Data-Driven Science and Health: Being F.A.I.R.

Gregory Butler

Department of Computer Science & Software Engineering
Data Science Research Centre
Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

November 2019 — Malaysia

Outline

Outline

Open Science and the F.A.I.R. Guidelines

Open Science

Open Access

Open Source Software

Open Data

Open Resources (like reagents, cell lines, etc)

Open Peer-Review

The Need for Open Science

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

of broad interest to a general medical audience.

among those tested in the field. R

DOI: 10.1371/journal.pmed.0020124

IS THERE A REPRODUCIBILITY CRISIS?

Reproducibility Crisis in Neuroimaging

- Noisy data and incomplete statistics can lead to spurious results (Bennett et al., 2011)
- Dominant software libraries have inflated false positive rates (Eklund et al., 2016)
- 1-voxel perturbations to inputs result in significantly different outputs (Lewis et al., 2016)
- Similar tools performing similar operations give different results (Bowring et al., 2018)
- Operating system differences have led to different results (Glatard et al., 2015)

What does it mean for a tool to be FAIR?

Findable

1. Globally persistent records
2. Described with rich metadata
3. Searchable

We leverage **Zenodo [2]** to create DOIs for Boutiques descriptors which can be accessed via the Zenodo API.

Accessible

1. Easily retrievable
2. Universal access
3. Persistent metadata beyond data lifetime

The retrievable tool descriptions contain **immutable** human- and machine-readable instructions for testing and launching each tool.

Interoperable

1. Formalized and shared metadata standard
2. Metadata standards adopted are FAIR
3. Linking between objects where appropriate

CARMIN [3] and **Boutiques [4]** standards are used to describe and launch tools, either locally or through a RESTful API.

Re-Usable

1. Multiple accurate and relevant attributes
2. Clearly licensed
3. Meets minimum domain standards

Docker [5] and **Singularity [6]** virtualization enable re-runability across platforms and enclosed testing. Simulation and querying allow runtime evaluation.

MD Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016

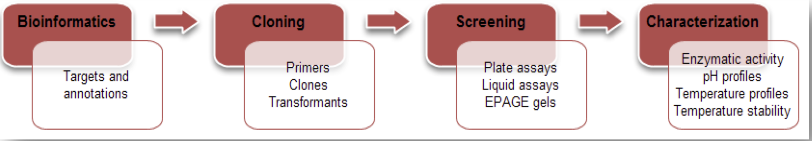
Outline

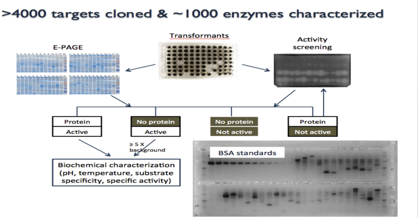
My Data-Driven Research

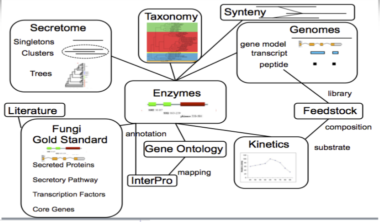
Greg Butler (CSE) – Centre for Structural & Functional Genomics

'Omics Analysis

LIMS





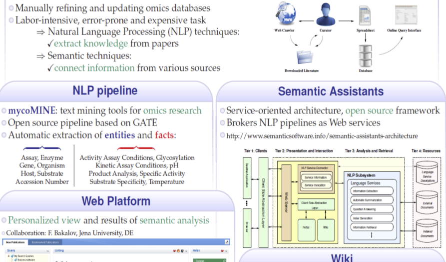


30 fungal genomes

500 "profiles"

50GB per dataset

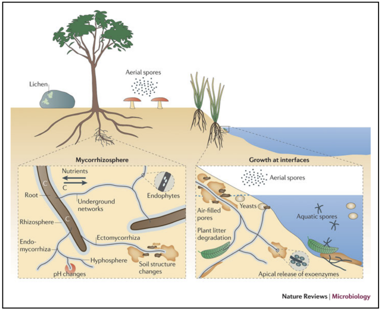
40 TB total



Data Integration

Curation & Text Mining with Drs Witte & Kosseim (CSE)

Fungi



Symbiosis

- plant roots
- lichen
- *"noble rot"*
- microbiome

Pathogens

- Plant *blight, smut, mould*
  red pine beetle
- Human *aspergillosis*, *C. albicans*
- Bacteria, insects, frogs, animal

Food

- yeast
- edible mushrooms

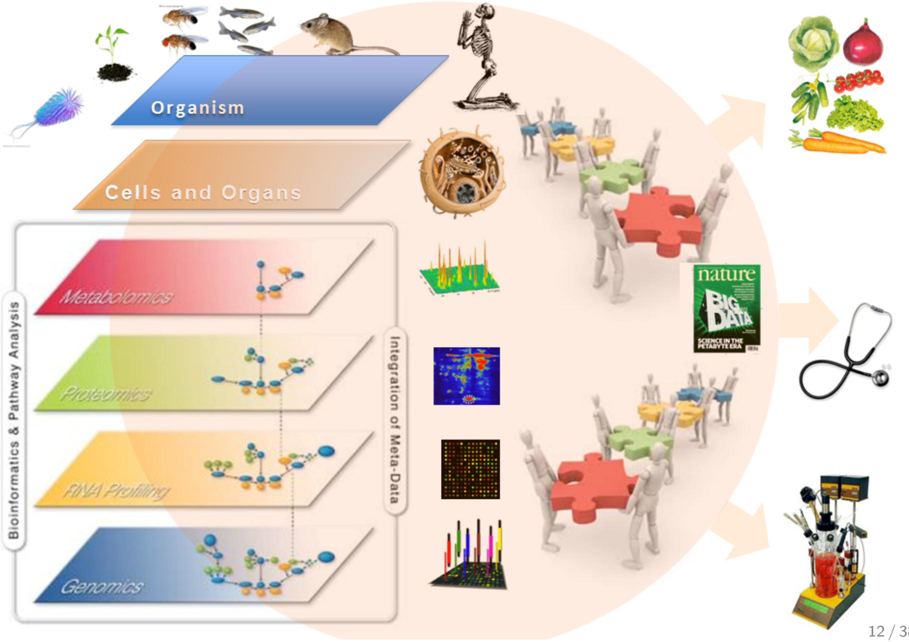
Degradation

- plant litter
- polyphenols

Microbiomes

- cattle rumen, elk, deer, muskoxen, etc
- termite gut

Life Science data: Multi-omics, multi-technology, multi organism, multi dimensional



The Toot Suite Project

Genome Canada BCB 2017 Competition

TooT Suite: Predication and classification of membrane transport proteins, Gregory Butler and Tristan Glatard, 2018–2021

Bioinformatics and Machine Learning

Develop predictors for transporter proteins and membrane proteins

Open Science

tools — open source
platform for experiments — Boutiques + bfx tools + ML tools
reproducible experiments

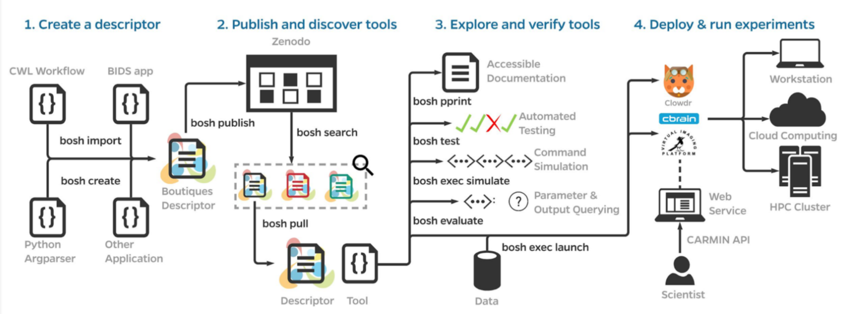
Scale to microbiomes

Motivation

Improve agricultural productivity
provide tools to help understand microbiome-host interaction

Toot Suite — Experimental Infrastructure

Boutiques using Docker



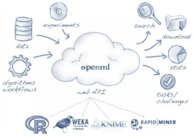
Compute Canada

MP2 cluster: 1632 nodes, 12 core/node, 32-512 GB/node

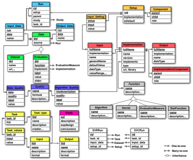
T Glatard et al, Boutiques: a flexible framework to integrate command-line applications in computing platforms. Gigascience. 2018 May 1;7(5)

Toot Suite — F.A.I.R.

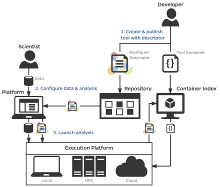
openML



openML



Boutiques



Data Sources

- UniProt
- TCDB
- GO, ChEBI, Pfam
- ABC, SLC

Findable

Zenodo, doi

Accessible

DockerHub

Interoperable

Boutiques, Galaxy, CWL

Reusable

Boutiques, openML

MD Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016

Outline

Data-Driven Science & Health

Actionable Data in Data-Driven (Clinical) Healthcare

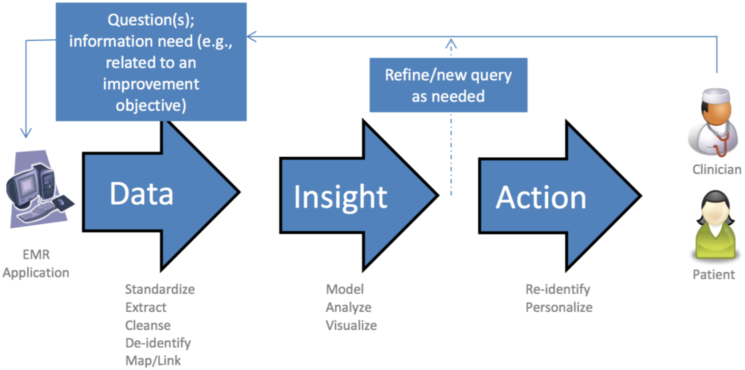
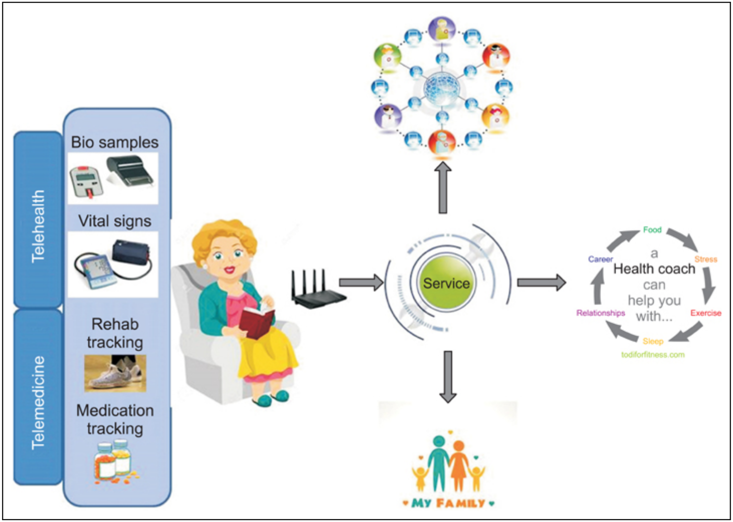


Figure 2: Framework to Create Clinical Value from EMR Data

(Infoway Health Canada 2016)

The Elderly or Remote Patient Perspective

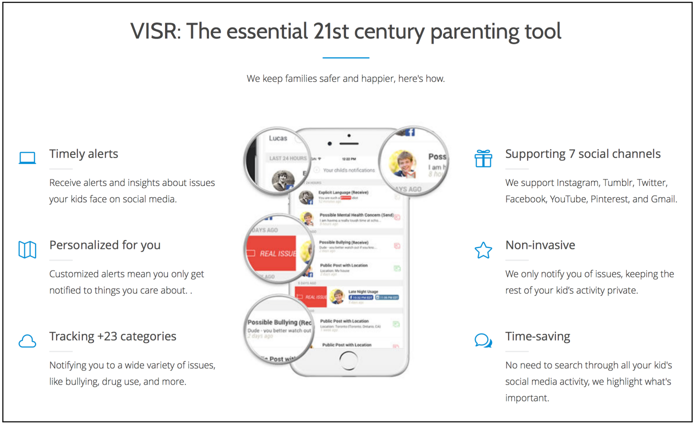


Dimitrov (Health Informatics Research, 2016)

VISR — A Canadian Company

Better mental and emotional health via social media data mining

"On a mission to help families better navigate technology, by notifying parents about safety and wellness issues their kids face on social media"



http://www.visr.co

Applications for Big Data in Healthcare

Diagnostics
Data mining and analysis to identify causes of illness

Preventative medicine
Predictive analytics and data analysis of genetic, lifestyle, and social circumstances to prevent disease

Precision medicine
Leveraging aggregate data to drive hyper-personalized care

Medical research
Data-driven medical and pharmacological research to cure disease and discover new treatments and medicines

Reduction of adverse medication events
Harnessing of big data to spot medication errors and flag potential adverse reactions

Cost reduction
Identificaton of value that drives better patient outcomes for longterm savings

Population health
Monitor big data to identify disease trends and health strategies based on demographics, geography, and socio-economics

Outline

Big Data and Data Analytics

Big Data (`http://dsrc.encs.concordia.ca/what-is-bigdata.html`)

Big Data

Definition of *"Big"* has changed as we have become more advanced

History

Hollerith Cards 1890 (US population census)

Economic Data 1952 (GDP etc)

Computers 1959 — The First Digital Data Tsunami

World Wide Web 1990's — The Second Digital Data Tsunami

Social Media 1985 — The Third Digital Data Tsunami

Internet of Things 2000 — The Fourth Digital Data Tsunami

Big Science — 1960's onwards

Deep Knowledge — 2011 onwards

A key notion is **actionable data** that is useful in supporting decisions, determining actions, and adding value to an endeavour.

Big Data

The 5 V's

Volume: amount of data
Variety: different types of data
Velocity: rate at which data is generated
Veracity: trustworthiness, level of noise
Value: usefulness of data to a business
plus Visualization, Viscosity (sticky), Virality (convey a message)

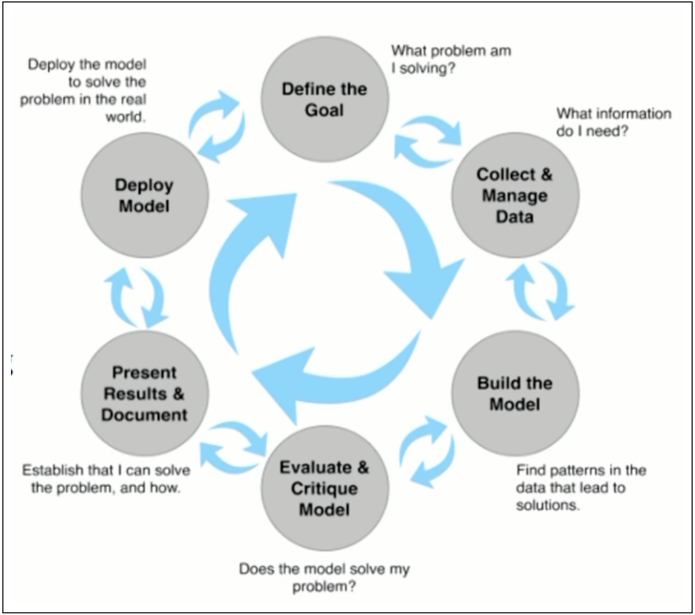
Drivers

Transactions
Mobile
Social Media
Internet of Things

MGI Report

McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, May 2011.

Data Analytics — Not (exactly) the Scientific Method



(Microsoft)

Data Analytics: Data Wrangling

Design a Data Collection Program

- Establish whether or not the data exists in the real world and is relevant to the question
- Devise a collection scheme to acquire it
  Logistical considerations? Cost? Privacy issues?
- Coordinate with departments or agencies needed for collection

Data Analytics: Data Wrangling

Collect and Review the Data

- Store the incoming data to allow modeling and reporting
- Join data from multiple sources in relevant & logical manner
- Check for anomalies or unusual patterns
  - Caused by the collection process?
  - Inherent to topic of investigation?
  - Correct them, or develop new collection scheme?

Data Analytics: Exploratory Data Analysis

Exploratory Data Analysis

Learn about the properties of the data

Steps

- Descriptive statistics: mean/median, variance/quartiles, outliers
- Correlation
- Fitting curves and distributions
- Dimension reduction
- Clustering

Data Analytics: Modeling

Modeling

Getting *"meaning"* from a clean data set

Steps

- Build a data model to fit the question
- Validate the model against the actual collected data
- Perform the necessary statistical analyses
- Machine-learning or recursive analysis
- Regression testing and other classical statistical analysis
- Compare results against other techniques or sources

Data Analytics: Modeling

The choice of a model affects (and is affected by)

- Whether the model meets the business goal
- How much pre-processing the model needs
- How accurate the model is
- How explainable the model is
- How fast the model is (in making predictions)
- How scalable the model is (building and predicting)

(Microsoft)

Approaches to Data Analysis

Scripting

Unix tools, eg
text files, csv files for inputs, outputs, intermediate steps
stepwise development of analysis
script captures steps, parameters
easy to replay

Notebooks

Jupyter, eg
interactive scripting with "literate programming"
keep track of thought processes during analysis
work with files to replay analysis

"Spreadsheet" Environments

OpenRefine, eg
lots of tools, little guidance
need macros, histories, to capture/replay work
often proprietary

Big Data Analytics — Compute Clusters & the Cloud

Map Reduce Approach

Hadoop, Spark
Distributed database support (HBase)

Knowledge Graphs

Linked data, ontologies & semantic web

Cloud

Flexible, distributed computing, as needed

noSQL Databases

Modern technology for varieties of data

Outline

Take Home Lessons

Technology and Computation is not the Goal

improved quality of life is!

Knowledge is key

not data!

Veracity (Trust, Traceability, Accountability) is essential!

cf chain of reasoning (math); traceability (SE);
provenance (to sci. literature); blockchain

Be open, transparent, and F.A.I.R.

Make data & knowledge
    Findable, Accessible, Interoperable, and Reuseable

Thank You!

Questions, Please?

Privacy and Security

*"**Privacy** refers to an individual's right to control the collection, use, and disclosure of his/her personal health information (PHI) and/or personal information (PI) in a manner that allows health care providers to do their work.*

***Security** is about ensuring the information gets to the right person in a secure manner."*

Ontario's Ehealth Blueprint `http://www.ehealthblueprint.com`

Privacy by Design 2009

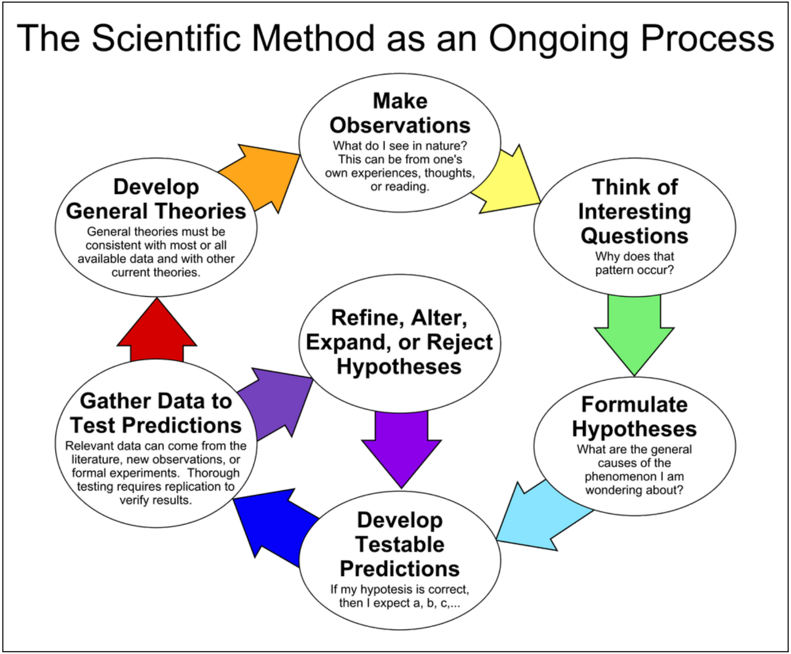
Seven Foundational Principles

1) being proactive not reactive;
2) having privacy as the default setting;
3) having privacy embedded into design;

4) avoiding the pretence of false dichotomies,
such as privacy vs. security;

5) providing full life-cycle management of data;
6) ensuring visibility and transparency of data; and
7) being user-centric

Prof. Ann Cavoukian, formerly Information and Privacy Commissioner of Ontario; now Ryerson University. `http://www.privacybydesign.ca`



https://raeonscience.weebly.com)

Scientific Method

Hypothesis-driven Experimental Design and Analysis

Not exploratory data analysis (EDA).

You have a single, specific hypothesis to accept or reject.

Steps

- Set null hypothesis H_0 and alternative hypothesis H_1
- Design experiment to collect data, and
- Design analysis of experimental data to accept/reject hypothesis
- Determine *statistical power* of experiment
  Do you have enough data points?
- Do experiment, do analysis, accept/reject hypothesis