Research Data Management @Harvard

Mercè Crosas, Ph.D.,
Chief Data Science and Technology Officer, IQSS
Harvard University

Why is Harvard concerned about research data management?

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Print edition | Leaders ›
May 6th 2017

DATA SCIENCE, BIG DATA

"Every two years, the amount of digitized data is equal to all of the data ever collected before. The world's knowledge is at our fingertips, and **data science** allows us to effectively and efficiently make use of that knowledge. This is facilitating a societal shift as big as the Industrial Revolution. "

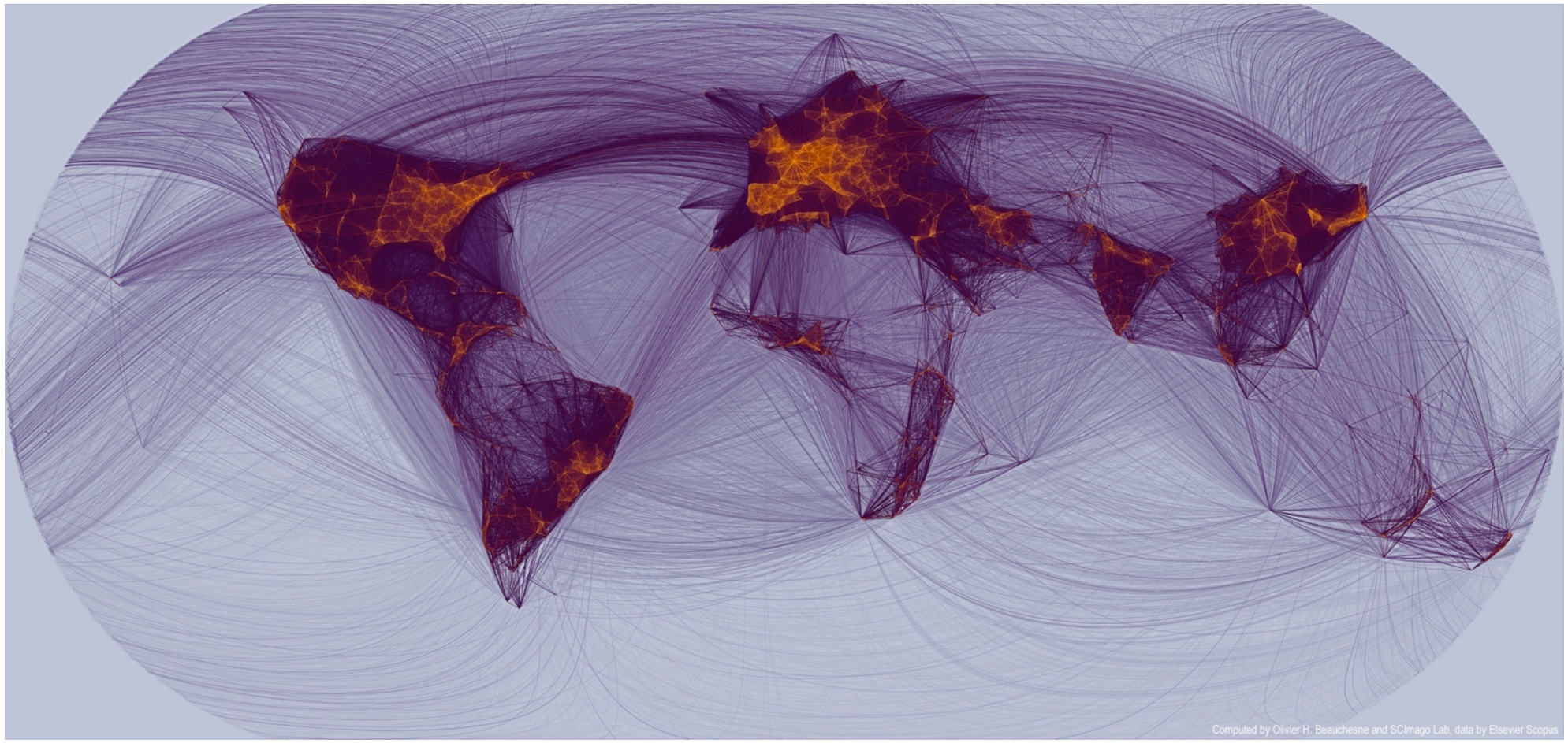
UVAToday Q&A, August 21, 2017

Phil Bourne

Data Science Director, UVA

Former Associate Director for Data Science, NIH

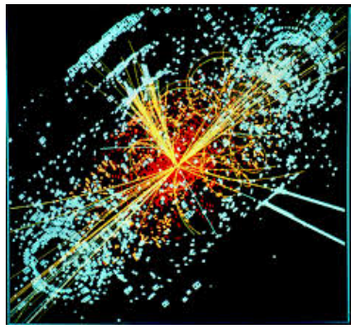
Research is Increasingly Collaborative



"Bright lines in this map of scientific collaborations between 2005 and 2009 show many joint publications."

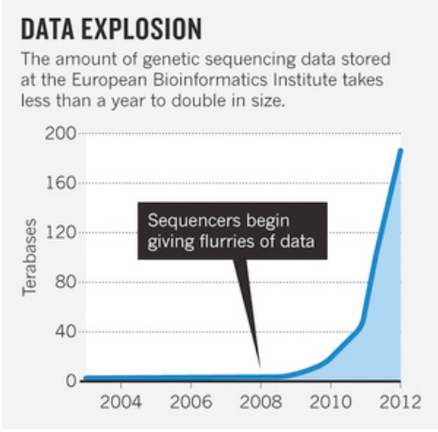
Collaborations: The Rise of Research Networks, The Nature **490**, 335–336 (18 October 2012) doi:10.1038/490335a

... Increasingly Big

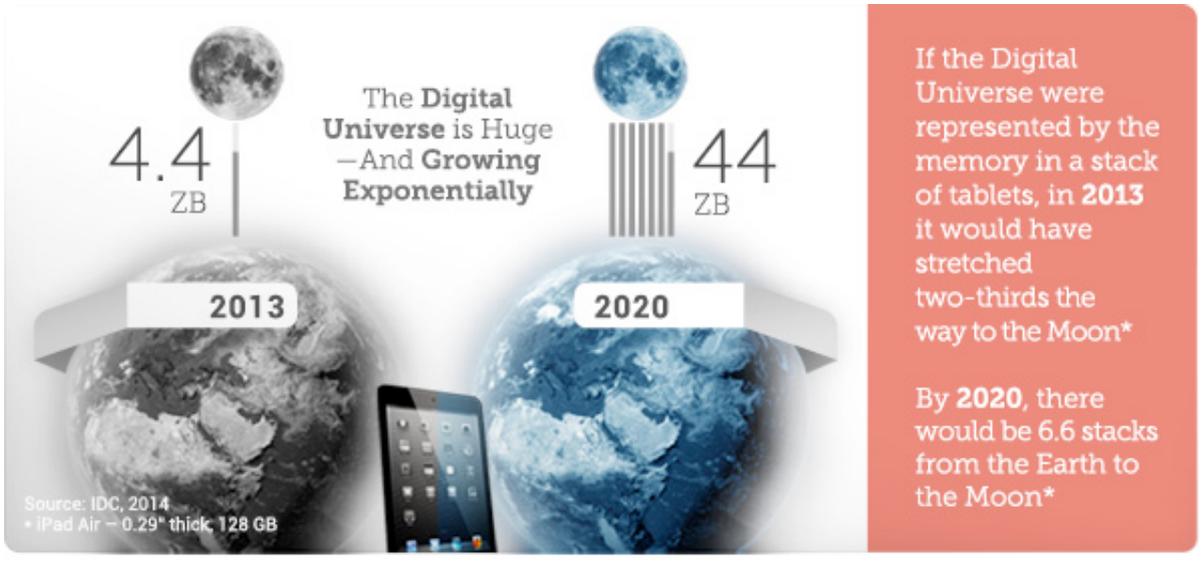


CERN: High-Energy Physics Data >100PB

EBI: Sequence Data > 20PB



Entire Data created by all of us >ZB = 1millionPB



... and some times difficult to Reproduce

ONLY 6 (11%) OUT OF 53 LANDMARK STUDIES THAT CLAIM TO TREAT CANCER COULD BE REPRODUCED

Begley & Ellis, Nature, 2012
(from Scientists at Amgen biotechnology)

Research Data Management facilitates Reuse and Reproducibility

Reproducibility and Replication (*National Science Foundation*)

- The ability for a researcher to replicate the results of a prior study using the same materials and procedures used by the original investigator (reproducibility)
- same procedures are followed but new data are collected (replication)

Empirical, Computational, Statistical Reproducibility (*Stodden, 2014*)

- Empirical: data and collection details are made freely available
- Computational: code, software, hardware, and implementations details are provided
- Statistical: details on choice of statistics tests, model parameters

"Research data management concerns the organization of data, from its entry to the research cycle through the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre

How does
Harvard help you?

Research Data Management Efforts at Harvard

- HMS Data Management Working Group (started 2014)
- Research Computing Council – Data group (started at 2016)
- Harvard Data Group (started at 2016)
  - Co-chaired by Ara Tahmassian (OVPR), Mercè Crosas (IQSS)
  - Connects and includes members of HMS and RCC groups
  - New Data User Agreement Group

Releasing in 2018

- Harvard-wide research data management website: http://datamanagement.harvard.edu (Q1)
- Single contact: datamanagement@harvard.edu (Q1)
- RDM training modules (Q1 basic, Q3/Q4 advanced)
- DUA decision tree

HARVARD UNIVERSITY HARVARD.EDU

Research Data Management @Harvard

Home | Data Lifecycle | Vision | Contact

A reference guide with information and resources to help you manage your data and make it **FAIR**: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable.

Quick Start to your Data Lifecycle ➤

Data Acquisition and Planning

What do I need to know before bringing research data into Harvard? How do I prepare for a data management plan?

- Data Use Agreement, Data Management Plan, Harvard Policies, licensed data.

Data Storage

Where and how should I store my research data? What are the options at Harvard? What do I need to know about security?

- Data files, documentation, logbooks, notebooks, security levels, and permits.

Analysis and Computation

What are the options for research computing at Harvard? Which tools or methods should I use for my research?

- Harvard Research Computing, data science and computational help.

Data Sharing and Archiving

What is Data Sharing and why is it important? What do Funders and Journals require? Can I get help on data curation?

- Harvard Dataverse repository, domain repositories, Open Data policies.

Preservation Services

What is long-term preservation? What services does Harvard offer for preservation of data collections?

- Harvard Library services, format migration, suitable medium.

Data Disposal

Are there some cases where I need to destroy my data? How should I do it? What services does Harvard offer?

- Contractual obligations, method of disposal, documentation.

Coming soon: http://datamanagement.harvard.edu

Data Acquisition and Planning

- Data Management Plan required by funding agencies - use DMPTool
- Data User Agreements - contact new DUA group
- Human subject data - contact IRB, Harvard IT Security (HIPAA, FERPA, and other regulations may apply)
- Animal data - contact IACUC
- Harvard Retention Policy: 7 years
- Planning for reuse:
  - Common formats and file/data structure, variable/attributes metadata, documentation, annotations
  - Share in data repository
  - Version control for code (GitHub)

Data Storage

Analysis and Computation

	FAS/SEAS Odyssey	HMS O2	IQSS RCE	HBS RC
Storage	37 PB	27 PB	100 TB	150 TB
Cores	75,000 cores	7000 shared; 1300 dedicated cores	1000 cores	400 cores
Security	Level 3/4	Level 3	Level 3	Level 3/4
For	FAS/SEAS & beyond	HMS and HSPH	Social Science & beyond	HBS

Releasing in 2018: New SID project

- Single access to all RC resources, including access to commercial Cloud resources (AWS, Azure, Google Cloud)

Data Storage

Analysis and Computation

DataTags	Harvard Security Levels
Blue	**Level 1:** No sensitive data; open data
Green	**Level 1:** Low-risk, de-identified data
Yellow	**Level 2:** Confidential information by University standards; no material harm
Orange	**Level 3:** Confidential information that could cause material harm (non-level 4 FERPA)
Red	**Level 4:** High-risk confidential information (SSN)
Crimson	**Level 5*** (Level 4.5, on the network) Information that would cause severe harm

Training and Consulting on Data Science - R, Python, Stata, GIS, data cleaning, computing, visualizations, and more:

- IQSS Data Science Services
- Harvard Chan Bioinformatics Core
- HBS Research Computing Services
- Center for Geographic Analysis
- CfA - Wolbach Library
- Digital Scholarship Support Group

New Harvard Data Science Initiative (DSI)

Sharing data, documentation, and code in a public repository:

- Required by funding entities
- Required (or strongly recommended) by many journals
- Harvard Dataverse repository: http://dataverse.harvard.edu
  - Data citation, with global persistent identifier
  - Extensive metadata support
  - Terms of use and restrictions
  - Versioning
- Domain-specific repositories

FAIR data: Findable, Accessible, Interoperable, and Reusable

Preservation Services | Data Disposal

- Harvard Library provides services for special, thorough preservation of research collections
- In same cases, research data must be destroyed after use (based on DUA)
- Otherwise, should be retained 7 years in RC facility
- And shared in public repository (such as Harvard Dataverse) forever

Thanks

Coming soon: http://datamanagement.harvard.edu

Thanks to Ara Tahmassian, Alan Wolf, Caroline Shamu, Ceilyn Boyd, Radhika Khetani, Jessica Pierce, Julie Goldman, Jennifer Pionting, Helen Page, Robert Freeman, Julian Gautier, Tania Schlatter and many more.