



Unlocking Data from the Web

<http://training.theodi.org/UnlockingData>

David Tarrant · @davetaz

Open Data Science

Day 1: Discovering Data Science

Day 2: Statistics and Machine Learning

Day 3: Big data and interactive infographics



Introductions



Your name

What interests you most about data science?

What do you want to do differently after the course?

Course aim

Equip you with the knowledge and tools
to help you upskill as modern data
scientists.

Session 1

Discovering Data Science

Session 2

Gathering and preparing data

Session 3

Publishing insight

Session 1

Discovering Open Data Science



Outcomes

Define open data science

Describe a number of key data science stories

Identify the characteristics in open data science projects





Prof. Dr. Frank Bensberg

Osnabrück University of Applied Science
Faculty of Business Management &
Social Sciences
Capriviustrasse 30a
D-49076 Osnabrück
F.Bensberg@hs-osnabrueck.de

List the 4 aspects of

Scientist

Analyst

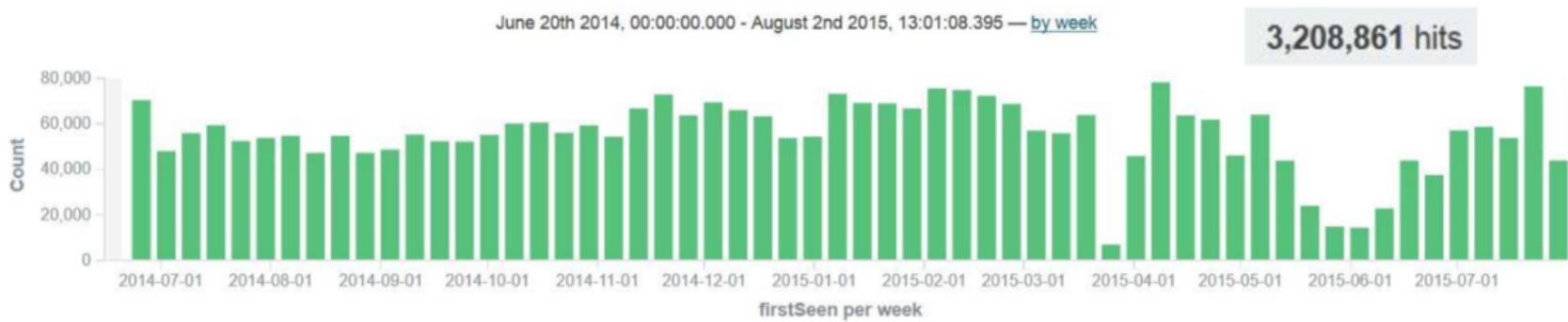
- Domains
- Concepts
- Products
- Soft skills

Engineer

Visualiser

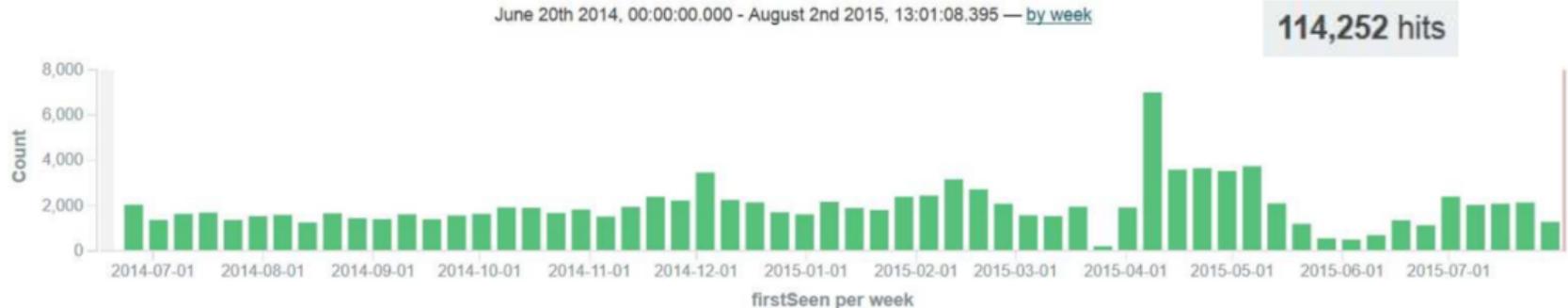
Database

Total Ads Collected



Big Data Ads

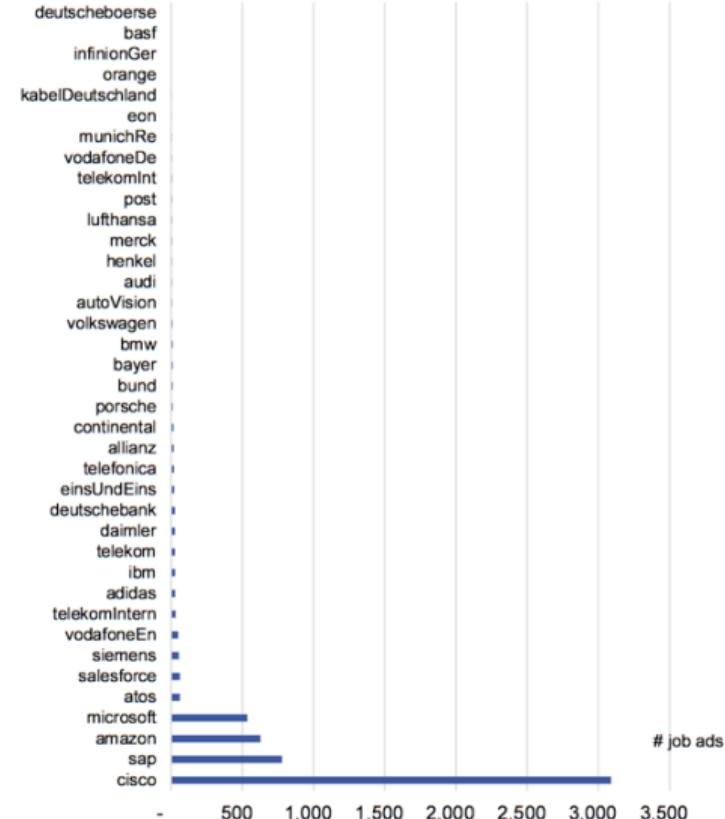
description:("big data" OR "data analytics" OR "data scien")



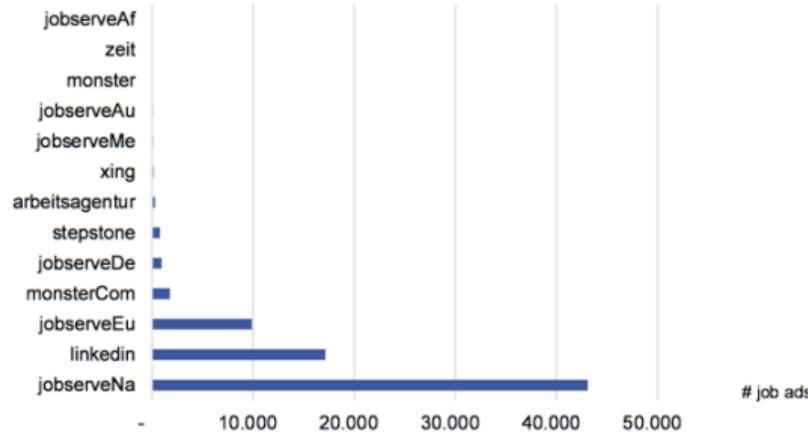
Data Set

- Basic data pool comprises about 3m of job ads for the ICT sector collected since June 2014
- Job ads are periodically collected from public job portals (jobserve, monster, ...) and from company-specific portals (Cisco, SAP, ...)
- Selection of job ads which contain at least one of the keywords *big data*, *data science*, *data scientist*, *smart data* or *fast data* (n=80.014)

Company-specific Portals



Public Job Portals



Data Scientist

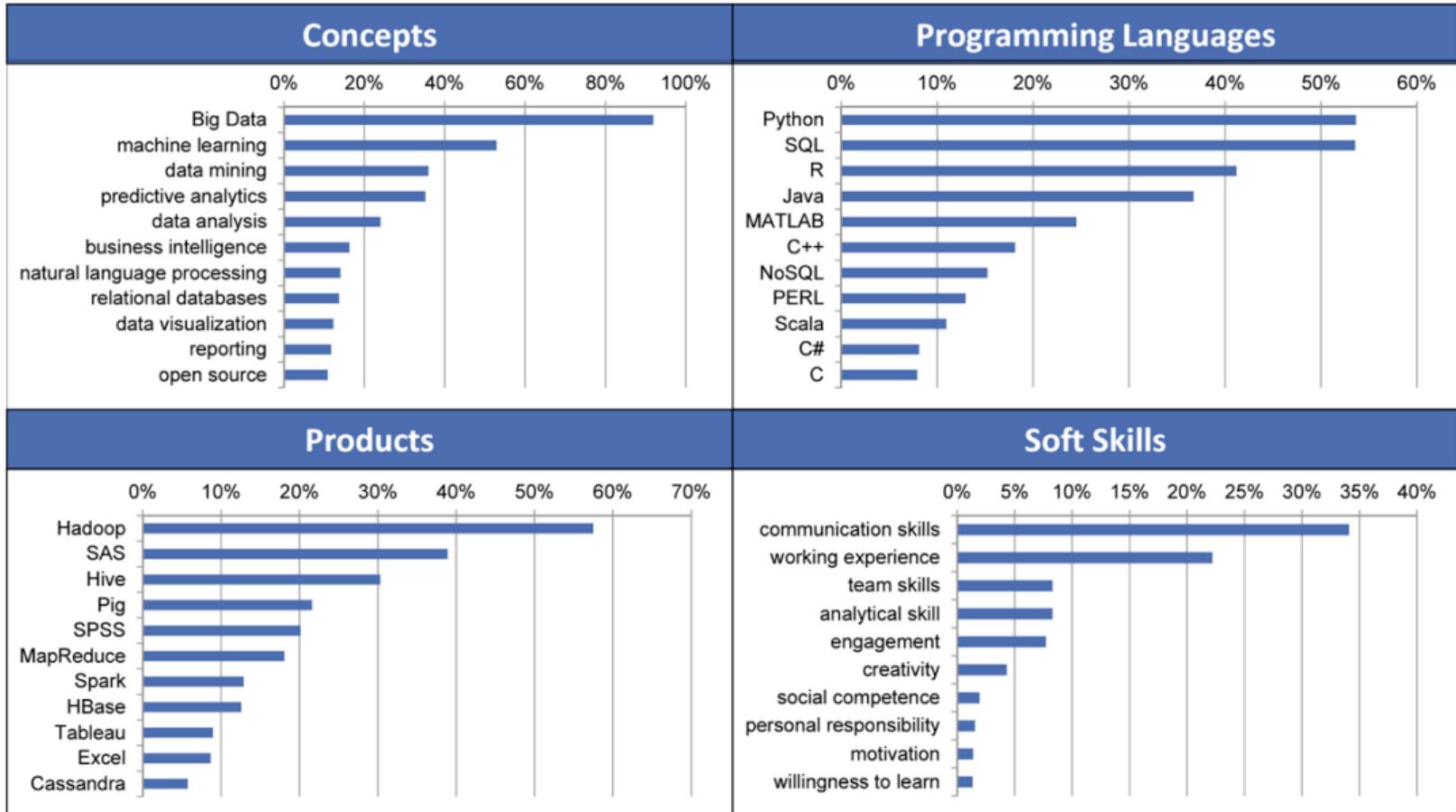
Data Set

# Job Ads	1.942	Sample Job Titles
Organizations (Top 10)	Liberty Personnel Services; Opus Recruitment Solutions Ltd; BigR.io; Career Brokers; Next Ventures Ltd; ITBM Consulting Inc;A3Logics US;Engage3;All In Analytics;Alivia technology	Data Scientist Big Data Scientist Senior Data Scientist Principal Data Scientist (R, Matlab, [...], Spark) Senior Data Scientist (Python, Matlab, R, SQL)

Professional Activities

Position	Freq	Object	Activities
1	39%	analytics	work, develop, use, include, provide, require, deliver, create, apply
2	39%	model	work, develop, provide, include, unite, use, deliver, look, analyze, create
3	37%	analysis	work, develop, use, provide, deliver, look, create, design, apply
4	35%	statistic	work, develop, use, provide, require, deliver, analyze, apply, solve
5	27%	algorithm	work, develop, use, provide, solve, apply, require, deliver, build
6	27%	machine learning	work, develop, use, apply, solve, build, design, deliver
7	24%	data set	work, develop, use, include, analyze, provide, require, design, deliver, create
8	11%	decision tree	develop, work, design, create, join, require, provide, utilize, deliver, grow
9	10%	data source	work, use, develop, require, provide, apply, help, prefer, understand
10	4%	data analysis	work, use, develop, provide, require, bring, build, drive, apply

Data Scientist

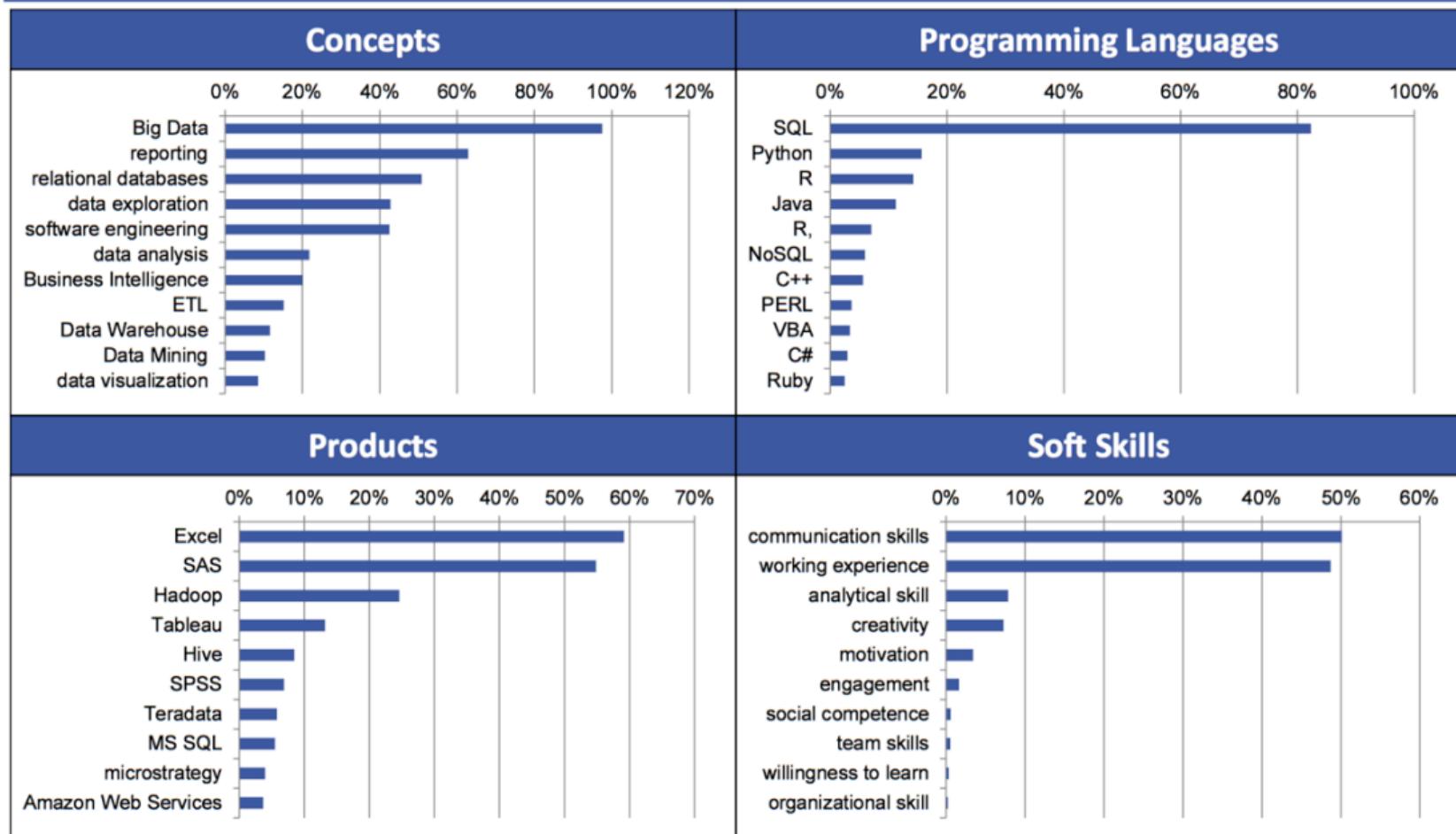


Data Analyst

Data Set			
# Job Ads	1.690	Sample Job Titles	
Organizations (Top 10)	MedeAnalytics, Liberty Personnel Services, FILD, TalentRISE, R Square, Libertyjobs.com, LS Direct Marketing, The Judge Group, UnitedHealth Group Cognius	Business Data Analyst - Health Plan Solutions Data Analyst Senior Data Analyst Big Data Analyst Data Analyst (BIG DATA - Finance Analytics)	
Professional Activities			
Position	Freq	Object	Activities
1	59%	analytics	work, develop, use, provide, require, support, deliver, solve, look, make, apply
2	55%	model	use, work, provide, develop, analyze, support, make, deliver, apply, require
3	53%	statistic	work, use, provide, develop, require, look, deliver, solve, analyze, make
4	48%	decision	make, analyze, develop, deliver, apply, require, solve, leverage, gather
5	46%	dataset	work, provide, develop, use, deliver, support, amke, apply, require, look, solve
6	44%	workflow	use, provide, analyze, develop, deliver, make, apply, gather, look, leverage
7	43%	risk	analyze, identify, reduce, look, plan, improve, change, demonstrate
8	42%	data exploration	use, develop, help, apply, provide, solve, offer, demonstrate, support, make

Annotation: data set shows a strong sectoral bias (Healthcare).

Data Analyst



Data Engineer

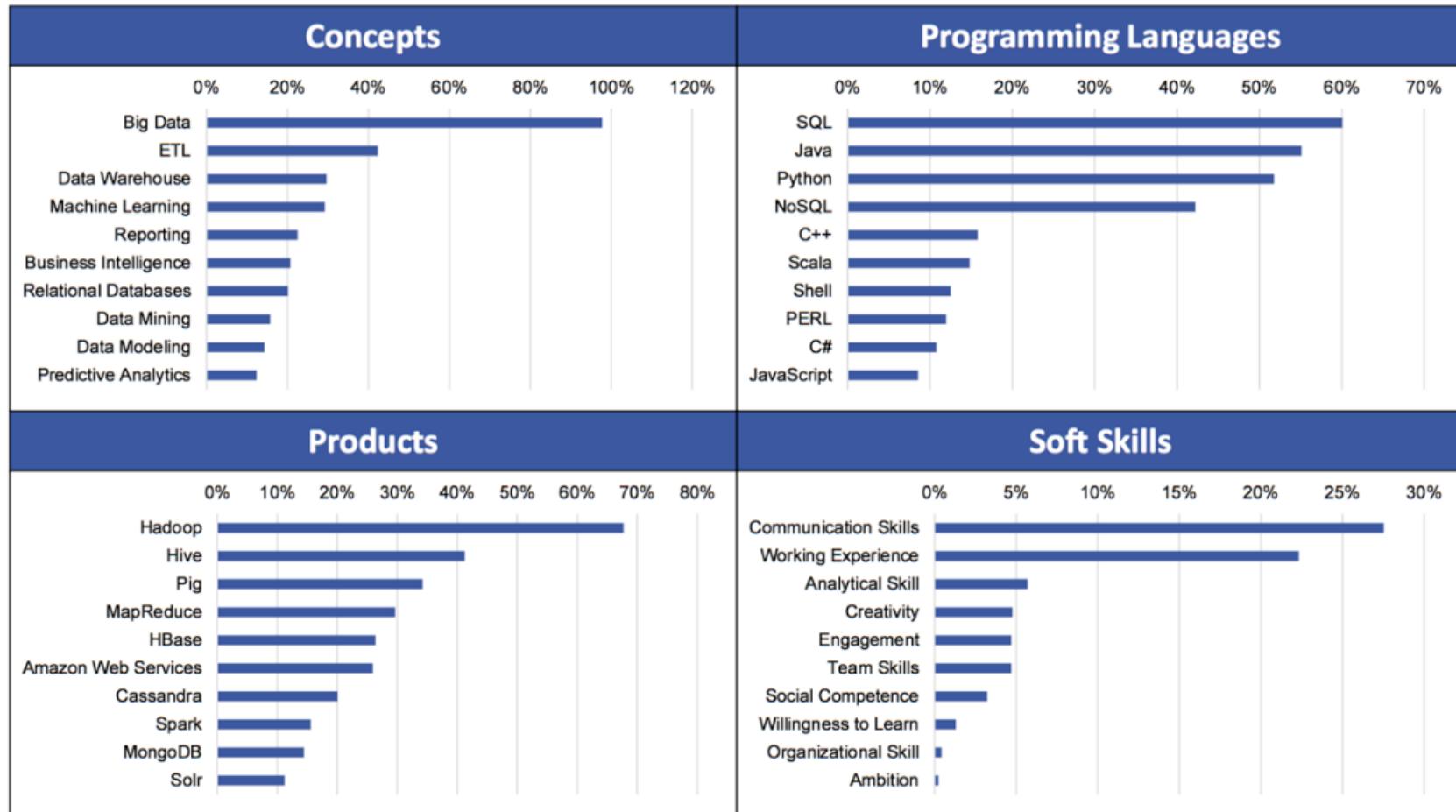
Data Set

# Job Ads	1.613*	Sample Job Titles
Organizations (Top 10)	FILD, Prosum, Liberty Personnel Services, Praedicat, Libertyjobs, Emprise Technologies, Amazon, Teradata, ColdLight, ARC IT Recruitment	Big Data Engineer Data Engineer Senior Data Engineer Senior Big Data Engineer Director of Data Engineering

Professional Activities

Position	Freq	Object	Activities
1	29%	Analytics	work, develop, include, use, design, build, provide, implement, grow
2	28%	ETL	work, develop, design, unite, provide, include, build, implement, create
3	21%	Machine Learning	work, unite, use, develop, design, build, include, grow, create, implement, provide
4	20%	Algorithm	work, use, develop, build, unite, include, create, grow, design, analyze, implement
5	16%	Data Set	work, use, develop, design, unite, include, build, process, create, grow, make, implement, analyse, provide
6	15%	Large Data	work, use, develop, unite, process, create, design, grow, build
7	13%	Unstructured Data	work, develop, include, design, use, unite, build, create, analyze, provide
8	11%	Data Platform	develop, design, work, unite, support, analyze, build, implement, grow, provide
9	11%	Predictive Analytics	work, build, develop, create, use, unite, grow

Data Engineer



Visualization Engineer

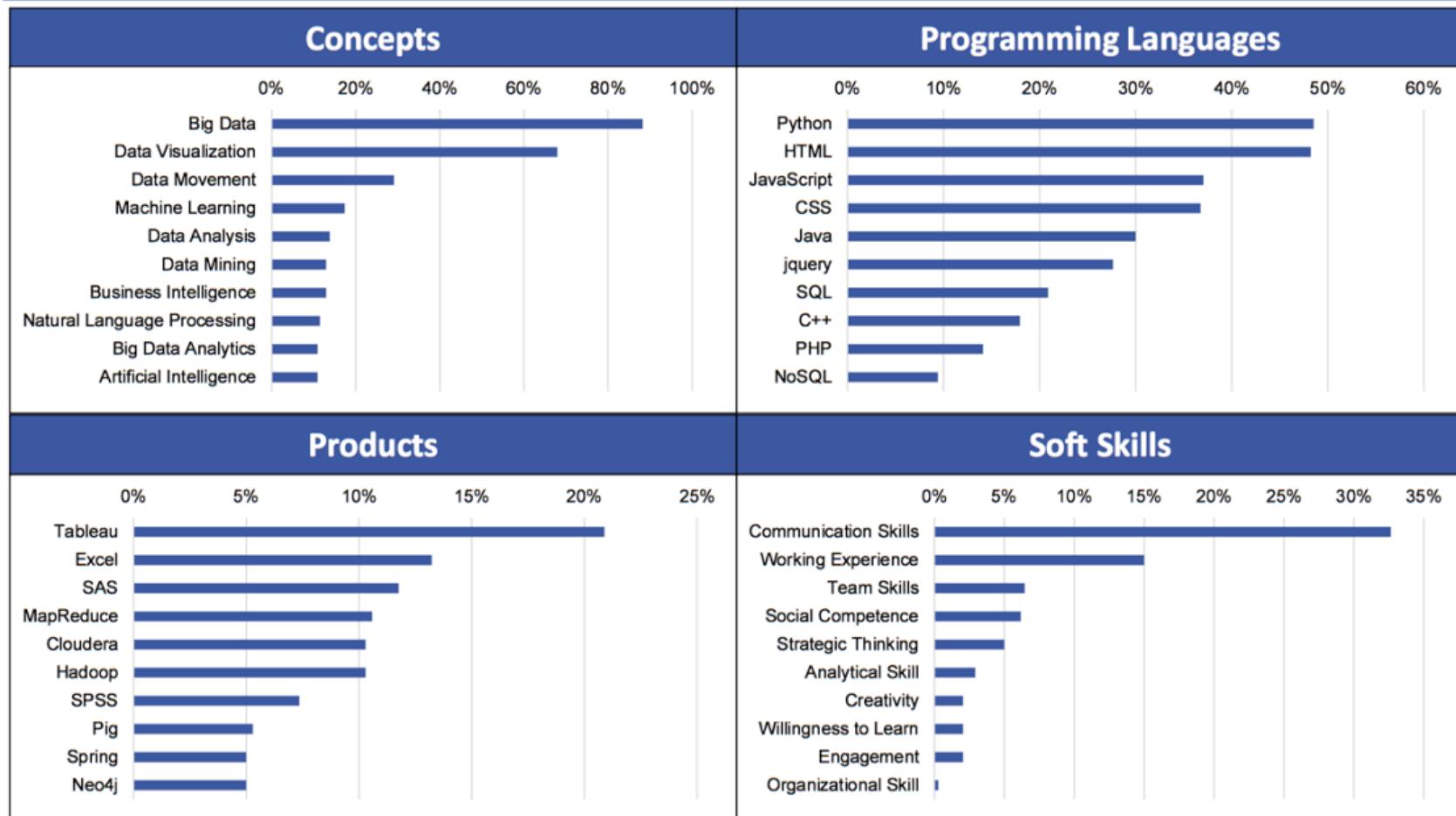
Data Set

# Job Ads	340	Sample Job Titles
Organizations (Top 10)	BigR.io, Infinite Resources, MGRS Group, FIELD, Analyze Corporation, Opus Recruitment, TASC, Splunk, N-Tier Solutions, Globys, AT&T	Data Visualization Engineer Software Engineer - Data Visualization/Graphic Design Java Visualization Software Eng. JavaScript Engineer - Python - Data Visualization - Big Data Challenge Visual & UI Designer - Adobe Creative, HTML, CSS

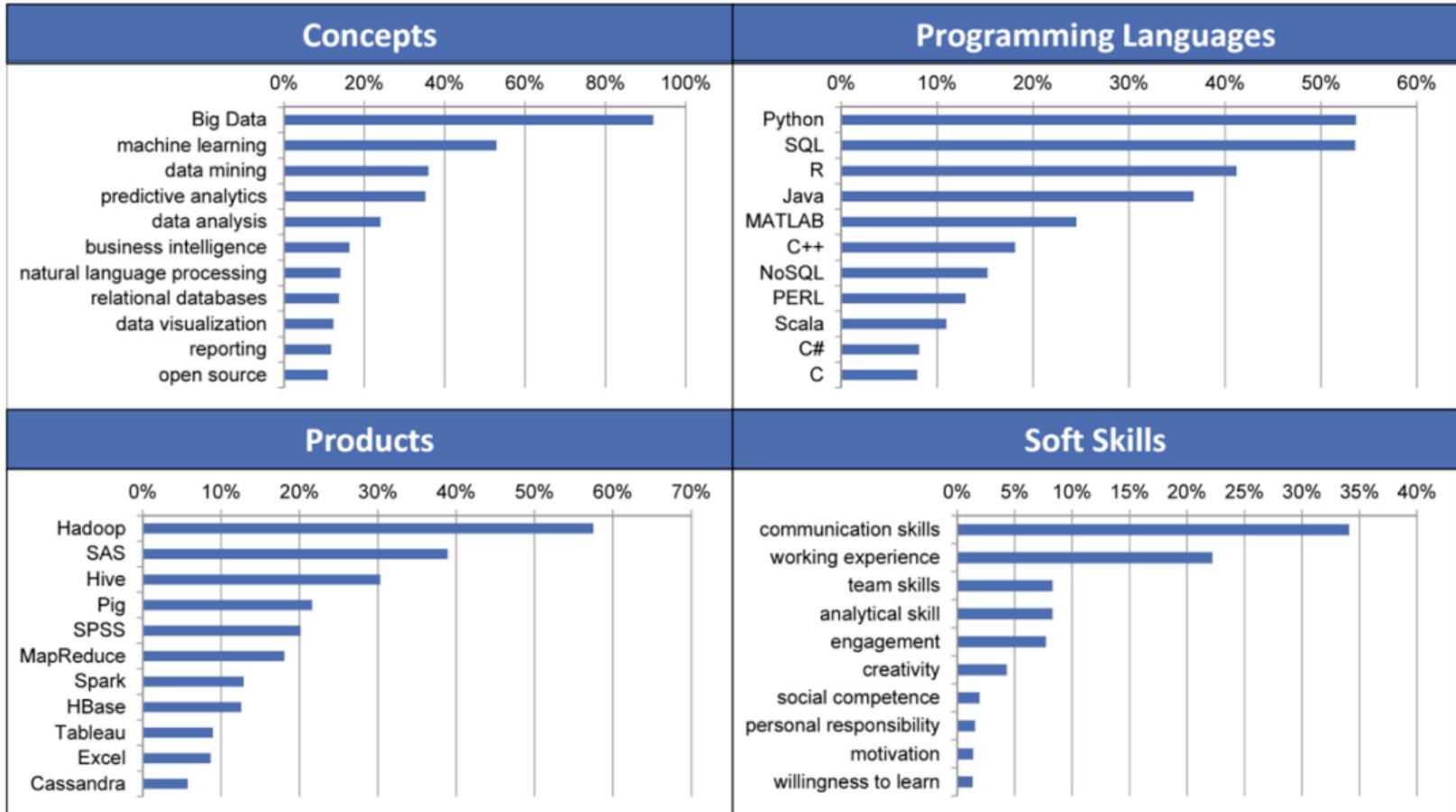
Professional Activities

Position	Freq	Object	Activities
1	44%	analytics	work, use, help, unite, develop, deliver, provide, focus, bring
2	42%	data visualization	work, use, develop, build, deliver, provide
3	31%	interface	provide, deliver, focus, use, bring, poise
4	26%	data model	use, deliver, unite, look, help, work, provide
5	22%	complex data	work, unite, deliver, provide, use
6	22%	chart	use, deliver, provide, build, bring, focus
7	21%	performance data	pinpoint, assure, dig, detail
8	15%	machine learning	develop, work, use
9	14%	interaction	work, develop, design, provide
10	9%	artificial intelligence	program, develop, design, conceptualize, implement

Visualization Engineer



Data Scientist



Key areas

- 1) Big Data
- 2) Machine Learning and Prediction
- 3) Data Collection and Analysis
- 4) Maths and Statistics
- 5) Interpretation and Visualisation
- 6) Advanced Computing and Programming
- 7) *Business Intelligence and Domain Expertise*
- 8) Open Source Tools and Concepts

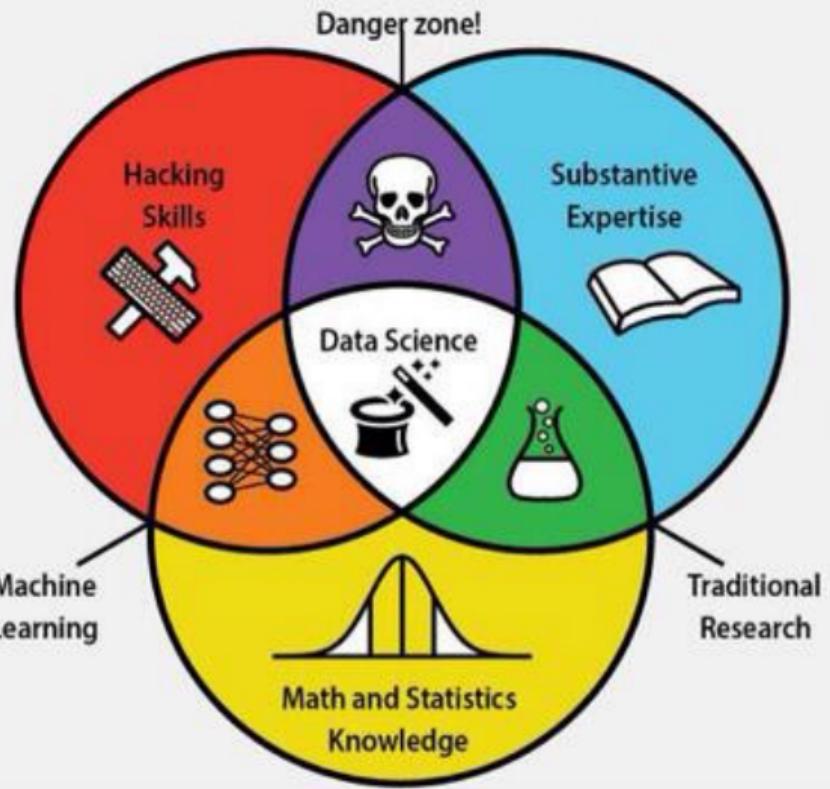


Key areas

- 1) Big Data
- 2) Machine Learning and Prediction
- 3) **Data Collection and Analysis**
- 4) Maths and Statistics
- 5) **Interpretation and Visualisation**
- 6) **Advanced Computing and Programming**
- 7) *Business Intelligence and Domain Expertise*
- 8) Open Source Tools and Concepts



DATA SCIENCE SKILLSET



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills, math and statistics knowledge**, and **substantive expertise** in a field of science.



Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

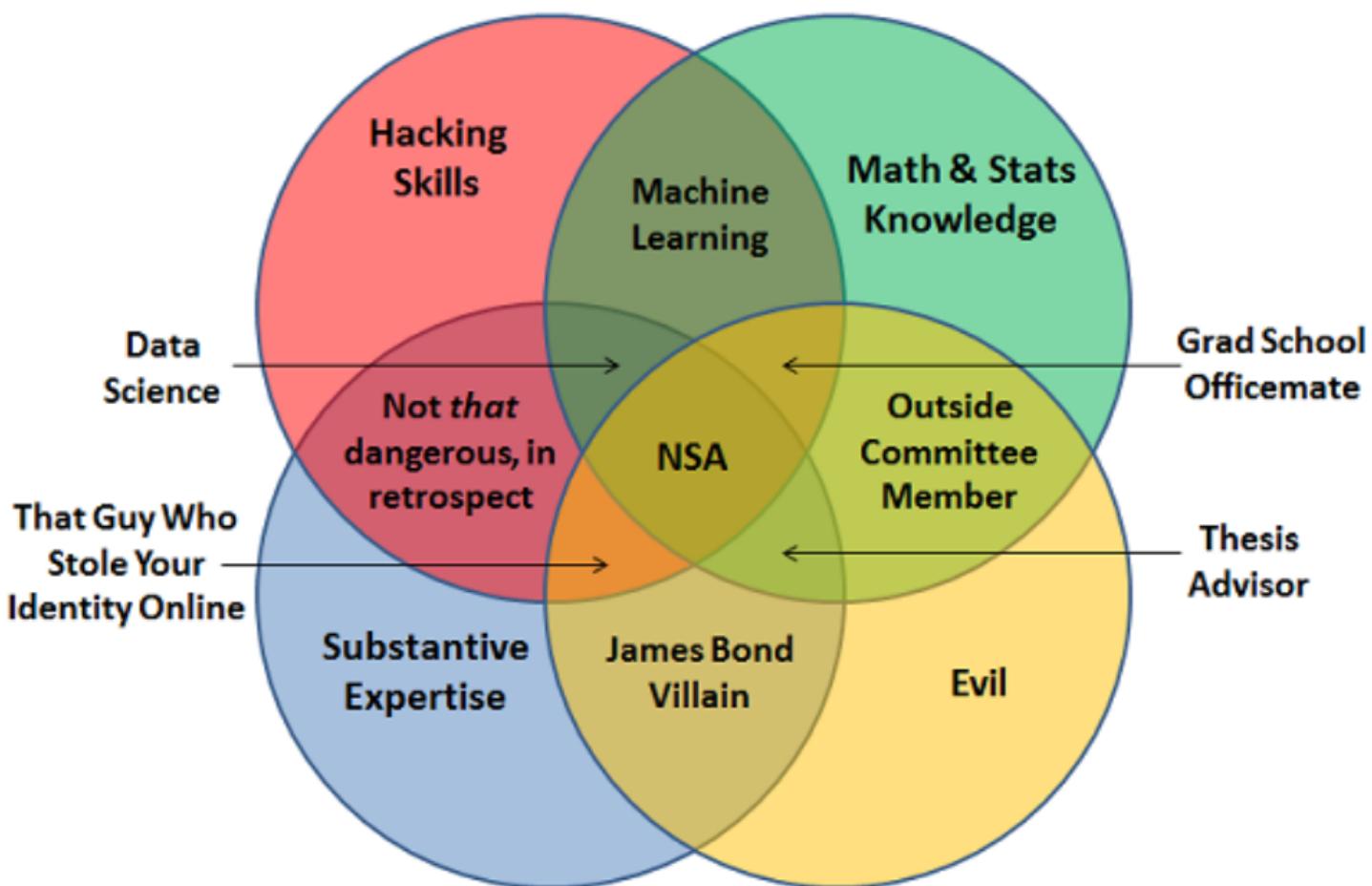


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.





Data Science

“part analyst, part artist”

-- Anjul Bhambhani (VP for Big Data, IBM)



Outcomes

Define open data science

Describe a number of key data science stories

Identify the processes in open data science projects

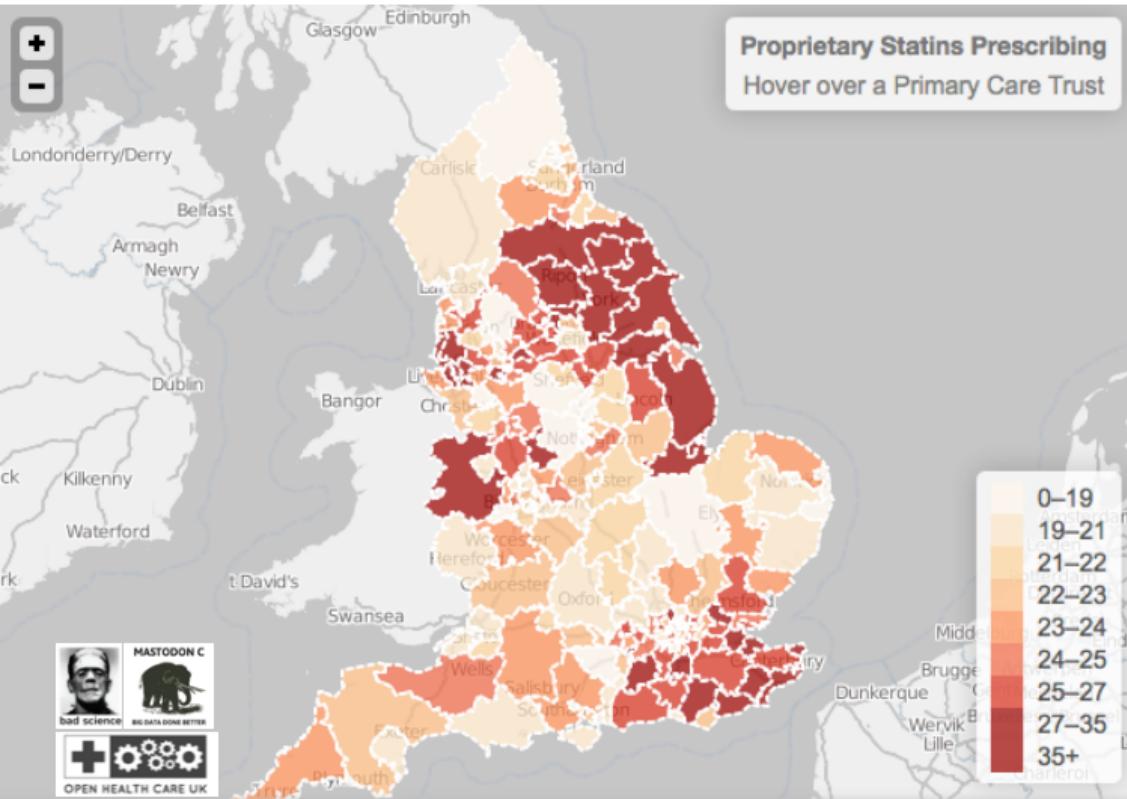


Stories that combine areas?

- 1) Big Data
- 2) Machine Learning and Prediction
- 3) Data Collection and Analysis
- 4) Maths and Statistics
- 5) Interpretation and Visualisation
- 6) Advanced Computing and Programming
- 7) Business Intelligence and Domain Expertise
- 8) Open Source Tools and Concepts



£200m potential saving identified in 6 weeks



Convened domain-experts

- + health & data analytics
- + communications

Analysed 35m records

- + all the data & clinical facts

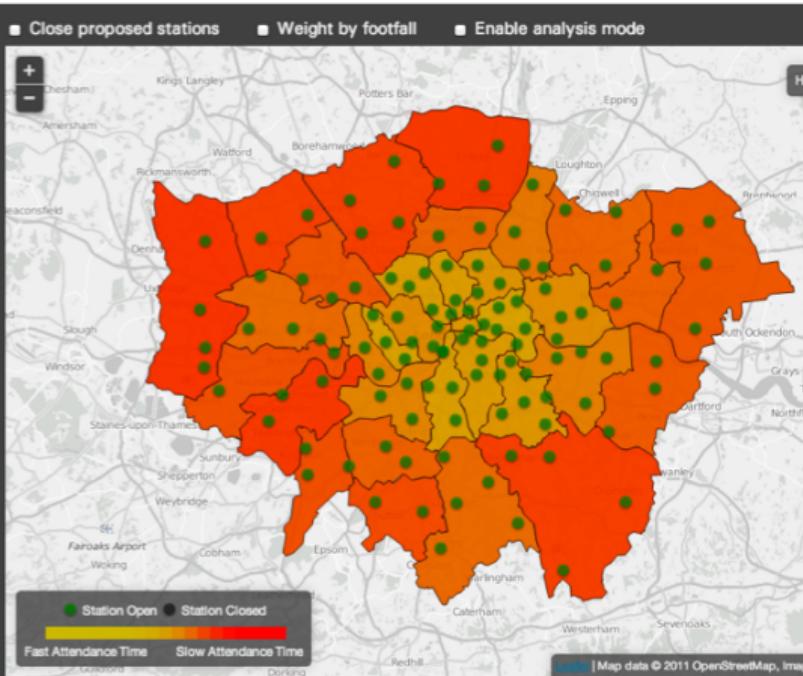
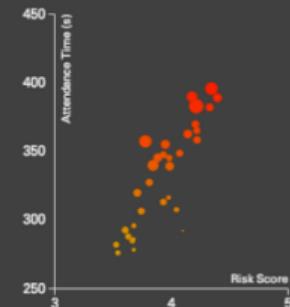
National & international reach

- + Economist & FT
- + broadsheets & tabloid press
- + cited in G8 & govt. reports

Scalable to £1bn [Bn]

Tools that aid policy makers and are open to the public

London
Average attendance time:
5m34s



<http://theodi.org/stories>

Convened domain experts

- + Fire service
- + Smart-steps intelligence (Telefonica)
- + Data analytics (ODI)

Real-time big data processing

- + 509,000 incidents over (4y+)
- + 120,000 network stations
- + 600,000,000 location records

1 expert analysis tool

- + Making cities smarter
- + View impact on people, the borough, and whole city

MARYLAND STATESTAT

WAR ROOM | 14:00



Outcomes

Define open data science

Describe a number of key data science stories

Identify the processes in open data science projects



DATABLOG

Facts are sacred

[Previous](#)[Blog home](#)[Next](#)

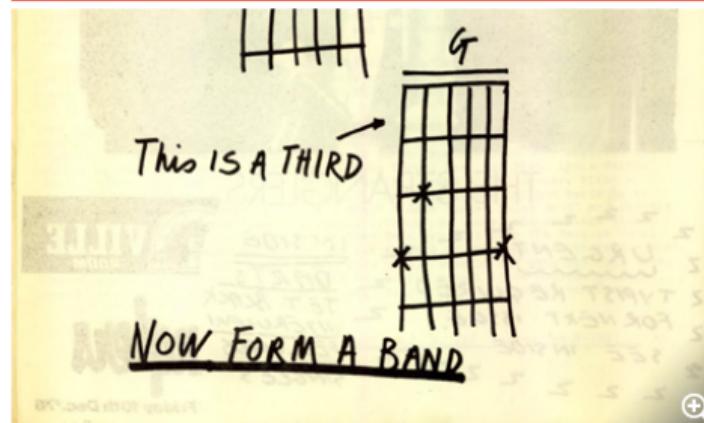
Anyone can do it. Data journalism is the new punk

Can anyone be a data journalist? **Simon Rogers** on what we can learn from a 1977 diagram

- Another view: [What data can and cannot do](#) by Jonathan Gray



Posted by
Simon Rogers
Thursday 24 May 2012
13.00 BST
theguardian.com
[Jump to comments \(8\)](#)



Page two of *Sideburns*, January 1977

●● This is a chord... this is another... this is a third. NOW FORM A BAND



[Article history](#)

Media
Data journalism · Open journalism

Technology

<http://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk>



The 80/20 of data

In **any data project** you will spend 80% of your time preparing, analysing, cleaning, enriching etc... and 20% of the time actually producing the output.



Guardian data blog takes a 70/30 view, but the point remains:

<http://www.theguardian.com/news/datablog/2011/apr/07/data-journalism-workflow>

Data percolation: The stages of data analysis



See also: the Data Journalism Handbook

How should I budget my time?



- 1.1 **FIND** reliable data sources
- 1.2 Understand your **RIGHTS**
- 1.3 Visualise and **UNDERSTAND** your data
- 2.1 **CLEAN** your data
- 2.2 **TRANSFORM** it where useful
- 2.3 **COMBINE** it with other data sets
- 3.1 **REDUCE** and find the story
- 3.2 Think and understand the **CONTEXT**
- 3.3 Do your results pass a **SENSE-CHECK?**

Session 1

Discovering Data Science

Session 2

Gathering and preparing data

Session 3

Publishing insight

Session 2

Gathering and preparing data



Outcomes

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web

Clean a dataset ready for use



How should I budget my time?



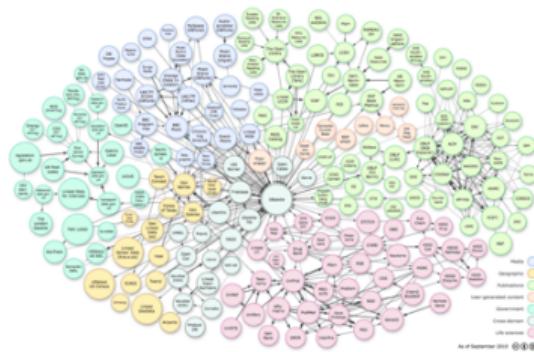
- 1.1 **FIND** reliable data sources
- 1.2 Understand your **RIGHTS**
- 1.3 Visualise and **UNDERSTAND** your data
- 2.1 **CLEAN** your data
- 2.2 **TRANSFORM** it where useful
- 2.3 **COMBINE** it with other data sets
- 3.1 **REDUCE** and find the story
- 3.2 Think and understand the **CONTEXT**
- 3.3 Do your results pass a **SENSE-CHECK?**

Approaches to publishing data

ON the web



IN the web

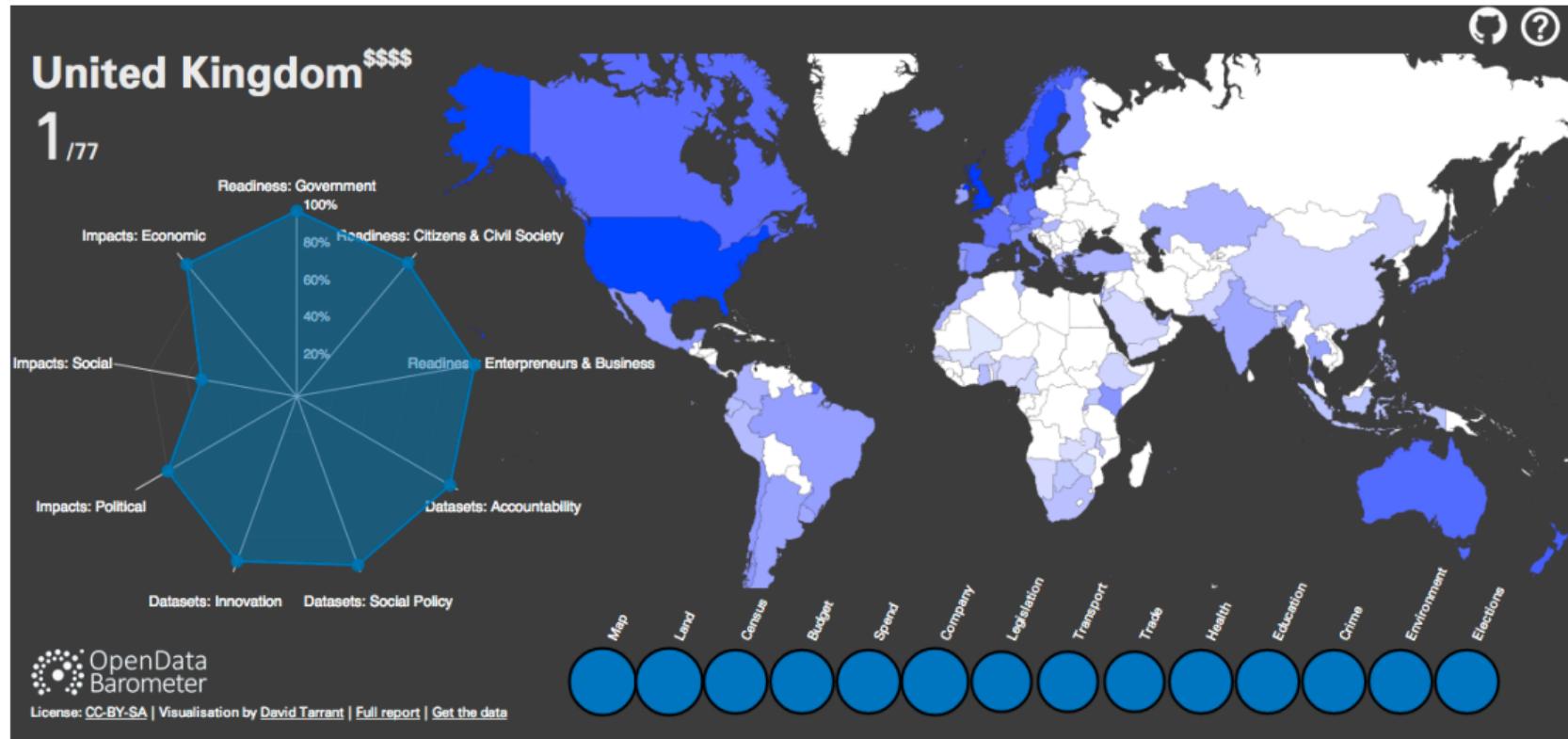


Finding data on the web (of documents)

- Government data
- Private sector data
- Google advanced
- Aggregators and portals
- Scraping



Government data



data.gov.XX

DATOS.GOB.MX BETA OPEN TO PARTICIPATE

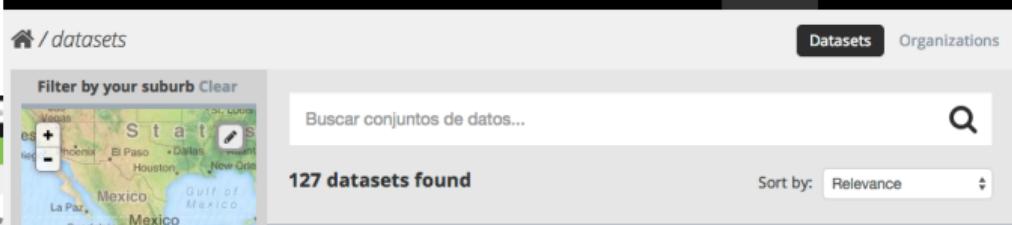
DataStoriesAdvances

/ datasets

Buscar conjuntos de datos...

127 datasets found

Sort by: Relevance



RATING SOCIAL PROGRAMS

Database containing the evaluation of the results of social programs from the federal government subject to the annual assessment.

data.gov.my

Portal Rasmi Open Data Kerajaan Malaysia
The Government of Malaysia's Open Data Official Portal



Latest Datasets | Top Publishers | Feedback

Latest News and Events

Currently, there are no latest event.

Rank	Dataset Name	Last Updated
1	Applikasi Pemetaan Belia Malaysia	2014-06-26 11:30:19
2	Applikasi Pemetaan Belia Malaysia	2014-06-26 11:27:30
3	Applikasi Pemetaan Belia Malaysia	2014-06-26 11:24:52

DATA.GOV.UK Beta Opening up Government

Home Data Apps Interact

Datasets Map Search Data Requests Publishers Public Roles & Salaries Spend Reports Site Analytics Reports

/ Datasets

Search for data...

19266 Results

Sort by

Show only...
Highways Agency

Published datasets (15180)

OpenData Burkina Faso

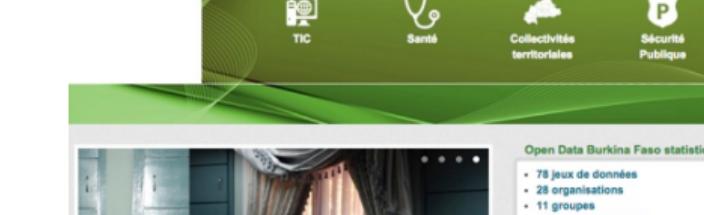
Accueil Comprendre l'Open Data Thèmes Producteurs Jeux de données Applications Partenaires A propos

THEMES

Agriculture	Éducation	Diplomatie	Infrastructures	Eau et Environnement
Tous les thèmes				
TIC	Santé	Collectivités territoriales	Sécurité Publique	Culture

Open Data Burkina Faso statistiques

- 78 jeux de données
- 28 organisations
- 11 groupes



CC BY SA

Suppliers



X OPEN DATA

<http://manufacturingmap.nikeinc.com/#>

You agree not to change or delete any ownership notices from materials downloaded or printed from the Platform. You agree not to modify, copy, translate, broadcast, perform, display, distribute, frame, reproduce, republish, download, display, post, transmit or sell any Intellectual Property or Content appearing on the Platform



Google advanced

Google site:gov filetype:xls

Web Images Maps Shopping More Search tools

About 4,150,000 results (0.22 seconds)

[XLS] [Code List or Concept \(Acronym\)](#)

www.acquisition.gov/short_codelistsTS.xls Share

File Format: Microsoft Excel - [View as HTML](#)

A, B, C, D, F, G, H, 1, Code List or Concept (Acronym), Definition, Authoritative or Maintenance Agency, Current Version, Update Frequency and/or ...

[XLS] [Approps - Foreign Assistance.gov](#)

www.foreignassistance.gov/Full_ForeignAssistanceData.xls

File Format: Microsoft Excel

A, B, C, D, E, F, G, H, 1, Planning Data, 2, 3, 4, Fiscal Year, Fiscal Year Type, Account Name, Agency Name, Operating Unit, Category, Sector, Amount ...

[XLS] [TSB Monthly Cash Flow Projection](#)

www.dia.iowa.gov/tsb/cashflow.xls

site: Get results only from certain sites or domains

link: Find pages that link to a certain page

related: Find sites similar to one you already know

filetype: Find certain file types only

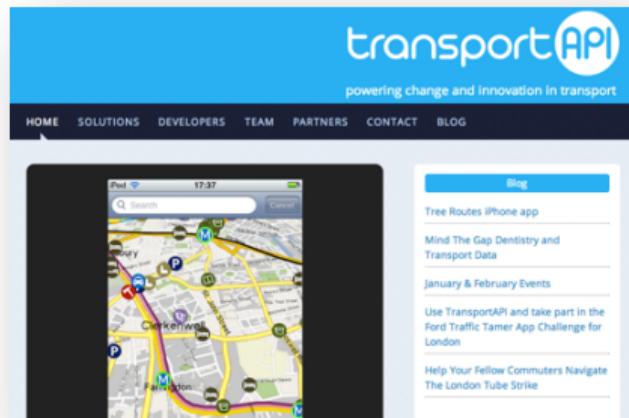


Aggregators and portals

Collect together data from across the web into one place.



enigma.io



transportAPI



Scraping

If you can't obtain usable data (csv, xls) then you may have to resort to scraping.



“excellent, so excited beyond description”
George Ofosu, Doctoral Student, UCLA

pdftables.com

A screenshot of the import.io web interface. At the top, there's a search bar with the placeholder "Enter a URL, for a list page" and a pink "Extract Data" button. Below the search bar, there are several examples of websites being scraped, each with a preview image and the name of the website below it: "Reseller Ratings", "Zoopla", "500px", "Growth Hackers", "Udemy", and "Stack Exchange".

magic.import.io

Exercise

Explore some of the tools and methods to discover new data on the **web of documents** relevant to you.

10 mins

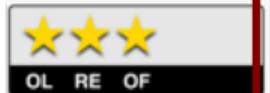


5-Stars



<http://5stardata.info/>

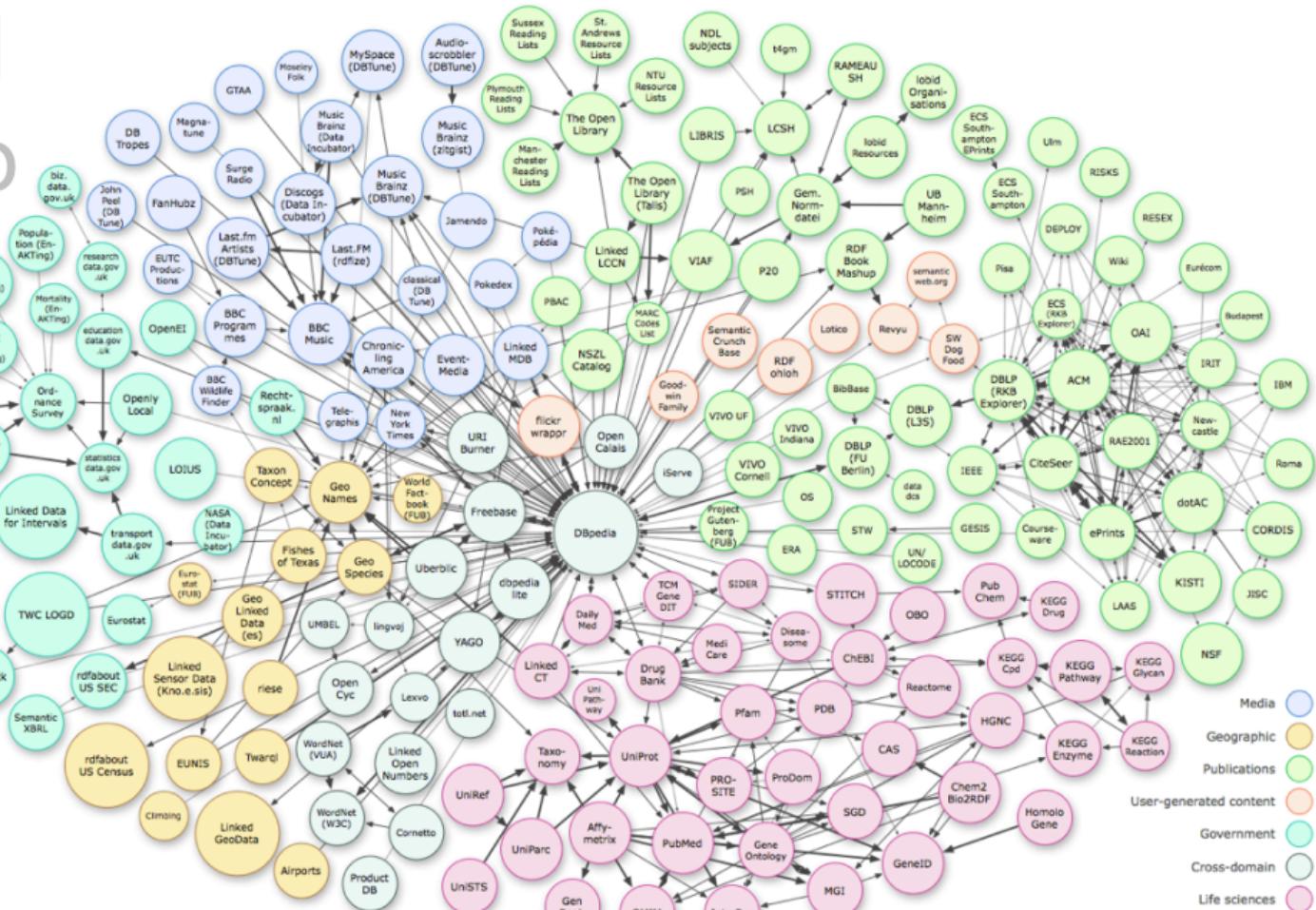
ON THE WEB



IN THE WEB



Data IN the web



Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data **IN THE WEB**
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)

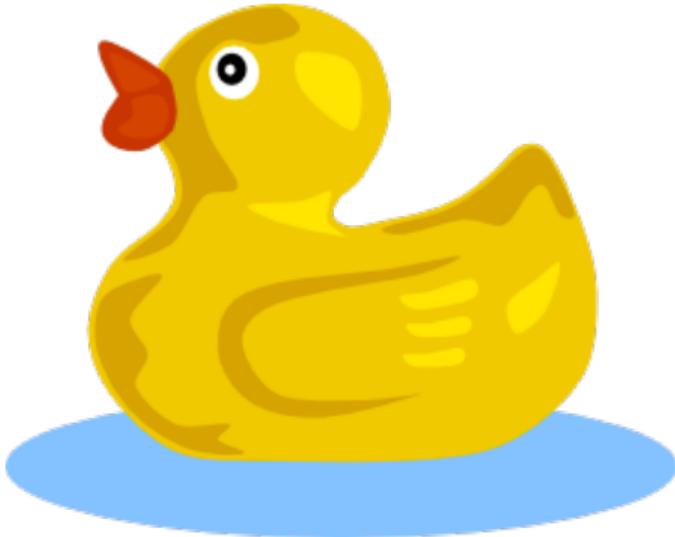


How the web should work,
but people forgot that Tim
put this in when he
invented it!

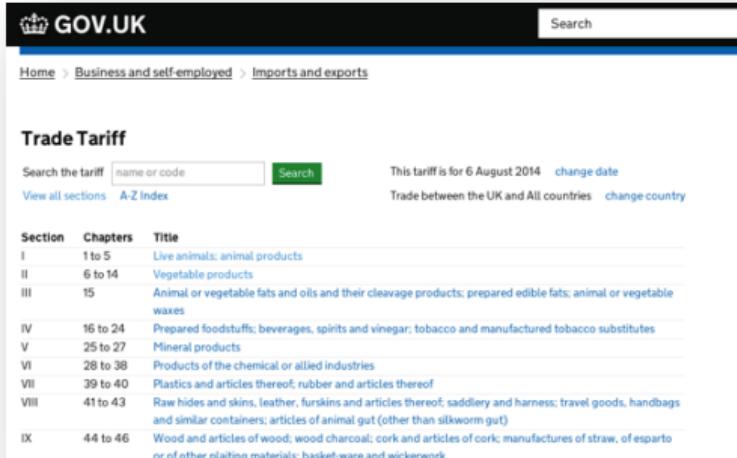
Duck typed data

If it looks like a duck
and quacks like a duck,
then it's probably a duck.

Basically, keep an eye out for tables,
lists and other stuff that looks like data.



1. Adding random extensions



The screenshot shows the UK Trade Tariff page on the GOV.UK website. At the top, there's a search bar and a navigation menu with links to Home, Business and self-employed, and Imports and exports. Below this, the title "Trade Tariff" is displayed. A search bar allows users to search by name or code, with a "Search" button. To the right, a note says "This tariff is for 6 August 2014" and provides options to change date and country. A sidebar lists "View all sections" and "A-Z Index". The main content area is a table with three columns: Section, Chapters, and Title. The sections are numbered I through IX, with their corresponding chapters and titles listed below.

Section	Chapters	Title
I	1 to 5	Live animals; animal products
II	6 to 14	Vegetable products
III	15	Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes
IV	16 to 24	Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes
V	25 to 27	Mineral products
VI	28 to 38	Products of the chemical or allied industries
VII	39 to 40	Plastics and articles thereof; rubber and articles thereof
VIII	41 to 43	Raw hides and skins; leather, furskins and articles thereof; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)
IX	44 to 46	Wood and articles of wood; wood charcoal; cork and articles of cork; manufactures of straw, of esparto or of other plaiting materials; basket-ware and wickerwork

UK Trade Tariff



The screenshot shows the BBC Music and Programmes website for the TV show Doctor Who. The header features the BBC One logo and the show's title "DOCTOR WHO" with the TARDIS icon. Below the header, there are links for Home, Episodes, Clips, Galleries, Latest News, Characters, Monsters, Fun and Games, and More. The main content area has two sections: "On iPlayer" and "On TV". The "On iPlayer" section shows a thumbnail of the Doctor and a woman, with text about the latest launch. The "On TV" section shows a thumbnail of the Doctor and another character, with text about the Day of the Doctor broadcast.

BBC Music and Programmes

Try using the following: .csv .json .xml .rss .rdf

2. Look for alternative links



Business Insight - NEWSASIA

NEWS TV WATCH LIVE

Wed, Aug 06 2014

ASIA PACIFIC SINGAPORE WORLD BUSINESS SPORT ENTERTAINMENT TECHNOLOGY HEALTH LIFESTYLE VIDEOS WEATHER MORE ▾

CHANGINGLIVES LUMINARY AWARDS START-UP

Scroll down!

SINGAPORE STORIES

Raise of up to 12% for Home Team officers, with sign-on bonuses of up to S\$30,000

National Day Award 2014

SP (2) Ng (SP)

MEDIACORP

Officers from the Home Team, both past and present were recognised at the Home Team National Day Observance Ceremony on Wednesday.

9 hours ago

Pay rise, special bonus for about 23,000 nurses

10 hours ago

50,000 openings on Jobs Bank for Singaporeans, PRs

1 hour ago

NUS University Town identified as a high-risk dengue cluster

10 hours ago

LIFESTYLE VIDEOS

2. Look for alternative links



 CHANNEL NEWSASIA MediaCorp News Group. © 2014 MediaCorp Pte Ltd. All Rights Reserved. Terms and Conditions Privacy Policy About MediaCorp Pte Ltd	NEWS Asia Pacific Singapore World Business Sport Entertainment Technology Health Lifestyle Videos Photos Special Reports Archives	TV Live TV TV Videos TV Schedule SERVICES Weather ADVERTISE WITH US Online Advertising Mobile Advertising TV Advertising Contact Sales	ABOUT US About Channel NewsAsia Our Logo Our Coverage Our Tagline Presenters and Correspondents Contact Us GET OUR NEWS 
---	---	---	---



RSS



3. Look for embedded data

ODI Experiment

Hidden data extractor

open
data
institute

Hidden data extractor

Enter the URL of any webpage to see what JSON data is hidden within it.

Submit

Try these

[Products from Marks and Spencer UK](#)

[Products from ASOS](#)

CC BY SA

Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data
4. Do some content negotiation
5. **Spot the API**
6. Scrape (or search google again)



Exercise

What is an API?

Any examples?



What is an API

Defines how one application
can **consistently** interact with
another.



Teaching nerds passion

```
function makeOut(passionLevel, partsOfBody) {  
    for (each partOfBody in partsOfBody) {  
        partOfBody.kiss(passionLevel);  
        lookIntoEyes();  
        sighDeeply();  
    }  
    moanDaintily();  
    sleep();  
}
```



Making your call

```
function makeOut(passionLevel, partsOfBody) {  
    for (each partOfBody in partsOfBody) {  
        partOfBody.kiss(passionLevel);  
        lookIntoEyes();  
        makeOut(10, ["neck", "ear", "mouth"]);  
    }  
    moanDaintily();  
    sleep();  
}
```



The Webs API

```
class Photos {  
    function search(api_key, tags) {  
        if (!validKey(api_key)) {return(401);}  
        ...  
        ...  
        return results;  
    }  
}
```



Using the Webs API

`http://api.flickr.com/services/rest/?method=flickr.photos.search&api_key=a8c42ef2ac8d88aed7e351f95e93160f&tags=fox`

`http://api.flickr.com/services/rest/
 ?method=flickr.photos.search
 &api_key=a8c42ef2ac8d88aed7e351f95e93160f
 &tags=fox`



A better solution?

`https://www.flickr.com/photos/{user}/{photo}/comments
/tags
/licence`

`https://www.flickr.com/search?tags=...`





<http://docs.transportapi.com/>

```
http://fcc.transportapi.com/{ api_version } /  
    /uk/train/station/{station_code}/live.json  
    /uk/train/station/{station_code}/{date}/{time}/t  
imetable.json
```

Exercise

Explore some of the tools and methods to discover new data on the **web of data** relevant to you.

15 mins



Outcomes

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web

Clean a dataset ready for use



Introducing Open Refine

Google Refine 2.0 - Introduction (1 of 3) (vide...)

Facet / Filter Union / Redo ▾ Refresh ▾ Reset All Remove All

2250 rows Show 5 10 25 50 rows ▾ first ▾ previous 1 - 50 next ▾ last ▾

Contract ID Contractor Name Type of Contract Date of Award Start Date End Date Total value of Contract Contract Awarded

1. 1038	ASAP SOFTWARE EXPRESS INC D/B/A CONTRACTS	Minimum Entitlement Agreement	04/01/2009	04/01/2009	06/03/2011	1,302 year
2. 1040	WAC SOFTWARE DISTRIBUTION INCORPORATED	Revolving Service Desk Maintenance	04/01/2009	04/01/2009	03/01/2010	0.001 year
3. 1041	GO-COMMUNICATIONS INCORPORATED	Cisco SnapShot	25/01/2009	05/01/2009	04/03/2011	0.007 year
4. 1042	IT'S CORPORATION	Time & Materials	12/01/2008	01/01/2009	12/03/2011	20 year
5. 1043	ISBET INTERNATIONAL INCORPORATED		05/01/2009	05/05/2009	27/03/2009	0.040157 year
6. 1044			01/06/2009	01/06/2009	18/03/2010	0.738 year
7. 1045	IT FEDERAL SERVICES LIMITED LIABILITY COMPANY	Firm Fixed Price	10/01/2008	10/01/2008	09/03/2010	0.342 year
8. 1046		Firm Fixed Price	08/05/2009	10/01/2009	08/05/2010	0.004 year
9. 1047		Firm Fixed Price	11/05/2009	11/05/2009	05/03/2010	0.002 year
10. 1048	RECHUNK IT SOLUTIONS LLC	Firm Fixed Price	01/03/2009	01/01/2010	12/01/2010	0.912 year

0:00 / 6:48

YouTube

<http://openrefine.org>



Data processing pipelines

ODI Experiment

Refine AutoBot

Refine AutoBot

Enter the URL of the CSV file for cleaning.

Enter the refine operation history.

Submit

The screenshot shows a web-based application titled "Refine AutoBot". At the top left is the title "Refine AutoBot". To its right is the "ODI Experiment" logo. On the far right is the "open data institute" logo, which consists of the letters "odi" in a bold, white, sans-serif font, with "open data institute" in a smaller, white, sans-serif font below it. The main body of the page has two input fields. The first field is a long rectangular input box with the placeholder text "Enter the URL of the CSV file for cleaning.". Below it is another long rectangular input box with the placeholder text "Enter the refine operation history.". At the bottom center is a large blue rectangular button with the word "Submit" in white. The entire application is set against a white background.

<http://theodi.github.io/refine-autobot/>

Session 1

Discovering Data Science

Session 2

Gathering and preparing data

Session 3

Publishing insight

Session 3

Publishing insight



Outcome

Build simple web pages that bring together a number of datasets to reveal new insight





World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) .

What's out there?

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

Help

on the browser you are using

Software Products

A list of W3 project components and their current state. (e.g. [Line Mode](#) ,[X11 Viola](#) , [NeXTStep](#) , [Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

Technical

Details of protocols, formats, program internals etc

Bibliography

Paper documentation on W3 and references.

People

A list of some people involved in the project.

History

A summary of the history of the project.

How can I help ?

If you would like to support the web..

Getting code

Getting the code by [anonymous FTP](#) , etc.

Human
Readable
Web



Question

What is significant about the first web page?

What actually did Tim invent?



Markup Links & Tags

World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) .

[What's out there?](#)

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) , [X11 Viola](#) , [NeXTStep](#) ,
[Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

If you would like to support the web..

[Getting code](#)

Getting the code by [anonymous FTP](#) , etc.



World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) .

<list>

What's out there?

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) , [X11 Viola](#) , [NeXTStep](#) ,
[Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

If you would like to support the web..

[Getting code](#)

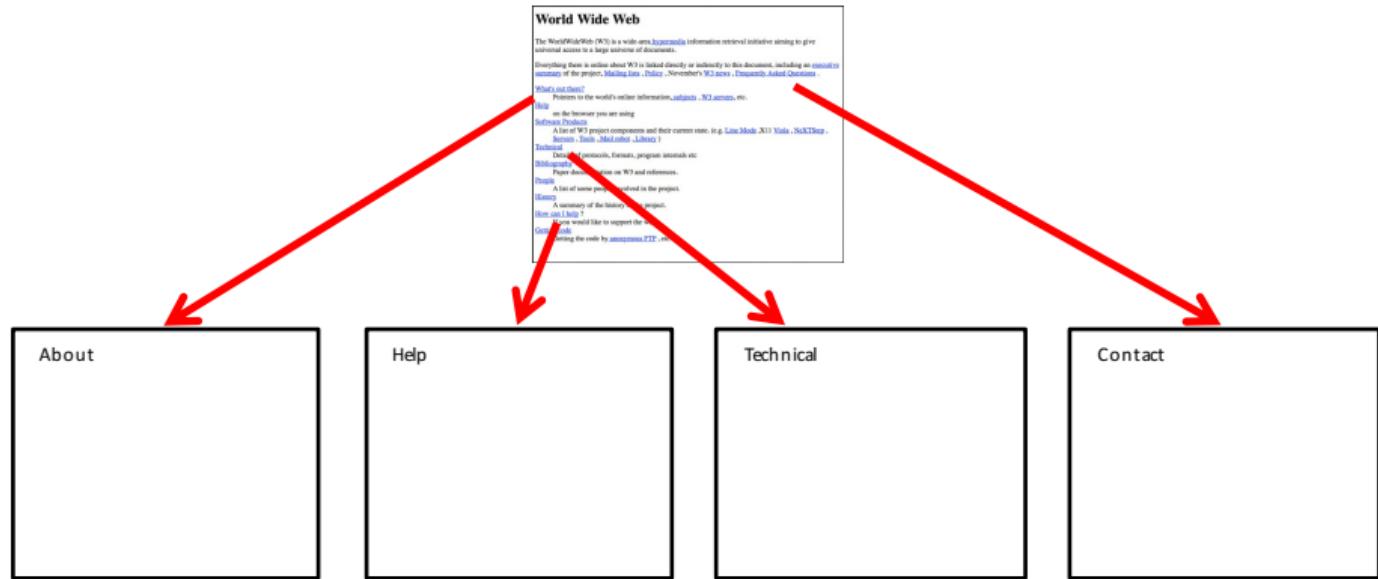
Getting the code by [anonymous FTP](#) , etc.

</list>

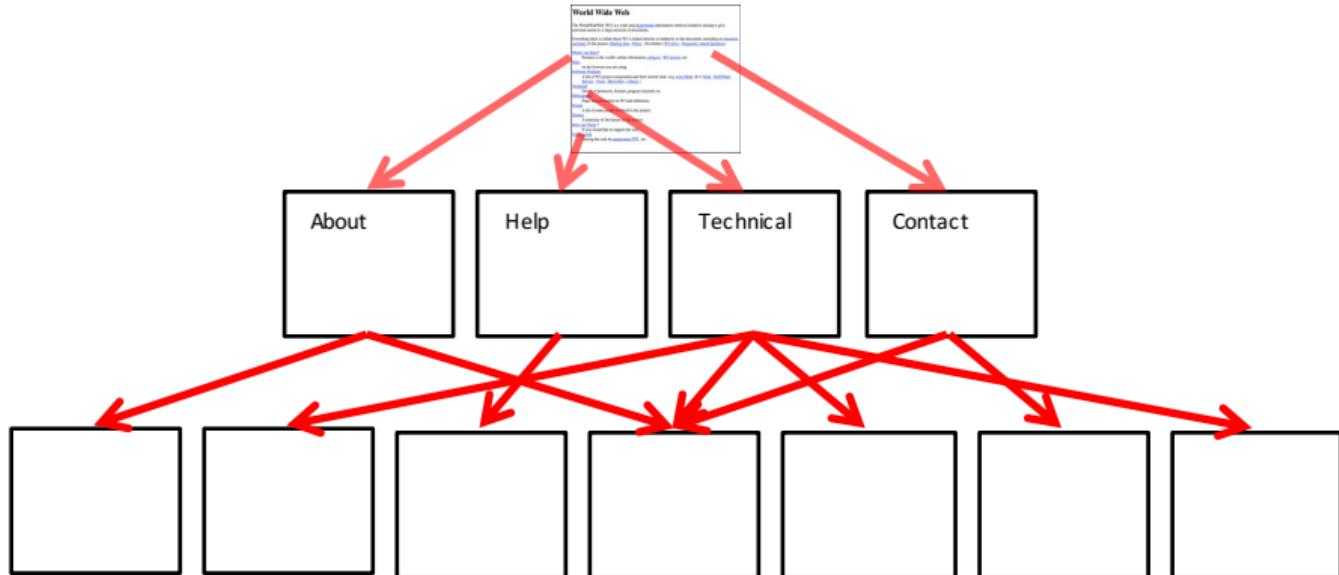
Markup Links & Tags



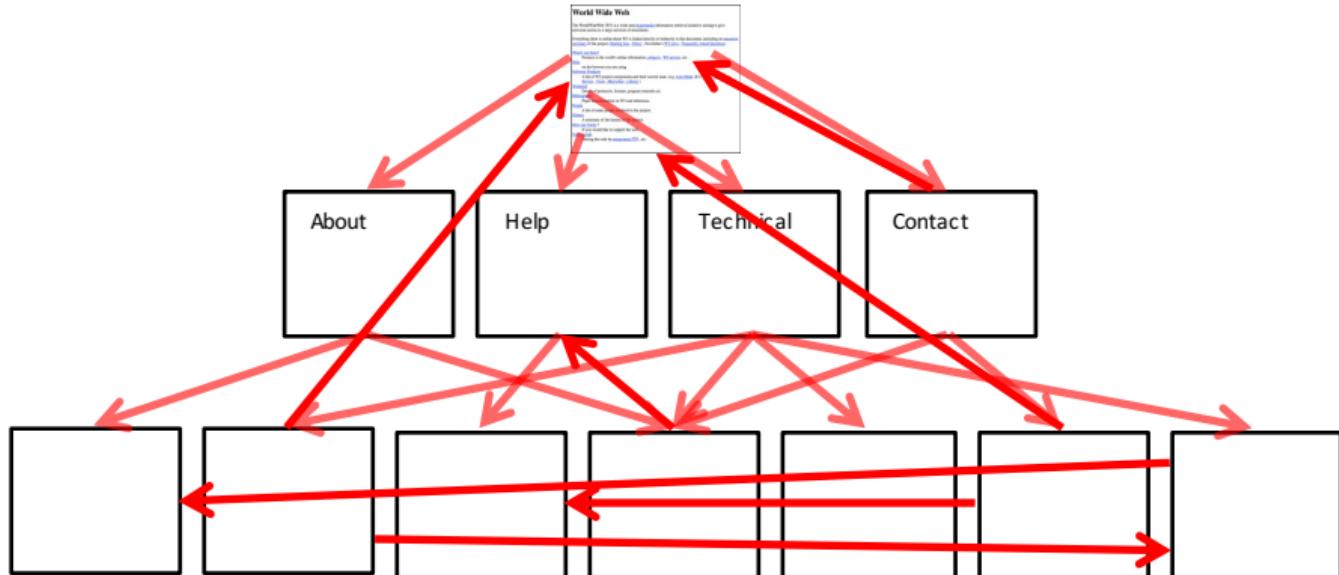
Nodes and Links

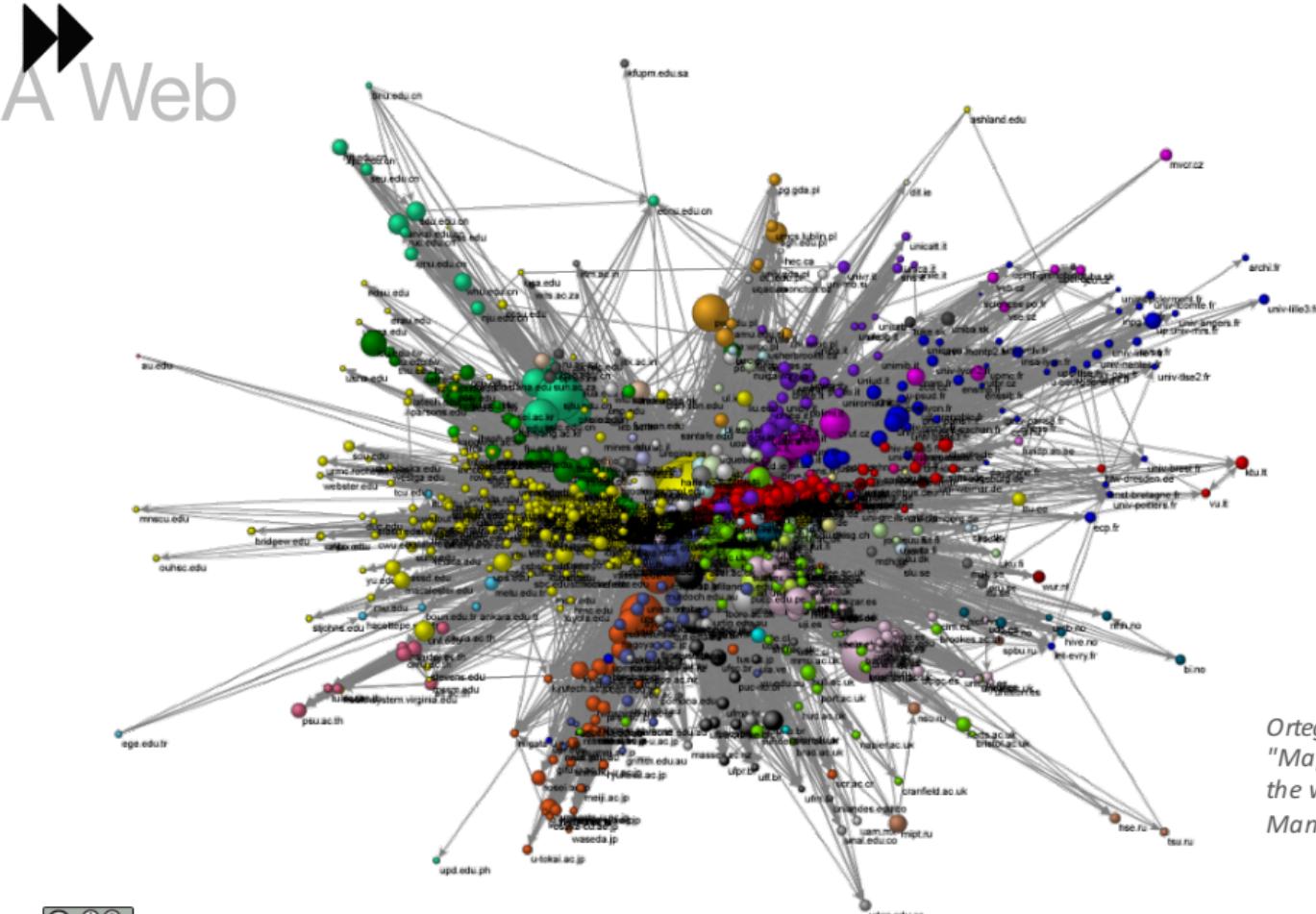


Nodes and Links



Nodes and Links





Ortega, Jose Luis, and Isidro F. Aguillo.
"Mapping world-class universities on the web." *Information Processing & Management* 45.2 (2009): 272-279.

HTML

HyperText (Links) Markup Language

Here is some *really important* text!

Remember that in the morning we start at 9:30am

<blink>

This text is likely to annoy you.

</blink>

HTML

<h>World Wide Web </h>

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

<h1>This is heading 1</h1>

<h2>This is heading 2</h2>

<h3>This is heading 3</h3>

<h4>This is heading 4</h4>

<h5>This is heading 5</h5>

<h6>This is heading 6</h6>



A Link



Link to BBC News



HTML 5

<menu>

<summary>

<figure>

<details>

<nav>

<legend>

↔blink↔

<input>

<label>

<marquee>

<header>

<title>

<section>

<option>

<footer>

<a>





HTML 5

html																			col	table															
head	span																			div	fieldset	form	body	h1	section	colgroup	tr								
title	a																			pre	meter	select	aside	h2	header	caption	td								
meta	rt	dfn	em	i	small	ins	s	br	p	blockquote	legend	optgroup	address	h3	nav	menu	th																		
base	rp	abbr	time	b	strong	del	kbd	hr	ol	dl	label	option	datalist	h4	article	command	tbody																		
link	noscript	q	var	sub	mark	bdi	wbr	figcaption	ul	dt	input	output	keygen	h5	footer	summary	thead																		
style	script	cite	samp	sup	ruby	bdo	code	figure	li	dd	textarea	button	progress	h6	hgroup	details	tfoot																		
																					img	area	map	embed	object	param	source	iframe	canvas	track	audio	video			

HTML5



Why Markup?

Aids your browser to render the page

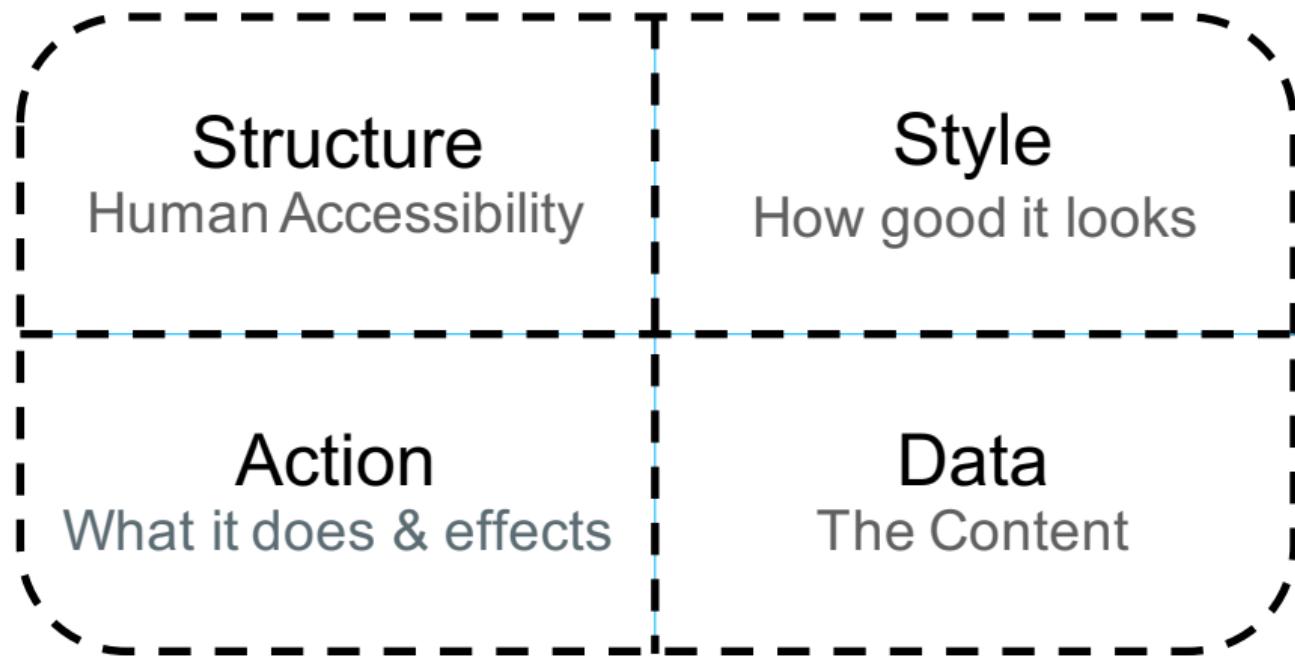
Critical for screen readers!

Adds semantics about importance of elements.

Aids search engines



Building Blocks



The Language of the Web

HTML5



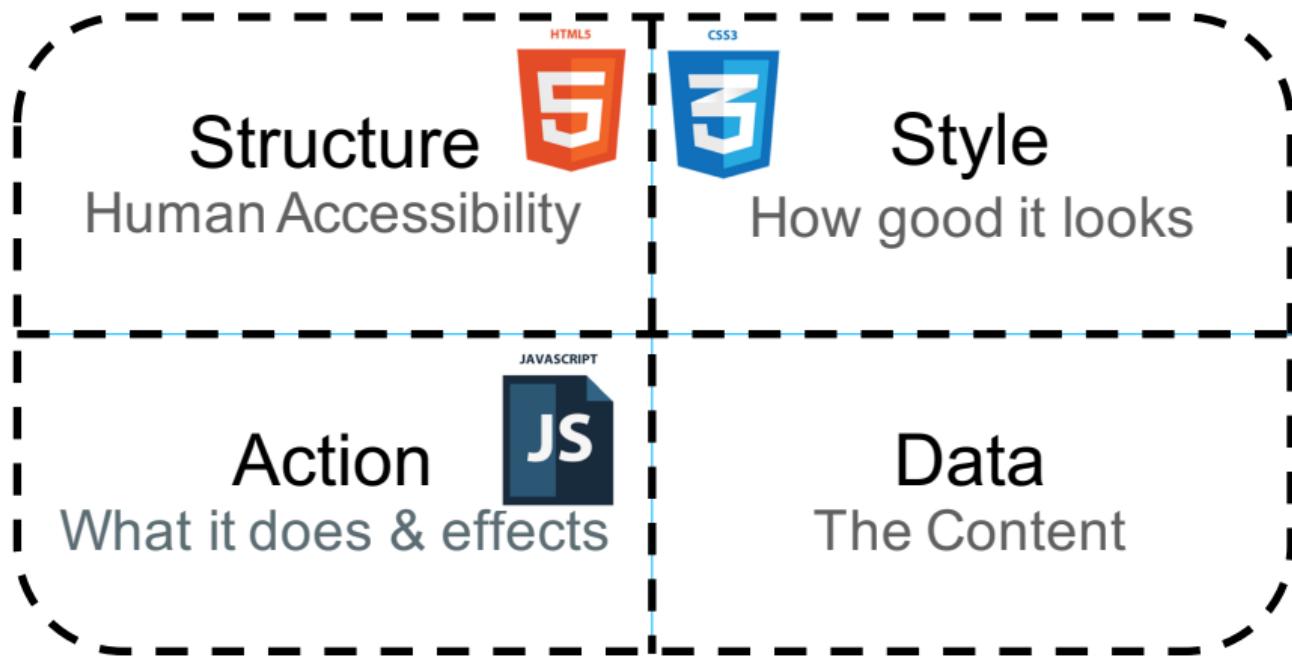
CSS3



JAVASCRIPT



Building Blocks



JSON: A Natural Fit



Last exercise

Build a web page that shows the time and platform for your train home*.

<http://training.theodi.org/UnlockingData/>

* I've made a massive assumption that someone catches a train home.



Session 1

Discovering Data Science

Session 2

Gathering and preparing data

Session 3

Publishing insight

Outcomes

Define open data science

Describe a number of key data science stories

Identify the characteristics in open data science projects



Outcomes

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web

Clean a dataset ready for use



Outcome

Build simple web pages that bring together a number of datasets to reveal new insight



Homework

Complete your web page and then build a new one to consume other data from another service and display it.





Thank-you