Lezione 6

Semantica documentale e visualizzazione

Luigi Di Caro

Semantica documentale e visualizzazione

- Panoramica
  - Topic Modeling
  - Dynamic Topic Modeling
- Text Visualization
  - Tag clouds, correlation circles, radviz, parallel coordinates, heatmaps, etc.
- Lab

Semantica documentale e visualizzazione

- Cos'è un *topic model*
  - Modello *statistico* o *probabilistico* che analizza l'uso del linguaggio ed individua automaticamente gli *argomenti* in una collezione di testi (o base documentale)
- Modello *non supervisionato*
  - Nessuna necessità di annotazioni manuali

Semantica documentale e visualizzazione

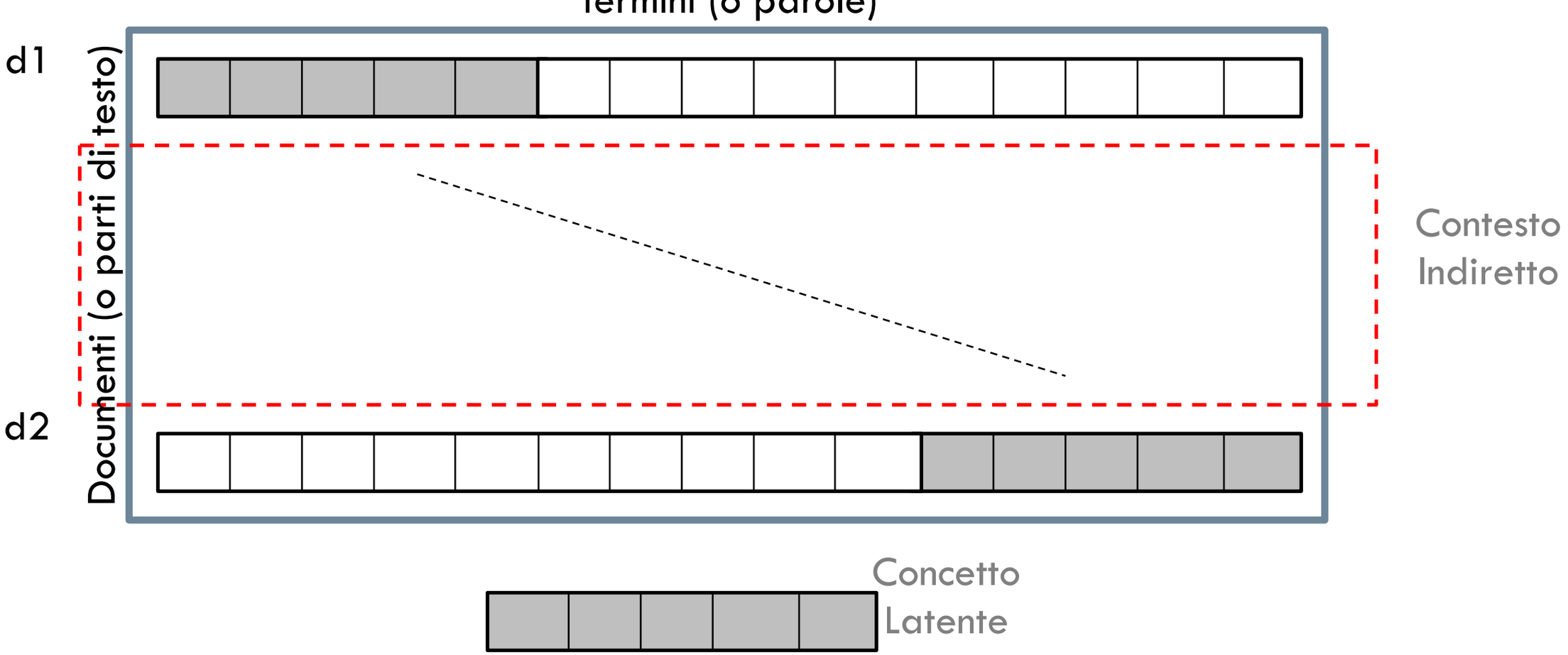
- Topic Modeling
  - Meccanismi non supervisionati per estrarre semantica del tipo
    - Di fatto, un topic è una *lista pesata di parole*
- Problemi:
  - Interpretabilità non ovvia
  - Topic estratti non sempre utili (es. function words, eventi anomali, numeri, combinazioni lessicali accidentali e non di contenuto)

Semantica documentale e visualizzazione

- Latent Semantic Analysis (LSA)
  - basata su una fattorizzazione matriciale chiamata *Singular Value Decomposition (SVD)*
  - Passaggio da matrice sparsa a densa, le cui dimensioni rappresentano combinazioni lineari delle features originarie
  - Cattura “concetti latenti”

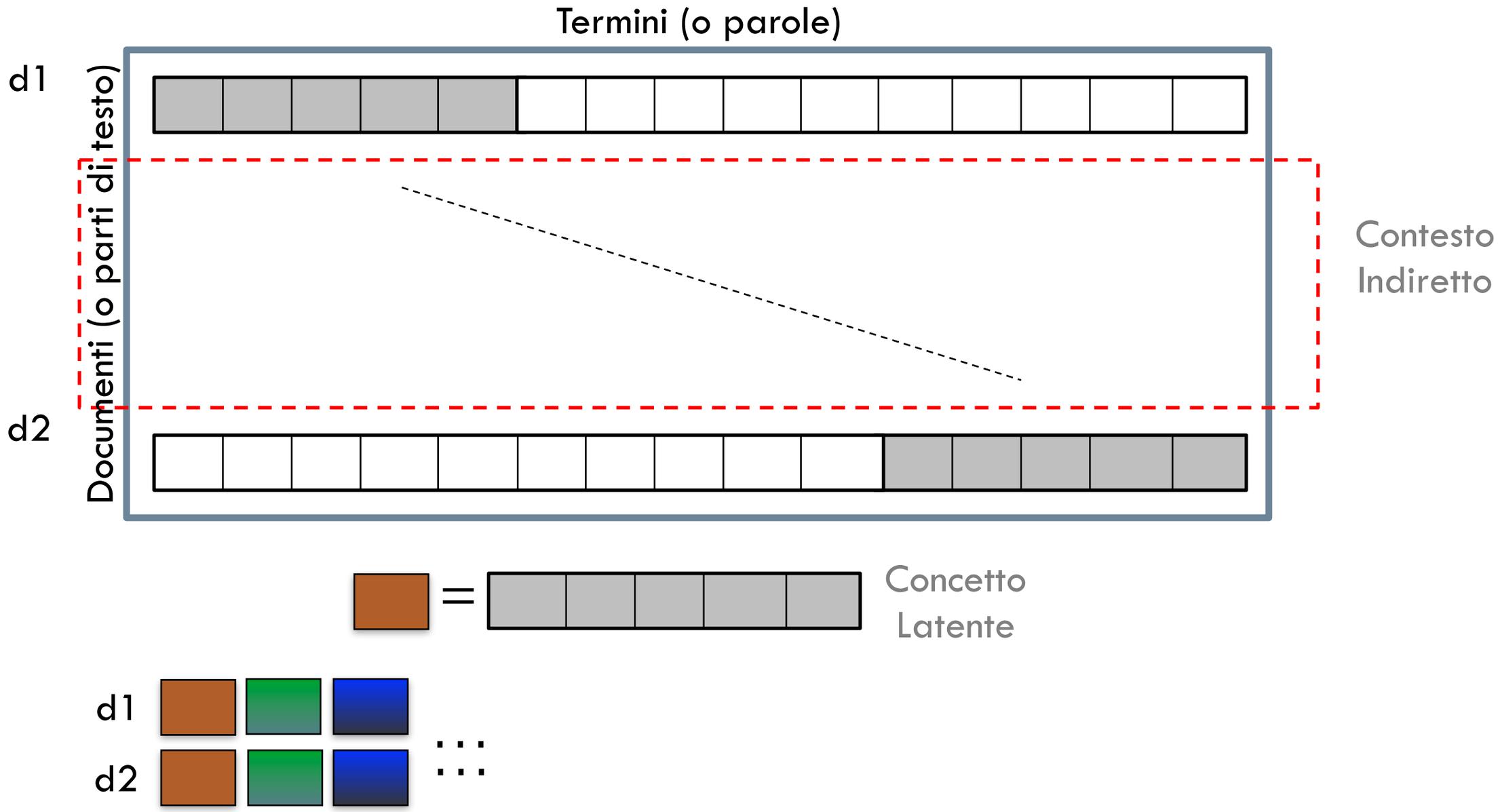
Trasformazioni matriciali

- Latent Semantic Analysis (LSA)
  - Analisi dei *contesti indiretti* e dei *concetti latenti*



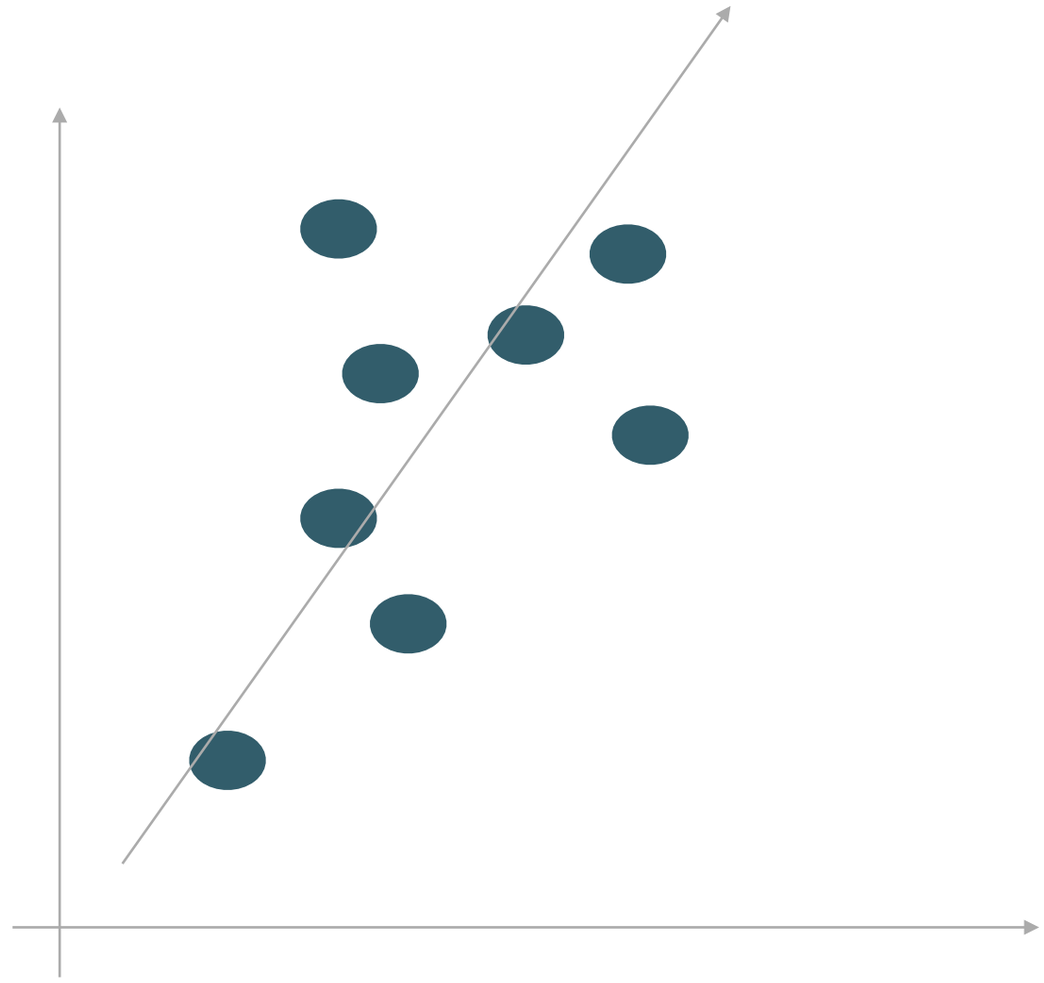
Trasformazioni matriciali

- Latent Semantic Analysis (LSA)



Trasformazioni matriciali

- Latent Semantic Analysis (LSA)



Semantica documentale e visualizzazione

- Latent Semantic Analysis
  - Problema: modello che non generalizza su documenti non visti
  - Problema: valori negativi dopo la fattorizzazione non di facile interpretazione
    - Esistono varianti: ad es. NNMF non-negative matrix factorization

Semantica documentale e visualizzazione

- Latent Dirichlet Allocation (LDA)
  - Estende una versione probabilistica della LSA, chiamata pLSA
  - Sfrutta la statistica Bayesiana
  - Si basa sull’assunto che un documento è un mix di topics, e ogni parola ha una certa probabilità che compaia in ogni singolo topic

Semantica documentale e visualizzazione

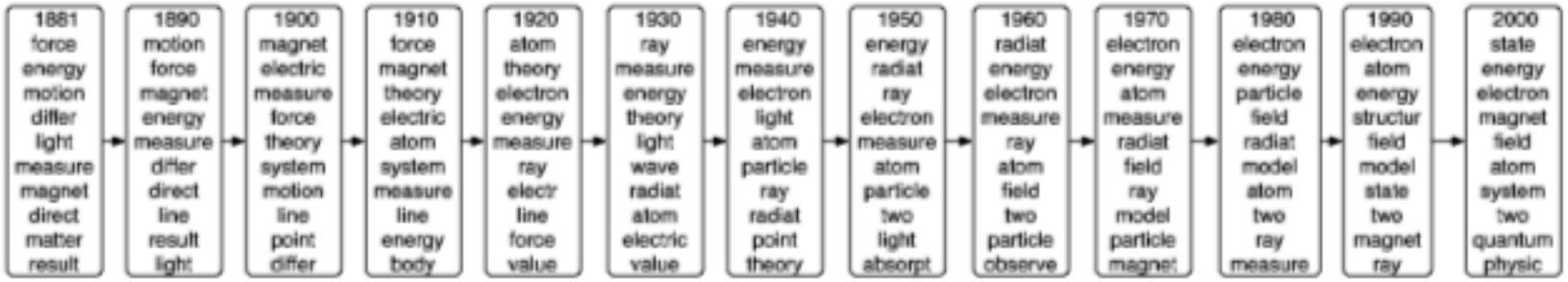
- Latent Dirichlet Allocation

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

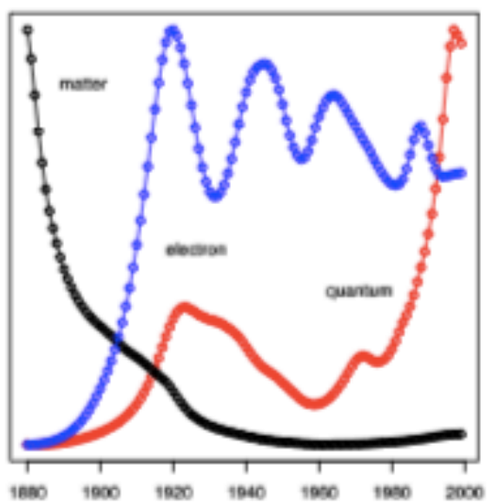
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Semantica documentale e visualizzazione

- Dynamic Topic Modeling
  - area di ricerca che si concentra sull'evoluzione dei topic nel tempo, avendo a disposizione un corpus annotato con valori temporali



"Atomic Physics"



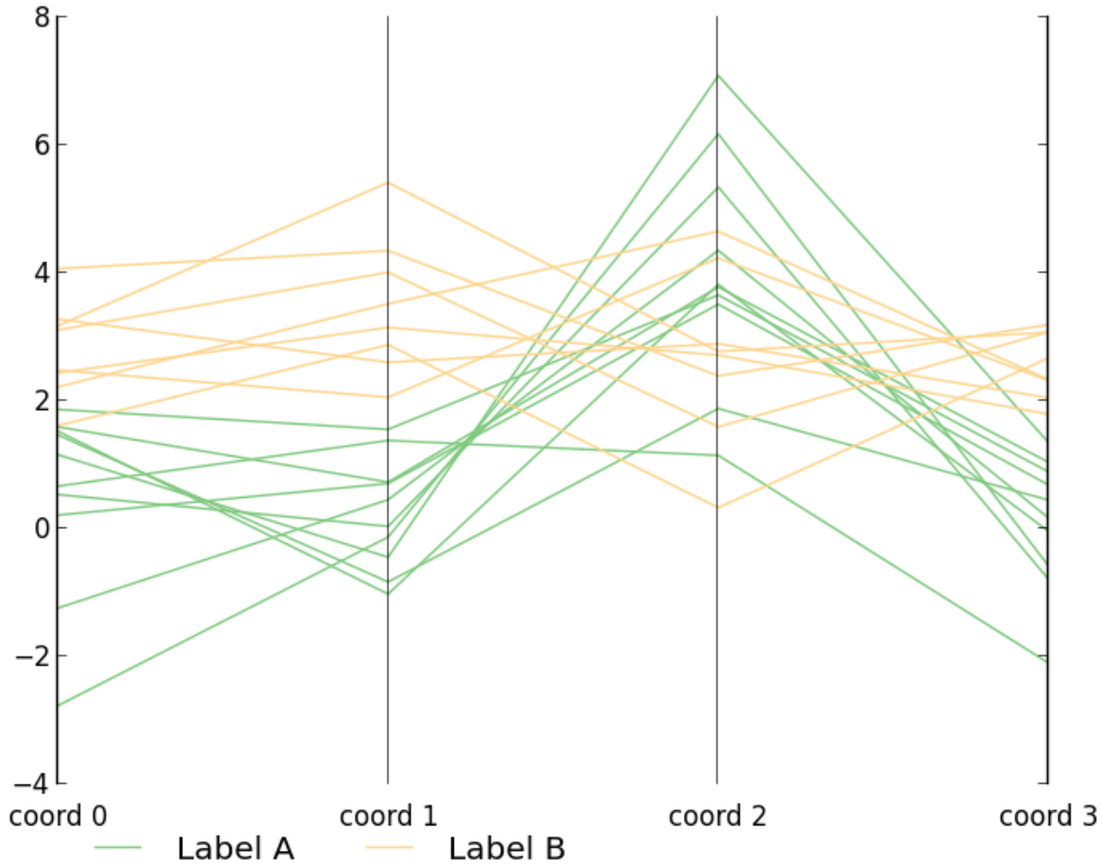
1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Semantica documentale e visualizzazione

- Text Visualization
  - Visualization è un'area di ricerca a parte, che prende spazio anche con dati di natura testuale
  - Diverse "*strategie*" di mapping multidimensionale
    - Fattorizzazioni matriciali, PCA, Multi Dimensional Scaling, SOM, etc.
    - Approcci grafici: Parallel Coordinates, RadViz, HeatMap, Correlation Circle, …

Semantica documentale e visualizzazione

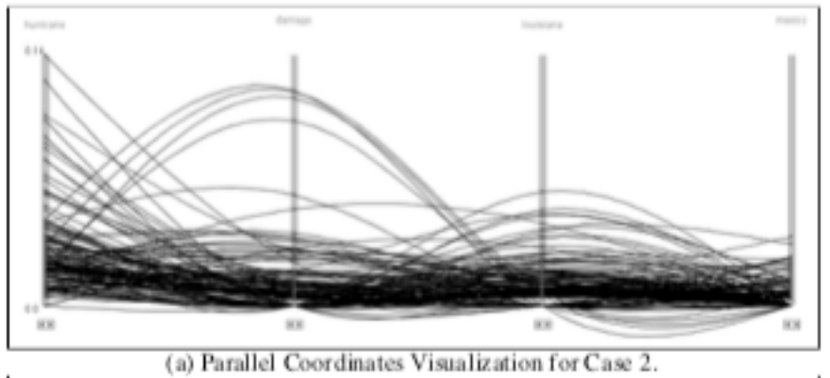
- Text Visualization
    - Approcci grafici
        - **Parallel Coordinates**
        - RadViz
        - HeatMap
        - Correlation Circle



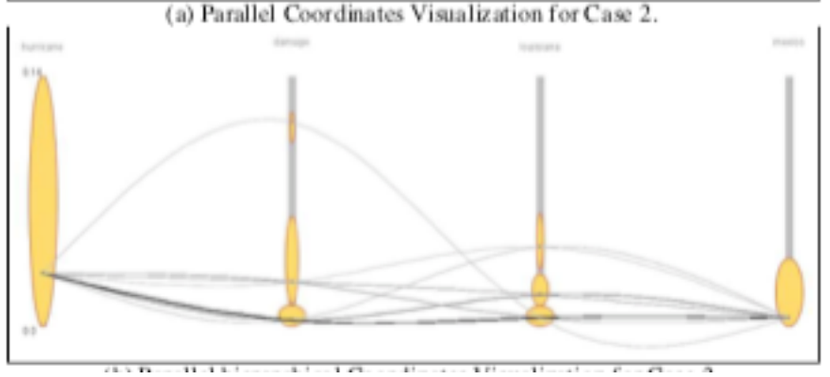
Semantica documentale e visualizzazione

- Text Visualization
  - Approcci grafici
    - **Parallel Coordinates**

      Estensioni
    - RadViz
    - HeatMap
    - Correlation Circle



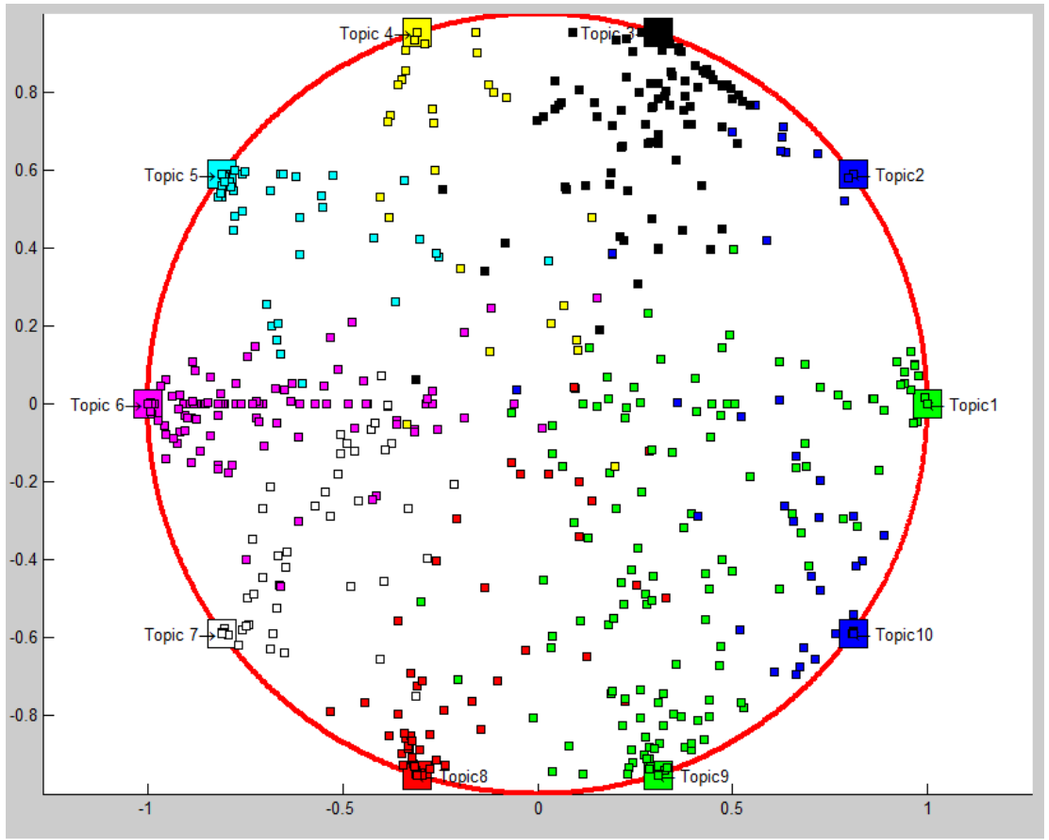
(a) Parallel Coordinates Visualization for Case 2.



KS Candan, Luigi Di Caro, Maria Luisa Sapino. PhC: Multi-resolution Visualization and Exploration of Text Corpora with Parallel Hierarchical Coordinates. Transactions of Intelligent Systems and Technology, 2012

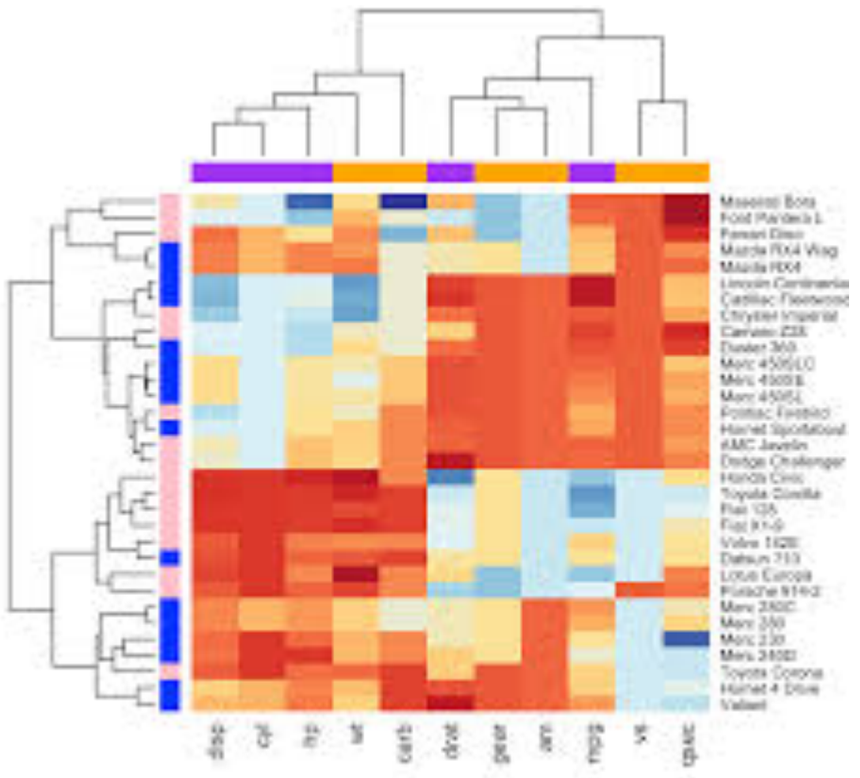
Semantica documentale e visualizzazione

- Text Visualization
  - Approcci grafici
    - Parallel Coordinates
    - **RadViz**
    - HeatMap
    - Correlation Circle



Semantica documentale e visualizzazione

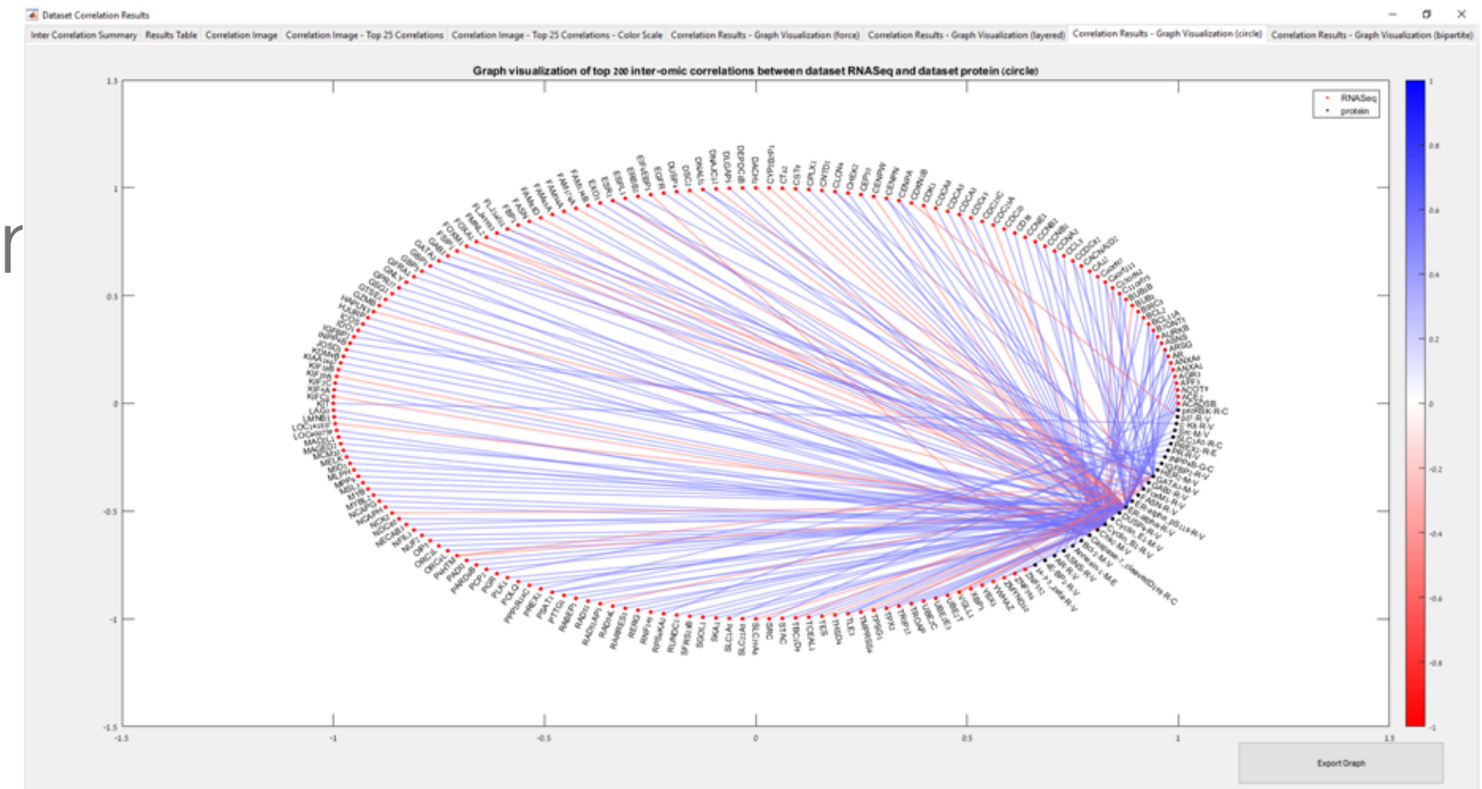
- Text Visualization
  - Approcci grafici
    - Parallel Coordinates



    - RadViz
    - **HeatMap**
    - Correlation Circle

Semantica documentale e visualizzazione

- Text Visualization
  - Approcci grafici
    - Parallel Coordir
    - RadViz
    - HeatMap
    - **Correlation Circle**



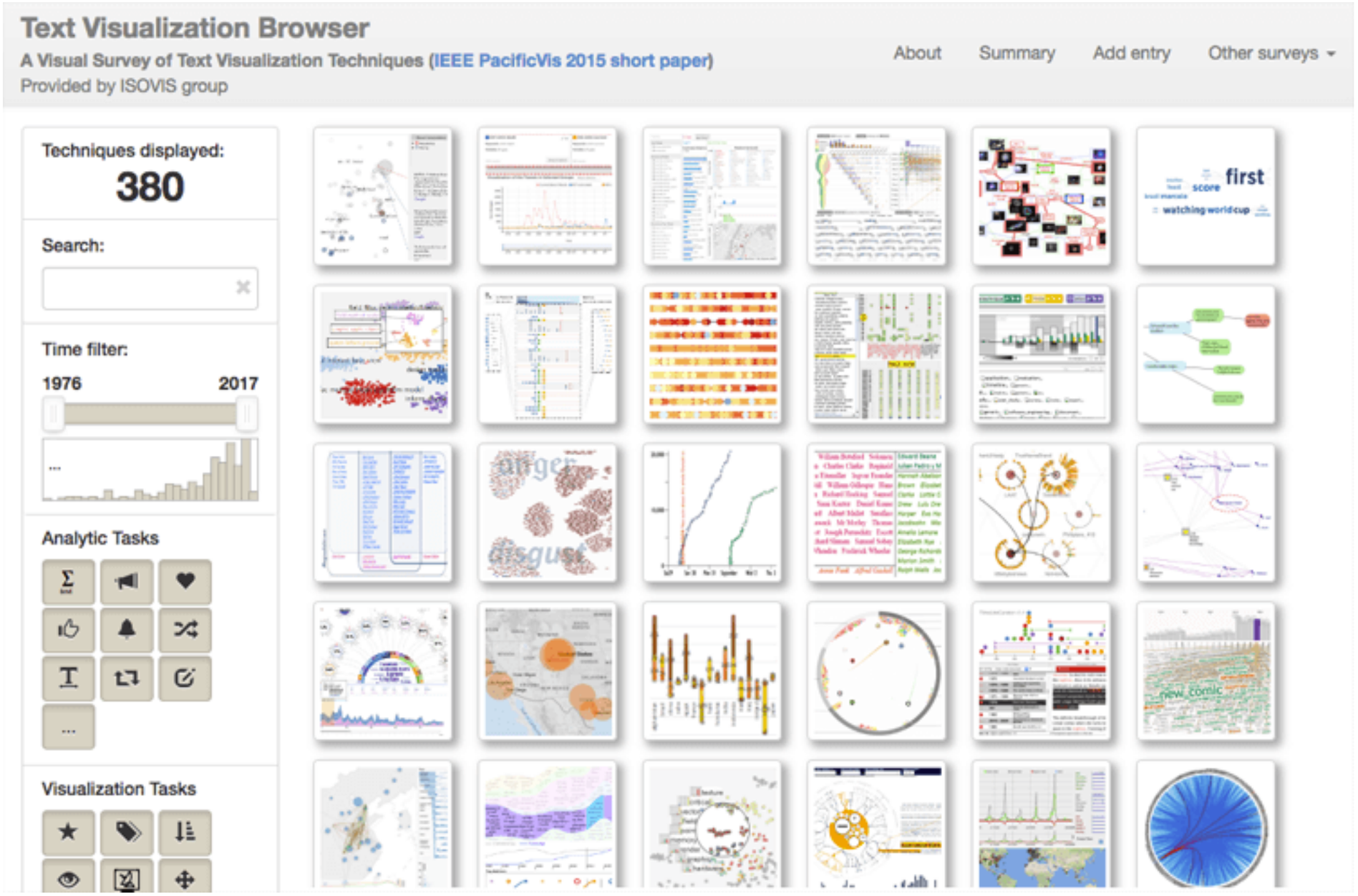
Semantica documentale e visualizzazione

- Text Visualization
  - Approcci grafici
    - Parallel Coordinates
    - RadViz
    - HeatMap
    - **Correlation Circle**

      Varianti

Semantica documentale e visualizzazione

- Text Visualization http://textvis.lnu.se



Laboratorio - *TM/TV*

- A scelta Topic Modeling e Text Visualization:
  - Topic modeling: partendo da un corpus, estrarre topics
    - Suggerimento: uso della libreria Gensim https://radimrehurek.com/gensim/
  - Visualizzazione: scelta di librerie di visualizzazione di dati testuali (qualsiasi!)