Defining the gap in neuroscience expertise around data handling and analysis

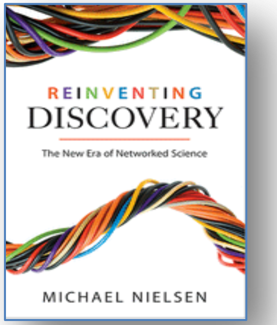
Maryann E. Martone, Ph. D.
Principal Investigator, Neuroscience Information Framework
President: FORCE11
Future of Research Communications and e-Scholarship

The duality of modern scholarship

Observation: Those who build information systems from the machine side don’t understand the requirements of the human very well

Those who build information systems from the human side, don’t understand requirements of machines very well

And then there’s the internet...



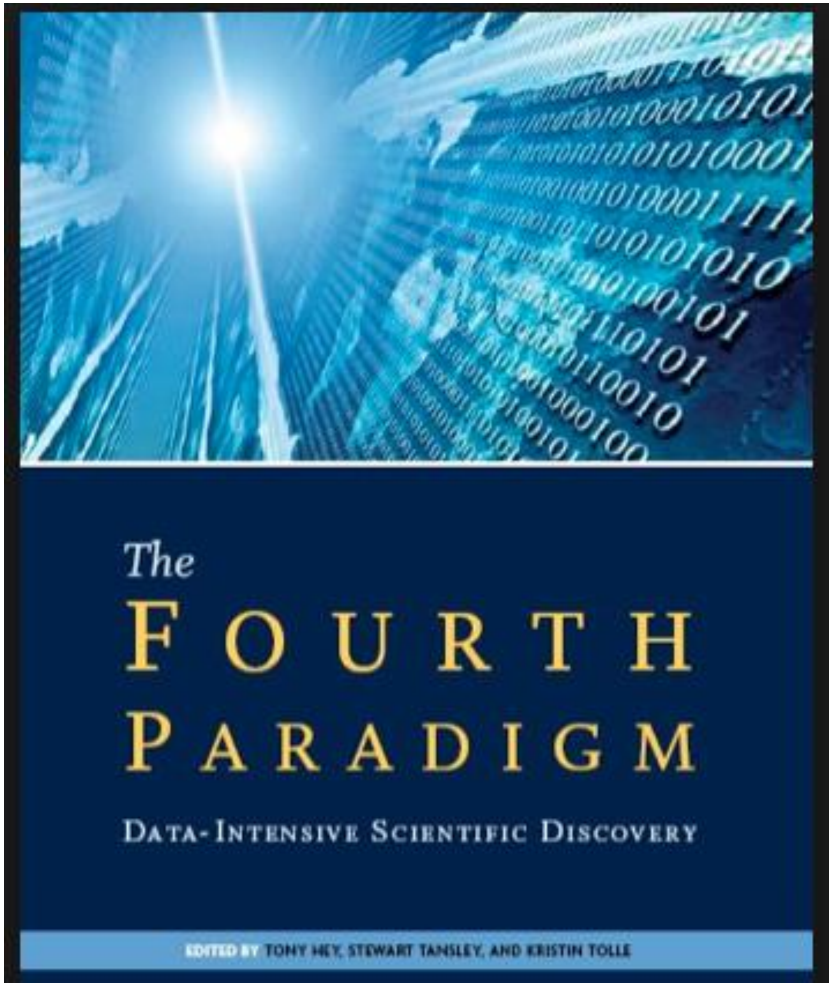
REINVENTING

DISCOVERY

The New Era of Networked Science

MICHAEL NIELSEN

Solving the large problems of science?


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

...starts in the laboratory

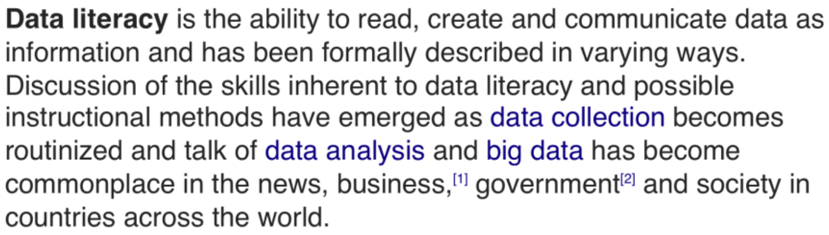
- Observation
- Experimentation
- Modeling
- Cooperative data intensive science

"An unaided human's ability to process large data sets is comparable to a dog's ability to do arithmetic, and not much more valuable." –Michael Nielson, *Reinventing Discovery*, 2011.

Basic data literacy

- Concepts
  - Data type
  - Data structure
  - Databases
  - Metadata
  - Query languages
  - Data formats

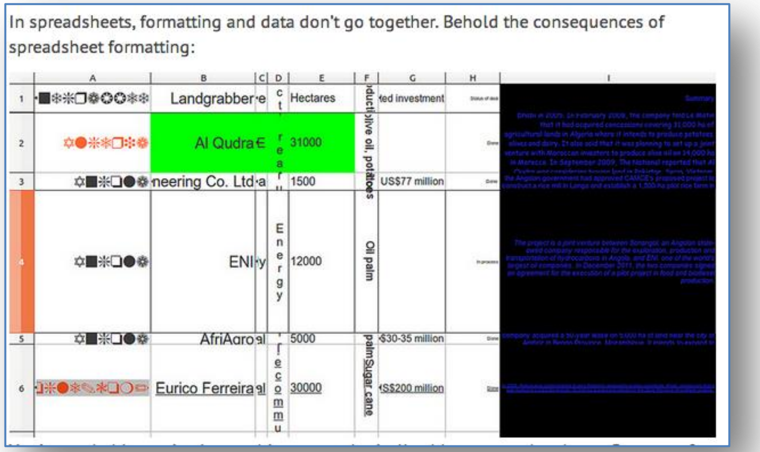
→ Producing *actionable* data



Data literacy is the ability to read, create and communicate data as information and has been formally described in varying ways. Discussion of the skills inherent to data literacy and possible instructional methods have emerged as data collection becomes routinized and talk of data analysis and big data has become commonplace in the news, business,[1] government[2] and society in countries across the world.

This is not text

Wikipedia



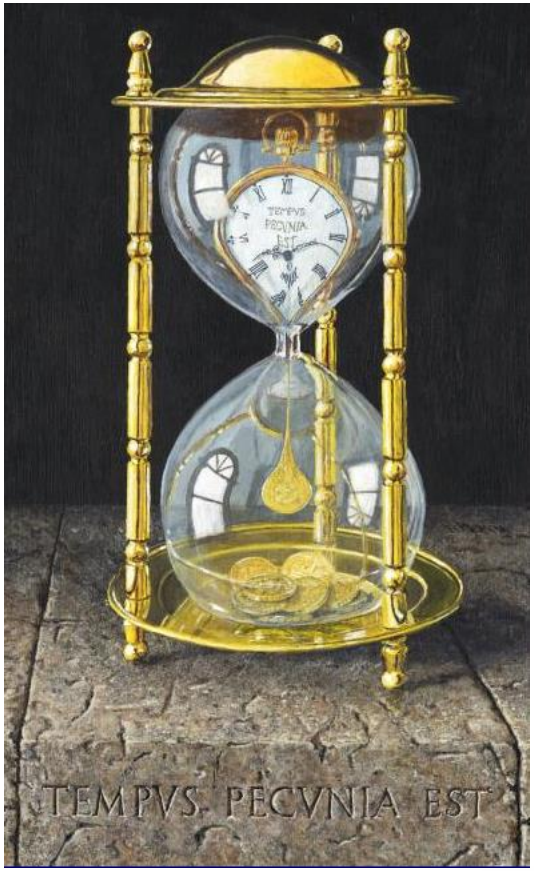
In spreadsheets, formatting and data don't go together. Behold the consequences of spreadsheet formatting:

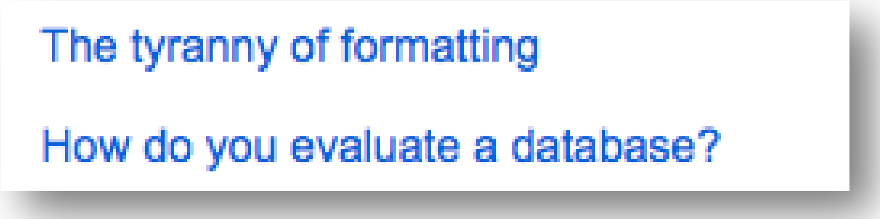
	A	B	C	D	E	F	G	H	I
1	■❄✳❒❁✪✪❄❄	Landgrabber·e		ct	Hectares	ductive oil, potatoes	ted investment	Status of deal	Summary
2	✡●✳✳❒❄❁	Al Qudra E		rea	31000			Done	[illegible]
3	✡■✳❒●❁	neering Co. Ltd·a		r u	1500		US$77 million	Done	[illegible]
4	✡■✳❒●❁	ENI·y		Energy	12000	Oil palm		In process	[illegible]
5	✡■✳❒●❁	AfriAgro·al			5000	palmSugar cane	$30-35 million	Done	[illegible]
6	❒✳●✳✎✳❒○▭	Eurico Ferreira·al		recommu	30000		S$200 million	Done	[illegible]

From: School of Data

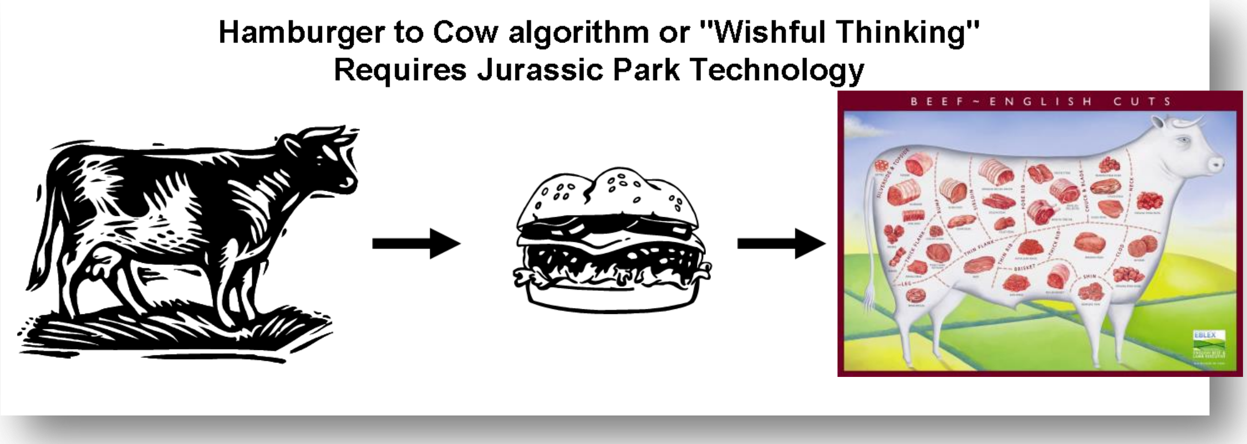
“This sounds like a very interesting discussion, Dr. Martone, but what is XML?”

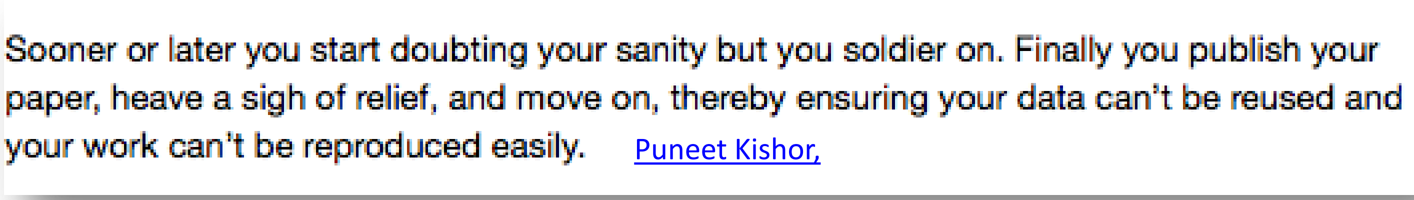
Current methods are inefficient and result in a non-computational product




The tyranny of formatting

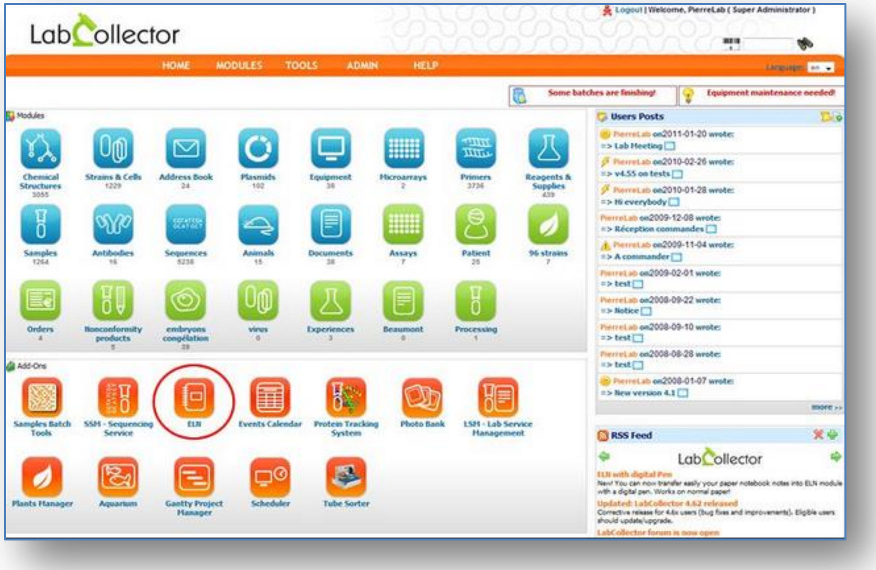
How do you evaluate a database?


Hamburger to Cow algorithm or "Wishful Thinking"
Requires Jurassic Park Technology


Sooner or later you start doubting your sanity but you soldier on. Finally you publish your paper, heave a sigh of relief, and move on, thereby ensuring your data can't be reused and your work can't be reproduced easily. Puneet Kishor,

Data management

- Electronic laboratory notebooks
- Laboratory Information Management Systems
- Institutional repositories
- Spreadsheets
  - Enhanced spreadsheets
- On-line tools
- Data mark up and annotation



Some batches are finishing!

Equipment maintenance needed!

PierreLab on2011-01-20 wrote:
=> Lab Meeting

PierreLab on2010-02-26 wrote:
=> v4.55 on tests

PierreLab on2010-01-28 wrote:
=> Hi everybody

PierreLab on2009-12-08 wrote:
=> Réception commandes

PierreLab on2009-11-04 wrote:
=> A commander

PierreLab on2009-02-01 wrote:
=> test

PierreLab on2008-09-22 wrote:
=> Notice

PierreLab on2008-09-10 wrote:
=> test

PierreLab on2008-08-28 wrote:
=> test

PierreLab on2008-01-07 wrote:
=> New version 4.1

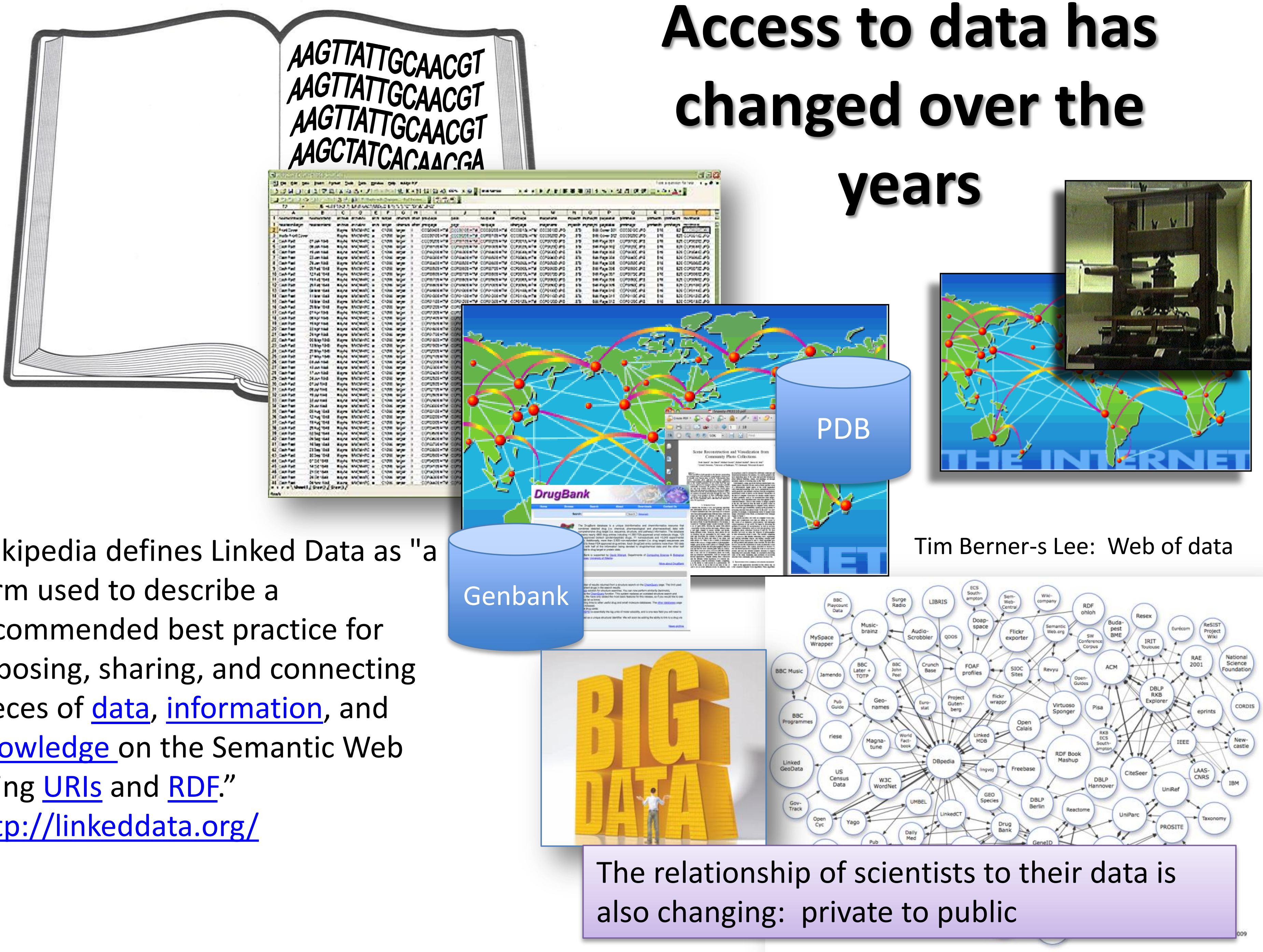
ELN with digital Pen
New! You can now transfer easily your paper notebook notes into ELN module with a digital pen. Works on normal paper!

Updated: LabCollector 4.62 released
Corrective release for 4.6x users (bug fixes and improvements). Eligible users should update/upgrade.

LabCollector forum is now open

Funding agencies are requiring data management plans

Access to data has changed over the years



Tim Berner-s Lee: Web of data

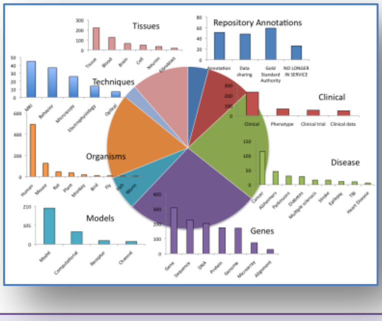
Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."
http://linkeddata.org/

The relationship of scientists to their data is also changing: private to public

Data sharing

- Why share data
- What data should be shared
- Use of data standards
- Role of data repositories
  - Institutional repositories
- Data papers/data journals
- Data curation

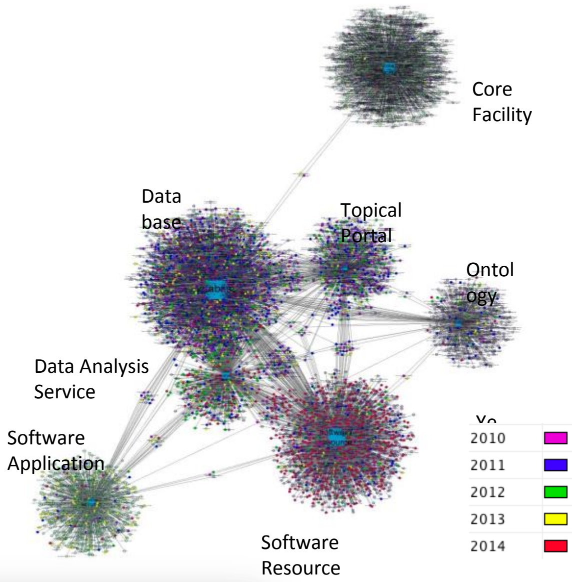
...but it's also necessary



Make your data FAIR: Findable, Accessible, Interoperable, Reusable
https://www.force11.org/group/fairgroup

Finding and using public data

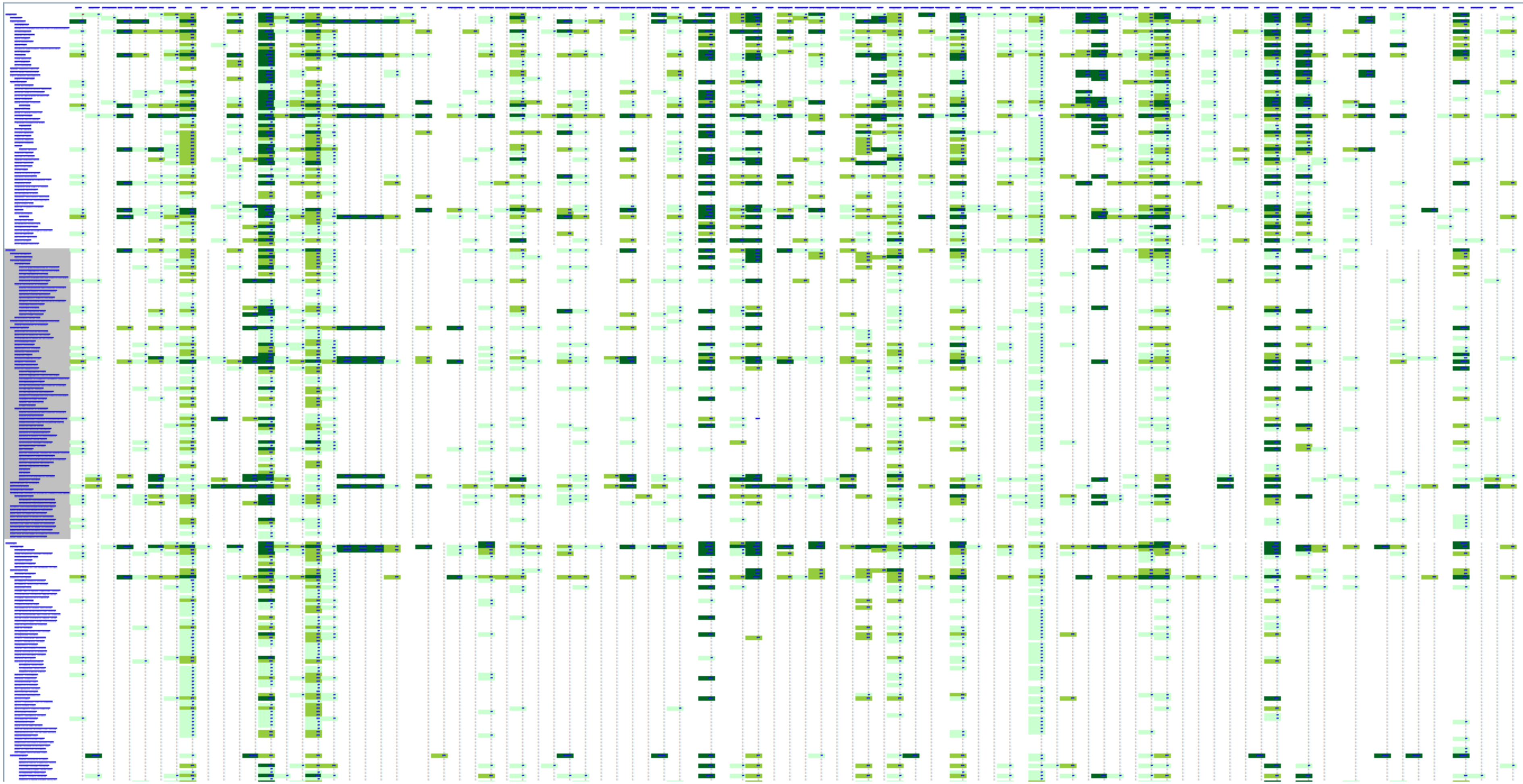
- How to find data
  - NIF
  - NIH Data Commons/Data Discovery Index
  - On line portals, e.g., DataHub
- How to evaluate an on-line resource
  - Coverage
  - Provenance
  - Curation
- Data licenses



ALLEN INSTITUTE *for* BRAIN SCIENCE
Fueling Discovery

Data and knowledge gaps

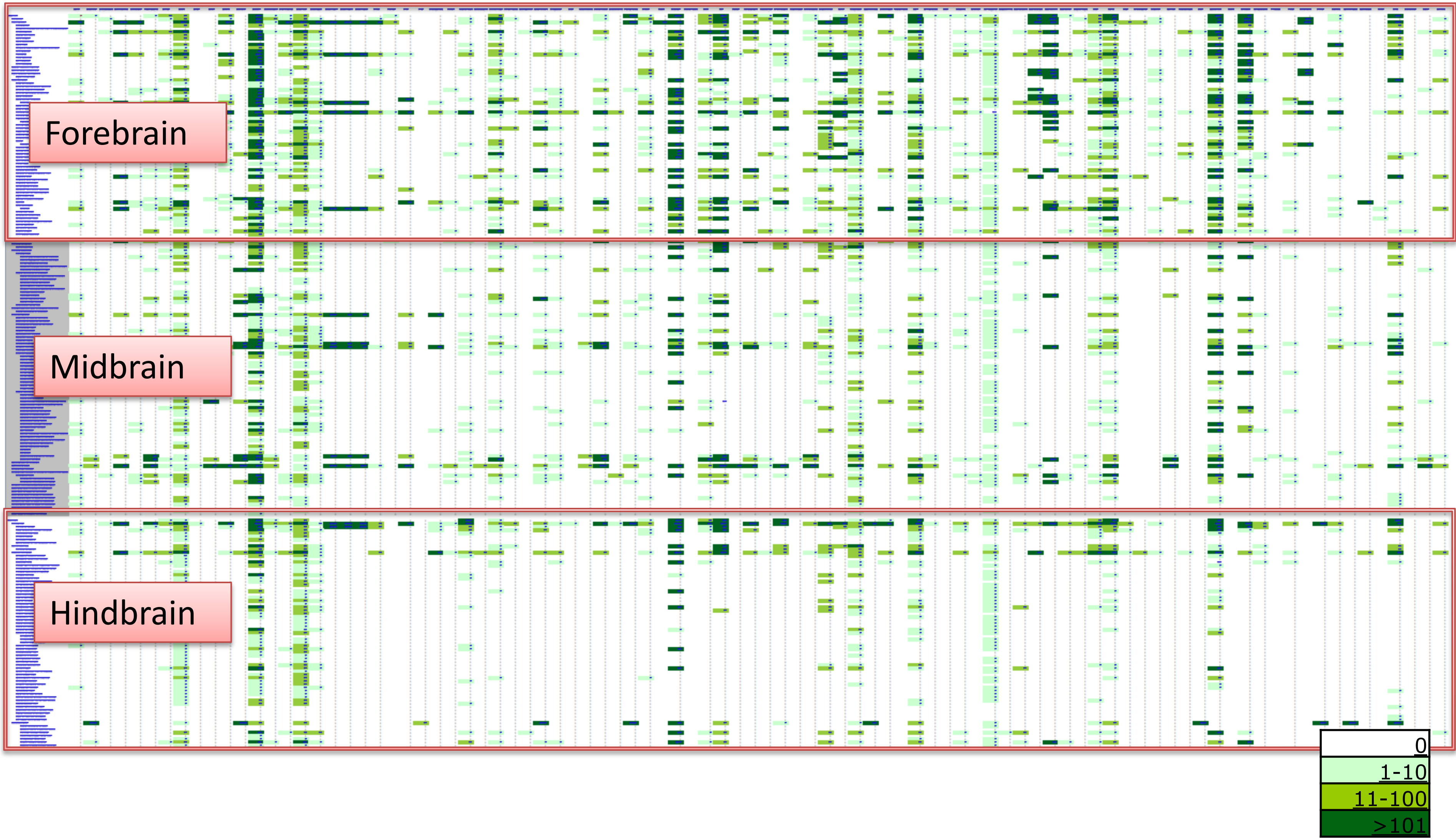
Data Sources



0
1-10
11-100
>101

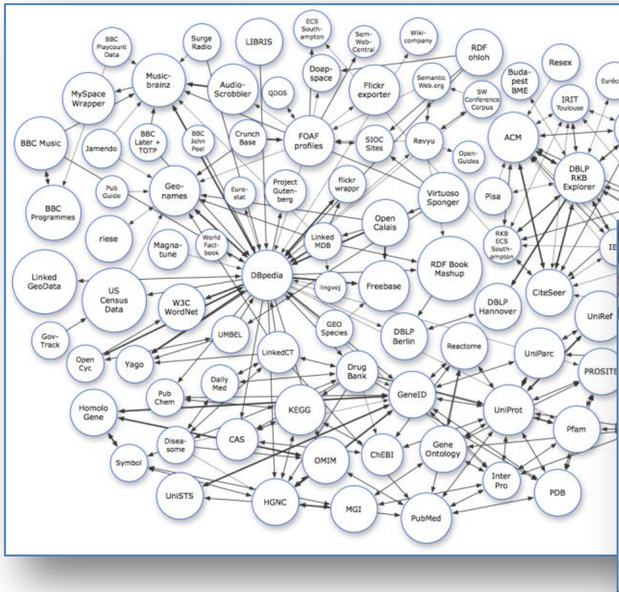
Neuroscience Information Framework: Mapping the data space

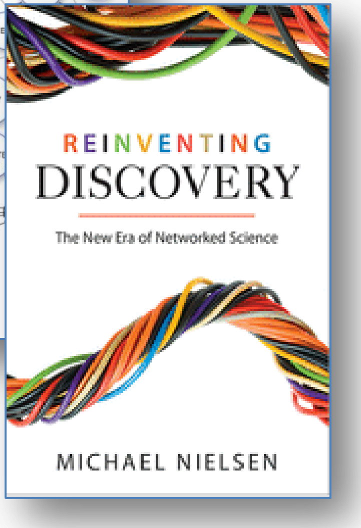
Data Sources



Basic web and web-data literacy

- Data scraping
- Web services/API’s
- URI’s
- DOI’s
- “Web of Data”:
  - RDF
  - Semantic web
  - Linked data

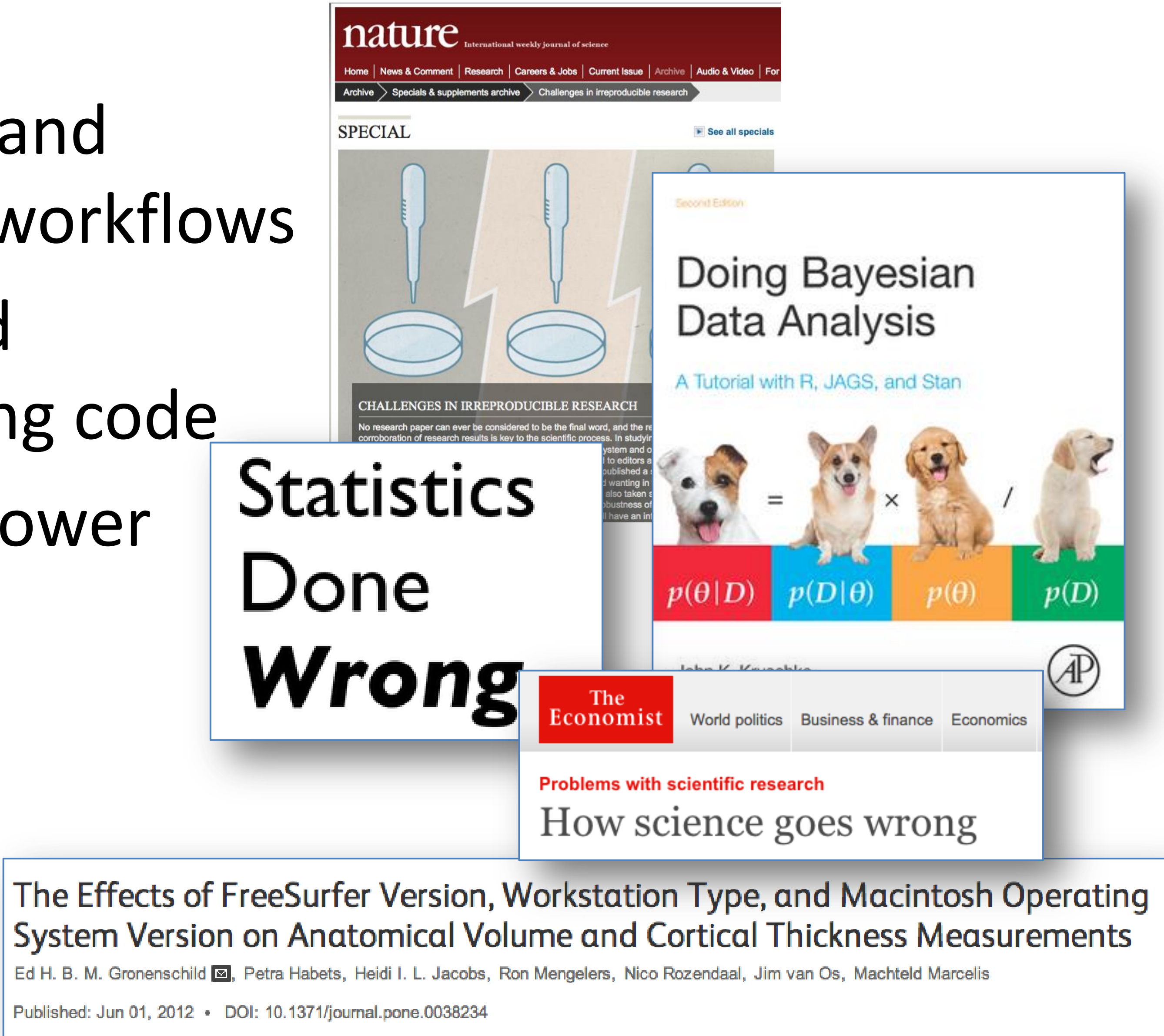




“Whichever technology wins broad adoption will become, by default, the data web. That’s why we don’t need to know which technological vision of the data web will win to conclude that the data web is inevitable”-Michael Nielson

Reproducibility

- Workflows and publishing workflows
- Sharing and documenting code
- Statistics, power analysis



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Specials & supplements archive > Challenges in irreproducible research

SPECIAL

See all specials

CHALLENGES IN IRREPRODUCIBLE RESEARCH

No research paper can ever be considered to be the final word, and the re
corroboration of research results is key to the scientific process. In studyir
ystem and o
l to editors a
published a s
l wanting in
also taken s
obustness of
l have an int

Second Edition

Doing Bayesian Data Analysis

A Tutorial with R, JAGS, and Stan

$p(\theta|D) = p(D|\theta) \times p(\theta) / p(D)$

John K. Kruschke

Statistics Done ***Wrong***

The Economist | World politics | Business & finance | Economics

Problems with scientific research

How science goes wrong

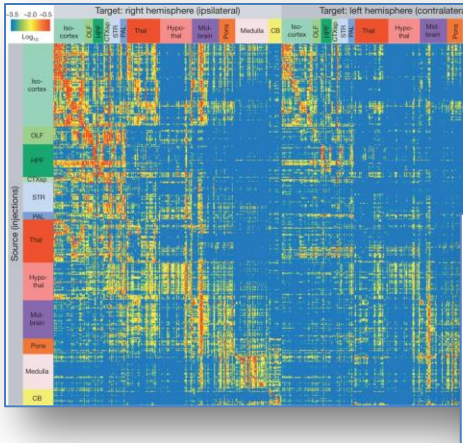
The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

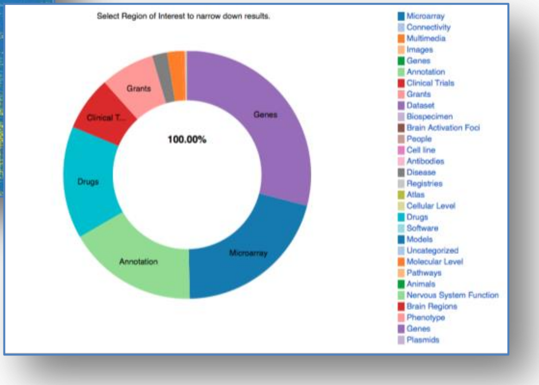
Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: Jun 01, 2012 • DOI: 10.1371/journal.pone.0038234

Analysis and Analytics

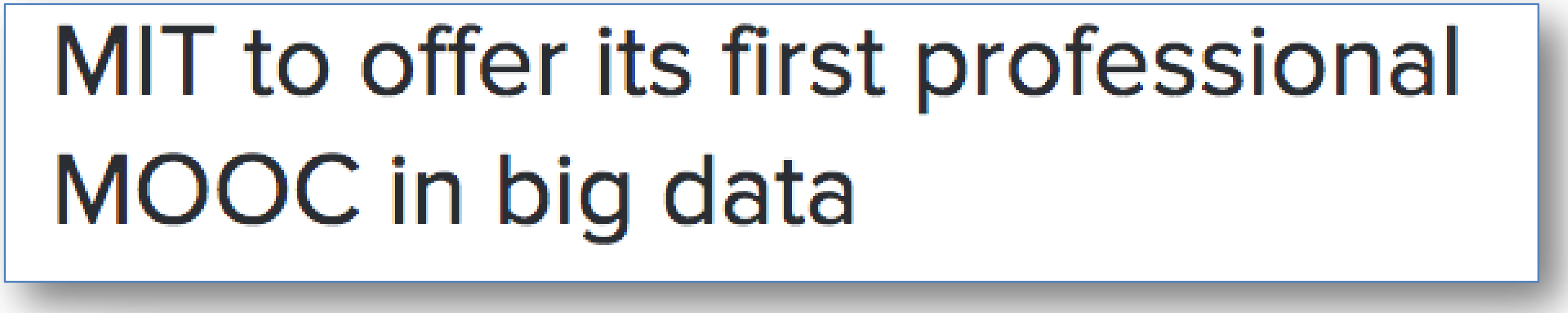
- Analysis:
  - Matlab
  - R
  - Workflows
- Analytics: discovery and communication of meaningful patterns in data
  - Many packages, opportunities and pitfalls



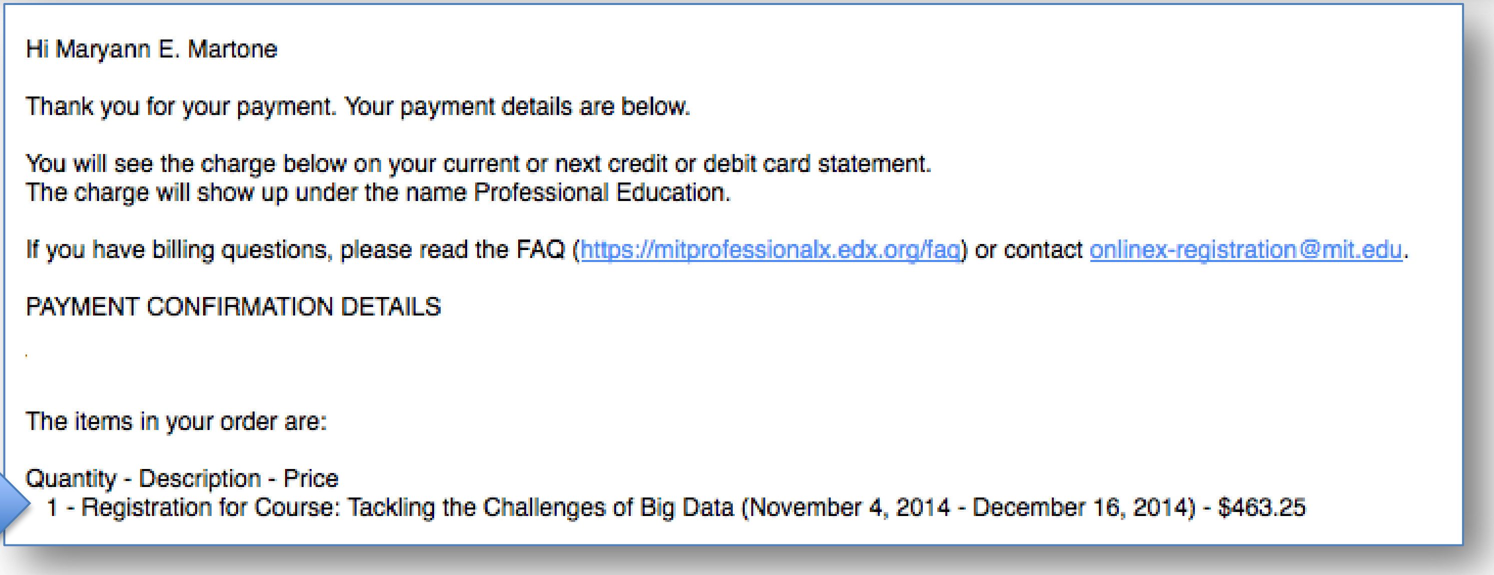

Select Region of Interest to narrow down results.

Training in Big Data

"Big data comes to us all" –Chronicle of Higher Education



MIT to offer its first professional MOOC in big data



Hi Maryann E. Martone

Thank you for your payment. Your payment details are below.

You will see the charge below on your current or next credit or debit card statement.
The charge will show up under the name Professional Education.

If you have billing questions, please read the FAQ (https://mitprofessionalx.edx.org/faq) or contact onlinex-registration@mit.edu.

PAYMENT CONFIRMATION DETAILS

The items in your order are:

Quantity - Description - Price
1 - Registration for Course: Tackling the Challenges of Big Data (November 4, 2014 - December 16, 2014) - $463.25

Issues: Specializations

- Will all neuroscientists need to be data scientists or will they need to communicate with data scientists?
  - Will academia be able to retain any data scientists?
- Will all neuroscientists need to be data managers and data curators, or will there be opportunities for other specializations?
  - Library sciences
  - Publishers
  - Government
- Will these other specialists have a career path, or will they live in our current patronage-based, soft-money academic system?
- Better tools will remove some of the burdens
  - Designing academic tools and data: make them FAIR