Data Visualisation and Dashboarding

Week 10 – Networks and unstructured data

UNIVERSITY OF WESTMINSTER

What shape is your data?

Rectangular data (key-value pairs, tables, matrix, etc.)

Hierarchical (org chart, etc)

Graphs (networks, processes, etc.)

Unstructured (Text, images, audio, film, etc.)

What’s a graph?

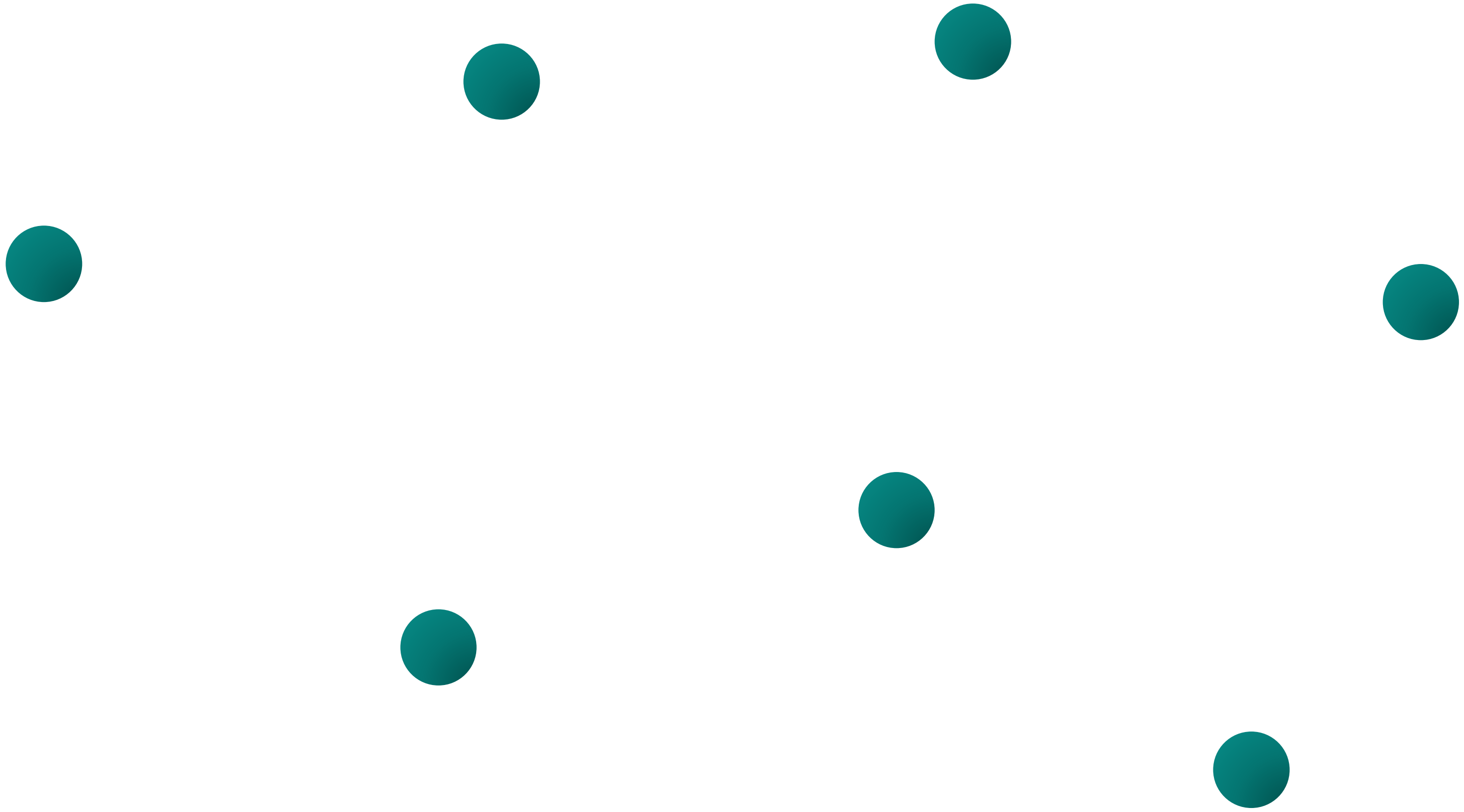
Graph (*noun*): *“a collection of vertices [nodes] and edges that join pairs of vertices”*

Vertex (*noun*): *“a point (as of [a…] graph […]) that terminates a line […]”*

(Merriam-Webster)

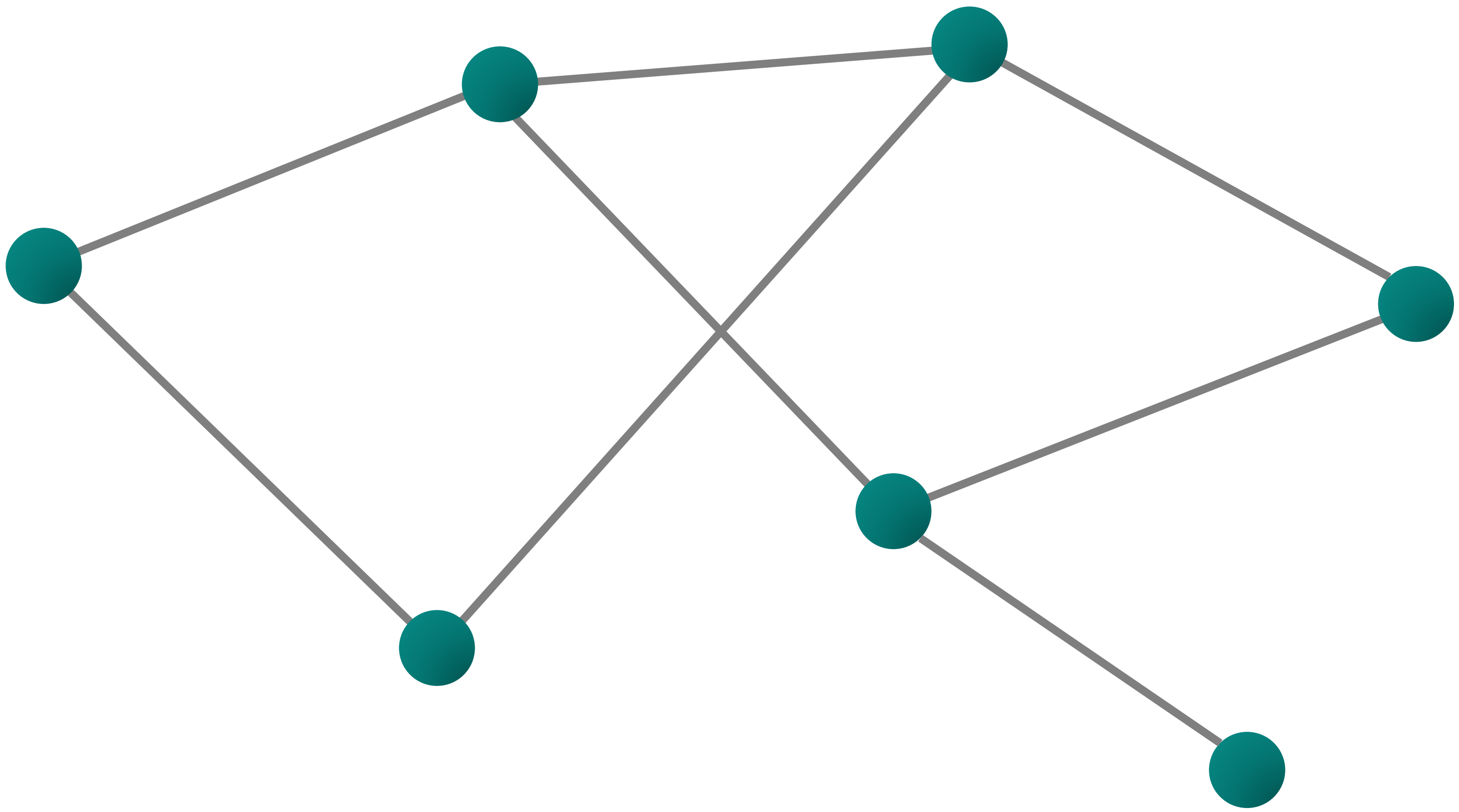
What is a graph?

Nodes



What is a graph?

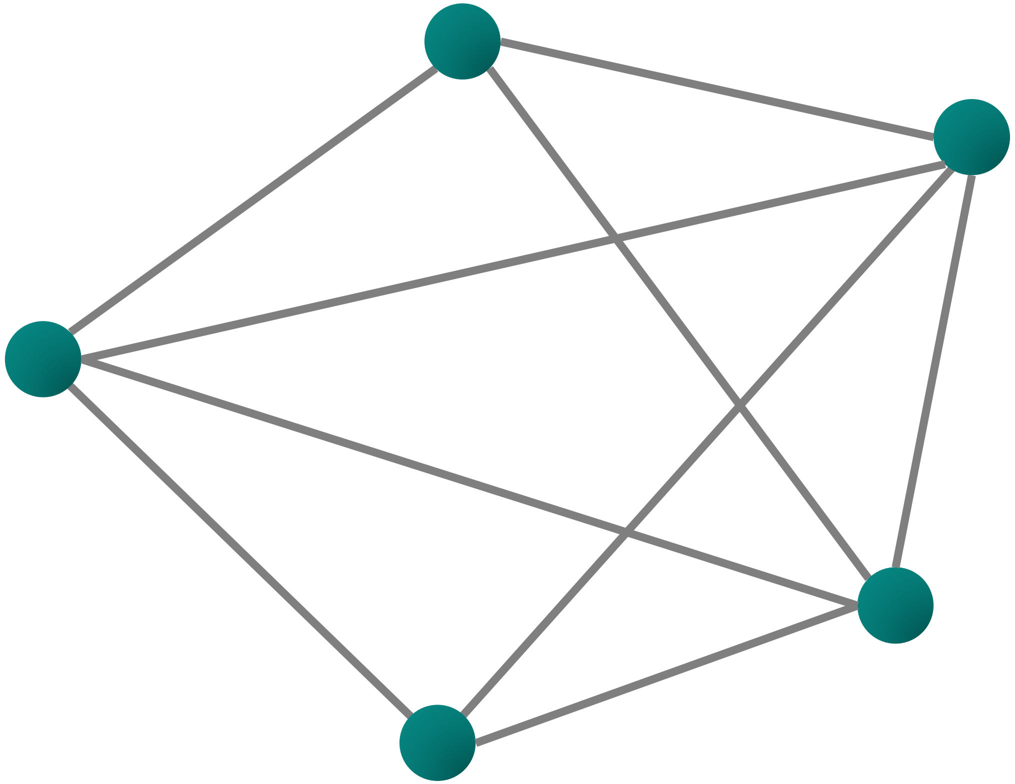
Edges



Complete Graph

Fully connected graph

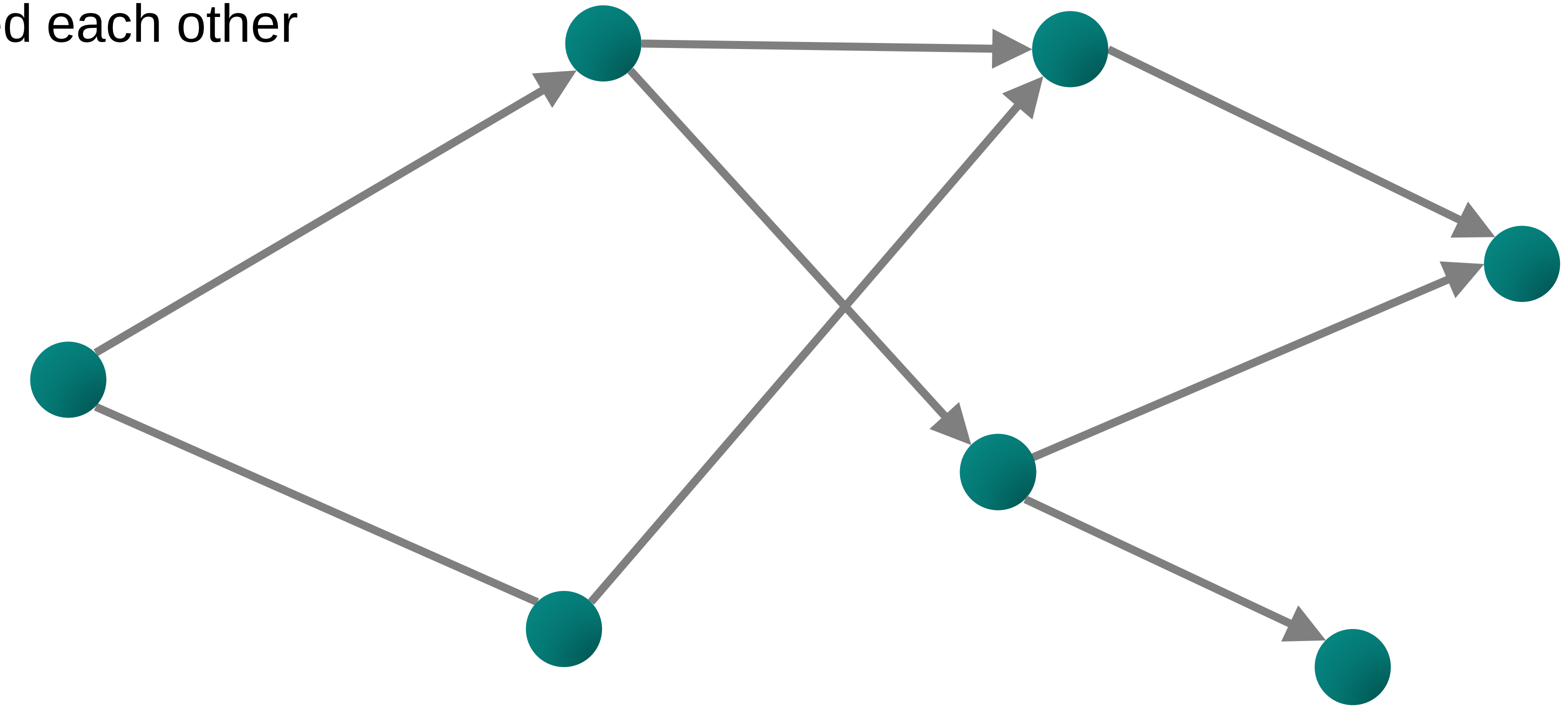
Each node is connected to all other nodes



Directed Graph

Edges have a direction

Nodes precede/succeed each other

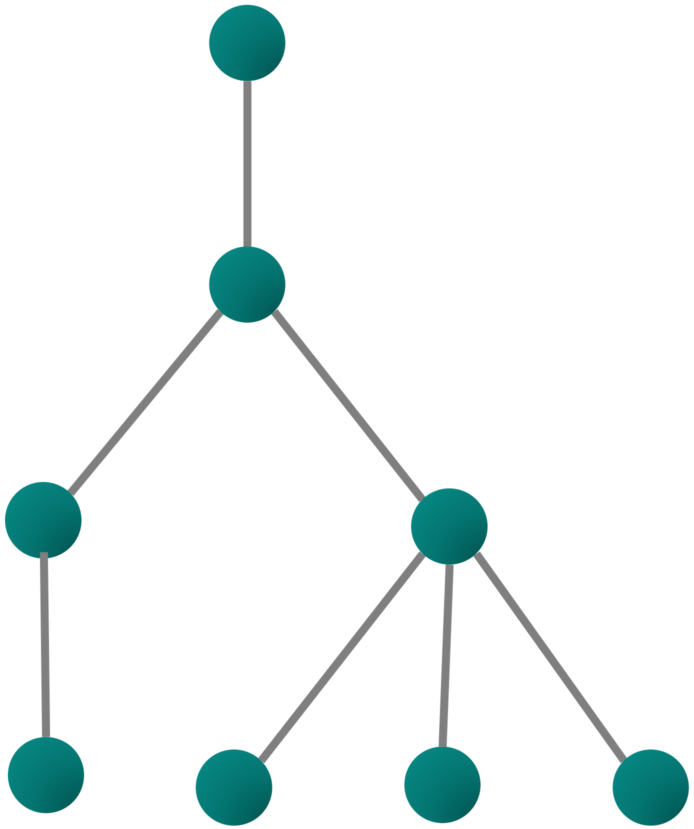


Acyclic graph (tree)

Graph without any cycles

Can describe hierarchical relationships

Nodes have a parent/child relationship

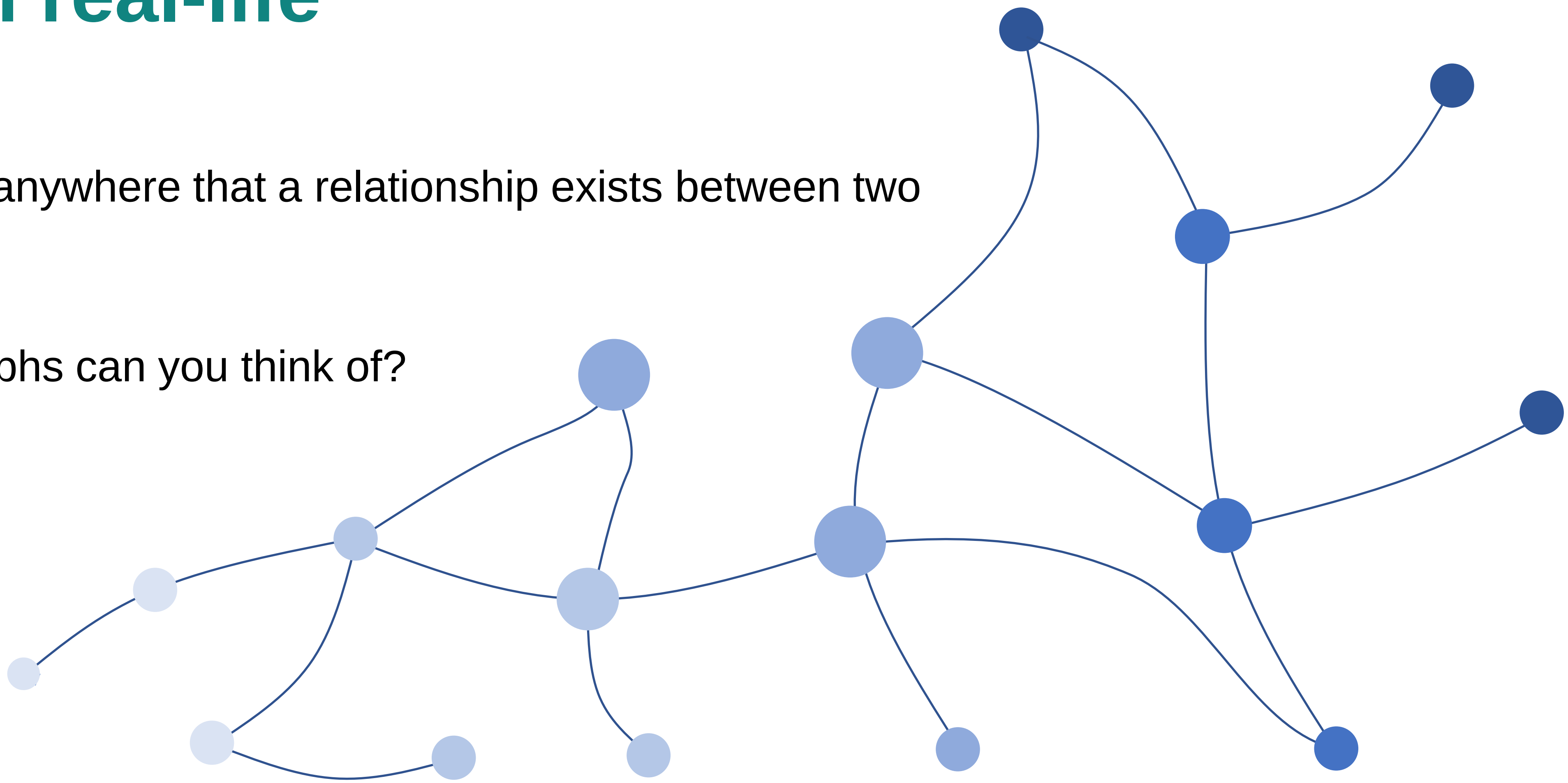


UNIVERSITY OF WESTMINSTER

Graphs in real-life

Graphs can exist anywhere that a relationship exists between two entities.

What types of graphs can you think of?



Social network

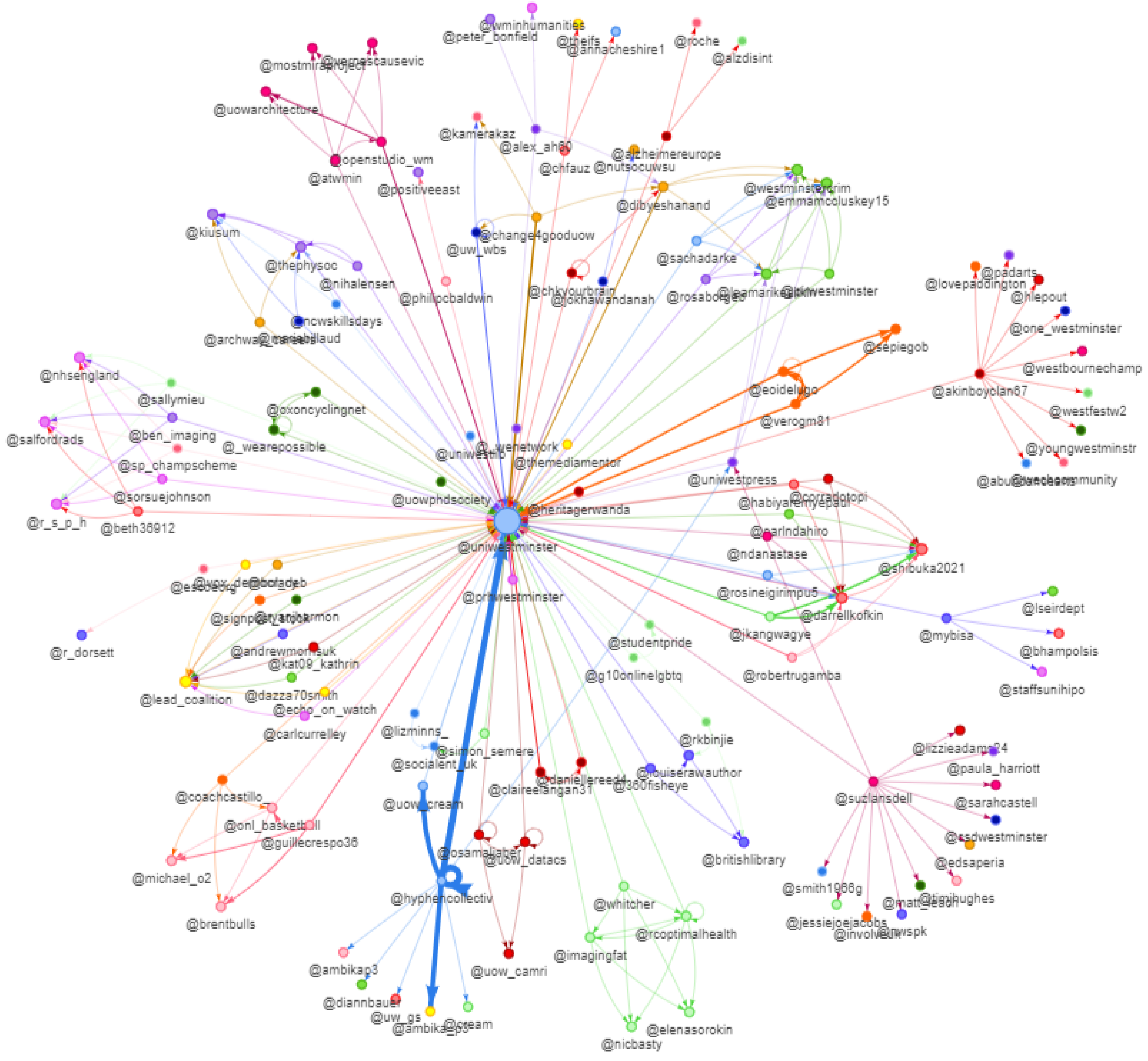
@uniwestminster tweets

Nodes are users

Edges are interactions (RT or mentions)

Helps to identify communities

Generated by SocioViz.net



UNIVERSITY OF WESTMINSTER

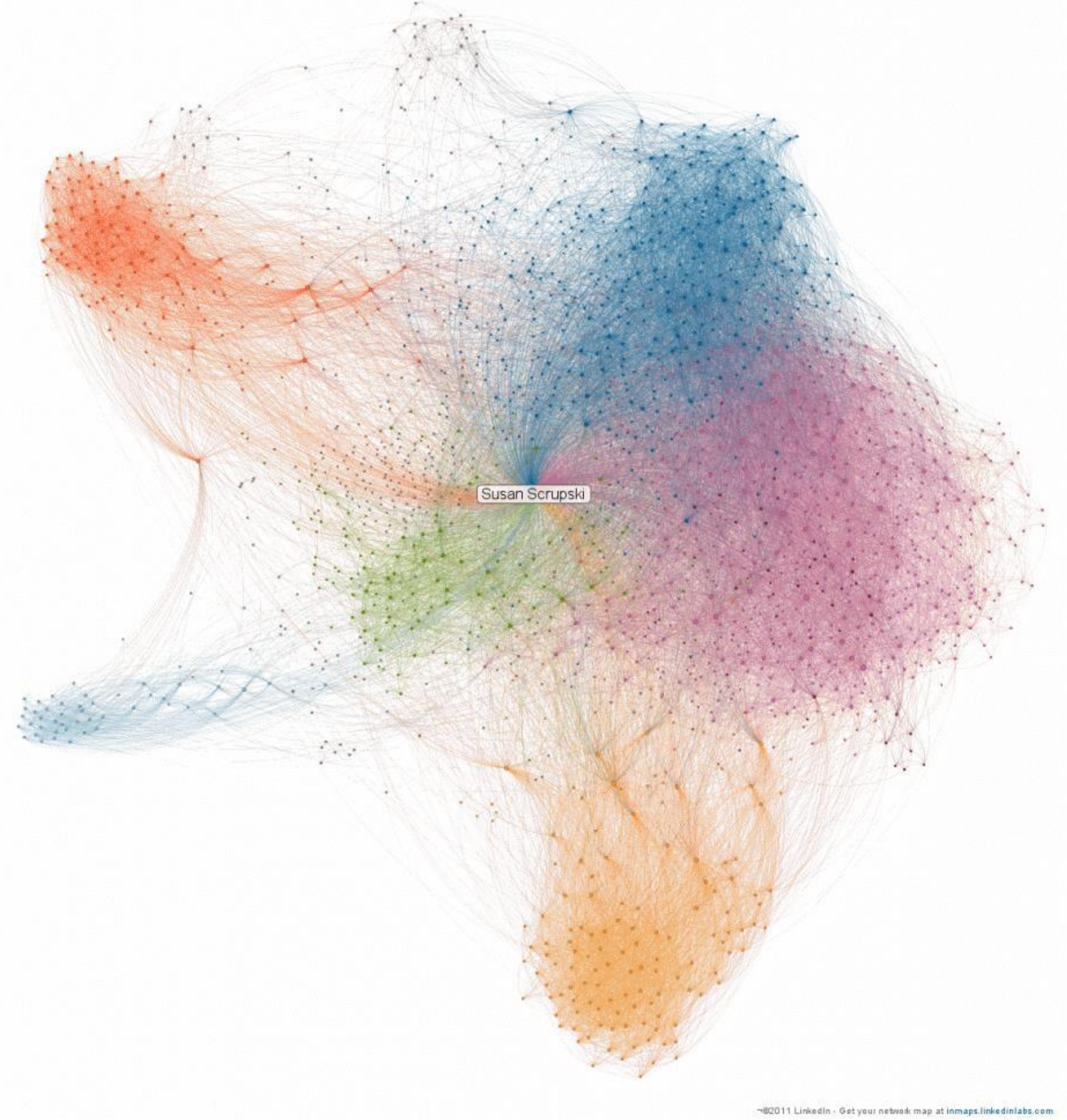
Social Network

LinkedIn profile connections

Clusters are colour-coded

Strengths?

Weaknesses?



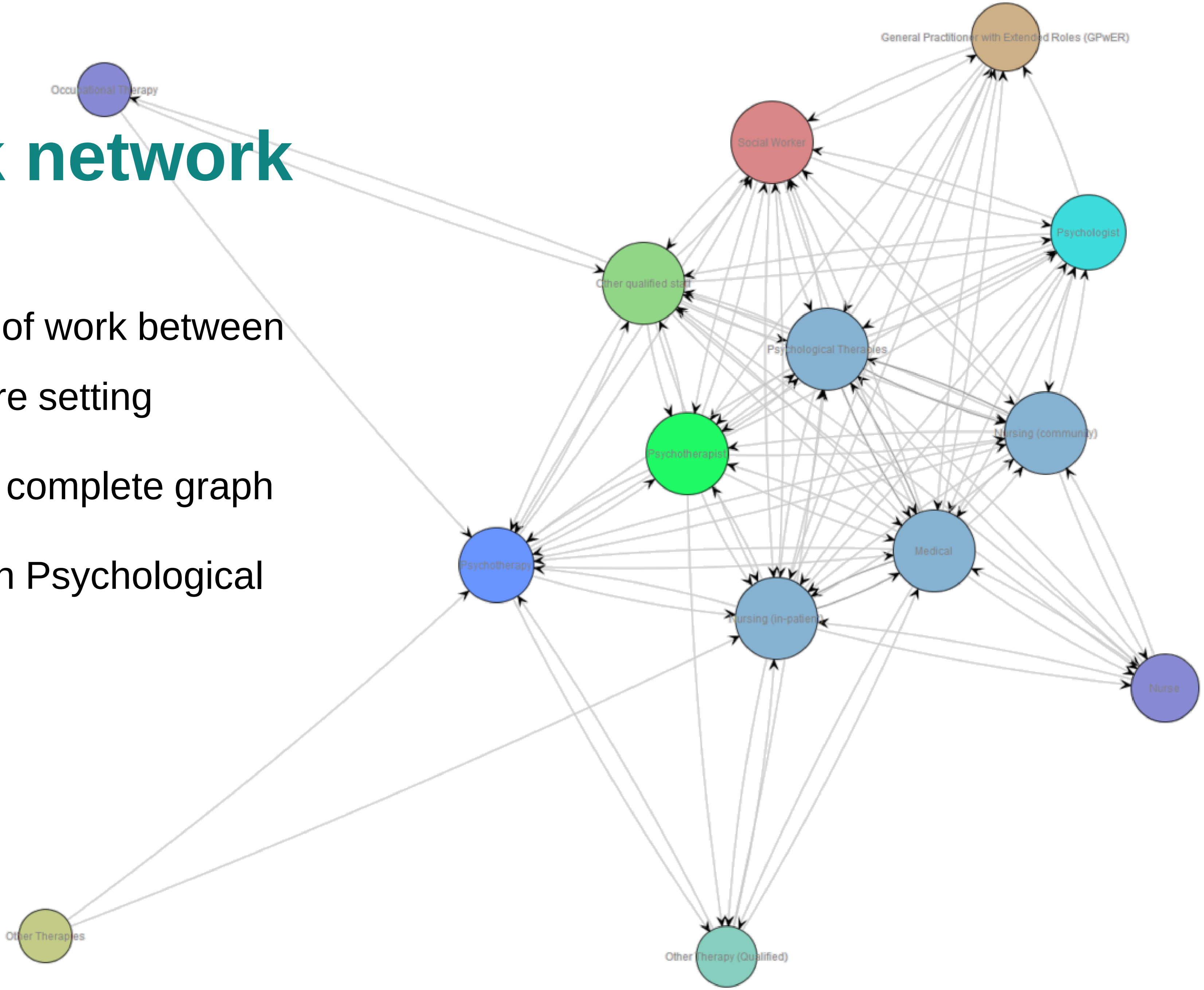
UNIVERSITY OF
WESTMINSTER

Handover of work network

Visualisation shows the handover of work between various roles in a mental healthcare setting

Very complex scenario – almost a complete graph

Most work is handed over between Psychological Therapies, Nursing and Medical.



H.C. Beck original London Underground map

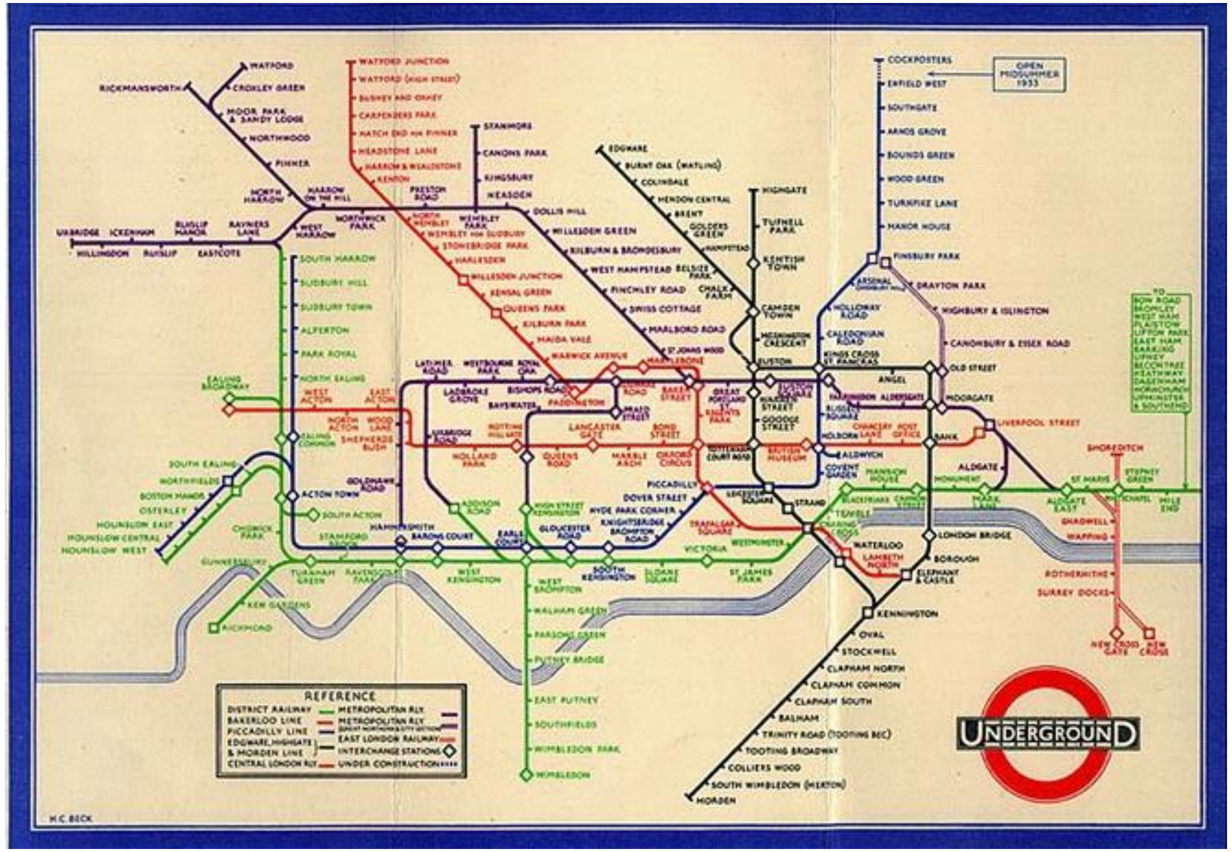
No immediate geospatial frame of reference when underground

Only interested in connections and stations

H.C. Beck designed map in 1933 inspired by electrical wiring diagrams

Drawback: No information about actual distance between stations.

254 people per day travel from Covent Garden to Leicester square (taking 6 minutes by train or 4 minutes walking)



Facebook connections

1.1 billion users and their connections (Facebook, 2013)

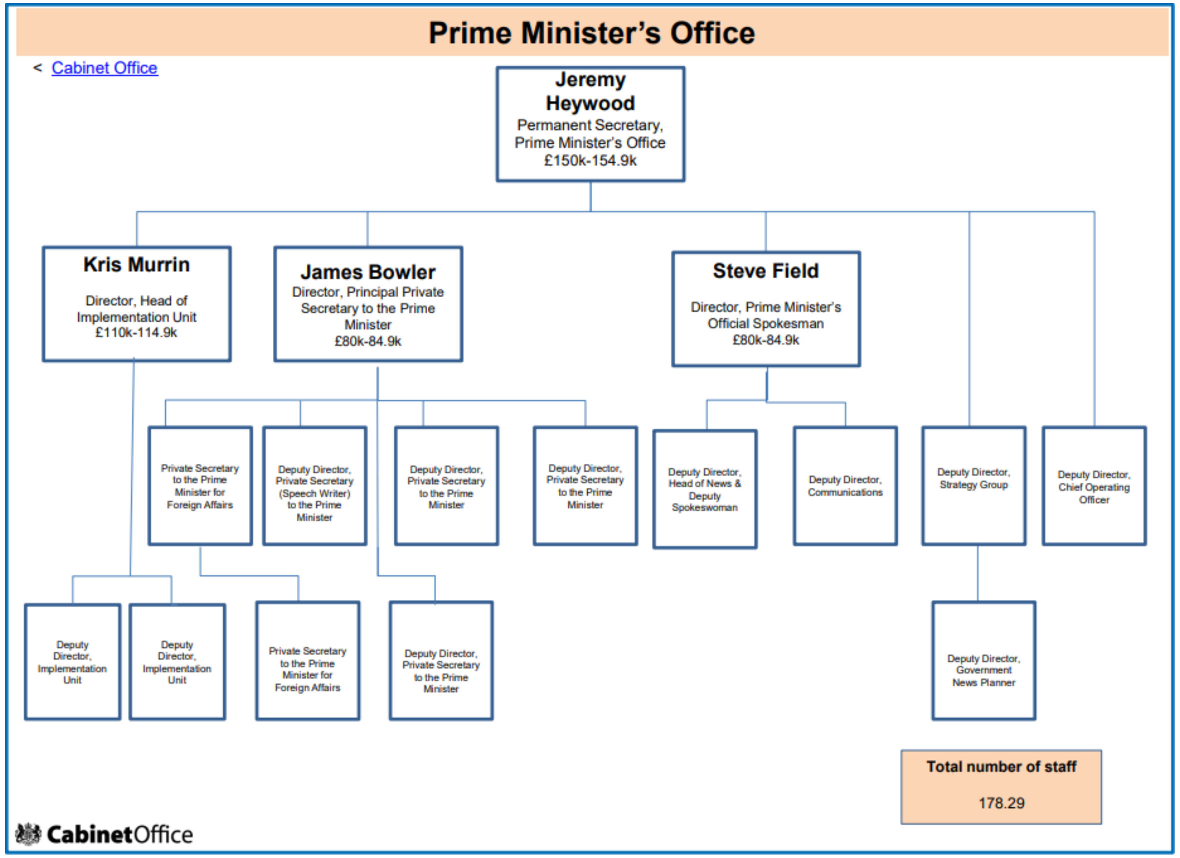
Shows global dominance of FB as social media platform

Also shows areas where FB is not dominant (Russia, Belarus, North Korea)

Organisation charts

Organisation chart of the Prime Minister's Office

(March 2022, Cabinet Office)



UNIVERSITY OF WESTMINSTER

Treemaps

UNAIDS Treemap showing people with HIV infection

Each nested rectangle represents child-node of the enclosing rectangle

Size of rectangle is proportionate to value

Uses colour to group continents

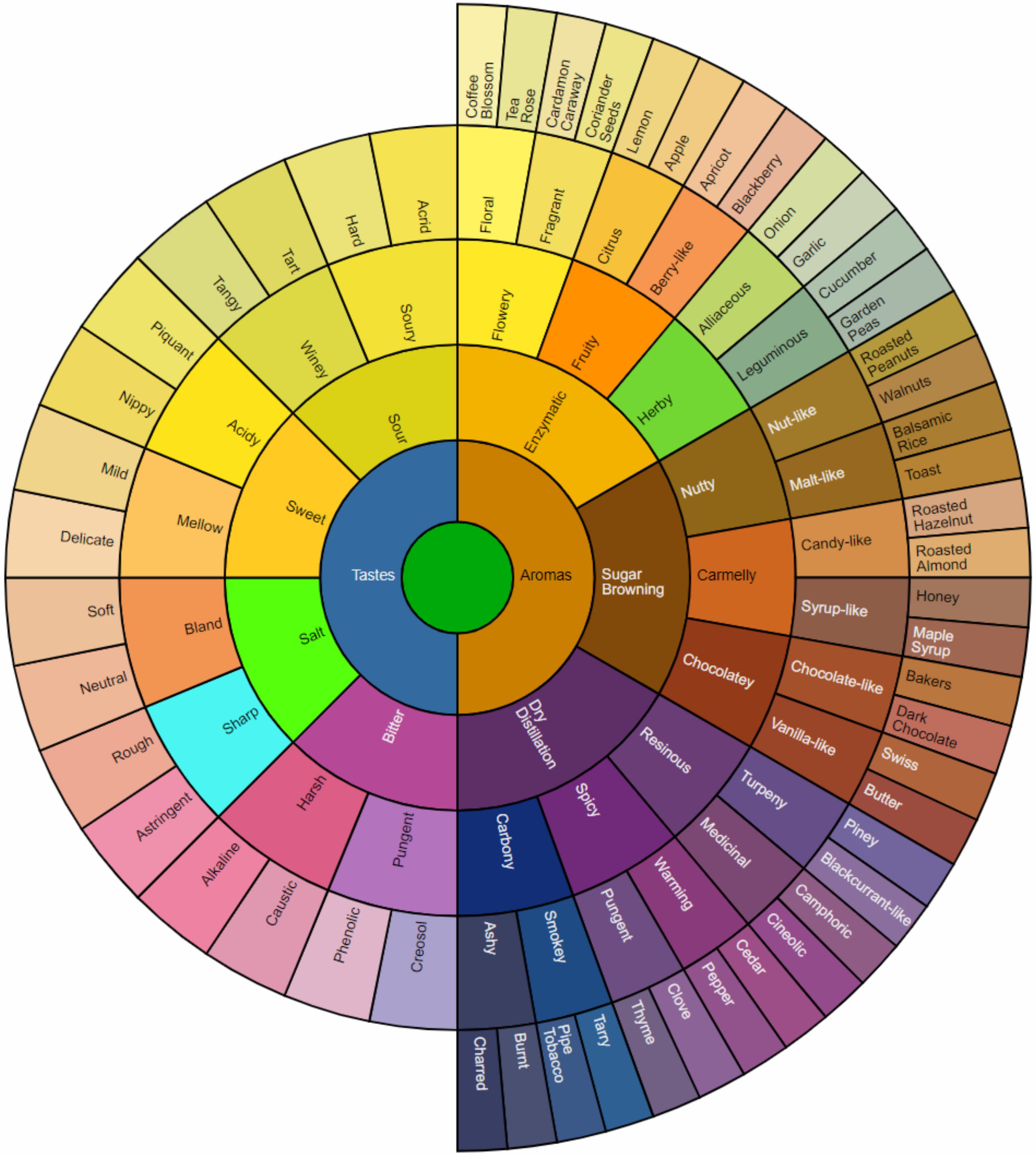
Layered hierarchy

Sunburst Tree derived from Hierarchical data

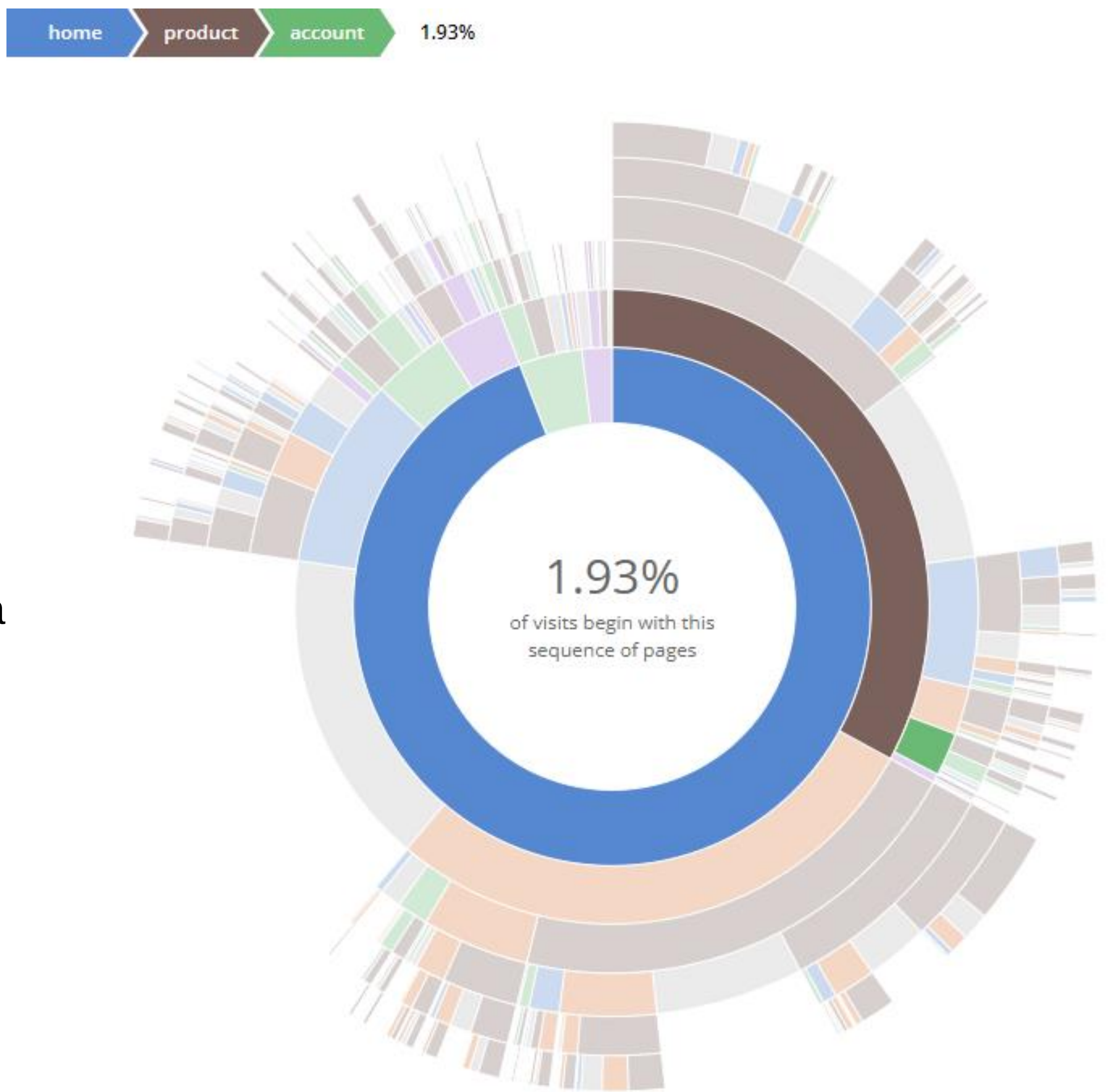
Each concentric ring shows another layer of detail

Colours on outermost layer are selected by designer, inner layers are coloured using average colours

[Coffee Flavour Wheel (jasondavies.com)](https://www.jasondavies.com/coffee-wheel/)



UNIVERSITY OF WESTMINSTER



Sequence of events

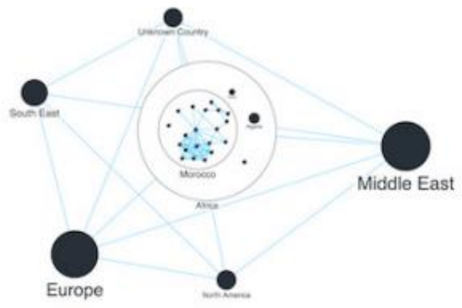
Visualisation shows web site visits

Inner most circle represents first page visited, second circle the second page, etc.

Shows all navigation paths in one view (up to a certain depth)

[Sequences sunburst - bl.ocks.org]

Applications

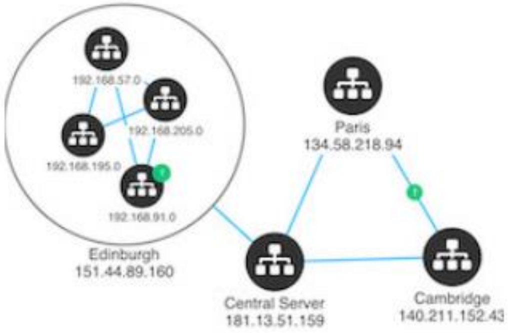


Security & intelligence

Distilling complex connected data into critical intelligence and insight

Anti-fraud

Detecting or investigating fraud in finance, insurance or online activity



Cyber security

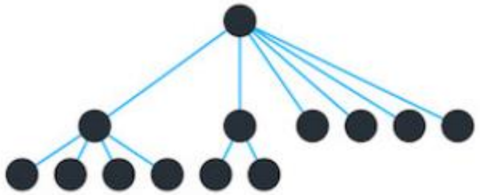
Tracking the behavior of cyber threats and analyzing incident forensics

Cambridge Intelligence (n.d.): The ultimate guide to graph visualization

Applications

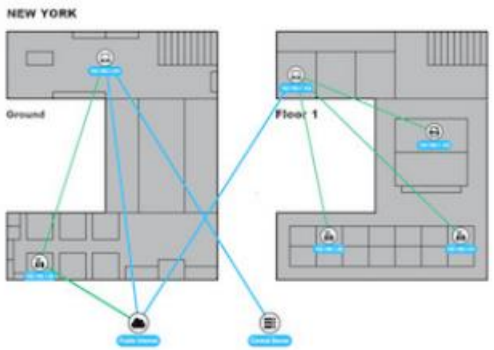
Law enforcement

Enabling detailed pattern of life and behavioral analysis



Compliance

Ensuring regulatory compliance through effective data analysis

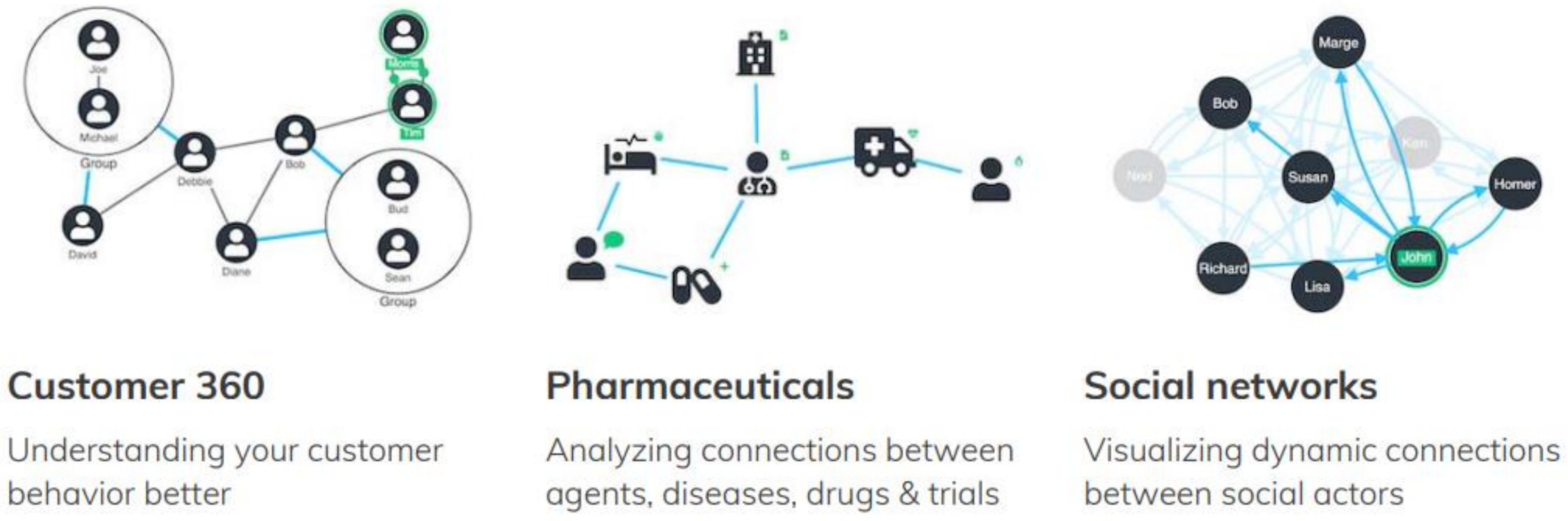


Infrastructure

Monitoring performance and faults plus root cause analysis

Cambridge Intelligence (n.d.): The ultimate guide to graph visualization

Applications



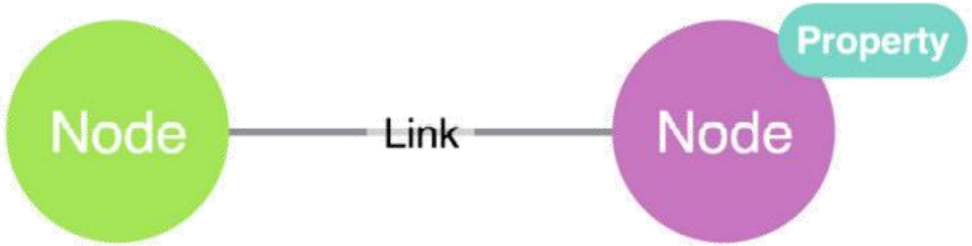
Cambridge Intelligence (n.d.): The ultimate guide to graph visualization

Graph data modelling

Nodes are the fundamental units of your data

Links are relationships between nodes

Properties are descriptive characteristics of nodes and links, but aren't important enough to become nodes themselves



Cambridge Intelligence (n.d.): The ultimate guide to graph visualization

Graph data modelling example

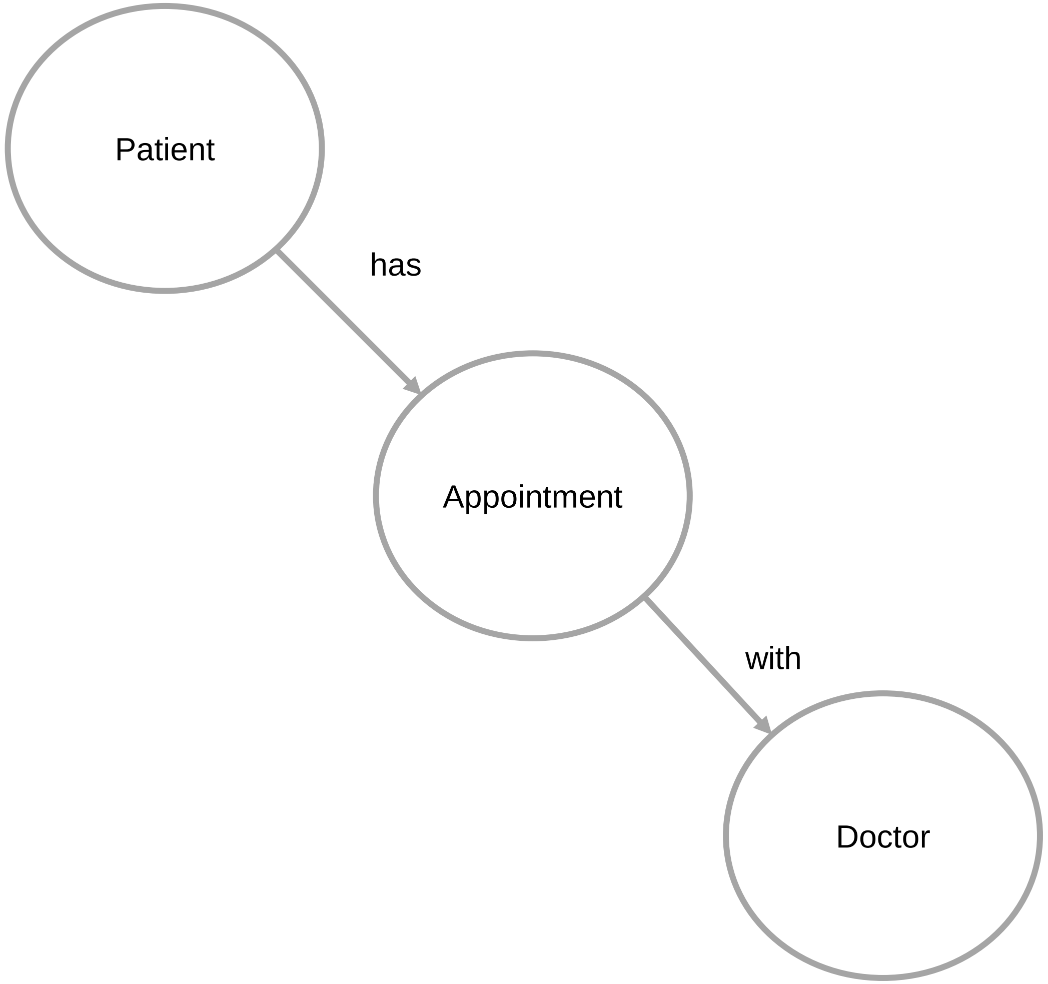
Let’s consider a healthcare graph visualisation with doctors, patients and appointments

We could model each entity as a node

What does the user want to know?

- How many patients did a doctor see?
- How many appointments has a patient had?

Is there an easier way?

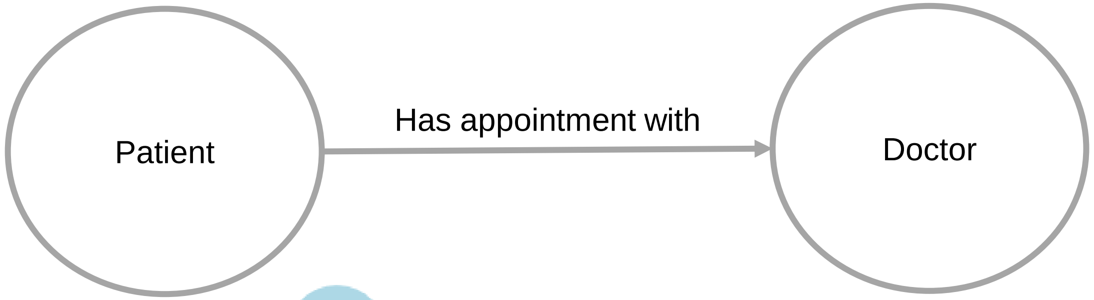


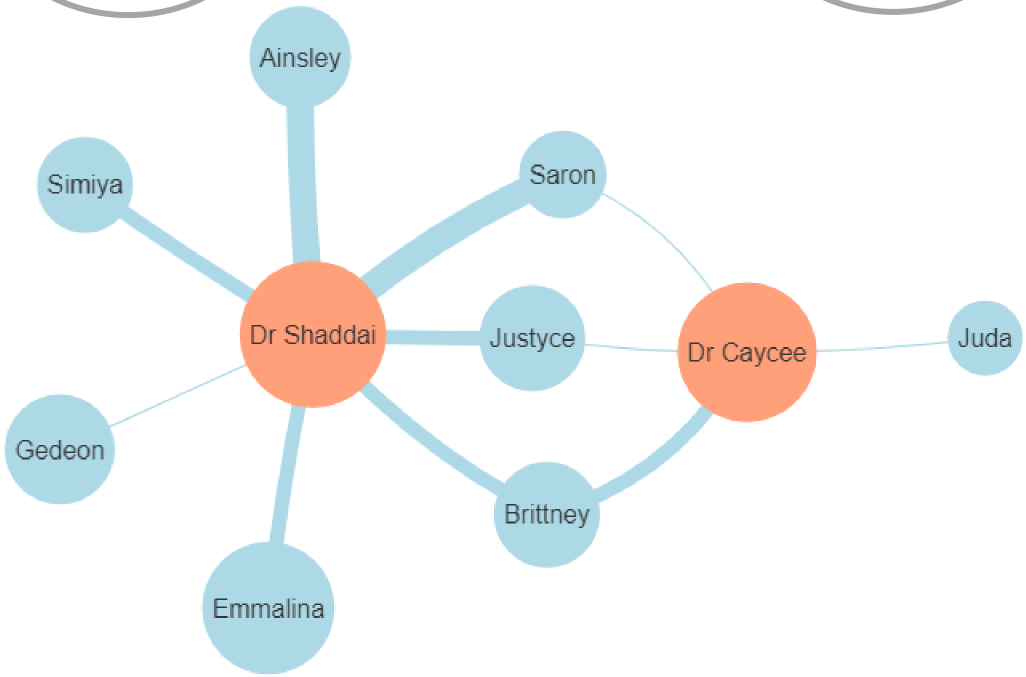
Graph data modelling makeover

Modelling appointments reduces the number of nodes

Resulting graph is easier to read

Colours help separating doctors from patients





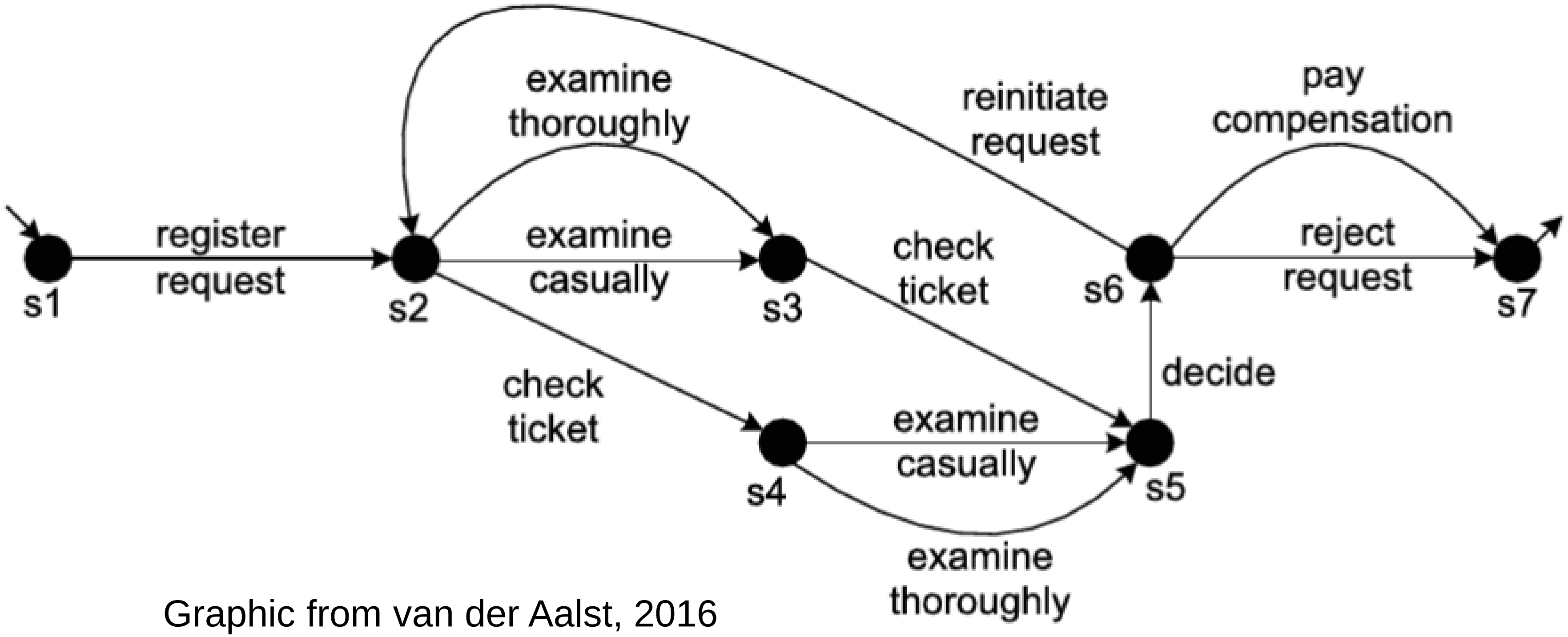
Visualising Business Processes

UNIVERSITY OF
WESTMINSTER

Visualising Business Processes

Most basic process modelling notation is a **transition system**

States are nodes, transitions are edges



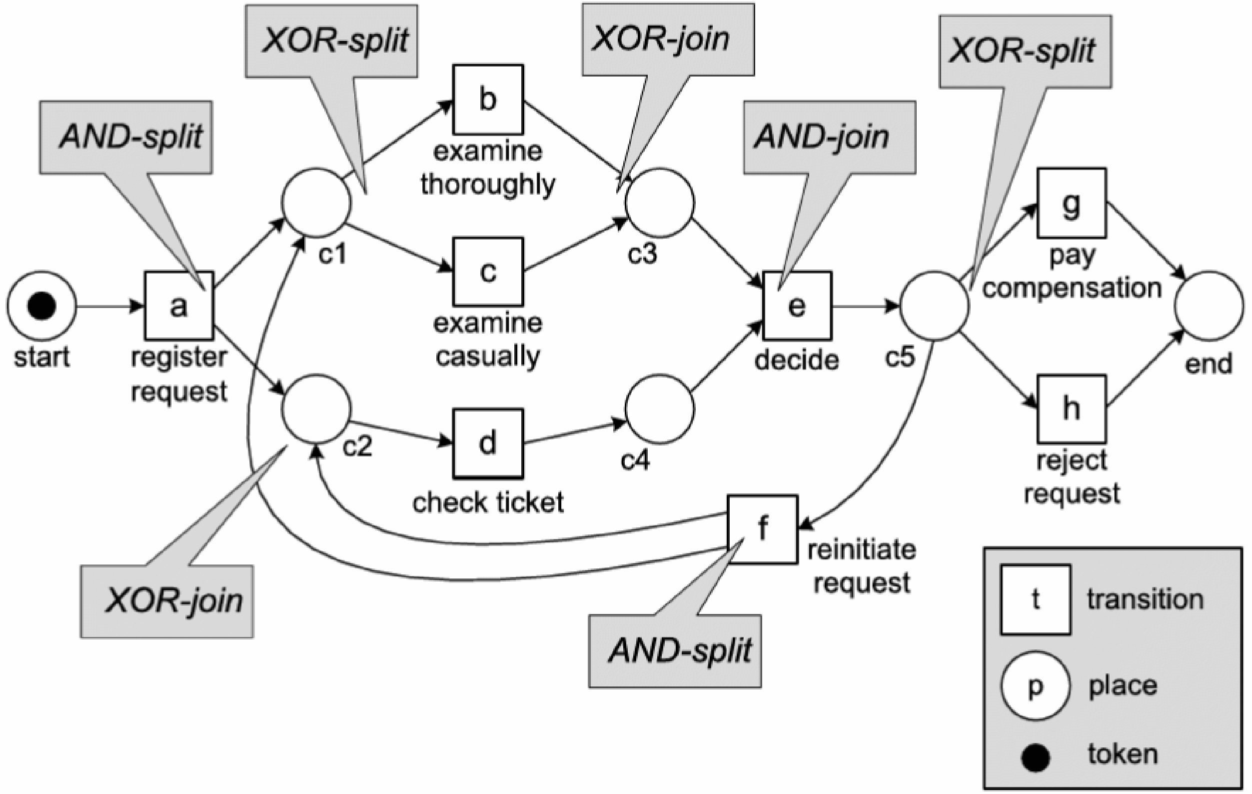
Graphic from van der Aalst, 2016

Petri Nets

Petri Nets are the oldest process modelling language allowing concurrency.

Activities are nodes

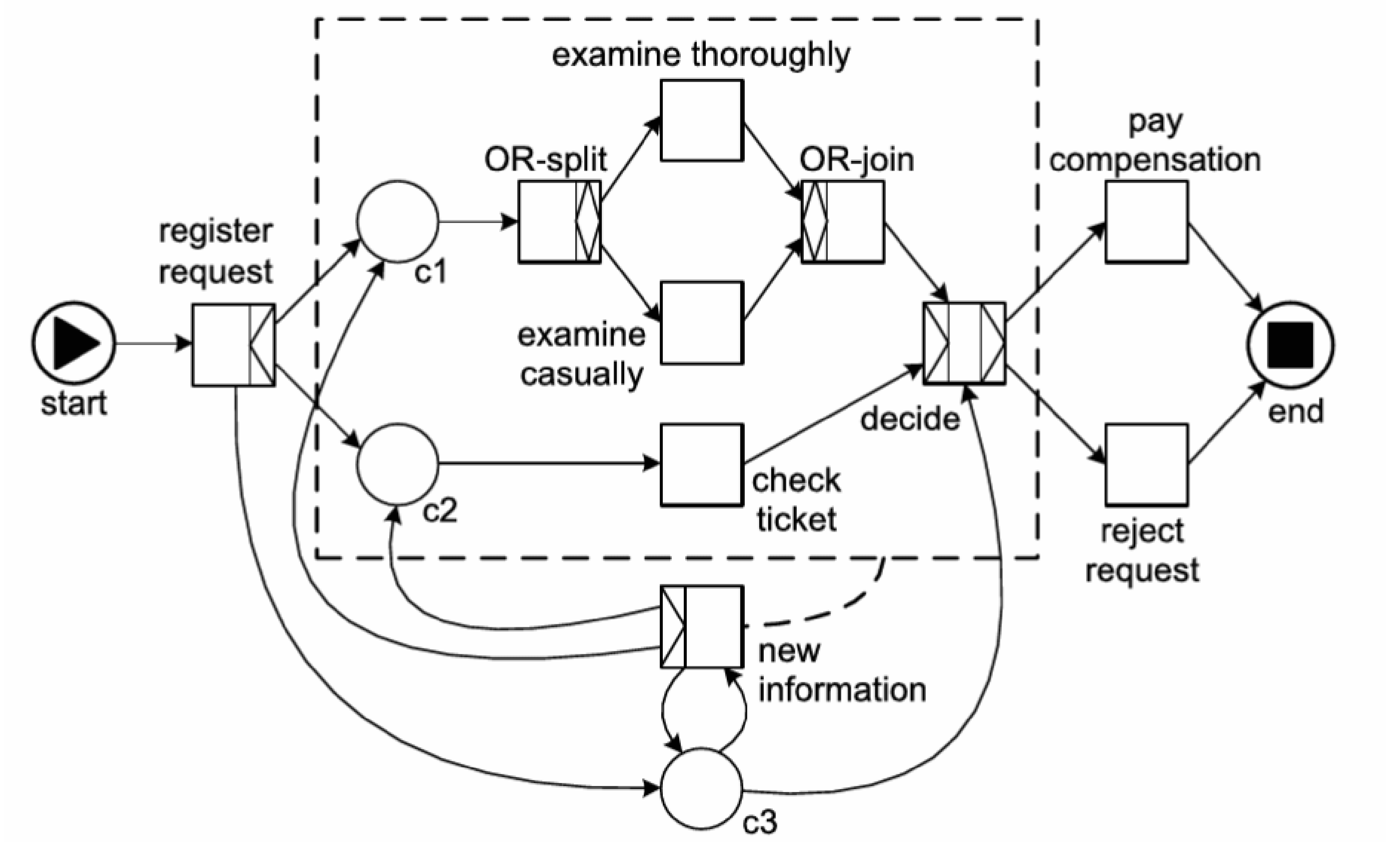
Graphical notation is intuitive but also executable



Graphic from van der Aalst, 2016

YAWL

Yet Another Workflow Language



Graphic from van der Aalst, 2016

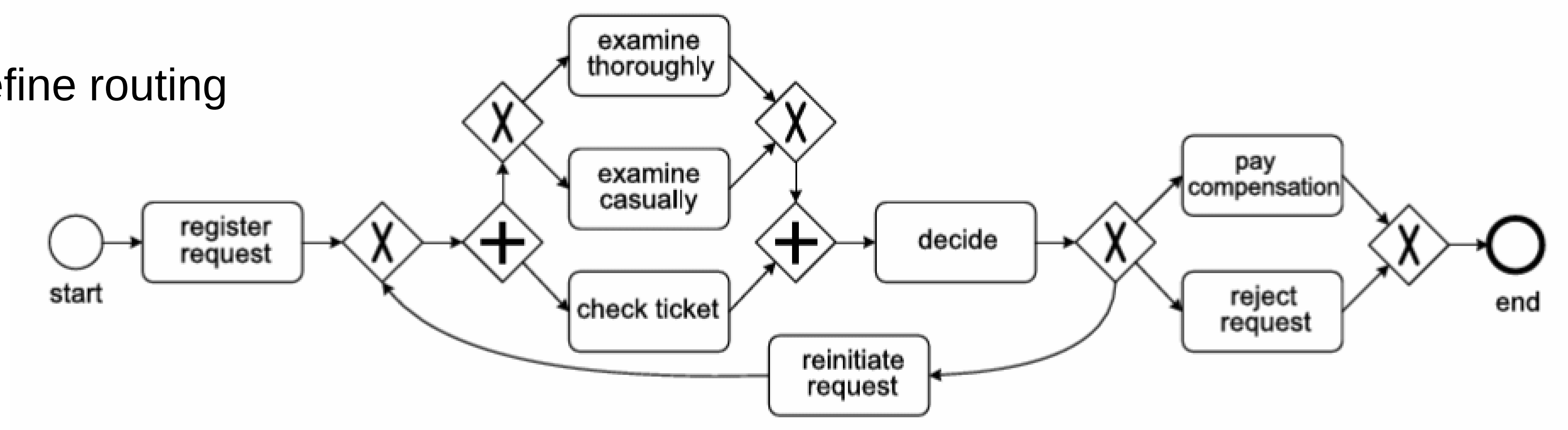
BPMN

Business Process Modelling Notation

One of the most widely used language to model business processes

Atomic activities are tasks

Uses *gateways* to define routing



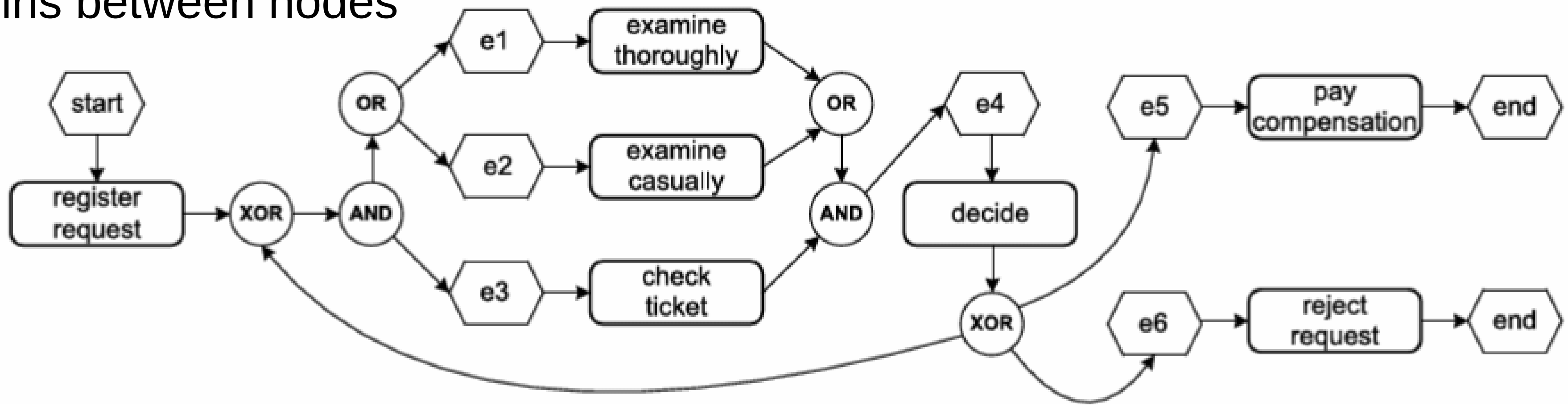
Graphic from van der Aalst, 2016

Event-Driven Process Chains (EPCs)

Supported by ARIS and SAP R/3

Bipartite graph: Events are followed by functions, which are followed by events again

Can also have splits and joins between nodes



Graphic from van der Aalst, 2016

Process Mining

Typically, Processes are identified by Business Analysts using interviews and similar techniques

Error-prone, time-intense and costly

Process Mining uses event data by IT systems to discover processes

Discovers real & correct process models

Discovers process highways

Able to understand complex systems

Tools

ProM

- Free
- Research-driven
- Not the most user-friendly tool

Disco

- Commercial
- Simple, good for ad-hoc process analyses

Celonis

- Commercial
- Industry leader
- Fully fledged Process Mining suite

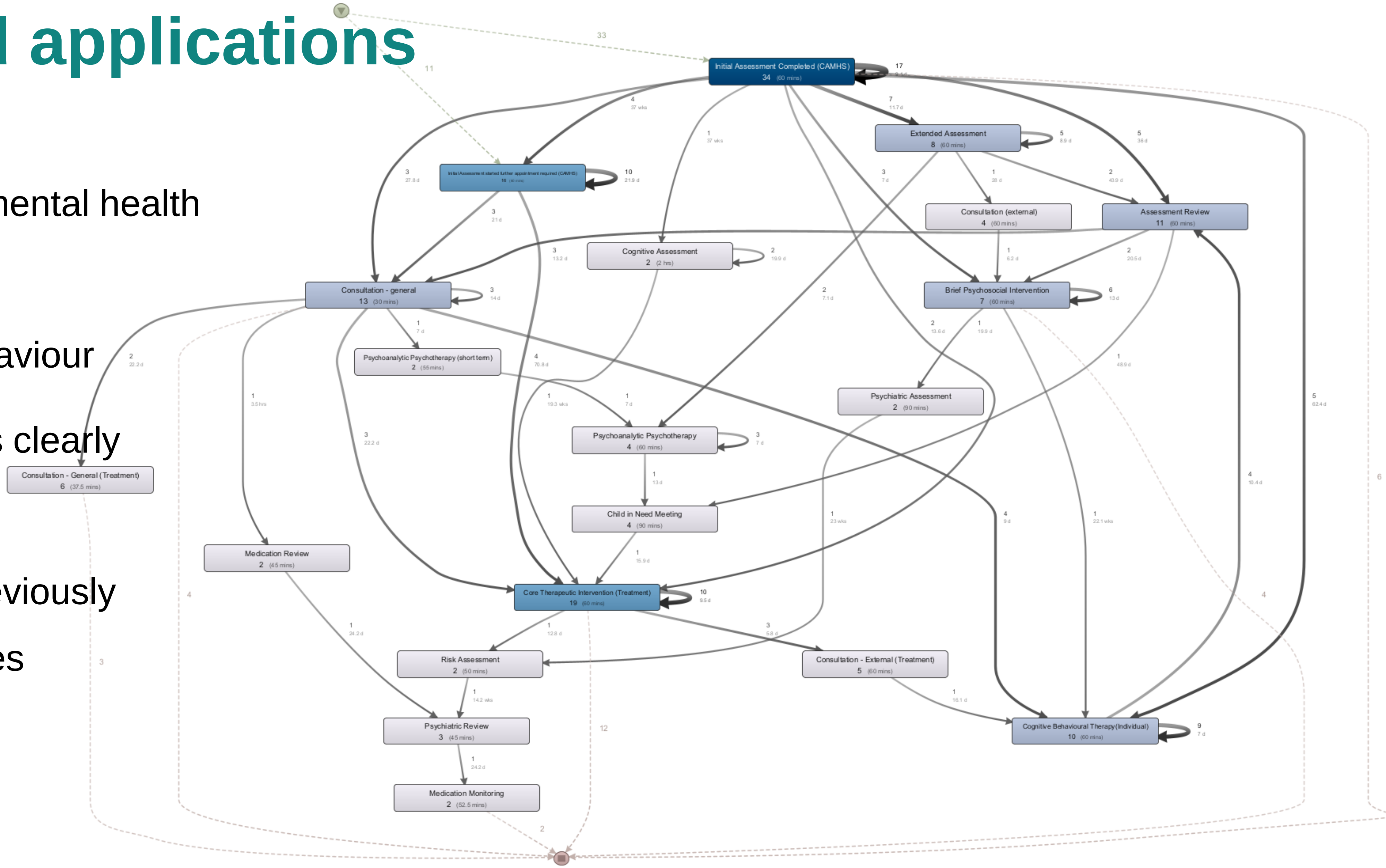
Real-world applications

Care pathway in a mental health care setting

Highly complex behaviour

Most common paths clearly identifiable

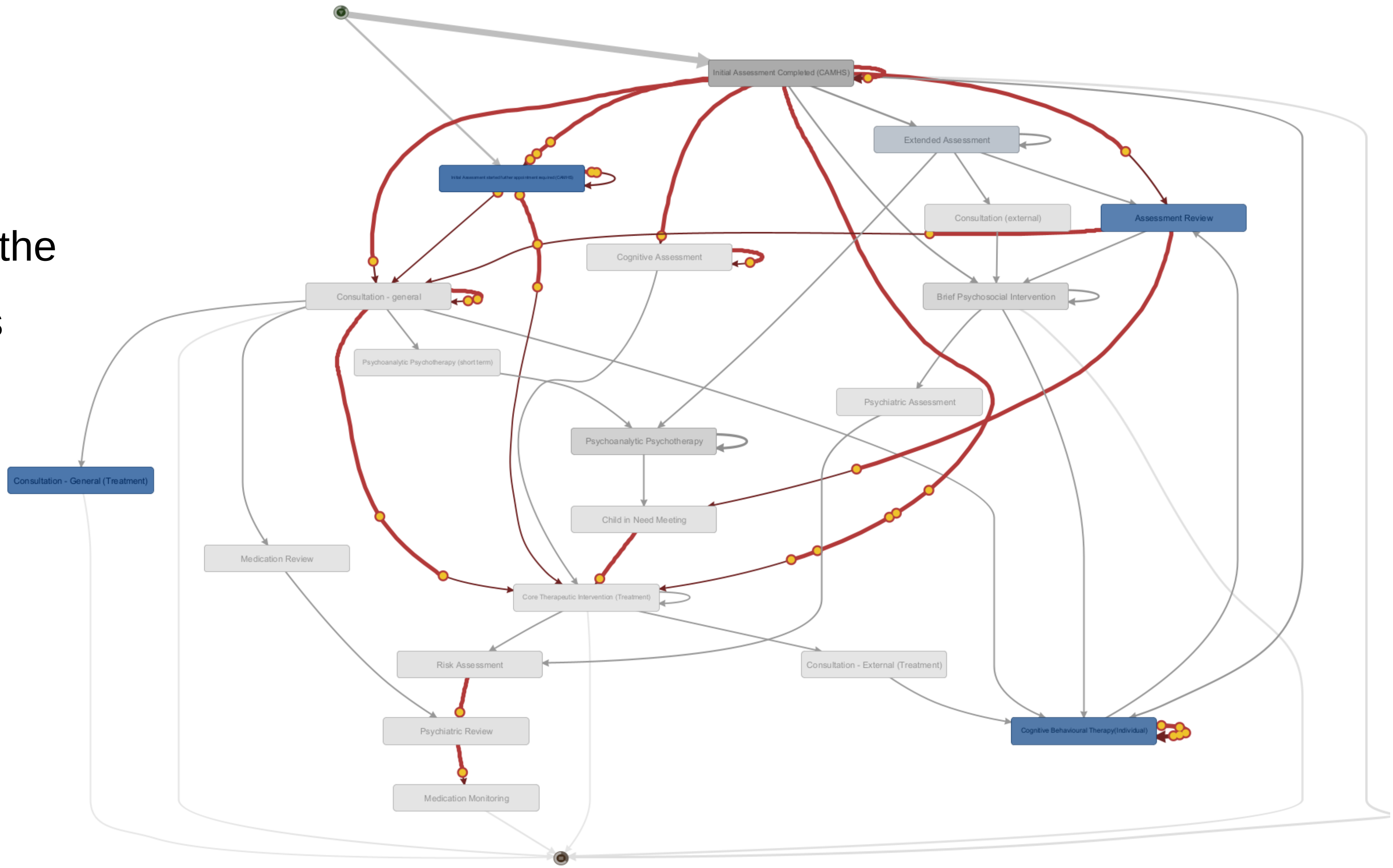
Not as formal as previously mentioned languages



UNIVERSITY OF WESTMINSTER

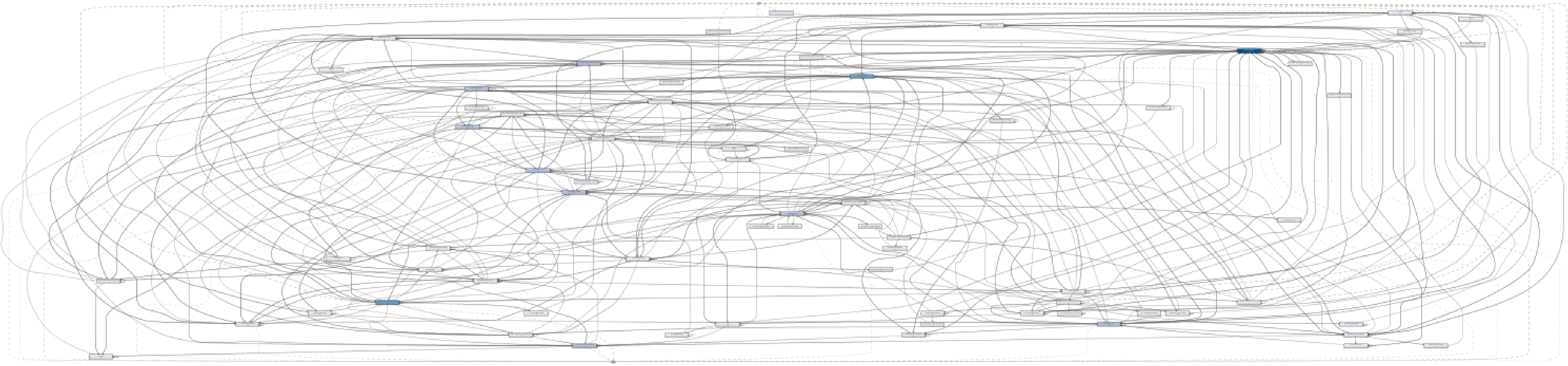
Animate it!

Animation often helps with the understanding of a process



UNIVERSITY OF WESTMINSTER

Keep it simple!



UNIVERSITY OF WESTMINSTER

Visualising Text Data

UNIVERSITY OF
WESTMINSTER

Why visualising text data?

Vast amount of text data (Web Pages, emails, SMS, progress notes, transcribed calls, etc.)

To large to read all of it

Visualising text data can give a quick insight

A quick introduction to Text Mining

Text data needs to be transformed to be analysed.

The most commonly used model is the bag-of-words model.

We simply count the occurrence of each word.

Consider the sentence “John likes to watch movies. Mary likes movies too.”

This can be represented structurally as a bag of words:

$$BoW_1 = \{\text{John}: 1, \text{likes}: 2, \text{to}: 1, \text{watch}: 1, \text{movies}: 2, \text{Mary}: 1, \text{too}: 1\}$$

We may choose to ignore “stop-words”, such as “to”, “too”, etc.

Multiple documents form a Document Term Matrix

Term	Doc 1	Doc 2
John	1	0
Likes	2	1
To	1	1
Watch	1	1
Movies	2	0
Mary	1	1
Too	1	0
Also	0	1
Football	0	1
games	0	1

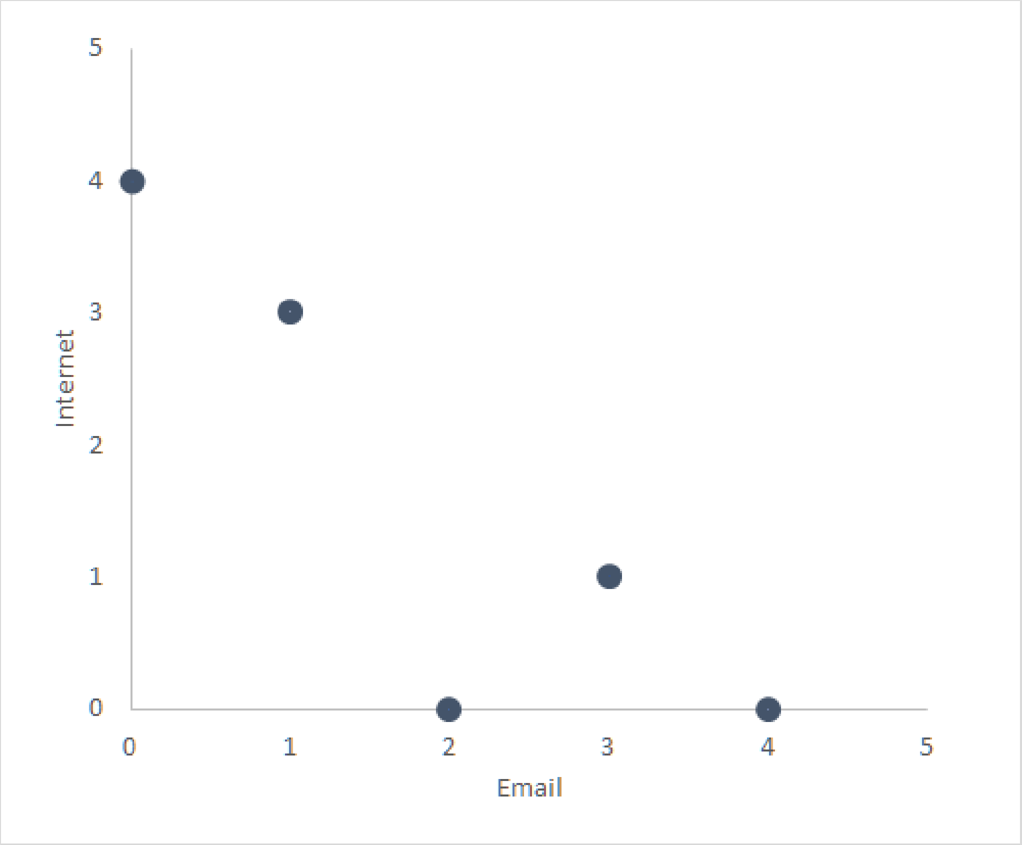
UNIVERSITY OF WESTMINSTER

Working with text data: Challenges

Dimensionality: Each term is its own dimension

Sparsity: Many documents don't use all terms

Abstraction



Word variations

Stemming: truncate a word to leave only its stem.

Porter stemmer is the commonly used algorithm for English

Lemmatisation: uses a dictionary to look up the basic form of a word

```
> SnowballC::wordStem(c("are", "be", "walked", "walking"), "porter")
[1] "ar"    "be"    "walk" "walk"
> textstem::lemmatize_words(c("are", "be", "walked", "walking"))
[1] "be"    "be"    "walk" "walk"
>
```

Word count vs significance

Most basic Bag-of-Words model counts word occurrence (Term Frequency)

We often prefer to work with the term significance.

Tf-Idf can be used to calculate term signicifance

Term-frequency $tf_{t,d}$: Number of occurrences of a specific term

Document-frequency df_t: Number of documents a specific term occurs

Inverse-document-frequency $idf_t = log\left(\frac{N}{df_t}\right)$

The rarer the term, the higher is the idf, the more frequent the term, the lower it is

Tf-idf: $tf\text{–}idf_{t,d} = tf_{t,d} \times idf_t$

Word Cloud

Simplest way to visualise a document or a corpus

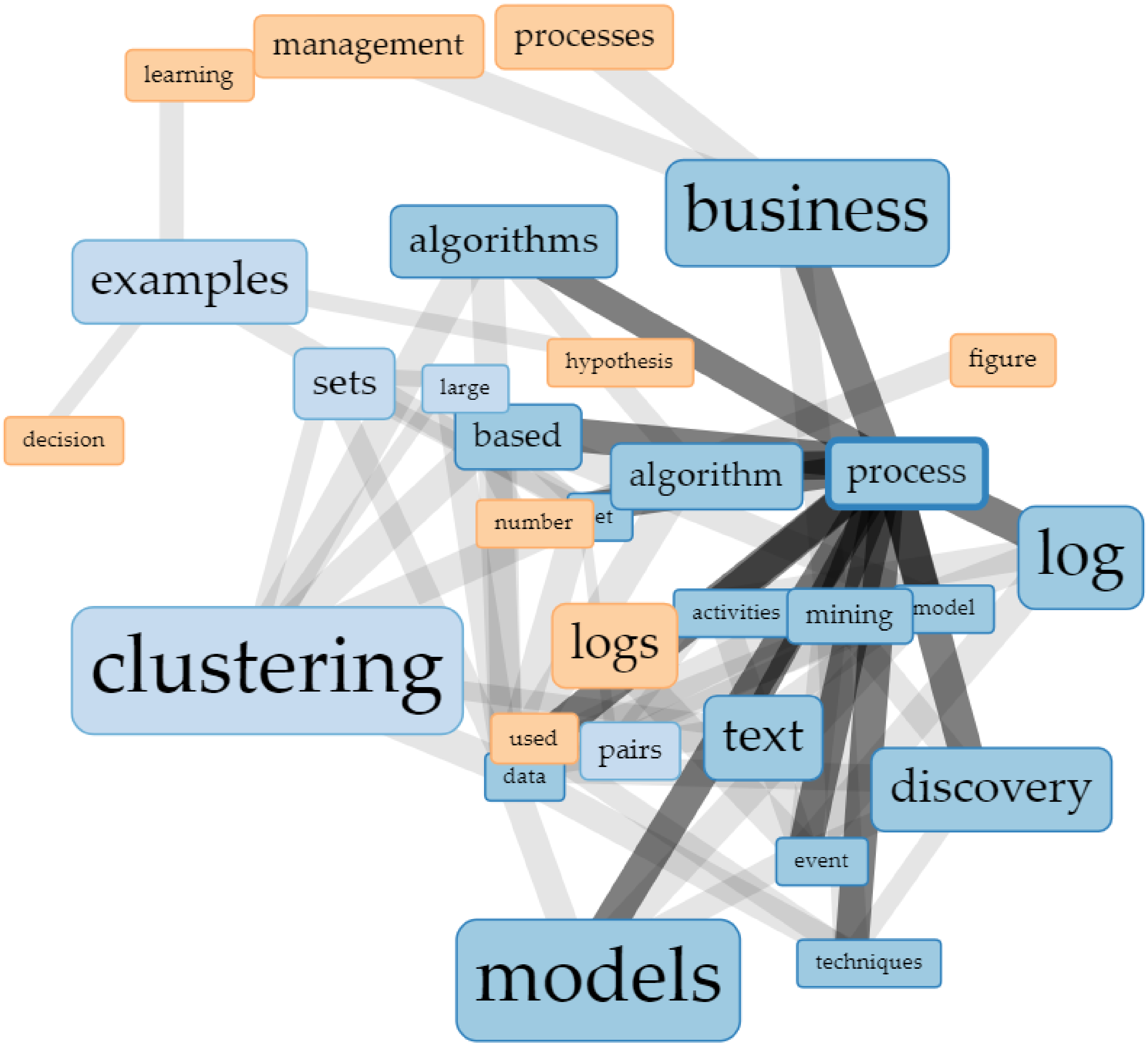
Shows most important/frequent terms bigger

Cloud shows the terms in the presenters' paper library and is generated by [Voyant Tools]

UNIVERSITY OF WESTMINSTER

Word nets

Which terms occur together



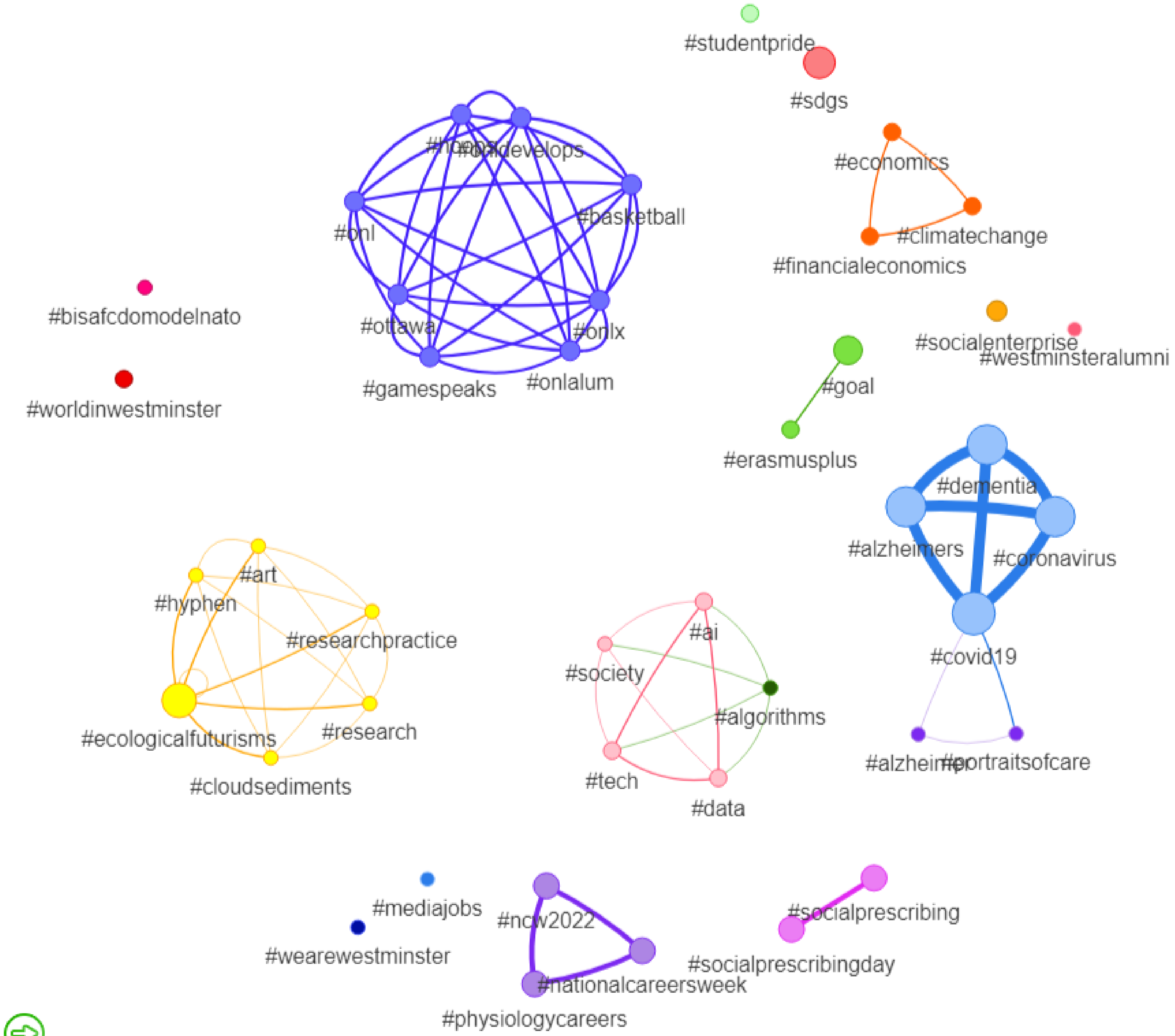
UNIVERSITY OF WESTMINSTER

Hashtag map

Shows co-occurring hashtags

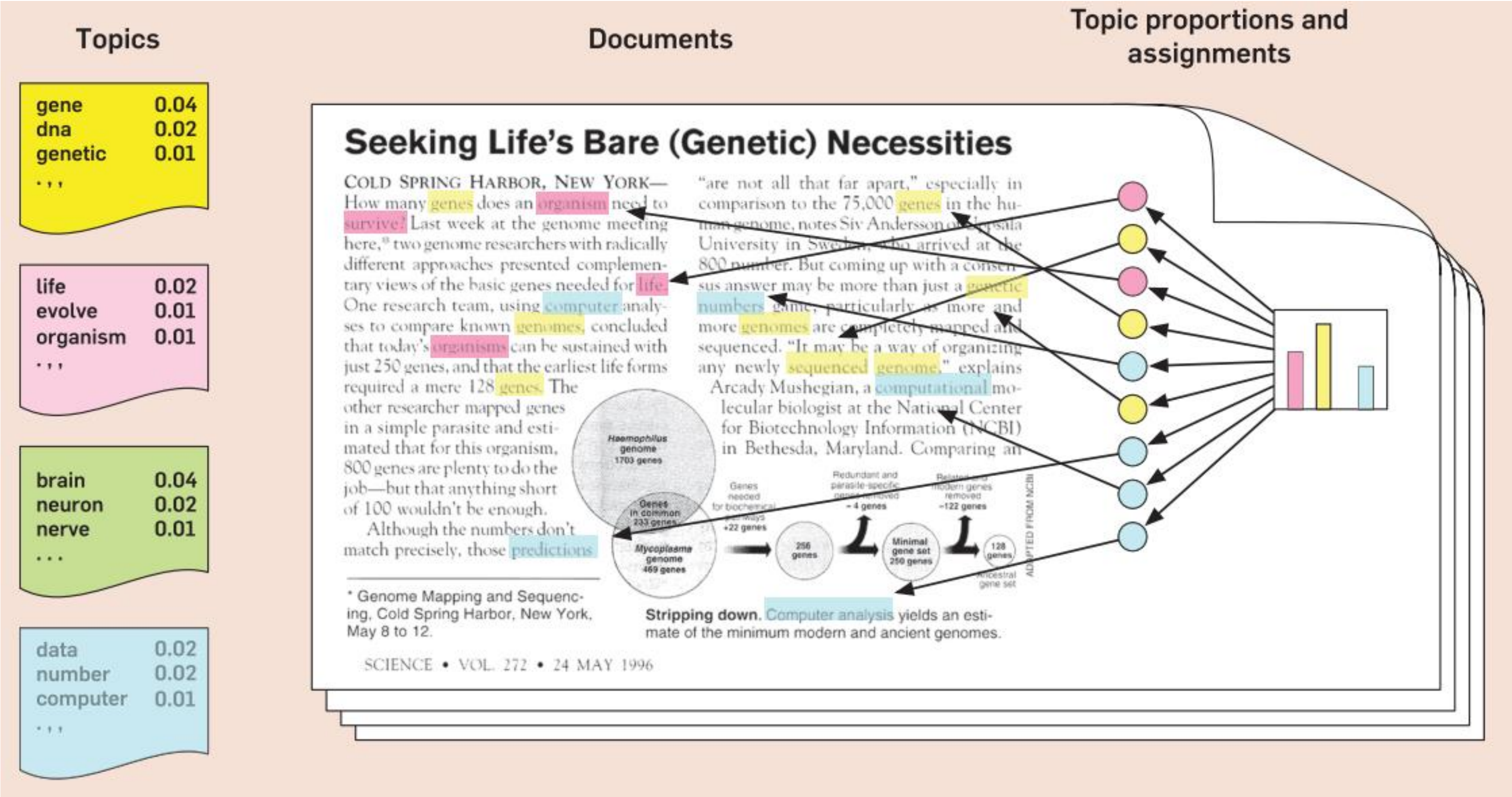
Useful to identify hashtags

Useful to identify similar topics



UNIVERSITY OF WESTMINSTER

Beyond terms: Topic modelling



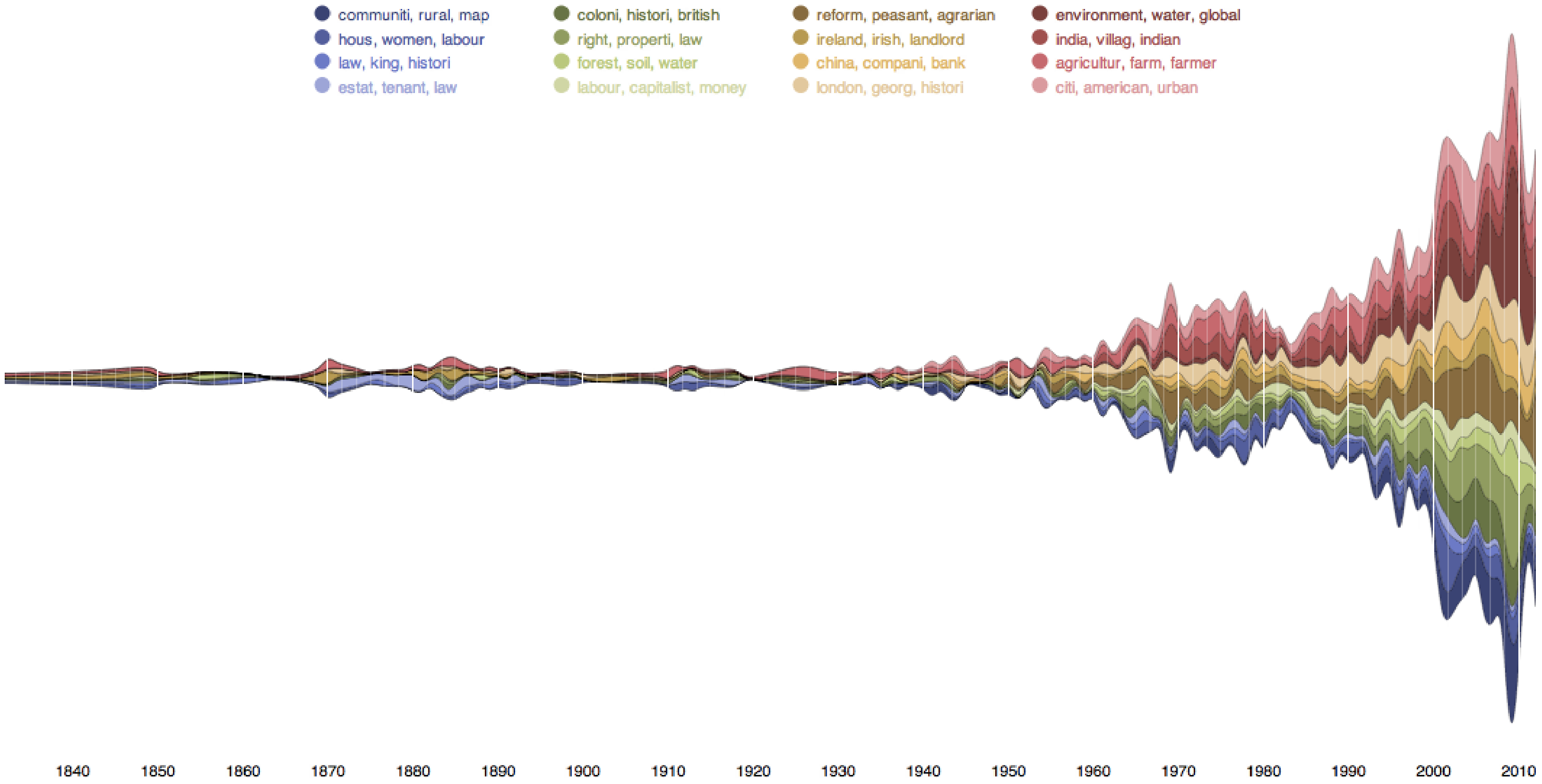
A topic is a collection of weighted terms

A document is a collection of weighted topics

A term is a collection of weighted topics

Blei, D., Carin, L. and Dunson, D. (2010) 'Probabilistic topic models', *IEEE Signal Processing Magazine*, 27(6), pp. 55–65. doi: 10.1109/MSP.2010.938079.

Paper topics over time



Stream graph shows number of papers by topic over time

Topics are identified by the most important terms

http://journalofdigitalhumanities.org/2-1/review-papermachines-by-adam-crymble/

Please remember to submit the Student Module Evaluation!

UNIVERSITY OF WESTMINSTER

Further Reading

Aalst, W. van der (2016) Process mining : data science in action. Second edition. Berlin, [Germany] ;: Springer.

Manning, C. D. et al. (2008) Introduction to information retrieval. Cambridge: Cambridge University Press.

UNIVERSITY OF WESTMINSTER