Please scan this code to register your presence on Ilias
for contact-tracing purposes. Your Ilias identity will be used.
– Only do this if you are *physically present*! –

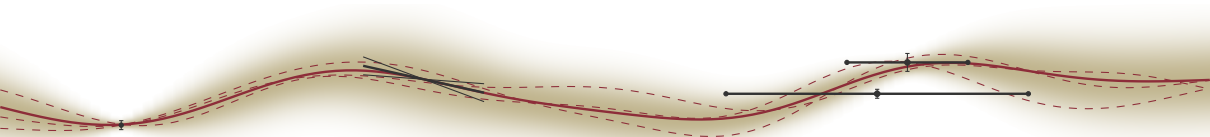
Data Literacy
Lecture 11
Working with Data

Philipp Hennig

17 January 2022

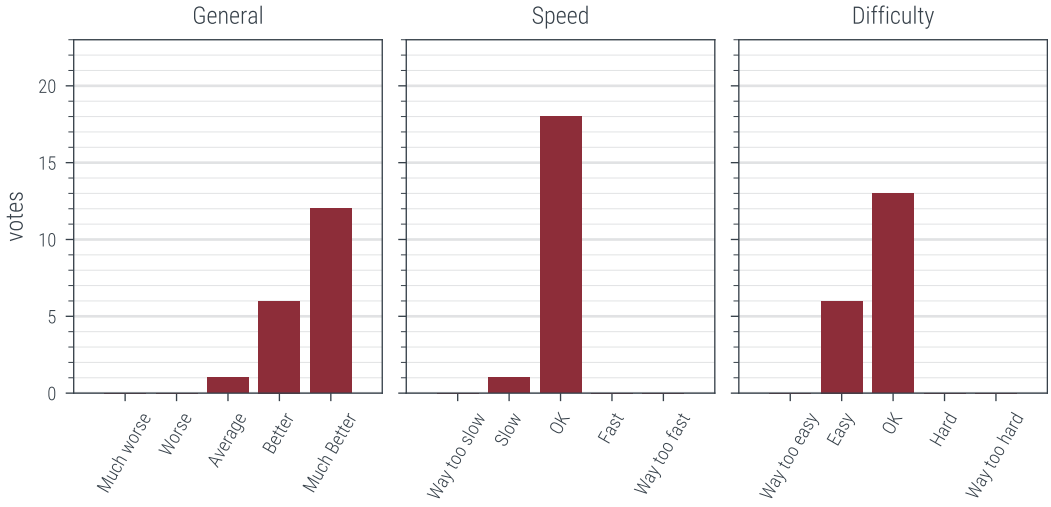
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science



Feedback

quantitative



Detailed Feedback

your answers

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you *didn't* like

- A little bit of overoptimism regarding people being able to control the increase in usage
- Would have loved even more information
- The power consumption example

Things you liked

- The very selection of the topic, it's so important and personally close
- Eye opening moment at the end of the lecture „giving job advice"
- Data storytelling in action. Back-of-the-envelope calculations. The very smooth advertising for the coming numerics course.

Things you didn't understand

- It was very clear today, I had no problems understanding.

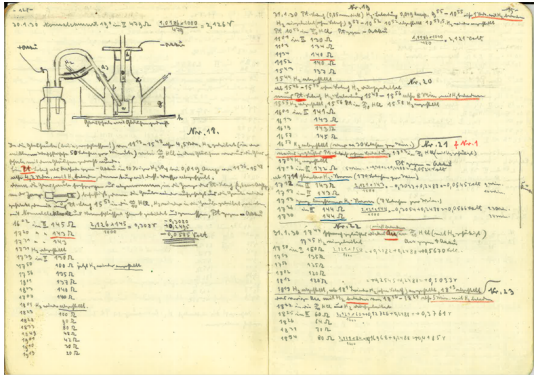
Working with data

- Properly *documenting* data collection and analysis, to make results *reproducible*
- Pitfalls of data use in industrial and scientific applications

image: MPG

- Document all your experiments, even the failed ones!
- Even document the stuff you don't think you need to document!
- Numbers should ideally be stored directly in digital format, but with traceable edits
- Meta-document your files and folders

Data and Code are not that different!



Much of this also applies to Corporate Research!

SPIEGEL ONLINE DER SPIEGEL SPIEGEL TV Anmelden

Menü | Politik Meinung Wirtschaft Panorama Sport Kultur Netzwelt Wissenschaft mehr ▾

WISSENSCHAFT Schlagzeilen | Wetter | DAX 12.548,62 | TV-Programm | Abo

Nachrichten › Wissenschaft › Technik › Forschungsskandale › Leibniz-Preis: Britta Nestler nachträglich ausgezeichnet

Kein wissenschaftliches Fehlverhalten

Britta Nestler erhält Leibniz-Preis nachträglich

Materialforscherin Britta Nestler bekommt den renommierten Leibniz-Preis doch noch verliehen. Die Auszeichnung war verschoben worden, nachdem es anonyme Hinweise im Zusammenhang mit ihrer Arbeit gegeben hatte.

Über die Vorwürfe, die der namenlose Tippgeber am Freitag vor der Preisverleihung an die DFG schickte, will Nestler nicht im Einzelnen reden, nur dass sie bis 1999 zurückreichen. Die jüngsten beziehen sich auf 2013. Klar ist: Der Verleumder verfolgt ihre Arbeit und die ihrer Forschergruppe seit langem. Der Stoß Papier, den er an die DFG schickte, umfasste offenbar hunderte Seiten inklusive zahlreicher Anlagen.

https://www.jmwiarda.de/2017/08/01/man-fragt-sich-wieso-wird-dem-überhaupt-nachgegangen/

It **will** happen to you!

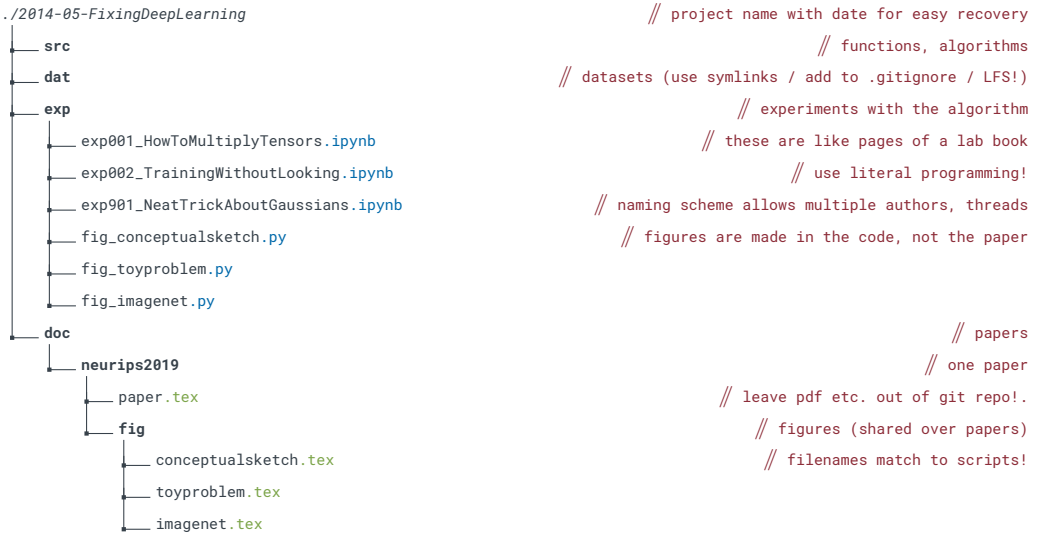
document your work!

From: Winfried Denk <Winfried.Denk@mpimf-heidelberg.mpg.de>
Subject: daten fuer fig 5 von Hennig&Denk 2007
Date: 17 August 2014 at 11:52:38 CEST
To: phennig@tuebingen.mpg.de

Hallo Philipp,
has Du die Daten fuer Fig 5 von Hennig&Denk 2007 greifbar?
Gruesse
w.

A template for research projects

an electronic lab book for Computer Scientists



Literal Programming

Do it, but don't overdo it

```
' ' '
exp001_HowToMultiplyTensors.py

The standard order for collapsing indices in tensor multiplication in deep-learning packages
(i.e. batch-first) is inconvenient for inference during optimization. Quantities like
batch-variances can only be extracted when the order is exchanged.

This experiment shows that the two formulations are in fact equivalent.

Philipp Hennig, April 2014
' ' '
```

Research code is not the same as production code.
Acknowledge that your code has a short life-span, and will evolve.

Examples

there is no unique correct solution

Some examples:

- precondGP (https://github.com/JonathanWenger/preconditioning-gps) (private), a
- capos (`https://github.com/nathanaelbosch/capos`): A basic "minimum viable product" documenting a single paper, later superseded by a proper code package (`https://nathanaelbosch.github.io/ProbNumDiffEq.jl`)
- Cockpit Separation of Product (`https://github.com/f-dangel/cockpit`) and publication with experiments (`https://github.com/fsschneider/cockpit-experiments`)
- Laplace (`https://aleximmer.github.io/Laplace/`) An "end-user" package, with lots of examples to complement the paper
- Probnum (`http://probnum.org`) full open-source "product" with community processes

Summary: Reproducible Data Science

- document *all* your experiments (including failed ones)
- record as much meta-data as you can (it's never enough)
- use open and uncompressed formats (future proof, easy to inspect) if it makes sense
- connect data, code and publication. But you don't have to release all of them publicly. Nest the releasable code (`src, dat, fig`) within your private meta-information (communications, notes)
- create figures & tables procedurally to prevent stale figures (making figures: next lecture)
- expect having to return to the project *years* later

Deploying Machine Learning in Industrial Practice

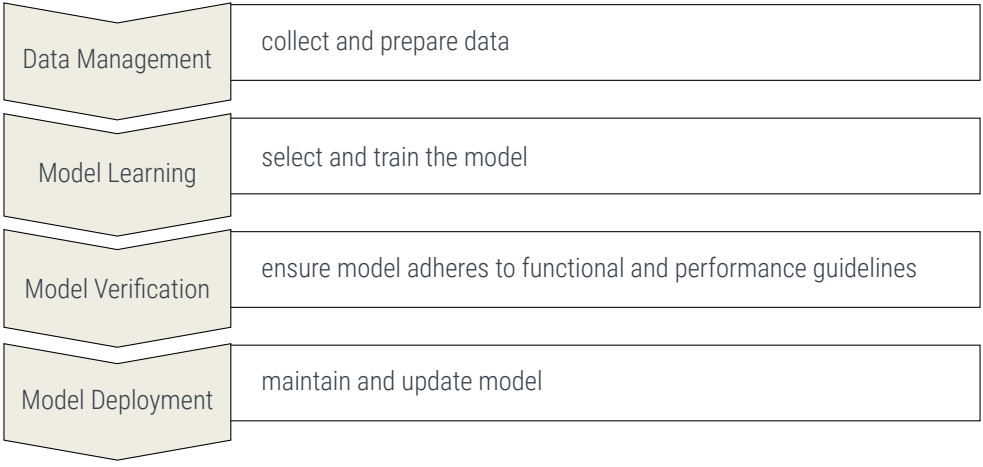
Common challenges and processes

- In business settings, ML and data science has to be part of a complex production process
- Data brings new challenges not well addressed (or opposed to) established software development concepts
- Many of these challenges have no widely accepted solution yet

Warning: Business Fluff! Take with a grain of salt. But do not ignore, either!

Data Engineering

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)



These steps are functionally analogous to academic research, albeit with different performance goals (speed & profit vs. correctness & openness)

- Data collection
  - Data is often not available centrally, but distributed across microservices
  - It may be stored in terrible formats (log files)
  - Worse, it may not be stored at all, and new API calls must be created or hacked

- Data collection
  - Data is often not available centrally, but distributed across microservices
  - It may be stored in terrible formats (log files)
  - Worse, it may not be stored at all, and new API calls must be created or hacked
- Data preprocessing
  - Data may be stored in differing schemata (e.g. addresses)
  - …and differing formats (e.g. xlsx files, csv, pdf-scans, SQL db)

- Data collection
  - Data is often not available centrally, but distributed across microservices
  - It may be stored in terrible formats (log files)
  - Worse, it may not be stored at all, and new API calls must be created or hacked
- Data preprocessing
  - Data may be stored in differing schemata (e.g. addresses)
  - …and differing formats (e.g. xlsx files, csv, pdf-scans, SQL db)
- Data augmentation
  - labels are generally not available
  - labeling experts are difficult to access (e.g. medical doctors)
  - available training data may lack variance / have bias (e.g. simulation, test track for autonomous cars)

- Data collection
  - Data is often not available centrally, but distributed across microservices
  - It may be stored in terrible formats (log files)
  - Worse, it may not be stored at all, and new API calls must be created or hacked
- Data preprocessing
  - Data may be stored in differing schemata (e.g. addresses)
  - …and differing formats (e.g. xlsx files, csv, pdf-scans, SQL db)
- Data augmentation
  - labels are generally not available
  - labeling experts are difficult to access (e.g. medical doctors)
  - available training data may lack variance / have bias (e.g. simulation, test track for autonomous cars)
- Data analysis
  - visualization is an art, not a science
  - visualization tools are not widely deployed, and need expert knowledge (remember *t*-SNE)

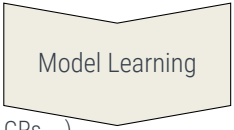
"Data scientists think about data issues as the main reason to doubt the quality of their overall work"

Kim, Zimmermann, DeLine, and Begel. *Data scientists in software teams: State of the art and challenges*. IEEE T Software Engineering, 2017.

Model Learning

select and train the model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

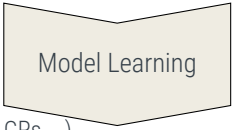


- **Model selection**
  - the main goal of academic research
  - but *simple models often work well!* (PCA, decision trees, logistic regression, SVMs, GPs, …)
  - this is not a question of *potential* performance. Simple models are easier to set up, and tend to be easier to explain to users, management, experts

Model Learning

select and train the model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)



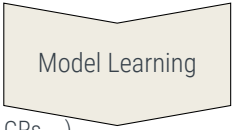
- **Model selection**
  - the main goal of academic research
  - but *simple models often work well!* (PCA, decision trees, logistic regression, SVMs, GPs, …)
  - this is not a question of *potential* performance. Simple models are easier to set up, and tend to be easier to explain to users, management, experts
- **Training**
  - Inefficient trainig is inefficient in both *energy* and *human design time*
  - Training a SotA BERT model costs 50K$ to 1.6M$ [1]

[1] Sharir, Peleg, and Shoham. arXiv:2004.08900, 2020

Model Learning

select and train the model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)



- **Model selection**
  - the main goal of academic research
  - but *simple models often work well!* (PCA, decision trees, logistic regression, SVMs, GPs, …)
  - this is not a question of *potential* performance. Simple models are easier to set up, and tend to be easier to explain to users, management, experts
- **Training**
  - Inefficient trainig is inefficient in both *energy* and *human design time*
  - Training a SotA BERT model costs 50K$ to 1.6M$ [1]
- **Hyperparameter tuning**
  - random search is widely used, but inefficient
  - outer-loop methods like Bayesian optimization are increasingly popuplar. Inner loop methods, held back by rapid churn, hold promise, but are currently still a research goal.

Expert knowledge is required, and has high economic value.

[1] Sharir, Peleg, and Shoham. arXiv:2004.08900, 2020

Model Verification

ensure model adheres to functional and performance guidelines

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

- **Requirement encoding**
  - training loss is not a good KPI
  - business units must set clear, value-drifen measures (customer conversion, production precision, …)
  - indirect KPIs also matter (customer satisfaction, service tickets, uptime, …)

Model Verification

ensure model adheres to functional and performance guidelines

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

- **Requirement encoding**
  - training loss is not a good KPI
  - business units must set clear, value-drifen measures (customer conversion, production precision, …)
  - indirect KPIs also matter (customer satisfaction, service tickets, uptime, …)
- **Formal verification**
  - formal proofs of correctness are rarely possible for ML systems
  - convergence guarantees, PAC bounds are mostly of academic interest, because they tend to be loose
  - but regulatory rules will arrive soon (with unclear outcome)

Model Verification

ensure model adheres to functional and performance guidelines

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

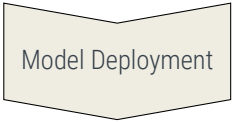
- **Requirement encoding**
  - training loss is not a good KPI
  - business units must set clear, value-drifen measures (customer conversion, production precision, …)
  - indirect KPIs also matter (customer satisfaction, service tickets, uptime, …)
- **Formal verification**
  - formal proofs of correctness are rarely possible for ML systems
  - convergence guarantees, PAC bounds are mostly of academic interest, because they tend to be loose
  - but regulatory rules will arrive soon (with unclear outcome)
- **Test-based verification**
  - testing should ideally be done on real-world data, but this is not always possible
  - if simulations are used, careful and slow pre-flighting is needed
  - the dataset itself also needs to be validated!

Testing and debugging code that uses data poses peculiar challenges (lecture 13)

Model Deployment

maintain and update model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

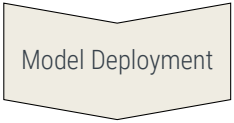


- **Integration**
  - Code / model reuse must be carefully considered
  - Researchers and developers must interact as closely as possible
  - no established best practices for DevOps with AI yet, but emerging (“AIOps”)

Model Deployment

maintain and update model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)

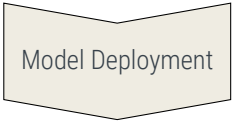


- Integration
  - Code / model reuse must be carefully considered
  - Researchers and developers must interact as closely as possible
  - no established best practices for DevOps with AI yet, but emerging (“AIOps”)
- Monitoring
  - Feedback loops can show up in deployment that were not foreseeable in development
  - outlier inputs must be detected and handled
  - monitoring, similar to visualization, may require custom-built tools to detect hidden failures in data-integrity, model performance

Model Deployment

maintain and update model

after Paleyes, Urma, Lawrence, arXiv 2011.09926 (on Ilias)



- Integration
  - Code / model reuse must be carefully considered
  - Researchers and developers must interact as closely as possible
  - no established best practices for DevOps with AI yet, but emerging (“AIOps”)
- Monitoring
  - Feedback loops can show up in deployment that were not foreseeable in development
  - outlier inputs must be detected and handled
  - monitoring, similar to visualization, may require custom-built tools to detect hidden failures in data-integrity, model performance
- Updating
  - Concept drift / dataset-shift can invalidate or deprecate solution (‘data-rot’).
  - Tracking such changes is not always algorithmically easy (re-train a deep network?)
  - continuous delivery (CD) is difficult in ML, because an ML systems changes along data, model and code (i.e. not “just” code).

ML systems are one of the worst sources of technological debt.

Sculley, et al. *Machine Learning: The High Interest Credit Card of Technical Debt*, 2014

- **Ethics** (cf. lecture 9)
  - bias can arise from training data in complicated ways, even without malicious intentions by anyone along the way
  - fairness and ethical decision making are not universal concepts, but can be defined in mutually exclusive ways
  - the decision itself to use certain kind of data, produce certain products must be ethically questioned

Cross-Cutting Issues

Non-Technical Challenges to deployment of data-driven tools and services

- **Ethics** (cf. previous two lectures)
  - bias can arise from training data in complicated ways, even without malicious intentions by anyone along the way
  - fairness and ethical decision making are not universal concepts, but can be defined in mutually exclusive ways
  - the decision itself to use certain kind of data, produce certain products must be ethically questioned
- **Security** By nature, ML systems are vulnerable to adversarial attacks
  - *Data poisoning:* Attempts to inject training data that corrupt the model's predictions
  - *Model stealing:* probing the model through the API to steal IP like support vectors, weights, etc.
  - *Model inversion:* extracting parts of the training set through model queries
    (cf. `https://arxiv.org/abs/2012.07805`)

Cross-Cutting Issues

Non-Technical Challenges to deployment of data-driven tools and services

- **Ethics** (cf. previous two lectures)
  - bias can arise from training data in complicated ways, even without malicious intentions by anyone along the way
  - fairness and ethical decision making are not universal concepts, but can be defined in mutually exclusive ways
  - the decision itself to use certain kind of data, produce certain products must be ethically questioned
- **Security** By nature, ML systems are vulnerable to adversarial attacks
  - *Data poisoning:* Attempts to inject training data that corrupt the model's predictions
  - *Model stealing:* probing the model through the API to steal IP like support vectors, weights, etc.
  - *Model inversion:* extracting parts of the training set through model queries
    (cf. `https://arxiv.org/abs/2012.07805`)

Cross-Cutting Issues

Non-Technical Challenges to deployment of data-driven tools and services

- **Ethics** (cf. previous two lectures)
  - bias can arise from training data in complicated ways, even without malicious intentions by anyone along the way
  - fairness and ethical decision making are not universal concepts, but can be defined in mutually exclusive ways
  - the decision itself to use certain kind of data, produce certain products must be ethically questioned
- **Security** By nature, ML systems are vulnerable to adversarial attacks
  - *Data poisoning:* Attempts to inject training data that corrupt the model's predictions
  - *Model stealing:* probing the model through the API to steal IP like support vectors, weights, etc.
  - *Model inversion:* extracting parts of the training set through model queries
    (cf. `https://arxiv.org/abs/2012.07805`)
- **Building Trust & Cooperation with Stakeholders**
  - *Data Analyst* role necessary to facilitate communication between management and engineers
  - Engineers must understand user requirements beyond predictive scores
  - Management must understand formal and algorithmic limitations, follow function instead of fad

Some useful high-level approaches

- build *data-centric architectures* instead of micro-services
- start with simple, basic ML models, iterate
- hire well-educated experts ;)
- new role for *Data Analysts* to facilitate communication between engineers and management

Summary: Reproducible Data Science

- document *all* your experiments (including failed ones)
- record as much meta-data as you can (it's never enough)
- use open and uncompressed formats if possible
- connect data and code
- create figures & tables procedurally to prevent stale figures
- expect having to return to the project *years* later

Summary: Working with Data

- getting and preparing data is often the hardest task
- simple ML models should not be dismissed, but they must be understood too!
- AI/ML poses new challenges for operations. Solutions are still to be settled

Please provide feedback:

I'm looking for German-speaking Tutors (HiWis) for "Informatik III" next summer. Interested? Get in touch!

