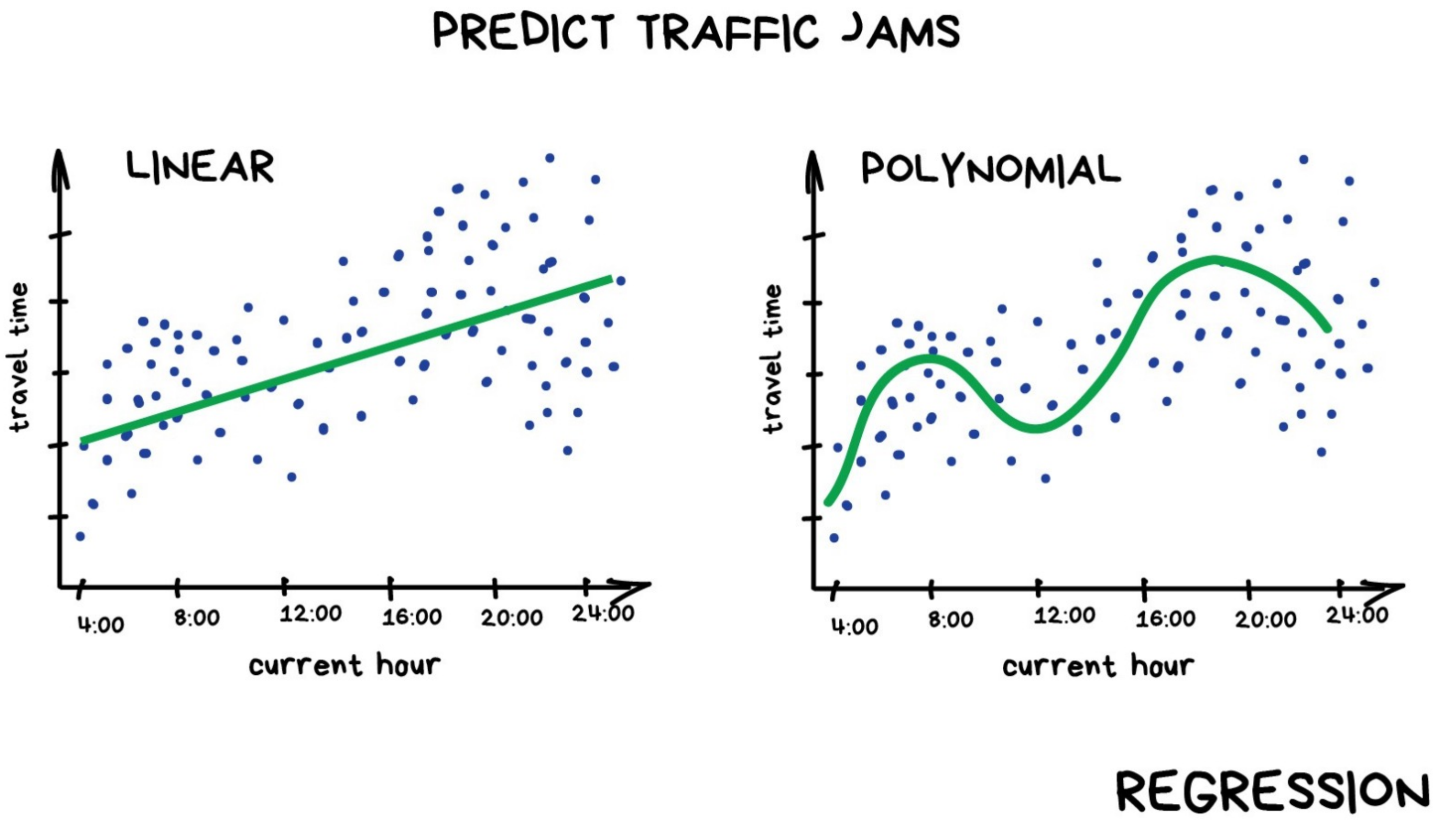
Today used for:

- Spam filtering
- Language detection
- A search of similar documents
- Sentiment analysis
- Recognition of handwritten characters and numbers
- Fraud detection
- Users based on interests (as algorithmic feeds do)



Today this is used for:

- Stock price forecasts
- Demand and sales volume analysis
- Medical diagnosis
- Any number-time correlations
- Car price by its mileage
- Traffic by time of the day



When the line is straight — it's a linear regression, when it's curved – polynomial. These are two major types of regression. The other ones are more exotic. Logistic regression is a different sheep in the flock. Don't let it trick you, as it's a classification method, not regression.

Predicting a variable

Let’s imagine a scenario where we'd like to predict one variable using another (or a set of other) variables.

Examples:

- Predicting the number of views, a YouTube video will get next week based on video length, the date it was posted, the previous number of views, etc.
- Predicting which movies, a Netflix user will rate highly based on their previous movie ratings, demographic data, etc.

Data

The **Advertising data set** consists of the sales of a particular product in 200 different markets, and advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) **with** permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Response vs. Predictor Variables

There is an asymmetry in many of these problems:

- The variable we would like to predict may be more difficult to measure, is more important than the other(s), or maybe directly or indirectly influenced by the other variable(s).

Thus, we'd like to define two categories of variables:

- variables whose values we want to predict
- variables whose values we use to make our prediction

Response vs. Predictor Variables

X
predictors
features
covariates

Y
outcome
response variable
dependent variable

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

n observations

p predictors

Response vs. Predictor Variables

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

Predictors, features, covariates

$Y = y_1, \dots, y_n$

outcome

response variable

dependent variable

n observations

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

p predictors

Statistical Model

True vs. Statistical Model

We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

$$Y = f(X) + \varepsilon$$

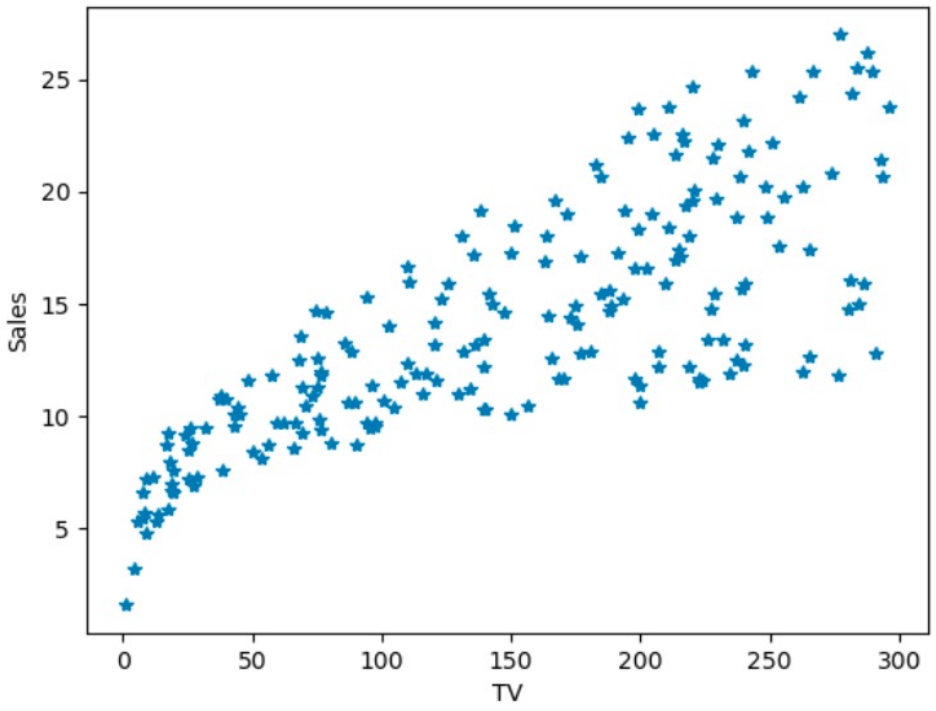
Here, f is the unknown function expressing an underlying rule for relating Y to X, ε is the random amount (unrelated to X) that Y differs from the rule $f(X)$.

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\widehat{f}$.

Example: predicting sales

Motivation: Predict Sales

Build a model to **predict** sales based on TV budget



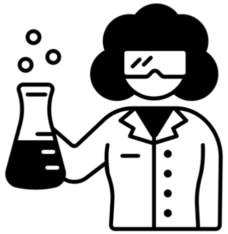
The response, y, is the sales

The predictor, x, is TV budget

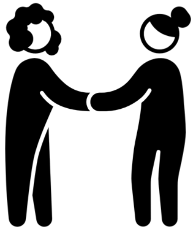
Break

CHILL

WALK



COFFEE OR TEA

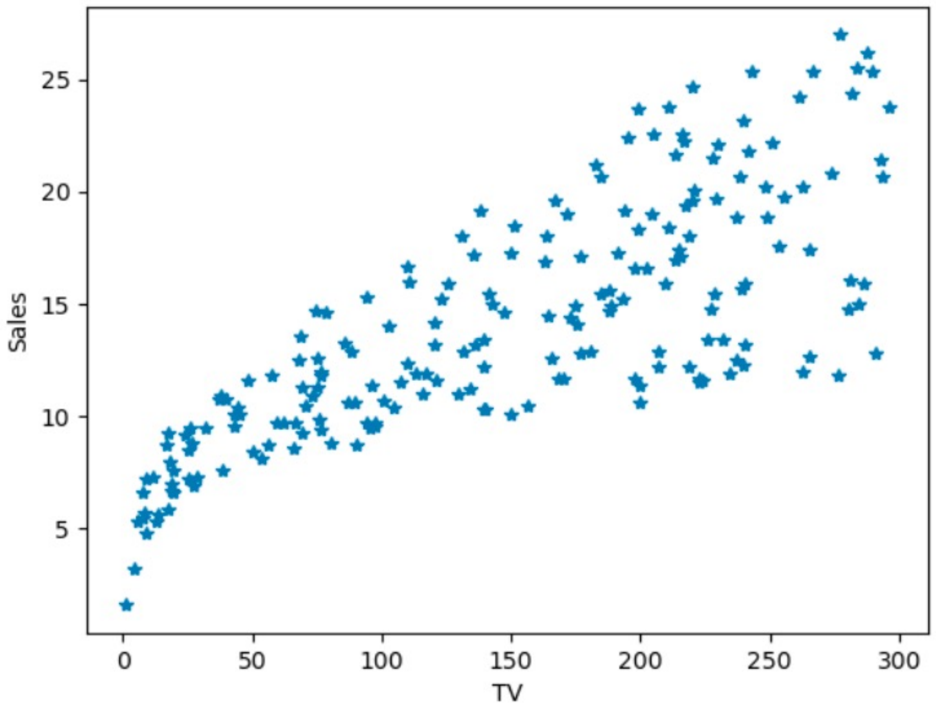


MAKE FRIENDS

Example: predicting sales

Motivation: Predict Sales

Build a model to **predict** sales based on TV budget

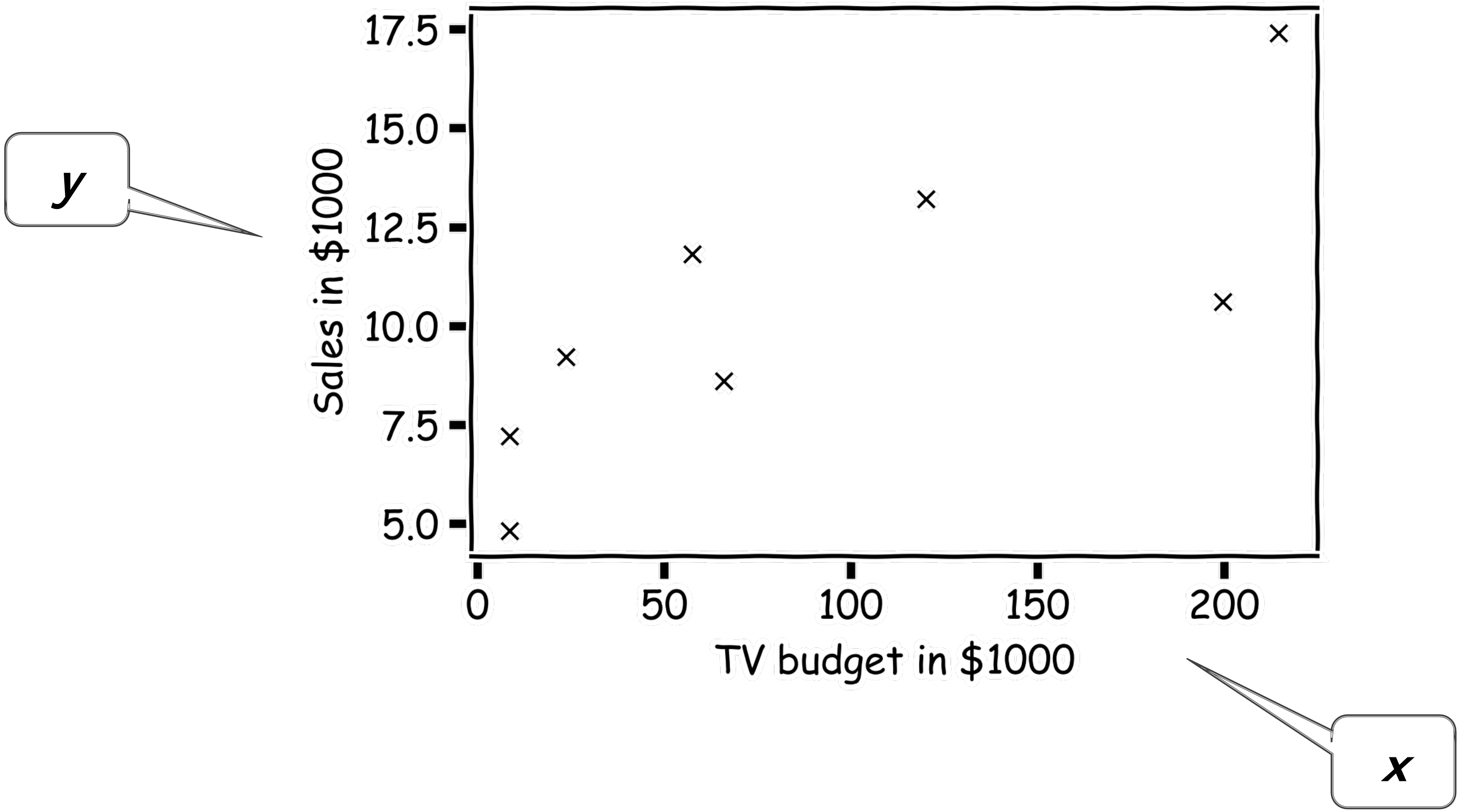


The response, y, is the sales

The predictor, x, is TV budget

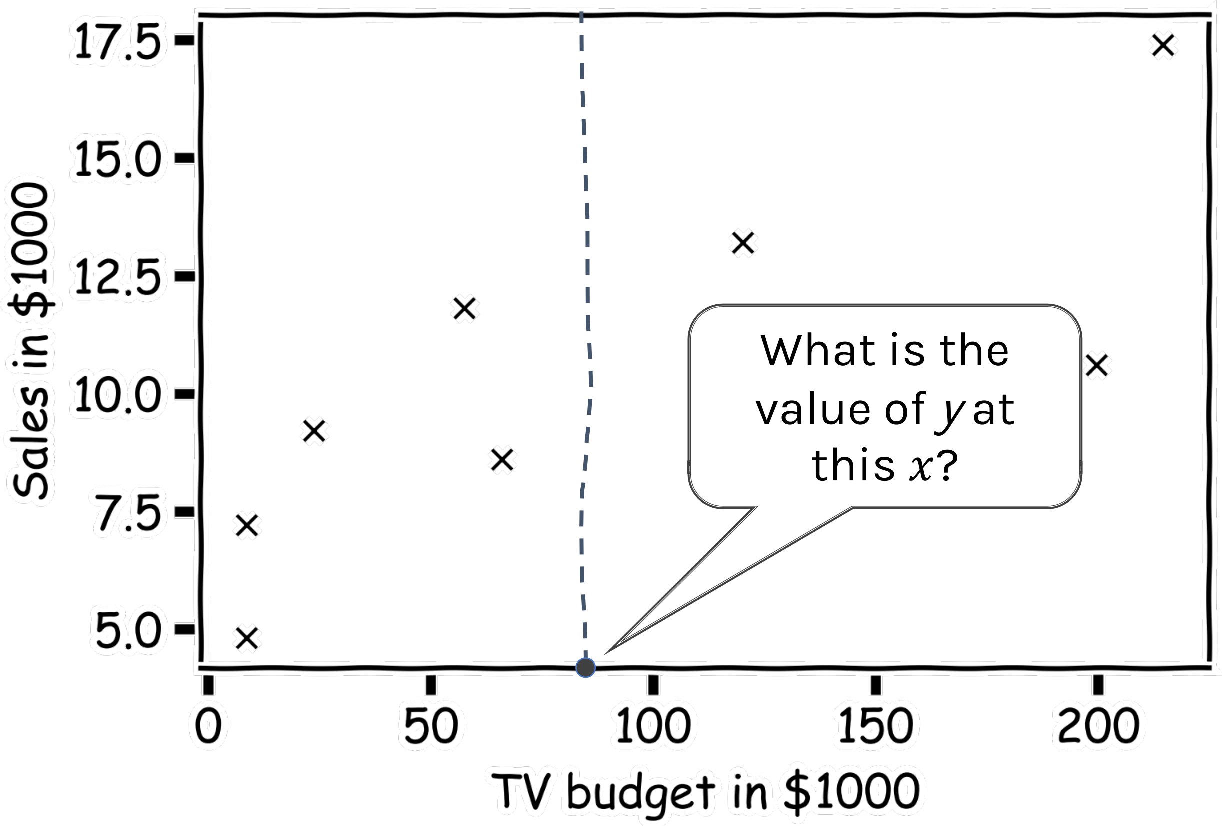
This example and all material are provided by the Data Science course at Harvard University

Statistical Model



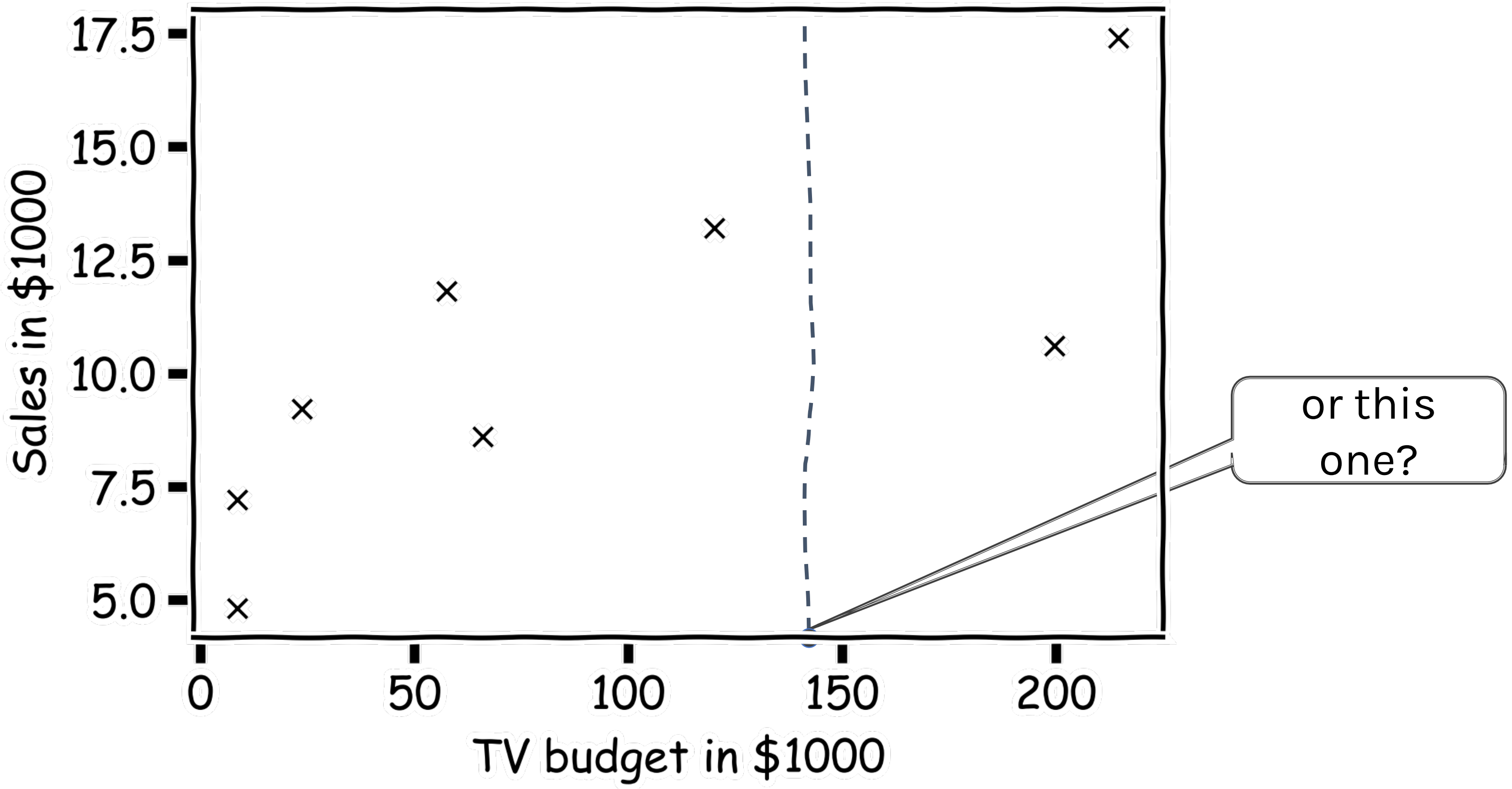
Statistical Model

How do we predict y for some x



Statistical Model

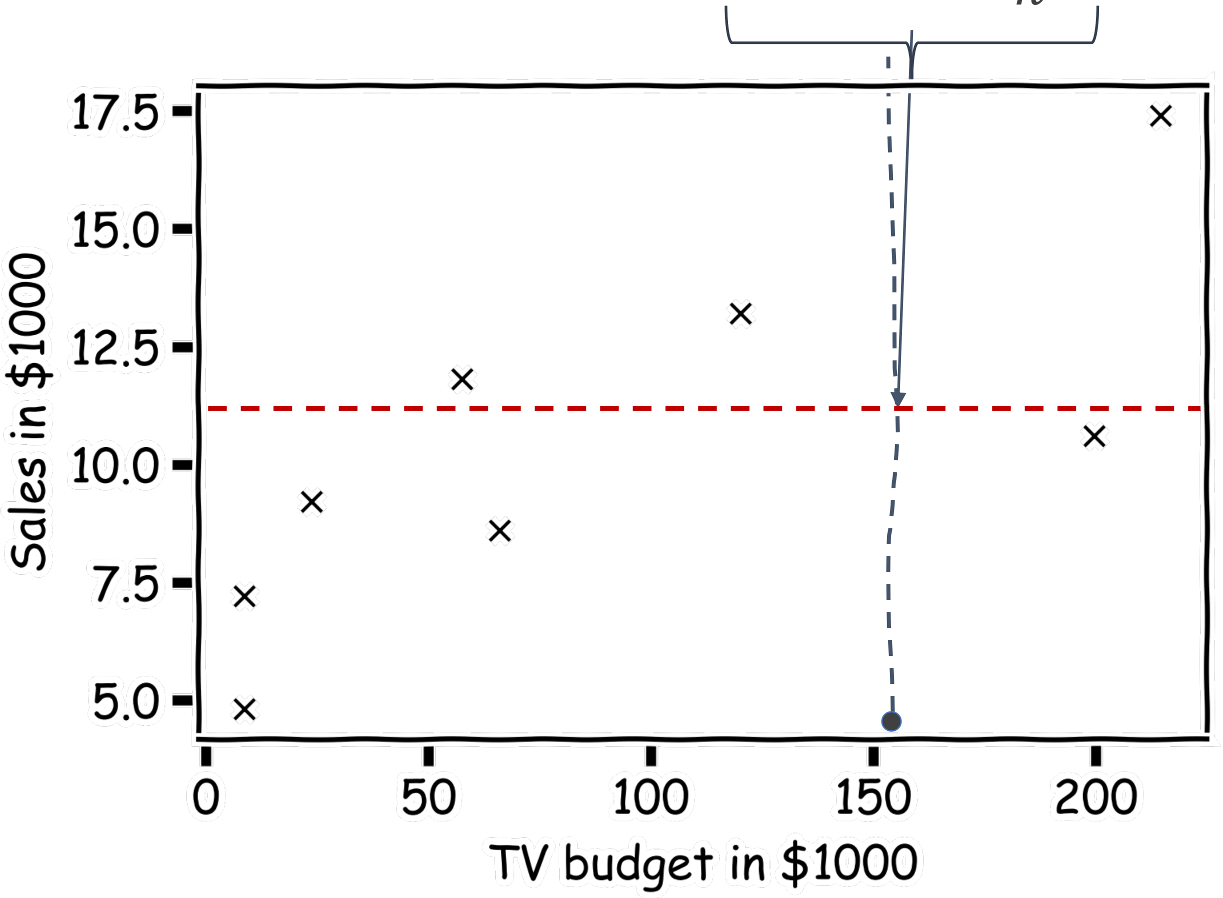
How do we predict y for some x



This example and all material are provided by the Data Science course at Harvard University

Statistical Model

Simple idea is to take the mean of all y's: $\frac{1}{n}\sum_{1}^{n} y_i$



This example and all material are provided by the Data Science course at Harvard University

Prediction vs. Estimation

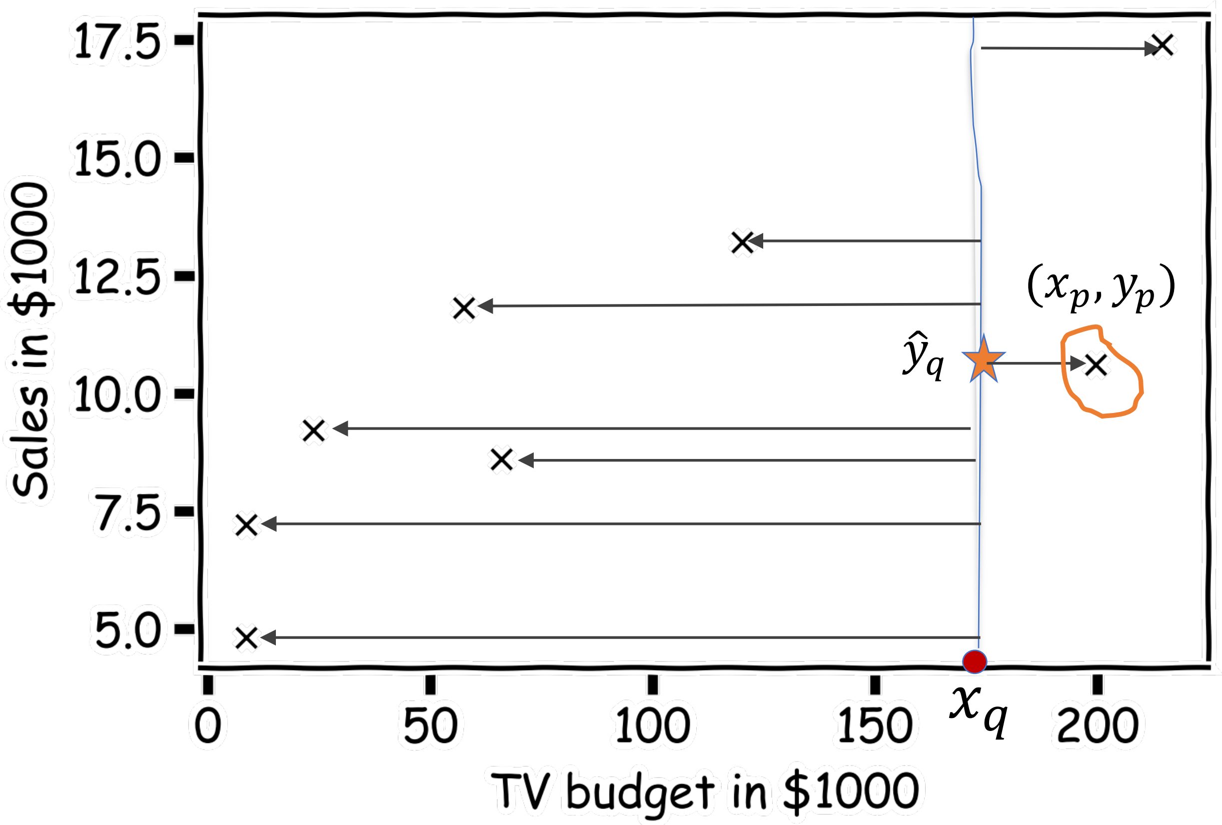
For some problems, what's important is obtaining $\hat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements, $(x_{i,1}, \dots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p}).$$

For some problems, we don't care about the specific form of $\hat{f}$, we just want to make our predictions $\hat{y}$'s as close to the observed values y's as possible. These are called ***prediction problems***.

Simple Prediction Model



What is $\hat{y}_q$ at some x_q ?

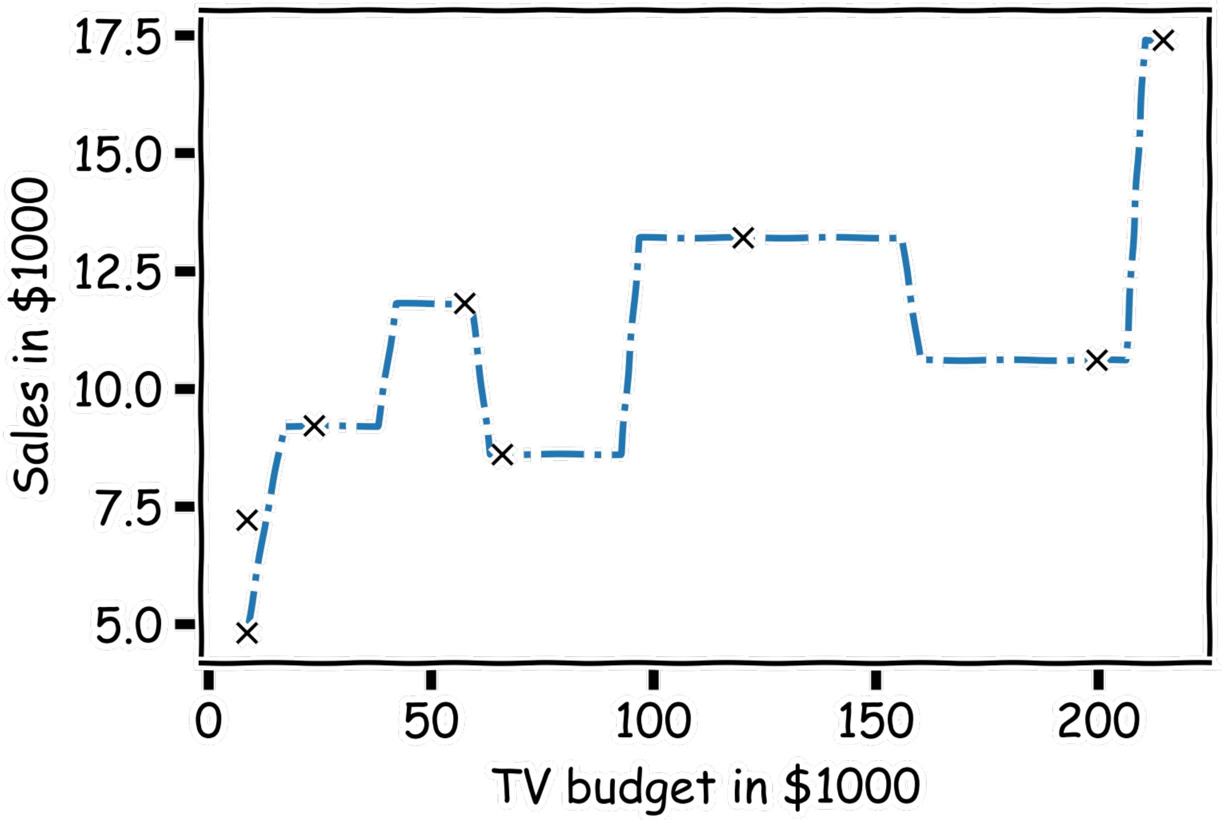
Find distances to all other points $D(x_q, x_i)$

Find the nearest neighbor, (x_p, y_p)

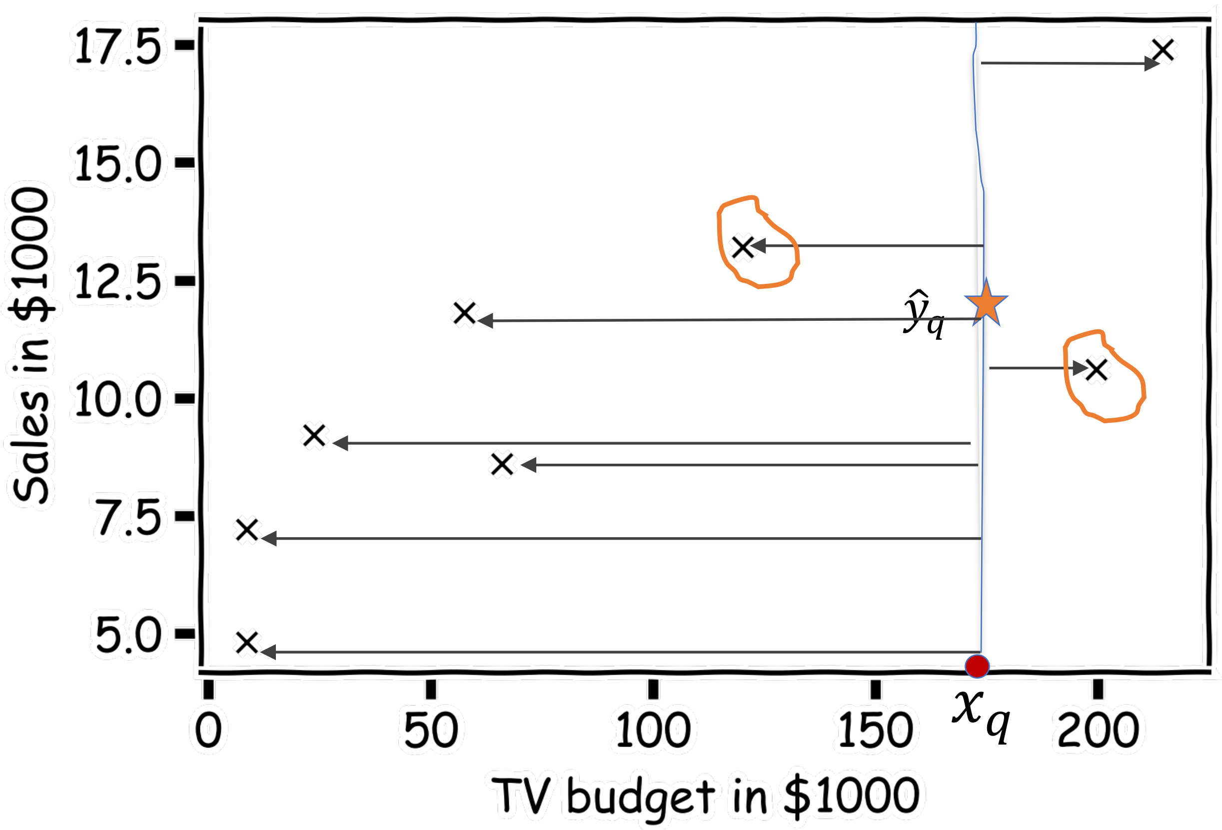
Predict $\hat{y}_q = y_p$

Simple Prediction Model

Do the same for "all" $x's$



Extend the Prediction Model



What is $\hat{y}_q$ at some x_q ?

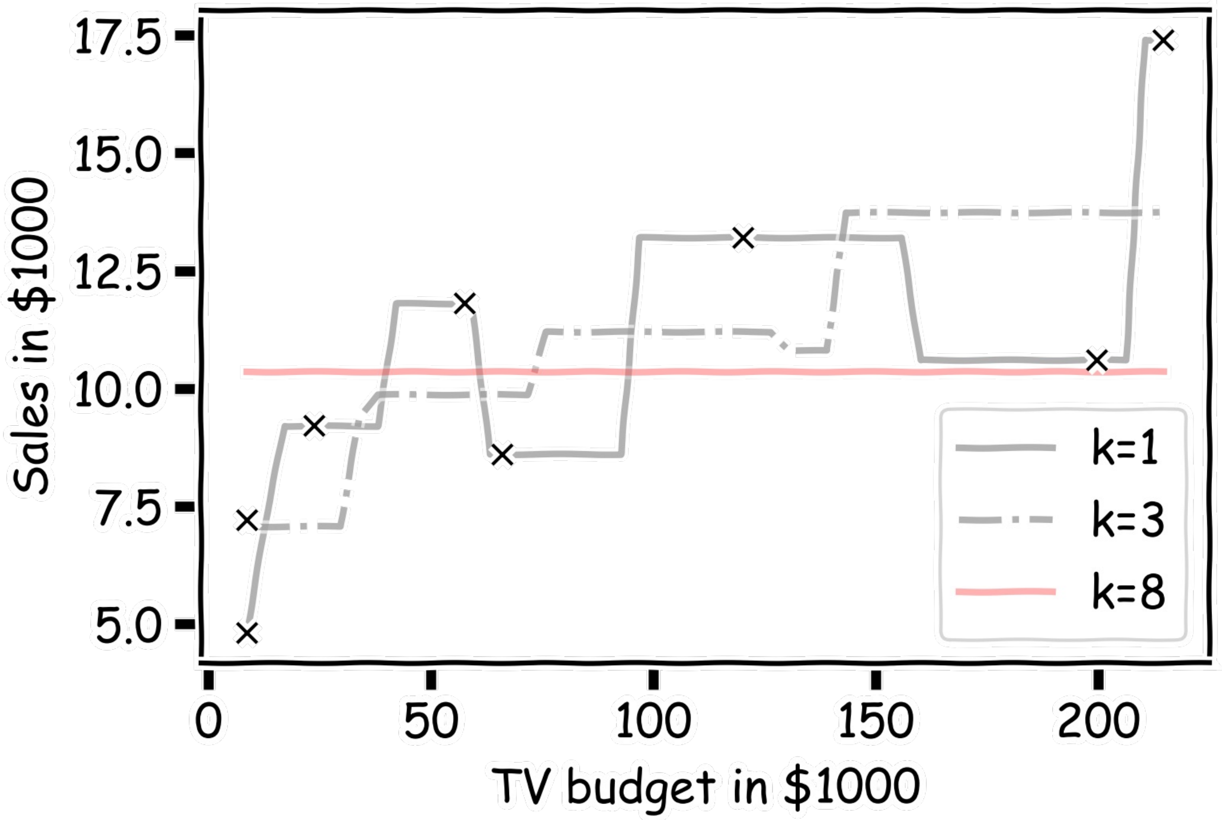
Find distances to all other points $D(x_q, x_i)$

Find the k-nearest neighbors, $x_{q_1}, \dots, x_{q_k}$

Predict $\hat{y}_q = \frac{1}{k} \sum_i^k y_{q_i}$

This example and all material are provided by the Data Science course at Harvard University

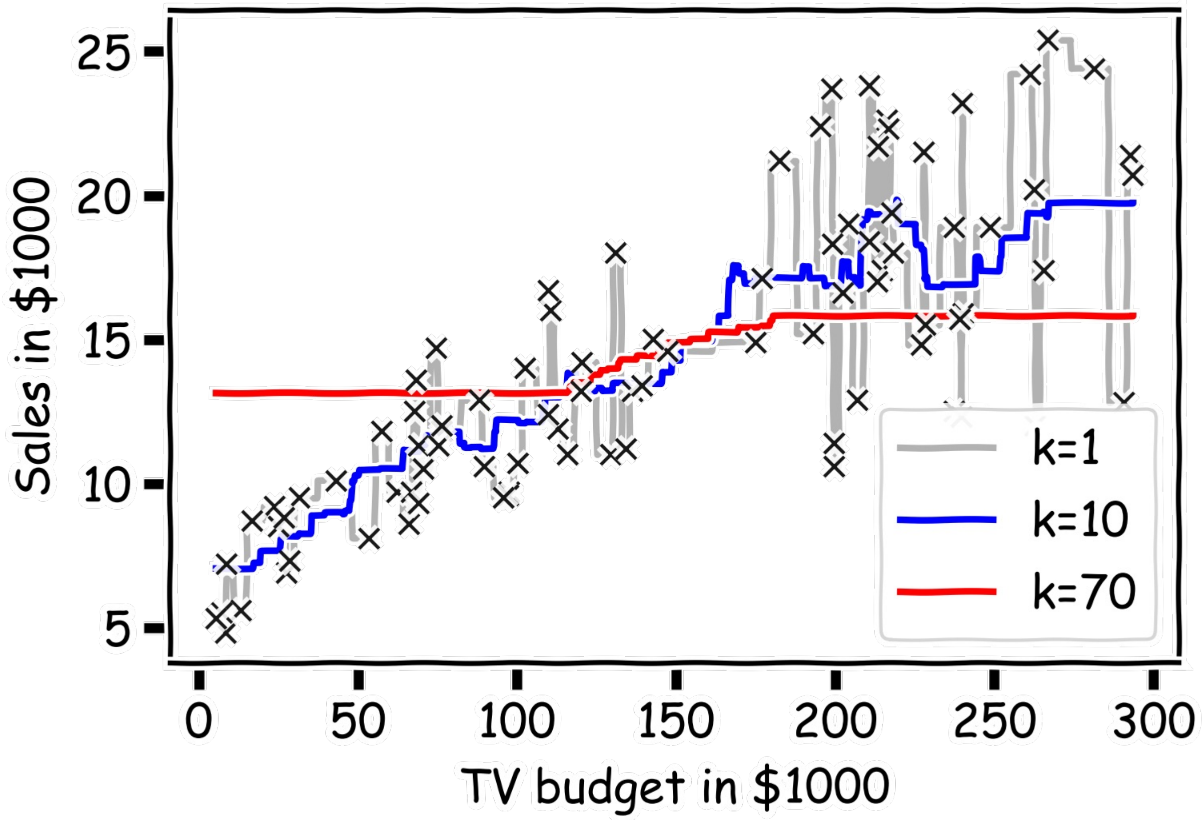
Simple Prediction Models



This example and all material are provided by the Data Science course at Harvard University

Simple Prediction Models

We can try different k-models on more data



k-Nearest Neighbors

The ***k-Nearest Neighbor (kNN) model*** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most like it*

kNN is a nonparametric learning algorithm. When we say a technique is nonparametric , it means that it does not make any assumptions on the underlying data distribution.

Note: this strategy can also be applied in classification to predict a categorical variable. But classification is not in the scope of this course.

TU Delft

k-Nearest Neighbors

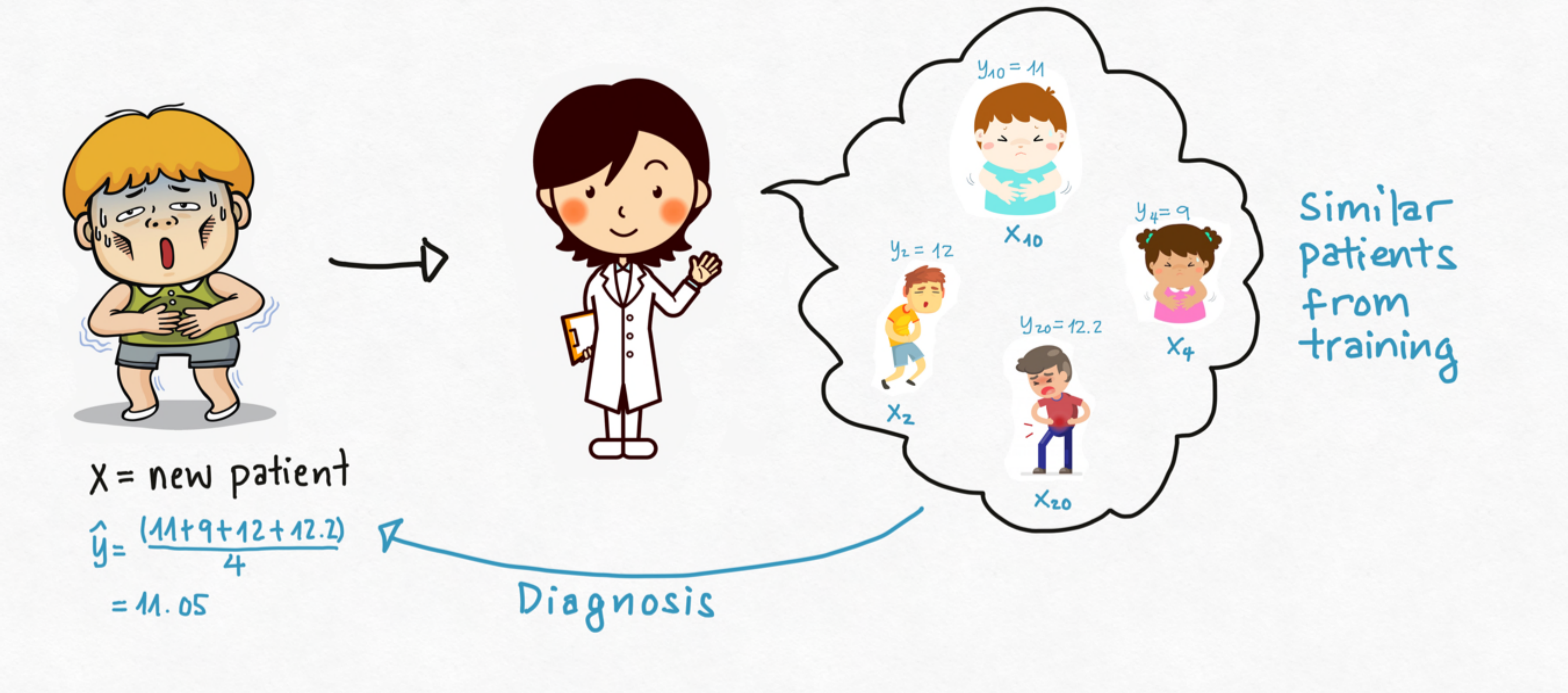


Image taken from the Data Science course at Harvard University

k-Nearest Neighbors

The very human way of decision making by similar examples. kNN is a nonparametric learning algorithm.

The k-Nearest Neighbor Algorithm:

Given a dataset $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$. For every new X:

1. Find the k-number of observations in D most like X:

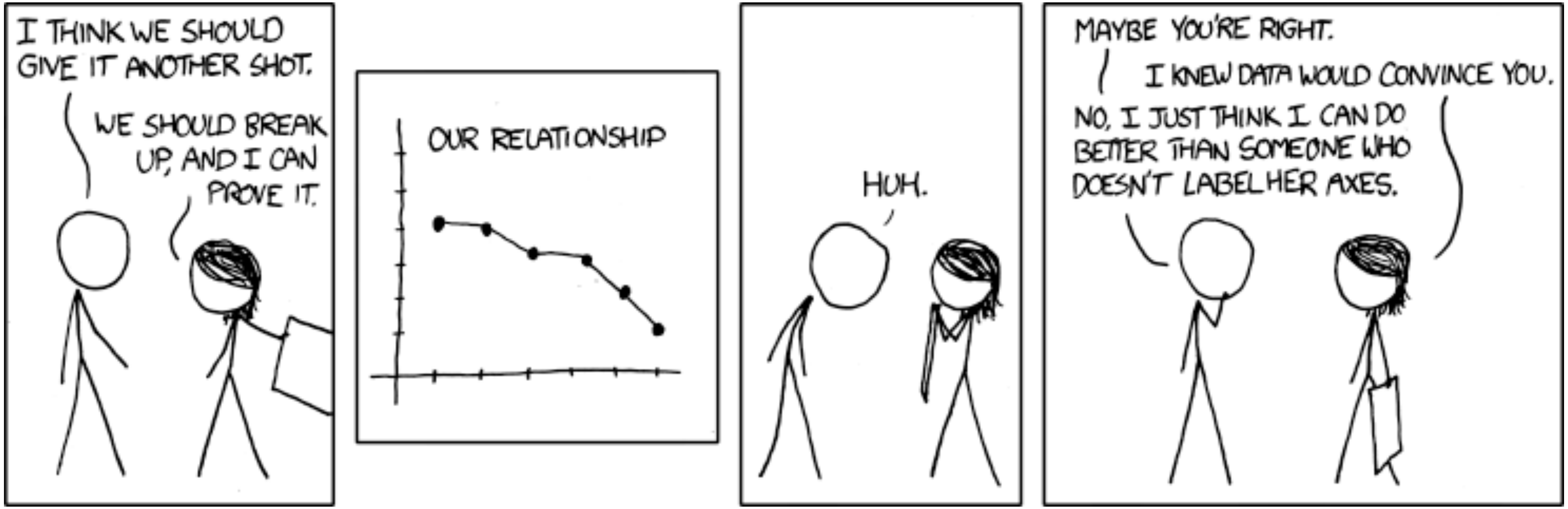
$$\{(x^{(n_1)}, y^{(n_1)}), \ldots, (x^{(n_k)}, y^{(n_k)})\}$$

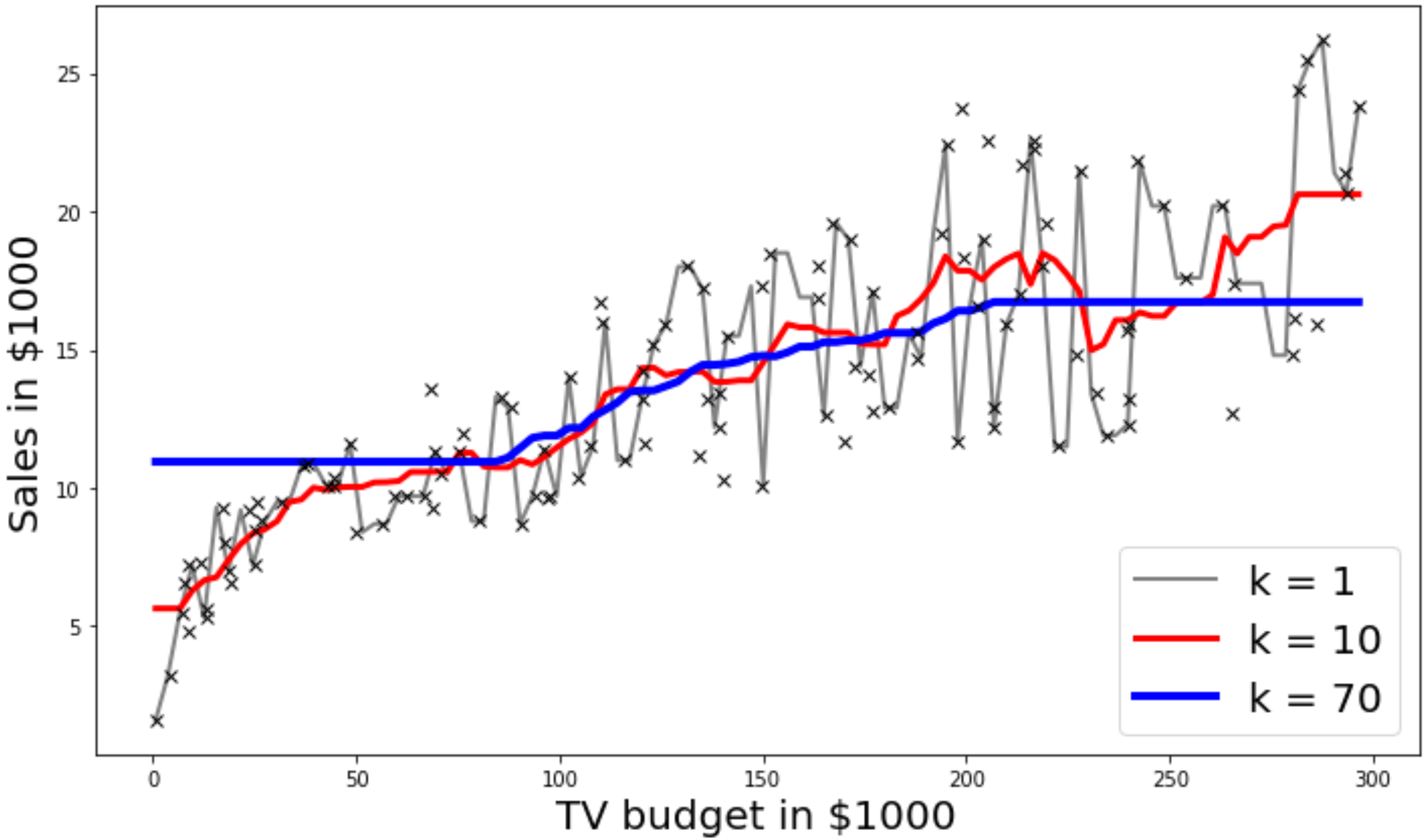
   These are called the k-nearest neighbors of x

2. Average the output of the k-nearest neighbors of x

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} y^{(n_k)}$$

Error Evaluation and Model Comparison


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

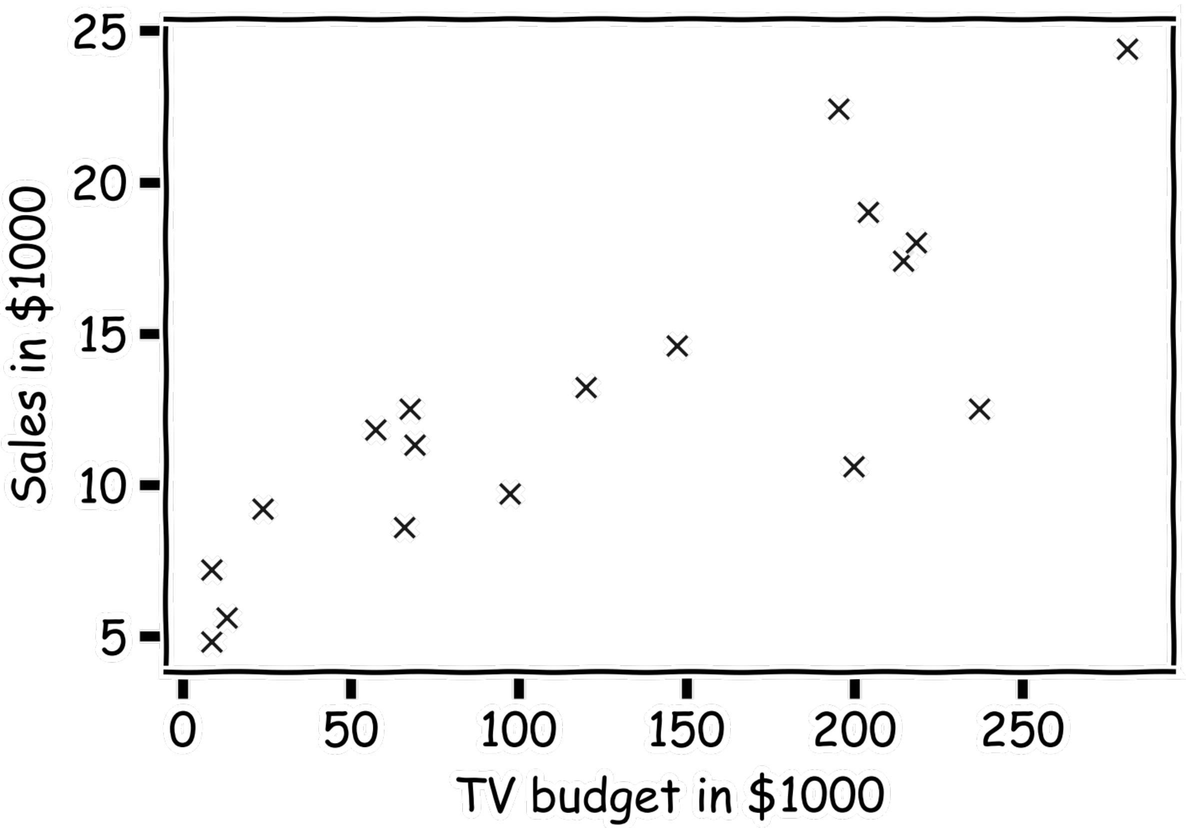


25
20
15
10
5
Sales in $1000
0
50
100
150
200
250
300
TV budget in $1000
k = 1
k = 10
k = 70

Error Evaluation

Error Evaluation

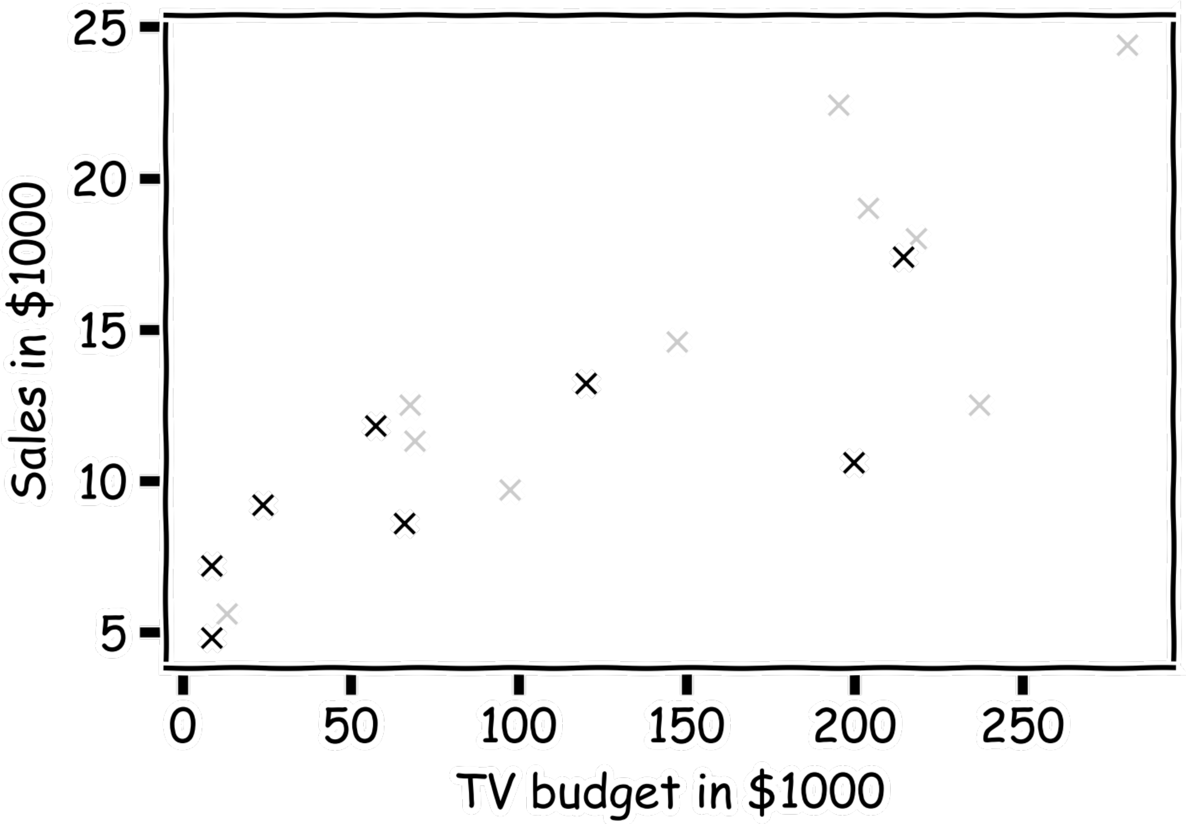
Start with some data.



This example and all material are provided by the Data Science course at Harvard University

Error Evaluation

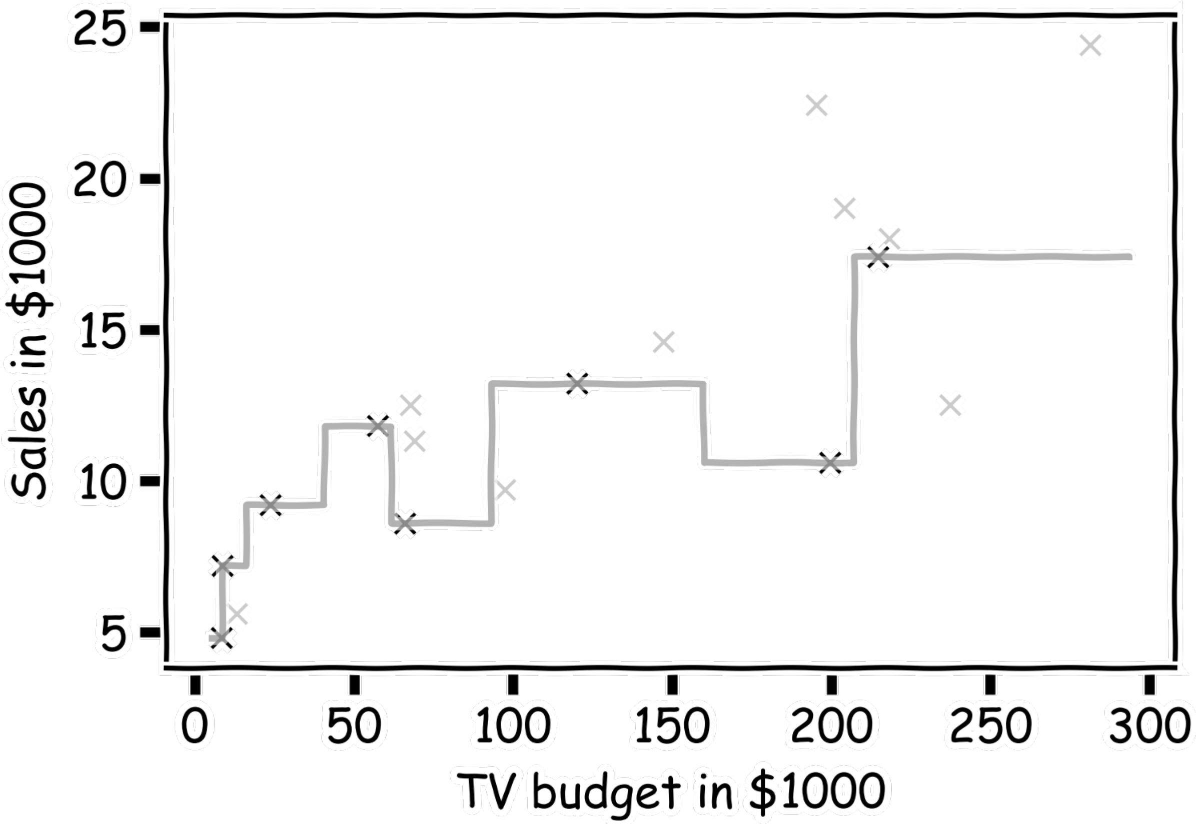
Hide some of the data from the model. This is called **train-test** split.



We use the train set to estimate $\hat{y}$, and the test set to evaluate the model.

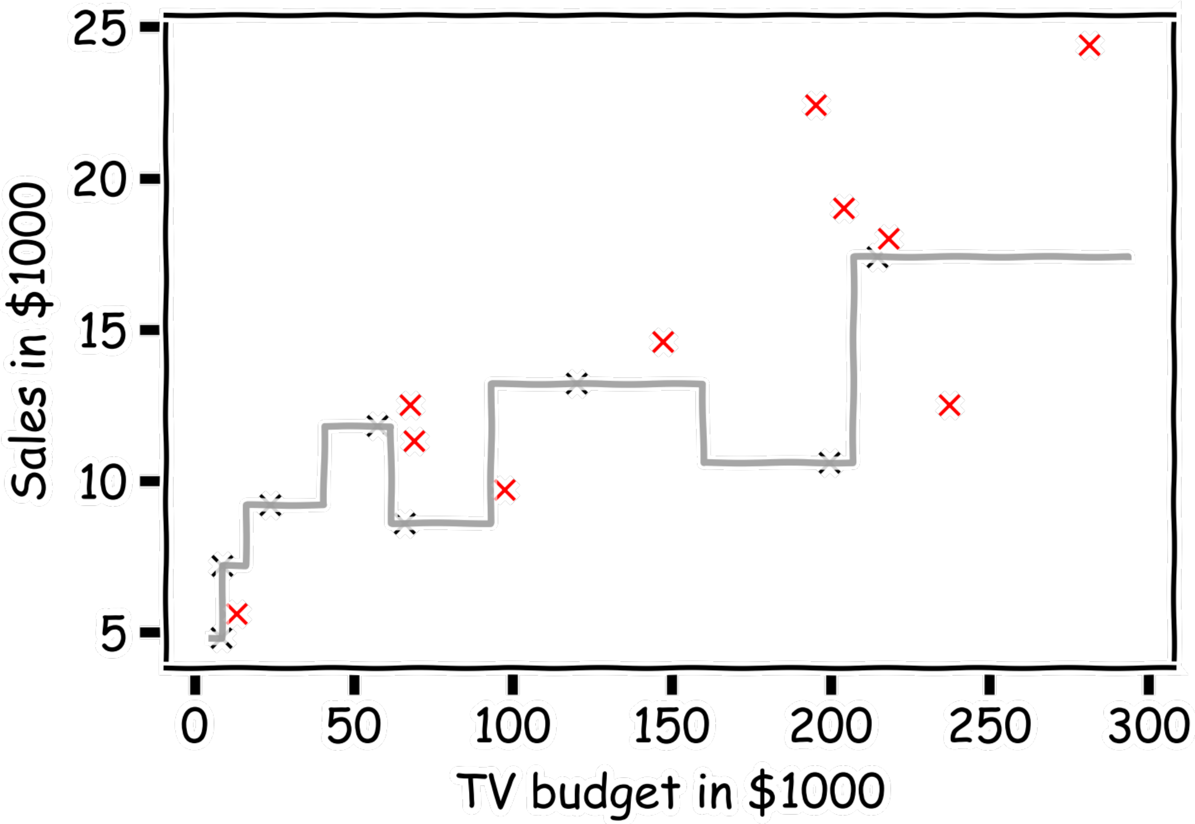
Error Evaluation

Estimate $\hat{y}$ for *k=1* .



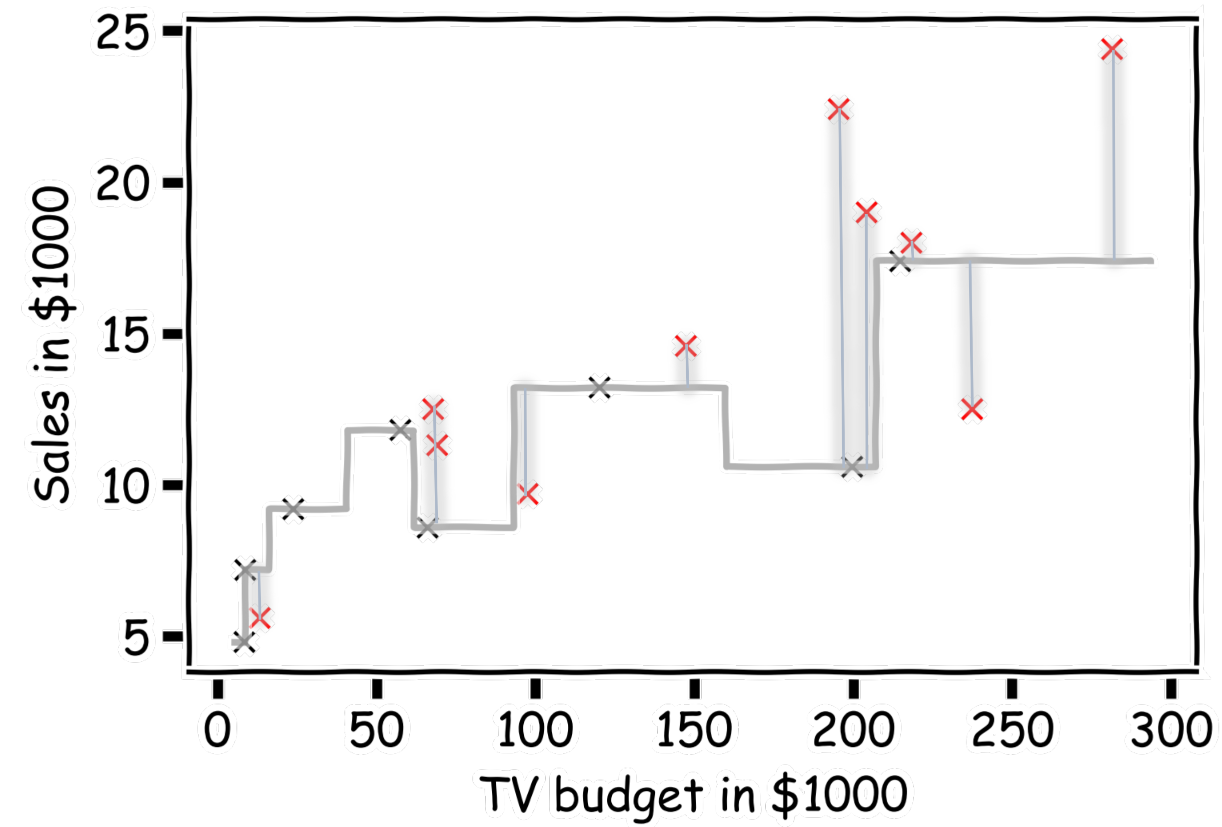
Error Evaluation

Now, we look at the data we have not used, the **test data** (red crosses).



Error Evaluation

Calculate the **residuals** $(y_i - \hat{y}_i)$.



This example and all material are provided by the Data Science course at Harvard University

Error Evaluation

In order to quantify how well a model performs, we aggregate the errors, and we call that the ***loss*** or ***error*** *or* ***cost function.***

A common loss function for quantitative outcomes is the **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Note: Loss and cost function refer to the same thing. Cost usually refers to the total loss where loss refers to a single training point.

Error Evaluation

Caution: The MSE is by no means the only valid (or the best) loss function!

1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error

We can choose other functions here (but out of scope of this course).

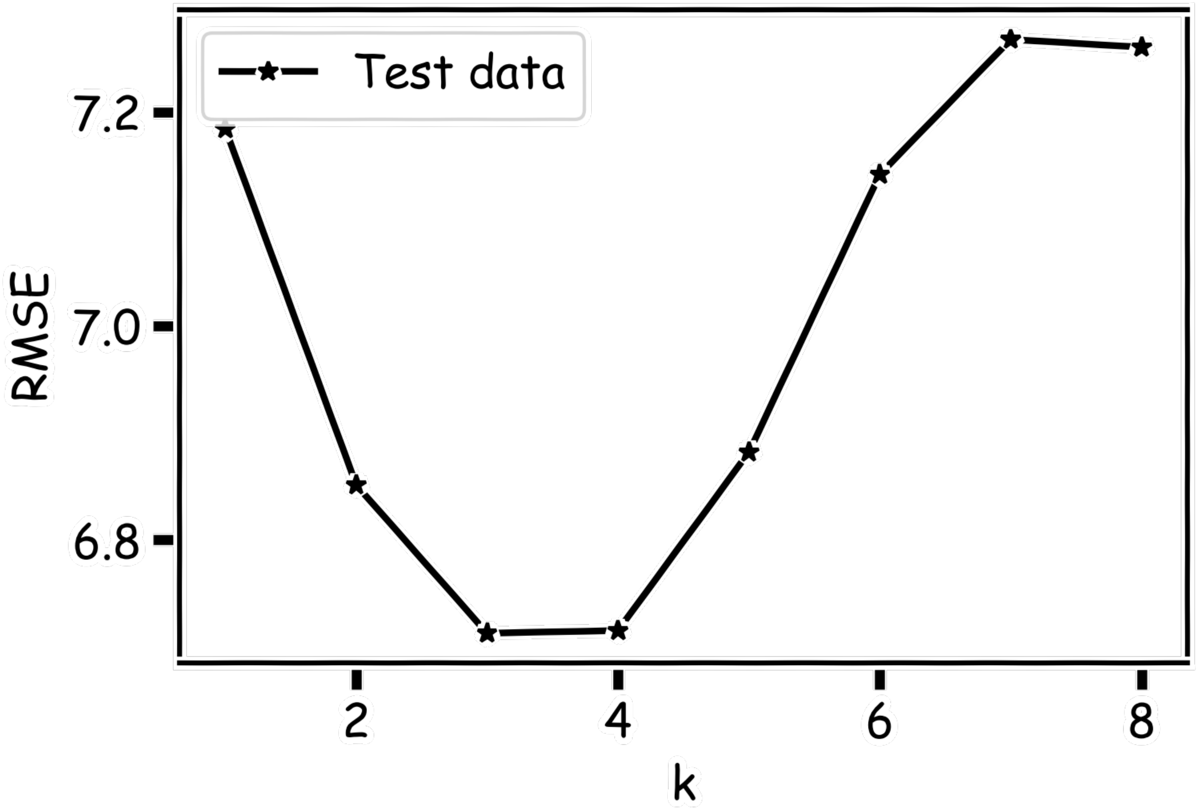
Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Model Comparison

Model Comparison

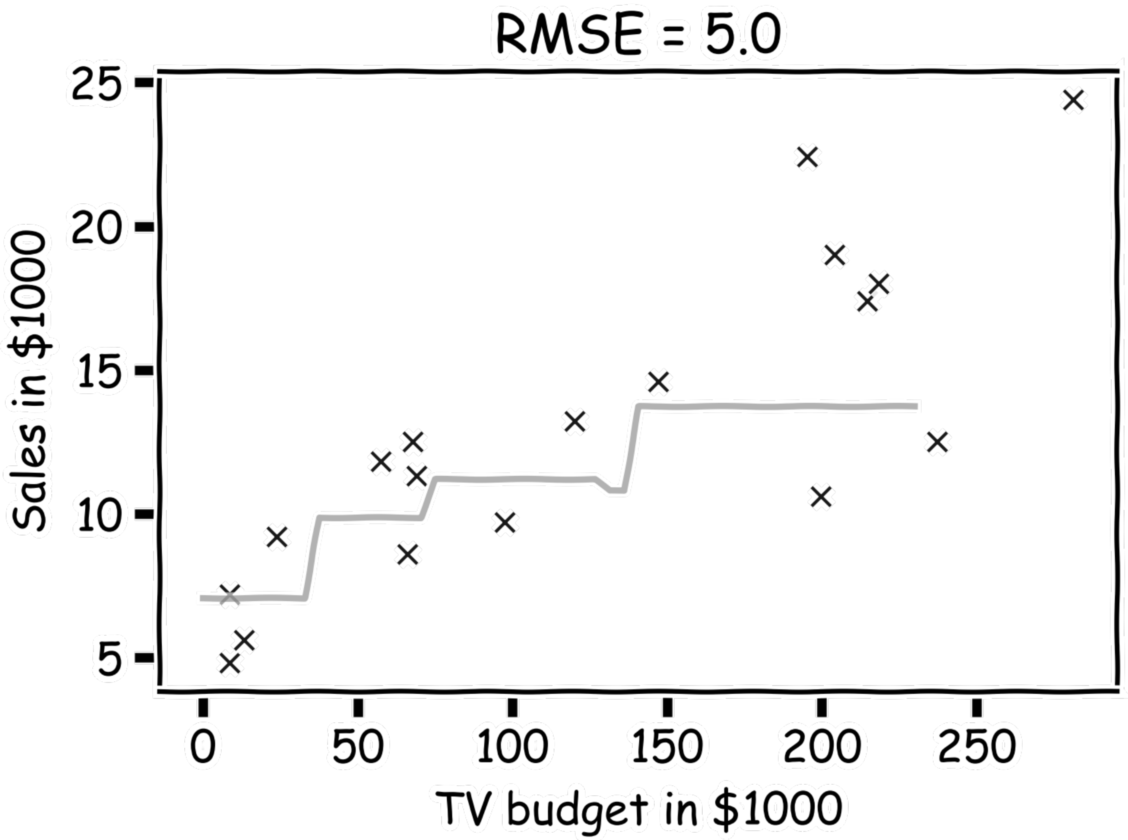
Do the same for all *k*'s and compare the RMSEs. *k=3* seems to be the **best model**.



Model Fitness

Model fitness

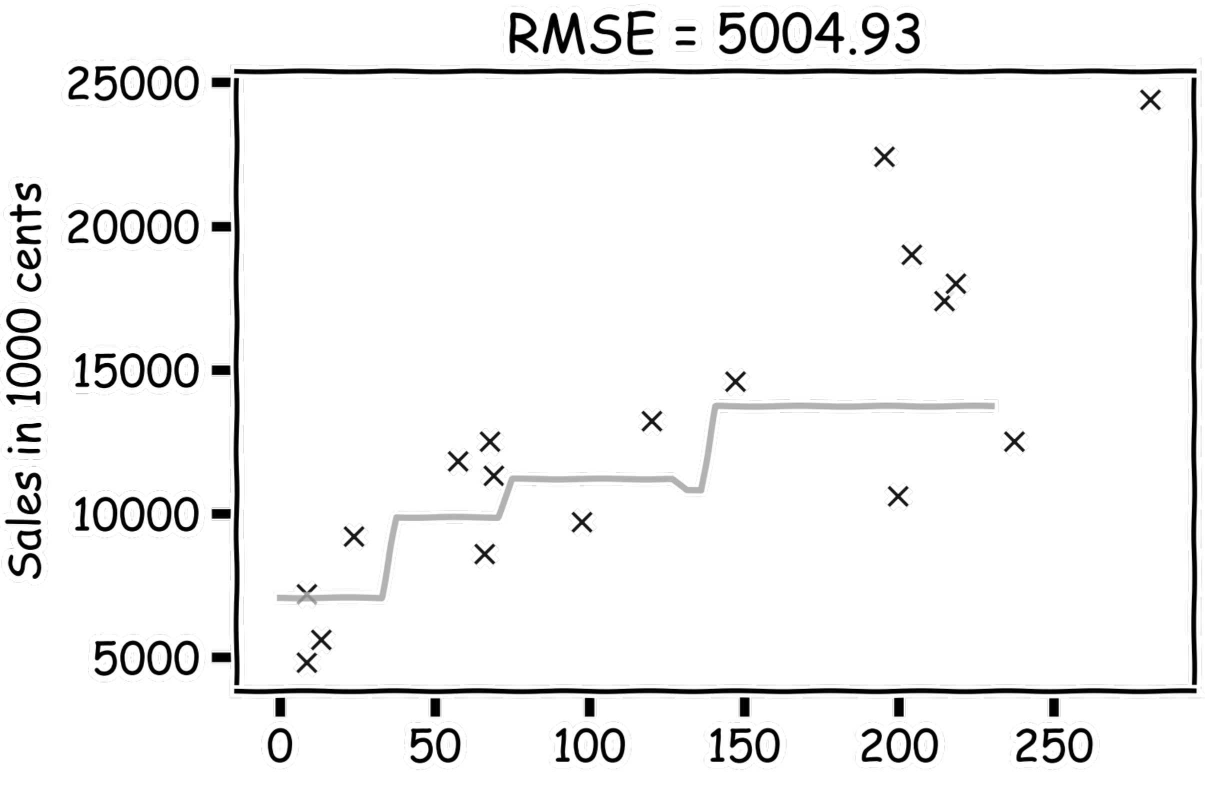
For a subset of the data, calculate the RMSE for *k=3.*



Is RMSE = 5.0 good enough?

Model fitness

What if we measure the Sales in cents instead of dollars?

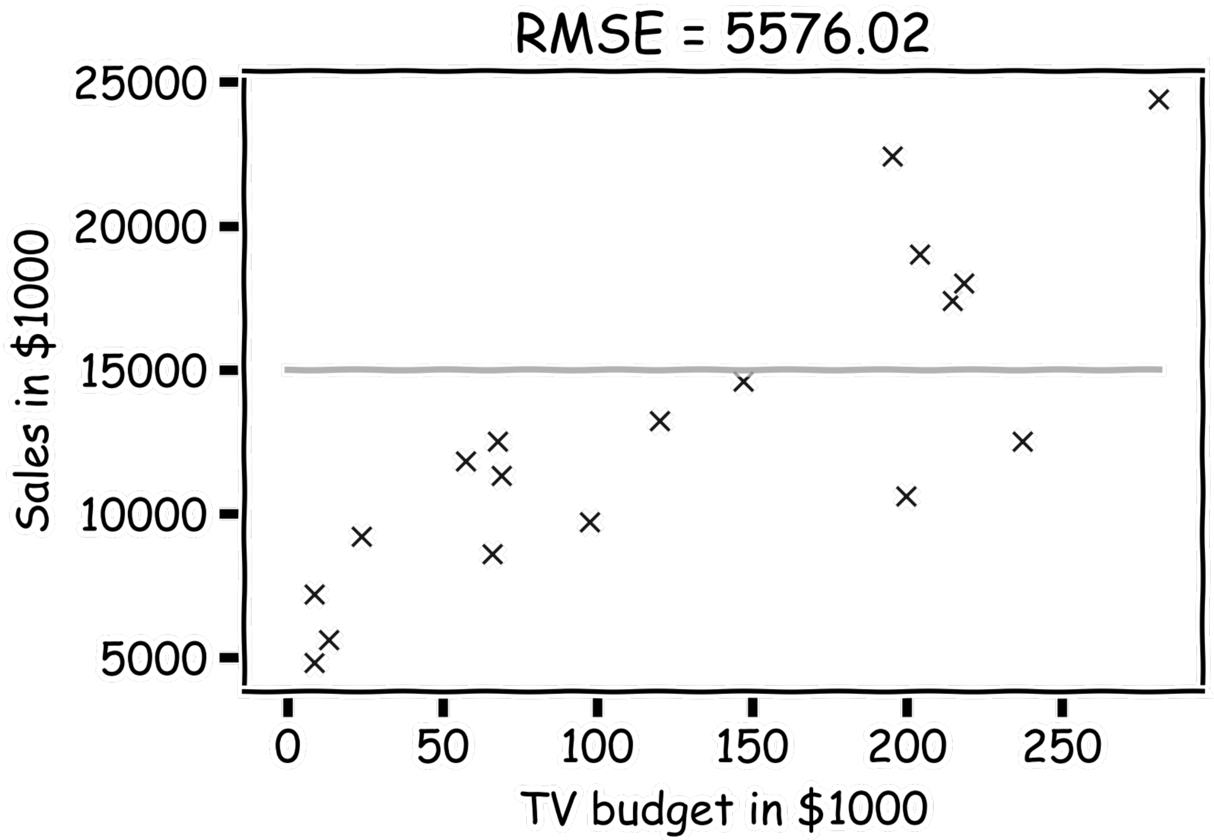


RMSE is now 5004.93.

Is that good?

Model fitness

It is better if we compare it to something.



We will use the simplest model:

$$\hat{y} = \bar{y} = \frac{1}{n}\sum_i y_i$$

as the worst possible model and

$$\hat{y_i} = y_i$$

as the best possible model.

R-squared

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

- If our model is as good as the mean value, $\bar{y}$, then $R^2 = 0$
- If our model is perfect, then $R^2 = 1$
- R^2 can be negative if the model is worst than the average. This can happen when we evaluate the model in the test set.
- It explains the percentage of variance in y explained by our features x.

We will use the simplest model:

$$\hat{y} = \bar{y} = \frac{1}{n}\sum_i y_i$$

as the worst possible model and

$$\hat{y}_i = y_i$$

as the best possible model.

Recapitulation

Machine learning is for everyone, and it only predicts results based on incoming data!

For next class..

Finish Labs to practice programming

Complete Homework for more practice

Check Assignment contents and due date

See “To do before class” for next lecture (~ 1 hour of self-study)