



**MET CS688 C1**

# ***WEB ANALYTICS AND MINING***

**ZLATKO VASILKOSKI**

INTRODUCTION

# General Course Information

- Instructor: Zlatko Vasilkoski (email: [zlatko@bu.edu](mailto:zlatko@bu.edu) or [zlatko.vasilkoski@gmail.com](mailto:zlatko.vasilkoski@gmail.com))
- Office hours: by appointment (you can email me with questions at any time)
- Class Time: Mondays from 6:00pm to 8:45pm, PSY B53
- Course Prerequisites:
  - MET CS 544 - Foundations of Analytics or
  - MET CS 555 - Data Analysis and Visualization
- Course Grading Policy:
  - Quizzes/Lab Projects (10%)
  - Assignments (20%)
  - Midterm exam (35%)
  - Term project (35%)
- Course Topics:
  - Web Analytics and Web Analytics Tools
  - Text Mining
  - Web Mining
  - Mining the Social Web, Twitter, Game Analytics
  - Data Visualization, Google visualization APIs illustrated on above mining examples
  - Basics of machine learning – **Let me know if there is an interest for it!**
- No specific textbook (reference textbooks are listed in the syllabus)

# My Background

- PhD. in physics from Tufts University working with David Weaver and Martin Karplus on computational implementation of the diffusion collision multi scale model of protein folding, to which the 2013 Nobel Prize for chemistry was awarded. The algorithms I designed as part of my thesis, greatly expanded the applicability of the the diffusion collision model.
- My college teaching experience includes BU, Tufts, MIT, Suffolk, Wentworth and Bentley.
- I have been developing the curriculum and working as a lecturer at the Metropolitan College Computer Science department since 2012.
- My work experience includes
  - Chief Data Scientist at FacilityConneX
  - Senior research scientist at Neurala Inc.
  - Postdoctoral research work at MIT and Northeastern
  - Worked in the area of Neural Network's learning laws at Department of Cognitive and Neural Systems, at BU.
  - Worked as data scientist at Harvard Medical School.
- My current research interests include algorithm development in computational physics, biomedical image processing, computer graphics, computer vision, machine learning and neural network systems for adaptive complex behavior in robots.

# Class overview

- This course covers the theoretical and practical aspects of
  - Web Analytics
  - Text Mining
  - Web Mining
  - Internet of Things (IoT)
  - Mining the Social Web
  - Game and Sports Analytics
- The web analytics part of the course studies
  - the metrics of web sites
  - their content
  - user behavior during web site visit
  - reporting
- In this class Google analytics tool is used for collection of web site data and implementing the analysis. The use of Google Trends and Google Correlate will be also illustrated.

# Class overview

- The text mining part covers the analysis of text and it includes
  - Preprocessing and content extraction from various file types
  - The mathematical (matrix) representation of the extracted text
  - String matching, fuzzy string matching, and their measures of closeness
  - Documents matching in the “concept space” and the simple math behind it
  - Aspects of supervised learning, Tagging, Classification, and Categorization
- The web mining (structure & content) part covers aspects such as
  - Web crawling (gathering pages from the web )
  - Indexing (to support a search engine)
  - Understanding Search Performance and how to measure it
  - The graph representation of the web pages and ranking the web pages
  - Practical applications to the social web and game data
- Illustrations of these concepts are given using R.
- Please indicate how many of you have used R before.

# Class overview

## Some Term Projects Examples

- Search engine:

### Popcorn DB

---

PopCorn DB <http://popcorn-db.net> is a personal project which aims at recreating **from scratch** an IMDB like website with a machine learning layer on top of it. **Therefore it includes the following features:**

#### **A fast & scalable web crawler.**

*I used Apache Spark for parallel computing, and InfluxDB for logging in live its activity. To reuse the CPU idle time when waiting for network responses I configured Spark to create 8 times more executors than CPU cores for each machine of the cluster.*

#### **A blazingly fast custom built search-engine with fuzzy search and autocompletion.**

*The average query time for 100K movies is 0.03ms. The speed is obtained by indexing every possible ngram of each movie title. The fuzzy search is done by building & exploring Levenstein automata on the go.*

#### **A movie genre & nationality predictor**

*I used a naive bayes network approach as it seemed after experimentation to be the best Machine Learning model adapted to this case.*

#### **A web-server, socket-server and front-end**

*The search engine and machine learning layers are written in C++. So I decided to build a web-server also in my C++ program. No need of Apache or nginx, less overhead = more speed.*

# Class overview

## Popcorn DB

PopCorn DB <http://popcorn-db.net> is a personal project which aims at recreating **from scratch** an IMDB like website with a machine learning layer on top of it. Therefore it includes the following features:

**A fast & scalable web crawler.**

*I used Apache Spark for parallel computing, and InfluxDB for logging in live its activity. To reuse the CPU idle time when waiting for network responses I configured :*

## Tags analytics

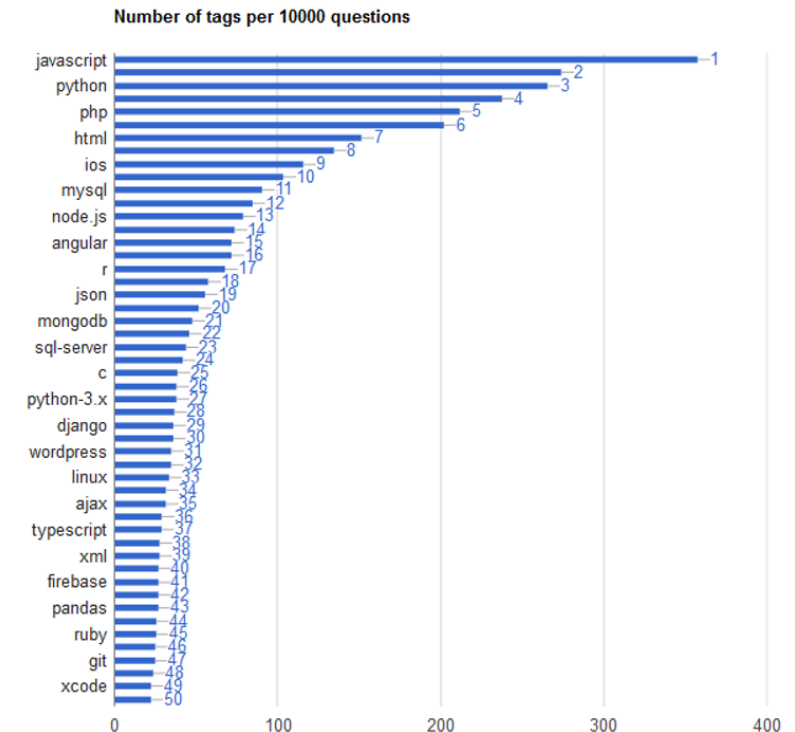
**A blazingly fast custom built**

*The average query time for 100K movies  
The fuzzy search is done by building*

TOTAL: 36534

**A movie genre & nationality**

*I used a naive bayes network approach*



## Some Term Projects Examples

- Data mining Stackoverflow
- Places Analysis : Mining
  - Analysis on the basis of the place given by user
  - Shows the data related to the images on that location
  - Analyze the Images to get the attributes like age, gender, smile
  - Sentiment Analysis on Comments from Instagram
  - Visualization of data on Dashboard

## Instagram API Call

localhost:63342/Places\_Analysis/index.html?\_ijt=62hm3e9jp3ivjkquodu9svf7m0

**Enter any Tourist Place**

Instagram Token: 3968699125.e029fea.1c080b992e314386a61be54a1b7a2555

Instagram API Count: 5

Tourist Place for Analysis: Theater District, New York, NY, USA

Submit Process Post

localhost:63342 says:  
Instagram API successfully gave the Images!

OK

# Please Introduce Yourself

- Introduce yourself to me and the other students
- Tell us about your background, your interests, hobbies, etc., so that we can get to know each other better.
- Please describe two or three objectives you hope to accomplish by the end of the course, e.g.
  - How does this course fit into your academic and professional objectives;
  - What do you hope to gain from the course.
- Please describe the type of data you work with and what pattern you typically look for in it.



# Introduction

- Most of the information we use today is stored online. There are claims that the data generated over the last 2 years is few fold larger than the data generated previously in the history of mankind.
- Most of this newly generated data is text, images and videos in a form of email, Google, YouTube, Facebook, Twitter, blogs, and most of the other technologies that define our digital age. To this we should also add the new communication tools such as social networks, instant messaging, Yammer, Twitter, Facebook, LinkedIn etc. too.
- By some general estimates a **third** of our time is spent on searching for information and another **quarter** analyzing it. It is widely believed that more and more data will be generated in the near future (IoT) and the time managing this data must be as productive as possible.
- This is just one aspect of what this course is about! We have an exciting journey ahead as we acquire the skills regarding this subject step-by-step.

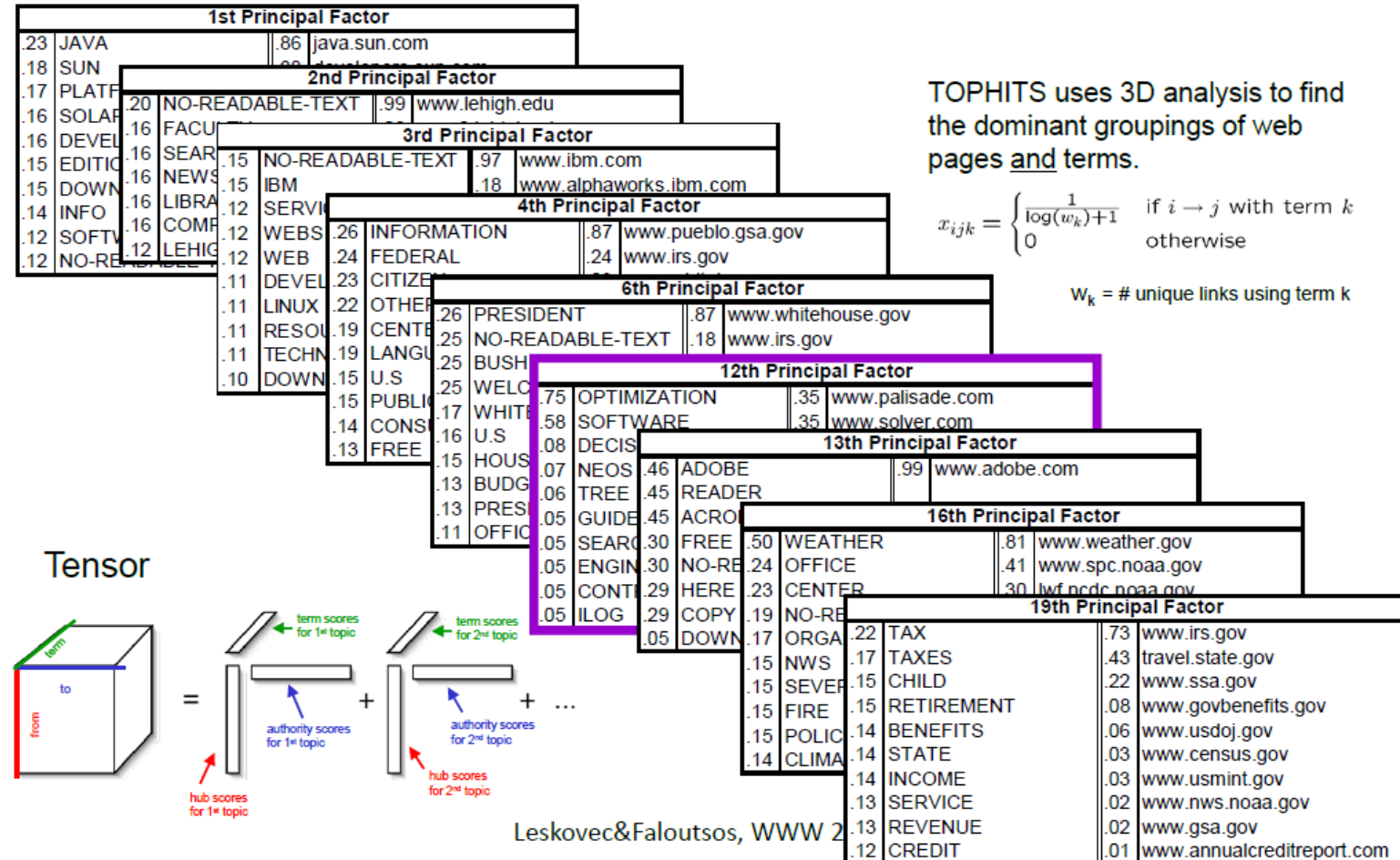
# Examples of Web analytics use

Web analytics is commonly used to give:

- Real-time visibility into web site performance,
- Order status,
- Inventory levels,
- Warehouse management systems.

# Data Analytics

- Real data are often in high dimensions with multiple aspects (modes)
- Matrices and tensors provide elegant theory and algorithms



# Singular Value Decomposition (SVD)

- [https://en.wikipedia.org/wiki/Latent\\_semantic\\_analysis](https://en.wikipedia.org/wiki/Latent_semantic_analysis)

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

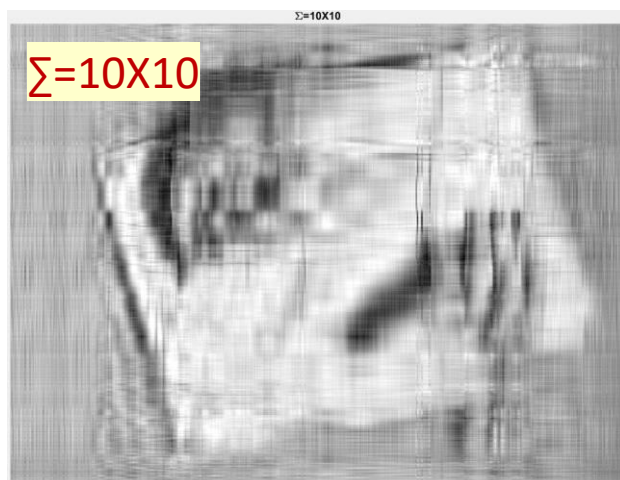
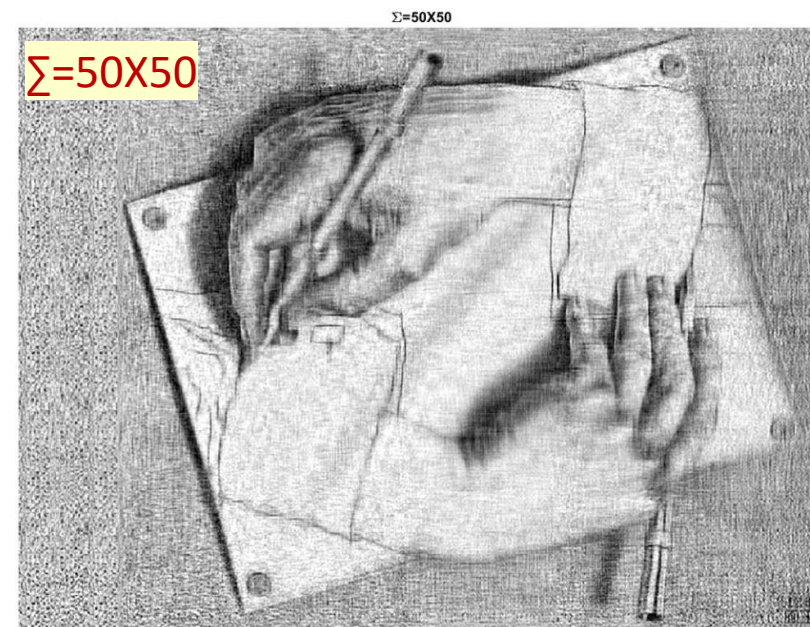
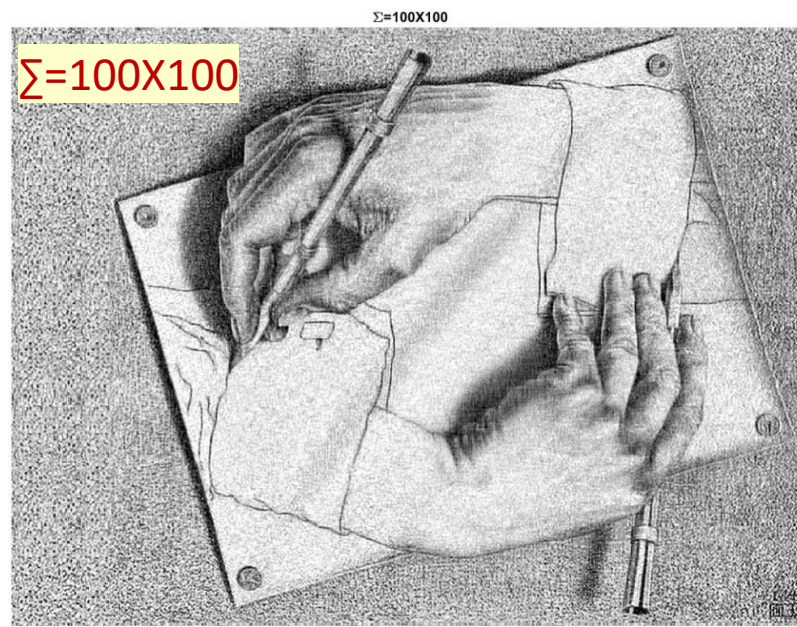
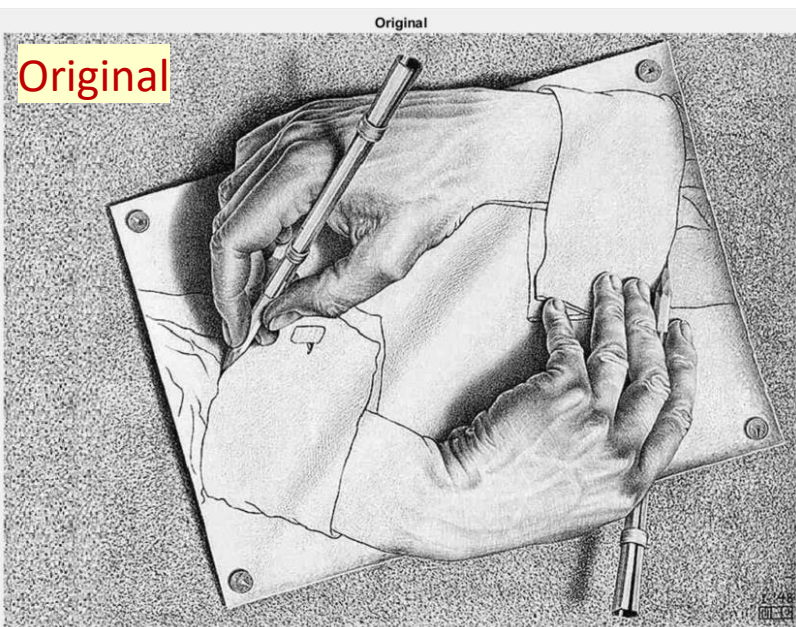
The diagram illustrates the SVD equation  $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$  with visual representations of the matrices:

- Matrix  $\mathbf{X}$  (input data):** Represented as a collection of vertical bars (columns) labeled  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(M)}$ . A bracket below it is labeled "input data".
- Matrix  $\mathbf{U}$  (left singular vectors):** Represented as a collection of vertical bars (columns) labeled  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$ . A bracket below it is labeled "left singular vectors".
- Matrix  $\mathbf{\Sigma}$  (singular values):** Represented as a diagonal matrix with elements  $\sigma_1, \sigma_2, \dots, \sigma_k$  on the diagonal. A bracket below it is labeled "singular values".
- Matrix  $\mathbf{V}^T$  (right singular vectors):** Represented as a collection of horizontal bars (rows) labeled  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ . A bracket below it is labeled "right singular vectors".

The equation is shown as  $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ , with the matrices  $\mathbf{X}$ ,  $\mathbf{U}$ ,  $\mathbf{\Sigma}$ , and  $\mathbf{V}^T$  arranged from left to right, separated by multiplication dots.

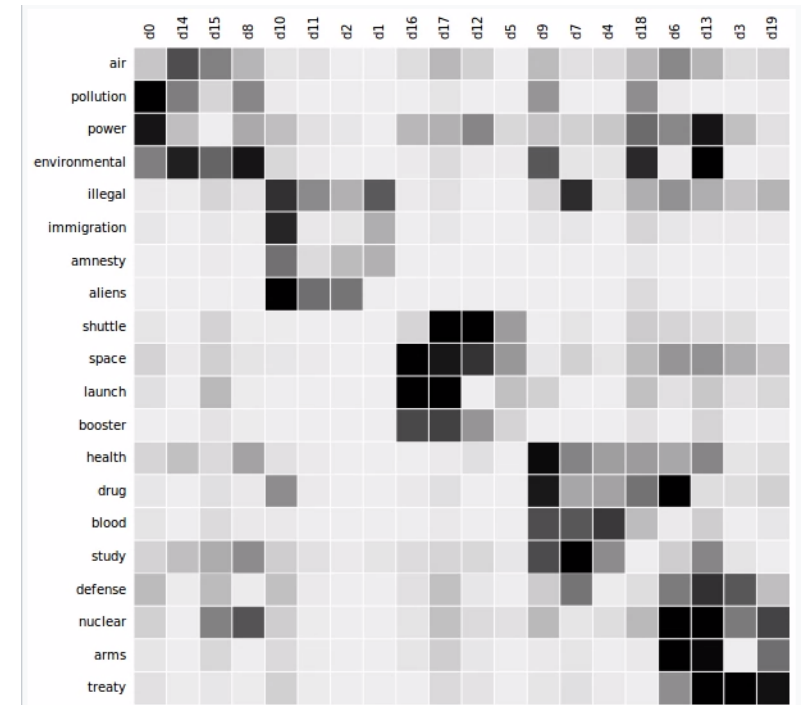
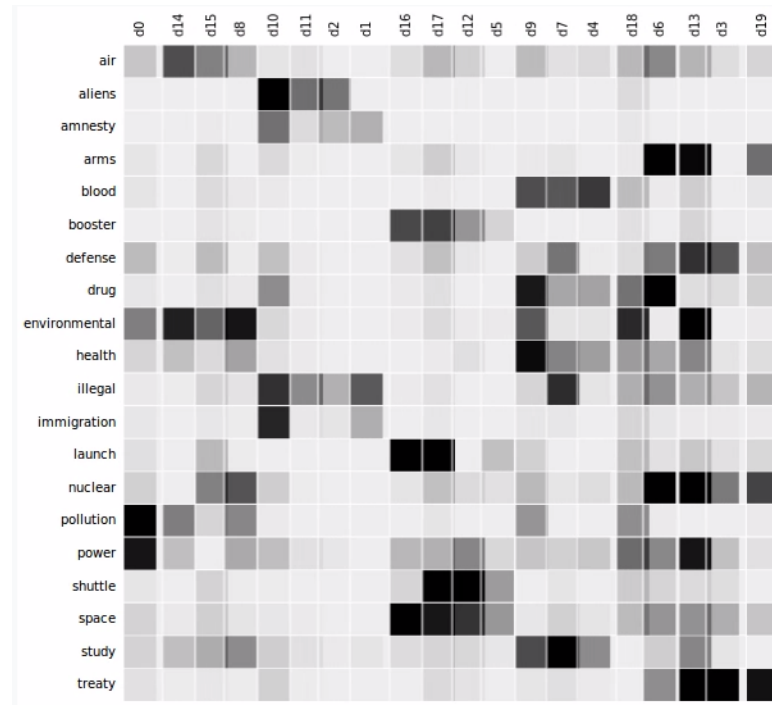
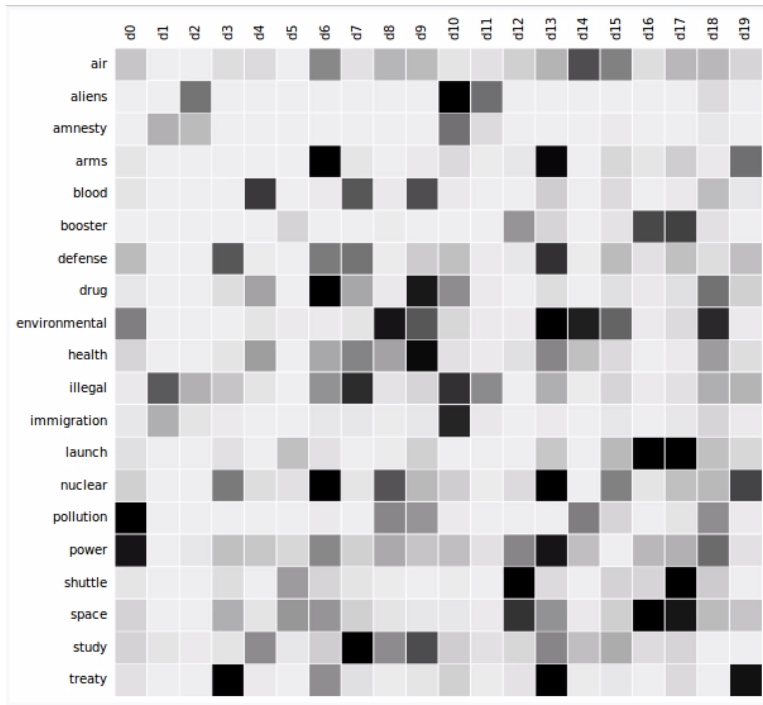


# Singular Value Decomposition (SVD)

[illegible]

# SVD and Text Mining

- [https://en.wikipedia.org/wiki/Latent\\_semantic\\_analysis](https://en.wikipedia.org/wiki/Latent_semantic_analysis)



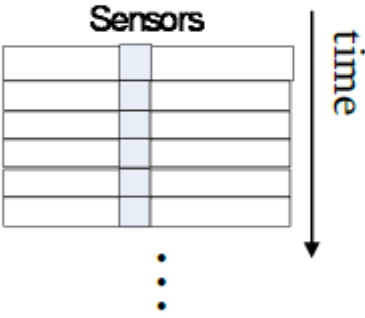
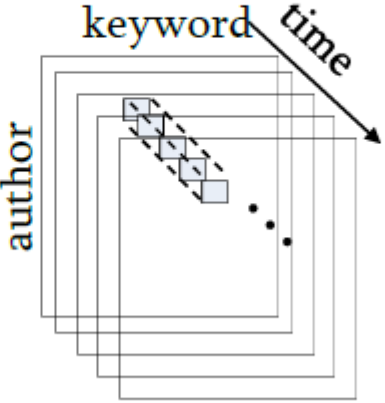
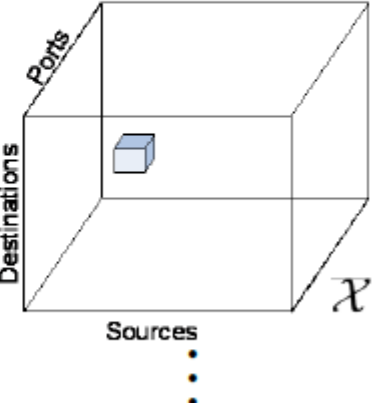
$A$  is document term  $[m \times n]$  matrix (m rows, n columns).

$A^T A$  is term to term similarity  $[m \times m]$  matrix.

$A A^T$  is document to document similarity  $[n \times n]$  matrix.

# Time Series Data

- Dynamical (streaming) data such as in IoT
  - A sequence of  $M^{\text{th}}$  order tensors.

Order	1st	2nd	3rd
Correspondence	Multiple streams	Time evolving graphs	3D arrays
Example			

# Introduction: Working with data

- Lots of digital data generated over the last few years with expectations that more and more data will be generated in the near future.
- Questions arise how to optimally manage it in less time.
- Typically at some point all this unstructured data is reduced to digital text as most practical form.
- **Analytics** - discovery of relevant patterns in data.
- **Mining** - extracting useful information from data.
- Either way the goal is to learn from data.
- In the language of **machine learning** (to learn without explicitly being programmed) these learning categorizations can be subdivided as
  - **Supervised learning** (we know the categories into which we need to separate the data). Task: for new feature  $x$ , predict  $y$ )
    - Regression (continuous  $y$  data predictions from features  $x$ )
    - Classification (discrete  $y$  data predictions from features  $x$ )
  - **Unsupervised learning** (we don't have any knowledge into what categories the data can be subdivided ) - find structure in large data set.
    - Clustering



# What is an analytic?

- Methodology of revealing a meaningful pattern in recorded information (data) to quantify a performance.
- Relies on the simultaneous application of several steps and techniques including research and data storage managed by a computer system.
  - Database
    - Typically used for large, long-lived data.
  - The knowledge base data storage (called an ontology) is
    - A dynamic resource
    - An object model with classes, subclasses, and instances.
    - Benefits - being able to store, analyze, and reuse knowledge
    - Typically used to arrive at a specific answer to a problem.
- Use of computer systems to implement
  - Statistics techniques
  - Including machine learning
- Analytics is used to drive decision making
  - By identifying which data is useful and meaningful

# Components of Analytics

- **Descriptive** analytics – describe what is happening in your system.
  - Simply describes past events and can allow for interpretation in preventing future negative impacts.
- **Predictive** analytics – predict what could happen.
  - Utilizes a variety of statistical, modeling, data mining and machine learning techniques to study historical and recent data.
  - Allows analysts to make predictions about future (positive or negative) events.
  - Currently being able to foresee positive and negative events is extremely powerful feature used to derive marginal advantages over competitors, if not to gain competitive advantage.
  - Predictive analytics takes into account all historical data, allowing for linking of key data points over time to provide predictive features related to operational cost effectiveness and possible downtrends.
- All this is used to prescribe solutions to help mitigate issues along the way.
- This new and growing area of predictive analytics gives the probability of an event and gives the data points needed to mitigate it.
- The combination of analytics and predictive analytics (coding) offers more advanced and cost-effective operation.
- In addition, machine based predictive analytics capabilities can help find the meaning and subset useful data based on input from the user.
- An intuitive predictive coding workflow can validate performance and allow weighing the costs and benefits based on specific measure thresholds (such as precision and recall).
- It can recommend one or more courses of action – and show the likely outcome of each scenario.