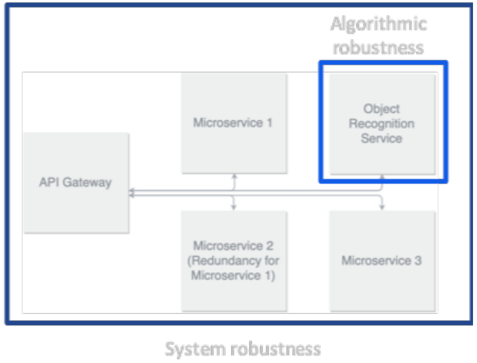
Engineering best practices for machine learning

Alex Serban

Radboud University, Leiden University, Software Improvement Group
The Netherlands

Machine learning robustness



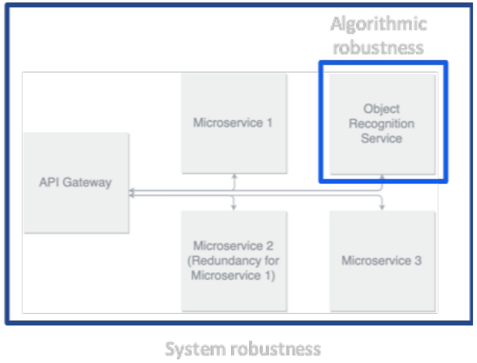
Robustness has multiple facets, e.g., algorithmic robustness, system or software robustness

Algorithmic robustness is the ability to maintain training performance when tested on new and noisy samples

System robustness is the ability to cope with errors and erroneous inputs during execution

In ML, the boundaries between robustness and trustworthiness erode, s.t. robustness may include fairness, privacy, transparency, etc.

Machine learning robustness

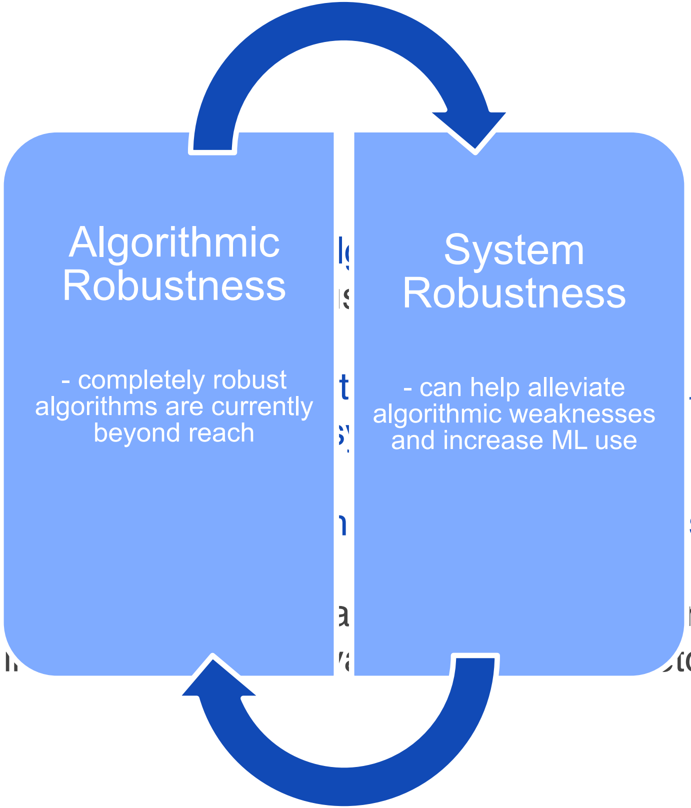


Robustness has m … , system or software

Algorithmic robustness … nce when tested on new

System robustness is th … s inputs during execution

In ML, the boundaries b … rode, s.t. robustness may … tc.



Robustness in the wild

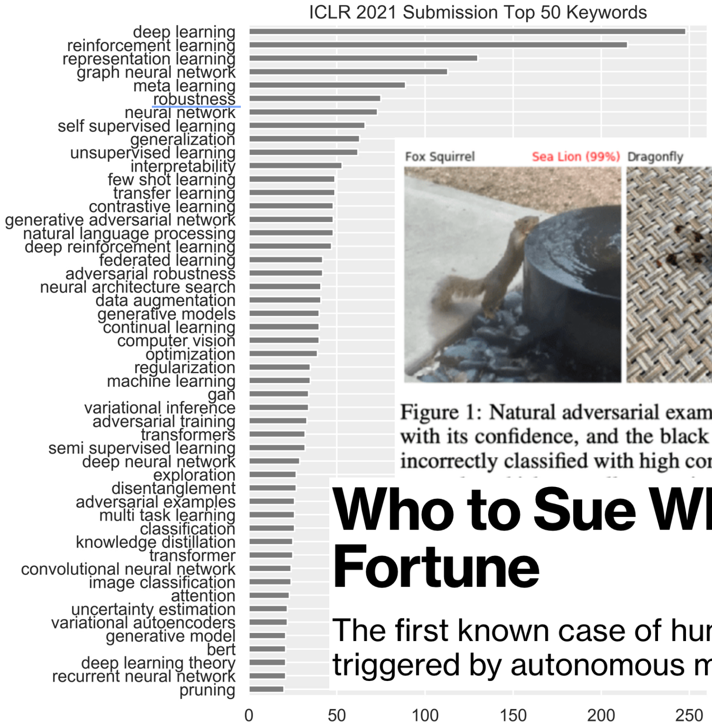
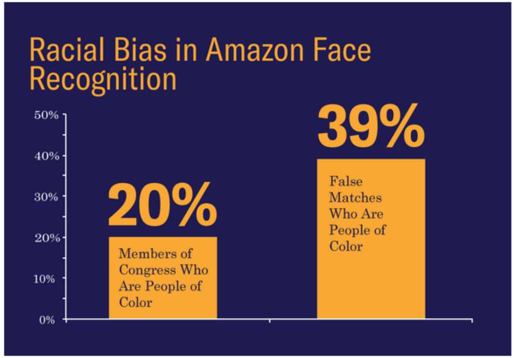


Figure 1: Natural adversarial examples from IMAGENET-A. The red text is a ResNet-50 prediction with its confidence, and the black text is the actual class. Many natural adversarial examples are incorrectly classified with high confidence, despite having no adversarial modifications as they are



Who to Sue When a Robot Loses Your Fortune

The first known case of humans going to court over investment losses triggered by autonomous machines will test the limits of liability.

Robustness in policy

On 8 April 2019, the High-Level Expert Group on AI presented the **Ethics Guidelines for Trustworthy Artificial Intelligence**.

Trustworthy means:

- Lawful
- Ethical
- **Robust**

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Alex Serban
cs.ru.nl/~aserban

Good engineering, a prerequisite for building robust ML systems

How are software engineering practices **impacted** by use of ML components in software systems?

What best practices are being **proposed** by researchers and practitioners?

To what extent are practices **adopted** by engineering teams?

What are the **effects** of practices adoption on the quality of systems with ML components?

Investigating ML engineering best practices

SE4ML
se-ml.github.io

Academic and grey literature → **Review literature** → Awesome reading list
awesome

Create practice catalog → **29** practices in fixed format

400+ practitioners → **Survey adoption and effects** → **ranking** of practices and links to effects

Serban et al, "Adoption and effects of software engineering best practices in machine learning", ESEM 2020

Interview practitioners → "State of ML engineering practices" **report**

Add **trustworthiness** → **+14** practices and link to **seven** requirements

Add **AutoML** → "State of AutoML" report

Add **architecture** → **Tactics** linked to **quality** aspects

Online catalog of engineering practices for ML

Originally, **29** practices. Now grown to **45**.

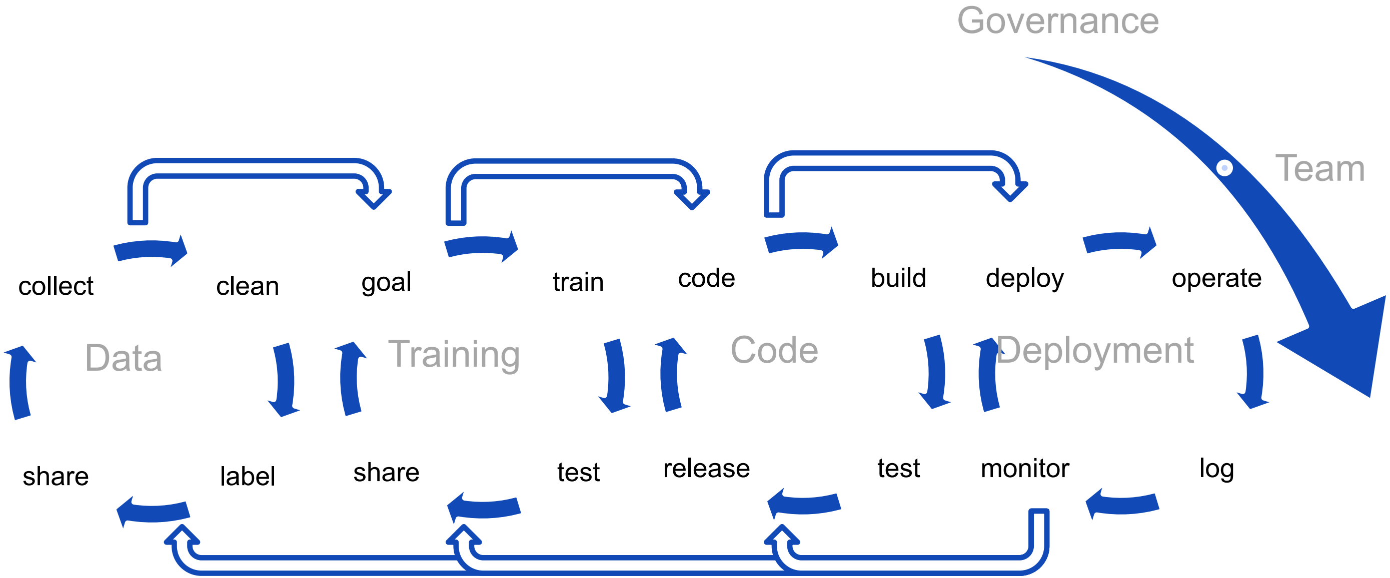
Grouped into **6** categories.

- Intent
- Motivation
- Applicability
- Description
- Adoption
- Related practices
- References

basic medium advanced

Online catalog of engineering practices for ML

The practice grouping can also be seen as a process mapping.



Alex Serban
cs.ru.nl/~aserban

Example practice

Title

Nr • Category • Difficulty

- Intent
- Motivation
- Applicability
- Description
- Adoption
- Related practices
- References

Use Sanity Checks for All External Data Sources

January, 2021 • Alex Serban, Koen van der Blom, Joost Visser

← 1 / 45 • Data • medium →

Difficulty

Category

Intent

Avoid invalid or incomplete data being processed.

Motivation

Data is at the heart of any machine learning model. Therefore, avoiding data errors is crucial for model quality.

Applicability

Data quality control should be applied to any machine learning application.

Description

Whenever external data sources are used, or data is collected that may be incomplete or ill formatted, it is important to verify the data quality. Invalid or incomplete data may cause outages in production or lead to inaccurate models.

Start by checking simple data attributes, such as:

- data types,
- missing values,
- data min. or max. values,
- histograms of continuous values,

and gradually include more complex data statistics, such as the ones recommended here.

Missing data can also be substituted using data imputation; such as imputation by zero, mean, median, random values, etc.

Also, make sure the data verification scripts are reusable and can be later integrated in any processing pipeline.

Measuring practice adoption

Survey among teams building software that incorporates ML components.

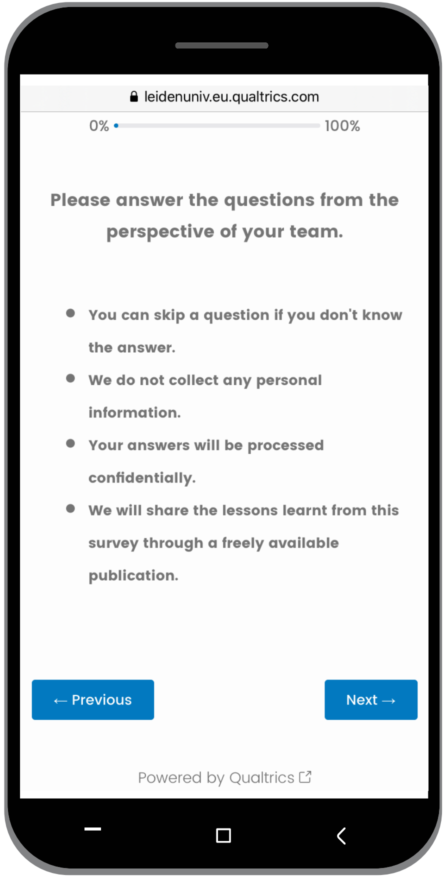
Questions:

- **General**
  ex. Team size, team experience, country, type of organization, type of data, tools used.

- **Practices**
  ex. ”Our process for deploying our ML model is fully automated.”

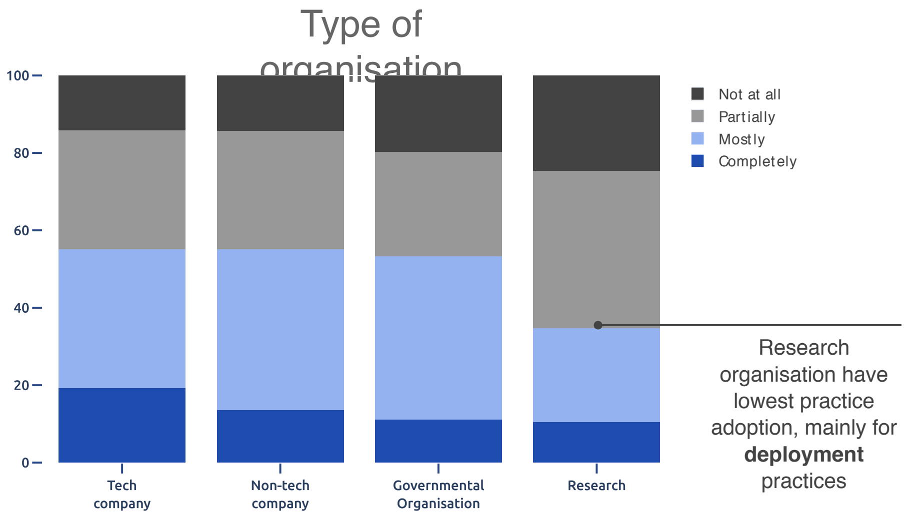
- **Effects**
  ex. “We are able to easily and precisely reproduce past behavior of our models and applications.”

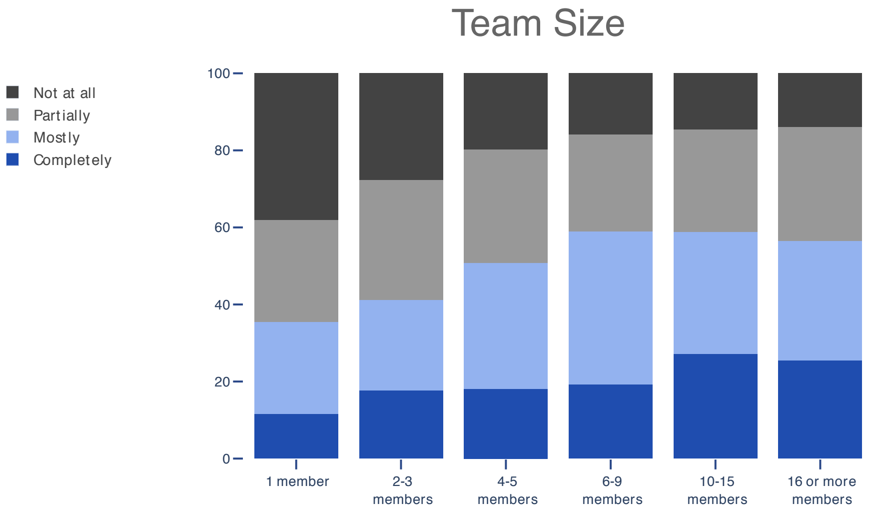
Not at all
Partially
Mostly
Completely



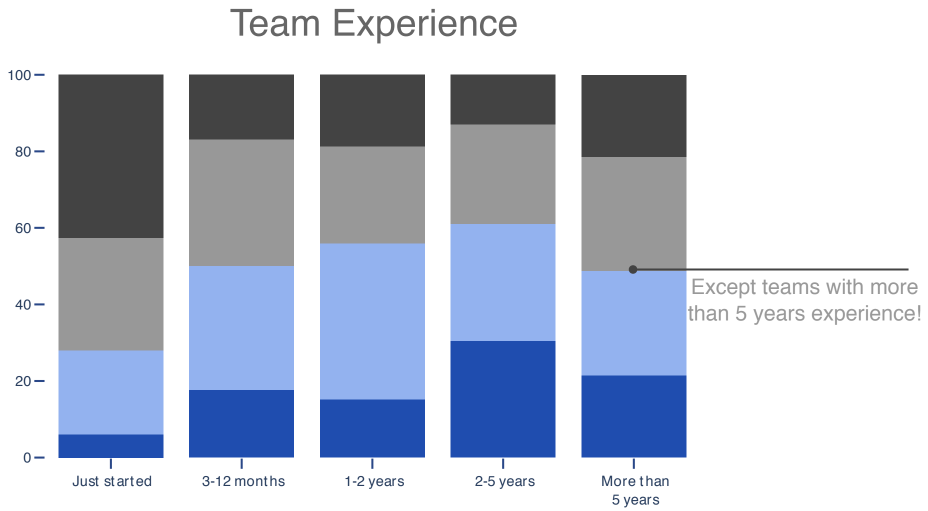
Tech companies lead practice adoption

The adoption of best practices by tech companies is higher than by non-tech companies, governmental organizations, and research labs.





Larger teams tend to adopt more practices.

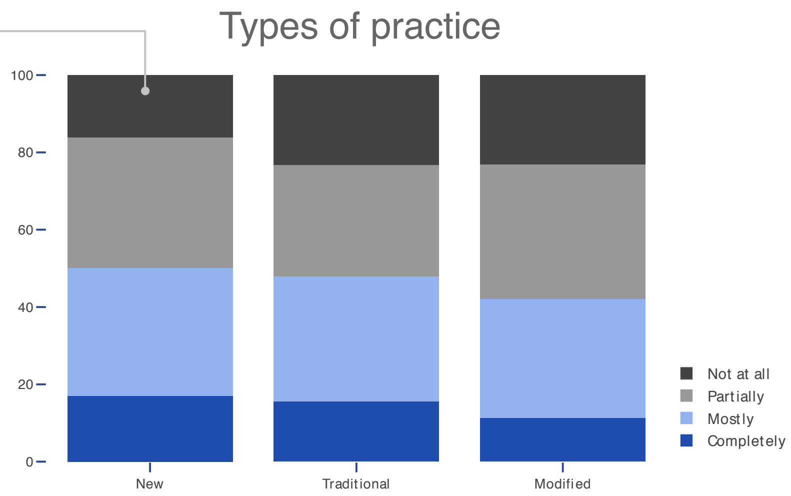


More experienced teams tend to adopt more practices.

Practice adoption increases with team size and experience

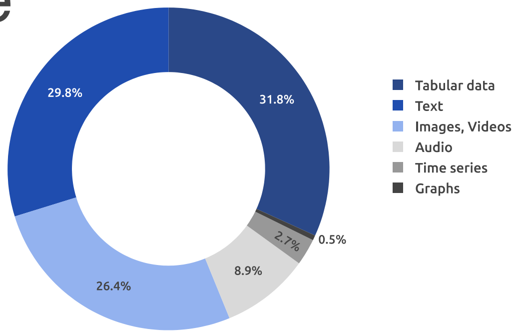
ML-specific practices are adopted slightly more than traditional SE practices

ML-specific practices enjoy the highest degree of adoption

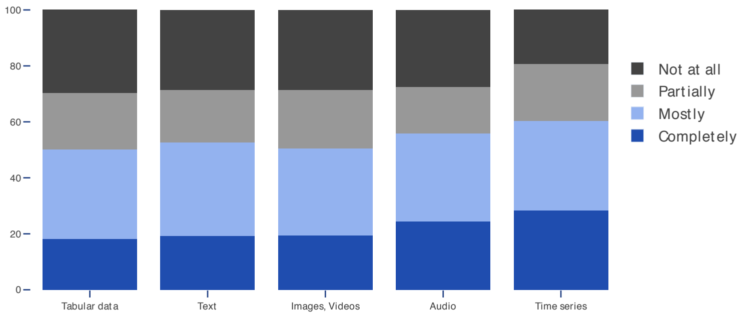


Among ML teams, the adoption of ML-specific practices is highest, followed by general Software Engineering (SE) practices and SE practices adapted to ML.

Practice adoption by data type



The adoption of practices is largely **independent** of the data type used



back to our

Example practice

Title

Nr • Category • Difficulty

- Intent
- Motivation
- Applicability
- Description
- Adoption
- Related practices
- References

Use Sanity Checks for All External Data Sources

January, 2021 • Alex Serban, Koen van der Blom, Joost Visser

Difficulty

1 / 45 • Data • medium

Category

Intent

Avoid invalid or incomplete data being processed.

Motivation

Data is at the heart of any machine learning model. Therefore, avoiding data errors is crucial for model quality.

Applicability

Data quality control should be applied to any machine learning application.

Description

Whenever external data sources are used, or data is collected that may be incomplete or ill formatted, it is important to verify the data quality. Invalid or incomplete data may cause outages in production or lead to inaccurate models.

Start by checking simple data attributes, such as:

- data types,
- missing values,
- data min. or max. values,
- histograms of continuous values,

and gradually include more complex data statistics, such as the ones recommended here.

Missing data can also be substituted using data imputation; such as imputation by zero, mean, median, random values, etc.

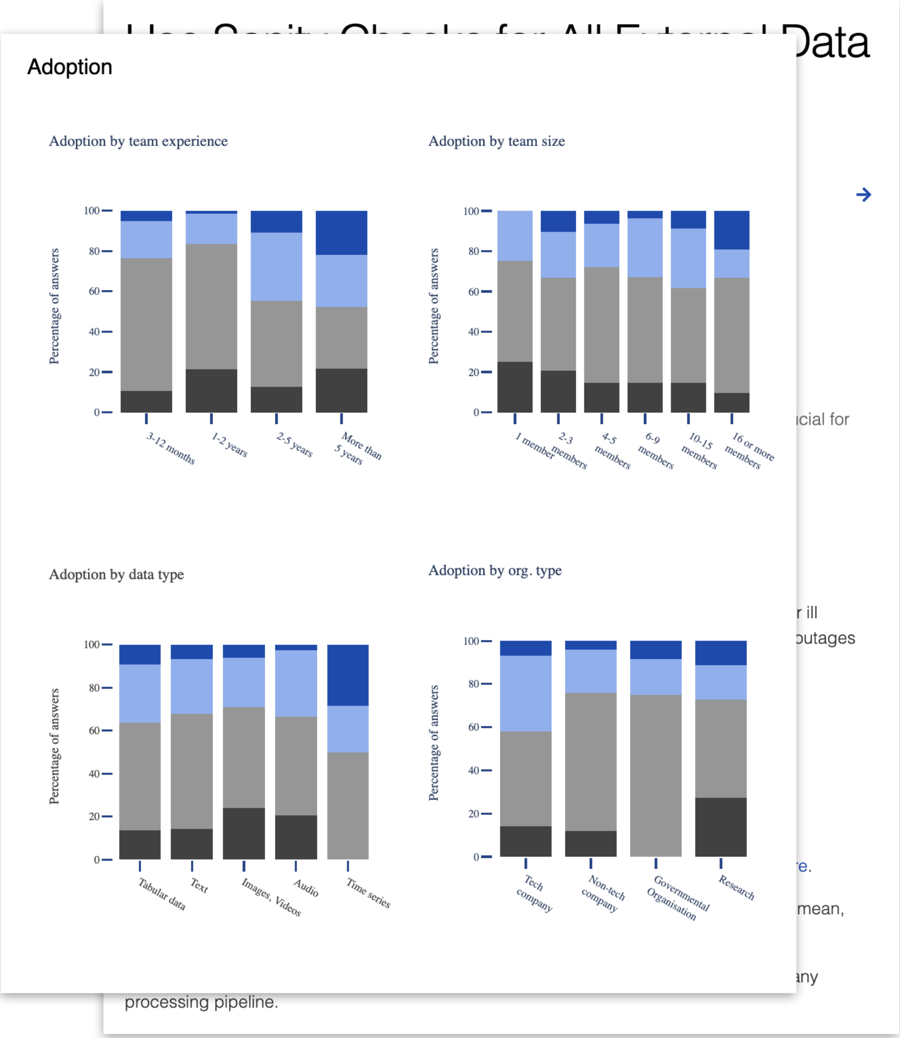
Also, make sure the data verification scripts are reusable and can be later integrated in any processing pipeline.

Example practice

Title

Nr • Category • Difficulty

- Intent
- Motivation
- Applicability
- Description
- **Adoption**
- Related practices
- References



- Not at all
- Partially
- Mostly
- Completely

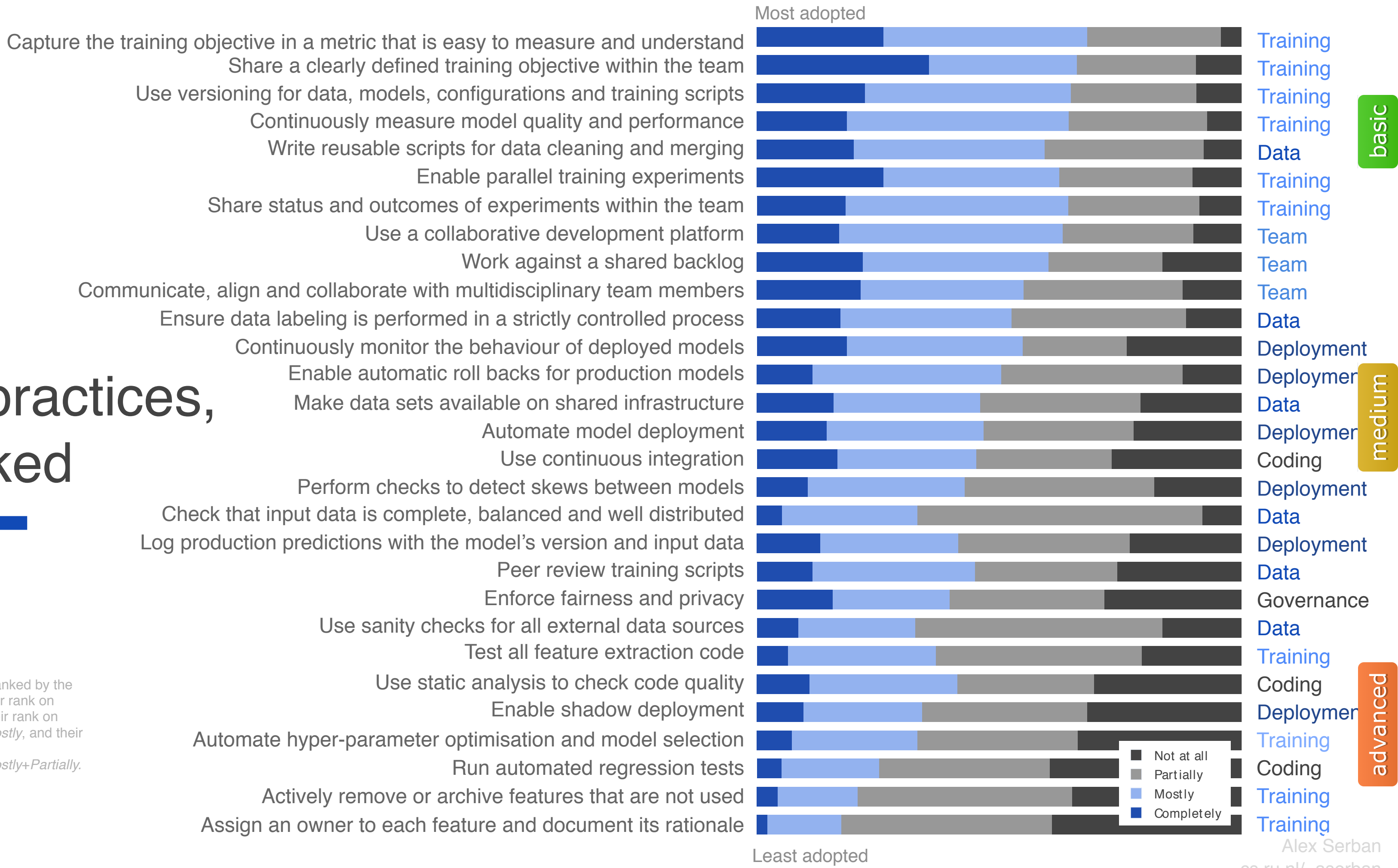
Example practice

Title

Nr • Category • Difficulty

- Intent
- Motivation
- Applicability
- Description
- Adoption
- Related practices
- References

29 practices, ranked



Practices are ranked by the average of: their rank on *Completely*, their rank on *Completely+Mostly*, and their rank on *Completely+Mostly+Partially.*

Alex Serban
cs.ru.nl/~aserban

Most adopted practices

Practices related to **measurement** and **versioning**
are widely adopted.

The top 4 adopted practices are all related to **model training.**

Top 5

1. Capture the training objective in a metric that is easy to measure and understand
2. Share a clearly defined training objective within the team
3. Use versioning for data, model, configurations and training scripts
4. Continuously measure model quality and performance
5. Write reusable scripts for data cleaning and merging

Alex Serban
cs.ru.nl/~aserban

Least adopted practices

The two most neglected practices are related to **feature management**.

Outside research, **Automated ML** through automated optimisation of hyper-parameters and model selection, is not (yet) widely applied.

Bottom 5

1. Assign an owner to each feature and document its rationale
2. Actively remove or archive features that are not used
3. Run automated regression tests
4. Automate hyper-parameter optimisation and Model Selection
5. Enable shadow deployment

Alex Serban
cs.ru.nl/~aserban

Measuring effects of practice adoption

For **four** effects, we hypothesized a relation with a specific selection of practices.

- **Linear regression**
  Confirmed hypotheses.
- **Non-linear regression – Random Forest**
  Demonstrated non-linear influence.
- **Importance of each practice – Shapley**
  Some very important practices have low adoption.

Effects	Description
Agility	The team can quickly experiment with new data and algorithms, and quickly assess and deploy new models
Software Quality	The software produced is of high quality (technical and functional)
Team Effectiveness	Experts with different skill sets (e.g., data science, software development, operations) collaborate efficiently
Traceability	Outcomes of production models can easily be traced back to model configuration and input data

Alex Serban
cs.ru.nl/~aserban

Different practices, different outcomes

Analysis of survey responses shows that desired outcomes such as **traceability**, **agility**, team **effectiveness**, and software **quality** are each related to specific sets of practices.

Per desired outcome, we list the three practices with the largest influence.

Agility

1. Automate model deployment
2. Communicate, align, and collaborate with multidisciplinary team members
3. Enable parallel training experiments

Traceability

1. Log production predictions with the model's version and input data
2. Continuously monitor the behavior of deployed models
3. Use versioning for data, model, configurations and training scripts

Team Effectiveness

1. Work against a shared backlog
2. Use a collaborative development platform
3. Share a clearly defined training objective within the team

Software Quality

1. Use continuous integration
2. Run automated regression tests
3. Use static analysis to check code quality

Practice importance for each effect

Using the importance of each practice to the effects we can suggest improvements.

Using the practice adoption as a proxy to difficulty we can plan and prioritize practice adoption

Engineering best practices for ML

How are software engineering practices **impacted** by incorporation of ML components in software systems?

What new practices are being **proposed** by researchers and practitioners?

To what extent are practices **adopted** by engineering teams?

What are the **effects** of practices adoption on the quality of systems that incorporate ML components?

Answers lead to new questions …

- **Trustworthiness**
  More practices? Link to **policy**?
- **Architecture**
  Practices as **tactics** to reach architectural goals.
- **AutoML**
  Transfer from research to broad adoption?

Seven key requirements

Evaluate and address these continuously throughout the AI system's lifecycle, via:

- **Technical methods**
  e.g., Constraints in the software architecture, embedded in design and implementation. Explanation functionality. Deliberate testing and validation. Measure algorithm quality indicators.

- **Non-technical methods**
  e.g., Regulations, code of conduct, standardization, certification, governance, education, awareness, stakeholder participation, diversity in design teams.

New practices, mapped to trustworthiness requirements

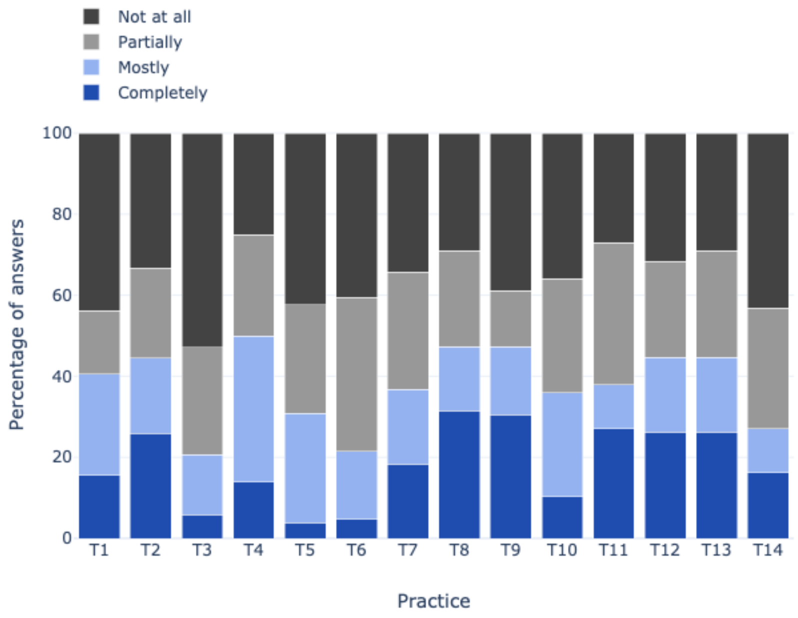
- Lawful
- Ethhical
- Robust

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, fairness
- Societal, environmental well-being
- Accountability

- T1 - Test for social bias in training data
- T2 - Prevent discriminatory data attributes as model features
- T3 – Use privacy-preserving ML techniques
- T4 – Employ interpretable models whenever possible
- T5 – Assess and manage subgroup bias
- T6 – Assure application security
- T7 – Provide audit trails
- T8 – Decide trade-offs through an established team process
- T9 – Establish responsible AI values
- T10 – Perform risk assessments
- T11 – Inform users on ML usage
- T12 – Explain results and decision to users
- T13 – Provide safe channels to raise concerns
- T14 – Have your application audited

Adoption of practices for trustworthy ML

- T1 - Test for social bias in training data
- T2 - Prevent discriminatory data attributes as model features
- T3 – Use privacy-preserving ML techniques
- T4 – Employ interpretable models whenever possible
- T5 – Assess and manage subgroup bias
- T6 – Assure application security
- T7 – Provide audit trails
- T8 – Decide trade-offs through an established team process
- T9 – Establish responsible AI values
- T10 – Perform risk assessments
- T11 – Inform users on ML usage
- T12 – Explain results and decision to users
- T13 – Provide safe channels to raise concerns
- T14 – Have your application audited



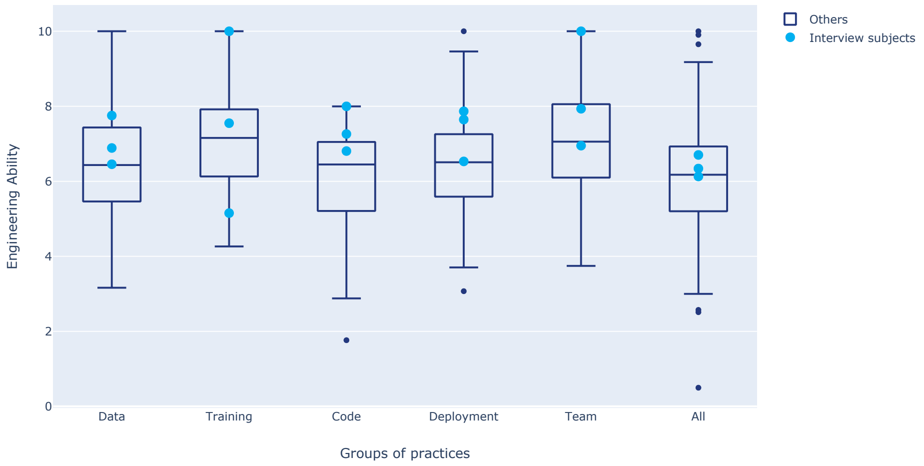
Adoption of practices as a proxy to ML engineering ability

We used psychometrics (IRT) to evaluate teams' ML engineering ability based on the practice adoption rate.

- **Difficulty of adoption per practice**
  Based on all answers.

- **Engineering ability per class of practices**
  And benchmarks against other teams.

- **Suggestions for improvements**
  Based on difficulty of adoption and importance for the effects.



Take away

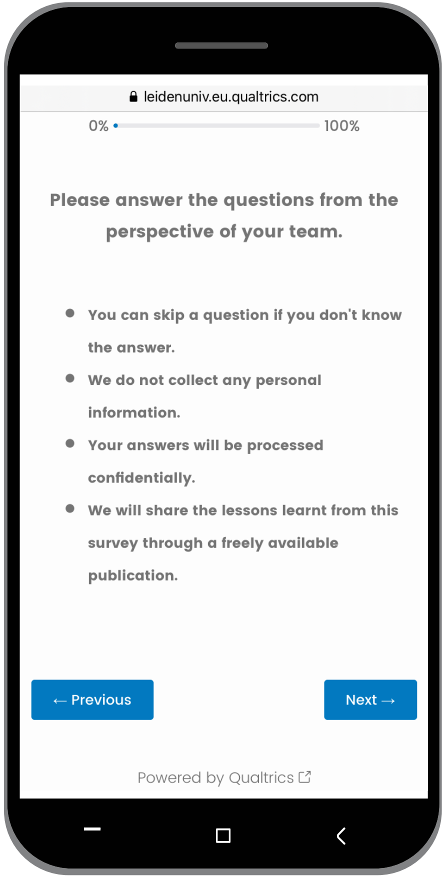
Demand for **robust** development and use are not unique to ML, but become more acute.

Good **engineering practices** are a **prerequisite** for quality attributes such as robustness or agility

Engineering **practices** are being modified and developed at a quick pace. **Adoption** varies and **effects** are not completely understood.

Robustness and **trustworthiness** get wide attention by policy makers and advisers, although **practitioners do not** adopt these practices.

You can help



Take the Survey

If you have not done so yet, please take our 10-min survey!

We will use your answers for our next report on the State of Engineering Practices for Machine Learning.

https://se-ml.github.io/survey

31

Learn more

Reading list

We reviewed scientific and popular literature to identify recommended practices. Check out this Awesome List with relevant literature.

Catalogue

The best practices that we identified are describe in more detail in this Catalogue of ML Engineering Best Practices.

Preprints

Full details of the methodology behind our survey are described in scientific articles. Read the preprints here.

se-ml.github.io

Visit our project website for more details, to take the survey yourself, and to stay up-to-date with our latest results.