Drawing safety insights and foresight from free-text reports

Combining mathematical and safety expertise

Corinne Bieder, ENAC
Thierry Klein, ENAC

CONTENT

Context & problem statement

Objectives & challenges

Research project approach

Results & perspectives

Overall objectives

- Get safety insights and foresight from the wealth of data collected on operations (normal, incidents, accidents)
- Start with the non-structured data available in reported events

Context

- A wealth of data, a limited use

FDM and the like for ATC, weather data…	Accidents/Critical incidents	Event reports (ECCAIRS)
- Normal operations - Systematic recording - Technical - Restricted access	- Safety critical outcome - Small numbers - Holistic (variety of aspects) - Public reports	- Heterogeneous safety impact - Representativity? - Quality? - Non-structured - Potentially contextualized and holistic - Centralized repository

Problem statement

- How to use a potentially rich source of non-structured data for safety (ASRs and ATC reports)?
- How to cross-feed structured and non-structured data?

Objective & challenges

- Cluster reports according to common issues (insight)
- Detect "isolated" atypical record and emerging concerns (foresight)

➢ Challenges
  - Linguistic: "normalizing" free-text
  - Mathematical: finding an appropriate 'distance'
  - Managerial: choose the most appropriate one & clustering rules

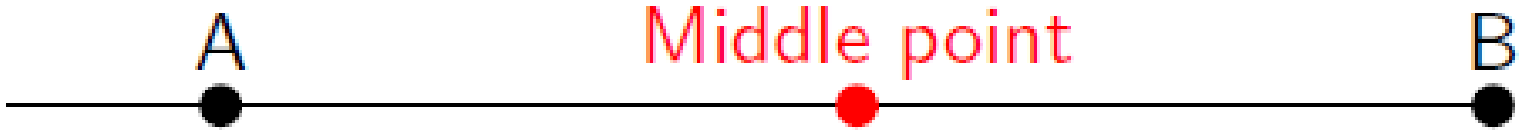
A preliminary linguistic processing

- Free-text challenges
  - Fail // failed…
  - Takeoff // Take-off // TO // Take-of…
  - Landing gear // gear…

- A pre-processing to associate a report with a list of "normalized" terms reducing the natural variability of narratives

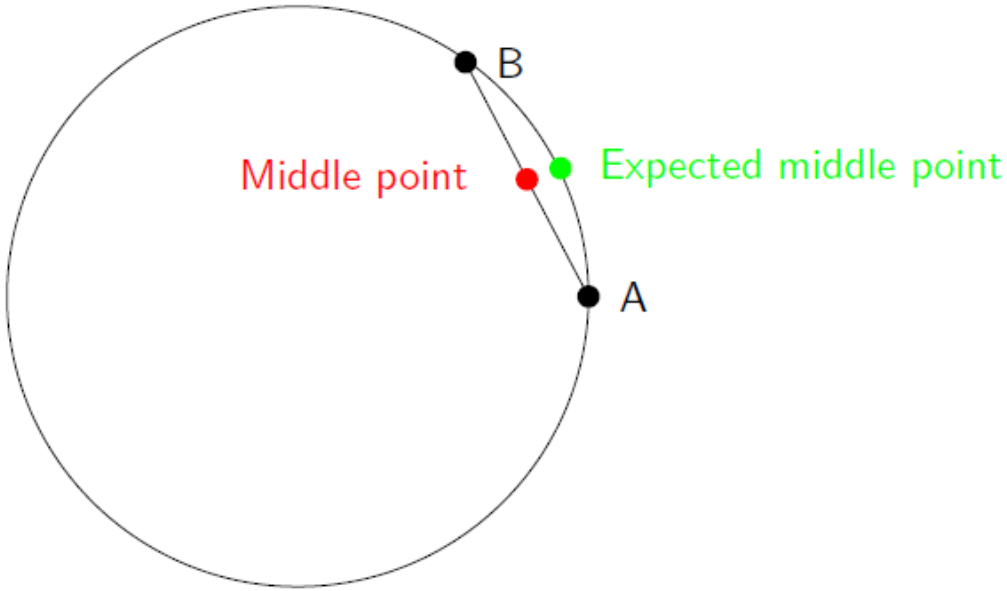
Determining an appropriate Distance - 1

- Easy to find in a Flat space



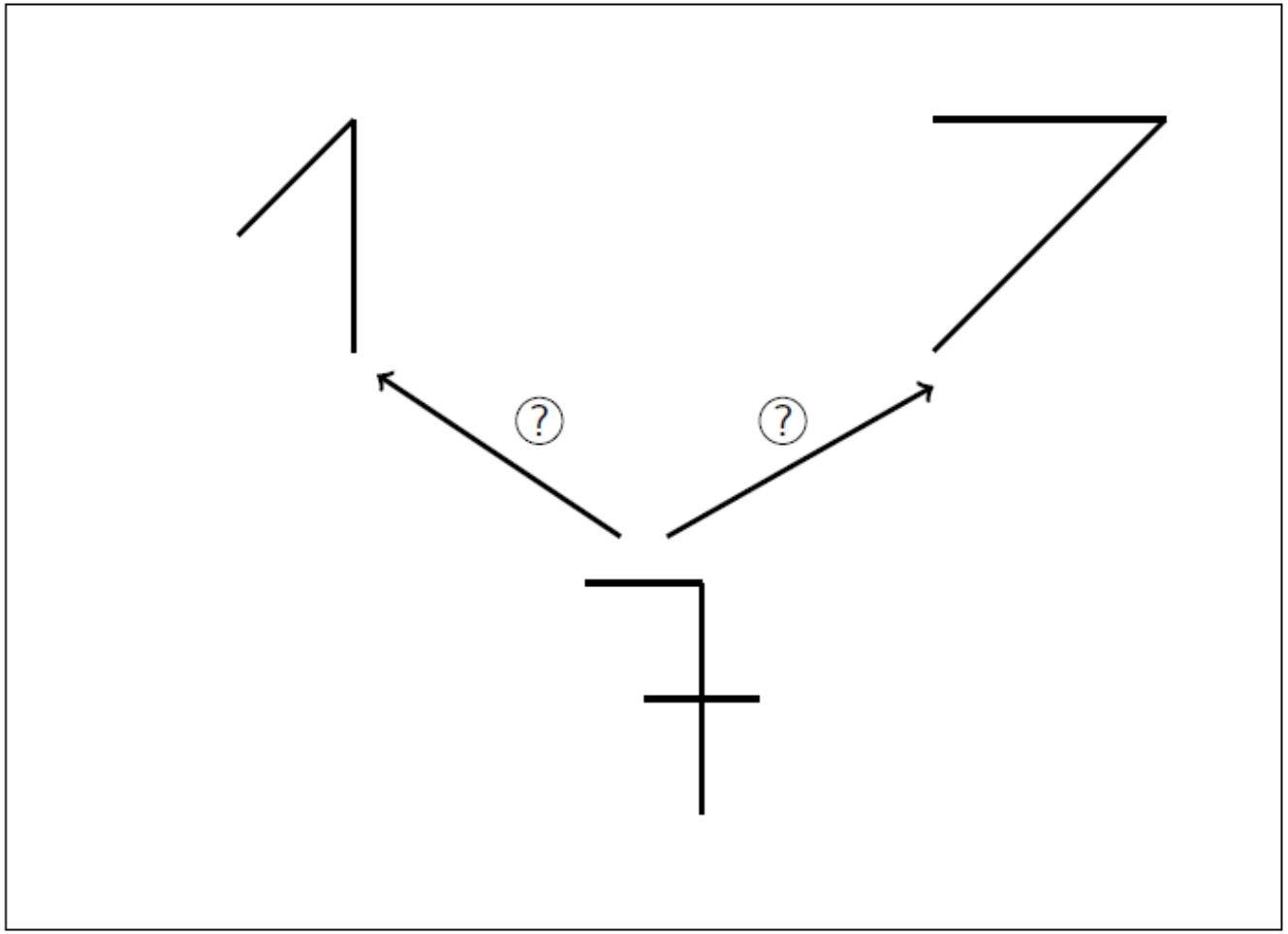
Determining an appropriate Distance - 2

- Relatively easy for a sphere

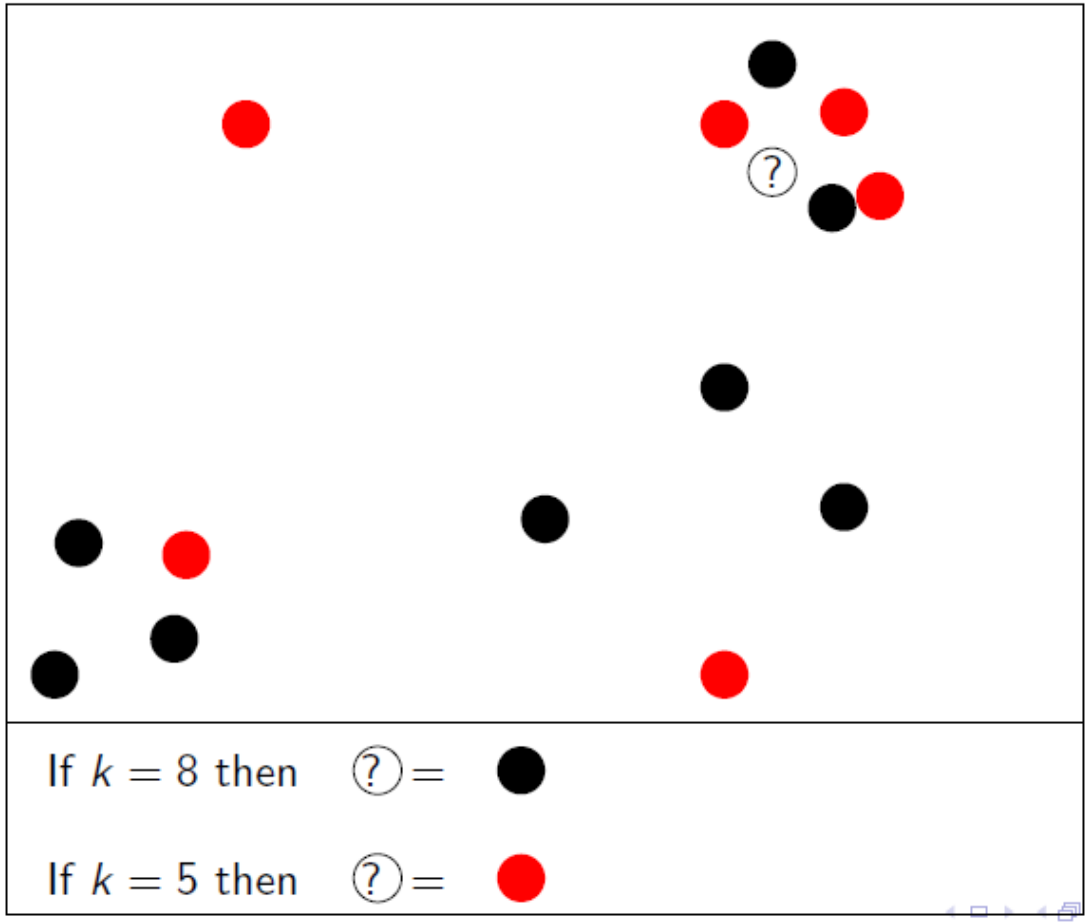


Determining an appropriate Distance - 3

- A real challenge sometimes
- An even more complicated challenge for text



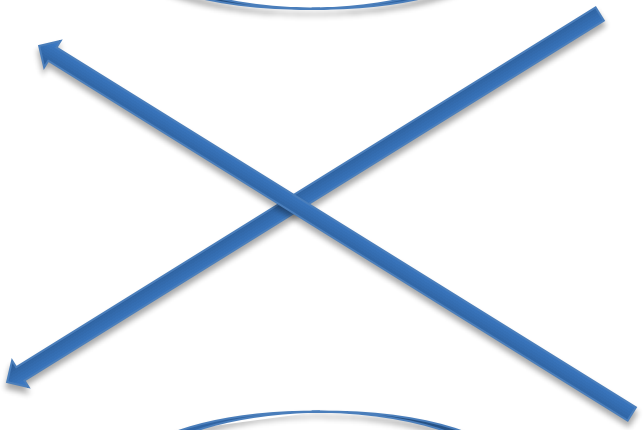
Distance is not everything: the clustering challenge



Mathematics & safety management cooperation

Topics development
Mathematical model

Topics Safety relevance check
Safety expertise



Distance development
Mathematical model

Safety distance sense making check
Safety expertise

Research Project-Step1

Study an existing text mining algorithm in the particular case of events reports.

- From a set a free text to a matrix of normalized terms (SD-CFH http://www.safety-data.com/)
- From the matrix to distribution of words and topics (Algorithm LDA)
- Study the distance used in the LDA Algorithm

Research Project-Step2

Use adavanced probability theory to define the good notion of distance between texts.

- Use the wassertein distance between distributions
- No natural ordering of the words or topics \color{black} appearing in the distributions\color

Results & Perspectives

- Results
  - Work in progress on French event reports (~10 000 events from 1st Jan. to 12 March 2017)
- Perspectives
  - Come up with consistent and safety meaningful mathematical model & algorithms
  - Explore the combination of structured & non-structured data