

# Overlay of science and technology patterns with unsupervised learning: Case of thermal management system

**Tech Mining Conference 2015**

Samira Ranaei<sup>1</sup>, Arho Suominen<sup>2</sup>, Mika Lohtander<sup>1</sup> and Tuomo Kassi<sup>1</sup>

Lappeenranta University of Technology<sup>1</sup>, Finland

VTT research center<sup>2</sup>, Finland

# Background



Open your mind. LUT.  
Lappeenranta University of Technology

Methods for measuring science and technology interaction: (Mayer 2000, Narin et al 1995, 1997)

1. Industrial publication
2. University patenting
3. Non-patent literature

Disagreement over the reliability of patent citation analysis to assess science and technology relationship:

- Patent citations only reference novel arts or limited output and cannot reveal the complete knowledge transfer flows of patent innovation . ([Jaffe and Trajtenberg , 2002](#), [Criscuolo & Verspagen, 2008](#)).
- Patents Citation patterns vary significantly by firm, industry, and even country characteristics ([Alcacer et al. 2009](#))
- Firms' citation choices can be strategic ([Hegde and Sampat 2009](#)), Meaning of citation behaviour ([Bornmann & Daniel, 2008](#), [Martyn, 1964](#)).
- NPL citations not only scientific citations: Mixed set of other type of publications. conference proceedings, books, and many other non-scientific sources such as disclosure bulletins, abstract services, and so forth.

# Research objectives



Open your mind. LUT.  
Lappeenranta University of Technology

**Purpose:** Study the linkage between science and technology with content-based approach

**Research Question:**

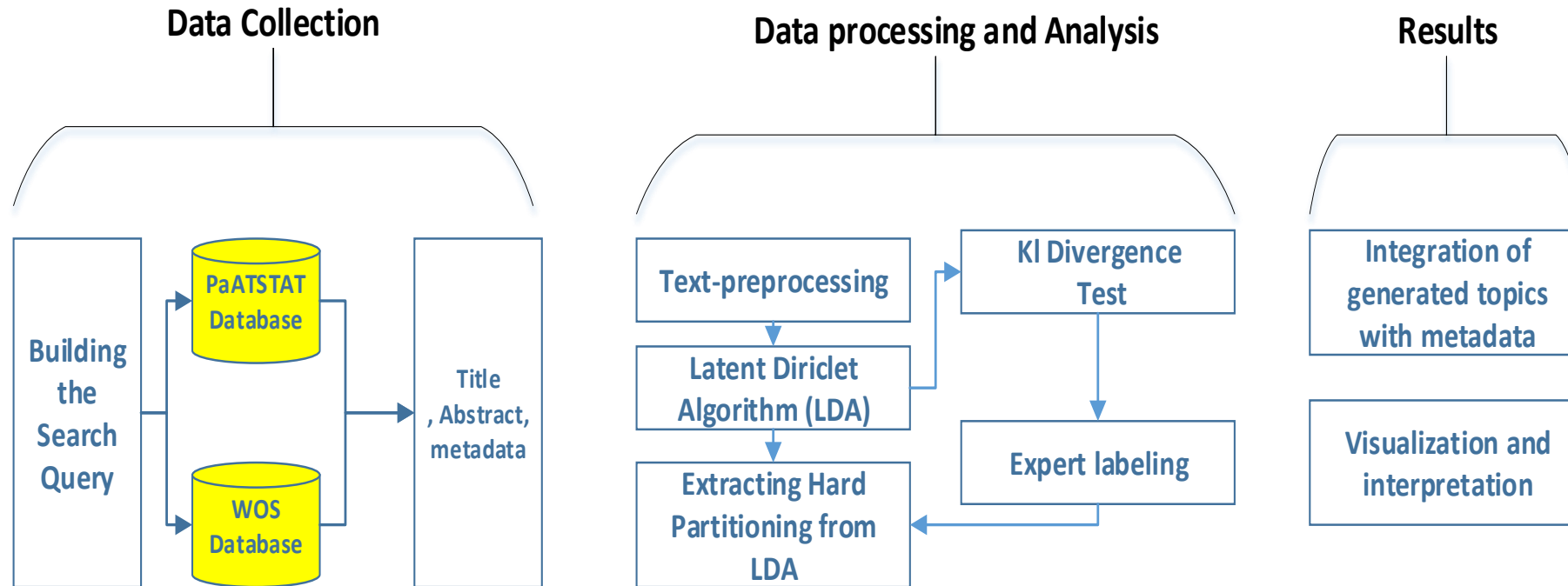
- How the content based indicator can be used to assess the science and technology relationship ?
- How feasible it is to apply unsupervised algorithms on textual data to measure science and technology linkage

**Case :** Thermal management systems- cooling methods used in electronic devices

# Data collection and methodology



- Data: 610 patents, 625 papers from 1980-2013
- Gensim Python library designed for topic modeling



# Latent Dirichlet Algorithm



### Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

### Documents

#### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,<sup>8</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions** "are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

### Topic proportions and assignments

# Challenges ...

- How many topics should be learned?
- How many learned topics are useful?
  
- How to measure the similarity between patents and publications?
- How different would be the result if we cluster patents and publication in one data set or separately?
- Mixed corpus
- Separated corpus

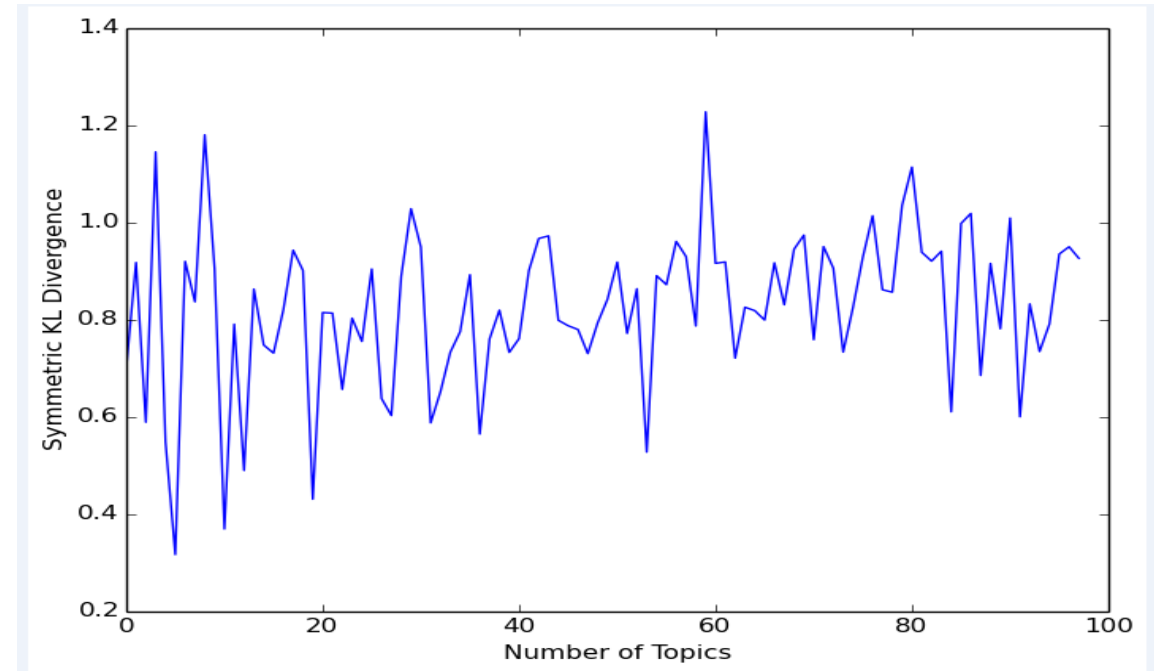


# Optimum number of topics

- Kullback -Leibler divergence measure
  - Less than 5
- Expert review
  - LDA generated 20 topics
  - Expert reviews the top keywords of each cluster and proposed 5 topics



Open your mind. LUT.  
Lappeenranta University of Technology



**Source :** Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N.(2010). [On finding the natural number of topics with latent dirichlet allocation: Some observations.](#) In Advances in Knowledge Discovery and Data Mining (pp. 391-402).



# LDA on mixed data set

1. Title and abstract of 1235 docs
2. Pre-processing = tokenization, downcasing, stopwords removal, removing words appearing once
3. Symmetric Dirichlet priors  $\alpha=0.5$ ,  $\beta=0.1$ , 1000 iteration
4. Number of topics = 5
5. Hard Clustering
  
6. Manual screening of the results
  - 10 docs from each cluster
  - Only 4 documents were clasified by experts differently

Doc Type	Topic a	Topic b	Topic c	Topic d	Topic e	Top topics
Patents1	0,0259	0,7479	0,0262	0,0258	0,1743	b
Publication857	0,0204	0,0209	0,0205	0,0206	0,9177	e
Publication862	0,0279	0,8899	0,0275	0,0273	0,0274	b
Patents3	0,0309	0,0308	0,8765	0,0311	0,0307	c
Patents4	0,0338	0,0338	0,8649	0,0337	0,0337	c
Publication858	0,0203	0,0200	0,9196	0,0201	0,0201	c
Patents6	0,0251	0,9003	0,0249	0,0248	0,0249	b
Patents7	0,0323	0,0322	0,8716	0,0320	0,0320	c
Publication859	0,8300	0,1094	0,0201	0,0201	0,0203	a
Publication860	0,0235	0,9062	0,0235	0,0234	0,0234	b

Number of topics	Patents	Publication	Grand Total
Topic a	162	135	297
Topic b	151	95	246
Topic c	164	125	289
Topic d	79	125	204
Topic e	54	145	199
<b>Grand Total</b>	<b>610</b>	<b>625</b>	<b>1235</b>



# Top keywords in each cluster



## Topic a =Performance and efficiency

velocity

temperature

droplets

impact

evaluation

mass

experiment

## Topic b= cooling Methods variation

Air-cooling

Two phase cooling

Spray cooling

jet

Jet impingement

Dry cooling

laser

Sprary

Liquid characteristics

## Topic c=componets and parts

nozzle

water

condenser

channel

duct

chamber

plate

pump

transfer

## Topic d=application of methods

Skin

laser

treatment

patient

treat

device

human

epidermal

pulse

## Topic e= arrangementsd of components

duct

transfer

heat

flux

stage

surface

plate

steel

particle



# LDA on separated dataset

1. Title and abstract of 610 patents, 625 publications
2. Pre-processing = tokenization, downcasing, stopword removal, removing words appearing once
3. Symmetric Dirichlet priors  $\alpha=0.5$ ,  $\beta=0.1$ , 1000 iteration
4. Number of topics = 5 for each set
5. Hard Clustering
6. Cosine similarity measure

$$\text{Cosine similarity} = \frac{T_1 \cdot T_2}{\|T_1\| \cdot \|T_2\|} = \frac{\sum_{i=1}^n T_{1i} \cdot T_{2i}}{\sqrt{\sum_{i=1}^n T_{1i}^2} \cdot \sqrt{\sum_{i=1}^n T_{2i}^2}}$$

		Patent Topics				
		0	1	2	3	4
Paper Topics	0	0,8180	0,8460	0,8841	<b>0,9821</b>	0,9911
	1	0,7847	0,6030	<b>0,9879</b>	<b>0,9730</b>	0,8788
	2	0,6841	0,8060	0,9779	0,7700	<b>0,9102</b>
	3	0,5211	<b>0,9916</b>	0,8060	0,7073	0,8102
	4	<b>0,9911</b>	0,7567	0,9017	0,8516	0,8522

# Discussion



Open your mind. LUT.  
Lappeenranta University of Technology

- Patents and papers are different but share similar features (topics). They can be compared based on their similarities
- Both approaches show that semantic relationship exists between patents and publications of cooling systems. Therefore it is possible to measure S&T relationship using unsupervised algorithm.

**Limitations:** In the case of cooling methods, we learnt that patent and publications are utilizing almost similar vocabularies. The cosine similarity signals this issue. Therefore, more cases should be studied in terms of generalization.

Future work:

- Study the relationship of documents in each cluster.
- Combination of meta data with generated topics



Thank you for your attention



ABSTRACT. The analysis of citation networks of patents or papers has been extensively used to define the knowledge structure or linkage between science and technology(S&T). However, citation approach is limited due to the time lag, data coverage to cited or citing documents, and may under-represent the possible knowledge flow between S&T data sources. In this paper, it is assumed that the linguistic pattern of patents and publications illustrate their topical overlaps and would spot the potential growing fields in research or practice. The novelty of our approach is the utilization of topic modeling and expert opinion, in order to cluster patents and articles based on their content rather than citations. Applicability and accuracy of our method is tested on a corpus of documents in field of thermal management system.