THE ROOTS

LINKED DATA AND THE FOUNDATIONS OF SUCCESSFUL AGRICULUTURE DATA

Dr. Paul Groth | @pgroth | pgroth.com

Disruptive Technology Director

Elsevier Labs | @elsevierlabs

G20 Workshop Linked Open Data and Agriculture

September 27, 2017

QUESTIONS FOR THIS WORKSHOP

1. How can Linked Open Data make a difference in agriculture?
2. What technical obstacles stand in the way?
3. What policies are needed to achieve the potential?

DATA IS CENTRAL IN PRECISION AGRICULTURE

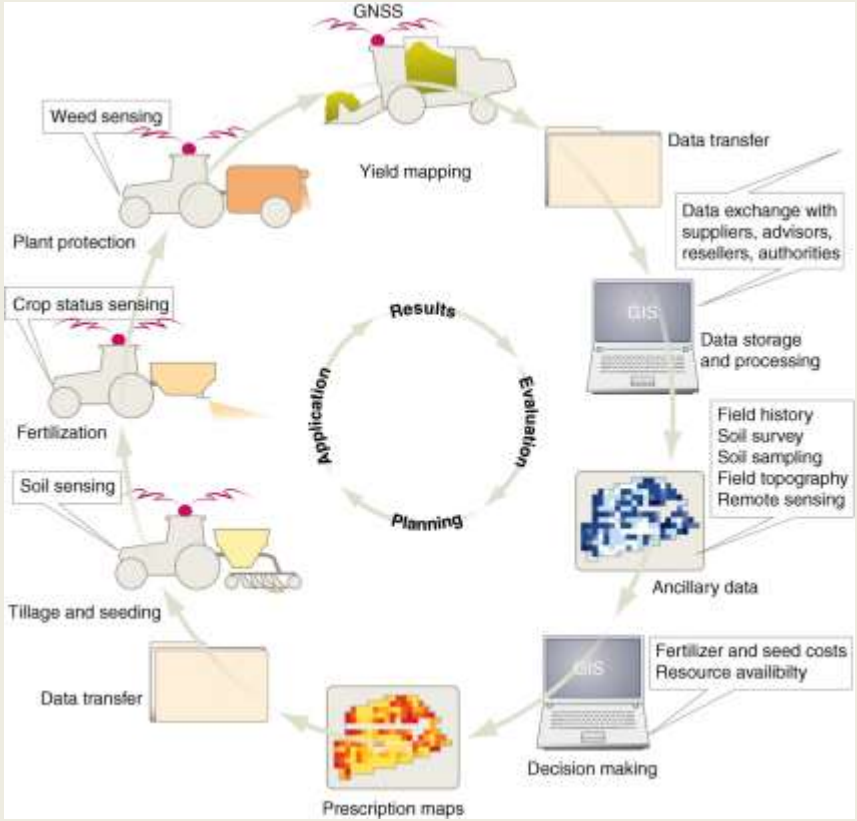


Fig. 2 Precision agriculture information flow in crop production [after (19), modified].

Robin Gebbers, and Viacheslav I. Adamchuk
Science 2010;327:828-831

Published by AAAS

THE DATA SUPPLY CHAIN IN AGRICULTURE

Table 4. State of the art of Big Data applications in Smart Farming and key issues.

Stages of the data chain	State of the art	Key issues
Data capture	Sensors, Open data, data captured by UAVs (Faulkner and Cebul, 2014) Biometric sensing, Genotype information (Cole et al., 2012) Reciprocal data (Van 't Spijker, 2014)	Availability, quality, formats (Tien, 2013)
Data storage	Cloud-based platform, Hadoop Distributed File System (HDFS), hybrid storage systems, cloud-based data warehouse (Zong et al., 2014)	Quick and safe access to data, costs (Zong et al., 2014)
Data transfer	Wireless, cloud-based platform (Karim et al., 2014; Zhu et al., 2012), Linked Open Data (Ritaban et al., 2014)	Safety, agreements on responsibilities and liabilities (Haire, 2014)
Data transformation	Machine learning algorithms, normalize, visualize, anonymize (Ishii, 2014; Van Rijmenam, 2015)	Heterogeneity of data sources, automation of data cleansing and preparation (Li et al., 2014)
Data analytics	Yield models, Planting instructions, Benchmarking, Decision ontologies, Cognitive computing (Van Rijmenam, 2015)	Semantic heterogeneity, real-time analytics, scalability (Li et al., 2014; Semantic Community, 2015)
Data marketing	Data visualization (Van 't Spijker, 2014)	Ownership, privacy, new business models (Orts and Spigonardo, 2014)

Sjaak Wolfert, Lan Ge, Cor Verdouw, Marc-Jeroen Bogaardt, Big Data in Smart Farming – A review, In Agricultural Systems, Volume 153, 2017, Pages 69-80, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2017.01.023.

Table 4. State of the art of Big Data applications in Smart Farming and key issues.

Stages of the data chain	State of the art	Key issues
Data capture	Sensors, Open data, data captured by UAVs (Faulkner and Cebul, 2014) Biometric sensing, Genotype information (Cole et al., 2012) Reciprocal data (Van 't Spijker, 2014)	Availability, quality, formats (Tien, 2013)
Data storage	Cloud-based platform, Hadoop Distributed File System (HDFS), hybrid storage systems, cloud-based data warehouse (Zong et al., 2014)	Quick and safe access to data, costs (Zong et al., 2014)
Data transfer	Wireless, cloud-based platform (Karim et al., 2014; Zhu et al., 2012), Linked Open Data (Ritaban et al., 2014)	Safety, agreements on responsibilities and liabilities (Haire, 2014)
Data transformation	Machine learning algorithms, normalize, visualize, anonymize (Ishii, 2014; Van Rijmenam, 2015)	Heterogeneity of data sources, automation of data cleansing and preparation (Li et al., 2014)
Data analytics	Yield models, Planting instructions, Benchmarking, Decision ontologies, Cognitive computing (Van Rijmenam, 2015)	Semantic heterogeneity, real-time analytics, scalability (Li et al., 2014; Semantic Community, 2015)
Data marketing	Data visualization (Van 't Spijker, 2014)	Ownership, privacy, new business models (Orts and Spigonardo, 2014)

WHERE LINKED DATA CAN HELP

Sjaak Wolfert, Lan Ge, Cor Verdouw, Marc-Jeroen Bogaardt, Big Data in Smart Farming – A review, In Agricultural Systems, Volume 153, 2017, Pages 69-80, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2017.01.023.

STARTING FROM THE GROUND UP

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

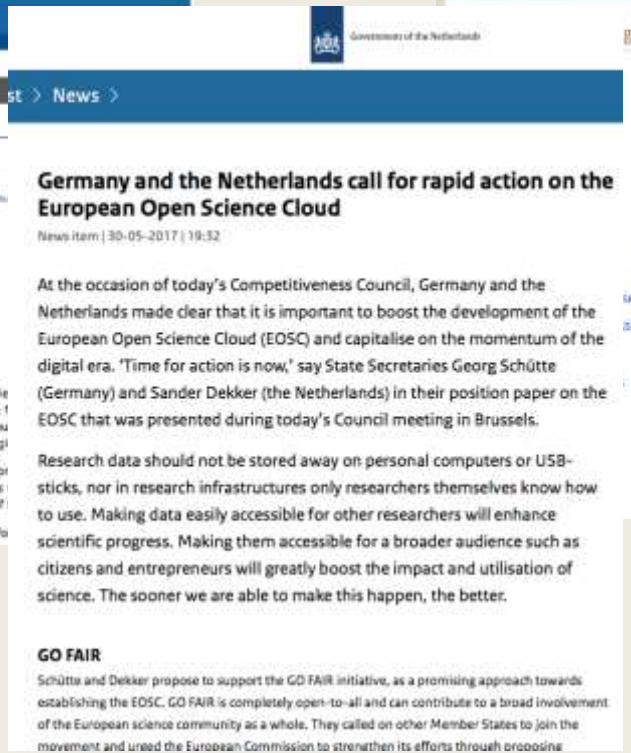
Received: 10 December 2015

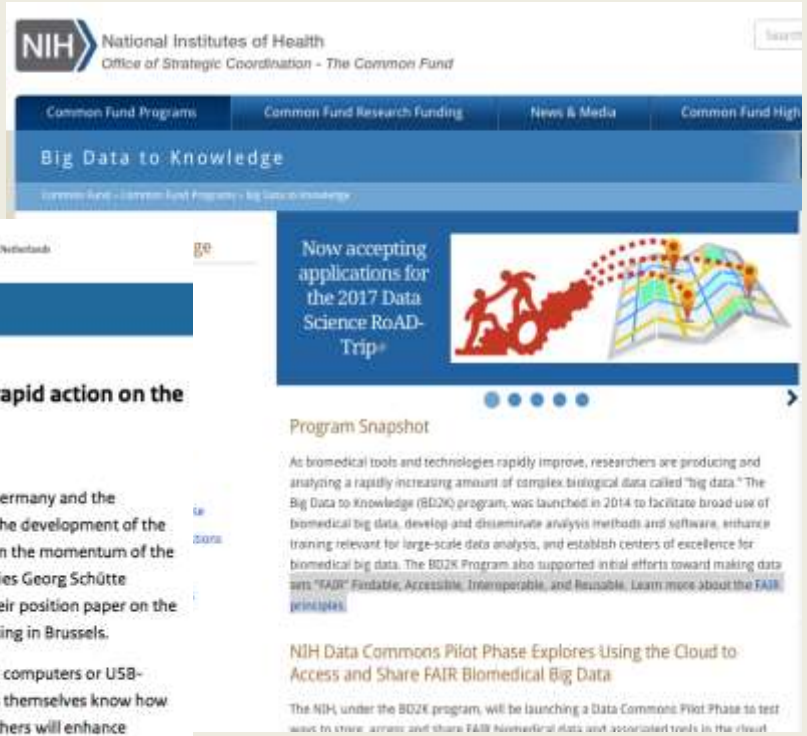
Accepted: 12 February 2016

Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

FAIR EVERYWHERE





Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

CREATING SUCCESSFUL DATA

ENCOURAGING THE RESEARCHER

FORCE11

The Future of Research Communications and e-Scholarship

Search

English

ABOUT ▾ COMMUNITY ▾ GROUPS RESOURCES ▾ NEWS + BLOGS ▾ EVENTS ▾ PUBLICATIONS ▾ MEDIA ▾ DONATE

JOINT DECLARATION OF DATA CITATION PRINCIPLES - FINAL

When citing please use: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [/datacitation].

ENDORSEMENT LIST

PREAMBLE

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and

DC1

Data Citation Principles

Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices
"The TOP Guidelines"

Version 1.0.1

Reproducibility of research can be improved by increasing transparency of the research process and products. This document provides template guidelines to enhance transparency in the science that journals publish. With minor adaptation of the text, funders can adopt these guidelines for research that they fund.

There are eight transparency standards covered by these guidelines. The guidelines are modular so they can be adopted singly or collectively:

1. Citation
2. Data transparency
3. Analytic methods (code) transparency
4. Research materials transparency
5. Design and analysis transparency
6. Preregistration of studies
7. Preregistration of analysis plans
8. Replication

Each category template text for three levels of transparency: Level 1, Level 2, and Level 3. Adopting journals select among the levels based on readiness to adopt milder to stronger transparency standards for authors and researchers. There are many factors that will influence

Single-molecule analysis of mtDNA replication uncovers the basis of the common deletion

Published: 4 Jan 2017 | **Version 1** | **DOI:** 10.17632/hctwmmpj9r.1

Contributor(s): Agnel Sfeir, Aaron Phillips, Marco tigano, erika brunet

Description of this data

Data included in Phillips et al.,

Experiment data files

Download all files (337)

3r347q hets copy.tiff	3 MB
Figure 3_gel guide.pptx	8 MB
Figure 3_wt_fwd-R374QFwd copy.ab1	250 KB
Figure 3_61_1-R374Q_cDNA_HET_T_FWD copy.ab1	214 KB
Figure 3_66_8-R374Q_cDNA_HET_ii_FWD copy.ab1	210 KB
Figure 3_r347q hets copy.tiff	3 MB

Associated article

peer reviewed

This data is associated with the following peer reviewed publication:

Single-Molecule Analysis of mtDNA Replication Uncovers the Basis of the Common Deletion

Cite this article

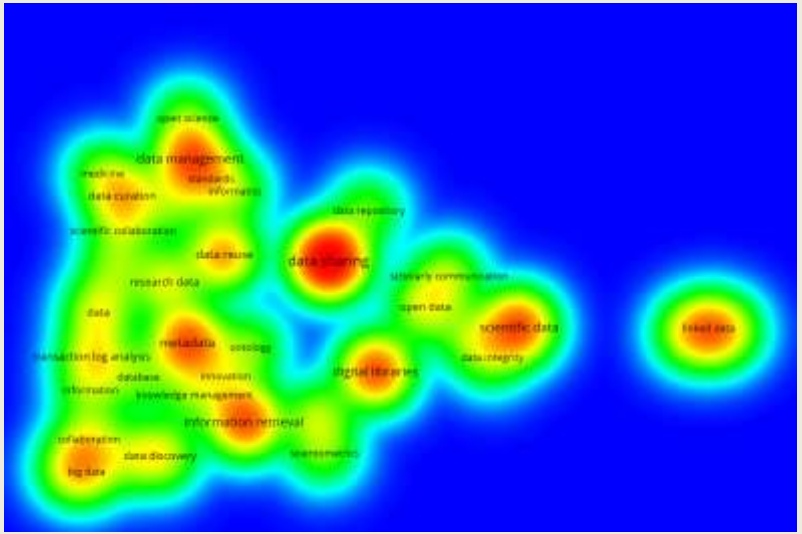
Published in:

Molecular Cell

HOW DO RESEARCHERS SEARCH FOR DATA?

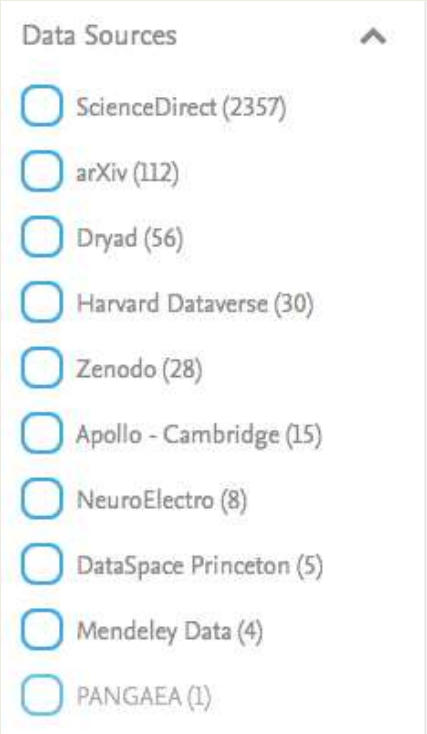
Some observations from @gregory_km survey:

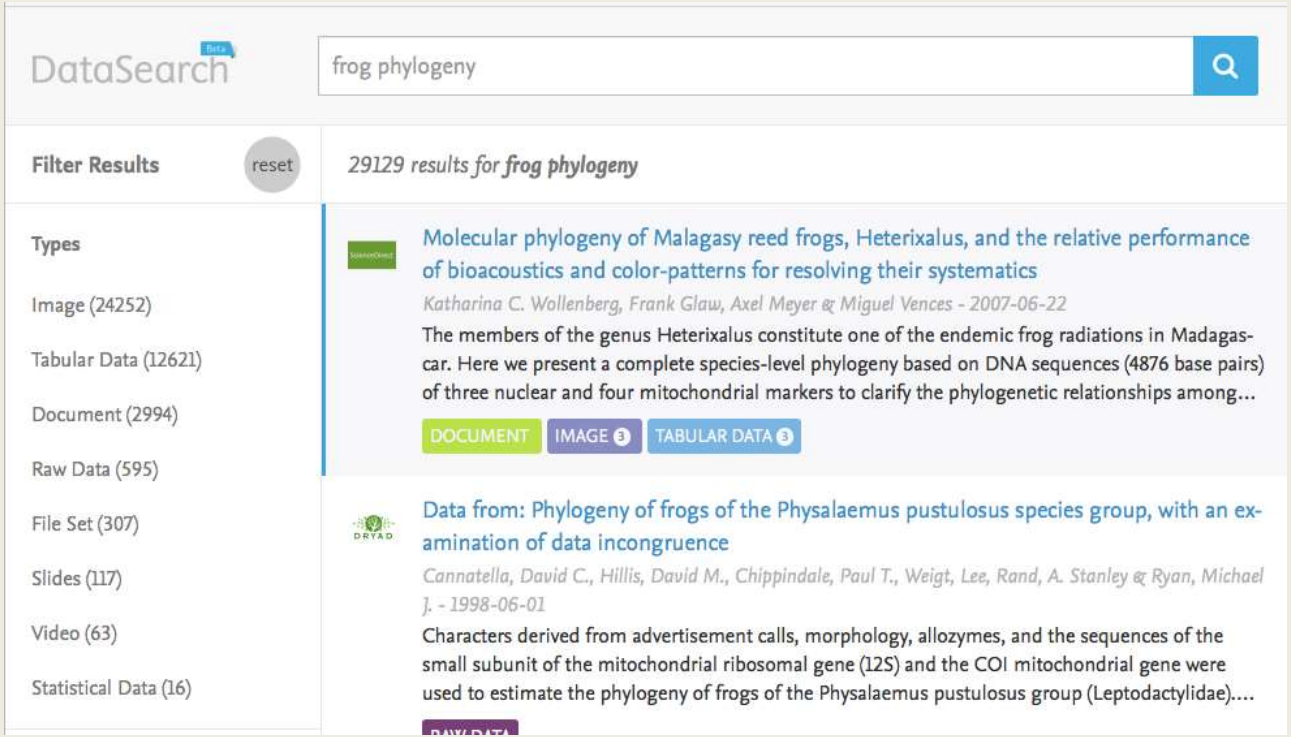
1. The needs and behaviours of specific user groups (e.g. early career researchers, policy makers, students) are not well documented.
2. Background uses of observational data are better documented than foreground uses.
3. Reconstructing data tables from journal articles, using general search engines, and making direct data requests are common.



Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2017). Searching Data: A Review of Observational Data Retrieval Practices. *arXiv preprint arXiv:1707.06937*.

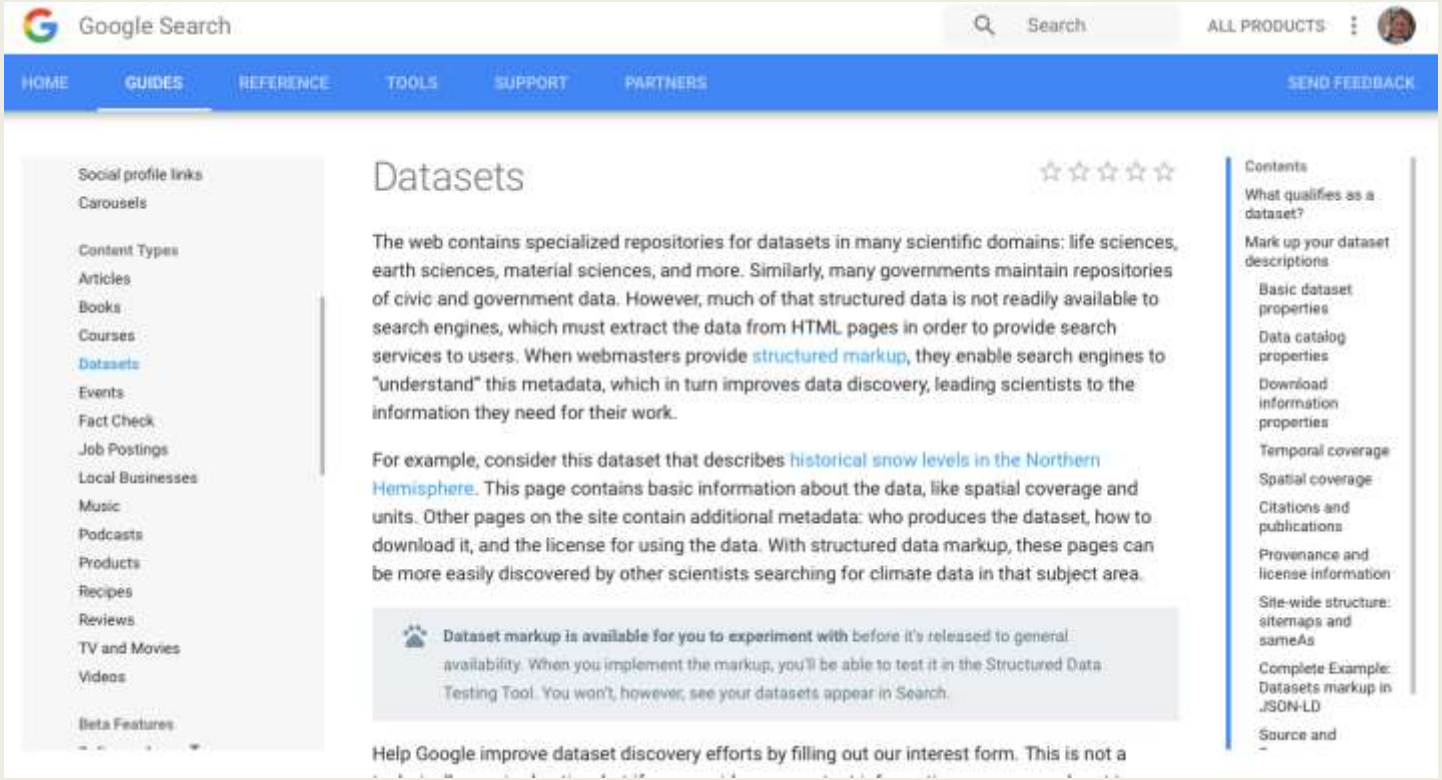
DATA SEARCH





Antony Scerri, John Kuriakose, Amit Ajit Deshmane, Mark Stanger, Peter Cotroneo, Rebekah Moore, Raj Naik, Anita de Waard; Elsevier's approach to the bioCADDIE 2016 Dataset Retrieval Challenge, *Database*, Volume 2017, 1 January 2017, bax056, https://doi.org/10.1093/database/bax056

ENABLING DATASET DISCOVERY



<script type="application/ld+json">

```
{
"@context": "http://schema.org/",
  "@type": "Dataset",
  "name": "Single-molecule analysis of mtDNA replication uncovers the basis of the common deletion",
  "description": "Data included in Phillips et al., ",
  "url": "http://dx.doi.org/10.17632/hctwmmpj9r.1",
  "sameAs": "https://data.mendeley.com/datasets/hctwmmpj9r",
  "version": "1",
  "keywords": "Molecular Biology",
  "publisher": "Mendeley Data",
  "mainEntityOfPage": {
    "@type": "WebPage",
    "@id": "https://data.mendeley.com/datasets/hctwmmpj9r/1"
  },
  "datePublished": "2017-01-04T00:15:20.232Z",
  "dateModified": "2017-01-04T00:15:20.232Z",
  "author": {
    "name": "Agnel Sfeir"
  },
  "includedInDataCatalog": "https://data.mendeley.com",
  "distribution": {
    "fileFormat": "application/octet-stream",
    "contentURL": "/archiver/hctwmmpj9r?version=1"
  },
  "citation": {
    "@type": "ScholarlyArticle",
    "text": "Sfeir, Agnel; Phillips, Aaron; tigano, Marco ; brunet , erika  (2017), 'Single-molecule a
    "headline": "Single-molecule analysis of mtDNA replication uncovers the basis of the common deleti
    "image": "https://data.mendeley.com/journal-images/10972765.jpg",
    "datePublished": "2017",
    "dateModified": "2017",
    "url": "10.1016/j.molcel.2016.12.014"
  },
  "license": {
    "@type": "Dataset",
    "text": "CC0 1.0",
    "url": "http://creativecommons.org/publicdomain/zero/1.0/"
  }
}
```

INTEROPERABILITY & INTEGRATION

What is ISOBUS?

Ag equipment manufacturers around the world have agreed on ISOBUS as the universal protocol for electronic communication between implements, tractors and computers.

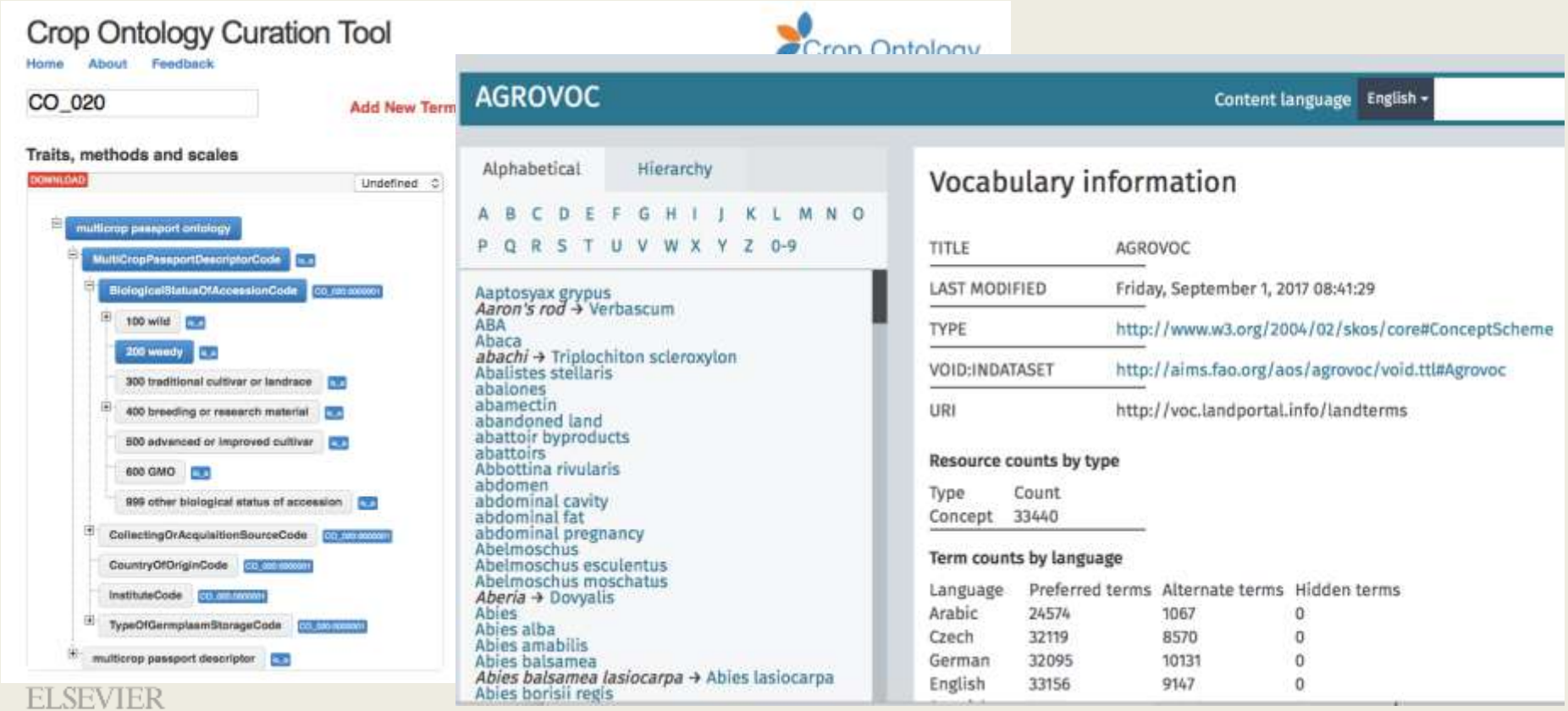
The primary goal of ISOBUS data technology is to standardize the communication which takes place between tractors and implements while ensuring full compatibility of data transfer between the mobile systems and the office software used on the farm.

The basis is the international ISO 11783 standard – "Tractors and machinery for agriculture and forestry - Serial control and communications data network".



+ click image to enlarge

MOVING UP THE STACK



INTEGRATION

Home ▸ Semantics ▸ AGROVOC ▸

AGROVOC Linked Data

ABOUT | SEARCH | ACCESS | COMMUNITY | LINKED DATA | PUBLICATIONS | GU

Linked Data is a method of web publication in which each individual piece of data is:

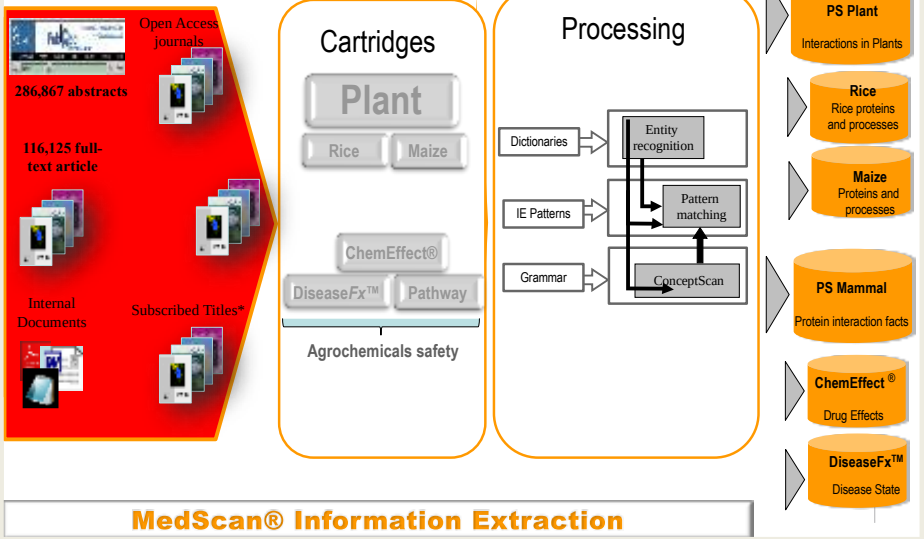
- uniquely identified using HTTP URIs (that is, URLs, or 'web addresses'),
- available both as 'machine readable' data and as 'human readable pages, and
- linked to other resources.

AGROVOC is now available as a linked data (LD) set published, aligned (linked) with several vocabularies. The Linked Data version of AGROVOC is in RDF/SKOS-XL, and is stored in Allegrograph triple store. Data is accessible to machines through a SPARQL endpoint, and to humans by means of a HTML pages generated with Loddy.

	Ryż (ziarno) 稻米
skos:inScheme	http://aims.fao.org/aos/agrovoc
skos:exactMatch	http://www.eionet.europa.eu/gemet/concept/7214 http://eurovoc.europa.eu/3732 http://d-nb.info/gnd/4049271-0 http://cat.ali.caas.cn/concept/c_8549 http://cat.ali.caas.cn/concept/c_7599 http://lod.nal.usda.gov/nalt/56293 http://id.loc.gov/authorities/sh85113862#concept http://zbw.eu/stw/descriptor/14095-0
skos:closeMatch	http://purl.org/bncf/tid/17341 http://purl.org/bncf/tid/38716 http://dbpedia.org/resource/Rice
void:inDataset	http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc

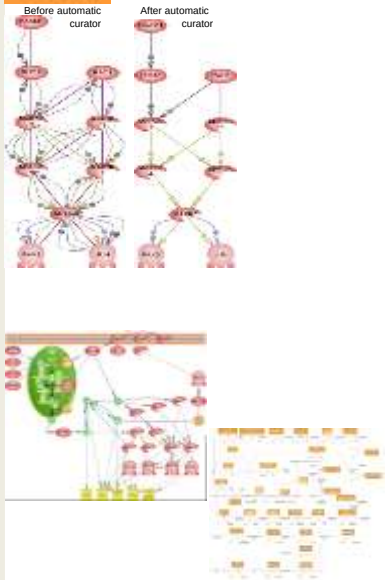
INTEGRATION ACROSS DOMAINS

Pathway Studio technology overview



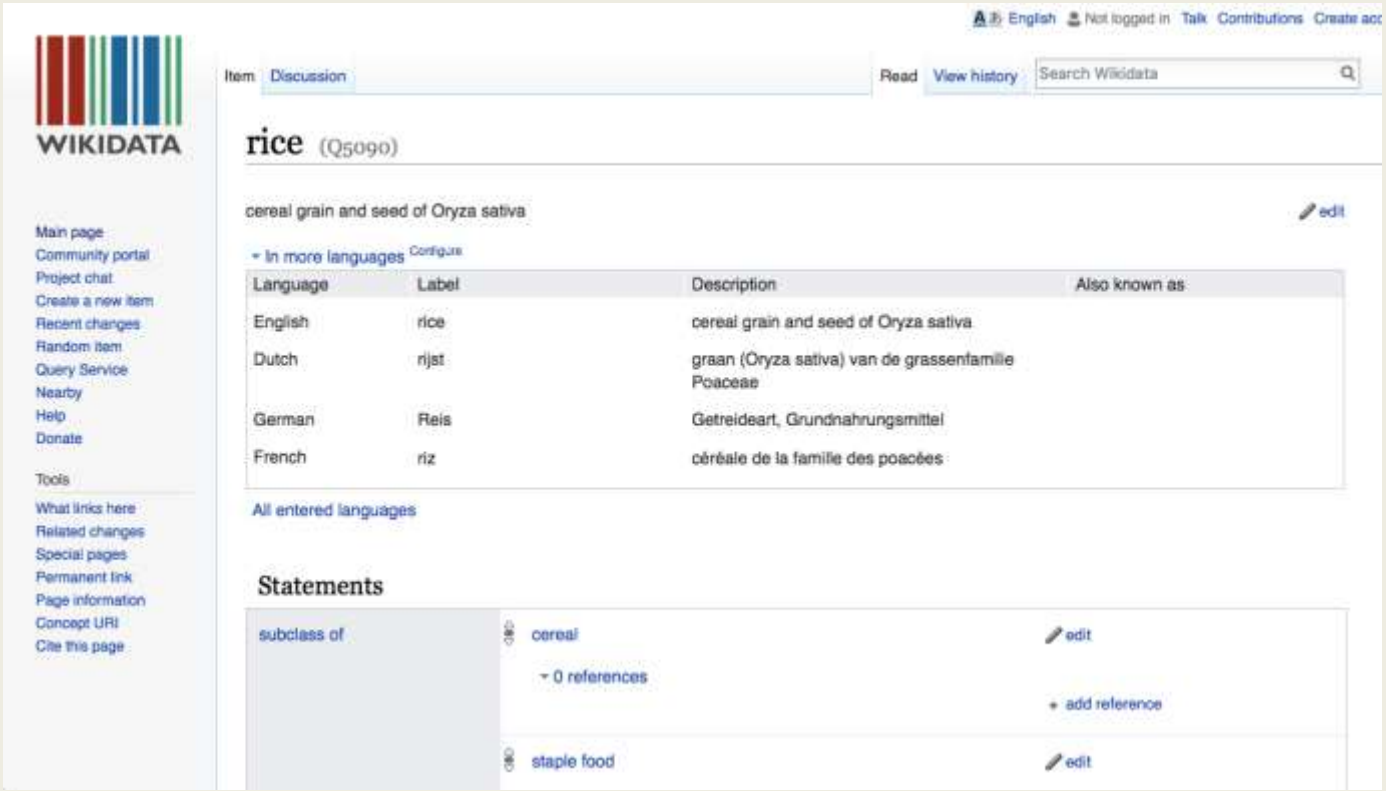
Pathway Studio Plant Knowledgebase

>778,99 mln unique relations supported by >576,083 references



- **Automatically curated MedScan data**
  - Compressed and purified by automatic curation
    - ✓ removes historical redundancy (>30%)
    - ✓ removes false positives (~5%)
- **Entity annotation**
  - Entrez Gene for proteins
  - Pubchem for chemicals
  - Aliases from MedScan dictionaries
  - Protein functional annotation from Gene Ontology
- **Ontologies**
  - Pathway Studio Ontology of intracellular signaling
  - Gene Ontology
- **Curated pathways**
  - Cell signaling pathways
  - Metabolic pathways (AraCyc)

DATA SUSTAINABILITY



THINGS TO THINK ABOUT

ARE WE MISSING A USER?

The significance of machines in data-rich research environments

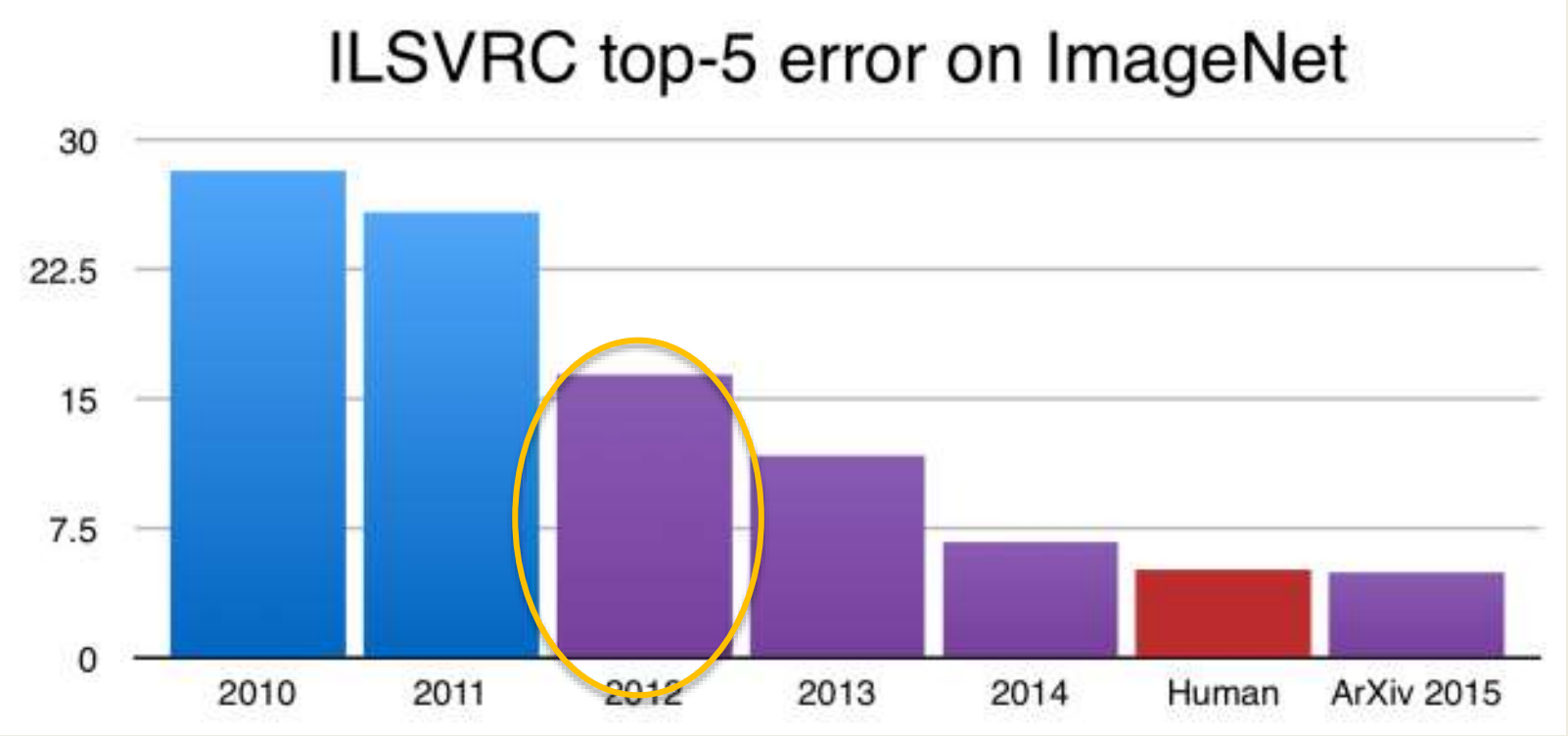
The emphasis placed on FAIRness being applied to both human-driven and machine-driven activities, is a specific focus of the FAIR Guiding Principles that distinguishes them from many peer initiatives (discussed in the subsequent section). Humans and machines often face distinct barriers when attempting to find and process data on the Web. Humans have an intuitive sense of 'semantics' (the meaning or intent of a digital object) because we are capable of identifying and interpreting a wide variety of contextual cues, whether those take the form of structural/visual/iconic cues in the layout of a Web page, or the content of narrative notes. As such, we are less likely to make errors in the

WHAT CAN MACHINE INTELLIGENCE DO TODAY?

If there's a task that a normal person can do with less than one second of thinking, there's a very good chance we can automate it with deep learning.

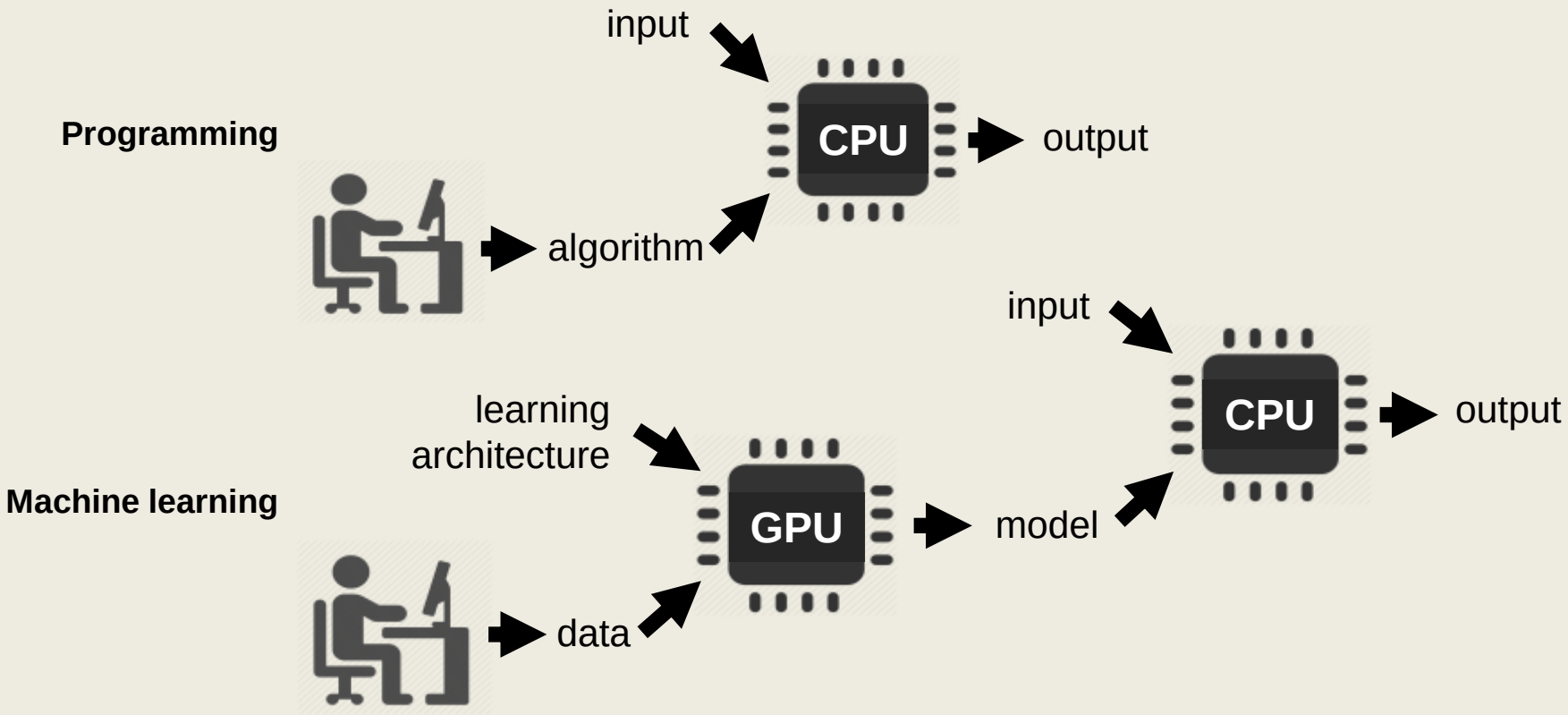
Andrew Ng, Chief Scientist, Baidu (lecture at Bay Area Deep Learning School, Stanford, CA, September 24, 2016)

IMAGE RECOGNITION



https://devblogs.nvidia.com/parallelforall/author/czhang/

ADVANCES ARE ENABLED BY MACHINE LEARNING



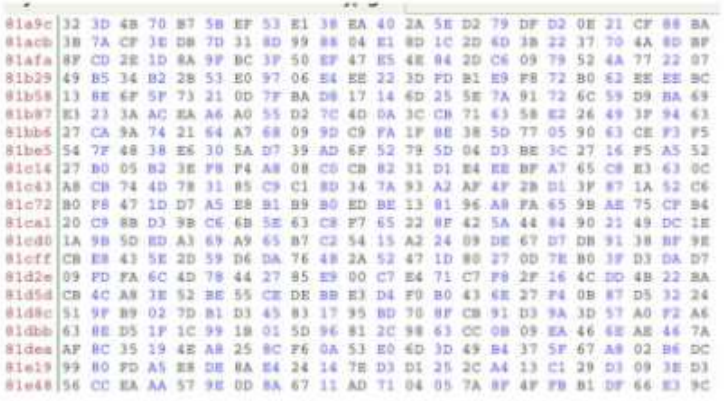
THESE RESULTS ARE DRIVEN BY DATA

“The paradigm shift of the ImageNet thinking is that while a lot of people are paying attention to models, let’s pay attention to data, …”

– Prof. Fei-Fei Li [1]

[1] The data that transformed AI research—and possibly the world https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

RAW DATA



(a) hex dump of picture of a lion

(b) same lion in human-readable format

Figure 1: The hex dump represented on the left has more information contents than the image on the right. Only one of them can be processed by the human brain in time to save their lives. Computational convenience matters. Not just entropy.

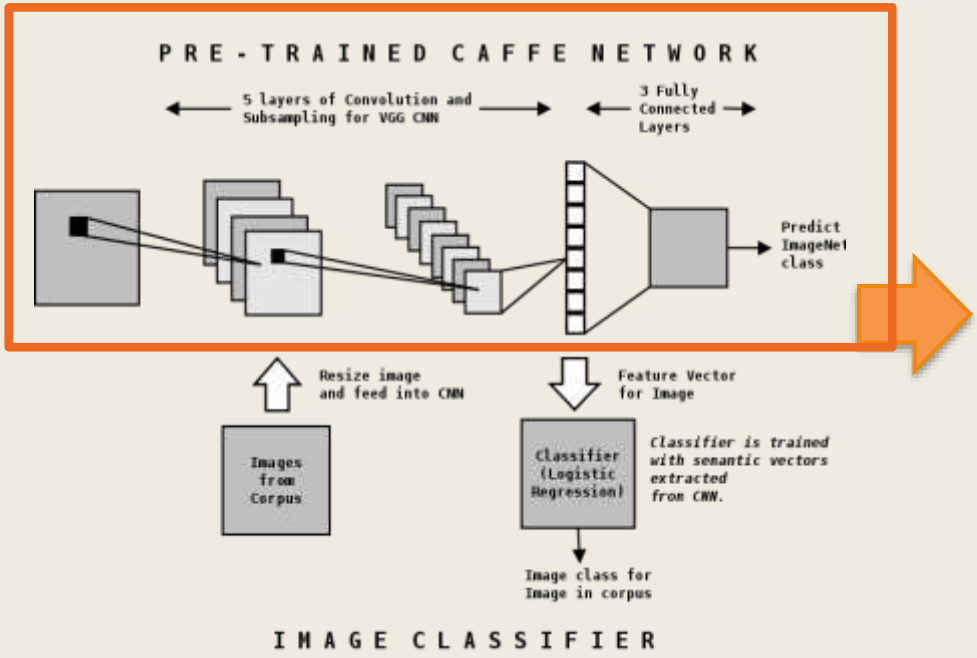
From: Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv:1610.01644v1.

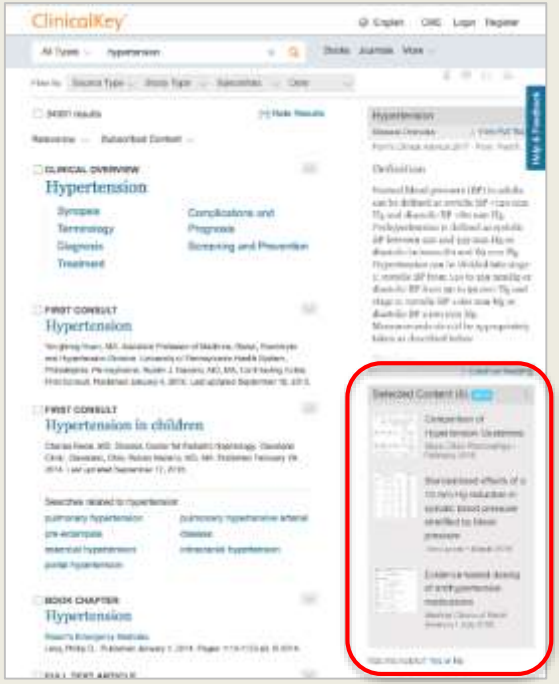
VOCABULARIES ARE SETS OF VECTOR EMBEDDINGS

Figure 3: Emoji vector embeddings, projected down into a 2-dimensional space using the t-SNE technique. Note the clusters of similar emoji like flags (bottom), family emoji (top left), zodiac symbols (top left), animals (left), smileys (middle), etc.

From: Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M. and Riedel, S. (2016). Emoji2vec: learning emoji representations from their description. arXiv:1609.08359v1.

MODELS AS REUSABLE COMPONENTS



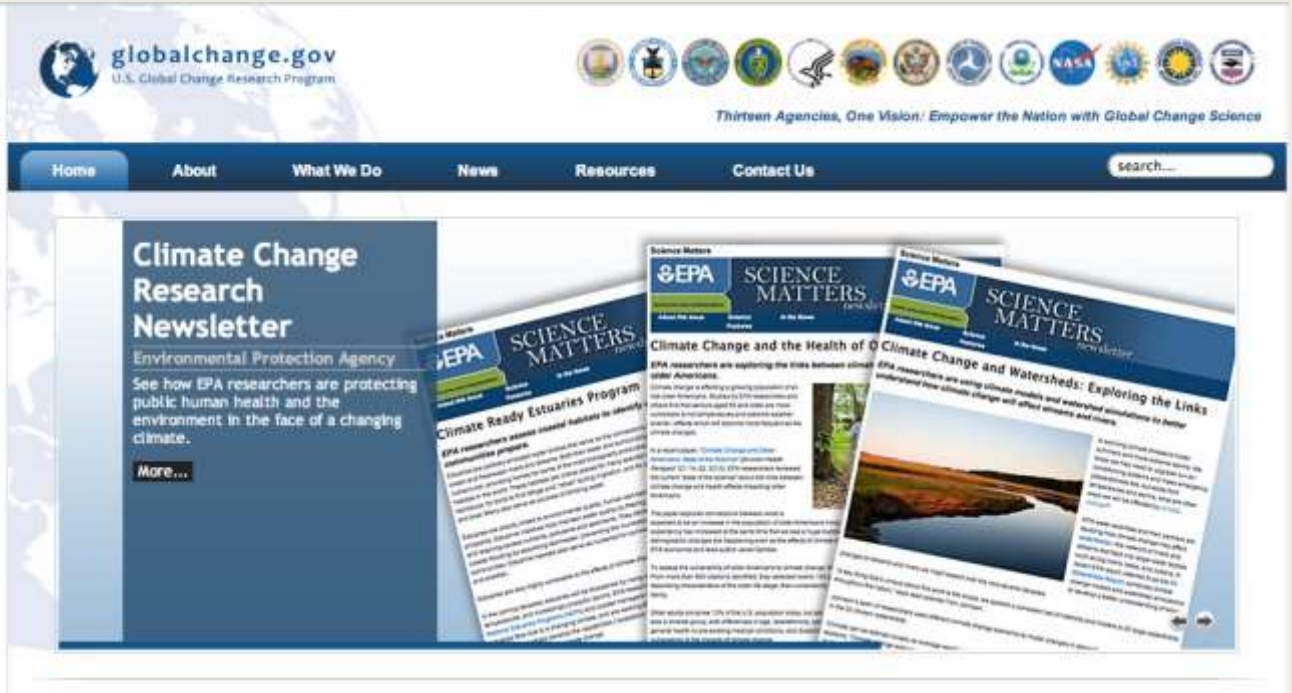


Check out: sujitpal.blogspot.com for more

LINKED DATA & MACHINE LEARNING

- Machines' proficiency in learning to answer questions from text, audio, images and video will depend on our ability to train them effectively to read information from the Web
- How machines read the Web today
    - Crawling and indexing Web resources, possibly semantically tagged (e.g. using schema.org)
    - Find-and-follow crawling of open linked data resources for ontology and data sharing and reuse
    - Programmatic access to APIs mediated through HTTP/S and other Internet protocols
- Need to think about supporting ML oriented data

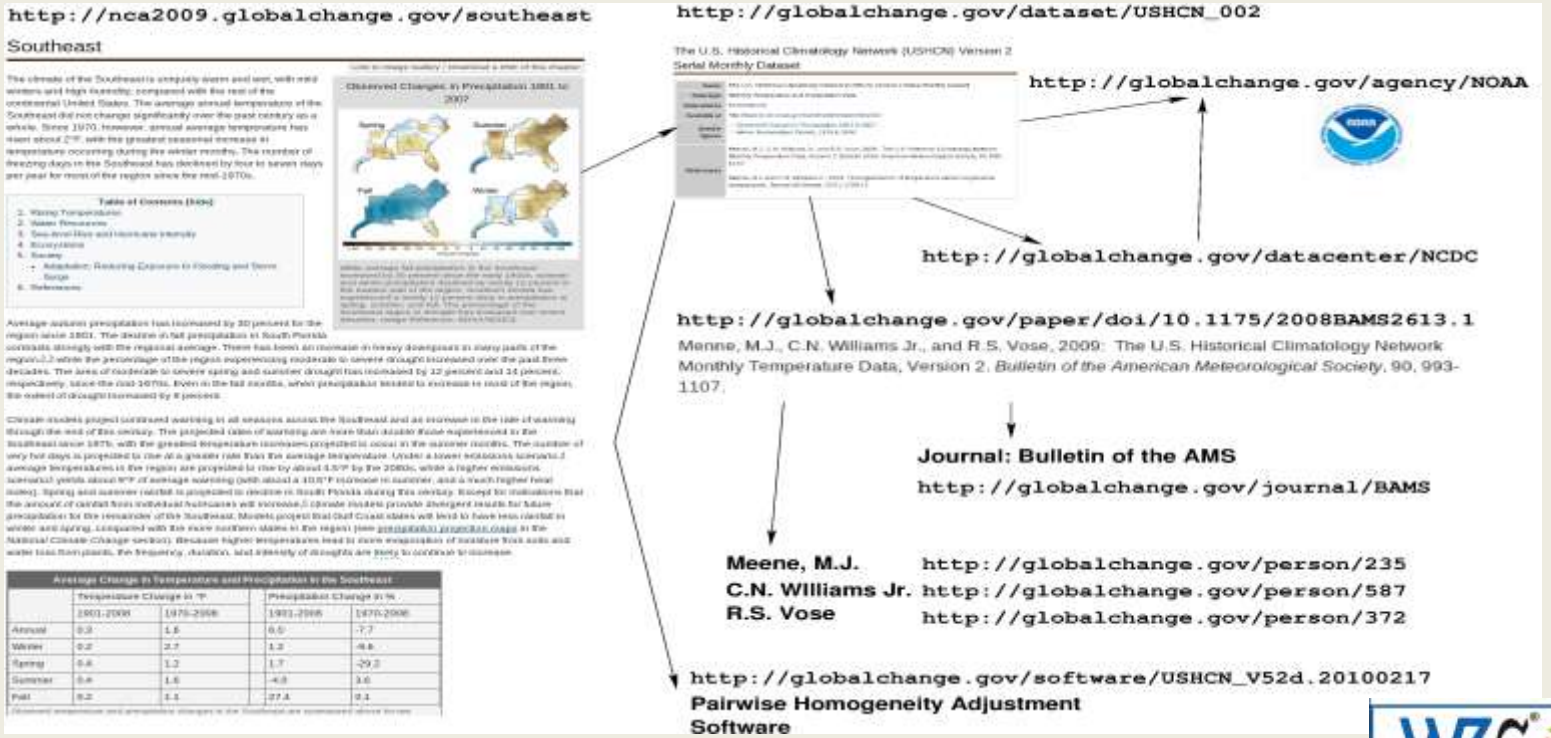
PROVENANCE FOR DATA



Credits: Curt Tilmes, Peter Fox

Tilmes, C.; Fox, P.; Ma, X.; McGuinness, D.L.; Privette, A.P.; Smith, A.; Waple, A.; Zednik, S.; Zheng, J.G., "Provenance Representation for the National Climate Assessment in the Global Change Information System," *Geoscience and Remote Sensing, IEEE Transactions on* , vol.51, no.11, pp.5160,5168, Nov. 2013

NATIONAL CLIMATE CHANGE ASSESSMENT PROVENANCE



FAIR TRADE + FAIR TRADE DATA?

Fair Trade Data

Given the data supply chain's length and the complexity of the procedures involved, the provenance of any one result can be huge and easily overwhelm users. A key challenge is to develop coherent abstractions for provenance that provide insight into the data's quality on the basis of how it was produced. Additionally, we need good mechanisms for communicating the resulting summaries. Essentially, what we need is a fair trade certificate for data — a seal of approval that says our data is produced and derived in a way that we as data consumers think is correct.

Groth, Paul, "Transparency and Reliability in the Data Supply Chain," Internet Computing, IEEE, vol.17, no.2, pp.69,71, March-April 2013 doi: 10.1109/MIC.2013.41

GOAL: SUCCESSFUL FAIR AGRICULTURE DATA

1. How can Linked Open Data make a difference in agriculture?
2. What technical obstacles stand in the way?
3. What policies are needed to achieve the potential?

THANK YOU

Dr. Paul Groth | @pgroth | pgroth.com

labs.elsevier.com