Introduction to Open & Reproducible Science (IORDS)

Michael Dayan, Data Scientist Manager

Methods & Data facility
Human Neuroscience Platform
Foundation Campus Biotech Geneva

Virtual machine info

To get an IP, please fill the form at:
https://tinyurl.com/IORDS2021-IP-ML1

START VS CODE AND JUPYTER ON YOUR BROWSER:

- *Start an internet browser on your own machine*
- *VS Code:* `<your_IP>:8080`
- *Jupyter:* `<your_IP>:8888`

PASSWORD: braincode!

Connecting your:	WIFI SSID	WIFI Password
Laptop (no phones)	NIDS_course	reproduciblescience
Phone	CAMPUS_VISITORS	welcomecampus

PLEASE CONNECT TO THE VM
→ Login: brainhacker
→ Password: brainhack!

Connect to Slack and download the exercise slides

ANY PROBLEM? Please raise your hand or ask questions on Slack: channel **#machine-learning**

On site support (including coding):


Maël


Nathan

Remote support (including coding):


Serafeim

HNP METHODS & DATA

LECTURE OBJECTIVES

Introduction to machine learning lectures objectives (you should be able to…):

- Understand the typical form of the data characterizing a machine learning problem (N samples x p features, with p labels for supervised machine learning)
- Understand what fitting a model means
- Understand how to score predictions and why an independent test set is essential
- Know how to implement cross-validation, and know why this is useful
- Know how to implement regularization, and know when it can be useful

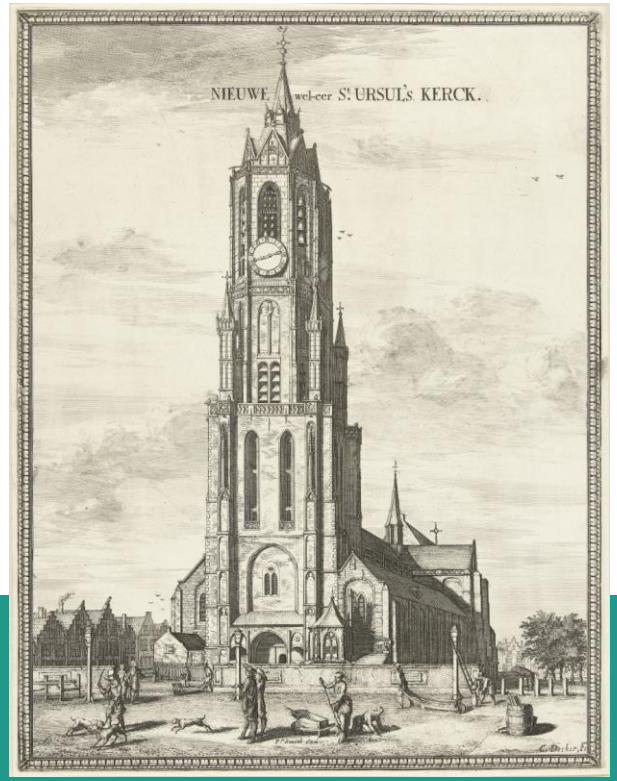
ML
Part 1

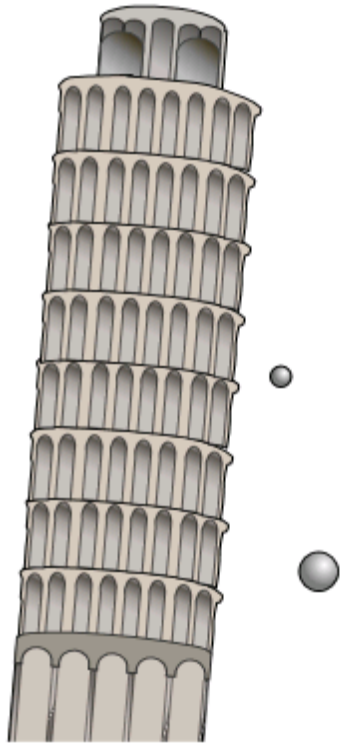
SOME HALLMARKS OF SCIENCE

➢ Testable predictions

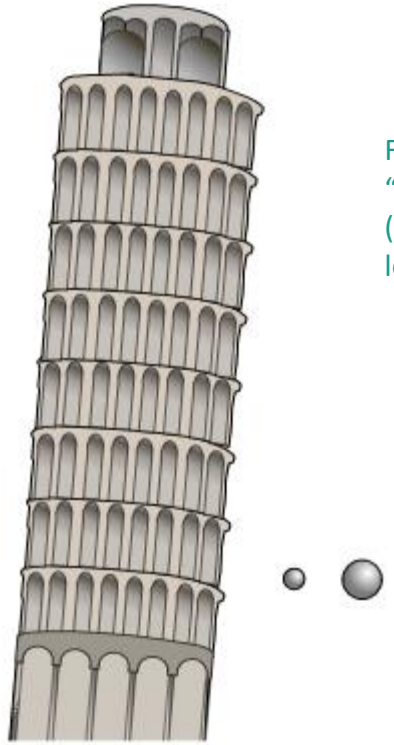
Aristotle's theory of gravity: objects fall at a speed proportional to their mass

Galileo's theory of gravity: objects fall at the same speed





Aristotle’s theory



Galileo’s theory

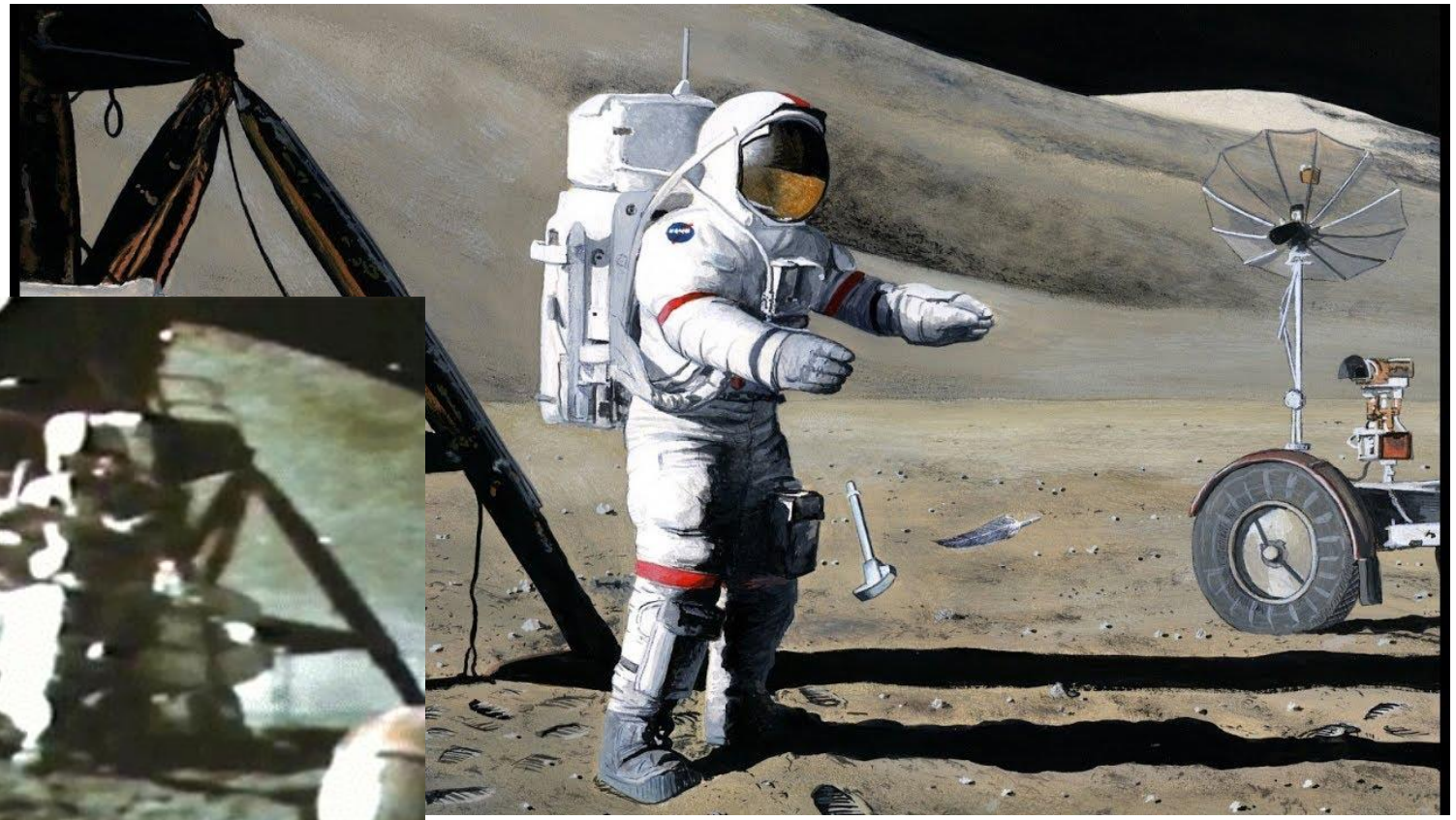
Figure adapted from “Layers of Learning” (https://layers-of-learning.com)

Simon Stevin & Jan Cornets de Groot in 1586:
Two balls with different weights, dropped from a height of 30 feet
[Coenraet Decker, Pieter Smith, Arnold Bon (1667)]

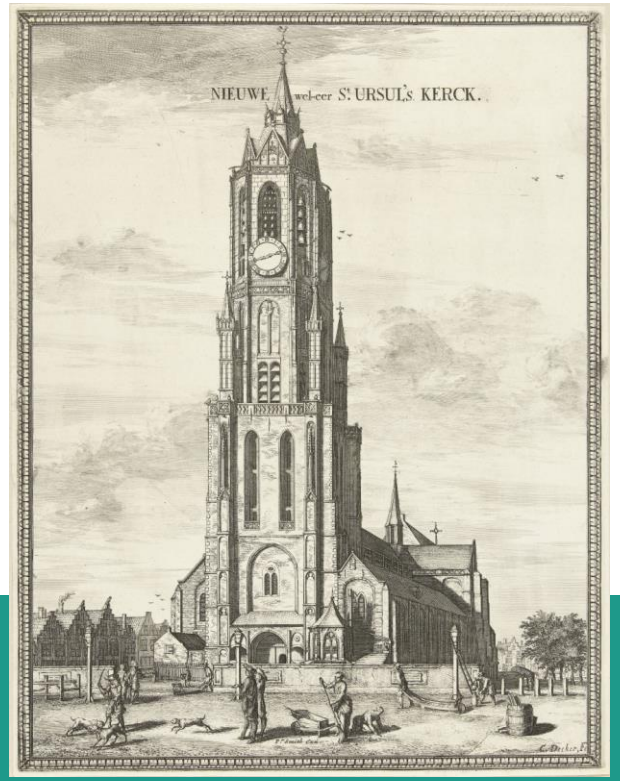
SOME HALLMARKS OF SCIENCE

- Generalization

  Derive findings that also apply to other experiments

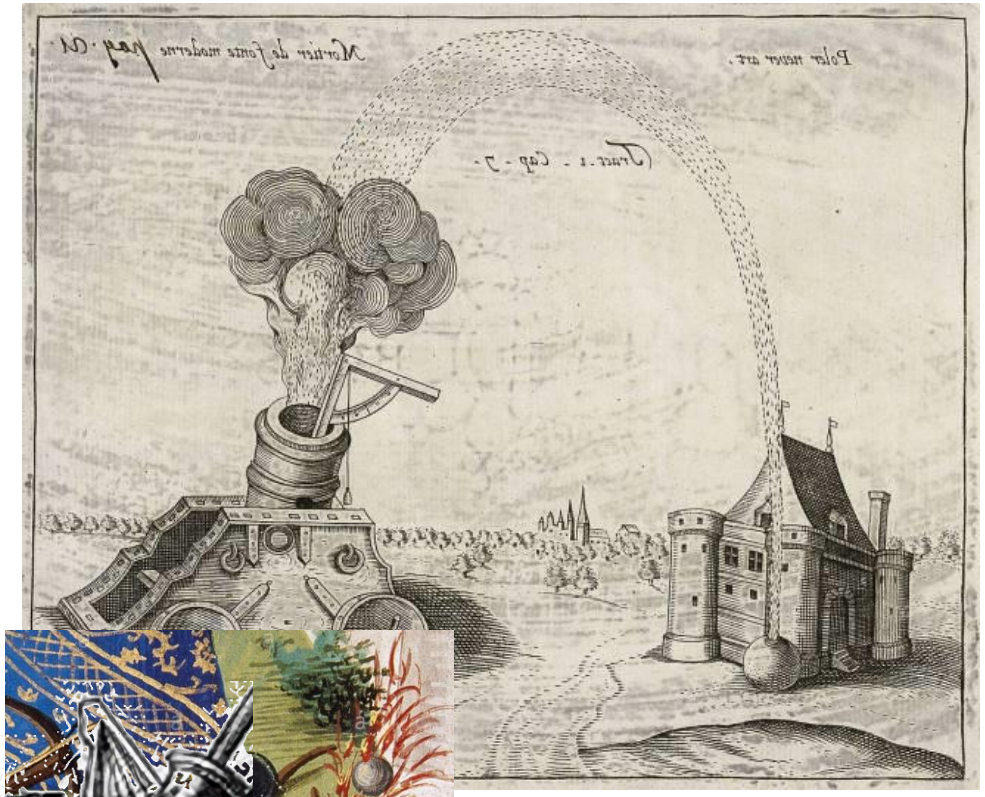


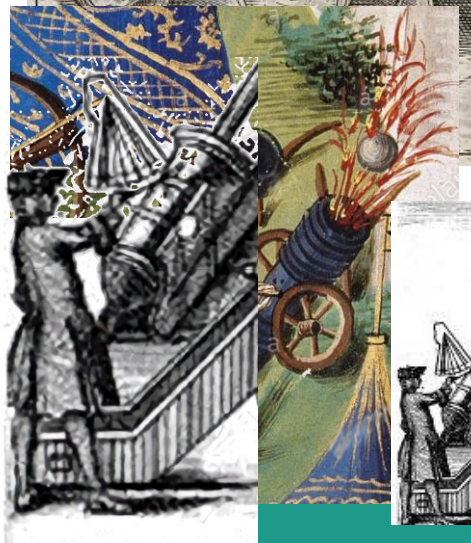
David Scott (taking part in NASA's Apollo 15 mission) reproducing on the moon an experiment demonstrating objects falling at the same speed. He released simultaneously a hammer and a feather from the same height: they landed at the same time.

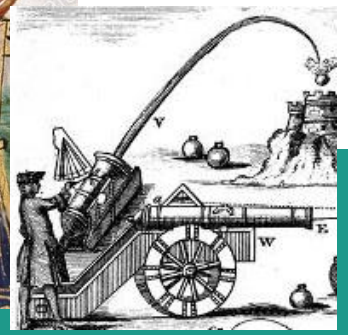


Simon Stevin & Jan Cornets de Groot in 1586:
Two balls with different weights, dropped from a height of 30 feet
[Coenraet Decker, Pieter Smith, Arnold Bon (1667)]

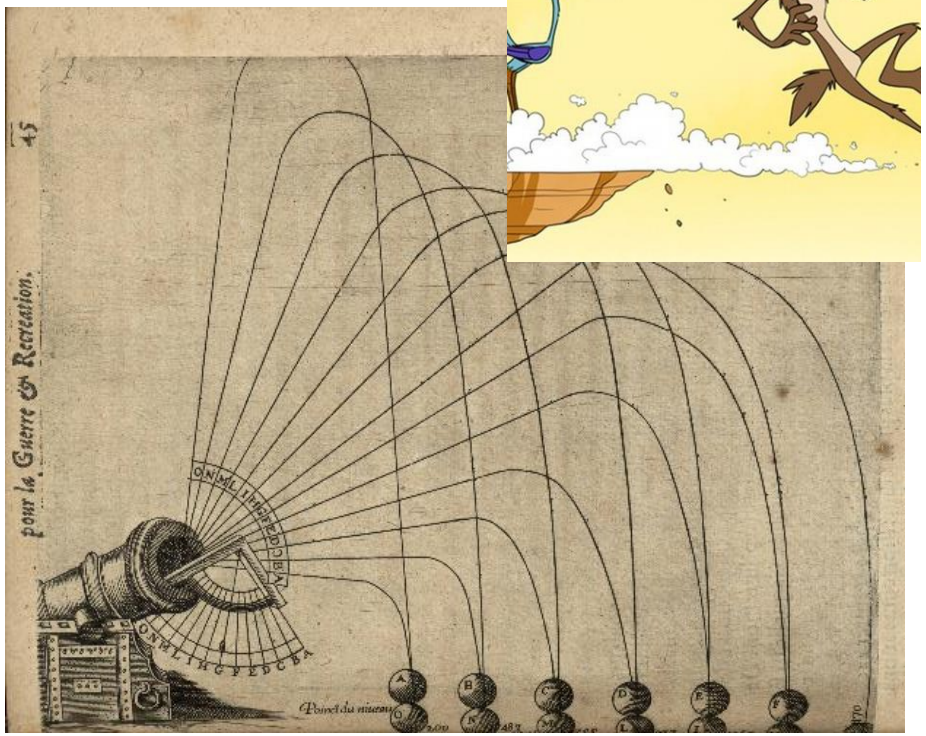
EXAMPLE OF MACHINE LEARNING APPLIED TO PROJECTILES



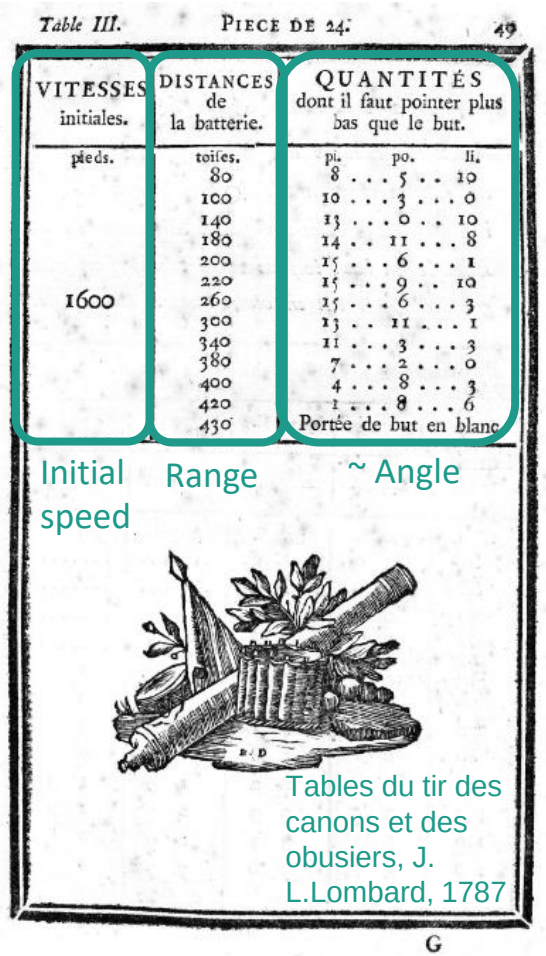




Mortier de fonte moderne, J. T. de Bry, 1613



La pyrotechnie de Hanzelet Lorrain, J. A. Hanzelet, 1630



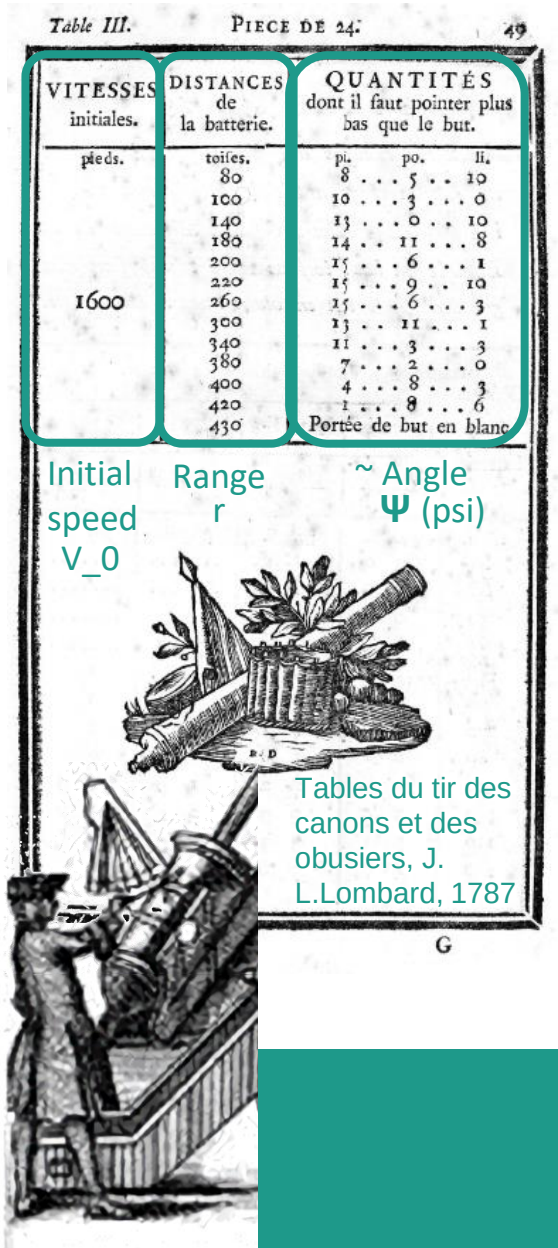
Initial speed

Range

~ Angle

Tables du tir des canons et des obusiers, J. L.Lombard, 1787

EXAMPLE OF MACHINE LEARNING APPLIED TO PROJECTILES



- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)

The choice of outcome(s) (and features) is fully part of the research design

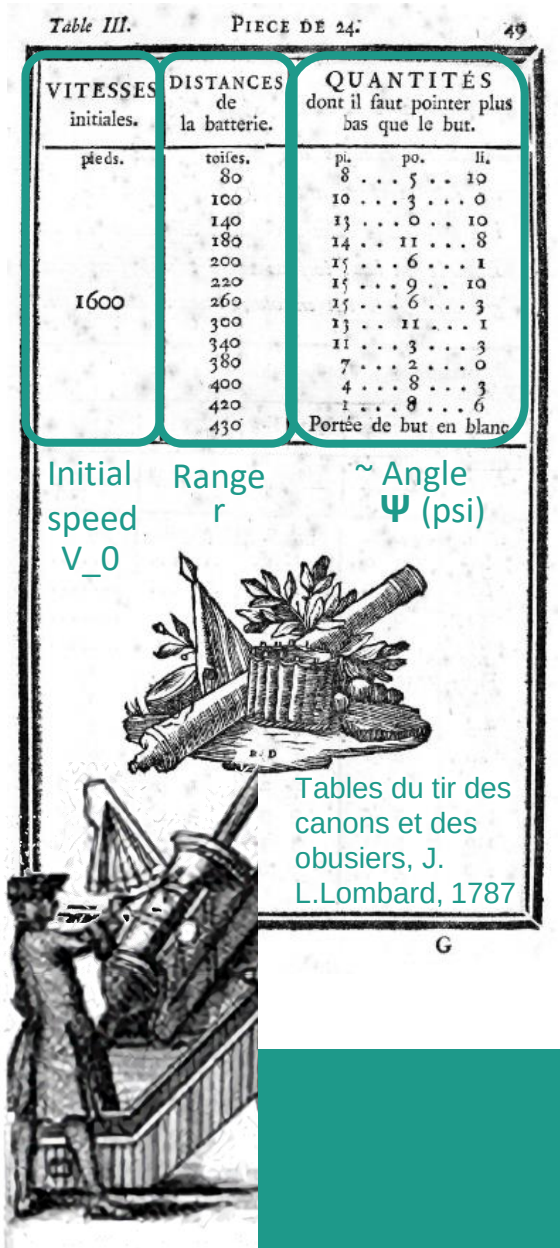
DATASET

Features | Labels (outcomes)

Observations (data points / samples)

	Angle Ψ (psi)	Initial speed	Mass	Radius	…	Hit Target at 100 m
Shot 1	60	20	3	0.10	…	No
Shot 2	45	20	3	0.10	…	Yes
Shot 3	45	20	5	0.12	…	No
Shot 4	35	20	4	0.08	…	Yes
…	…	…	…	…	…	…

CHOICE OF OUTCOME (AND FEATURES)



- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)

The choice of outcome(s) (and features) is fully part of the research design

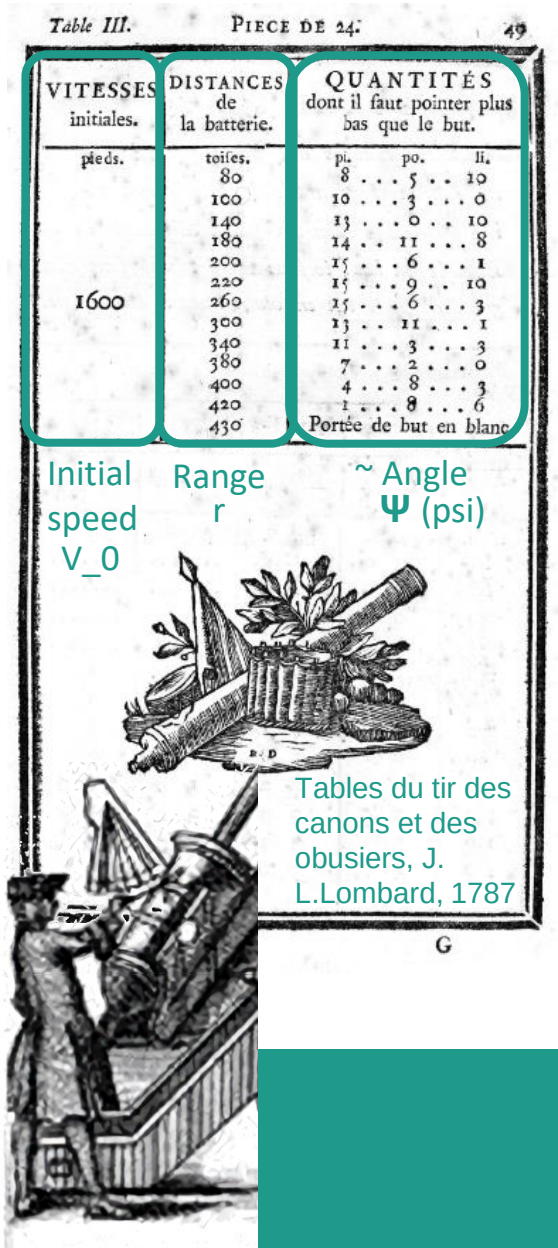
DATASET

Features | Labels (outcomes)

Observations (data points / samples)

	Age	Sex	ROI 1	ROI 2	…	Has disease
Subj 1	60	F	42.0	0.15	…	No
Subj 2	45	M	29.1	0.11	…	Yes
Subj 3	45	F	31.7	0.12	…	No
Subj 4	35	F	25.4	0.14	…	Yes
…	…	…	…	…	…	…

CHOICE OF OUTCOME (AND FEATURES)



Initial speed V_0

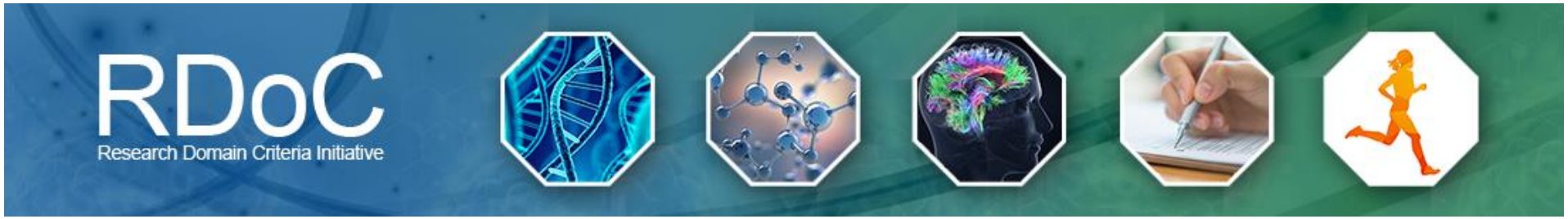
Range r

~ Angle Ψ (psi)

Tables du tir des canons et des obusiers, J. L.Lombard, 1787

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)

The choice of outcome(s) (and features) is fully part of the research design



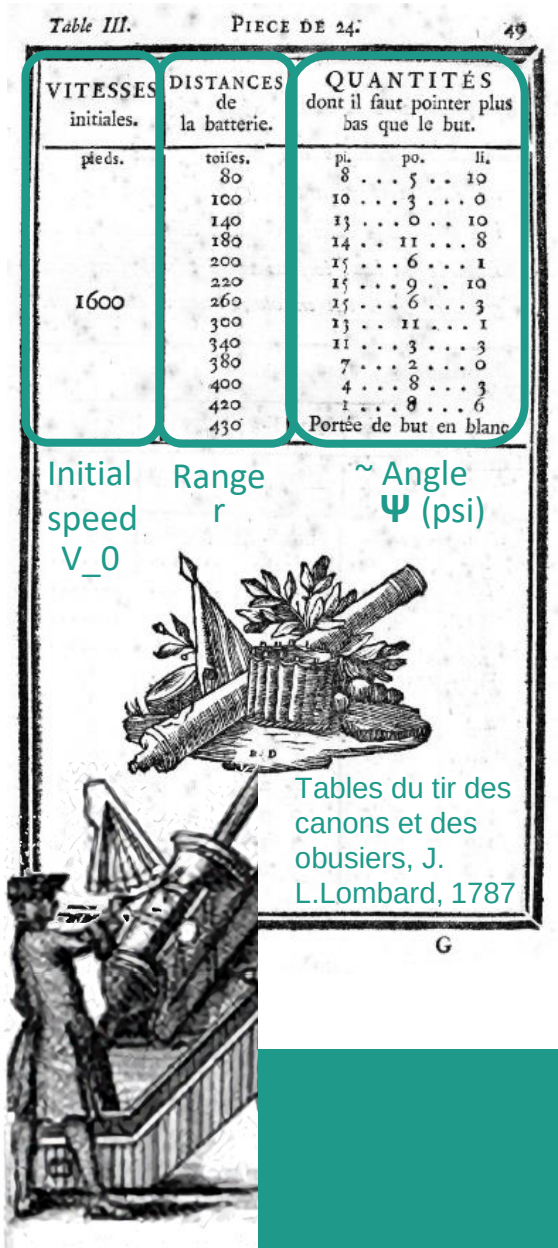
Binary diagnostic can impair understanding of disease:

- People with the same diagnostic can have different symptoms
- People with given symptoms can be likely to have an additional disorder
- Difficulty of clear diagnosis may cause to exclude patients from studies
- Criteria to be diagnose with a disorder can be arbitrary

➔ Choose biological, physiological, and behavioral dimensions as outcome

https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc.shtml

EXAMPLE OF MACHINE LEARNING APPLIED TO PROJECTILES



- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)

The choice of outcome(s) (and features) is fully part of the research design

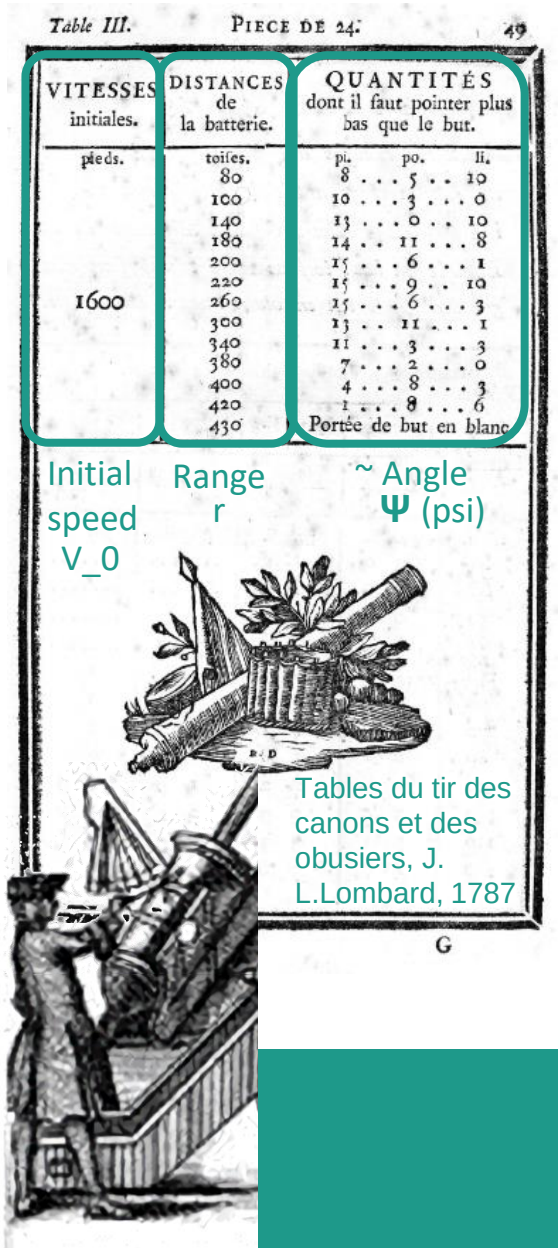
DATASET

Features | Labels (outcomes)

Observations (data points / samples)

	Angle Ψ (psi)	Initial speed	Mass	Radius	…	Hit Target at 100 m
Shot 1	60	20	3	0.10	…	No
Shot 2	45	20	3	0.10	…	Yes
Shot 3	45	20	5	0.12	…	No
Shot 4	35	20	4	0.08	…	Yes
…	…	…	…	…	…	…

EXAMPLE OF MACHINE LEARNING APPLIED TO PROJECTILES



- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)

The choice of outcome(s) (and features) is fully part of the research design

DATASET

Features | Labels (outcomes)

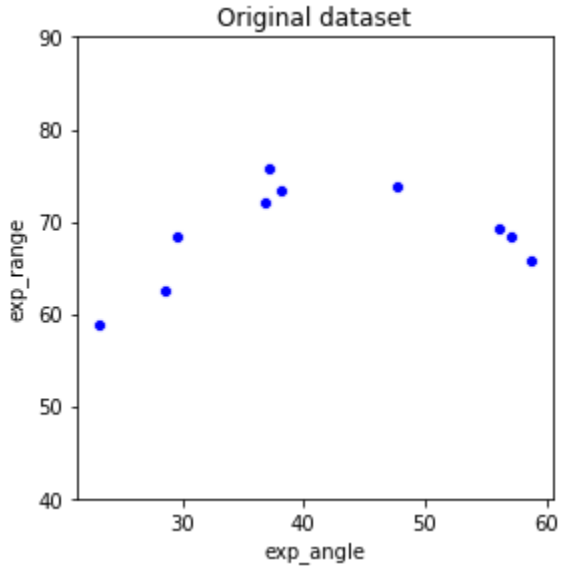
Observations (data points / samples)

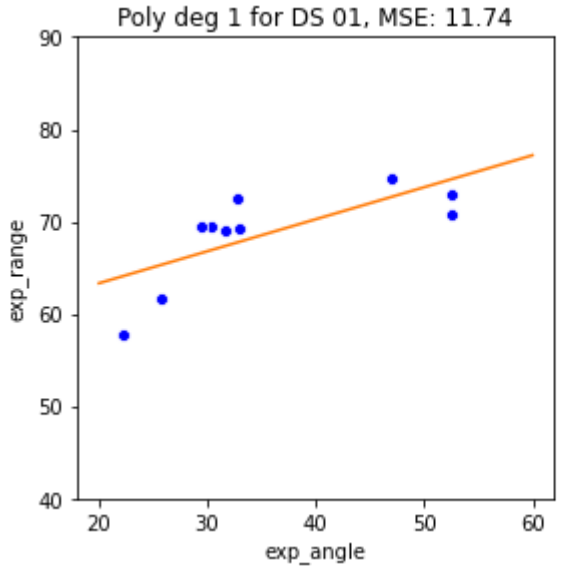
	Angle Ψ (psi)	Initial speed	Mass	Radius	…	Range r
Shot 1	60	20	3	0.10	…	32.4
Shot 2	45	20	3	0.10	…	36.2
Shot 3	45	20	5	0.12	…	37.7
Shot 4	35	20	4	0.08	…	36.4
…	…	…	…	…	…	…

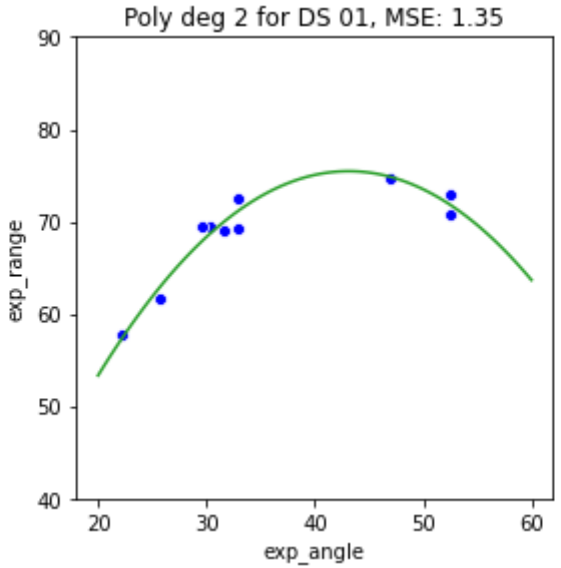
MAKING PREDICTIONS

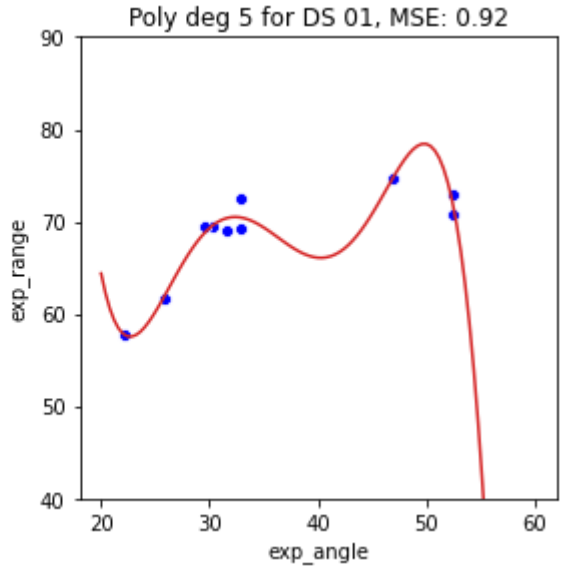
"LEAST BAD" MODEL?

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions (accuracy)





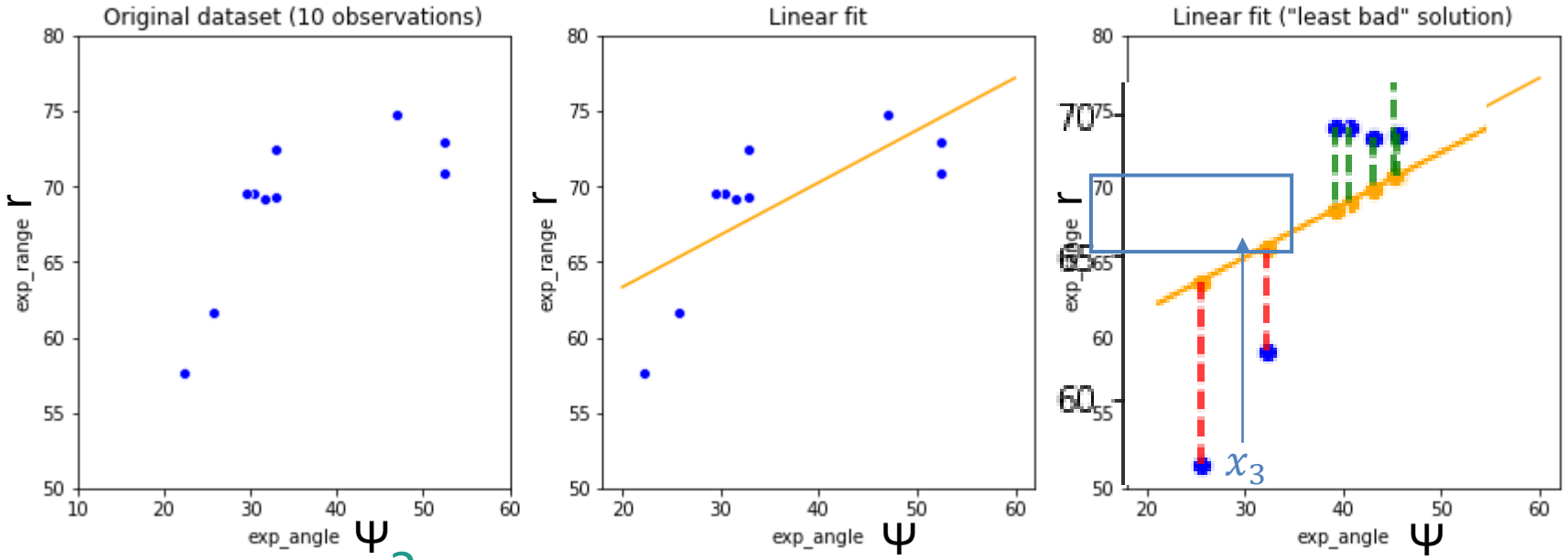


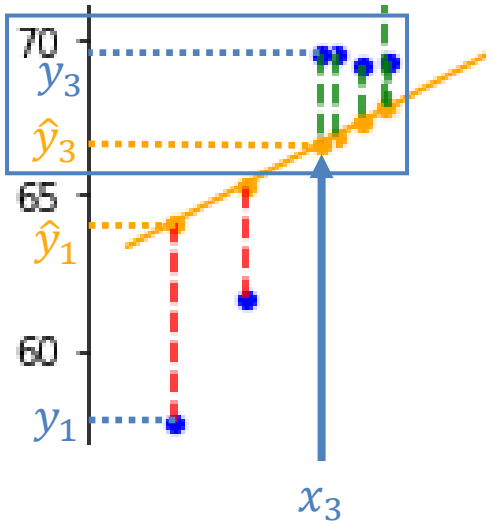


MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

"LEAST BAD" MODEL?





?

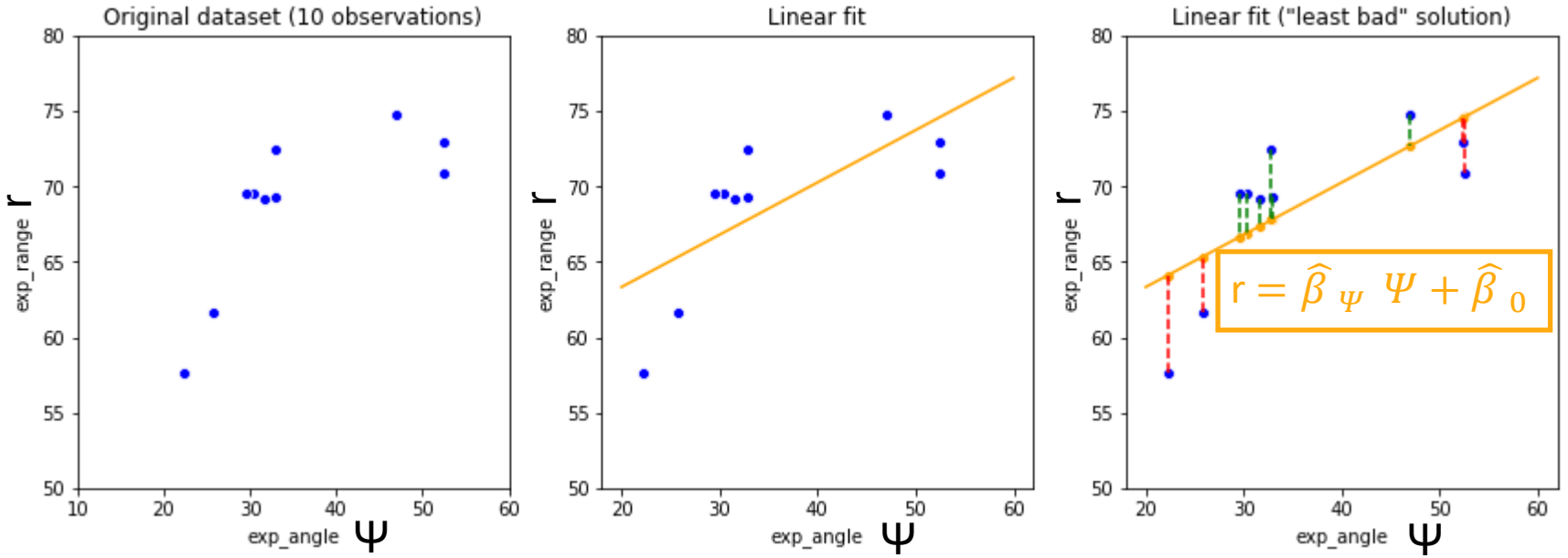
$$Error = (y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + (y_3 - \hat{y}_3) + \cdots + (y_N - \hat{y}_N)$$

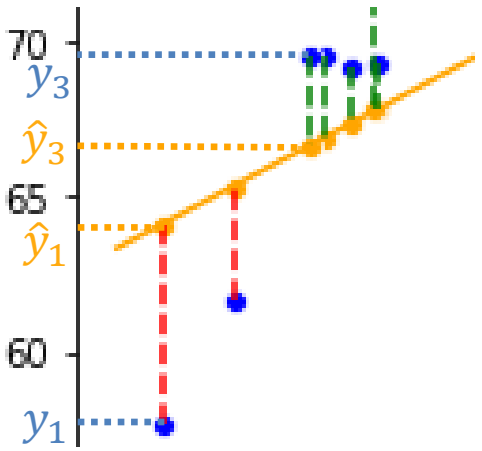
$$Squared\ Error = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \cdots + (y_N - \hat{y}_N)^2 = \sum_{i=0}^{N}(y_i - \hat{y}_i)^2$$

MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

"LEAST BAD" MODEL?





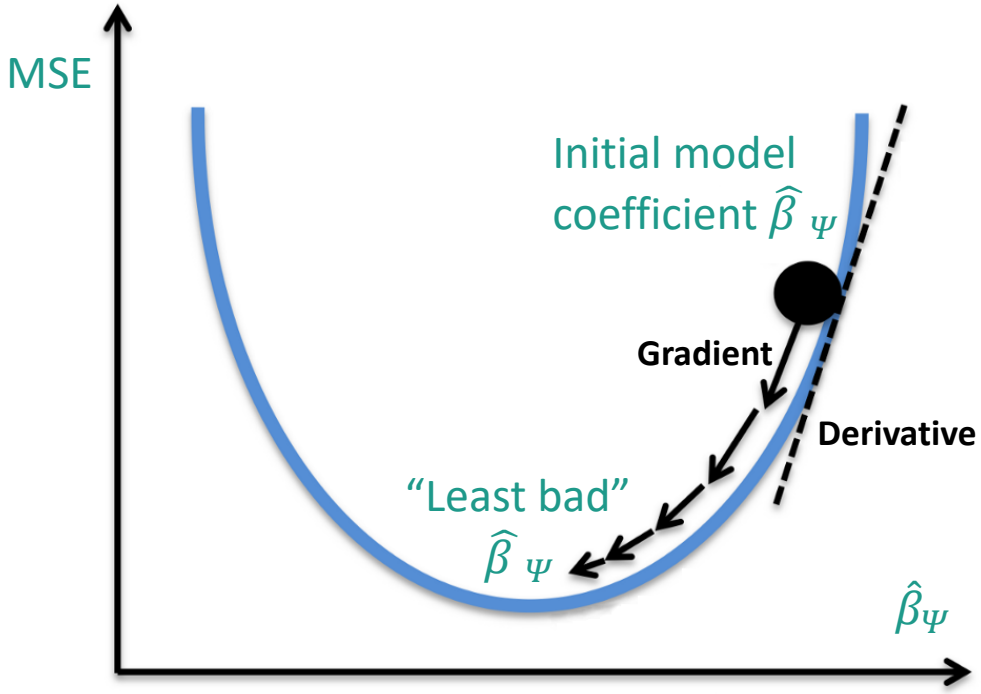
$$Mean\ Squared\ Error\ (MSE) = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2$$

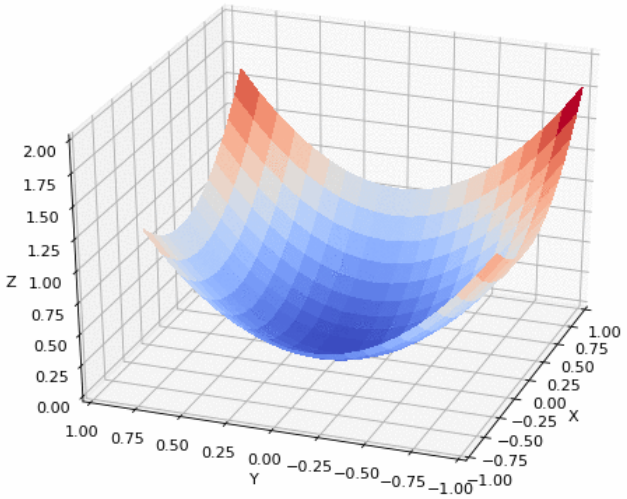
$$Error = (y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + (y_3 - \hat{y}_3) + \cdots + (y_N - \hat{y}_N)$$

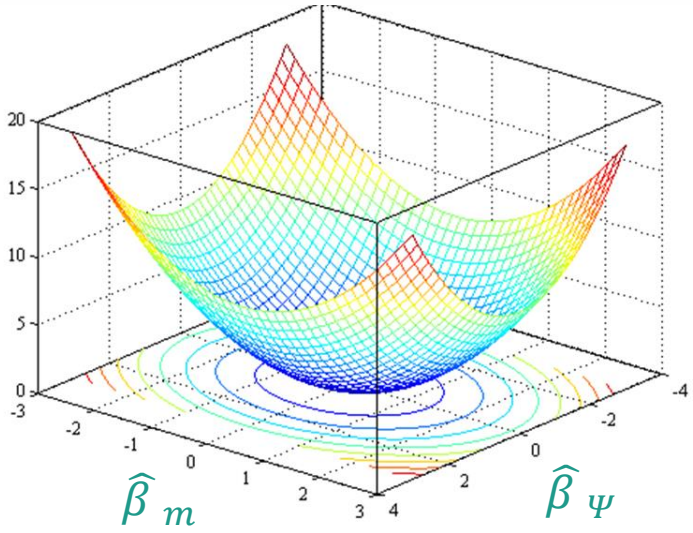
$$Squared\ Error = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \cdots + (y_N - \hat{y}_N)^2 = \sum_{i=0}^{N}(y_i - \hat{y}_i)^2$$

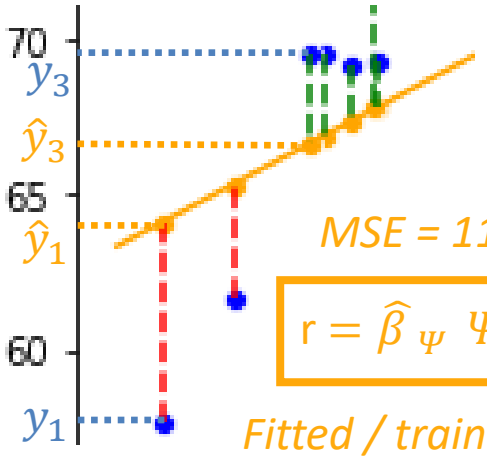
MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of prediction









MSE = 11.74 (m^2)

$r = \hat{\beta}_\Psi \Psi + \hat{\beta}_0$

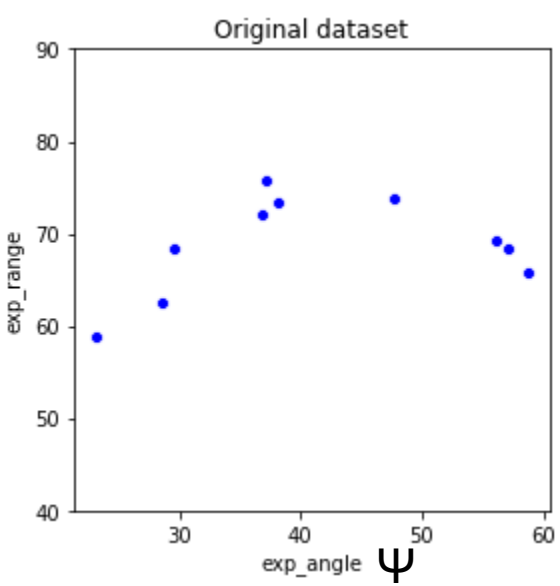
Fitted / trained model

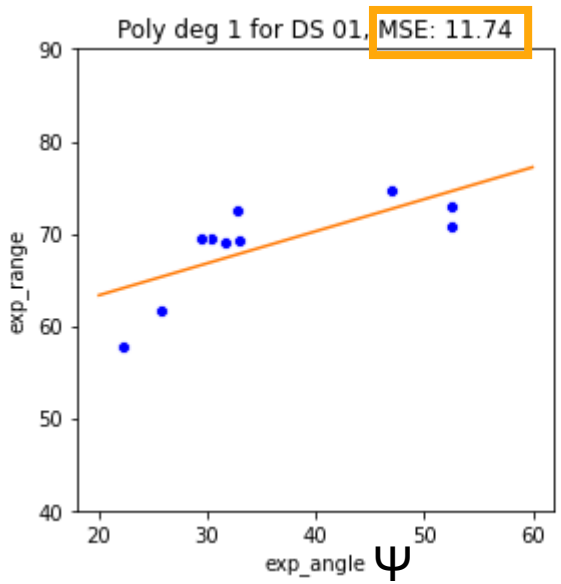
$$Mean\ Squared\ Error\ (MSE) = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2$$

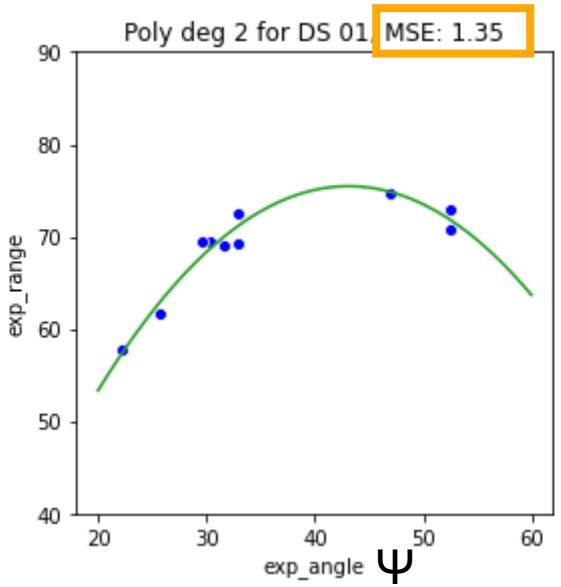
MAKING PREDICTIONS

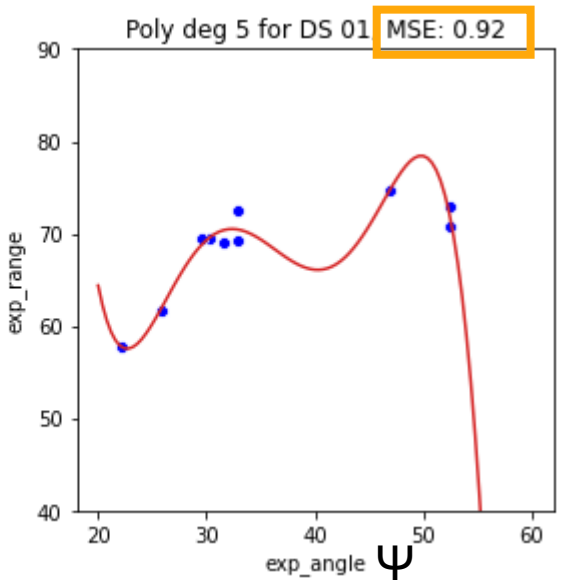
- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

"LEAST BAD" MODEL?









MSE = 11.74 (m^2)

$$r = \hat{\beta}_{\Psi} \Psi + \hat{\beta}_{0}$$

Fitted / trained model

$$r = \hat{\beta}_{\Psi^2} \Psi^2 + \hat{\beta}_{\Psi} \Psi + \hat{\beta}_{0}$$

$$r = \hat{\beta}_{\Psi^5} \Psi^5 + \hat{\beta}_{\Psi^4} \Psi^4 + \cdots$$

MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

.EAST BAD" MODEL?

Fitted model

$r = \hat{\beta}_{\Psi}\ \Psi + \hat{\beta}_{0}$

X → y_{pred}

Test score (e.g. MSE)

ORIGINAL TRAINING DATASET

X: Features — y: Labels

Observations (data points / samples)

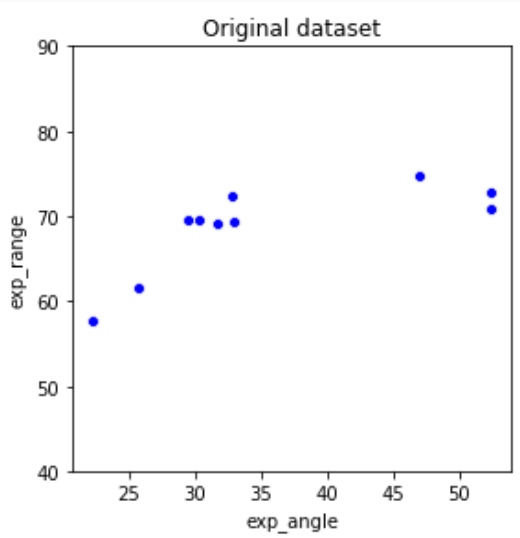
	Angle Ψ (psi)	Range r
Shot 1	60	32.4
Shot 2	45	36.2
Shot 3	45	37.7
…	…	…

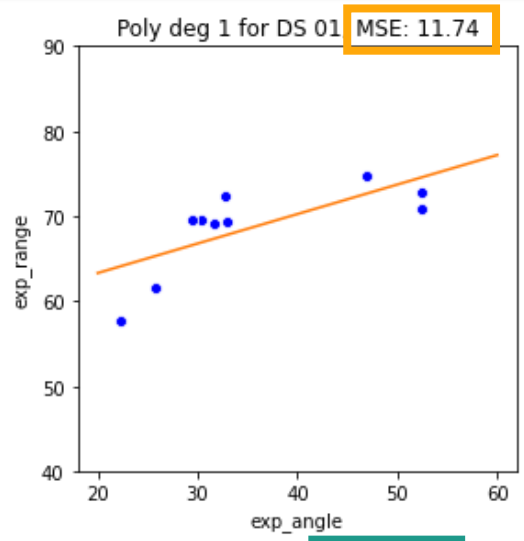
NEW TESTING DATASET

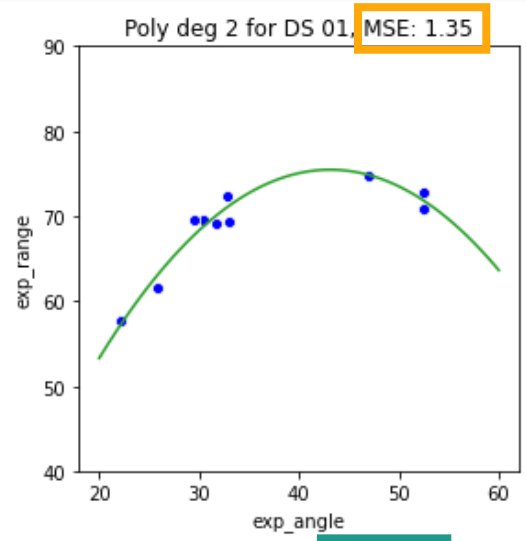
Features — Labels — y

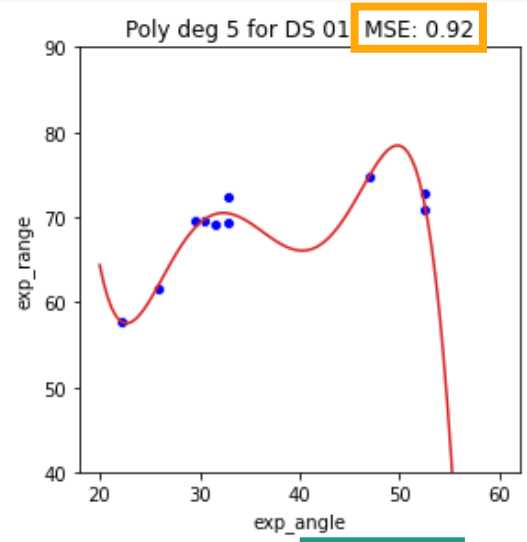
	Angle Ψ (psi)	Range r
Shot 1	37	33.1
Shot 2	24	29.6
Shot 3	54	41.4
…	…	…

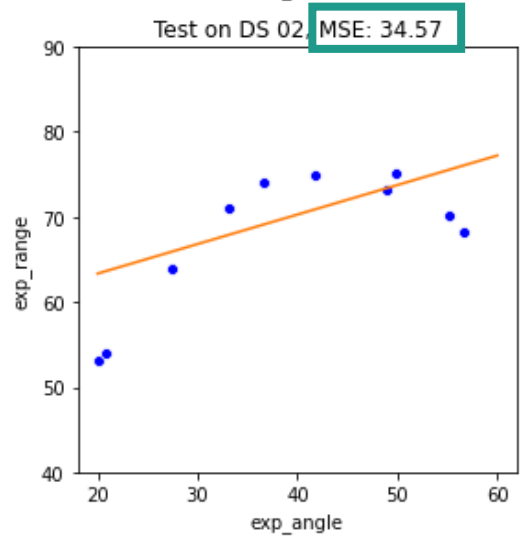
MAKING PREDICTIONS

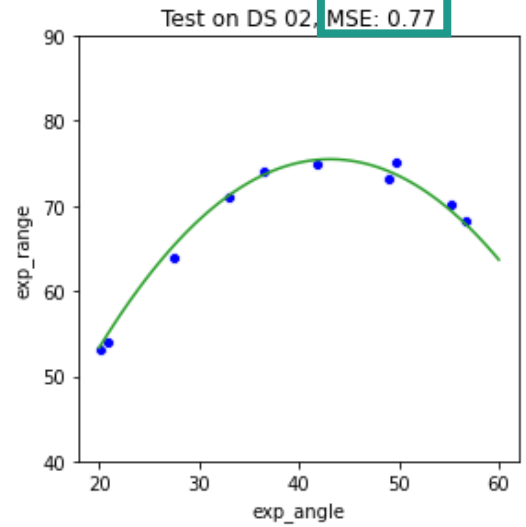


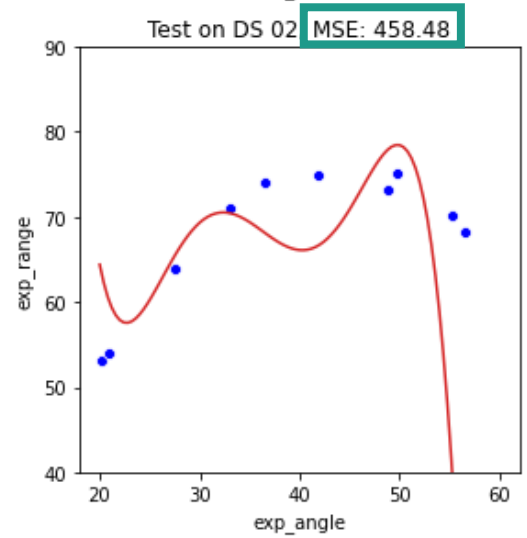






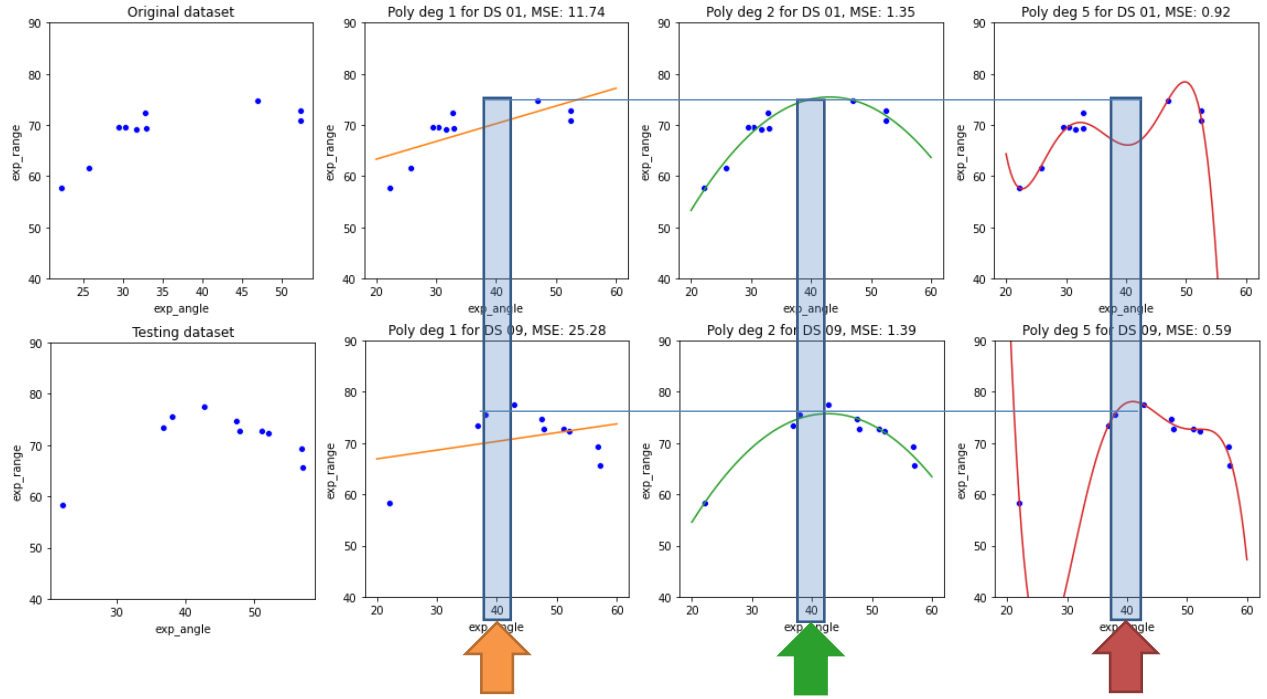


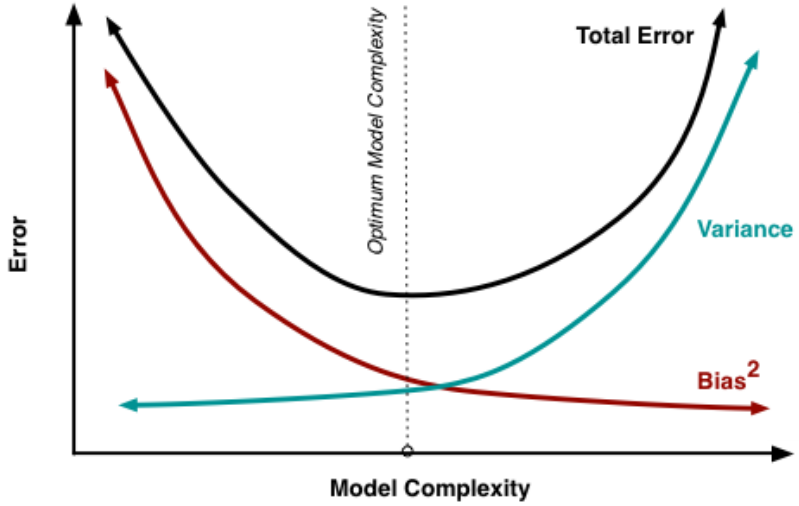


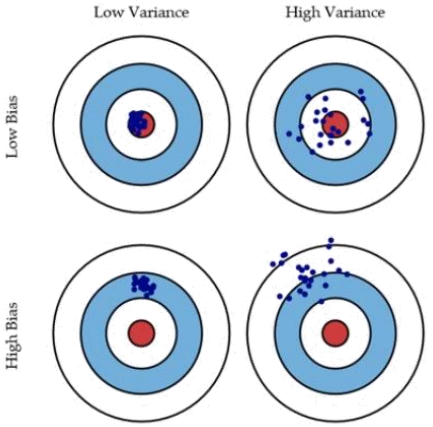


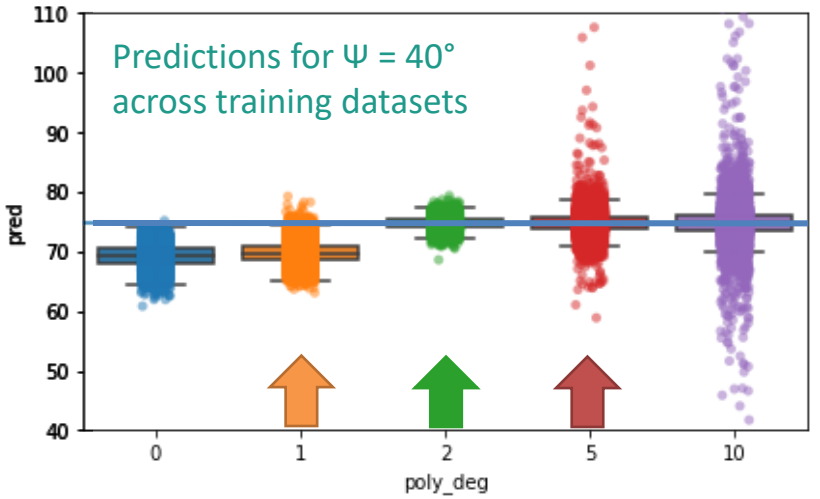
MAKING PREDICTIONS

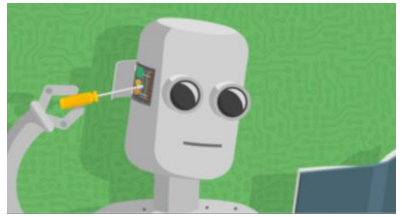
- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions











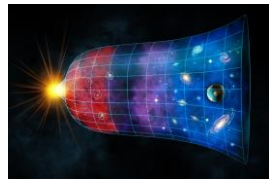
Ground Truth

→ **Need a testing set: split your data in training and testing set**
Let's practice!

GENERALIZABILITY OF ML MODELS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

"Non-reproducible single occurrences are of no significance to science." *Karl Popper*



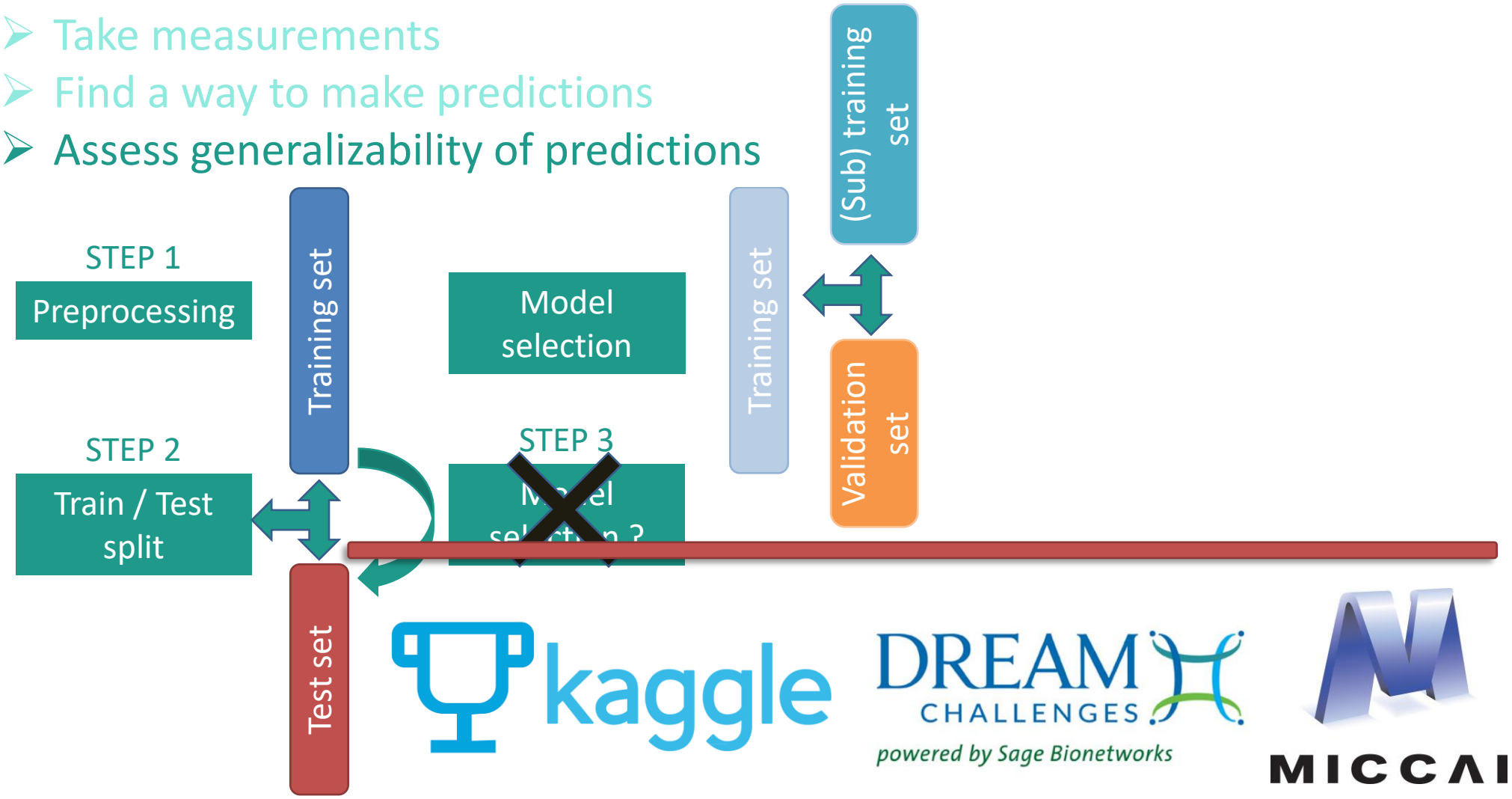
A hallmark of science is **generalization**:

→ derive findings that also apply to other experiments

➔ derive neuroimaging findings that apply to other population samples

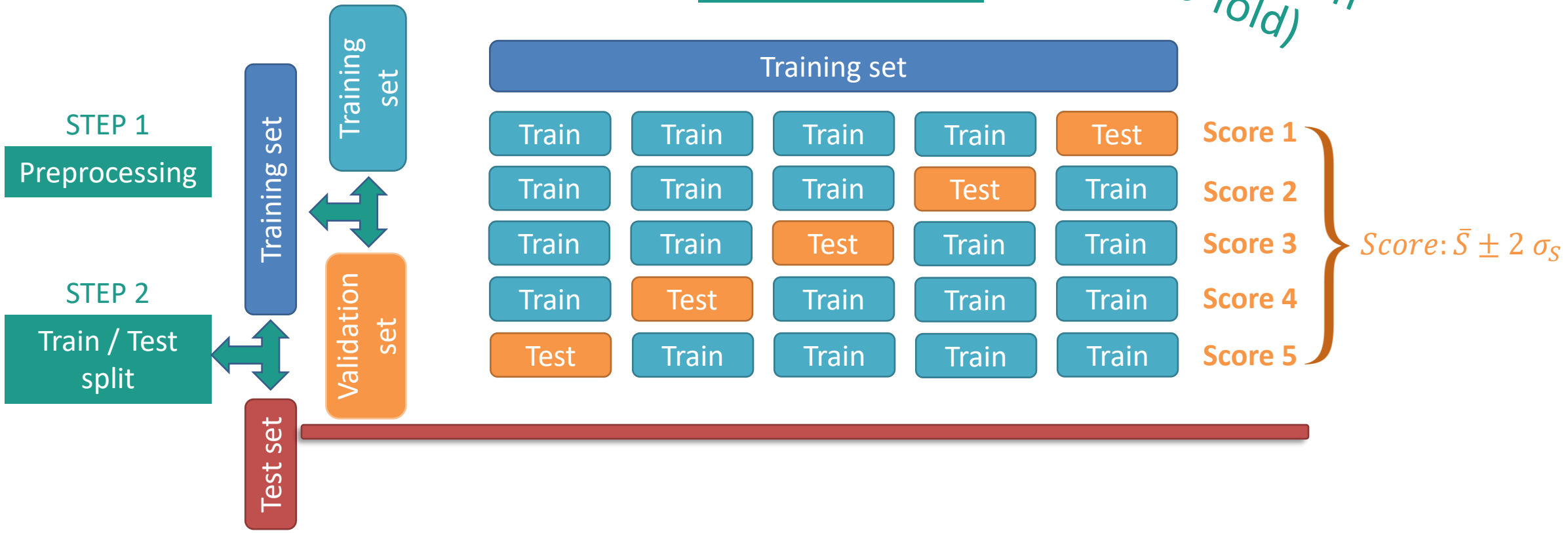
MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions



GENERALIZABILITY OF ML MODELS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions



→ **Let's practice !**

COURSE SUPPORT

SLACK (iords2021.slack.com)

- Course main channel: `#general`
- Topic channels: `#linux`, `#linux-capstone`, `#git`, `#git-capstone`, `#python`, `#full-example`, `#machine-learning`

→ Check regularly for course info (esp. pinned items)

→ Do not hesitate to ask questions (please reply "in thread")

1-to-1 OFFICE HOURS for course questions:

- 20-min slots every Friday morning between 9AM and 11AM

→ Book a time slot here: https://tinyurl.com/IORDS-office-hours

→ Do not hesitate to ask any kind of question, this is a beginner course !

EMAIL: methods@fcbg.ch ← Please whitelist!


Thank You!
Michael Dayan: methods@fcbg.ch
H N P
METHODS & DATA
campus biotech
Fondation Campus Biotech Geneva