nesta

Data science for Science, Technology and Innovation research

Juan Mateos-Garcia
April 2018

About today

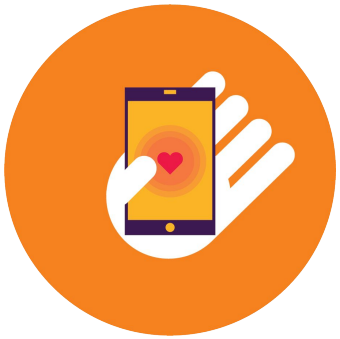
- Convince you of the potential of data science for Science, Technology and Innovation Research (DS4STIR)
- Review these methods and (hopefully) clarify some of the vocabulary
- Flag some of the gotchas and questions you should ask yourself
- Give you some advice about how/where to get started

1. **Why** do DS4STIR?
   - Different types of STIR knowldedge
   - Limitations of the status quo
   - New opportunities
   - A definition
2. **What** are the contents of the DS4STIR toolkit?
   - Data collection tools
   - Analysis tools
   - Reporting tools
3. **How** can you get started with DS4STIR (if you haven't already)?
4. **Group activity**
5. **Conclusion**

We are an innovation foundation.

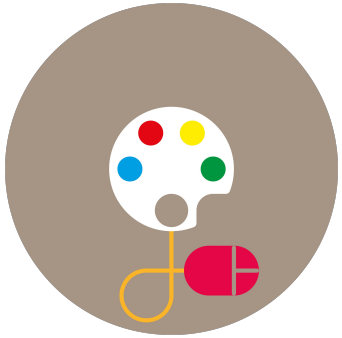
We back new ideas to tackle the big challenges of our time.

Our priority fields of work



Health

Government innovation



The creative economy, arts & culture

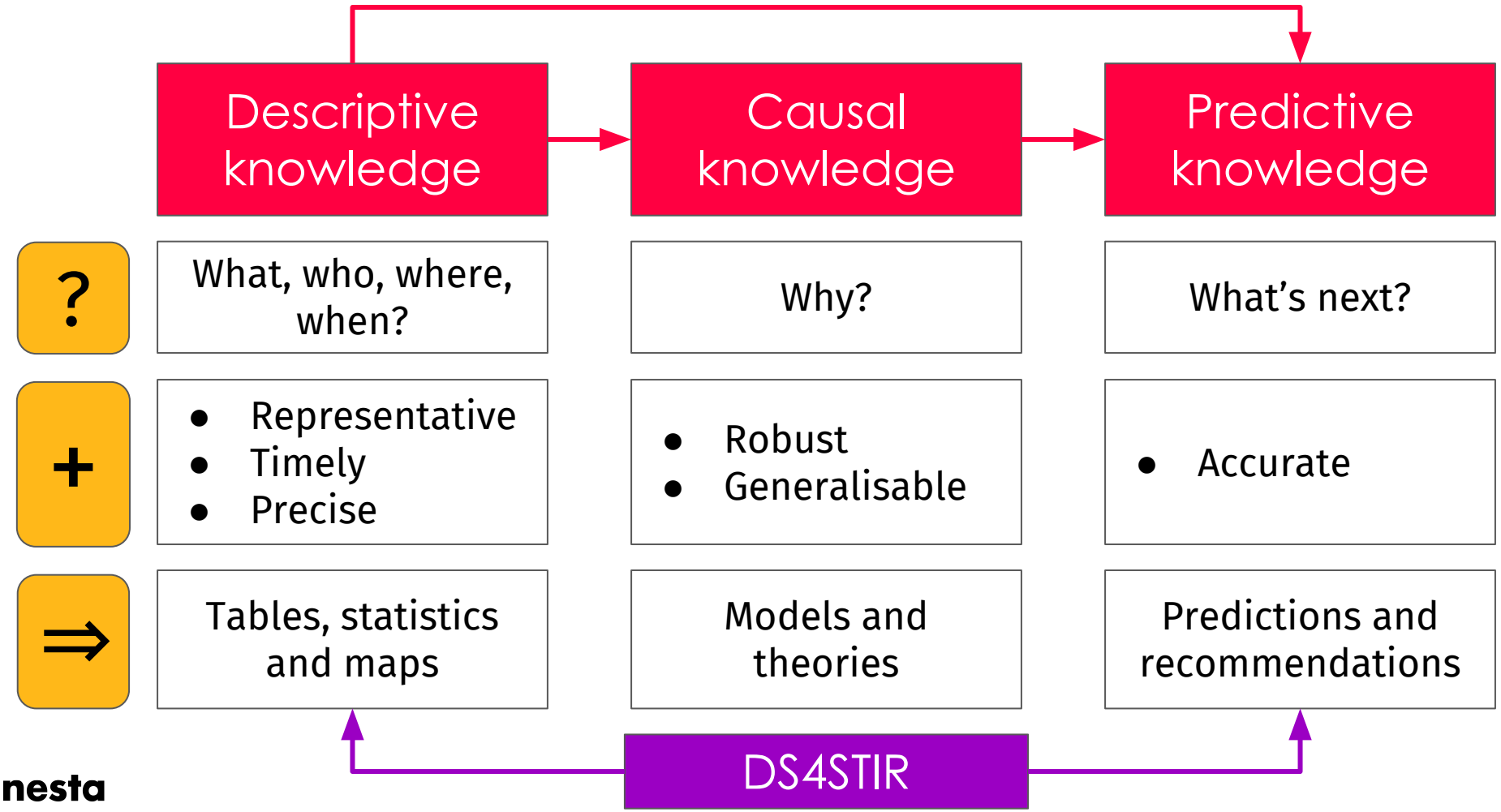
Education

Innovation policy

Why do DS4STIR?

nesta

STIR = Interdisciplinary field that analyses the processes through which new ideas are generated, diffused and applied. Some of this **knowledge** can be used to develop policies to support and steer innovation for the common good.

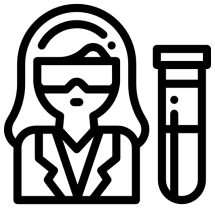


Data status quo (and its discontents)

HUMAN RESOURCES
1.1.1 New doctorate graduates
1.1.2 Population completed tertiary education
1.1.3 Lifelong learning
RESEARCH SYSTEMS
1.2.1 International scientific co-publications
1.2.2 Scientific publications among top 10% most cited
1.2.3 Foreign doctorate students
INNOVATION-FRIENDLY ENVIRONMENT
1.3.1 Broadband penetration
1.3.2 Opportunity-driven entrepreneurship (Motivational Index)
FINANCE & SUPPORT
2.1.1 Public R&D expenditure
2.1.2 Venture capital
FIRM INVESTMENTS
2.2.1 Business R&D expenditure
2.2.2 Non-R&D innovation expenditure
2.2.3 Enterprises providing ICT training
INNOVATORS
3.1.1 SMEs with product or process innovations
3.1.2 SMEs with marketing/organisational innovations
3.1.3 SMEs innovating in-house
LINKAGES
3.2.1 Innovative SMEs collaborating with others
3.2.2 Public-private co-publications
3.2.3 Private co-funding of public R&D expenditures
INTELLECTUAL ASSETS
3.3.1 PCT patent applications
3.3.2 Trademark applications
3.3.3 Design applications
EMPLOYMENT IMPACTS
4.1.1 Employment in knowledge-intensive activities
4.1.2 Employment in fast-growing firms of innovative sectors
SALES IMPACTS
4.2.1 Medium & high tech product exports
4.2.2 Knowledge-intensive services exports
4.2.3 Sales of new-to-market and new-to-firm innovations

EIS (2018)

Standard industry classifications, business surveys, administrative data and bibliometrics (legal proxies)

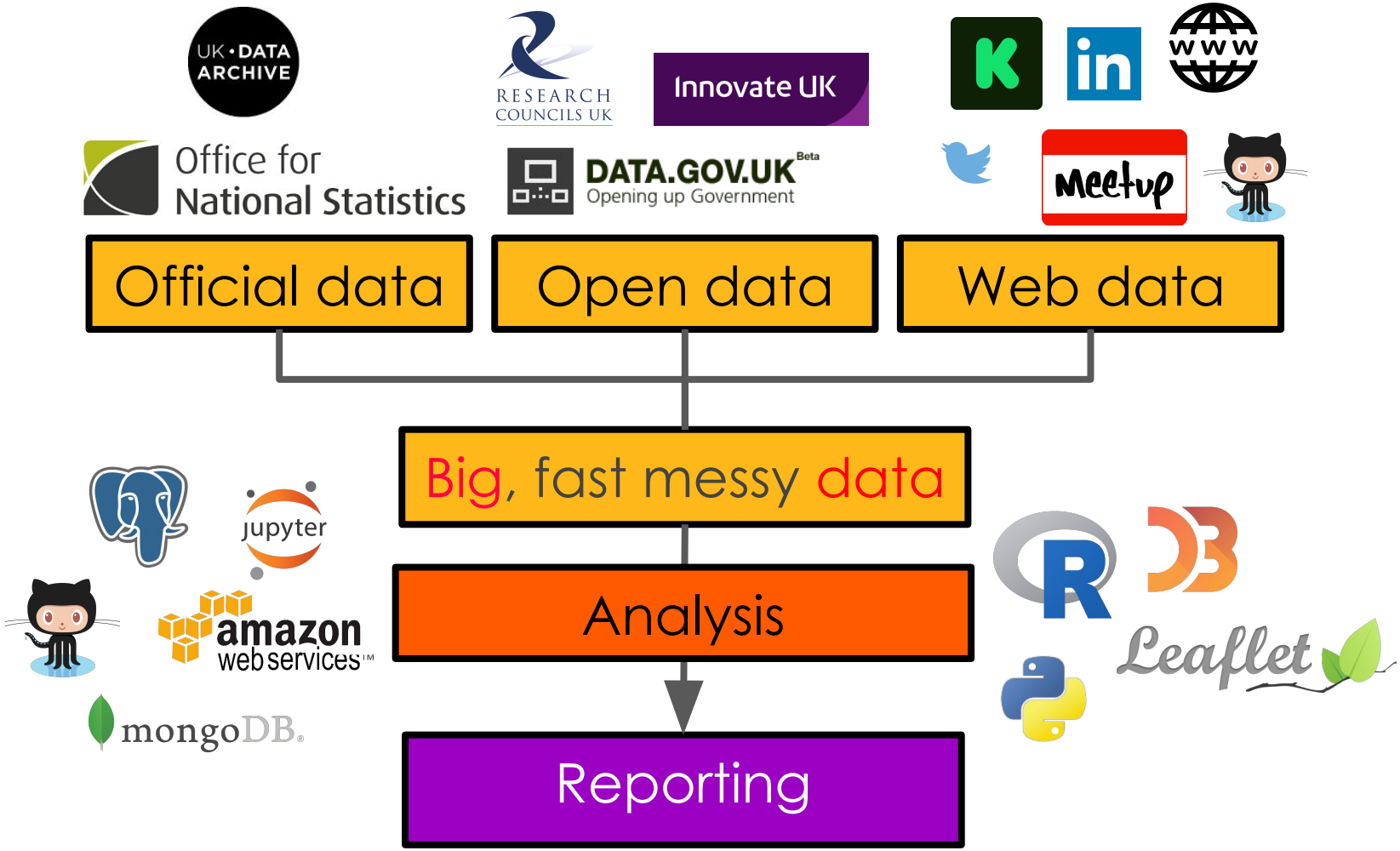


+	−
RELEVANT Captures STI activities and sectors	**BIASED** Misses hidden innovation in other industries
STRUCTURED Easy to understand categories	**RIGID** Unsuitable for analysing emergence and hybridisation
AGGREGATE Removes noise and reflects policy audiences	**COARSE** Black box, irrelevant for many audiences

Risk: skewed research and ineffective policy

nesta

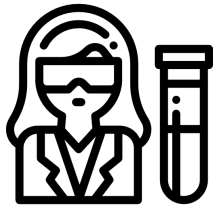
New data paradigm



Some examples

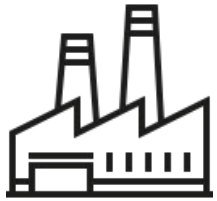
Nesta has been working in this area for ~6 years. We currently have a team of 8 people working in projects exploiting different opportunities opened up by new data sources.

+

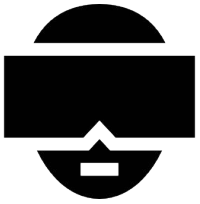


INCLUSIVE
Captures innovation outside science & tech

Mapping Health Innovation (w/ Robert Woods Johnson Foundation): 15 month project to map non-science based health innovations with new data

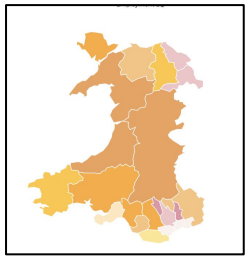


FLEXIBLE
Work with unstructured data to understand emergence



Mapping the Immersive economy: Forthcoming report with Innovate UK using web data to identify VR/AR companies in the UK

GRANULAR
Generate outputs relevant for more audiences

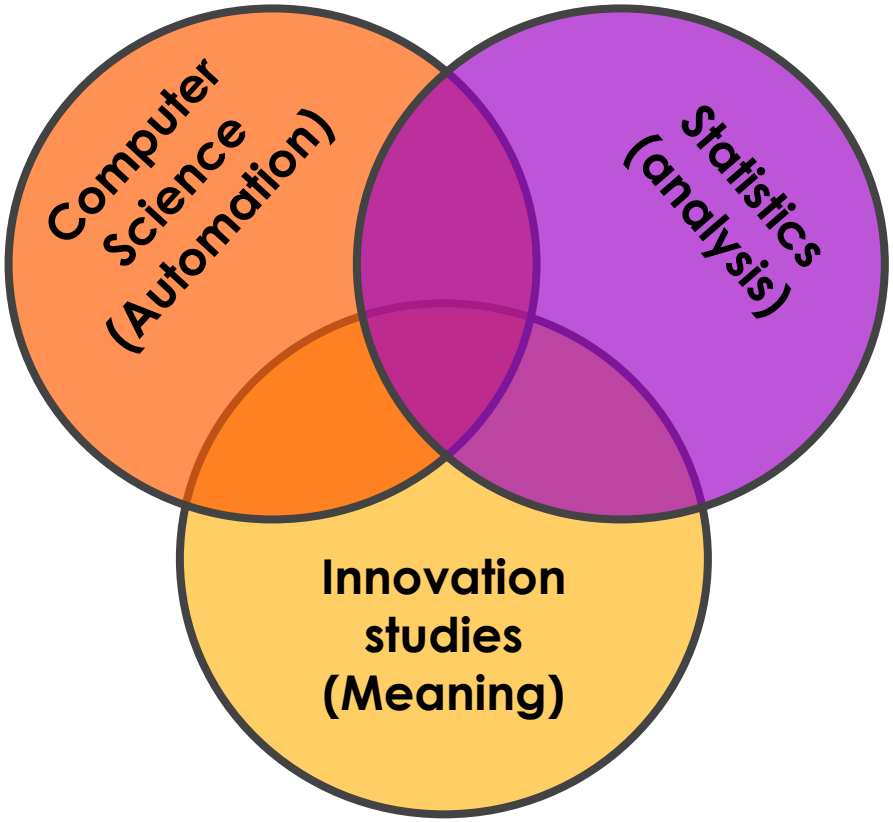


Arloesiadur.

Platform measuring innovation in Wales and its local economies (now being extended to Scotland)

nesta

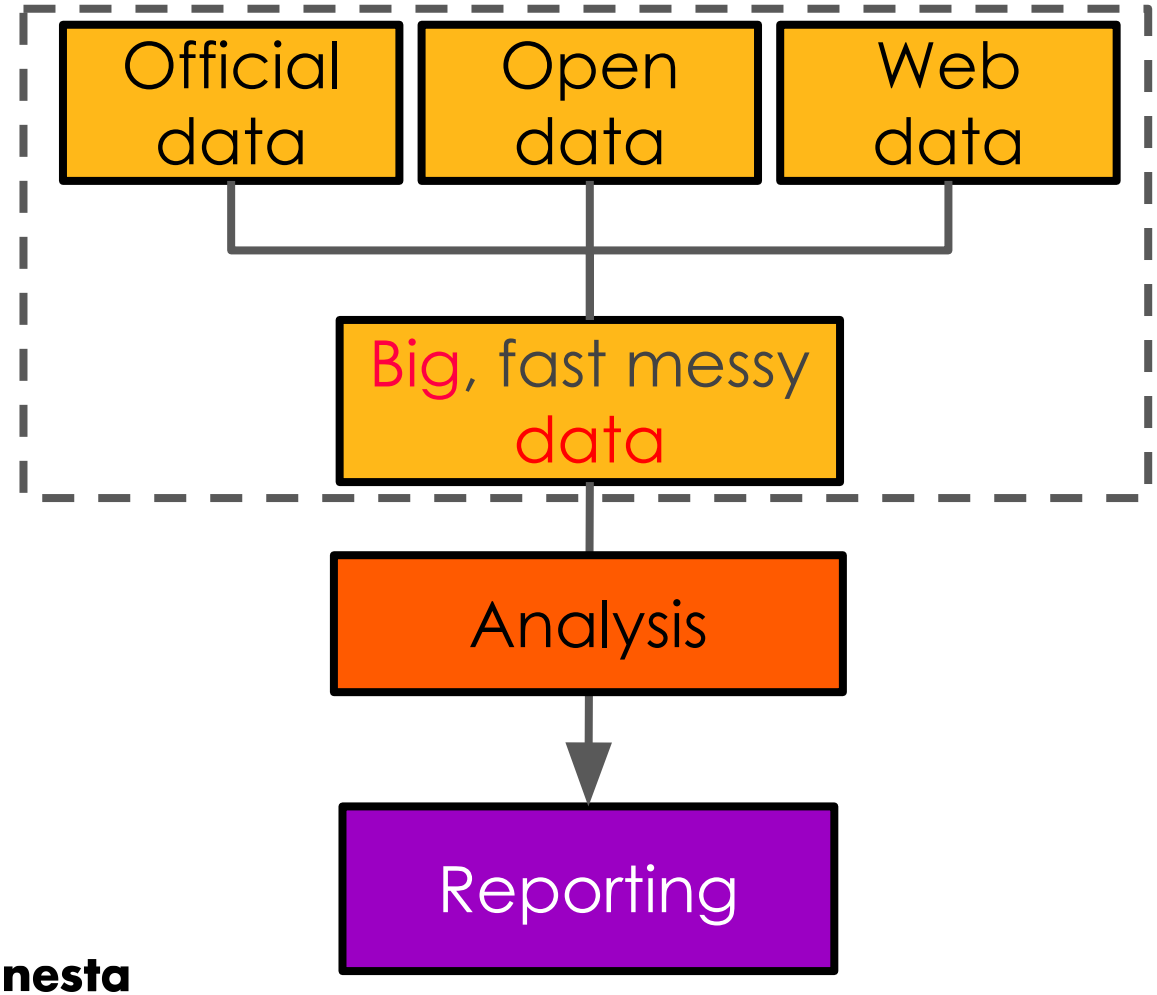
DS4STIR ⇒ Combinations of computer science, statistics and subject (innovation) domain knowledge to create value from new data sources, analytics methods and data products



What are the contents of the DS4STIR toolkit?

Data pipeline = collection of activities through which we transform raw data into valuable information / knowledge / products and services

Note: It is not linear, lots of feedback loops!!

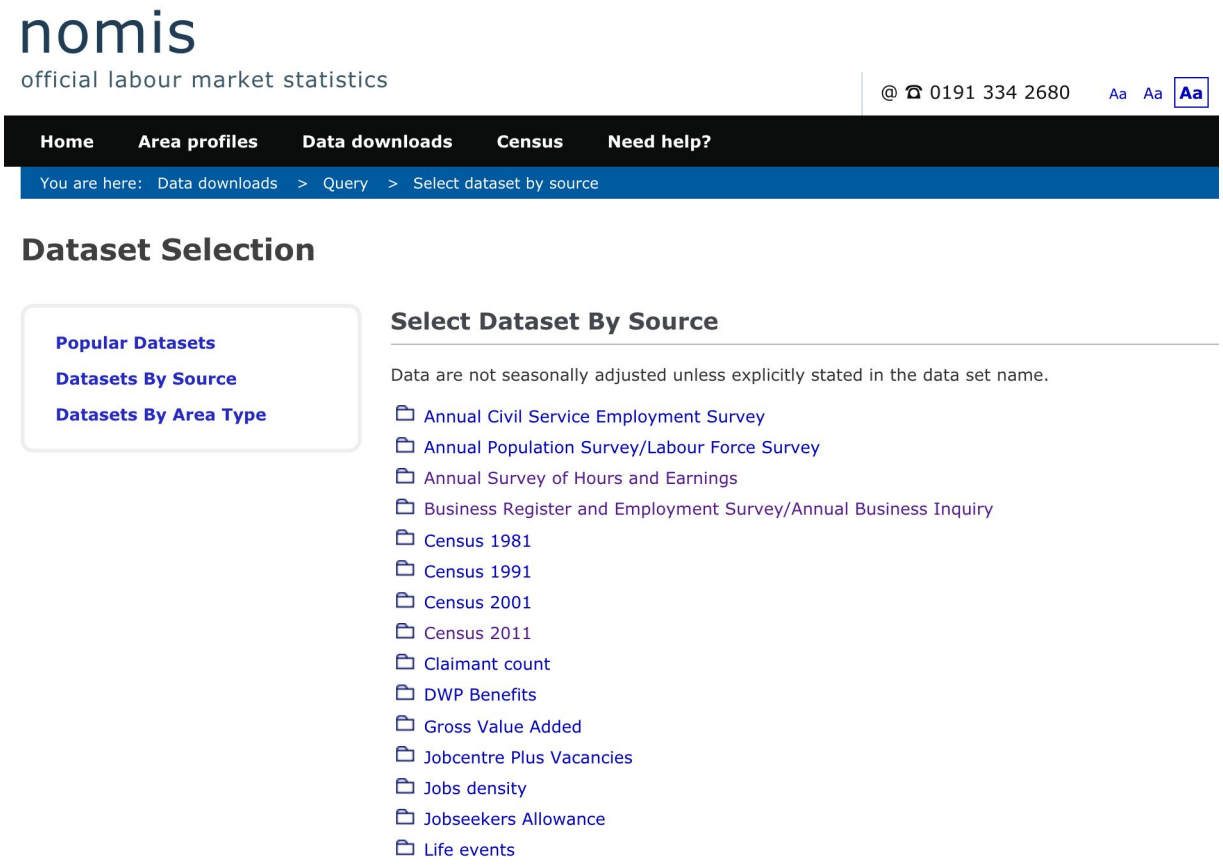


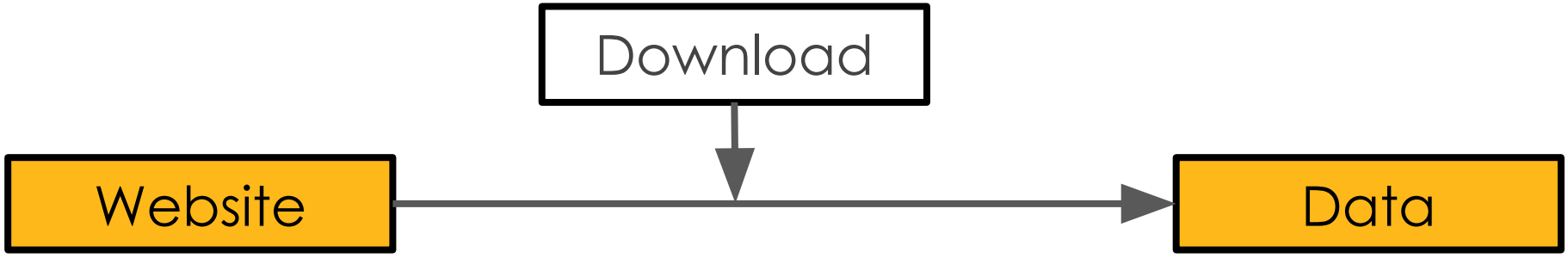
The first stage of the data science pipeline is to collect the data.

I assume that you have already identified an interesting research question and reviewed the literature. **Without that, you are lost!**

Question: How easy or hard has the 'provider' made it to for you to get the data?

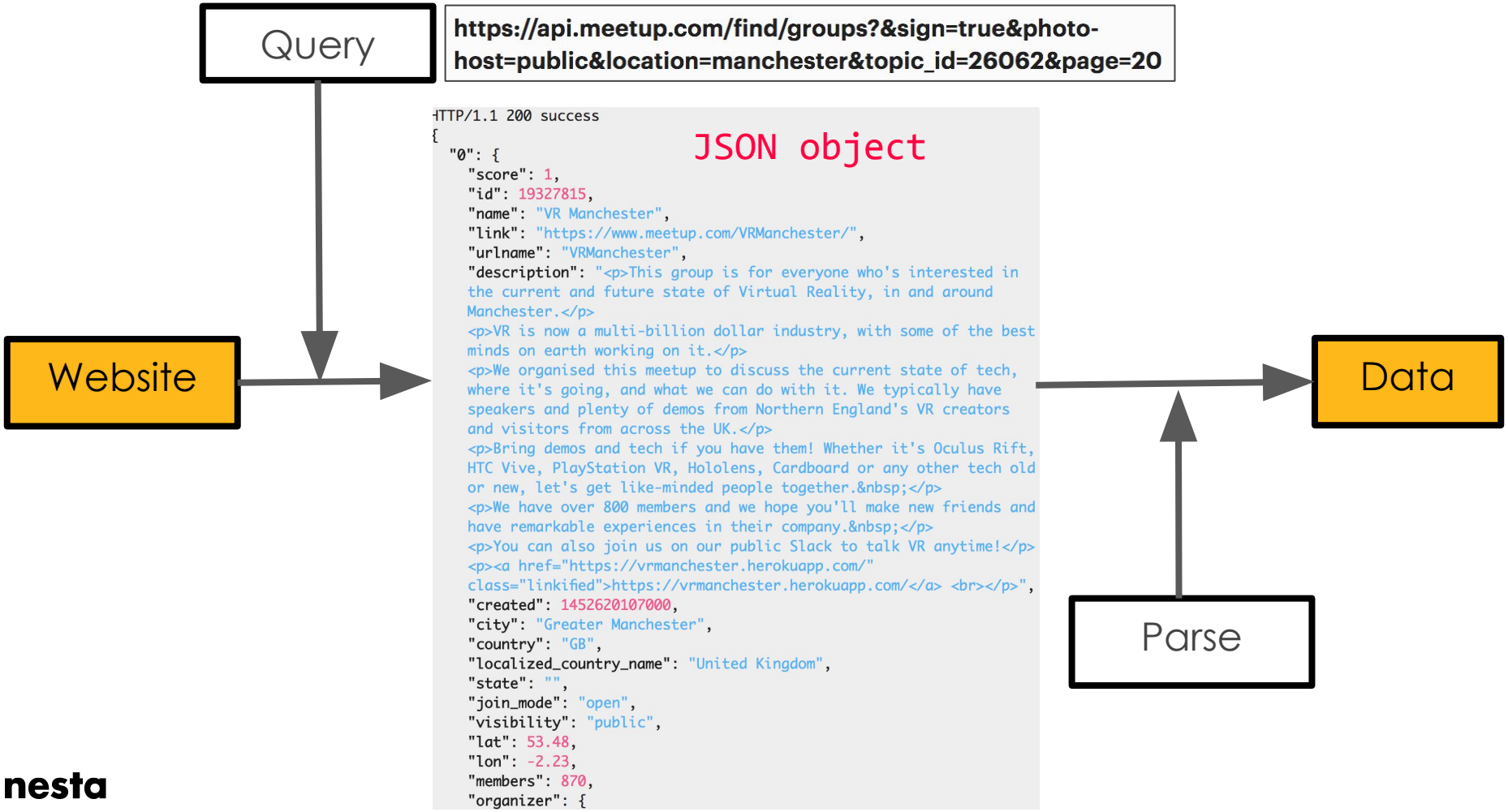
Data collection [1]: Downloads (the easy case!)





nesta

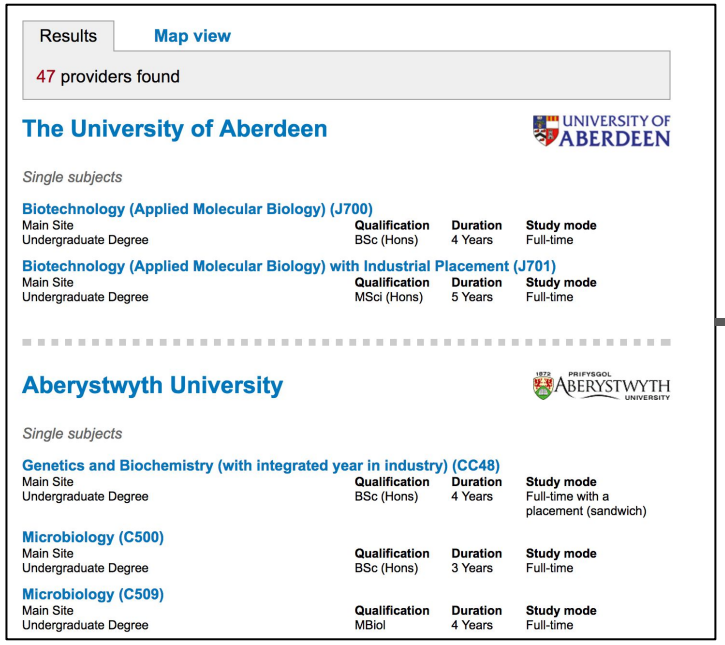
API = Application Programming Interface, a set of clearly defined methods of communication between various software components




```
HTTP/1.1 200 success
{
  "0": {
    "score": 1,
    "id": 19327815,
    "name": "VR Manchester",
    "link": "https://www.meetup.com/VRManchester/",
    "urlname": "VRManchester",
    "description": "<p>This group is for everyone who's interested in
    the current and future state of Virtual Reality, in and around
    Manchester.</p>
    <p>VR is now a multi-billion dollar industry, with some of the best
    minds on earth working on it.</p>
    <p>We organised this meetup to discuss the current state of tech,
    where it's going, and what we can do with it. We typically have
    speakers and plenty of demos from Northern England's VR creators
    and visitors from across the UK.</p>
    <p>Bring demos and tech if you have them! Whether it's Oculus Rift,
    HTC Vive, PlayStation VR, Hololens, Cardboard or any other tech old
    or new, let's get like-minded people together. </p>
    <p>We have over 800 members and we hope you'll make new friends and
    have remarkable experiences in their company. </p>
    <p>You can also join us on our public Slack to talk VR anytime!</p>
    <p><a href="https://vrmanchester.herokuapp.com/"
    class="linkified">https://vrmanchester.herokuapp.com/</a> <br></p>",
    "created": 1452620107000,
    "city": "Greater Manchester",
    "country": "GB",
    "localized_country_name": "United Kingdom",
    "state": "",
    "join_mode": "open",
    "visibility": "public",
    "lat": 53.48,
    "lon": -2.23,
    "members": 870,
    "organizer": {
```

Data collection [3]: Web scraping (the hard case!)

Web scraping = Automated downloading of data from a website



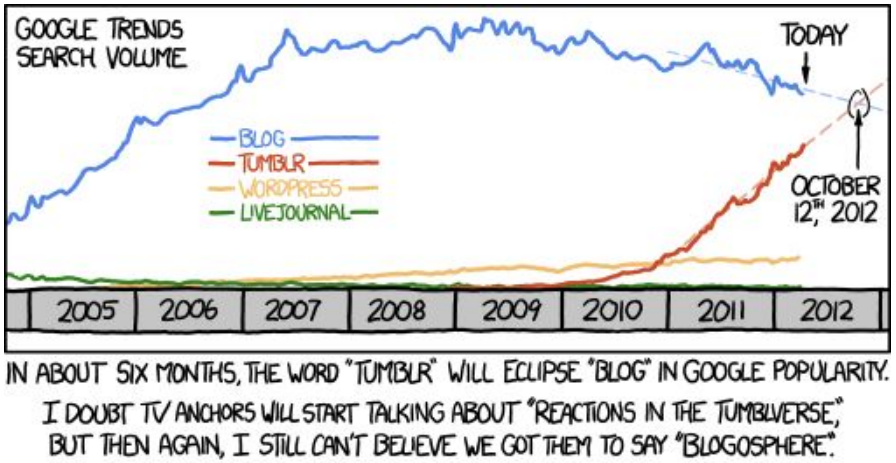
Download

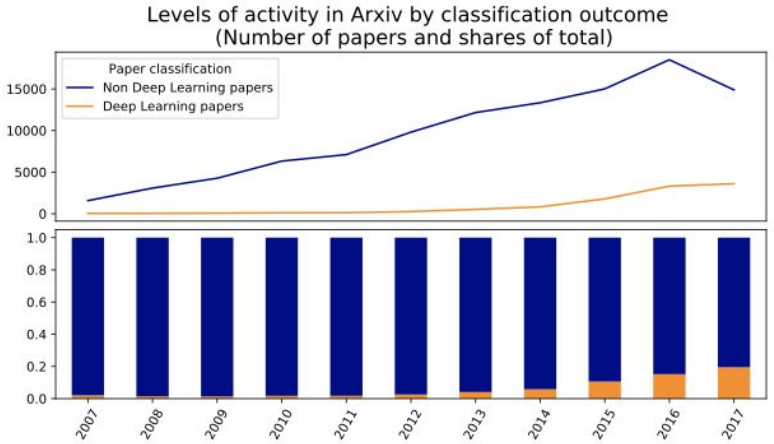
Parse

Data

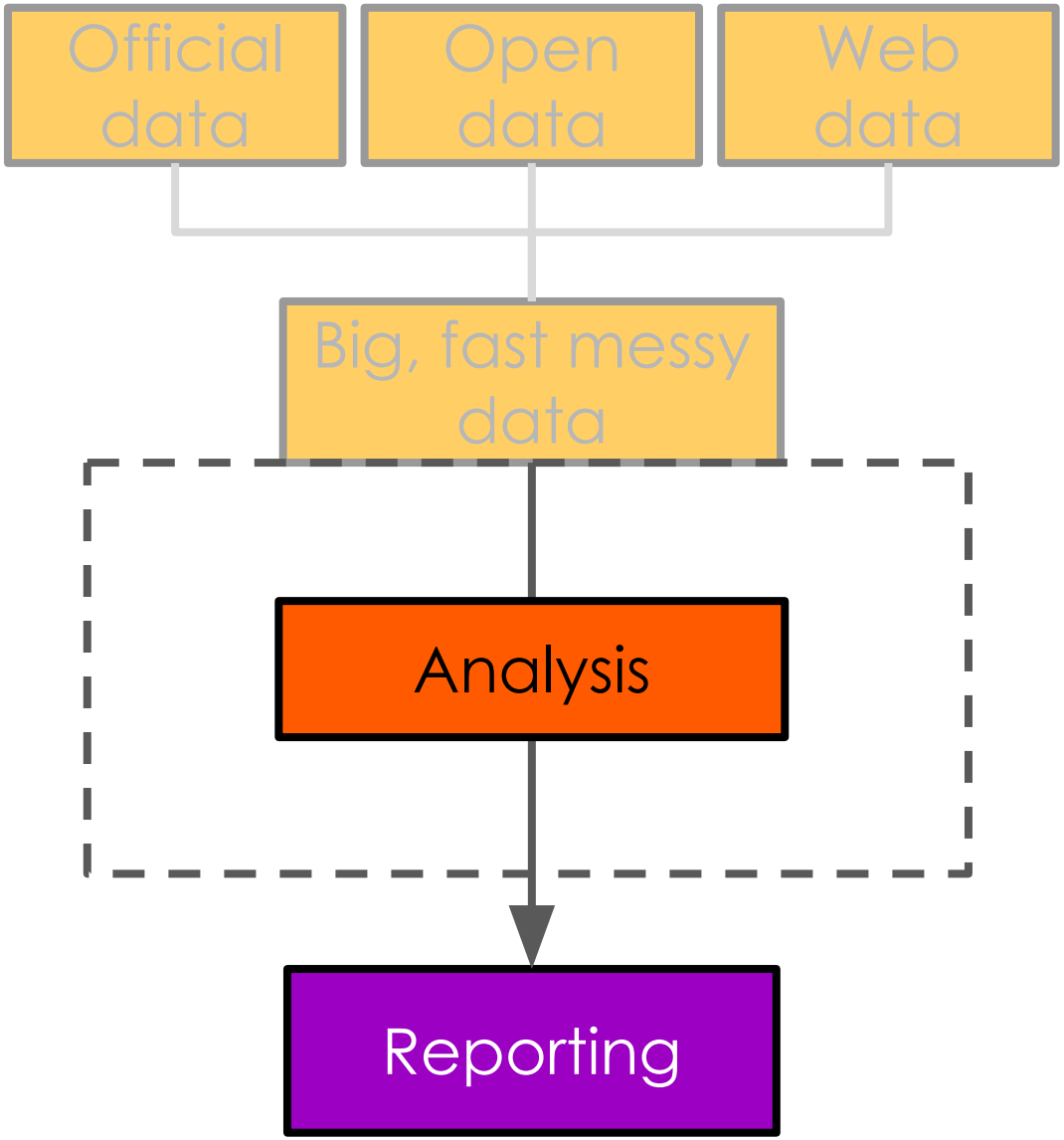
Things to watch out for

- Are you allowed to scrape a dataset?
- Are the users of the site you are studying representative?
- Are your data robust to freak user behaviours?
- Are you measuring changes in the phenomenon you are interested in, or changes in the design of the platform or its popularity?





Analysis of research trends in arXiv (a Computer Science pre-prints database: we need to normalise by total levels of activity to understand changes in activity).



Before analysing the data, you need to pre-process it. Remember that the data will very often be:

- Unstructured (no categories, no numbers)
- Incomplete
- Missing important information

Example: Mapping health innovation.

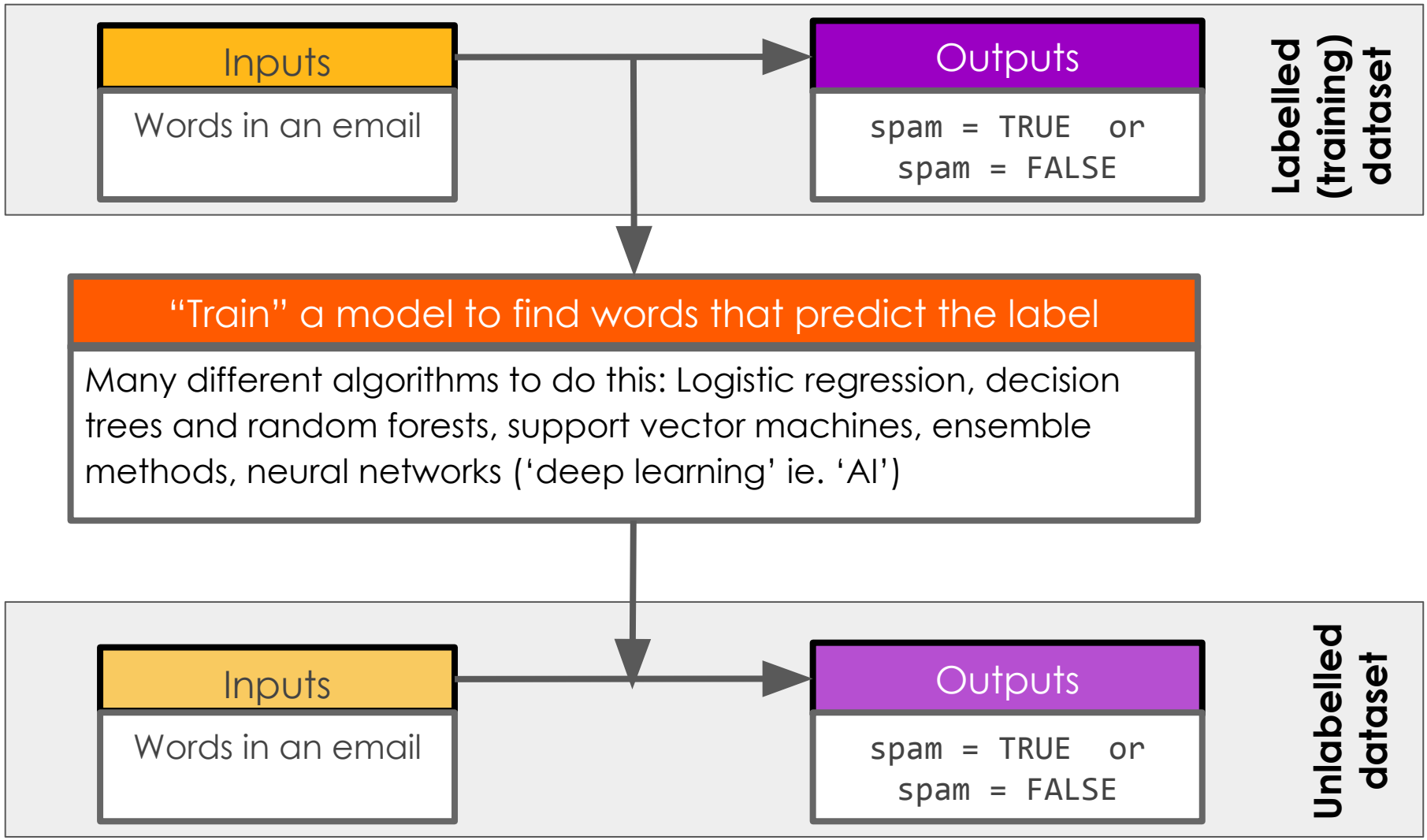
title	about	topics	address	website
Providing general operating support for the Bi...	NaN	NaN	Rockefeller Family Fund Inc.\n475 Riverside Dr...	http://www.rffund.org/
Supporting development of new strategies to ad...	To promote and protect the ability of local go...	[Social Determinants of Health]	Rockefeller Family Fund Inc.\n475 Riverside Dr...	http://www.rffund.org/
Spreading learning by RWJF grantees from the C...	To facilitate conversations between patients a...	[Public and Community Health, Health Care Cost...	Avalere Health, LLC\n1350 Connecticut Avenue, ...	http://www.avalerehealth.net/
Analyzing the work of the U.S. Preventive Serv...	To (1) examine how the U.S. Preventive Service...	[Public and Community Health, Disease Preventi...	New York University School of Medicine\n550 1s...	http://www.med.nyu.edu/
Supporting RWJF's Upstream Action Acceleration...	To support innovative community-based projects...	[Built Environment and Health]	Third Sector New England\n89 South Street, Sui...	http://www.tsne.org/
Examining whether providing patients with info...	NaN	[Health Coverage, Health Care Access, Disease ...	Indiana University\nBryan Hall 200\n107 South ...	http://www.indiana.edu/
Developing a digital resource to advance under...	To develop a digital resource for evaluation a...	[Health Care Quality, Built Environment and He...	Measured Lab\n77 East 110th Street, Suite 5B\n...	http://measured.design/
Continuing Healthy Marketplace Index metrics w...	To fund projects related to tracking health ca...	[Health Coverage, Health Care Cost and Value, ...	Health Care Cost Institute, Inc.\n1100 G Stree...	http://www.healthcostinstitute.org/
Supporting the New Jersey State Theatre Region...	To continue to support performance and educati...	[Health Disparities, Public and Community Health]	State Theatre Regional Arts Center at New Brun...	http://www.statetheatrenj.org/
Providing general operating support for the Bi...	To provide general operations support to the B...	NaN	Bipartisan Policy Center, Inc.\n1225 I Street,...	http://www.bipartisanpolicy.org/

- Can we predict missing labels in the data?
- Can we extract categories from text?
- Can we identify meaningful groups in the data?
- Can we enrich through further data collection and merging with other sources?

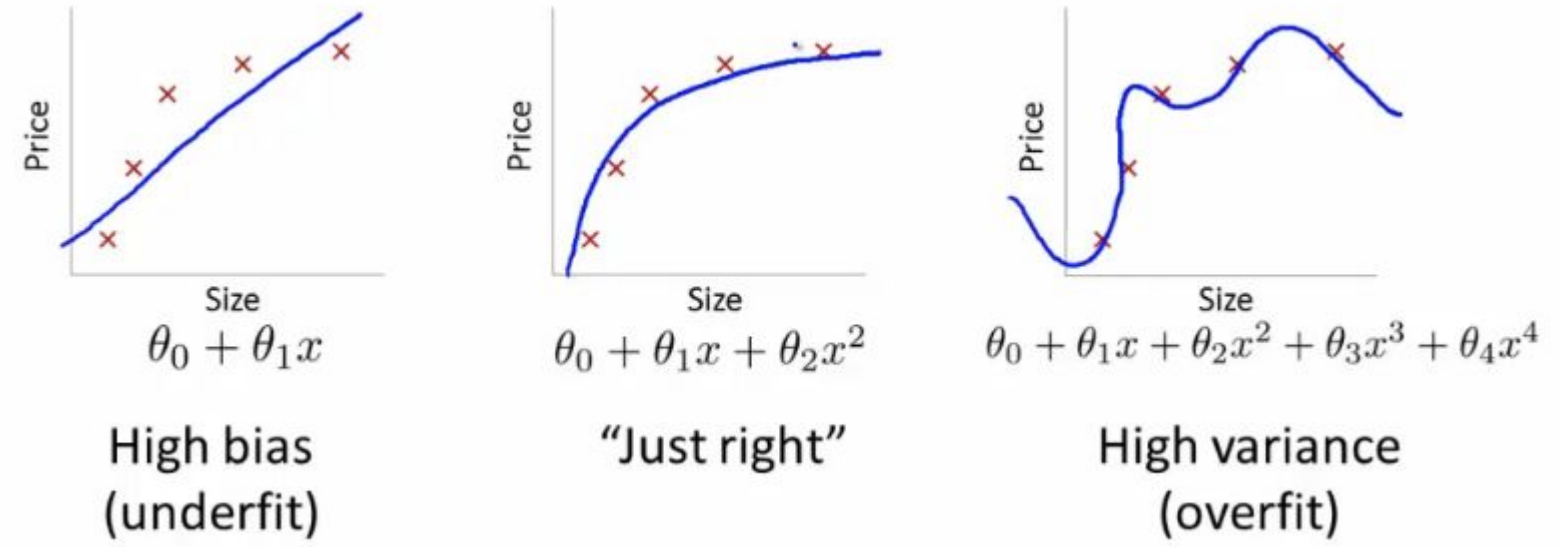
```
To develop a digital resource for evaluation and scaling the impact of social design on human health
by (1) looking across sectors and cataloging its many definitions, forms, and approaches in
social-change efforts directly affecting health; (2) tracking its history and capturing case studies
to explain how different sectors are using social design and the conditions for success; (3)
developing consistent definitions and terminology to establish shared language and concepts; (4)...
```

nesta

Learning from examples: Supervised learning



The best model minimises some loss function (measure of prediction/classification error) in the training set. An important risk is that the model learns to predict noise in the training set but does not generalise outside.



Source: https://datascience.stackexchange.com/questions/361/when-is-a-model-underfitted

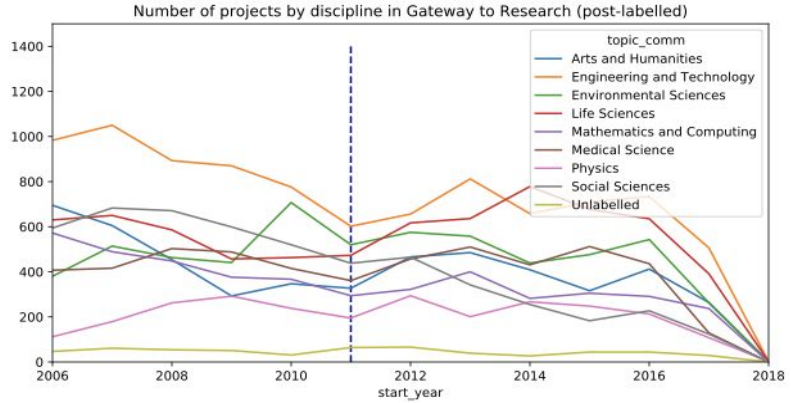
This is addresses through regularisation (penalise complex models) and cross-validation (choose the models that generalise better across subsets ('folds') of the training set).

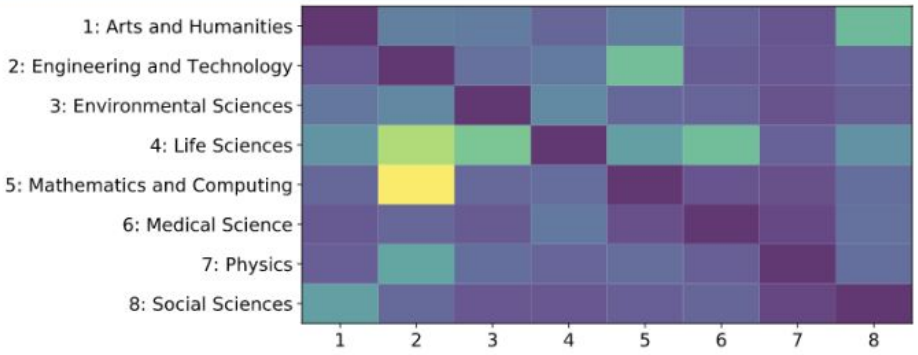
We have used supervised ML to:

- Predict missing labels in open data (Analysis of Research Council open data, Mateos-Garcia and Stathoulopoulos, 2017)
- Expand survey responses from sample to population (Immersive economy analysis)

Other interesting examples:

- Nowcasting and placecasting of entrepreneurial activity in the USA (Guzman and Stern, 2016).



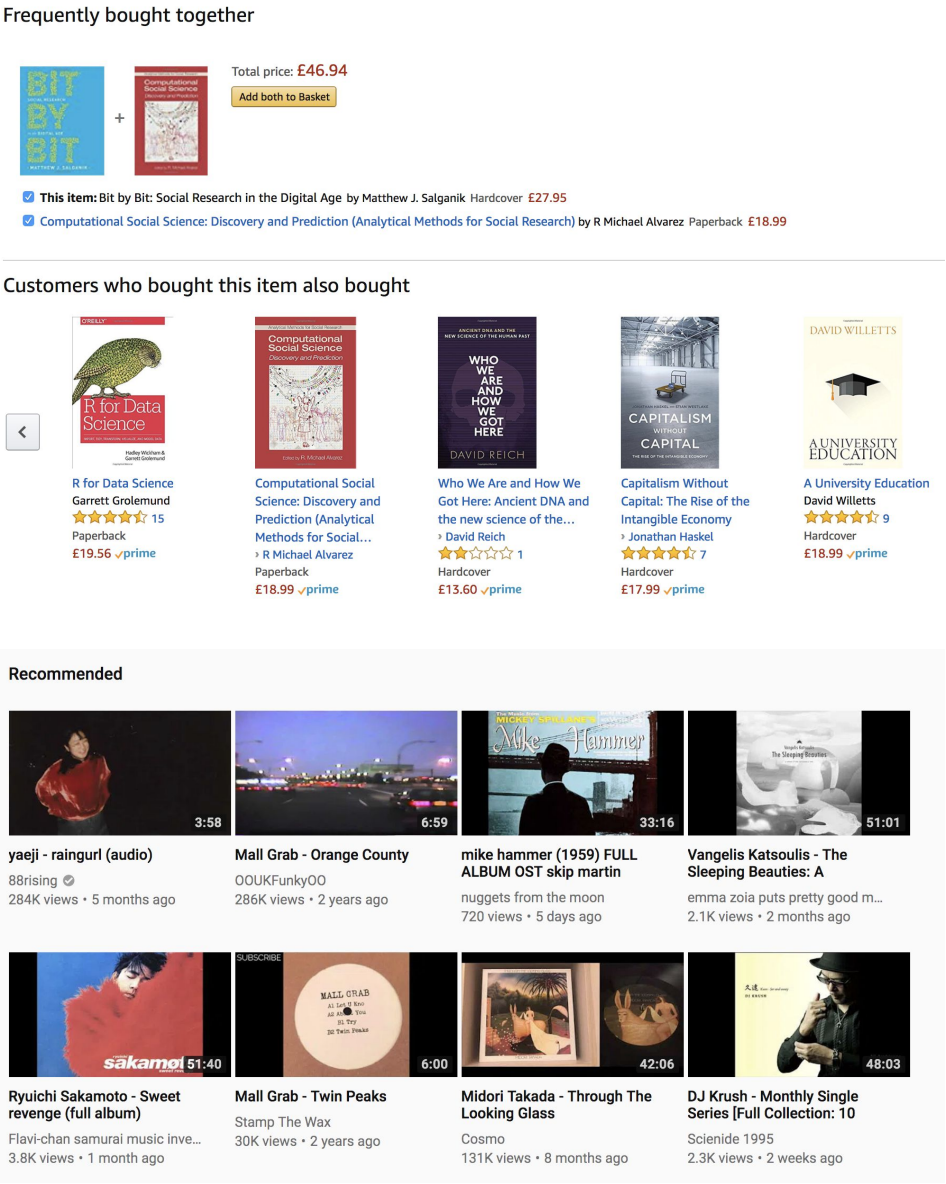


BSRC and MSRC did not label their data over the whole period. We used ML to back-cast the data and identify 'crossover' projects (with high probabilities in multiple disciplines).

Supervised ML is big business

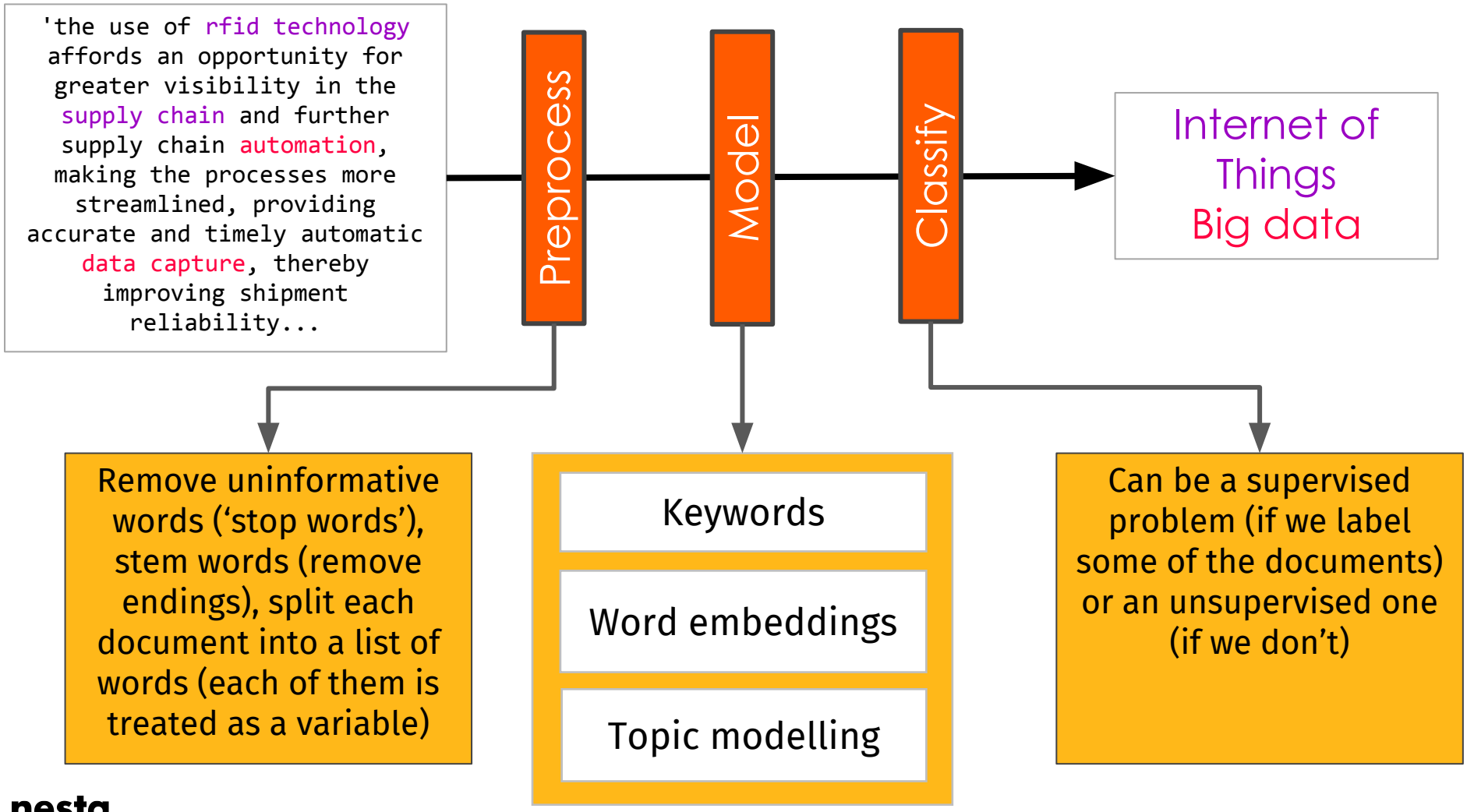
Supervised Machine Learning is very useful whenever you want to make a prediction but don't care so much about explaining: you want predictive knowledge instead of causal knowledge.

Not often the case in social science settings but can be importance in policy (although there is growing demand for interpretable models).



nesta

Can we classify research papers, content in business websites, text in job ads, into relevant categories?



Quick primer on NLP algorithms

Keywords

```
dig_terms = ['app','digital','software','computer','ict','web','internet']

_data = rwj_has_ab

_data.loc[:,'dig_n'] = [len(set(dig_terms) & set(x)) for x in _data.about_tokenised]

_data.sort_values('dig_n',ascending=False)[['title','dig_n']].head(n=10)
```

	title	dig_n
14945	Developing a mobile phone and Internet applica...	4
2483	Educational research into the effectiveness of...	3
5861	Expanding mothers' advocacy and support group ...	3
11206	Measuring the public health impact of increase...	2
7865	Web site redevelopment for the Dartmouth Atlas...	2

- Nice and easy
- Requires an initial vocabulary.
- Risk of low recall if we are missing important words. Only captures the categories we already know about.

Word embeddings

```
w2v = models.Word2Vec(rwj_has_ab.about_tokenised)

w2v.wv.most_similar('digital')

[('multimedia', 0.808357298374176),
 ('marketing', 0.7946643829345703),
 ('messages', 0.7729908227920532),
 ('interactive', 0.7460197806358337),
 ('message', 0.7453526258468628),
 ('eshe', 0.7391177415847778),
 ('food_beverage', 0.7251322269439697),
 ('messaging', 0.7122663259506226),
 ('distribution', 0.7024813294410706),
 ('content', 0.7019100189208984)]
```

- Represents words as vectors based on their semantic similarity and measures their distance.
- Can be used to augment an initial vocabulary.
- Can be used to estimate document similarity (useful for clustering)

Topic modelling

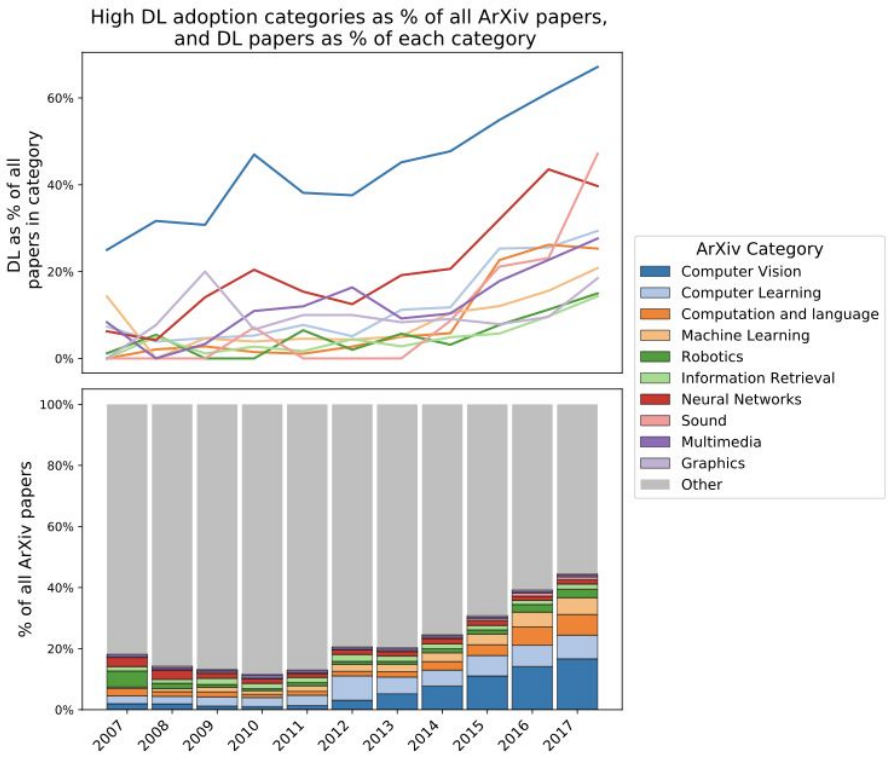
```
test_norm.lda_topics[2:5]

[(2,
  '0.075*"services" + 0.058*"community" + 0.039*"health" + 0.034*"project" + 0.028*"coalition" + 0.022*"conditions" + 0.020*"care" + 0.020*"develop" + 0.020*"foundations" + 0.019*"designed"'),
 (3,
  '0.082*"evaluation" + 0.024*"innovations" + 0.024*"innovation" + 0.022*"online" + 0.018*"open" + 0.018*"sustainability" + 0.017*"plan" + 0.015*"implementation" + 0.015*"development" + 0.012*"organizational"'),
 (4,
  '0.072*"state" + 0.048*"states" + 0.046*"coverage" + 0.033*"health" + 0.031*"medicaid" + 0.029*"insurance" + 0.025*"care" + 0.018*"enrollment" + 0.015*"federal" + 0.012*"access"')]
```

- Generate documents with a probabilistic model.
- Returns k ‘topics’ (clusters of keywords) and a probabilities over all topics for each document
- Covers all the data
- Hard to interpret, noisy

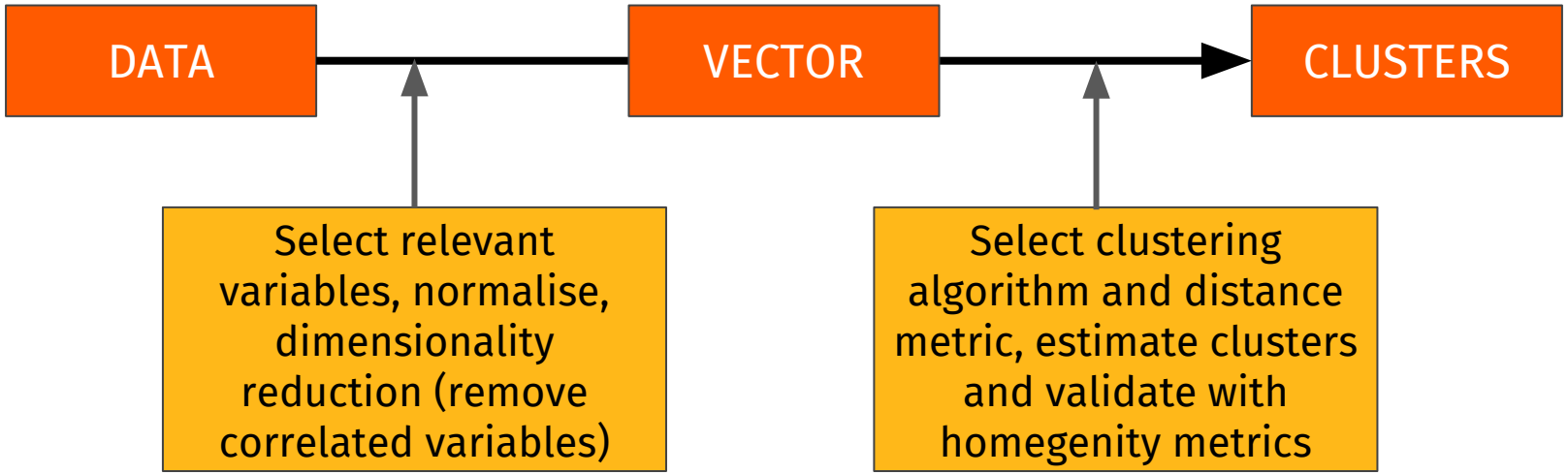
We use NLP all the time.

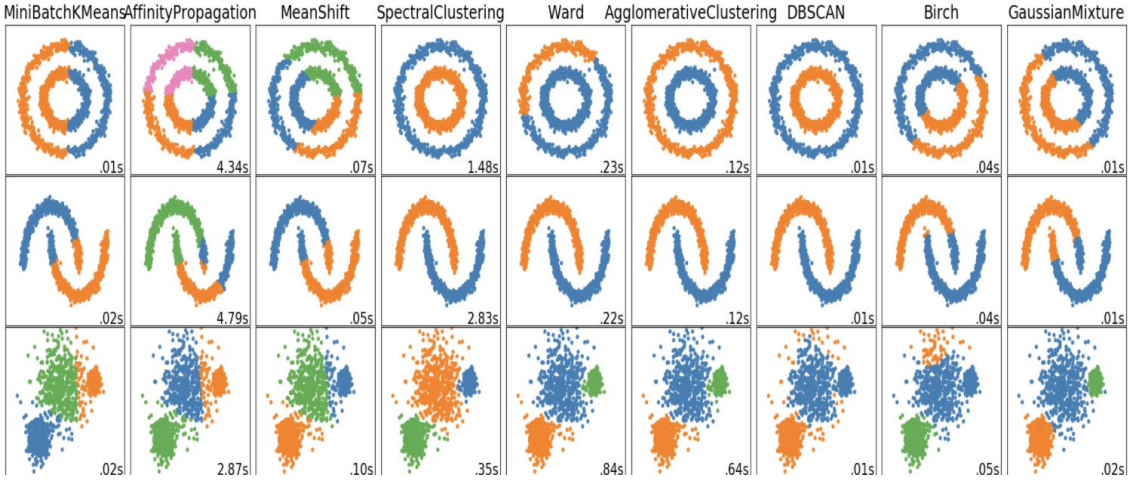
- Use topic modelling to detect Deep Learning papers (a flavour of AI) in a corpus of pre-prints and analyse their evolution and geography
- We have written a search engine ('Clio') that queries funding databases and returns lists of projects related to an initial query. We are using this approach in multiple projects...



This chart shows levels of 'diffusion' of deep learning in different sub-disciplines of Computer Science (Klinger et al, 2018).

We can also use unsupervised learning (clustering) to identify groups of similar observations in the data.

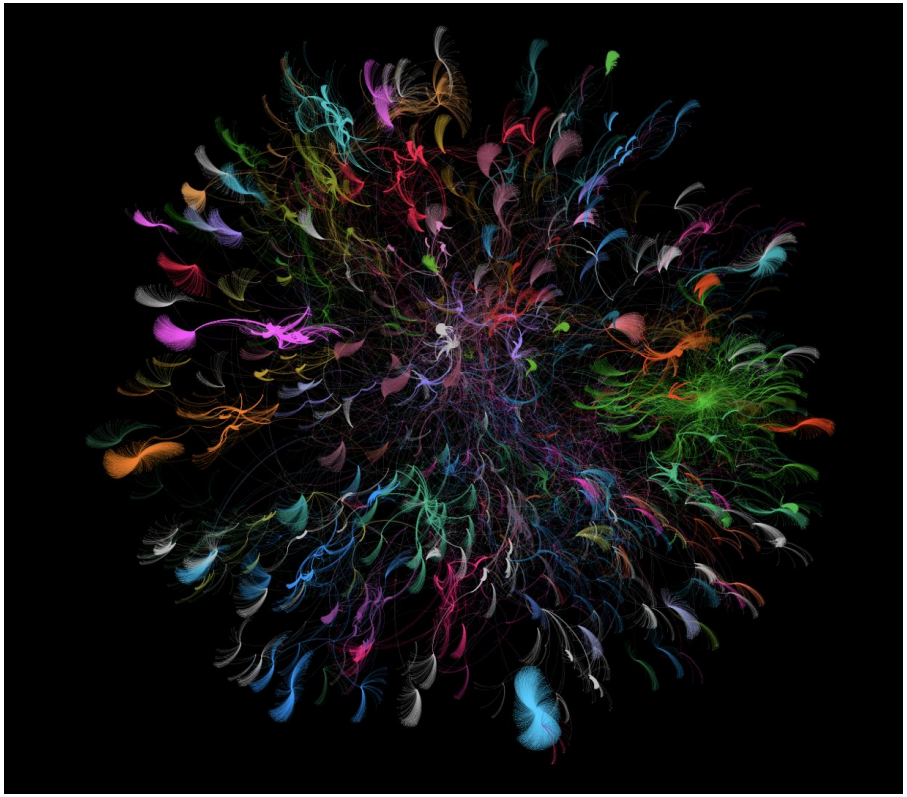




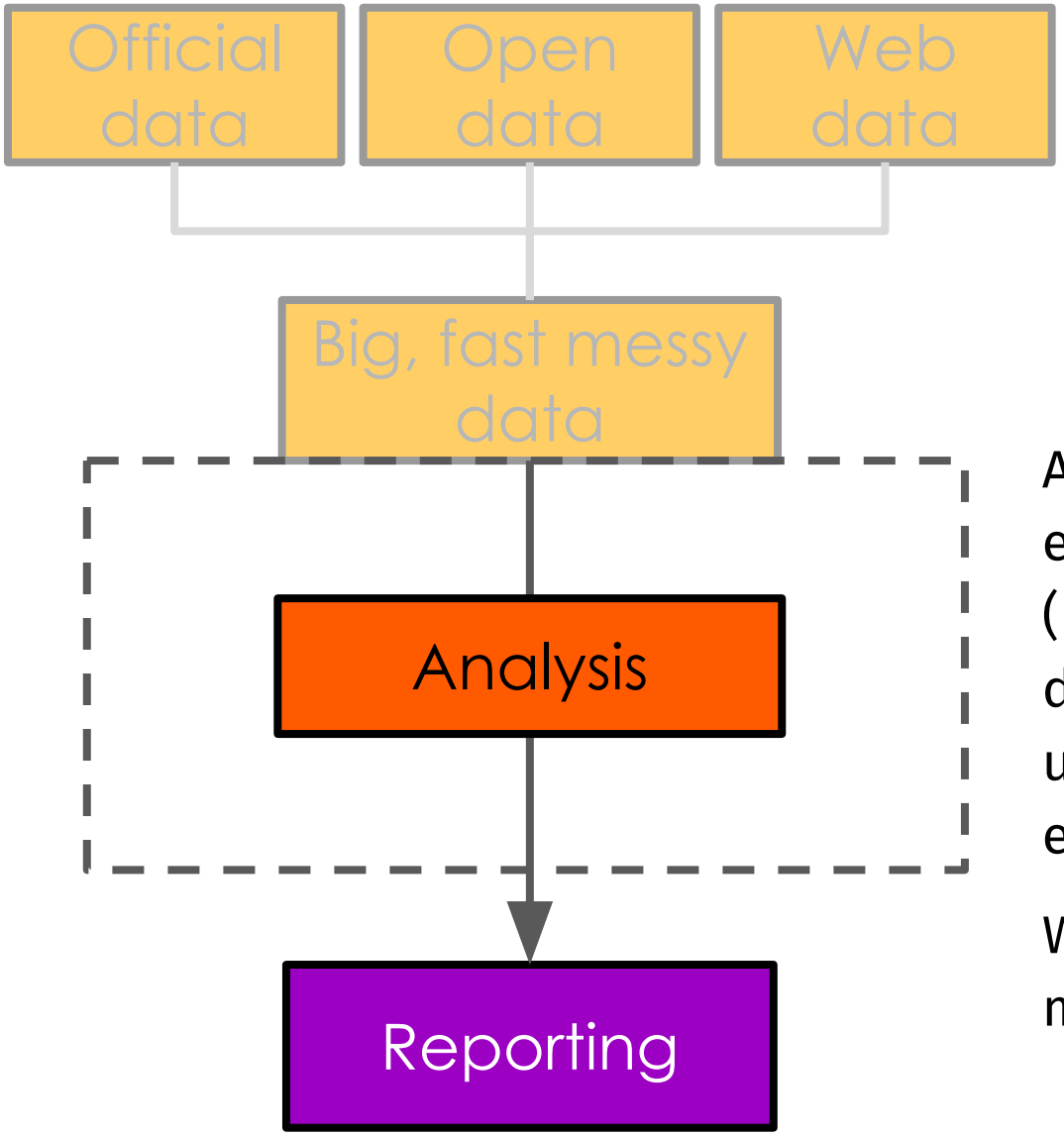
Source: http://scikit-learn.org/stable/modules/clustering.html

Clusters inside networks

- Community detection is another flavour of clustering: the algorithm looks for tightly connected groups in network structures.
- Many of the methods we discussed before give you measures of similarity / distance you can represent as networks.
- All these methods interconnect in interesting and creative ways



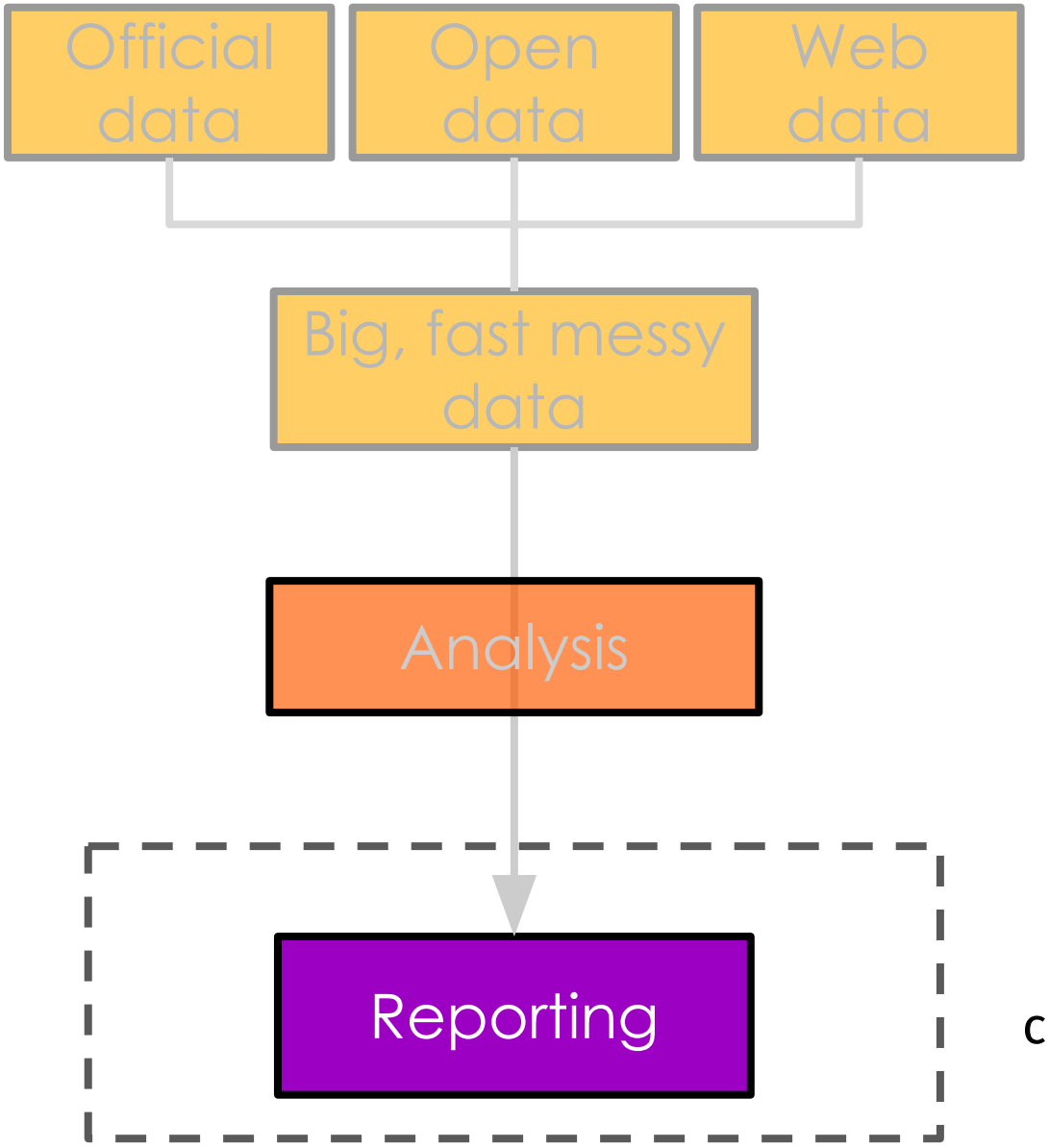
Representation of a university website: connections are links between pages, and colours are communities. We are using this in an analysis of skills supply from universities for the EU (Klinger, forthcoming).



After pre-processing and enriching the data, you (hopefully) end with a novel dataset you can analyse using statistics, econometrics etc.

We won't focus on those methods here.

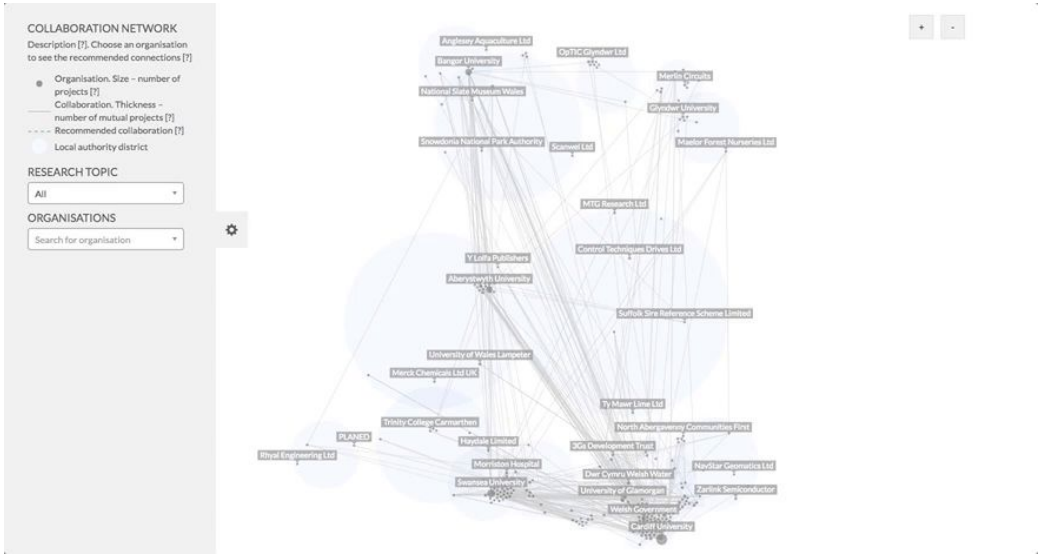
...and reporting



DS4STIR also multiplies channels for dissemination

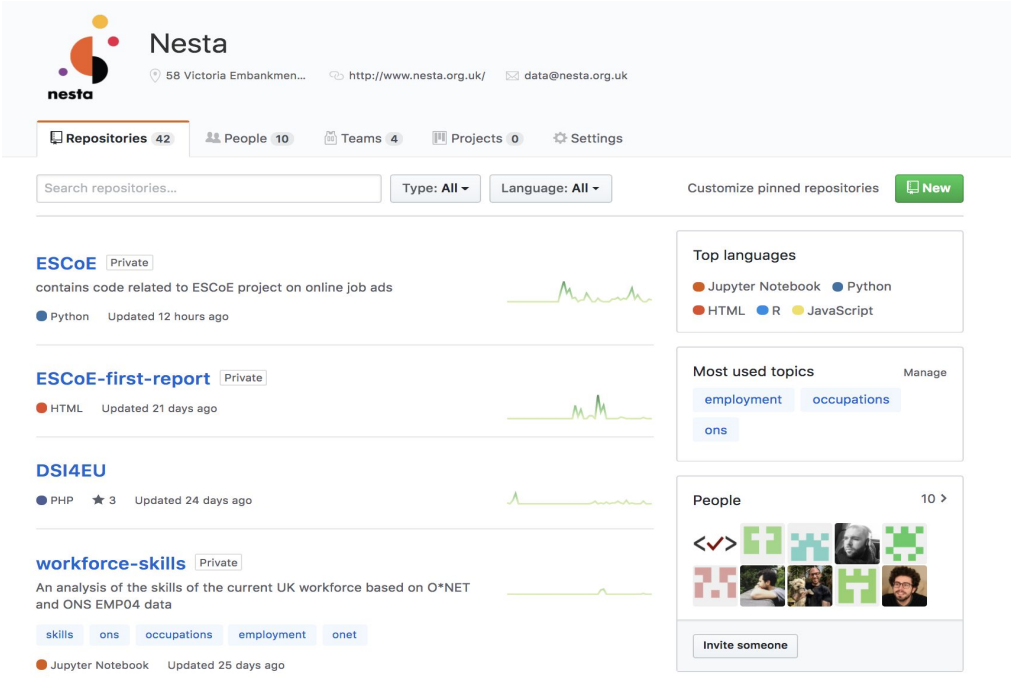
Interactive visualisations and tools

Lower barriers to access the data and enable users to discover new insights, democratising the STIR process



Open source code, data and notebooks

Reproducibility is critical. Make data and code openly available for others to review and use.



Things to watch out for

- Is the labelled dataset representative of the unlabelled dataset? For how long?
- Do the results make sense?
- Is the model good for explaining as well as predicting?
- Have you considered ethical / Intellectual Property issues from sharing data ?
- Have you user-tested and documented your outputs?

XKCD cartoon

How can you get started with DS4STIR?

(if you haven't already)

nesta

If you want to do DS4STIR, forget Stata, Matlab, SPSS...open source software is the way to go.

- Free as in beer and speech
- Active development, with thousands of packages in CRAN / Pip / Anaconda
- Stand on the shoulders of tech / research giants...
- ...and let others stand on yours. Anyone can reproduce your work and build on your code

The difference that matters

This technical engineering dimension is not the only one we should use to compare the proprietary and open models. There is an independent social dimension, where the metrics assess the interactions between people. Does it increase trust? Does it increase the importance that people attach to a reputation for integrity?

It is along this social dimension that open source unambiguously dominates the proprietary model. Moreover, at a time when trust and truth are in retreat, the social dimension is the one that matters.

Jupyter rewards transparency; Mathematica rationalizes secrecy. Jupyter encourages individual integrity; Mathematica lets individuals hide behind corporate evasion. Jupyter exemplifies the social systems that emerged from the Scientific Revolution and the Enlightenment, systems that make it possible for people to cooperate by committing to objective truth; Mathematica exemplifies the horde of new Vandals whose pursuit of private gain threatens a far greater pubic loss—the collapse of social systems that took centuries to build.

https://paulromer.net/jupyter-mathematica-and-the-future-of-the-research-paper/

Where to learn?

MOOCs and tutorials to get over the initial learning curve / cliff

- Coursera
- Datacamp
- Udacity
- New Microsoft AI MOOCs

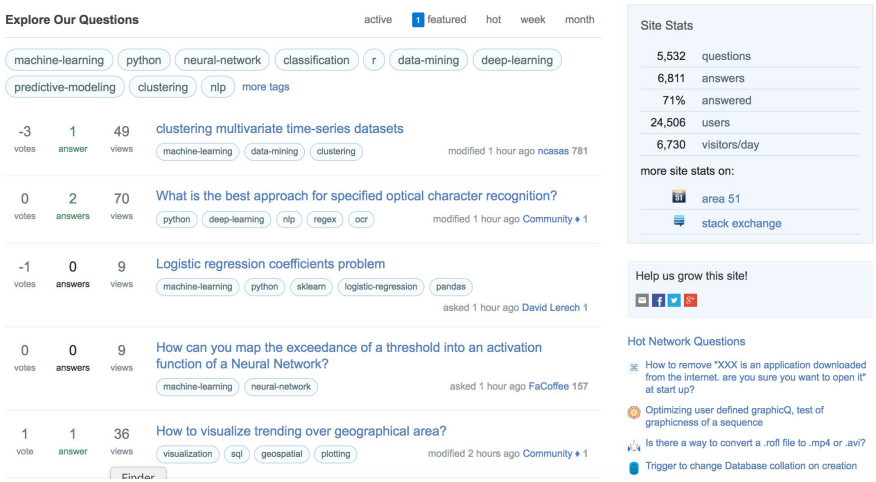
Open datasets and competition sites to practice with

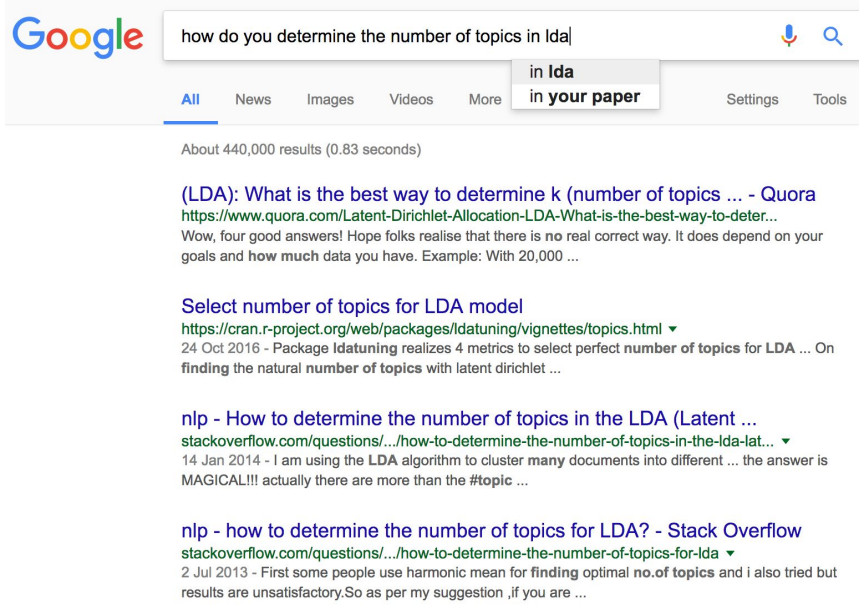
- Kaggle

Q&A sites for when you get stuck

- Stack Overflow
- Quora
- Google

Or e-mail us!





nesta

Why bother?


Highly sought after portable skills
More and better knowledge about STIR
New opportunities to have an impact & change the world

nesta.org.uk

@nesta_uk

Juan.mateos-garcia@nesta.org.uk
@Jmateosgarcia