

# Linking and Exploring Open Government Data

## *Business Investment, Research and Grants*

BSc Thesis Project Video

4 May 2020

William Vigolo da Silva

Student ID: 14279225

Email: [psywv@nottingham.ac.uk](mailto:psywv@nottingham.ac.uk)



**University of  
Nottingham**

UK | CHINA | MALAYSIA

# Introduction

## Objective:

- Case study of applying **data aggregation, analysis and network visualization** to government open data to derive insights into UK government investment in research and innovation

## Open data:

- Free to access, use, and free from restrictions
  - Often published by the government for transparency
- Novel in-depth analysis of the entire UKRI Gateway to Research data set.

# UKRI's Gateway to Research Portal

- UK Research and Innovation & Subsidiaries fund research and innovation projects
  - Yearly budget > £7 billion
- Provides data on funded projects since 1973
  - ~100,000 projects
  - ~80,000 individuals
  - ~50,000 organisations

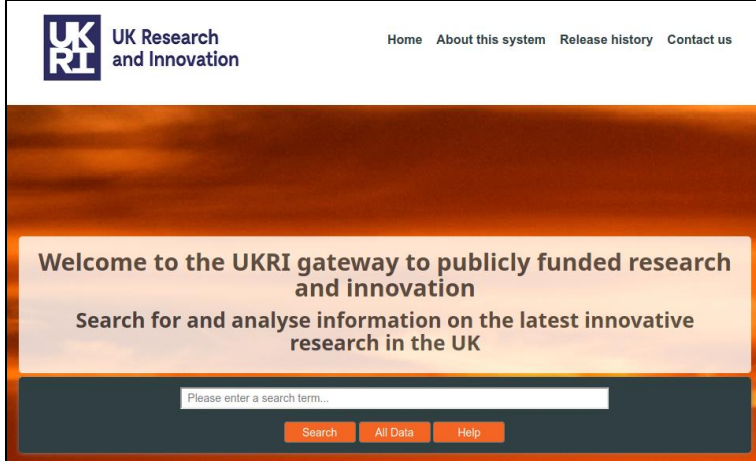
Lead Research Organisation: [City University London](#)  
Department Name: Faculty of Management

[Go back](#)

Overview **Organisations** People Publications Outcomes

## Organisations

[City University London, United Kingdom \(Lead Research Organisation\)](#)  
[ESRC, United Kingdom \(Co-funder\)](#)  
[Tech City UK \(Collaboration\)](#)  
[London School of Economics & Pol Sci, United Kingdom \(Collaboration\)](#)  
[CREATe \(RCUK Centre for Copyright and New Business Models in the Creative Economy\) \(Collaboration\)](#)



UKRI UK Research and Innovation

Home About this system Release history Contact us

Welcome to the UKRI gateway to publicly funded research and innovation

Search for and analyse information on the latest innovative research in the UK

Please enter a search term...

Search All Data Help



UKRI UK Research and Innovation

Home About this system Release history Contact us

## Building Better Business Models: Capturing the Transformative Potential of the Digital Economy

Lead Research Organisation: [City University London](#)  
Department Name: Faculty of Management

[Go back](#)

Overview **Organisations** People Publications Outcomes

### Abstract

This research project is exploring how firms are applying and engaging with new digital technologies to become more efficient, profitable and dynamic. While there is considerable understanding about how digital technologies allow firms to create value, there is much less understanding of how firms can use DE to sense what consumers and society needs and monetize that value and turn it into financial returns for investors, entrepreneurs and shareholders. This is part of a more general concern that the UK economy is relatively good at invention but less good at producing firms that capture its benefits in new, fast growing markets. By exploring how digital technology is transforming the three elements that make up a business model - how firms understand customers' needs, how they create value for customers, and how they capture and monetize this value - this project will generate new understanding about how digital technology can be commercialised more effectively. This knowledge will help firms in the UK generate more jobs, more economic growth and improved services to firms and the general public.

The empirical part of the project will conduct research on (a) sectors that generate digital technology such as open-source software

**Funded Value:**  
£939,389

**Funded Period:**  
Jul 13 - Jun 17

**Funder:**  
EPSRC

**Project Status:**  
Closed

# Process

- Data warehousing: Incorporate multiple sources of data into a single data store
- Linking: Associate related or identical entities
- Data Analysis:
  - What is the structure of the ecosystem of publicly-funded research?
  - How does the ecosystem change over time?
  - What are the significant factors that influence collaboration between organisations?

## Data Acquisition

- Retrieve & store data

## Data Preparation

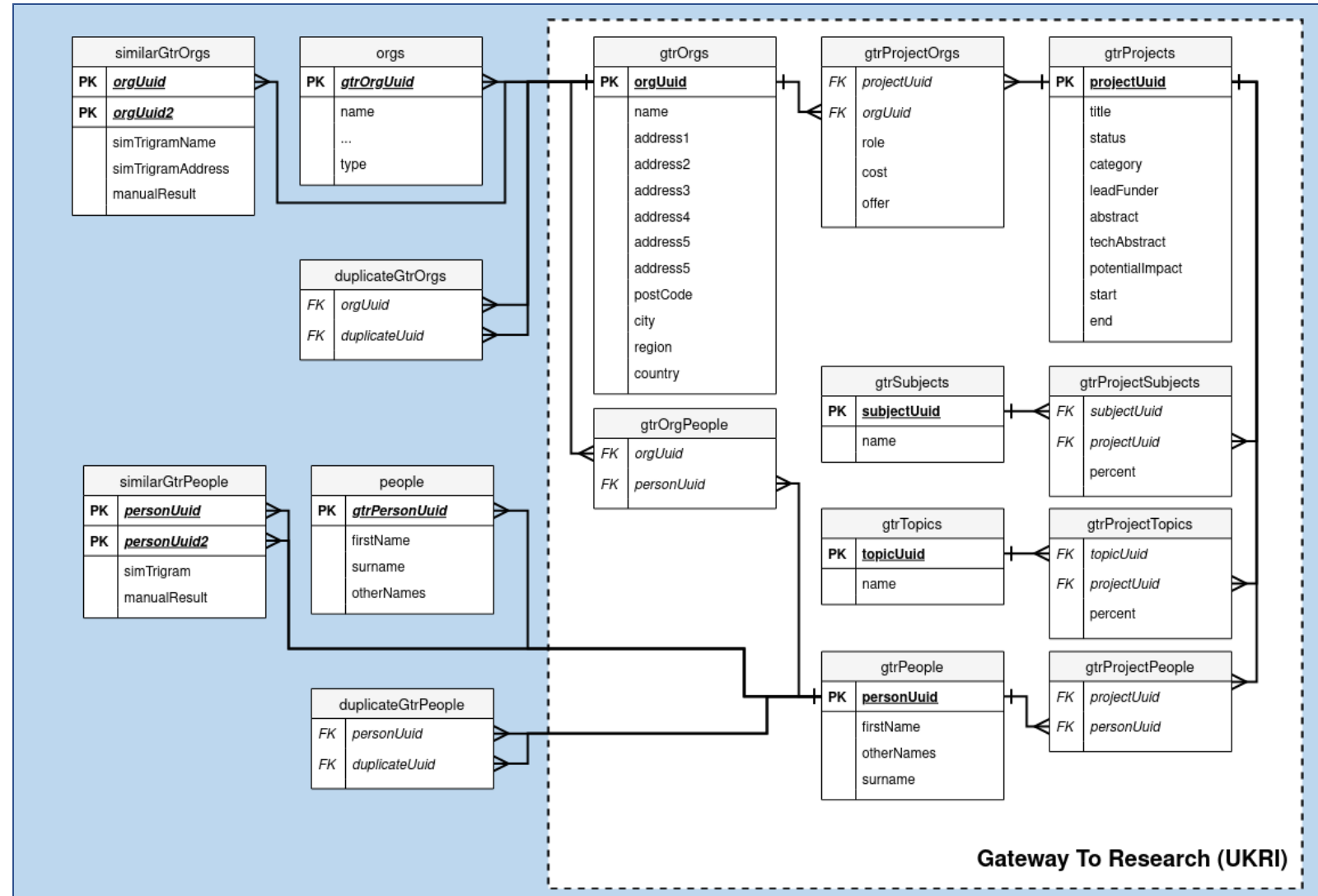
- Clean up data
- Linking

## Data Analysis & Visualization

- Apply computational techniques to data

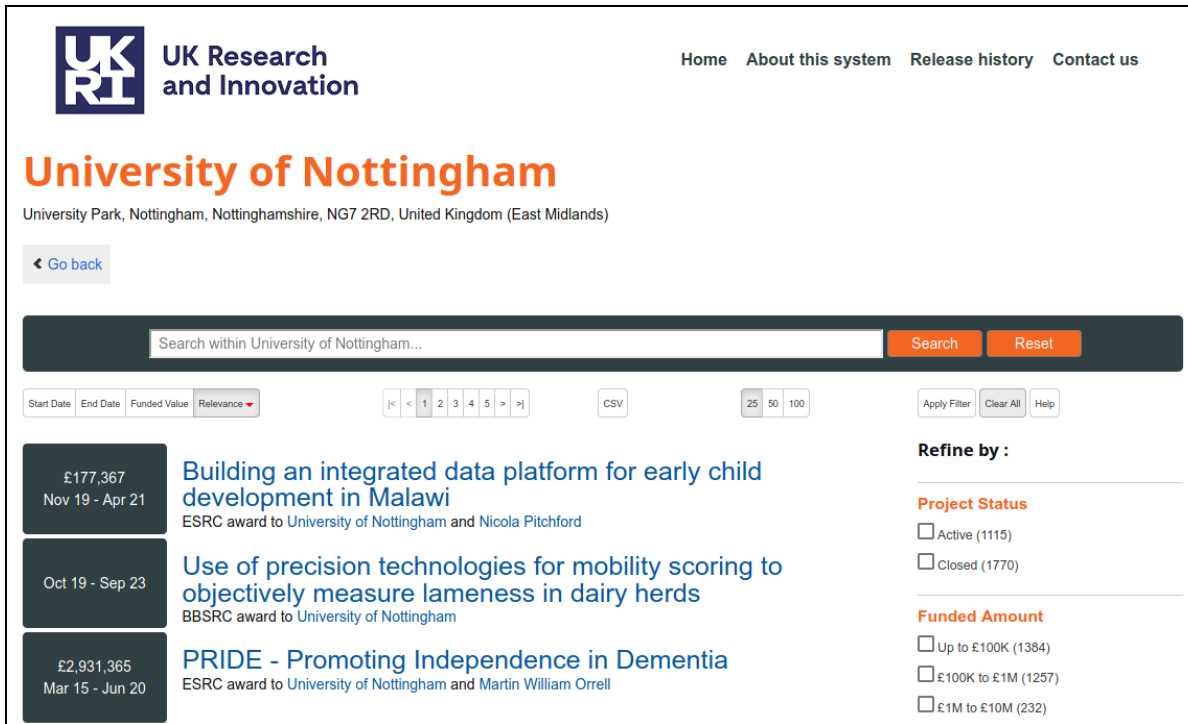
# Data Acquisition

- Download XML dataset via Web API
- Design a **unified database schema**
- Apply to a relational Database Management System (DMS) - **PostgreSQL**
- Import XML through PostgreSQL procedures
- Keep all original data for reporting, e.g., impact of cleaning
- Keep some calculated data for use in analysis



# Preparation: Eliminating invalid records

- Invalid records are identified and removed:
  - Contain names like 'Unknown', or 'Unlisted'
- Records are filtered out by name, or by detecting a lack of links to projects



UKRI UK Research and Innovation

Home About this system Release history Contact us

## University of Nottingham

University Park, Nottingham, Nottinghamshire, NG7 2RD, United Kingdom (East Midlands)

[Go back](#)

Search within University of Nottingham... [Search](#) [Reset](#)

Start Date End Date Funded Value Relevance CSV 25 50 100 Apply Filter Clear All Help

**Refine by :**

**Project Status**

- ☐ Active (1115)
- ☐ Closed (1770)

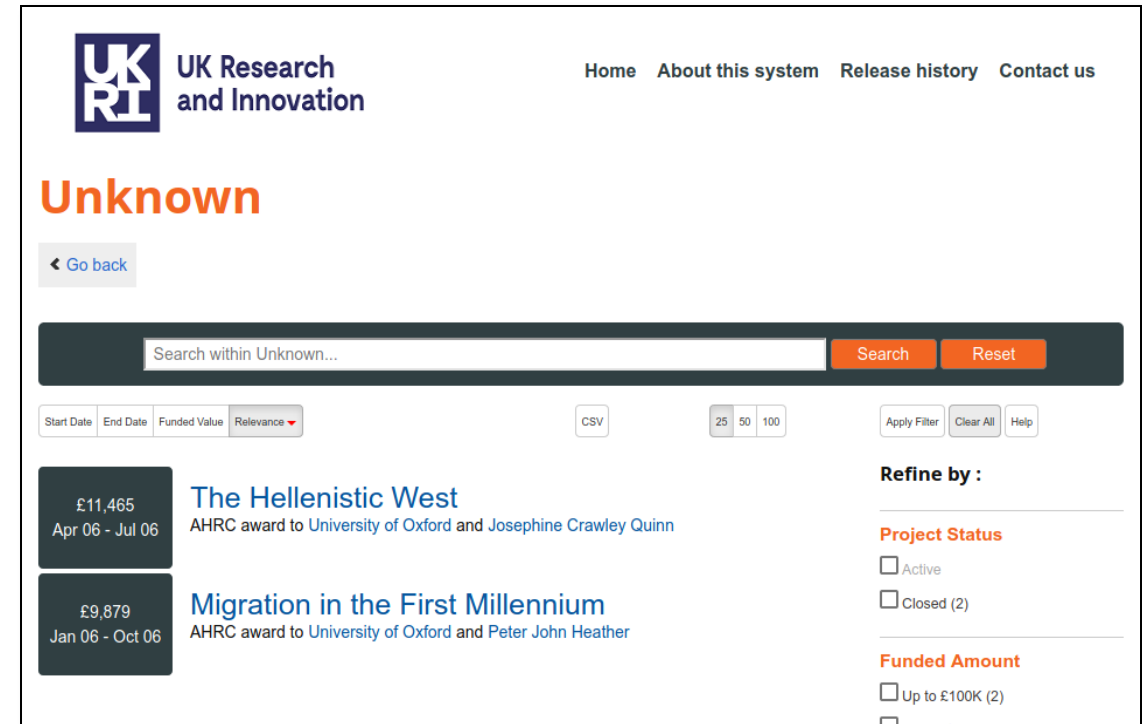
**Funded Amount**

- ☐ Up to £100K (1384)
- ☐ £100K to £1M (1257)
- ☐ £1M to £10M (232)

£177,367 Nov 19 - Apr 21 **Building an integrated data platform for early child development in Malawi**  
ESRC award to University of Nottingham and Nicola Pitchford

Oct 19 - Sep 23 **Use of precision technologies for mobility scoring to objectively measure lameness in dairy herds**  
BBSRC award to University of Nottingham

£2,931,365 Mar 15 - Jun 20 **PRIDE - Promoting Independence in Dementia**  
ESRC award to University of Nottingham and Martin William Orrell



UKRI UK Research and Innovation

Home About this system Release history Contact us

## Unknown

[Go back](#)

Search within Unknown... [Search](#) [Reset](#)

Start Date End Date Funded Value Relevance CSV 25 50 100 Apply Filter Clear All Help

**Refine by :**

**Project Status**

- ☐ Active
- ☐ Closed (2)

**Funded Amount**

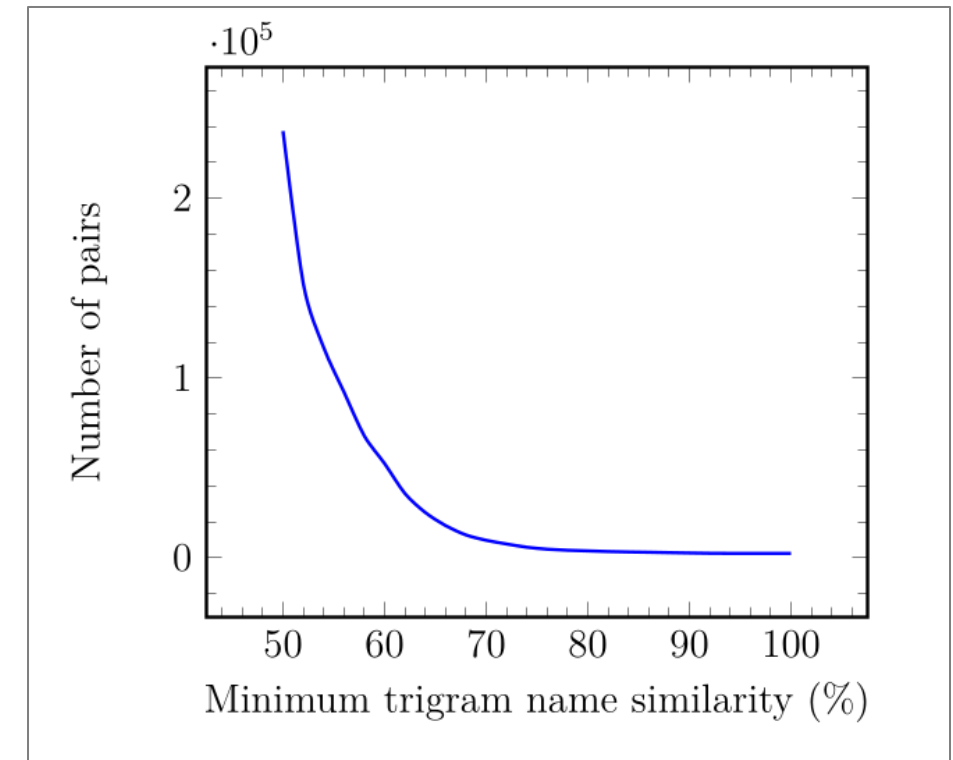
- ☐ Up to £100K (2)

£11,465 Apr 06 - Jul 06 **The Hellenistic West**  
AHRC award to University of Oxford and Josephine Crawley Quinn

£9,879 Jan 06 - Oct 06 **Migration in the First Millennium**  
AHRC award to University of Oxford and Peter John Heather

# Preparation: De-duplication

- Duplicate records caused by human error & multiple data sources
- De-duplicate by merging similar records
- Similarity detected through trigrams – triplets of sequential characters in text
  - "trigram"  $\Rightarrow$  ("tri", "rig", "igr", "gra", "ram")
  - Choose a threshold, e.g., 90% similarity
- Used contextual information: postcodes for organisations, or employers for people
- Replaced common abbreviations/acronyms
  - "Uni"  $\equiv$  "University"



# Analysis: Grouping organisations by type

- Identify type using their names
  - E.g. 'Universities' are academic organisations, 'Limited' companies are privately-owned
- Detect variations
  - University  $\equiv$  Uni.
  - Limited  $\equiv$  Ltd.
- 54.8% records were assigned a type
- *Possible improvement*: Cross-references with complementary information sources
  - E.g. the United Kingdom's Companies House

Registered office address

**The Sir John Peace Building Experian Way, Ng2 Business Park,  
Nottingham, NG80 1ZZ**

Company status

**Active**

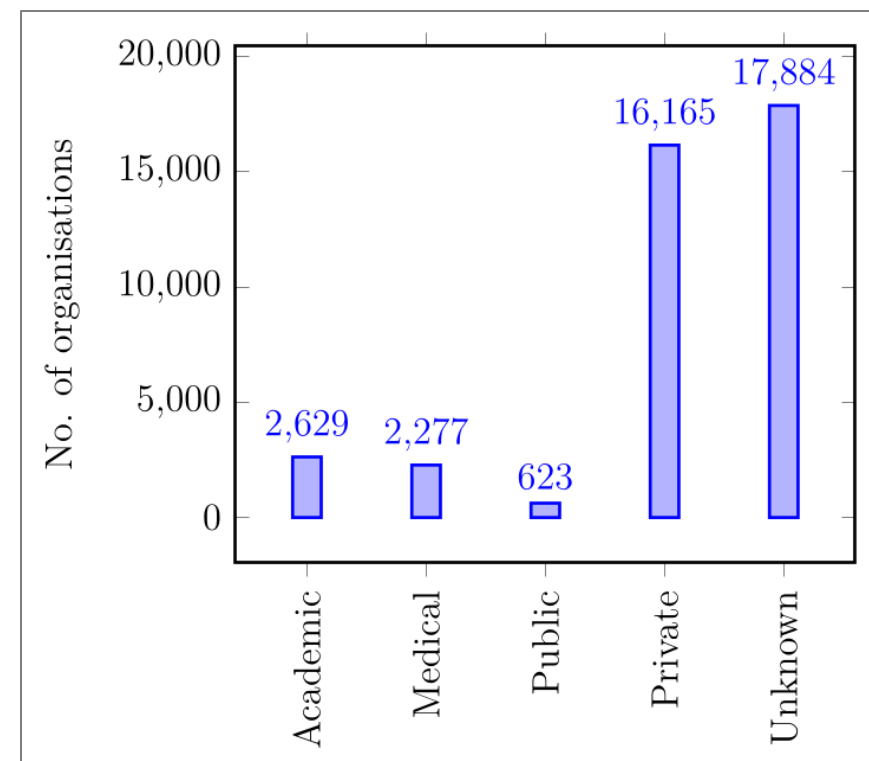
Company type

**Private limited Company**

Incorporated on

**22 March 1960**

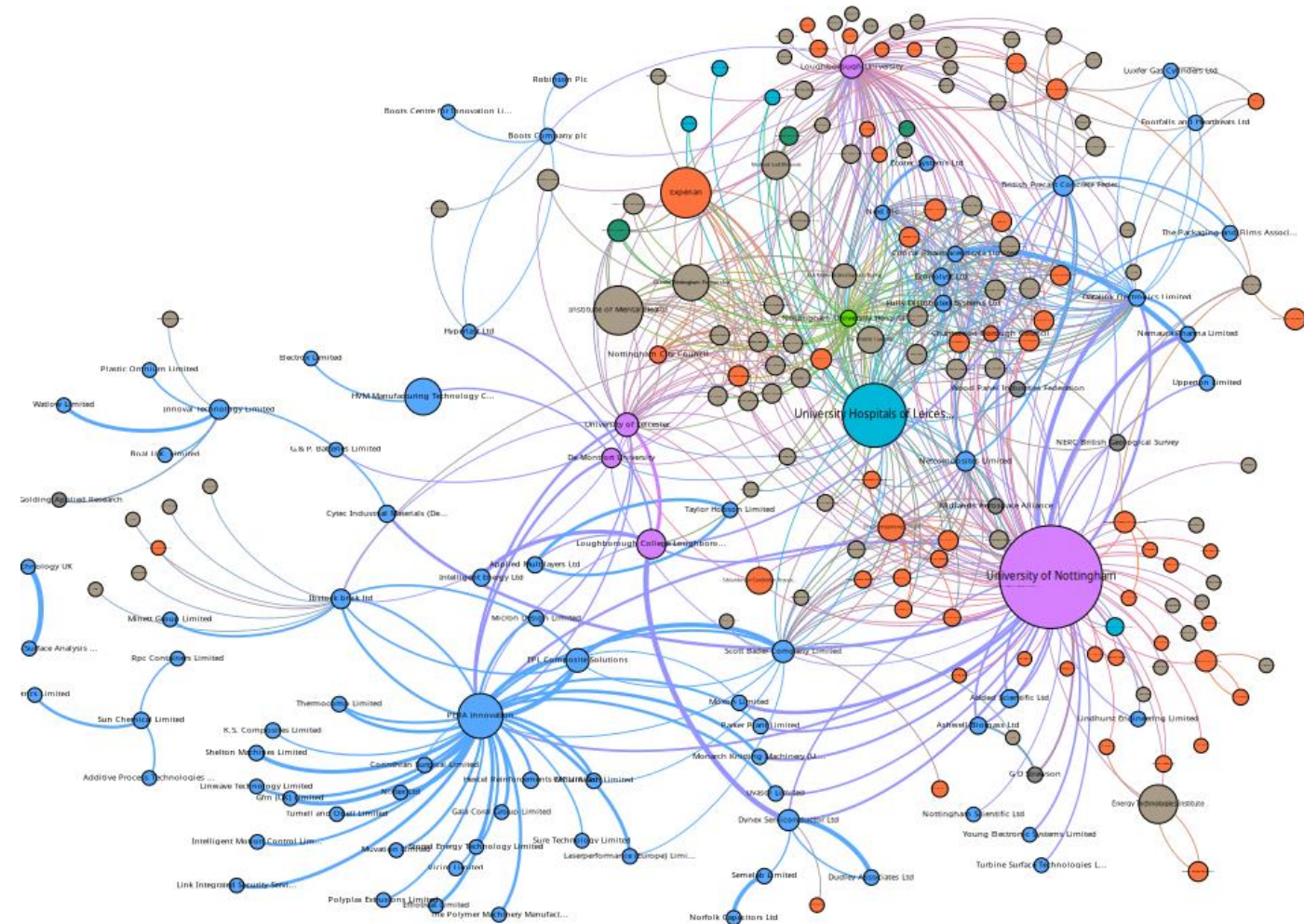
Type	Exemplary keywords
Academic	University, College, Academy, Academic
Medical	Hospital, NHS, Medical, Health
Private	Limited/LTD, Corporation, Company, Incorporated
Public	Council, Government, Governorate



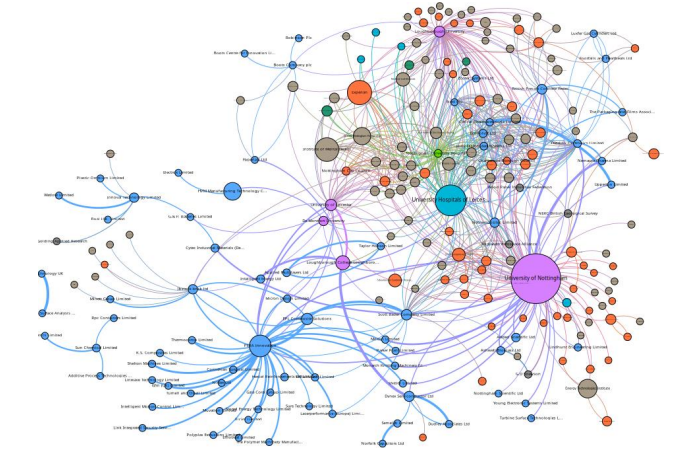
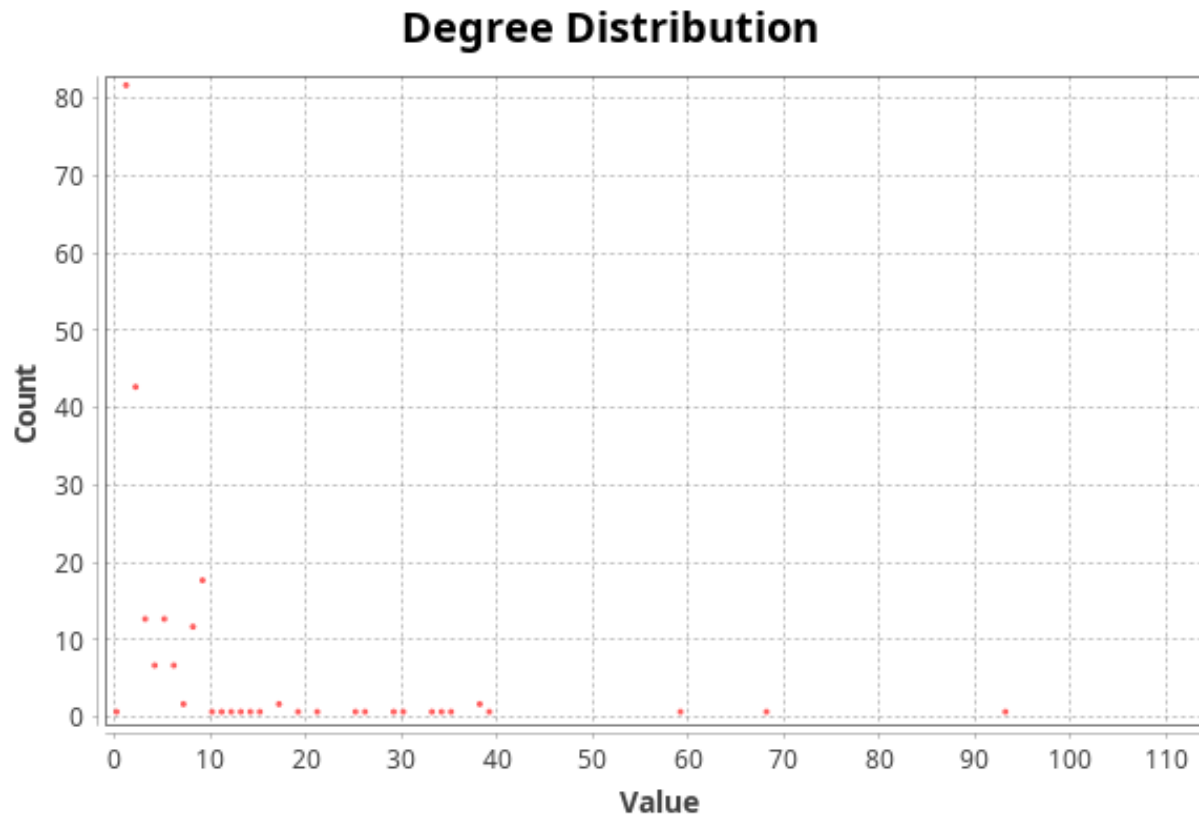


# Network Analysis and Visualization with Gephi: East Midlands ecosystem

- Gephi: A free and open source, Java-based network analysis tool
- Analysed East Midlands organisations between 2002-2008
  - Links are project collaborations
- Network analysis including (weighted) degree distribution, clustering coefficients
- Results:
  - Academic & private hubs are large, but little cross-collaboration
  - Lots of cross-collaboration in medical hubs



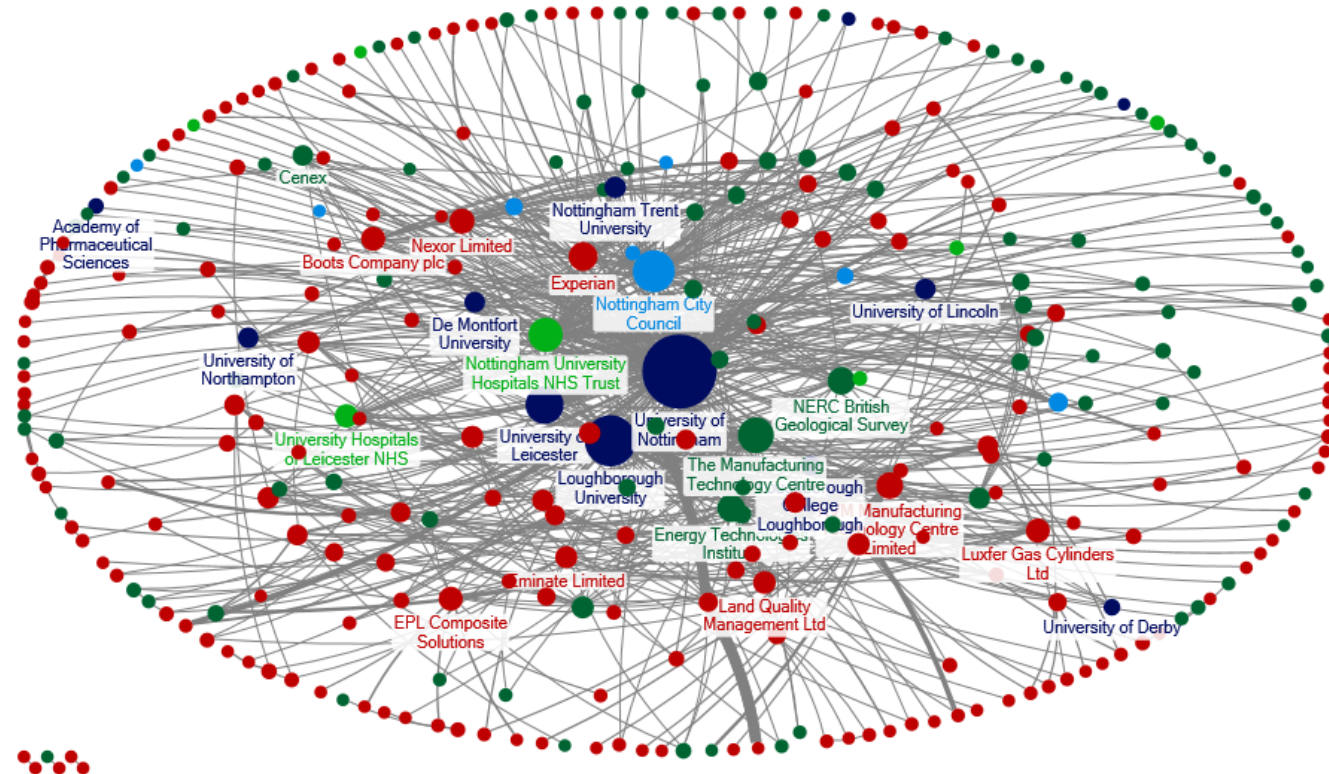
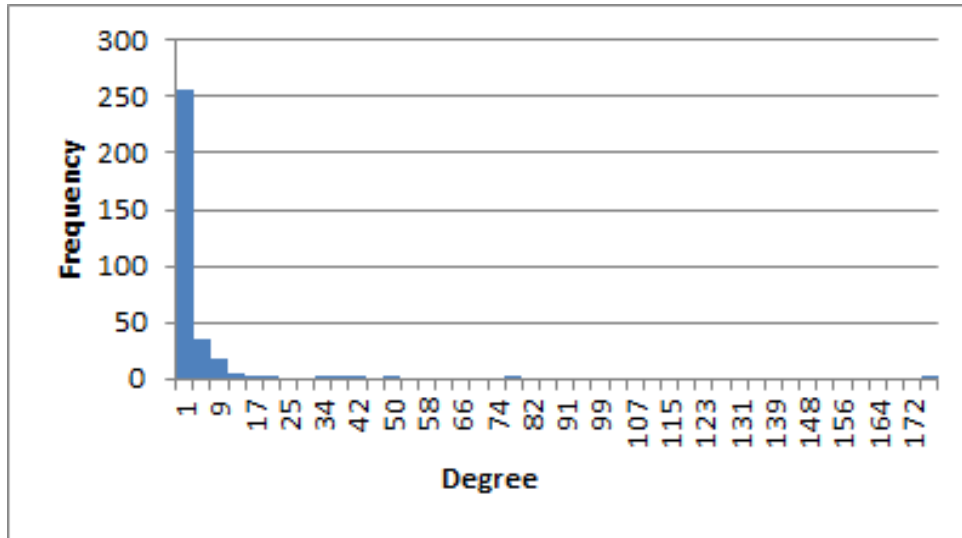
# Network Analysis and Visualization with Gephi: East Midlands Hubs



- Most organisations collaborate with only a few others
- 'Hub' organisations have a lot of collaborations:
  - University of Nottingham: 114
  - Loughborough University: 93
  - University Hospitals of Leicester: 68
  - Nottingham University Hospitals: 59
  - PERA Innovation: 38
  - Experian: 17

# Network Analysis and Visualization with NodeXL: East Midlands ecosystem

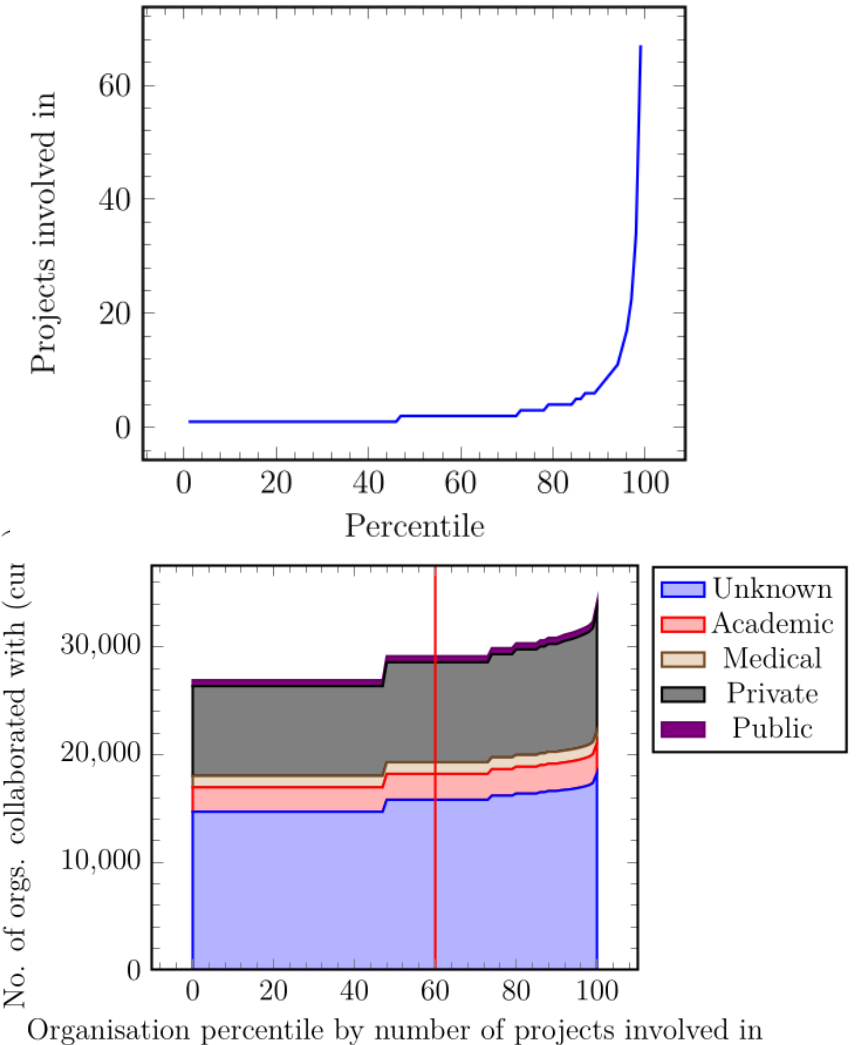
- NodeXL: An add-on for Microsoft Office Excel for (social) network analysis
- Analysed East Midlands organisations between 2005-2010
  - Links are projects collaborations
  - Explore changes from 2002-2008
- Network analysis of degree distribution
- Results: very similar pattern, but the top





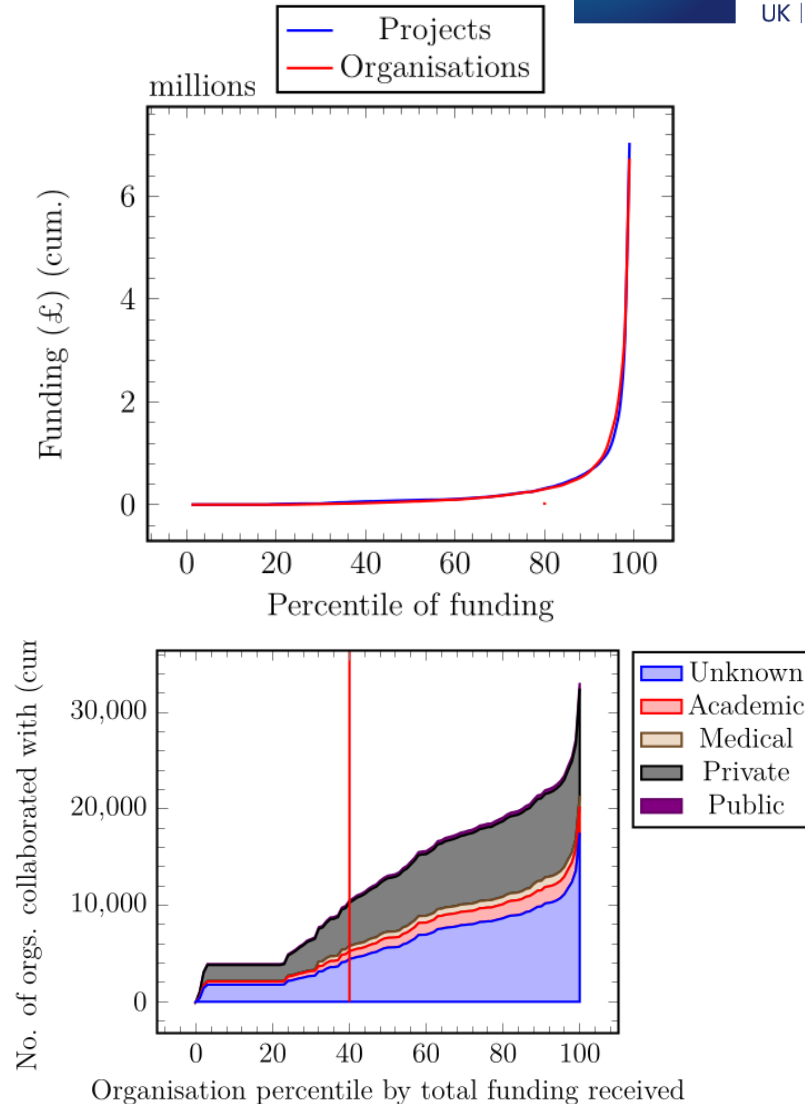
# Analysis: Influence of the number of projects in collaboration

- What's the relationship between number of projects an organisation is involved in and the collaboration with other organisations?
- Findings:
  - The top 10% of organisations are involved in the vast majority of projects
  - 78.9% of organisations collaborate with others ranked in the bottom 10%, by the amount of previous research
  - 9.6% of organisations have collaborated only with the top organisations



# Analysis: Influence of funding in collaboration

- What's the relationship between total funding received and collaboration with other organisations?
- Findings:
  - The top 10% of both organisations and projects receive the vast majority of funding
  - 35.1% of organisations have only collaborated with the top 10% by funding received



# Summary



- Successfully completed data warehousing
  - Designed a single schema for incorporating all data in the datasets
  - Applied existing techniques for linking and de-duplication
- Applied data normalization, analysis and visualization to answer specific questions about entities within the dataset
- Provided a baseline for further research into UKRI's data
- Provided an illustrative case study for researchers working on open data
- Highlighted difficulties and approaches in processing data from open data platforms.

# Thank you!

William Vigolo da Silva

Student ID: 14279225

Email: [psywv@nottingham.ac.uk](mailto:psywv@nottingham.ac.uk)