PADERBORN UNIVERSITY

Evaluating Machine Learning

Pitfalls, Rabbit Holes and Hints

Michael Röder

DICE

Data Science Group
Paderborn University

October 23, 2024

Disclaimer

- This presentation mainly comprises
  -  Pitfalls
  -  Rabbit holes
  -  (hopefully helpful) hints

Disclaimer

- This presentation mainly comprises
  -  Pitfalls
  -  Rabbit holes
  -  (hopefully helpful) hints

- My view is biased by my previous research
  → Some statements may not hold for your area
  → Please ask if something is unclear

- This is not a one man show
  → Feel free to add your own views and experiences

Introduction

Why do we need an evaluation?

Introduction

Why do we need an evaluation?

What do we expect from a good evaluation?

Introduction

Evaluation?

- Measure success / answer a research question
- Comparability
- Objectivity / Fairness
- Reproducibility / Repeatability

Evaluation

1. Dataset

2. Algorithm Execution

3. KPIs

4. Combination

Section 1

Datasets

Datasets

- Do you have data?

Datasets

- Do you have data?
  -  Starting without having data is a good first step to waste your time.
  -  Ensure that you have a good understanding which data you actually need.
    $\rightarrow$  Write down the (formal) definition of the problem that you want to tackle

Datasets

- Do you have data?
  -  Starting without having data is a good first step to waste your time.
  -  Ensure that you have a good understanding which data you actually need.
    →  Write down the (formal) definition of the problem that you want to tackle
- Where can we get data from?

Datasets

- Do you have data?
  -  Starting without having data is a good first step to waste your time.
  -  Ensure that you have a good understanding which data you actually need.
    →  Write down the (formal) definition of the problem that you want to tackle
- Where can we get data from?
  - Use an existing dataset
  - Create your own dataset
  - Use synthetic data

Datasets

Reuse Existing Datasets

- Advantages:
  - Low effort (typically)
  - Easier comparison to related work
    ( but not always possible)

[1] https://archive.org/web/

Datasets

Reuse Existing Datasets

- Advantages:
  - Low effort (typically)
  - Easier comparison to related work
    ( but not always possible)
- Dataset sources:
  - Related Work
  - Data portals
  - Challenges / competitions

[1] https://archive.org/web/

Datasets

Reuse Existing Datasets

- Advantages:
  - Low effort (typically)
  - Easier comparison to related work
    ( but not always possible)
- Dataset sources:
  - Related Work
  - Data portals
  - Challenges / competitions

 Datasets can disappear over time!

 Download and store it (with a README).

 Ask the Internet Archive to archive the page.[1]

[1] https://archive.org/web/

Datasets

Reuse Existing Datasets

Check the dataset!

- Datasets may...
  -  miss (meta) data

Example taken from [Vogel and Jiang, 2019]

Datasets

Reuse Existing Datasets

Check the dataset!

- Datasets may...
  -  miss (meta) data
  -  not contain the data you expect

```
1 { "Date": "2017-08-30",
2   "URL": "https://schluesselkindblog.com/2017/[...]",
3   "Title": "Prozess beginnt: Mord an Freiburger
        Studentin",
4   "Teaser": "Prozessbeginn gegen [...]",
5   "False_Statement_1": "Die Pflegefamilie [...]",
6   "Ratio_of_Fake_Statements": "1",
7   "Overall_Rating": "0.7" }
```

Some datasets will never be perfect

Example taken from [Vogel and Jiang, 2019]

Datasets

Reuse Existing Datasets

Analyze datasets

- How it has been created?
- What are its features? (has to be part of your paper)
  - How many instances are there?

Maybe just an example dataset or an excerpt?

Is the data balanced?

Figure from [Ribeiro et al., 2016]

Datasets

Reuse Existing Datasets

Analyze datasets

- How it has been created?
- What are its features? (has to be part of your paper)
  - How many instances are there?

Maybe just an example dataset or an excerpt?

Is the data balanced?

Does it have biases?

(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

→ Come back to the dataset during the analysis of the results

Figure from [Ribeiro et al., 2016]

Datasets

Reuse Existing Datasets

Gold standards are not always "golden" [Jha et al., 2017]

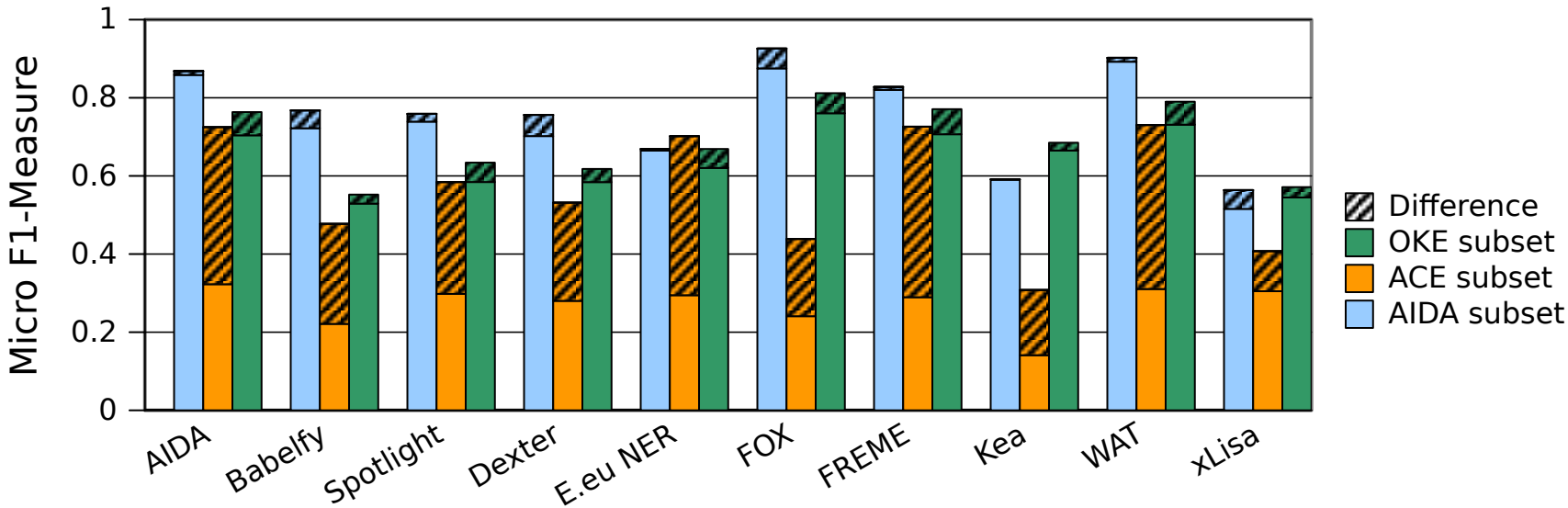
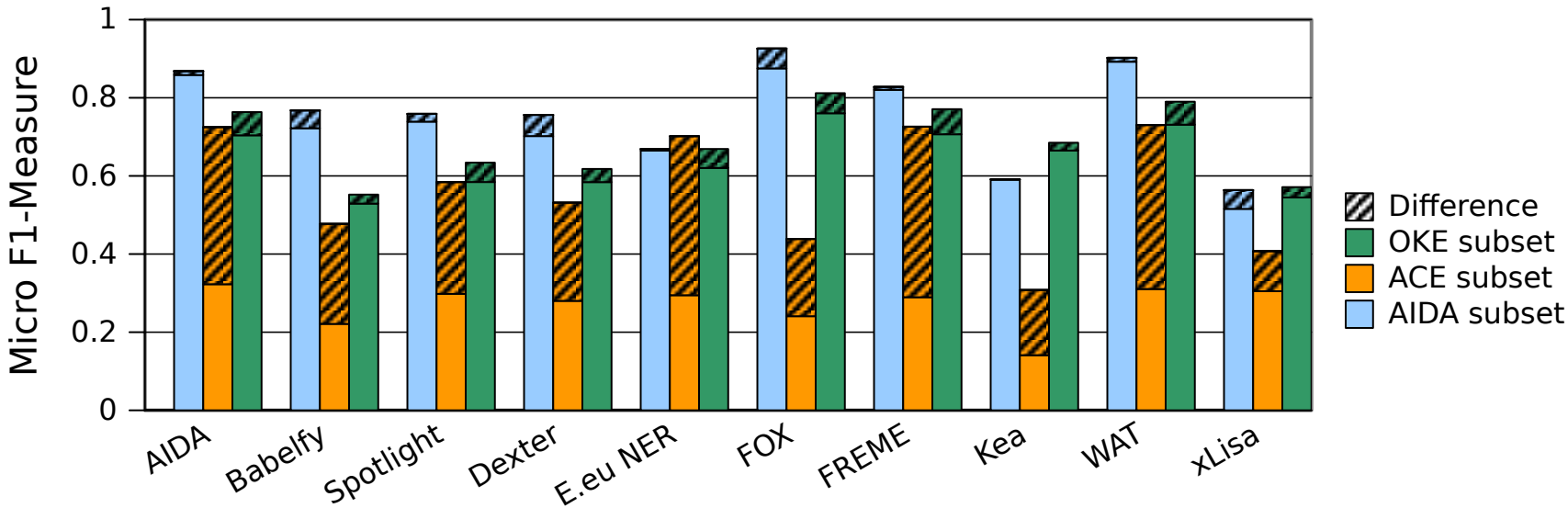


Figure from [Jha et al., 2017]

Datasets

Reuse Existing Datasets

Gold standards are not always "golden" [Jha et al., 2017]



- Datasets might be bound to a certain point in time
  -  Datasets can be outdated [Jha et al., 2017]
  -  A dataset may not fit to data that you use in your algorithm

Figure from [Jha et al., 2017]

Datasets

Creating a Dataset

Don’t underestimate the effort

Datasets

Creating a Dataset

Don’t underestimate the effort

- If you use humans as annotators...
  - Objectivity: use more than one person to annotate data
  - 
    Ensure that all have the same understanding of the task
  - Measure and report the interrater agreement
    (Cohen’s kappa [Cohen, 1960], Fleiss’ kappa [Fleiss and Cohen, 1973], F1-measure [Hripcsak and Rothschild, 2005], Percent Agreement [Shweta et al., 2015])
  - Define how you handle conflicts
  - 
    Think about using a tool to pre-annotate data [Jha et al., 2017]

Datasets

Creating a Dataset


Don’t underestimate the effort

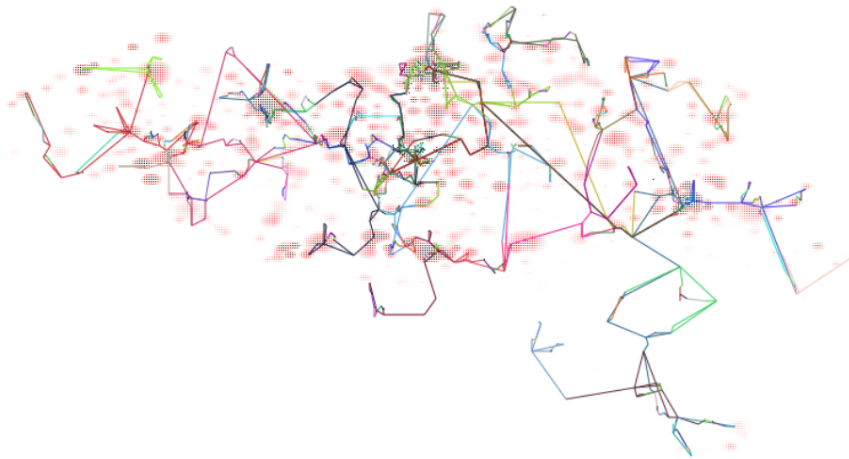
- If you use humans as annotators...
  - Objectivity: use more than one person to annotate data
  - 
    Ensure that all have the same understanding of the task
  - Measure and report the interrater agreement
    (Cohen’s kappa [Cohen, 1960], Fleiss’ kappa [Fleiss and Cohen, 1973], F1-measure [Hripcsak and Rothschild, 2005], Percent Agreement [Shweta et al., 2015])
  - Define how you handle conflicts
  - 
    Think about using a tool to pre-annotate data [Jha et al., 2017]


Reminder Do not rely on internet resources to stay where they are!
(Download the data, use the internet archive)

Datasets

Dataset Generators

- Domain-dependent
  (LUBM [Guo et al., 2005],
  PoDiGG [Taelman et al., 2019])

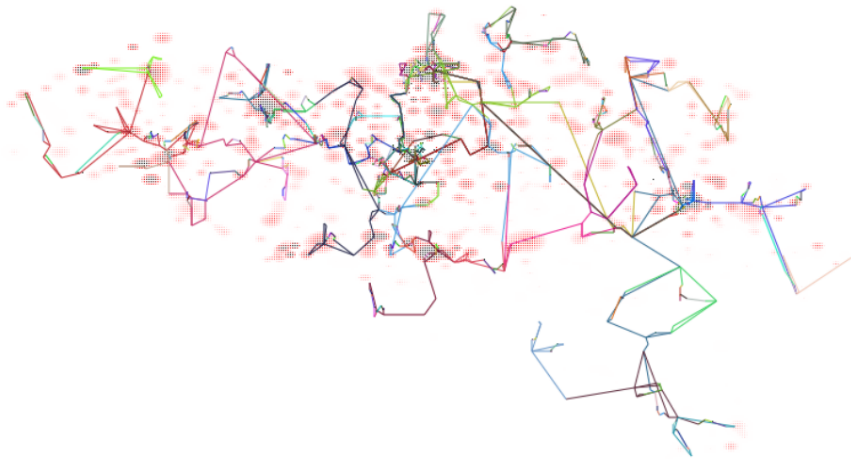


Figures from [Taelman et al., 2019] and [Duan et al., 2011]

Datasets

Dataset Generators

- Domain-dependent
  (LUBM [Guo et al., 2005],
  PoDiGG [Taelman et al., 2019])
- Domain-independent
  (Lemming [Röder et al., 2021])

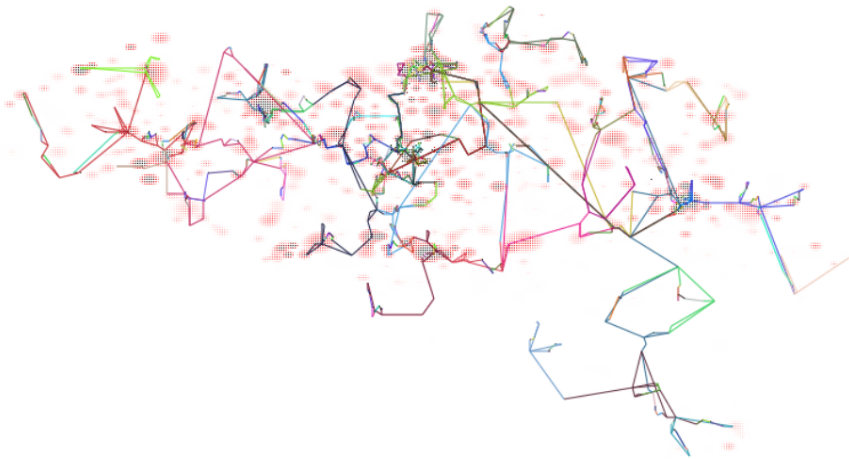


Figures from [Taelman et al., 2019] and [Duan et al., 2011]

Datasets

Dataset Generators

- Domain-dependent
  (LUBM [Guo et al., 2005],
  PoDiGG [Taelman et al., 2019])
- Domain-independent
  (Lemming [Röder et al., 2021])
- Advantages
  - Scalable datasets
  - Focus on a specific scenario

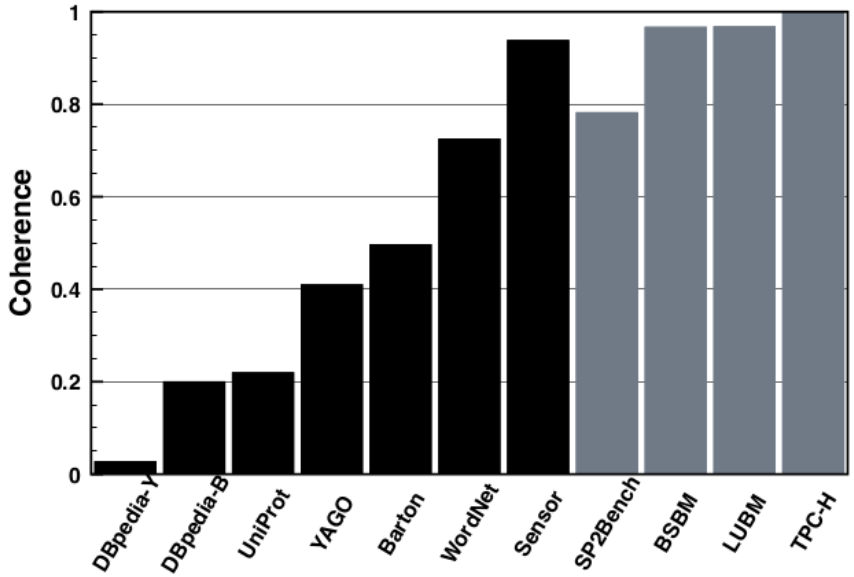


Figures from [Taelman et al., 2019] and [Duan et al., 2011]

Datasets

Dataset Generators

- Domain-dependent
  (LUBM [Guo et al., 2005],
  PoDiGG [Taelman et al., 2019])
- Domain-independent
  (Lemming [Röder et al., 2021])
- Advantages
  - Scalable datasets
  - Focus on a specific scenario
- Limitations
  - Bound to assumptions
  - Might generate unrealistic data

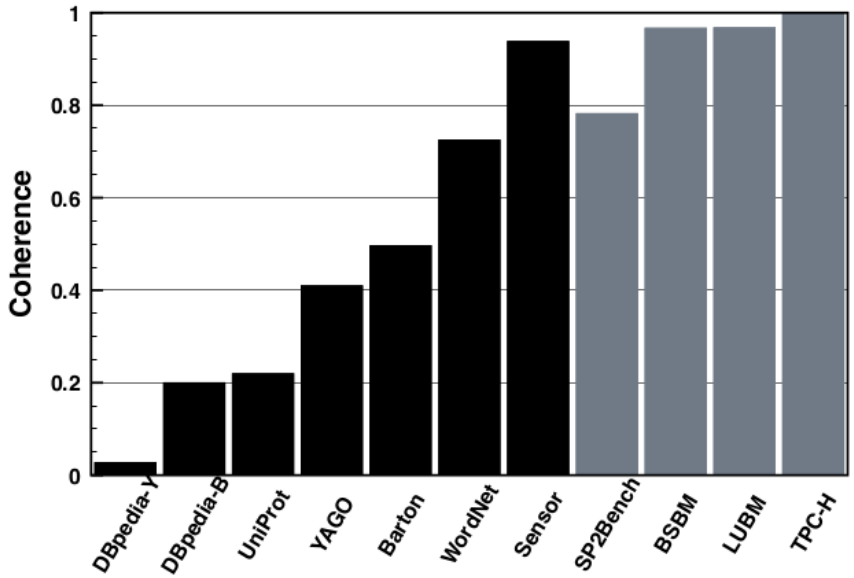


Figures from [Taelman et al., 2019] and [Duan et al., 2011]

Datasets

Dataset Generators

- Domain-dependent
  (LUBM [Guo et al., 2005],
  PoDiGG [Taelman et al., 2019])
- Domain-independent
  (Lemming [Röder et al., 2021])
- Advantages
  - Scalable datasets
  - Focus on a specific scenario
- Limitations
  - Bound to assumptions
  - Might generate unrealistic data



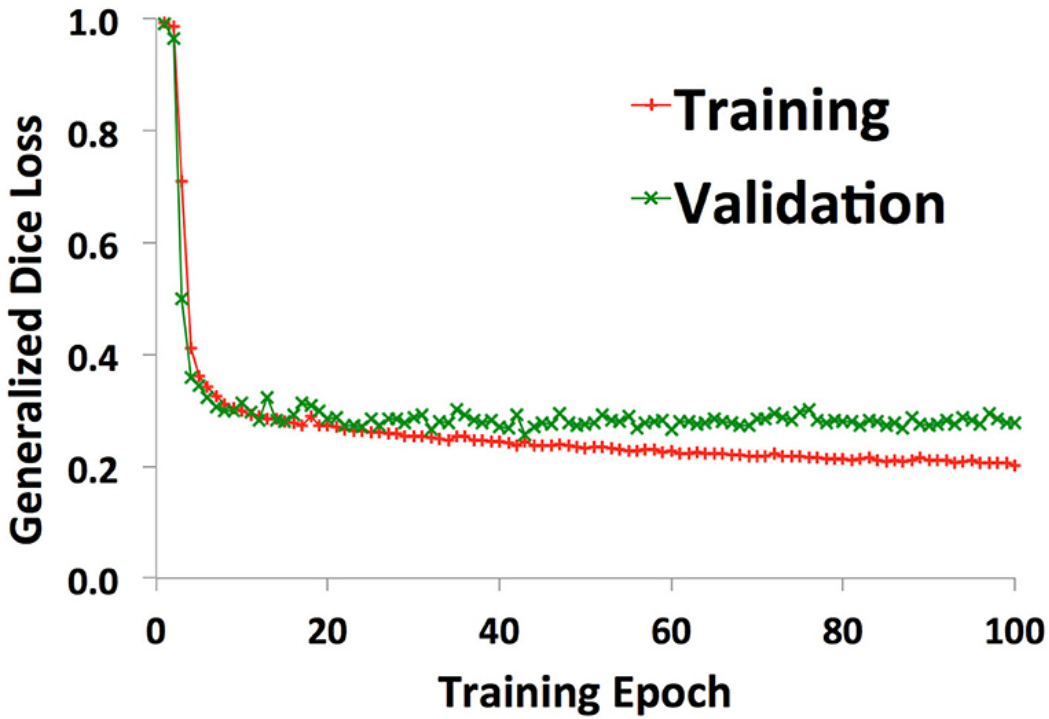
→ Can be helpful; real-world datasets might be necessary

Figures from [Taelman et al., 2019] and [Duan et al., 2011]

Datasets

Dataset Splits

	Train	Validation	Test
System gets Ground Truth	Yes	Yes	No
Model trained on	Yes	No	No
Used for KPIs	Maybe (sep)	Maybe (sep)	Yes



Datasets

Dataset Splits

	Train	Validation	Test
System gets Ground Truth	Yes	Yes	No
Model trained on	Yes	No	No
Used for KPIs	Maybe (sep)	Maybe (sep)	Yes

Splits may have an influence on the result [Shchur et al., 2018]

- Cross validation

Figure from `https://www.frontiersin.org/files/Articles/517375/fsurg-07-517375-HTML/image_m/fsurg-07-517375-g002.jpg`

Datasets

Dataset Splits

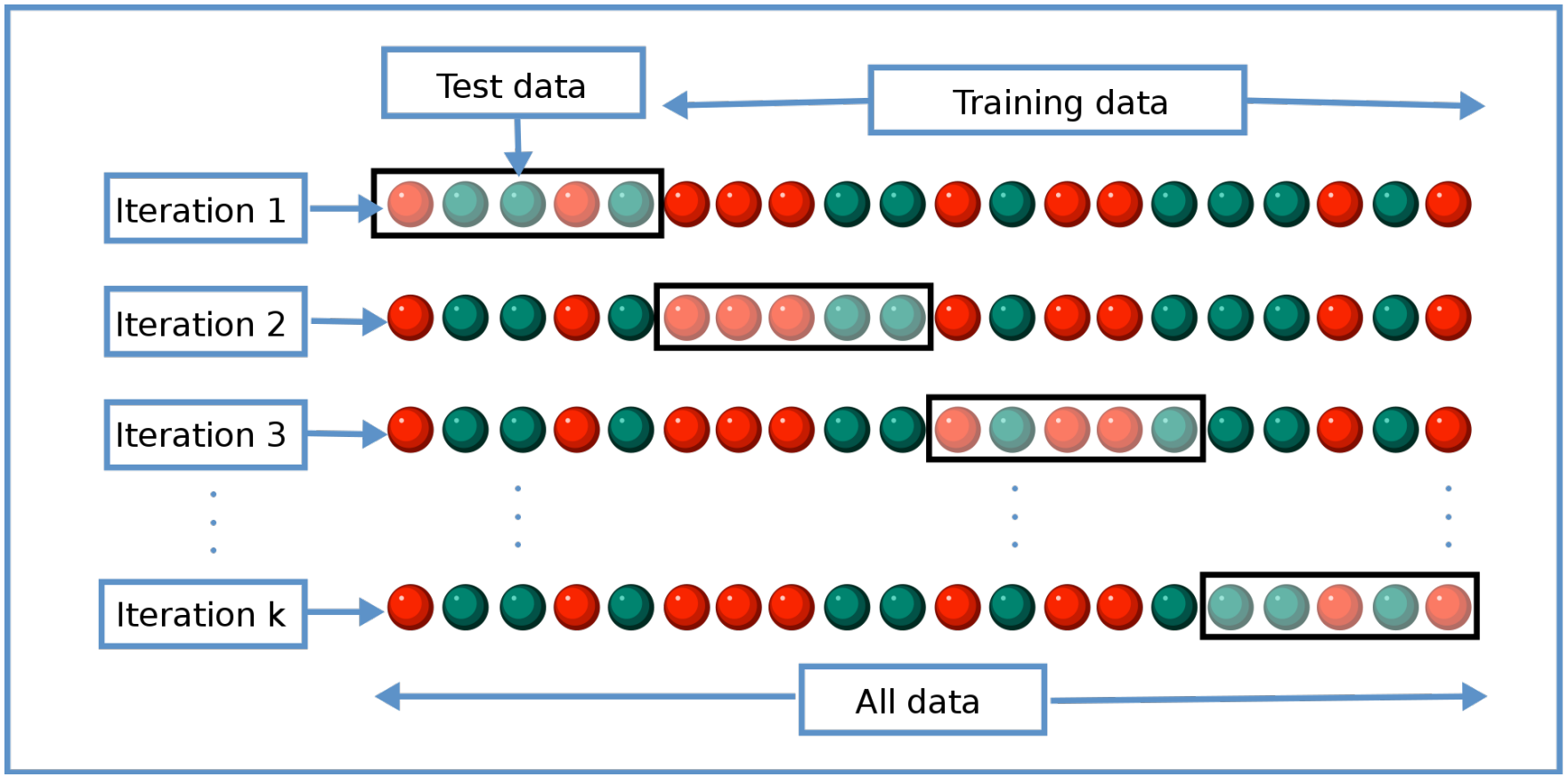


Figure from
https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Section 2

Algorithm Execution

Algorithm Execution

Different Views

Algorithm Developer

Benchmark Developer

Algorithm Execution

Algorithm Developer

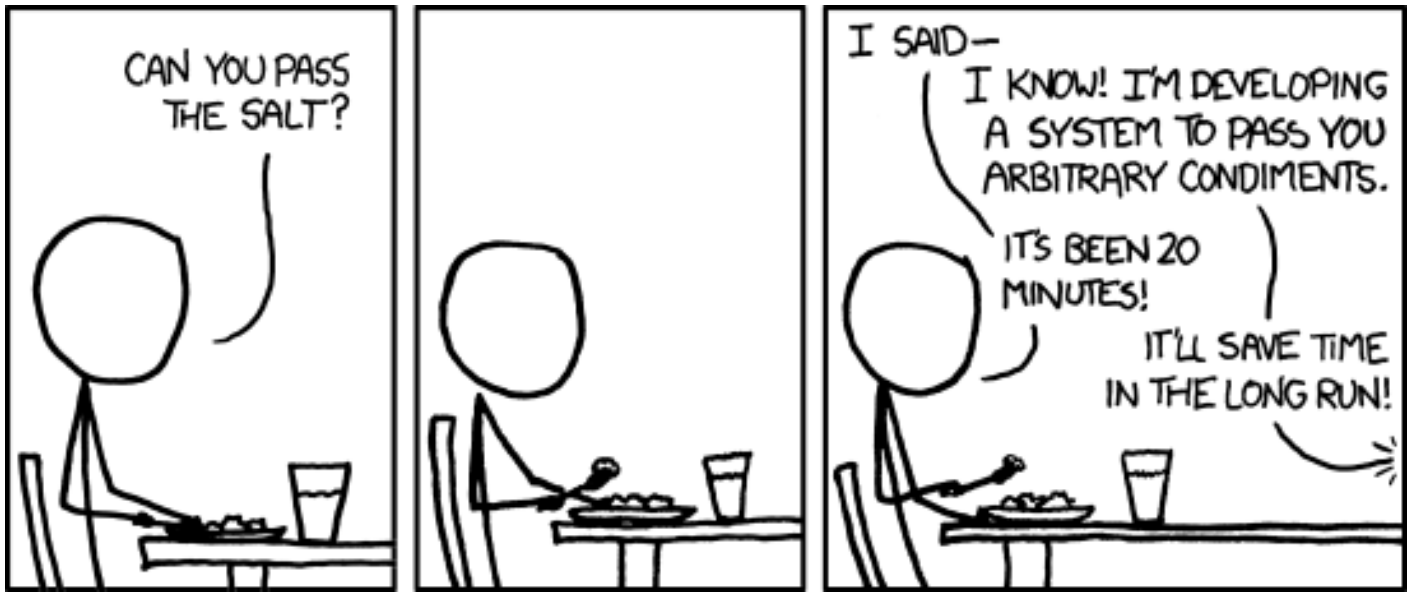
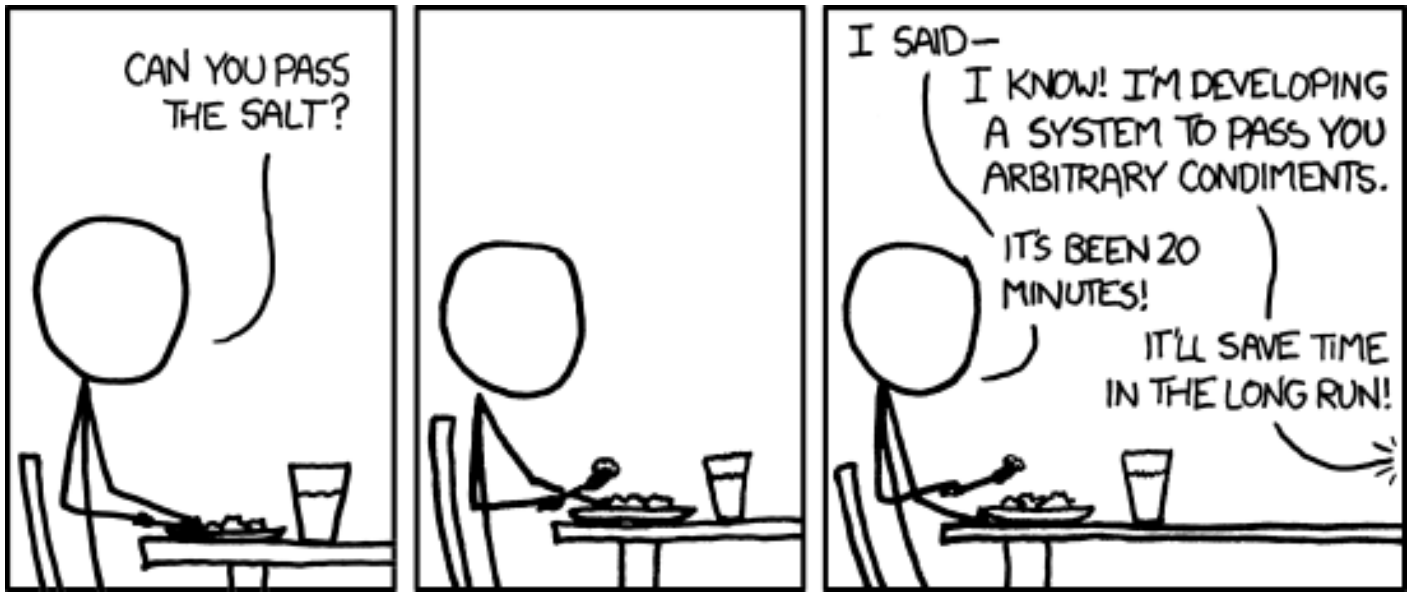


Figure from `https://imgs.xkcd.com/comics/the_general_problem.png`

Algorithm Execution

Algorithm Developer



Do not try to implement the perfectly integrated system.

- You (typically) develop a research prototype!

Focus on the important part

Try to find a balance between hacking and necessary software engineering.

Figure from `https://imgs.xkcd.com/comics/the_general_problem.png`

Algorithm Execution

Algorithm Developer

Check your approach!

Check preprocessing

Use a small example set

Write unit tests

Do not try to achieve the perfect first step
(i.e., do not investigate the first interessting question / problem that you find)

Algorithm Execution

Algorithm Developer

- Check your approach!
  - 
    Check preprocessing
  - 
    Use a small example set
  - 
    Write unit tests
  - 
    Do not try to achieve the perfect first step
    (i.e., do not investigate the first interessting question / problem that you find)
- 
  Store intermediate results (saves tons of runtime!)

Algorithm Execution

Algorithm Developer

- Check your approach!
  - 
    Check preprocessing
  - 
    Use a small example set
  - 
    Write unit tests
  - 
    Do not try to achieve the perfect first step
    (i.e., do not investigate the first interessting question / problem that you find)
- 
  Store intermediate results (saves tons of runtime!)
- 
  Store final results for later analysis

Algorithm Execution

Algorithm Developer

Does your approach make use of random numbers?

- How to ensure repeatability?

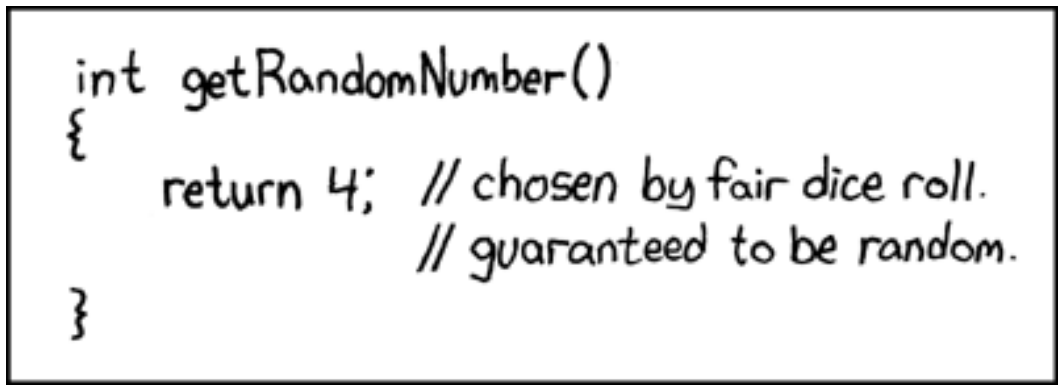
Figure from `https://imgs.xkcd.com/comics/random_number.png`

Algorithm Execution

Algorithm Developer

Does your approach make use of random numbers?

- How to ensure repeatability?




Make use of seed values to ensure repeatability


Different random number generators shouldn't get the same seed

- More details at `https://dice-research.org/news/2020-09-03-RandomnumbersJava/`

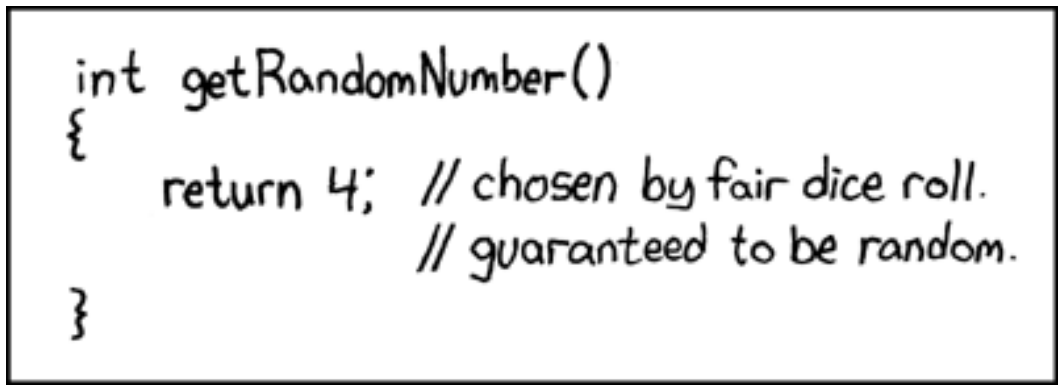
Figure from `https://imgs.xkcd.com/comics/random_number.png`

Algorithm Execution

Algorithm Developer

Does your approach make use of random numbers?

- How to ensure repeatability?




Make use of seed values to ensure repeatability


Different random number generators shouldn't get the same seed

- More details at `https://dice-research.org/news/2020-09-03-RandomnumbersJava/`


Do not only run it once!


Run it *n* times and report average values with standard deviation

---

Figure from `https://imgs.xkcd.com/comics/random_number.png`

Algorithm Execution

Algorithm Developer

Does your approach have (hyper-)parameters that should be optimized?

- Grid search
- Random search [Bergstra and Bengio, 2012]

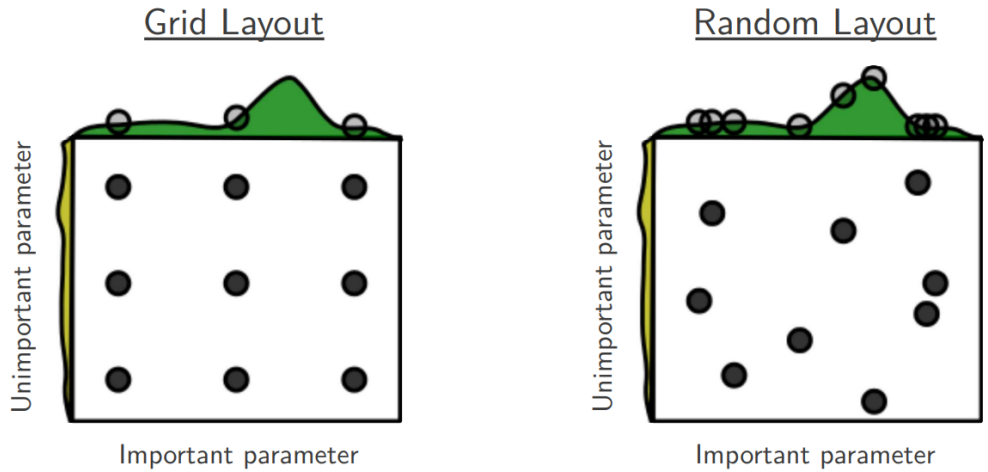


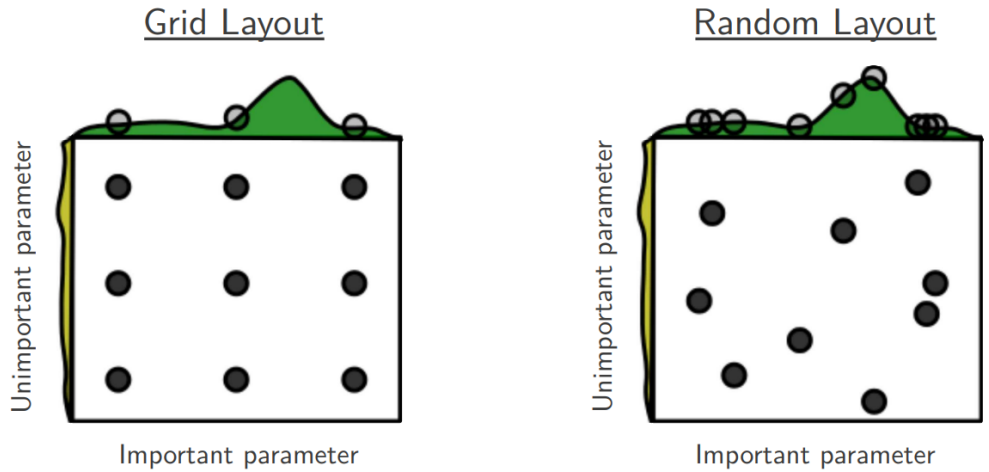
Figure from [Bergstra and Bengio, 2012]

Algorithm Execution

Algorithm Developer

Does your approach have (hyper-)parameters that should be optimized?

- Grid search
- Random search [Bergstra and Bengio, 2012]



- Bayesian optimization
  (Nice depiction at
  https://en.wikipedia.org/wiki/Bayesian_optimization)

Figure from [Bergstra and Bengio, 2012]

Algorithm Execution

Algorithm Developer

⚠ Don't let your test data leak into the training / optimization process

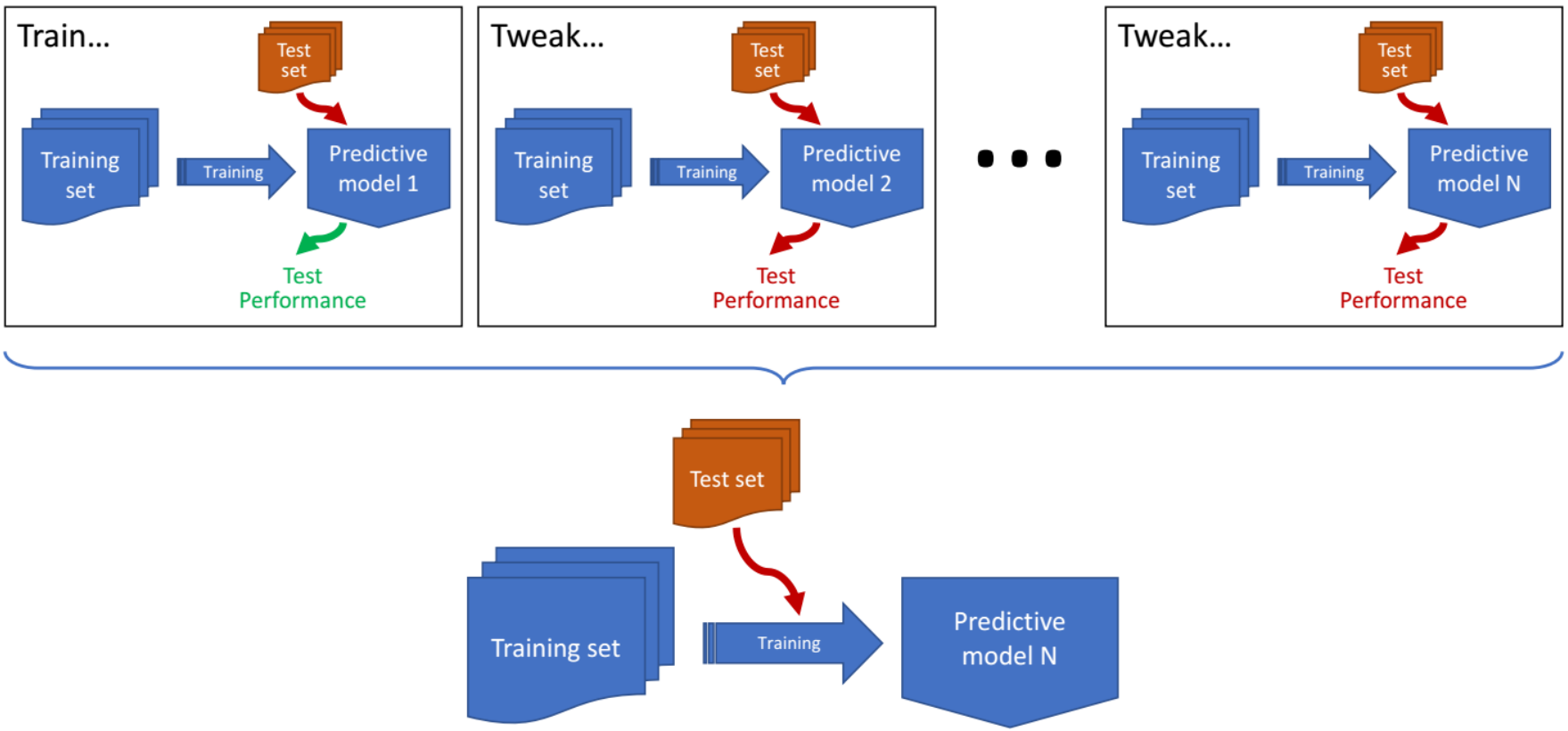


Figure from [Lones, 2024]

Algorithm Execution

Algorithm Developer

Use validation data

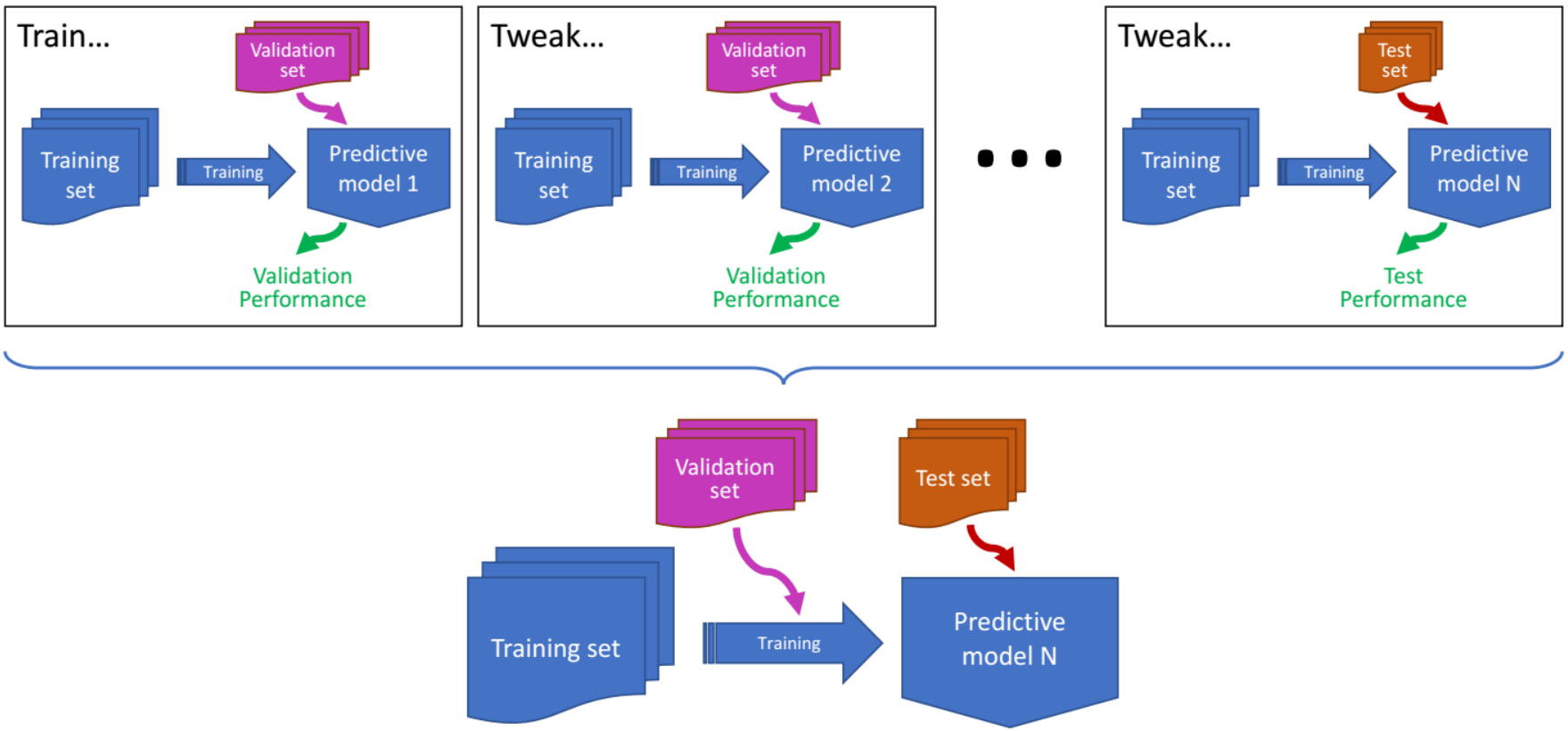
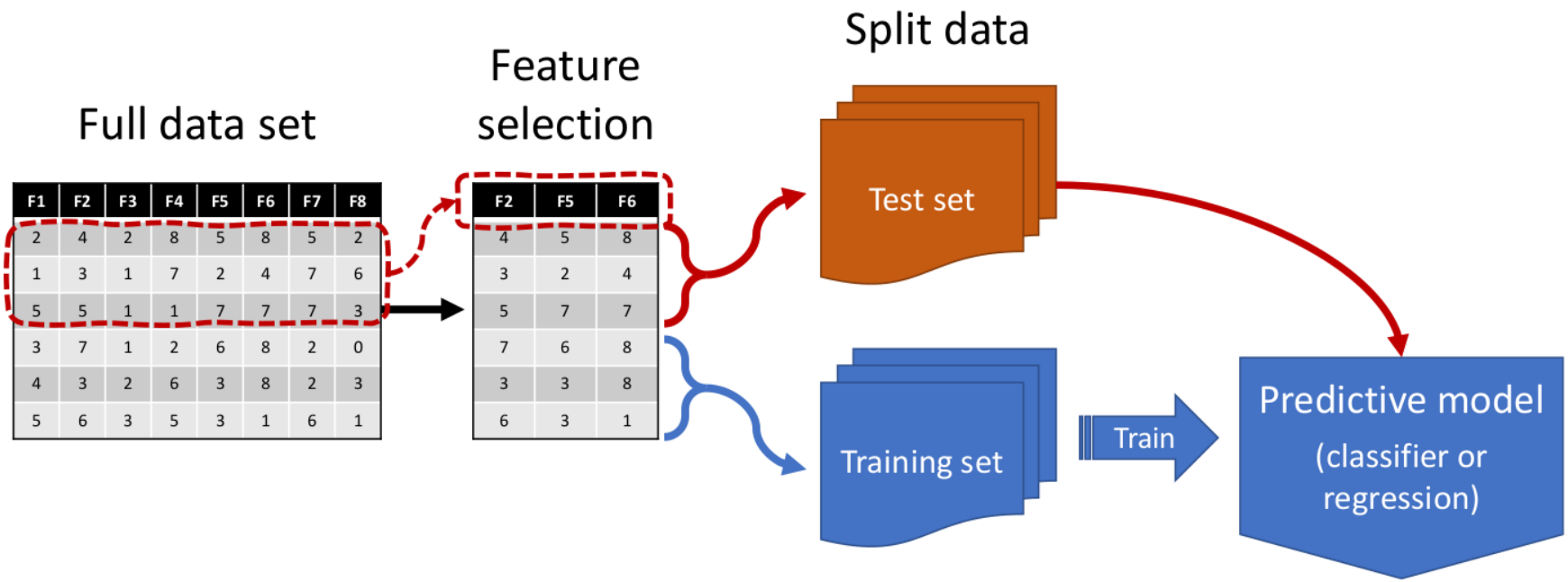


Figure from [Lones, 2024]

Algorithm Execution

Algorithm Developer

Data Split vs. Feature Selection



Test data leaks into the training process

Figure from [Lones, 2024]

Algorithm Execution

Algorithm Developer

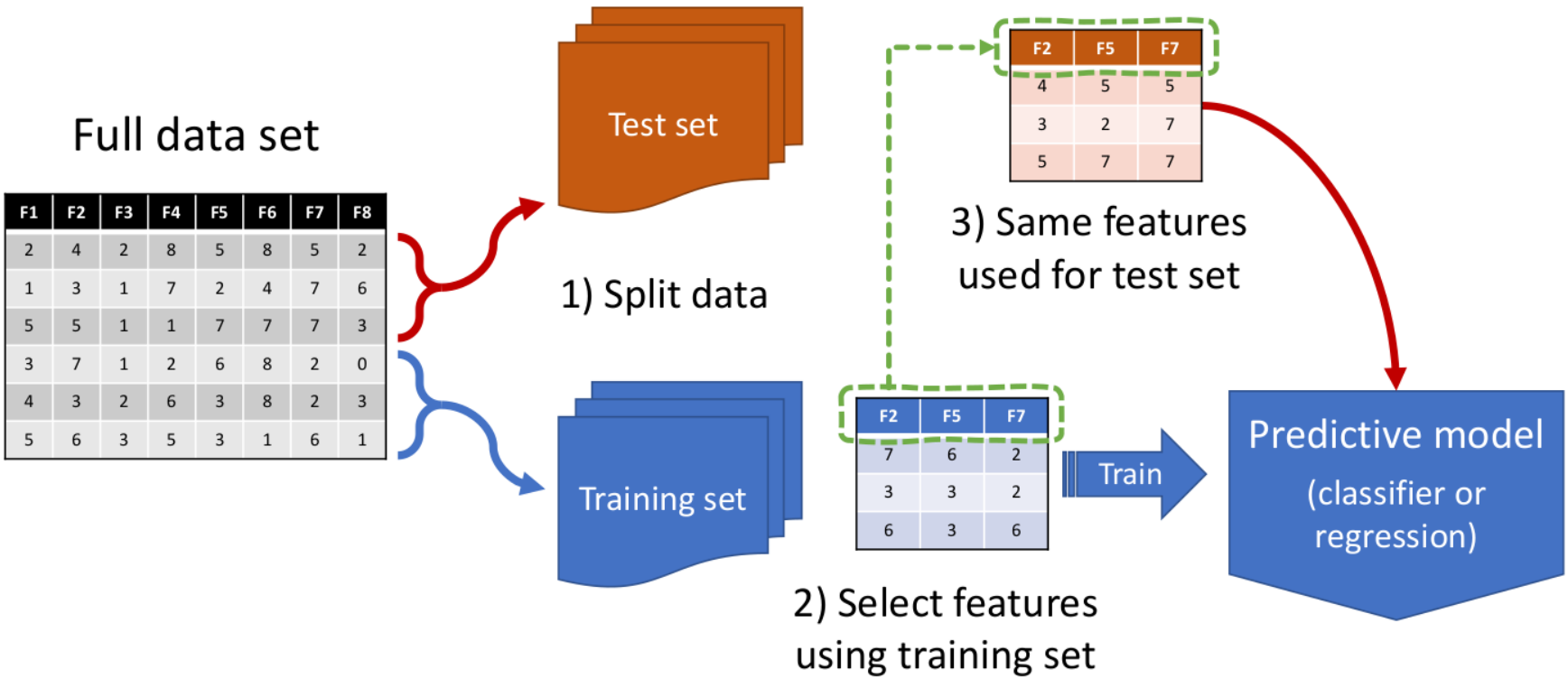


Figure from [Lones, 2024]

Algorithm Execution

Algorithm Developer

Cross Validation vs. Feature Selection

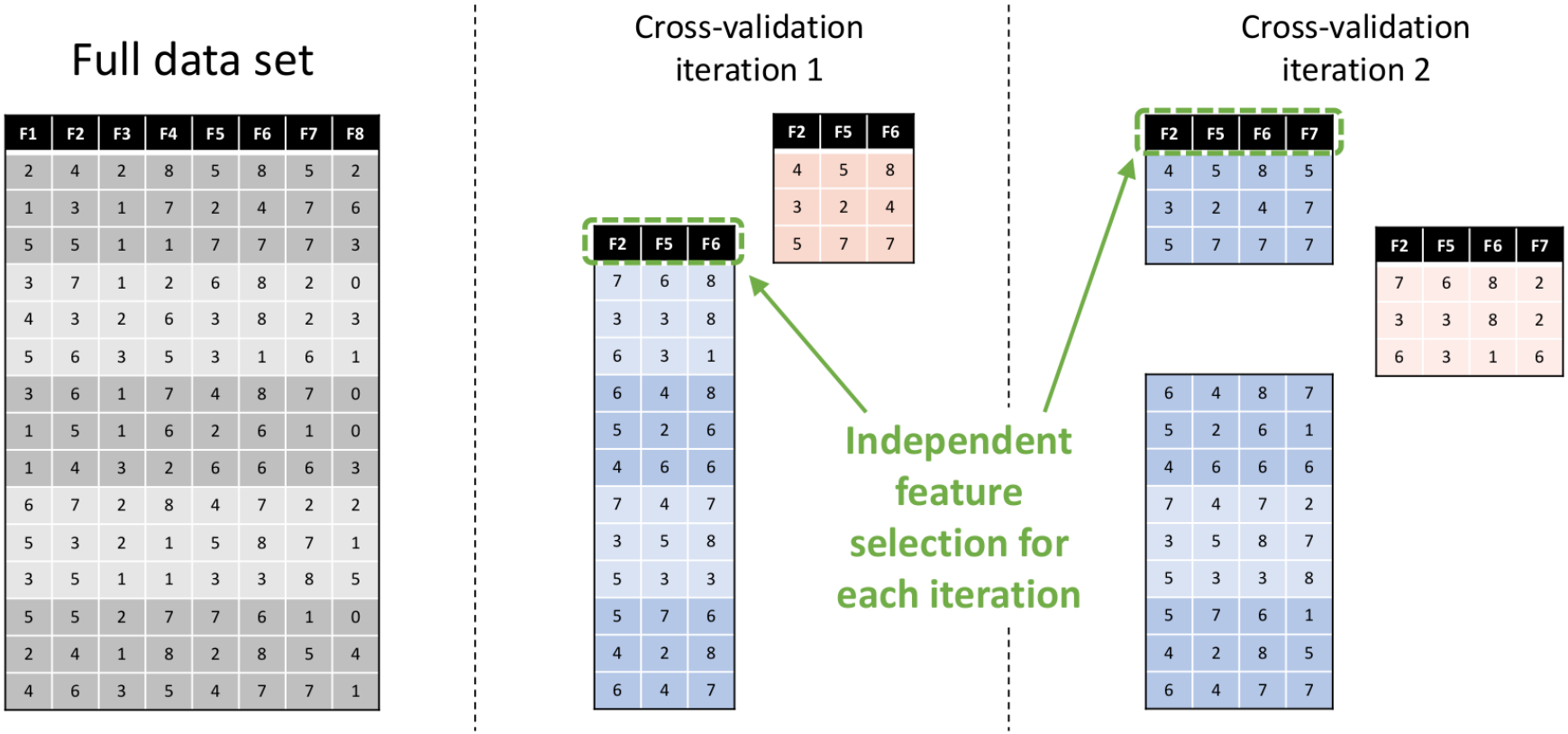


Figure from [Lones, 2024]

Algorithm Execution

Benchmark Developer

Why "benchmarks"?

- Can support introducing a "quasi standard" for
  - Task definition
  - Evaluation execution
  - KPI measurement
- Improves repeatability
- Eases the research in an area

Algorithm Execution

Benchmark Developer

- Benchmarks should be
  - Based on the problem definition
  - Independent of the system implementation

[1] https://projects.ics.forth.gr/isl/RDF-Benchmarks-Tutorial/index.html

Algorithm Execution

Benchmark Developer

- Benchmarks should be
  - Based on the problem definition
  - Independent of the system implementation

Do not trust the benchmarked system!

Do not rely on measures provided by the system

- Systems may
  - Answer with an unexpected value
  - Not answer at all
  - Misbehave

[1] https://projects.ics.forth.gr/isl/RDF-Benchmarks-Tutorial/index.html

Algorithm Execution

Benchmark Developer

- Benchmarks should be
  - Based on the problem definition
  - Independent of the system implementation

- Do not trust the benchmarked system!

  - Do not rely on measures provided by the system
  - Systems may
    - Answer with an unexpected value
    - Not answer at all
    - Misbehave
- Different colored systems...
  - Blackbox (Typically, the easiest way)
  - Grey/Whitebox
    - Allows for additional analysis
    - Choke-point-based design[1]

    - Assumptions might be unfair

[1] https://projects.ics.forth.gr/isl/RDF-Benchmarks-Tutorial/index.html

Section 3

KPIs

KPIs

Overview

Effectiveness vs. Efficiency

KPIs

Overview

Effectiveness vs. Efficiency

 An evaluation should cover both

KPIs

Effectiveness

Examples

- Accuracy, Precision, Recall, F1
- ROC curve, REC curve
- MRR, Hits@N
- (R)MSE
- Perplexity
- Correlations (Pearson, Spearman, Kendall's Tau)
- Similarity / distance measures (Cosine, KL-divergence, BLUE, ROUGE, Meteor)

KPIs

Efficiency

Examples

- Qps, QMpH
- CPU, user or wall clock time
- RAM
- Disk space
- $
- CO_2 equivalent

Wall clock time is an easy way to cover Efficiency

KPIs

Hints and Pitfalls

Be aware of a KPI's features and limitations: Accuracy

F1	F2	F3	F4	F5	Class
2	4	2	8	5	A
1	3	1	7	2	A
5	5	1	1	7	A
3	7	1	2	6	A
4	3	2	6	3	A
5	6	3	5	3	B
3	6	1	7	4	B
1	5	1	6	2	B
1	4	3	2	6	B
6	7	2	8	4	B

Always predict class A

Predicted	Correct?
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	FALSE
A	FALSE
A	FALSE
A	FALSE
A	FALSE

Number of correct classifications

$$\text{Accuracy} = \frac{5}{10} = 50\%$$

Total number of classifications

F1	F2	F3	F4	F5	Class
2	4	2	8	5	A
1	3	1	7	2	A
5	5	1	1	7	A
3	7	1	2	6	A
4	3	2	6	3	A
5	6	3	5	3	A
3	6	1	7	4	A
1	5	1	6	2	A
1	4	3	2	6	A
6	7	2	8	4	B

Always predict class A

Predicted	Correct?
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	TRUE
A	FALSE

Number of correct classifications

$$\text{Accuracy} = \frac{9}{10} = 90\%$$

Total number of classifications

Figure from [Lones, 2024]

KPIs

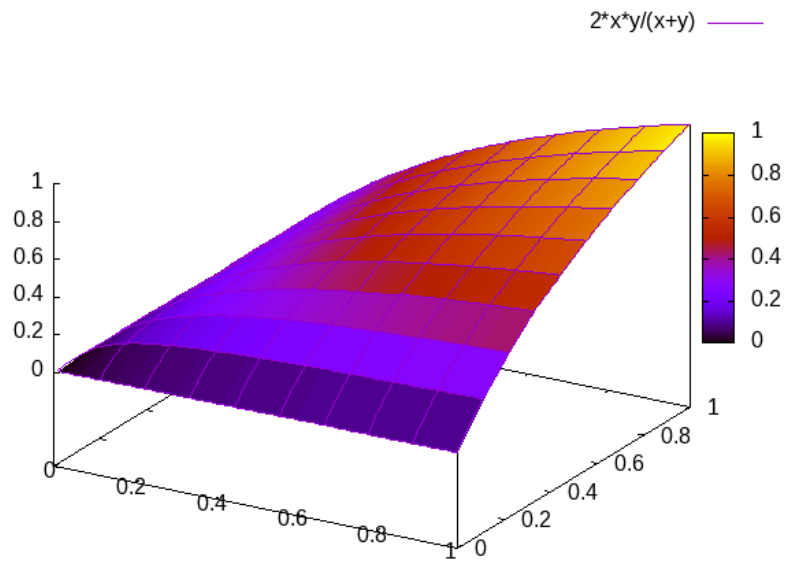
Hints and Pitfalls

⚠ Be aware of a KPI's features and limitations: F1-score

		Ground truth	
		True	False
Pred.	True	TP	FP
	False	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

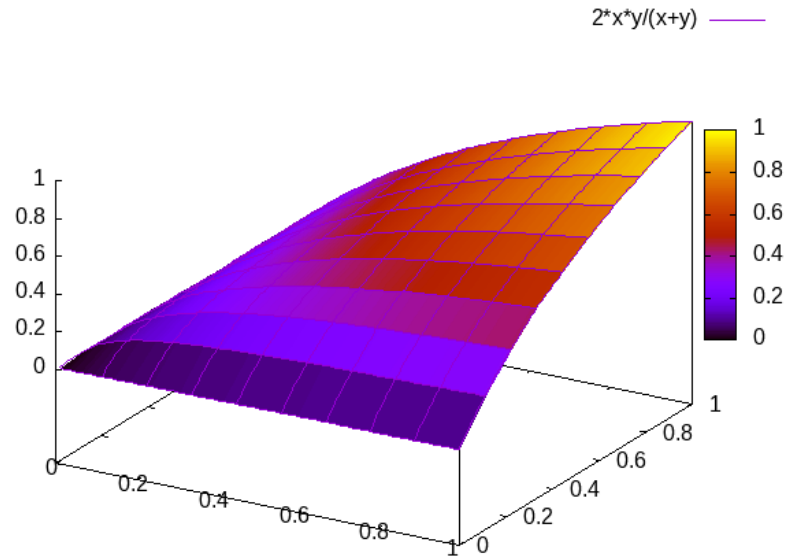


KPIs

Hints and Pitfalls

Be aware of a KPI's features and limitations: F1-score

		Ground truth	
		True	False
Pred.	True	TP	FP
	False	FN	TN



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

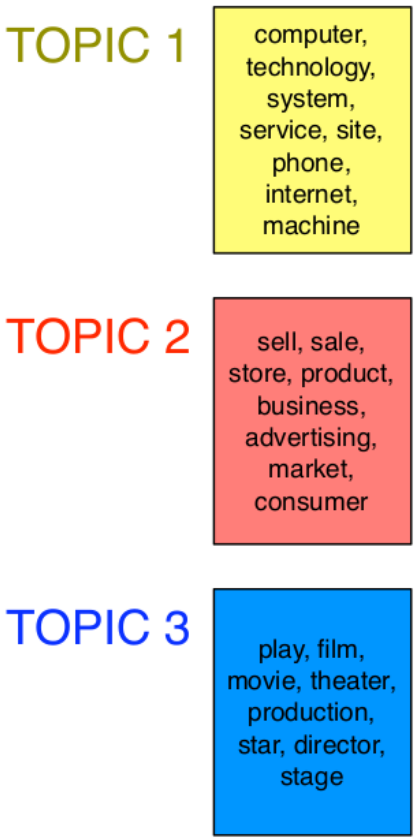
Be aware of corner cases: Precision and Recall

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

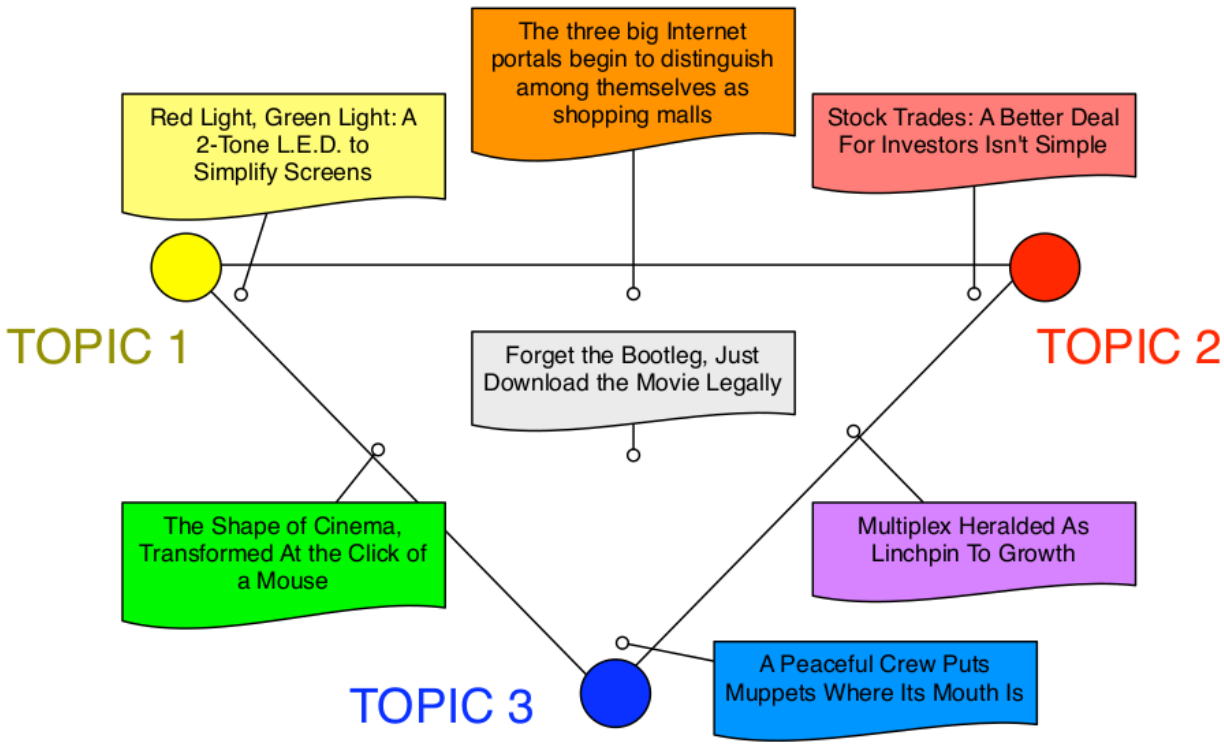
KPIs

Hints and Pitfalls

Ensure that the KPI fits to your goal



(a) Topics



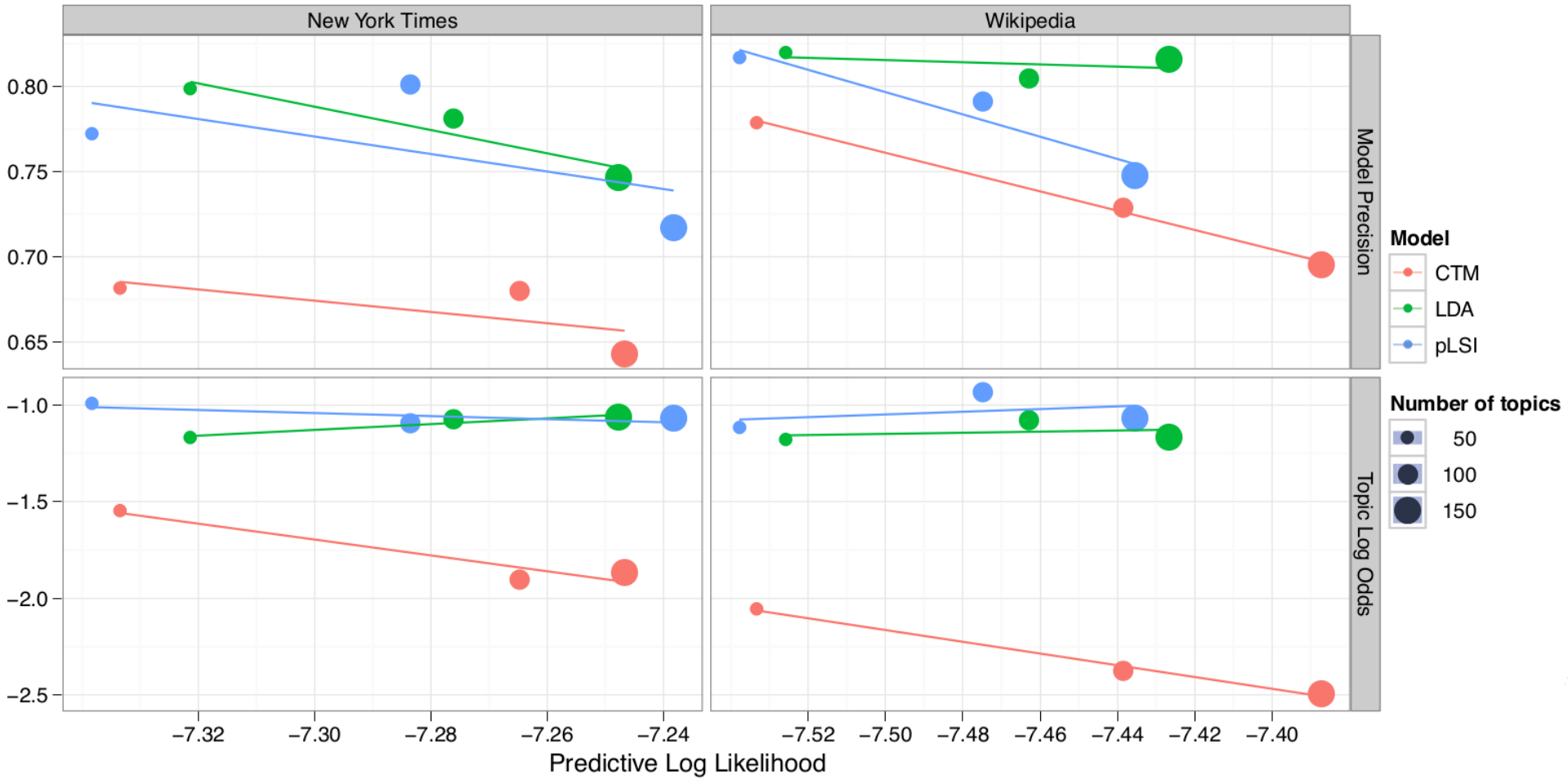
(b) Document Assignments to Topics

Figures from [Chang et al., 2009]

KPIs

Hints and Pitfalls

Ensure that the KPI fits to your goal



Figures from [Chang et al., 2009]

KPIs

Hints and Pitfalls

- Summarizations
  - Micro, macro, weighted average
  -  Arithmetic mean should always come with the standard deviation

KPIs

Hints and Pitfalls

- Summarizations
  - Micro, macro, weighted average
  -  Arithmetic mean should always come with the standard deviation
-  Comparison is necessary
  - Related work
  -  Baselines
    - Random guesser
    - Most frequent class

KPIs

Human Evaluation

What about humans?

Figure from `https://i.redd.it/7sritiz8e3951.jpg`

KPIs

Human Evaluation

What about humans?

However, human evaluation is sometimes needed...

Figure from `https://i.redd.it/7sritiz8e3951.jpg`

KPIs

Human Evaluation

- If your experiment is more complex, please get help / look for other sources to prepare it!

The description is very important; even "experts" might not always know what you want

KPIs

Human Evaluation

- If your experiment is more complex, please get help / look for other sources to prepare it!

The description is very important; even "experts" might not always know what you want

Try to use established evaluation shemes, e.g.,

- System Usability Scale [Brooke, 1996]
- Pairwise comparisons might be easier than absolute ratings [Saaty, 2008]
- Word or topic intruder experiment [Chang et al., 2009]

An established scheme gives no guarantee of success

KPIs

Human Evaluation

- If your experiment is more complex, please get help / look for other sources to prepare it!

The description is very important; even "experts" might not always know what you want

Try to use established evaluation shemes, e.g.,

- System Usability Scale [Brooke, 1996]
- Pairwise comparisons might be easier than absolute ratings [Saaty, 2008]
- Word or topic intruder experiment [Chang et al., 2009]

An established scheme gives no guarantee of success

Test your evaluation with some participants

Use your network to get participants

KPIs

Human Evaluation

- If your experiment is more complex, please get help / look for other sources to prepare it!

The description is very important; even "experts" might not always know what you want

Try to use established evaluation shemes, e.g.,

- System Usability Scale [Brooke, 1996]
- Pairwise comparisons might be easier than absolute ratings [Saaty, 2008]
- Word or topic intruder experiment [Chang et al., 2009]

An established scheme gives no guarantee of success

Test your evaluation with some participants

Use your network to get participants

It can be possible to mimik human raters [Röder et al., 2015]

KPIs

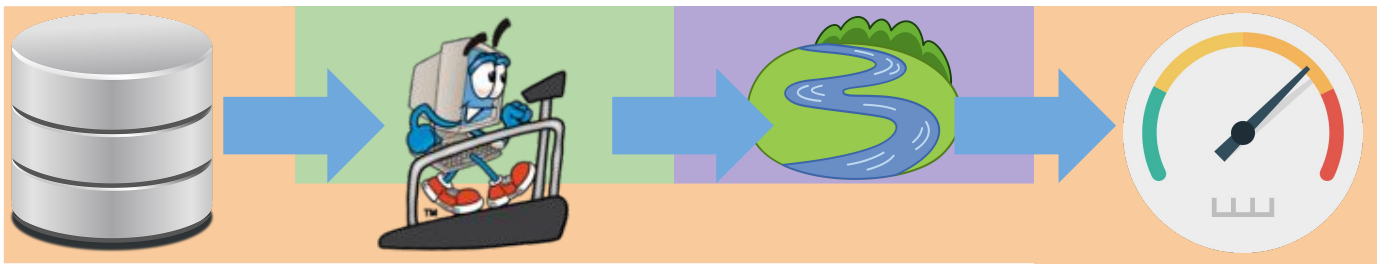
Intrinsic vs. Extrinsic Evaluation

- What if we cannot find any good KPI?

KPIs

Intrinsic vs. Extrinsic Evaluation

- What if we cannot find any good KPI?
- Extrinsic evaluation with a downstream task [Jurafsky and Martin, 2020]



- Examples:
  - Link prediction for knowledge graph embeddings

Section 4

Combination

Experiment Planning

Plan your experiments beforehand!

Do you need specific hardware?

Experiment Planning

Plan your experiments beforehand!


Do you need specific hardware?


How long will your experiment(s) take?

- Try to estimate the runtime
  - Try on a “typical” dataset
  - 
    If you tried it on a small dataset, take runtime complexity into account

Experiment Planning

Plan your experiments beforehand!


Do you need specific hardware?


How long will your experiment(s) take?

- Try to estimate the runtime
  - Try on a “typical” dataset
  - 
    If you tried it on a small dataset, take runtime complexity into account
- Example: Experiments measuring the robustness of knowledge graph embeddings against non-adversarial attacks
  - 7 Datasets
  - 5 Knowledge Graph Embedding algorithms
  - 4 Pertubation ratios
  - 3 Attack surfaces
  - Attacks are random $\rightarrow$ repeat 5 times

Experiment Planning

Plan your experiments beforehand!


Do you need specific hardware?


How long will your experiment(s) take?

- Try to estimate the runtime
  - Try on a “typical” dataset
  - 
    If you tried it on a small dataset, take runtime complexity into account
- Example: Experiments measuring the robustness of knowledge graph embeddings against non-adversarial attacks
  - 7 Datasets
  - 5 Knowledge Graph Embedding algorithms
  - 4 Pertubation ratios
  - 3 Attack surfaces
  - Attacks are random $\rightarrow$ repeat 5 times
  - $=$ 2100 Experiments
  - with $\sim$1h per $=$ 2100h $=$ 87.5 days

Experiment Planning

How long will your experiment(s) take? (cont.)

Solution 1: Run experiments in parallel

Don’t run multiple experiments in parallel on the same machine

Ensure that the hardware and software are equal (comparability)

Experiment Planning

How long will your experiment(s) take? (cont.)

Solution 1: Run experiments in parallel

Don't run multiple experiments in parallel on the same machine

Ensure that the hardware and software are equal (comparability)

Solution 2: Prioritization

- Which experiments are the most important?
- Example above: start with the most important dataset / model combination

Is one of the experiments more important than the others?

Prioritize as explained above

(Also known as "fail fast")

Summary Over Datasets

coherences		Name	C_V	C_P	C_{UMass}	$C_{\text{one-any}}$	C_{UCI}	C_{NPMI}	C_A
		$\mathcal{S}$	S^{one}_{set}	S^{one}_{pre}	S^{one}_{pre}	S^{one}_{any}	S^{one}_{one}	S^{one}_{one}	S^{one}_{one}
		$\boldsymbol{\mathcal{P}}$	$\mathcal{P}_{sw(110)}$	$\mathcal{P}_{sw(70)}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{cw(5)}$
		$\mathcal{M}$	$\tilde{m}_{cos(nlr,1)}$	m_f	m_{lc}	m_d	m_{lr}	m_{nlr}	$\tilde{m}_{cos(nlr,1)}$
		Σ	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a
using corpus		20NG	0.665	0.756	0.395	0.563	0.312	0.486	0.563
		Genomics	0.671	0.652	0.514	0.549	0.624	0.630	0.632
		RTL-Wiki	0.627	0.615	0.272	0.545	0.527	0.573	0.542
		Movie	0.548	0.549	0.093	0.453	0.473	0.438	0.431
		average	0.628	0.643	0.319	0.528	0.484	0.532	0.542
using the Wikipedia	$N = 10$	20NG	0.859	0.825	0.562	0.822	0.696	0.780	0.739
		Genomics	0.773	0.721	0.442	0.452	0.478	0.594	0.530
		NYT	0.803	0.757	0.543	0.612	0.783	0.806	0.747
		average	0.812	0.768	0.516	0.629	0.652	0.727	0.672
	$N = 5$	RTL-NYT	0.728	0.720	0.106	0.438	0.631	0.678	0.687
		RTL-Wiki	0.679	0.645	0.350	0.499	0.558	0.606	0.602
		Movie	0.544	0.533	0.143	0.454	0.447	0.452	0.465
		average	0.650	0.633	0.200	0.464	0.545	0.579	0.585
		average	**0.731**	0.700	0.358	0.546	0.599	0.653	0.628

Figure from [Röder et al., 2015]

Summary Over Datasets

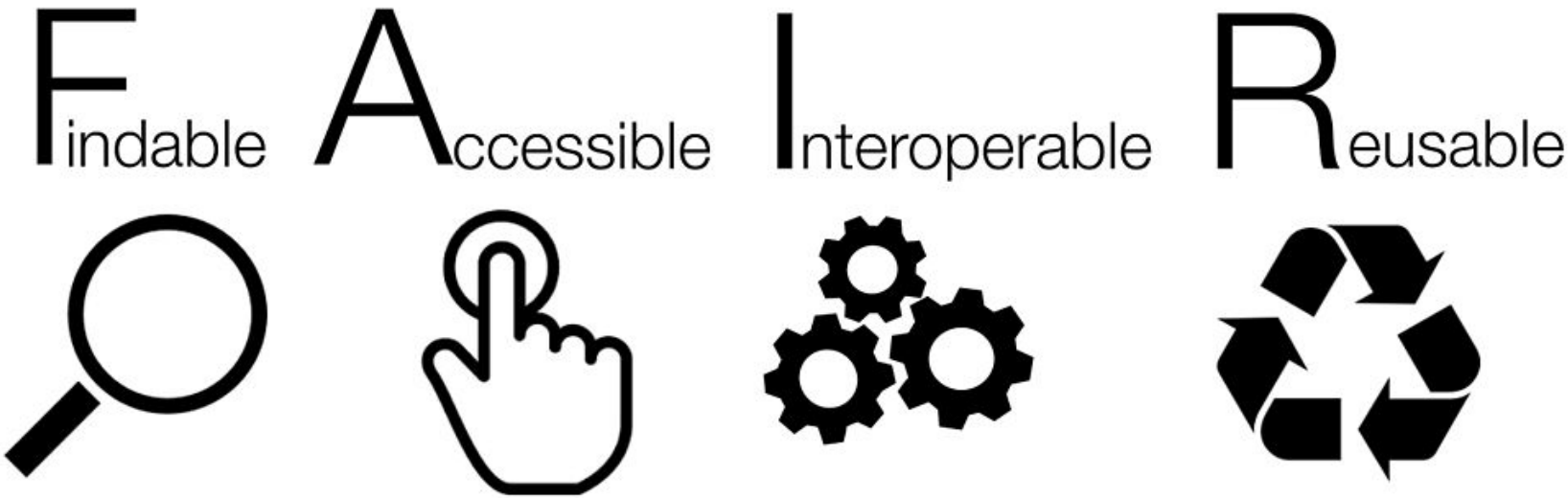
			C_V	C_P	C_{UMass}	$C_{\text{one-any}}$	C_{UCI}	C_{NPMI}	C_A
coherences		Name							
		$\mathcal{S}$	S^{one}_{set}	S^{one}_{pre}	S^{one}_{pre}	S^{one}_{any}	S^{one}_{one}	S^{one}_{one}	S^{one}_{one}
		$\mathcal{P}$	$\mathcal{P}_{sw(110)}$	$\mathcal{P}_{sw(70)}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{cw(5)}$
		$\mathcal{M}$	$\tilde{m}_{cos(nlr,1)}$	m_f	m_{lc}	m_d	m_{lr}	m_{nlr}	$\tilde{m}_{cos(nlr,1)}$
		Σ	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a
using corpus		20NG	0.665	0.756	0.395	0.563	0.312	0.486	0.563
		Genomics	0.671	0.652	0.514	0.549	0.624	0.630	0.632
		RTL-Wiki	0.627	0.615	0.272	0.545	0.527	0.573	0.542
		Movie	0.548	0.549	0.093	0.453	0.473	0.438	0.431
		average	0.628	0.643	0.319	0.528	0.484	0.532	0.542
using the Wikipedia	$N = 10$	20NG	0.859	0.825	0.562	0.822	0.696	0.780	0.739
		Genomics	0.773	0.721	0.442	0.452	0.478	0.594	0.530
		NYT	0.803	0.757	0.543	0.612	0.783	0.806	0.747
		average	0.812	0.768	0.516	0.629	0.652	0.727	0.672
	$N = 5$	RTL-NYT	0.728	0.720	0.106	0.438	0.631	0.678	0.687
		RTL-Wiki	0.679	0.645	0.350	0.499	0.558	0.606	0.602
		Movie	0.544	0.533	0.143	0.454	0.447	0.452	0.465
		average	0.650	0.633	0.200	0.464	0.545	0.579	0.585
		average	**0.731**	0.700	0.358	0.546	0.599	0.653	0.628

Averaging over datasets is in many cases not allowed [Demšar, 2006]

Use statistical tests instead (e.g., Wilcoxon Signed-Rank test)

Figure from [Röder et al., 2015]

Be "FAIR"



That's all!

References I

[Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012).
Random search for hyper-parameter optimization.
J. Mach. Learn. Res., 13(null):281–305.

[Brooke, 1996] Brooke, J. (1996).
SUS: a "quick and dirty" usability.
Usability evaluation in industry, 189(3):189–194.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009).
Reading tea leaves: How humans interpret topic models.
In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

References II

[Cohen, 1960] Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46.

[Demšar, 2006] Demšar, J. (2006).
Statistical Comparisons of Classifiers over Multiple Data Sets.
The Journal of Machine Learning Research, 7:1–30.

[Duan et al., 2011] Duan, S., Kementsietsidis, A., Srinivas, K., and Udrea, O. (2011).
Apples and oranges: A comparison of rdf benchmarks and real rdf datasets.
In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, page 145–156, New York, NY, USA. Association for Computing Machinery.

References III

[Fleiss and Cohen, 1973] Fleiss, J. L. and Cohen, J. (1973).
The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.
Educational and Psychological Measurement, 33(3):613–619.

[Guo et al., 2005] Guo, Y., Pan, Z., and Heflin, J. (2005).
LUBM: A benchmark for OWL knowledge base systems.
Journal of Web Semantics, 3(2):158–182.
Selected Papers from the International Semantic Web Conference, 2004.

[Hripcsak and Rothschild, 2005] Hripcsak, G. and Rothschild, A. S. (2005).
Agreement, the f-measure, and reliability in information retrieval.
Journal of the American Medical Informatics Association, 12(3):296–298.

References IV

[Jha et al., 2017] Jha, K., Röder, M., and Ngonga Ngomo, A.-C. (2017).
All That Glitters is not Gold – Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking.
In *The Semantic Web. Latest Advances and New Domains: 14th International Conference, ESWC 2017, Proceedings*. Springer International Publishing.

[Jurafsky and Martin, 2020] Jurafsky, D. and Martin, J. H. (2020).
Speech and Language Processing.
3rd (draft) edition.
Last time accessed, July 20th, 2021.

[Lones, 2024] Lones, M. A. (2024).
Avoiding common machine learning pitfalls.
Patterns, 5(10):101046.

References V

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

[Röder et al., 2015] Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

References VI

[Röder et al., 2021] Röder, M., Nguyen, P. T. S., Conrads, F., da Silva, A. A. M., and Ngomo, A.-C. N. (2021).
LEMMING – Example-based Mimicking of Knowledge Graphs.
In *Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC)*, pages 62–69. IEEE Computer Society.

[Saaty, 2008] Saaty, T. L. (2008).
Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process.
RACSAM - Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 102:251–318.

References VII

[Shchur et al., 2018] Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. (2018).
Pitfalls of graph neural network evaluation.
CoRR, abs/1811.05868.

[Shweta et al., 2015] Shweta, Bajpai, R., and Chaturvedi, H. C. (2015).
Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods.
Journal of the Indian Academy of Applied Psychology, 41(3):20–27.

[Taelman et al., 2019] Taelman, R., Colpaert, P., Mannens, E., and Verborgh, R. (2019).
Generating Public Transport Data based on Population Distributions for RDF Benchmarking.
Semantic Web Journal, 10(2):305–328.

References VIII

[Vogel and Jiang, 2019] Vogel, I. and Jiang, P. (2019).
Fake news detection with the new german dataset “germanfakenc”.
In Doucet, A., Isaac, A., Golub, K., Aalberg, T., and Jatowt, A., editors, *Digital Libraries for Open Knowledge*, pages 288–295, Cham. Springer International Publishing.