Map of Technology: Topic modeling full-text patent data

GTM Conference 2015

Atlanta, GA

Arho Suominen & Hannes Toivanen

16.9.2015

INTRODUCTION

JUSTIFICATION

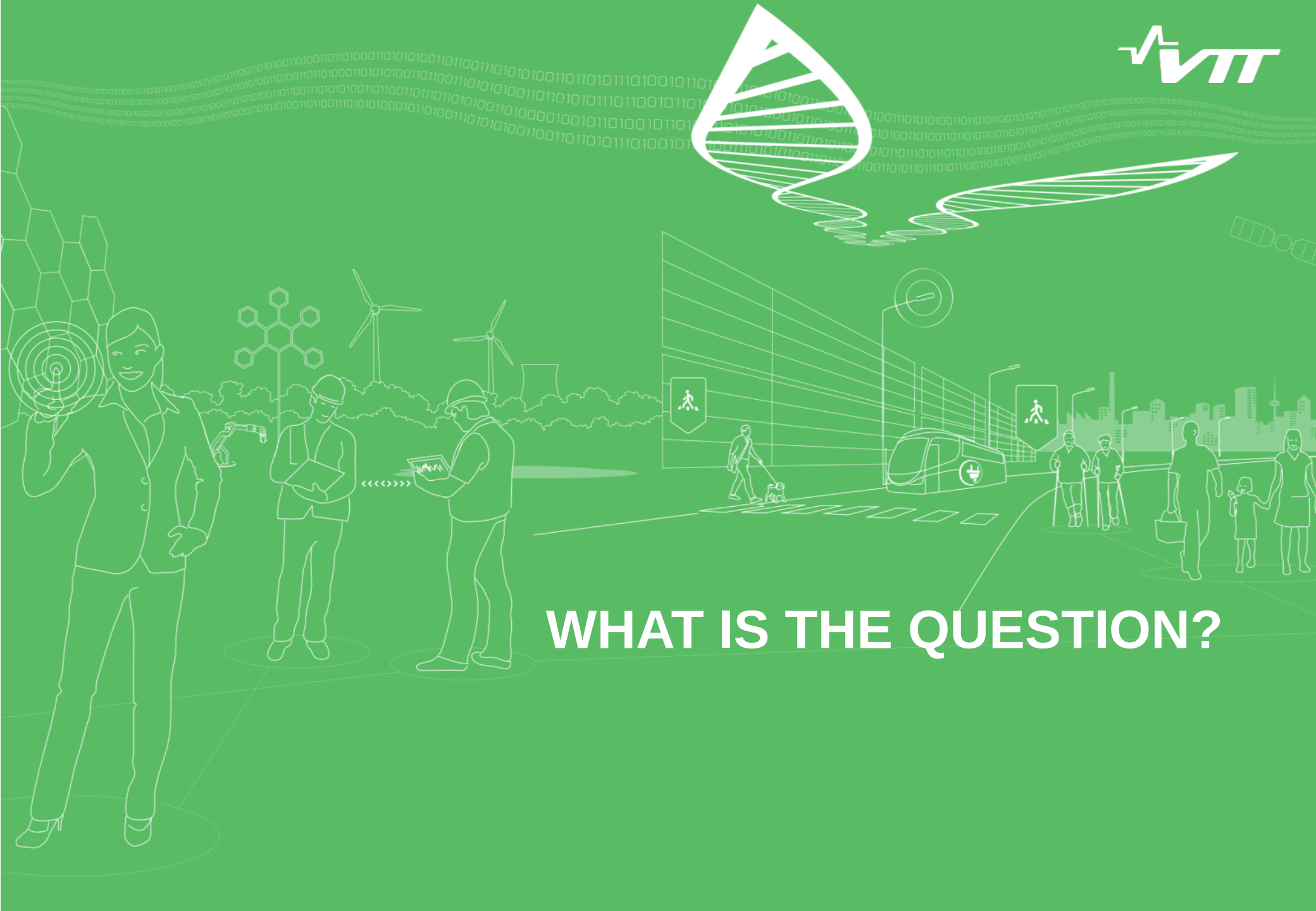
- A central challenge for the mapping patents is the creation of valid and accurate coordinates.

CONTRIBUTION

- Our study discusses the choice of the origin of coordinates in order to make a map of technology, and, in particular, demonstrates the advantages of unsupervised learning-assigned coordinates over those created by human reasoning.

INTRODUCTION

- Previous studies on mapping patent information:
  - specifically focusing on patent maps e.g. Yoon et al. (2002); Lee et al. (2009); Kim et al. (2008)
  - citation analysis e.g. Karki (1997) and Daim et al. (2006)
  - technology roadmapping e.g. Yoon and Phaal (2013)
  - Text mining e.g. Tseng, Lin and Lin (2007)
- Above example are not an exhaustive list.


WHAT IS THE QUESTION?

INTRODUCTION

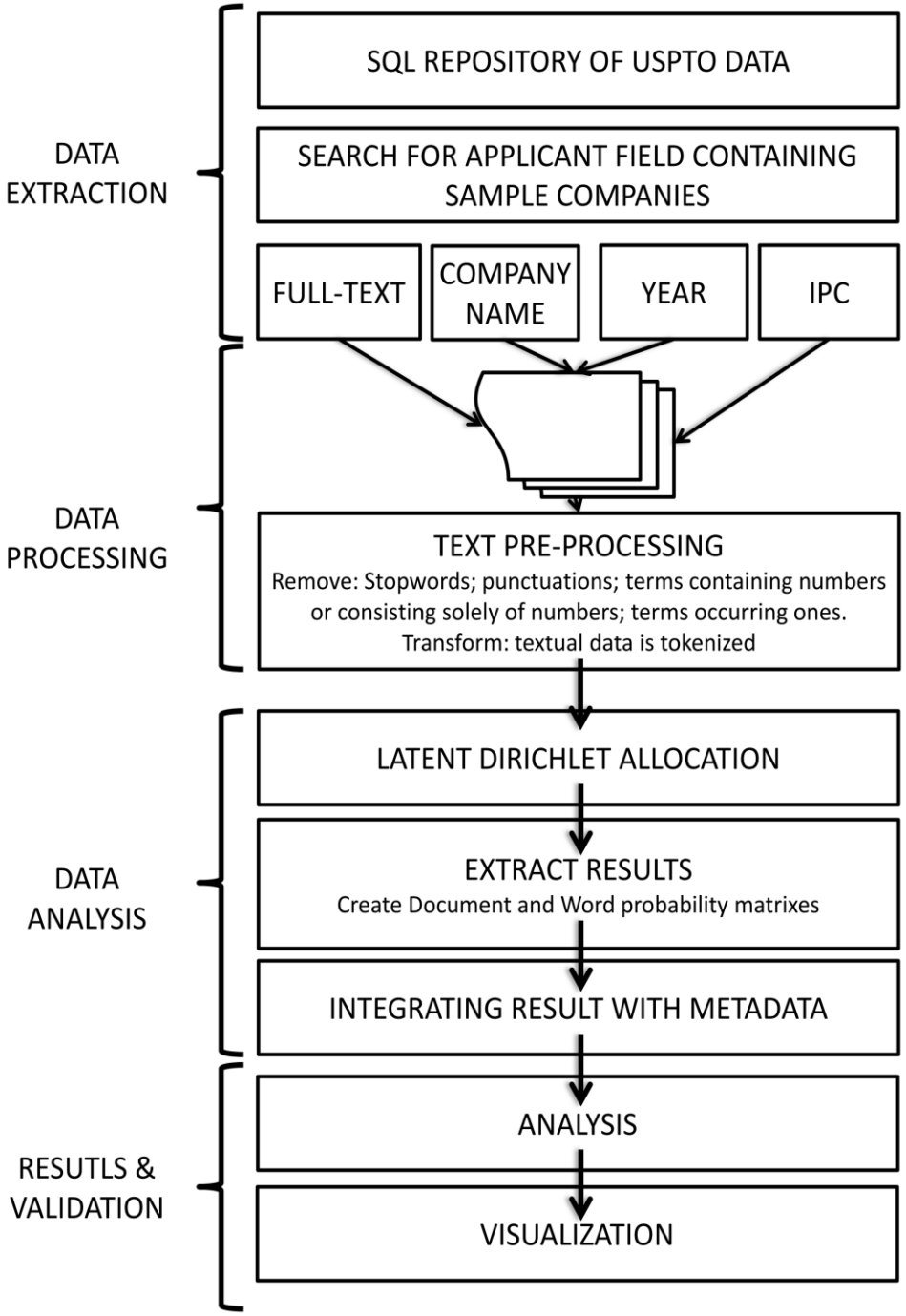
- Using patent classifications to form a patent map is not without limitations.
- Connecting patent classifications directly to industry sectors is challenging.
  - Classifications are also of limited value in directing inventive effort
  - The human process related to assigning classes to patents is valuable in the patenting process, even to the extent that automated classifications fall short of providing similar results.
- ARGUMENT: Machine-learning can offer an alternate structure of coordinates for patent data that creates deeper understanding on how technology is created

UNSUPERVISED LEARNING

- Produces an outcome based on an input while not receiving any feedback from the environment.
  - reliance on a formal framework that enables the algorithm to find patterns.
- Topic models " ...can extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computer". (Blei & Lafferty 2009)
- As a simplification each document in a corpus is a random mixture over latent topics, and each latent topic is characterized by a distribution over words.

DATA, PRE-PROCESSING AND ANALYSIS

- SAMPLE: one year of patent applications
  - We assume one year of patenting representing a adequate sample of patents to use as a basis of analysis.
- DATA: full-text patent descriptions filed in the USPTO containing approximately 6 million patents. The repository, hosted by Teqmine Analytics Ltd
  - The analysis was limited to 2014.
  - Final data contains 374,704 full-text records.



ALGORITHM: LDA

The algorithm is based on an online variational Bayes algorithm for LDA [9]

Number of Topics used was set using a trial-and-error & Arun et al. (2010) approach to 200.

IMPLEMENTATION: Python

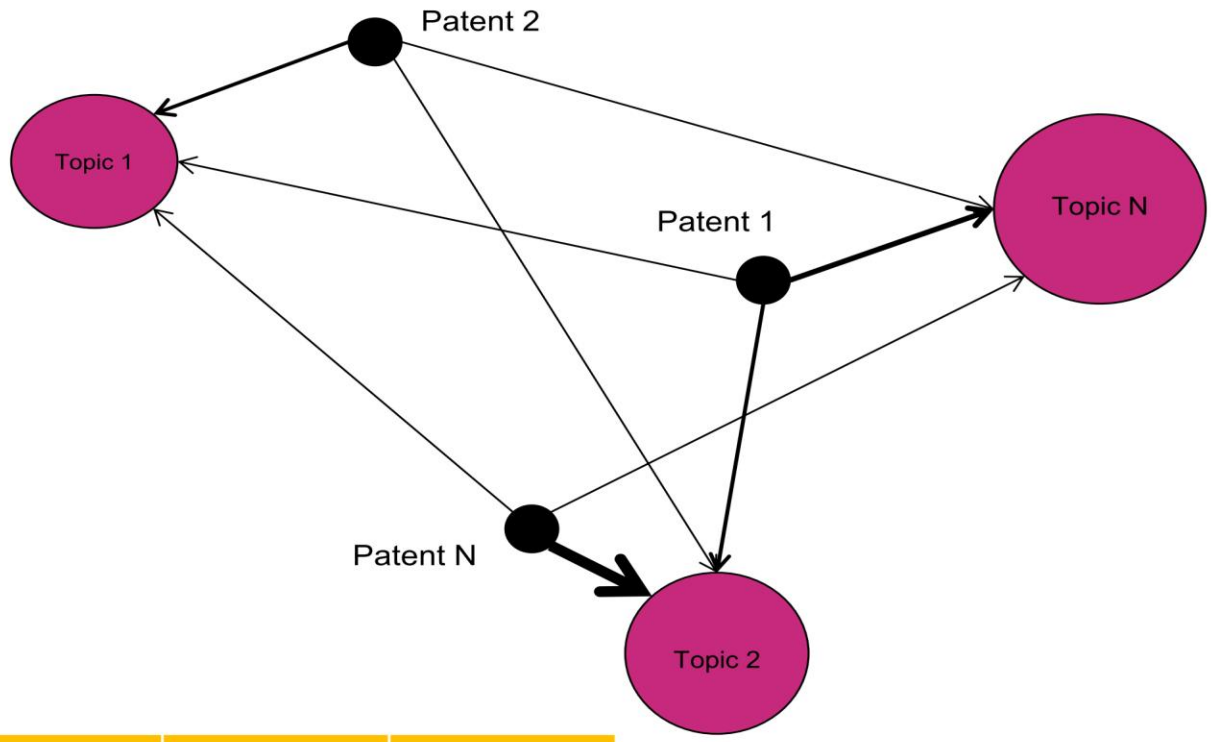
Python implementation included pre-processing

ANALYSIS: Python, Gephi & R

Networkx package in Python was used to decompose two-mode network data

Gephi was used to create visuals from the soft classification created by the algorithm. The modularity algorithm within Gephi was used to cluster network data.

R was used for additional calculations



	Topic 1	Topic 2	…	Topic N
Patent 1	0.10	0.24		0.40
Patent 2	0.40	0.01		0.10
…				
Patent N	0.01	0.80		0.01

RESULTS

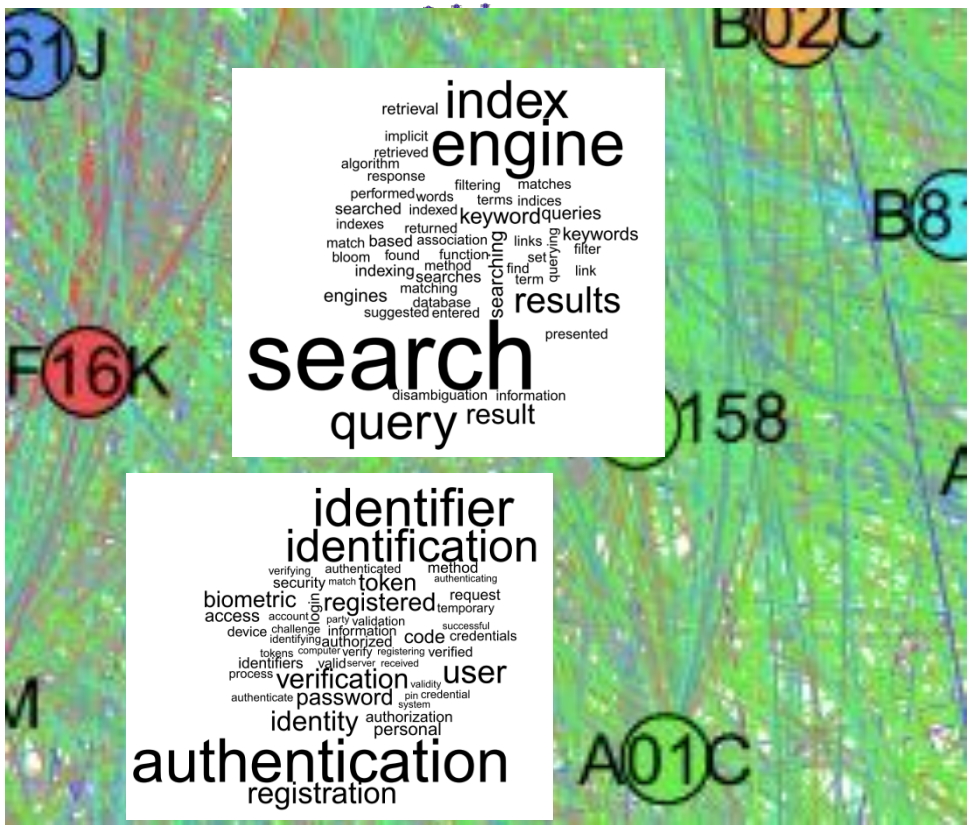
RESULTS

- RAW OUTPUT FROM ALGORITHM:
  - A soft classification of document probabilities to belong in a topic
  - A probability matrix of token probability to belong in a topic.
- RAW DATA TRANSFORMATION:
  - Document probability data was transformed to bipartite networks data using a Python script
    - Nodes are defined as Unsupervised learning based topics and IPC classifications
    - To diminish complexity patent data classes were limited to the third sub-classification
    - Edge weight is derived from topic IPC class co-occurrences. Counting was done with a fractionalized counting scheme.
    - Networkx package was used to create a one-mode projection of the data.

RESULTS

- VISUALIZATION:
  - Gephi was used to visualize networks.
    - OpenOrd and Fruchterman-Reingold algorithms were used to define node positions.
    - Modularity algorithm by Blondel et al. (2008) was used to derive latent clusters in the networks. Algorithm was run with several resolutions.
- ADDITIONAL ANALYSIS
  - Case by case analysis of unsupervised learning topics and IPC classes

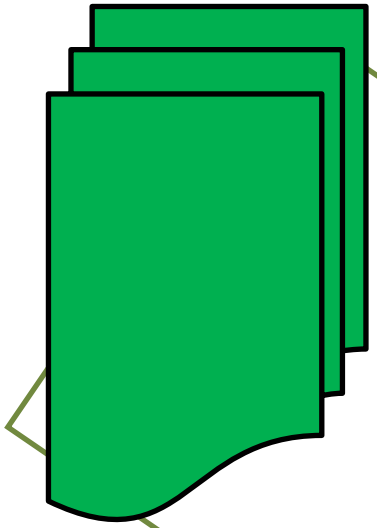
Bipartite network

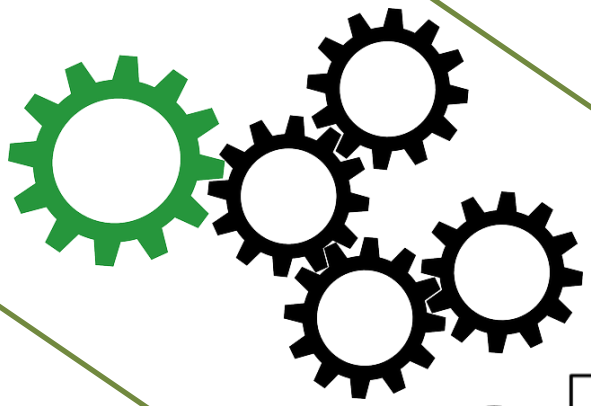


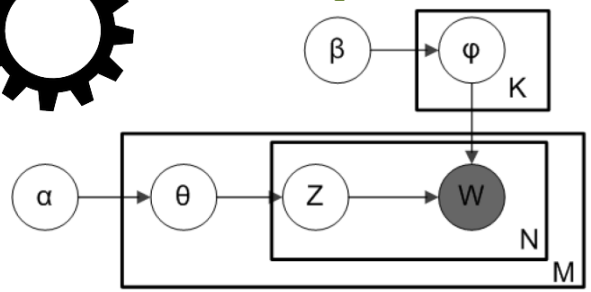
- Bipartite network show unsupervised learning and IPC class nodes and weighted edges based on a fractional counting scheme.
- Figures are drawn from a 818 nodes (topics and IPC classes) and 18373 edges between nodes.
- Merging the two modes are challenging – what is the counting scheme versus probability scoring of documents?

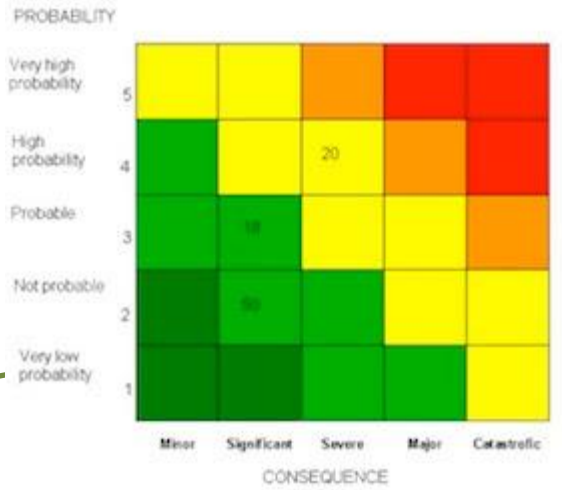
One-mode projection

- Python Networkx package was used to project the bipartite graph to one-mode projection
- Figures are drawn from 200 unsupervised learning based topics (nodes) and 19773 edges between nodes.






β
φ
K
α
θ
Z
W
N
M


PROBABILITY
Very high probability
High probability
Probable
Not probable
Very low probability
5
4
3
2
1
20
Minor
Significant
Severe
Major
Catastrofic
CONSEQUENCE

Conclusions

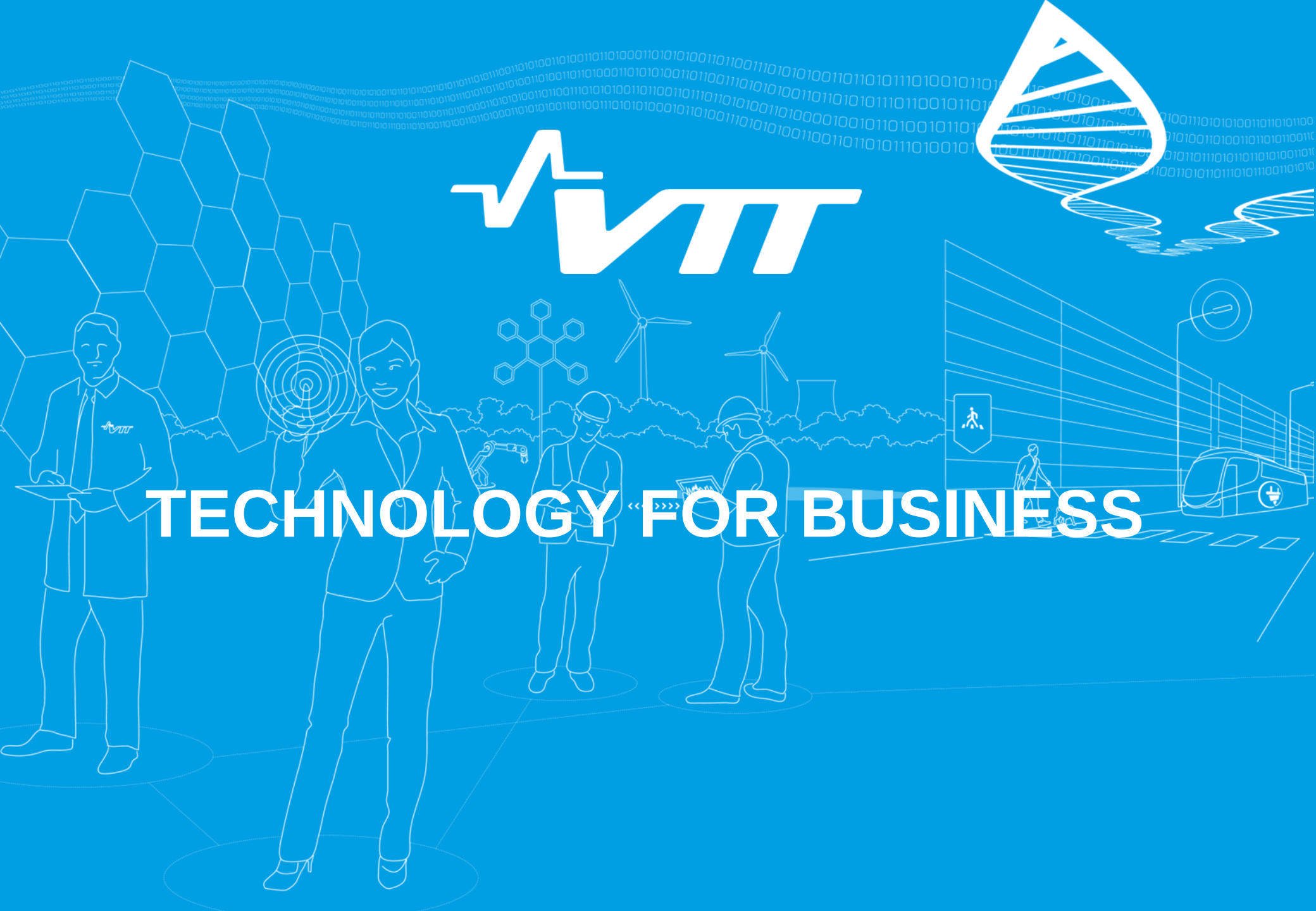
- Unsupervised learning offers an approach to classify large semantic datasets.
  - Optimally uncovering latent patterns without human intervention.
- Data represents a key challenge
  - What is the amount of preprocessing done that keeps semantic variability but loses noise.
- Merging automated classes and human given labels should be unproblematic – keeping that both are in a sense correct
  - However, weighting schemes used in this study seem to through the analysis of at several stages.

QUESTIONS?

Dr. Arho Suominen
Senior Scientist
VTT Technical Research Centre of Finland
Postdoctoral Researcher
Academy of Finland
arho.suominen@vtt.fi
@ArhoSuominen
https://fi.linkedin.com/in/arhosuominen

Dr. Hannes Toivanen
Principal Scientist
VTT Technical Research Centre of Finland
Adjunct Professor
Lappeenranta Technical University
hannes.toivanen@vtt.fi

The Academy of Finland supported this work with the research grant (288609) awarded to Dr. Arho Suominen for the project "Modeling Science and Technology Systems Through Massive Data Collections ". The project was also supported by Tekes—the Finnish Funding Agency for Technology and Innovation research grant awarded for the project "Radical and Incremental Innovation in Industrial Renewal". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript


TECHNOLOGY FOR BUSINESS