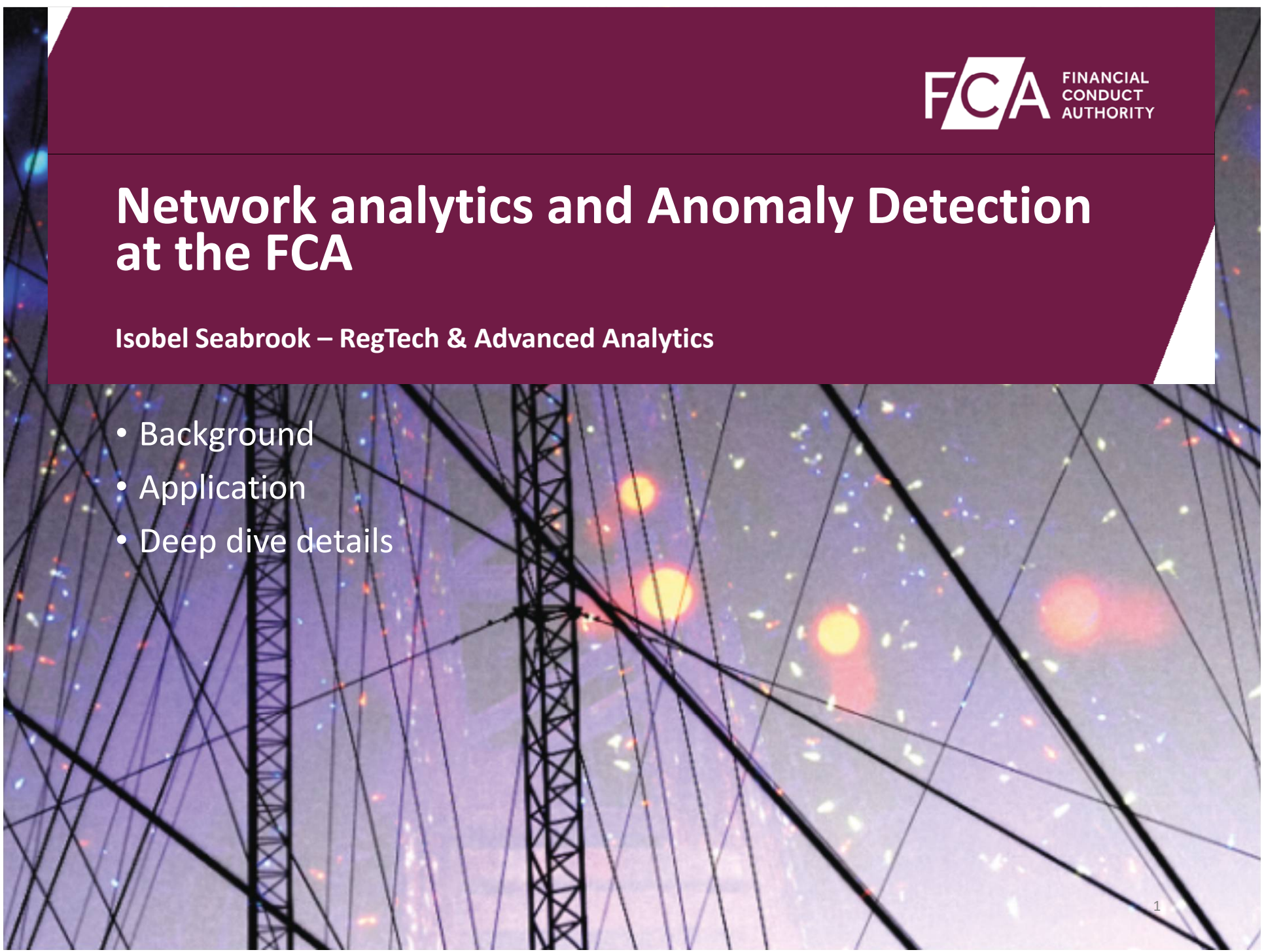
Network analytics and Anomaly Detection at the FCA

Isobel Seabrook – RegTech & Advanced Analytics



- Background
- Application
- Deep dive details

Why Advanced Analytics?

AA Team Timeline

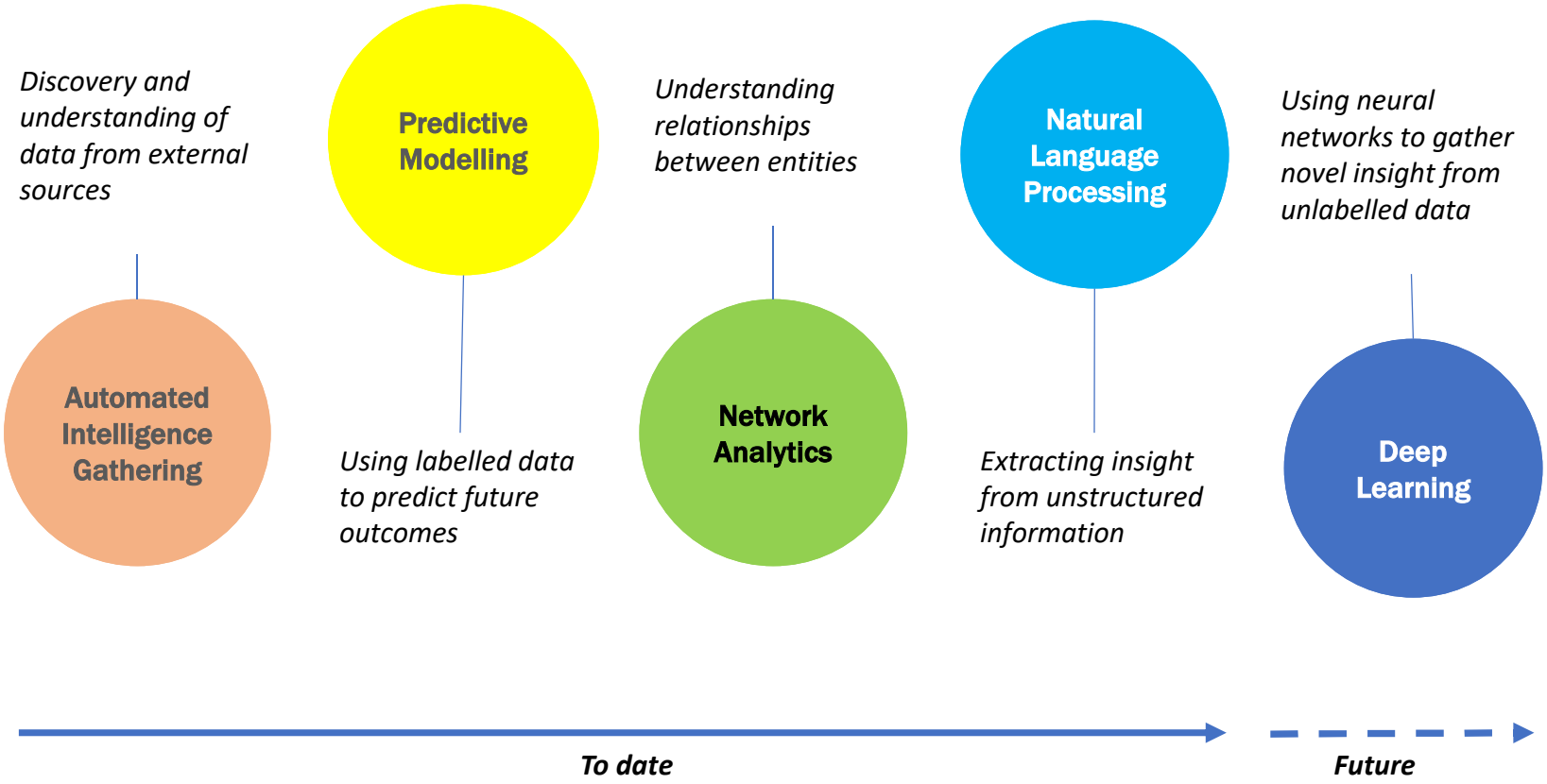
October 2017
Advanced Analytics Team set up
3 permanent staff
3 consultants

January 2018
First permanent Data Scientist hired!
4 permanent staff
6 consultants

January 2019
All staff permanent
10 data scientists
3 data engineers
2 Bas/PMs

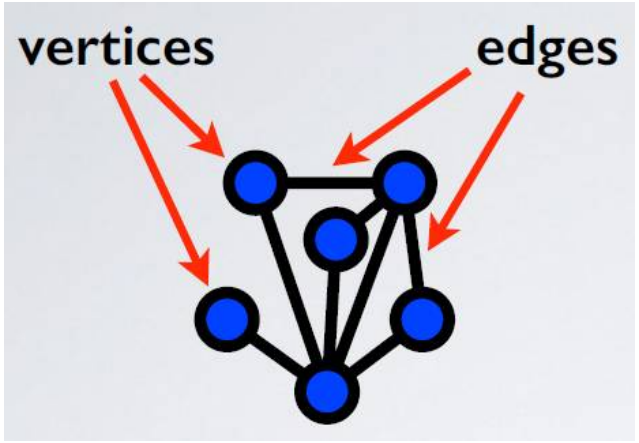
April 2019 –
Build out of hub and spokes
e.g. 20 data scientists

Overview of key techniques



What are Networks?

- Collection of vertices and edges
- Vertices are connected to each other by edges
- The whole system can be described mathematically by an adjacency matrix **A**
- **A** can be weighted or unweighted



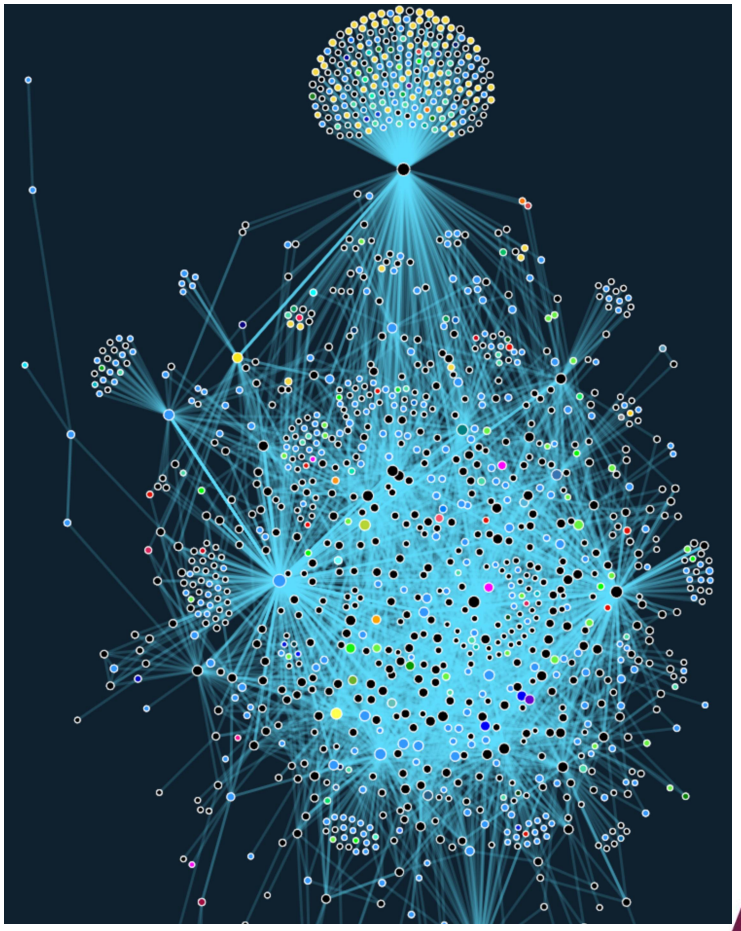
$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Social Networks

- Vertices are people, edges are their relationships or communications.
- Why we are interested: identification of hot topics → targeted advertising and/or regulation, control of fake news, identification of influencers

Transaction networks

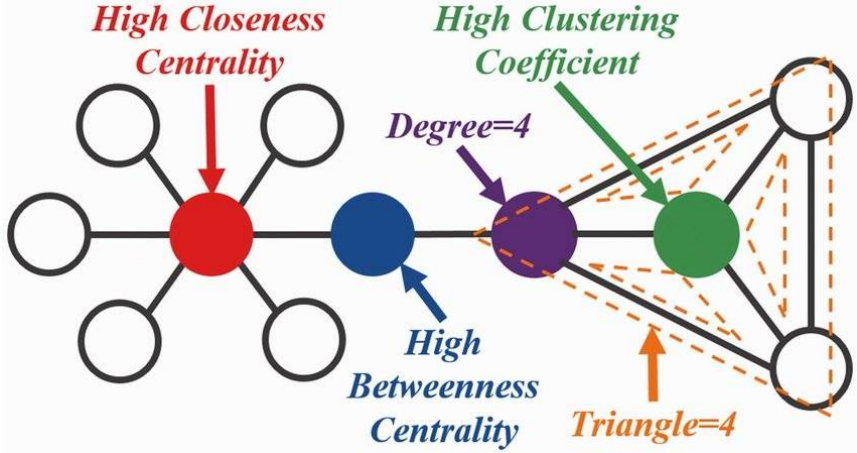
- Vertices are individuals/institutions
- Edges are the trades - movement of money or product, can be weighted by price or amount
- Can be very complex – multi-layer, multipartite (different types of vertices), multi-edge, and very large e.g. billions of transactions in the UK capital markets
- Can vary over a huge range of different time scales – from high frequency trading to trading in illiquid products

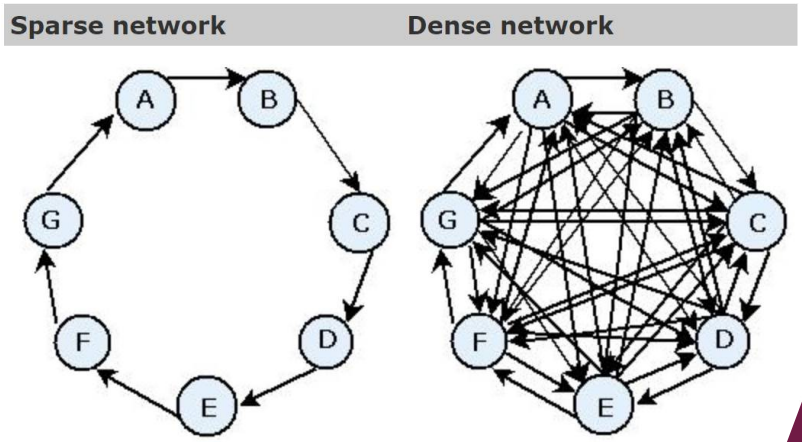


Example of a bitcoin transaction network

Describing networks

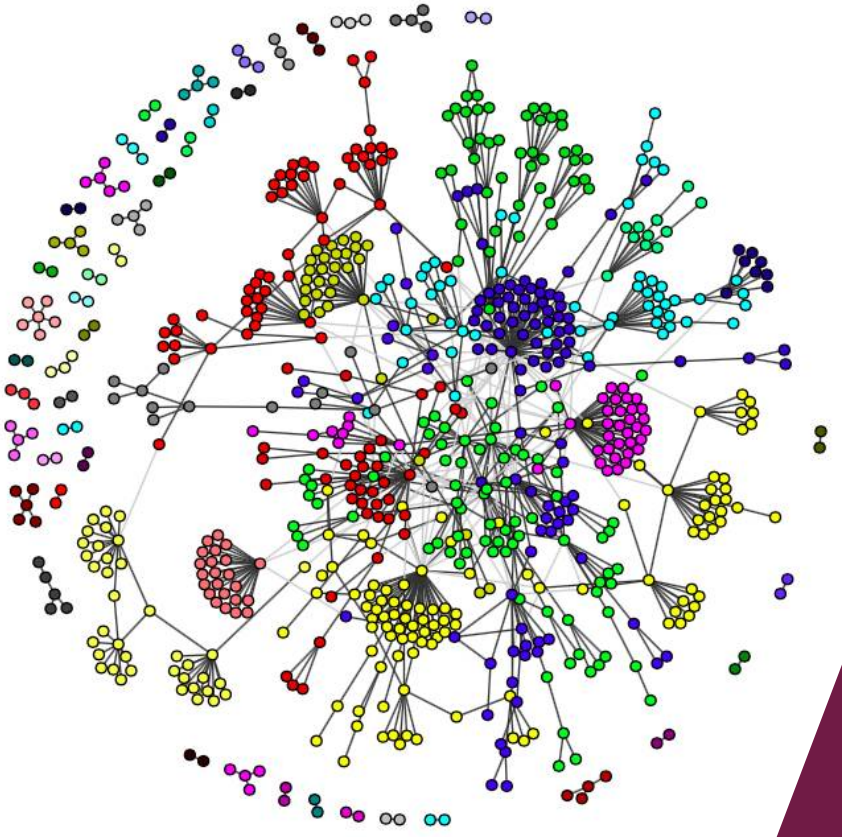
- Questions to address:
  1. How are the edges organised?
  2. How do the vertices differ?
  3. Do geographic properties matter?
  4. Are there underlying patterns?
  5. What processes shape the network?
- Degree: number of connections (in- or out- for directed) per node
- Network density: represents the proportion of possible relationships in the network that are actually present
- Centrality: 'importance' score for each node, based on degree, betweenness, closeness etc.





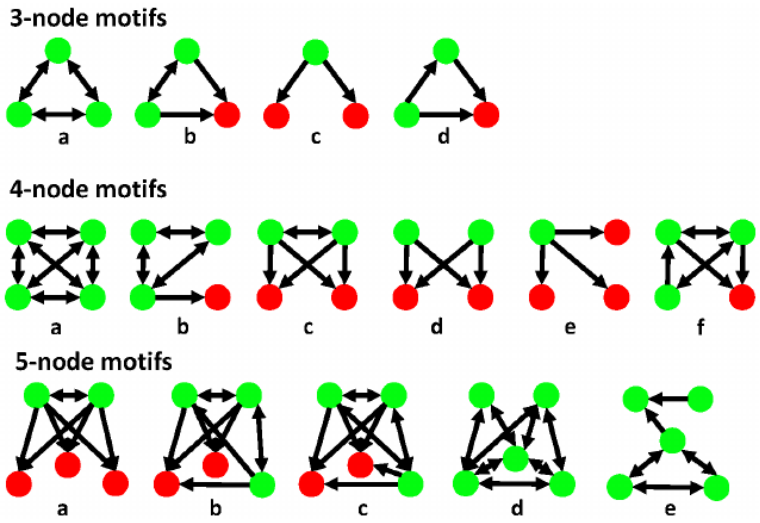
Community detection

- Connected components: groups of vertices that are all reachable from each other
- Spectral methods: Look for clusters in the eigenvalue spectrum of the graph
- Statistical Inference: fit a network model on the data and look at partitions that are different from the model
- Dynamics: identify communities by running dynamical processes on the network e.g. random walks and finding where they spend a lot of time



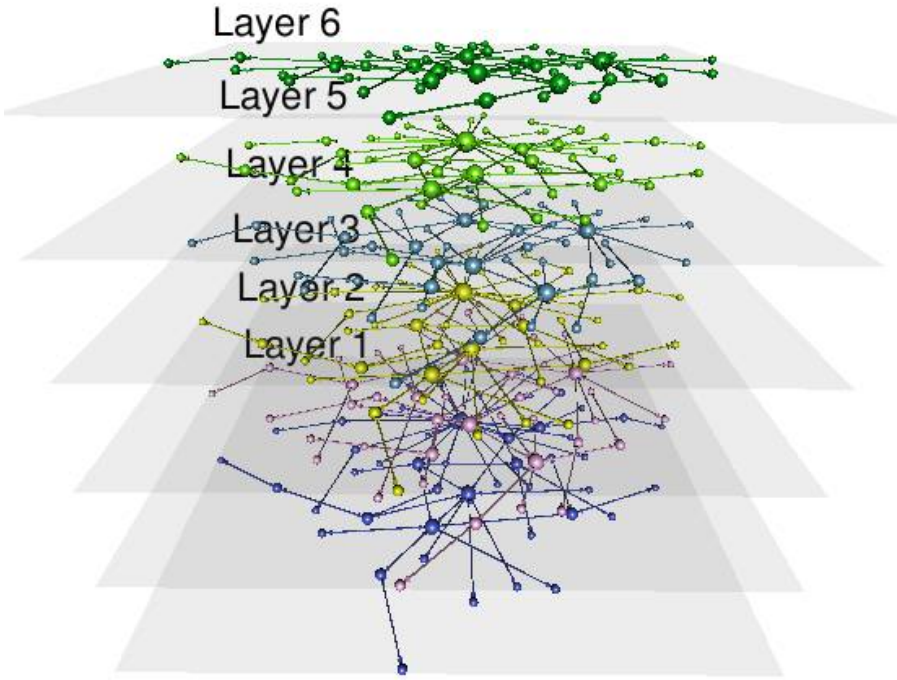
Pattern finding

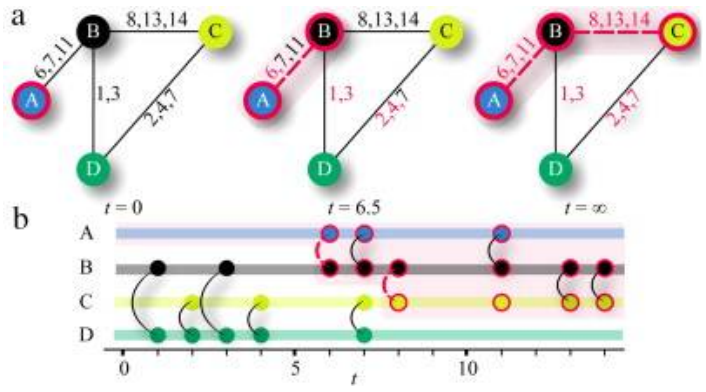
- **Network motifs** are recurrent and statistically significant sub-graphs or patterns
- Of notable importance largely because they may reflect functional properties
- Difficult to find: algorithms generally comprise of two main steps:
  1. Calculate the number of occurrences of a sub-graph and then,
  2. Evaluate the sub-graph significance.
- The recurrence is significant if a pattern appears far more than expected.



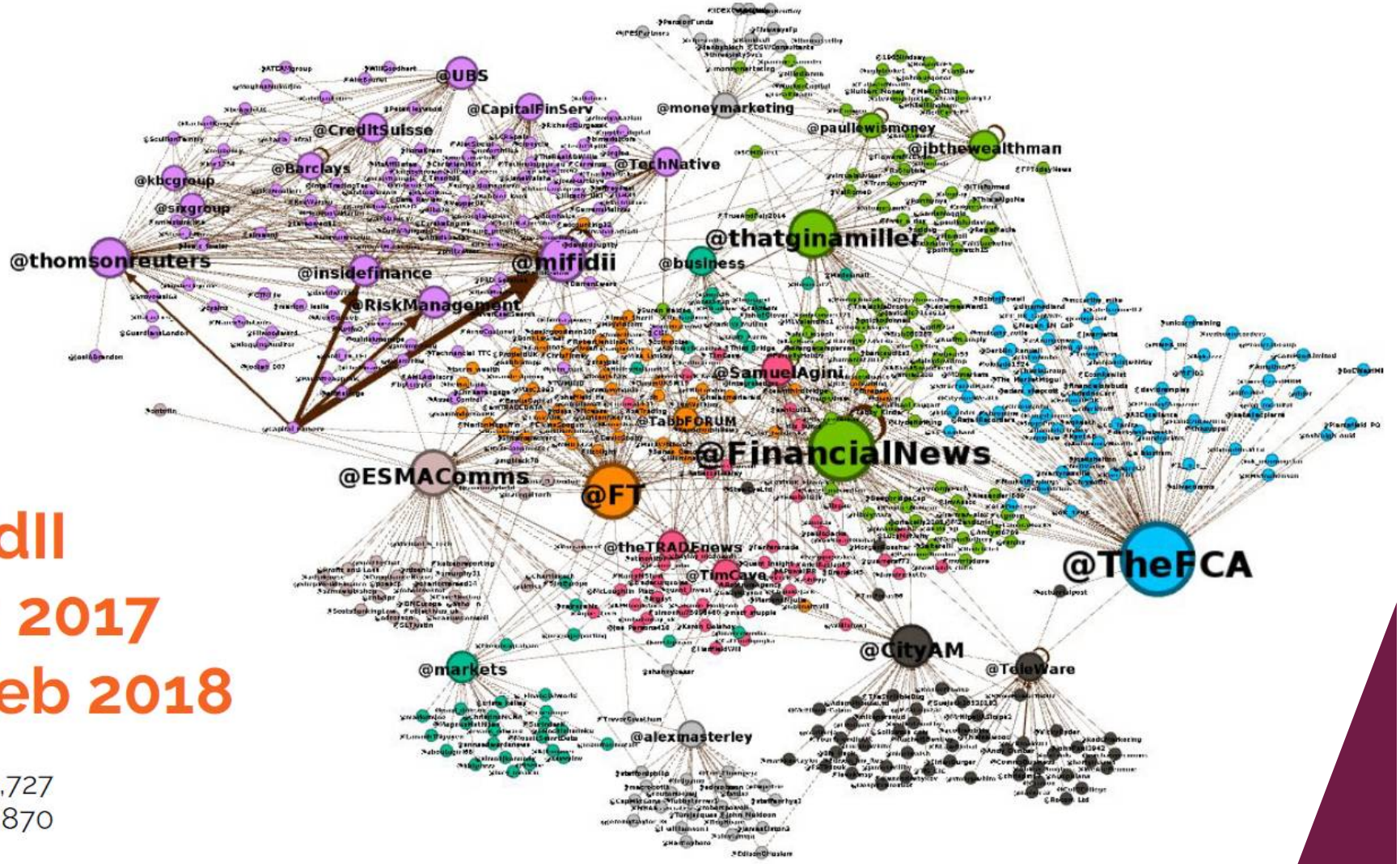
Multi-Layer and temporal networks

- Multi-layer – can have inter- and intra-layer links
- Temporal networks – relationships must obey the ‘arrow of time’
- Ordinary community detection algorithms and network descriptions need to be redefined for multi-layer and temporal networks.
- There are several different projection methods to condense these to simple (single layer) networks.





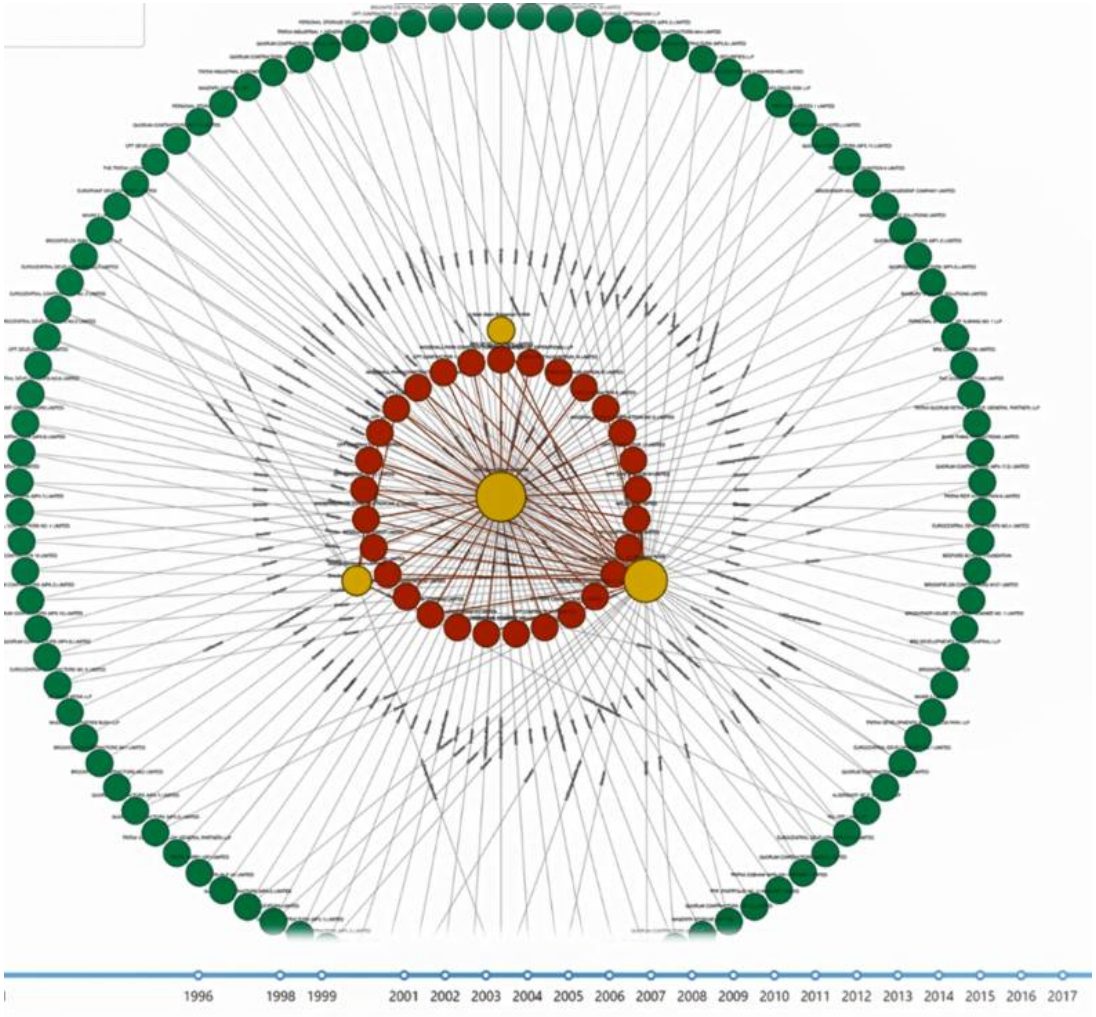
Example: Twitter Analysis



MifidII
Nov 2017
- Feb 2018

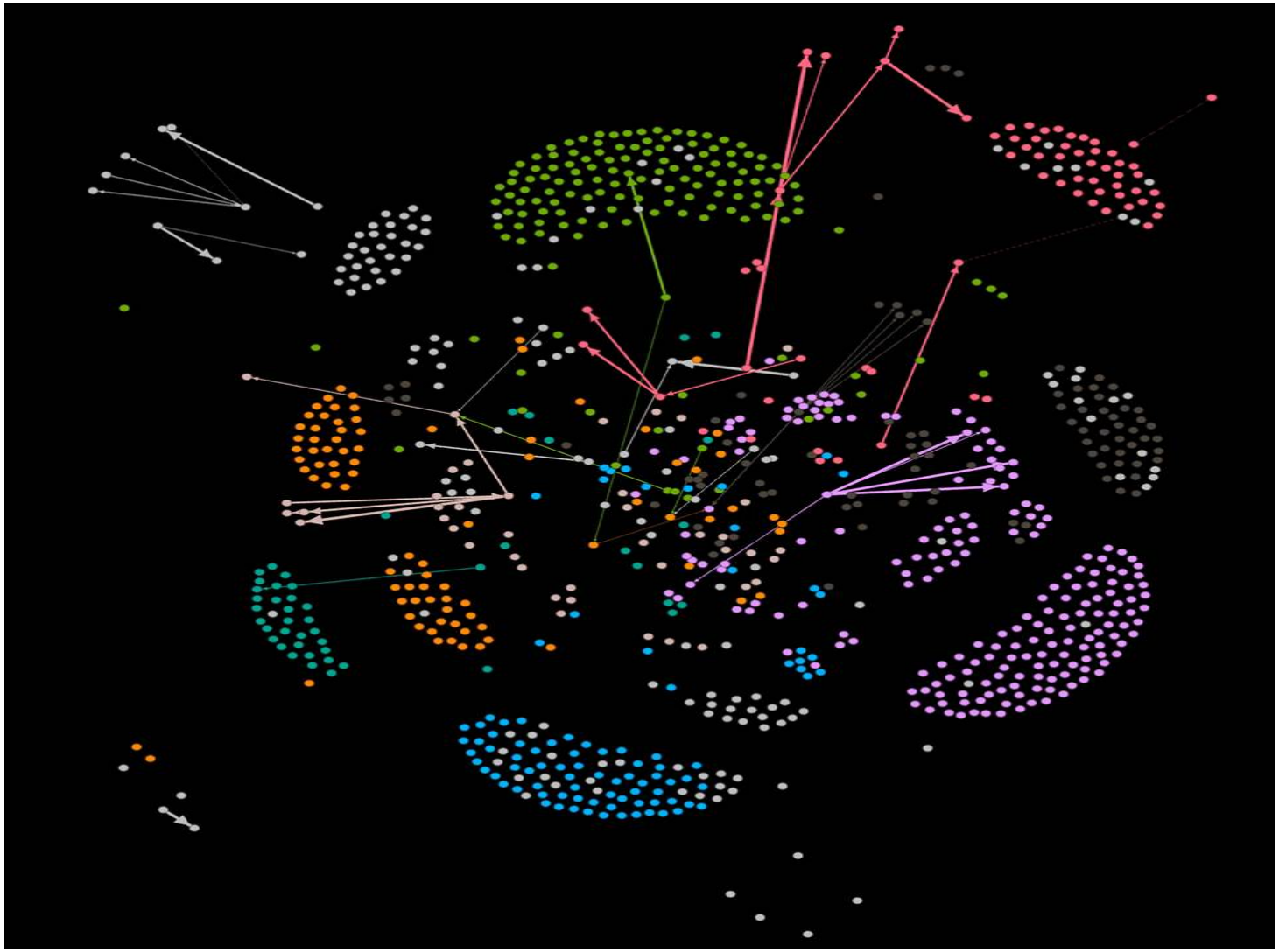
Nodes: 4,727
Edges: 5,870

Companies House Network Analytics



- For detection of phoenixing through flagging patterns of insolvency
- Companies House data - companies and their directors in an interconnected network.
- Reduce manual workload **and** provide valuable **new intelligence**

Example: Temporal transaction network



Part 2: Time Series Anomaly Detection

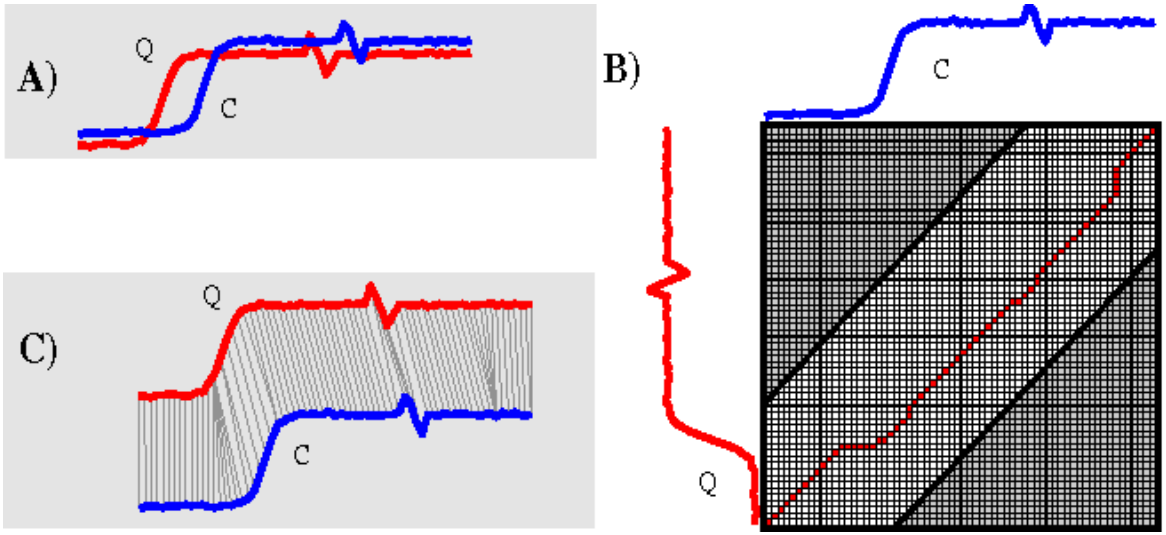
- Dynamic Time Warping
- Other anomaly detection techniques
- What can we find?

Our data

- Retail Mediation Activities return - regulatory return for firms who provide intermediary services arranging and/or advising on mortgages, non-investment insurance or investment products.
- Firms report minimum twice yearly.
- The dataset contains data for ~20,000 firms for ~60 variables, dating back as early as 2005.
- Data can be highly non-linear and may contain anomalous data points.

Dynamic Time Warping

- Dynamic time warping looks at how the time series can be locally stretched to optimize a global fit.
- This means that similar shapes can be matched, even if they have a time-phase difference.
- DTW is used to calculate a distance score between each time series with every other time series in a dataset → distance matrix.
- The distance matrix used as input for clustering and outlier detection algorithms.

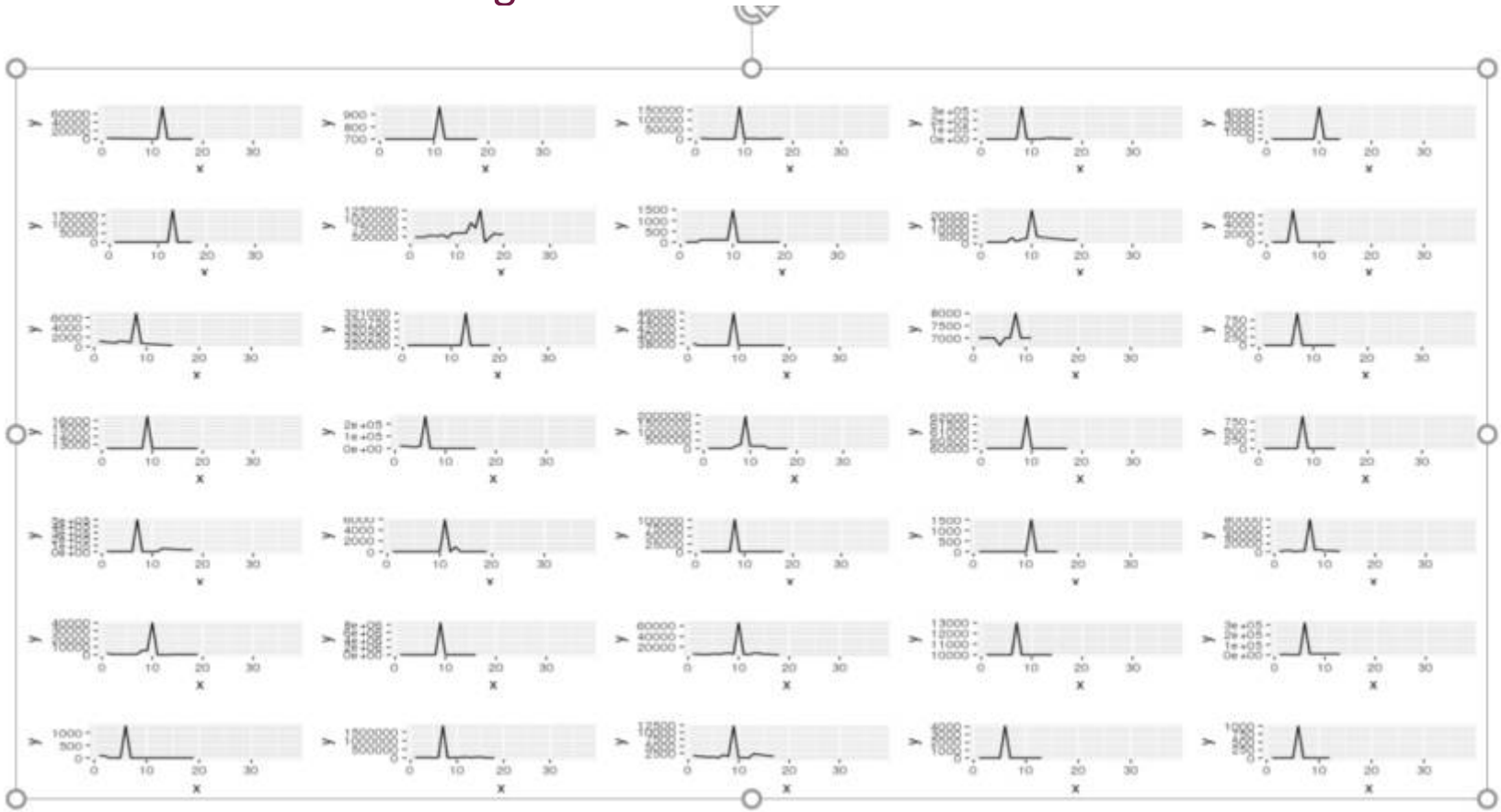


Other anomaly detection methods

- Alternative similarity metrics: Euclidean, Jaccard, Malanobis
- Similarity metrics → outlier detection (or clustering): LOF, DBSCAN, K-medoids clustering
- Isolation forest on firm level features (e.g. mean, variance, max value)
- Random forest regression to predict a firm's latest return, in order to flag unexpected values

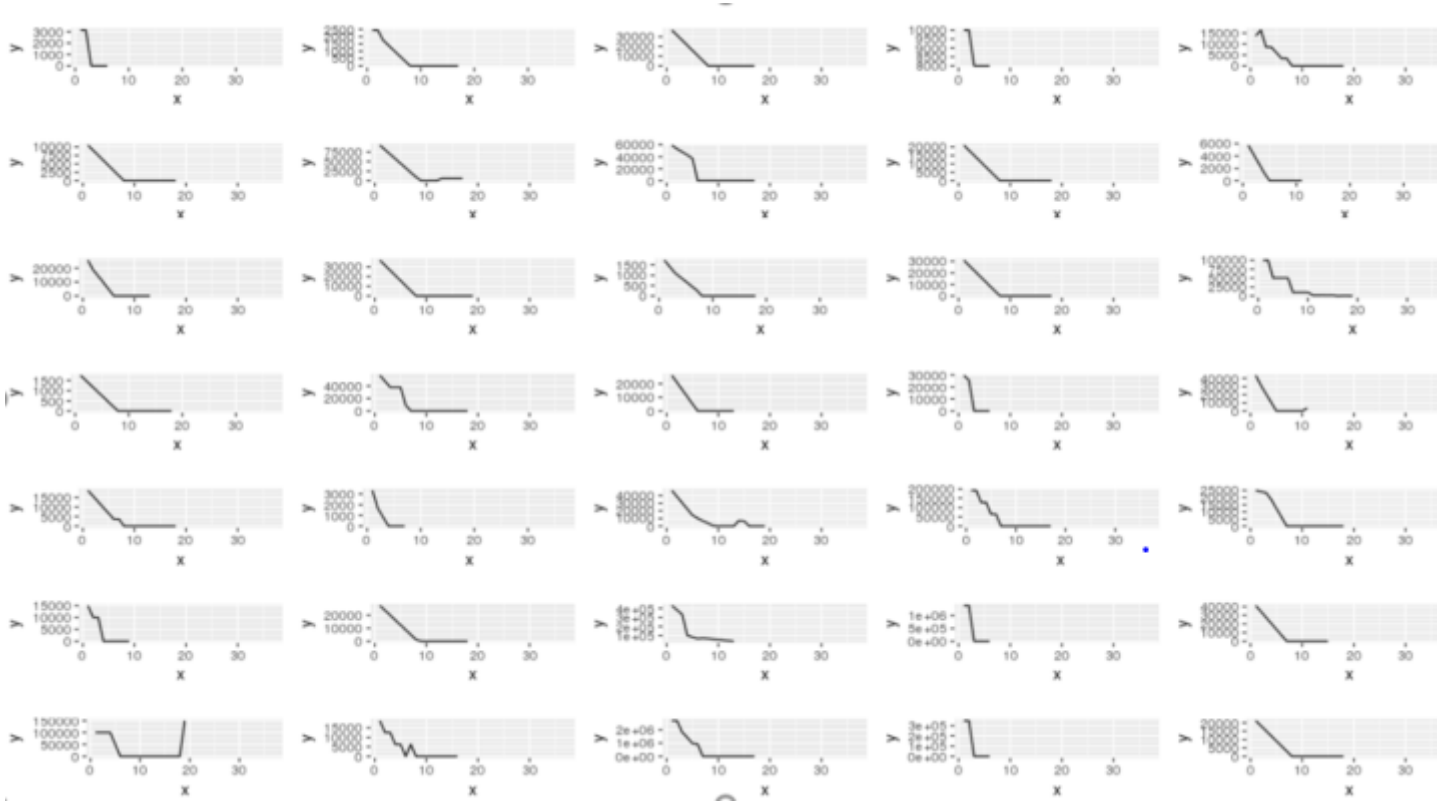
What do we want to identify?

- Similarity matrices and clustering: These algorithms will cluster the time series into groups with similar shapes. Some of these can be identified as containing outliers.



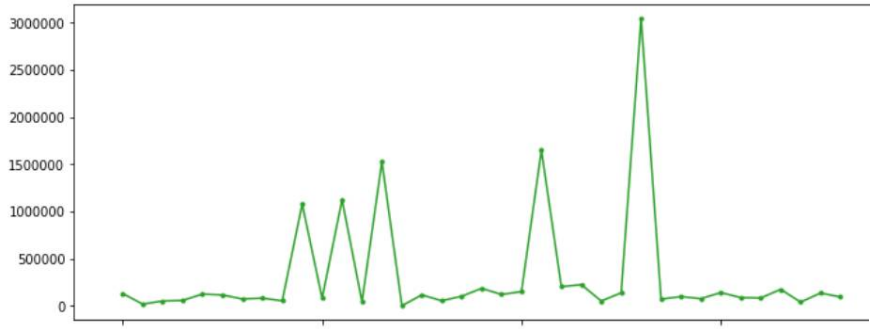
What can it identify?

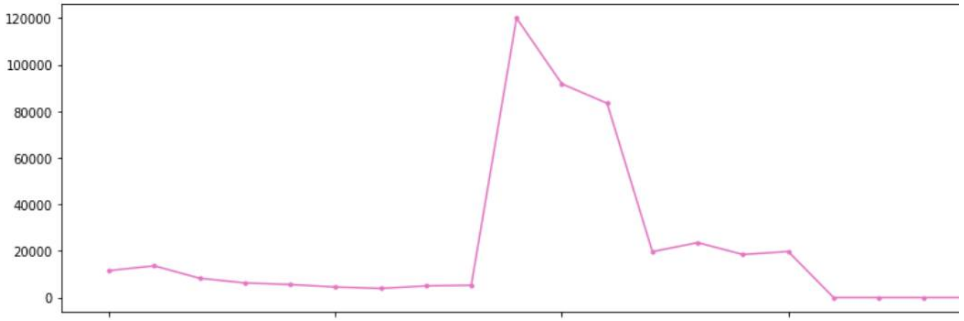
- We can also identify clusters of firms following similar trajectories. This could be used to select firms that demonstrate known problem trends.

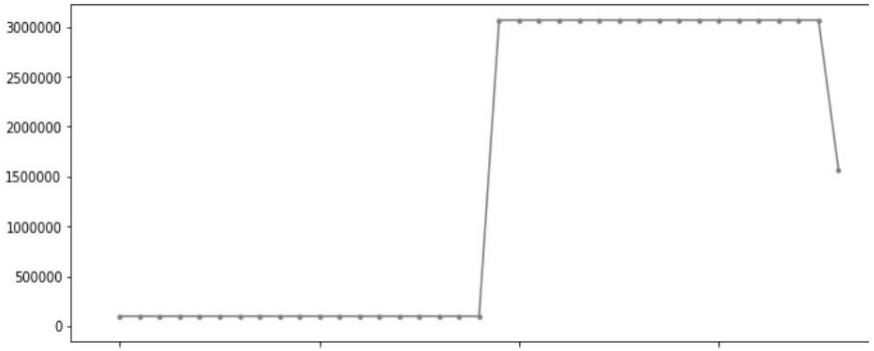


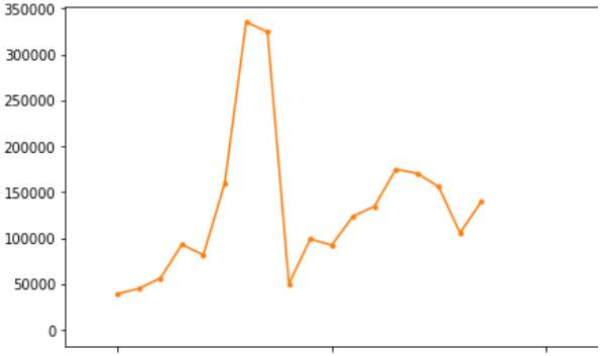
Further anomaly examples

- Unusual looking time series may be incorrect/concerning data, or may be as a result of business processes.
- Data currently unlabelled → need to be validated by SMEs
- Patterns may only be anomalous in the context of a multivariable analysis

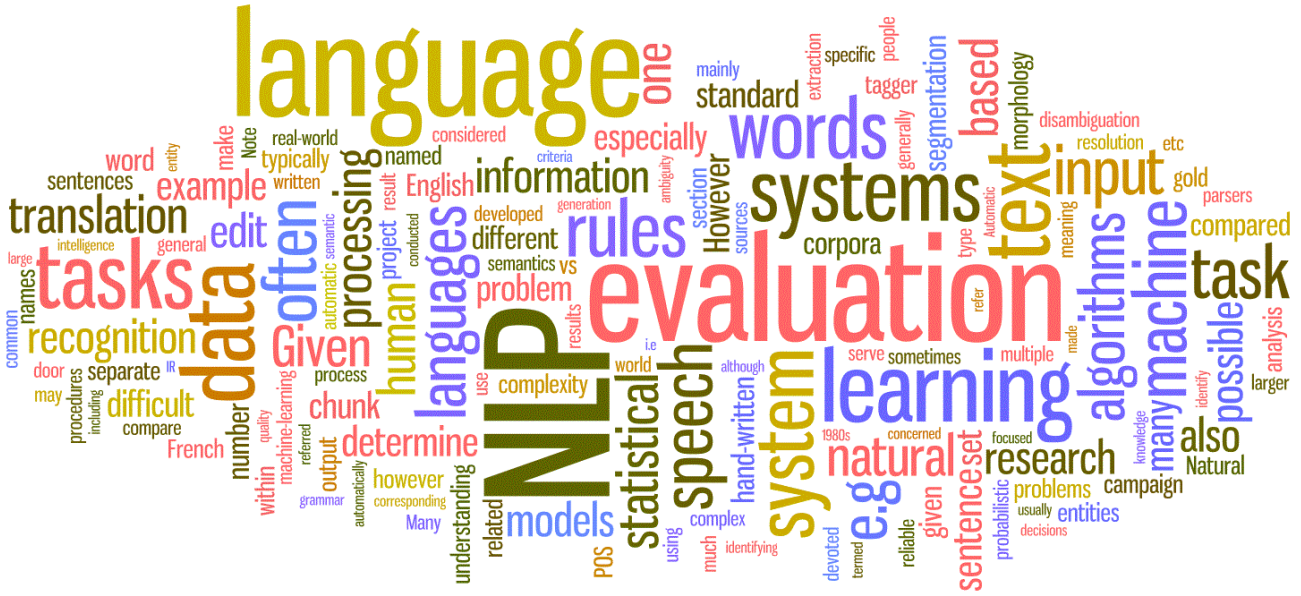








Natural Language processing (NLP) at the FCA



Marek Biernacki

What is NLP?

Natural language processing (NLP) is a sub-field of Machine Learning / Artificial Intelligence that focuses on enabling computers to understand and process human language.

NLP is everywhere even if we don’t know it.

Autocomplete

Better Search using NLP

Machine Translation

Chatbots / Personal Assistants

amazon alexa

Types of NLP

Speech Recognition—The translation of spoken language into text.

Natural Language Understanding—The computers ability to understand what we say.

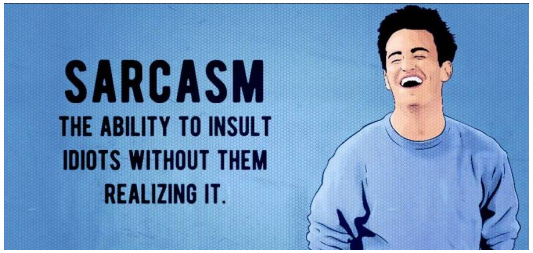
Natural Language Generation—The generation of natural language by a computer.

Why is NLP difficult?

Why is NLP difficult ?

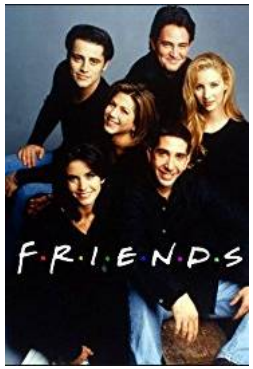
- **Context:** Pressing a suit

- **Sarcasm:** Yeah, right!



- **Understanding:** *common sense/knowledge*

- **Understanding *named entities*:** I'm going to watch friends later.



Why is NLP difficult ?

Errors: *"chicken" and "chiiccen"*

Neologisms: *youthquake*

Idioms*:* *Over the moon*

Ambiguity: *"Call me a cab!"*

Co-reference: Mentioning specifics earlier

Why is NLP difficult?: Syntactic and Semantic Analysis

Two main techniques that lead to understanding of language:

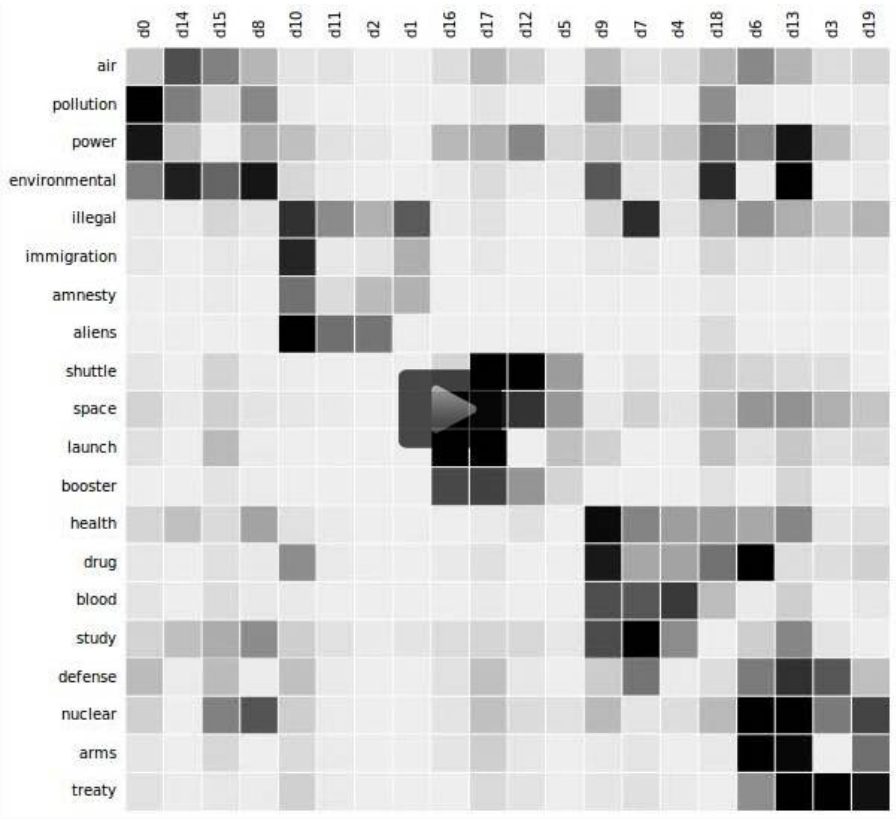
Syntactic analysis - grammatical structure of the text
Semantic analysis - the meaning that is conveyed by it

Example:
"cows flow supremely"

This is grammatically valid but does not make any sense.

Techniques to understand Text: Topic Modelling

- In NLP a **topic model** is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.
- If a document is about a particular topic, we would expect particular words to occur more frequently.
- “Dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats.

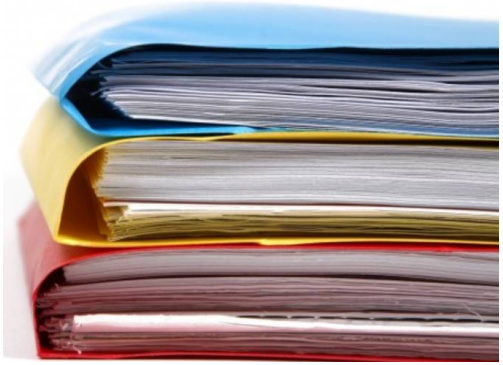


Project 1: Topic modelling

Project:

- Documents contain some categorical data, which is already collected and analysed.

- Additionally there free text answers in 2 sets of different documents (1424 of Document A, 1841 of Document B).



- Free text currently read and analysed manually, which is time consuming.

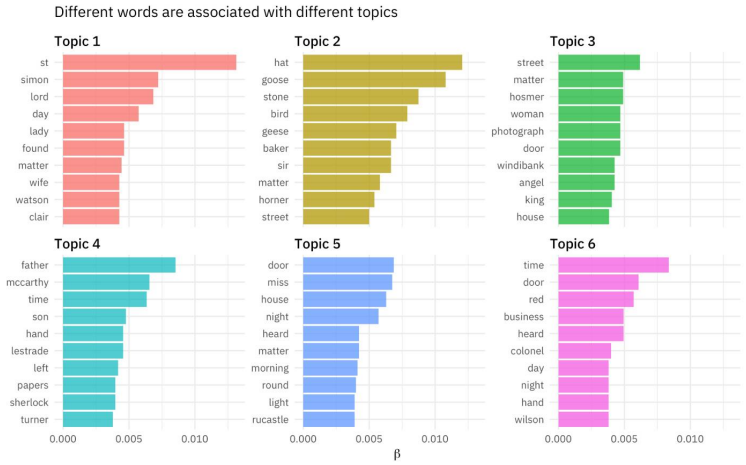
Project 1: Techniques considered

- **Latent Dirichlet Allocation (LDA) -** probabilistic model, and to obtain cluster assignments, it uses two probability values: P(word | topics) and P(topics | documents).

- **Non-negative Matrix Factorization (NMF)** - Linear-algeabreic model, that factors high-dimensional vectors into a low-dimensionality representation. Similar to Principal component analysis (PCA)

Project 1: LDA

- Unsupervised machine learning algorithm: Latent Dirichlet Allocation (LDA)
- Applied LDA to convert the documents into a set of topics
- Each document is represented as a distribution over topics
- Each topic is represented as a distribution over words



Techniques to understand Text: Sentiment Analysis

In Sentiment Analysis, we want to determine the attitude (the sentiment) of a person to an entity.

The sentiment is mostly categorized into positive, negative and neutral categories. With the use of Sentiment Analysis, we want to predict for example a customers opinion and attitude about a product based on a review he wrote about it.

Because of that, Sentiment Analysis is widely applied to things like reviews, surveys, documents and much more.

Project 2: Sentiment Analysis

Project:

- A dataset of 3500 documents, containing free text.

- Free text currently read and analysed manually, which is time consuming and boring.

- FCA team wanted to analyse sentiment of people to around 15 entities.

Project 2: Sentiment Analysis

- Extracted sentences containing 7 entities

- Used VADER (Valence Aware Dictionary and sEntiment Reasoner) which is a lexicon and rule-based sentiment analysis tool.

- Contains words which are labelled according to their semantic orientation as either positive or negative and how positive / negative they are.

Project 2: Sentiment Analysis

Frequency vs Sentiment plot for Entity Stratford issues (2018)

- Issues raised in relation to Entity:
  - Time
  - Travel/commute
  - Technology
  - Building
  - Work life balance
  - Flexible working

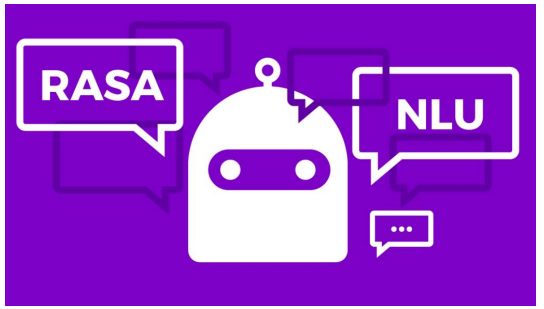
Individuals appear to be less positive about the impact of the entity on "X" and "Y".

Individuals appear to be more positive about the impact of the entity on "A" and "B".
In total there were 207 documents containing mentions of the entity.

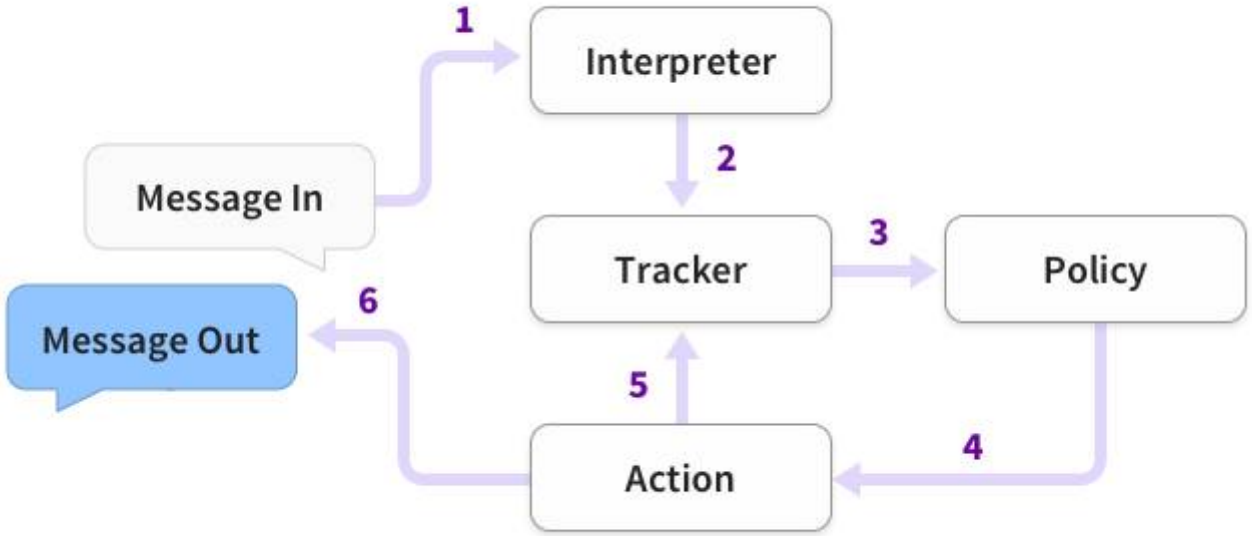
- **In general individuals refer to "C" and "D" as being poor in relation to the current entity when discussing the new entity.**
- **Although these sentences are *positive* they are expressed in a negative way and so generate a negative association with the new entity.**

Project 3: Chatbot

- Currently the AA team have a mailbox to which FCA staff can send emails, typically asking one of four main things:
    - Topic 1
    - Topic 2
    - Topic 3
    - Topic 4
- **Proposition**: use Python package Rasa to create a chatbot trained on previous emails



Project 3: High-level structure



Two main parts:

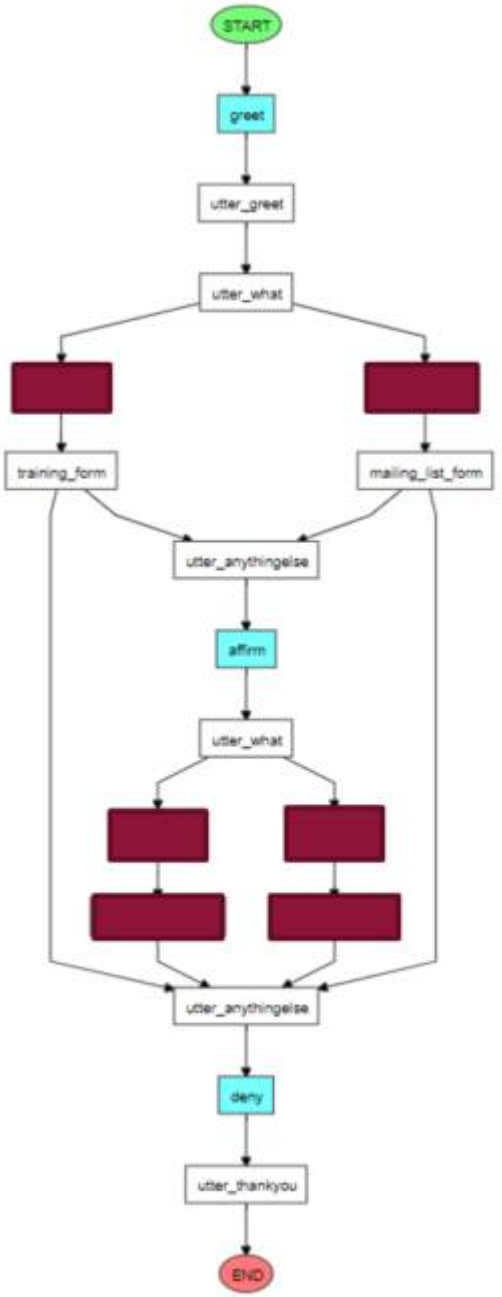
1. Interpreter ("Natural Language Understanding")
2. Core (made up of Tracker, Policy and Action)

Project 3: Interpreter

- Interpreter takes in user message and identifies:
    - the ‘intent’ of the message
    - any entities in the message
- Interpreter is trained on previous emails where intent and entities have been labelled
- Rough pipeline:
    1. Tokenize text
    2. Create features for intent classification
    3. Named entity recognition
    4. Classify intent (SVM)

Project 3: Core

- Core dictates how the bot deals with the intents and entities extracted by the interpreter from each message
- ‘Tracker’ stores the information yielded by the conversation
- ‘Policy’ predicts the next action the bot should take after seeing the Tracker
- ‘Actions’ are the things the bot runs in response to user input, including sending a message back
- Core is then trained on a set of example ‘stories’ (made up of intents, entities and actions) mapping out possible conversations



Other Projects:

1. Mission Embedding

2. Minerva – NLP platform – Change of regulation monitoring.

3. Simple unstructured search including OCR.