DCML-CPS - Module 5

Data Analysis

Tommaso Zoppi
University of Trento - Povo (IT)
tommaso.zoppi@unifi.it
tommaso.zoppi@unitn.it

Course Map

Monitoring

- 1. Basics and Metrology
- 2. Monitoring

Testing

- 3. Fault Injection
- 4. Robustness Testing

Anomaly Detection

- 5. Data Analysis
- 6. Supervised ML
- 7. Unsupervised ML
- 8. Meta-Learning
- 9. Error/Intrusion Detection

Tools & Libs

Deep Learning

Data Analysis

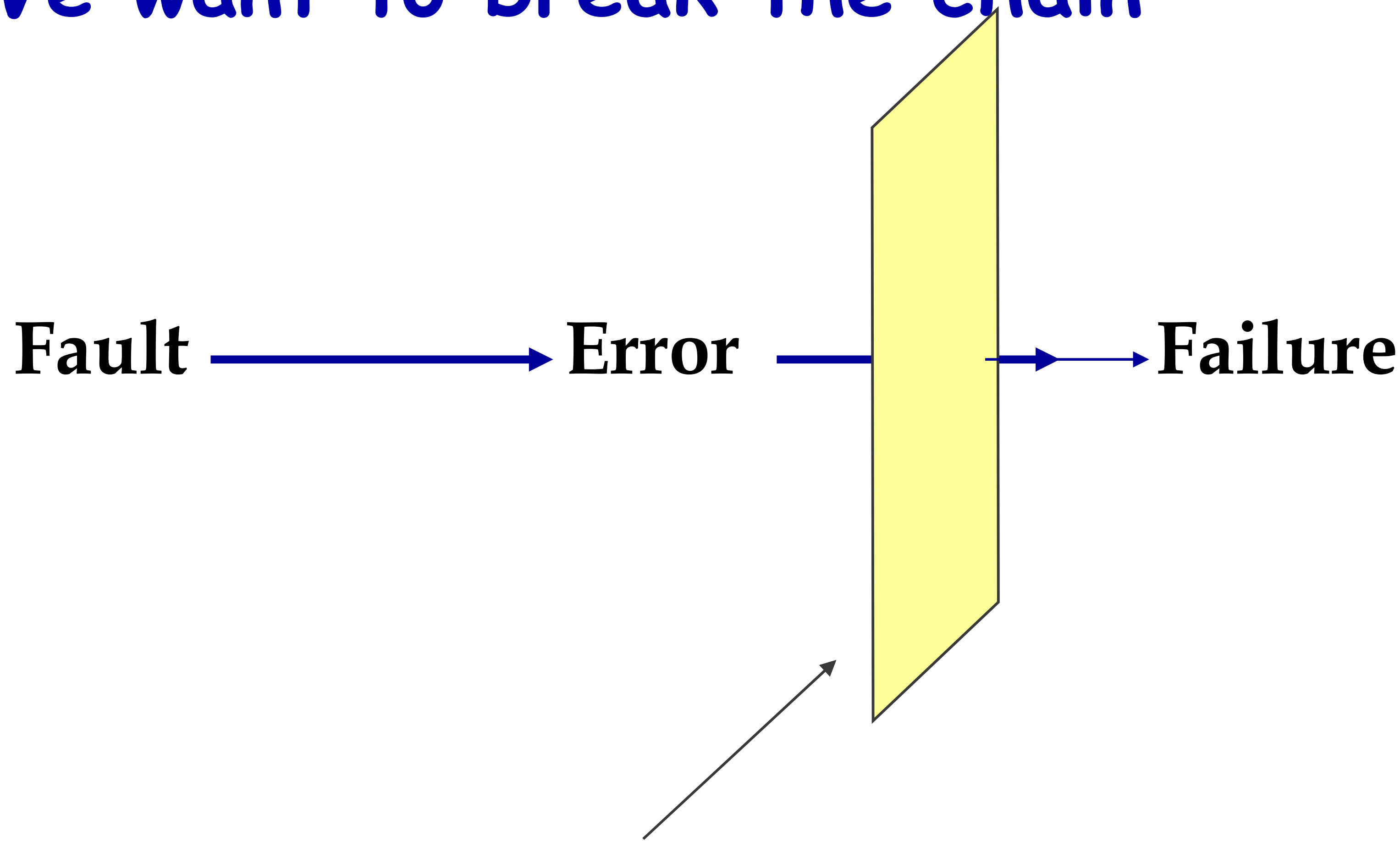
Data Analysis

- We understood how to gather a nice amount of data
  - As a result of monitoring activities
  - application logs
  - Also performing fault injection or robustness testing

- But, then… What are we going to do with these?

Remember!

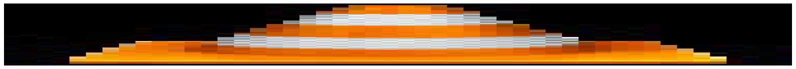
- We want to break the chain



How can I detect the activation of a fault to activate fault tolerance mechanisms?

Basics of Machine Learning

- Video from
  - Machine Learning Basics | What Is Machine Learning? | Introduction To Machine Learning | Simplilearn
    - https://www.youtube.com/watch?v=ukzFI9rgwfU



Machine Learning Basics _ What Is Machine Learning_ _ Introduction To Machine Learning _ Simplilearn.mp4

Video Answers (I)

- Facebook recognizes your friend in a picture from an album of tagged photographs
  - **Supervised** learning. Here Facebook is using tagged photos to recognize the person. Therefore, the tagged photos become the labels of the pictures and we know that when the machine is learning from labeled data, it is supervised learning.

Video Answers (II)

- Recommending new songs based on someone's past music choices
  - **Supervised** learning. The model is training a classifier on pre-existing labels (genres of songs). This is what Netflix, Pandora, and Spotify do all the time, they collect the songs/movies that you like already, evaluate the features based on your likes/dislikes and then recommend new movies/songs based on similar features.

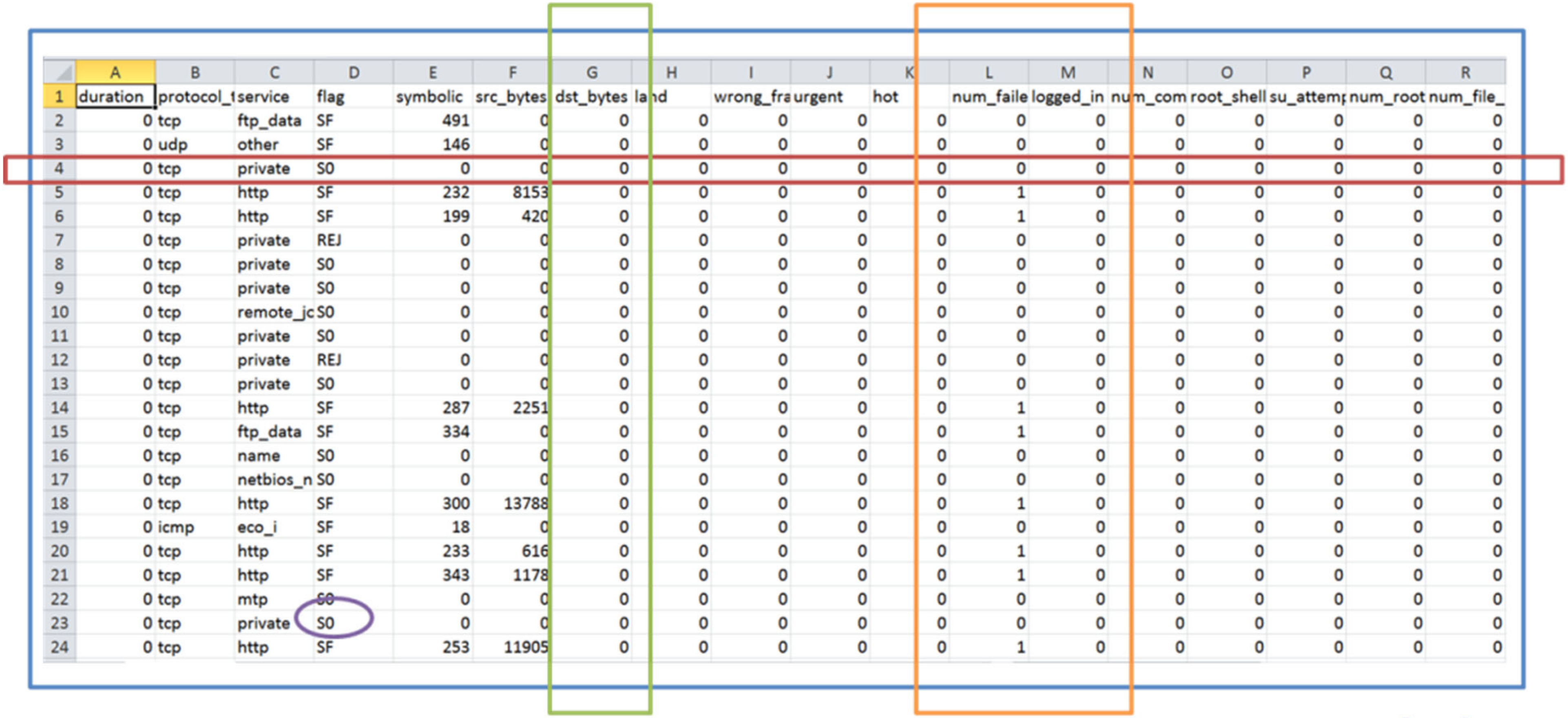
Video Answers (III)

- Scenario 3: Analyze bank data for suspicious-looking transactions and flag the fraud transactions
  - **Unsupervised** learning. In this case, the suspicious transactions are not defined, hence there are no labels of "fraud" and "not fraud". The model tries to identify outliers by looking at anomalous transactions and flags them as 'fraud'.

General Structure of a Dataset

Feature (F)

Feature Set (FS)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	duration	protocol_t	service	flag	symbolic	src_bytes	dst_bytes	land	wrong_fra	urgent	hot	num_faile	logged_in	num_com	root_shell	su_attemp	num_root	num_file_
2	0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	udp	other	SF	146	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	tcp	http	SF	232	8153	0	0	0	0	0	1	0	0	0	0	0	0
6	0	tcp	http	SF	199	420	0	0	0	0	0	1	0	0	0	0	0	0
7	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	tcp	remote_jc	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	tcp	http	SF	287	2251	0	0	0	0	0	1	0	0	0	0	0	0
15	0	tcp	ftp_data	SF	334	0	0	0	0	0	0	1	0	0	0	0	0	0
16	0	tcp	name	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	tcp	netbios_n	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	tcp	http	SF	300	13788	0	0	0	0	0	1	0	0	0	0	0	0
19	0	icmp	eco_i	SF	18	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	tcp	http	SF	233	616	0	0	0	0	0	1	0	0	0	0	0	0
21	0	tcp	http	SF	343	1178	0	0	0	0	0	1	0	0	0	0	0	0
22	0	tcp	mtp	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	tcp	http	SF	253	11905	0	0	0	0	0	1	0	0	0	0	0	0

Data Point (DP)

Feature Value (FV)

Dataset (D)

Steps of a Data Analysis Process

- Overall, steps of a data analysis process are
  - Generating/Getting the dataset
  - Exploratory study:
    - which are the features, label …
  - Feature Selection
    - Understanding which features are more informative
    - Often embedded in the algorithm: representation learning
  - Algorithm(s) Selection
    - Based on experimental analyses, availability of implementations, computational complexity, …
  - Choice of the Metric + Gather and Discuss Results

Feature Selection and Training

- **Dataset** (D): data structured as a set of **data points** (DP), described through a set of **Features** (F)
  - Each data point is composed by features values (FV)
  - A group of features defines a Feature Set (FS).
  - Algorithms can be defined for the data analysis process
- They are being trained by feeding them with a Training Set T
  - Each algorithm extracts the feature values from T
- The result of this process is an Algorithm **Model**

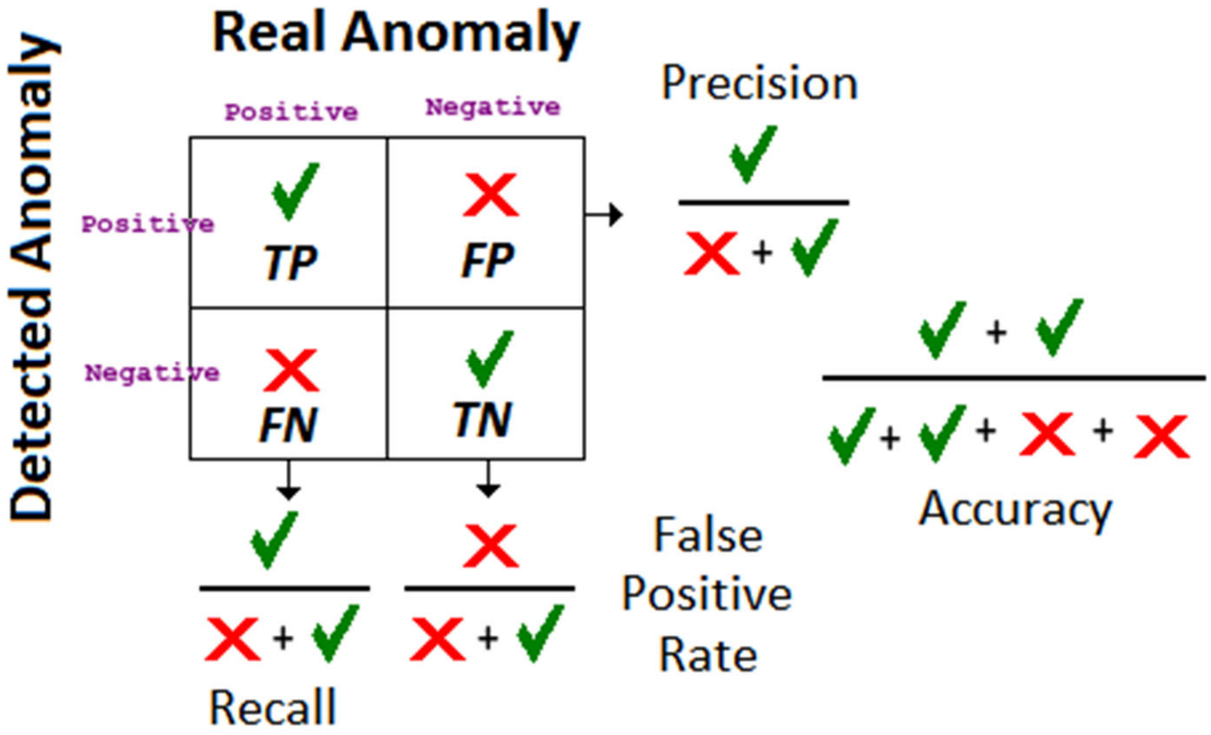
Testing

- The model is used for testing.
  - Data points in the **validation set** V are given as input to the model separately
  - The model outputs a **numeric score** AS that allows to decide on the «class» of the data point
  - AS can then be converted into a boolean score (for binary classification), or into categories (for multi-class) through a **Decision Function** DF
- If **Ground Thruth** (label) is available, it is possible to calculate Metric Scores

Classification Metrics

Scoring Metrics - I

- The suitability and the effectiveness of such techniques are usually evaluated and compared depending on specific metrics.
  - True Positives (TP)
  - True Negatives (TN)
  - False Positives (FP)
  - False Negatives (FN)



Real Anomaly

Detected Anomaly

	Positive	Negative
Positive	✓ TP	✗ FP
Negative	✗ FN	✓ TN

Precision

$$\frac{✓}{✗ + ✓}$$

$$\frac{✓ + ✓}{✓ + ✓ + ✗ + ✗}$$

Accuracy

$$\frac{✓}{✗ + ✓}$$

Recall

$$\frac{✗}{✗ + ✓}$$

False Positive Rate

Scoring Metrics - II

- Single metrics might be aggregated in order to build complex ones
- F-Score(β) is obtained by combining Precision and Recall,
  - weighting them by using a parameter β.
- F-Score(1) is often referred as F-Measure
  - β > 1 favours recall (low FNs)
  - β < 1 favours Precision (low F

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P) + R}$$

Scoring Metrics - III

However…. With unbalanced datasets the metrics above have severe weaknesses

- Example
  - a dataset which contains 5% of normal data and 95% of anomalies
  - A silly algorithm which always answers "anomaly"
- What do we get?
  - TP: 95%, FP: 5%, FN: 0%, TN: 0%
  - Accuracy: 95%, Precision: 95%, Recall: 100%, F1: 97.5%
- They seem very good scores, correct?

Scoring Metrics – III (cont.)

- They seem very good scores, correct?
  - But the algorithm does not "deserve" very good scores (it is a very silly one)
- Therefore, with unbalanced datasets where the ratio of normal/anomaly data points differs a lot, it is better to use a different metric

Scoring Metrics – III (cont.)

- Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

  - Ranges from -1 to 1
  - 1 means "perfectly correct classification"
  - -1 means "perfectly wrong classification"
    - Desirable as well, we just need to do the opposite with respect to what the algorithm says
  - 0 means "random guessing", or very bad classification
- In the example, MCC: 0 -> appropriate for a silly algorithm!

More on Metrics

- Further readings on metrics
  - Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.
  - Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 1-13.