OPENING SCIENCE IN A DATA-DRIVEN WORLD

Olivier Verscheure, PhD
Swiss Data Science Center
EPFL & ETH Zurich

SRCE e-infrastructure Days, Zagreb, Croatia – April 11, 2018
About me

Academia

New York

17y

1999

2016

Dublin

2010
Data Science – A fragmented ecosystem

What is the hyperplane that best separates two classes of points in multidimensional space?

How can I best match the right drug with the right dosage to the right patient at the right time?
How is my data protected?
How private is it?
How exactly is it used?
What is the hyperplane that best separates two classes of points in multidimensional space?

How can I best match the right drug with the right dosage to the right patient at the right time?

The Swiss Data Science Center (SDSC)

Foster adoption of data science both in academia and industry

Data scientists

Domain experts

Data providers

ETH Zurich + EPFL

Multi-disciplinary team of 40 full-time computer and data scientists, and domain experts

Foster adoption of data science in both academia and industry

What is the hyperplane that best separates two classes of points in multidimensional space?
What do you see?
A fantastic source of data!
Data is a valuable resource

Crunch time in France
Ten years on: banking after the crisis
South Korea’s unfinished revolution
Biology, but without the cells

The world’s most valuable resource

Data and the new rules of competition
“YEP... GOT MY CELLPHONE, MY PAGER, MY INTERNET LINK, MY WIRELESS FAX, AND THANKS TO THIS NIFTY SATELLITE NAVIGATING SYSTEM, I KNOW PRECISELY WHERE I AM AT ALL TIMES!”

BY LOWE FOR THE SUN-SENTINEL, FLO
And it is a resource that is often not shared
• 2017 survey of academics conducted by Elsevier and Leiden University
• Collected 1,162 responses from all scientific fields

• 73% agreed that having access to researchers’ data would be beneficial
• 36% not willing or undecided to let other researchers’ access their data

• Only 13% of scholars published their data in a data repository
Yet, sharing may foster innovation
… leading to multidisciplinary collaborations
Why are researchers hesitant to open data?

74%
- I am concerned that my research will be scooped
- I am concerned about misinterpretation or misuse
- I am concerned about being given proper citation credit or attribution

57%
- I did not know where to share my data
- Insufficient time and/or resources
- I did not know how to share my data

42%
- Intellectual property rights, privacy concerns and ethical issues

2014 Wiley’s Researcher Data Insights Survey
Why are researchers hesitant to open data?

74%
- I am concerned that my research will be scooped
- I am concerned about misinterpretation or misuse
- I am concerned about being given proper citation credit or attribution

57%
- I did not know where to share my data
- Insufficient time and/or resources
- I did not know how to share my data

42%
- Intellectual property rights, privacy concerns and ethical issues

2014 Wiley’s Researcher Data Insights Survey
Why are researchers hesitant to open data?

- 74%: I am concerned that **my research will be scooped**
- 74%: I am concerned about **misinterpretation or misuse**
- 74%: I am concerned about being given **proper citation credit or attribution**

- 57%: I did not know where to share my data
- 57%: Insufficient time and/or resources
- 57%: I did not know how to share my data

- 42%: Intellectual property rights, privacy concerns and ethical issues

2014 Wiley’s Researcher Data Insights Survey
Challenges & Opportunities – Instill trust

• **Trust the source** – Traceability and transparency
  • Can I trust the quality and the veracity of this research result?

• **Trust the consumers** – Accreditation and attribution
  • Is my research properly credited?
  • Who is using my data? for what purpose? is it used as agreed?

• **Trust the system** – Secure federated environment
  • Is the system secure?
  • Am I in full control of the system?
  • Am I disclosing more information than I intend?
### Why are researchers hesitant to open data?

- **74%**
  - I am concerned that my research will be scooped
  - I am concerned about misinterpretation or misuse
  - I am concerned about being given proper citation credit or attribution

- **57%**
  - I did not know where to share my data
  - Insufficient time and/or resources
  - I did not know how to share my data

- **42%**
  - *Intellectual property rights, privacy concerns and ethical issues*

---

*2014 Wiley’s Researcher Data Insights Survey*
Beyond Open Data – The rise of data sharing

The Data Spectrum

- Internal access
  - Employment contract + policies
  - Sales reports
- Named access
  - Explicitly assigned by contract
  - Driving licences
- Group-based access
  - Via authentication
  - Medical research
- Public access
  - Licence that limits use
  - Twitter feed
- Open
  - Open licence
  - Bus timetable

Small / Medium / Big data
Personal / Commercial / Government data
Anecdote: The (bad) politics of Open Data

• The US Environmental Protection Agency (EPA) environmental regulations are set using the “best possible science”

• Until the last couple of years, “best possible science” was understood as evidence from peer-reviewed scientific publications in leading journals

• Recently, several bills were pushed in Congress to eliminate what the so-called “Secret Science”: science based on data not shared in its entirety

• For a scientific study to be used to inform EPA regulations, all data would have to be made available. The objective of the bills is for the studies to be easily “reproducible”.
• Environmental health studies rely on patient cohorts or health claims data (such as Medicare or Medicaid in the US), protected by stringent privacy requirements.

• Former EPA requirements were about replication (= have a consistent body of evidence that a regulation would have a positive impact).

• Current legal direction is only about reproducibility.

• Open data, even if wishable from a general scientific viewpoint, is not an option from a legal perspective.

• Open data + health data = NO SCIENCE, unless privacy-protecting data sharing tools are deemed acceptable by Congress.

Anecdote: The (bad) politics of Open Data
But sharing data is not enough

Data collected from multiple sources:
- Difficult to find
- Difficult to access
- Difficult to understand
- Difficult to reuse
Most research is not reproducible today

1 year later ...
From Open Science to Collective Intelligence

- Reproducible Research
  - See the (versioned) algorithms
  - See the (versioned) data
  - Replay a workflow
  - Compare workflows, validate robustness

- Reusability, replicability
  - Reuse data on new workflows
  - Clone and modify workflows

- Credit and attribution
  - Data popularity, H-index
  - Who is using the data?
  - For what?

- IP Protection, confidentiality
  - Decide who sees the data,
  - The algorithms,
  - The data I use,
  - And how I use it
RENGA - 連歌

- Provide scientists means to create **reproducible** science
- Facilitate the **sharing** and **reuse** of research artifacts
- Enable **rapid discovery** of relevant data and methods
- Foster a **collaborative environment** for rapid prototyping of novel approaches
- Adhere to **FAIR principles** and **DMP** mandate
- Allow federated access across institutions, giving each the freedom to impose its own access controls over resources
Renga (連歌), plural renga, a genre of Japanese linked-verse poetry in which two or more poets supply alternating sections of a poem linked by verbal and thematic associations.

—Encyclopædia Britannica
Capturing and recording data provenance is the core of Renga

1. Inputs and outputs of analysis steps are recorded into a knowledge graph
2. Steps can be repeated or integrated into more complex workflows
3. Provenance of all data products always accessible via simple tools
4. Version control built-in for data, code, and workflows
1. Tracking research provenance

- JSON-LD for better interpretability and interoperability
- Extended Dublincore ontology
- Provenance graphs based on PROV-O W3C recommendation
- CWL for representing all computational steps
- Capture individual steps from user input
- Tools for constructing workflows from basic pieces
- Rely on container technologies to ensure reproducibility
2. Discovery

Graph-based search...

Search for a publication, obtain a full view of the relevant *data lineage*
3. Reusability

- Explore workflows and data interactively
- Query the knowledge graph for data and code usage
- Easily reuse work from others, linking lineage and metadata
- Identify popular datasets and algorithms across the platform; audit

... and reuse in an entirely new context
Coming soon: RENGA Federation

Federated Parties

Genome Center
Proteome Center
USZ

public
private
Coming soon: RENGA Federation

I will trust no one but my Renga instance.

I make no assumption about other instances.

Anything could happen to my data once it is out.

Admin

Admin

Admin

Genome Center

Proteome Center

USZ

RENGA Federation

SDSC
Coming soon: RENGA Federation

One stop shop

login

Federated Parties

Genome Center
Proteome Center
USZ
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ
One stop shop

Federated Parties

Proteome Center

Genome Center

USZ

Coming soon: RENGA Federation
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ

RENGA

SDSC
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ

SDSC
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ

RENGA

Reproduce
Results
Coming soon: RENGA Federation

One stop shop

Reusability

Federated Parties

Data

Infrastructure

Lineage

Algorithms

Genome Center

Proteome Center

USZ
Coming soon: RENGA Federation

One stop shop

Federated Parties

Genome Center
Proteome Center
USZ

SDSC
Expected new release end of April 2018

RENGA 連歌
Highly-scalable & secure open software platform designed to foster multidisciplinary data science collaborations across mutually-interesti
IN GOD WE TRUST.

ALL OTHERS MUST BRING DATA.

- W. EDWARDS DEMING, STATISTICIAN, PROFESSOR, AUTHOR
RENGA Adherence to FAIR Principles

• **Findable**
  • “data and meta-data should be easy to find by both humans and computers”
  • In Renga’s knowledge representation, all entities, including data sets and methods, are uniquely identifiable, and described using standard ontologies (Dublin-core, Prov-o). The entities are searchable by id, key terms and extended relationships.

• **Accessible**
  • “data and meta-data should be stored for the long-term, such that they can be easily accessed and downloaded using standard communication protocols.”
  • Renga provides a one-stop shop secure HTTP REST endpoint to access data and code, with respect to access policies.

• **Interoperable**
  • “data is ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets”
  • Renga’s knowledge representation supports standard ontologies capable to describe the schema of the data content and encoding methods in a machine and human interpretable way.

• **Reusable**
  • “Data and metadata are well-described and can be reused in future research. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans”
  • Renga guarantees that data is always reusable with confidence. It also (1) provide an enviornment allowing the reproducibility of the research that created the data, and (2) can be used to enforce the conditions under which the data is accessed or the research reproduced.