Data Quality Assurance

Or „How to get good data“, by Florian Netzer & Lars Wolf

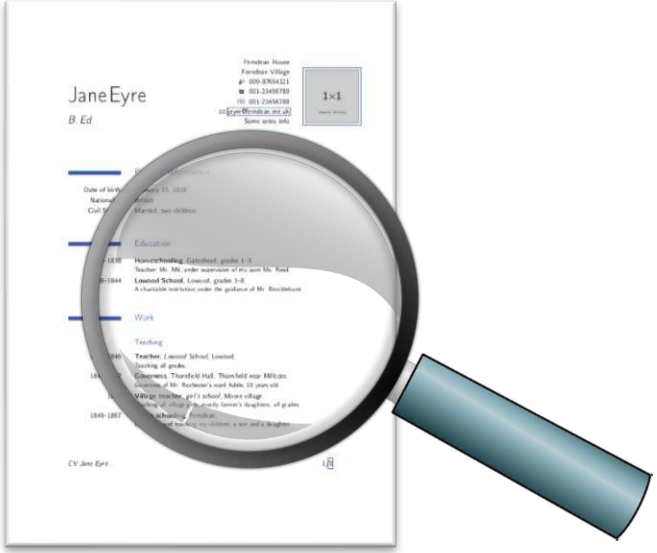


Image sources: stackexchange.com texwelt.de stackexchange.com CC BY-SA

Overview

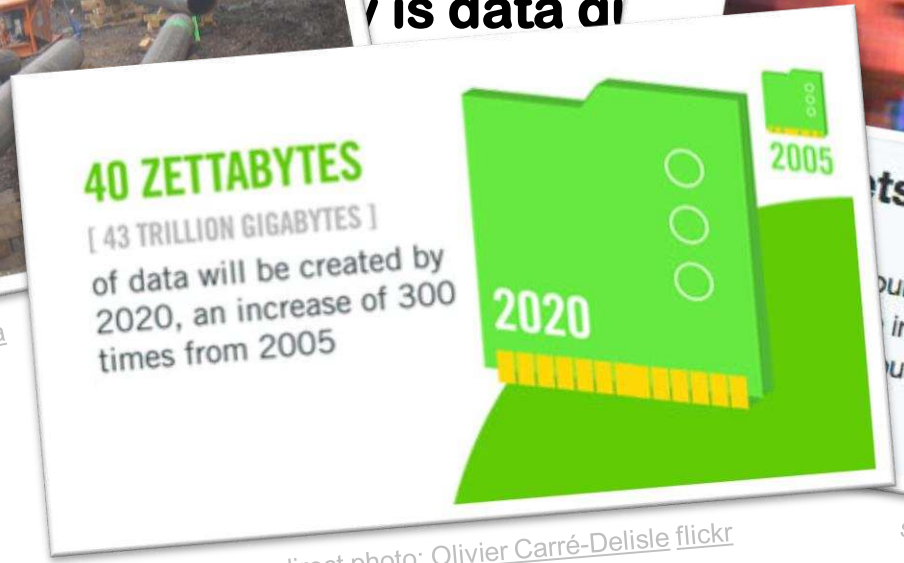
TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Data Collection**
   - What are potential problems?
   - How do you get good data?

2. **Data Cleaning**
   - What are potential problems?
   - How do you get clean data?

3. **Take-Aways**
   - Tools for your use
   - Summary

source: Peter Gaßner, wikipedia



source: sciencedirect photo: Olivier Carré-Delisle flickr

source: spiegel.de

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

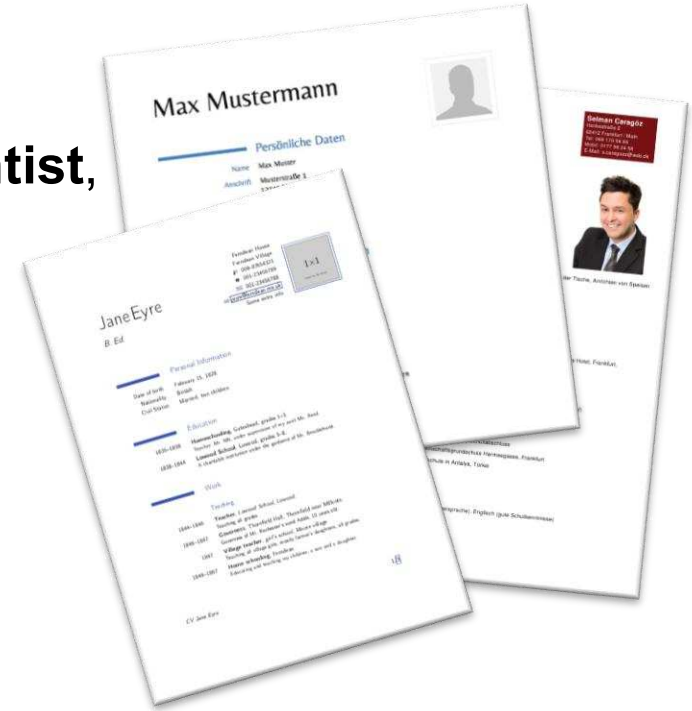
1. **Data Collection**
   - What are potential problems?
   - How do you get good data?

2. **Data Cleaning**
   - What are potential problems?
   - How do you get clean data?

3. **Take-Aways**
   - Tools for your use
   - Summary

Let‘s start with an example!

We tell **five people** on amazon mechanical turk to rate job applications of people applying as **data scientist**, on a **scale from 1 to 5**.



What could be potential problems?

Tell us on menti.com

Possible Issues

TECHNISCHE UNIVERSITÄT DARMSTADT

biases
e.g. men vs. woman

feedback-loops
„more men have been hired in the past, so they are better at SE“

interrater reliability
e.g. people using 5 stars only for the most perfect applications

out-of-date
„people knowing flash are great web developers“

noise
rater accidentally clicked wrong button

wrong data
e.g. people who are 18 yo and have 25 years of working experience“

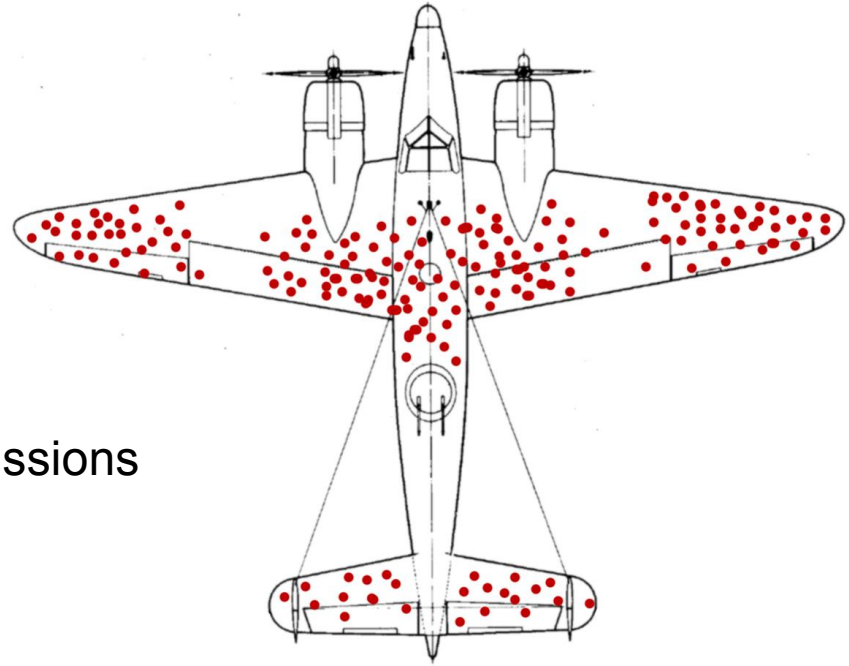
...and many more!

Where does your data come from?

TECHNISCHE UNIVERSITÄT DARMSTADT

US-Military in WW2:

„We need more armour in the areas that were hit most“



Data came from planes that returned from missions

➡ Survivorship Bias

biases

Source: Wikipedia, McGeddon, CC BY-SA 4.0

What is your data not showing?



Is the dataset showing **boats**? …or **the sea**?

It depends on the negative examples!

→ Negative Set Bias (in section 3.2)

biases

Image sources: ImageNet dataset

Other types of biases

- **Selection bias** (e.g. camera angle)

ImageNet cars

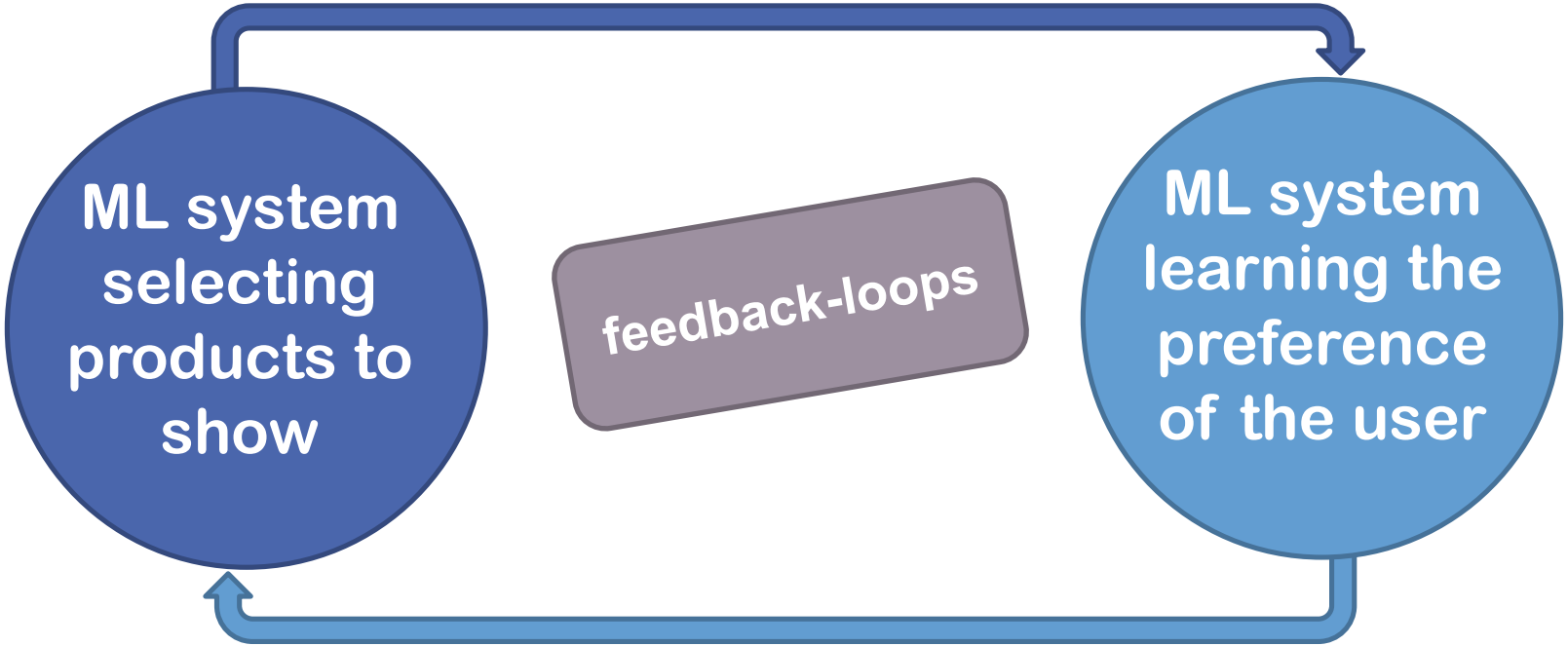
- **Bias in reality**

  e.g. searching for „3 black teenagers“

  vs. „3 white teenagers“

biases

Feedback Loops



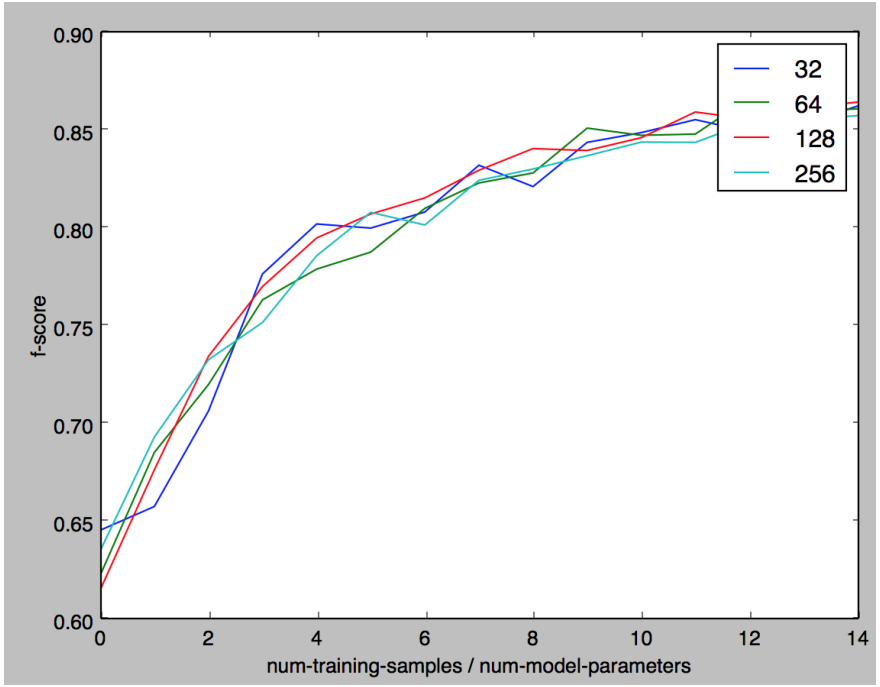
Source: Hidden Technical Debt in Machine Learning Systems

How much data do you need?

Depends on…

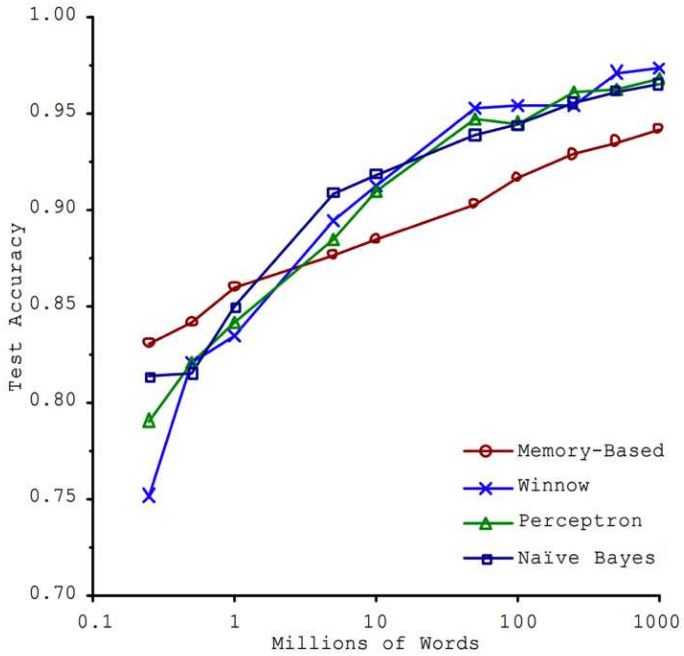
- Model-type
- Number of parameters
- Number of features

About 10 times more samples than parameters is a good place to start.



Source: Malay Haldar: How much training data do you need? medium.com

More data is almost always better



More data
~ logarithmic increase of performence

Source: Scaling to Very Very Large Corpora for Natural Language Disambiguation

Value of a Dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

…but only if the bias matches the test data!

	SUN09 market	LabelMe market	PASCAL market	ImageNet market	Caltech101 market
1 SUN09 is worth	1 SUN09	0.91 LabelMe	0.72 pascal	0.41 ImageNet	0 Caltech
1 LabelMe is worth	0.41 SUN09	1 LabelMe	0.26 pascal	0.31 ImageNet	0 Caltech
1 pascal is worth	0.29 SUN09	0.50 LabelMe	1 pascal	0.88 ImageNet	0 Caltech
1 ImageNet is worth	0.17 SUN09	0.24 LabelMe	0.40 pascal	1 ImageNet	0 Caltech
1 Caltech101 is worth	0.18 SUN09	0.23 LabelMe	0 pascal	0.28 ImageNet	1 Caltech
Basket of Currencies	0.41 SUN09	0.58 LabelMe	0.48 pascal	0.58 ImageNet	0.20 Caltech

Measuring Dataset's Value (in section 4)

How do you get good data?

- Contains: **context, action & outcome**
- Avoid **feedback loops**
- Collected in **interactions that users care about**

- Best: **implicit actions** on **real usage** (avoiding interrater reliability issues)
- Test on other data as well! (cross dataset generalization) [5]

As in: Building Intelligent Systems: A Guide to Machine Learning Engineering by Hulten et al.

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Data Collection**
   What are potential problems?
   How do you get good data?

2. **Data Cleaning**
   What are potential problems?
   How do you get clean data?

3. **Take-Aways**
   Tools for your use
   Summary

Possible Issues

biases
e.g. men vs. woman

feedback-loops
„more men have been hired in the past, so they are better at SE“

interrater reliability
e.g. people using 5 stars only for the most perfect applications

out-of-date
„people knowing flash are great web developers“

noise
rater accidentally clicked wrong button

wrong data
e.g. people who are 18 yo and have 25 years of working experience“

...and many more!

Possible Issues

TECHNISCHE UNIVERSITÄT DARMSTADT

special standard-values
e.g. -9999

unmatched dependencies
e.g. zip codes & cities

outliers

duplicates

noise
rater accidentally clicked wrong button

wrong data
e.g. people who are 18 yo and have 25 years of working experience“

...and many more!

How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Let us introduce you to 3 examples:

Data Linting
…for simple errors.

Unit Tests for Data
…for complex constraints.

HoloClean
…for automatic cleaning.

„level of aggression“

How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Linting

detects potential (simple) errors:

1. miscodings of data **2. outliers** **3. packaging errors**

e.g.:

special string types
e.g. „2020-06-16“

unnormalized features
e.g. wide ranges of values

duplicates
e.g. duplicate rows

Data Linting

Unit Tests for Data

HoloClean

→ The Data Linter
(paper by Hynes et al.)

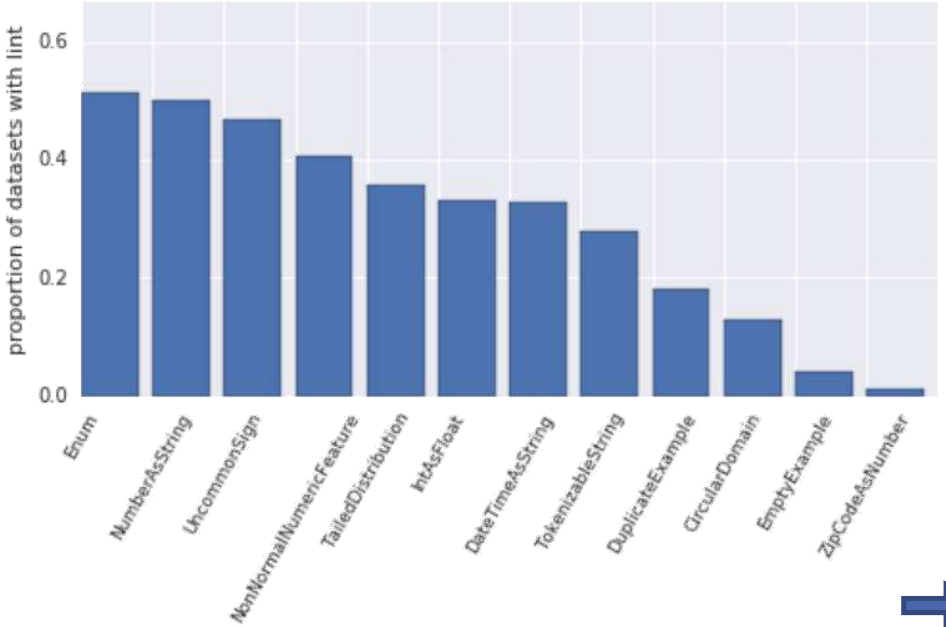
How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Linting

Frequency of
Data Lints

Across
600 Kaggle Data Sets



Data Linting

Unit Tests for Data

HoloClean

The Data Linter
(paper by Hynes et al.)

How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Unit Tests for Data

tests potential constraints in **incrementally growing** datasets:

1. completeness

e.g.:

IsComplete
Check for missing values in column

2. consistency

isUnique
Check for duplicates in this column

3. statistics

hasStandardDeviation
Check if standard deviation matches expected value

Data Linting

Unit Tests for Data

HoloClean

→ Automating Large-Scale Data Quality Verification
(paper by Schelter et al.)

How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Unit Tests for Data

```
5
6  checks += Check(Level.Error)
7   .isComplete("customerId", "title",
8    "impressionStart", "impressionEnd",
9    "deviceType", "priority")
10  .isUnique("customerId", "countryResidence",
11   "deviceType", "title")
12  .hasCountDistinct("title", _ <= numTitles)
13  .hasHistogramValues("deviceType",
14    _.ratio("phone") <= maxExpectedPhoneRatio)
15
```

+ Constraint Suggestion **+ Anomaly Detection**

Automating Large-Scale Data Quality Verification
(paper by Schelter et al.)

Data Linting

Unit Tests for Data

HoloClean

How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

HoloClean

automatically cleans online data, combining:

1. integrety constraints
2. statistics
3. External data

e.g. recognizing zip ⇔ city misscodings

e.g. repairing outliers like typos

e.g. data for zip ⇔ city mappings

Data Linting

Unit Tests for Data

HoloClean

HoloClean: Holistic Data Repairs with Probabilistic Inference
(paper by Rekatsinas et al.)

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Data Collection**
   What are potential problems?
   How do you get good data?

2. **Data Cleaning**
   What are potential problems?
   How do you get clean data?

3. **Take-Aways**
   Tools for your use
   Summary

Take-Aways | Your toolbox

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Visualize your Data**
   e.g. by using Facets

2. **Find mistakes in your Data**
   e.g. by using a Data Linter

3. **Automatically clean your Data**
   e.g. by using HoloClean

Take-Aways | Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Watch out for biases!**
2. **If possible, test on other data!**
3. **Get enough data!**
4. **Good data should contain context, action & outcome!**
5. **Don‘t just use your data, look for fixable errors!**
6. **In online learning systems: test your data continually!**

References

[1] Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). **Automating large-scale data quality verification** *Proceedings of the VLDB Endowment, 11*(12), 1781–1794.

[2] Rekatsinas, T., Chu, X., Ilyas, I., & Ré, C. (2017). **HoloClean : Holistic Data Repairs with Probabilistic Inference**(i).

[3] Hynes, N., Sculley, D., Brain, G., Google Brain, M., & Terry, M. (2017). **The Data Linter: Lightweight, Automated Sanity Checking for ML Data Sets** *NIPS MLSys Workshop*(Nips).

[4] Gao, J., Xie, C., & Tao, C. (2016). **Big data validation and quality assurance - Issuses, challenges, and needs** *Proceedings - 2016 IEEE Symposium on Service-Oriented System Engineering, SOSE 2016*, 433–441.

[5] Torralba, A., & Efros, A. (2011). **Unbiased look at dataset bias** *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1521–1528.

[6] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., & Dennison, D. (2015). **Hidden Technical Debt in Machine Learning Systems**, 1–9.

[7] Hulten, G. (2018). **Building Intelligent Systems: A Guide to Machine Learning Engineering**. Apress.

Icons: Font Awesome by Dave Gandy - http://fontawesome.io

4. Questions & Discussion

5. Additional Material

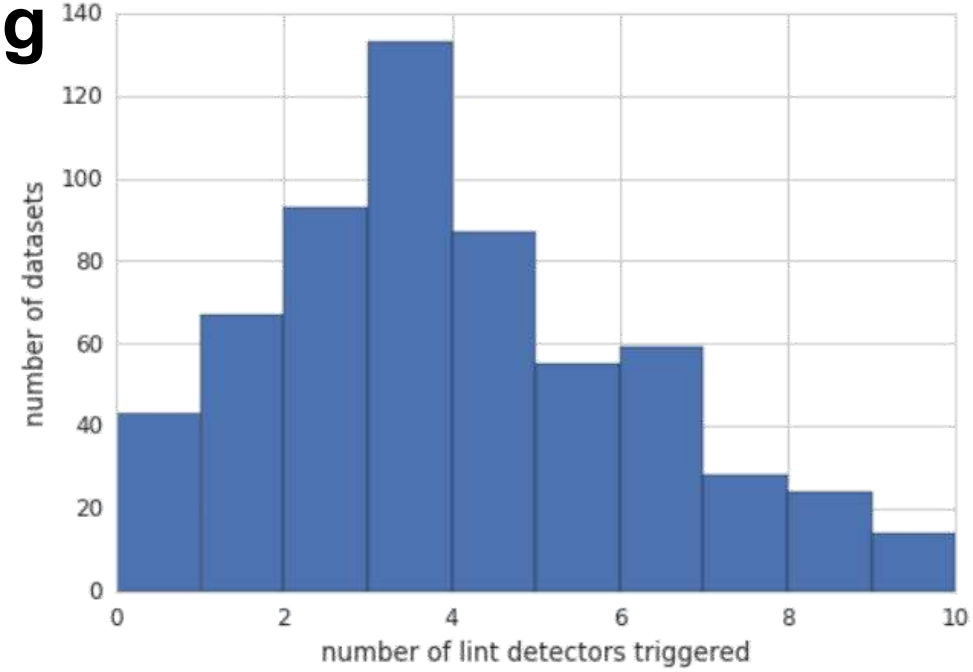
How do you get clean data?

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Linting

Historgram of
# of Lints per Data Set

Across
600 Kaggle Data Sets



Data Linting

Unit Tests for Data

HoloClean

The Data Linter
(paper by Hynes et al.)