Data Visualization and Social Sciences

Dae-Jin Lee
BCAM Researcher

First Workshop on Interactions between Mathematics and Social Sciences

September 26, 2014

Outline

Motivation: no more boring data

What is Data Visualization?

Some examples

Final comments

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

Motivation

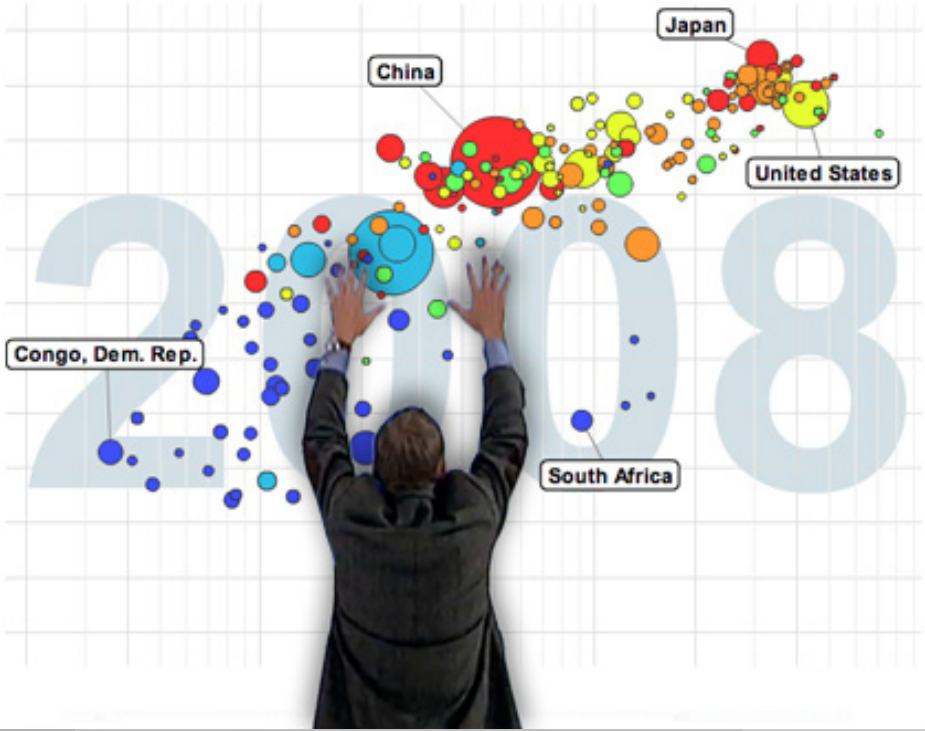
No more boring data

- In 2006 Hans Rosling (Prof. of Public Health Science at Karolinska Institute) gave an inspiring talk at **TED** about social and economic developments in the world over the last 50 years, which challenged the views and perceptions of many listeners.

Motivation

No more boring data

- In 2006 Hans Rosling (Prof. of Public Health Science at Karolinska Institute) gave an inspiring talk at **TED** about social and economic developments in the world over the last 50 years, which challenged the views and perceptions of many listeners.
- **"Stats that reshape your world-view"**
  http://www.youtube.com/watch?v=hVimVzgtD6w

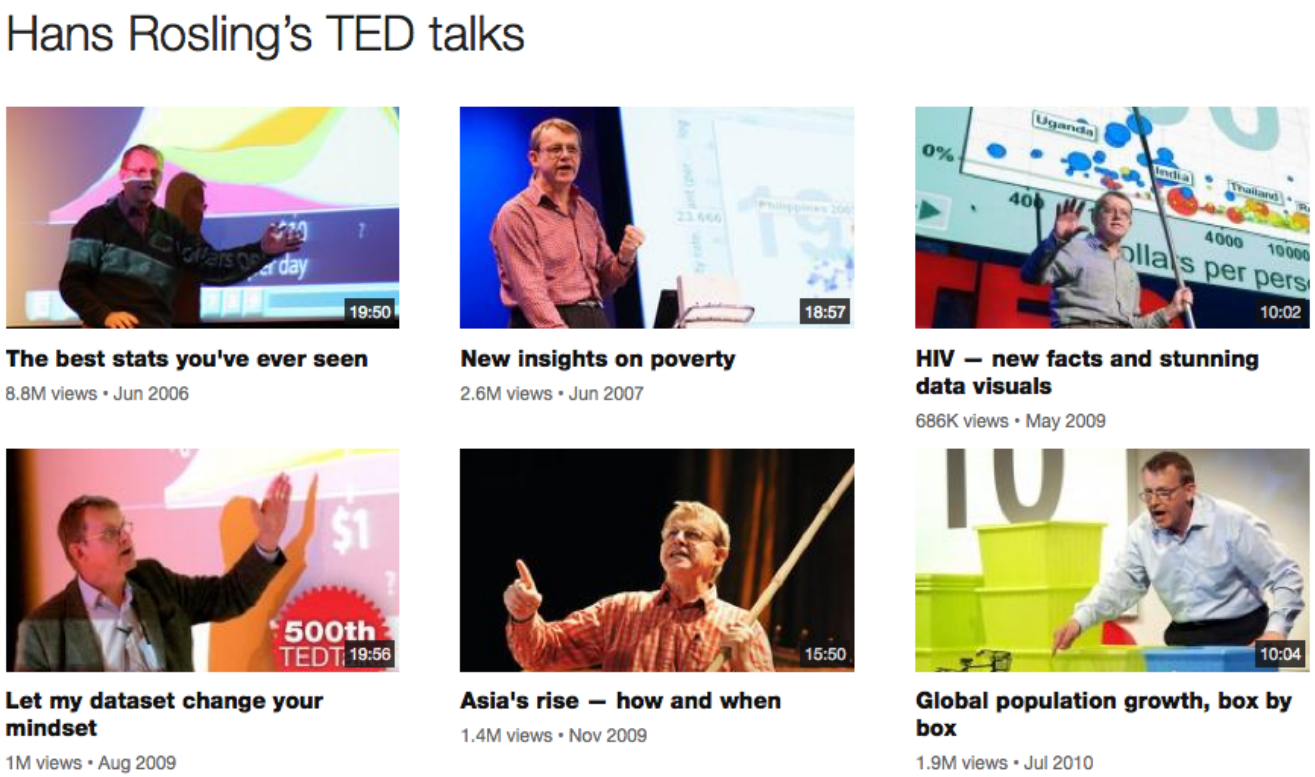


TED
IDEASWORTHSPREADING

Motivation

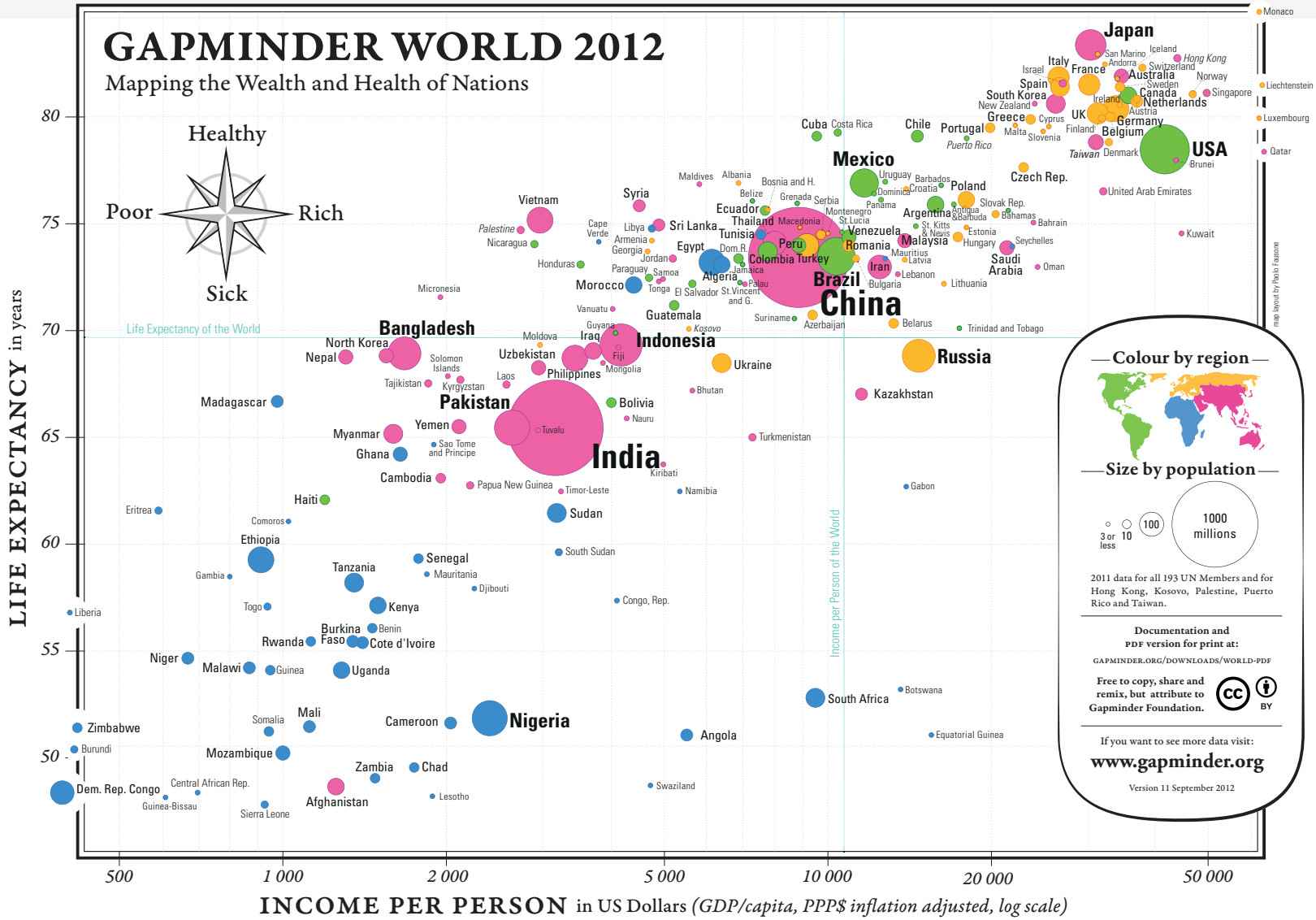
Professor Hans Rosling's TED Talk (2006)

- To visualise his talk, he and his team at `gapminder.org` used a public domain software
- **Gapminder** software was purchased by **Google** and integrated as motion chart into their **Visualization API**



Motivation

Gapminder



Data Science

The Age of Data

- **Data science** is the study of the generalizable extraction of **knowledge from data**
- Includes elements from many fields:
  - signal processing, mathematics, probability models, machine learning, computer programming, data engineering, high performance computing and **visualization**
- **Data science techniques** affect research in many domains, including the biological sciences, medical informatics, health care, social sciences/politics and the humanities. It heavily influences economics, business and finance ("business analytics").

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

Data Visualization (DataVis)

What is it?

- **Data visualization** is the use of tools to represent data in the form of charts, maps, plots, animations, or any graphical means that make content easier to understand.
- Instead of presenting information in **numerical tables**, it is a graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents.
- Graphic representations of data are popular because:
  - they open up the way we think about complex data,
  - reveal hidden patterns/structures in the data, and
  - highlight connections among elements.
  - check assumptions or support hypotheses from the data

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

Characteristics of Data

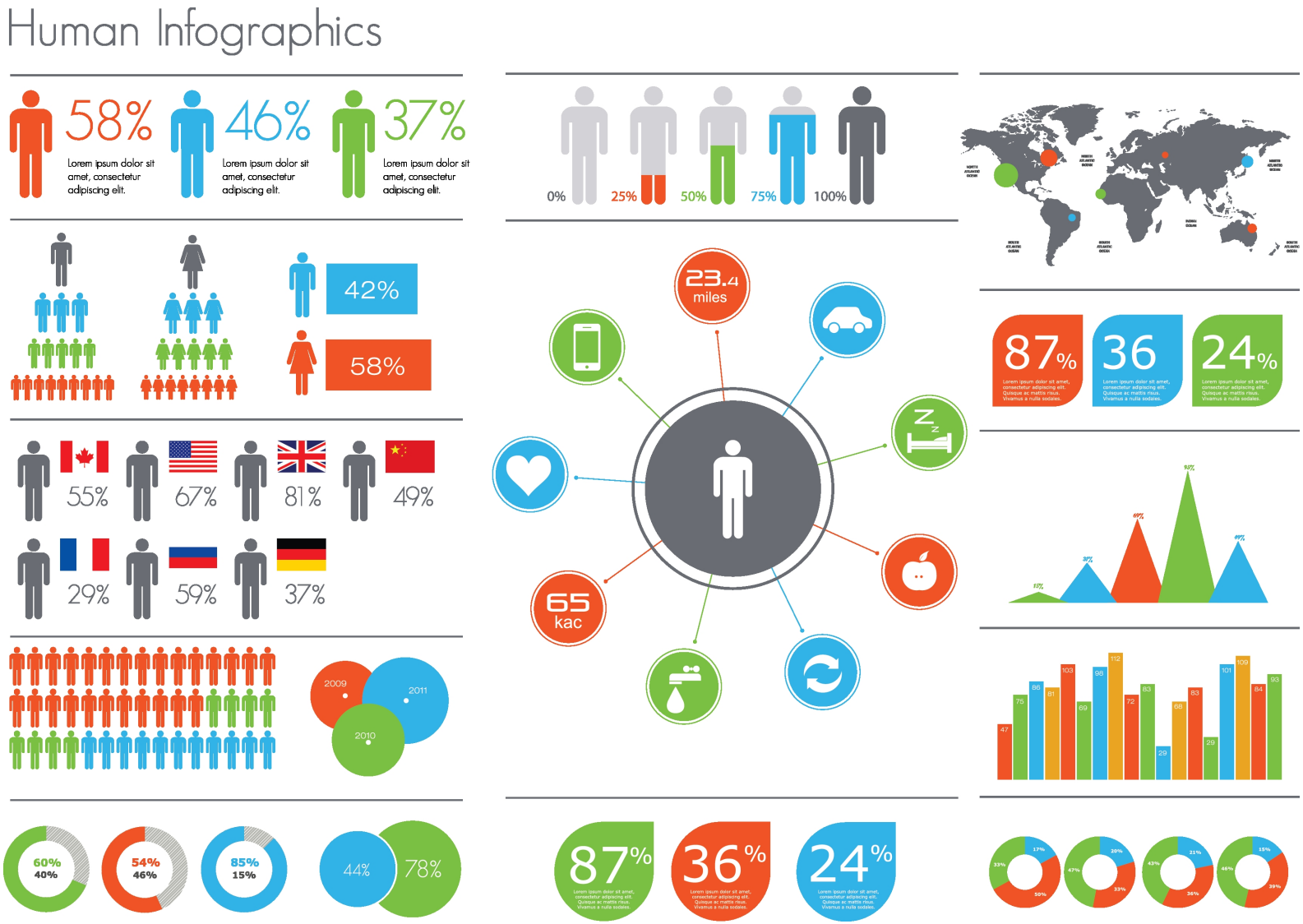
- Numeric, symbolic (or mix)
- Scalar, vector, or complex structure
- Various units
- Discrete or continuous
- Spatial, quantity, category, temporal, relational, structural
- Dense or sparse
- Ordered or non-ordered
- Disjoint or overlapping
- Binary, enumerated, multilevel
- Independent or dependent
- Multidimensional
- ...

Data Visualization (DataVis)

Descriptive Graphs

Data Visualization (DataVis)

Infographics



Data Visualization (DataVis)

Why ?

- **Why visualize social science research?**
  - Social Science research seeks to understand, explain, and predict human behavior by observing, reflecting, and/or measuring social phenomena.
  - Social Sciences are gradually becoming a more quantitative scientific field

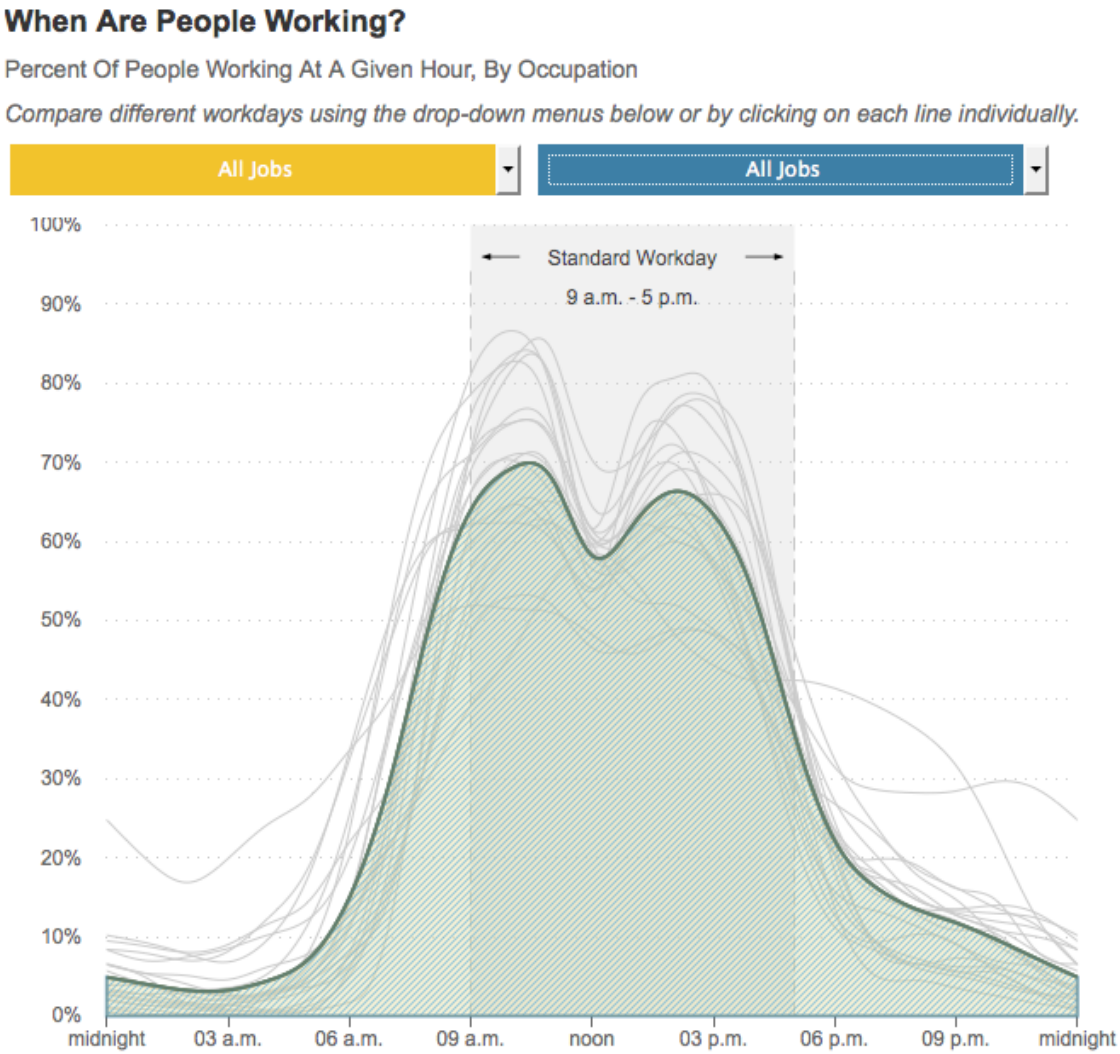
Data Visualization (DataVis)

Why ?

- **Why visualize social science research?**
  - Social Science research seeks to understand, explain, and predict human behavior by observing, reflecting, and/or measuring social phenomena.
  - Social Sciences are gradually becoming a more quantitative scientific field
  - Indeed, most of the top influencial news agencies and Public Institutions (The Guardian, New York Times, Washington Post, OECD,...) have their own DataVis Sections
    http://datajournalismhandbook.org

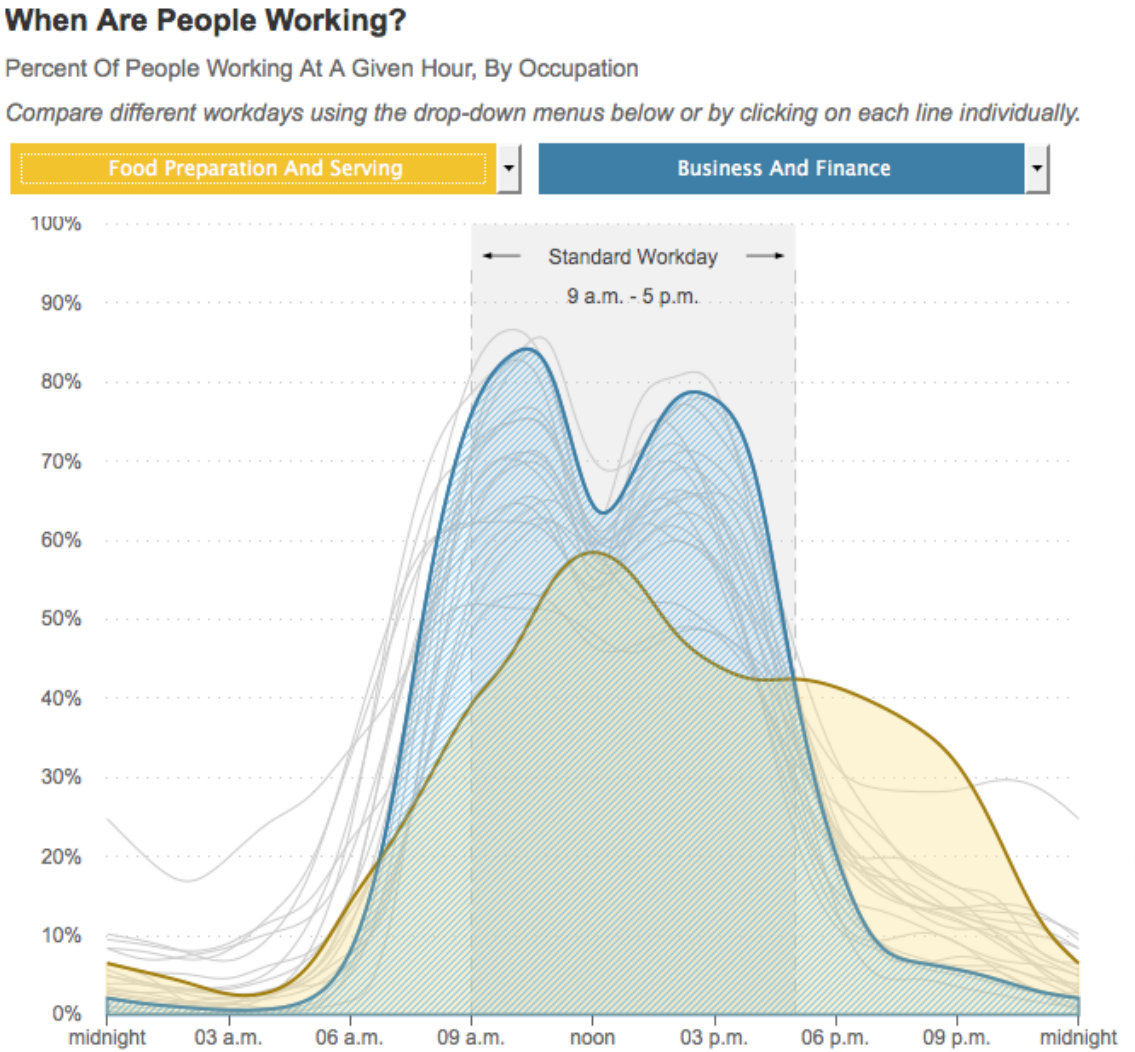
Some examples

Who's In The Office? The American Workday In One Graph (www.npr.org)



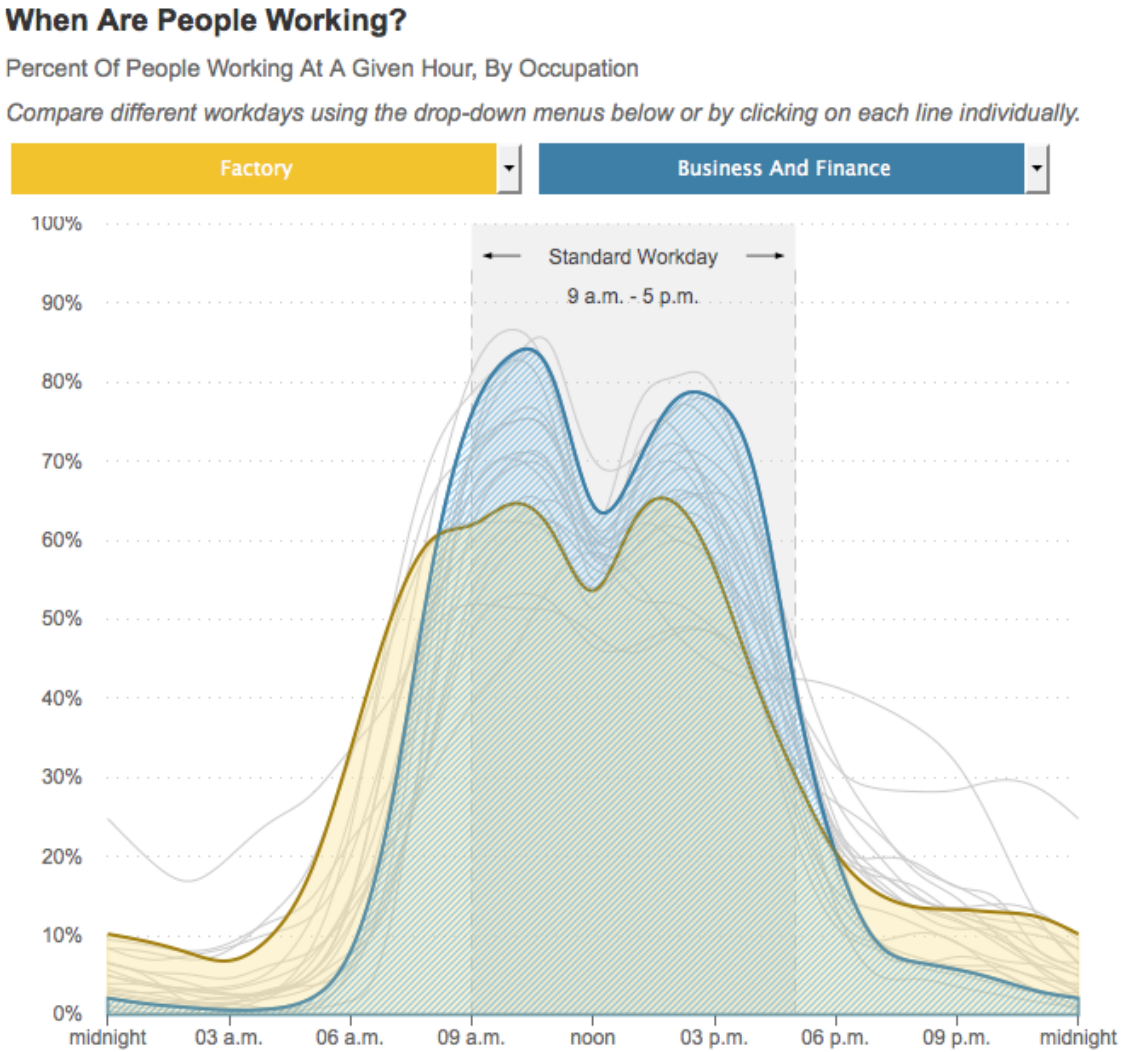
Some examples

Who's In The Office? The American Workday In One Graph (www.npr.org)



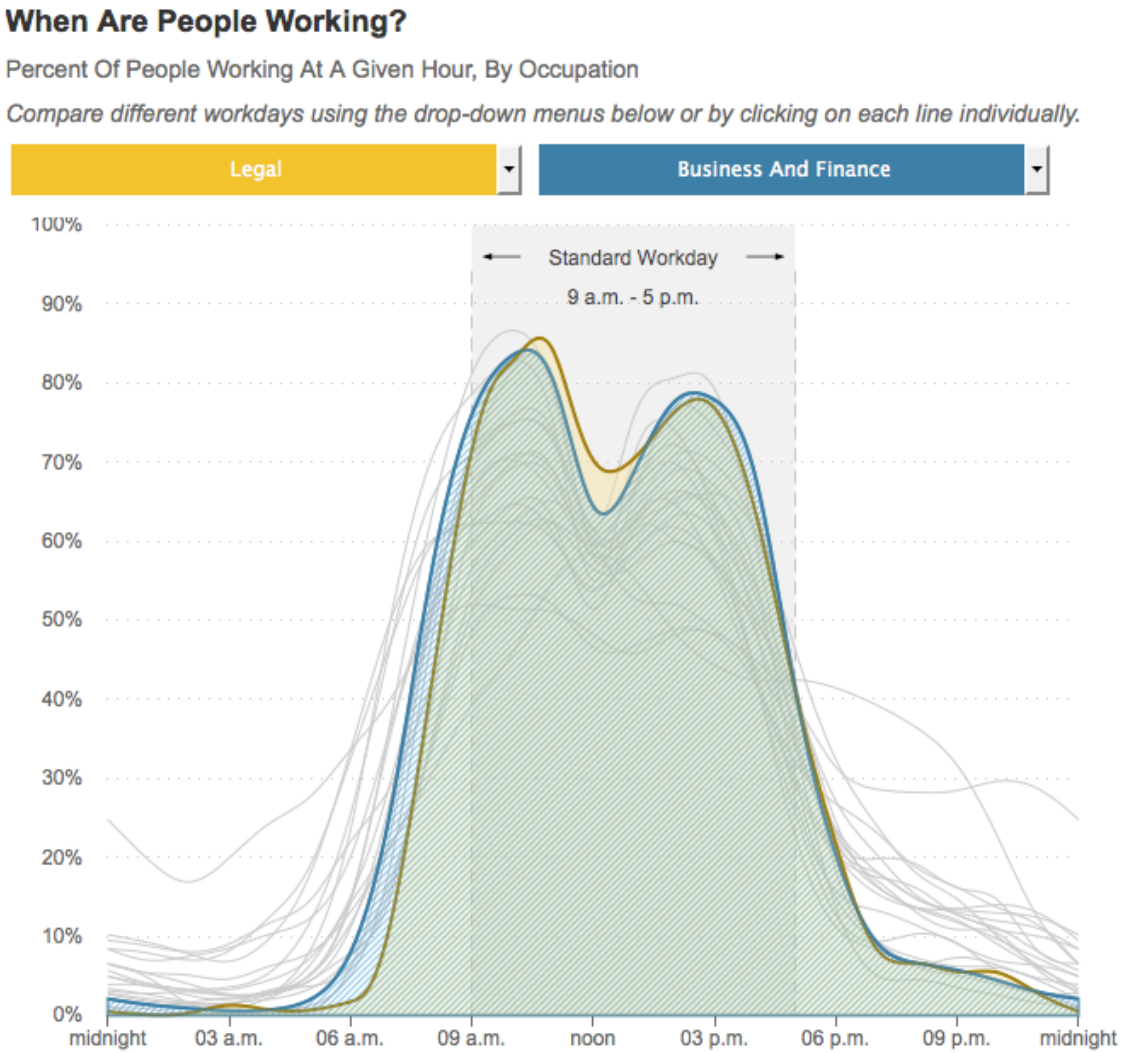
Some examples

Who's In The Office? The American Workday In One Graph (www.npr.org)



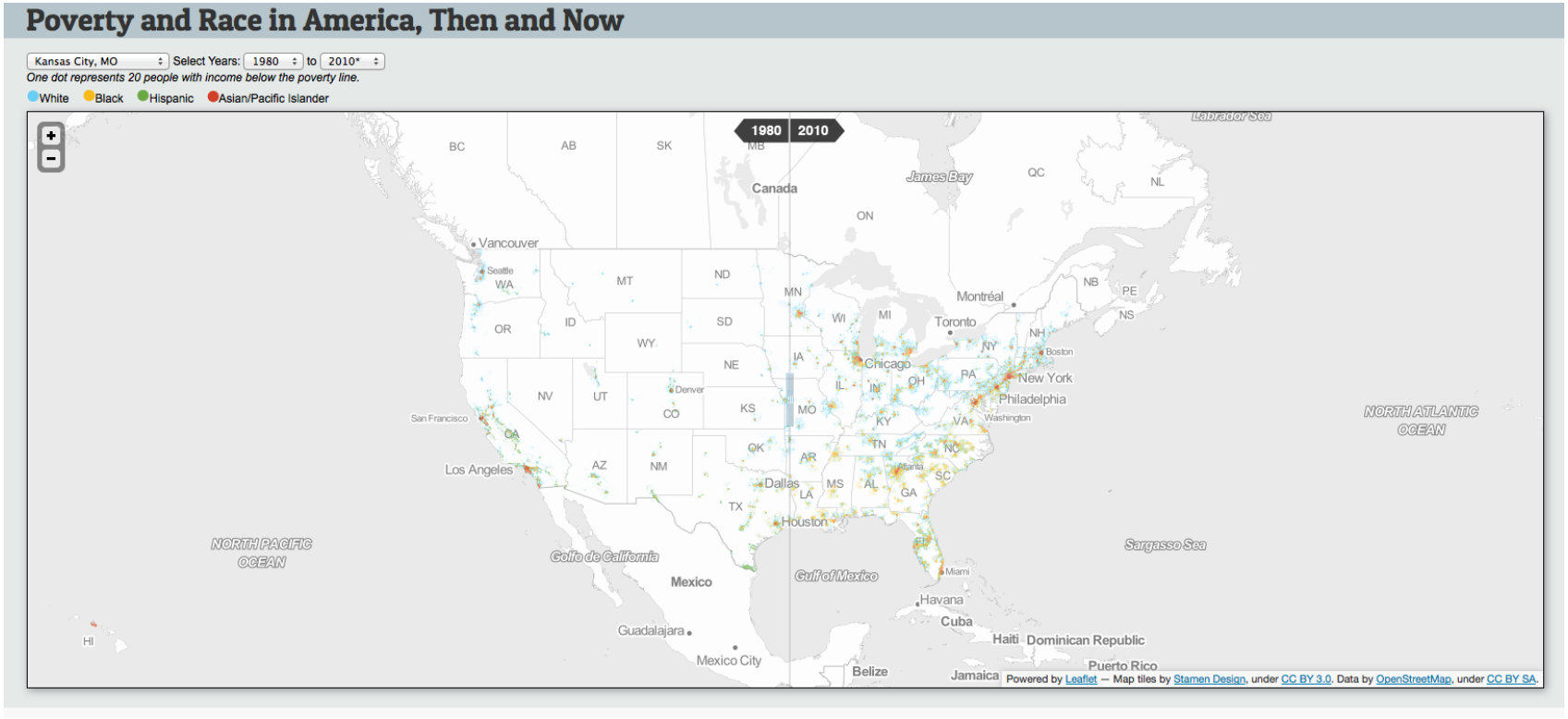
Some examples

Who's In The Office? The American Workday In One Graph (www.npr.org)



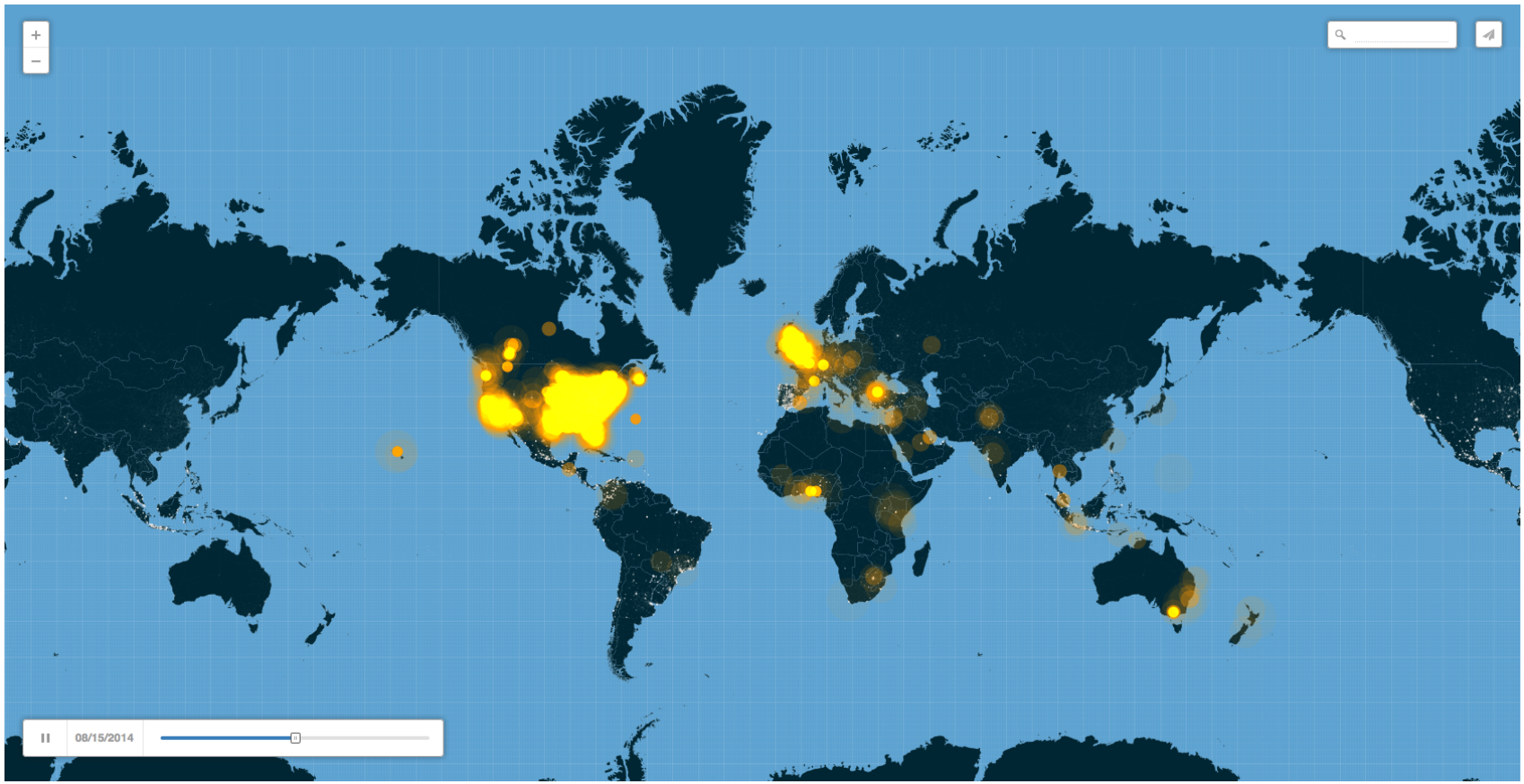
Some examples

Metro trends: Poverty and Race in America, Then and Now



Some examples

cartoDB: How news of #Ferguson spread across Twitter



Data Visualization (DataVis)

Social Networks Analysis

- Visual representations of social networks are important to understand network data
- It facilitates qualitative interpretation of network data
- **Network maps** are widely used for visualizing social network structures, trading relations between countries, migration rates, groups, individuals etc ...
- Help to study patterns of relationships that connect members of a social system at all scales.
- It is a powerful method for conveying complex information

Data Visualization (DataVis)

Social Networks Analysis

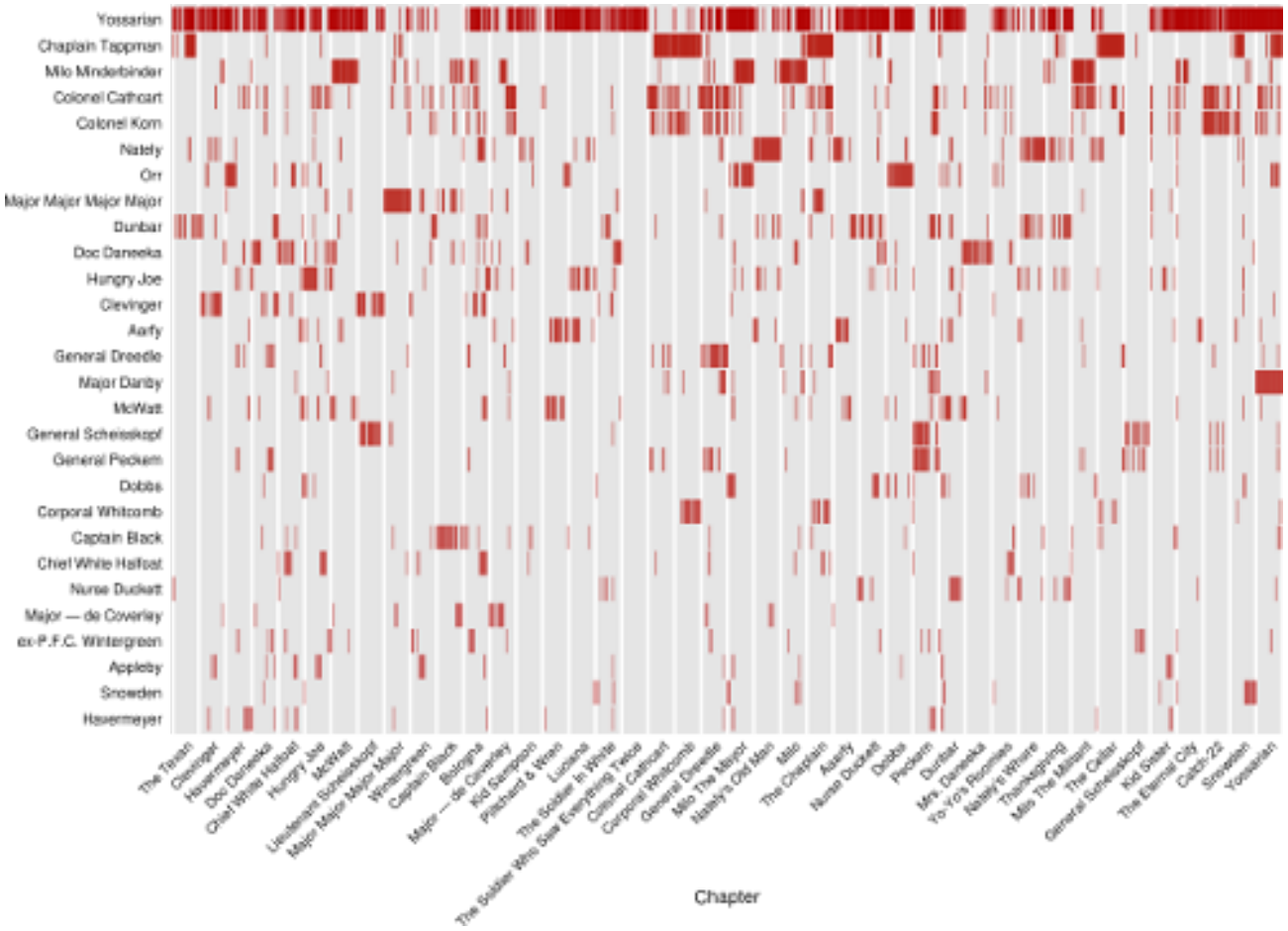
- Visual representations of social networks are important to understand network data
- It facilitates qualitative interpretation of network data
- **Network maps** are widely used for visualizing social network structures, trading relations between countries, migration rates, groups, individuals etc ...
- Help to study patterns of relationships that connect members of a social system at all scales.
- It is a powerful method for conveying complex information
- **Example:** *Les Misérables*
  - Network representation of character co-occurrence in the chapters of Victor Hugo's classic novel.
  - The original data set is from **The Stanford GraphBase: A Platform for Combinatorial Computing** (by Donald Knuth)

    * Inspired by Jeff Clark's Novel Views: Les Miserables series

Social Networks Analysis

Example: Les Misérables

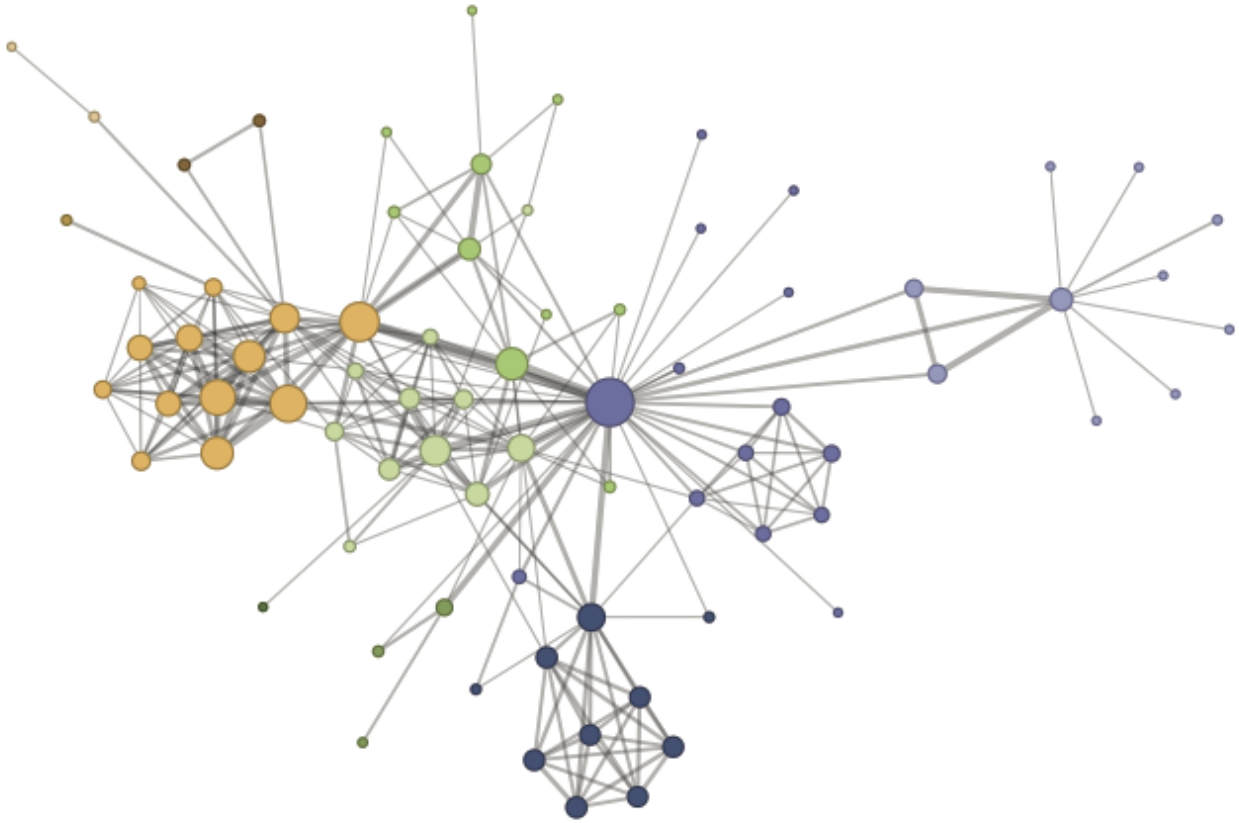
- This plot essentially represents a time series of when different characters were mentioned in the book.



Social Networks Analysis

Example: Les Misérables

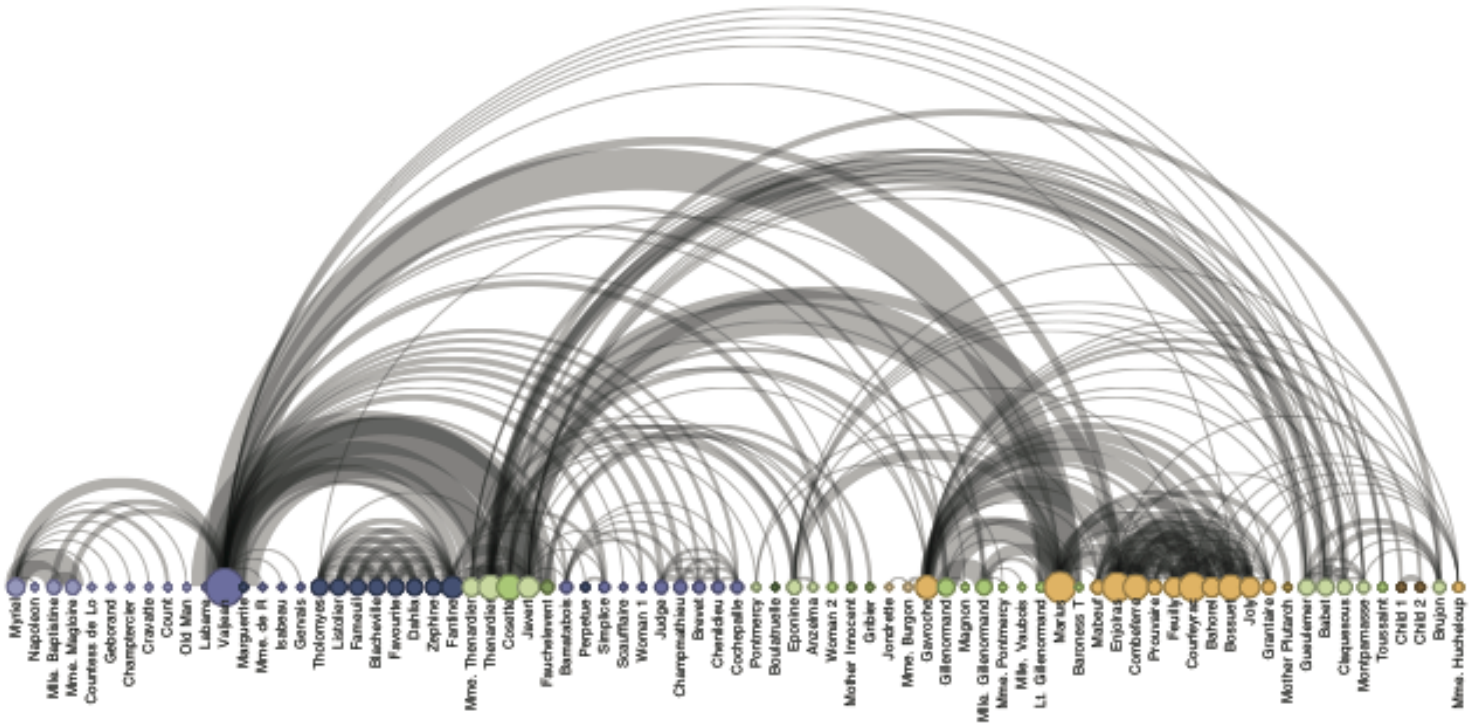
- Co-ocurrence of characters with node colors depicting **cluster memberships** computed by a **community-detection algorithm**



Social Networks Analysis

Example: Les Misérables

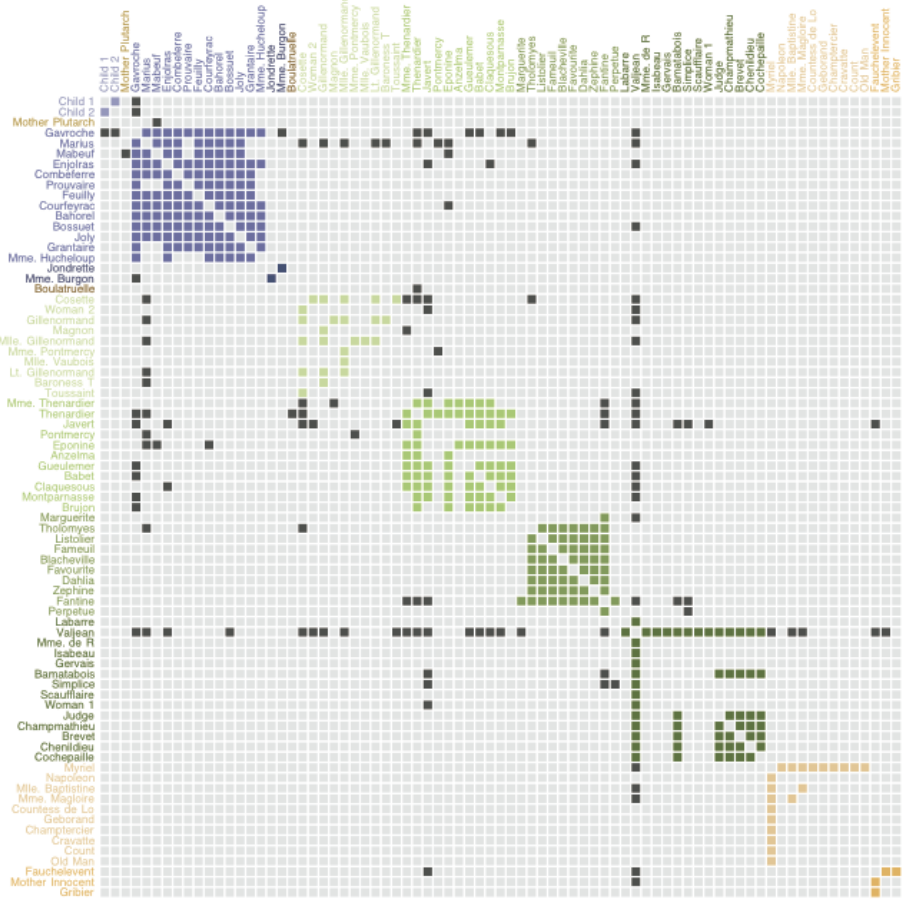
- Co-ocurrence of characters with **Arc-diagrams**



Social Networks Analysis

Example: Les Misérables

- **Adjacency matrix** where each value in row i and column j corresponds to the link from node i to node j



Social Networks Analysis

Practical applications

- Applications include:
  - Data aggregation and mining, network propagation modelling
  - Behavior analysis
  - In business it support activities such as customer interaction, marketing, and business intelligence
  - National Security: Determine crime networks and leaders within the network

Social Networks Analysis

Sentiment analysis in Twitter

- Twitter data has been used to predict and explain a variety of real-world phenomena, including opinion polls, elections, the spread of contagious diseases, and the stock market.
- Twitter messages in aggregate contain useful information that can be exploited with statistical methods.
- The hypothesis is that we can obtain high accuracy on classifying sentiment in Twitter messages using machine learning techniques.
- **Sentiment** is defined to be a personal positive or negative feeling
- There are different approaches. **E.g.:** Bayesian Models, Maximun Entropy Classifiers, ...
- and tools such as **Sentiment Viz** (NC State University, CS Department)

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

Search term:

mathematics

Emotions are represented as a psychology model by Russell (1980)

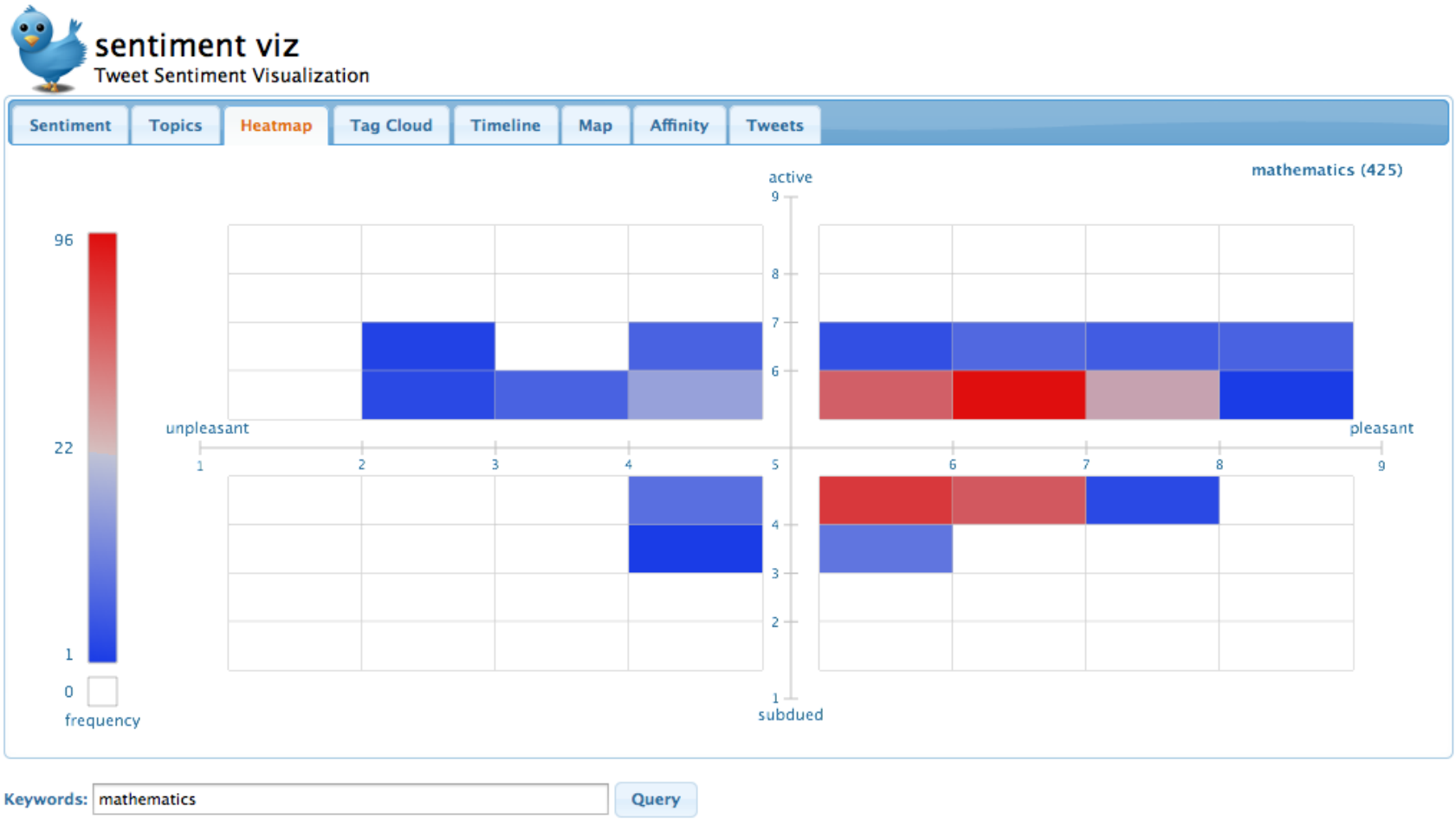
Search term:

mathematics

Clustering emotions

Search term:

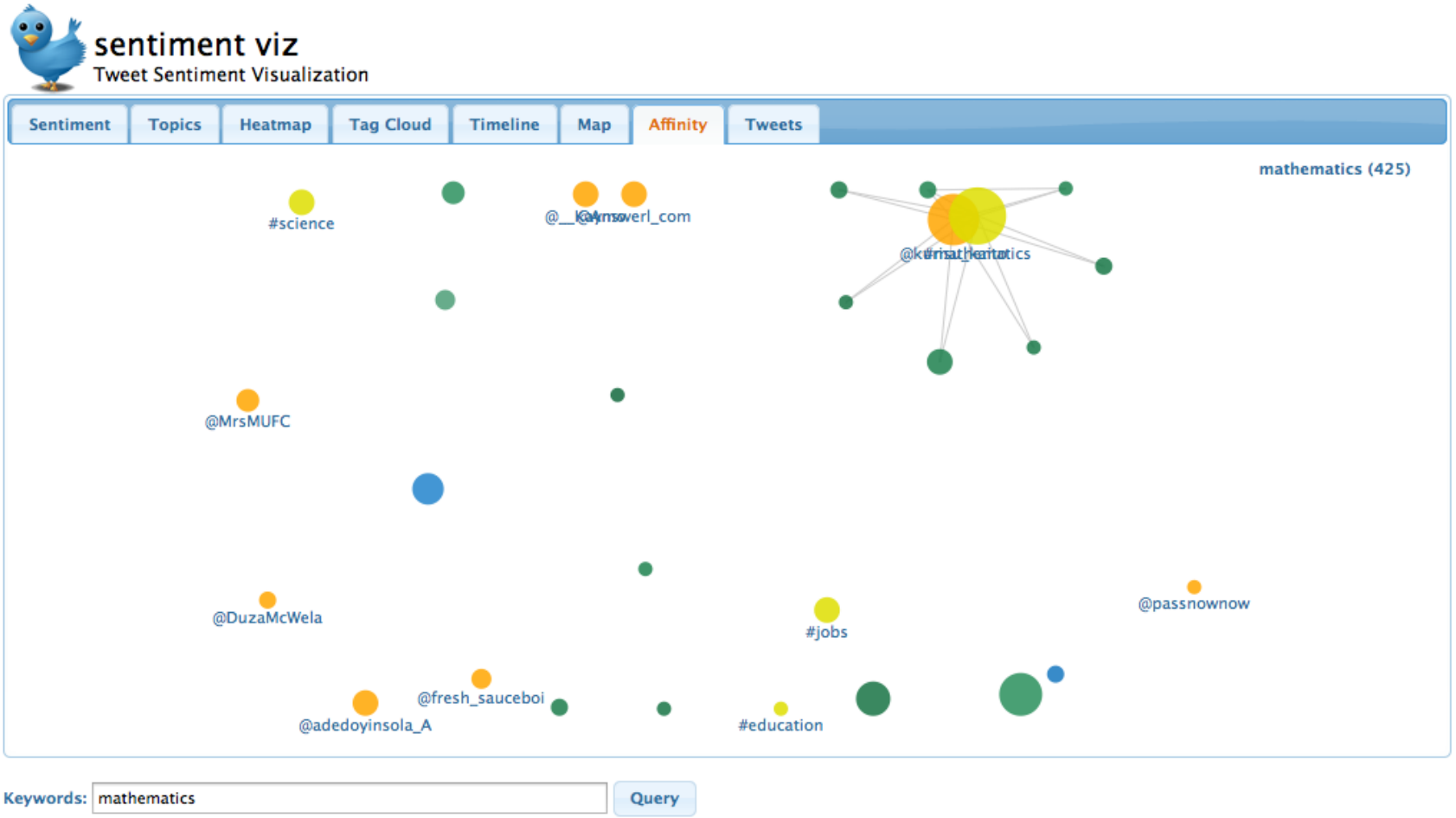
mathematics



Frequency of emotions on a 8×8 grid of emotion regions

Search term:

mathematics



Relathionships of affinities between Tweets

Politics in the US

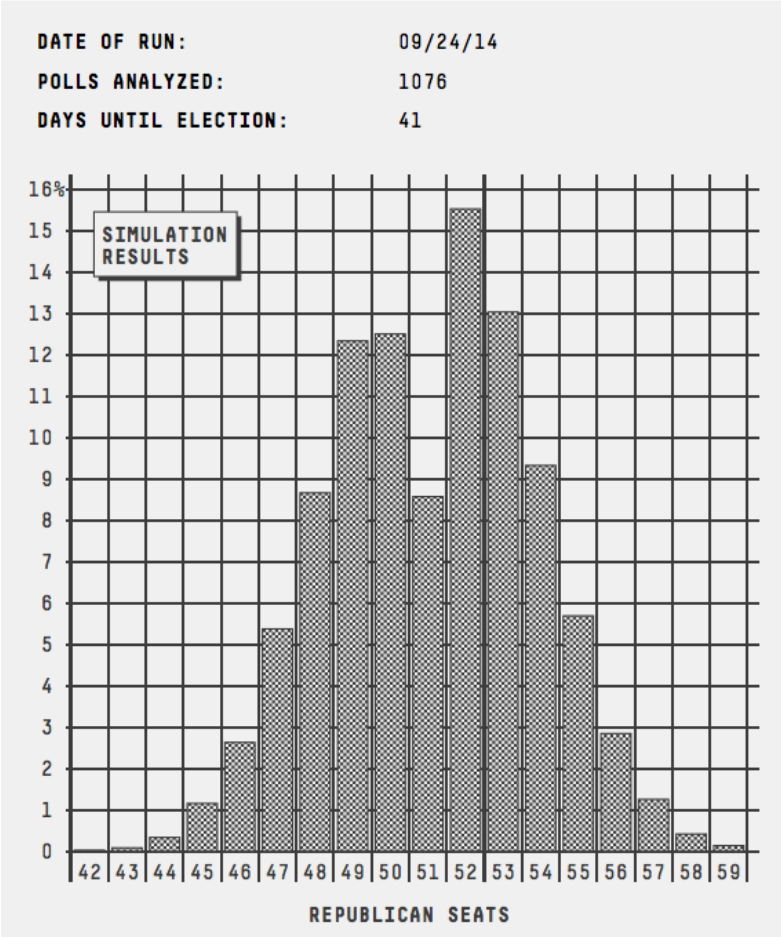
FiveThirtyEight election forecasting model

- **Nate Silver** (and american statistician) created www.FiveThirtyEight.com and predicted in 2008 presidential elections in the U.S. the winner of **49 of the 50 states**, the next year he correctly predicted the winner of **all 35** U.S. Senate races.
- **FiveThirtyEight's election forecasting model** combines hundreds of opinion polls with historical and demographic information to calculate odds for each Senate race.
- They estimate the probability that each party will win control of the Senate.
  - Based on a **probabilistic model** and **historical evidence** (empirical data)
  - ... and **simulating outcomes** to estimate probabilities
  - The forecast is updated regularly

Politics in the US

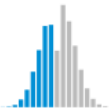
FiveThirtyEight election forecasting model

The Model's Output

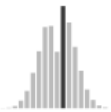


What It Means

Republicans have a **56.9%** chance of winning a majority.



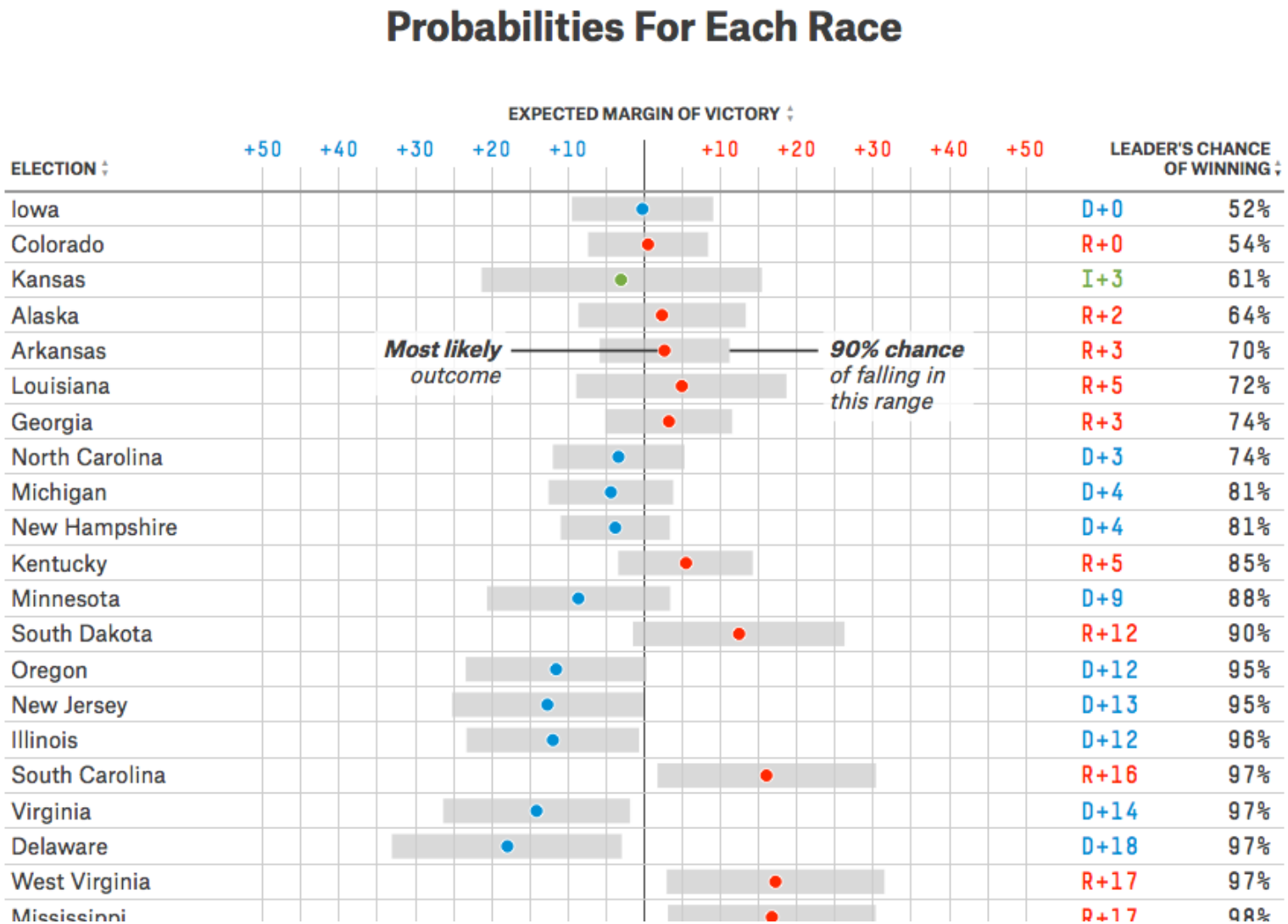
Democrats have a **43.1%** chance of keeping the majority.



There is a **15.5%** chance Republicans will control **52** seats and Democrats will control **48** seats.

Politics in the US

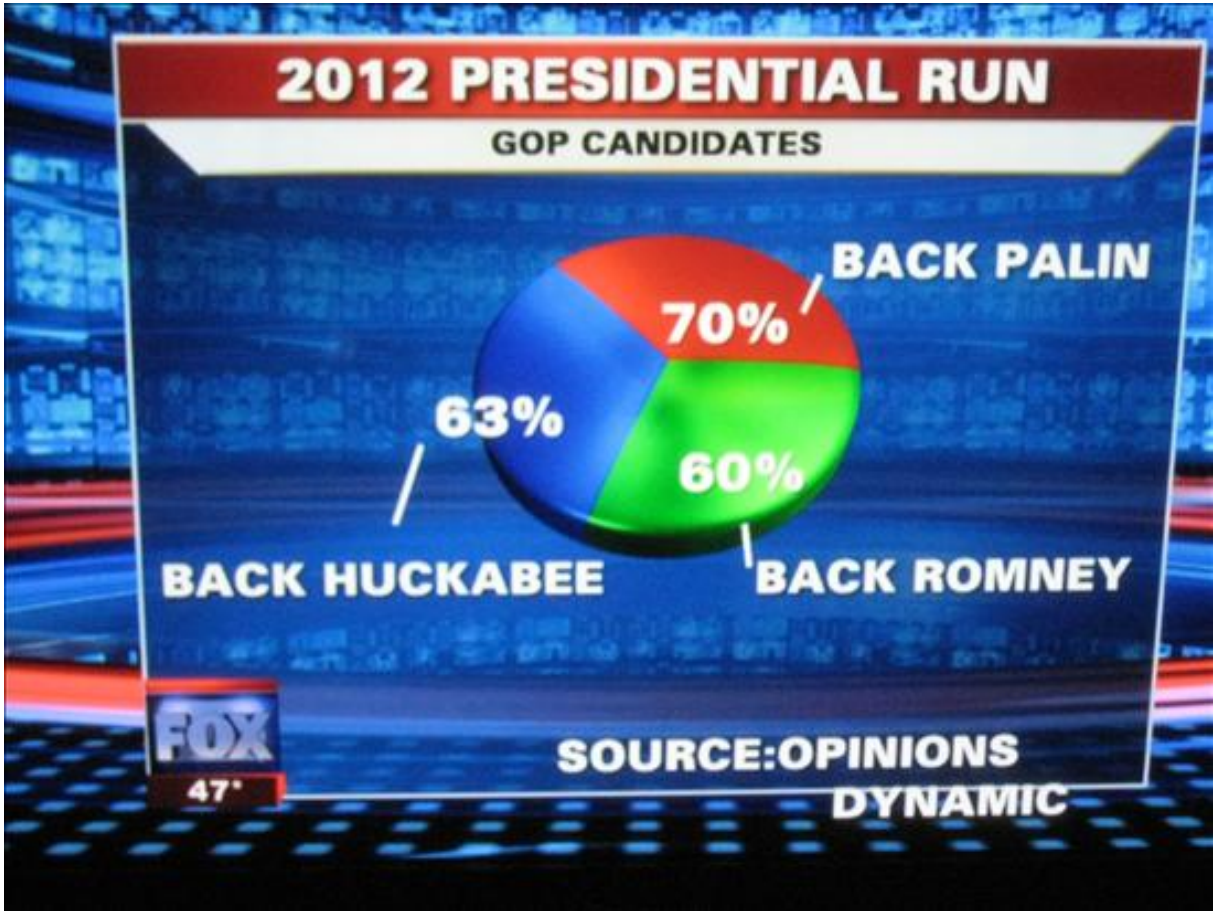
FiveThirtyEight election forecasting model



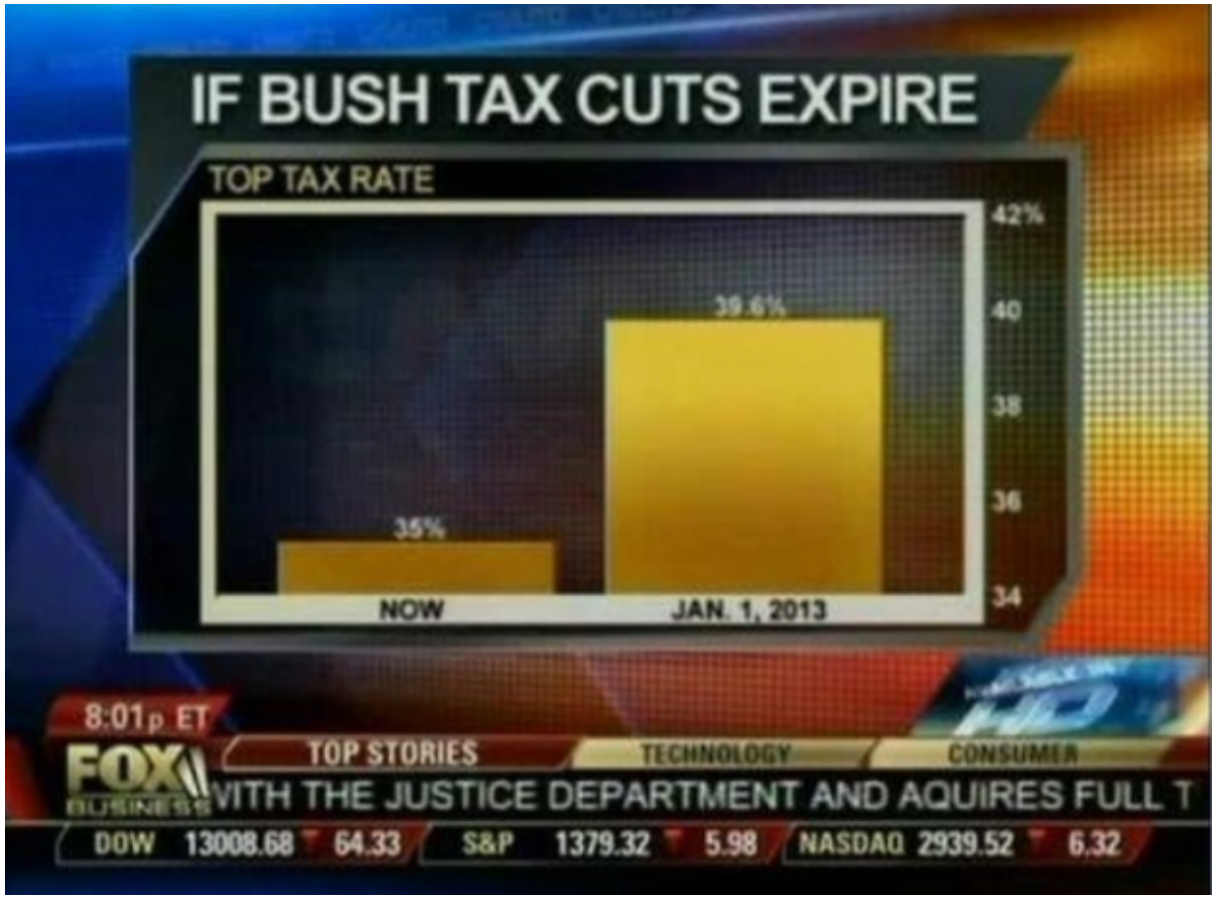
Final comments

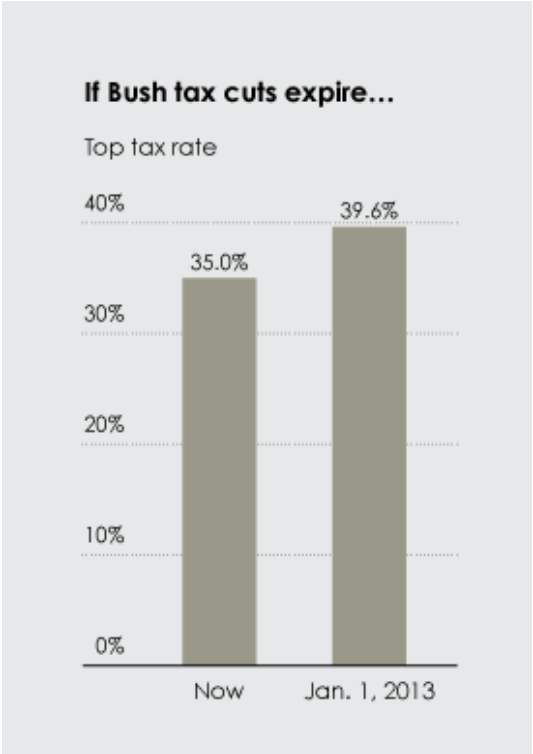
- More and more **Open Data** $\Rightarrow$ **Open Science** for publishing results of scientific activities available for analyze and reuse
- **DataVis** takes advantage of recent developments in
  - Computer Science and Computer Graphics,
  - Statistical methods,
  - Methods of information visualization, visual design
  - Social Networks (graph visualization, network maps, ...)
  - Geographic Information Systems (GIS)
- However, there is a risk related to possible misuse of graphical data displays (by mistakes or intentionally)

Final comments

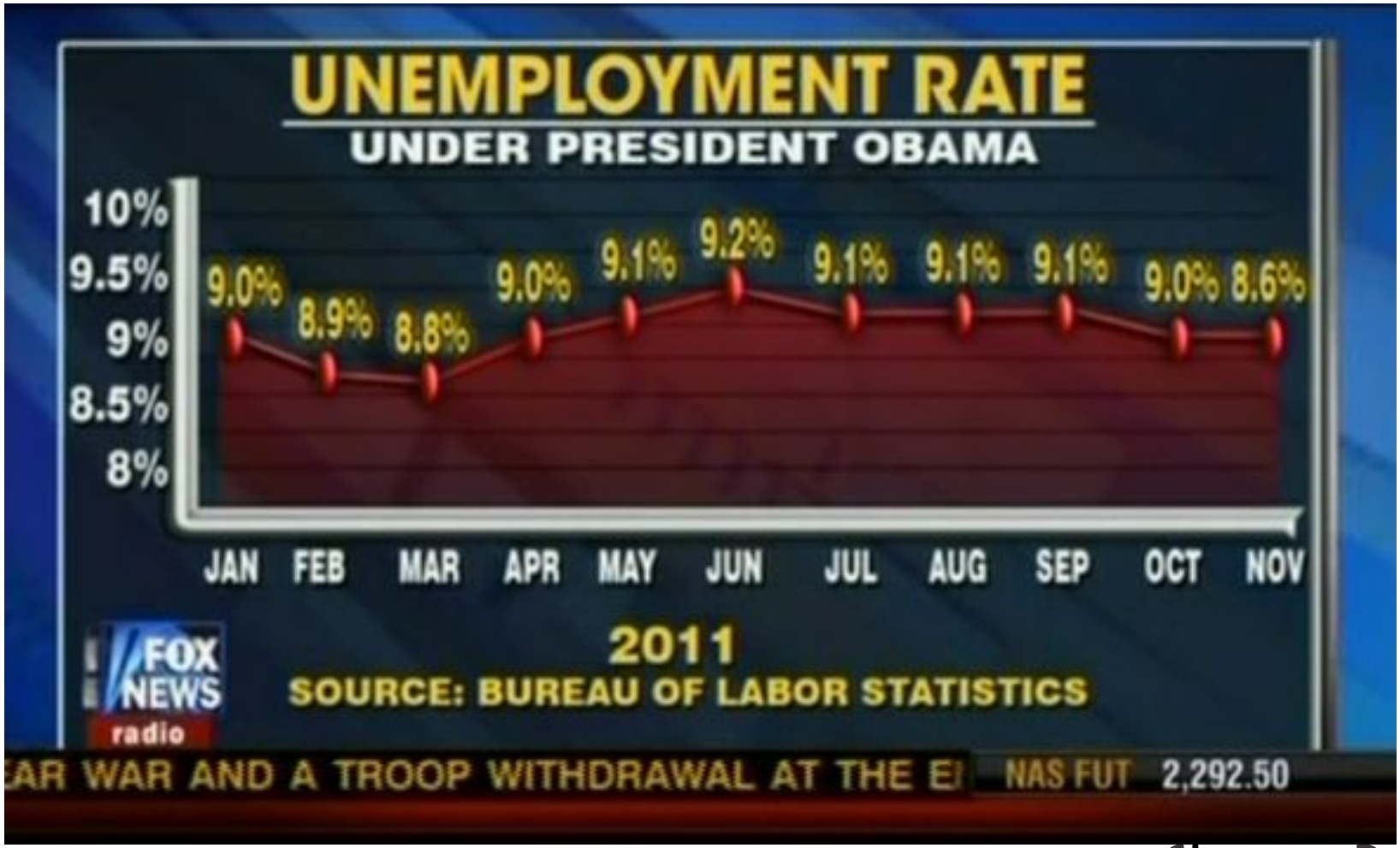


Final comments

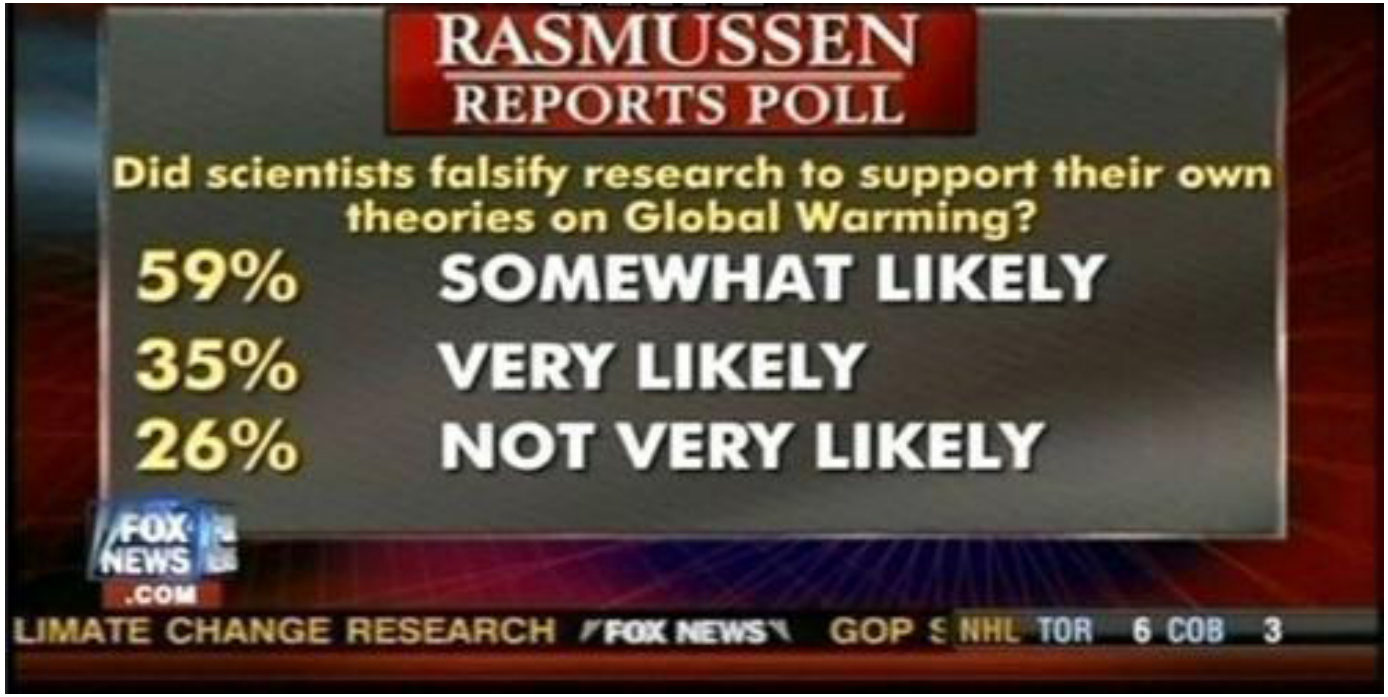




Final comments



Final comments



ESKERRIK ASKO

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA