UNIVERSITÀ DI TORINO

Tracing and Modelling Data and Social Dynamics via Big Data and AI

Prof. Fabio Ciravegna
Full Professor (Pervasive Computing)
Dipartimento di Informatica
Università di Torino
fabio.ciravegna@unito.it

Copyright Notice

- All the content of these slides
  - Is copyrighted by Fabio Ciravegna, Università di Torino
  - Any externally sourced images or text is - as far as I know - appropriately referenced
    - If you find anything that is not appropriately referenced, please let me know and I will amend the slides with apologies
- Redistribution is allowed only within Intesa Sanpaolo for the purpose of
  - Documenting the 23.02.2023 seminar when the slides are being presented
- Do not reuse reproduce or redistribute the presentation, single slides or images or text for any other purposes
- The slides contain personal opinions
  - they do not represent the opinions or data of any of the users or customers referenced in the slides
  - Errors, if any, are of course mine

This talk

- About myself
- My claim to impact in the real world
- Aerospace:
  - Mining and analysing data in large enterprises
  - Identifying knowledge communities
- Defence and Security:
  - Identifying rumours and fake news
  - Identifying bots and automated responses via behavioural analysis
  - Responding to requests for information using large scale semantics
- Conclusions

About myself

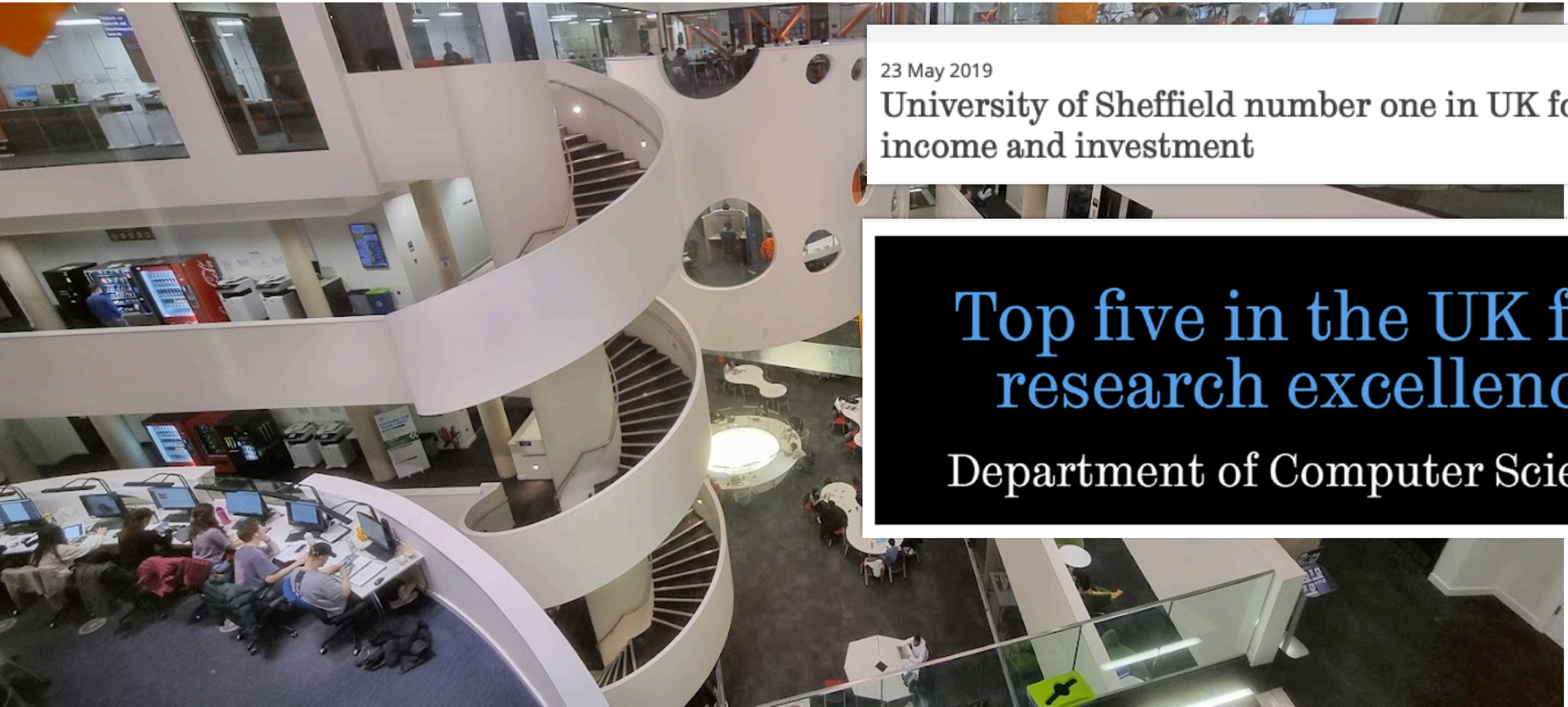
- Eduction:
  - Degree in Computer Science, University of Torino
  - PhD in Computer Science, University of East Anglia
- Work:
  - 1988-1993: Centro Ricerche Fiat, Researcher
  - 1993-2000 ITC/IRST (FBK), Trento - Senior Researcher
  - 2000-2022: The University of Sheffield
    - Professor of Pervasive Computing
      - 2009-2012: Director of R&I for the University (Digital World):
        - £8.9m of new projects in my last year
      - 2020-2022: Director of the University Technology Centre on AI for Defence and Security
    - 2020-2021: CEO, Aeqora Ltd (start up)
    - 2002-2019: Director of EU projects for €25m
  - 2022-present: Università di Torino

UNIVERSITÀ DI TORINO

The University of Sheffield

- The largest Engineering Faculty in the UK
- 75th in the QS World Universities Table
- 11th in Europe int THE's Teaching Quality Table

A world-class university – a unique student experience



23 May 2019

University of Sheffield number one in UK for engineering research income and investment

Top five in the UK for research excellence

Department of Computer Science

REF 2014

Research Excellence Framework 2014

UNIVERSITÀ DI TORINO

About My Research

- Pervasive computing with a focus on large scale data management.
  - **Data capturing**
    - Over large scale from multiple devices and sources
  - **Data analytics and Prediction**
    - To inform final users, problem owners, etc.
- Application areas:
  - From aerospace, to smart cities, environmental monitoring, emergency services, health, sports, photography, etc.
- Major partners:
  - Public Health England, Kodak, JustGiving, Rolls-Royce, Glastonbury Festival, City Councils, Football Whispers…

My Claim to Impact

- **Startups**
    - 2007: K-Now Ltd
    - 2012: The Floow Ltd $69m exit in April 2022
    - 2020: Aeqora Ltd

- **Intellectual Property** sold or released to industry and government
    - Rolls-Royce, JustGiving,
    - Public Health England, UK Ministry of Defence
    - Kodak, Football Whispers

- **Technology** released to millions of users
    - 1 million users for Public Health England
    - 2.5 million users served for Football Whispers
    - 1 million users monitored in emergency control rooms

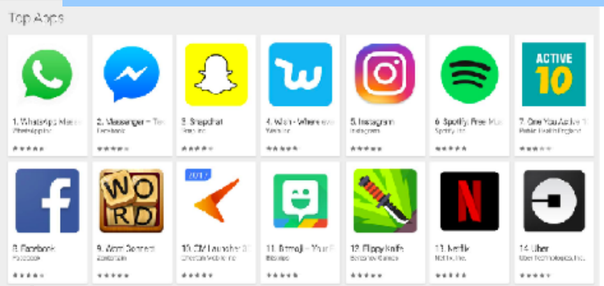
Health

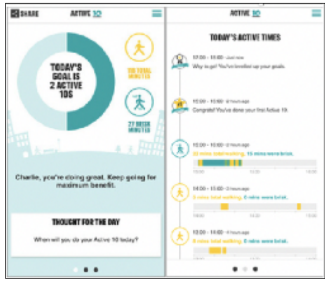
- Public Health England
  - Lifestyle tracking via Mobile Phones
  - 1 million users
  - 1 billion mobility data points collected
- Technology released in TV
  - 5 Hospitals in UK, Germany and Israel
- Moreover:
  - >6,000 people (MoveMore Sheffield)
  - >5,000 of bikes with Birmingham City Council
  - >1,000 people in Santander (SP)

ACTIVE 10

Public Health England

7th most downloaded app in the UK





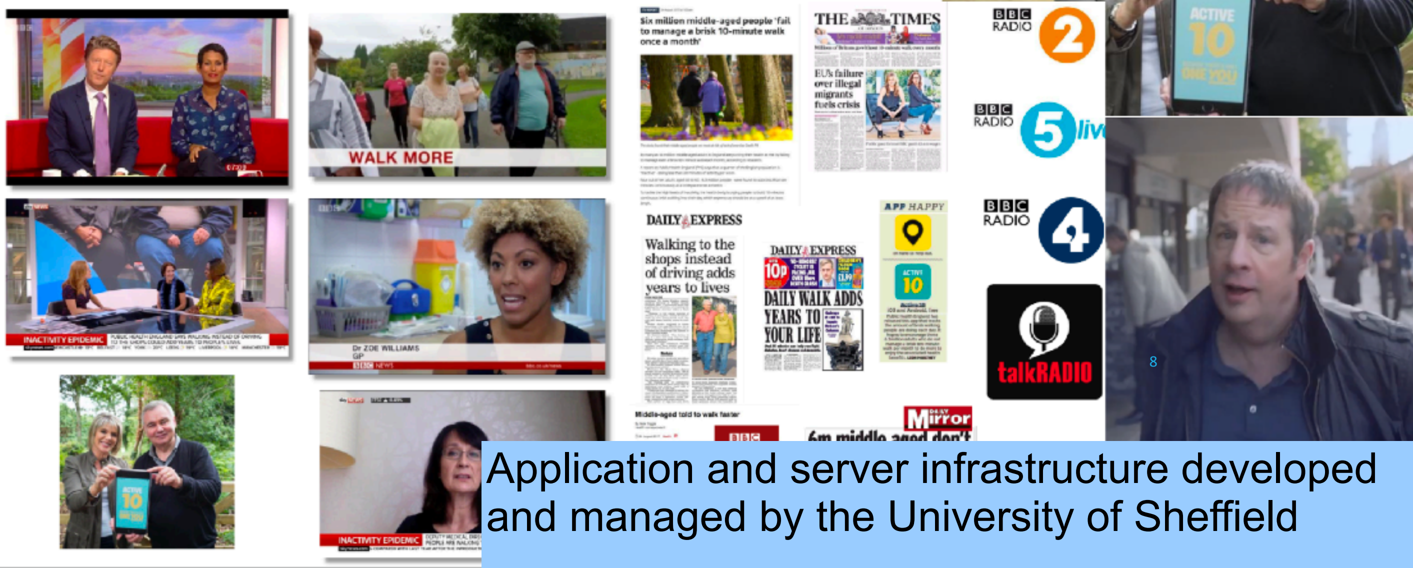
Hello magazine sponsored Facebook live event

Home | Football | Formula 1 | Cricket | Rugby U | Rugby L | Tennis | Golf

Get Inspired > England & FA People's Cup | Scotland | Wales | N. Ireland

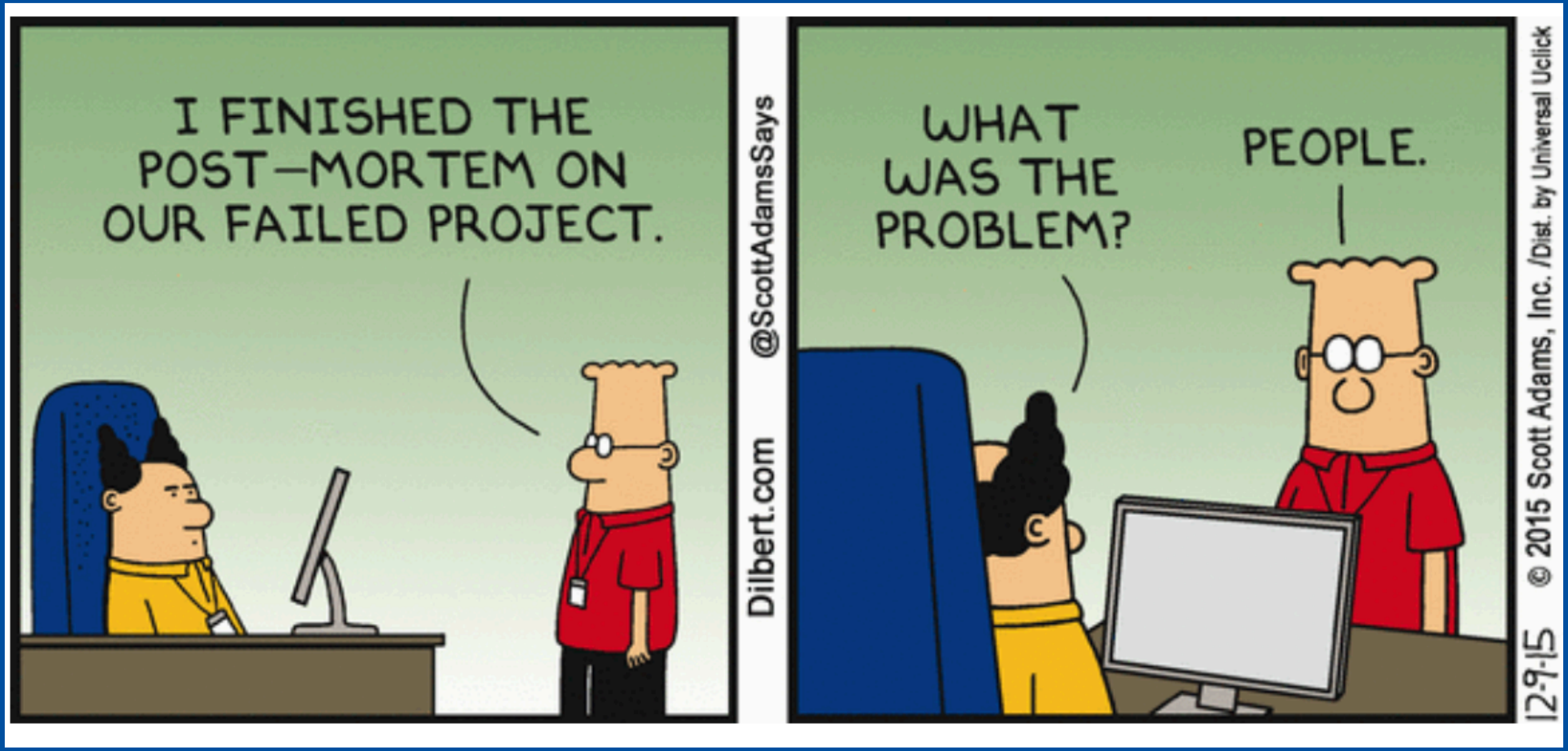
Active 10: Eamonn Holmes joins sup... of 10-minute fitness app

24 August 2017 Get Inspired



Application and server infrastructure developed and managed by the University of Sheffield

> *PHE were able to develop and launch the first free-to-use mobile app that provided the user with information on time, intensity and periodicity [of physical activity]. The app played a significant role [...] and made a major contribution to the overall success of the One You campaign*
>
> *Anand Amlani, Head of Marketing – Living Well @ Public Health England*

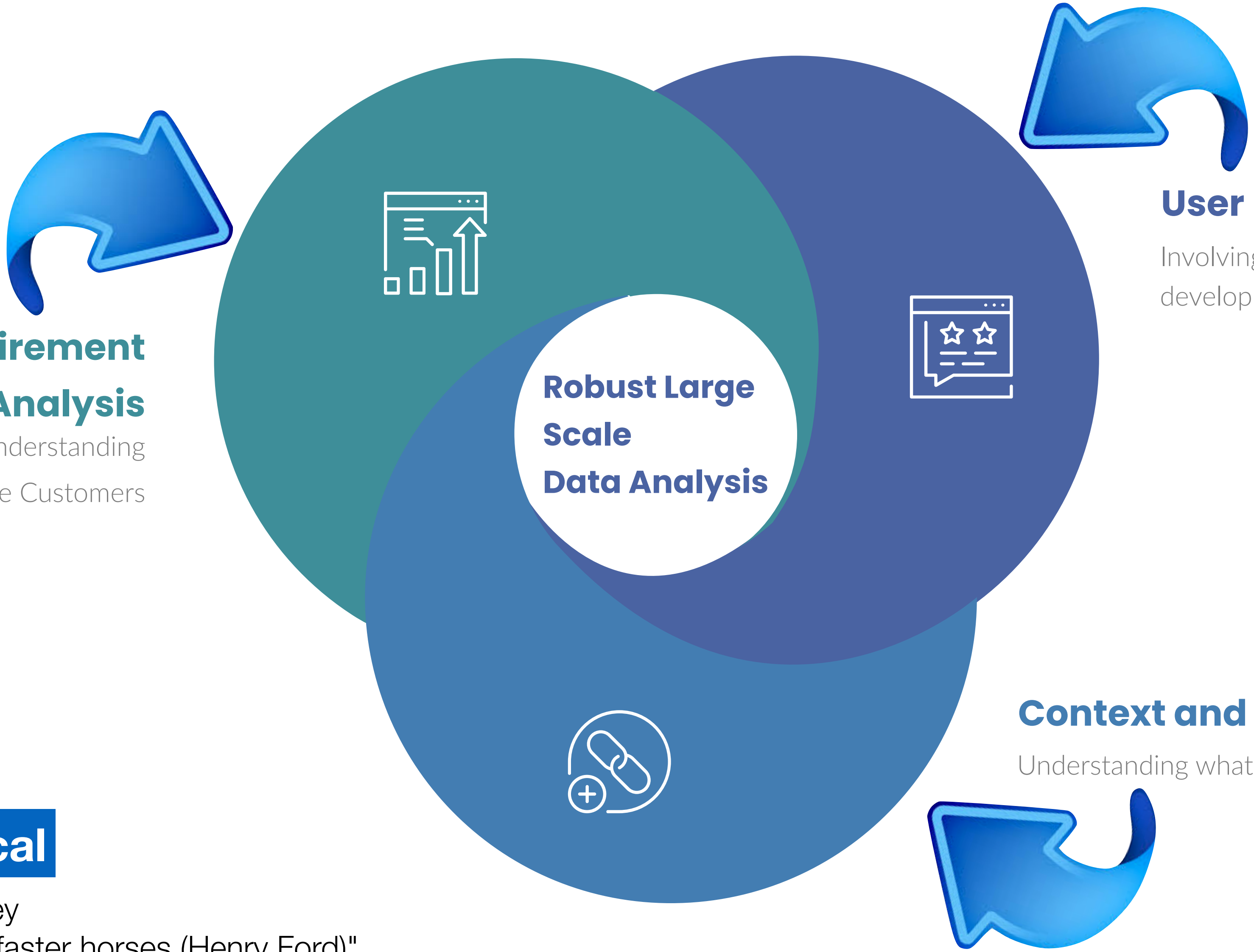


First and Foremost Understand the Users and Their Context

Designing a Solution

Robust Large Scale Data Analysis

Prof. Daniela Petrelli



Requirement and Impact Analysis

Understanding the Needs of the Customers

User Engagement

Involving the users in all phases of development

Context and Behaviour

Understanding what, why and when

But dare to be critical

"If I had asked people what they wanted, they would have said faster horses (Henry Ford)"

UNIVERSITÀ
DI TORINO

Knowledge Management In Large Enterprises

Rolls-Royce, Tata Steel
and many others

The University Of Sheffield.

Aerospace

- 10 year of research with Rolls Royce plc
  - Shortlisted twice for the Rolls Royce Director of Research Creativity Award

The nomination is given to solutions which can sensibly change the future way of working of the company and it is selected by vote by senior employees

Colin Cadas, Rolls Royce Associate Fellow Knowledge Management

  - Terminology recognition
    - 10,000 users at Rolls Royce plc
    - Part of a KM suite saving RR £14m/year

TR is the core component of a Knowledge Management improvements programme focussing on information extraction and data mining thousands of documents. It was strategically productionised as part of our corporate search strategy, delivered to over 10,000 engineers and with cost savings in the £14Millions

Colin Cadas, Rolls Royce Associate Fellow Knowledge Management

UNIVERSITÀ DI TORINO

Solutions Vs Products

- Modern manufacturing companies are selling complete service solutions instead of physical goods
    - Aircraft power Vs jet engines
    - 7 year warranty on cars means selling mobility
- Servitisation requires taking charge of the whole product life-cycle
    - Designing better products to have larger margins
        - As opposed to design to manufacture at low cost
    - Design products to minimise service requirements
        - As opposed to profit on service provision
- Seeking, processing and communicating information takes a considerable amount of a knowledge worker's time,
    - e.g. 55% of an aerospace designer's time
- 75-85% of information unstructured and doubling every year
- Unstructured information difficult to find and retrieve

X-MEDIA

Sense Making in Aerospace

jet engines are completely serialised
- every piece has a serial number (excepts nuts and bolts)
- the history of each part is recorded
    - e.g. part transferred between engines

- a jet engine can produce ~1Gbyte of vibration data per hour of flight;
    - if irregularities are found, part of the data can be stored
    - reports can be written (event reports)
    - pictures can be taken

When engine is serviced (e.g. overhaul)
- financial information is produced.
- if issues are found,
    - pictures are taken
    - reports are written
    - engine is tested

X-MEDIA

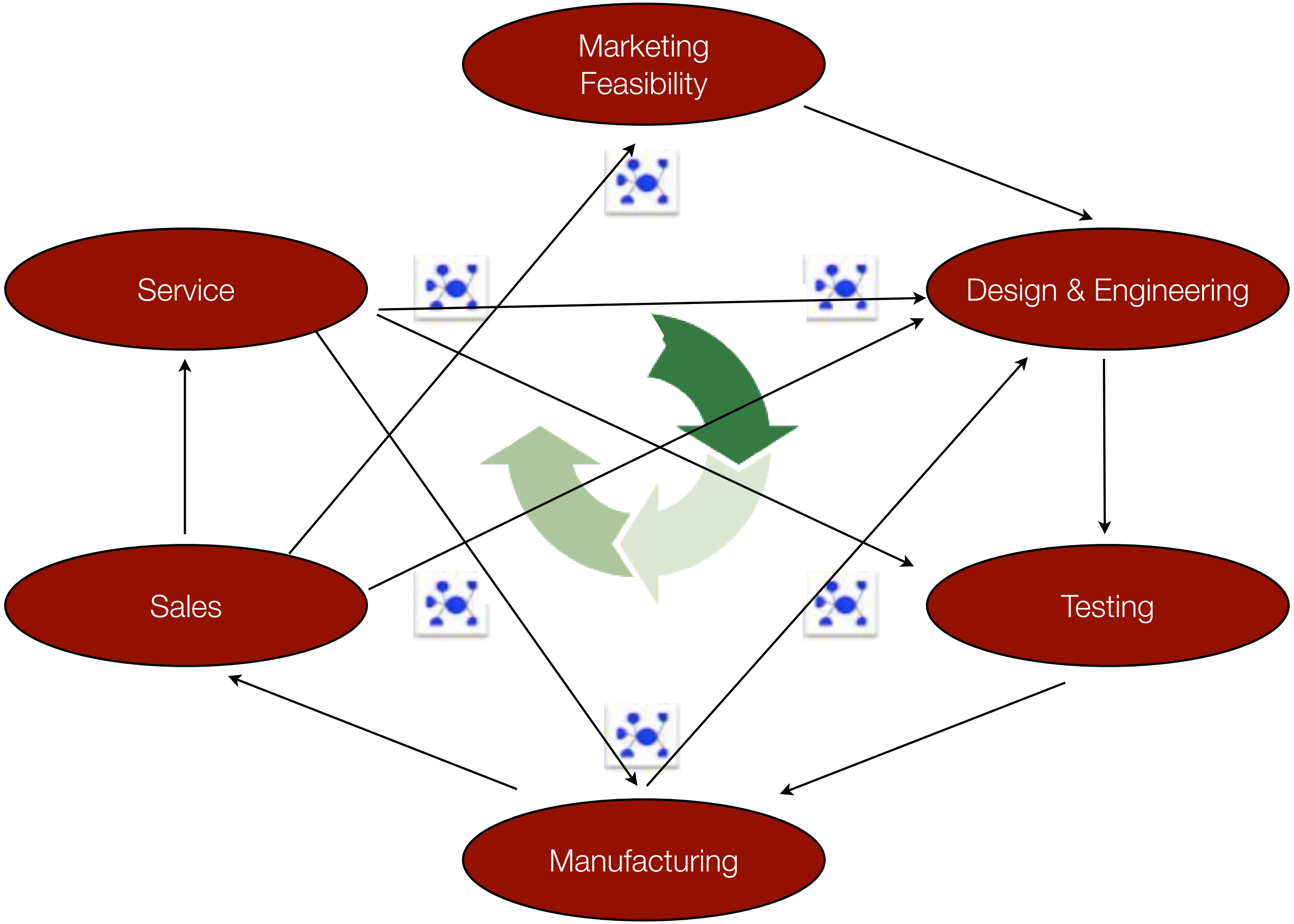
Jet engine example (3)

- If problem is recurring (or suspected so)
  - a problem resolution group is established
    - existing evidence is retrieved
    - further evidence is collected
    - a learned lesson is generated
    - same problem is investigated across models

Document Type
AROC proforma
AROC results
Development
EHM data
Emails
ONWING emails
Images
Lab findings
Monitoring Requir
Presentations
Procedures
RCP
Risk Assessment
Solution Reports
Technical Reports
TS&O Reports

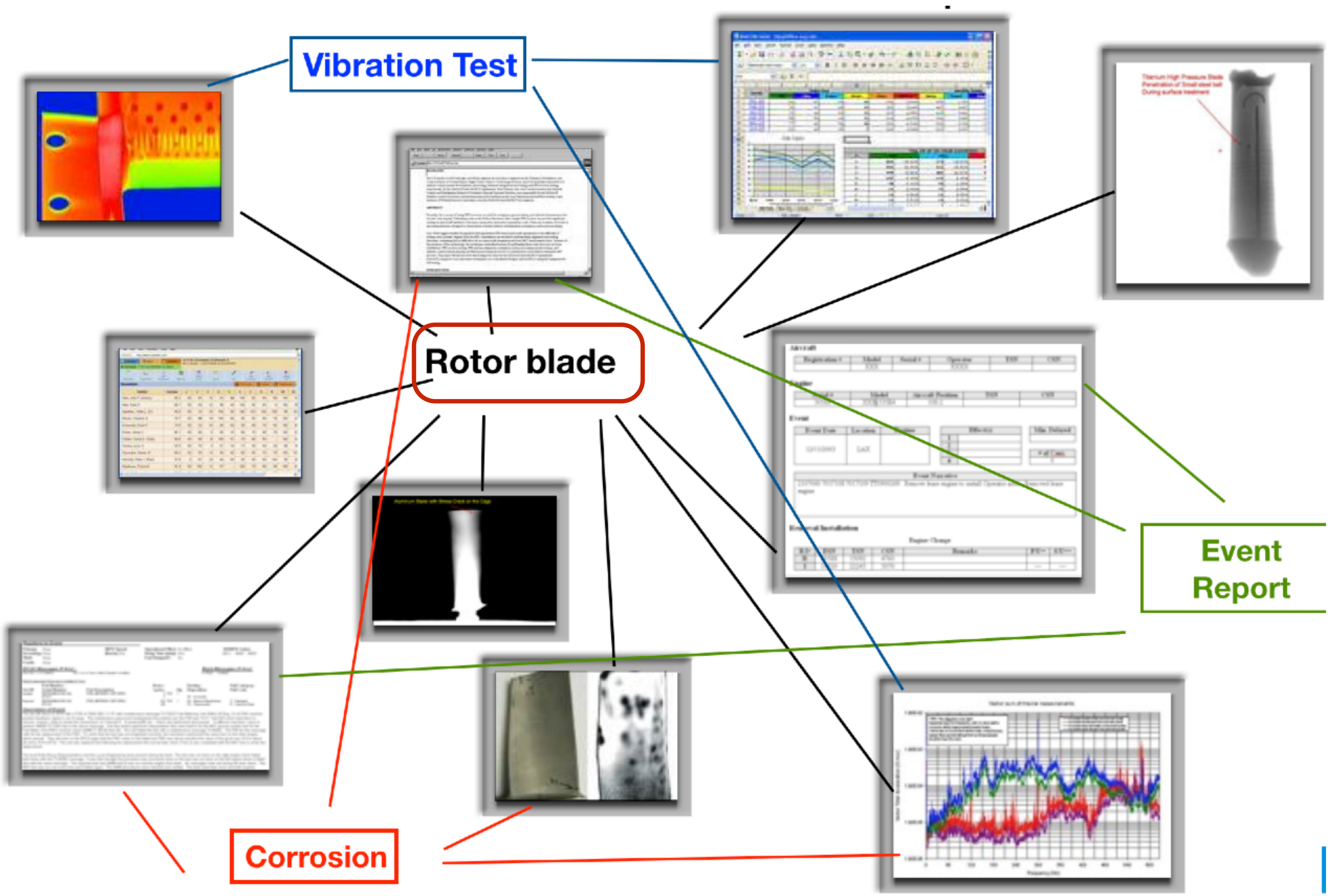
Different repositories represent different communities point of view!!

Closing the information loop



The University Of Sheffield.

A single rotor blade, much data

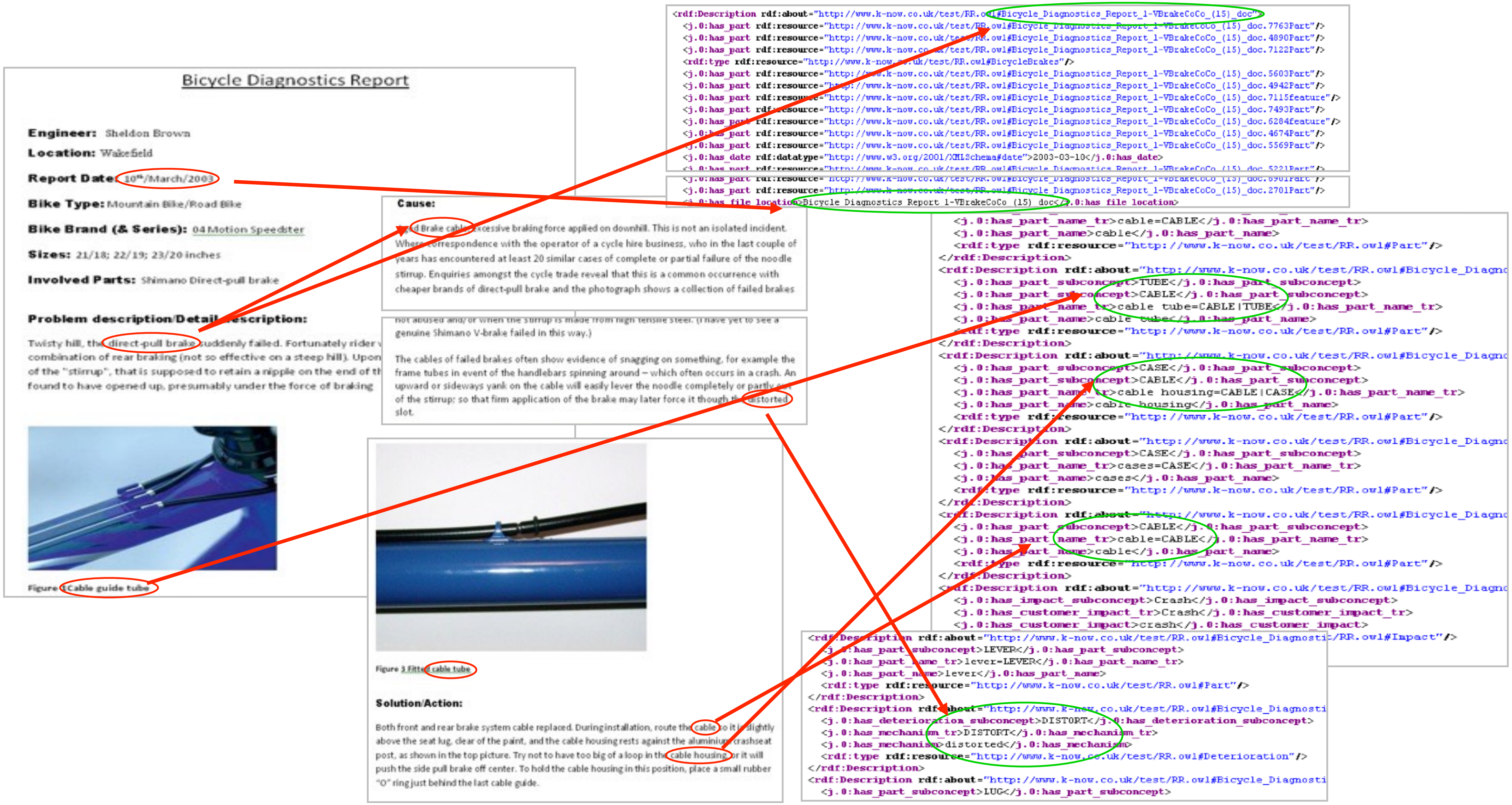


Terminology Recognition

"Low Pressure Turbine Stage 2 Rotor Blade"
"LP2 Blade"
"FK42164"
"LPT 2 Blade"
"72-41-12"
"T800 LP Turbine Blade Stage 2"
"Turbine Blade"
"72-41-12-400"
"Blade, Turb l2"
"Blade, LPT"
"TurbinneBladee"
"FK12548"

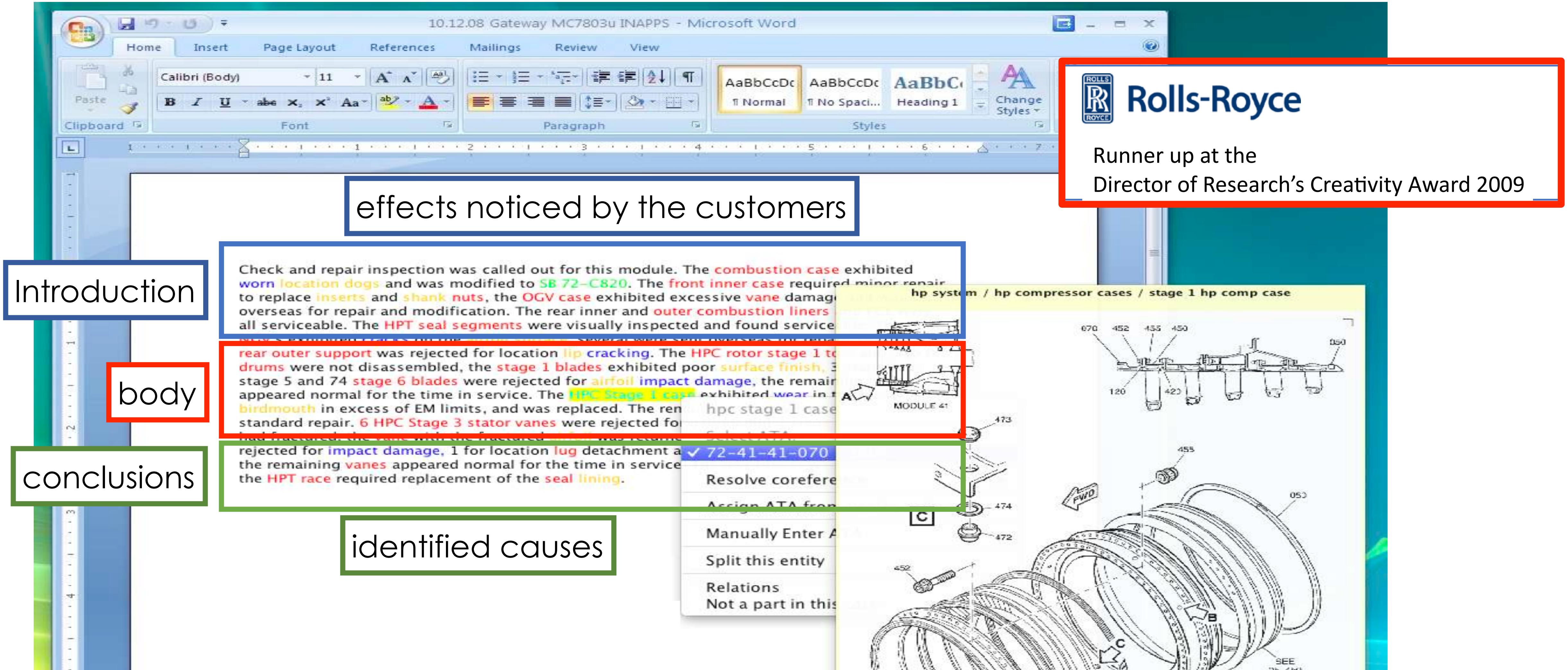
- Task of reducing all these terms to a unique identifier no matter how it is represented in documents or archives
  - Approach: a cascade of HMM and SVM models

Linking via Terminology Recognition



A Creative Use of TR

- Initially developed at the Department of Computer Science of the University of Sheffield
- Certified for use and part of the official knowledge management suite with thousands of users



UNIVERSITÀ DI TORINO

Finding the Needle in the Stack

(and quickly)

Social Media Analysis

- Emergency control rooms of events involving >1M people
  - Including the Glastonbury festival (200k people) (twice)
  - Evacuation of 30,000 people from Vicenza (Italy)
  - Italy invested €3.5 in a followup project (thank you, Brexit!)

Rolling Stones make Glastonbury debut

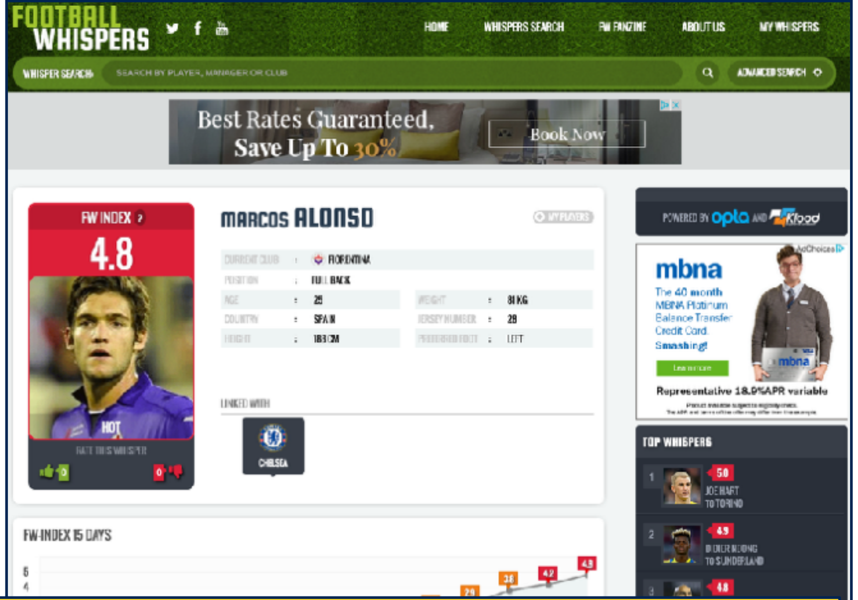
Michael Eavis's lifetime aim to see the band on the Pyramid stage is finally realised 43 years after festival first took place

"The contribution of the **OAK** group in this process was key. The project made the concept real and applicable; the technology developed by **OAK** provided concrete proof of the power of the citizen observatories as well as a powerful benchmark for requirement analysis and for the development of the final production technology"

Dr Michele Ferri, Projects Manager, Eastern Alps District River Authority

And More…

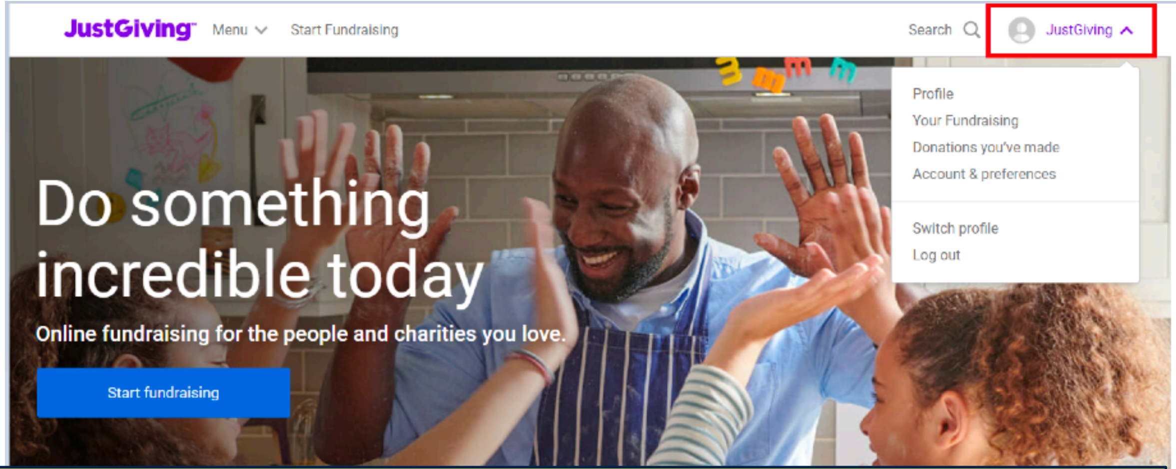
- Football Whispers:
  - Social media analysis
    - 70M messages a month analysed
  - 35 international leagues, hundreds of teams, thousands of players
  - Major customers: Sky Sports and 4-4-2
  - From 0 to 2.5m users in 6 months
  - Project delivered in 1.5 months



Thanks to the work of the OAK group, we were able to launch on time in January 2016 and with our full service offering – something that we would not have been able to accomplish without their input. […] In that time our business grew from 0 to 2,500,000 unique monthly users

Vivion Cox, CEO and Founder

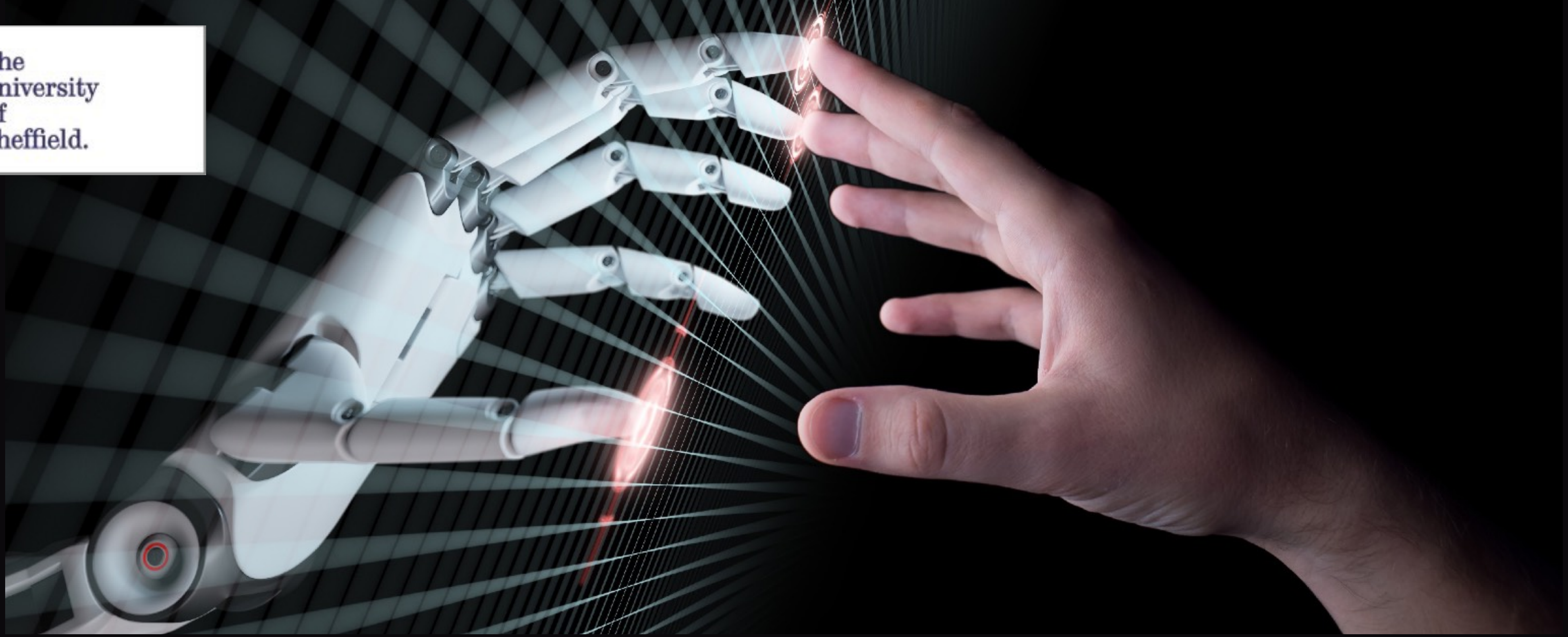
- JustGiving
  - The largest donation company in the world
    - Income: £3B a year
  - Recommender system via social media mining
    - Increased followup visits by 378%



We measured a 378% rise in the likelihood that a user would visit a page when they see it in the "you might be interested in" card in their feed, clearly demonstrating an improvement in the selection of related causes.

Richard Freeman, PhD, Lead Data and Machine Learning Engineer, JustGiving

The University Of Sheffield.



University Technology Centre on AI for Defence and Security

Prof. Fabio Ciravegna
Academic Director
Department of Computer Science
University of Sheffield
f.ciravegna@shef.ac.uk

Some Activities

- Explainable AI
  - e.g. extraction from learning models
- Dependable AI
  - e.g. monitoring distribution shifts, monitoring activation patterns, trusted datasets
- Deception in media
  - e.g. disinformation detection, bot detection
- Rapid large scale information identification
  - e.g. rapid response to RFIs
- Deception in real life
  - e.g. behaviour and mobility prediction, facial recognition bypass, audio spoofing

Disclaimer

- No confidential information was used to prepare this talk
- Examples used in this talk are not real life examples
  - They are just personal opinions on what could be or used for

(Credit: https://psychology-spot.com/rumors-gossip-and-fake-news)

Rumours Detection

The University Of Sheffield.

Rumours and emergencies

- Cause panic and distress
- Impair the effective and timely allocation of resources and police
- Debunking misinformation in the early stages of events is important

- Early rumour detection is a requirement in many applications

Early Rumour Detection (ERD)

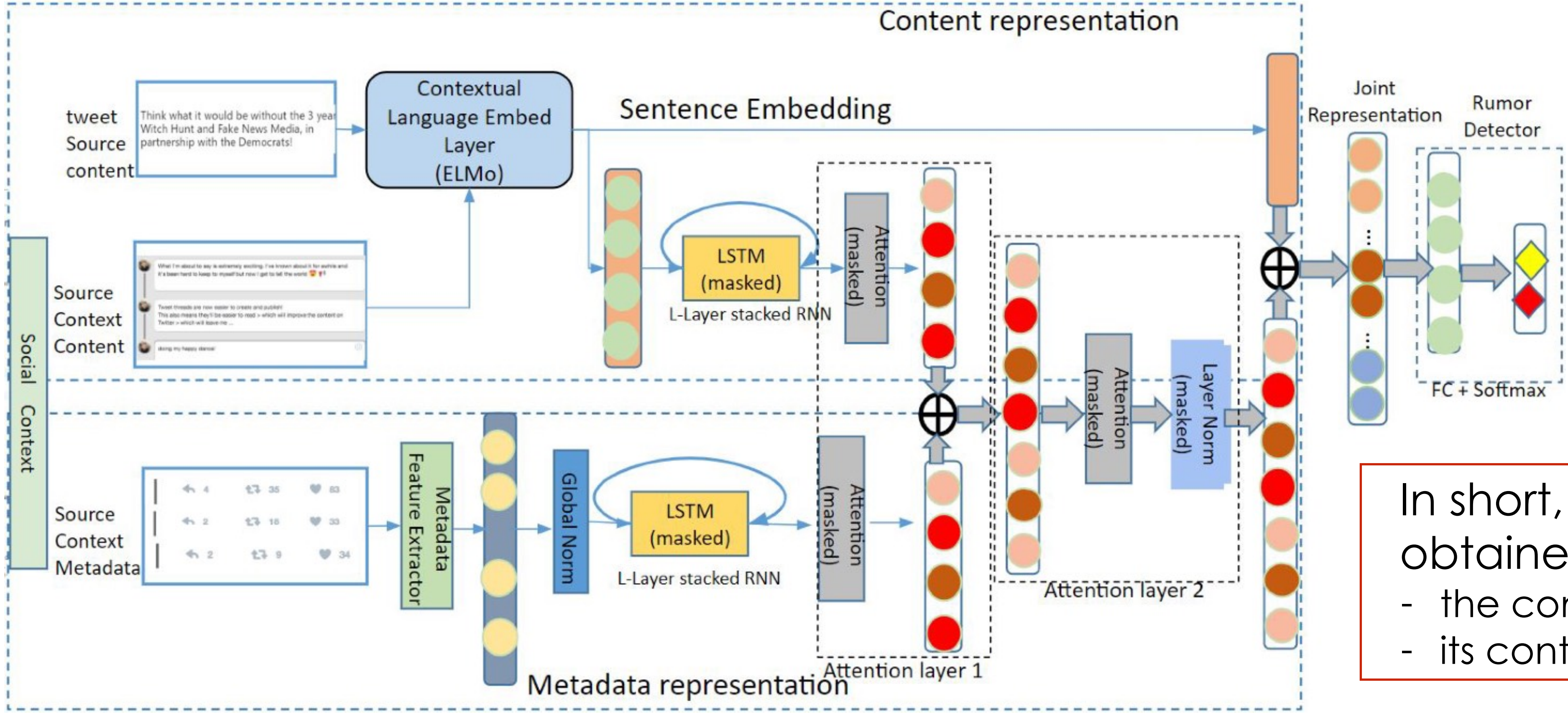
Rumour Propagation-Based Deep Neural Networks (RP-DNN)

- A hybrid deep learning architecture for tweet-level rumour detection
  - while the majority of recent work focuses on event-level classification.
- It advances state-of-the-art (SOTA) performance on tweet-level ERD
- A context-aware model
  - learning a unified rumour representation from multiple correlated context inputs
    - including source content (SC), context content (CC) and context metadata (CM) beyond the word-level modelling
- Stacked LSTM networks with multi-layered attention mechanisms
- Extensive experiments based on an ablation study and LOOCV are conducted to examine its effectiveness and generalisability.

- Our model outperforms SOTA models in tweet-level ERD and achieves comparable performance with SOTA event-level rumour detection models

Gao, Jie, Sooji Han, Xingyi Song, and Fabio Ciravegna (May 2020). "RP- DNN: A Tweet Level Propagation Context Based Deep Neural Networks for Early Rumor Detection in Social Media." In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France.
Han, Sooji, Jie Gao, and Fabio Ciravegna (May 2019a). "Data augmentation for rumor detection using context-sensitive neural language model with large-scale credibility corpus." In: Proceedings of the 7th International Conference on Learning Representations. Learning from Limited Labeled Data: ICLR 2019 Workshop.
Han, Sooji, Jie Gao, and Fabio Ciravegna (2019b). "Neural language model based training data augmentation for weakly supervised early rumor detection." In: Proceedings of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

Early Rumour Detection Model



In short, best results are obtained using
- the content of the tweet and
- its context (e.g. replies)

Gao, Jie, Sooji Han, Xingyi Song, and Fabio Ciravegna (May 2020). “RP- DNN: A Tweet Level Propagation Context Based Deep Neural Networks for Early Rumor Detection in Social Media.” In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France.

Methods	Precision	Recall	F1	Accuracy
RP-DNN	0.852	0.989	0.915	0.872
Ma et al. (2017)	–	–	0.738	0.741
Liu and Wu (2018)	–	–	0.843	0.853
Ma et al. (2018a)	–	–	0.753	0.730

Methods	Precision	Recall	F1	Accuracy
RP-DNN	0.790	0.868	0.826	0.818
RPDNN - CXT	0.785	0.844	0.811	0.804
RPDNN - SC	0.730	0.839	0.780	0.762
RPDNN - CC	0.762	0.846	0.801	0.788

Weakly supervised data augmentation

- Rationale
  - New variants of rumours in the early stages are mostly textual variations (Maddock, 2015; Zhao, 2015).
  - 80% of publicly available social media rumor data are duplicated contents (Chen, 2018).
  - Variations share similar propagation patterns (Kwon, 2017; Liu 2017)
- Why is this important?
  - During emergencies, unexpected requests for information, services, help and clarifications are available.
  - Data augmentation makes a model for emergency response/ management robust

Han, Sooji, Jie Gao, and Fabio Ciravegna (May 2019a). "Data augmentation for rumor detection using context-sensitive neural language model with large-scale credibility corpus." In: Proceedings of the 7th International Conference on Learning Representations. Learning from Limited Labeled Data: ICLR 2019 Workshop.

Weakly supervised data augmentation

- Noisy and less precise sources (e.g. data patterns) are leveraged to learn limited high-quality labelled data
- Our method is based on a state-of-the-art neural language model and semantic relatedness
- Data augmentation helps to boost performance on rumour detection on Twitter

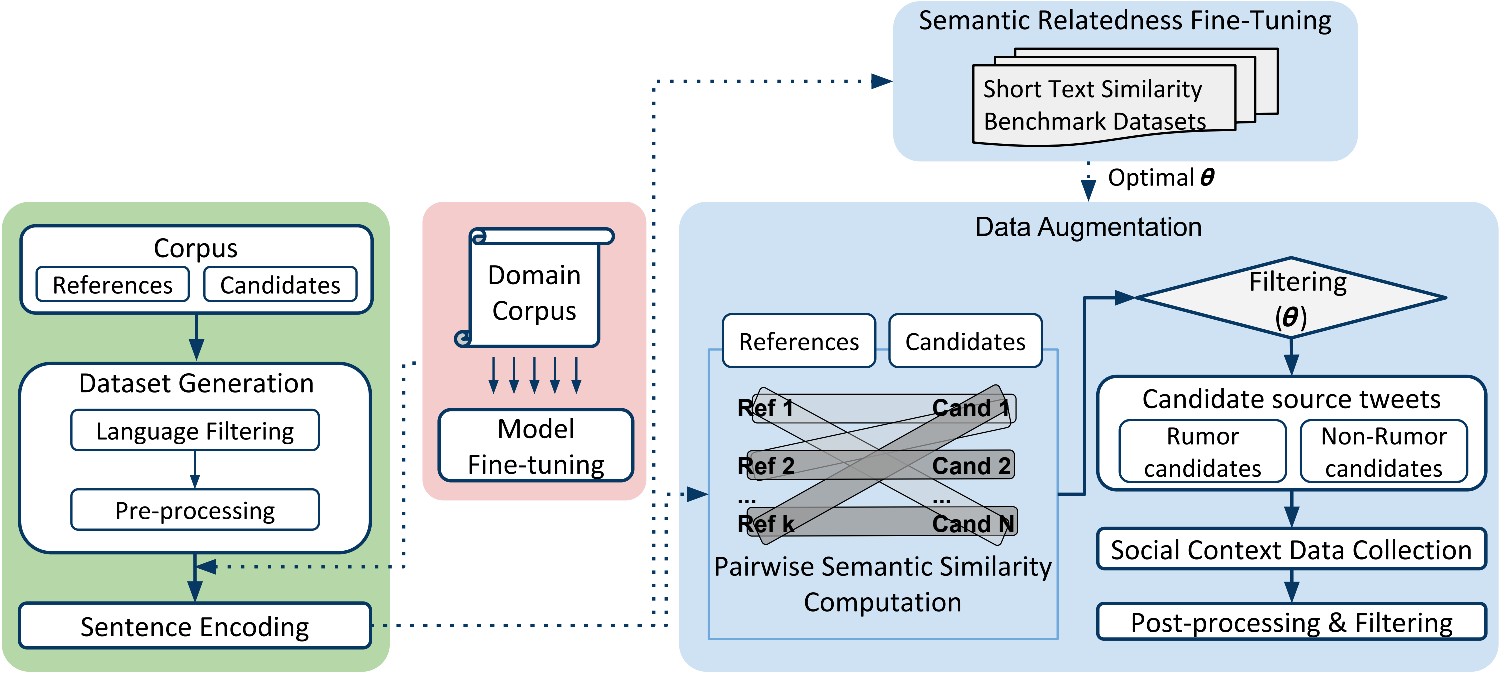


Table 4.13: Rumour detection results for different data sets.

Data	F	P	R	Acc.
PHEME5	0.535	0.650	0.484	0.622
Aug-PHEME-filtered -boston	0.625	0.688	0.585	0.664
Aug-PHEME-filtered	0.656	0.716	0.614	0.685

Table 4.14: LOOCV results for the PHEME5 and augmented data sets.

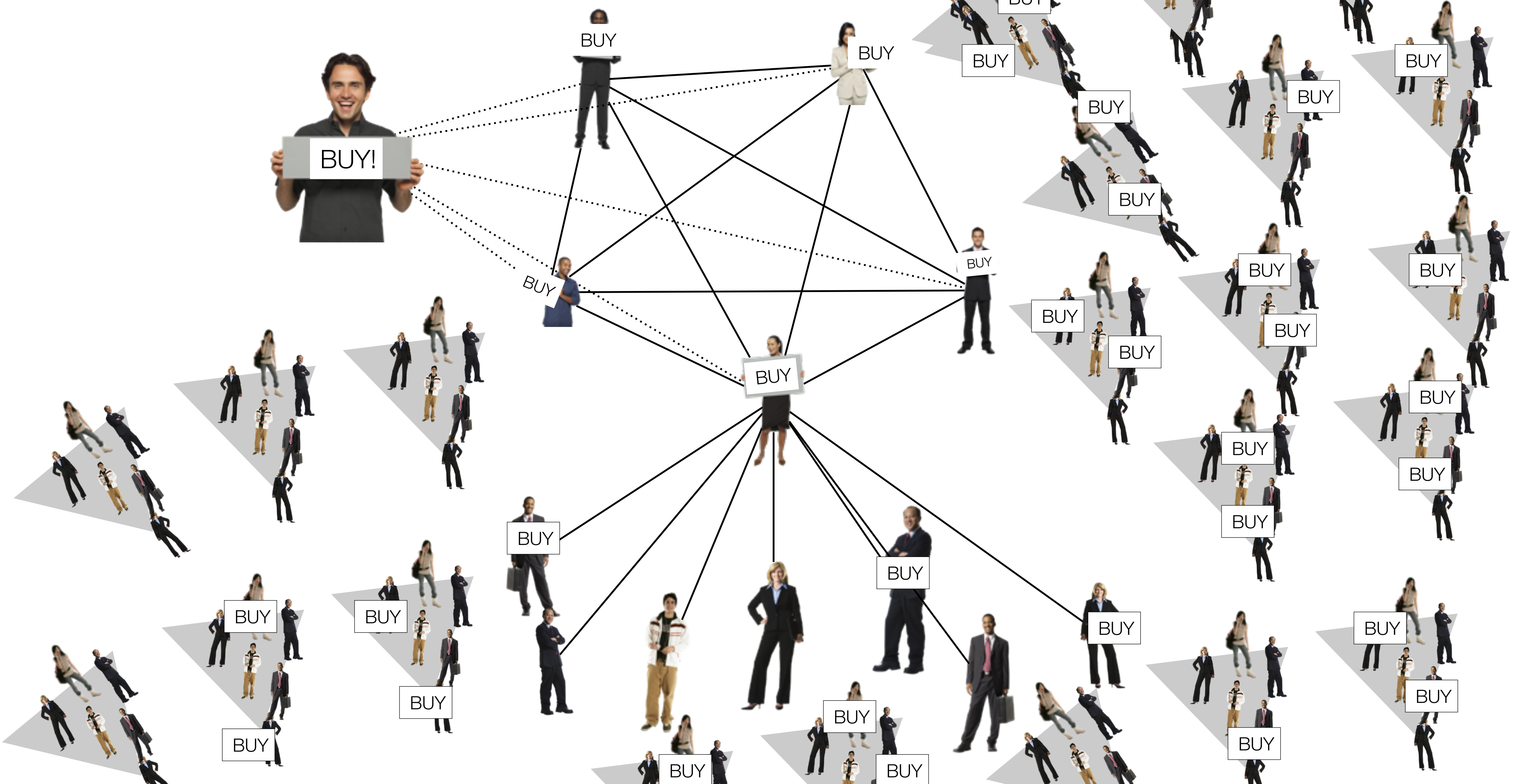
Event	Data	F	P	R	Acc.
germanwings	PHEME5	0.577	0.619	0.541	0.604
	Aug-PHEME-filtered -boston	0.601	0.652	0.558	0.630
	Aug-PHEME-filtered	0.575	0.650	0.515	0.619
sydneysiege	PHEME5	0.583	0.714	0.492	0.648
	Aug-PHEME-filtered -boston	0.695	0.755	0.644	0.717
	Aug-PHEME-filtered	0.632	0.759	0.542	0.685
fergusonunrest	PHEME5	0.242	0.550	0.155	0.514
	Aug-PHEME-filtered -boston	0.416	0.618	0.313	0.560
	Aug-PHEME-filtered	0.609	0.707	0.535	0.657
ottawashooting	PHEME5	0.516	0.653	0.426	0.600
	Aug-PHEME-filtered -boston	0.671	0.680	0.662	0.675
	Aug-PHEME-filtered	0.697	0.739	0.660	0.713
charliehebdo	PHEME5	0.758	0.714	0.808	0.742
	Aug-PHEME-filtered -boston	0.742	0.734	0.749	0.739
	Aug-PHEME-filtered	0.767	0.723	0.817	0.752

Han, Sooji, Jie Gao, and Fabio Ciravegna (May 2019a). "Data augmentation for rumor detection using context-sensitive neural language model with large-scale credibility corpus." In: Proceedings of the 7th International Conference on Learning Representations. Learning from Limited Labeled Data: ICLR 2019 Workshop.
Han, Sooji, Jie Gao, and Fabio Ciravegna (2019b). "Neural language model based training data augmentation for weakly supervised early rumor detection." In: Proceedings of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

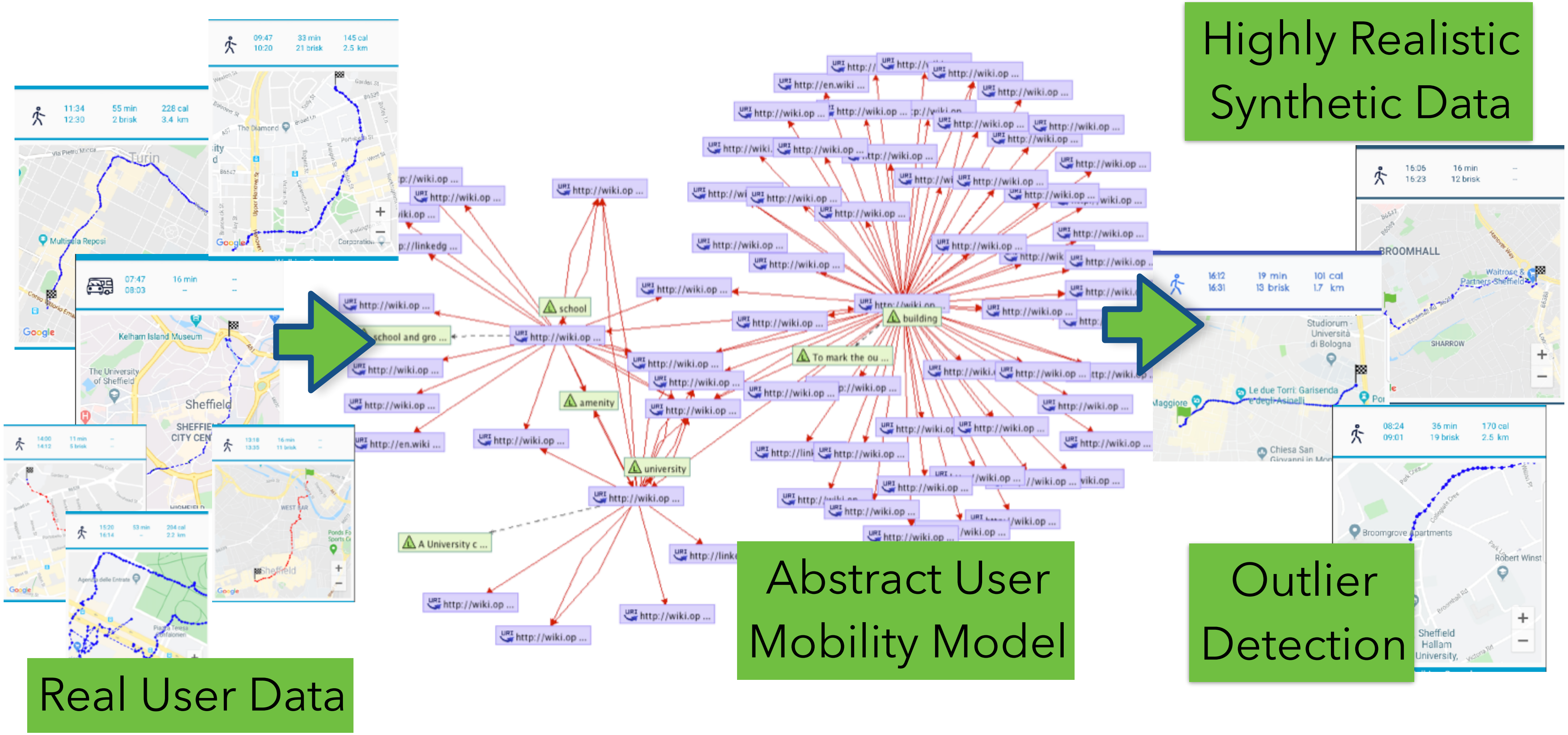
Detecting Bots

The University Of Sheffield.

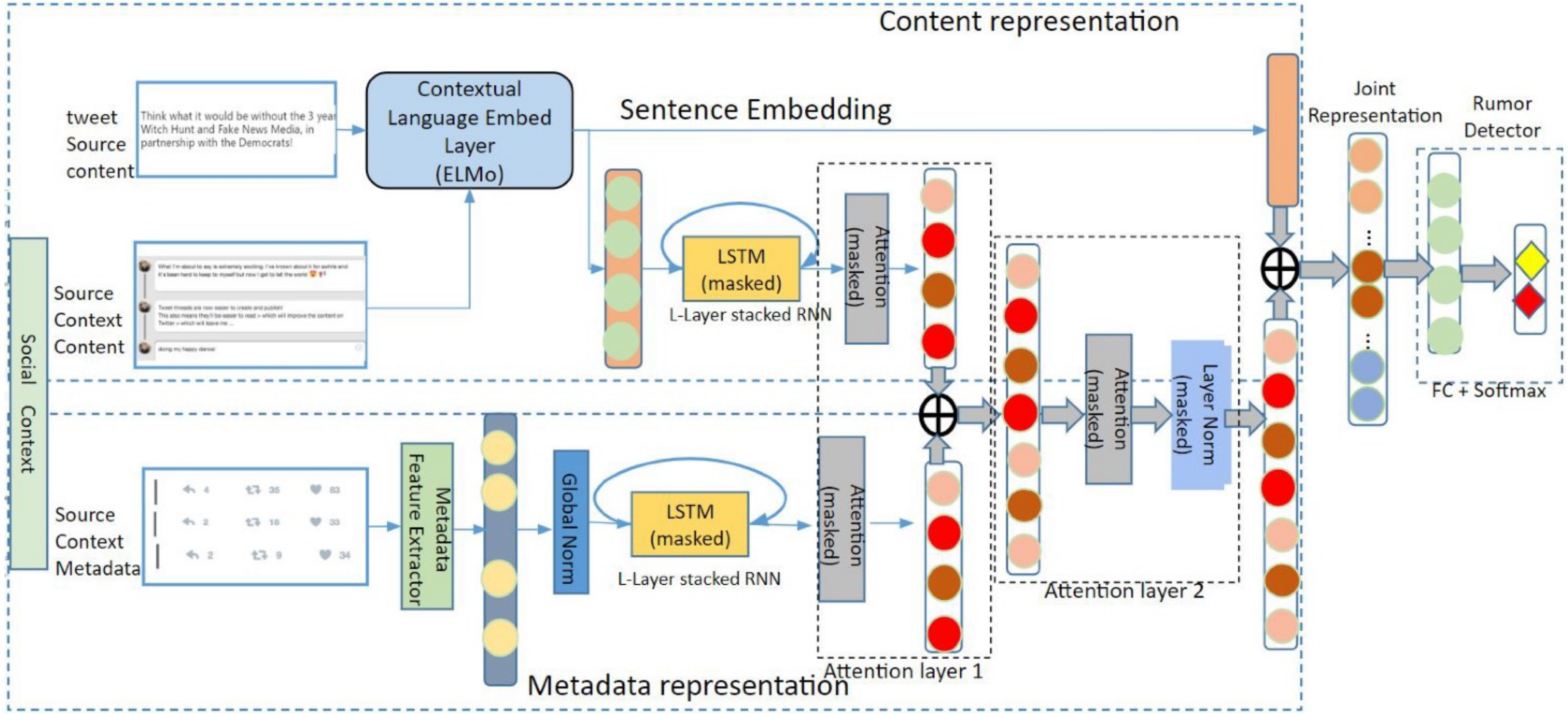
How bots work



The University Of Sheffield.

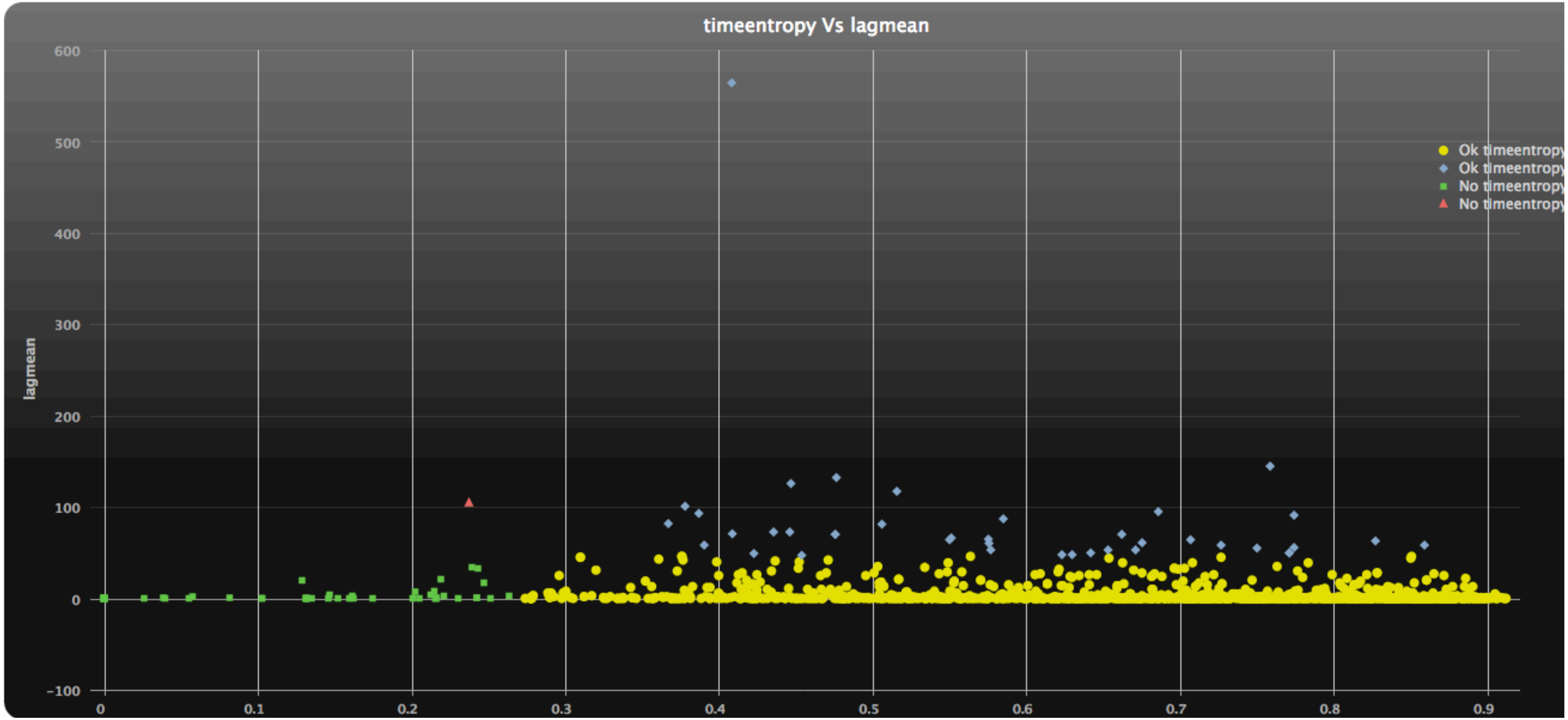


Cyborg Detection



https://www.deviantart.com/kittycat159/art/Honeytrap-2014-Movie-Folder-Icon-2-858191150

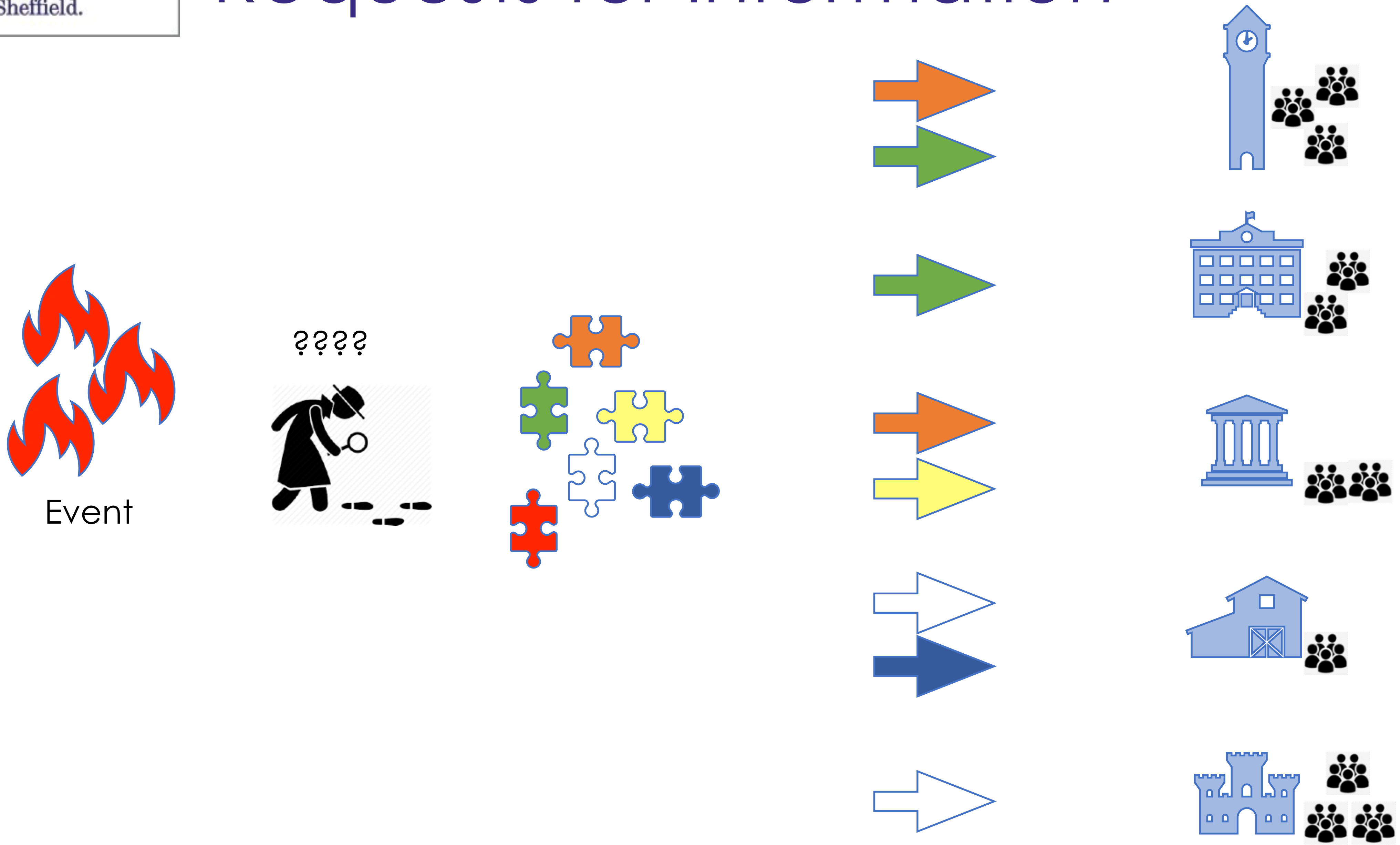
Meet 35,000 Beliebers



UNIVERSITÀ DI TORINO

Requests for Information

Requests for Information



Finding Experts

- As situations evolve quickly, decisions are informed
  - through the constant flow of questions and answers between decision-makers and intelligence units
  - to reduce uncertainty and manage decision risks.
- Through a Request for Information process, the questions are sent selectively to expert sub-units
  - who each contribute (in part) to answering the question,
    - drawing from their specific subsets of resources

From RFIs to Information Interests

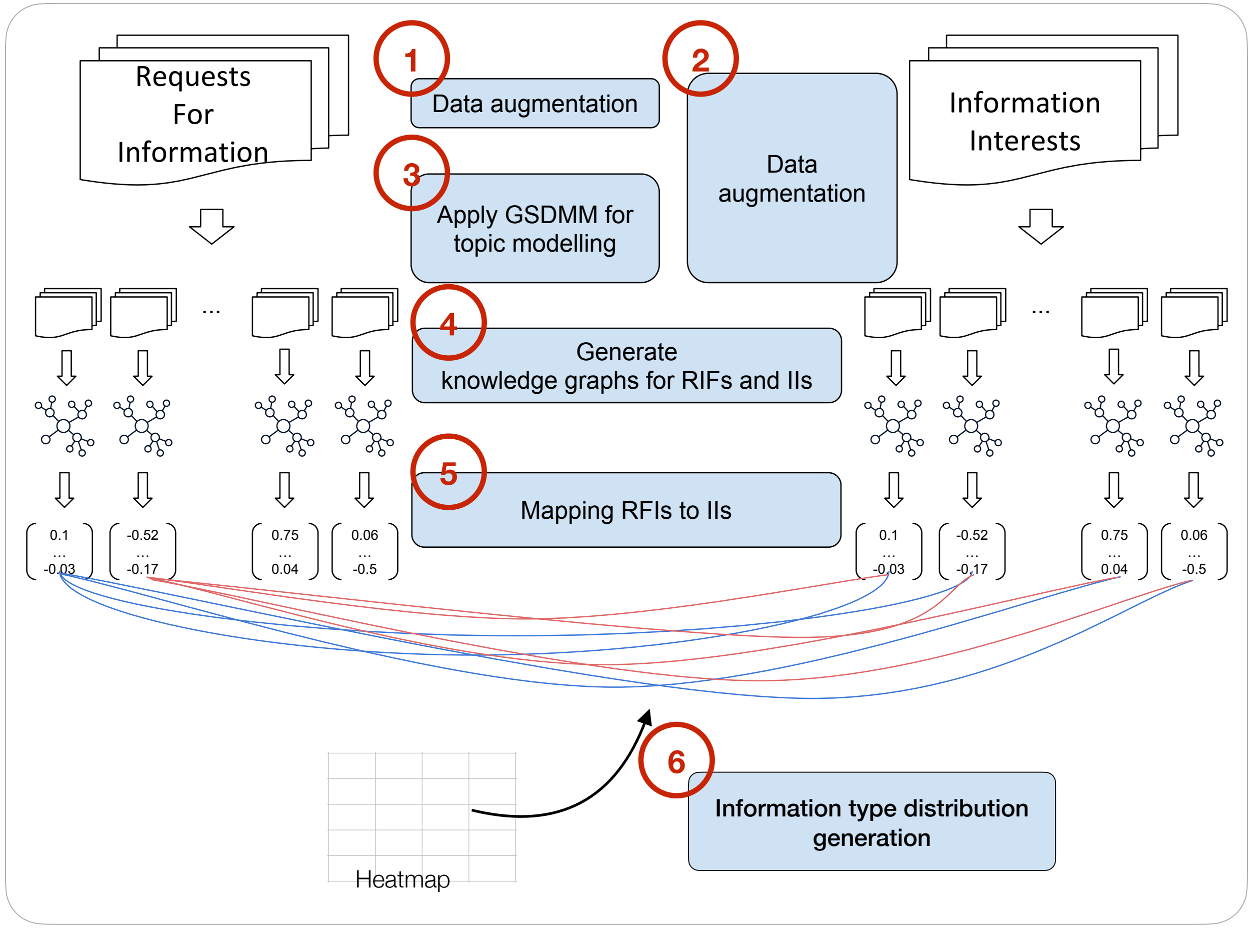


Figure. Pipeline for RFI decomposition

Tasks

1. RFI data augmentation
2. Information interest (II) data augmentation
3. Short text topic modelling over RFIs
4. RFI and II knowledge graph generation
5. Knowledge graph mapping
6. Information type contribution generation

An Edge Computing Task

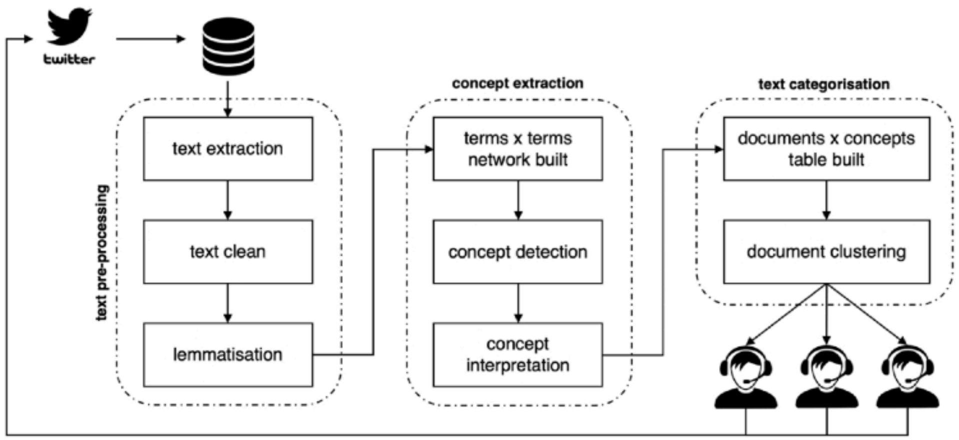
- We cannot mine all government's documents and data
  - you never know where you will end up
- We cannot identify the experts without their permission
  - you never know where you will end up
- Often you cannot identify yourself to those experts
- Solution:
  - Index IIs at the edges
  - Send a pre-processed request for information to the edge
  - Match the IR to the II at the edge
  - Identify the experts at the edge
  - Let the experts know you are looking for them
    - The experts will (in case) come back to you
    - you will however be protected as well by anonymity in the first instance
- Challenges:
  - Matching at the edges has a number of issues in terms of learnability (size), system maintenance and agility

Expert Finding

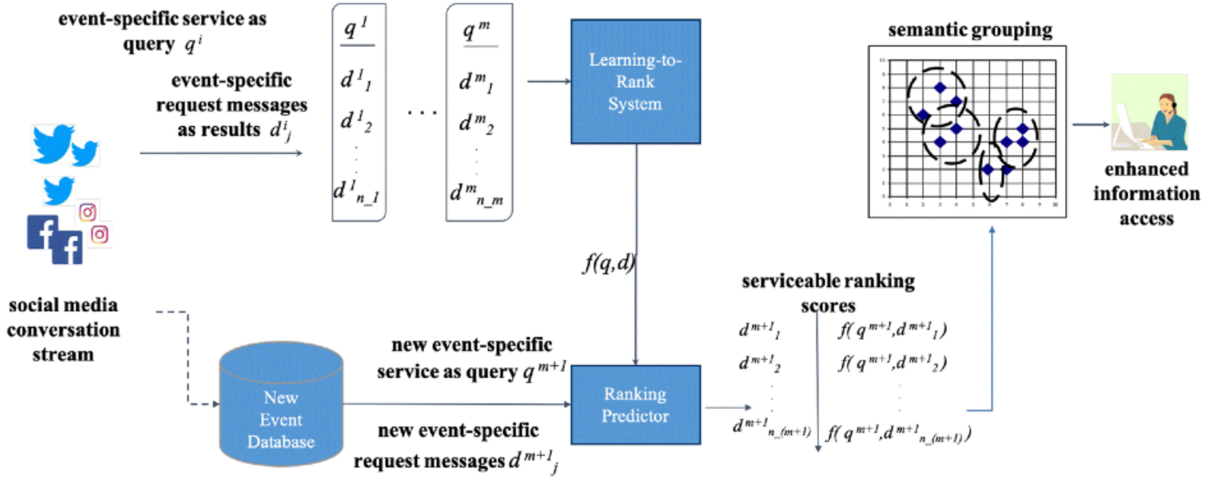
© Fabio Ciravegna, Università di Torino

Expert finding

- Mapping Requests for Information to Expertise
  - Short text clustering



Misuraca et al., 2020



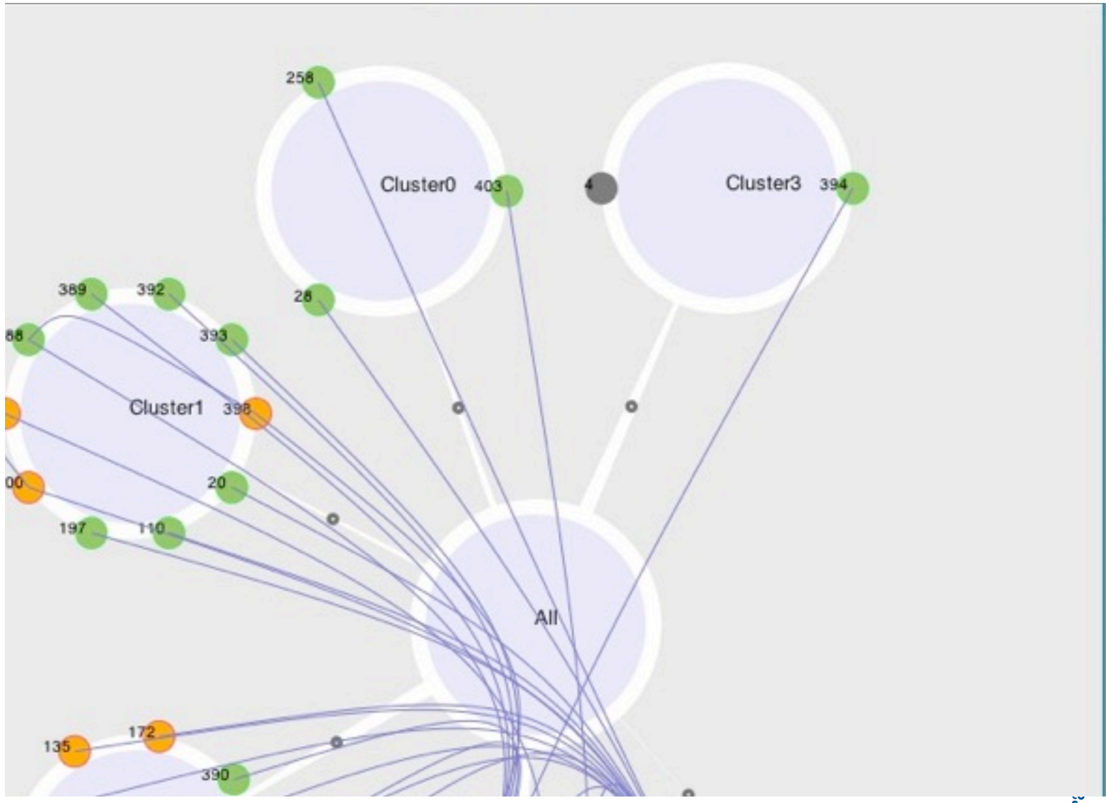
Purohit et al., 2020

Misuraca, M., Scepi, G., & Spano, M. (2020). A network-based concept extraction for managing customer requests in a social media care context. *International Journal of Information Management*, *51*, 101956.

Purohit, H., Castillo, C., & Pandey, R. (2020). Ranking and grouping social media requests for emergency services using serviceability model. *Social Network Analysis and Mining*, *10*(1), 1-17.

Modelling the Social Network

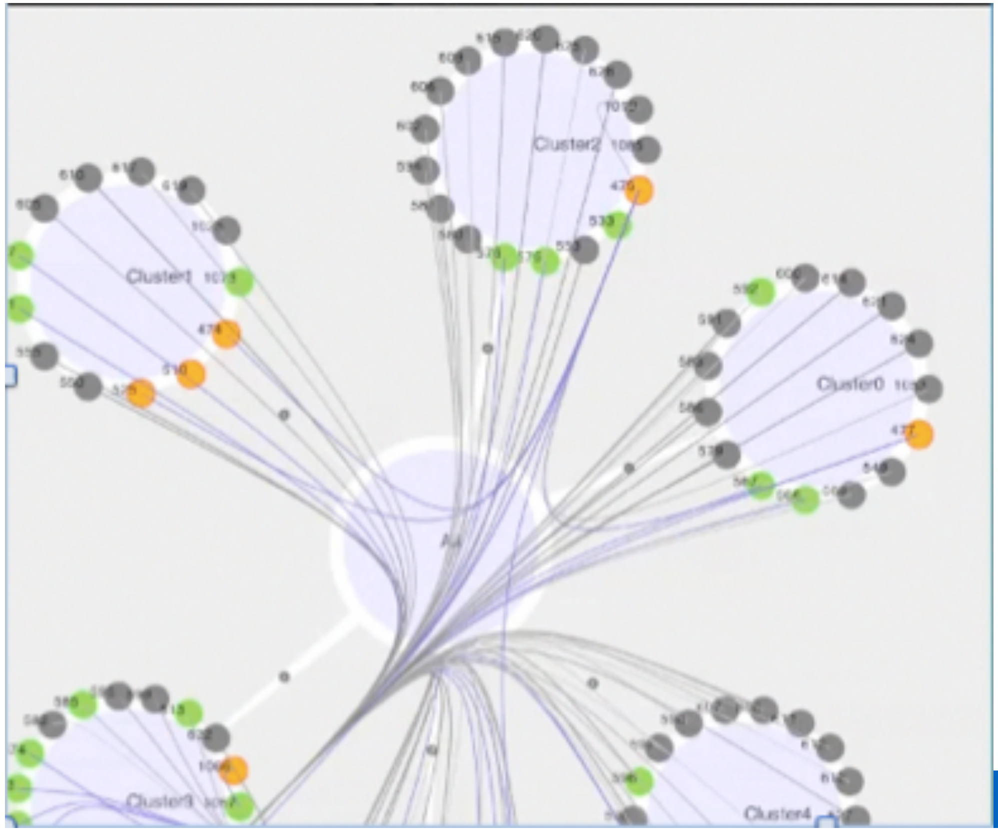
- Semantically Enriched Communication Network (SECN)
  - A formal representation of individuals and their communication exchanges
    - User profiles: a set of topics, weighted according to relevance to the user
    - Similarities between users based on their profiles
- A SECN is a typed, weighted graph:
  - Typed: nodes and edges within the graph are of several different types
  - Weighted: types of edges can be assigned a weight to boost importance of one type of connection or another



The University Of Sheffield.

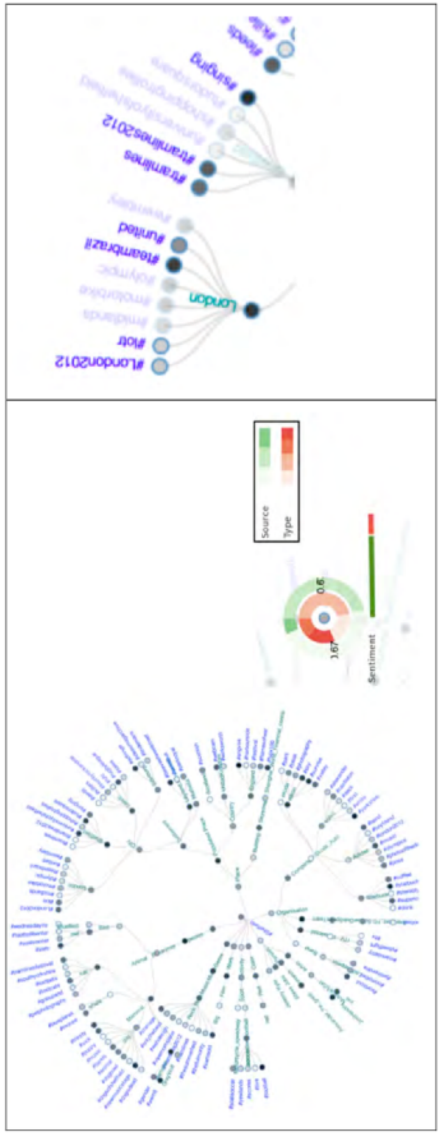
Semantic Social Networks

- Relationships are defined as communication exchanged by users on a specific topic
  - i.e. when selecting the topic "security" the network will show who talks with whom about security
- Profiles are built dynamically so are updated every time a new communication is sent



Social Influence Analysis

- We identify top rank influential users on the Twitter graph, given topics and/or an entities
- We use semantic trails left as side effect of tweeting, i.e.
  - the social relationship between a user retweeting a post and the author of the post
  - the relationship between a user and the topic of the post he retweeted
  - the relationship between a user and the entities (e.g. person, products) mentioned on the content of his posts or retweets



Elizabeth Cano Basave, Suvodeep Mazumdar, and Fabio Ciravegna: Social influence analysis in microblogging platforms-A topic-sensitive based approach in Journal of the Semantic Web 5(5):357-372 2011

UNIVERSITÀ DI TORINO

Conclusions

© Fabio Ciravegna, Università di Torino

Thanks

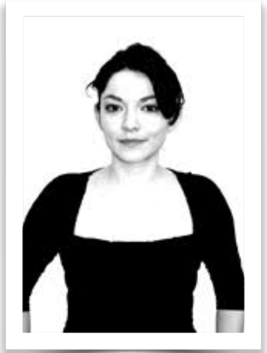
If I have seen, it was by standing on the shoulders of giants

(Isaac Newton)

- Dr. Sooji Han
- Dr. Vitaveska Lanfranchi
- Dr. Ziqi Zhang
- Dr. Anna Lisa Gentile
- Prof. Isabelle Augenstein
- Dr. Suvodeep Mazumdar
- Dr. Elizabeth Cano Basave
- Dr. Andrea Varga
- Dr. Sam Chapman
- Dr. Victoria Uren
- Prof. Daniela Petrelli
- Jie Gao
- Neil Ireson
- and around 45 others…









Thank you!

Prof. Fabio Ciravegna
Dipartimento di Informatica
Università di Torino

fabio.ciravegna@unito.it