Carnegie Mellon University

Data Discovery and Reuse: AI Solutions & the Human Factor

FEBRUARY 24, 2020

Huajin Wang, Ph.D.

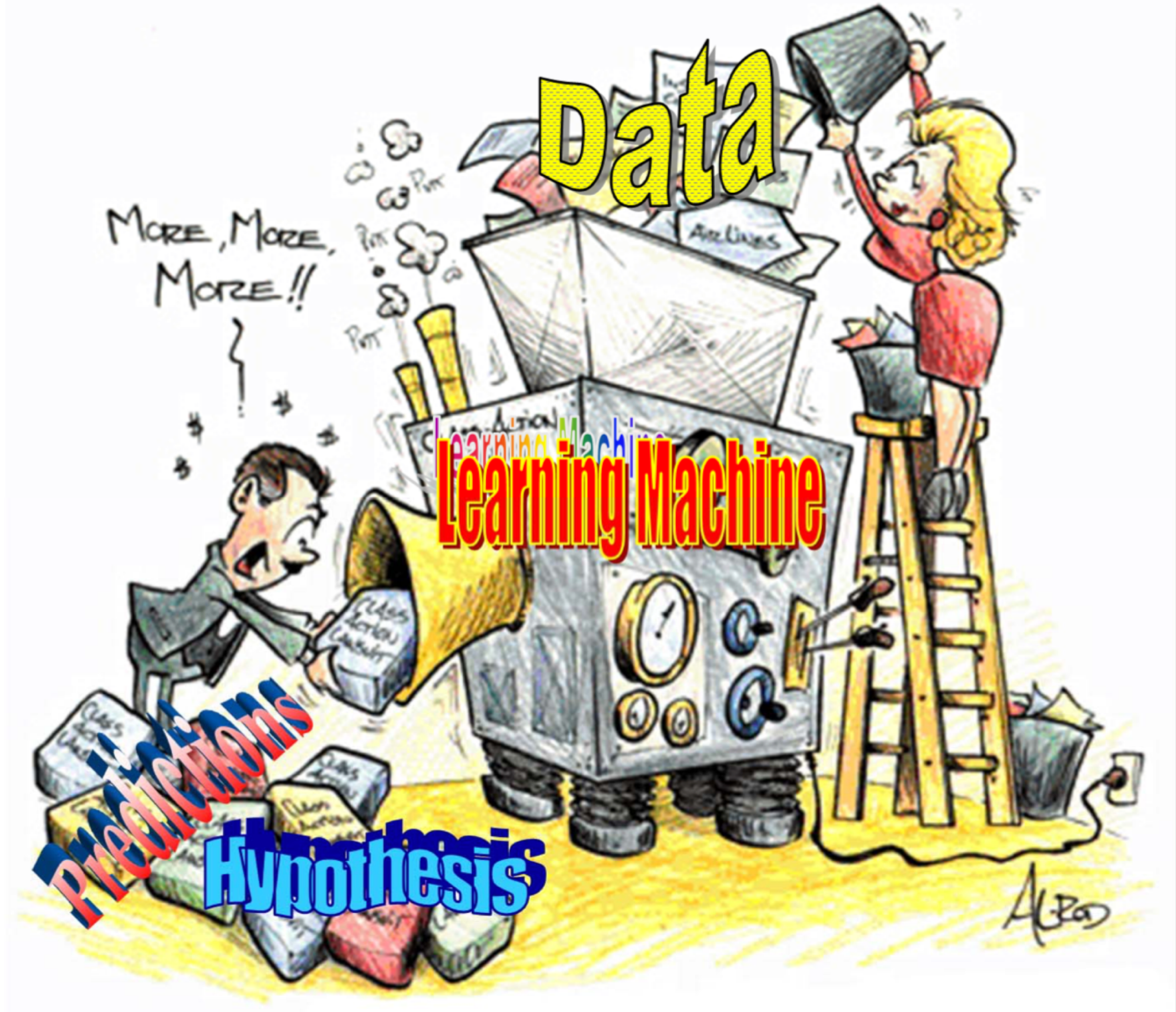
Liaison Librarian, Biology and Computer Science

Program Director, Open Science & Data Collaborations

huajinw@cmu.edu

The AI pipeline starts from data

“Garbage in, garbage out”: What does it mean?



© Eric Xing @ CMU, 2006-2016

AI is only as good as it’s data

- Poor data quality (accuracy, completeness, bias, consistency, validity, uniqueness, ….)

  → Bias or wrong prediction

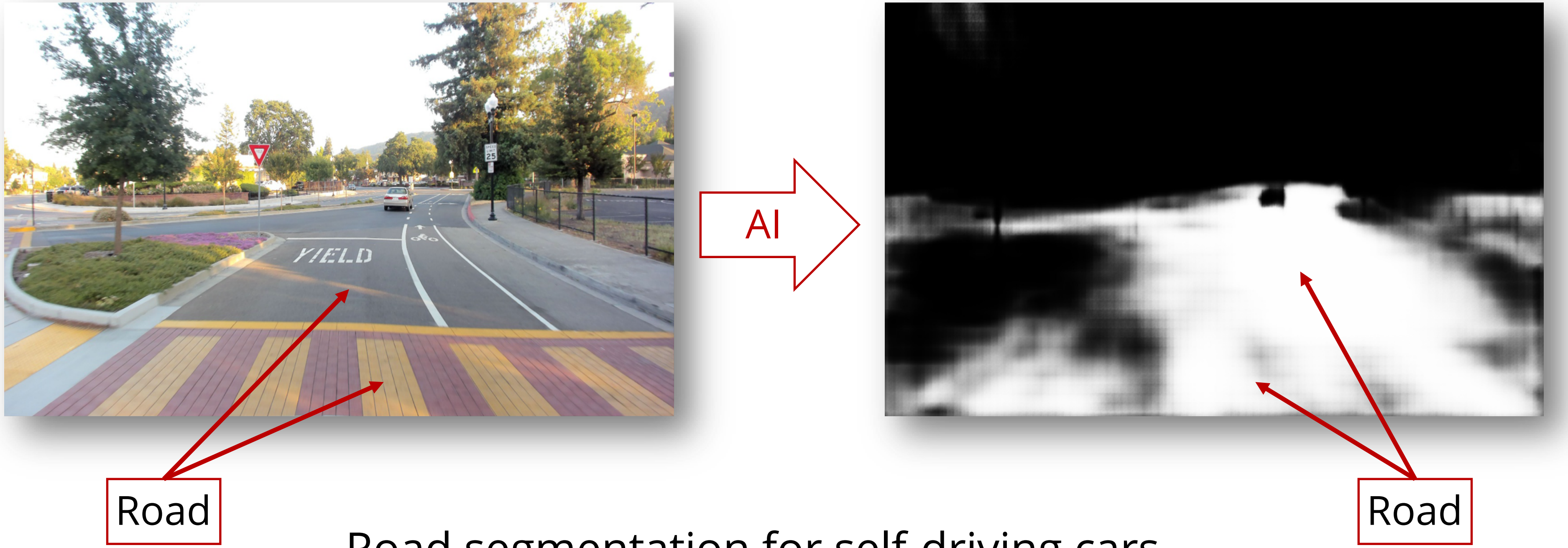
- Poor metadata or documentation

  → Data data (re)usability

- Not enough data

  → Model performance; overfitting

Key to successful AI systems:
accessible high quality labeled data



Road segmentation for self-driving cars

Adapted from Evgeny Toropov. AIDR ’19.

High quality data are hard to find

Lab website / Server

404

Page not found :(

The requested page could not be found.

Google Drive

aws

box

Dropbox

OneDrive

figshare

DRYAD

HARVARD Dataverse

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

zenodo

PubMed

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

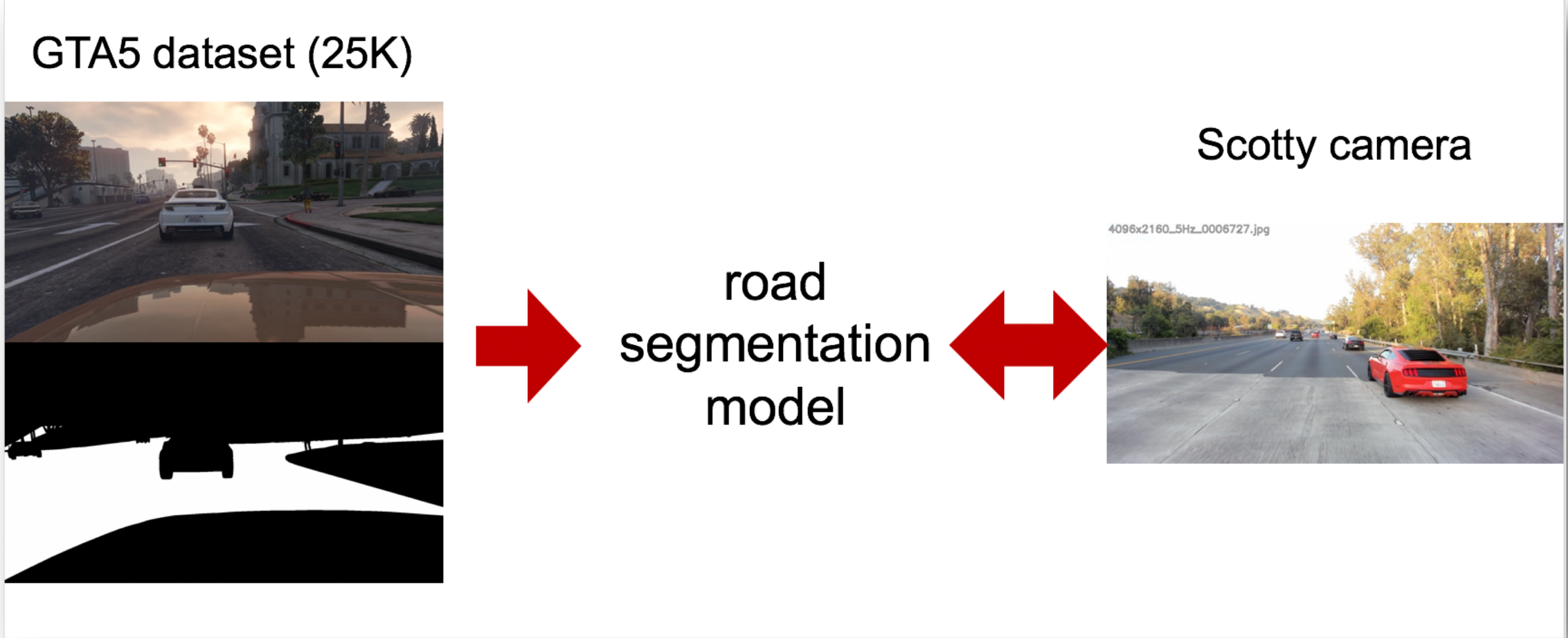
High quality data are hard to find

Technical solutions

- Search engines to find existing data
- Automation in evaluating data quality and integrating datasets
- Automation in data curation
- Model transfer and data augmentation

https://events.library.cmu.edu/aidr2019/

Technical use case 1: Domain adaptation to transfer model to new data



Road segmentation for self-driving cars

Technical use case 2: Finding existing datasets

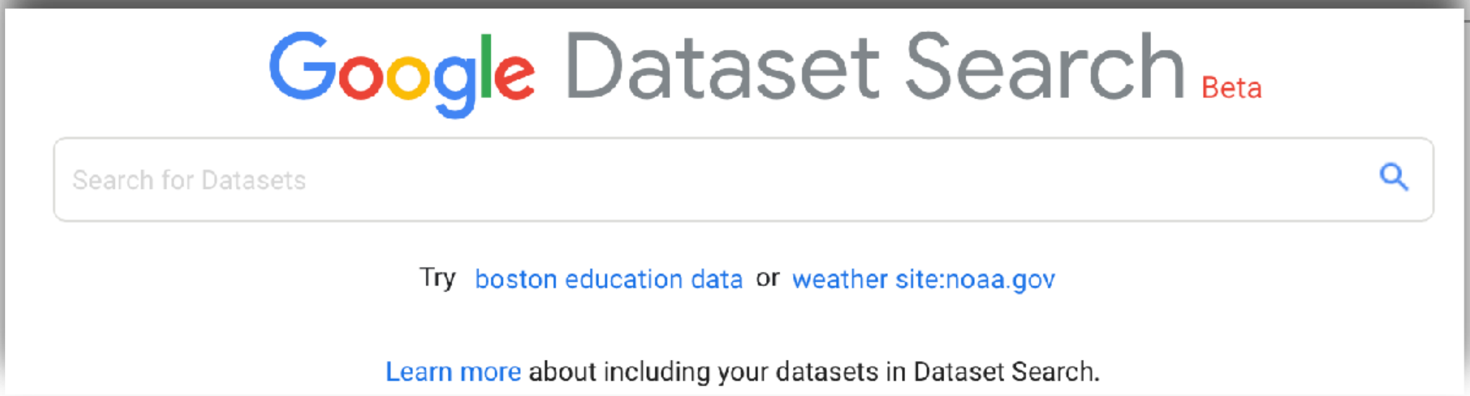
- Datasets are distributed and hard to find
- Structured data
  - Web (1%*)
  - Repositories
- Unstructured data
  - Web (99%)
  - Publications
  - Images
- Overall discovery layer missing

```
<html>
<head>
<title>Grandma's Holiday Apple Pie</title>
<script type="application/ld+json">
{
  "@context": "https://schema.org/",
  "@type": "Recipe",
  "name": "Grandma's Holiday Apple Pie",
  "author": "Elaine Smith",
  "image": "http://images.edge-generalmills.com/564592
  "description": "A classic apple pie.",
  "aggregateRating": {
    "@type": "AggregateRating",
    "ratingValue": "4",
    "reviewCount": "276",
    "bestRating": "5",
    "worstRating": "1"
  },
  "prepTime": "PT30M",
  "totalTime": "PT1H",
  "recipeYield": "8",
  "nutrition": {
    "@type": "NutritionInformation",
    "servingSize": "1 medium slice",
    "calories": "230 calories",
    "fatContent": "1 g",
    "carbohydrateContent": "43 g",
  },
  "recipeIngredient": [
    "1 box refrigerated pie crusts, softened as direct
    "6 cups thinly sliced, peeled apples (6 medium)",
    "..."
  ],
```

figshare
DRYAD
HARVARD Dataverse
re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES
zenodo
PubMed
NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

* https://www.bostonwebdesigners.net/news/structured-data-and-local-seo/

Technical use case 2:
Finding existing datasets (structured)

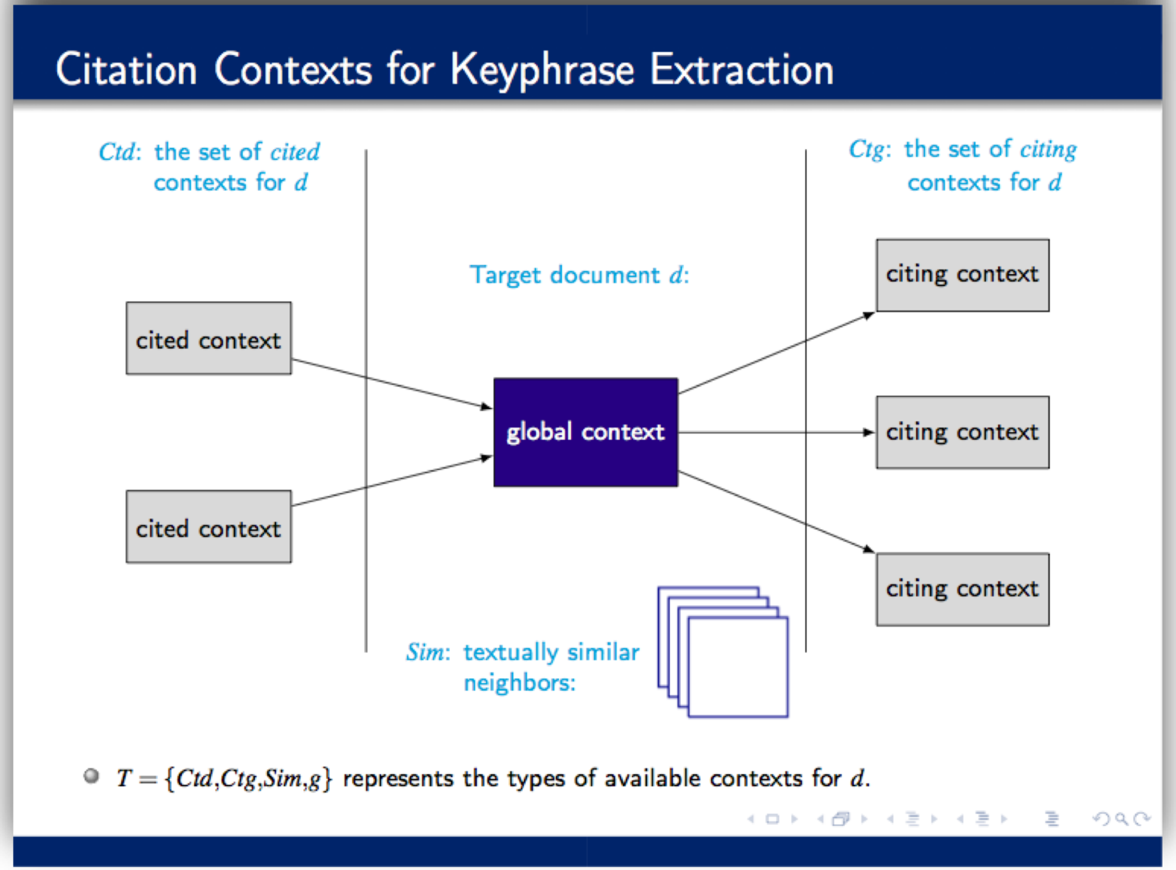


- Search engine powered by AI
- Simple keyword search for datasets across the web
- Searches over embedded **metadata**
  - Searches over metadata from data providers
  - schema.org data standards (embedded in html)
  - Dataset name, description, provider, temporal coverage, …

Technical use case 2:
Finding existing datasets (unstructured)

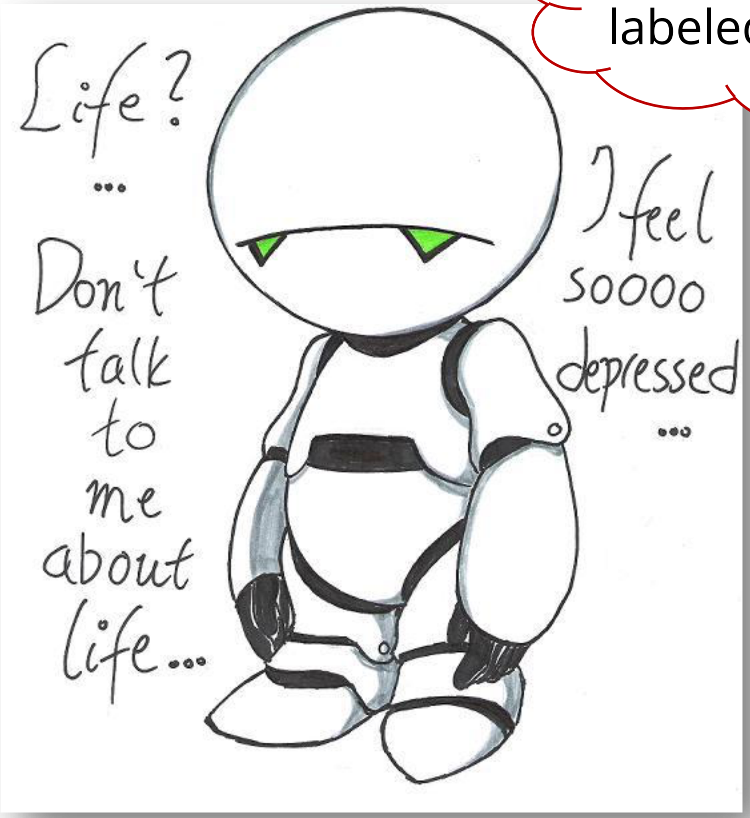
Unstructured data: need metadata tagging and data linking first

- Keyphrase extraction from scholarly documents



 Cornelia Caragea and C. Lee Giles. AIDR '19.

Why aren’t the
data already
labeled? …



 https://www.deviantart.com/stantti/art/MARVIN-35569632

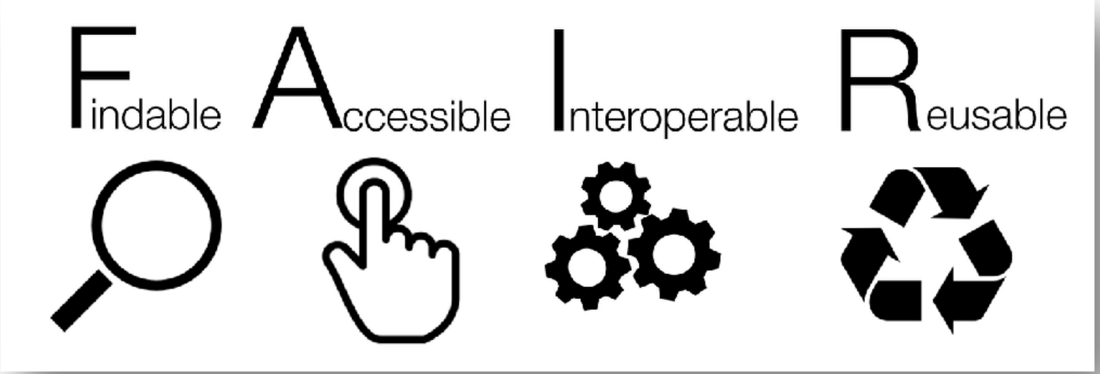
Non-technical solutions (the human factor)

Data stewardship

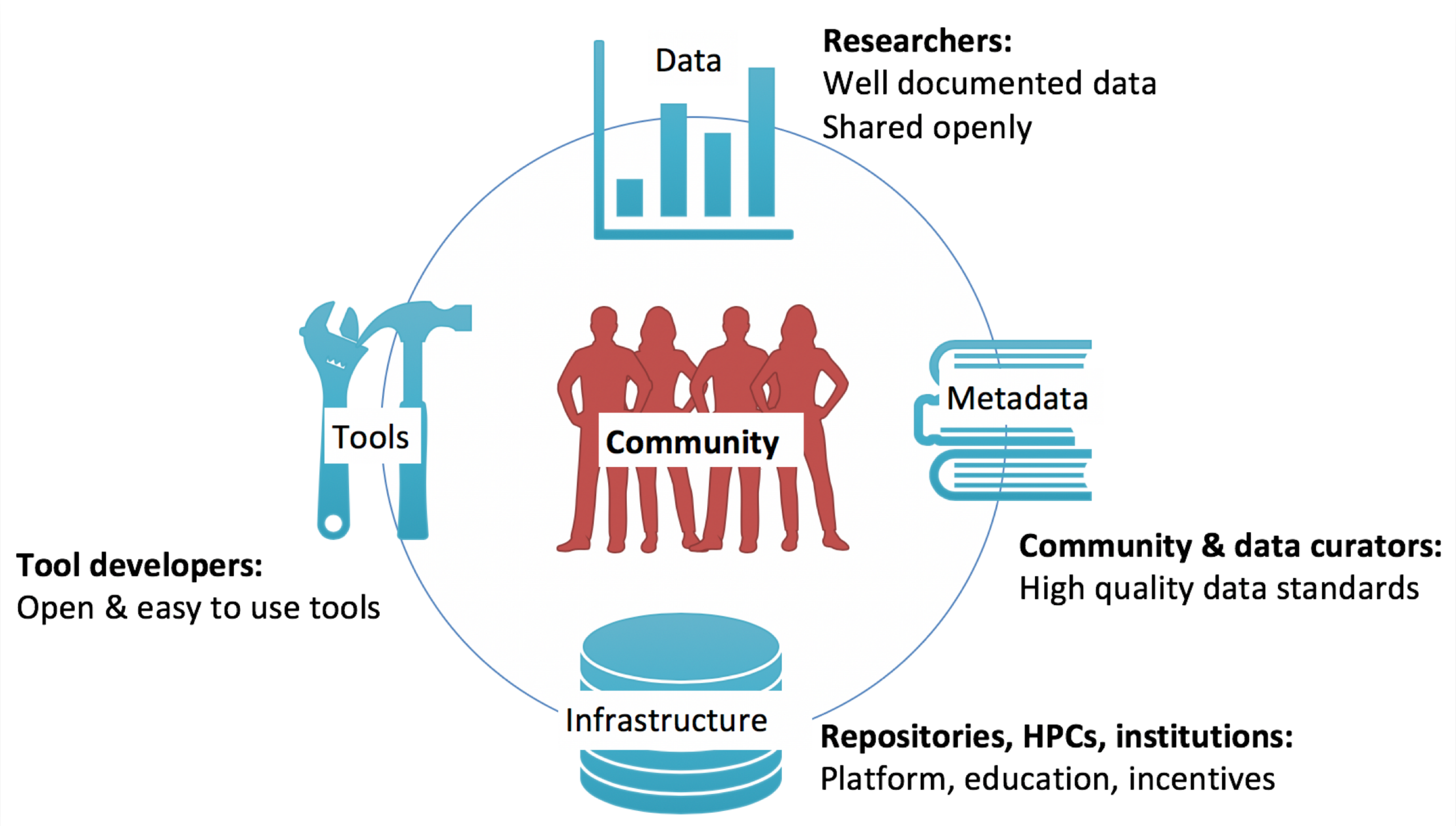
- Responsible data collection and documentation
- Data management best practices
- Metadata & **data standards**

Open science and data sharing

- Share data, code, workflow
- Follow FAIR principles
- Build easy to use and robust tools for data sharing and reuse
- Interdisciplinary collaborations

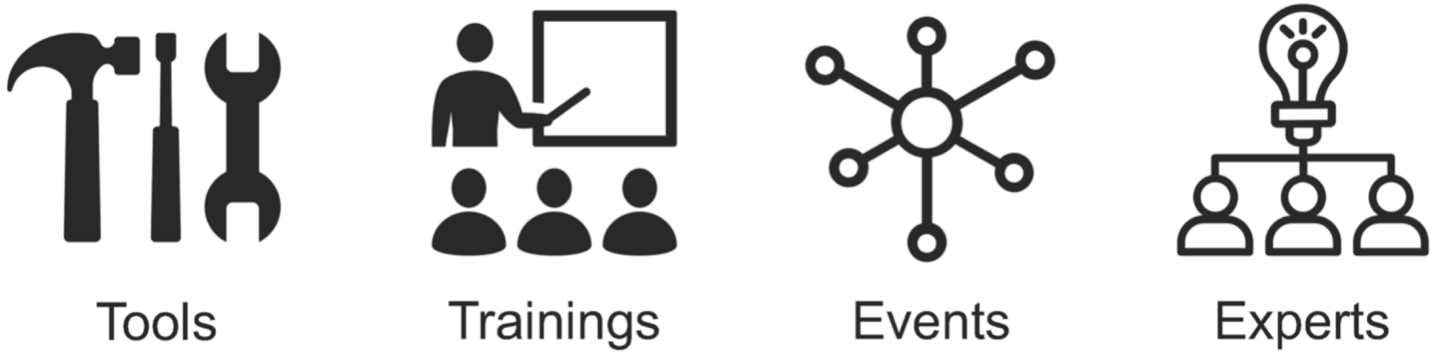


Work together to build a healthy data ecosystem



Open Science & Data Collaborations Program @Carnegie Mellon

Support open practices across the entire research life cycle – from project planning, DMPs, preregistration, through operational data management and documenting reproducible workflows and methods, to publicly sharing research products including data and code in a way that is discoverable and reusable (FAIR+).



 https://events.library.cmu.edu/aidr2020/

Thank you!

huajinw@cmu.edu

@HuajinBioLib

Photo: "Mobot" (MObile roBOTs) competition at CMU's spring carnival.