nesta

# From text to impact in 90 minutes

## Practical steps for using semantic technologies to inform innovation policy

**Joel Klinger, Juan Mateos-Garcia and Chantale Tippett**

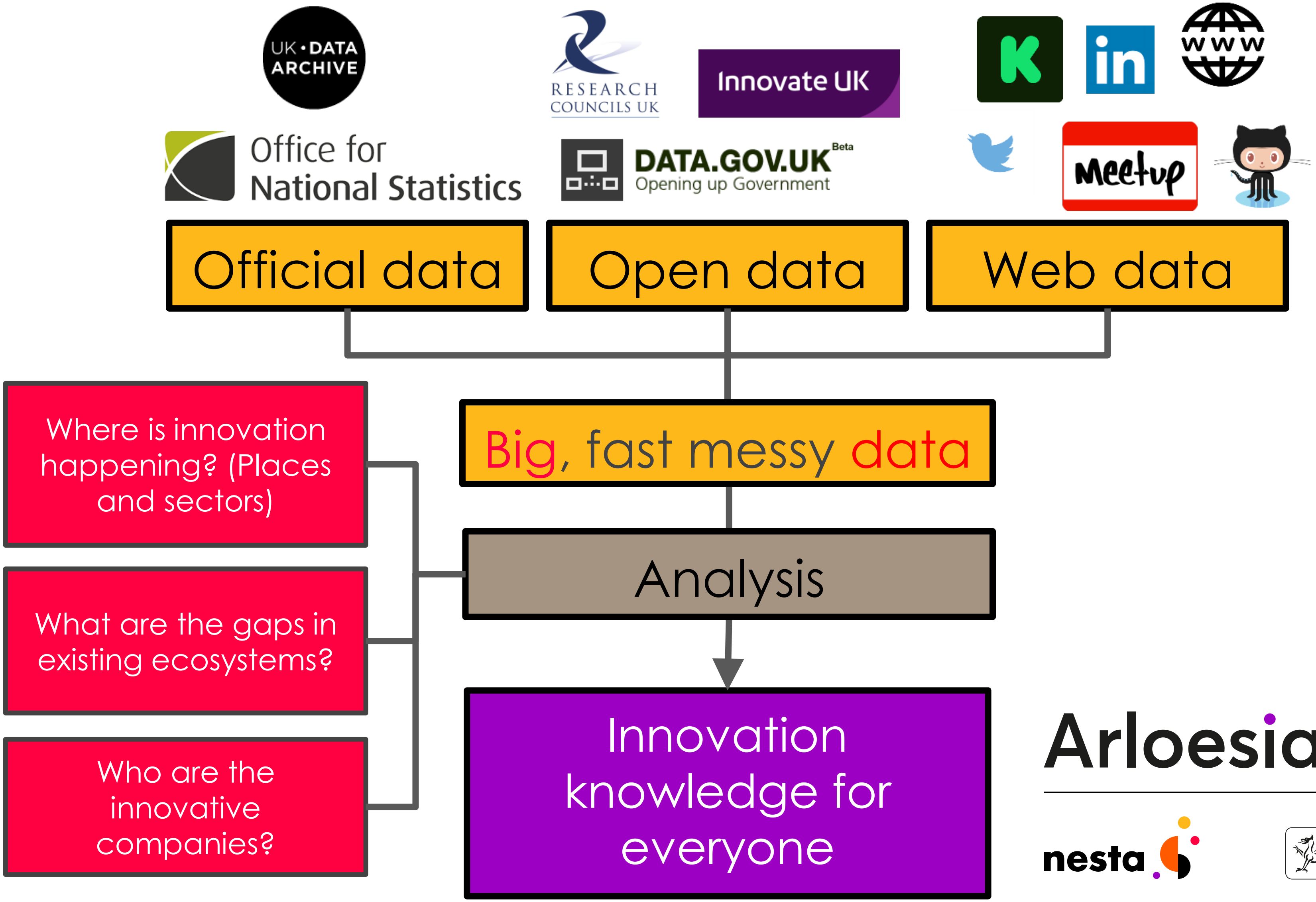
Paris

12 March 2018

## Objectives

To come away with an understanding of the steps, challenges and opportunities in the semantic analysis pipeline, as well as ideas on how you can apply it in your own context

# Motivation


UK • DATA ARCHIVE
RESEARCH COUNCILS UK
Innovate UK
Office for National Statistics
DATA.GOV.UK Beta Opening up Government
K
in
WWW
Meetup

Official data

Open data

Web data

Big, fast messy data

Analysis

Innovation knowledge for everyone

Where is innovation happening? (Places and sectors)

What are the gaps in existing ecosystems?

Who are the innovative companies?

Arloesiadur.

nesta

Ariennir yn Rhannol gan Lywodraeth Cymru
Part Funded by Welsh Government

# Step 1. Collecting data

UK • DATA ARCHIVE
Office for National Statistics
RESEARCH COUNCILS UK
Innovate UK
DATA.GOV.UK Beta
Opening up Government
www
Meetup

Official data

Open data

Web data

We need to identify data which address an innovation policy or problem, and access it. This might involve:

- Downloading a dataset from a website
- Working with an API
- Scraping websites!

# Step 1. Collecting data [examples]

- Gateway to Research: UK Research council-funded projects
- CORDIS: EU Framework Programme projects
- Federal RePORTER: US Science funding
- Ploteus: EU Learning opportunities and qualifications…
- CrunchBase: Tech companies
- Meetup: Tech networking events

…

# Step 1. Collecting data [exercise]

Identify an interesting dataset to achieve an innovation policy impact [or look at one of the above]

## Where

- Where are the data?

## What for?

- What is the policy / practical policy application?

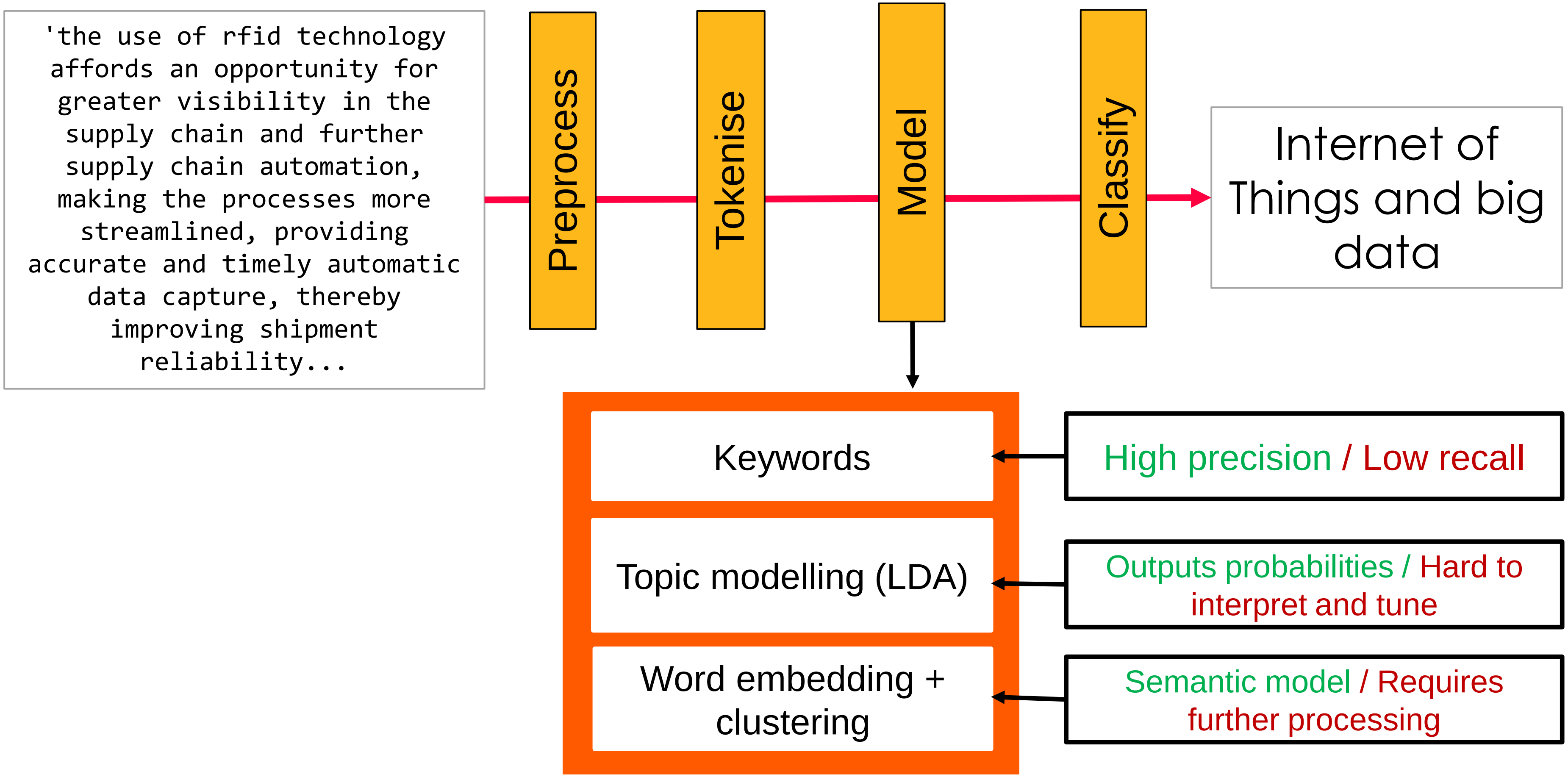
## How?

- How would you access them?
- What information do they contain?
- What is their coverage (time / sector)?
- Do you need to enrich them in some way?

**15 minutes**

# Step 2: Analysis

How do we transform textual information into policy-relevant categories (industries, technologies...)?



'the use of rfid technology affords an opportunity for greater visibility in the supply chain and further supply chain automation, making the processes more streamlined, providing accurate and timely automatic data capture, thereby improving shipment reliability...

Preprocess

Tokenise

Model

Classify

Internet of Things and big data

Keywords

Topic modelling (LDA)

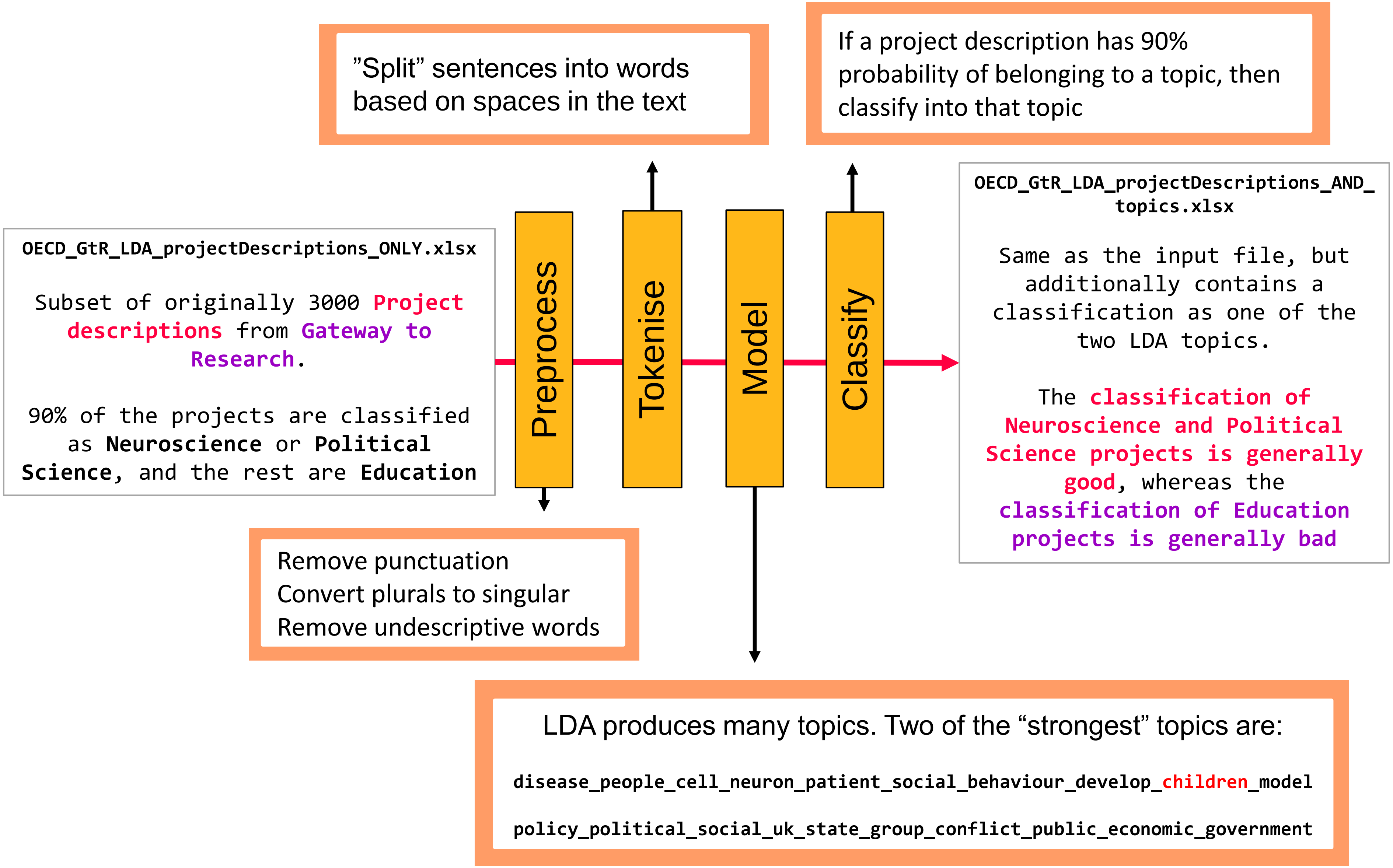
Word embedding + clustering

High precision / Low recall

Outputs probabilities / Hard to interpret and tune

Semantic model / Requires further processing

# Step 2: Analysis [Practical example]



**OECD_GtR_LDA_projectDescriptions_ONLY.xlsx**

Subset of originally 3000 **Project descriptions** from **Gateway to Research**.

90% of the projects are classified as **Neuroscience** or **Political Science**, and the rest are **Education**

Remove punctuation
Convert plurals to singular
Remove undescriptive words

”Split” sentences into words based on spaces in the text

LDA produces many topics. Two of the “strongest” topics are:

**disease_people_cell_neuron_patient_social_behaviour_develop_children_model**

**policy_political_social_uk_state_group_conflict_public_economic_government**

If a project description has 90% probability of belonging to a topic, then classify into that topic

**OECD_GtR_LDA_projectDescriptions_AND_topics.xlsx**

Same as the input file, but additionally contains a classification as one of the two LDA topics.

The **classification of Neuroscience and Political Science projects is generally good**, whereas the **classification of Education projects is generally bad**

# Step 2: Analysis [Exercise]

## Look at the data files we provided

### Raw data

OECD_GtR_LDA_projectDescriptions_ONLY.xlsx

- How would you identify a topic of interest in it?
- How would you classify projects into the right categories?
- What are some of the key challenges?

### Modelled data

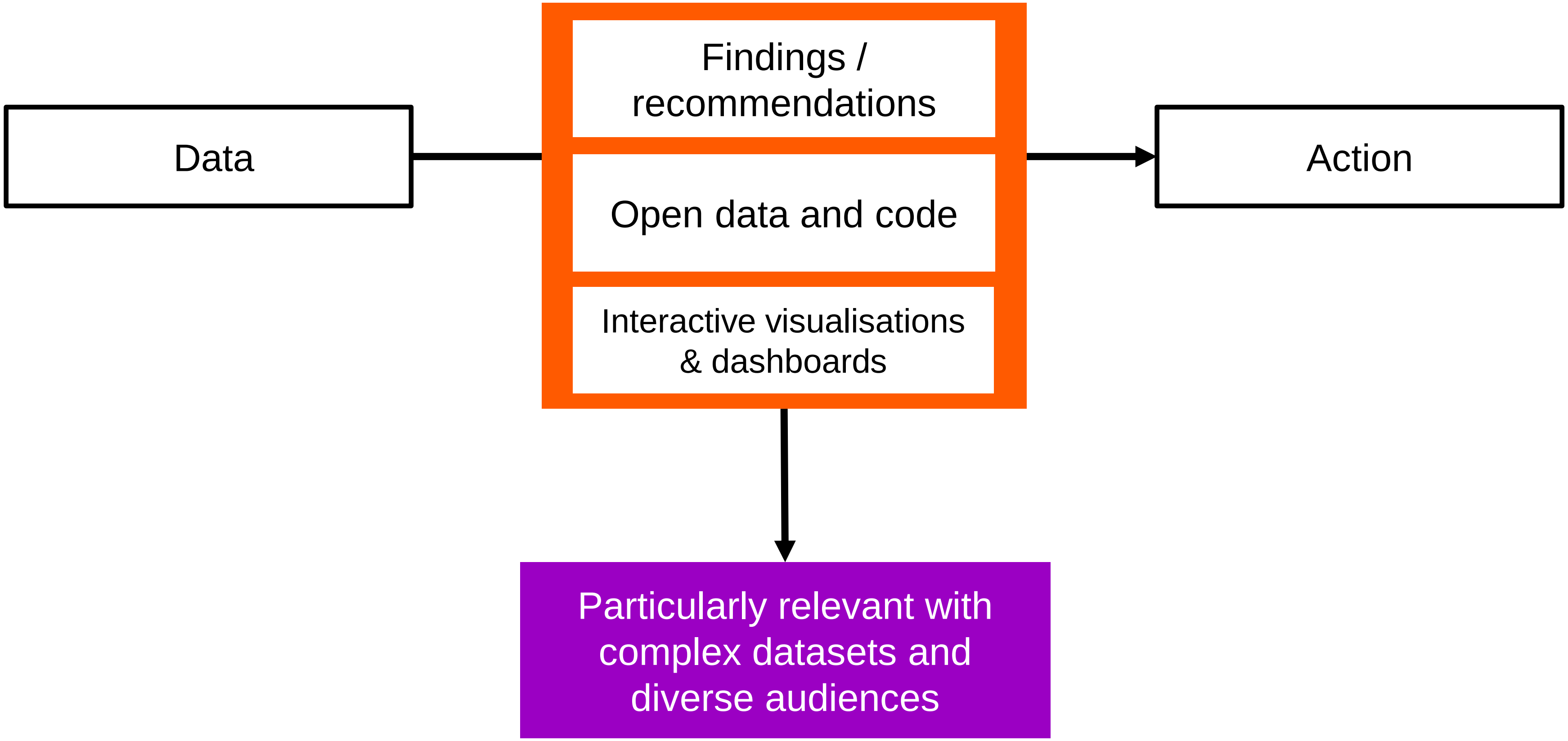
OECD_GtR_LDA_projectDescriptions_AND_topics.xlsx

- How do we go from model outputs to findings?
- What will be some of the challenges doing this?

15 minutes

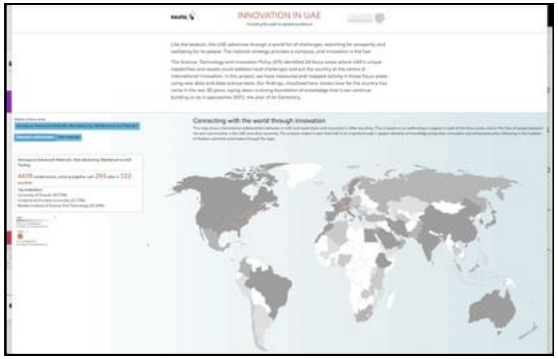
# Step 3: Communicating the results

How can we present our data to inform better action?


Data
Findings / recommendations
Open data and code
Interactive visualisations & dashboards
Action
Particularly relevant with complex datasets and diverse audiences

# Step 3. Communicating the results [options]

## Narrative

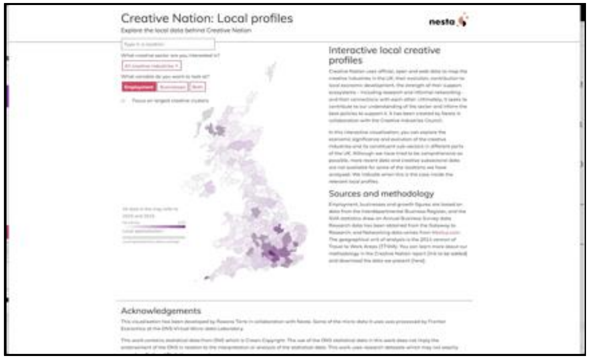

nesta

INNOVATION IN UAE

Connecting with the world through innovation

- Tell a story
- ✓ Easier to use
- ✓ Easier to create
- □ Harder to explore
- □ Short shelf-life

## Synthetic


Creative Nation: Local profiles

nesta

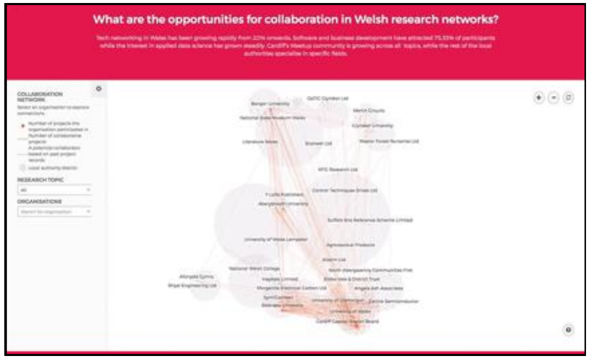
Interactive local creative profiles

Sources and methodology

Acknowledgements

- Summarise a situation
- Rich picture
- Benchmarking
- □ Data overload

## Exploratory


What are the opportunities for collaboration in Welsh research networks?

COLLABORATION NETWORK

RESEARCH TOPIC

ORGANISATIONS

- Explore a system
- Discover new patterns and actors
- □ Hardest to develop
- □ Privacy?

Best approach depends on data source / policy need: We need to discover them!

Also decide how to integrate in a single site.

# Step 3. Communicating the results [examples]

- Arloesiadur: Analysis of various datasets about innovation in Wales
- Creative Nation: Dashboard combining multiple sources
- OEC: Visualising economic complexity with trade data
- Startup Cartography project:  Maps of US startup based on predictive analytics.

…

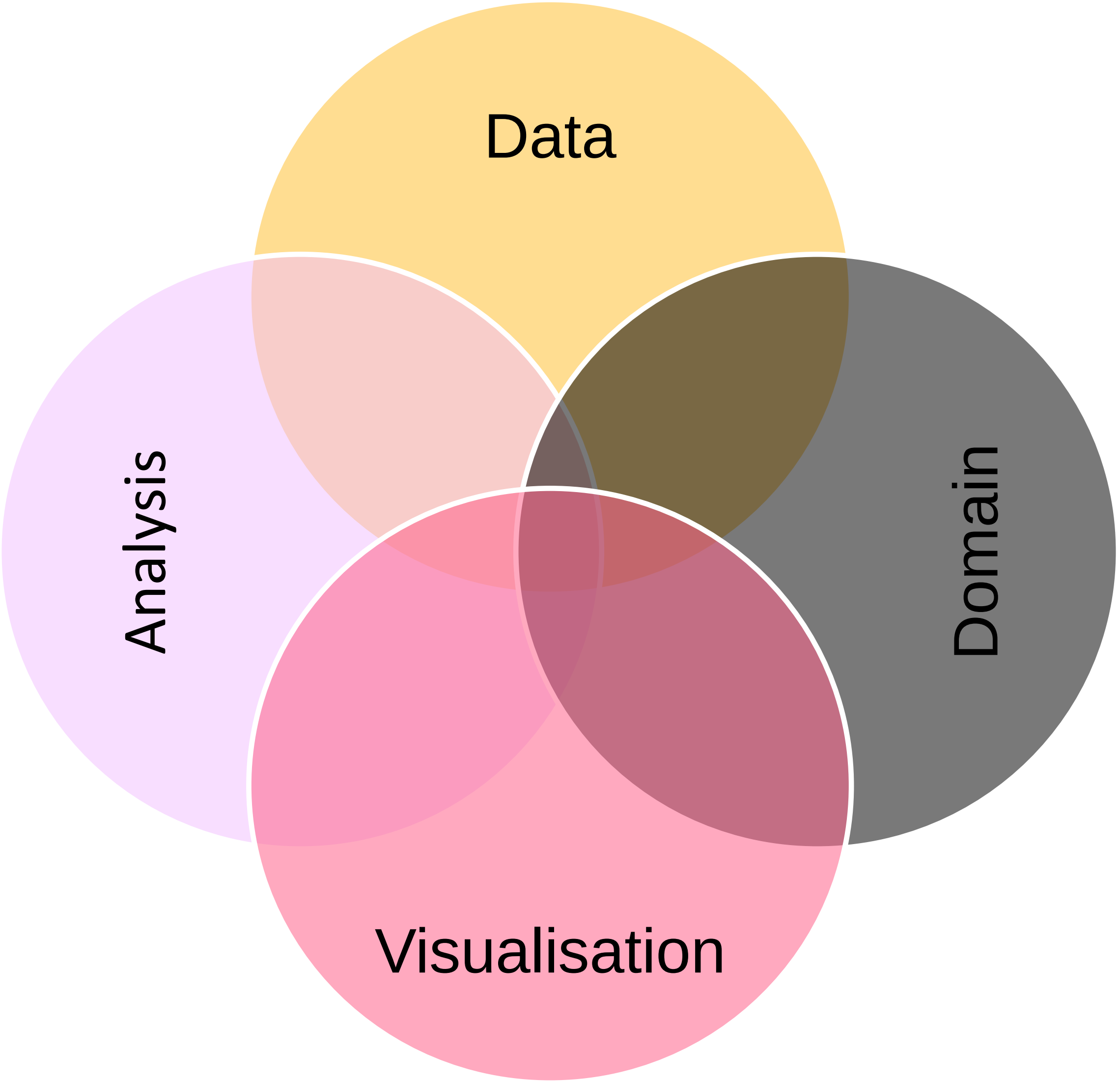
# Step 3. Communicating the results [Exercise]

## Thinking about your policy applications and data

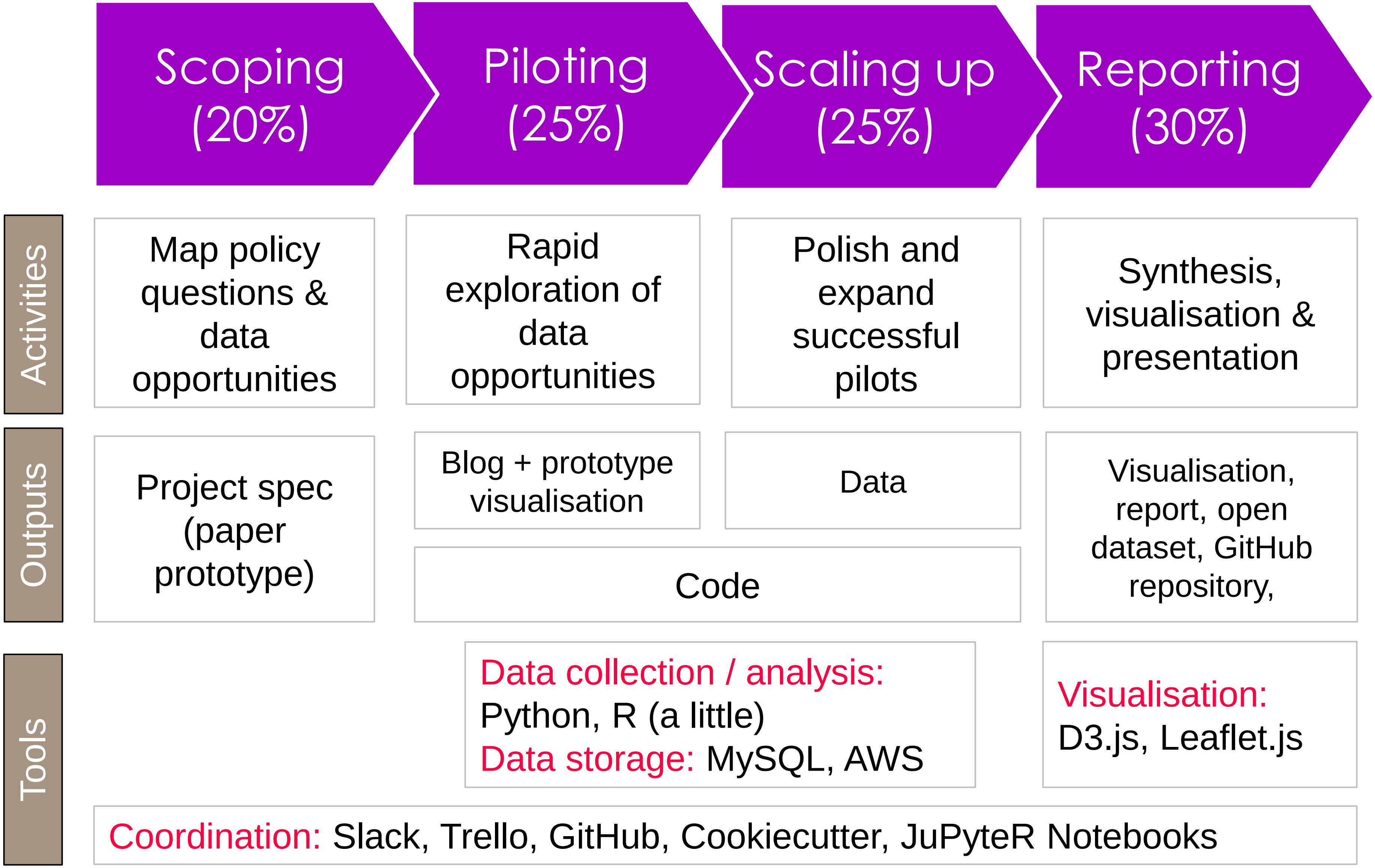
- Who are your audiences?
- What types of data formats would be more relevant to them? (think beyond tables!)
- How would you create these formats?
- What are some of the challenges you would expect to face as you do this?

15 minutes

# Capabilities and workflows



# Capabilities and workflows



| | Scoping (20%) | Piloting (25%) | Scaling up (25%) | Reporting (30%) |
|---|---|---|---|---|
| Activities | Map policy questions & data opportunities | Rapid exploration of data opportunities | Polish and expand successful pilots | Synthesis, visualisation & presentation |
| Outputs | Project spec (paper prototype) | Blog + prototype visualisation | Data | Visualisation, report, open dataset, GitHub repository, |
| | | Code | | |
| Tools | | Data collection / analysis: Python, R (a little)<br>Data storage: MySQL, AWS | | Visualisation: D3.js, Leaflet.js |
| | Coordination: Slack, Trello, GitHub, Cookiecutter, JuPyteR Notebooks | | | |

# Capabilities and workflows [Discussion]

Thinking about the activities we discussed

- Do you have the right capabilities to do this kind of project?
- How would you develop them / acquire them?
- What are the risks of different opportunities (eg. Internal vs. outsourcing vs. working with academic researchers?)

15 minutes

## Conclusion

- Projects taking you from text to impact have much potential but also risks
- These risks can be managed by following a structured approach and being mindful of the challenges

nesta

nesta.org.uk

@nesta_uk

Joel.Klinger@nesta.org.uk
Juan.mateos-Garcia@nesta.org.uk
Chantale.tippet@nesta.org.uk