High Value Datasets

(A passion project)

Ian Marsh

Feb. 2022

Jack of all trades, master of none.

- B.Sc. Physics & Mathematics
- M.Sc. Computer Science
- Ph.D Electrical engineering

- NVIDIA, UK
  - Graphics *and* Image processing
- CISRO, AU
  - Precursor to R
- IBM, DE
  - IP Version 5, broke it 😼

Dreamers.
Thinkers.
Doers.

A never ending story, receiving “dodgy data”.

Various ways of ‘getting’ data

1. In all (industrial) projects
2. Inherited by someone
3. Made available to us
4. Pestered someone for
5. Downloaded from http://
6. Emailed attached …
7. Have a look at … groan…

What's your (*data*) problem?

1. Too much?
2. Don't understand it?
3. How should I process the data?
   1. Batches? Records? Both?
4. Is the data secure?
   1. Can I process this data on …
      1. AW3? RISE North? elsewhere?
5. I don't understand the implications of the license!
6. Will AI work?
7. It doesn't produce the results I want…

OK, don’t panic, some guidance

- Establish pipelines that can be reused
- Process and transport data securely (see marketplace)
- Functional style of coding
- Let’s improve the quality of the data
  - Readiness
  - High Value

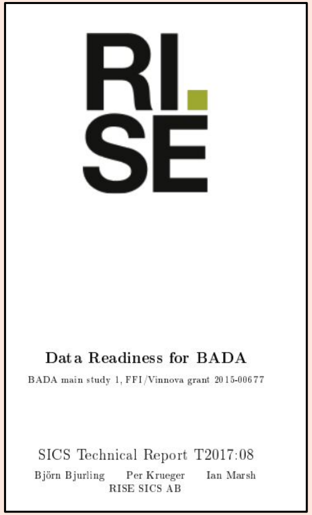
Data Readiness - a ‘TRL for Data’

Simple concept developed by Neil Lawrence when at Sheffield Uni.

SICS, (former RISE) wrote a paper further developing the ideas + interviews, see [IansPage, Neil, RISE]

Neil Lawrence

Basically divide the data into Bands, A, B, C and then classes within these 3.



Data Readiness for BADA

BADA main study 1, FFI/Vinnova grant 2015-00677

SICS Technical Report T2017:08

Björn Bjurling Per Krueger Ian Marsh
RISE SICS AB

TRLs Vs. DRLs[1] #1

Level	TRL	DRL
9	System proven in operational environment	Correct ML predictions with clean, exportable (DRL 8-9) data
8	System complete and qualified	Working system, customer happy
7	Integrated pilot system demonstrated	Data collection, imputation, processing and visualisation as expected.
6	Prototype system verified	“Demo”
5	Laboratory testing of integrated system	ML pipeline produces ‘sensible’ output. Training and cross validation as expected.
4	Laboratory testing of prototype component or process	Unit tests (functions, files) ok
3	Critical function, proof of concept established	Final results (regression, will work)
2	Technology concept and/or application formulated	Code works
1	Basic principles are observed and reported	Data algorithms identified

[1] This is not a definitive list :-)ailable

Data Quality

- Next step from Data Readiness
- Indication of usefulness in AI cases
- RISE and Ramboll evaluated Open Nordic Datasets
- Idea is to motivate others to improve quality, illustrating the better ones
  - Traffic, water, weather are good
  - Identify criteria for evaluation

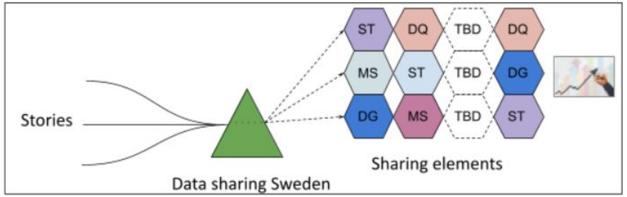
Nordic Council of Ministers

Nordic cooperation on data to boost the development of solutions with artificial intelligence

PDF, Site

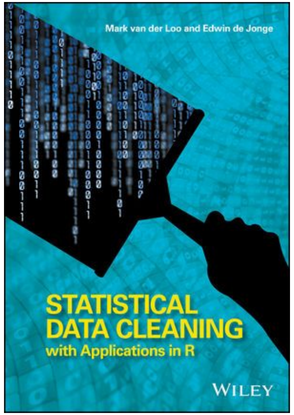
One project, 2 proposals

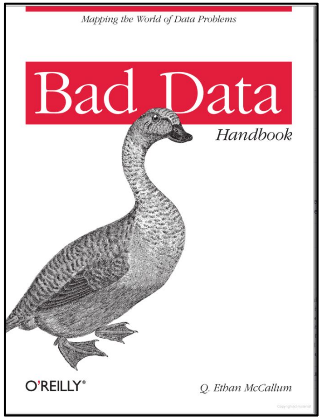
- ROADVIEW
- **Ro**bust **A**utomated **D**r**iv**ing in **E**xtreme **W**eather (just accepted)
- RISE will improve quality from LiDAR, RADAR and camera sensors to AI
- Proposals to Vinnoa
  - Data Marketplace
  - Data Coupons

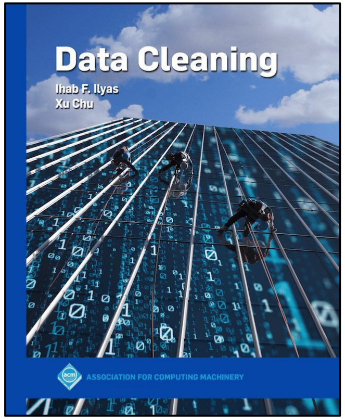


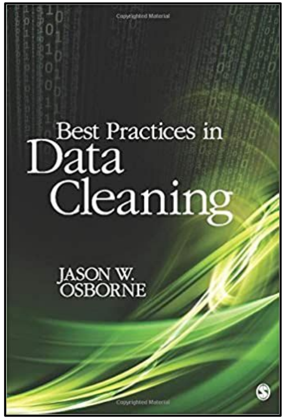
Stories, decomposed, into reusable tools, methods & technologies bootstrapping a data economy. Data sharing Sweden could be an independent expert organisation.

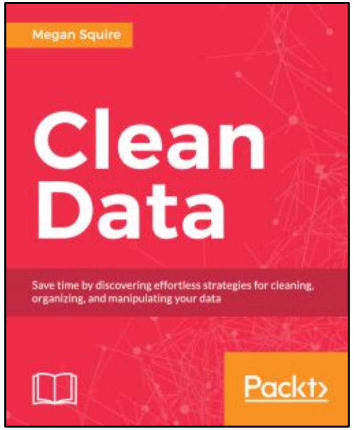
Reading (and coding) material


Mark van der Loo and Edwin de Jonge
STATISTICAL DATA CLEANING
with Applications in R
WILEY



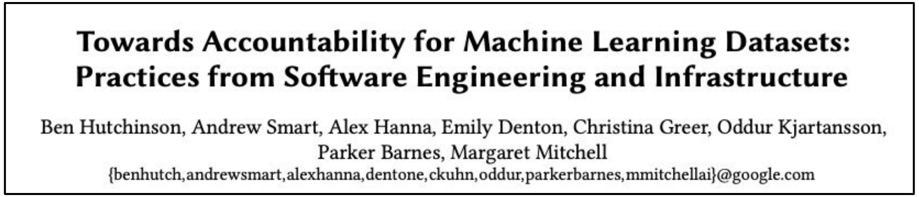





Megan Squire
Clean Data
Save time by discovering effortless strategies for cleaning, organizing, and manipulating your data
Packt>

Code:

HoloClean
Ours :-)



Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, Margaret Mitchell
{benhutch,andrewsmart,alexhanna,dentone,ckuhn,oddur,parkerbarnes,mmitchellai}@google.com

Thanks, Jacob Dexe.

Questions?

Extra slide: Wrangling woes

1. Commas / decimals points can mean different things
2. Location is not always GPS (SWEREF Trafikverket)
3. The ever existent missing value issue (remove line, add average value,etc. long list)
4. Poorly labelled fields (if at all)
5. Bad designs (e.g. speeds > 250 km / hr ‘were’errors)
6. In some cases missing lines (sensor data a common case)
7. Time zones, summer / winter time changes over long data
8. Data from different sources not time aligned
9. Choosing sec, min, hour resolution for visualisation, processing
10. Smoothing / averaging decision (lag parameter)