

THE ROOTS

LINKED DATA AND THE FOUNDATIONS OF SUCCESSFUL AGRICULTURE 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

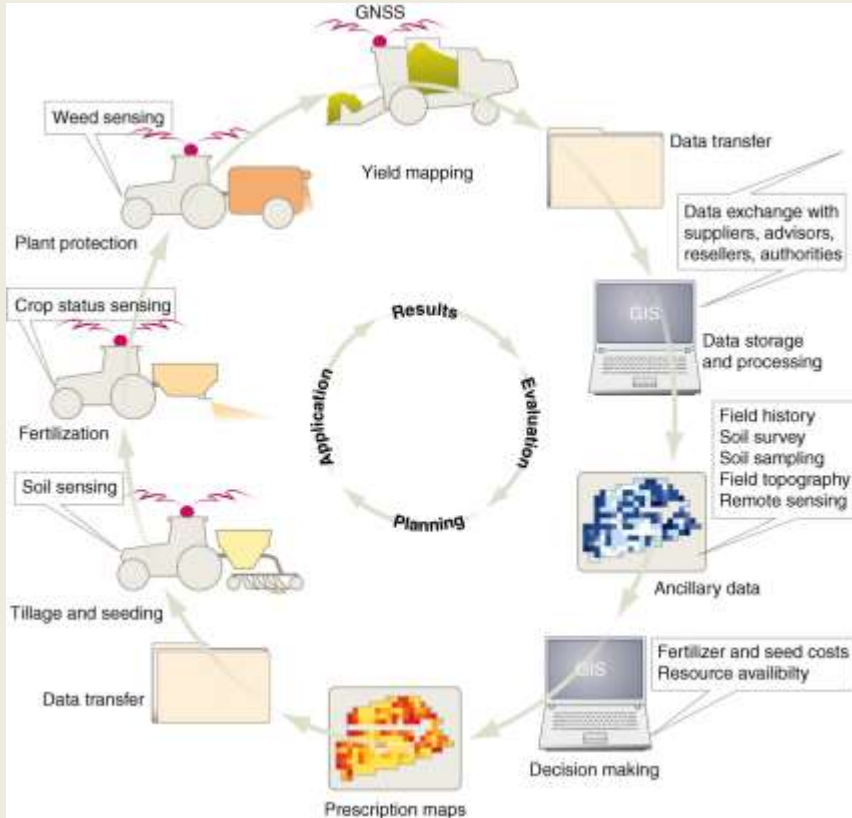


Fig. 2 Precision agriculture information flow in crop production [after (19), modified].

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



Published by AAAS

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)

THE DATA SUPPLY CHAIN IN AGRICULTURE

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.agry.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.*[#]

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 measurable 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.

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

FAIR EVERYWHERE

European Commission Press Release Database

European Commission > Press releases database > Press Release details

Latest updates Related links Contact Search

Other available languages: none

Back to the search results

European Commission - Statement

G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

1. We, the Leaders of the G20, met in Hangzhou, China on 4-5 September 2016.
2. We met at a time when the global economic recovery is progressing, resilience is improved in some economies are emerging, but growth is still weaker than desirable. Downside risks remain due to potential volatility in the oil and commodity prices, sluggish trade and investment, and slow productivity and employment growth in some countries from geopolitical developments, increased refugee flows as well as terrorism and conflicts also complicate the global economic recovery.
3. We also met at a time of continued shifts and profound transformations in the configuration of the global economy for growth. With these transformations come challenges and uncertainties as well as opportunities. The choices determine the effectiveness of our response to the challenges of today and help to shape the world economy of the future.
4. We believe that closer partnership and joint action by G20 members will boost confidence in, faster driving to

Government of the Netherlands

News

Germany and the Netherlands call for rapid action on the European Open Science Cloud

News item | 30-05-2017 | 19:32

At the occasion of today's Competitiveness Council, Germany and the Netherlands made clear that it is important to boost the development of the European Open Science Cloud (EOSC) and capitalise on the momentum of the digital era. 'Time for action is now,' say State Secretaries Georg Schütte (Germany) and Sander Dekker (the Netherlands) in their position paper on the EOSC that was presented during today's Council meeting in Brussels.

Research data should not be stored away on personal computers or USB-sticks, nor in research infrastructures only researchers themselves know how to use. Making data easily accessible for other researchers will enhance scientific progress. Making them accessible for a broader audience such as citizens and entrepreneurs will greatly boost the impact and utilisation of science. The sooner we are able to make this happen, the better.

GO FAIR

Schütte and Dekker propose to support the GO FAIR initiative, as a promising approach towards establishing the EOSC. GO FAIR is completely open-to-all and can contribute to a broad involvement of the European science community as a whole. They called on other Member States to join the movement and urged the European Commission to strengthen its efforts through brooding

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

Common Fund Programs Common Fund Research Funding News & Media Common Fund High

Big Data to Knowledge

Common Fund > Common Fund Programs > Big Data to Knowledge

Now accepting applications for the 2017 Data Science Road-Trip®

Program Snapshot

As biomedical tools and technologies rapidly improve, researchers are producing and analyzing a rapidly increasing amount of complex biological data called "big data." The Big Data to Knowledge (BD2K) program, was launched in 2014 to facilitate broad use of biomedical big data, develop and disseminate analysis methods and software, enhance training relevant for large-scale data analysis, and establish centers of excellence for biomedical big data. The BD2K Program also supported initial efforts toward making data sets "FAIR": Findable, Accessible, Interoperable, and Reusable. Learn more about the FAIR principles.

NIH Data Commons Pilot Phase Explores Using the Cloud to Access and Share FAIR Biomedical Big Data

The NIH, under the BD2K program, will be launching a Data Commons Pilot Phase to test ways to store, access and share FAIR biomedical data and associated tools in the cloud.

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



The screenshot shows the FORCE11 website header with the logo and tagline "The Future of Research Communications and e-Scholarship". A navigation bar includes links for ABOUT, COMMUNITY, GROUPS, RESOURCES, NEWS + BLOGS, EVENTS, PUBLICATIONS, MEDIA, and DONATE. The main content area features the title "JOINT DECLARATION OF DATA CITATION PRINCIPLES - FINAL" and a paragraph stating: "When citing please use: Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [datacitation]." Below this is an "ENDORSEMENT LIST" section with a "PREAMBLE" that begins: "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". To the right of the text is a large blue "DC¹" logo with "Data Citation Principles" written below it.

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](#)

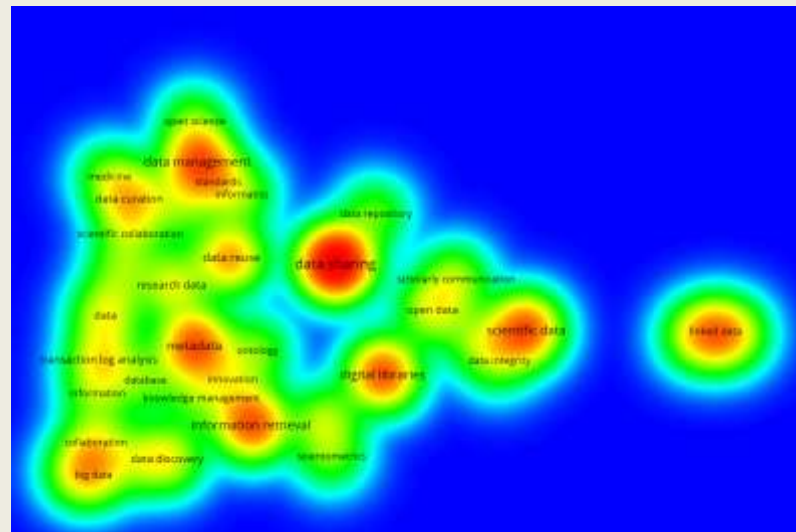


Published in:
Molecular Cell

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

Data Sources

- ☐ ScienceDirect (2357)
- ☐ arXiv (112)
- ☐ Dryad (56)
- ☐ Harvard Dataverse (30)
- ☐ Zenodo (28)
- ☐ Apollo - Cambridge (15)
- ☐ NeuroElectro (8)
- ☐ DataSpace Princeton (5)
- ☐ Mendeley Data (4)
- ☐ PANGAEA (1)

DataSearch

frog phylogeny

29129 results for *frog phylogeny*

Filter Results

reset

Types

Image (24252)

Tabular Data (12621)

Document (2994)


Raw Data (595)

File Set (307)

Slides (117)

Video (63)


Statistical Data (16)

 Molecular phylogeny of Malagasy reed frogs, *Heterixalus*, and the relative performance of bioacoustics and color-patterns for resolving their systematics

Katharina C. Wollenberg, Frank Glaw, Axel Meyer & Miguel Vences - 2007-06-22

The members of the genus *Heterixalus* constitute one of the endemic frog radiations in Madagascar. Here we present a complete species-level phylogeny based on DNA sequences (4876 base pairs) of three nuclear and four mitochondrial markers to clarify the phylogenetic relationships among...

DOCUMENT IMAGE 3 TABULAR DATA 3

 Data from: Phylogeny of frogs of the *Physalaemus pustulosus* species group, with an examination of data incongruence


Cannatella, David C., Hillis, David M., Chippindale, Paul T., Weight, Lee, Rand, A. Stanley & Ryan, Michael J. - 1998-06-01


Characters derived from advertisement calls, morphology, allozymes, and the sequences of the small subunit of the mitochondrial ribosomal gene (12S) and the COI mitochondrial gene were used to estimate the phylogeny of frogs of the *Physalaemus pustulosus* group (Leptodactylidae)....


RAW DATA

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

 Google Search

 Search

ALL PRODUCTS 

HOME GUIDES REFERENCE TOOLS SUPPORT PARTNERS SEND FEEDBACK

Social profile links
Carousels

Content Types
Articles
Books
Courses
Datasets
Events
Fact Check
Job Postings
Local Businesses
Music
Podcasts
Products
Recipes
Reviews
TV and Movies
Videos


Beta Features

Datasets

☆☆☆☆☆

The web contains specialized repositories for datasets in many scientific domains: life sciences, earth sciences, material sciences, and more. Similarly, many governments maintain repositories of civic and government data. However, much of that structured data is not readily available to search engines, which must extract the data from HTML pages in order to provide search services to users. When webmasters provide [structured markup](#), they enable search engines to “understand” this metadata, which in turn improves data discovery, leading scientists to the information they need for their work.

For example, consider this dataset that describes [historical snow levels in the Northern Hemisphere](#). This page contains basic information about the data, like spatial coverage and units. Other pages on the site contain additional metadata: who produces the dataset, how to download it, and the license for using the data. With structured data markup, these pages can be more easily discovered by other scientists searching for climate data in that subject area.

 Dataset markup is available for you to experiment with before it's released to general availability. When you implement the markup, you'll be able to test it in the Structured Data Testing Tool. You won't, however, see your datasets appear in Search.

Help Google improve dataset discovery efforts by filling out our interest form. This is not a

Contents

What qualifies as a dataset?

Mark up your dataset descriptions

Basic dataset properties

Data catalog properties

Download information properties

Temporal coverage

Spatial coverage

Citations and publications

Provenance and license information

Site-wide structure: sitemaps and sameAs

Complete Example: Datasets markup in JSON-LD

Source and

ELSEVIER

```

<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 an",
"headline": "Single-molecule analysis of mtDNA replication uncovers the basis of the common deletio",
"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

First priority: ISOBUS

What is ISOBUS?

Why ISOBUS?

AEF Functionalities

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

Crop Ontology Curation Tool

Home About Feedback

CO_020 Add New Term

Traits, methods and scales

DOWNLOAD Undefined

- multicrop passport ontology
 - MultiCropPassportDescriptorCode
 - BiologicalStatusOfAccessionCode CO_020:0000001
 - 100 wild
 - 200 weedy
 - 300 traditional cultivar or landrace
 - 400 breeding or research material
 - 500 advanced or improved cultivar
 - 600 GMO
 - 899 other biological status of accession
 - CollectingOrAcquisitionSourceCode CO_020:0000001
 - CountryOfOriginCode CO_020:0000001
 - InstituteCode CO_020:0000001
 - TypeOfGermplasmStorageCode CO_020:0000001
 - multicrop passport descriptor

AGROVOC

Content language English

Alphabetical Hierarchy

A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z 0-9

Aptosyax grypus
Aaron's rod → *Verbascum*
ABA
Abaca
abachi → *Triplochiton scleroxylon*
Abalistes stellaris
abalones
abamectin
abandoned land
abattoir byproducts
abattoirs
Abbottina rivularis
abdomen
abdominal cavity
abdominal fat
abdominal pregnancy
Abelmoschus
Abelmoschus esculentus
Abelmoschus moschatus
Aberia → *Dovyalis*
Abies
Abies alba
Abies amabilis
Abies balsamea
Abies balsamea lasiocarpa → *Abies lasiocarpa*
Abies borisii regis

Vocabulary information

TITLE	AGROVOC
LAST MODIFIED	Friday, September 1, 2017 08:41:29
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
VOID:INDATASET	http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc
URI	http://voc.landportal.info/landterms

Resource counts by type

Type	Count
Concept	33440

Term counts by language

Language	Preferred terms	Alternate terms	Hidden terms
Arabic	24574	1067	0
Czech	32119	8570	0
German	32095	10131	0
English	33156	9147	0

INTEGRATION

Home ▶ Semantics ▶ AGROVOC ▶

AGROVOC Linked Data

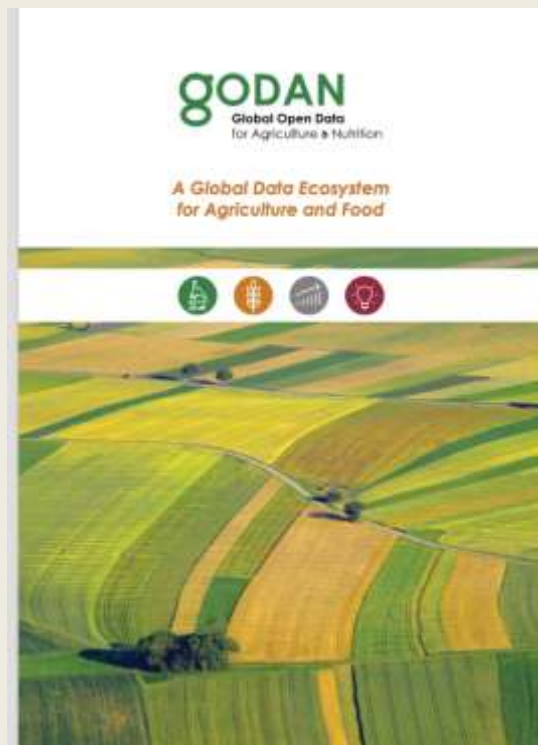
[ABOUT](#) | [SEARCH](#) | [ACCESS](#) | [COMMUNITY](#) | [LINKED DATA](#) | [PUBLICATIONS](#) | [GU](#)

[Linked Data](#) is a method of web publication in which each individual piece of data is:

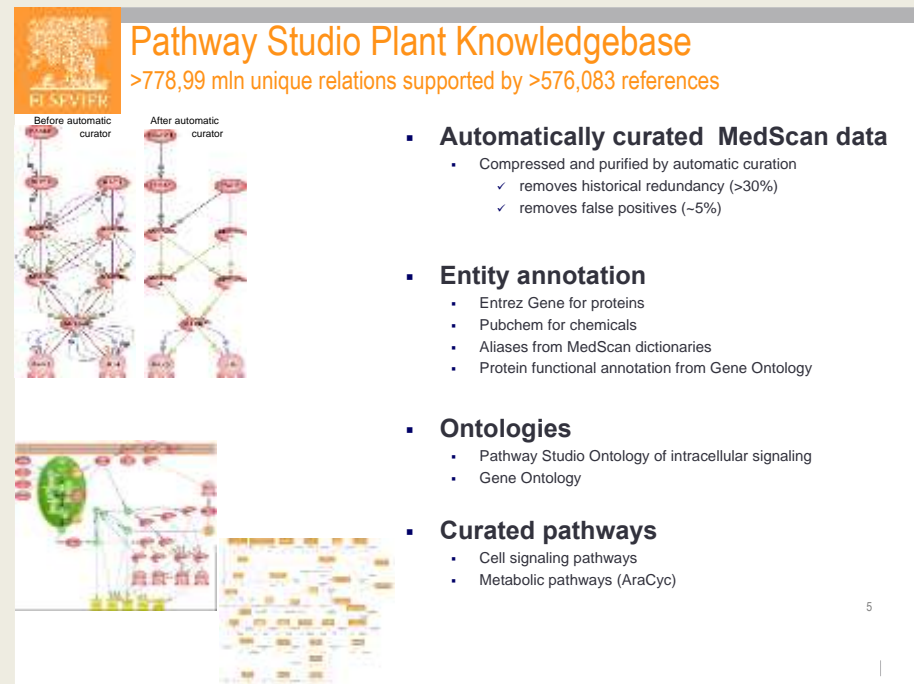
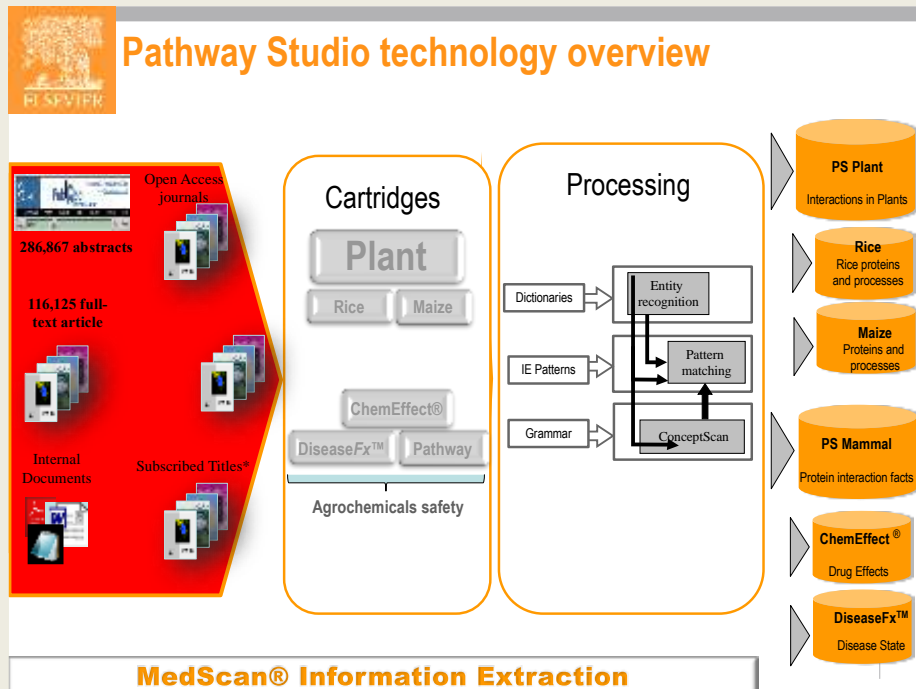
- uniquely identified using [HTTP URIs](#) (that is, URLs, or 'web addresses'),
- available both as 'machine readable' data and as 'human readable pages, and
- linked to other resources.

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](#).


	Ry2 (ziarno) 稻米
skos:inScheme	http://aims.fao.org/aos/agrovoc http://www.elonet.europa.eu/gemet/concept/7214 http://eurovoc.europa.eu/3732 http://d-nb.info/gnd/4049271-0
skos:exactMatch	http://cat.all.caas.cn/concept/c_8549 http://cat.all.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/bnct/tid/17341 http://purl.org/bnct/tid/38716 http://dbpedia.org/resource/Rice
void:inDataset	http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc



INTEGRATION ACROSS DOMAINS



DATA SUSTAINABILITY



Wikidata

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

English Not logged in Talk Contributions Create account

Item Discussion Read View history Search Wikidata

rice (Q5090)



cereal grain and seed of *Oryza sativa* [edit](#)

[+ In more languages](#) [Configure](#)

Language	Label	Description	Also known as
English	rice	cereal grain and seed of <i>Oryza sativa</i>	
Dutch	rijst	graan (<i>Oryza sativa</i>) van de grassenfamilie Poaceae	
German	Reis	Getreideart, Grundnahrungsmittel	
French	riz	céréale de la famille des poacées	

[All entered languages](#)

Statements

subclass of	 cereal edit
	+ 0 references + add reference
 staple food edit	

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

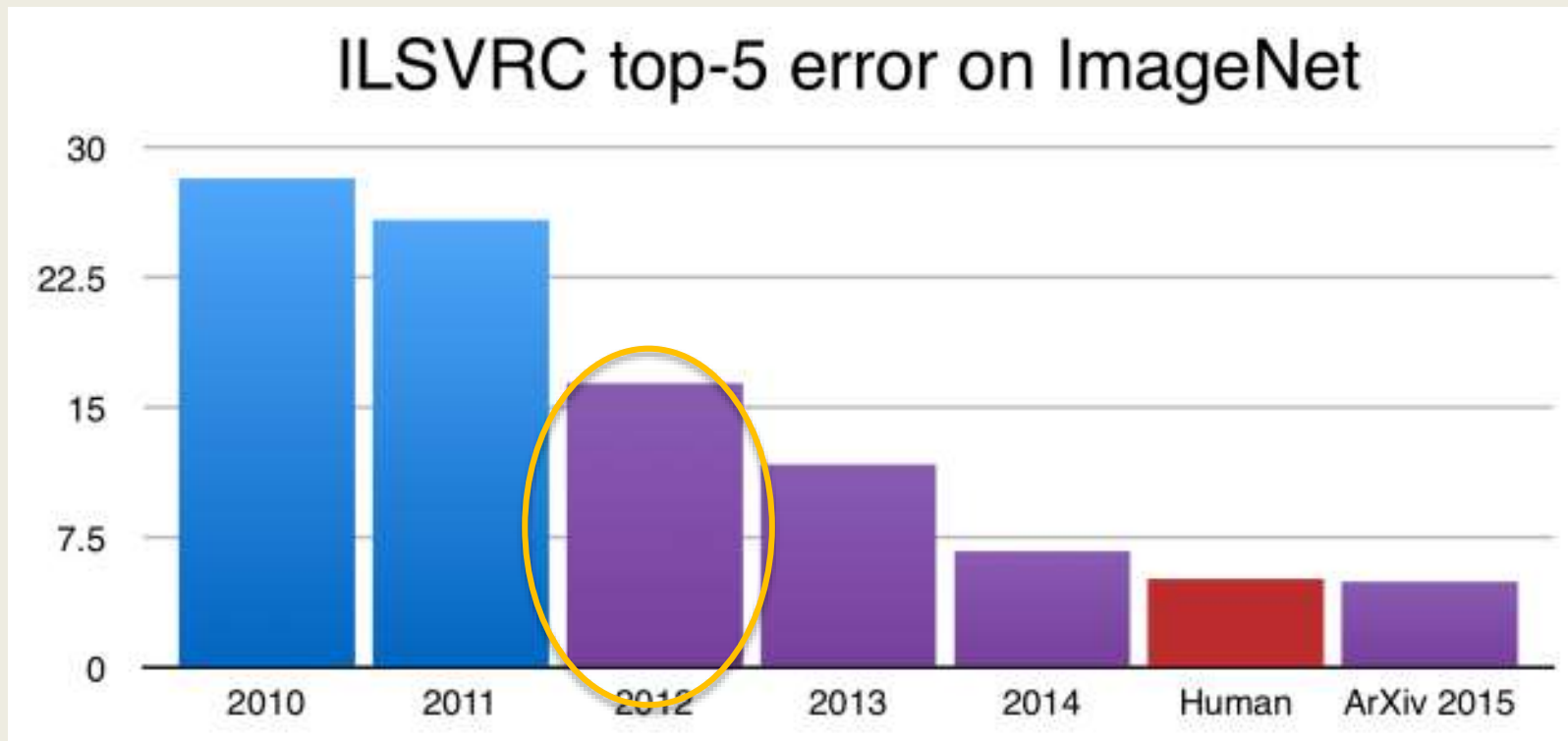
WHAT CAN MACHINE INTELLIGENCE DO TODAY?



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.

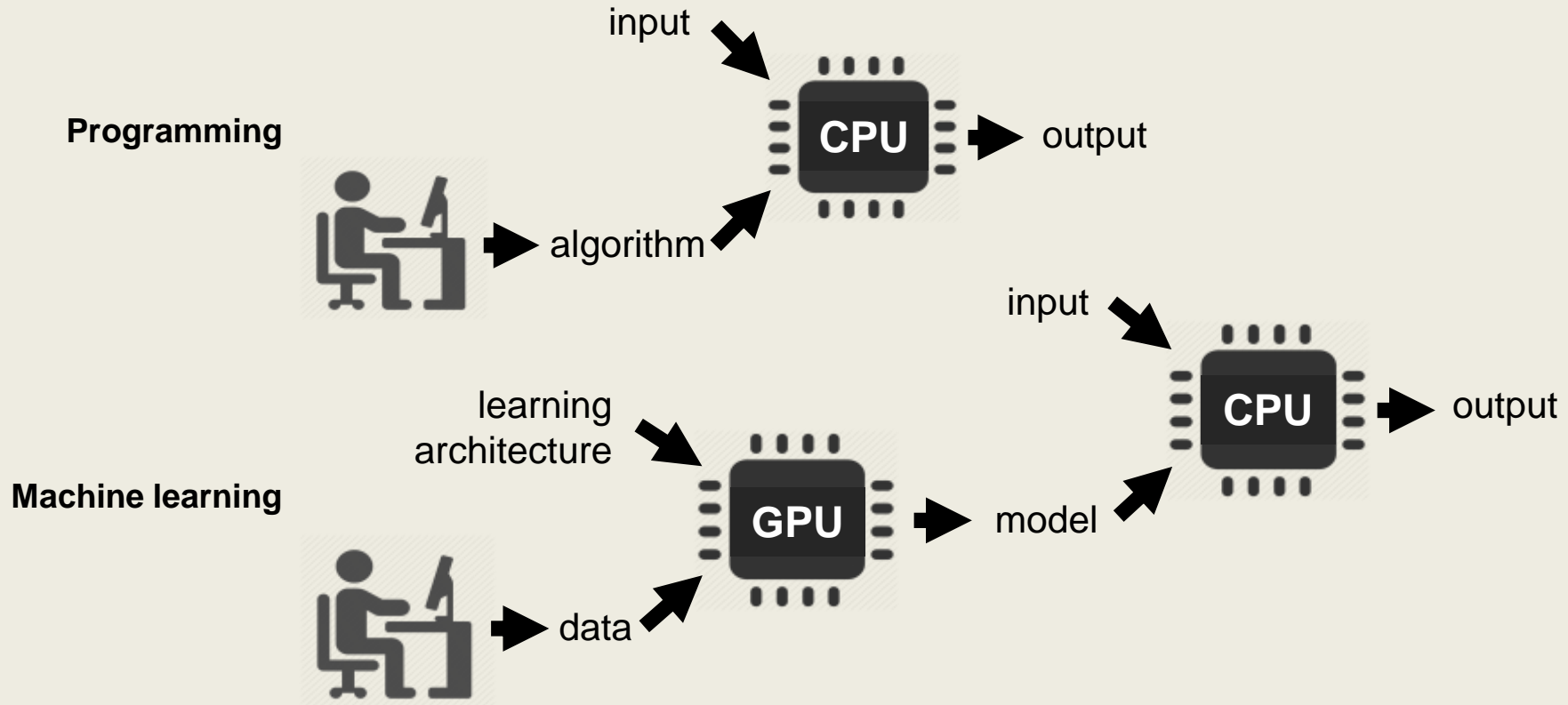
Andrew Ng, Chief Scientist, Baidu (lecture at Bay Area Deep Learning School, Stanford, CA, September 24, 2016)

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

```
01a9c 32 3d 4b 70 87 5b ef 53 e1 38 ea 40 2a 5e d2 79 df d2 0e 21 cf 88 ba
01acb 38 7a cf 3e db 7d 31 8d 99 88 04 e1 8d 1c 2d 6d 38 22 37 70 4a 8d bf
01afa 8f cd 2e 1d 8a 9f bc 3f 50 ef 47 85 4e 84 2d c6 09 79 52 4a 77 22 07
01b29 49 b5 34 b2 28 53 e0 97 06 e4 ee 22 3d fd b1 e9 f8 72 b0 e2 ee 8c
01b58 13 8e 6f 5f 73 21 0d 7f 8a d8 17 14 6d 25 5e 7a 91 72 6c 59 d9 ba 69
01b87 e3 23 3a ac ea a6 a0 55 d2 7c 4d da 3c cb 71 63 58 e2 26 49 3f 94 63
01bb6 27 ca 9a 74 21 64 a7 68 09 9d c9 fa 1f 8e 38 5d 77 05 90 63 ce f3 f5
01be5 54 7f 48 38 e6 30 5a d7 39 ad 6f 52 79 5d 04 d3 be 3c 27 16 f5 a5 52
01c14 27 b0 05 b2 3e f8 f4 a8 08 c0 cb 82 31 d1 e4 ee bf a7 65 c8 e3 63 0c
01c43 a8 cb 74 4d 78 31 85 c9 c1 8d 34 7a 93 a2 af 4f 28 d1 3f 87 1a 52 c6
01c72 80 f8 47 1d d7 a5 e8 b1 b9 b0 ed be 13 01 96 a8 fa 65 9b ae 75 cf b4
01ca1 20 c9 8b d3 3b ce 6b 5e 63 c8 f7 65 22 0f 42 5a 44 84 90 21 49 dc 1e
01ed0 1a 9b 5d bd a3 69 a9 65 87 c2 54 15 a2 24 09 de 67 d7 db 91 38 bf 9e
01eff cb e8 43 5e 2d 59 d6 da 76 48 2a 52 47 1d 80 27 0d 7e 80 3f d3 da d7
01d2e 09 fd fa 6c 4d 78 44 27 85 e9 00 c7 e4 71 c7 f8 2f 16 4c dd 4b 22 ba
01d5d cb 4c a8 3e 52 be 55 ce de bb e3 d4 f0 80 43 6e 27 f4 0b 87 d5 32 24
01d8c 51 9f b9 02 7d b1 d3 45 83 17 95 bd 70 8f cb 91 d3 9a 3d 57 a0 f2 a6
01dbb 63 8e d5 1f 1c 99 18 01 5d 96 81 2c 98 63 cc 0b 09 ea 46 62 ae 46 7a
01dea af 8c 35 19 4e a8 25 8c f6 0a 53 e0 6d 3d 49 b4 37 5f 67 a8 02 b6 dc
01e19 99 80 fd a5 ee de 8a e4 24 14 78 d3 d1 25 2c a4 13 c1 29 d3 09 3e d3
01e48 56 cc ea aa 57 9e 0d 8a 67 11 ad 71 04 05 7a 8f 4f f8 b1 df 66 e3 9c
```

(a) hex dump of picture of a lion

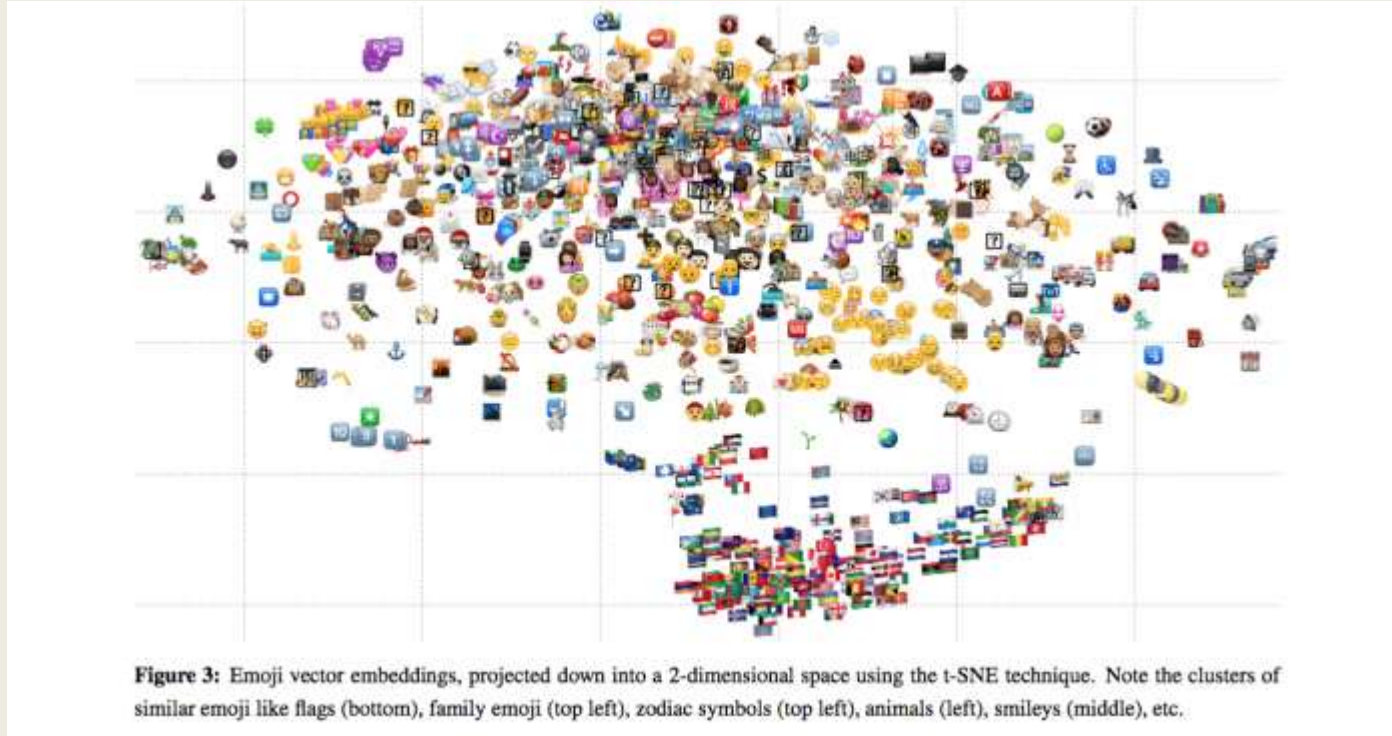


(b) same lion in human-readable format

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.

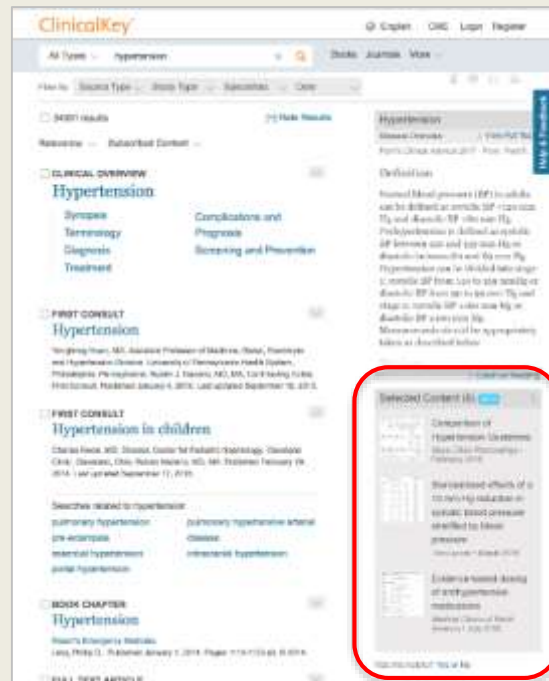
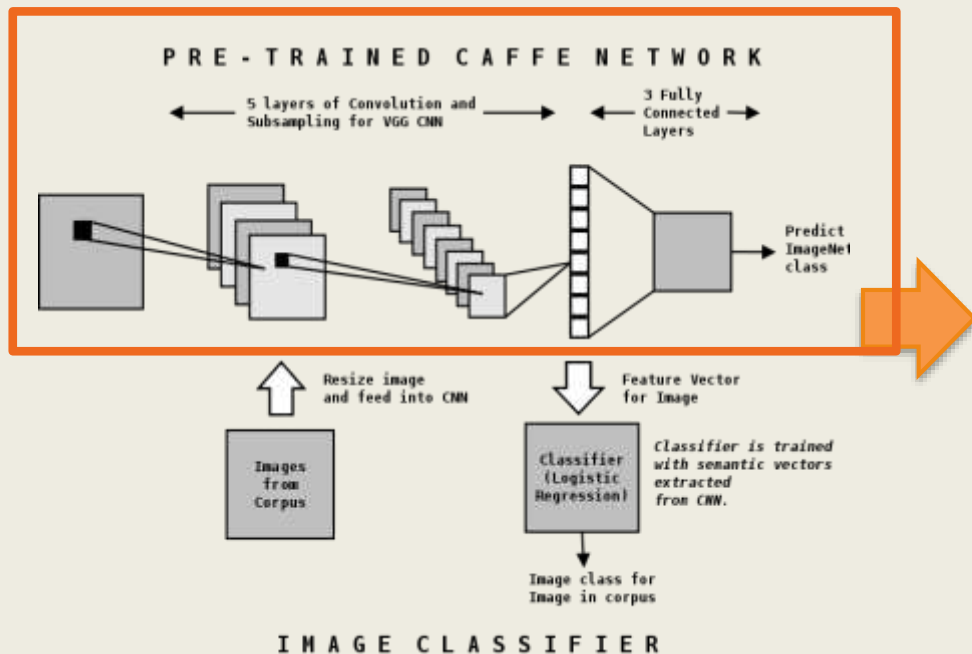
From: Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv:1610.01644v1.

VOCABULARIES ARE SETS OF VECTOR EMBEDDINGS



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



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

<http://nca2009.globalchange.gov/southeast>

Southeast

The climate of the Southeast is uniquely warm and wet, with mild winters and high humidity, compared with the rest of the continental United States. The average annual temperature of the Southeast did not change significantly over the past century as a whole. Since 1970, however, annual average temperature has risen about 2.0° F, with the greatest seasonal increase in temperatures occurring during the winter months. The number of freezing days in the Southeast has declined by four to seven days per year for most of the region since the mid-1970s.

Table of Contents [Back]

1. Water Temperature
2. Water Movement
3. Sea Area Size and Location
4. Windy Area
5. Safety
 - Adaptability: Reducing Exposure to Flooding and Storm Surge
6. Disasters

Average autumn precipitation has increased by 32 percent for the region since 1801. The decline in fall precipitation in North Florida contrasts strongly with the regional average. There has been an increase in heavy downpours in many parts of the region, 2.2 times the percentage of the heaviest superheavy moderate to severe droughts (classified under the fall from decades). The area of moderate to severe spring and summer drought has increased by 32 percent and 24 percent, respectively, since the last 100 years. Even in the fall months, when precipitation tended to increase in most of the region, the extent of drought increased by 8 percent.

[illegible]

	Temperature Change in °F		Precipitation Change in %	
	1801-1900	1975-2000	1901-2000	1977-2000
Annual	0.3	1.6	6.5	-7.7
Winter	0.2	2.7	1.2	-8.8
Spring	0.8	1.2	1.7	-29.2
Summer	0.6	1.6	-4.8	3.6
Fall	0.2	1.1	27.8	9.1

ELSEVIER

http://globalchange.gov/dataset/USHCN_002

The U.S. Historical Climatology Network (USHCN) Version 2
Serial Monthly Dataset

[illegible]

<http://globalchange.gov/agency/NOAA>



<http://globalchange.gov/datacenter/NCDC>

<http://globalchange.gov/paper/doi/10.1175/2008BAMS2613.1>

Menne, M.J., C.N. Williams Jr., and R.S. Vose, 2009: The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. *Bulletin of the American Meteorological Society*, 90, 993-1107.

Journal: Bulletin of the AMS

<http://globalchange.gov/journal/BAMS>

Meene, M.J.

<http://globalchange.gov/person/235>

C.N. Williams Jr.

<http://globalchange.gov/person/587>

R.S. Yose

<http://globalchange.gov/person/372>

http://globalchange.gov/software/USHCN_V52d.20100217

Pairwise Homogeneity Adjustment Software



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