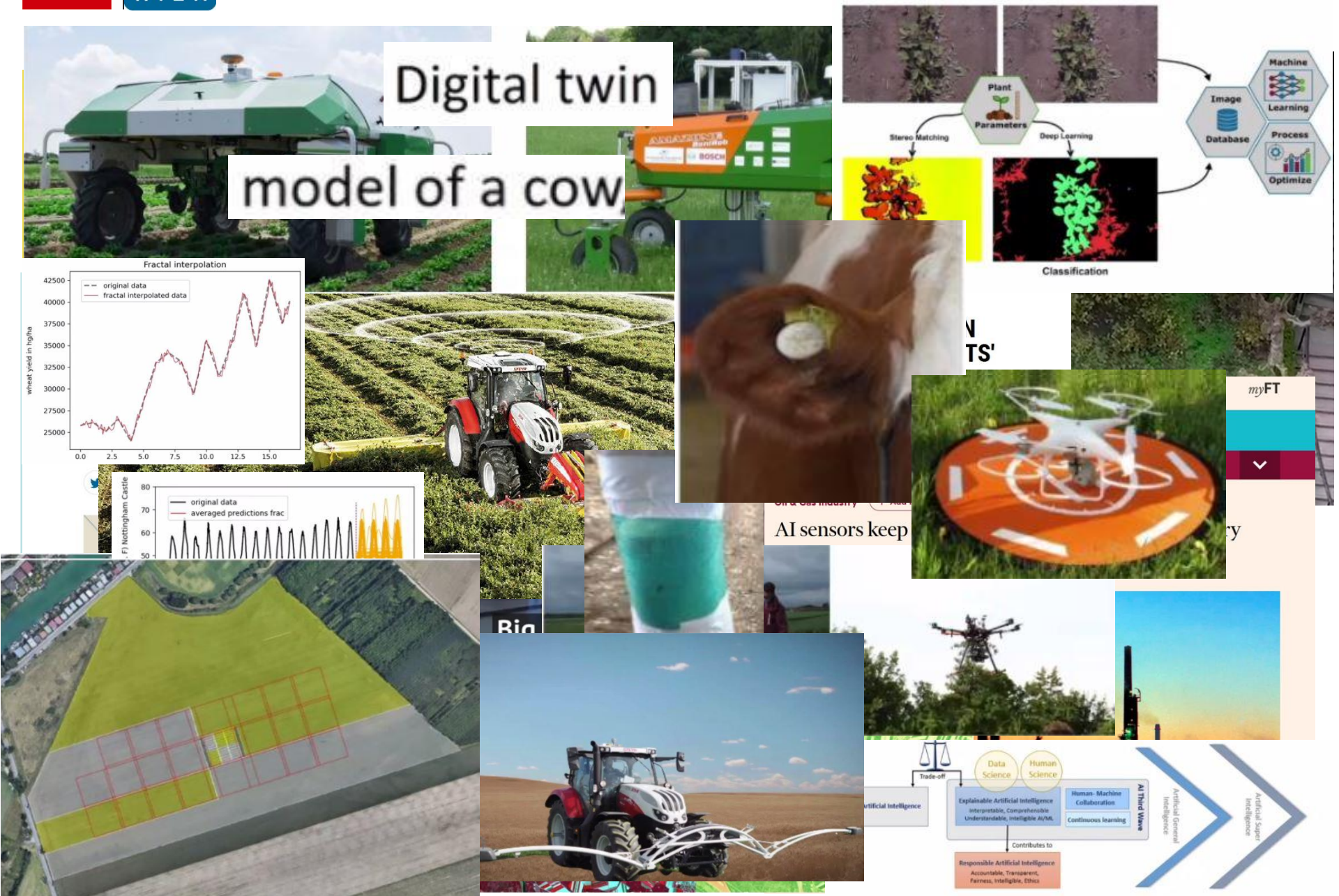
Reproduzierbarkeit im Bereich der Datenanalyse: Herausforderungen und Lösungen
---
Reproducibility in Data Analytics: Challenges and Solutions

Andreas Rauber

TU Wien
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
http://ww.ifs.tuwien.ac.at/~andi

FACULTY OF !NFORMATICS

Digitization and Big Data Promises



Digitization and Big Data Failures

The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Becam

BBC Sign in News

NEWS

Home Video World UK Business

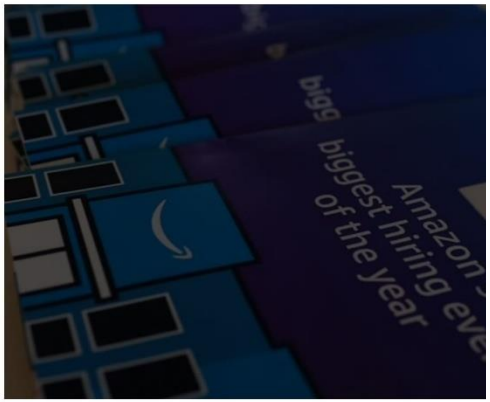
Technology

Fatal Tesla crash sp investigation

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN specialists uncovered a big problem: their new recruiting



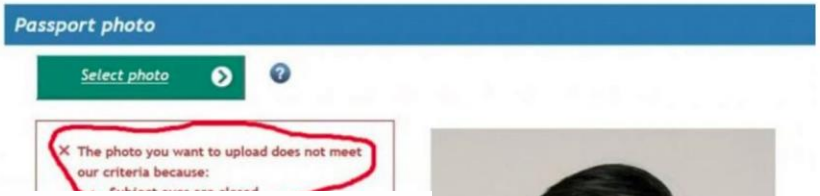
EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By CASEY ROSS @caseymross *and* IKE SWETLITZ / JULY 25, 2018

Robot passport checker rejects Asian man's application because "eyes are closed."



Apple Card is accused of gender bias. Here's how that can happen

By Evelina Nedlund, CNN Business

Updated 1904 GMT (0304 HKT) November 12, 2019

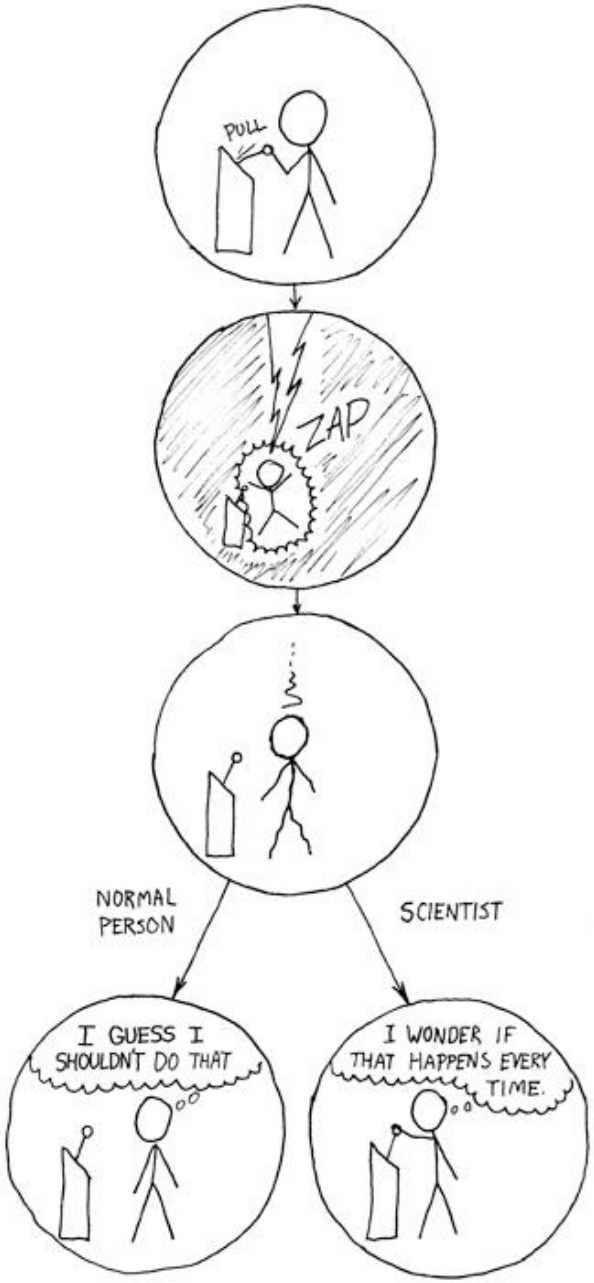
Image credit: Richard Lee | Facebook

Motivation

- Digitization promises huge benefits…
- … but also raises new challenges
- Need to avoid pitfalls
- Need to understand what is happening and why:
  - **Reproducibility**
  - Explainability

Reproducibility

- Reproducibility is core to the scientific method
- (and not just for science!)
- In computing should be easy (and thus also in digitization in agriculture):
  - Get the code, compile, run, …
  - Why is it difficult?



https://xkcd.com/242/

Outline

- What are the challenges in reproducibility?
- How to address the challenges of complex processes?
- How to deal with dynamically changing data?

Challenges in Reproducibility

- 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*

In this paper, we exploit a new multi-country historical dataset on public (government) debt to search for a systemic relationship between high public debt levels, growth and inflation.[1] Our main result is that whereas the link between growth and debt seems relatively weak at "normal" debt levels, median growth rates for countries with public debt over roughly 90 percent of GDP are about one percent lower than otherwise; average (mean) growth rates are several percent lower. Surprisingly, the relationship between public debt and growth is remarkably similar across emerging markets and advanced economies. This is not the case for inflation. We find no systematic relationship between high debt levels and inflation for advanced economies (albeit with individual country exceptions including the United States). By contrast, in emerging market countries, high public debt levels coincide with higher inflation.

Our hope would seem to be a timely one. Public debt has been soaring in the wake of the recent global financial meltdown, especially in the epicenter countries. This should not be surprising, given the experience of earlier severe financial crises.[2] Outstrized deficits and epic bank bailouts may be useful in fighting a downturn, but what is the long-run macroeconomic impact,

* Reinhart: Department of Economics, 4115 Tydings Hall, University of Maryland, College Park, MD 20742 (e-mail: creinhart@umd.edu); Rogoff: Economics Department, 216 Littauer Center, Harvard University, Cambridge MA, 02138-3001 (e-mail: kroguff@harvard.edu). The authors would like to thank Oliver Jonner and Vincent M. Reinhart for helpful comments.

[1] In this paper "public debt" refers to gross central government debt. "Domestic public debt" is government debt issued under domestic legal jurisdiction. Public debt does not include debts carrying a government guarantee. Total gross external debt includes the external debts of all branches of government as well as private debt that is issued by domestic private entities under a foreign jurisdiction.

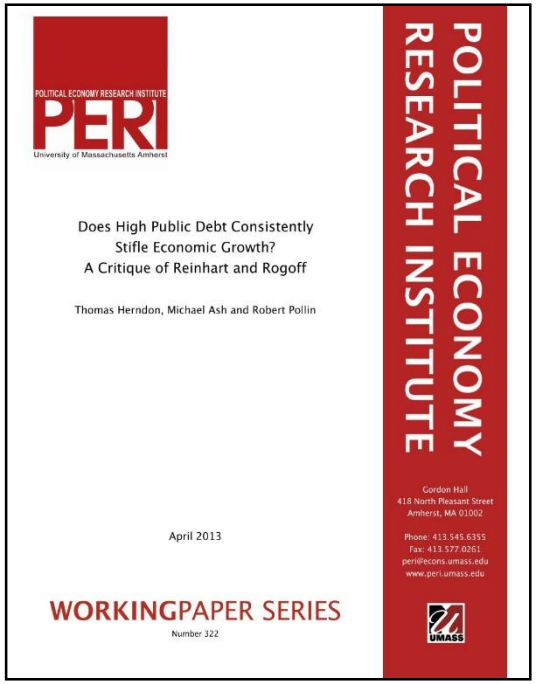
[2] Reinhart and Rogoff (2009a, b) demonstrate that the aftermath of a deep financial crisis typically involves a protracted period of macroeconomic adjustment, particularly in employment and housing prices. On average, public debt rose by more than 80 percent within three years after a crisis.

573

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

Challenges in Reproducibility

- 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

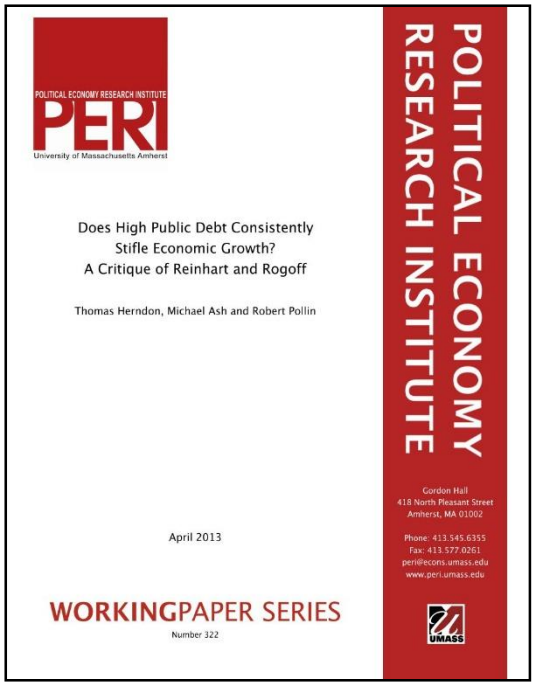
Phone: 413.545.6355
Fax: 413.577.0261
peri@econs.umass.edu
www.peri.umass.edu

UMASS

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

Challenges in Reproducibility

- 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


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

UMASS

- 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	10	124
VIEWS	CITATIONS	SAVES

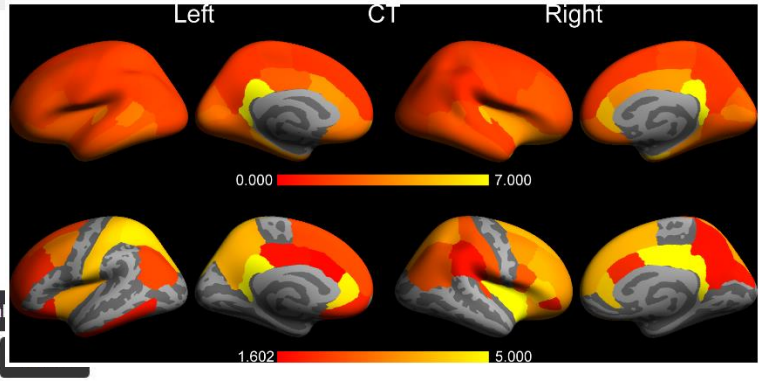
RESEARCH ARTICLE

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



Down

Print

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.

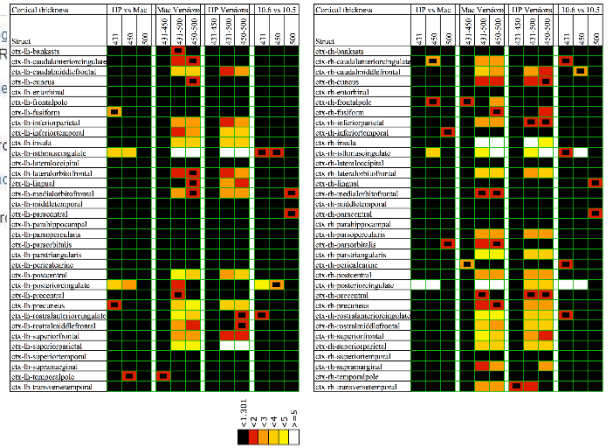
The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations 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.

Comments

In praise of prog
Posted by GedR

Media Coverage
Article
Posted by
PLoS_ONE_Gro

Comments mad
authors
Posted by EdGr



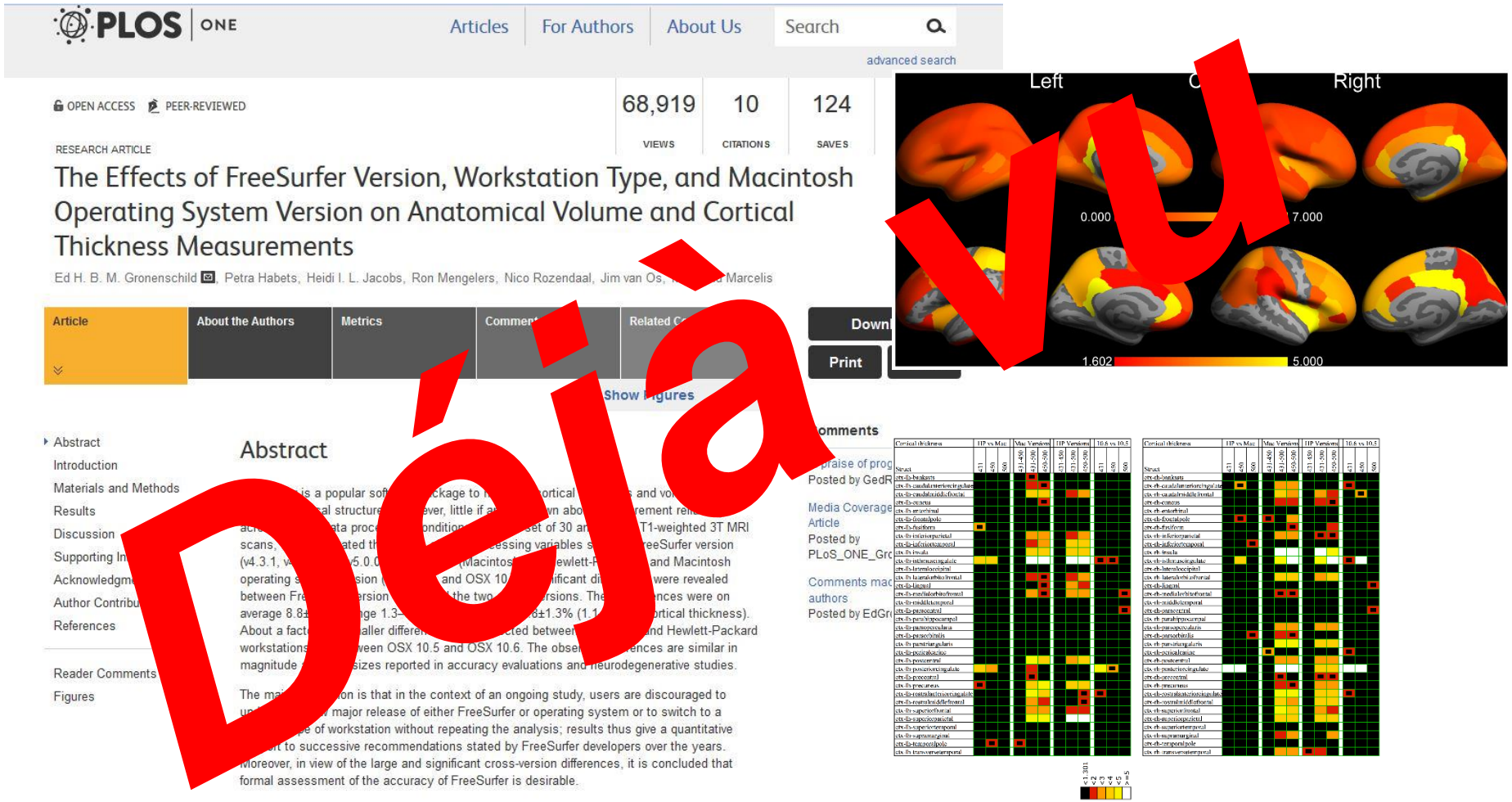
FACULTY OF !NFORMATICS

Outline

- What are the challenges in reproducibility?
- How to address the challenges of complex processes?
- How to deal with dynamically changing data?

Challenges in Reproducibility

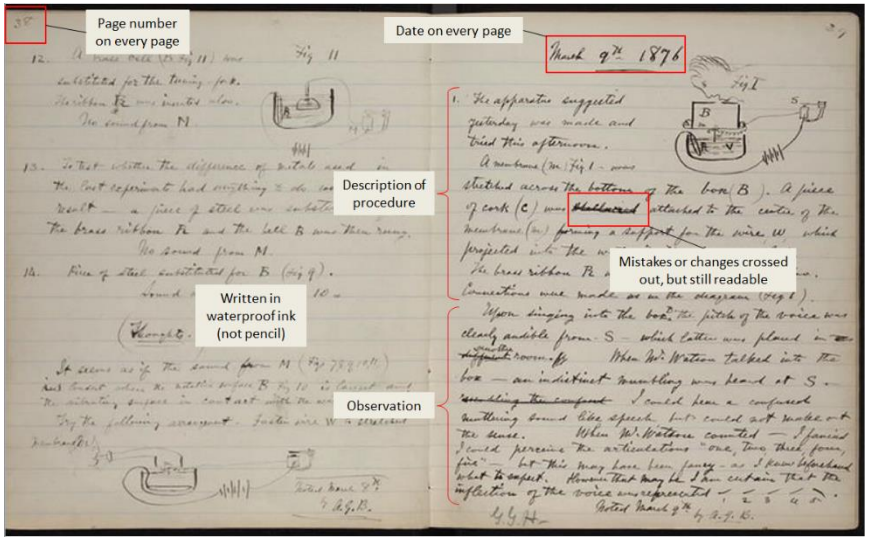
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234



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?

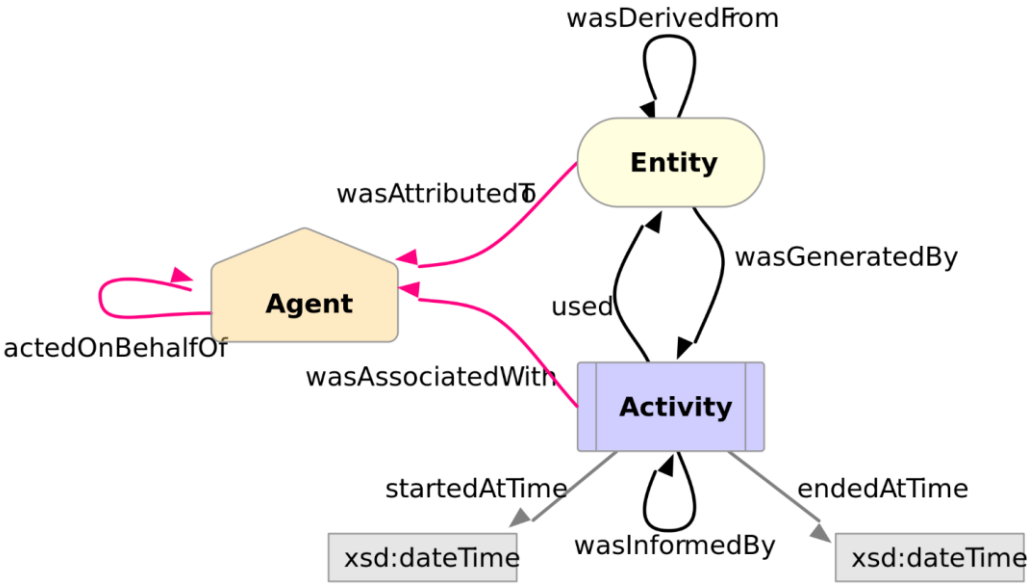


Alexander Graham Bell's Notebook, March 9 1876
https://commons.wikimedia.org/wiki/File:Alexander_Graham_Bell's_notebook,_March_9,_1876.PNG

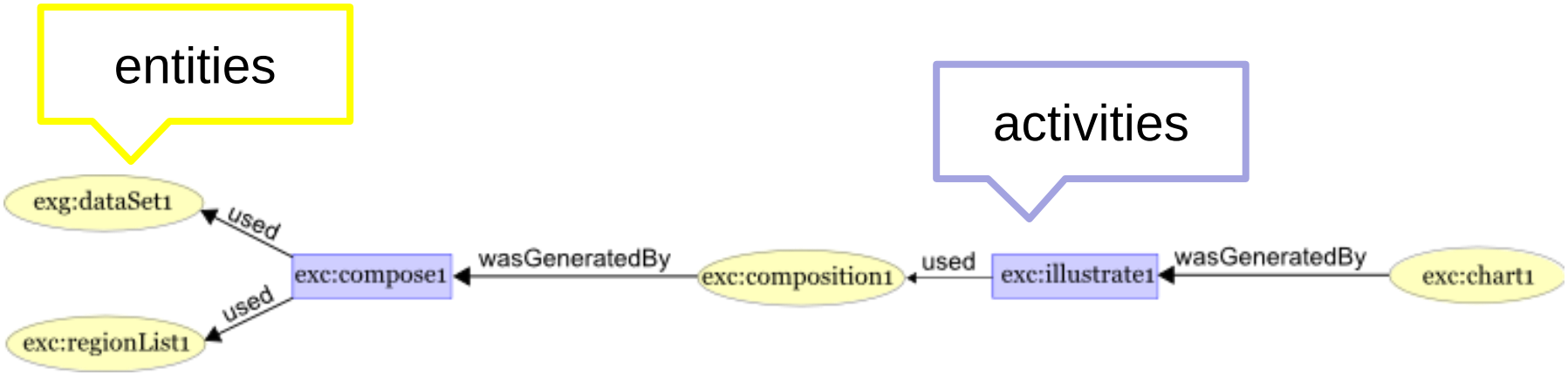
Pieter Bruegel the Elder: De Alchemist (British Museum, London)

PROV-O

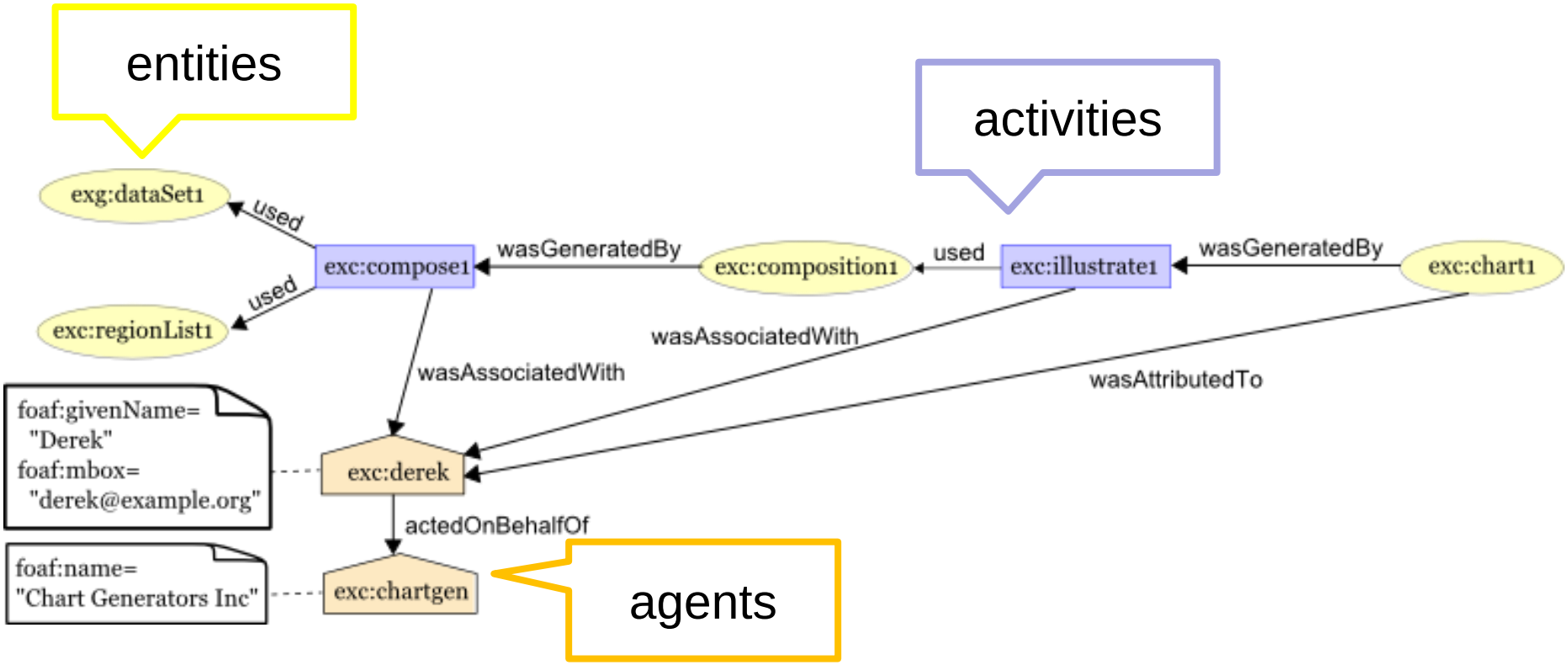
- W3C Recommendation
  https://www.w3.org/TR/prov-o/
- Ontology to represent provenance information
- May use other ontologies
  - FOAF (friends-of-a-friend)
  - Dublin Core
  - PREMIS



PROV-O

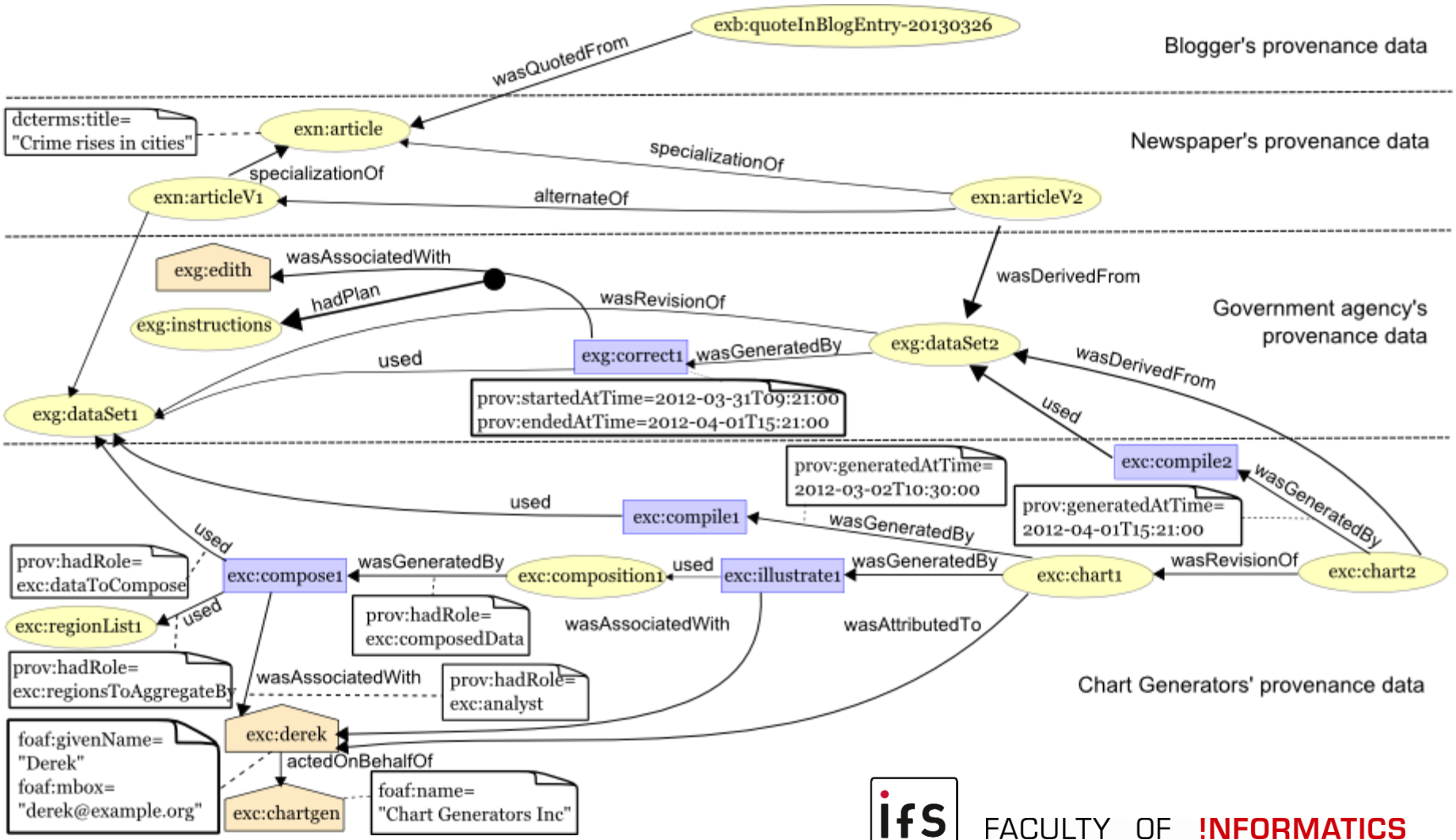

entities
activities
exg:dataSet1
used
exc:compose1
used
exc:regionList1
wasGeneratedBy
exc:composition1
used
exc:illustrate1
wasGeneratedBy
exc:chart1

PROV-O


entities
activities
exg:dataSet1
used
exc:compose1
wasGeneratedBy
exc:composition1
used
exc:illustrate1
wasGeneratedBy
exc:chart1
used
exc:regionList1
wasAssociatedWith
wasAssociatedWith
wasAttributedTo
foaf:givenName=
"Derek"
foaf:mbox=
"derek@example.org"
exc:derek
actedOnBehalfOf
foaf:name=
"Chart Generators Inc"
exc:chartgen
agents

PROV-O

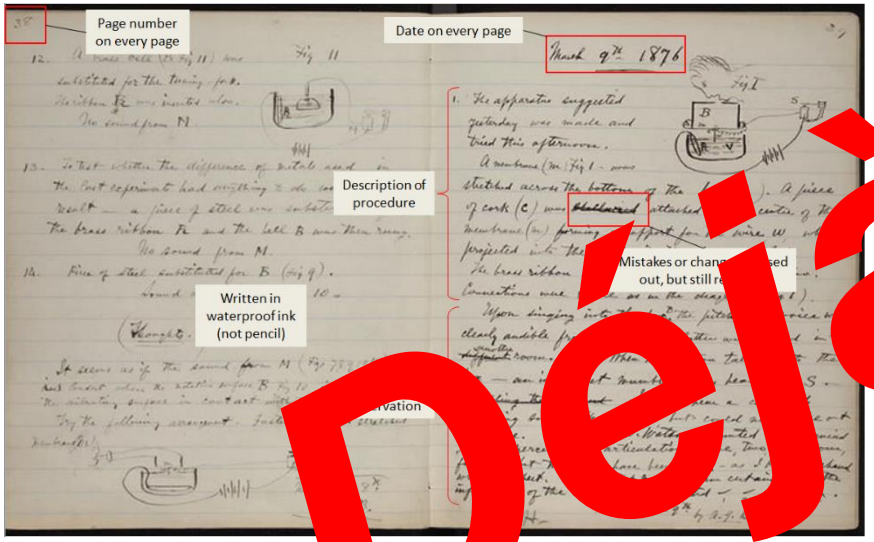
- Adding revisions, time dependencies, plans, …



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?**



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

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

Déjà vu

And the solution is…

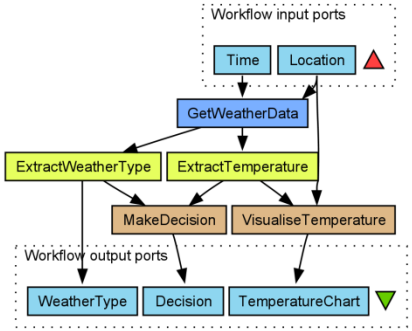
- Standardization and Documentation
  - Standardized components, procedures, workflows
  - Documenting complete system set-up across entire provenance chain
- **How to do this – efficiently!?**
- **Ideally:**
  - Processing pipeline documents provenance automatically
- **Reality: Combination of**
  -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

Static & Dynamic Analysis

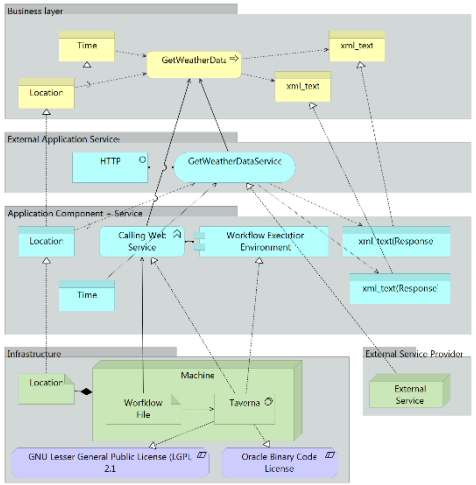
- Analyses steps, platforms, services, tools called
- Dependencies (packages, libraries)
- HW, SW Licenses, …

```
#!/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
```

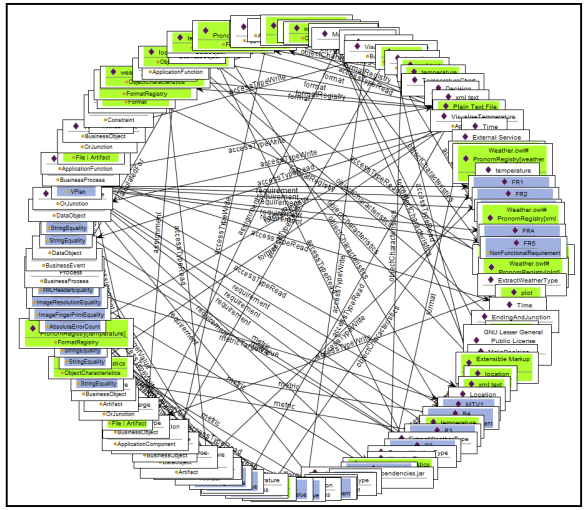
Script



Taverna Workflow

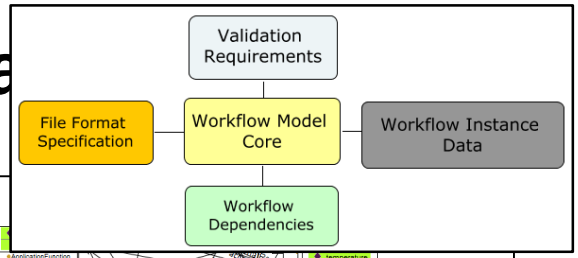


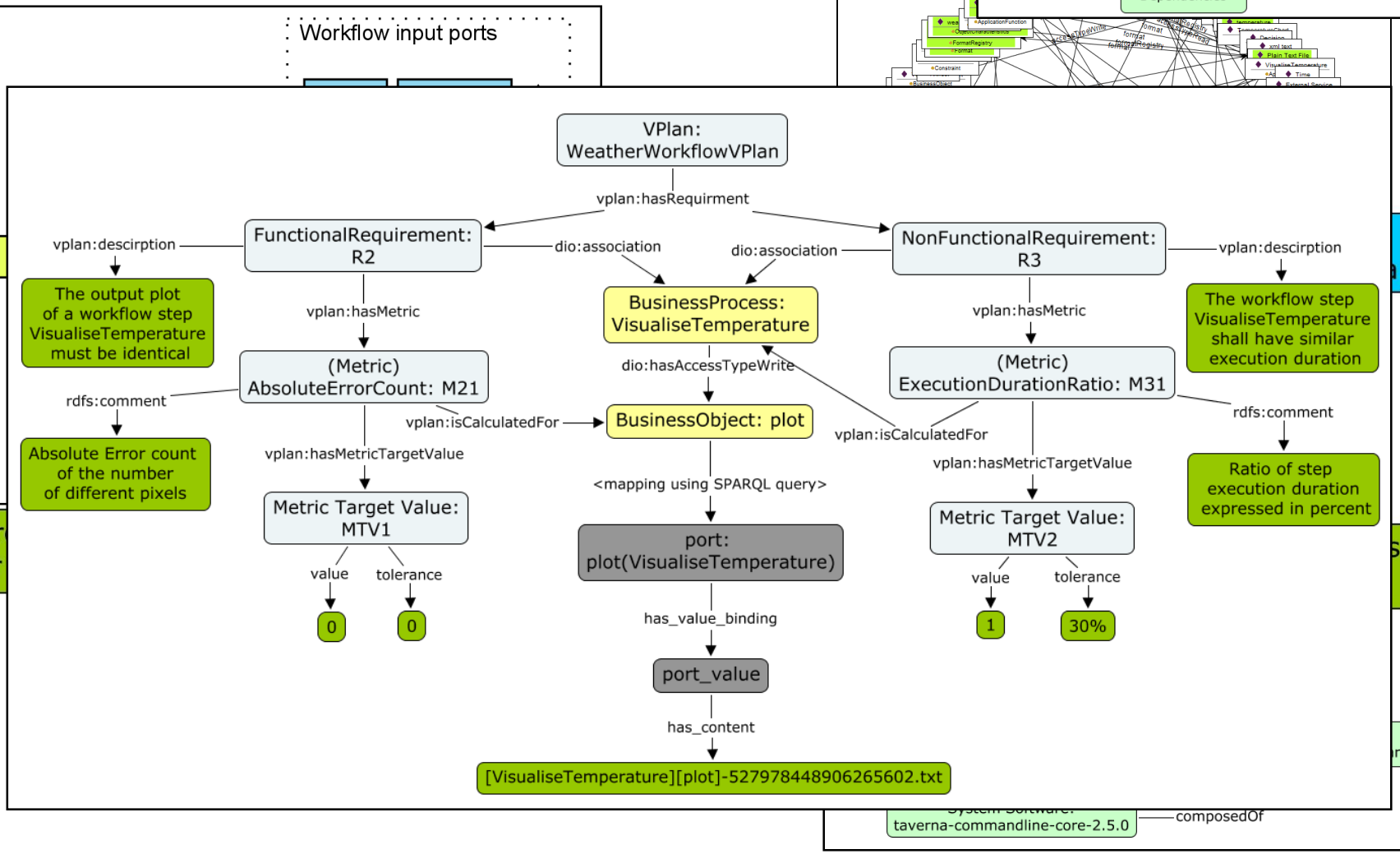
ArchiMate model



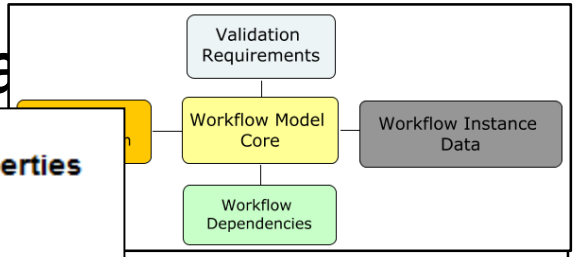
Context Model
(OWL ontology)

Static & Dynamic Ana





Static & Dynamic Ana



- 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

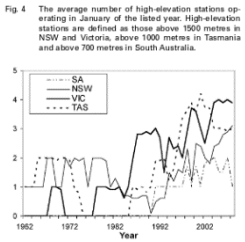
Required additonal Debian packages

base-files
cups-filters
fontconfig-config

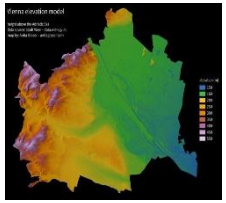
Outline

- What are the challenges in reproducibility?
- How to address the challenges of complex processes?
- How to deal with dynamically changing data?

Motivation



- Data is the key ingredient
  - Data serves as input for workflows and experiments
  - Data is the source for graphs and visualisations in publications
  - Decisions are based on data
- Data is needed for Reproducibility
  - Repeat experiments
  - Verify / compare results
- Need to provide specific data set
  - Service for data repositories



https://commons.wikimedia.org/w/index.php?curid=30978545

1. Put data in data repository,
2. Assign PID (DOI, Ark, URI, …)
3. Make is accessible
   → done!?

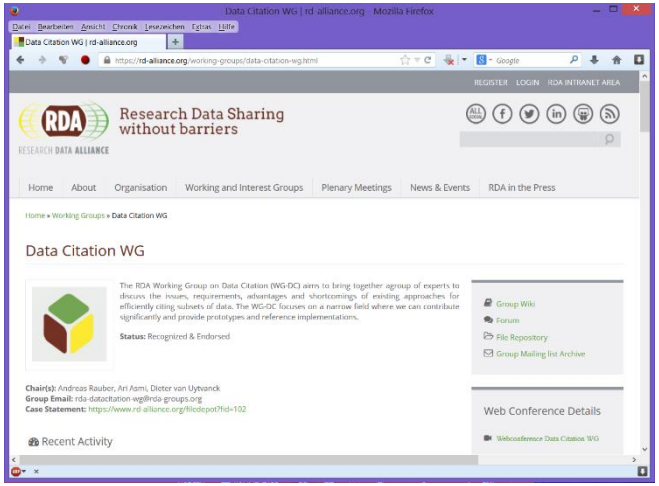
Identification of Dynamic Data

- Usually, datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
- But: (research) data is **dynamic**
  - Adding new data, correcting errors, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Current approaches
  - Identifying entire data stream, without any versioning
  - Using “accessed at” date
  - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to precisely identify the **data**
  as it existed at a specific point in time

Granularity of Subsets

- What about the **granularity** of data to be identified?
  - Enormous amounts of data
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire dataset, providing textual description of subset -> imprecise (ambiguity)
  - Storing list of record identifiers in subset -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to precisely identify the **subset of (dynamic) data used** in a process

RDA WG Data Citation



- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – September 2015
  - Concentrating on the problems of **large, dynamic (changing) datasets**
- Final version presented Sep 2015 at P7 in Paris, France
- Endorsed September 2016 at P8 in Denver, CO

- Adoption by standardization bodies, data centres, …

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

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & assign PID to "QUERY"**, enhanced with
   - **Time-stamping** for re-execution against versioned DB
   - **Re-writing** for normalization, unique-sort, …
   - **Hashing** result-set: verifying identity/correctness

   leading to landing page

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ubset of data
- er gets
  - Data (package, acce s API, …)
  - PID (e.g. DOI) (Que is time-stamped and stored)
  - Hash value compute over the data for local storage
  - Recommended citati n text (e.g. BibTeX)
- PID resolves to land g page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent provenance information on the data set!
- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. BibTeX)

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. BibTeX)
- PID resolves
  - Provides det
  - Option to ret
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

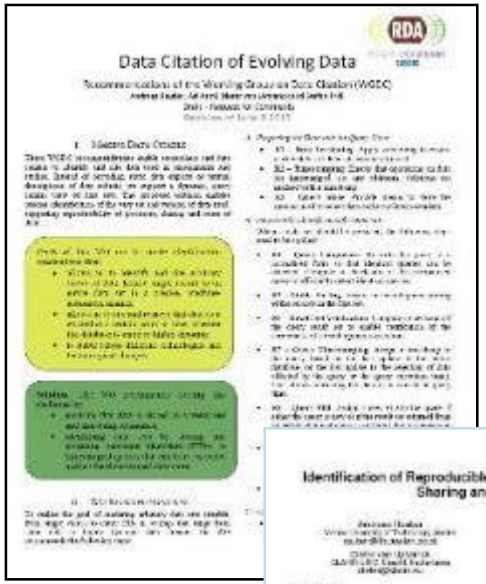
Note: query string provides excellent provenance information on the data set!

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

Identify which parts of the data are used. If data changes, identify which queries (studies) are affected

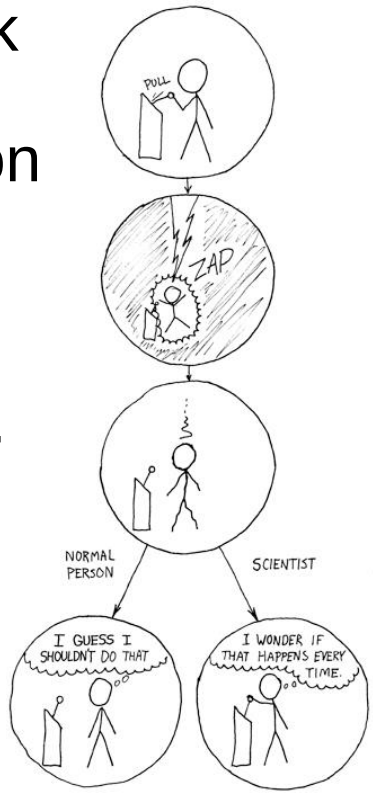
Data Citation – Output

- 14 Recommendations grouped into 4 phases:
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Resolving PIDs
  - Upon modifications to the data infrastructure
- 2-page flyer
  https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html
- More detailed report: Bulletin of IEEE TCDL 2016
  http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

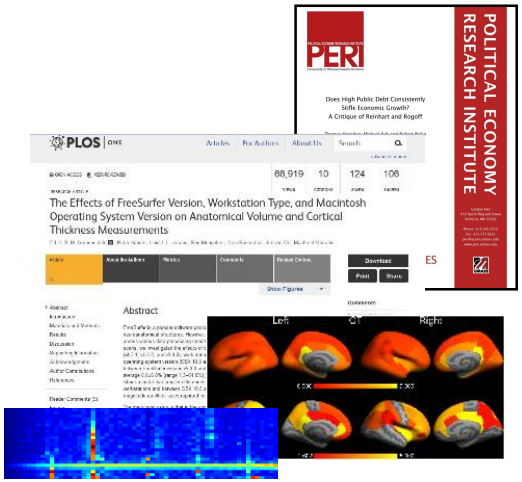


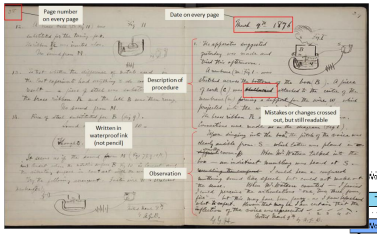
Conclusions

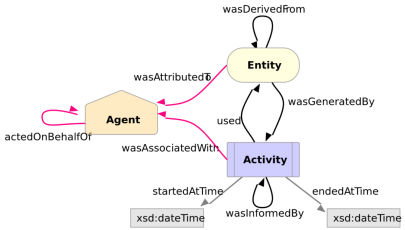
- Reproducibility is more challenging than we might assume
- Need to address it if we want to do proper work
- Standardization, documentation and automation
- Managing the dynamics in data and processes
- If not, we are closer to alchemy than science…
- …and may not reap the promised benefits of digitization

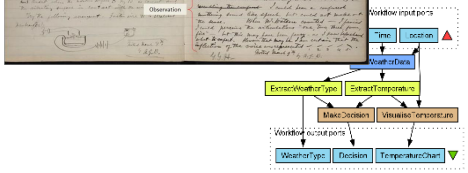


Thank you!










```
#!/bin/bash

# fetch data
java -jar
GestBarragensWSClientIQDa
ta.jar
unzip -o IQData.zip
```

Thanks!