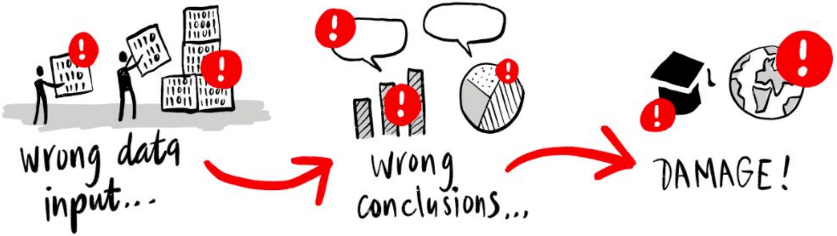
HCOMP 2020

Virtual Conference, 25-29 October 2020

FIRST Data Excellence WORKSHOP

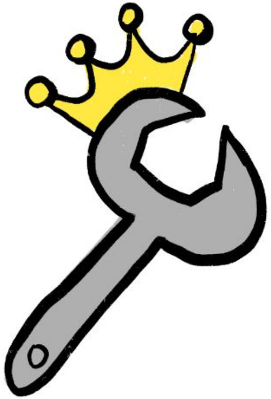
visuals by DENKSCHETS.NL

BAD DATA hurts SCIENCE & SOCIETY



COLLECT & TESTING DATA for A.I.

HOW TO GET Excellence?



HOW DO YOU BUILD Excellent Datasets?

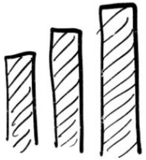
BUILD DATASETS WITHIN THE COMMUNITY OF USE

RELENTLESSY INTROSPECT

try to understand

GET BELOW THE SURFACE OF STATE OF THE ART (SOTA)

Challenges and lessons in ADJUDICATING MEDICAL DATA



Crowdsourcing medical data (diagnoses)

"Ask 3 doctors, get 4 opinions..."

CONCLUSIONS

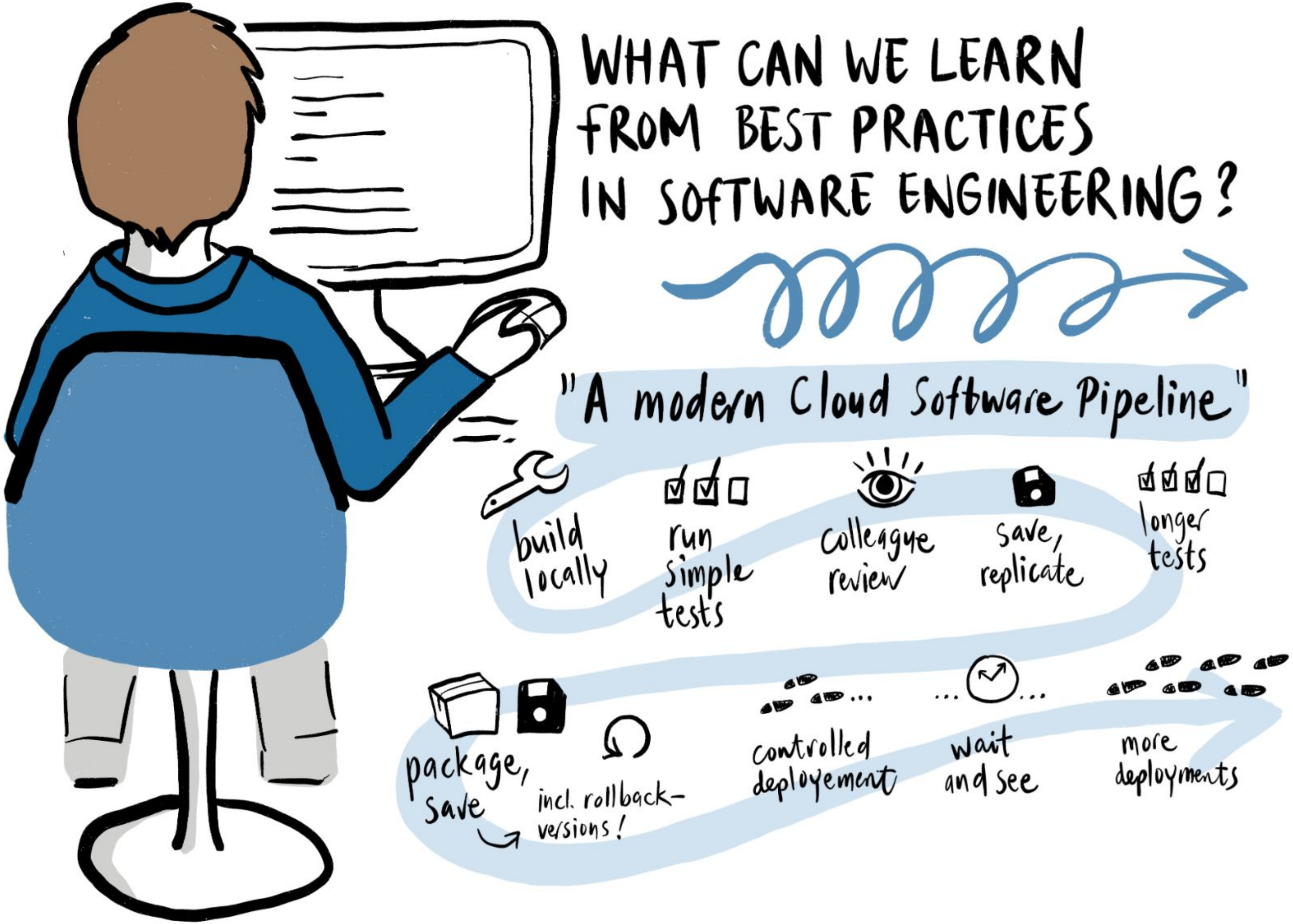
Medicine and Medical data are complex!

Improvement of
- * visibility
- * accountability
- * communication

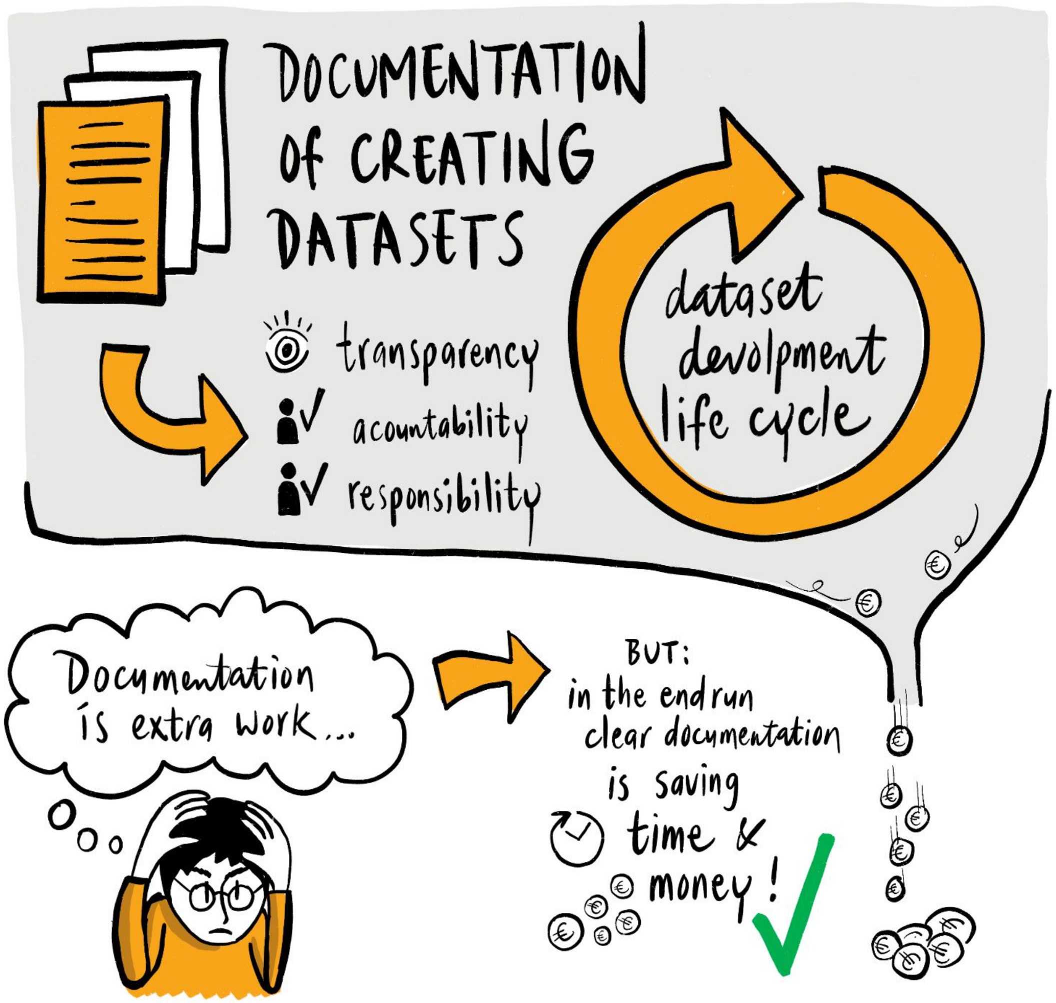
EM-inferred silver datasets may be an alternative...

Elimination of wait time!

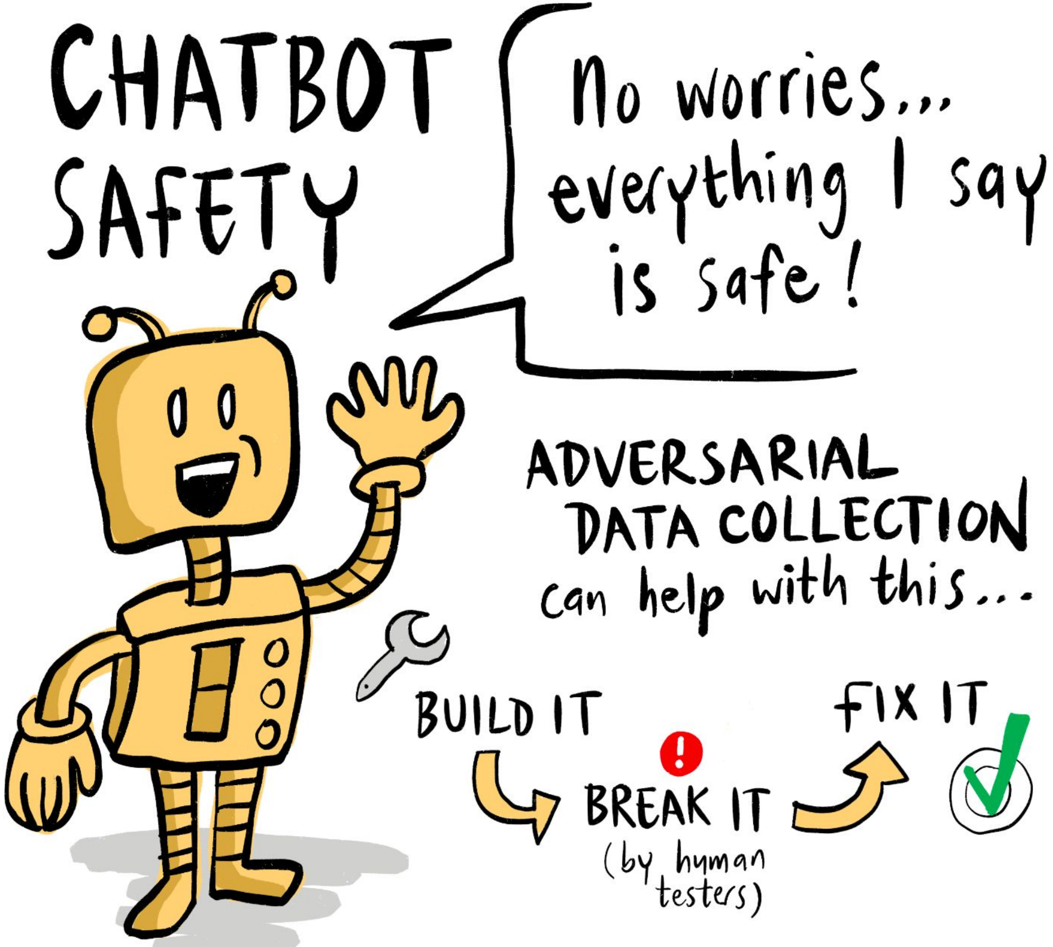
WHAT CAN WE LEARN FROM BEST PRACTICES IN SOFTWARE ENGINEERING?



"A modern Cloud Software Pipeline"


DOCUMENTATION OF CREATING DATASETS
transparency
acountability
responsibility
dataset devolpment life cycle
Documentation is extra work...
BUT:
in the endrun clear documentation is saving time & money!

CHATBOT SAFETY



ADVERSARIAL DATA COLLECTION can help with this...

BUILD IT

BREAK IT (by human testers)

FIX IT

WHAT DO OUR MODELS LEARN?

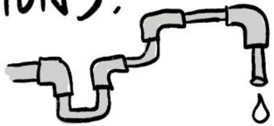
How well do our datasets reflect the real world?

WHERE DO BIASES COME FROM?

CONCLUSIONS:

???

What models do and do not learn is not always clear



Models are affected by biases of the world and our data pipelines

"But: we can improve the situation by accounting for them"

GETTING GOOD DATA* FASTER

*) human labeled data sets + machine learning solutions

MACHINE LEARNING DEVELOPMENT CYCLE

training → evaluation → "patching" production results

Sub cycle:

HLD PRODUCTION PROCES

define → pilot → produce

LABELING WORKFORCES

"How core is this to my business?"

PRODUCTION ARCHITECTURE

aka dataset pipeline