AI for Data Reuse - Tools, Challenges, and Opportunities

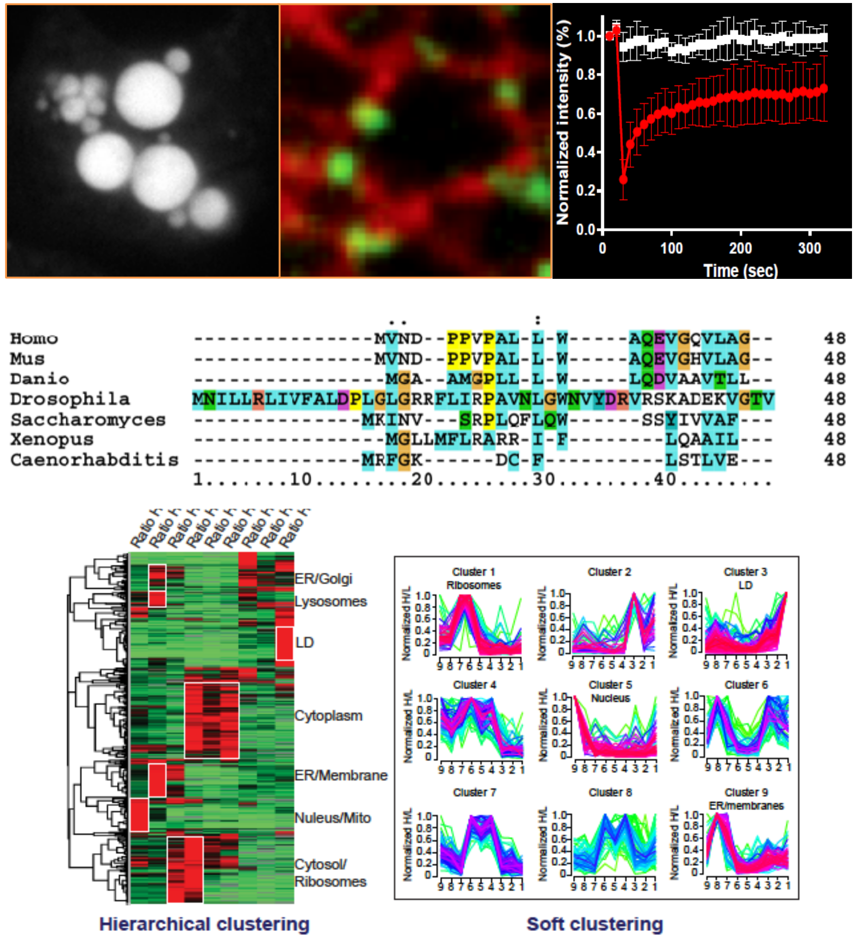
Huajin Wang, PhD
Carnegie Mellon University Libraries
Program Director, Research Data Collaborations

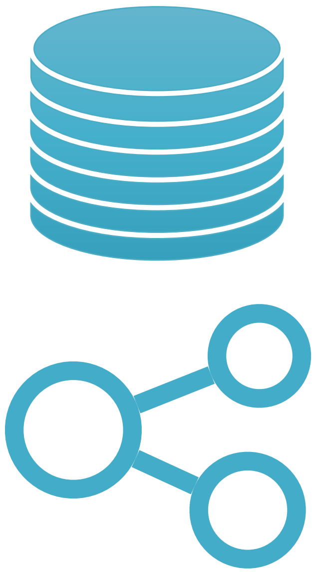
Reproducibility and Data Reuse in Life Science
@ SciLifeLab Data Centre
September 19, 2019

The research data life cycle

Creation → Processing + Analysis → Preservation + Sharing → Reuse

Research Data

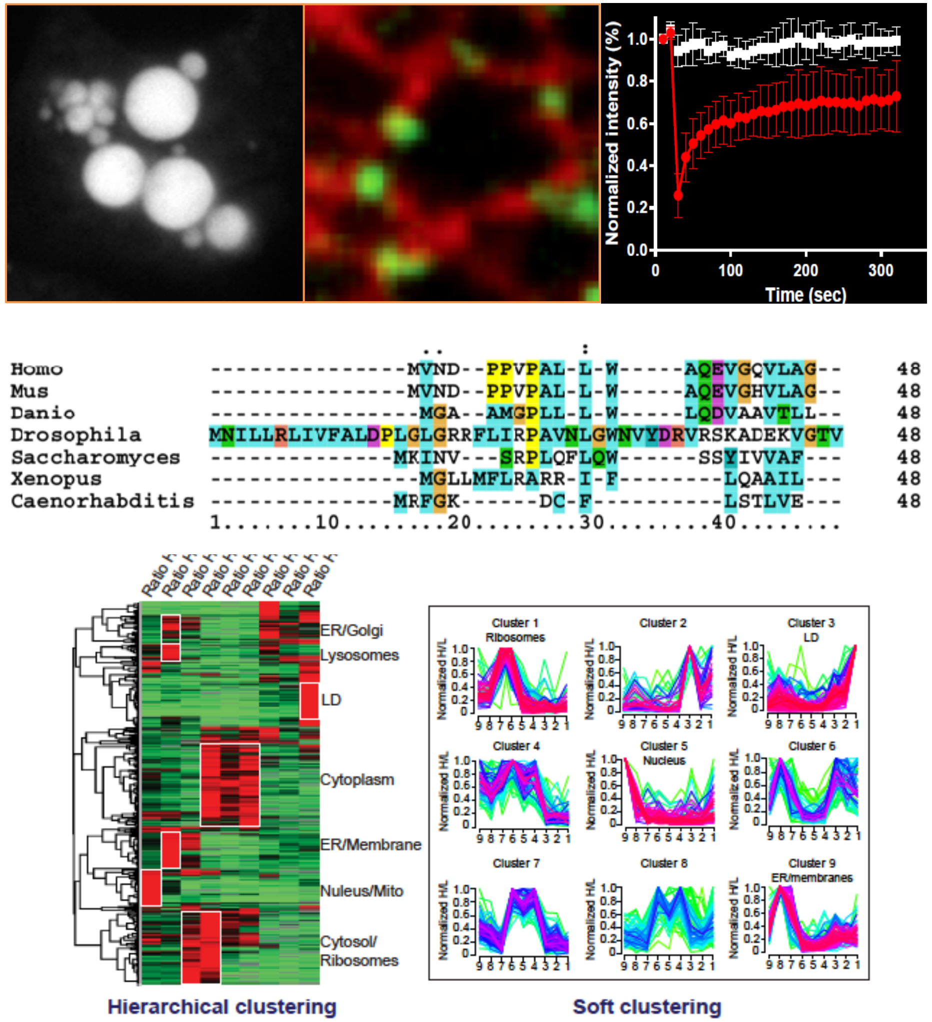




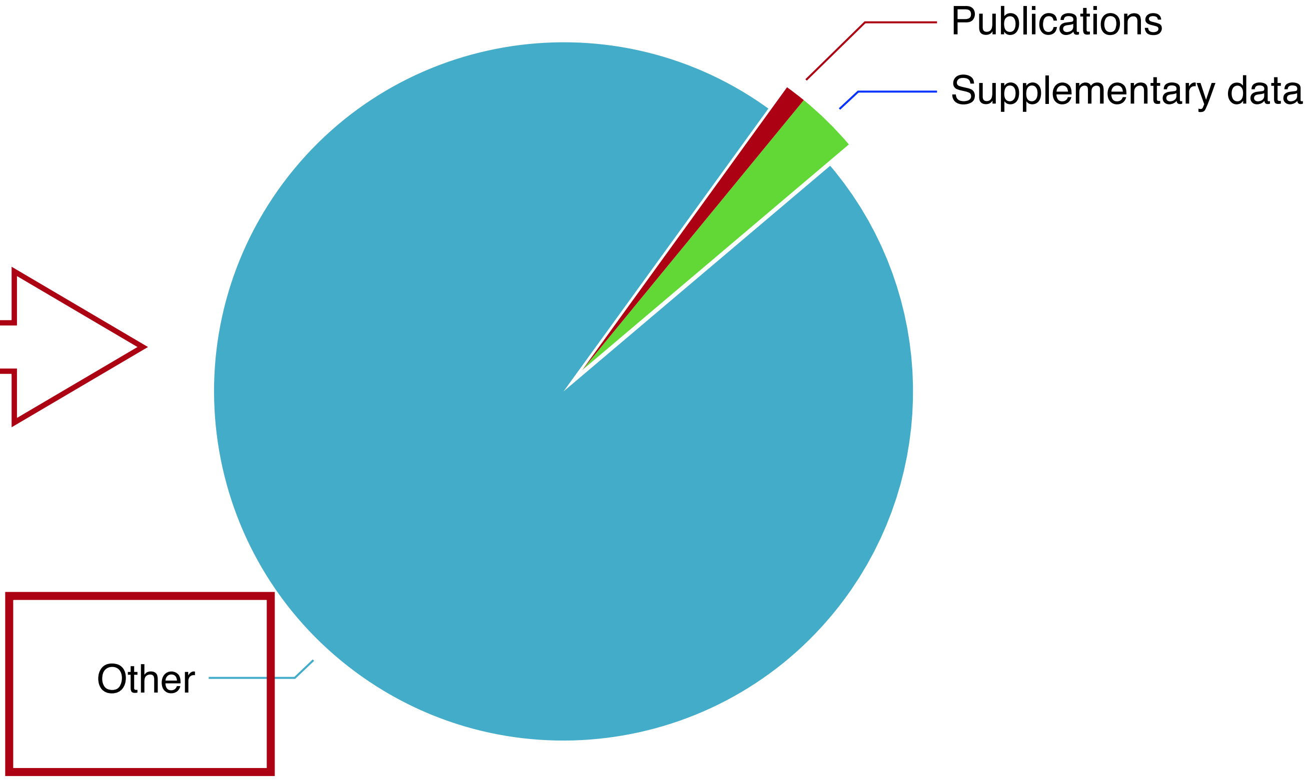
Test new hypotheses
Reproducibility
More training data (tool dev.)
Save $$$
Generate value

Where does research data go?

Research Data



Fate of Research Data Created



* Hypothetical values

Where does the "other" go?

Lab website / Server

404

Page not found :(

The requested page could not be found.

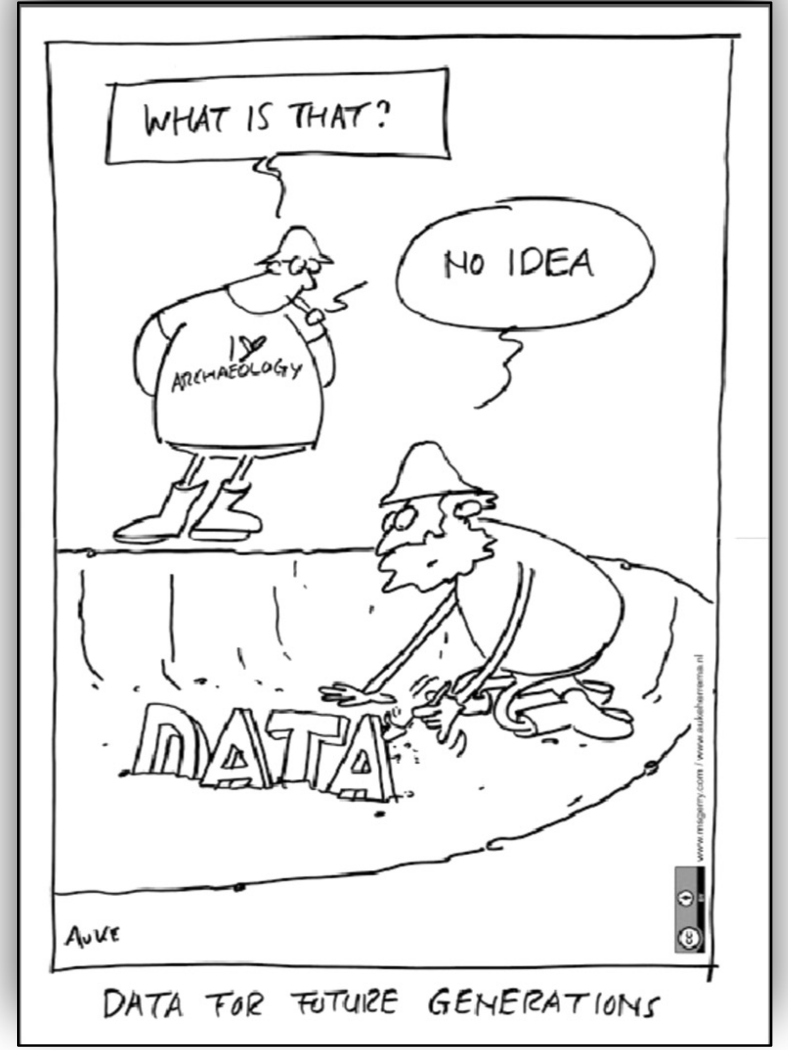


Image by Auke Herrema – Het Bouwteam (CC-BY)

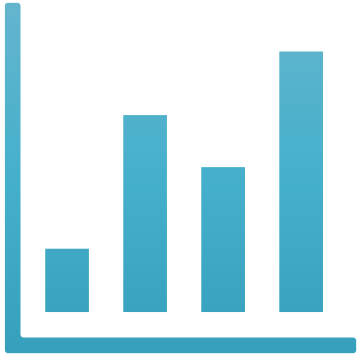
Growing demands in data sharing

- Funder mandates
- Publisher mandates
- University recommendations
- Communities, working groups, consortia
- Researchers' needs

Growth of data sharing in repositories

Creation → Processing + Analysis → Preservation + Sharing → Reuse

Data



KiltHub

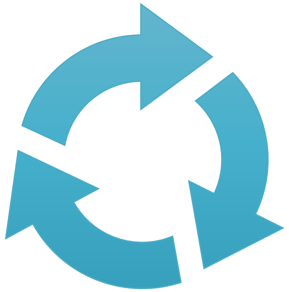
DRYAD

HARVARD Dataverse

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

zenodo

Great!
Can I reuse now?



Sharing ≠ Reusable

- Repositories lack discovery layer across platforms (F)
- Hard to retrieve data (A)
- Not machine / human readable (I)
- Proprietary format (I)
- Lack good metadata and data standards (R)
- Size, complexity, quality, and variability of data (R)

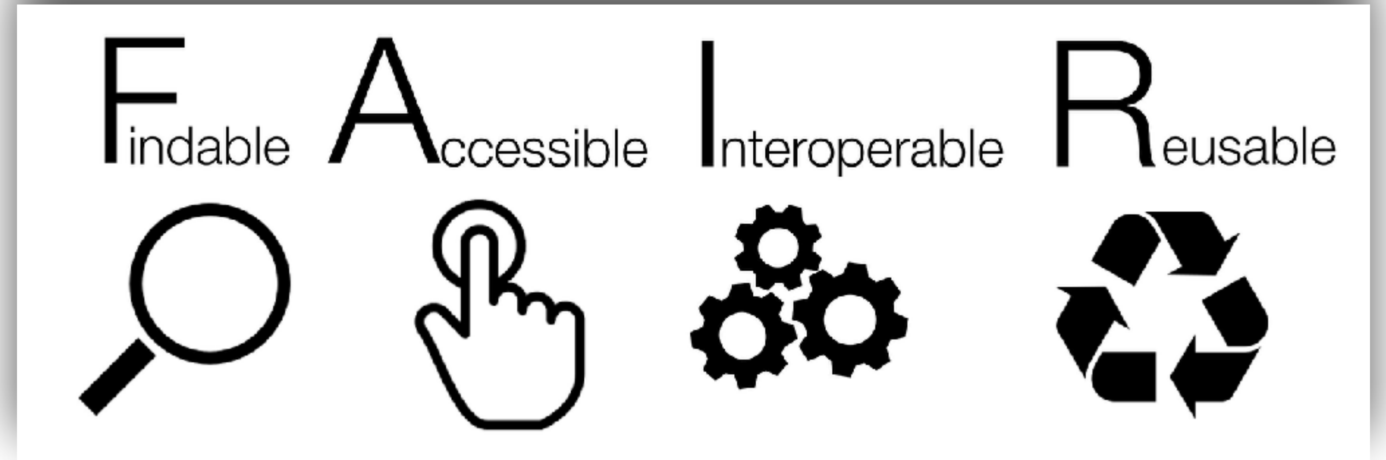


Image by SangyaPundir - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=53414062

NSF 18-060

Dear Colleague Letter: Advancing Long-term Reuse of Scientific Data

April 6, 2018

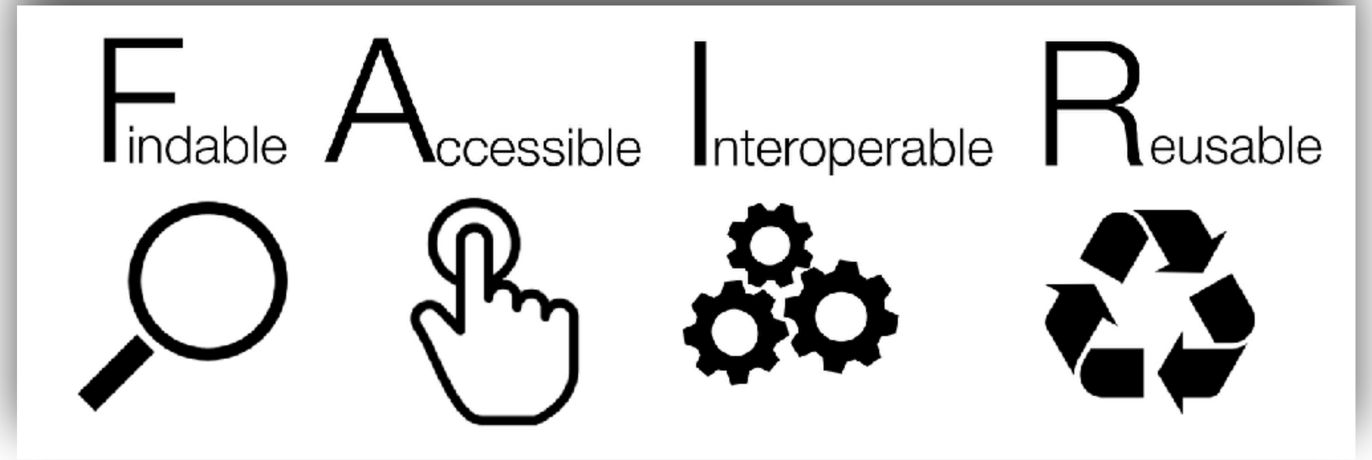
Dear Colleagues:

Through this Dear Colleague Letter (DCL), the National Science Foundation's (NSF) Office of Advanced Cyberinfrastructure (OAC) announces its intention to support initial exploratory activities toward the creation of social and technical infrastructure solutions that further NSF's commitment to public access. These solutions are a means to accelerate the dissemination and use of fundamental research results in the form of data that will advance the frontiers of knowledge and help sustain the Nation's prosperity well into the future.

NSF supports fundamental research grants that result in publications, primary data, samples, physical collections and other supporting materials created or gathered in the course of work performed under these grants [see NSF's Proposal and Award Policies and Procedures Guide (PAPPG) Chapter XI.D.4, https://www.nsf.gov/pubs/policydocs/pappg18_1/pappg_11.jsp#XID4 for details]. This particular DCL is focused on exploratory solutions that advance public access by reducing the barriers to data reuse within the scientific community, as guided by NSF's public access plan, Today's Data, Tomorrow's Discoveries (see https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15052).

Specifically, this DCL encourages two types of funding requests: (1) proposals for Conferences (i.e., community workshops and other events) that are designed to bring together stakeholders to explore opportunities to converge on innovative solutions to advancing public access; and (2) proposals for Early-Concept Grants for Exploratory Research (EAGER) for high-risk/high-reward innovative concepts and pilot projects

How can artificial intelligence help to reuse data?



AIDR 2019

Artificial Intelligence for Data Discovery and Reuse
May 13 -15, 2019
Carnegie Mellon University, Pittsburgh, PA

Event website: https://events.library.cmu.edu/aidr2019/
Slides & posters: https://f1000research.com/collections/aidr

An NSF-supported conference

Co-hosted by Carnegie Mellon University Libraries and Pittsburgh Supercomputing Center

Tom M. Mitchell
Interim Dean
E. Fredkin University Professor
School of Computer Science
Carnegie Mellon University

Glen de Vries
President and Co-founder
Medidata Solutions

Sean Davis
Senior Associate Scientist
National Cancer Institute, NIH

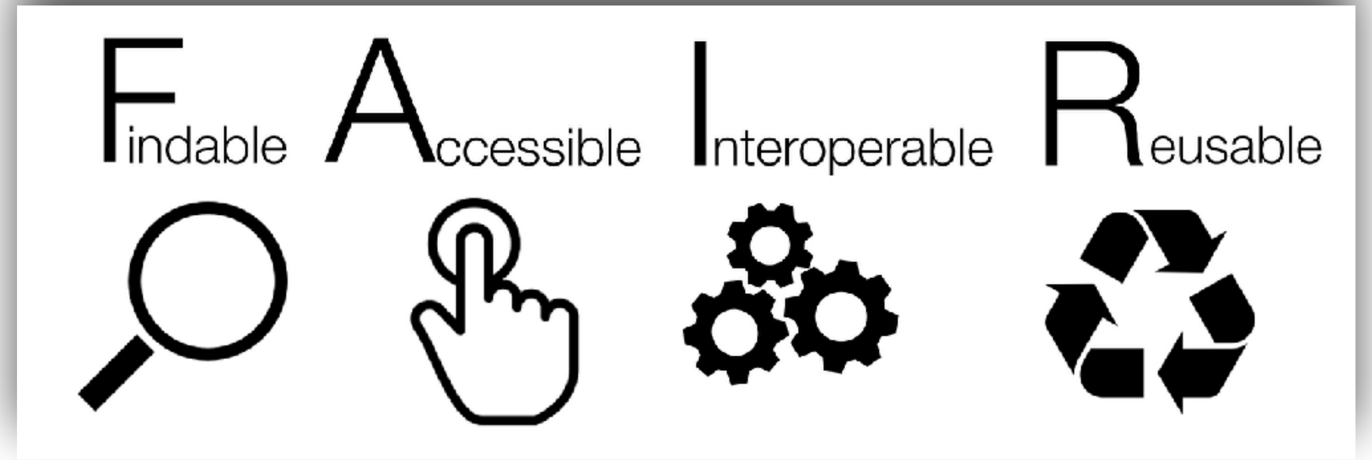
Casey Green
Assistant Professor of Systems Pharmacology and Translational Therapeutics
Perelman School of Medicine
University of Pennsylvania

Robert F. Murphy
Ray and Stephanie Lane Professor
Head of Computational Biology
School of Computer Science
Carnegie Mellon University

Fiona Nielsen
Founder and CEO
Repositive

Natasha Noy
Staff Scientist
Google AI

I. AI for data discovery - Findable



Datasets are distributed and hard to search

- Structured data
  - Web (1%*)
  - Repositores
- Unstructured data
  - Web (99%)
  - Publications
- 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 apple
    "..."
  ],
```

KiltHub

HARVARD Dataverse

DRYAD

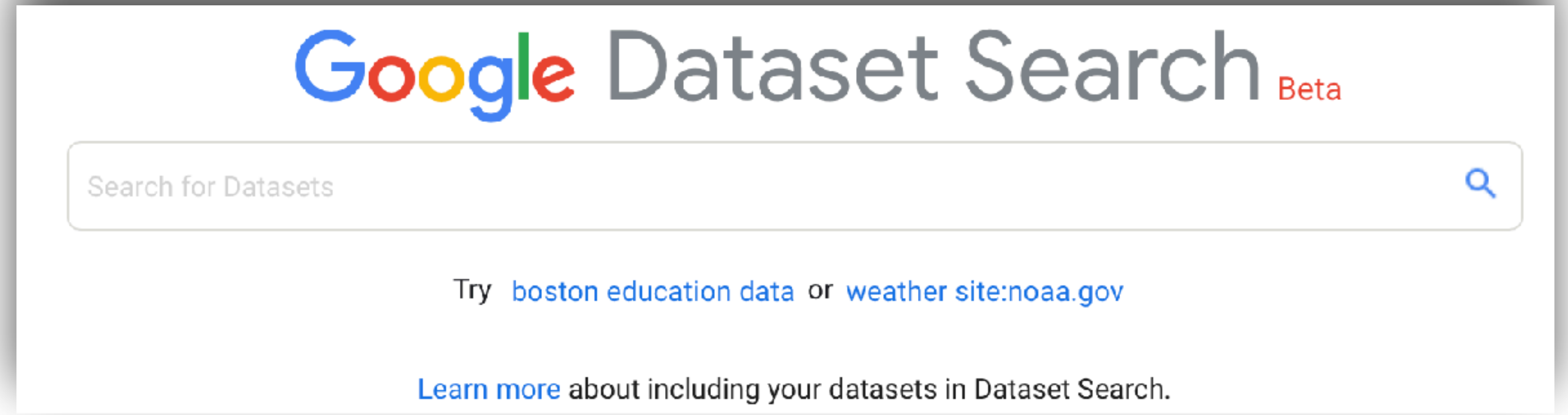
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

zenodo

PubMed

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

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



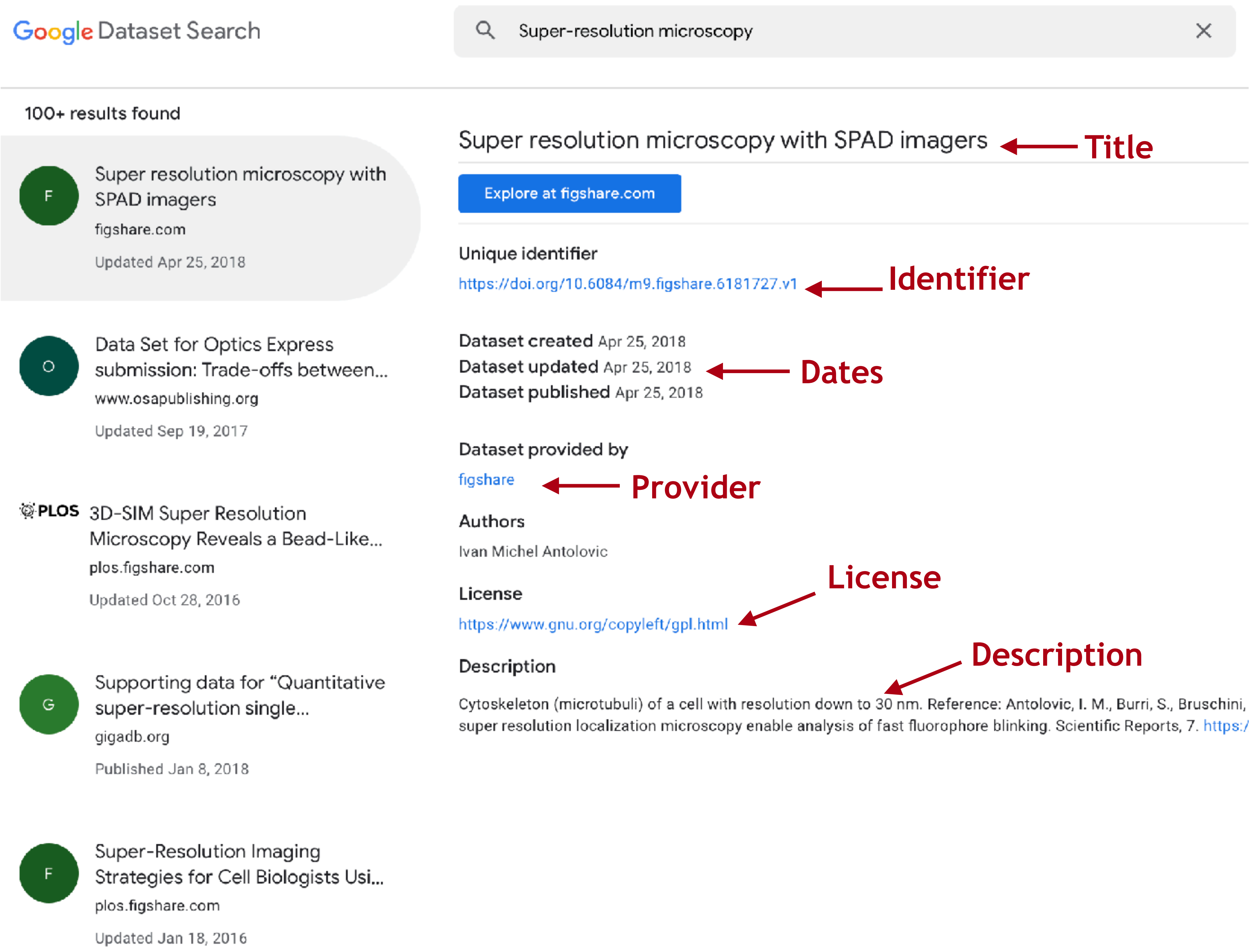
Natasha Noy

Staff Scientist

Google AI

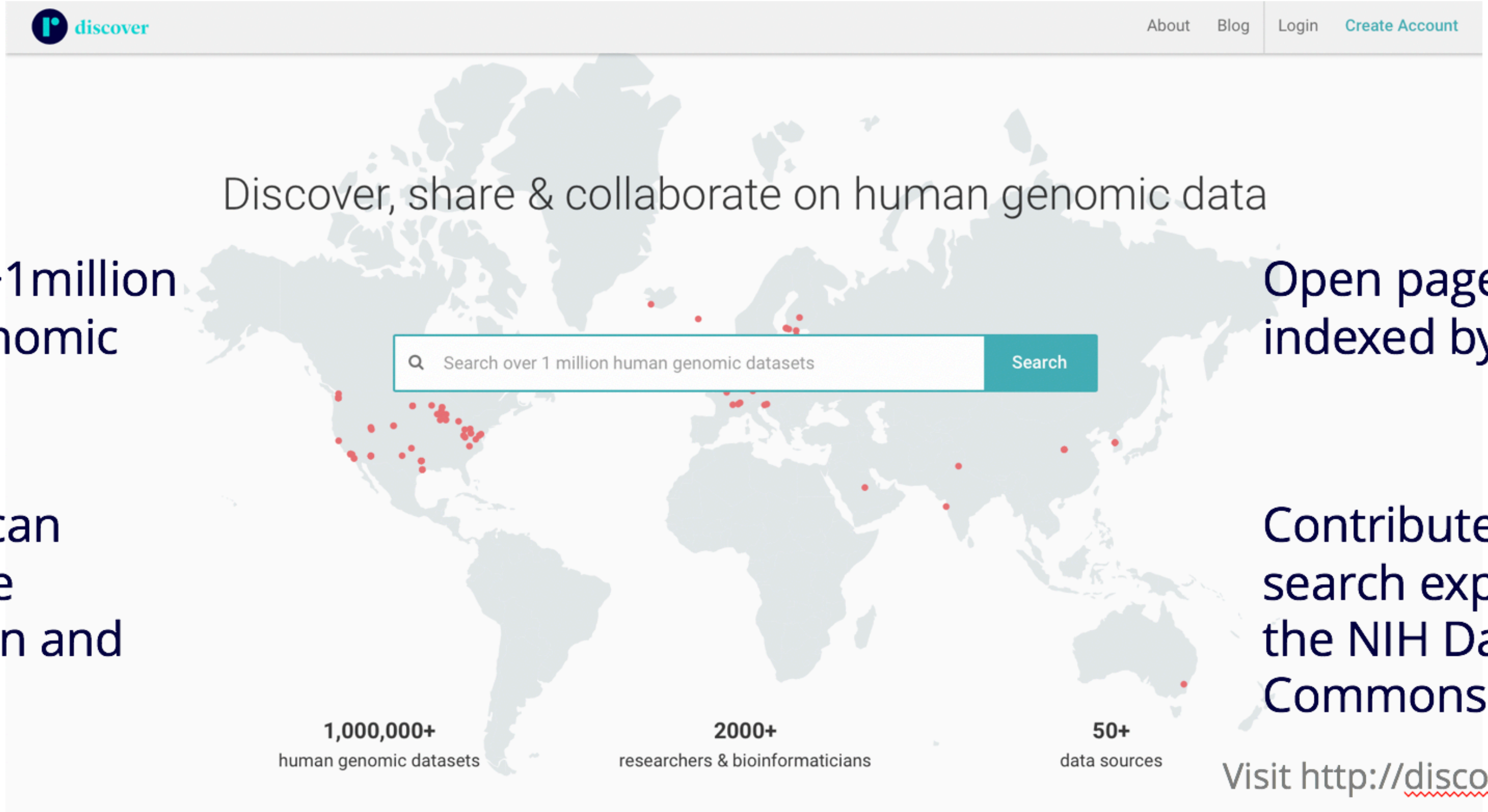
Structured data

- Simple keyword search for datasets
- Searches over embedded metadata
    - Searches over metadata from data providers
    - schema.org data standards (embedded in html)
    - Dataset name, description, provider, temporal coverage, …


Google Dataset Search
Super-resolution microscopy
100+ results found
Super resolution microscopy with SPAD imagers
figshare.com
Updated Apr 25, 2018
Data Set for Optics Express submission: Trade-offs between...
www.osapublishing.org
Updated Sep 19, 2017
PLOS
3D-SIM Super Resolution Microscopy Reveals a Bead-Like...
plos.figshare.com
Updated Oct 28, 2016
Supporting data for "Quantitative super-resolution single...
gigadb.org
Published Jan 8, 2018
Super-Resolution Imaging Strategies for Cell Biologists Usi...
plos.figshare.com
Updated Jan 18, 2016
Super resolution microscopy with SPAD imagers
Title
Explore at figshare.com
Unique identifier
https://doi.org/10.6084/m9.figshare.6181727.v1
Identifier
Dataset created Apr 25, 2018
Dataset updated Apr 25, 2018
Dataset published Apr 25, 2018
Dates
Dataset provided by
figshare
Provider
Authors
Ivan Michel Antolovic
License
https://www.gnu.org/copyleft/gpl.html
License
Description
Description
Cytoskeleton (microtubuli) of a cell with resolution down to 30 nm. Reference: Antolovic, I. M., Burri, S., Bruschini,
super resolution localization microscopy enable analysis of fast fluorophore blinking. Scientific Reports, 7. https:/

Leveraging structured web data

With Repositive we built a search engine for genomics



Index of >1million public genomic data sets

All users can contribute annotation and data sets

Open pages – all indexed by Google

Contributed our data search expertise to the NIH Data Commons Pilot

Visit http://discover.repositive.io

Fiona Nielsen

Founder and CEO

Repositive

repositive

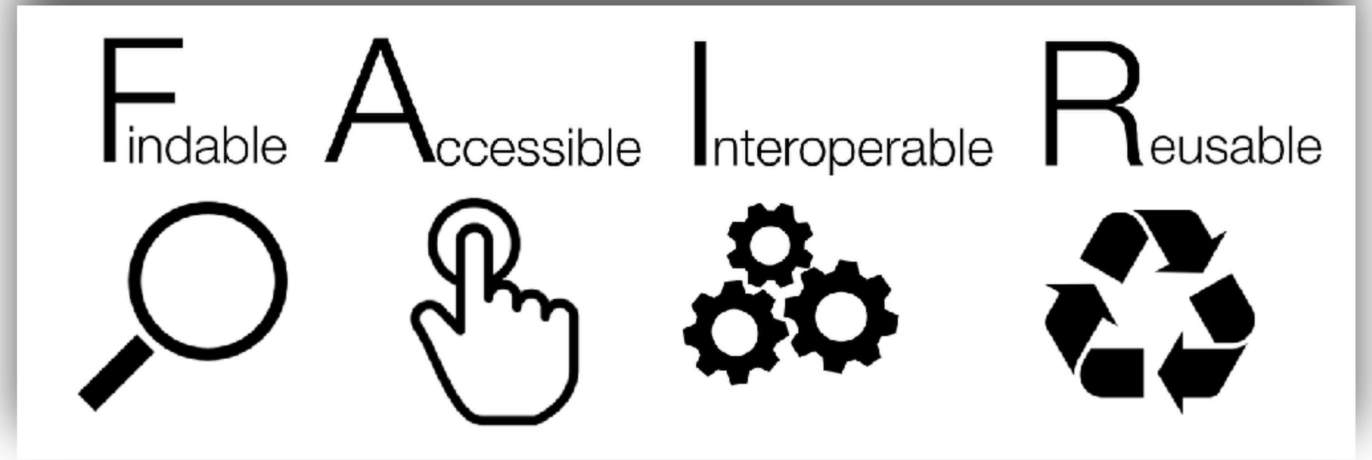
All metadata curated and indexed are findable in Google Dataset Search

What about unstructured data?

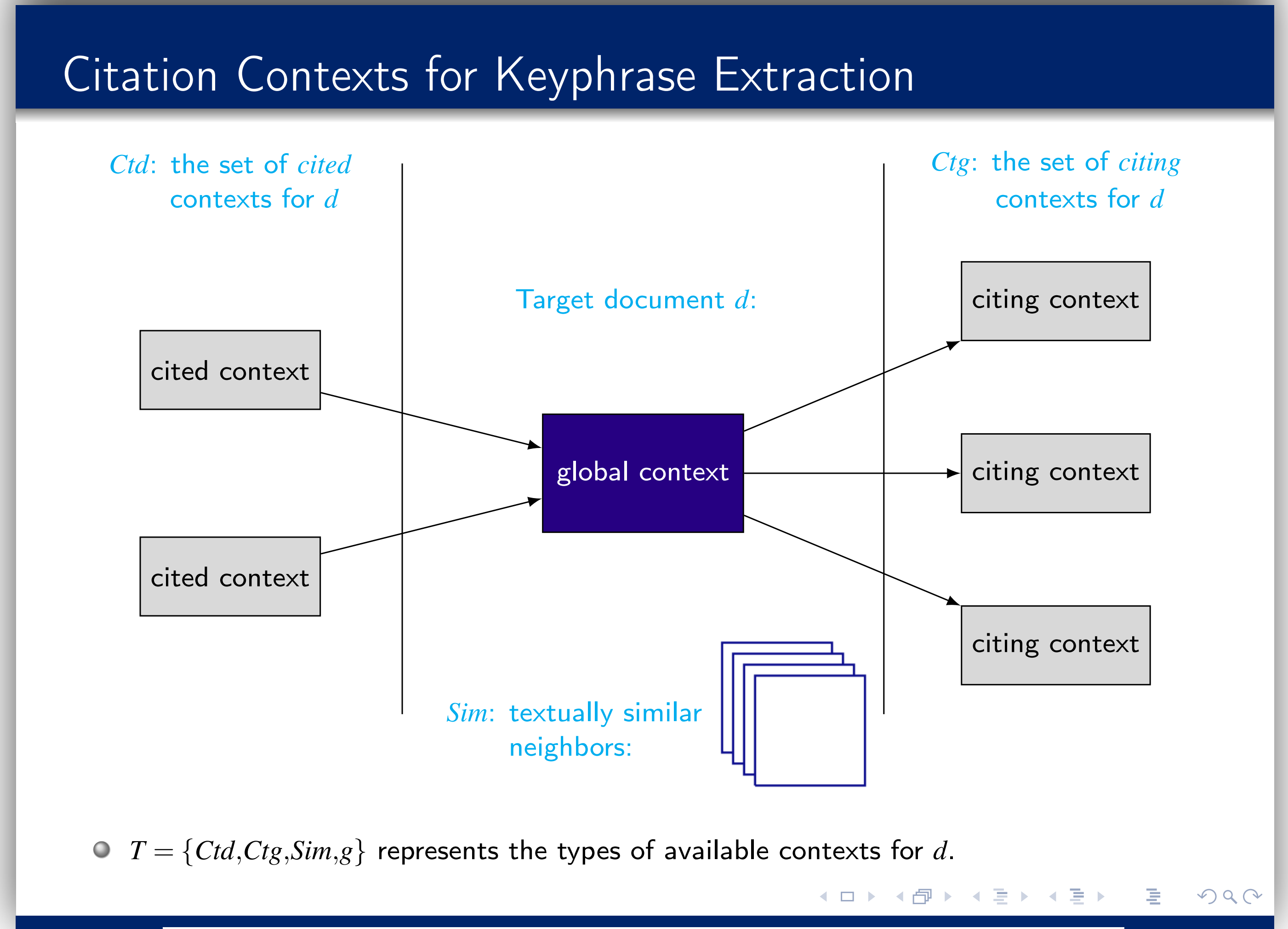
- Scholarly publications
- Images
- Unstructured websites
- Poor metadata

Need metadata tagging and data linking first

II. AI for data curation and metadata generation



Keyphrase extraction from scholarly documents



Cornelia Caragea and C. Lee Giles. AIDR 2019.

Reuse keyphrase:

Document discovery

Classification

Author characterization

Dataset discovery?

...

Image recognition for archaeological research

Large archaeology image data

With sparse metadata

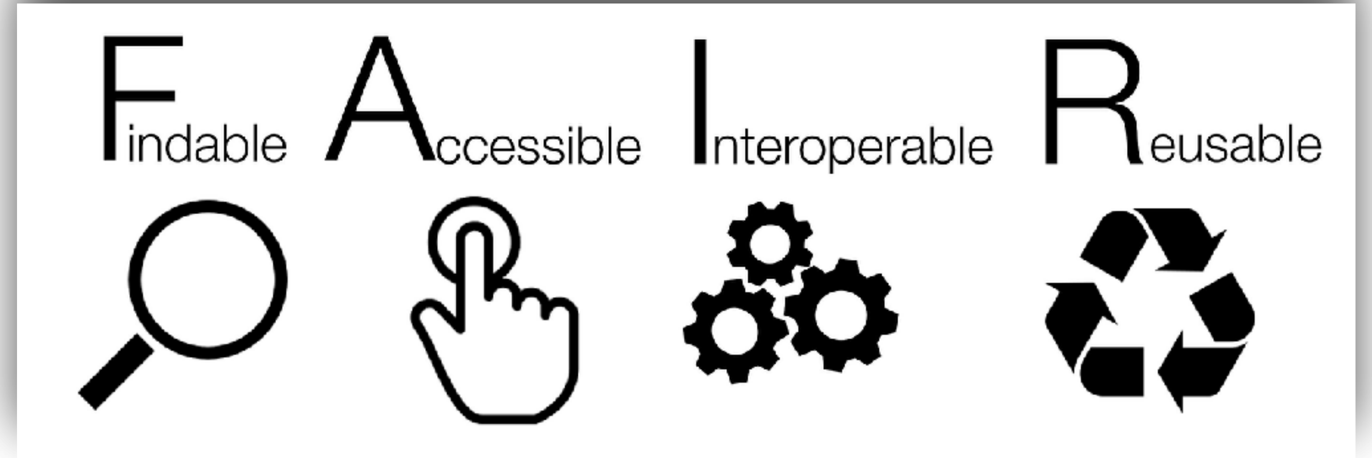
FD and FT and KW	5456
FD and FT only	2555
FD and KW only	58682
FD only	15196
FT and KW only	401
FT only	200
KW only	13832
none	49023

Need to extract metadata for archiving & classification

```
m@了 。4Ɑ4 CH unit: x find: 80 materia:bed initials/date:
50 cm
west trench 5 b 105 18372,18370 z1 08 ptw
hzoiu u 21722m 30.07.14 n. sunken floor
c.h 1996 1653 h. 5 cm
u12646 cut jpg 22 06 06
for oor
ch 09 oumy 1857 334 x2
1 0 cm
20 crm digplan 90 ɯ1301
c,h 2003 7s74x cm
ist ch06 unit 12(87 stace 294 tu 2307.04
20 cm 20 cm 2
4040 cho3 unit: x find: material: initials/date: BeH s 24.08
1834
çatal höyük 2004 15 10 0cm 1 2 3 4 5
chit south 32693 .1 f 8189 sp 620 oy-na 1201仔
se pae covering the ice sunce c extronce area coamel lotee
cho4 4 640
mellet. 225.3/30 c m
tm 13159. h
ch2012 d: 20379 b:97 sp :365 midden loyer at the se coner of sout…
b52
chit7 north
1st ch unit:1 892 x find: material: clay initials/date: . tu 30.0…
055
ch 2013 burial hill cri 29-713
kaui
20 cm 20 cmm
hob 13140,x3
çhii 40 40 f3615 sk 19 sol ru 11512. s. 336 ltot 02 / 08 / 11
ch08 4040 n 6751 fill space 93
su n s453 unit: find/sample: material: initials/date: 又 co-51 cm
tpc u32337 ovemi base
ho8 tp unit: find/sample material initials/date: oe a 16.os.o8 cm
939 959 40
```

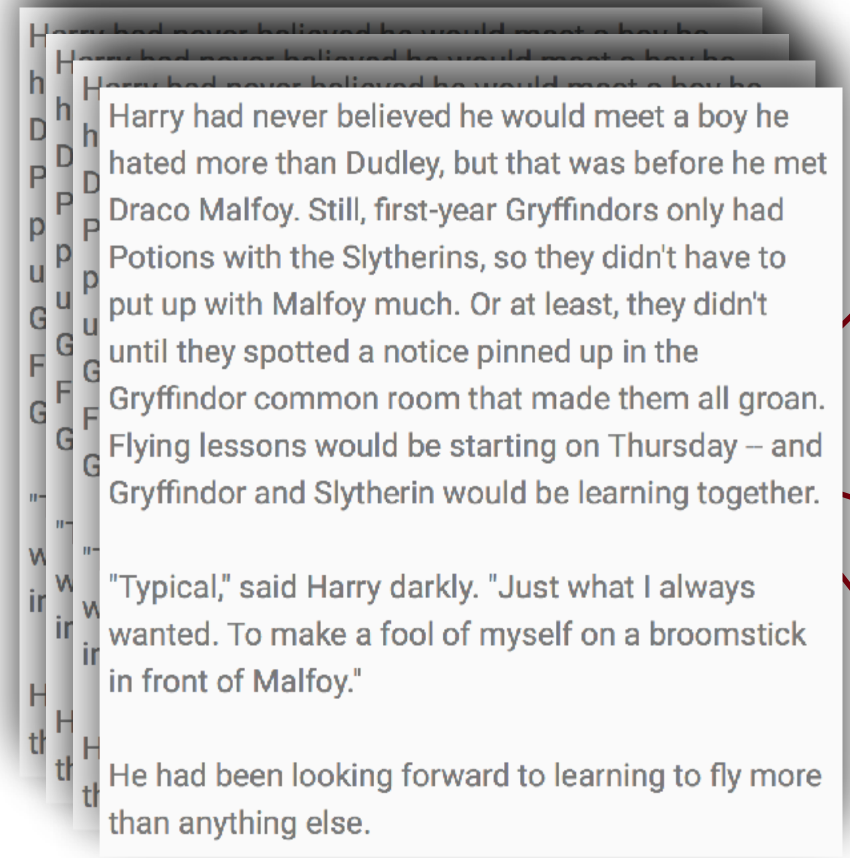
Claudia Engel & Peter Mangiafico. AIDR 2019.

III. AI for integrating datasets



Integrating heterogenous data sources to predict brain activity

Large text data

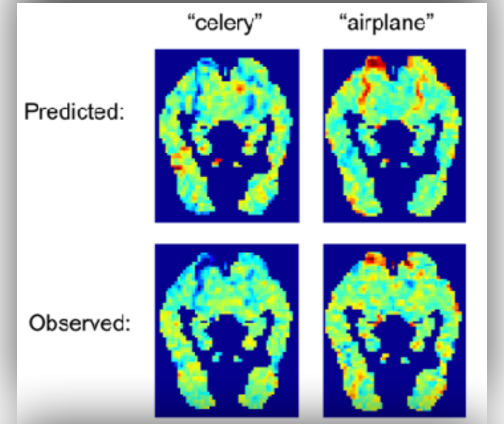

Harry had never believed he would meet a boy he hated more than Dudley, but that was before he met Draco Malfoy. Still, first-year Gryffindors only had Potions with the Slytherins, so they didn't have to put up with Malfoy much. Or at least, they didn't until they spotted a notice pinned up in the Gryffindor common room that made them all groan. Flying lessons would be starting on Thursday – and Gryffindor and Slytherin would be learning together.

"Typical," said Harry darkly. "Just what I always wanted. To make a fool of myself on a broomstick in front of Malfoy."

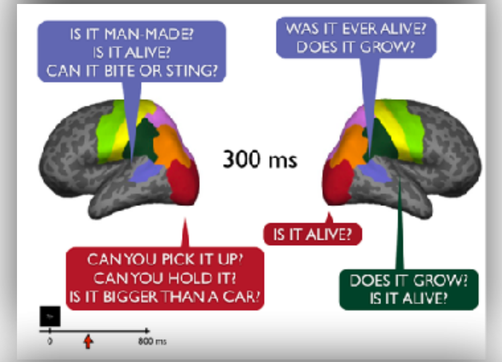
He had been looking forward to learning to fly more than anything else.

Moderate brain data

fMRI



MEG



Other

ECoG, EEG

Build a program that understands sentences, and predicts neural activity

Tom M. Mitchell

Interim Dean

E. Fredkin University Professor

School of Computer Science

Carnegie Mellon University

Sample scarcity - Integrating genomics data for rare diseases

- Problem: Machine learning model needs many samples; rare diseases have few samples. - Solution: Model transfer.

Casey Greene

Assistant Professor of Systems Pharmacology and Translational Therapeutics

Perelman School of Medicine

MultiPLIER

recount2

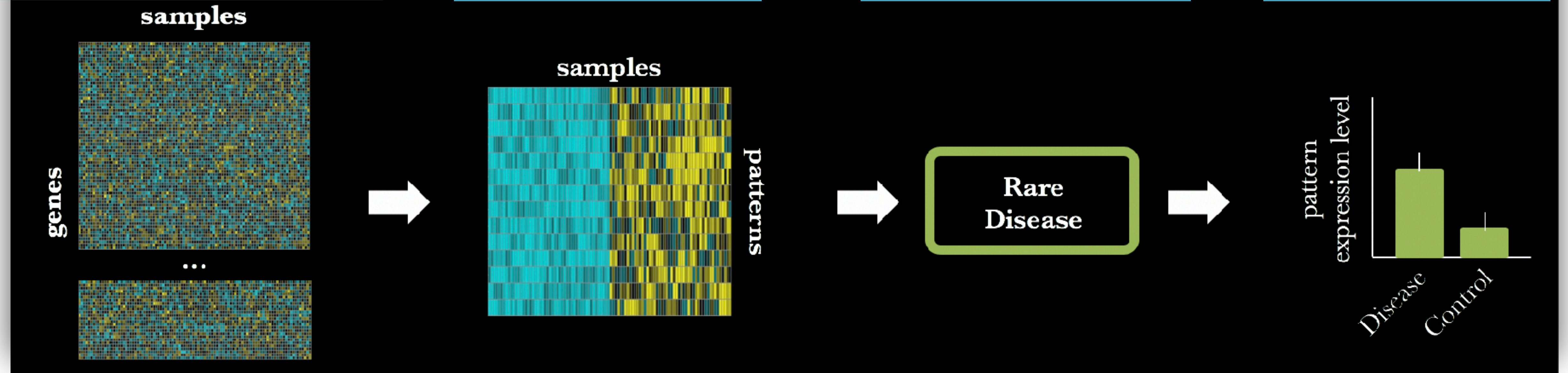
PLIER

Gene expression data from various public repositories (70,603 samples)

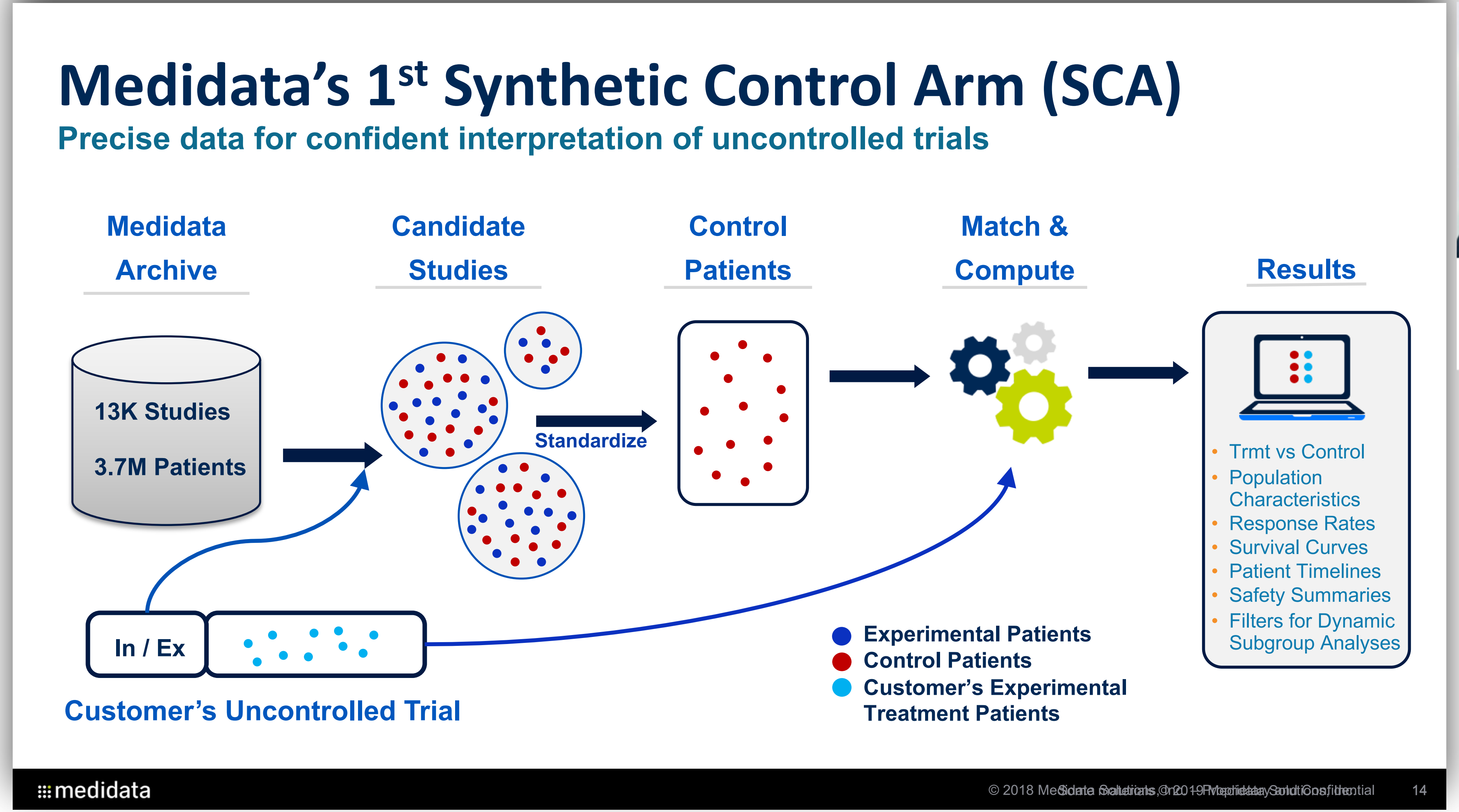
Extract pathway patterns

Model transfer to rare diseases data

Patterns ctrl vs disease



Clinical trials - data augmentation with synthetic controls

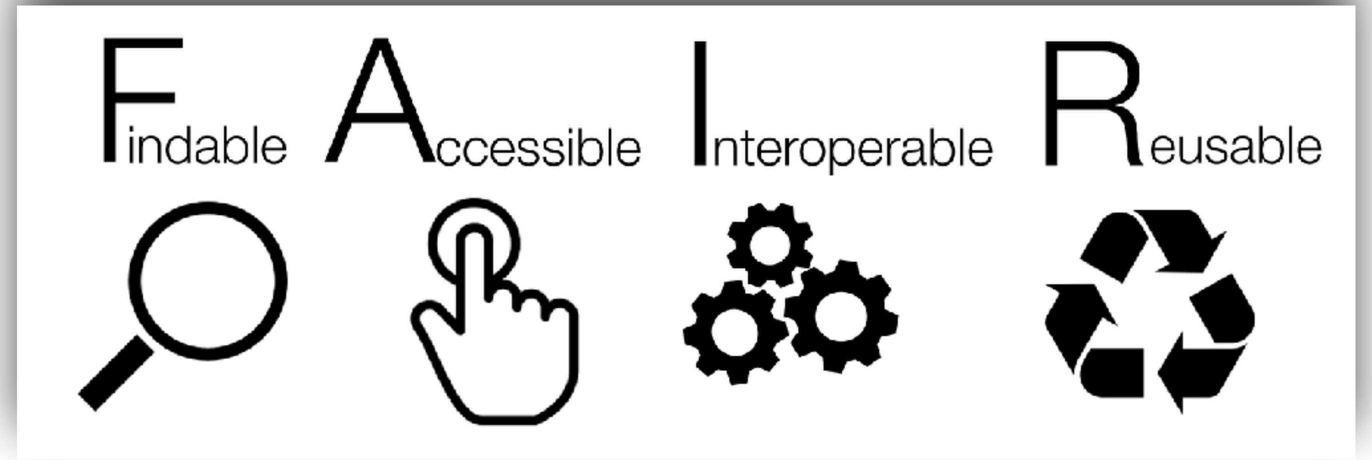


Glen de Vries

President and Co-founder

Medidata Solutions

IV. The future of scientific data and how we work together



Big opportunity in data integration

- Big opportunity 2: jointly analyze data from many experiments
  - have hundreds of fMRI, MEG data sets involving language processing
  - Hurdle: unclear how to jointly analyze
  - Hurdle: neuroscience culture of not sharing data
  - Hurdle: lack formal language to specify experiments

Need tools

Need an open culture

Data standards / Controlled vocabularies

"The greatest difficulty in cognitive neuroscience is to document every detail of the experiment, and to document in a way that computers can understand."

Tom M. Mitchell

Interim Dean

E. Fredkin University Professor

School of Computer Science

Carnegie Mellon University

Consensus on how to overcome hurdles

“Machine learning and AI is only the last piece of the puzzle; before we get there, we need to first build a healthy data ecosystem.”

“Culture and incentive for data stewardship, and open, non-proprietary data standards are the key.”

“If you care about the impact you want to make …, have to care about making easy to use tools and fixing the incentives.”

Open access tool development

Lightweight data engineering, tools, and approaches to facilitate data reuse and data science

...

Sean Davis, MD, PhD
National Cancer Institute, National Institutes of Health
AIDR 2019, Carnegie Mellon University

https://seandavi.github.io
@seandavis12
http://bit.ly/SD-AIDR2019

Biomedical Data Science in the 21st Century

Prototype Software for Machine Learning Analysis of Human Genomes, Variants, and Expression!

Ben Busby, Hackathon Participants
NCBI Hackathons Program

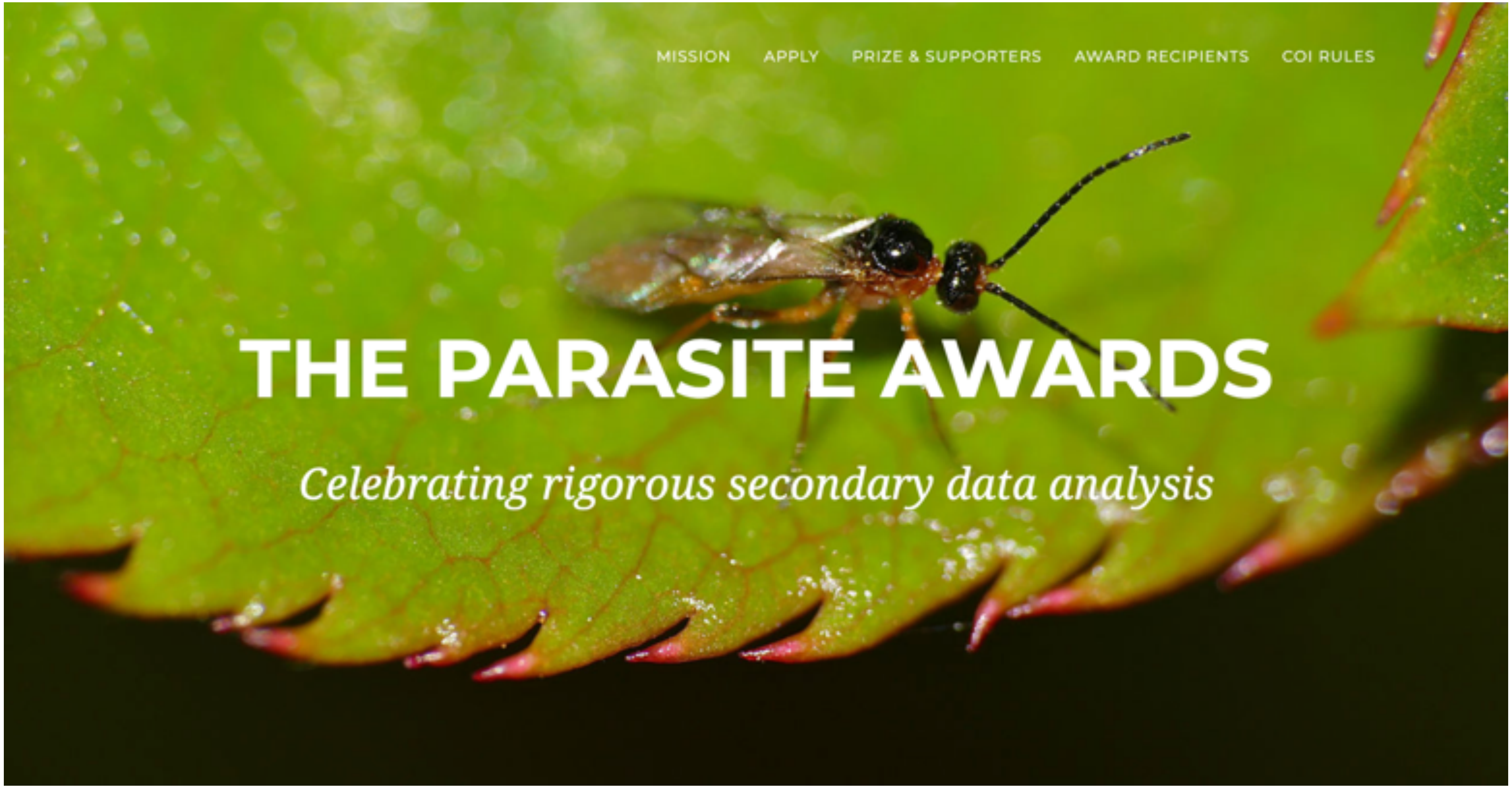
Ariel Precision Medicine, Johns Hopkins, Deloitte
@dcgenomics, LinkedIn

NIH U.S. National Library of Medicine

NCBI

https://f1000research.com/collections/aidr

Incentives for data sharing and reuse



https://researchparasite.com/

Developing community data standards is everyone's job

My small contributions:

- Data curation network - confocal images curation primer
- Data reuse initiative @ eLife ambassador program - survey for (eg. microscopy images) data standards to come!

Find your small (or big) contributions too!

DATA CURATION NETWORK

Home About

Confocal Microscopy Data: A Primer for Curators

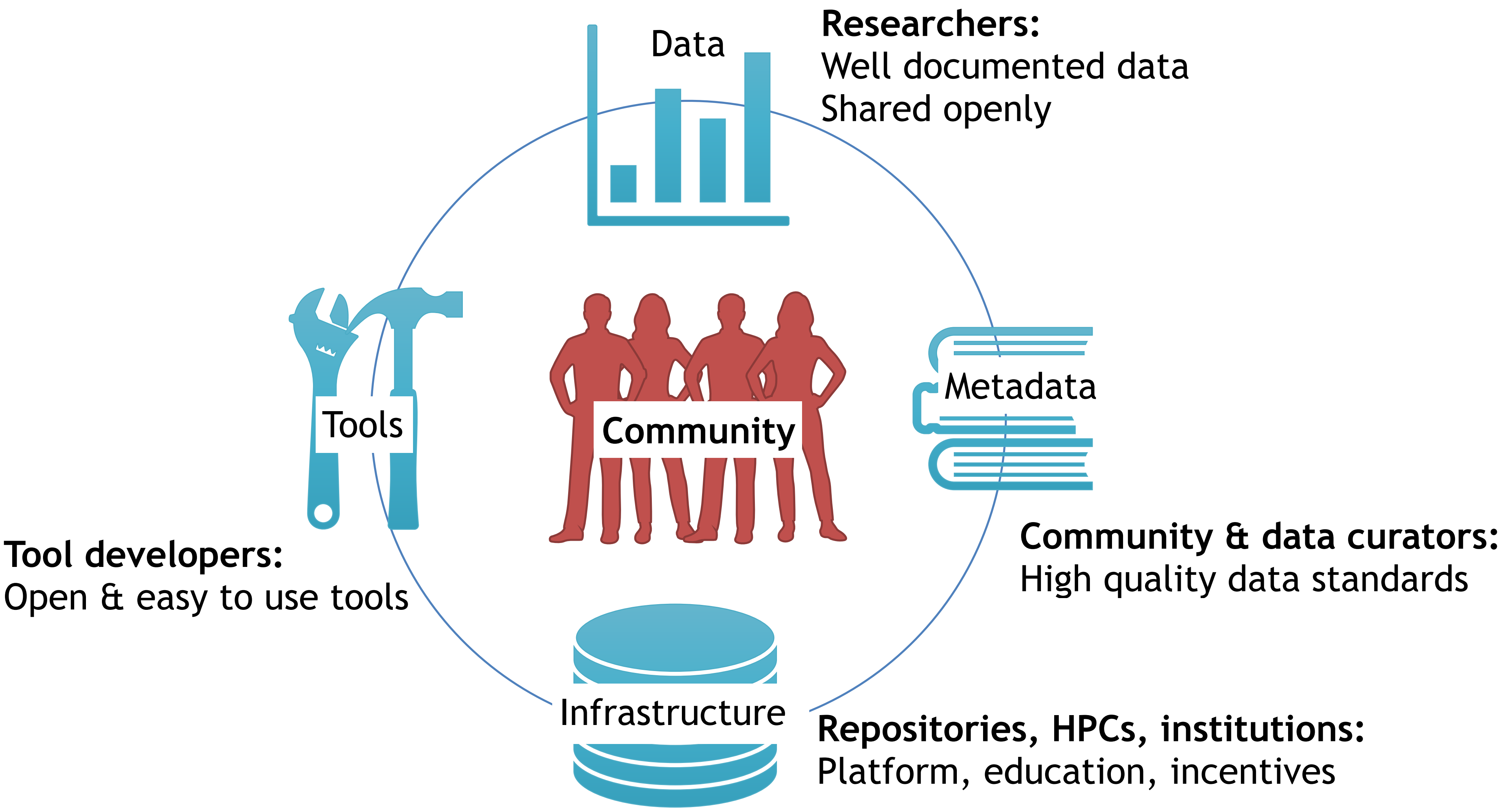
Susan Ivey- NC State University
Amy Koshoffer - University of Cincinnati
Gretchen Sneff - Temple University
Huajin Wang - Carnegie Mellon University
Team Mentor - Lisa Johnston - U. Minnesota

eLIFE Community Ambassadors

Data Reuse

Bradly Alicea @bradly.alicea
Marije Verhage @m.l.verhage
Bhavik Nathwani @bhavik.nathwani
Huajin Wang @huajinw
Sarvenaz Sarabipour @ssarabi2

Data reuse: work as a community to build a healthy data ecosystem



Thank you! Tack!

Join us at next AIDR: May 10-12, 2020

huajinw@cmu.edu

@HuajinBioLib