Digital Stewardship
–
Reproducibility, Data Management & Citation, and Explainable AI

Andreas Rauber

Vienna University of Technology
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/~andi

FACULTY OF !NFORMATICS

Outline

- Reproducibility:
  - Why do we need it? Why is it difficult? How can we achieve it?
- Data Management and Data Citation:
  - Why do we need it? Why is it difficult? How can we achieve it?
- Explainable AI:
  - Why do we need it? Why is it difficult? How can we achieve it?
- Summary

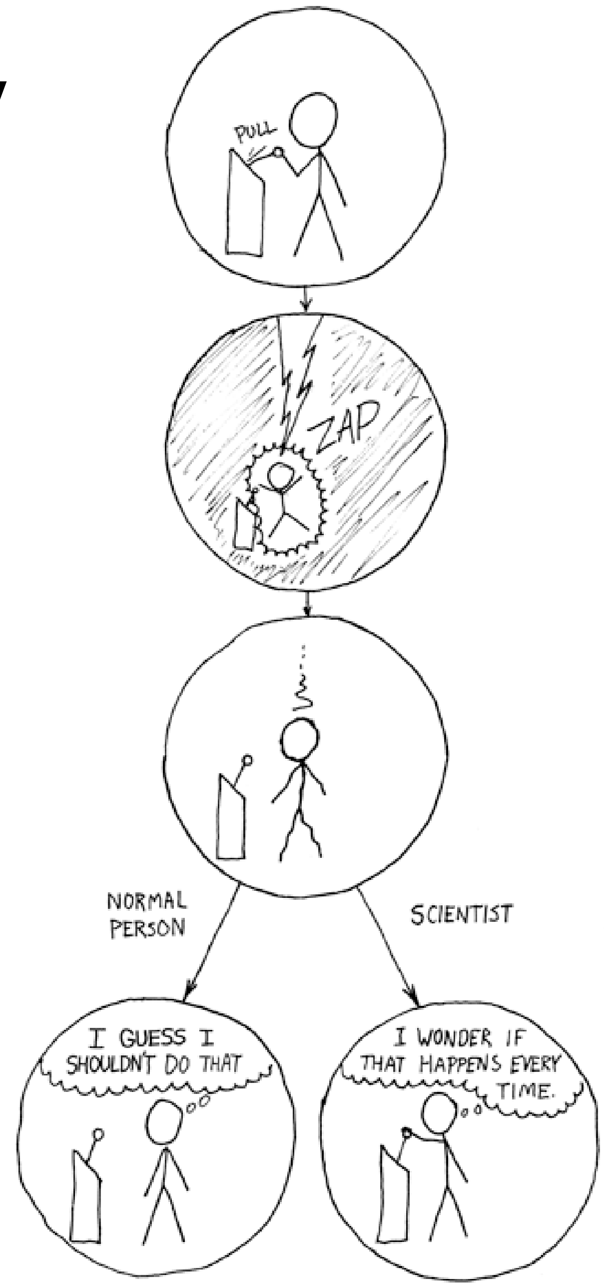
Outline

- Reproducibility
  - What are the challenges in reproducibility?
  - How to address the challenges of complex processes?
- Data Management & Citation
- Explainable AI
- Summary

Reproducibility

- Reproducibility is core to the scientific method
- Focus not on misconduct –
but on complexity and the will to produce good work

- Should be easy
  - Get the code, compile, run, …
  - Why is it difficult?



https://xkcd.com/242/

Reproducibility in “Small Data”

- Carmen M. Reinhart and Kenneth S. Rogoff: *Growth in a Time of Debt.* American Economic Review: Papers and proceedings 100:573-578, May 2010
- Study on relationship btw. debt and economic growth
  - Tipping point at 90% of government debt
  - Published after the Greek crisis
  - Analysis supporting budget cuts
  - Stimulus vs austerity
  - Strong political influence

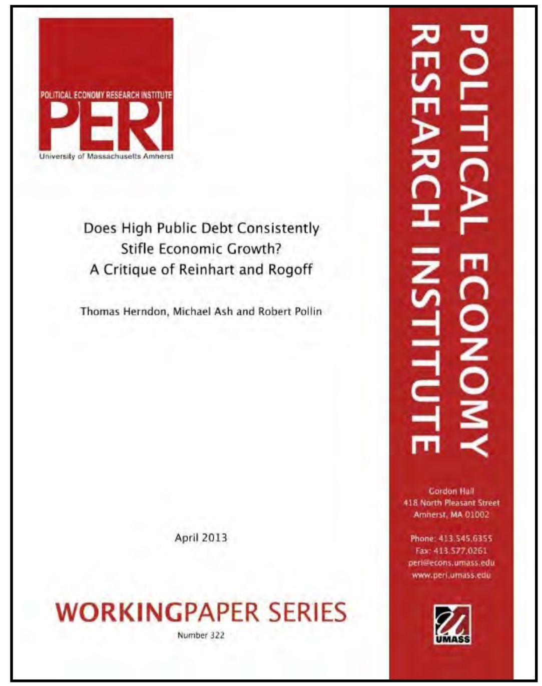

Growth in a Time of Debt

By Carmen M. Reinhart and Kenneth S. Rogoff

https://scholar.harvard.edu/files/rogoff/files/growth_in_time_debt_aer.pdf

Reproducibility in "Small Data"

- Carmen M. Reinhart and Kenneth S. Rogoff: *Growth in a Time of Debt.* American Economic Review: Papers and proceedings 100:573-578, May **2010***.*
- **Others could not reproduce results**: Thomas Herndon, Michael Ash, Robert Pollin: *Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff* UMASS Working Paper Series 322, April **2013**


POLITICAL ECONOMY RESEARCH INSTITUTE

PERI

University of Massachusetts Amherst

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon, Michael Ash and Robert Pollin

April 2013

WORKINGPAPER SERIES

Number 322

POLITICAL ECONOMY RESEARCH INSTITUTE

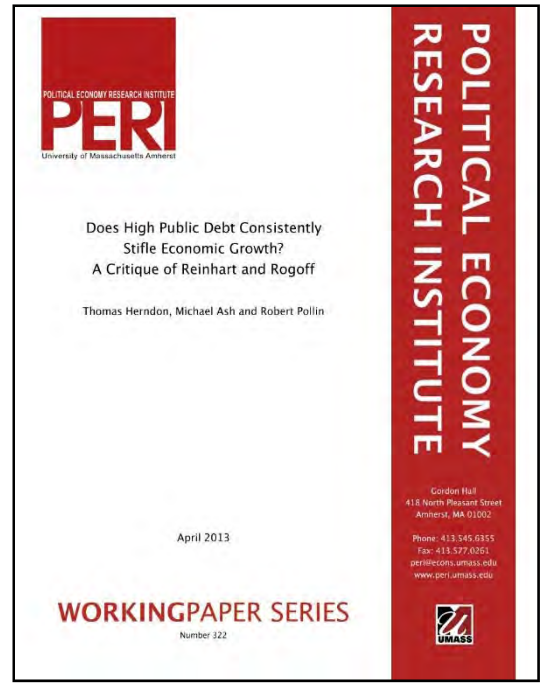
Gordon Hall
418 North Pleasant Street
Amherst, MA 01002

Phone: 413.545.6355
Fax: 413.577.0261
peri@econs.umass.edu
www.peri.umass.edu

https://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf

Reproducibility in "Small Data"

- Carmen M. Reinhart and Kenneth S. Rogoff (2010) *vs.* Thomas Herndon, Michael Ash, Robert Pollin (2013)
- **Original spreadsheet provided**
  - Some data excluded on purpose
  - Questionable statistical procedures
  - **Excel error**
    - Accidentally missed 5 rows of data!
    - Average Annual Growth changed from -0.1 to 2.2 after correction



- Lead to prominent coverage on importance of transparency, reproducibility

https://www.newyorker.com/news/john-cassidy/the-reinhart-and-rogoff-controversy-a-summing-up
https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html

Challenges in Reproducibility

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234

PLOS ONE | Articles | For Authors | About Us | Search | advanced search

OPEN ACCESS · PEER-REVIEWED

68,919 VIEWS · 10 CITATIONS · 124 SAVES

RESEARCH ARTICLE

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Article | About the Authors | Metrics | Comments | Related Content

Show Figures

Abstract
Introduction
Materials and Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

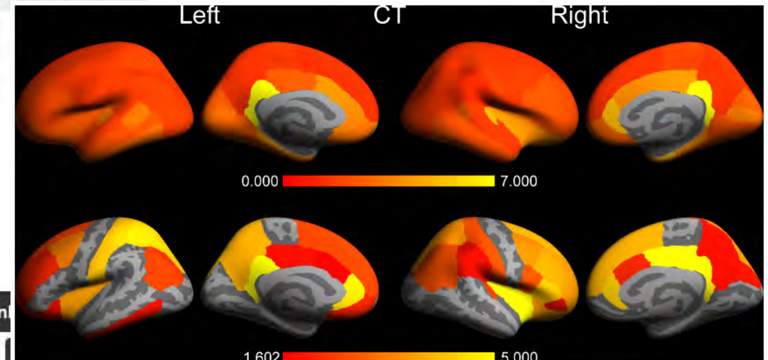
Reader Comments (5)
Figures

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Download | Print

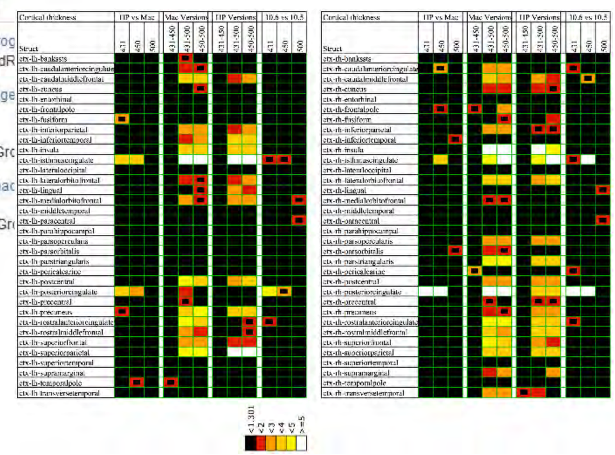


Comments

In praise of prog...
Posted by GedR...

Media Coverage ... Article
Posted by PLoS_ONE_Gro...

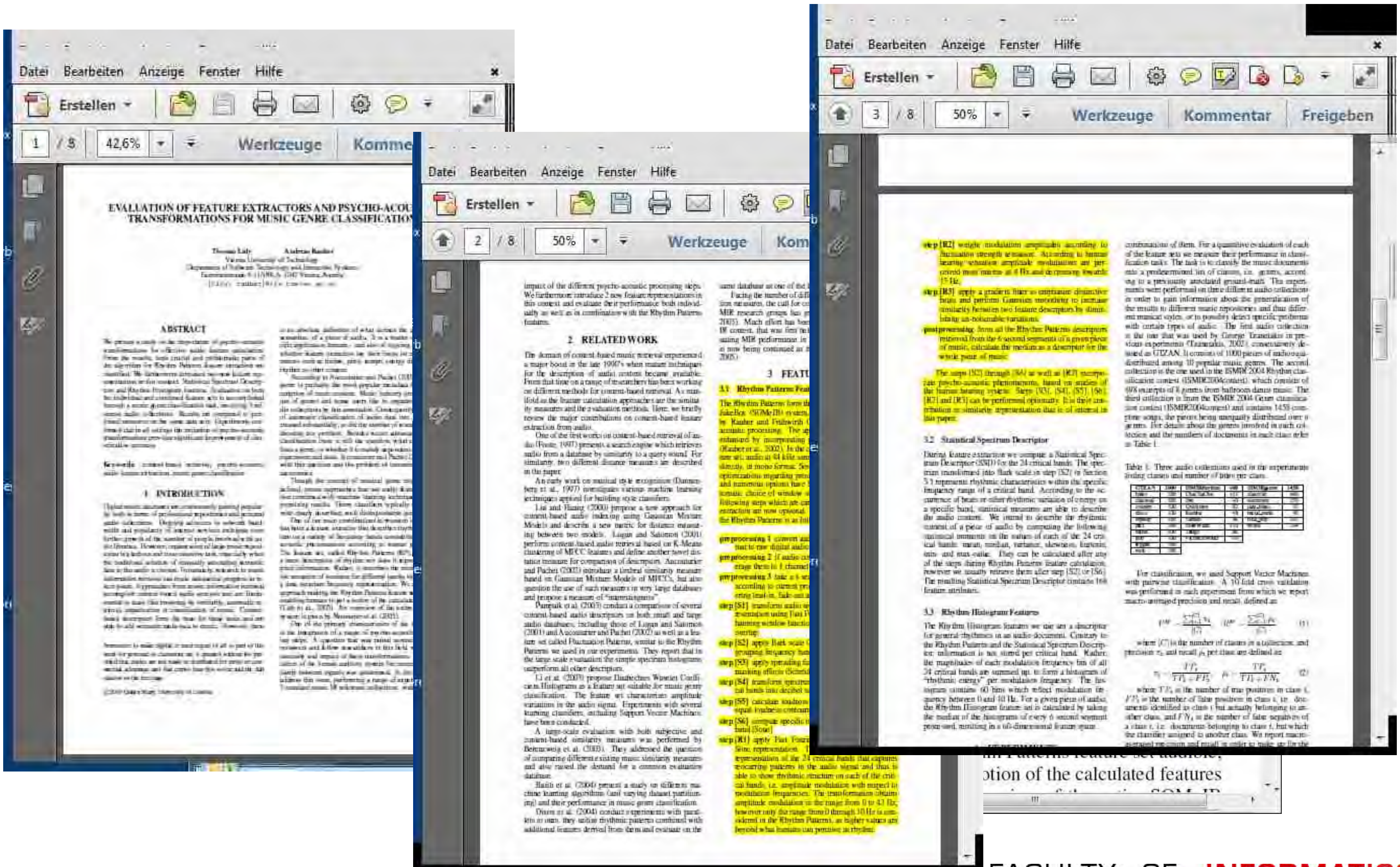
Comments mad... authors
Posted by EdGr...



FACULTY OF !NFORMATICS

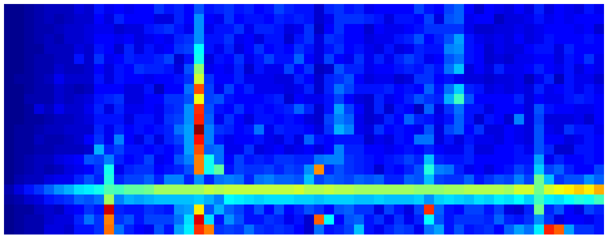
Challenges in Reproducibility

- Excursion: Scientific Processes

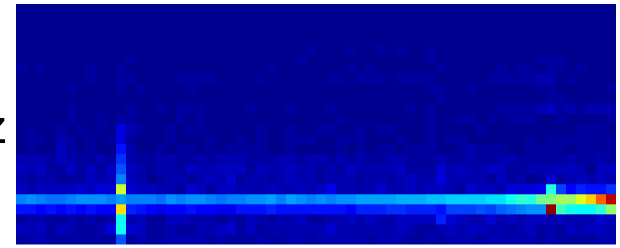


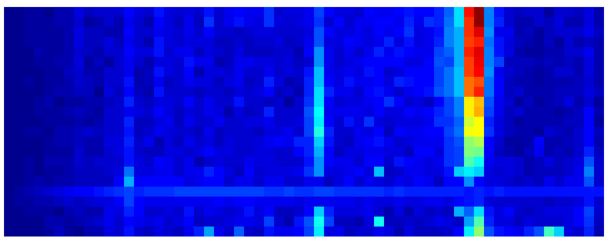
Challenges in Reproducibility

- Excursion: scientific processes

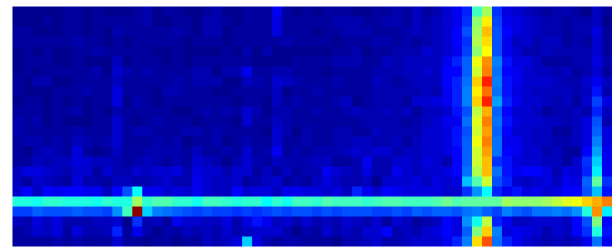


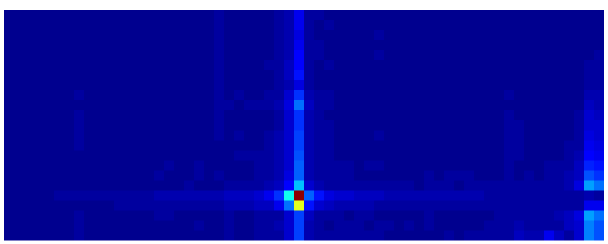
set1_freq440Hz_Am11.0Hz



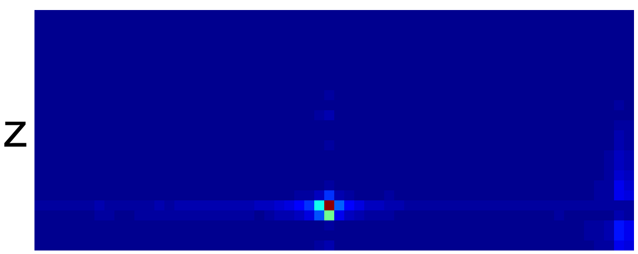


set1_freq440Hz_Am12.0Hz





set1_freq440Hz_Am05.5Hz

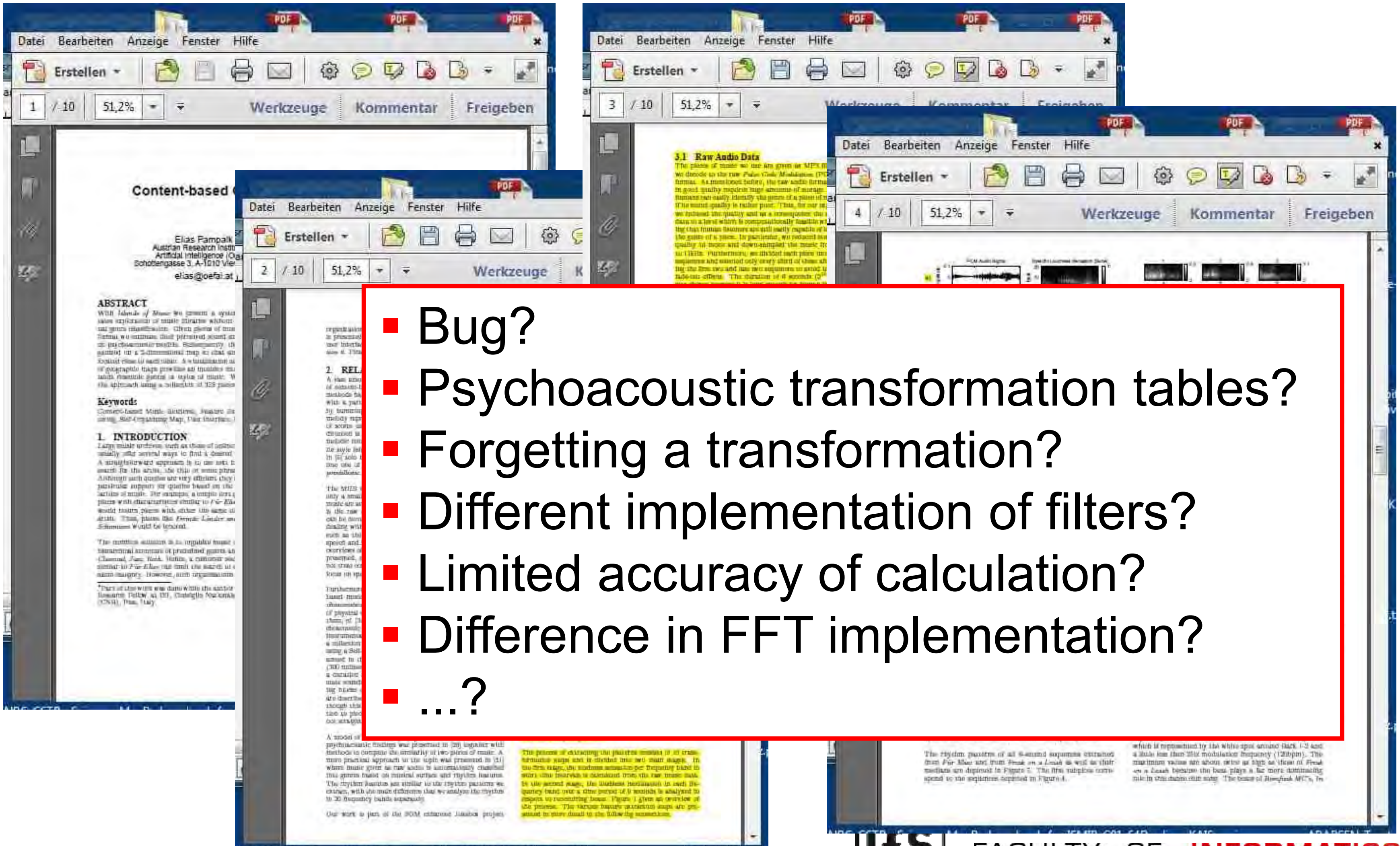


Java

Matlab

Challenges in Reproducibility

- Excursion: Scientific Processes



- Bug?
- Psychoacoustic transformation tables?
- Forgetting a transformation?
- Different implementation of filters?
- Limited accuracy of calculation?
- Difference in FFT implementation?
- ...?

Challenges in Reproducibility

A simpler example

- Image conversion from jpg to tiff using *ImageMagick*

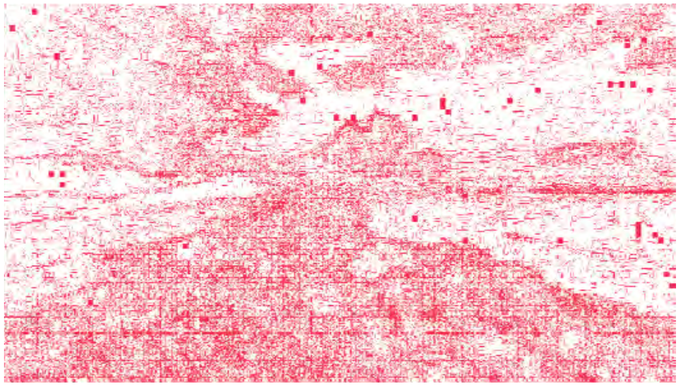
	View Path #1	*View Path #2*
Data formats	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)
Application	ImageMagick 6.8.9-7 Q16 Microsoft Visual C++ 2010	ImageMagick 6.8.9-7
JVM	Java SE 6 Update 45	Java SE 7 Update 10
Operating System	Windows 7 Enterprise SP1	OS X 10.9.4
Hardware	3,3GHz Intel Core i3 8GB 1600MHz DDR3 NVIDIA GT630 2GB	2,3GHz Intel Core i5 4GB 1333MHz DDR3 Intel HD Graphics 3000 384MB

Challenges in Reproducibility

Original jpg

TIFF
Migration on Windows7

TIFF
Migration on OSX

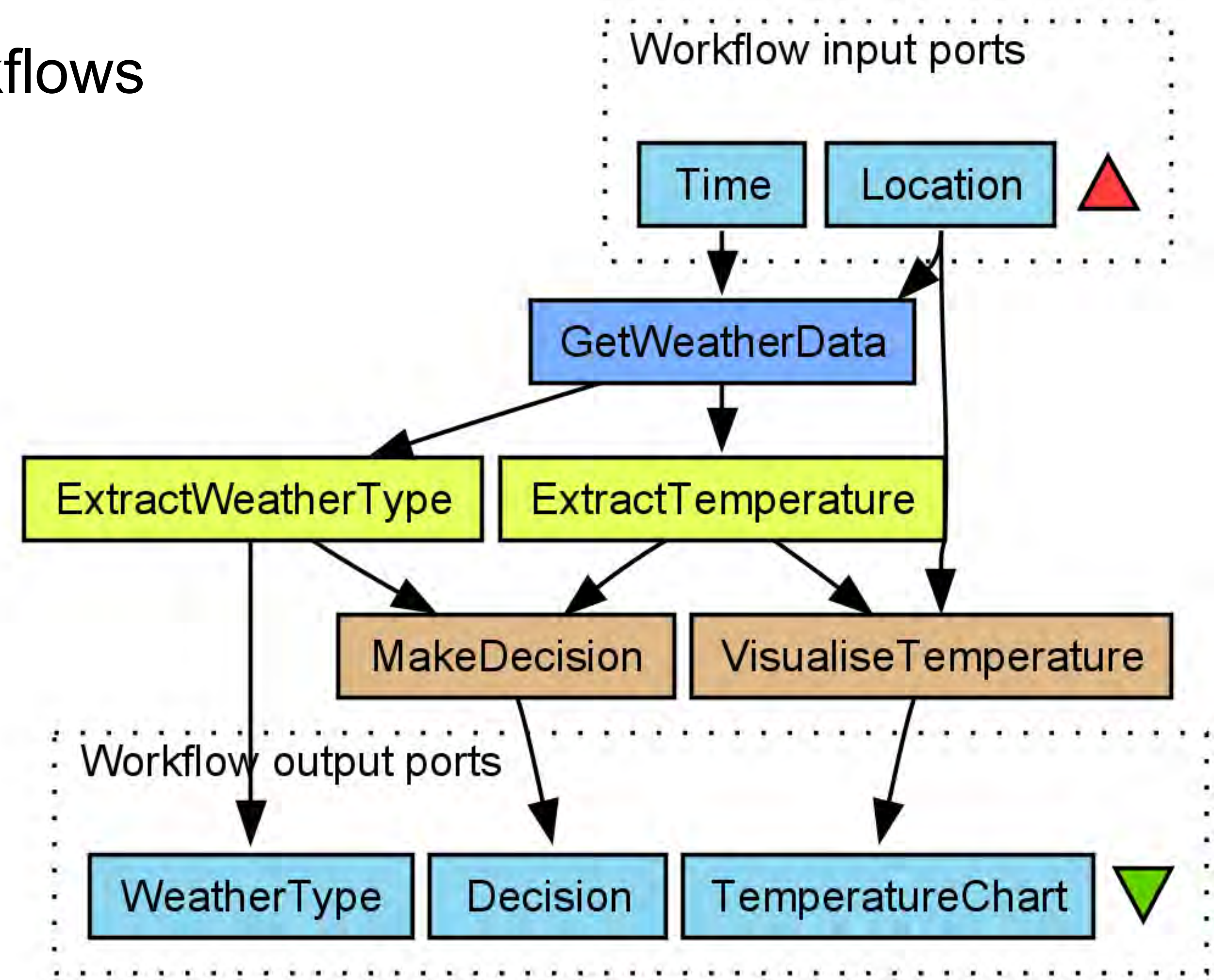


Diff

Challenges in Reproducibility

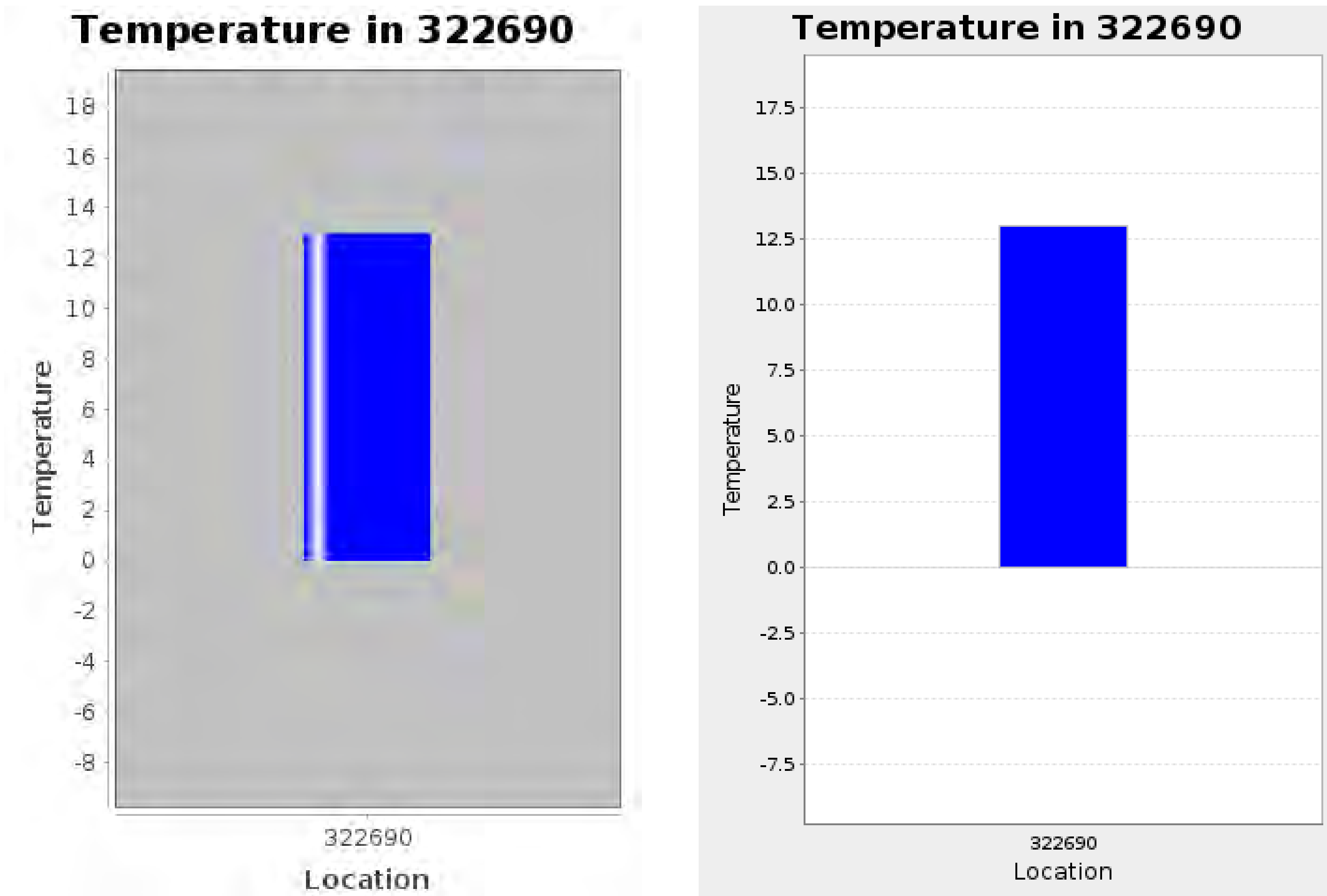
- Workflows

Taverna



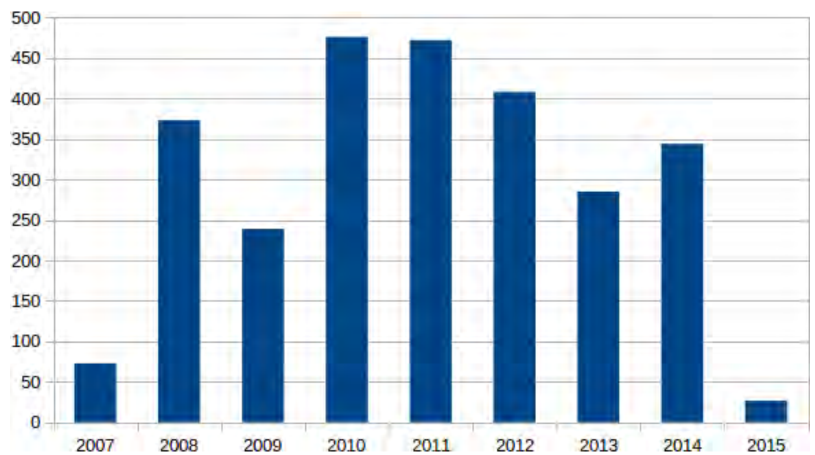
Activiti Kepler

Challenges in Reproducibility



Challenges in Reproducibility

- Large scale quantitative analysis
- Obtain workflows from MyExperiments.org
  - March 2015: almost 2.700 WFs (approx. 300-400/year)
  - Focus on Taverna 2 WFs: 1.443 WFs
  - Published by authors ➔ should be „better quality“
- Try to re-execute the workflows
  - Record data on the reasons for failure along
- Analyse the most common reasons for failures



Workflow Engine	%
Taverna 2	54.7
Taverna 1	20.9
RapidMiner	10.0
Galaxy	2.0
Others	12.4

Challenges in Reproducibility

Re-Execution results

- Majority of workflows fails
- Only 23.6 % are successfully executed
  - No analysis yet on correctness of results…

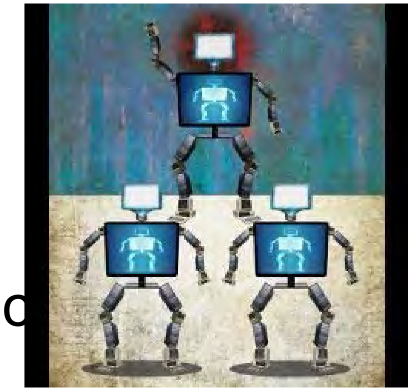
Processor	# WFs
Original Data Set	1443
- Missing input values	526
- Disabled processors (WSDL services)	180
- Not executable in test environment	6
Final Data Set	731

Processor	# WFs	% WFs
Not terminated >48hours	6	0.8
Execution failed	384	52.5
Execution successful	341	46.6

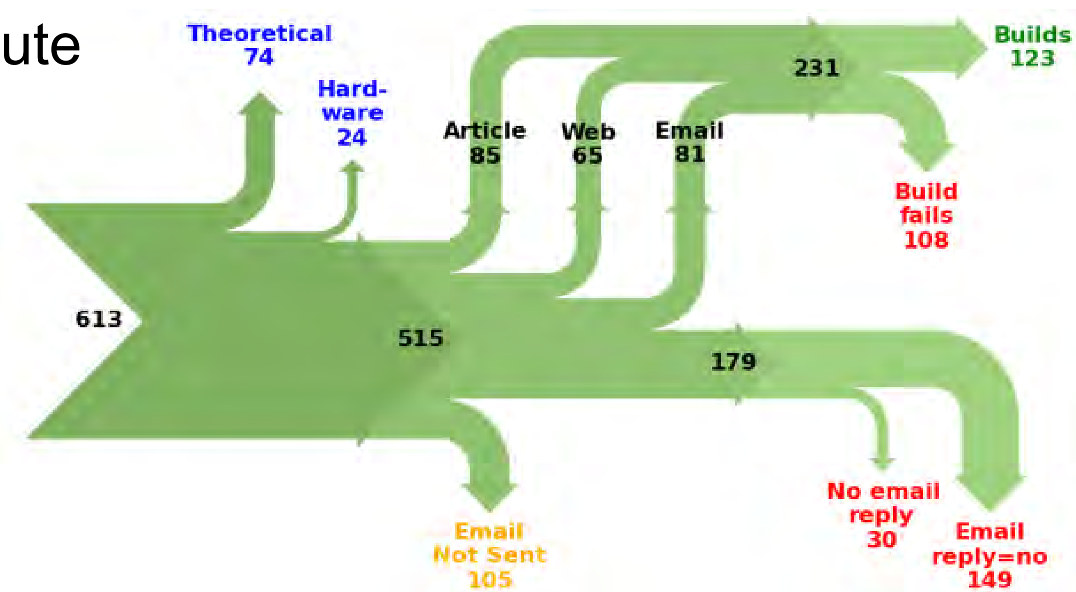
Processor	# WFs
REST unavailable	4
REST unauthenticated	5
Other unauthenticated	40
Missing Resources	14
Tool unavailable	19

Rudolf Mayer, Andreas Rauber, "A Quantitative Study on the Re-executability of Publicly Shared Scientific Workflows", 11th IEEE Intl. Conference on e-Science, 2015.

Challenges in Reproducibility



- 613 papers in 8 ACM conferences
- Process
  - download paper and classify
  - search for a link to code (paper, web, email twic
  - download code
  - build and execute



Christian Collberg and Todd Proebsting. “Repeatability in Computer Systems Research,” CACM 59(3):62-69.2016

Excursion: Ethics & Privacy

- **ACM Statement on Algorithmic Transparency and Accountability**, May 25 2017

  http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

1. **Awareness**: potential bias
2. **Access and redress**: for individuals and groups
3. **Accountability**: responsible for decisions made by algorithms
4. **Explanation**: encouraged to explain procedures, decisions
5. **Data Provenance**: data collection, bias analysis, …
6. **Auditability**: models, data, algorithms recorded
7. **Validation and Testing**: rigorous, routinely, public

Excursion: Ethics & Privacy

How can we address this, support us in proper behavior?

- ´Steps towards solutions:
  - Automated documentation, provenance
  - Data versioning, reproducibility
  - Monitoring data quality, data drift,
  - Defining triggers, roles and responsibilities
- Open questions
  - "Ethical algorithms by design" ?
  - Run-time monitoring for ethical behavior of algorithms?
  - Automated bias-testing for sensitive attributes?
  - Ontology of likely correlated attributes?
  - Can we encode ethical rules/behavior?
  - Role of randomness in human decision making?

Excursion: Ethics & Privacy

Examples

- Self-driving / connected cars
  - Minimizing the impact of accidents
  - Optimizing routing / driving behavior: global / local optimization
- Service provision
  - From elevators to self-driving cars
  - Infrastructure planning
  - Credit scoring
- Social media-based / crowd decision support (Manipulation and social dynamics)
  - Chatbots
  - Recommender Systems, Information retrieval / filters (hate speech)
  - Wikipedia (edit wars) -> input to algorithms -> …

Reproducibility – solved! (?)

- Provide source code, parameters, data, …
- Wrap it up in a container/virtual machine, …

docker LXC vmware …

- Why do we want reproducibility?
- Which levels or reproducibility are there?
- What do we gain by different levels of reproducibility?
- A simple “re-run” is usually not enough
  – otherwise, video would be sufficient….

Types of Reproducibility

- The **PRIMAD Model**[1]: which attributes can we “prime”?
  - Data
    - Parameters
    - Input data
  - Platform
  - Implementation
  - Method
  - Research Objective
  - Actors
- What do we gain by “priming” one or the other?

[1] Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of Data-Oriented Experiments in eScience. Dagstuhl Reports, 6(1):108-159, 2016.
http://drops.dagstuhl.de/opus/volltexte/2016/5817/pdf/dagrep_v006_i001_p108_s16041.pdf

Types of Reproducibility and Gains

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	**x**	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	**x**	-	-	-	-		Applicability across different settings
Port	-	-	**x**	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	**x**	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	**x**	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	**x**		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							**x**	Sufficiency of information, independent verification

Reproducibility Papers

- Aim for reproducibility: for one's own sake – and as Chairs of conference tracks, editor, reviewer, supervisor, …
  - Review of reproducibility of submitted work (material provided)
  - Encouraging reproducibility studies
  - (Messages to stakeholders in Dagstuhl Report)
- Consistency of results, not identity!
- Reproducibility studies and papers
  - Not just re-running code / a virtual machine
  - When is a reproducibility paper worth the effort / worth being published?
    → Issues with peer review and verification…

Challenges in Reproducibility

Peer Review and Verification

- Peer review is an established process
  - Focused on publications mainly
  - Hardly any data quality reviews
  - Even less reproducibility studies
- Reproducing or replicating experiments is not considered original research
  - No recognition
  - No money
  - A lot of work
- Encourage reproducibility studies
- **Needed beyond science!**

Challenges in Reproducibility

Peer Review and Verification

- Encourage reproducibility studies -> **How?**
- Dagstuhl Seminar:
  Reproducibility of Data-Oriented Experiments in e-Science, January 2016, Dagstuhl, Germany
  http://drops.dagstuhl.de/opus/volltexte/2016/5817/pdf/dagrep_v006_i001_p108_s16041.pdf
- Call for action to conference Organizers, Editors, …
- Several conferences include reproducibility tracks

Reproducibility Papers

Transparency, openness, and reproducibility are vital features of science. Scientists embrace these features as disciplinary norms and values, and it follows that they should be integrated into daily research activities. These practices give confidence in the work; help research as a whole to be conducted at a higher standard and be undertaken more efficiently; provide verifiability and falsifiability; and encourage a community of mutual cooperation. They also lead to a valuable form of paper, namely, reports on evaluation and reproduction of prior work. Outcomes that others can build upon and use for their own research, whether a theoretical construct or a reproducible experimental result, form a foundation on which science can progress. Papers that are structured and presented in a manner that facilitates and encourages such post-publication evaluations benefit from increased impact, recognition, and citation rates.

Experience in computing research has demonstrated that a range of straightforward mechanisms can be employed to encourage authors to produce reproducible work. These include: requiring an explicit commitment to an intended level of provision of reproducible materials as a routine part of each paper's structure; requiring a detailed methods section; separating the refereeing of the paper's scientific contribution and its technical process; and explicitly encouraging the creation and reuse of open resources (data, or code, or both).

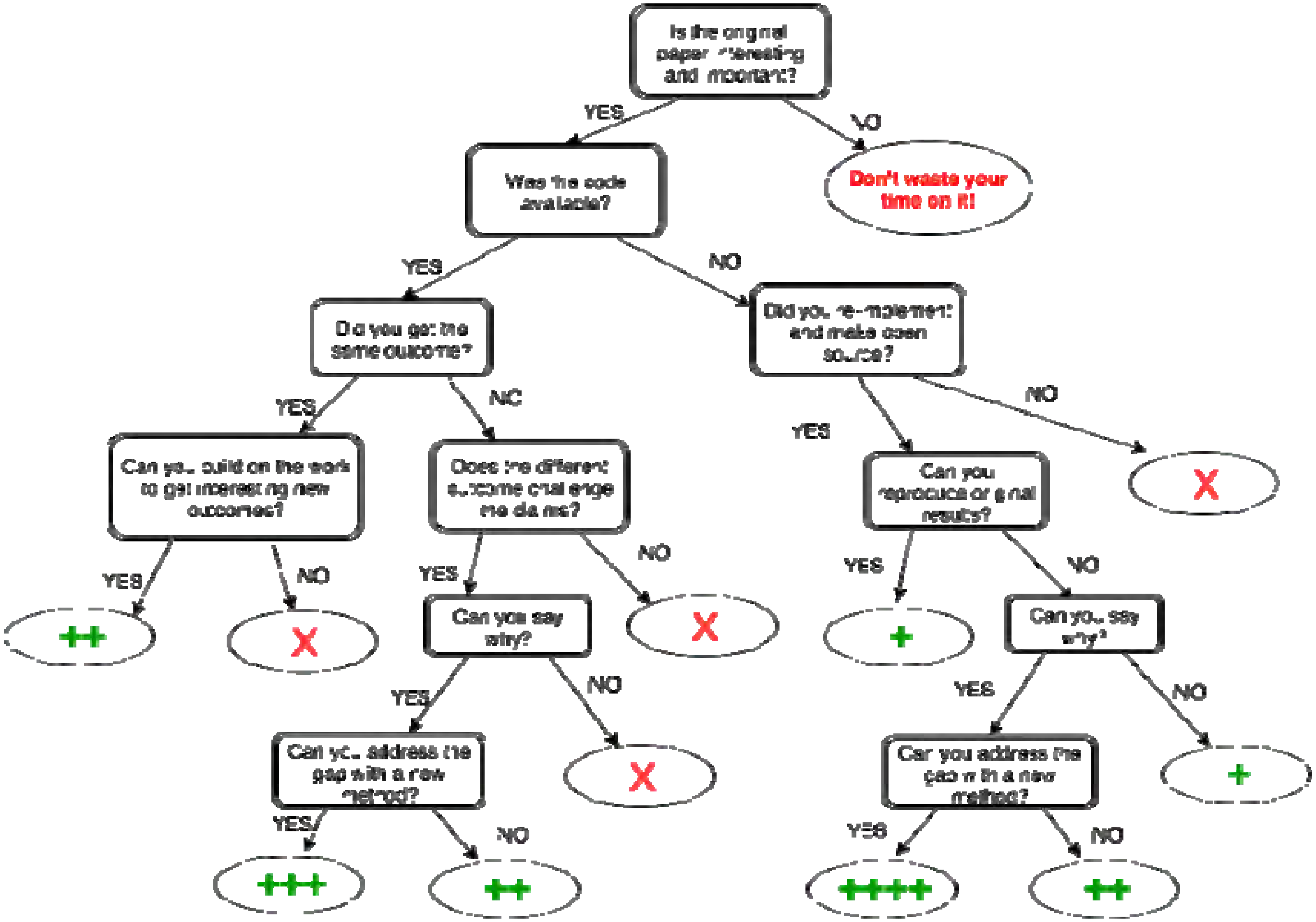
Reproducibility Papers

The [insert name of journal/conference] encourages authors to provide their work in a way that enables reproduction of their outcomes. Just as you have benefited as an author from the work you cite in your paper, and the tools and resources that others have provided, your efforts will also assist the community, including your future collaborators, if you provide access to and understanding of the tools and resources that you have used and created while carrying out your project. We therefore [encourage/request that] authors include in their papers detailed explanations of how their work might be reproduced by others in the field, and to accompany their papers with links to data and source code if it is possible to do so. Authors can request separate reviewing of the reproducibility of their work, a category of publication that we explicitly acknowledge.

In order to support these expectations authors are encouraged to include a detailed methods section in their paper that describes the techniques, tools, data resources, and code resources that enables readers to easily reproduce the work. Such a methods section is of greatest benefit to the reader when it is linked to materials stored in a trusted open repository, and these materials include illustrative or complete data, and code that can easily be re-used and understood.

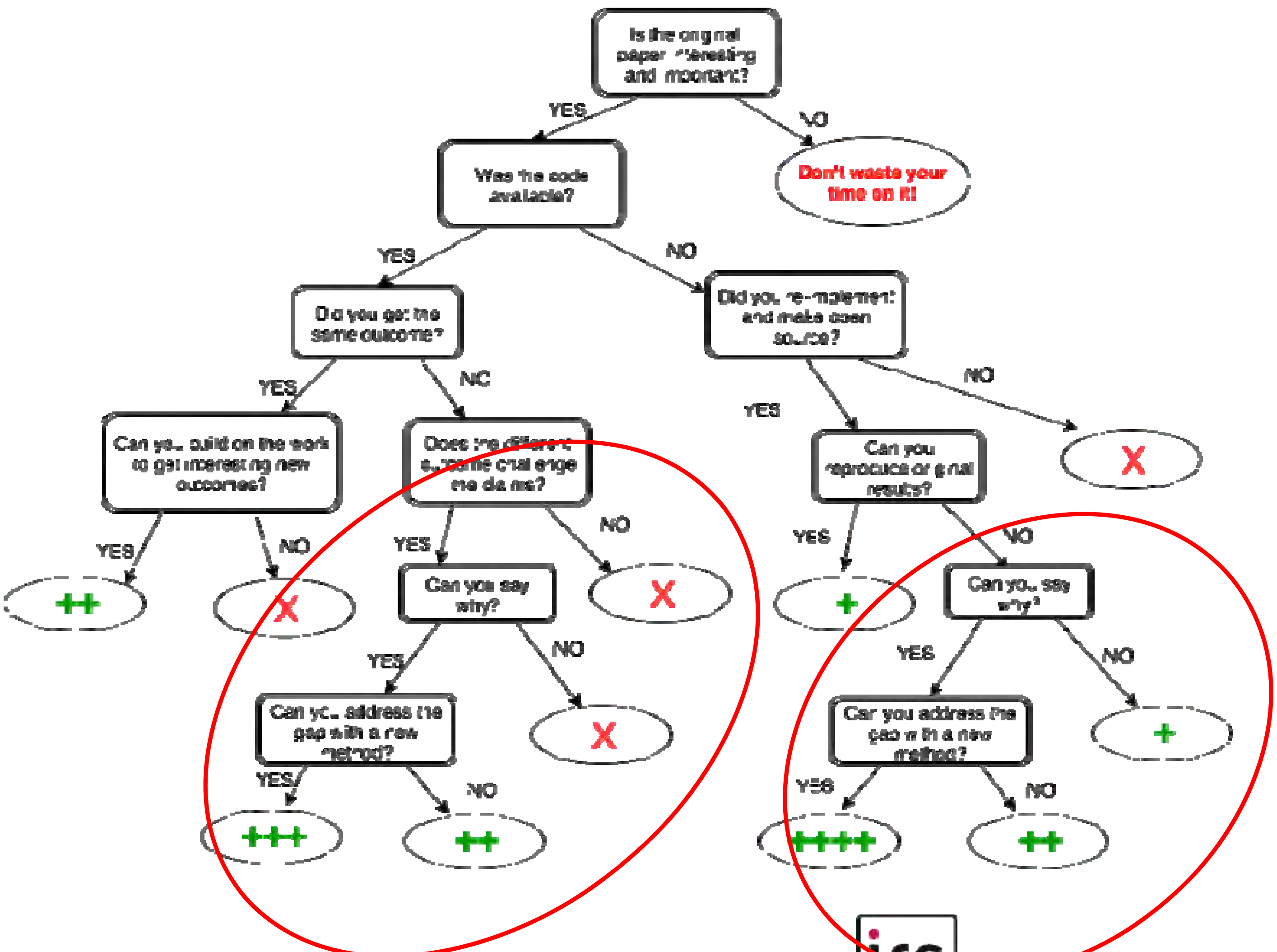
Reproducibility Papers

- When is a Reproducibility paper worth being published?

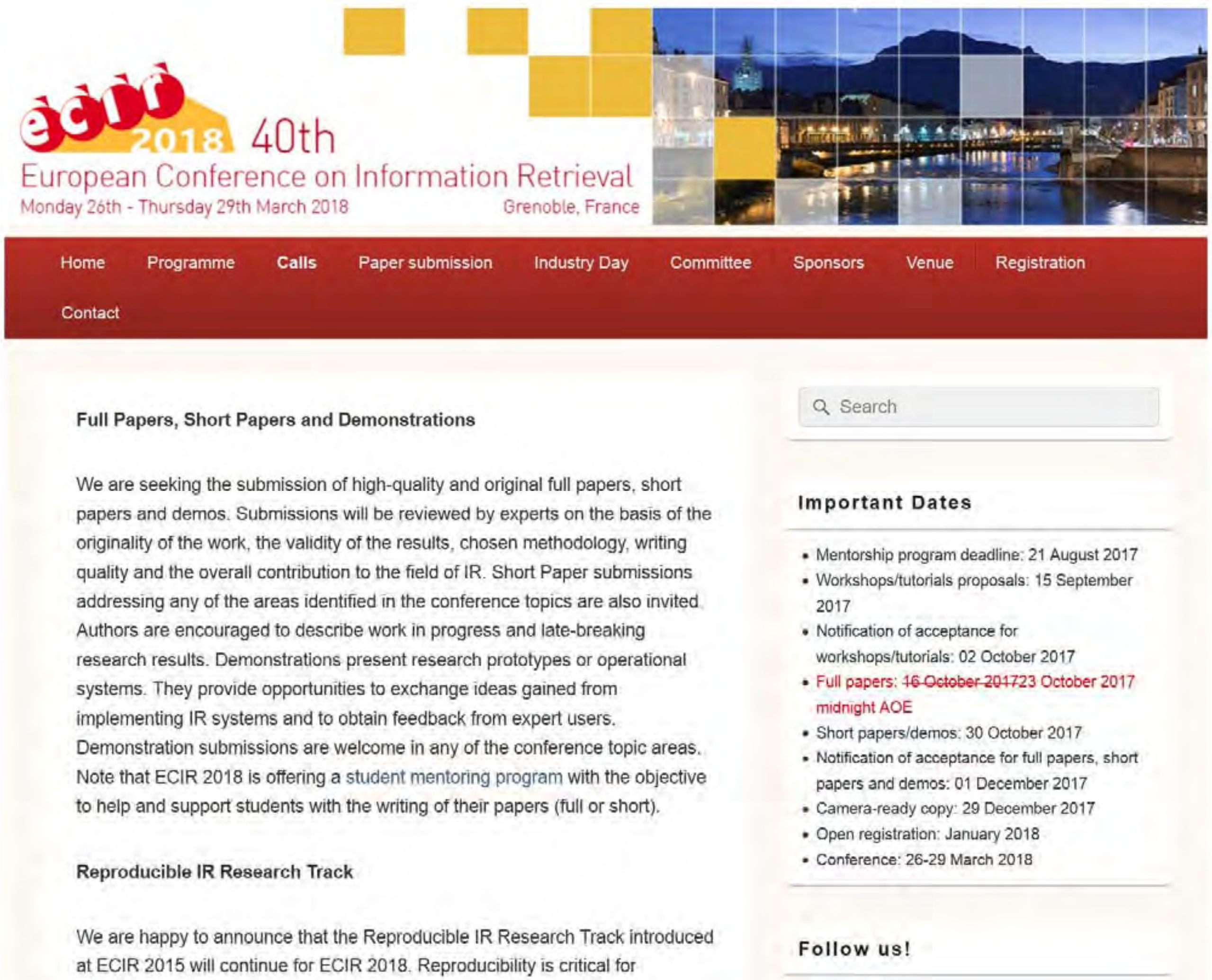


Reproducibility Papers

- When is a Reproducibility paper worth being published?



Reproducibility Papers


ecir 2018 40th

European Conference on Information Retrieval

Monday 26th - Thursday 29th March 2018 Grenoble, France

Home | Programme | Calls | Paper submission | Industry Day | Committee | Sponsors | Venue | Registration | Contact

Search

Full Papers, Short Papers and Demonstrations

We are seeking the submission of high-quality and original full papers, short papers and demos. Submissions will be reviewed by experts on the basis of the originality of the work, the validity of the results, chosen methodology, writing quality and the overall contribution to the field of IR. Short Paper submissions addressing any of the areas identified in the conference topics are also invited. Authors are encouraged to describe work in progress and late-breaking research results. Demonstrations present research prototypes or operational systems. They provide opportunities to exchange ideas gained from implementing IR systems and to obtain feedback from expert users. Demonstration submissions are welcome in any of the conference topic areas. Note that ECIR 2018 is offering a student mentoring program with the objective to help and support students with the writing of their papers (full or short).

Reproducible IR Research Track

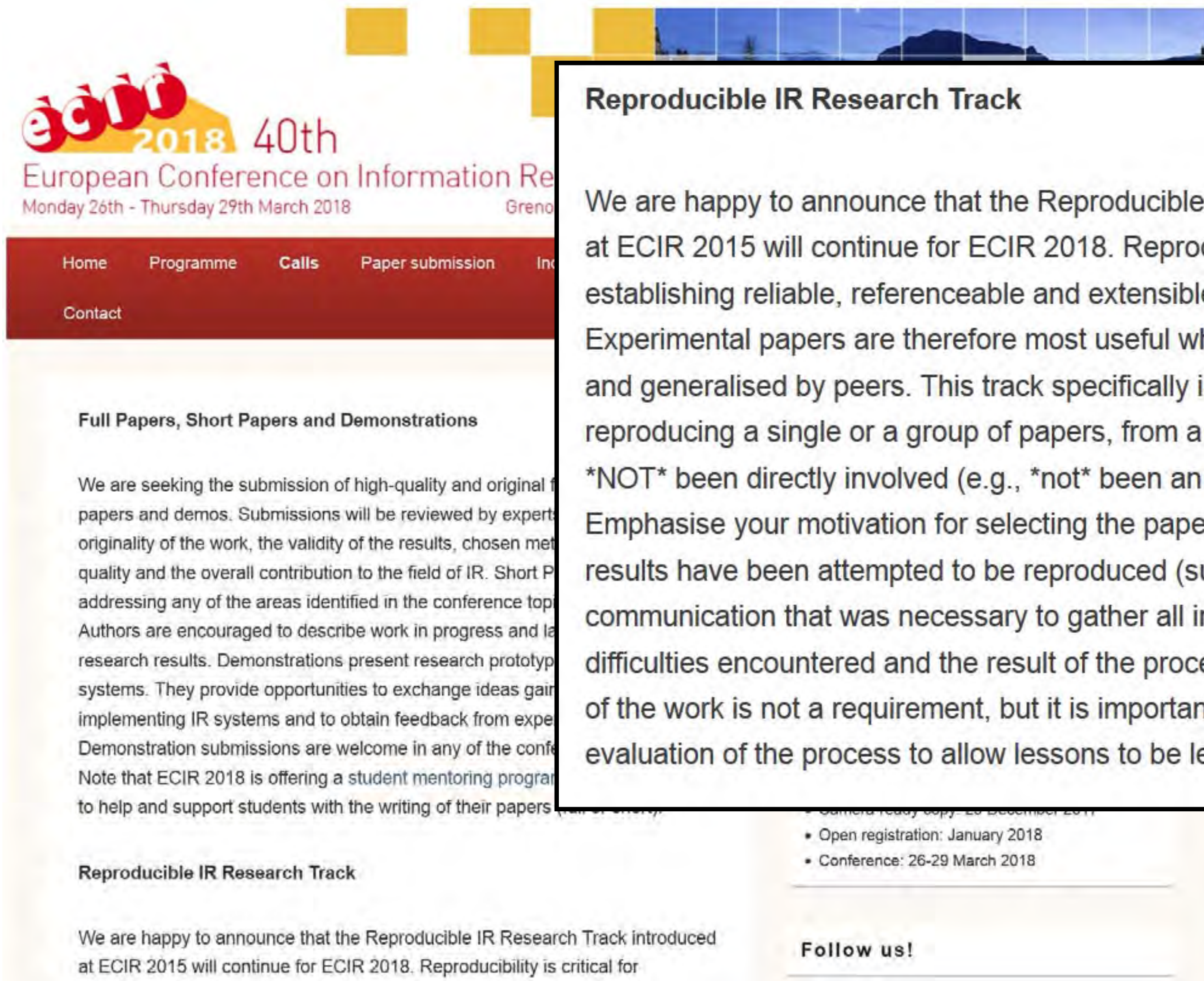
We are happy to announce that the Reproducible IR Research Track introduced at ECIR 2015 will continue for ECIR 2018. Reproducibility is critical for

Important Dates

- Mentorship program deadline: 21 August 2017
- Workshops/tutorials proposals: 15 September 2017
- Notification of acceptance for workshops/tutorials: 02 October 2017
- Full papers: ~~16 October 2017~~23 October 2017 midnight AOE
- Short papers/demos: 30 October 2017
- Notification of acceptance for full papers, short papers and demos: 01 December 2017
- Camera-ready copy: 29 December 2017
- Open registration: January 2018
- Conference: 26-29 March 2018

Follow us!

Reproducibility Papers



Reproducible IR Research Track

We are happy to announce that the Reproducible IR Research Track introduced at ECIR 2015 will continue for ECIR 2018. Reproducibility is critical for establishing reliable, referenceable and extensible research for the future. Experimental papers are therefore most useful when their results can be tested and generalised by peers. This track specifically invites submission of papers reproducing a single or a group of papers, from a third-party where you have *NOT* been directly involved (e.g., *not* been an author or a collaborator). Emphasise your motivation for selecting the paper/papers, the process of how results have been attempted to be reproduced (successful or not), the communication that was necessary to gather all information, the potential difficulties encountered and the result of the process. A successful reproduction of the work is not a requirement, but it is important to provide a clear and rigid evaluation of the process to allow lessons to be learned for the future.

Learning from Non-Reproducibility

- Do we always want reproducibility?
  - Scientifically speaking: yes!
- Research is addressing challenges:
  - Looking for and learning from non-reproducibility!
- Non-reproducibility if
  - Some (un-known) aspect of a study influences results
  - Technical: parameter sweep, bug in code, OS, … -> fix it!
  - Non-technical: input data! (specifically: “the user”)

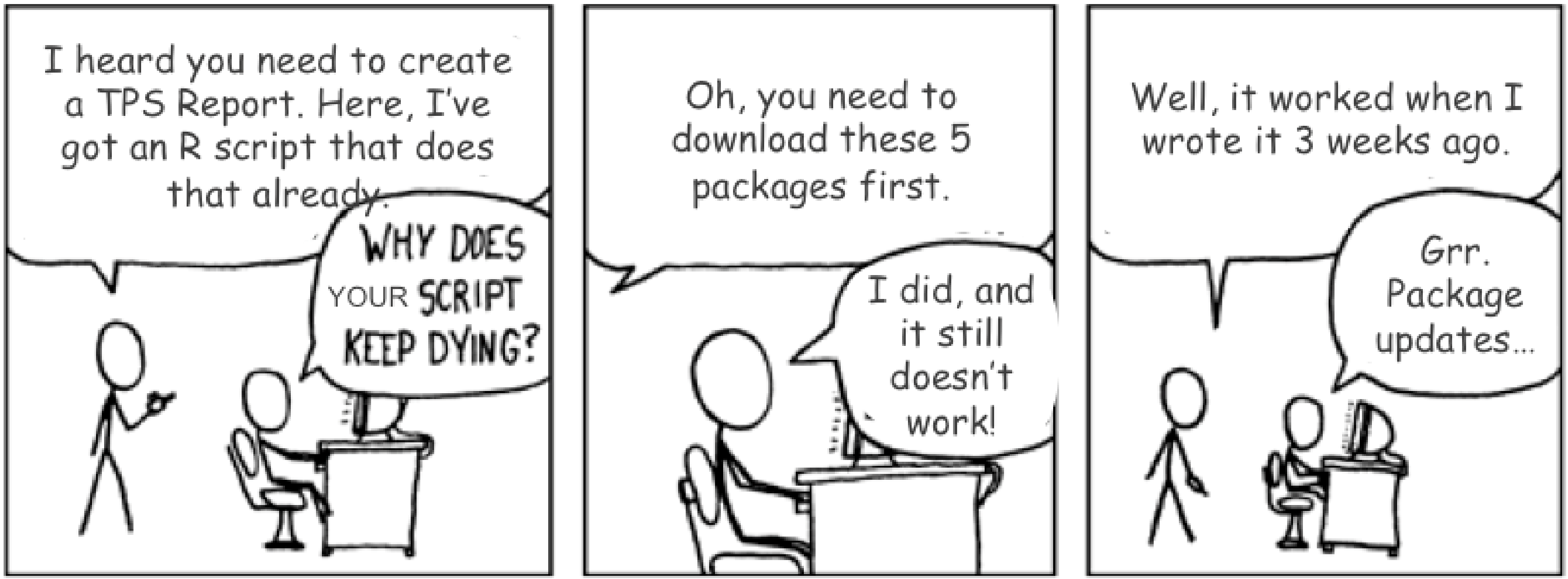
Learning from Non-Reproducibility

Challenges in MIR – "things don't seem to work"

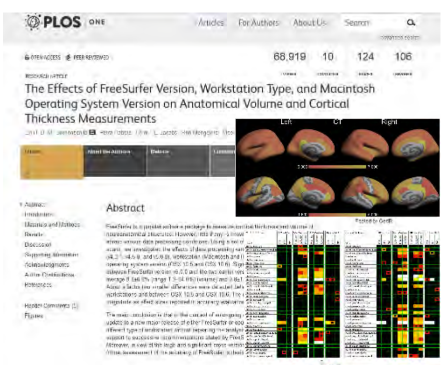
- Virtual Box, Github, <*your favourite tool*> are starting points
- Same features, same algorithm, different data ->
- Same data, different listeners ->
- Understanding "the rest":
  - Isolating unknown influence factors
  - Generating hypotheses
  - Verifying these to understand the "entire system", cultural and other biases, …
- Benchmarks and Meta-Studies

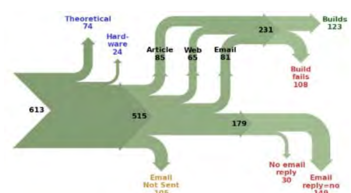
Challenges in Reproducibility

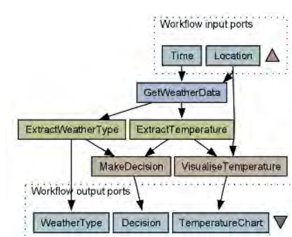
In a nutshell – and another aspect of reproducibility:



Source: xkcd





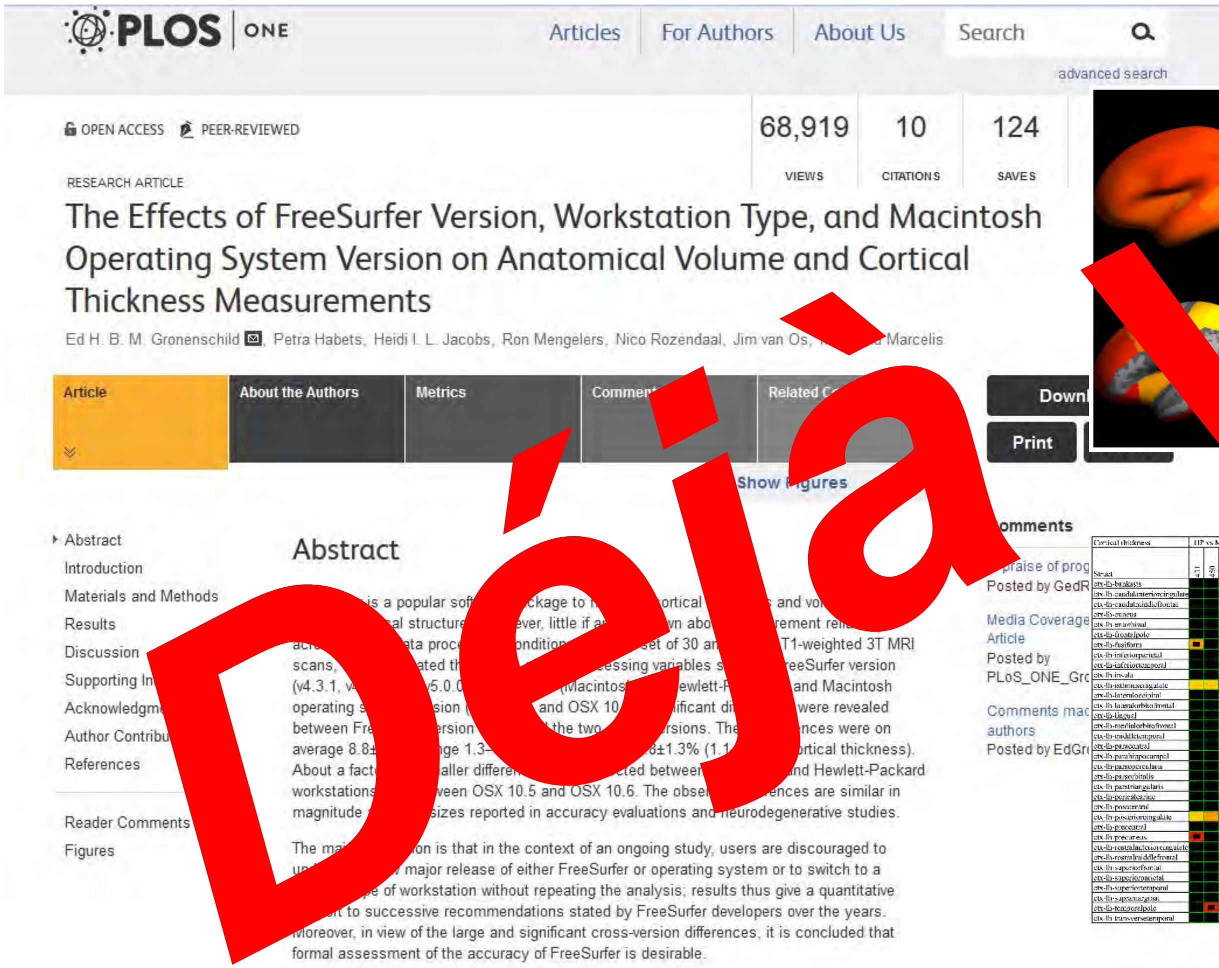


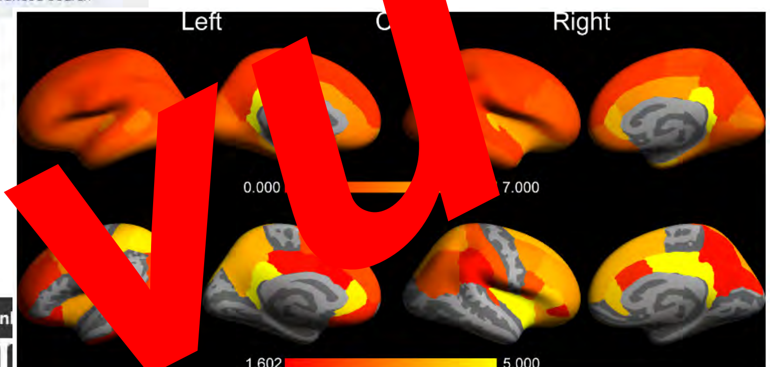
Outline

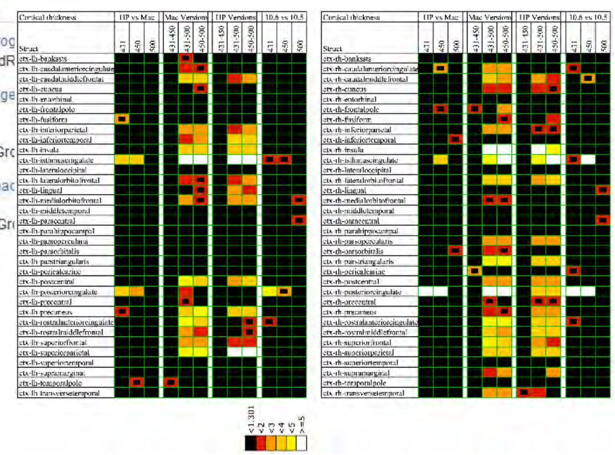
- Reproducibility
  - What are the challenges in reproducibility?
  - How to address the challenges of complex processes?
- Data Management & Citation
- Digital Preservation
- Summary

Challenges in Reproducibility

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234


PLOS | ONE
Articles | For Authors | About Us | Search
advanced search
OPEN ACCESS PEER-REVIEWED
68,919 VIEWS | 10 CITATIONS | 124 SAVES
RESEARCH ARTICLE
The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements
Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, ... Marcelis
Article | About the Authors | Metrics | Comment... | Related C...
Download
Print
Show Figures
Abstract
Introduction
Materials and Methods
Results
Discussion
Supporting In...
Acknowledgm...
Author Contribu...
References
Reader Comments
Figures
Abstract
... is a popular sof... ckage to ... ortical ... and vol... al structure... ver, little if a... wn abo... rement rel... acr... ta proc... ndition... et of 30 ar... T1-weighted 3T MRI scans, ... ated th... essing variables s... reeSurfer version (v4.3.1, v... 5.0.0... Macintos... ewlett-P... and Macintosh operating s... sion ... and OSX 10... ificant di... were revealed between Fre... ersion ... the two ... rsions. The... nces were on average 8.8± ... ge 1.3... 8±1.3% (1.1... rtical thickness). About a fact... aller differe... cted betwee... nd Hewlett-Packard workstations ... ween OSX 10.5 and OSX 10.6. The obse... ences are similar in magnitude ... sizes reported in accuracy evaluations and neurodegenerative studies.
The ma... on is that in the context of an ongoing study, users are discouraged to up... major release of either FreeSurfer or operating system or to switch to a ... e of workstation without repeating the analysis; results thus give a quantitative ... t to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.
...omments
...praise of prog...
Posted by GedR...
Media Coverage ... Article
Posted by PLoS_ONE_Gro...
Comments mad... authors
Posted by EdGr...

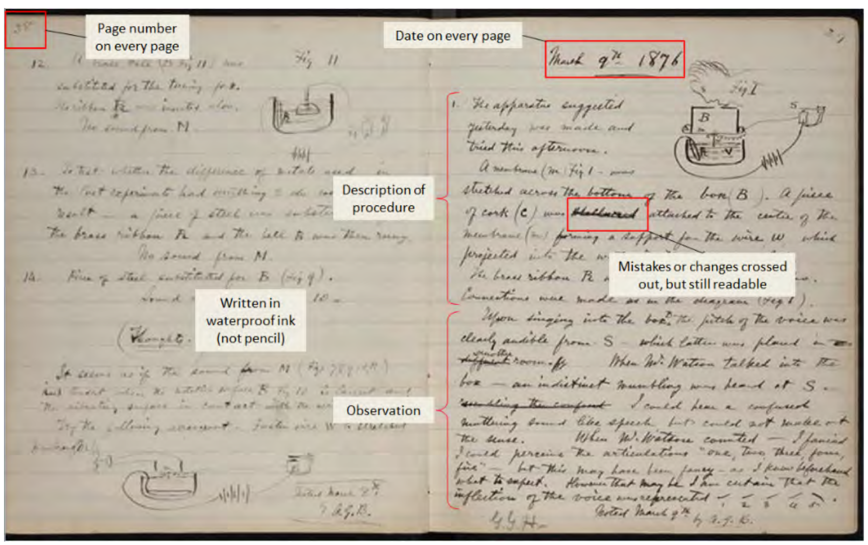




FACULTY OF !NFORMATICS

And the solution is…

- Standardization and Documentation
  - Standardized components, procedures, workflows
  - Documenting complete system set-up across entire provenance chain
- How to do this – efficiently?

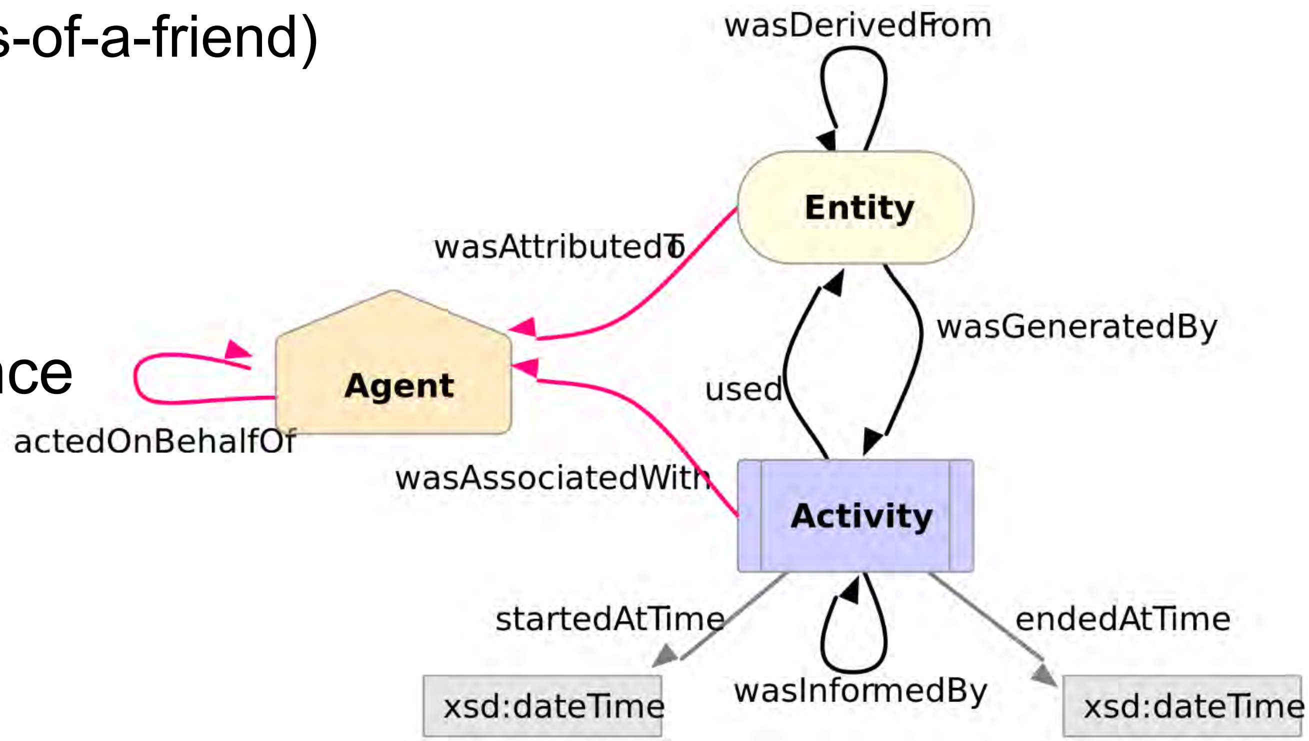


Alexander Graham Bell's Notebook, March 9 1876
https://commons.wikimedia.org/wiki/File:Alexander_Graham_Bell's_notebook,_March_9,_1876.PNG

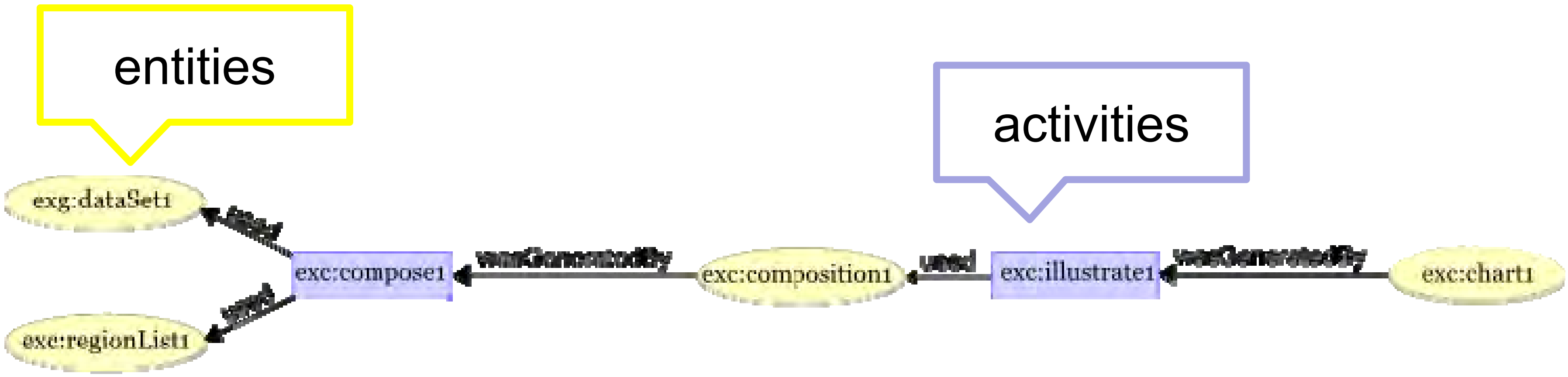
Pieter Bruegel the Elder: De Alchemist (British Museum, London)

PROV-O

- W3C Recommendation
  https://www.w3.org/TR/prov-o/
- Ontology to represent provenance information
- May use other languages
  - FOAF (friends-of-a-friend)
  - Dublin Core
  - PREMIS
- (Alternative:
  Open Provenance
  Model)



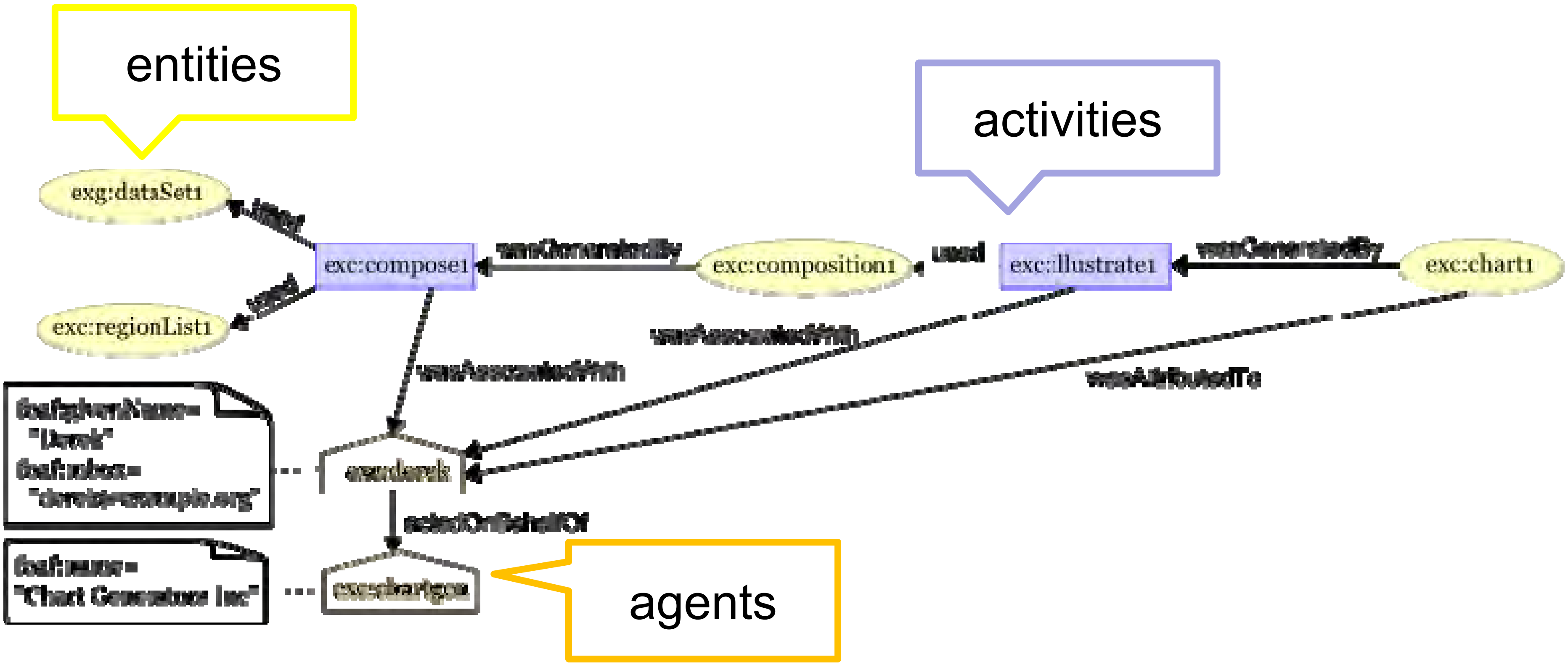
PROV-O



entities
activities
exg:dataSet1
exc:regionList1
exc:compose1
wasGeneratedBy
exc:composition1
used
exc:illustrate1
wasGeneratedBy
exc:chart1

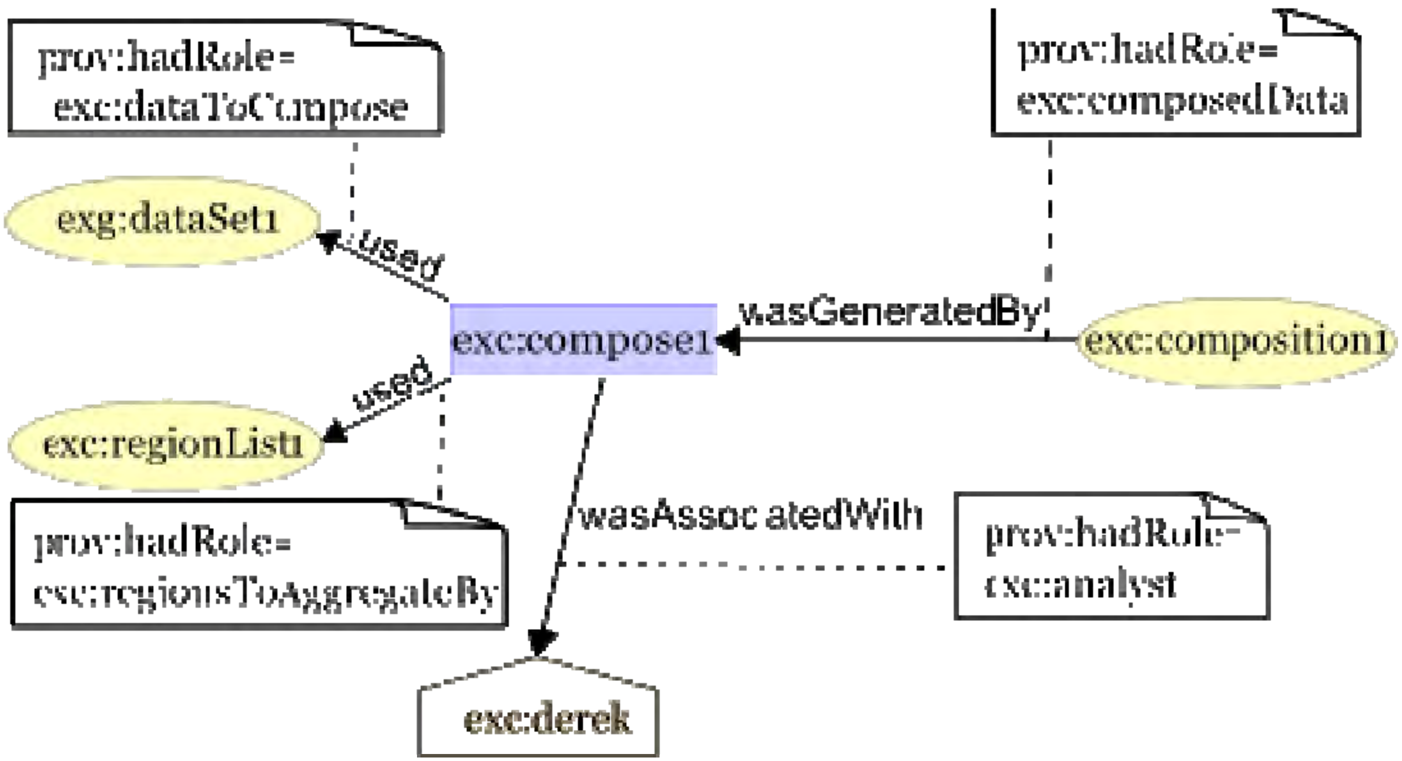
TU WIEN

PROV-O



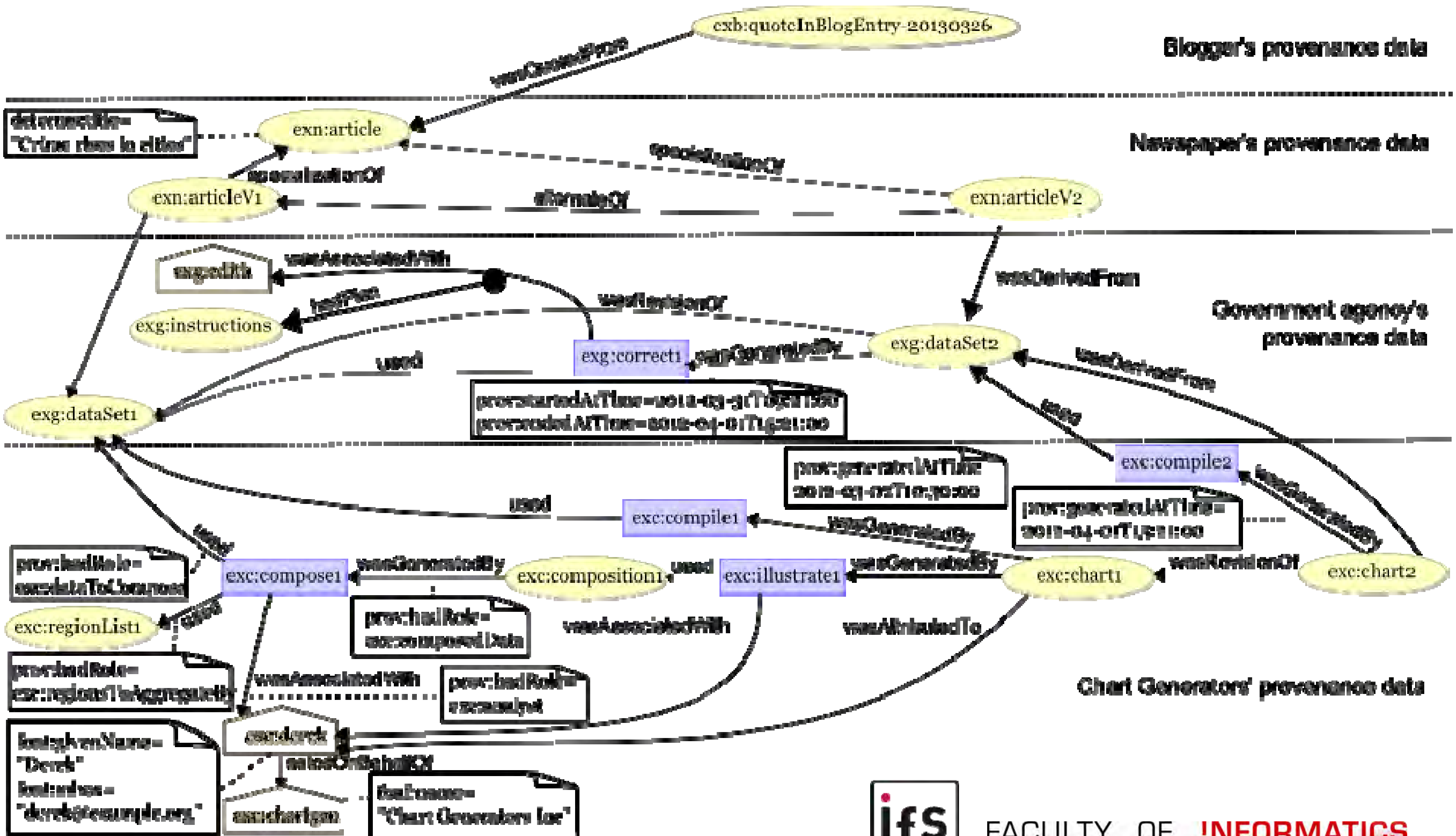
PROV-O

- Adding roles



PROV-O

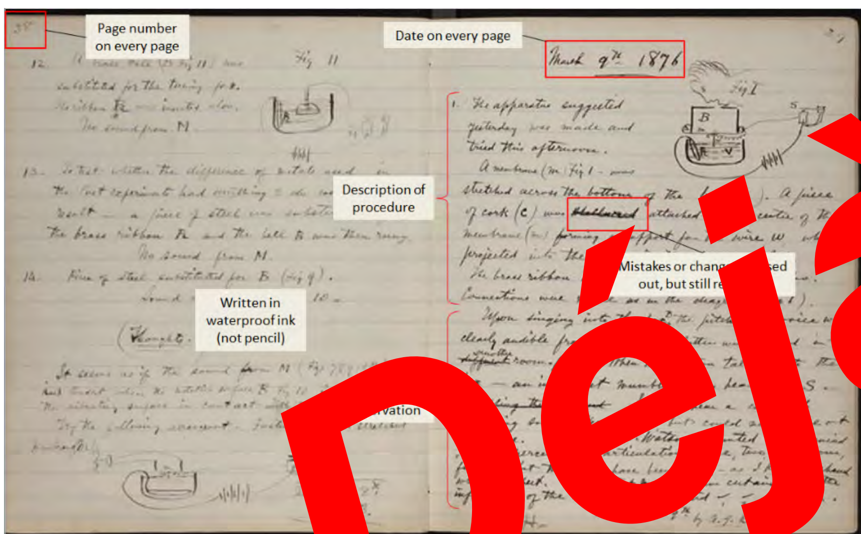
- Adding revisions, time dependencies, plans, …



FACULTY OF !NFORMATICS

And the solution is…

- Standardization and Documentation
  - Standardized components, procedures, workflows
  - Documenting complete system set-up across entire provenance chain
- **How to do this – efficiently?**



Déjà vu

Alexander Graham Bell's notebook, March 9 1876
https://commons.wikimedia.org/w…exander_Graham_Bell's_notebook,_March_9,_1876.PNG

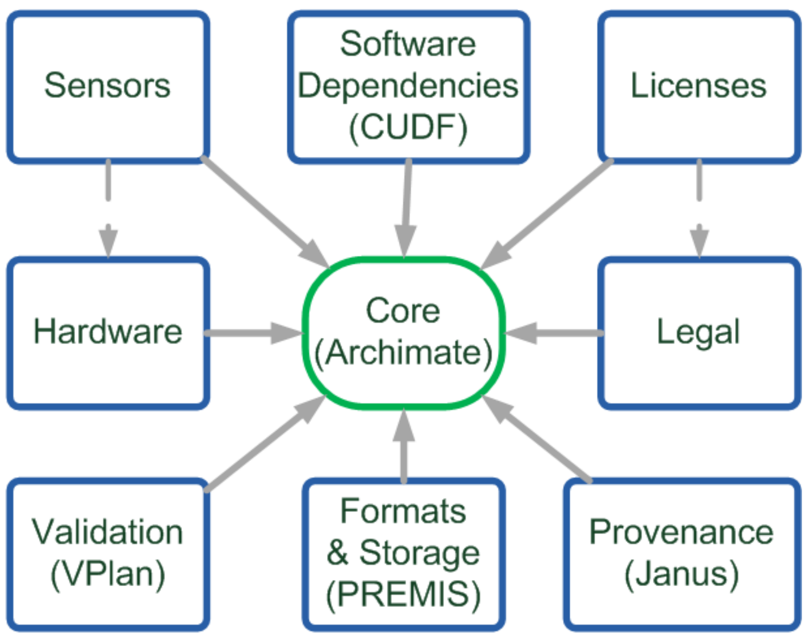
Pieter Bruegel the Elder: De Alchemist (British Museum, London)

And the solution is…

- Standardization and Documentation
  - Standardized components, procedures, workflows
  - Documenting complete system set-up across entire provenance chain
- **How to do this – efficiently!?**
- **Ideally:**
  - **Processing pipeline documents provenance automatically**
- **Reality:**
  - **Combination**
    - **automatic documentation / logging**
    - **monitoring behaviour of the system**

Documenting a Process

- Context Model: establish what to document and how
- Meta-model for describing process & context
  - Extensible architecture integrated by core model
  - Reusing existing models as much as possible
  - Based on ArchiMate, implemented using OWL
- Extracted by static and dynamic analysis

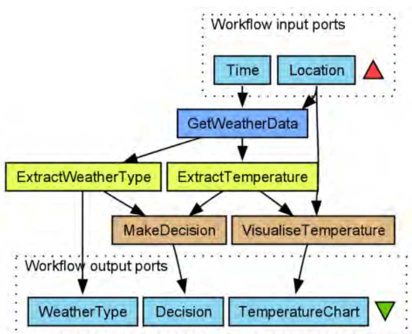


Context Model – Static Analysis

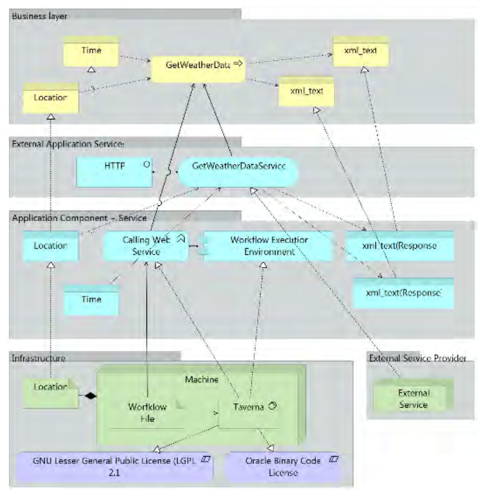
- Analyses steps, platforms, services, tools called
- Dependencies (packages, libraries)
- HW, SW Licenses, …

```
#!/bin/bash

# fetch data
java -jar GestBarragensWSClientIQData.jar
unzip -o IQData.zip

# fix encoding
#iconv -f LATIN1 -t UTF-8 iq.r > iq_utf8.r

# generate references
R --vanilla < iq_utf8.r > IQout.txt

# create pdf
pdflatex iq.tex
pdflatex iq.tex
```

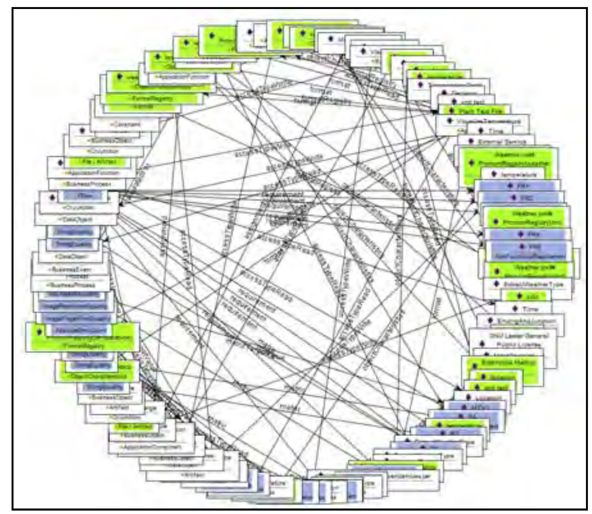
Script



Taverna Workflow



ArchiMate model

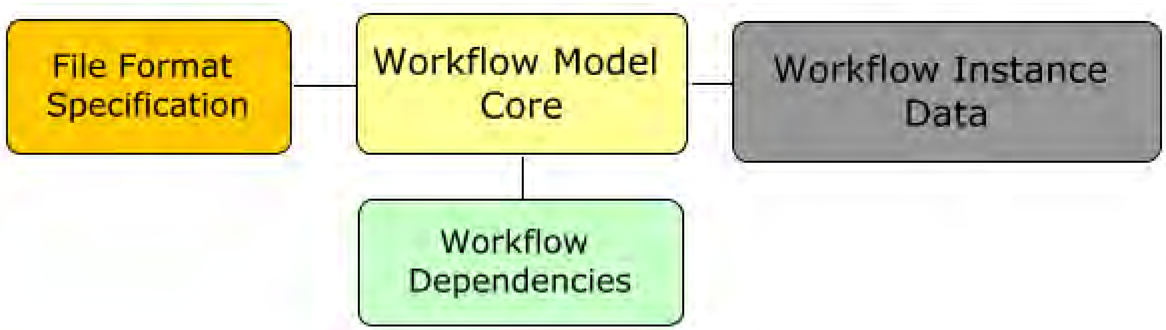


Context Model
(OWL ontology)

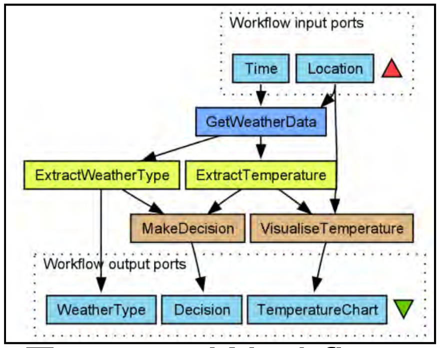
FACULTY OF !NFORMATICS

Context Model – Dynamic Analysis

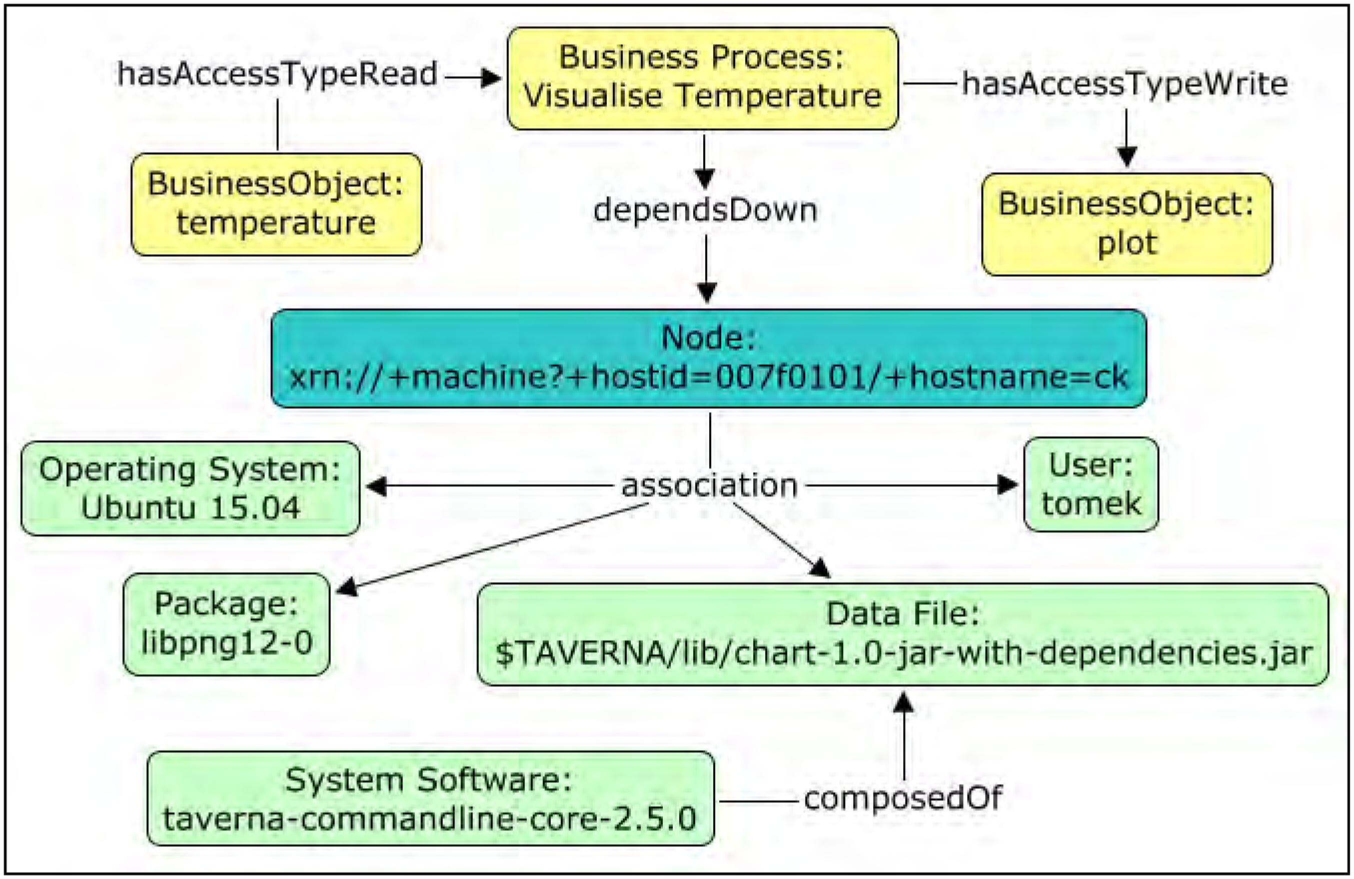
- Process Migration Framework (PMF)
  - designed for automatic redeployments into virtual machines
  - uses *strace* to monitor system calls
  - complete log of all accessed resources (files, ports)
  - captures and stores process instance data
  - analyse resources (file formats via PRONOM, PREMIS)



Context Model – Dynamic Analysis



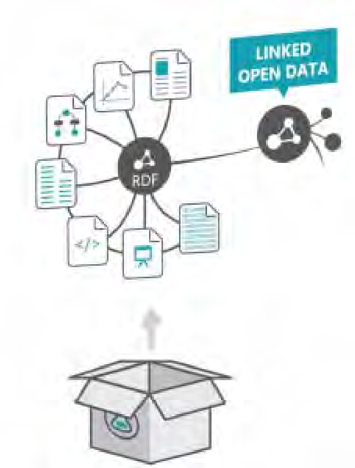
Taverna Workflow

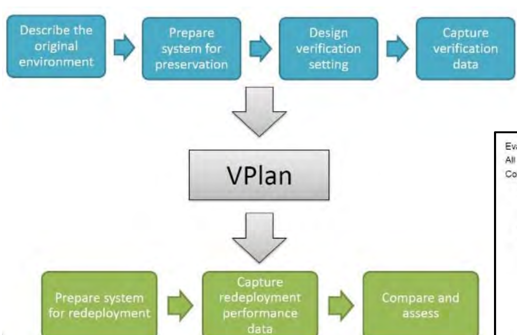


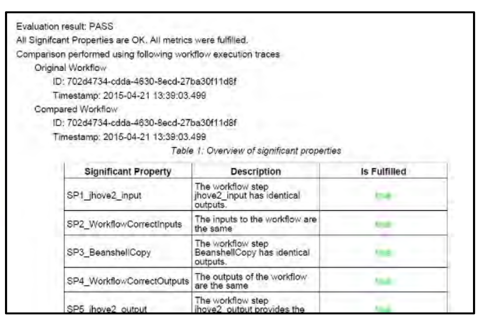
Process Capture

Preservation and Re-deployment

- „Encapsulate“ as complex Research Object (RO)
- DP: Re-Deployment beyond original environment
  - Format migration of elements of ROs
  - Cross-compilation of code
  - Emulation-as-a-Service
- Verification upon re-deployment







Evaluation result: PASS

All Signifcant Properties are OK. All metrics were fulfilled.

Comparison performed using following workflow execution traces

Original Workflow

ID: 702d4734-cdda-4630-8ecd-27ba30f11d8f

Timestamp: 2015-04-21 13:39:03.499

Compared Workflow

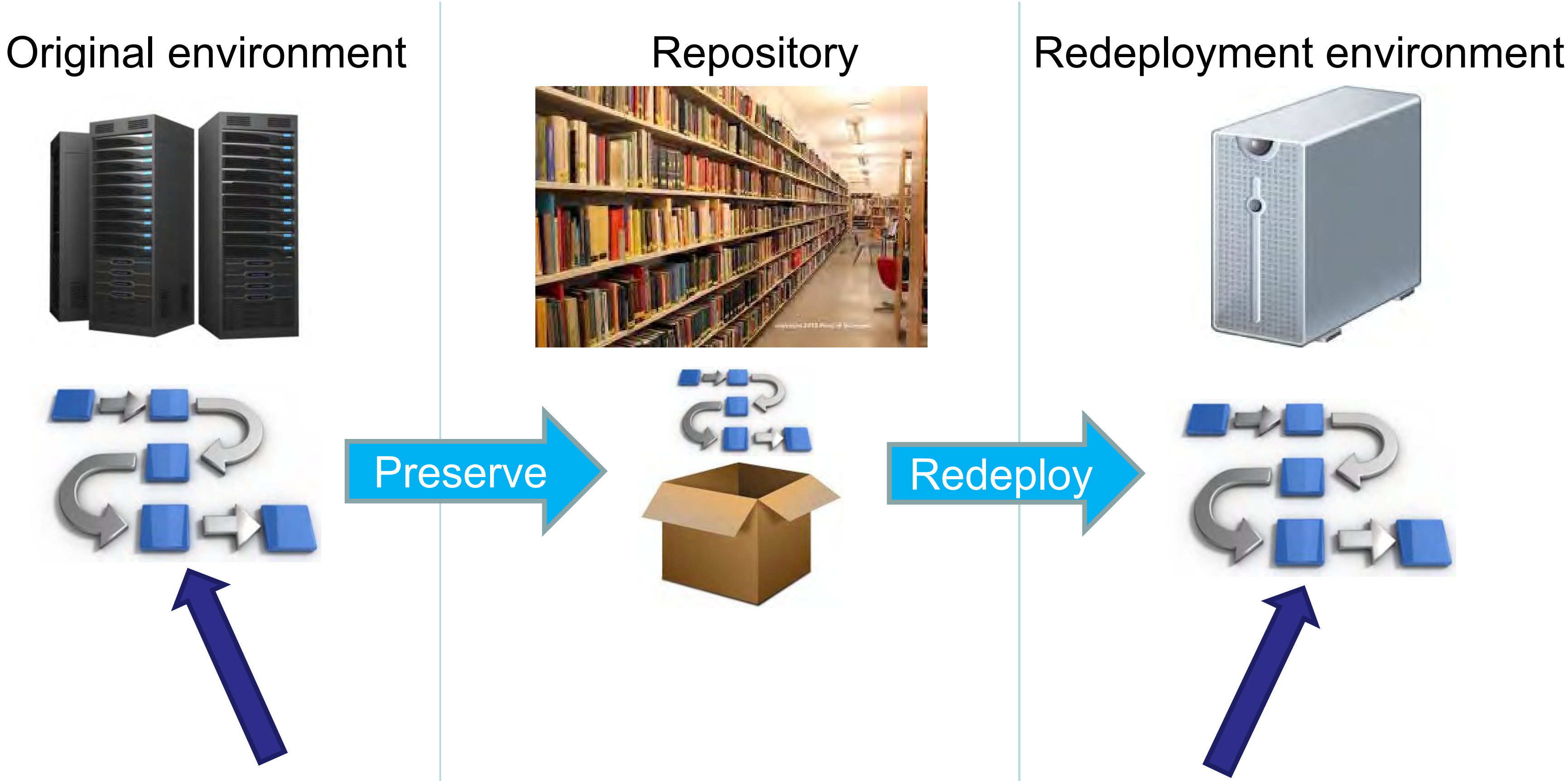
ID: 702d4734-cdda-4630-8ecd-27ba30f11d8f

Timestamp: 2015-04-21 13:39:03.499

Table 1: Overview of significant properties

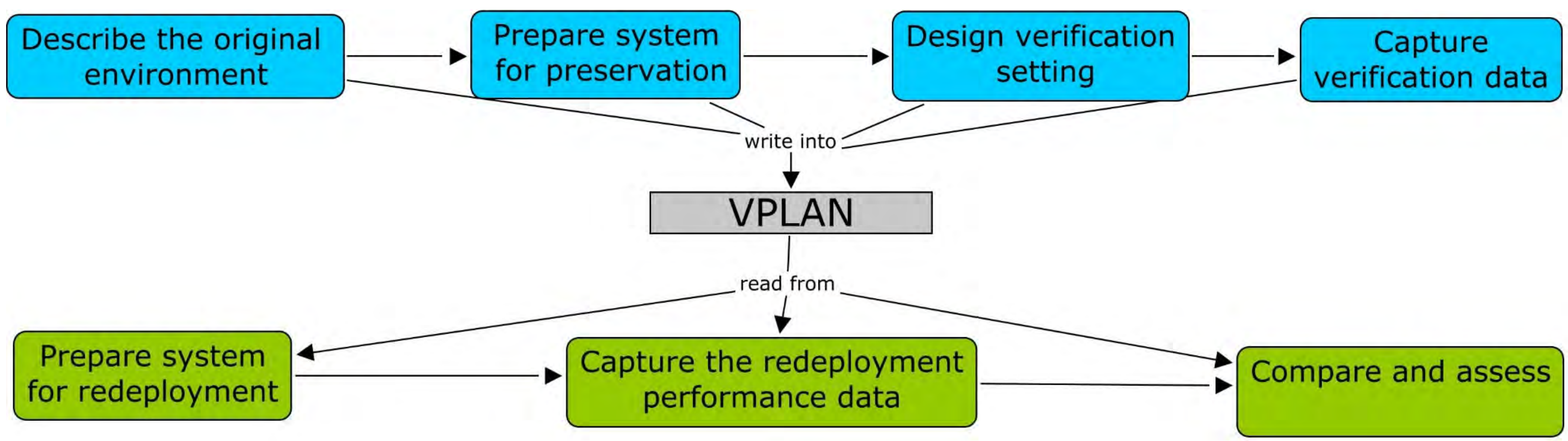
Significant Property	Description	Is Fulfilled
SP1_jhove2_input	The workflow step jhove2_input has identical outputs.	true
SP2_WorkflowCorrectInputs	The inputs to the workflow are the same	[illegible]
SP3_BeanshellCopy	The workflow step BeanshellCopy has identical outputs.	[illegible]
SP4_WorkflowCorrectOutputs	The outputs of the workflow are the same	[illegible]
SP5_jhove2_output	The workflow step jhove2_output provides the	[illegible]

VFramework



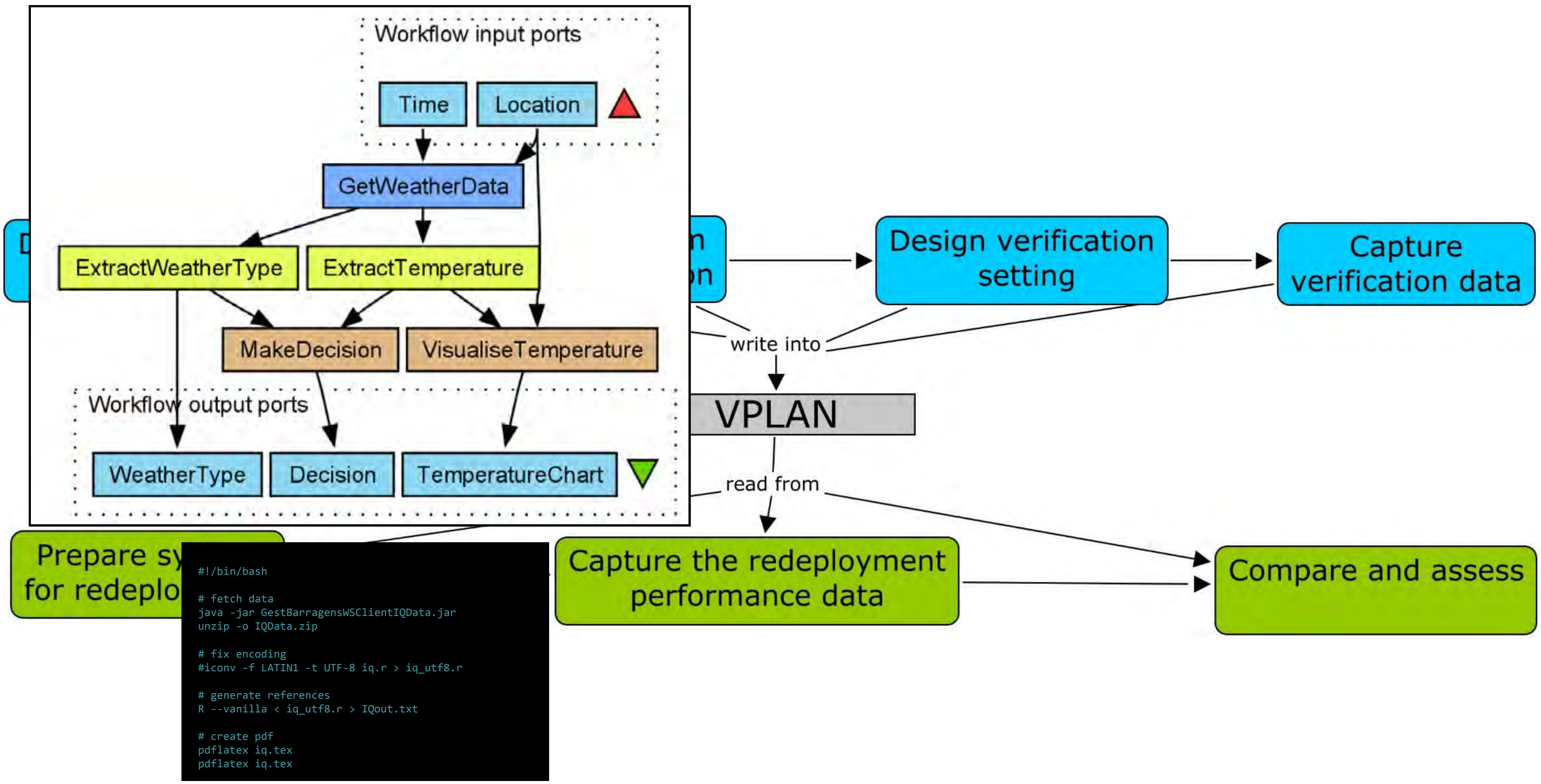
Are these processes the same?

VFramework



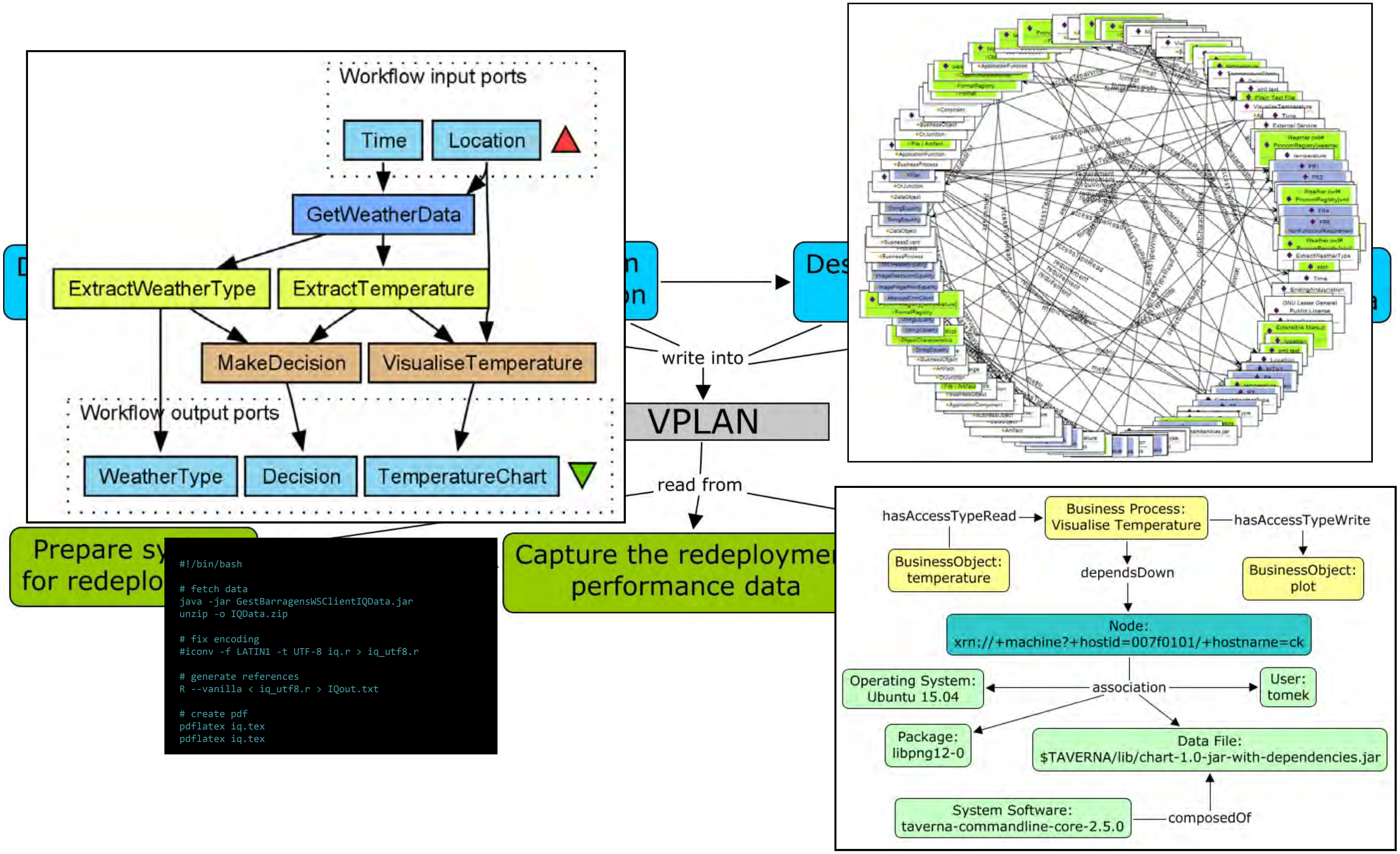
Describe the original environment
Prepare system for preservation
Design verification setting
Capture verification data
write into
VPLAN
read from
Prepare system for redeployment
Capture the redeployment performance data
Compare and assess

VFramework



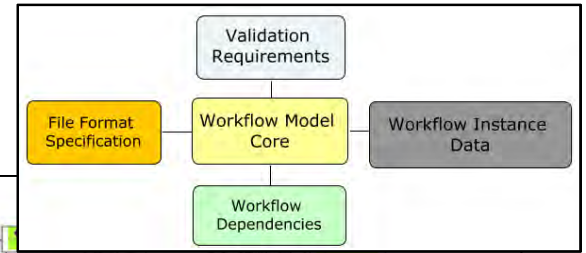
Workflow input ports
Time
Location
GetWeatherData
ExtractWeatherType
ExtractTemperature
MakeDecision
VisualiseTemperature
Workflow output ports
WeatherType
Decision
TemperatureChart
Design verification setting
Capture verification data
write into
VPLAN
read from
Prepare sy
for redeplo
Capture the redeployment performance data
Compare and assess
#!/bin/bash
fetch data
java -jar GestBarragensWSClientIQData.jar
unzip -o IQData.zip
fix encoding
#iconv -f LATIN1 -t UTF-8 iq.r > iq_utf8.r
generate references
R --vanilla < iq_utf8.r > IQout.txt
create pdf
pdflatex iq.tex
pdflatex iq.tex

VFramework

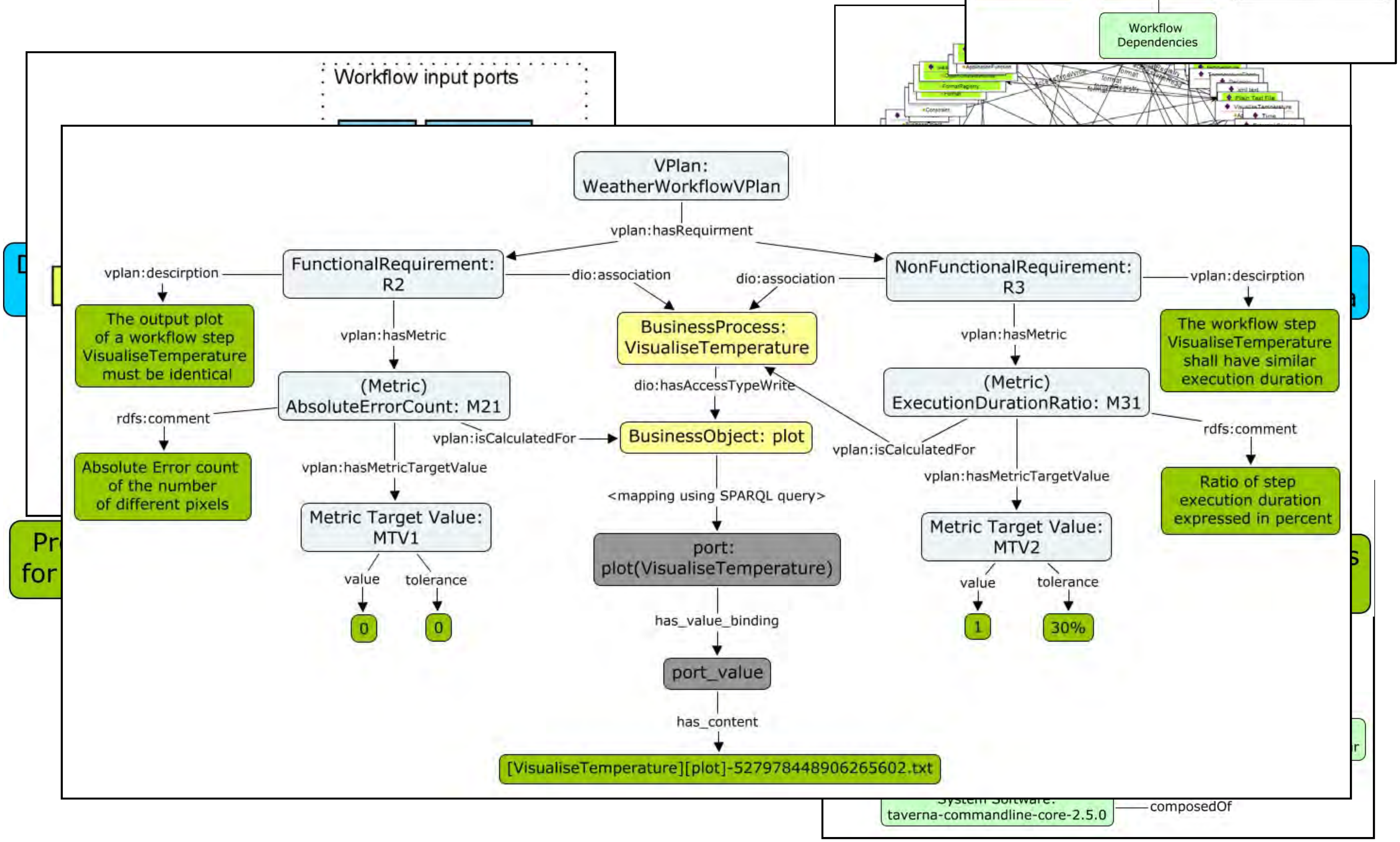


FACULTY OF !INFORMATICS

VFramework

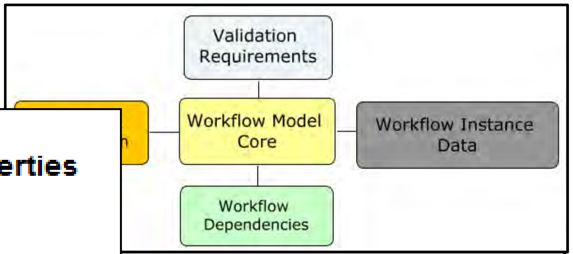


Validation Requirements
File Format Specification
Workflow Model Core
Workflow Instance Data
Workflow Dependencies



Workflow input ports
VPlan: WeatherWorkflowVPlan
vplan:hasRequirment
FunctionalRequirement: R2
NonFunctionalRequirement: R3
vplan:descirption
dio:association
dio:association
vplan:descirption
The output plot of a workflow step VisualiseTemperature must be identical
BusinessProcess: VisualiseTemperature
The workflow step VisualiseTemperature shall have similar execution duration
vplan:hasMetric
vplan:hasMetric
(Metric) AbsoluteErrorCount: M21
dio:hasAccessTypeWrite
(Metric) ExecutionDurationRatio: M31
rdfs:comment
vplan:isCalculatedFor
BusinessObject: plot
vplan:isCalculatedFor
rdfs:comment
Absolute Error count of the number of different pixels
vplan:hasMetricTargetValue
<mapping using SPARQL query>
vplan:hasMetricTargetValue
Ratio of step execution duration expressed in percent
Metric Target Value: MTV1
port: plot(VisualiseTemperature)
Metric Target Value: MTV2
value
tolerance
value
tolerance
0
0
1
30%
has_value_binding
port_value
has_content
[VisualiseTemperature][plot]-527978448906265602.txt
taverna-commandline-core-2.5.0
composedOf

VFramework



- DataFile (2)
  - **$USER_HOME/.java/fonts/1.8.0_45-internal/fcinfo-1-korona-LinuxMint-17-en.properties**
  - **$USER_HOME/taverna-commandline-core-2.5.0/lib/somtoolbox_full.jar**
- HTTPServiceInterface (4)
  - **127.0.1.1_interface**
  - **127.0.0.1_interface**
  - **::ffff:127.0.0.1_interface**
  - **'datapoint.metoffice.gov.uk/public/data/val/wxfcs/all/xml/322690?res=3hou**
- InfrastructureFunction (5)
- OperatingSystem (1)
  - **'Linux Mint 17 Qiana'**
- Service (4)
- User (1)
  - **timbus**

ADDED

- DataFile (2)
  - **$USER_HOME/.java/fonts/1.8.0_45-internal/fcinfo-1-ck-Ubuntu-15.04-en.**
  - **/usr/share/fonts/X11/Type1/fonts.dir**
- HTTPServiceInterface (3)
  - **23.0.174.40_interface**
  - **::ffff:23.0.174.40_interface**
  - **23.0.174.9_interface**
- OperatingSystem (1)
  - **'Ubuntu 15.04'**
- Package (48)
  - **fonts-takao-pgothic**
  - **libxcb1**
  - **language-selector-common**
  - **fonts-tlwg-typewriter**
  - **ttf-indic-fonts-core**
  - **fonts-tlwg-garuda**

NOT USED

Dependencies Overview

Shell calls	0
Remote services	3
Specific debian packages required	48
Specific file dependencies	1
Data files processed during workflow execution	7

Detailed results

OS specific command line invocations

There are no shell calls.

Workflow communication to external hosts

23.0.174.40_interface
23.0.174.9_interface
::ffff:23.0.174.40_interface

Required additional files and libraries

/home/tomek/taverna-commandline-core-2.5.0/lib/chart-1.0-jar-with-dependencies.jar

Data files used by the workflow

/home/tomek/Weather/
/home/tomek/Weather/log
/home/tomek/Weather/output/Decision/1/1
/home/tomek/Weather/output/TemperatureChart/1
/home/tomek/Weather/output/WeatherType/1
/home/tomek/Weather/workflow/Weather.t2flow
/home/tomek/Weather/workflowInvocation.sh

Required additonal Debian packages

base-files
cups-filters
fontconfig-config

VFramework

- Documents system set-up and process execution
- Represents data in ontology
- Can be used as provenance documentation
- Can be used to verify re-execution
- Can be used to trace causes for differing behaviour
  - Tomasz Miksa, Andreas Rauber. Using ontologies for verification and validation of workflow-based experiments, Web Semantics: Science, Services and Agents on the World Wide Web, 43:25-45, March 2017. https://doi.org/10.1016/j.websem.2017.01.002
  - Tomasz Miksa, Andreas Rauber, Eleni Mina. Identifying Impact of Software Dependencies on Replicability of Biomedical Workflows. Journal of Biomedical Informatics 64:232-254, 2016. https://doi.org/10.1016/j.jbi.2016.10.011

Outline

- Reproducibility
  - What are the challenges in reproducibility?
  - How to address the challenges of complex processes?
- Data Management & Citation
- Explainable AI
- Summary

Outline

- Reproducibility
- Data Management & Citation
  - Why should we cite data?
  - What is so difficult about citing data?
  - How should we do it?
- Explainable AI
- Summary

Not covered

Outline

- Reproducibility
- Data Management & Citation
  - Why should we cite data?
  - What is so difficult about citing data?
  - How should we do it?
- Explainable AI
- Summary

Outline

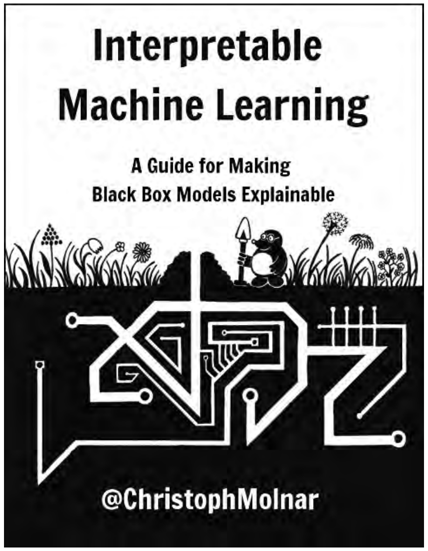
- Reproducibility
- Data Management & Citation
- Explainable AI
  - What is Explainability in ML and why do we need it?
  - Interpretable Models
  - Model-agnostic Approaches to Explainability
- Summary

Explainable ML

Reading Material

- [Riccardo Guidotti](), [Anna Monreale](), [Salvatore Ruggieri](), [Franco Turini](), Fosca Giannotti, [Dino Pedreschi]():
  A Survey of Methods for Explaining Black Box Models. [ACM Comput. Surv. 51(5)](): 93:1-93:42 (2019)
- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019.
  [https://christophm.github.io/interpretable-ml-book]()
- Further references in the slides

A Survey of Methods for Explaining Black Box Models



Explainability: What and Why?

- **Interpretability is the degree to which a human can understand the cause of a decision**
  (Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” arXiv Preprint arXiv:1706.07269)
- **Interpretability is the degree to which a human can consistently predict the model’s result**

Explainability: What and Why?

- Goal of Science:
  - Curiosity / learning (eat green berries -> sick)
  - Understanding the model
  - Detecting bias
  - Achieve / increase social acceptance
  - Debugging and auditing
  - Checking for essential characteristics
    - Fairness
    - Privacy
    - Reliability, robustness
    - Causality
  - Trust!

Excursion: Ethics & Privacy

- **ACM Statement on Algorithmic Transparency and Accountability**, May 25 2017

  http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

1. **Awareness**: potential bias
2. **Access and redress**: for individuals and groups
3. **Accountability**: responsible for decisions made by algorithms
4. **Explanation**: encouraged to explain procedures, decisions
5. **Data Provenance**: data collection, bias analysis, …
6. **Auditability**: models, data, algorithms recorded
7. **Validation and Testing**: rigorous, routinely, public

Explainability: What and Why?

- When do we not need explainability?

Explainability: What and Why?

- When do we not need explainability?
  - No impact (e.g. private use)
  - Well-studied and established (e.g. OCR)
    (but: beware changing world: adversary input, repurposing,…)
  - Risk of exposure: gaming the system
    (but: internal auditing, possibility of inspection!)

Explainability: What and Why?

- Types of explainability
  - Intrinsic: model-inherent (i.e. a linear model)
    (but: beware of complexity of the model!)
  - Post-hoc: extracting information
  - Ex-ante: data statistics, bias in data, definition of task
- Types of explanations
  - Feature statistics / visualizations
  - Model internals (e.g. weights)
  - Examples and counter-examples
  - Proxy models: simpler, easier to understand
    (but potentially wrong)

Explainability: What and Why?

- Types of approaches
  - Model-specific vs. model-agnostic
  - Local explanations vs. global explanations
    - Local: per instance, class, region, …
    - Global: holistic vs. modular: per attribute / per set of instances

- Model transparency vs. Algorithmic transparency
  - Knowing and UNDERSTANDING what the algorithm does
  - Source code is not sufficient!

Explainability: What and Why?

Quality criteria for explanations

- Contrastive: not just “*why x?*” but “*why x not y?*”
  - most similar with different outcome,
  - most influential characteristics
- Social setting: target audience
  - User / affected person vs. model builder/debugger vs. legal …
- Coherent with believes / intuition / knowledge (“Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” Review of General Psychology 2 (2). Educational Publishing Foundation: 175)
- Generalization: cover many cases
- Truthfullness: holds for other examples as well
- Selectiveness: not entire set of reasons, but most significant
- Abnormal features: prefer rare categorical values over frequent, outliers, …

Explainability: What and Why?

Evaluating explanations

- Real task
- Proxy task: simpler task, selected users judging quality
- Functional:
  - Explanation size
  - Sparsity (how many features?)
  - Feature complexity
  - Interaction of features
  - Monotonicity
  - Uncertainty part of the explanation
  - Cognitive processing time

Outline

- What is Explainability in ML and why do we need it?
- Interpretable Models
- Model-agnostic Approaches to Explainability

Interpretable Models

Algorithm	Linear	Monotone	Interaction	Task
Linear models	Yes	Yes	No	Regr.
Logistic regression	No	Yes	No	Class.
Decision trees	No	Some	Yes	Class. + Regr.
RuleFit	Yes	No	Yes	Class. + Regr.
Naive Bayes	Yes	Yes	No	Class.n
k-nearest neighbours	No	No	No	Class. + Regr.

Linear Regression

- Popular ML model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

- Many assumptions (often violated)
  - Linearity
  - Normal distribution of outcome
  - Homoscedascity: constant variance of error terms
    (e.g. variance of house prices is constant across different size ranges of houses)
  - Independence of instances:
    multiple measurements per data point (house, customer)
  - Fixed features, no errors
  - Absence of multicollinearity: no correlation across features
    (one will be picked as dominant, the other contributes variance)

Linear Regression

- Interpretation
  - Numerical feature: increase in x_j -> outcome changes by β_j
  - Binary feature: flip x_j from base level changes outcome by β_j
  - Categorical feature: on-hot encoding
  - Baseline / intercept β_0
  - R2 / Sum of Squared Errors (SSE): how much of the total variance in data is explained by model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Linear Regression

- Example: bike rental

	Weight estimate	Std. Error
(Intercept)	2399.4	238.3
seasonSUMMER	899.3	122.3
seasonFALL	138.2	161.7
seasonWINTER	425.6	110.8
holidayHOLIDAY	-686.1	203.3
workingdayWORKING DAY	124.9	73.3
weathersitMISTY	-379.4	87.6
weathersitRAIN/SNOW/STORM	-1901.5	223.6
temp	110.7	7.0
hum	-17.4	3.2
windspeed	-42.5	6.9
days_since_2011	4.9	0.2

Linear Regression

- Example: bike rental

	Weight estimate	Std. Error
(Intercept)	2399.4	238.3
seasonSUMMER	899.3	122.3
seasonFALL	138.2	161.7
seasonWINTER	425.6	110.8
holidayHOLIDAY		
workingdayWORKING DA		
weathersitMISTY		
weathersitRAIN/SNOW/STORM	-1901.5	223.6
temp	110.7	7.0
hum	-17.4	3.2
windspeed	-42.5	6.9
days_since_2011	4.9	0.2

increase of the temperature by 1 degree Celsius increases the expected number of bikes by 110.7, given all other features stay the same

Linear Regression

- Example: bike rental

	Weight estimate	Std. Error
(Intercept)	2399.4	238.3
seasonSUMMER	899.3	122.3
seasonFALL	138.2	161.7
seasonWINTER		
holidayHOLIDAY		
workingdayWORKING DAY		
weathersitMISTY	-379.4	87.6
weathersitRAIN/SNOW/STORM	-1901.5	223.6
temp	110.7	7.0
hum	-17.4	3.2
windspeed	-42.5	6.9
days_since_2011	4.9	0.2

estimated number of bikes is 1901.5 lower when it is rainy, snowing or stormy, compared to good weather, given that all other features stay the same

Linear Regression

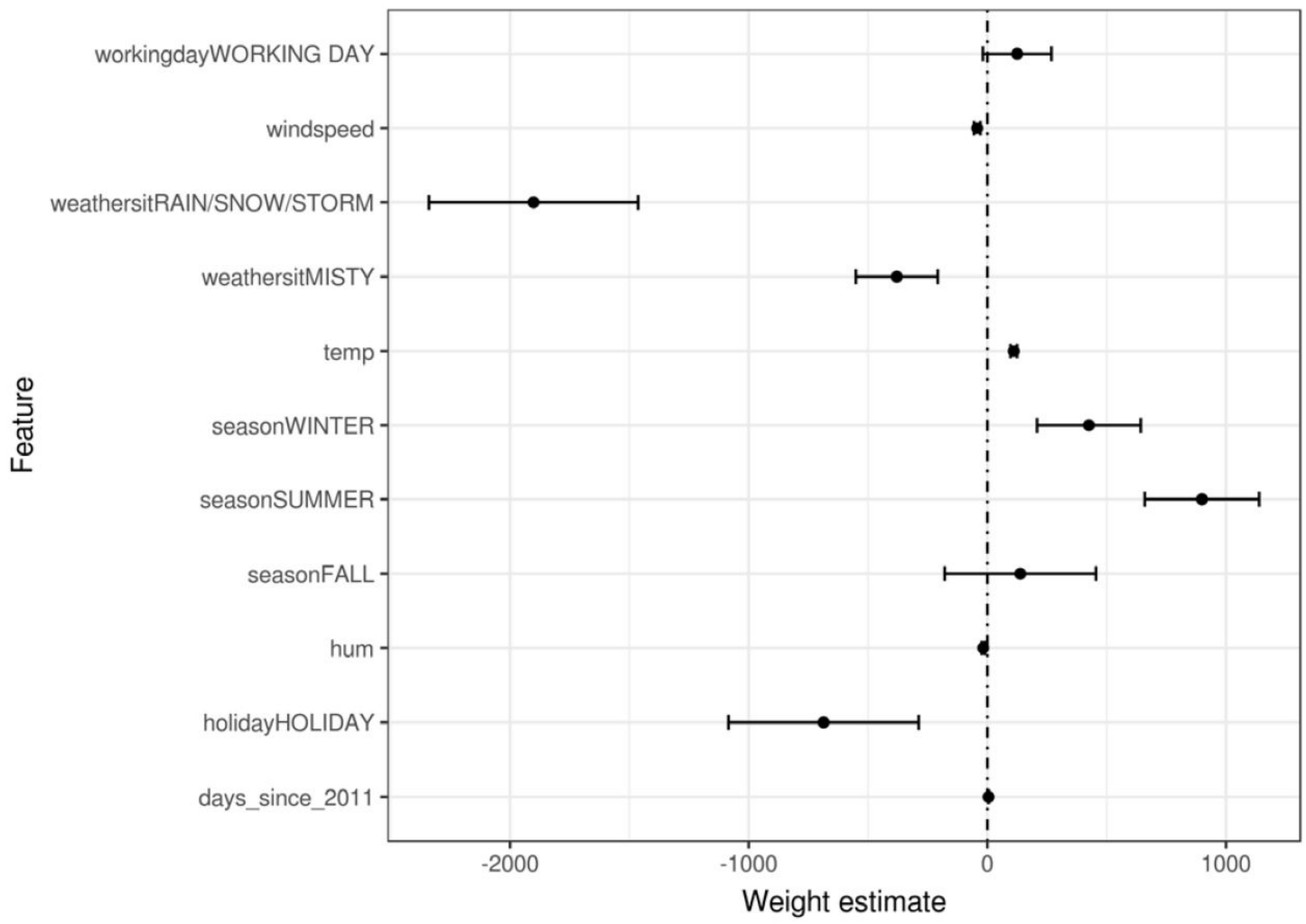
- Example: bike rental

	Weight estimate	Std. Error
(Intercept)	2399.4	238.3
seasonSUMMER	899.3	122.3
seasonFALL	138.2	161.7
seasonWINTER	[illegible]	[illegible]
holidayHOLIDAY	[illegible]	[illegible]
workingdayWORKING DAY	124.9	73.3
weathersitMISTY	-379.4	87.6
weathersitRAIN/SNOW/STORM	-1901.5	223.6
temp	110.7	7.0
hum	-17.4	3.2
windspeed	-42.5	6.9
days_since_2011	4.9	0.2

if the weather was misty, the expected number of bikes is 379.4 lower, compared to good weather, given that all other features stay the same

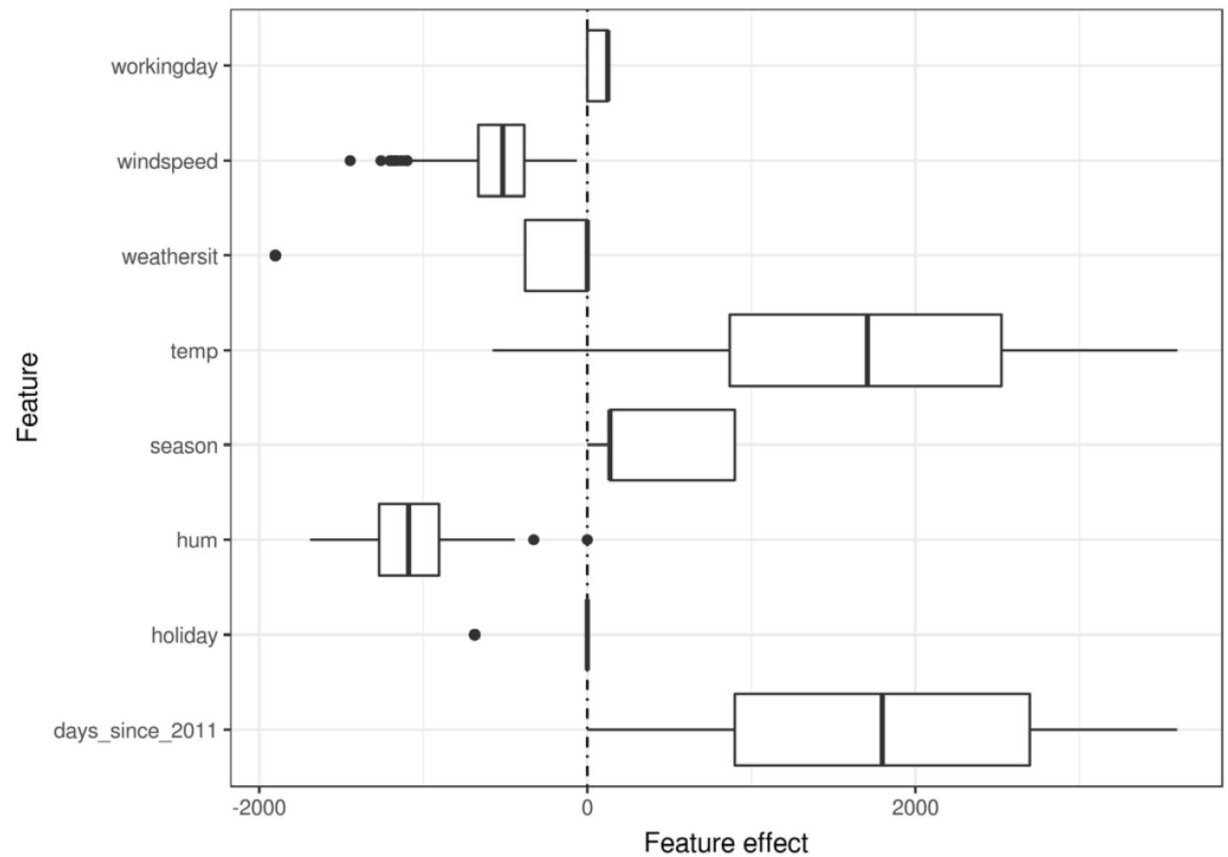
Linear Regression

- Example: bike rental:
  - Plotting feature weights + 95% confidence interval
  - Note: different scales!!



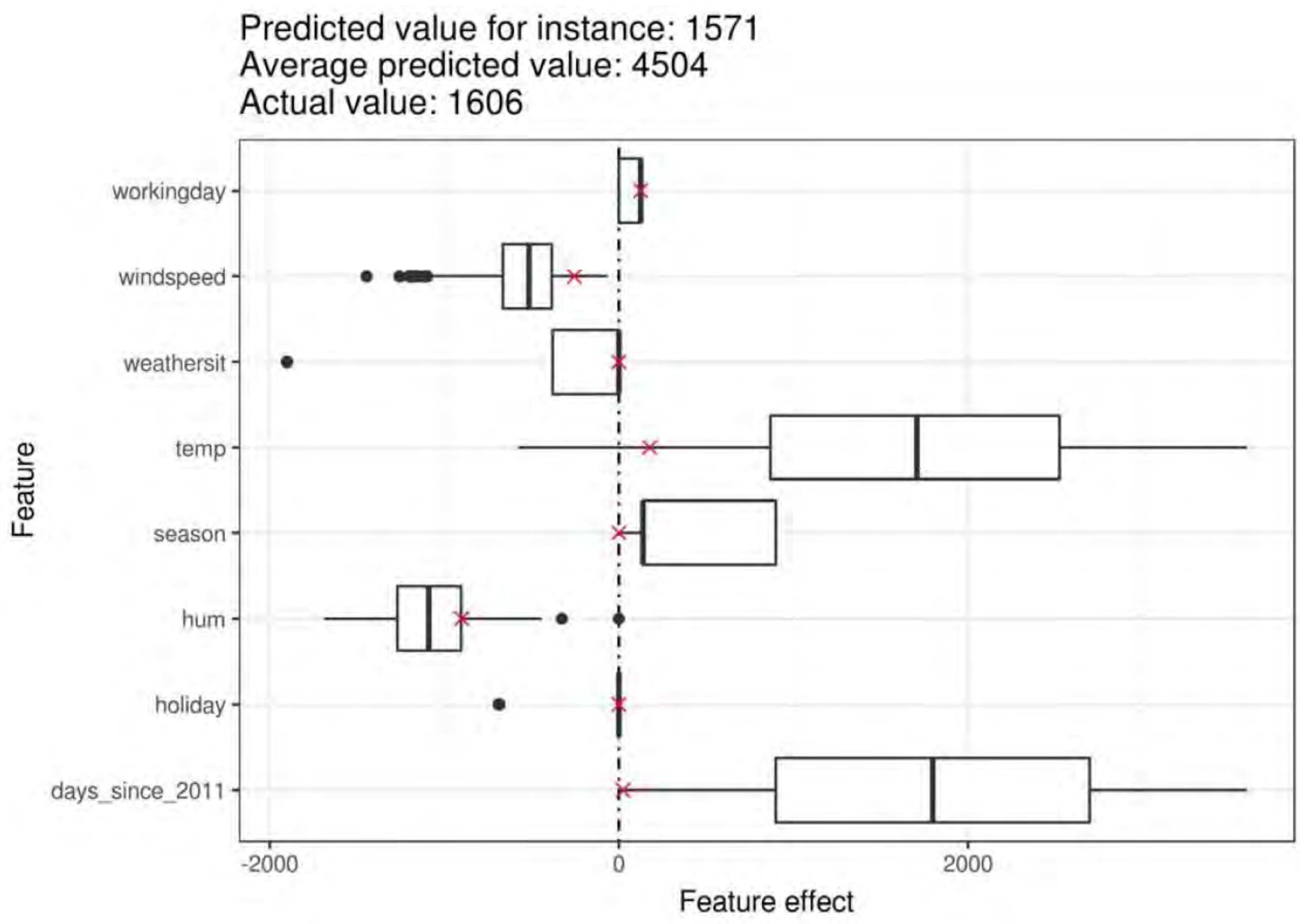
Linear Regression

- Example: bike rental:
  - Plotting feature effects + 95% confidence interval
  - Weight multiplied by feature values
  - Box-plot: Median, effect range 25%-75% of data, outliers



Linear Regression

- Example: bike rental:
  - Explaining single prediction (instance 6: early 2011, 2ºC



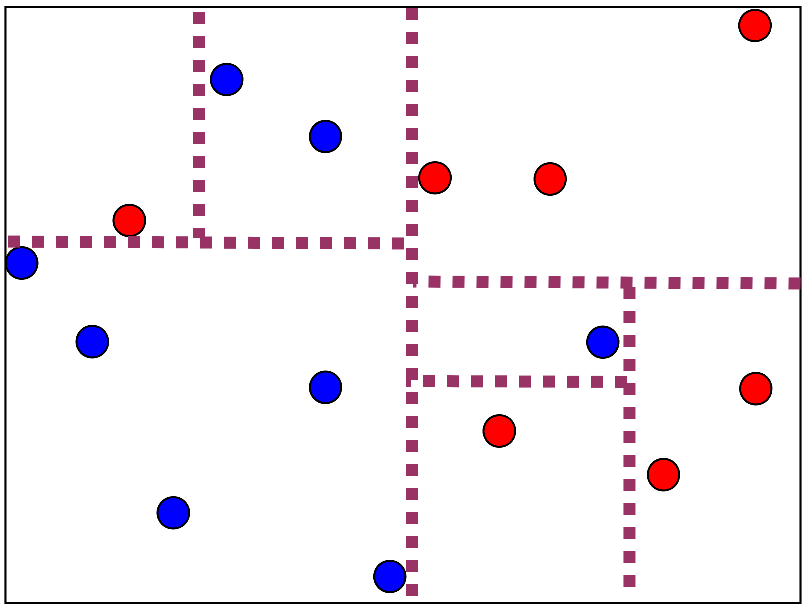
Linear Regression

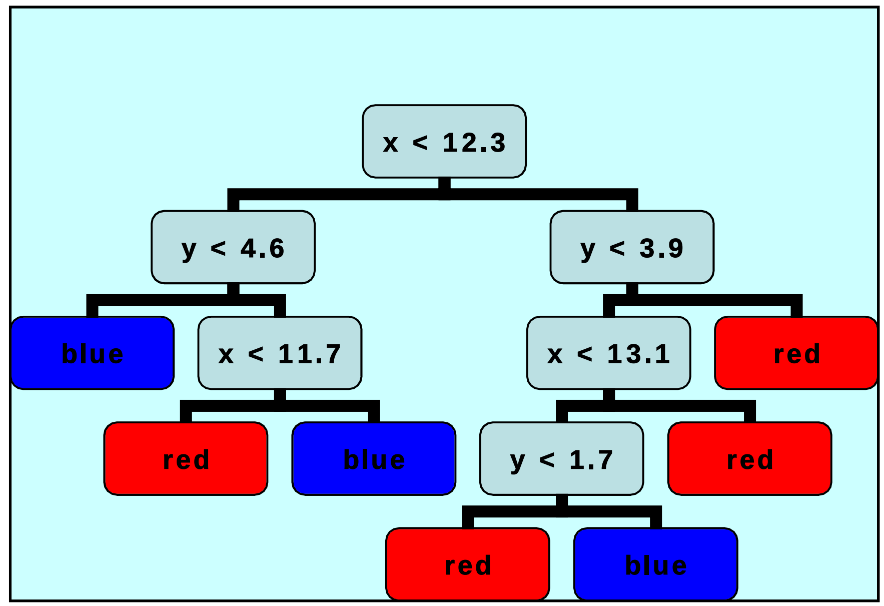
- Summary
  - Weighted sums are simple, highly transparent
  - High level of acceptance, experience, solid statistical theory
  - Only for linear relationships
  - Non-linearities have to be modelled as features
  - Low performance as many settings non-linear relationships
  - Unintuitive interpretation because of independence assumption
    (doesn’t hold in real world, e.g. size of house / nr of rooms)

Decision Trees

- Different algorithms
- Binary / non-binary splits
- Different splitting criteria
- Assigns each instance via branches to one leaf node
- Can be interpreted as rule set

Decision Trees



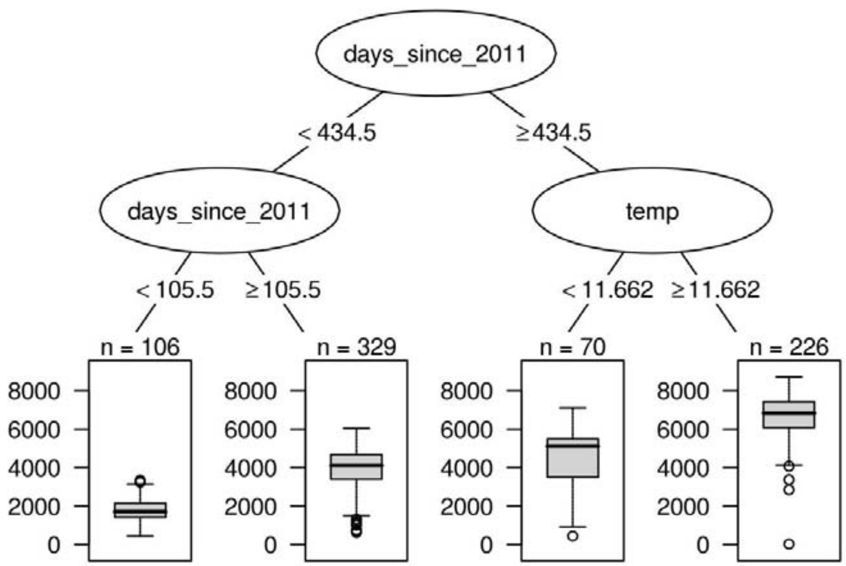


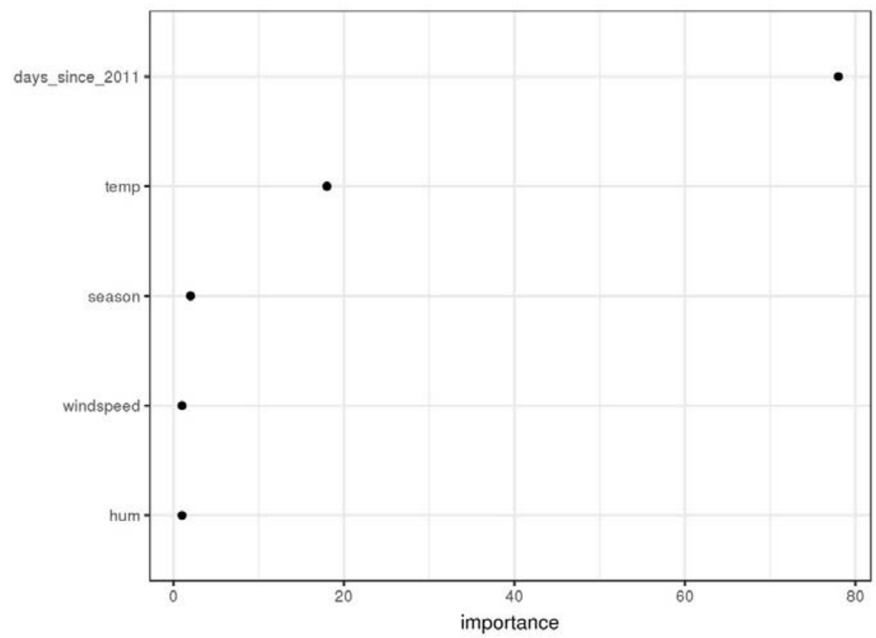
Decision Trees

- Interpretation
  - Reason for decision:
    - Rule set, sequence of decisions
    - Local explanation
    - Global explanation: usually too complex to grasp!
  - Feature importance:
    - All splits in which feature was used, compute contribution to quality measure (variance, Gini index, …)
    - Scale to 100%: share of each feature in decision

Decision Trees

- Example: bike rental (regression tree)
  - Splits plus variance in leaves
  - Feature importance: time trend higher than temperature





Decision Trees

- Summary
  - Captures interaction between data
  - Natural structure, visualization, intuitive
  - Allows identification of counterfactuals: “if x had been y”
  - No scaling needed
  - Not suitable for linear relationships (step-functions)
  - No smoothness –> small changes, big effects
  - Unstable: small changes in data, big effects
  - Complex for real-world settings

Outline

- What is Explainability in ML and why do we need it?
- Interpretable Models
- Model-Agnostic Approaches to Explainability
  - Partial Dependence Plots (PDP)
  - Accumulated Local Effects (ALE)
  - Local surrogate (LIME)
  - Shapley Values

Partial Dependence Plots

- Friedman, Jerome H. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232.
- Marginal effect one or two features x_S have on the predicted outcome of a machine learning model
- Estimated by calculating averages in the training data (Monte Carlo method)

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

Partial Dependence Plots

- Computation:
  1) Select feature
  2) Define grid
  3) Per grid value:
     1) replace feature with grid value and
     2) average predictions.
  4) Draw curve

Partial Dependence Plots

- Example: https://towardsdatascience.com/introducing-pdpbox-2aa820afd312
  Data set with 3 instances and 3 attributes & class Y

A	B	C	Y
A1	B1	C1	Y1
A2	B2	C2	Y2
A3	B3	C3	Y3

- Analyzing contribution of attribute A on prediction Y: generate new data set with all combinations of attributes

A	B	C	Y
A1	B1	C1	Y11
A1	B2	C2	Y21
A1	B3	C3	Y31
A2	B1	C1	Y12
A2	B2	C2	Y22
A2	B3	C3	Y32
A3	B1	C1	Y13
A3	B2	C2	Y23
A3	B3	C3	Y33

Partial Dependence Plots

- Generate **nrows * num_grid_points** number of predictions and averaged them for each unique value of Feature A

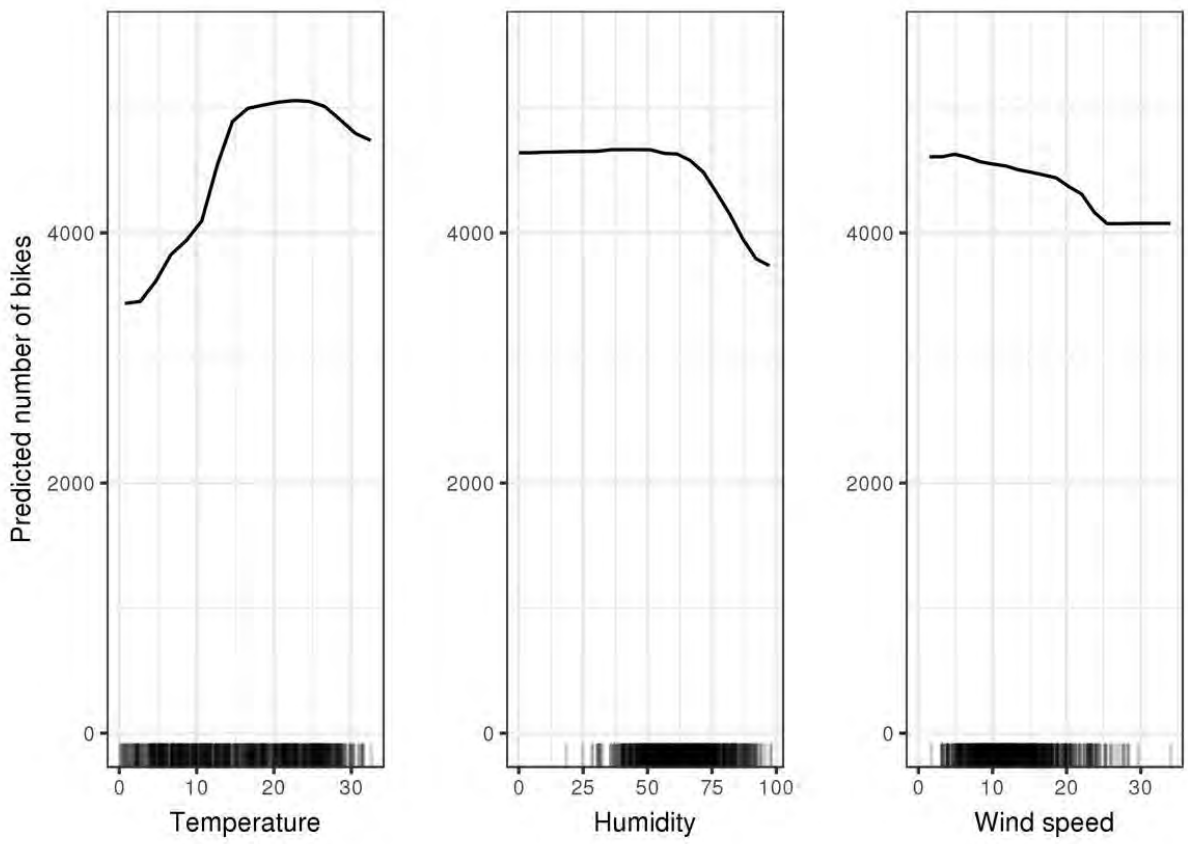
A	B	C	Y	mean
A1	B1	C1	Y11	Y(A1)
A1	B2	C2	Y21	
A1	B3	C3	Y31	
A2	B1	C1	Y12	Y(A2)
A2	B2	C2	Y22	
A2	B3	C3	Y32	
A3	B1	C1	Y13	Y(A3)
A3	B2	C2	Y23	
A3	B3	C3	Y33	

- Plot average predictions for each feature value of A

X	A1	A2	A3
Y	Y(A1)	Y(A2)	Y(A3)

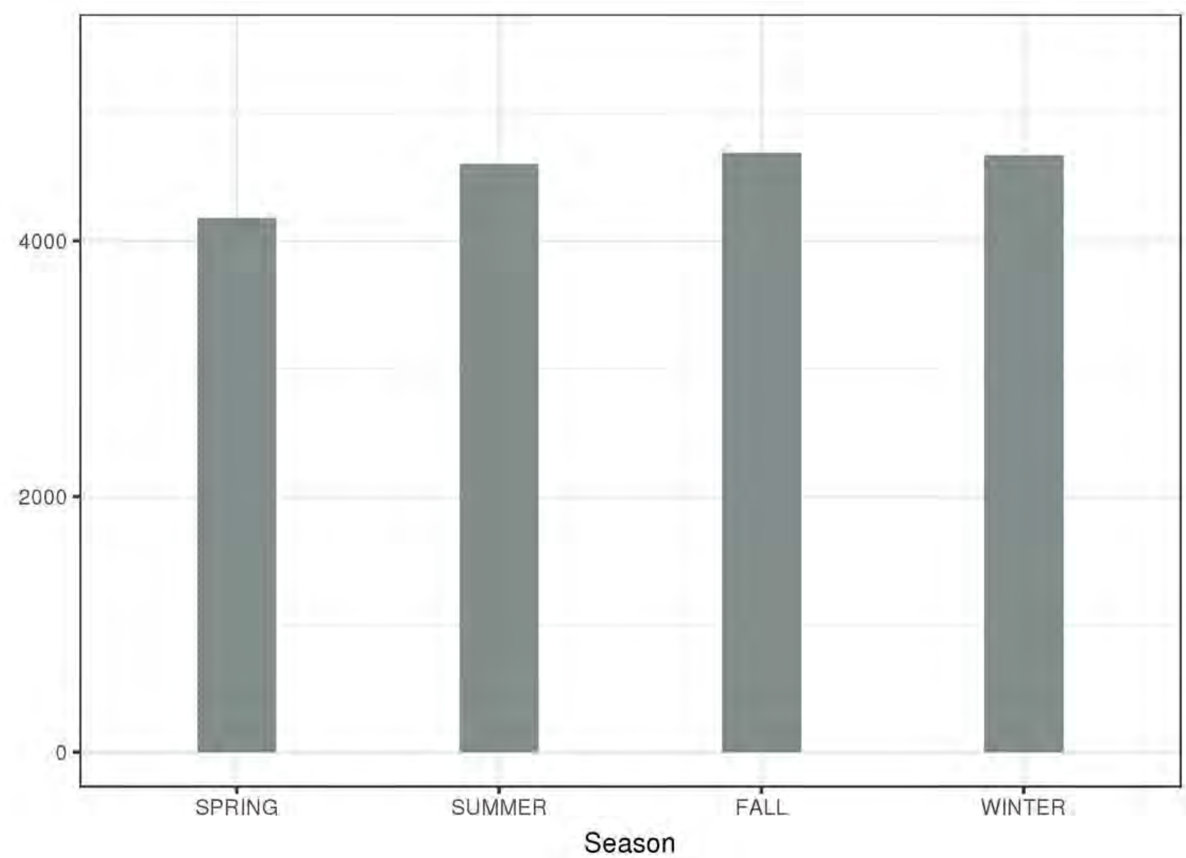
Partial Dependence Plots

- Example: bicycle rental (note histograms)



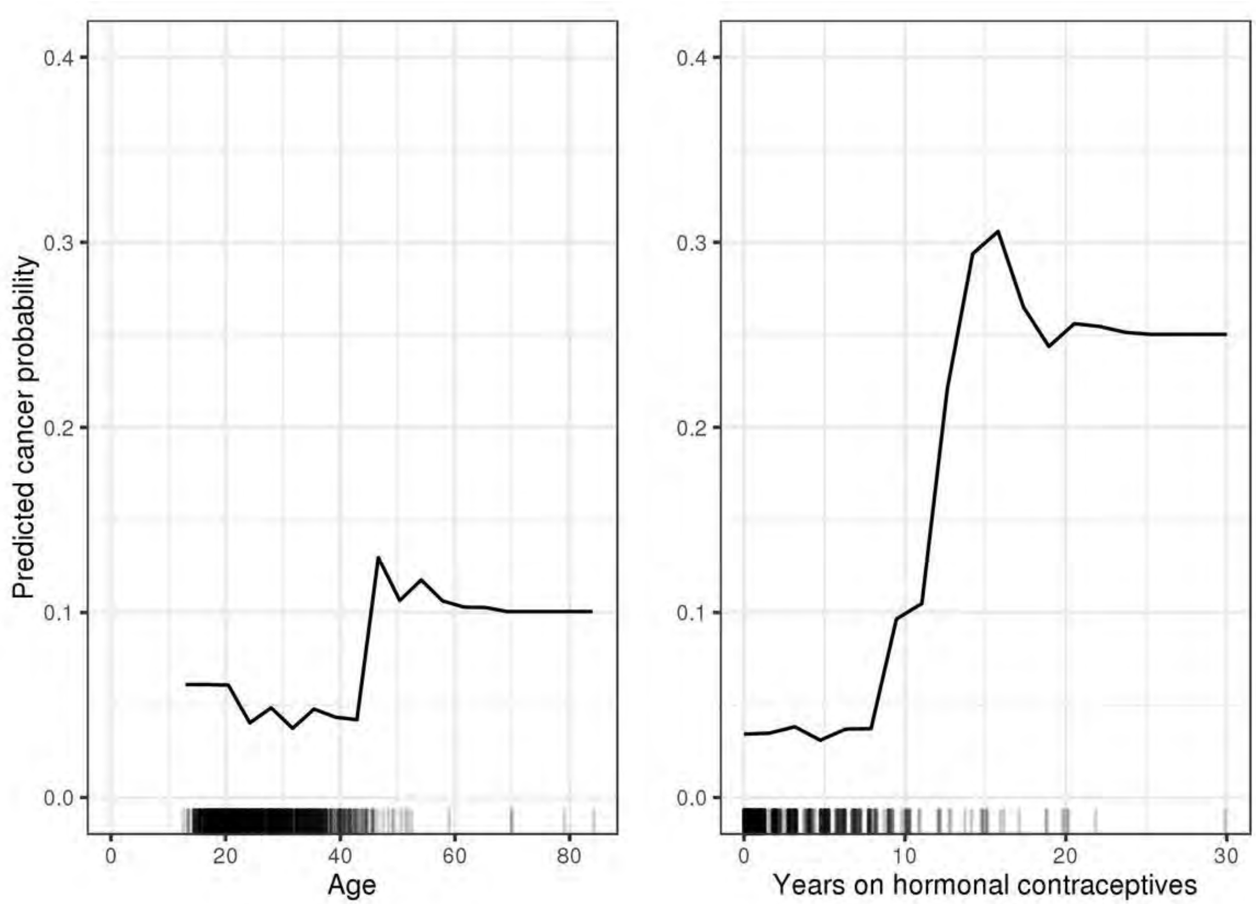
Partial Dependence Plots

- Example: bicycle rental (categorical features)



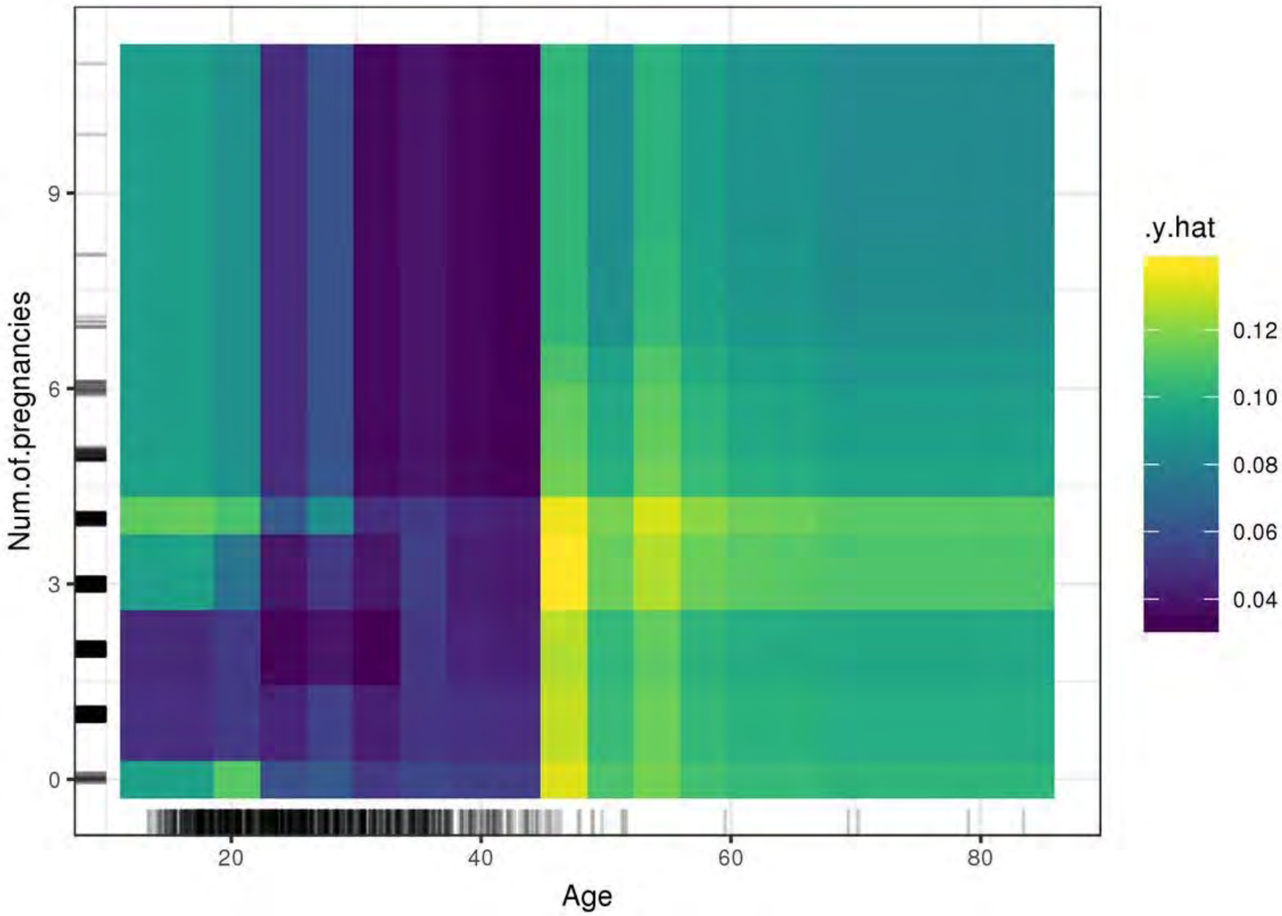
Partial Dependence Plots

- Example: Cervical Cancer (note histograms)



Partial Dependence Plots

- Example: Cervical Cancer (2 attributes)



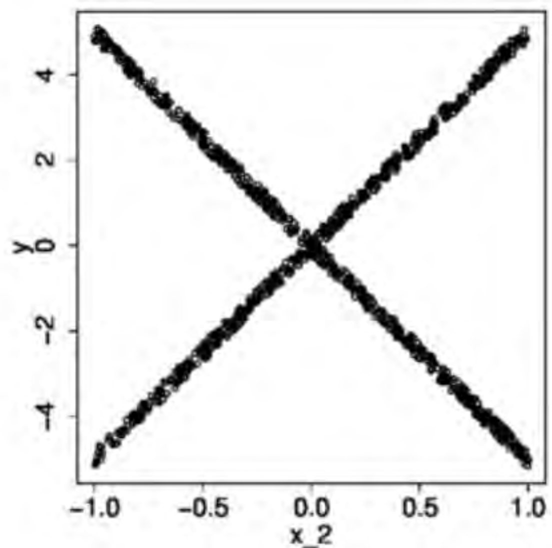
Partial Dependence Plots

- Summary
  - Intuitive
  - Provides causal interpretation of feature by the model
  - If features are not correlated, then perfect representation of feature influence
  - But: features are usually correlated
  - PDP computes over unrealistic feature combinations
    (30m2 flat with 10 rooms;
    person of 1,90m with weights btw. 45-120kg)
  - Heterogeneous effects may be hidden (PDP show average marginal effect – if half of data points have positive association, the other half negative, then zero effect is reported) -> Individual Conditional Expectation (ICE)

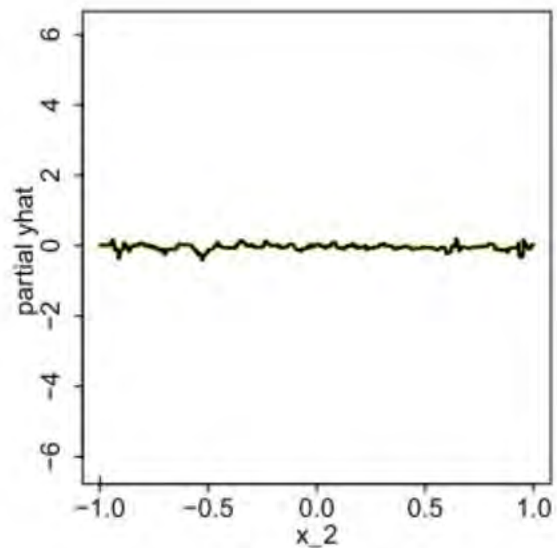
Individual Conditional Expectation (ICE)

- Like PDP, but
- Plot each data point separately instead of plotting averages

X	A1	A2	A3
Y1	Y11	Y12	Y13
Y2	Y21	Y22	Y23
Y3	Y31	Y32	Y33



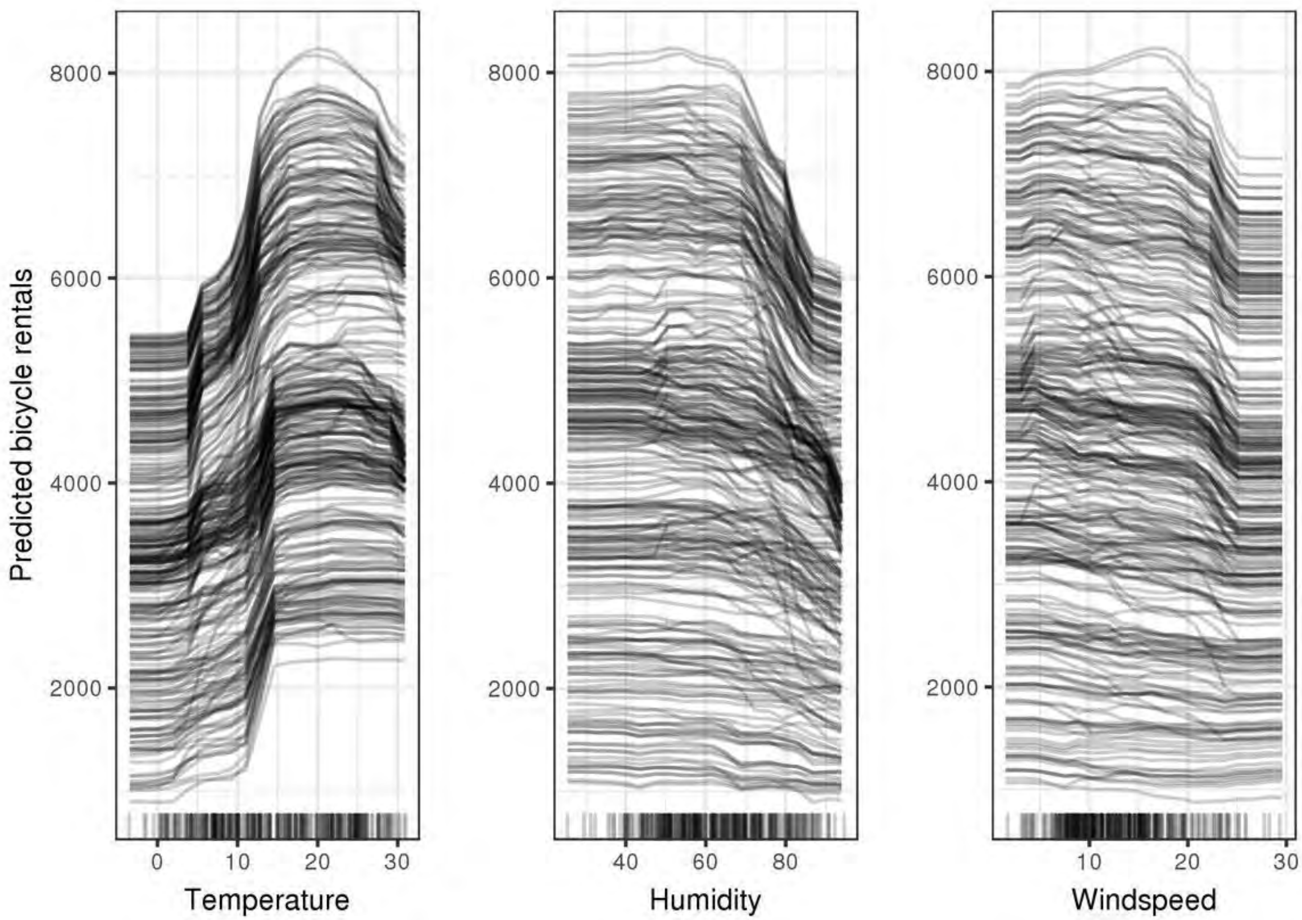
(a) Scatterplot of Y versus X_2



(b) PDP

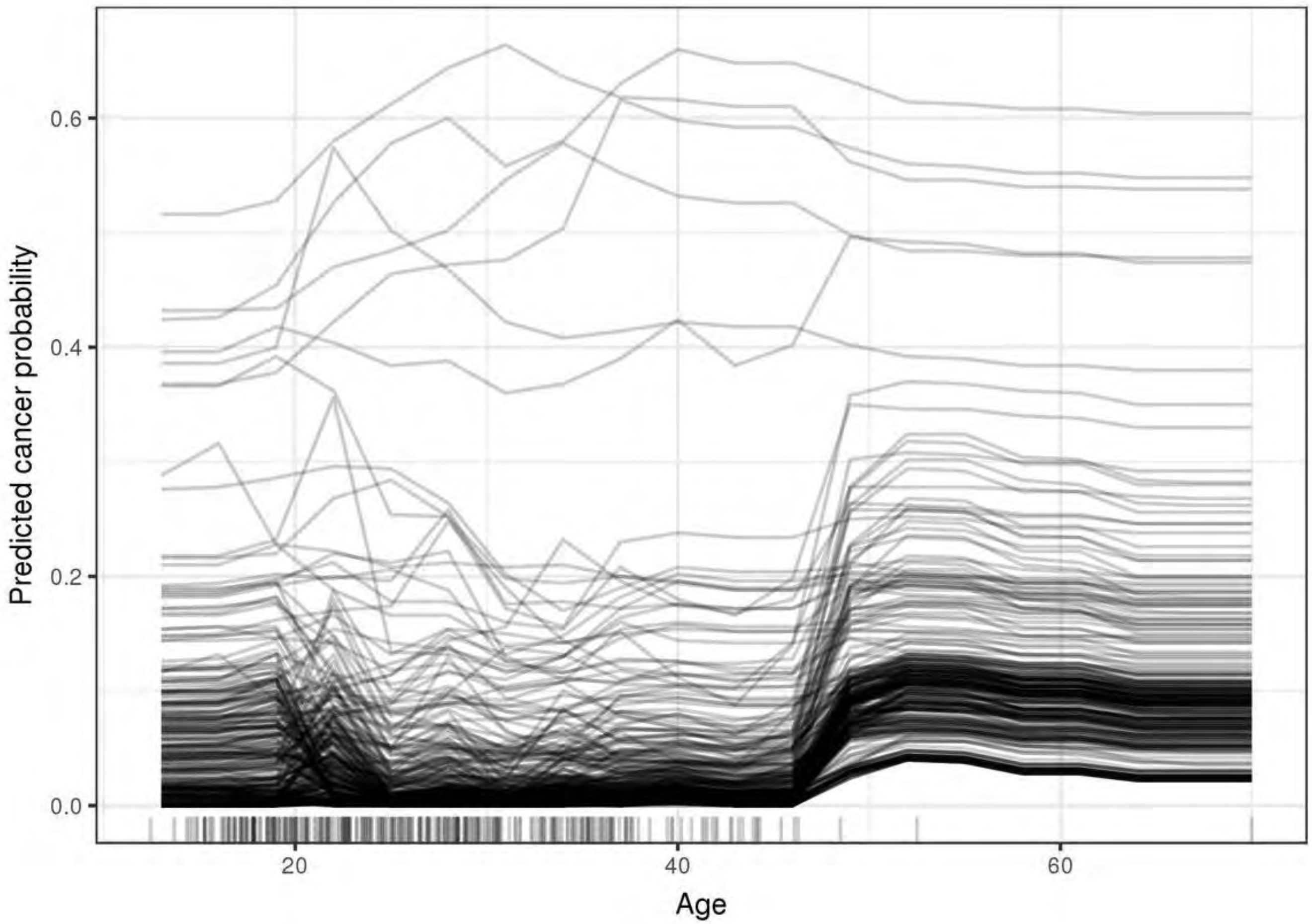
Individual Conditional Expectation (ICE)

- Example: bike rental



Individual Conditional Expectation (ICE)

- Example: cervical cancer



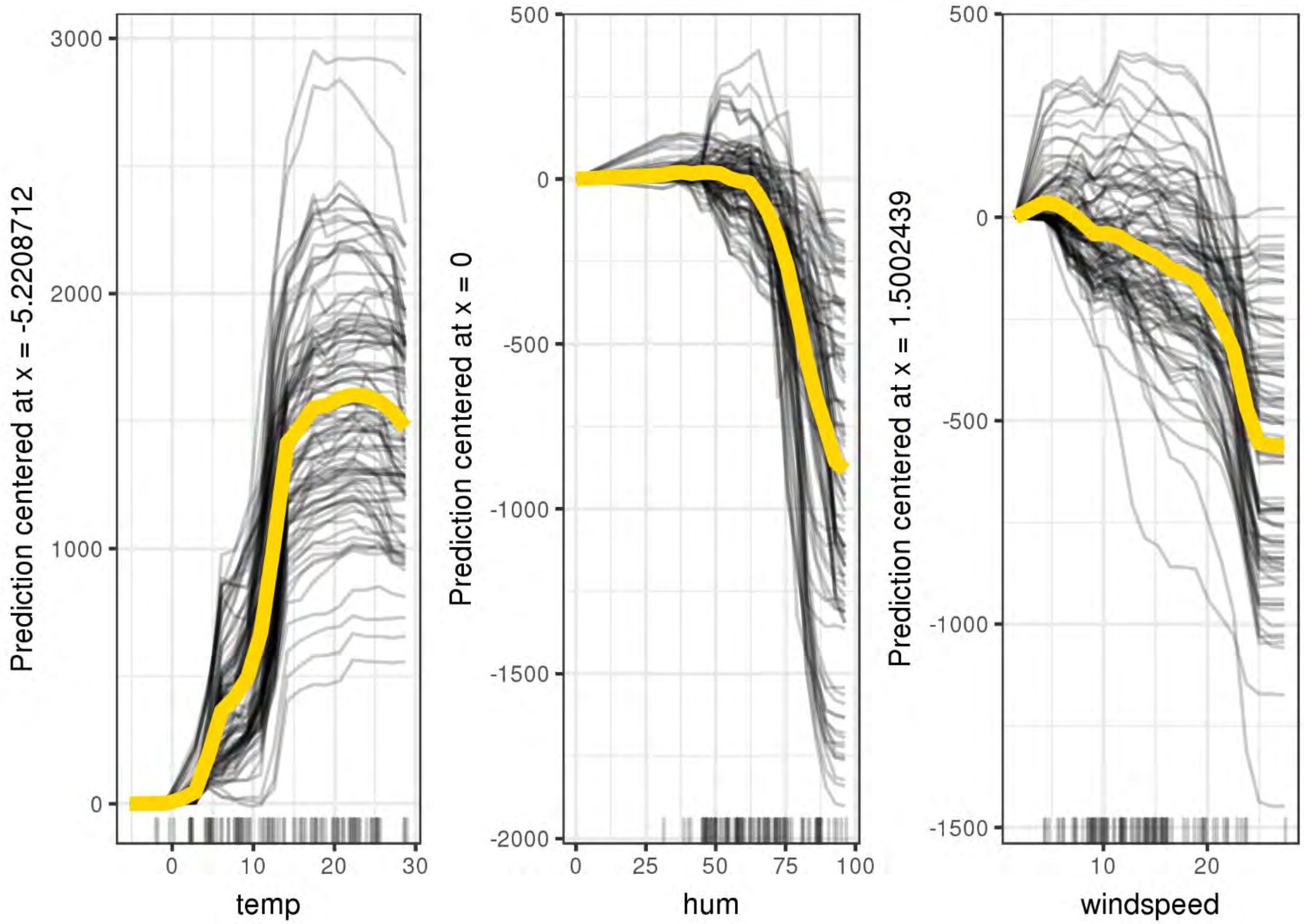
Individual Conditional Expectation (ICE)

- Centered ICE curve
  - ICE curve shows absolute variation
  - Interested in difference as value changes
  - Anchoring curve at certain (lower end) *i* of value range of attribute

$$\hat{f}^{(i)}_{cent} = \hat{f}^{(i)} - \mathbf{1}\hat{f}\,(x^a, x_C^{(i)})$$

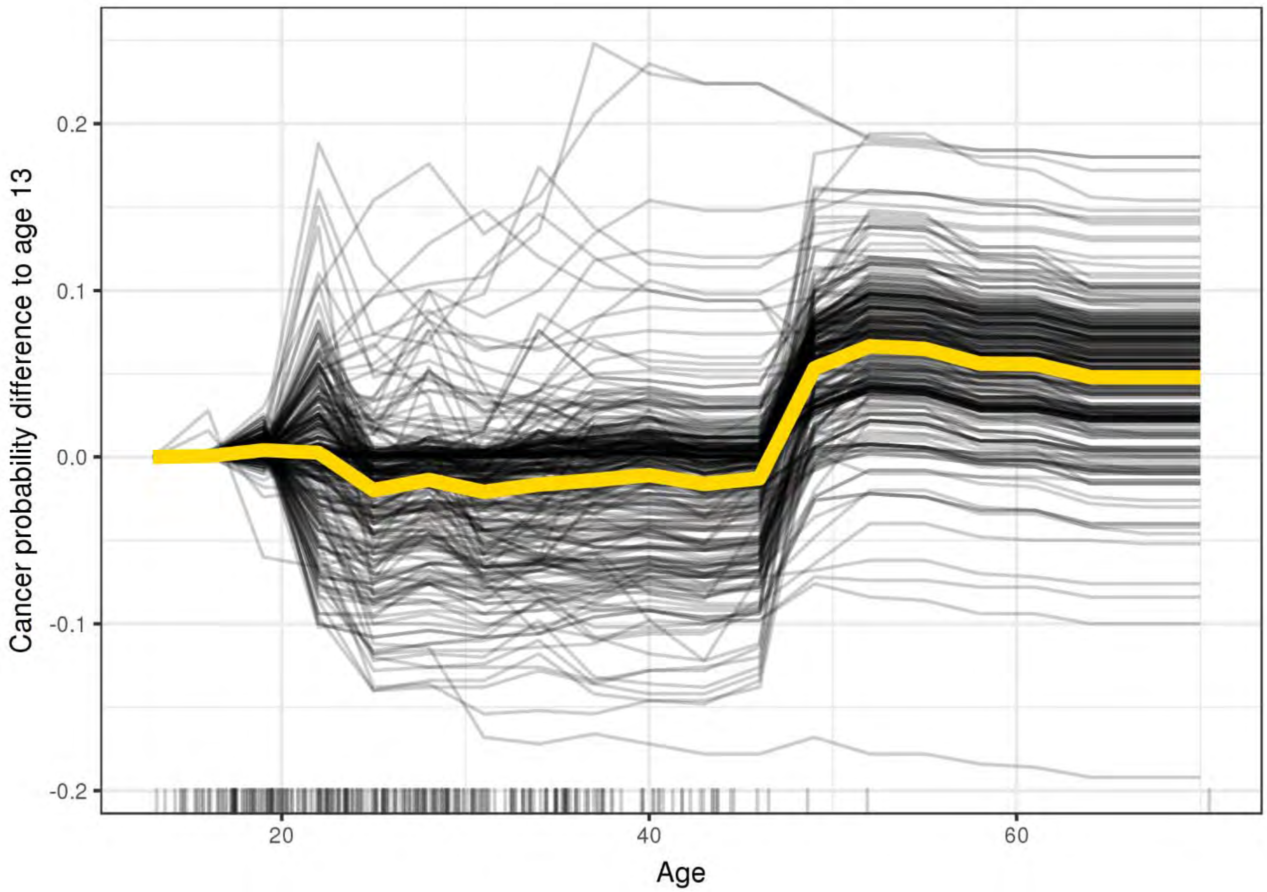
Individual Conditional Expectation (ICE)

- Example: bicycle rental



Individual Conditional Expectation (ICE)

- Example: cervical cancer



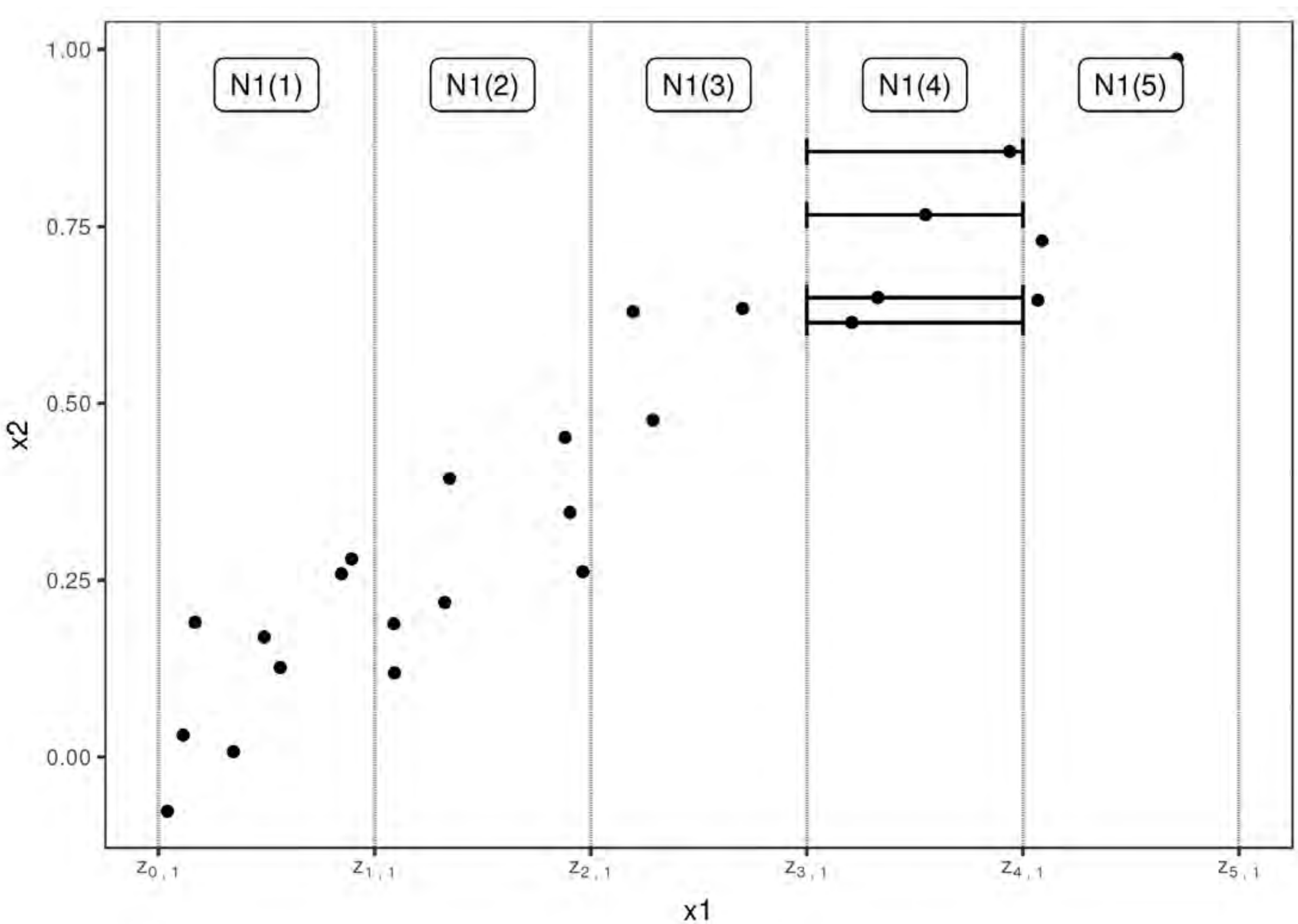
Individual Conditional Expectation (ICE)

- Summary
  - Clearer representation of actual distribution of feature contributions to prediction
  - Only for individual attributes, one at a time (no 2d-plots)
  - Still suffers from correlation btw. attributes: unrealistic combinations

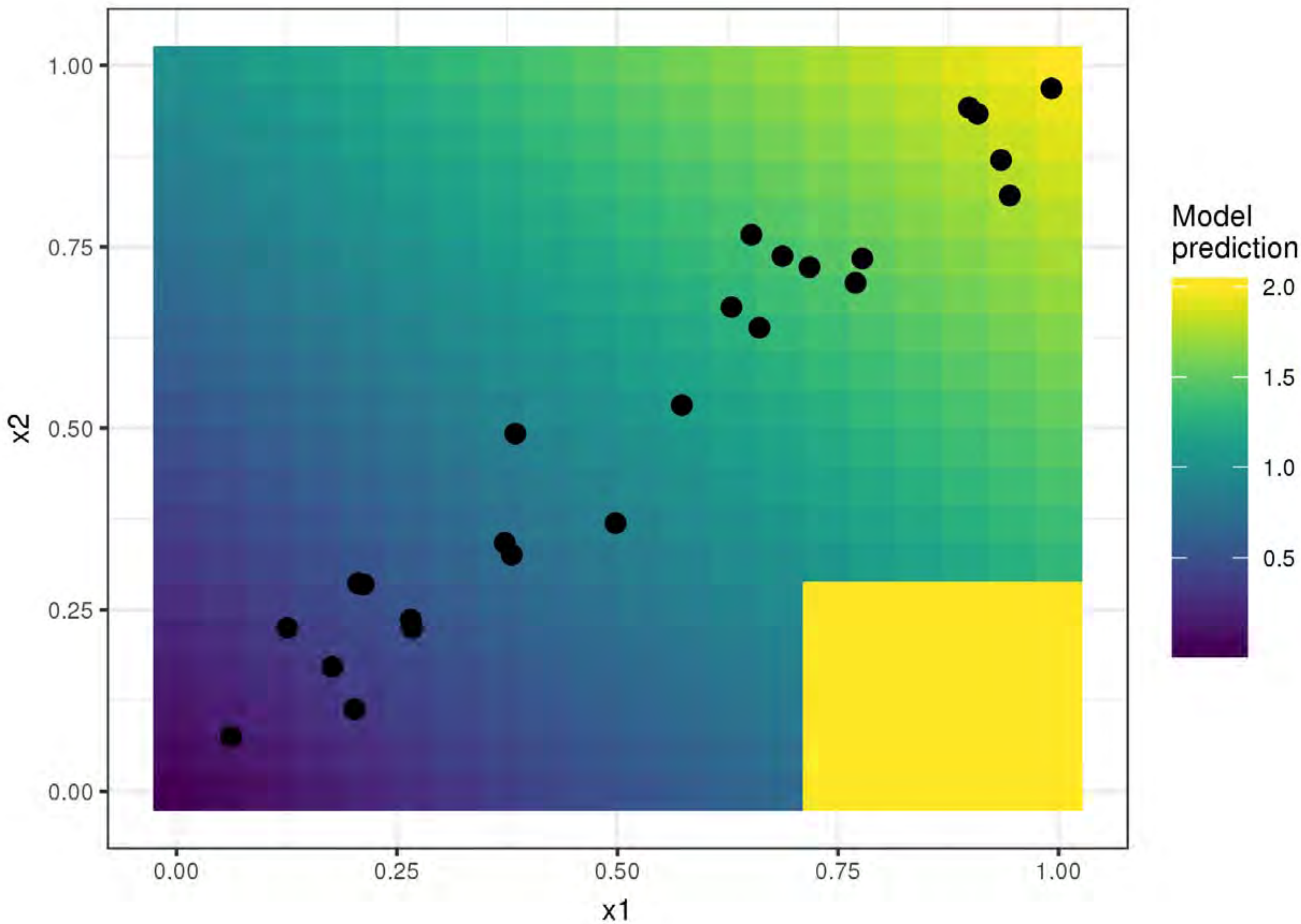
Accumulated Local Effects Plot (ALE)

- Overcomes feature dependency issue of PDP plots
- M-plots: average over the conditional distribution of the feature, meaning at a grid value of x1, we average the predictions of instances with a similar x1 value
- ALE plots: differences in predictions instead of averages
  1. divide the feature into intervals (vertical lines).
  2. for data instances in an interval, calculate difference in prediction when we replace the feature with the upper and lower limit of the interval (horizontal lines).
  3. differences are later accumulated and centered, resulting in the ALE curve

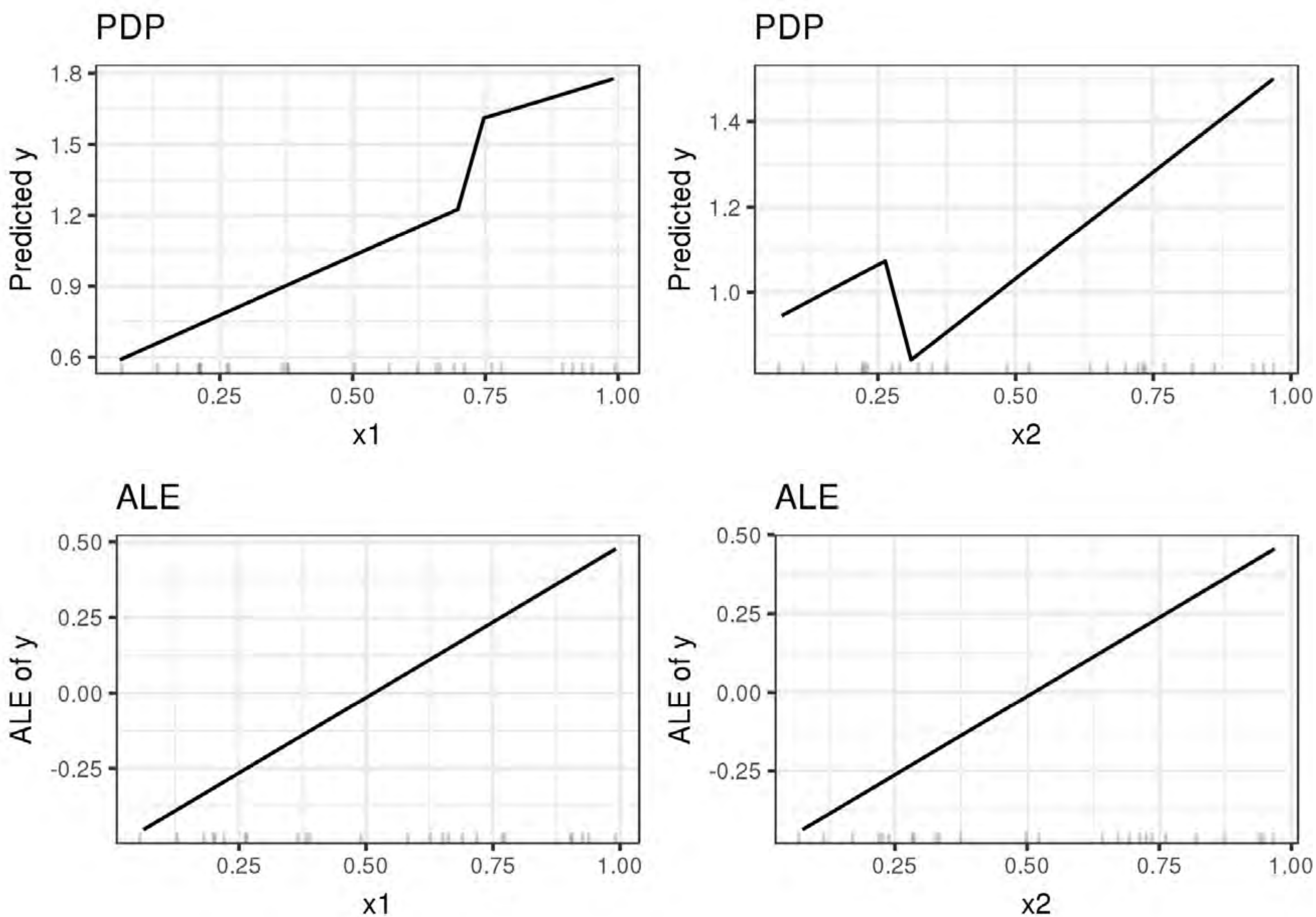
Accumulated Local Effects Plot (ALE)



Accumulated Local Effects Plot (ALE)

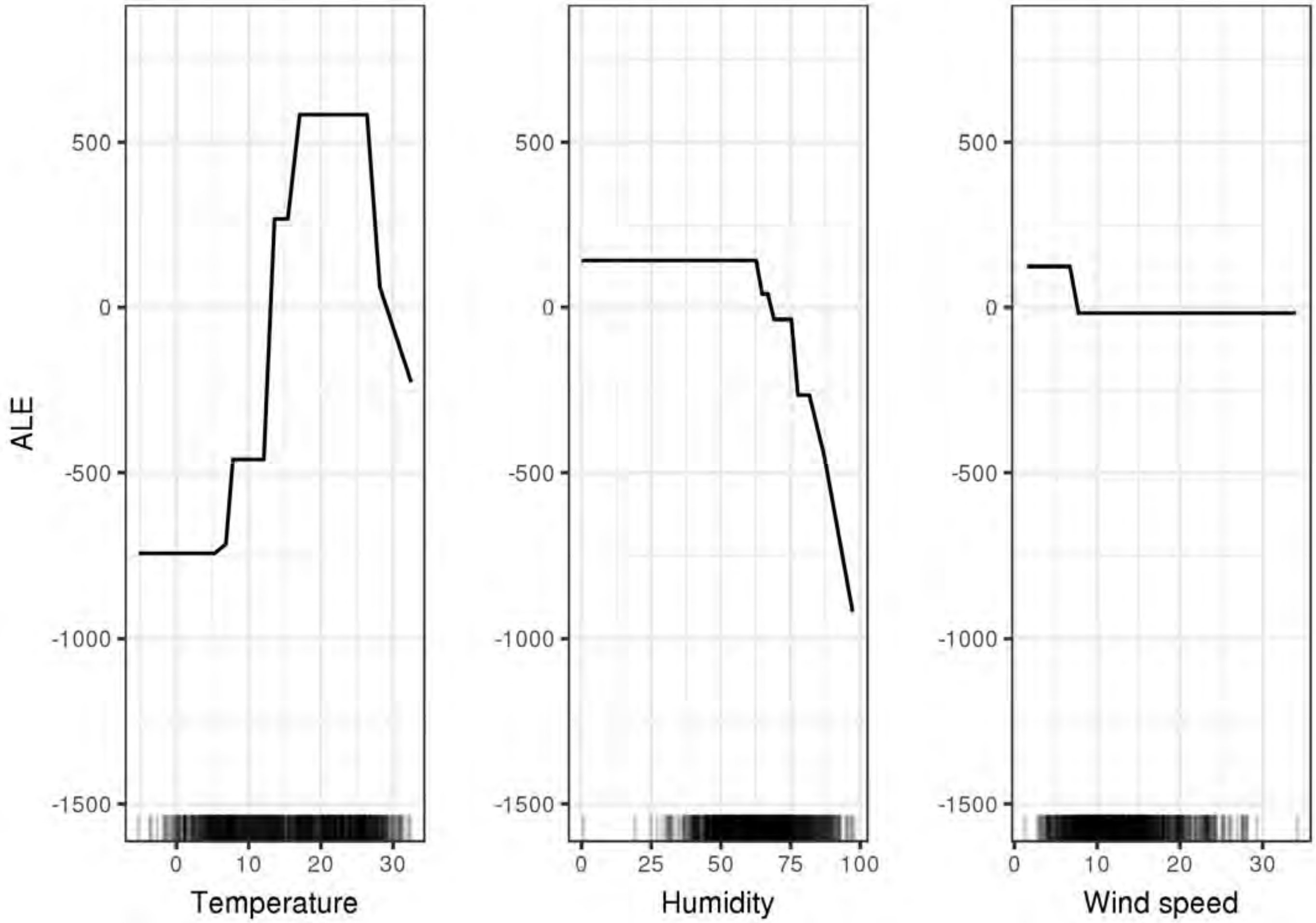


Accumulated Local Effects Plot (ALE)



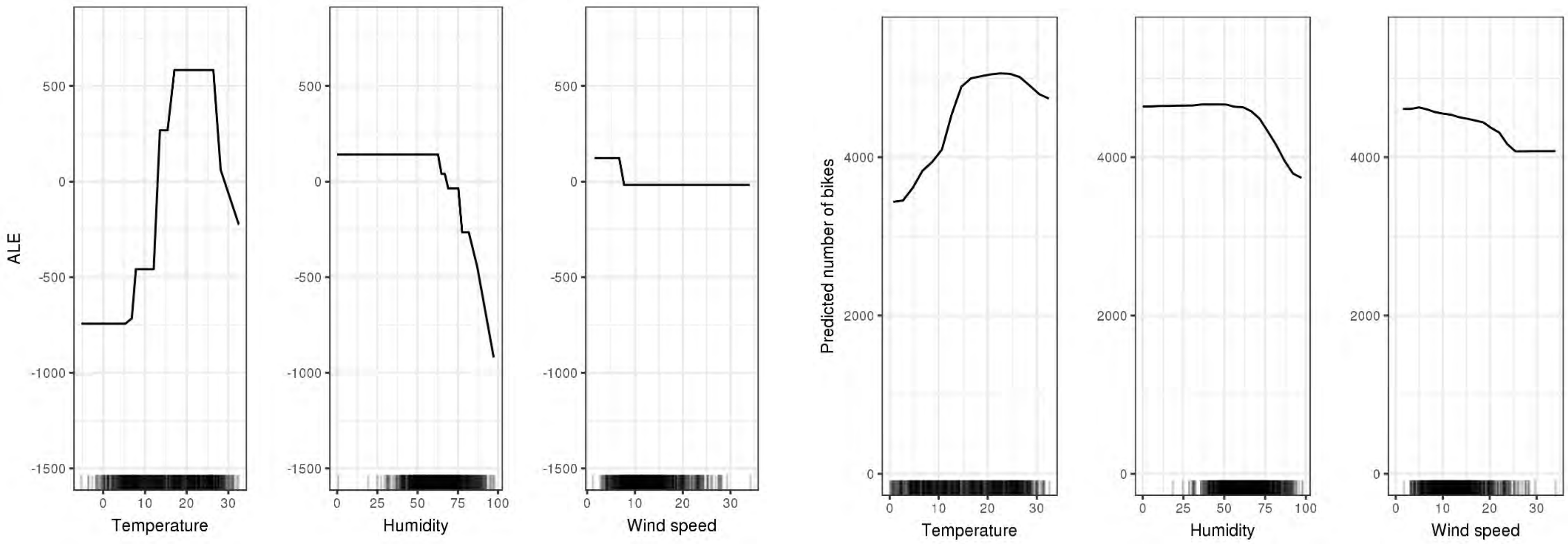
Accumulated Local Effects Plot (ALE)

- Example: bike rental



Accumulated Local Effects Plot (ALE)

- Example: bike rental: ALE vs. PDP



Accumulated Local Effects Plot (ALE)

- Summary
  - Unbiased across attribute correlations
  - Faster to compute: O(n) (nr. Intervals, max # data points)
  - Centered at 0, easy interpretation
  - Shaky with high number of intervals
  - No guidance of how many intervals to choose
  - No ICE curve equivalent to understand heterogeneous contributions

Global Surrogate Model

- Train interpretable model on predictions provided by black-box model:
  - Select a dataset X (same dataset as used for training the black box model or a new dataset from the same distribution)
  - Get the predictions of the black box model.
  - Select an interpretable model type (linear model, decision tree, …)
  - Train interpretable model on dataset X and its predictions
  - Measure how well the surrogate model replicates the predictions of the black box model
  - Interpret the surrogate model

Global Surrogate Model

- Summary
  - Flexible, works across all models, straightforward
  - Quality of surrogate model measured against its prediction of the original black-box model, not the ground truth labels!
  - Surrogate model quality may vary across data space
  - Limitations of interpretable models apply

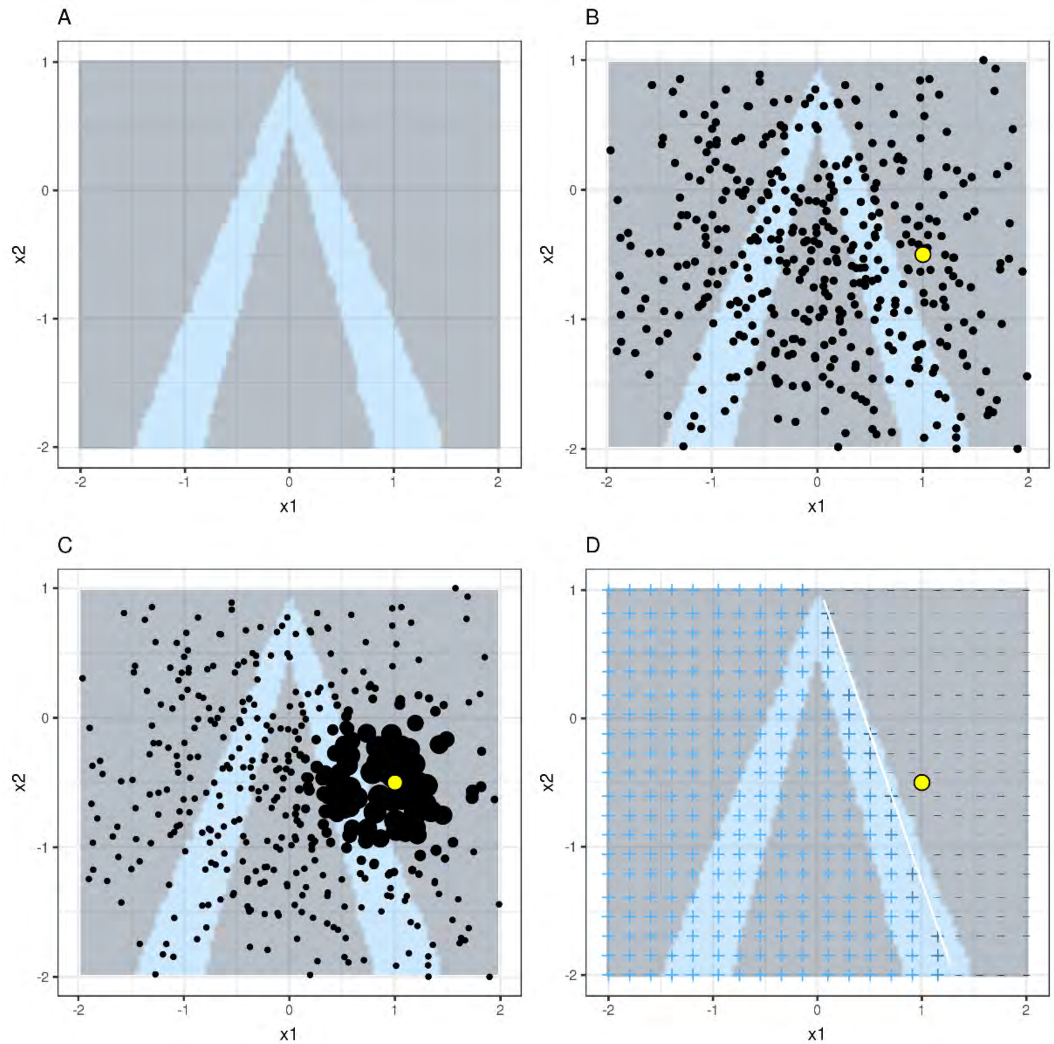
LIME

- Local interpretable model-agnostic explanations (LIME)
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why should I trust you?: Explaining the predictions of any classifier.” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)
- Around a data point of interest generate a new dataset consisting of permuted samples and the corresponding predictions of the black box model
- Train interpretable model on this data set

LIME

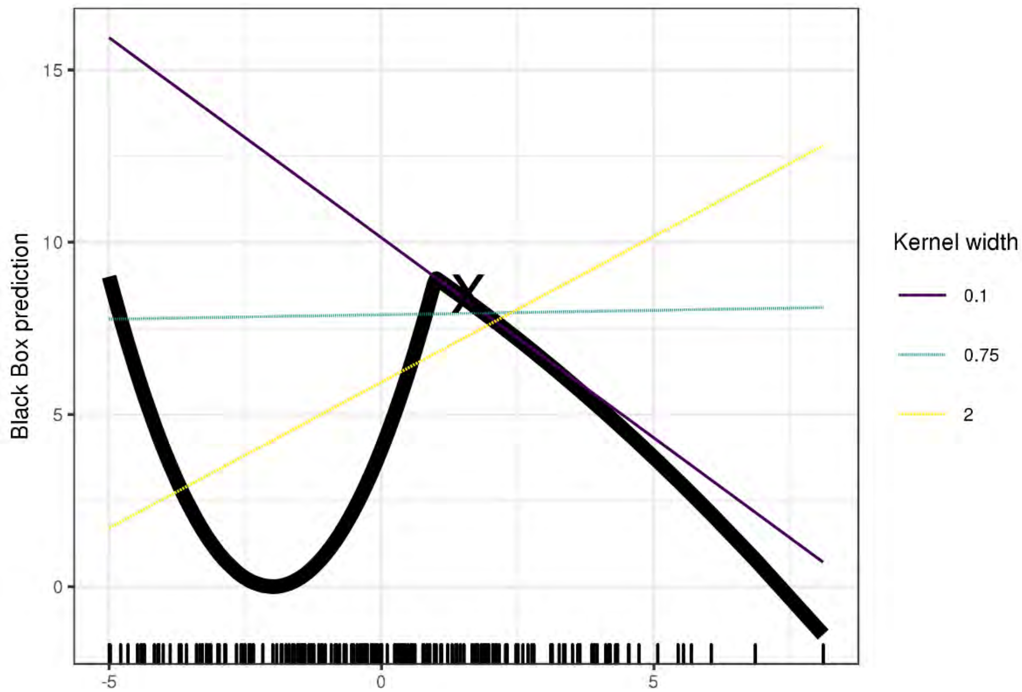
- Computation
  - Select instance of interest for which you want to have an explanation of its black box prediction
  - Perturb your dataset and get the black box predictions for these new points
  - Weight the new samples according to their proximity to the instance of interest
  - Train a weighted, interpretable model on the dataset with the variations
  - Explain the prediction by interpreting the local model
- Challenge: defining the perturbation neighborhood, wich influences the locality of the explanation

LIME



LIME

- Challenge: dertermining the perturbation neighborhood
- Example:
  - Black line: black-box prediction
  - Surrogate models (lin. Regr.) with 3 different kernel sizes



LIME

- Summary
  - Flexibly use any specific surrogate model
  - Fidelity measure provides information on how well the surrogate emodel explains the black-box model
  - Neighborhood kernel size is decisive and hard to estimate
  - Sampling within neighborhood kernel usually based on Gaussian, ignoring correlation btw. Attributes
  - Low stability in explanations for neighboring data points

Shapley Values

- Game-theoretic approach
- What is the contribution of each feature value to the prediction?
- Shapley, Lloyd S. “A value for n-person games.” Contributions to the Theory of Games 2.28 (1953): 307-317
- Each attribute is a “player”
- Evaluate coalitions of all players to the final outcome

Shapley Values

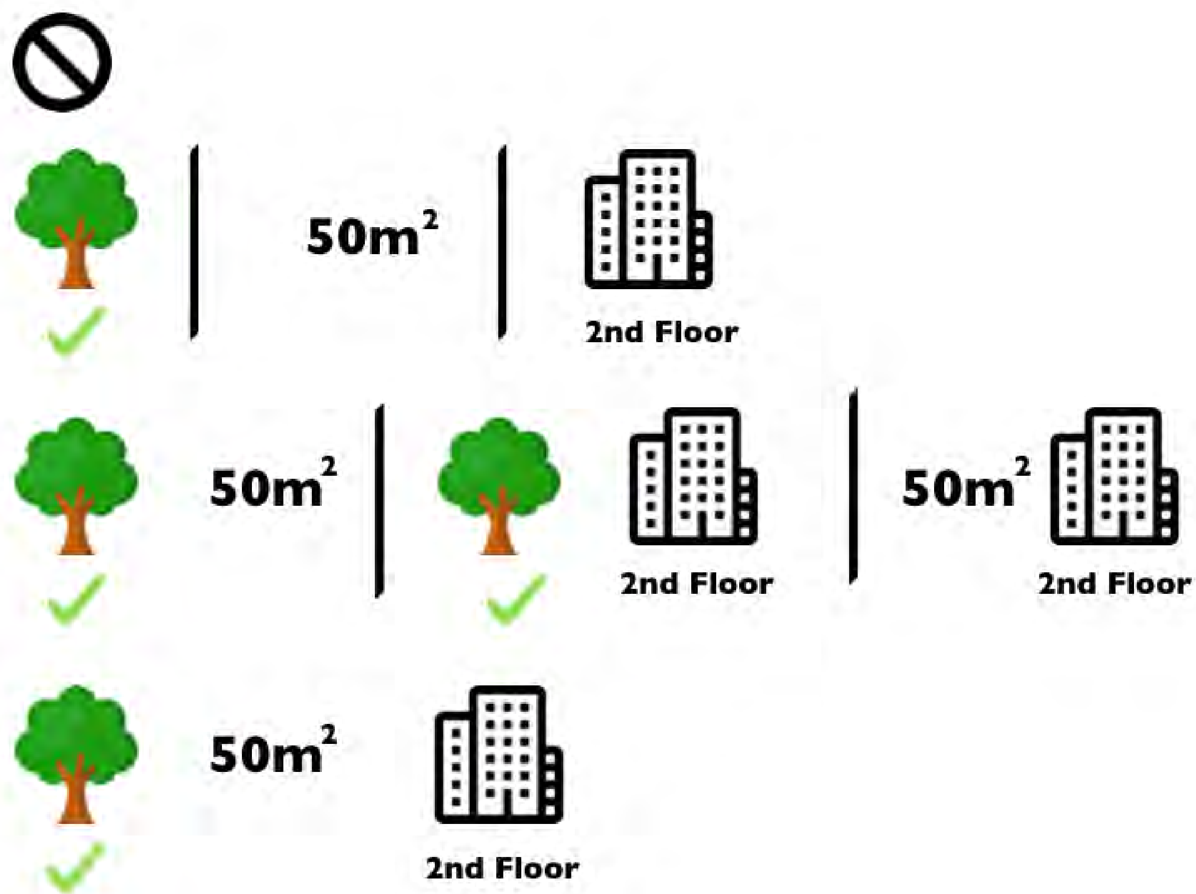
- Example: apartment price prediction
- Average price is 310.000
- For a specific instance, the predicted price is 300.000
- How much did each attribute contribute to increase or lower the price? (easy for linear regression models)

Shapley Values

- Contribution of one feature
  - Vary compared to instance of interest

Shapley Values

- Contribution of two features: cat banned in all permutations: 8 possibilities

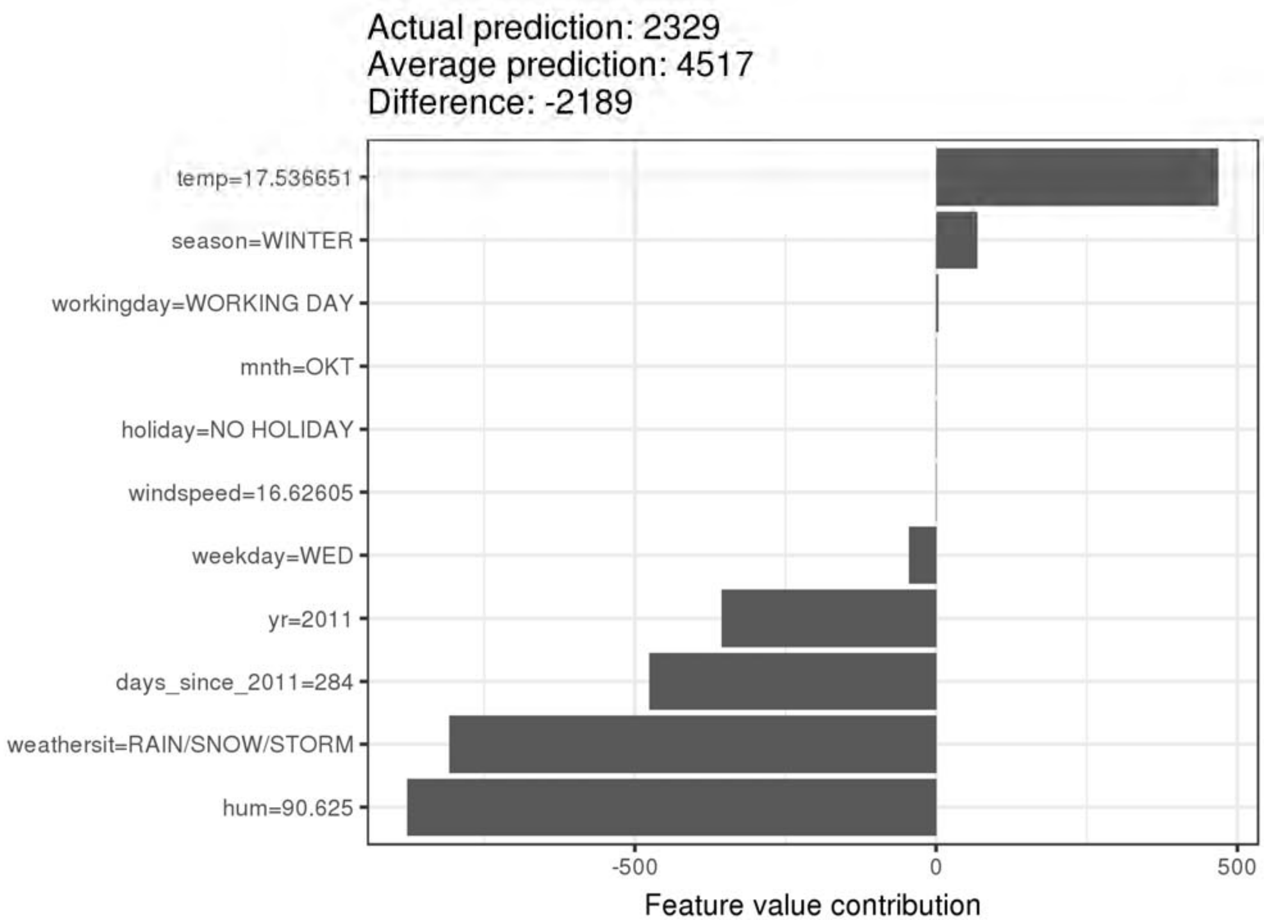


Shapley Values

- Compute prediction for all combinations with attribute in question turned on or off
- Take the difference as the marginal contribution of the attribute in the specific coalition
- Take random feature values for features not in coalition
- Take average across all predictions obtained that way
- Interpretation: the value of feature j contributed ϕ_j to the prediction.

Shapley Values

- Example: bike rental, day 285



Shapley Values

- Summary
  - Provides fair, full prediction
  - Effects distributed / analyzed fairly across all coalitions
  - Might be only legally permissible explanation
  - Interpretation: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value
  - Expensive to compute:
    - 2k possible coalitions + m random samples for instances not present
    - Sample only some coalitions
    - Reduce the number of m random instances (increases variance)
  - Always uses all features, no sparse explanations

Shapley Values

- Summary
  - Needs access to data (not just black-box model) to replace non-present features by random samples
  - Cannot be used to make statements about changes in prediction for changes in the input (If I were to earn €300 more a year, my credit score would increase by 5 points)
  - Inclusion of unrealistic data instances when features are correlated

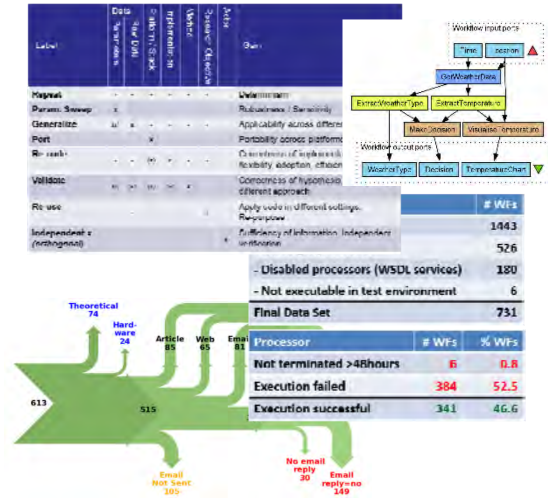
Outline

- What is Explainability in ML and why do we need it?
- Interpretable Models
- Model-Agnostic Approaches to Explainability
- Example-based Explanations
  - Counterfactual examples
  - Adversarial examples
  - Prototypes and Criticism
  - Influential instances

Outline

- Reproducibility
- Data Management & Citation
- Explainable AI
- Summary

Thank you!



Thanks!

https://rd-alliance.org/working-groups/data-citation-wg.html

RESEARCH DATA ALLIANCE

DC[1]
Data Citation Principles

docker

vmware

DMP ONLINE