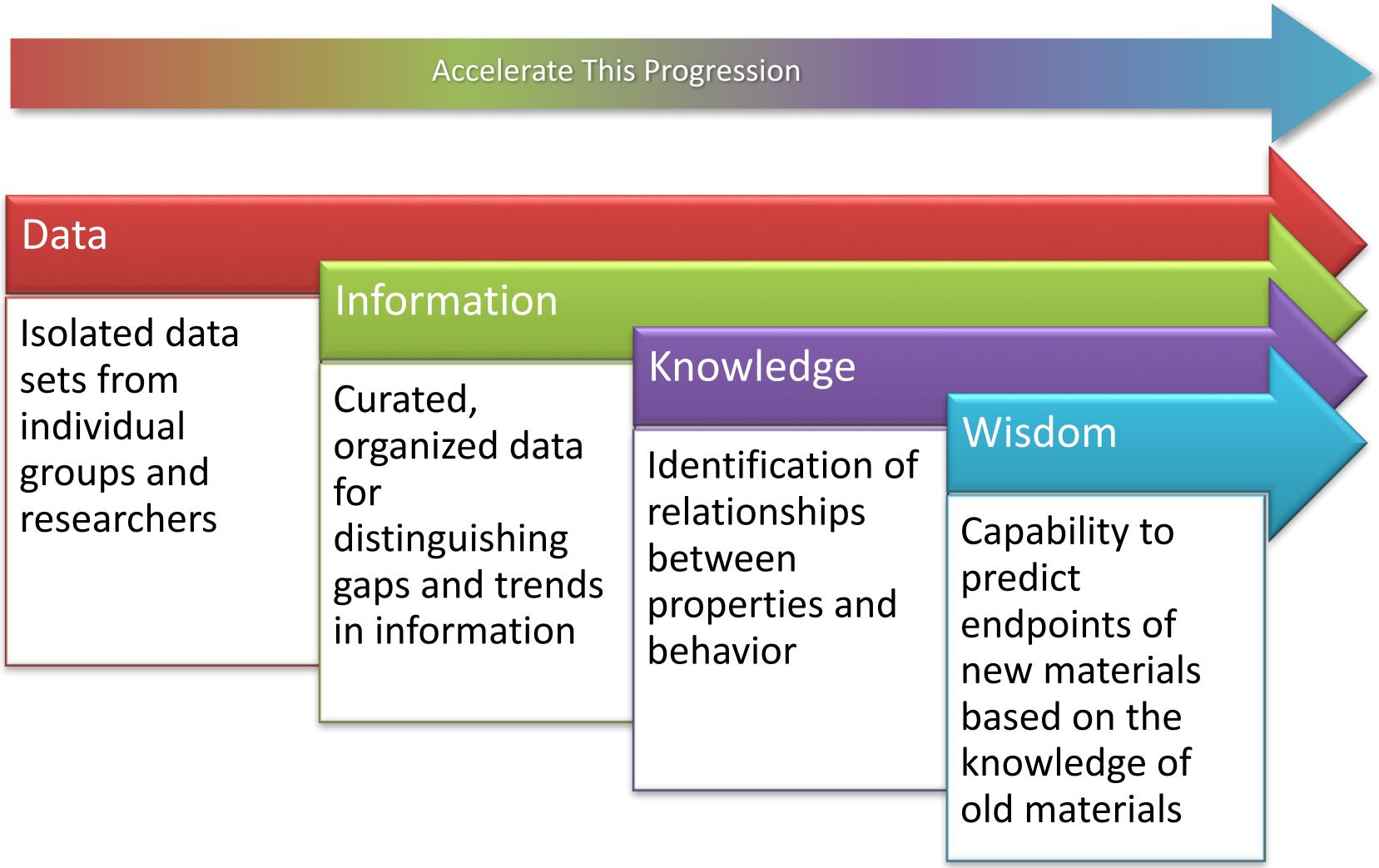
UNC
ESHELMAN
SCHOOL OF PHARMACY

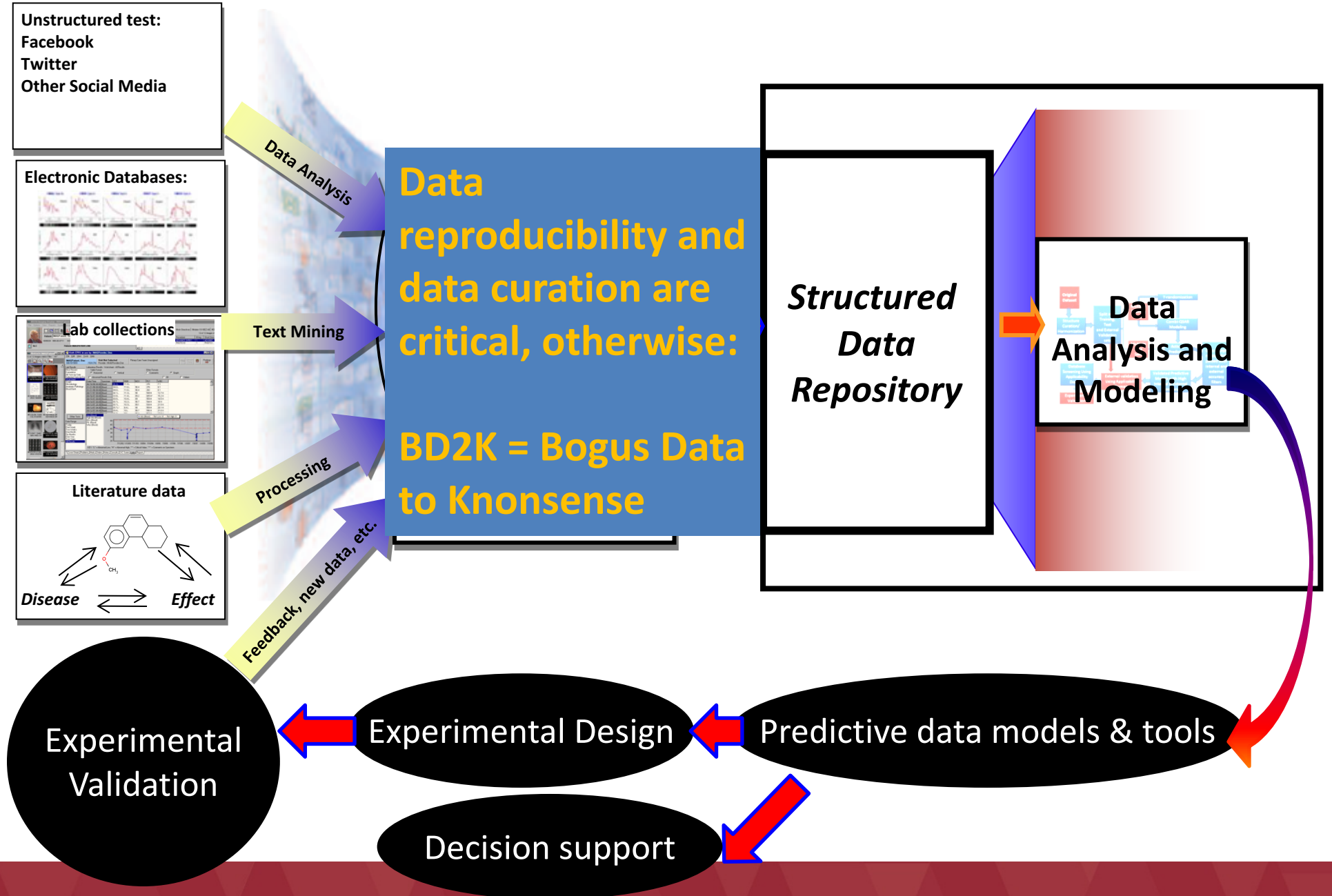
Data Sharing, Nanoinformatics, and eNanoBook

Alexander Tropsha, UNC-Chapel Hill
Fred Prior, University of Arkansas for Medical Sciences
With Contributions from Valery Tkachenko and Rick Zakharov, Science Data Software, LLC

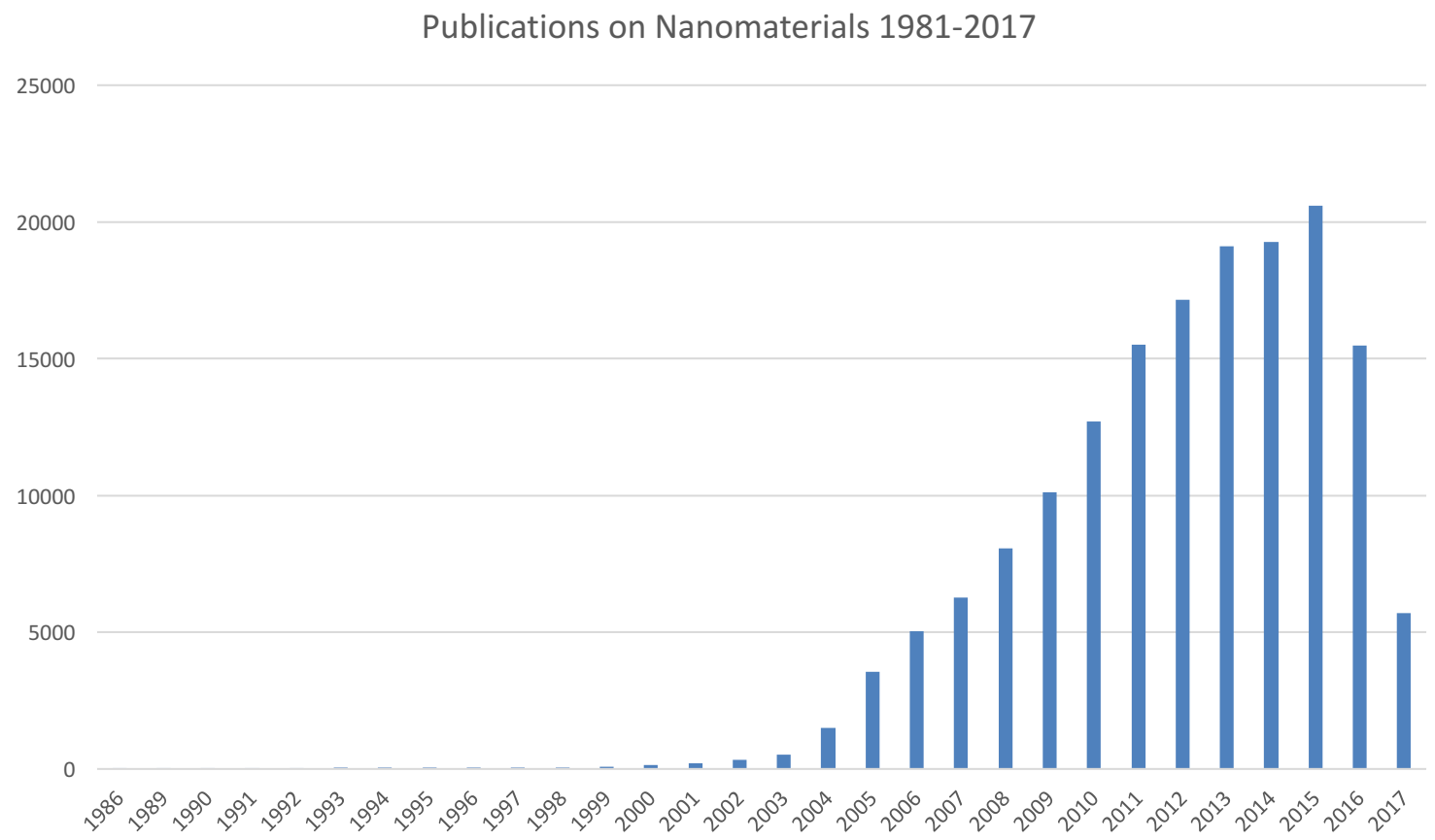
Data to wisdom progression in nanotechnology



Data Science and data cycle



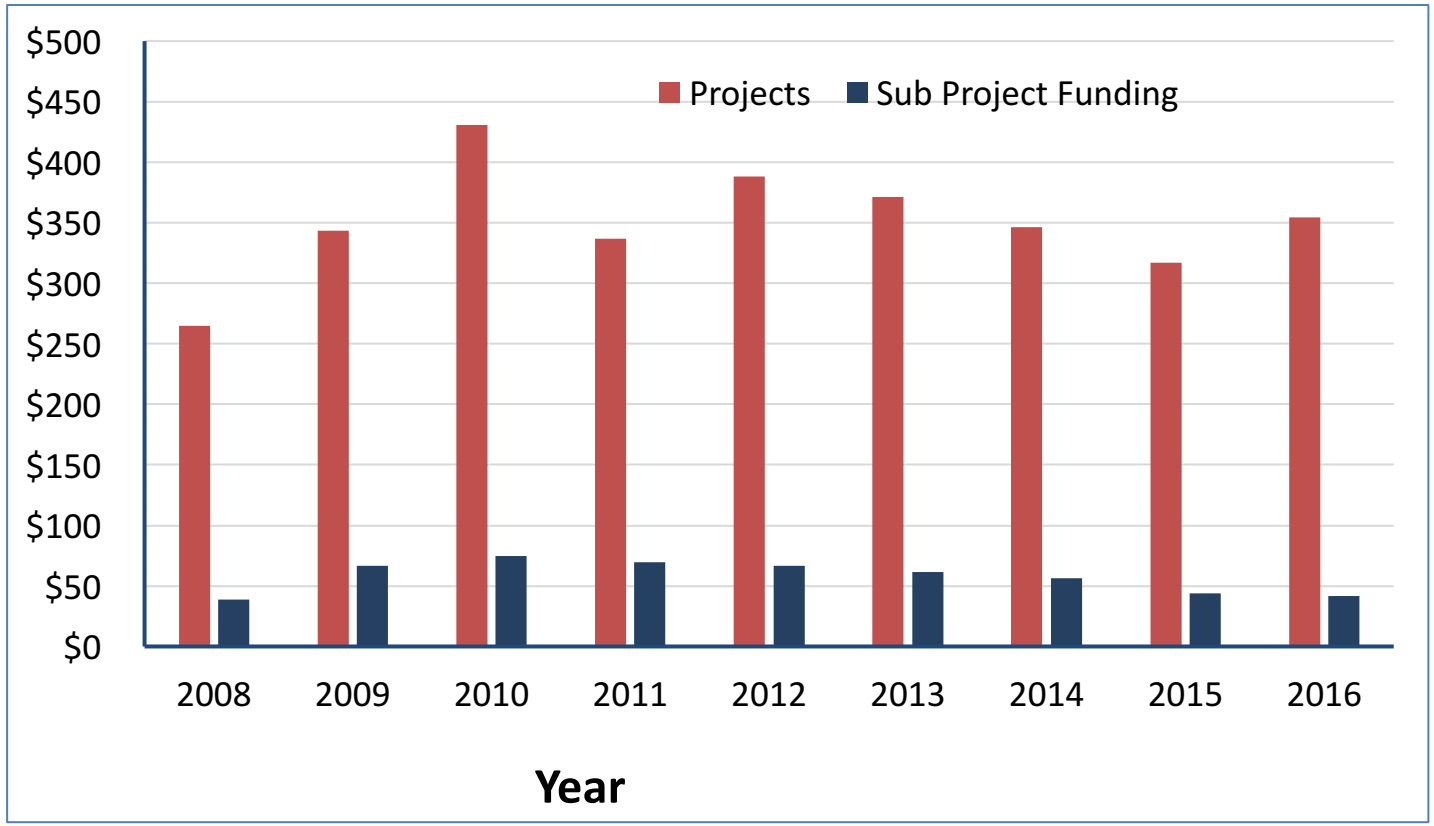
Growth in publications on nanomaterials from 1981 (1 paper) to 2017 (**161704** papers total)*



*Data based on Pubmed analysis as of 6/21/2017

NIH Funding for Nanotechnology

In Fiscal Years 2008-2016 NIH spent $3.7B to fund Nanotechnology-related projects ($2.2B on Cancer Nanotechnology); data was reported in 2,885 publications (590 on nanotechnology)



Data obtained from NIH reporter (https://projectreporter.nih.gov/reporter.cfm)

How much data across NIH*?

- Big Data
  - Total data from NIH-funded research currently estimated at 650 PB*
  - 20 PB of that is in NCBI/NLM (3%) and it is expected to grow by 10 PB this year
- Dark Data
  - Only 12% of data described in published papers is in recognized archives – 88% is dark data^
- Cost
  - 2007-2014: NIH spent ~$1.2Bn extramurally on maintaining data archives

*Courtesy of Dr. Phil Bourne, founding Assoc. Director of NIH for Data Science

* In 2012 Library of Congress was 3 PB

^ http://www.ncbi.nlm.nih.gov/pubmed/26207759

Transition To Fred

Principles and Guidelines for Reporting Preclinical Research

- Results of a 2014 NIH workshop with editors of major journals
- Consensus principles to enhance rigor and reproducibility
  - Rigorous statistical analysis
    - Testable hypotheses, appropriate statistical models and tests, justified sample sizes
  - Transparency in reporting
    - full description of methods
  - Data and material sharing
    - “all datasets on which the conclusions of the paper rely must be made available”
  - Consider establishing best practice guidelines for:
    - Image based data, description of biological materials

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

NIH: Rigor and Transparency in Research

To support the **highest quality science, public accountability, and social responsibility in the conduct of science**, NIH's Rigor and Transparency efforts are intended to clarify expectations and highlight attention to four areas that may need more explicit attention by applicants and reviewers:

- Scientific premise
- Scientific rigor
- Consideration of relevant biological variables, such as sex
- Authentication of key biological and/or chemical resources

Slide from NIH training materials for reviewers

Rigor + Transparency -> Reproducibility

Research Reproducibility: the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator.

"Documenting this kind of reproducibility thus requires, at minimum, the sharing of analytical data sets (original raw or processed data), relevant metadata, analytical code, and related software."

Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean?. Science translational medicine. 2016 Jun 1;8(341):341ps12-.

NIH Policy: Rigor and Transparency - Module 1 (02:22 / 33:24) Resources Exit

COMMENT

612 | NATURE | VOL 505 | 30 JANUARY 2014

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring[1,2]. As leaders of the US National Institutes of [IH), we share this concern and re some of the significant inter- hat we are planning.

has long been regarded as 'self- ', given that it is founded on the n of prior work. Over the long principle remains true. In the shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today's researchers to reproduce others' findings.

Let's be clear: with rare exceptions, we have no evidence to suggest that irreproducibility is about scientific misconduct. In 2011, the Office of Research Integrity of the US Department of Health and Human Services pursued only 12 such cases[3]. Even if this represents only a fraction of the actual problem, such papers are vastly

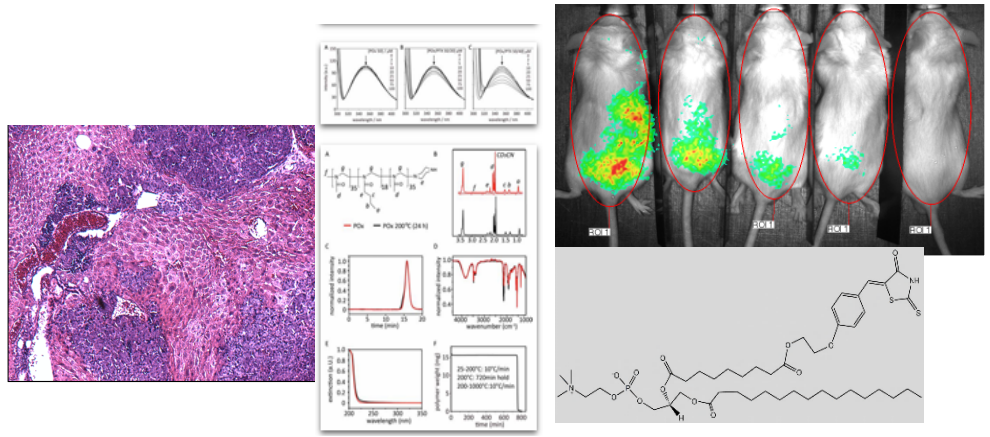
Collins FS, Tabak LA. Nature. 2014 Jan 30;505(7485):612-3.
Resource Guide is available internally to NIH Staff.

Data Reusability

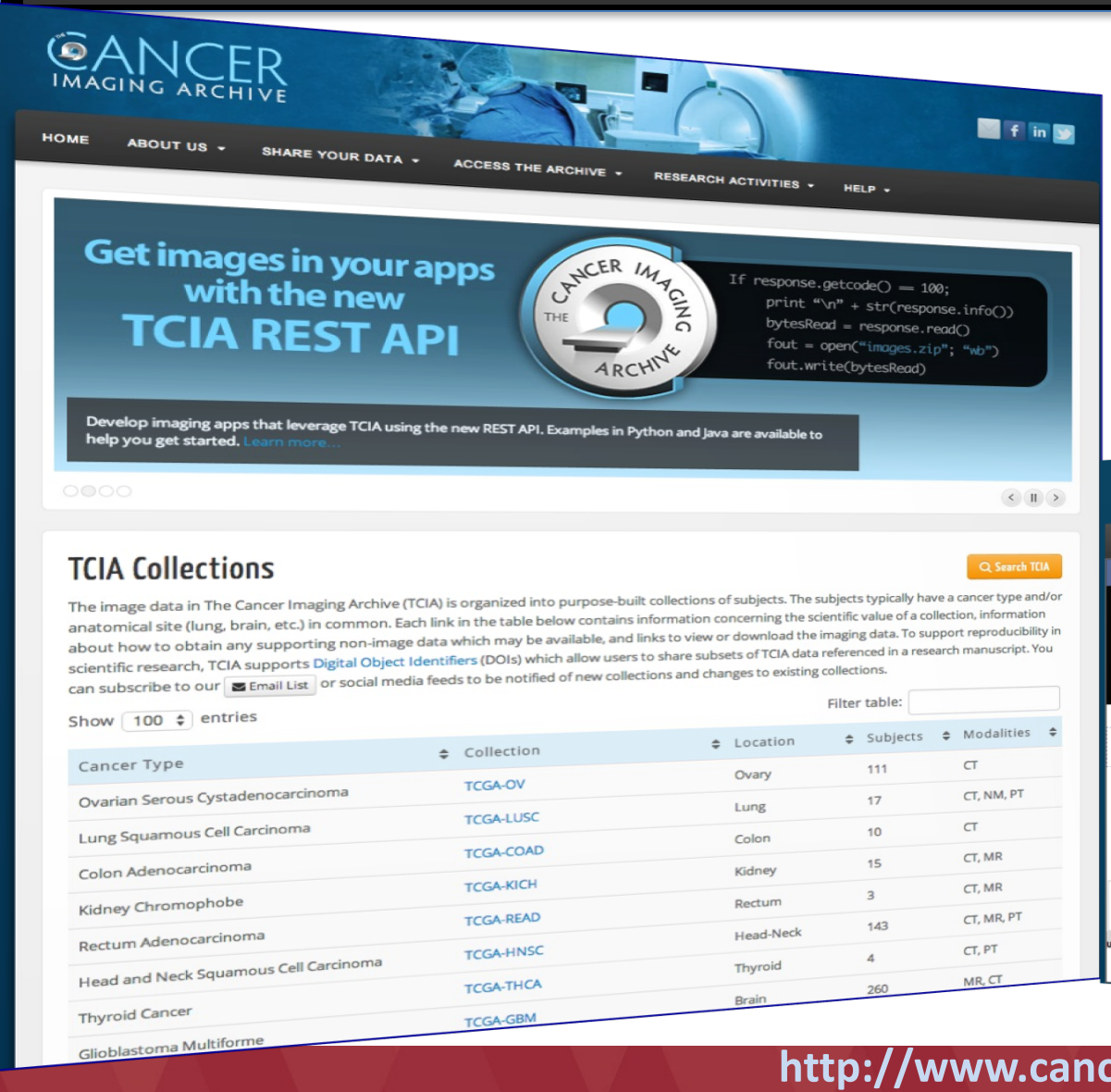
- Making biomedical research data more accessible also supports:
  - Exploration of secondary research aims
  - Testing and validation of new quantitative analysis algorithms
  - Establishment of larger patient cohorts from multi-site data sets
  - Development of methods to address variability in acquisition protocols and hardware

Data Management Perspective

- Manage Protocols for synthesis and characterization of nanomaterials and for pre-clinical studies
- Collect and manage in vitro and in vivo characterization data
  - Images from transmission electron microscope (TEM) and dynamic light scattering (DLS) experiments and analysis results for morphological characterization
  - Confocal microscopy images
  - HPLC Analysis
  - Clinical chemistry and hematology data
  - Histopathology results and images
  - Mass spectrometry results
  - Flow cytometry and other assays
  - PET images and related data including animal weight, disease burden, and clinical pathologies, including serum biomarkers
  - Images and analysis results of epifluorescence, fluorescence and bioluminescence imaging studies
- Cross-link experimental results, especially imaging findings (both pre-clinical and histopathology) to DNA and RNA sequencing results of MM clones
- Provide direct data access to the research team and the Biostatistics Resource Core.
- Protect Intellectual property by keeping information secure and releasing it to the public at the discretion of the PI

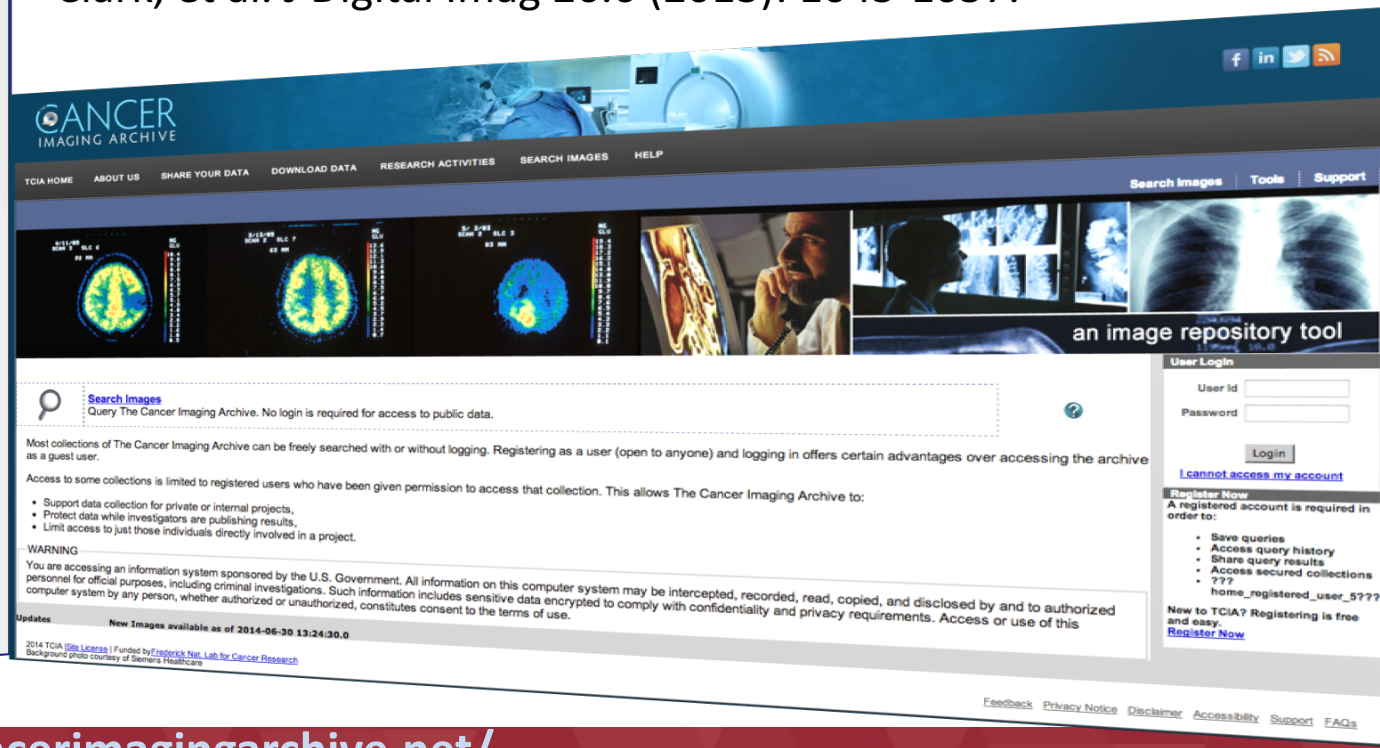


THE CANCER IMAGING ARCHIVE

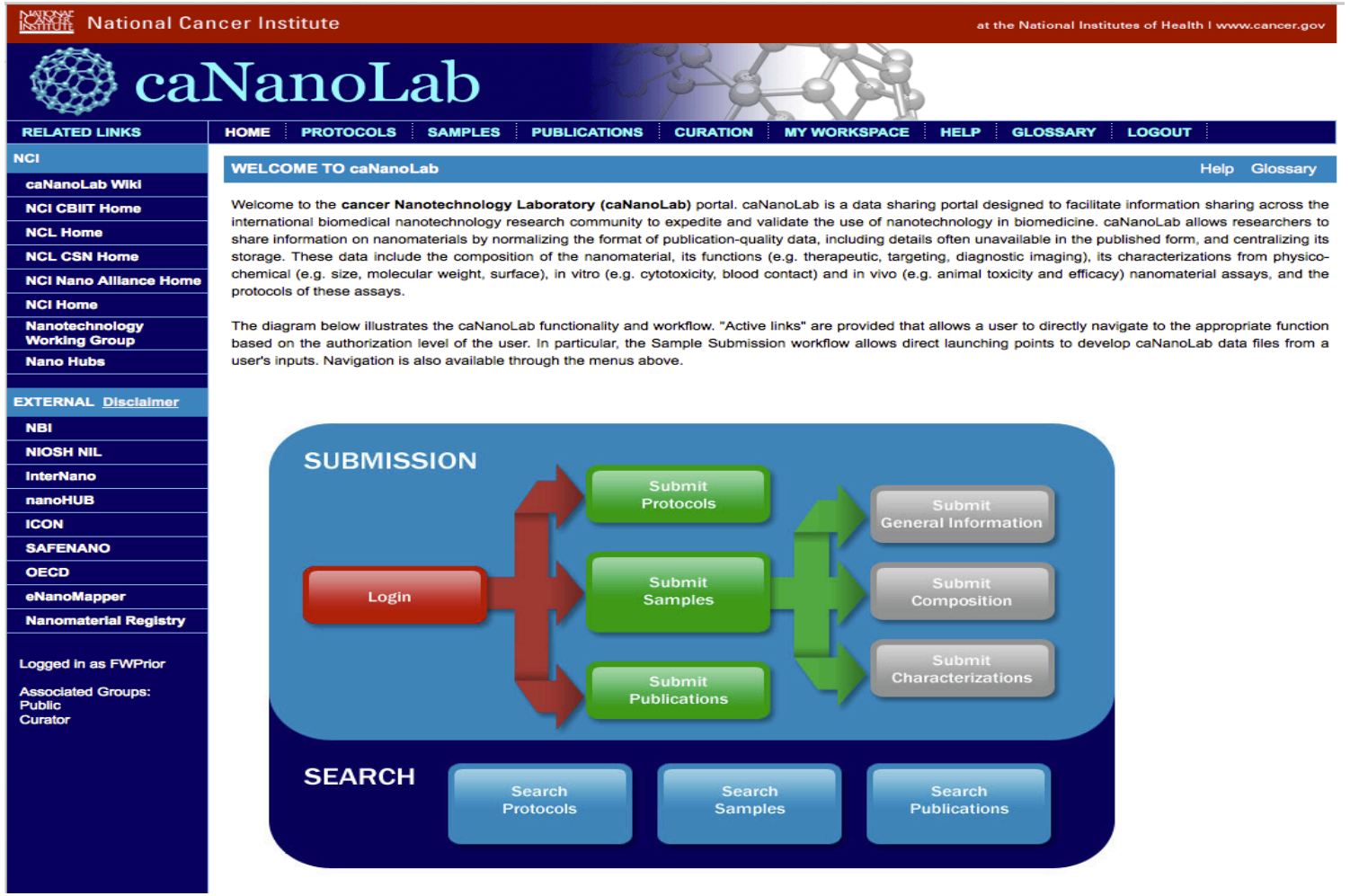


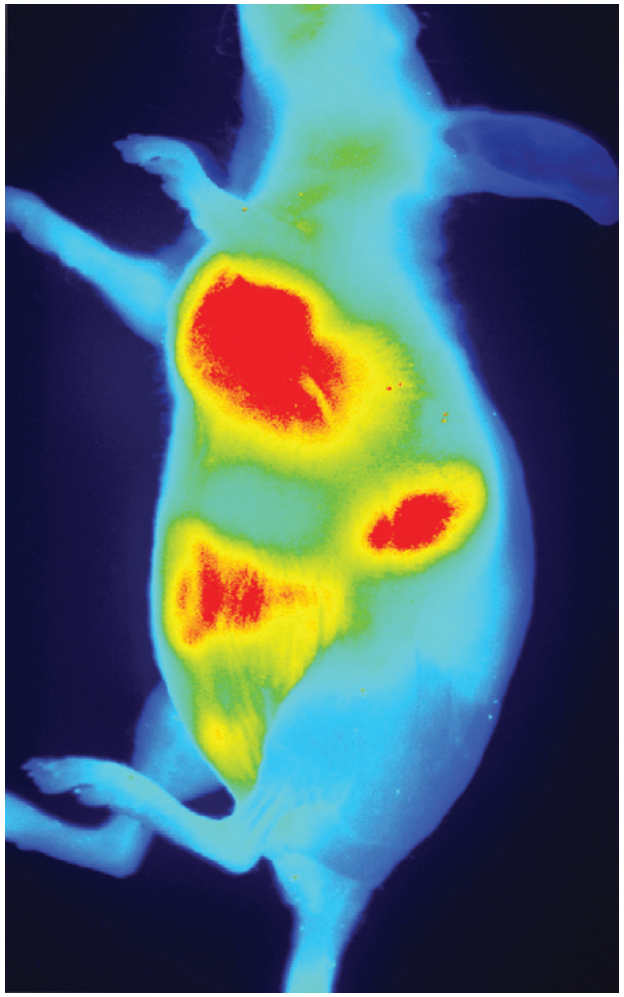
TCIA encourages and supports the cancer imaging open science community by hosting and managing Findable Accessible, Interoperable, and Reusable (FAIR) images and related data.

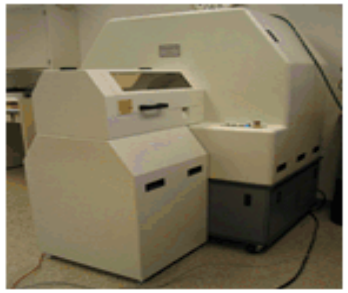
Clark, et al. J Digital Imag 26.6 (2013): 1045-1057.

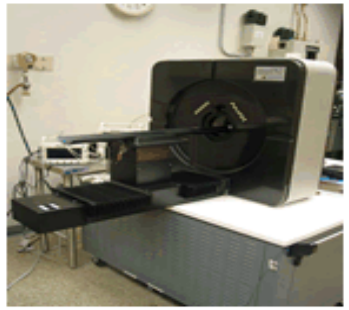


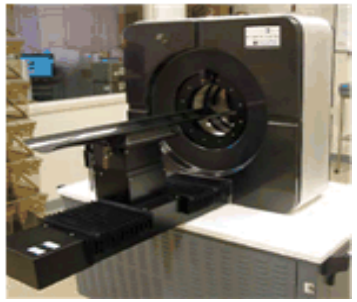
Integration of caNanoLab & TCIA







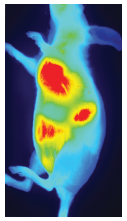


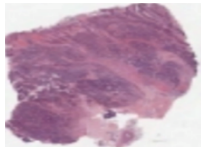


Integrated Infrastructure for CMMN

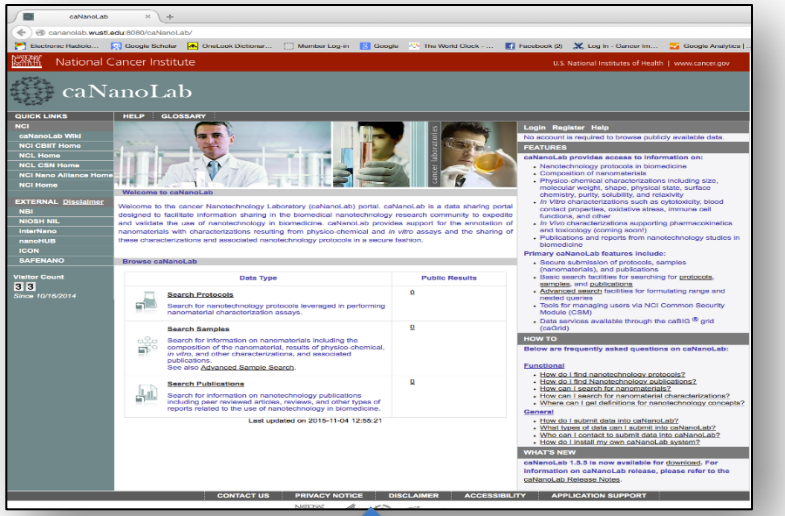
Synthesis protocols

in vitro characterization

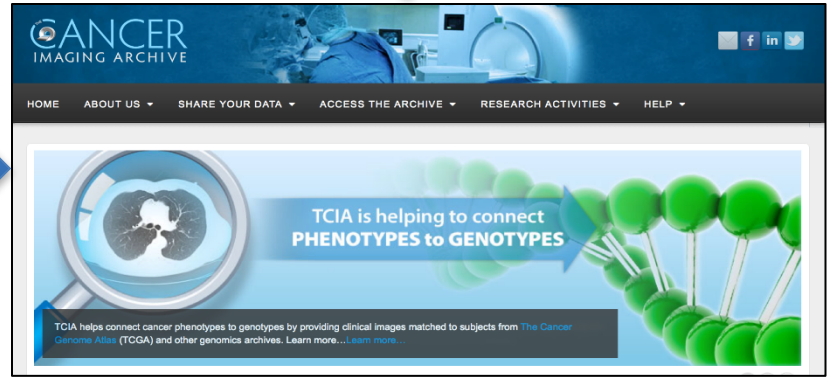




in vivo characterization (Images & animal model data)

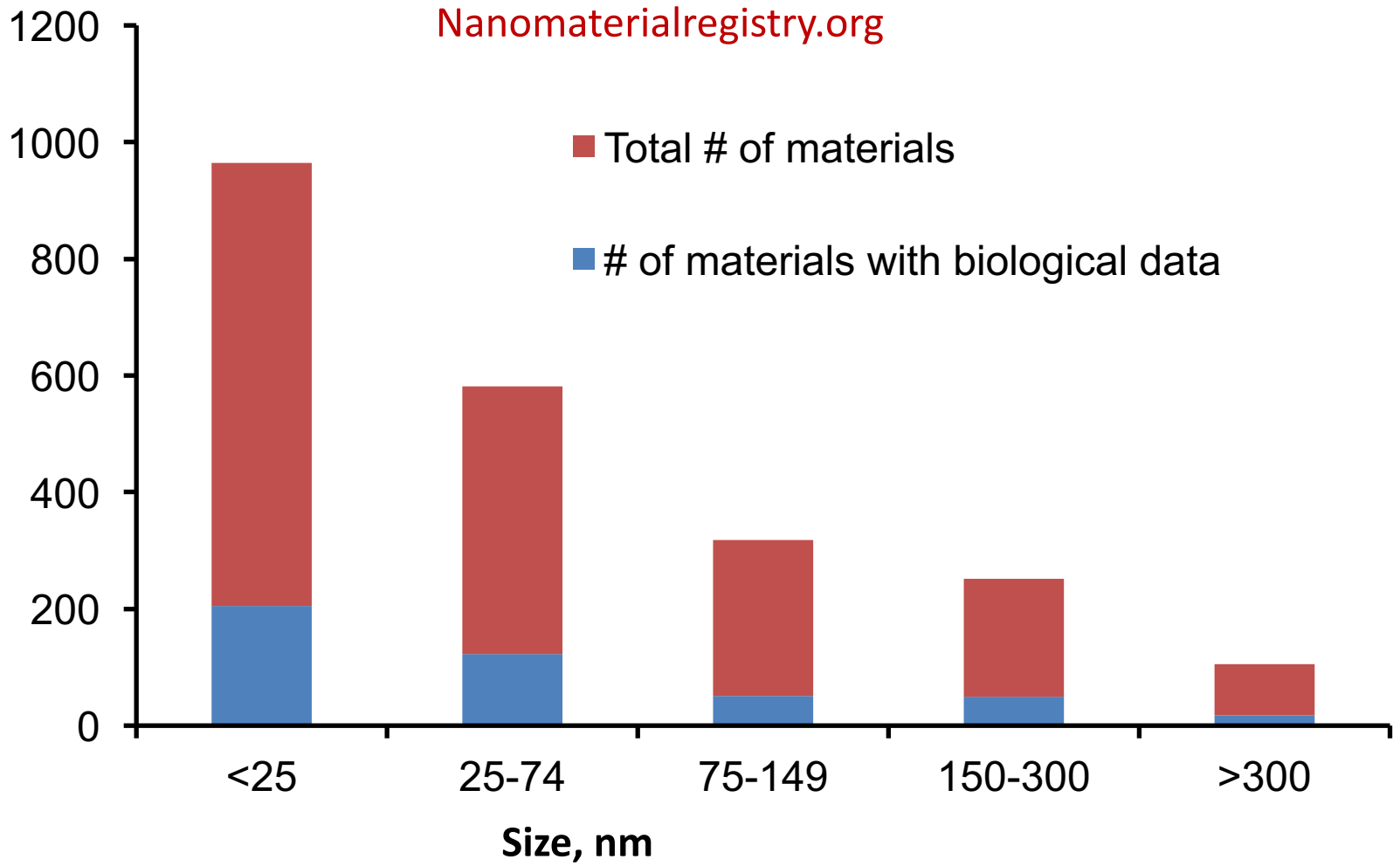


DOIs & URLs



Transition To Alex

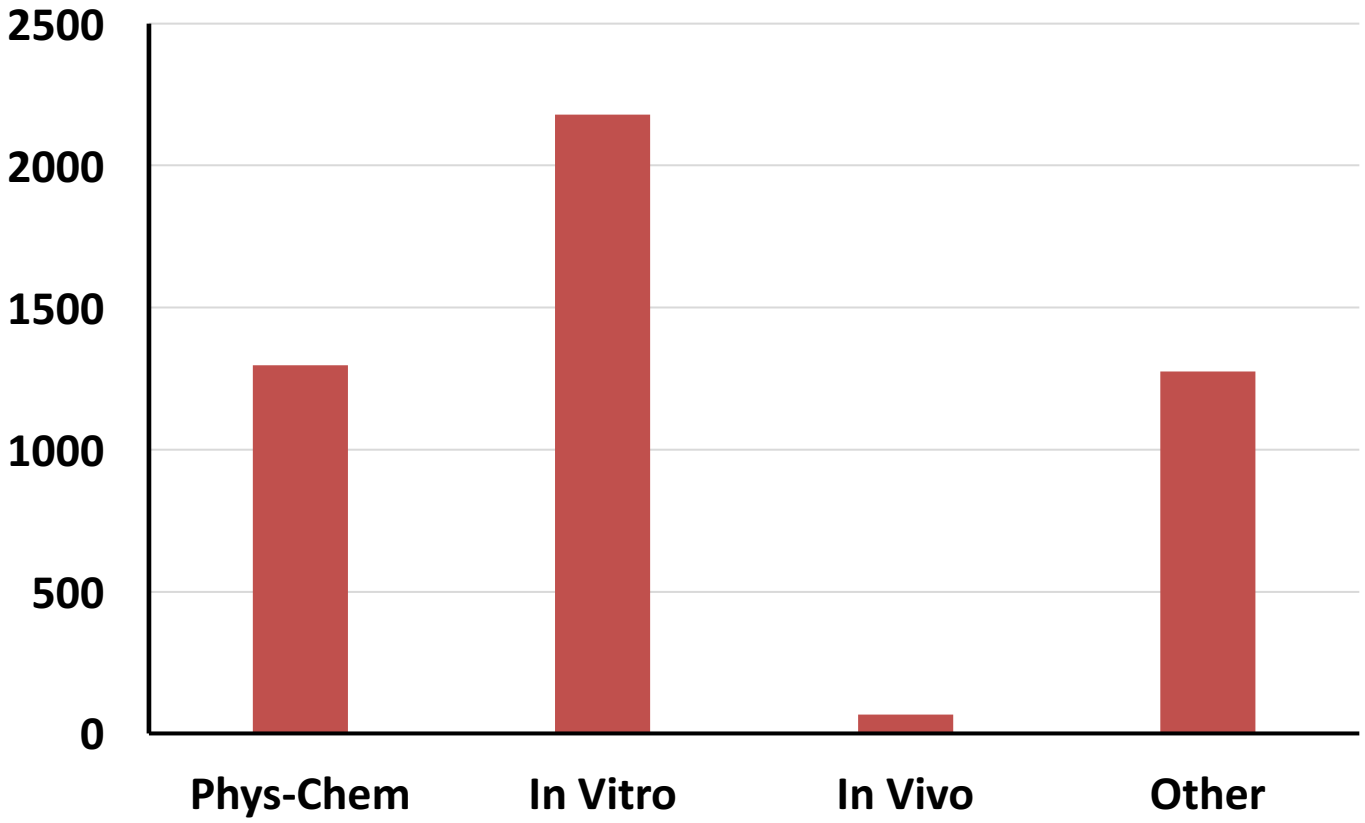
Nanomaterial Registry



445 out of 2000+ nanomaterials associated with biological data, mostly different types of toxicity, but also skin sensitization, mutagenicity, etc.

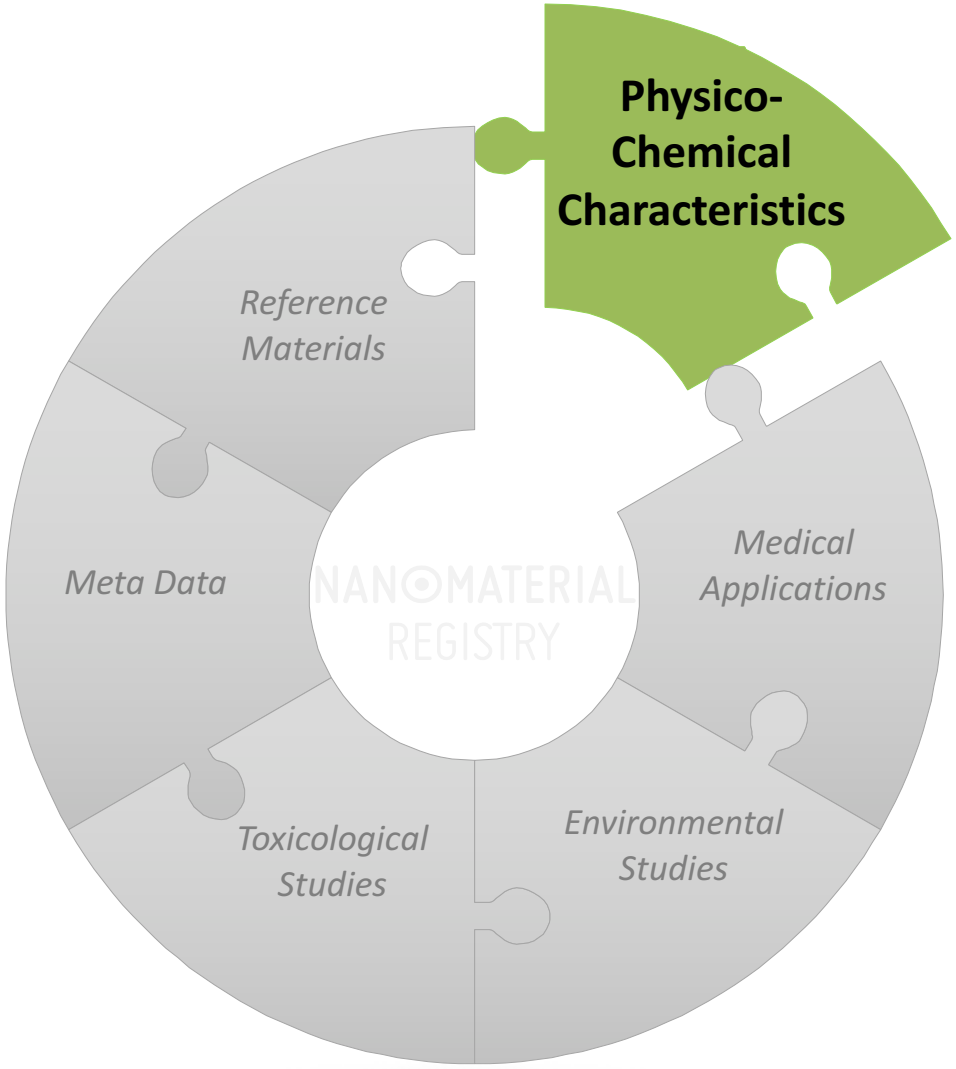
Nanomaterials Data in caNanoLab

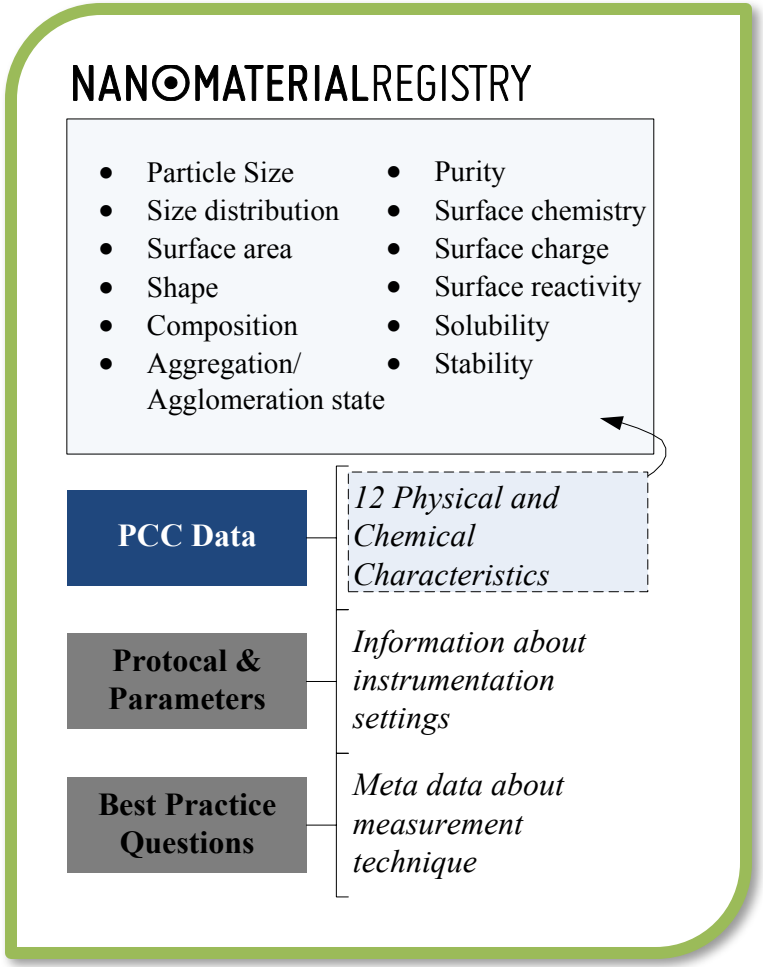
https://cananolab.nci.nih.gov/caNanoLab/#/



1217 Samples associated with 4817 data records in total

Minimal Information about Nanomaterials (MIAN)

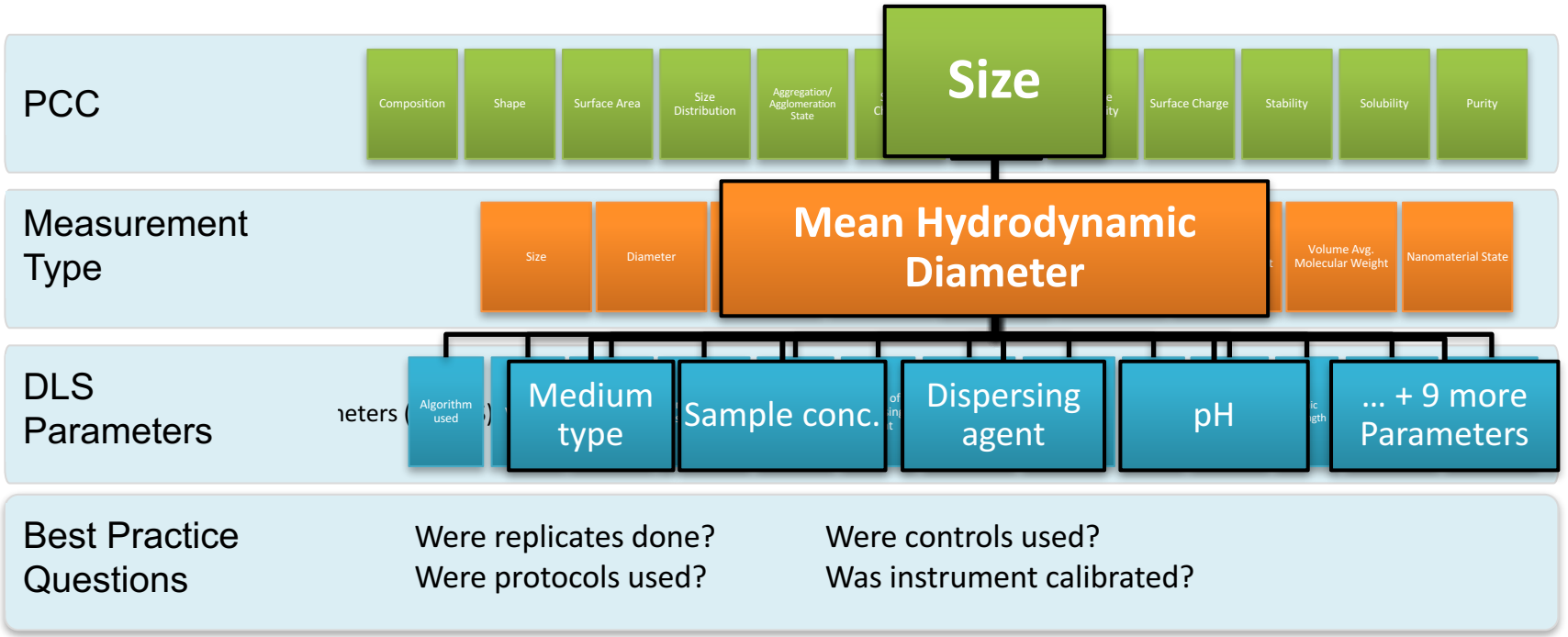




A **controlled vocabulary** of PCC & measurands have been identified (https://www.nanomaterialregistry.com/resources/**Glossary**.aspx)

Minimal Information About Nanomaterials*

NANOMATERIALREGISTRY

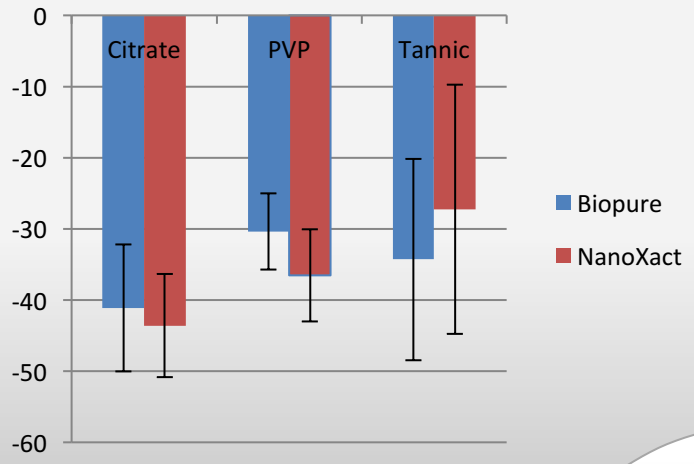


Mills, K. et al. Nanomaterial registry: database that captures the minimal information about nanomaterial physico-chemical characteristics. *J Nanopart Res* (2014) 16: 2219.

LOOKING ACROSS CHARACTERIZATION DATA

QUESTION

How are the Zeta Potentials of silver nanomaterial from specific product lines affected by capping agents?

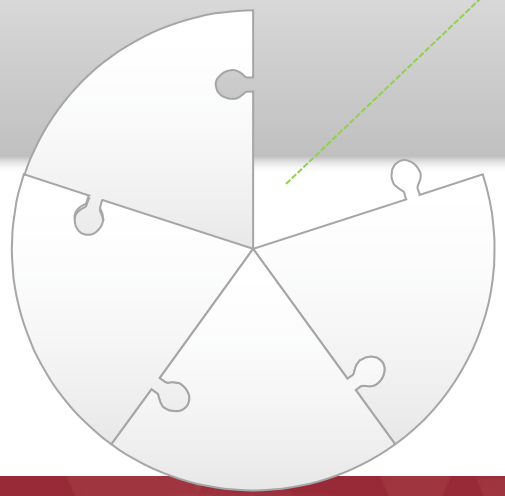


INTEGRATION

- ✓ The ability to look across data to see trends and linkages
- ✓ What questions can be answered?

IMPACT

Appropriate selection of proper in vitro toxicology assays



DATA CURATION PROCESS

1. Data identification
2. Data evaluation
3. Data entry

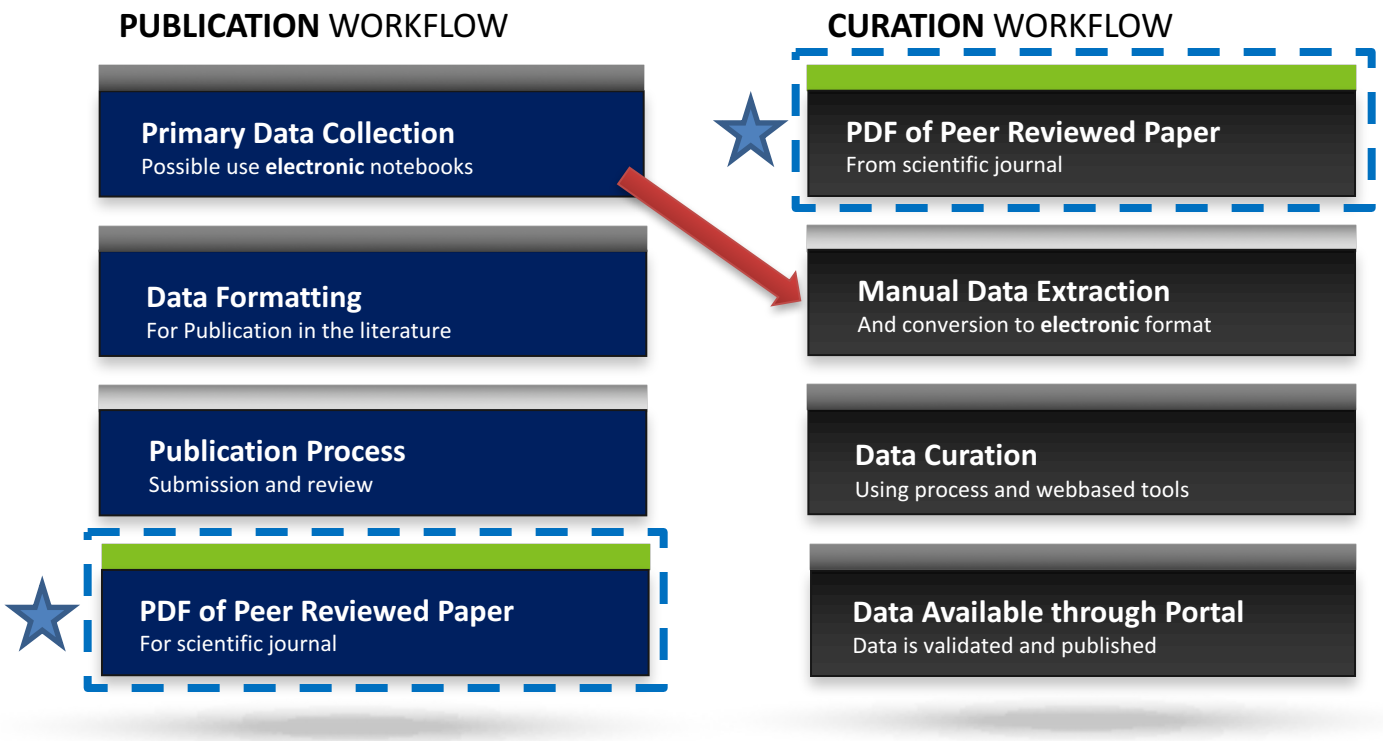
} Curation

4. Quality Assurance (transcription check)
5. Quality Control (scientific interpretation check)

Average of Time (min)	Database	Journal Article	Manufacturer	Other	Reference Material	Average Time (min)
Curation	15	62	8	13	16	23
QA	2	2	1	5	4	3
QC	12	22	2	16	15	13
Grand Total (min)	**29**	**86**	**11**	**33**	**35**	**39**

NANOMATERIALREGISTRY

CHALLENGE: STREAMLINING DATA COLLECTION



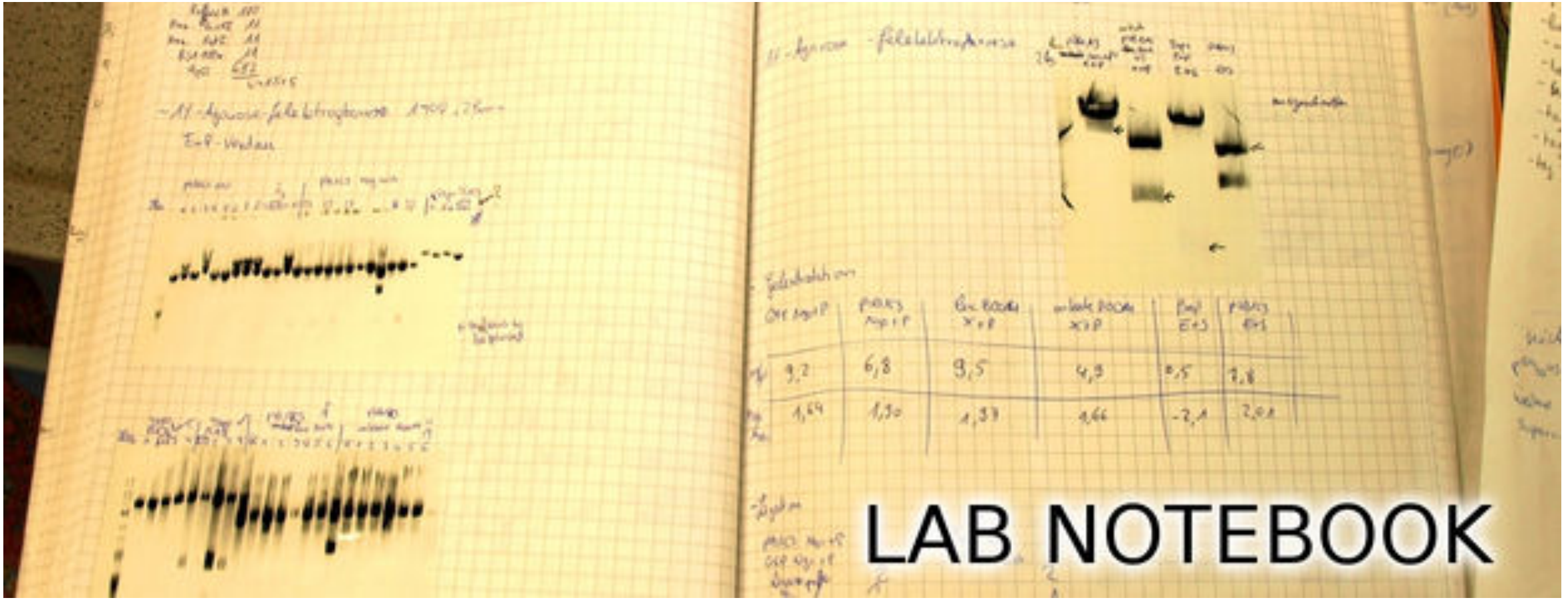
PURPOSE: GROW THE DATA REPOSITRY

Tropsha, Hickey, Mills. Nature Nano, 2017, in press

Transition to Fred

Fundamental Issue for Rigor and Reproducibility

- Capturing experimental data as it is being generated and organizing it in well curated repositories is key, BUT
- User interfaces that add a burden to already overburdened researchers DO NOT add value and are frequently NOT USED
  - Steep learning curve
  - Don’t fit with experimental workflow
  - No single repository captures everything


LAB NOTEBOOK

E-Notebook Interface to Facilitate Data Collection

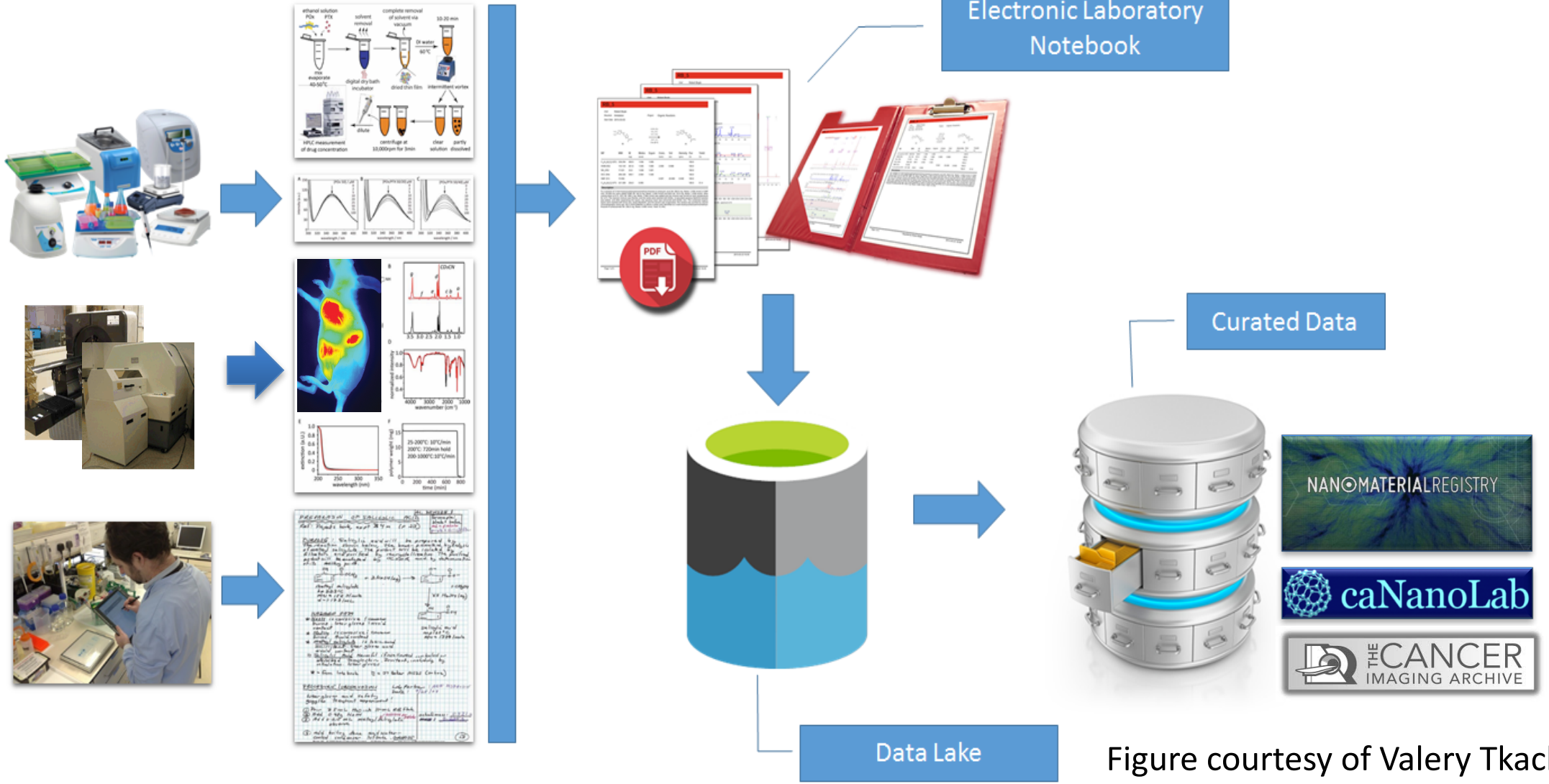
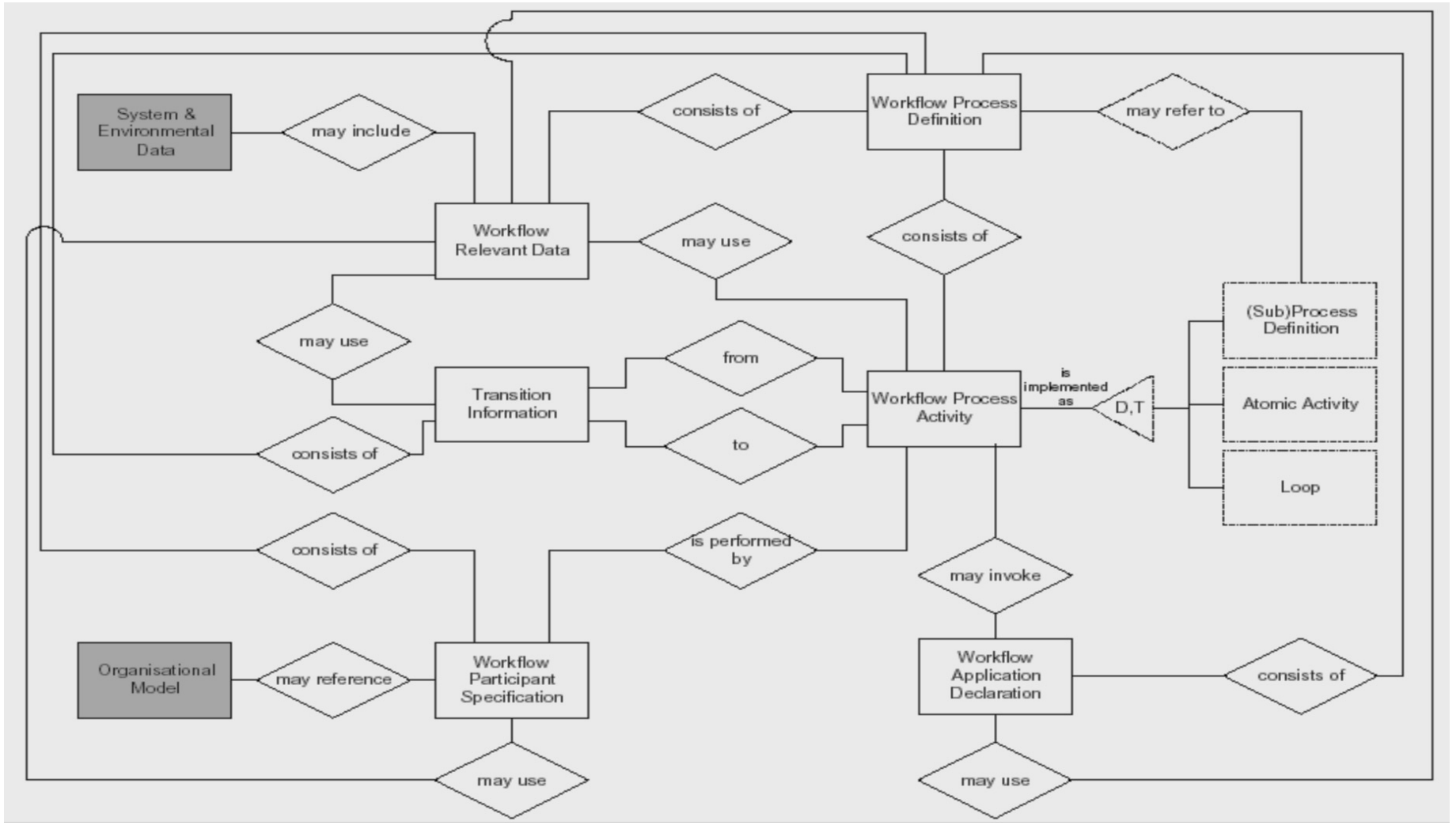


Figure courtesy of Valery Tkachenko

User Interface that Enables Efficient Research Workflow

- Structured data entry that matches experiment designs
- Usable on mobile devices and desktops
- Data retrieval from the same UI
  - Direct data to processing pipelines

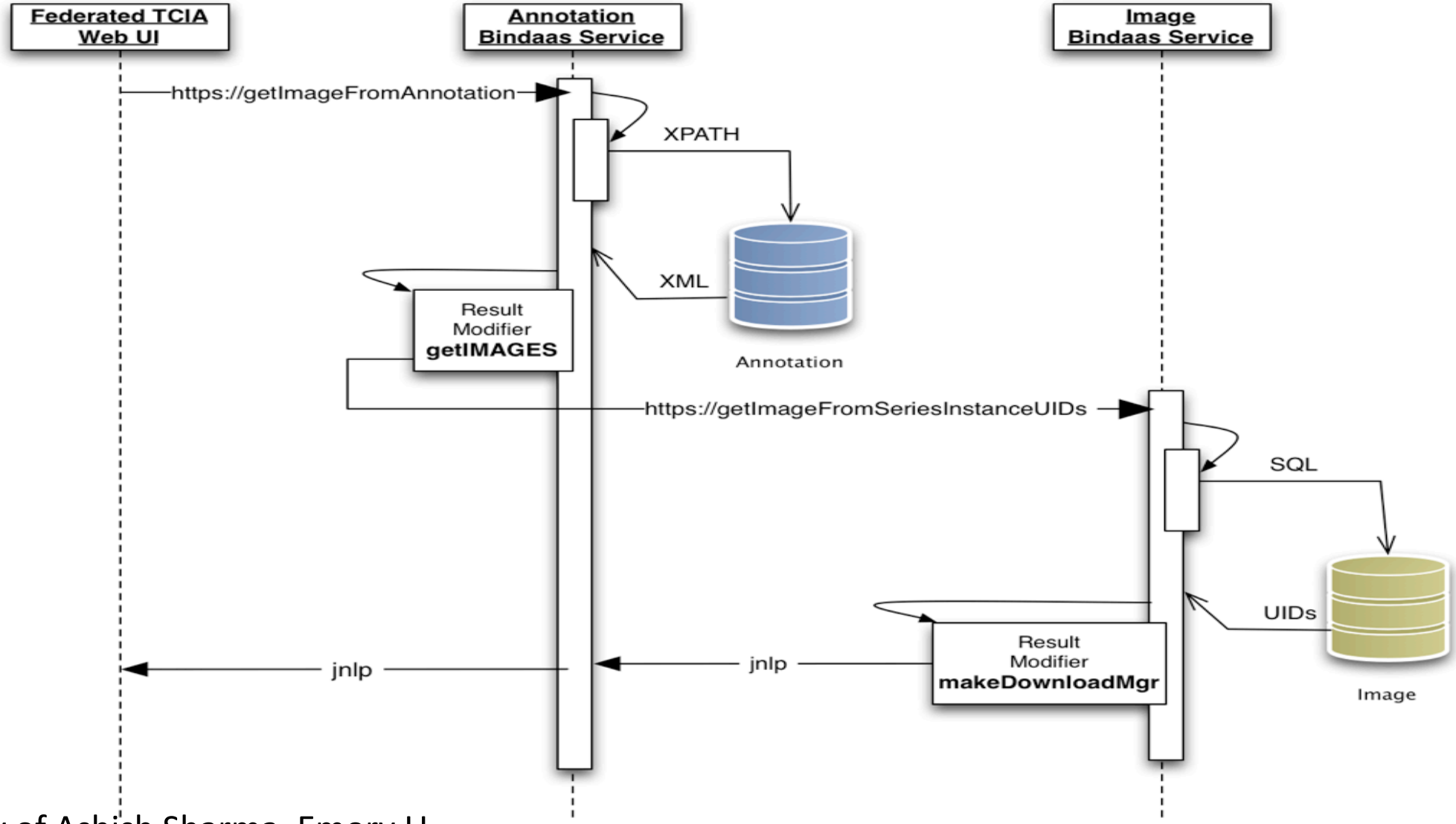
Capturing Experimental Workflows



Workflow Description Language and Workflow Patterns, Wang
https://www.cs.ucf.edu/~dcm/Teaching/ProcessCoordination/Fall02Class/ResearchPresentations/Yi%20Wang.ppt

- TCIA API provides access to images and non-image data
- APIs can be used to link with caNanoLab (and other repositories)
- APIs can be extended to support data exploration and integration of data from multiple repositories

Middleware can add Rest API to Existing Repositories



Slide Courtesy of Ashish Sharma, Emory U.

Transition to Alex

Data Science in Cancer Nanotechnology: challenges to be resolved in the next few years

- Implementation of Data to Knowledge to Wisdom (**D2KW**) Tools
  - Automated data extraction process, including text mining tools
  - Ontology-driven data collection, registration, direct deposition, complex querying, and views
  - Model-building tools
  - Model-driven experimental design
- Growth of use cases
- Access to *actual* materials via collaborations with manufacturers

A successful database should have the "right" answers to:

- What data is deposited?
- How to deposit data?
- How one tracks data usage?
- How to acknowledge data depositors (including points for promotion and tenure)?

- How to create data sharing continuum between researchers, publishers, funders?

Current Efforts to Promote Data Sharing and reproducibility

Funding Opportunity Title

Centers of Cancer Nanotechnology Excellence (CCNE) (U54)

Cooperative Agreement Terms and Conditions of Award:

Nanomaterial characterizations, protocols, and associated publications are **expected to be submitted to the caNanoLab data portal directly by awardees**. All CCNE investigators are strongly urged to work together to ensure that **all relevant data are deposited to caNanoLab (no later than upon publication of findings in scientific journals**.

NIH Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

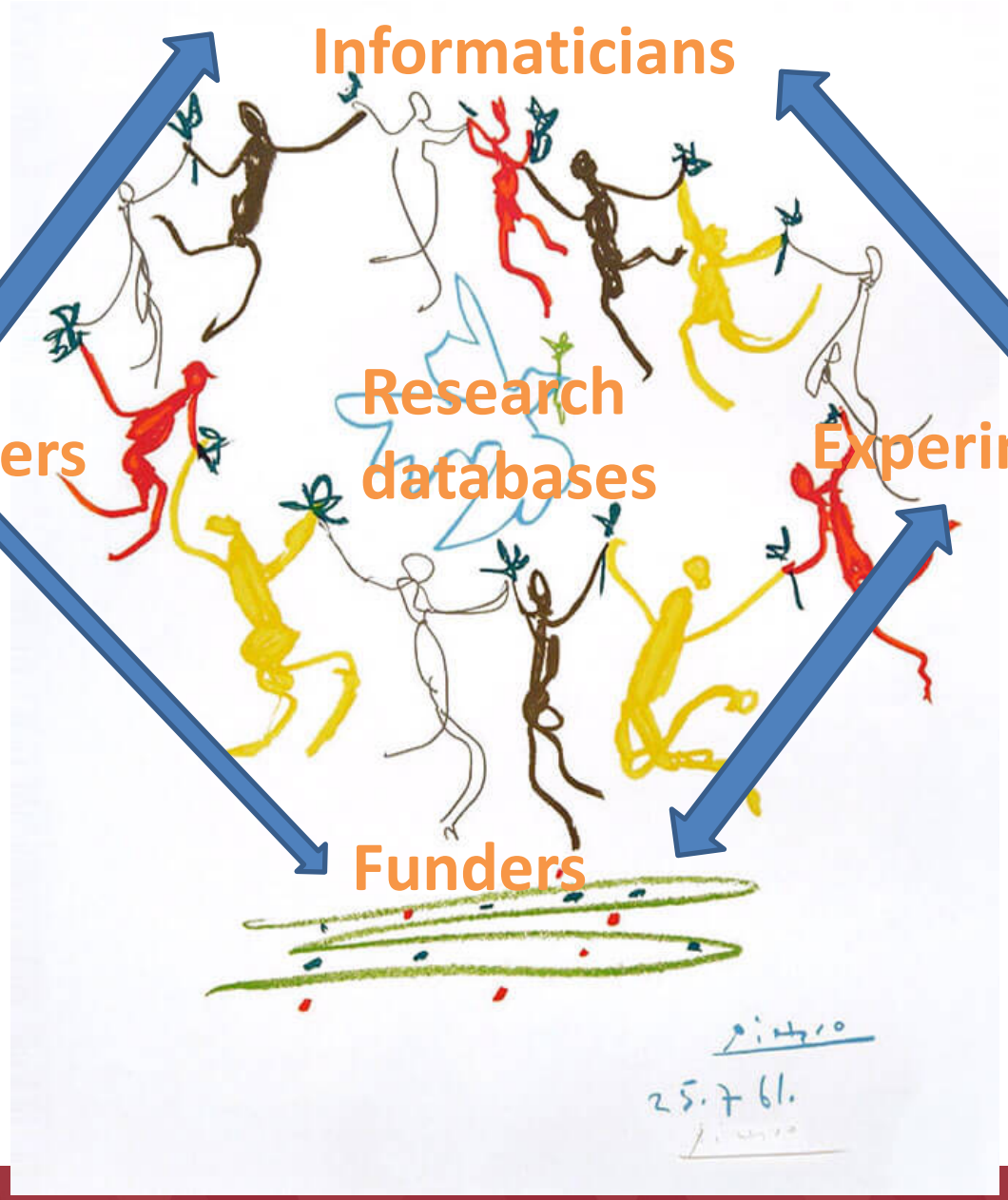
Home

NIH Data Sharing Repositories

This table lists NIH-supported data repositories that make data accessible for reuse. Most accept submissions of appropriate data from NIH-funded investigators (and

IC	Repository Name	Repository Description	Data Submission Policy	Access to Data
NCI	Cancer Nanotechnology Laboratory (caNanoLab)	caNanoLab is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical, in vitro, and in vivo assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.	How to submit your data to caNanoLab	How to access caNanoLab data
NCI	The Cancer Imaging Archive (TCIA)	The image data in The Cancer Imaging Archive (TCIA) is organized into purpose-built collections of subjects. The subjects typically have a cancer type and/or anatomical site (lung, brain, etc.) in common.	How to submit data to TCIA	How to access TCIA data
NIBIB	Nanomaterial Registry	By leveraging and developing a set of Minimal Information About Nanomaterials (MIAN), ontology and standards through a community effort, it has developed a data model for data collection and sharing in the nanotechnology field. It facilitates data validation and data quality improvement. It is a data-driven tool aimed at enabling researchers to close knowledge gap.	How to submit data to Nanomaterial Registry	How to access Nanomaterial Registry data

The data sharing interdependency circle



Questions

- Why share?
- What help should be provided to facilitate sharing?
- How can data science accelerate discovery?
- What practical actions can we take?
  - Working group?
  - Engagement of all major journals?
  - Funding for data stewards?
  - ???

User Acceptance Issues: Why are you resistant to CMMN-wide adoption of ENB*?

1. CMMN isn't my only source of funding.
2. I don't trust the internet in another state to protect my IP.
3. An external curator won't understand my data or contact me before making QA/QC edits.
4. I will lose data if this application fails.
5. I have job/idea security as a single-point expert.
6. The old ways keep chain of custody where I can reach all the records.
7. I dislike copying records into several formats.
8. Imperfect fit of the tool affects data harmony (import), workflow, & reporting (export)
9. Others?

*from a survey conducted by Prof. F. Prior's group.

User Acceptance Issues: Why would you like ENB?

1. ENB gives me raw lab updates from my students, so I can help when they get stuck, and identify projects to assign.
2. ENB produces my annual report for NCI almost automagically.
3. ENB keeps unfinished experiment state so I don't completely restart when someone leaves.
4. I don't have to re-enter all the details if the experiment only changed one parameter.
5. Training can be simplified
6. Complex data sharing is more complete with less effort at each step
7. Query of ENB is easier than flipping notebook pages to find information
8. Others?

*from a survey conducted by Prof. F. Prior's group.

DISCUSSION: How to make data sharing a reality in Cancer Nanotechnology?

NIH NATIONAL CANCER INSTITUTE

Home > Grants & Training > Research Grants

RESEARCH GRANTS

Funding Opportunities

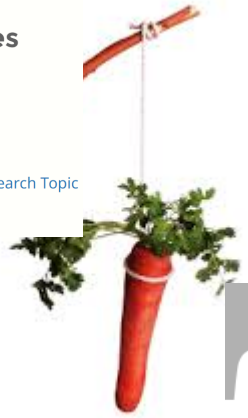
Outstanding Investigator Award +

NCI Research Specialist Award (R50)

Funding Opportunities

ON THIS PAGE

- Funding Opportunities by Type
- NCI Special Initiatives
- NCI Funding Opportunities by Research Topic



NIH National Institutes of Health
Office of Extramural Research

Research Performance Progress Report (RPPR)

Wed 9/13/2017 10:36 AM

Morris, Stephanie (NIH/NCI) [E] <morriss2@mail.nih.gov>

NCI Alliance Data Coordinators—Needed Participation in Working Group on Nanomedicine Data Reporting Guidelines

To: Alice Fan; Anh Ung; Dana Levine; Daniel Binzel; Hwang, Duhyeong; Eric Berns; Fotini Kouri; Fred Prior; Gokay Yamankurt ; Hongyu Zhou; Hsian-Rong Tseng; Huan Meng ; Hui Zhang; Jun Yue ; Kaiyuan Ni; Kate Hleb; Ketan Ghaghada; Malcolm Tobias; Sokolsky, Marina; Matthew Smalley; Heiskanen, Mervi (NIH/NCI) [E]; Michal Lijowski; Lijowski, Michal (NIH/NCI) [C]; Miles Miller; Zhang, Mingzhen; Morris, Stephanie (NIH/NCI) [E]; Barnes, Philippa (NIH/NCI) [C]; Pinar Kanlikilicer;

Cc: Harper, Stacey <Stacey.Harper@OREGONSTATE.EDU> (Stacey.Harper@OREGONSTATE.EDU); 'christine.hendren@duke.edu' (christine.hendren@duke.edu) (christine.hendren@duke.edu); Tropsha, Alexander; Fred Prior; Heiskanen, Mervi (NIH/NCI) [E]; Liu, Christina (NIH/NCI) [E]; Morris, Stephanie (NIH/NCI) [E]

Suggested Meetings + Get more app

Dear Data Coordinators,

I am contacting all of you about an important opportunity to provide information and also present your research as part of a working group focused on developing nanomedicine data reporting

nature nanotechnology

Home | Current issue | Comment | Research | Archive ▾ | Autho

Summary

- Rigor and Transparency are essential components of research
- NIH and major journals are enforcing increased scientific rigor and research reproducibility
- Well curated Information repositories are essential enablers of reproducibility
  - No Single Repository can manage the data even for a single discipline
- Current user interfaces are complex and do not map well into research workflow
- We believe Electronic Lab notebooks that capture all data within a particular domain and transparently distribute it to multiple repositories are essential

Chief talking/discussion points

- Everyone is producing data but most of this data is not accessible
- Data science is all over us …. but we are not all over data science … yet
- There are many challenges in making data work for us: we need to work on solutions together
  - Standards for data collection and dissemination
  - Establishing data sharing culture
  - Community-driven research databases based on FAIR principles
- Actionable Ideas?