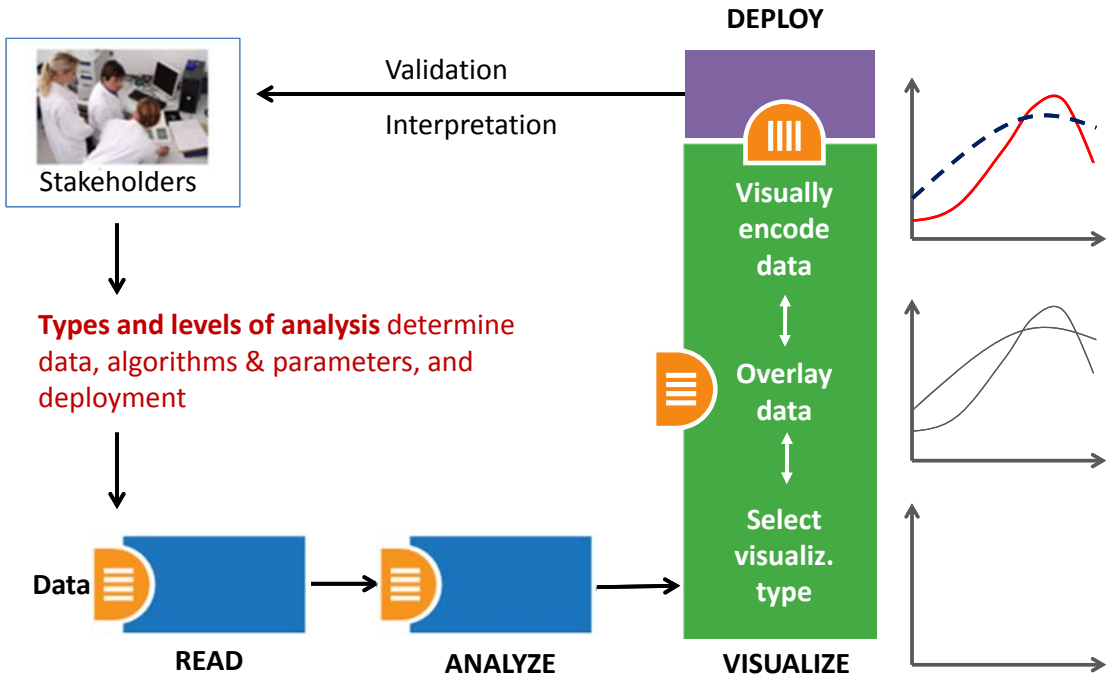
Needs-Driven Workflow Design



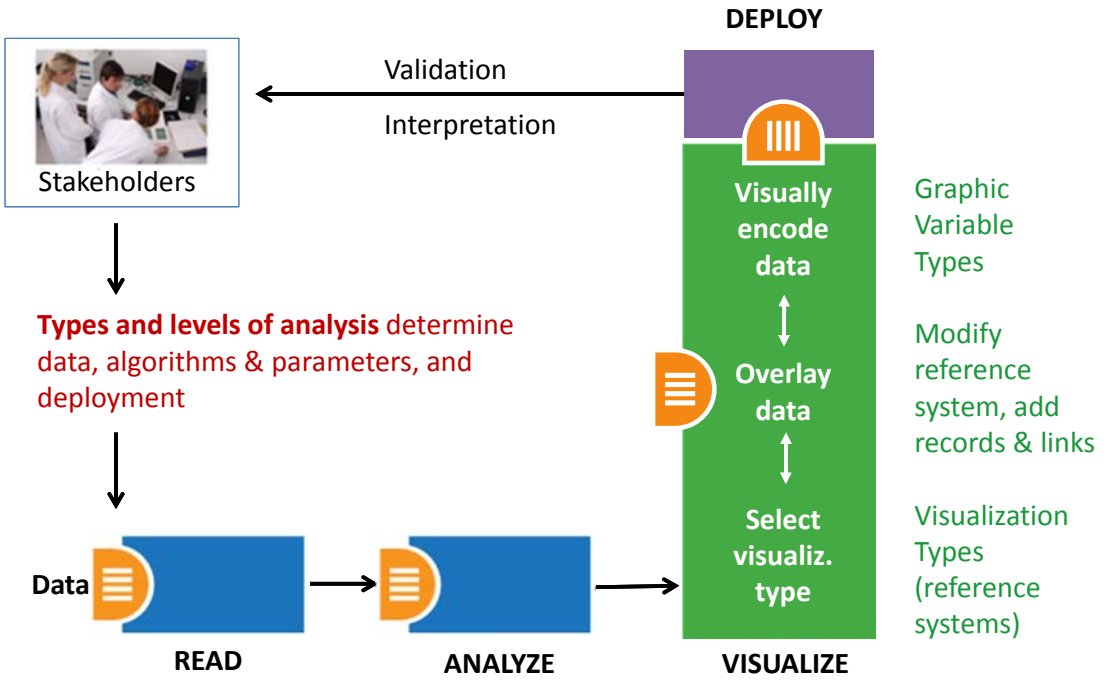
Information Visualization MOOC

Unit 2 – "When": Temporal Data

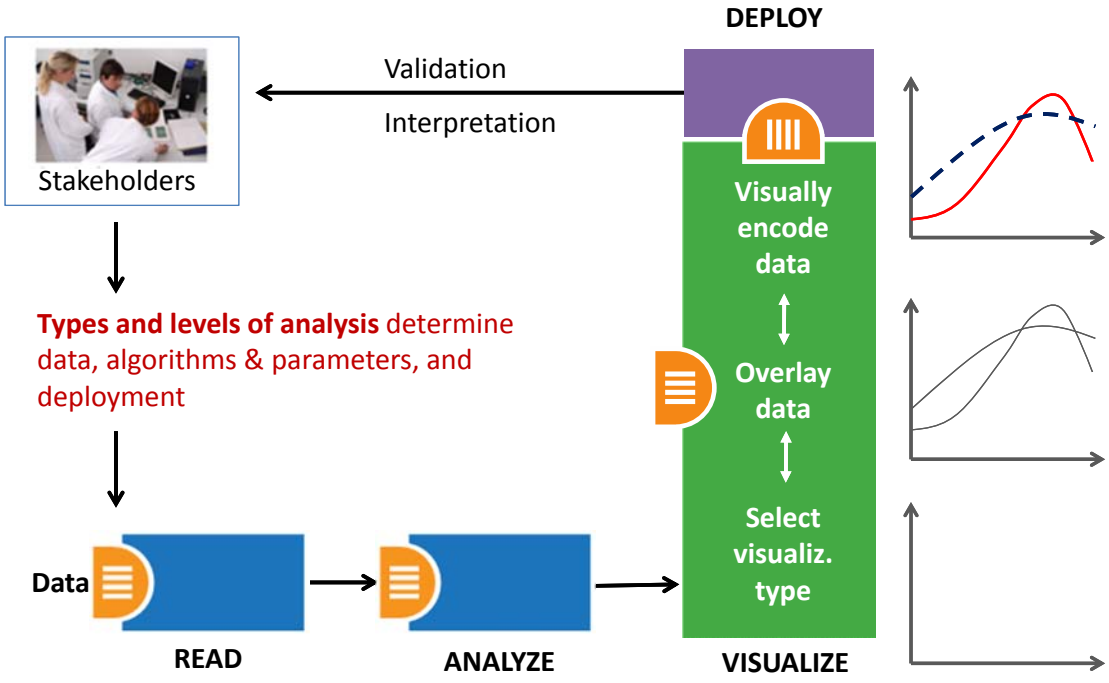
Workflow Design

Relevant Research Disciplines:
Mathematics, Statistics, Information Visualization

Needs-Driven Workflow Design



Needs-Driven Workflow Design



Read Data

Data Repositories:

- Gapminder data, http://www.gapminder.org/data/
- UCI Machine Learning Repository, http://archive.ics.uci.edu/ml (Time Series)
- Scholarly Database, http://sdb.cns.iu.edu
- IBM ManyEyes Datasets, http://www-958.ibm.com/software/data/cognos/manyeyes/datasets (350,976)
- Eurostat Data Market, http://datamarket.com

Data Formats:

- TXT
- XLS, CSV
- Databases

Data Formats

Time-series events are commonly represented by lists—e.g., emails

Subject header	**Date**	**Time**
Meeting to get key	1/1/2012	2:01 PM
CNS Talk–Meet Speaker?	1/1/2012	2:05 PM
Review paper draft	1/1/2012	3:01 PM
Happy New Year!	1/1/2012	3:01 PM
Action List Reminder	1/1/2012	3:04 PM

or Amazon book rankings over time:

Month	**Day**	**Year**	**Time**	**Rank**
10	23	2011	22:00	63,290
10	23	2011	23:00	76,128
10	24	2011	0:00	76,128
10	24	2011	1:00	95,578
10	24	2011	2:00	104,756

Each event is represented by a data record (row) that has temporal attribute(s) and possibly other attributes.

Data Preprocessing

Filtering—e.g., time slicing (see next slide)

Test for and remove large spikes in the data, but report this.

Normalization

- Deduplication
- Unit conversation
- Adjust (dollars) for inflation
- Adjust for time zones

Integration/interlinkage of different data sources

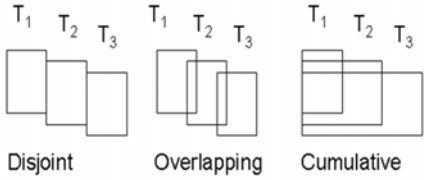
Classification/aggregation

Data Preprocessing – Time Slicing

Resolution: Milliseconds, seconds, minutes, hours, days, weeks, fortnights (fourteen days /two weeks), months, quarters, years, decades, and centuries.

Type:

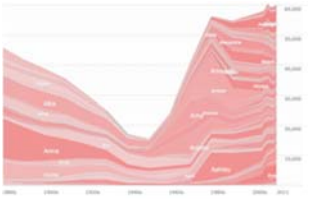
- Disjoint: Every row in the original table is in exactly one time slice.
- Overlapping: Selected rows are in multiple time slices.
- Cumulative: Every row in a time slice is in all later time slices.



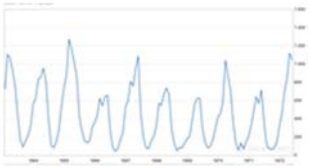
Alignment with calendar: If first event is June 7th, 2006, and yearly slices are chosen, then the first slice will be from

- June 7th, 2006, to June 6th, 2007 (if not aligned)
- January 1st, 2006, to December 31st, 2006 (if aligned)

Analyze Data



Trends—e.g., in baby names

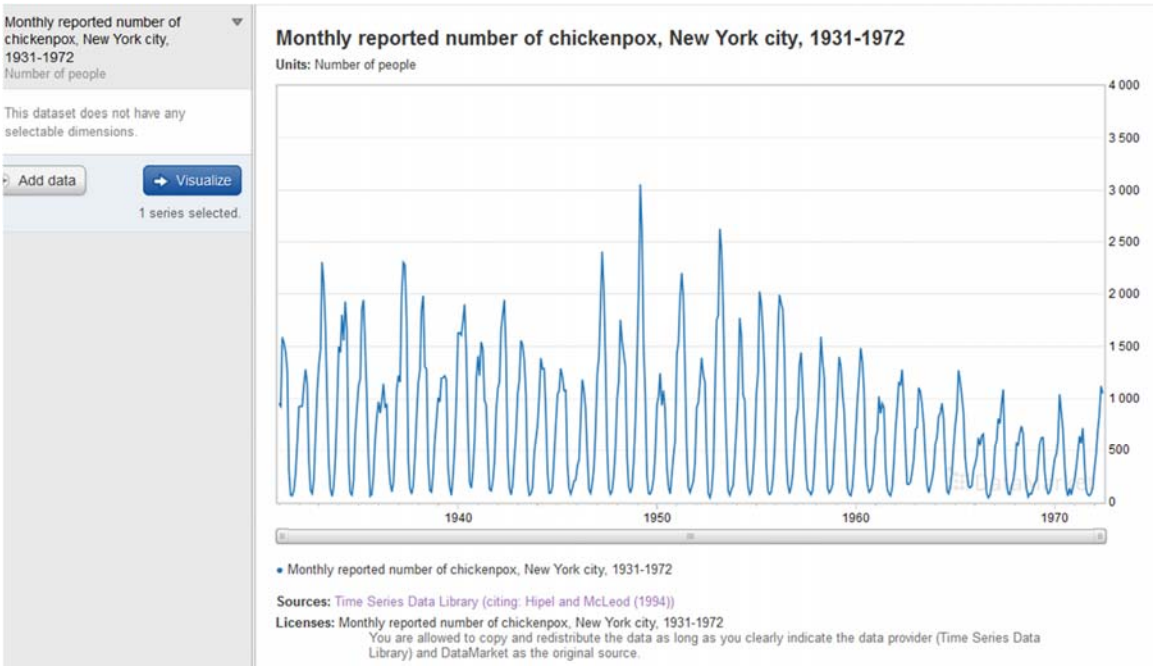


Patterns—e.g., seasonality in chickenpox cases

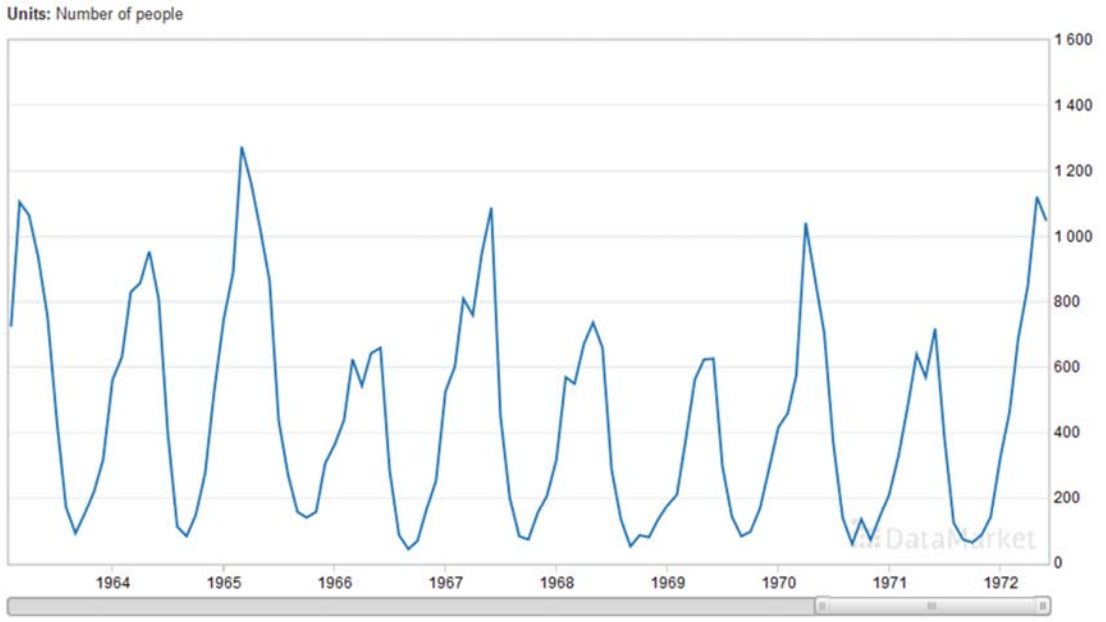
Correlations—e.g., in gas prices or movie ratings

Bursts—see final Unit 3 slide set.

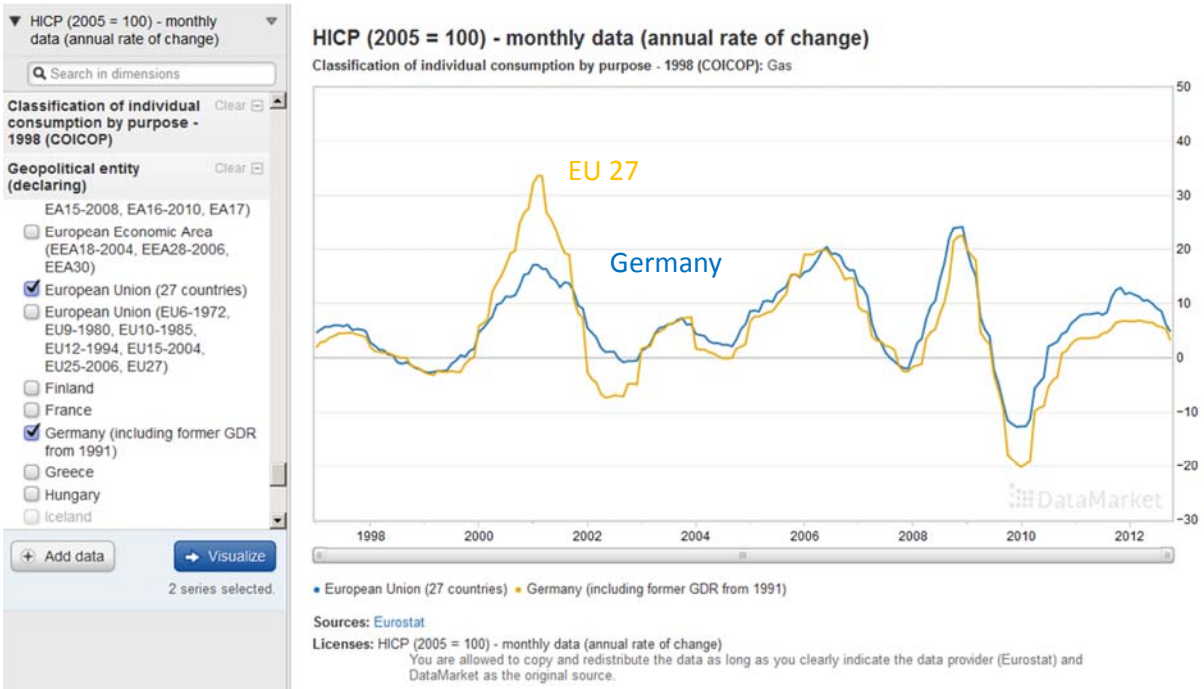
DataMarket: Monthly reported number of chickenpox, New York City, 1931-1972
http://datamarket.com/data/set/22v7/monthly-reported-number-of-chickenpox-new-york-city-1931-1972#!display=line&ds=22v7=provider:tsdl



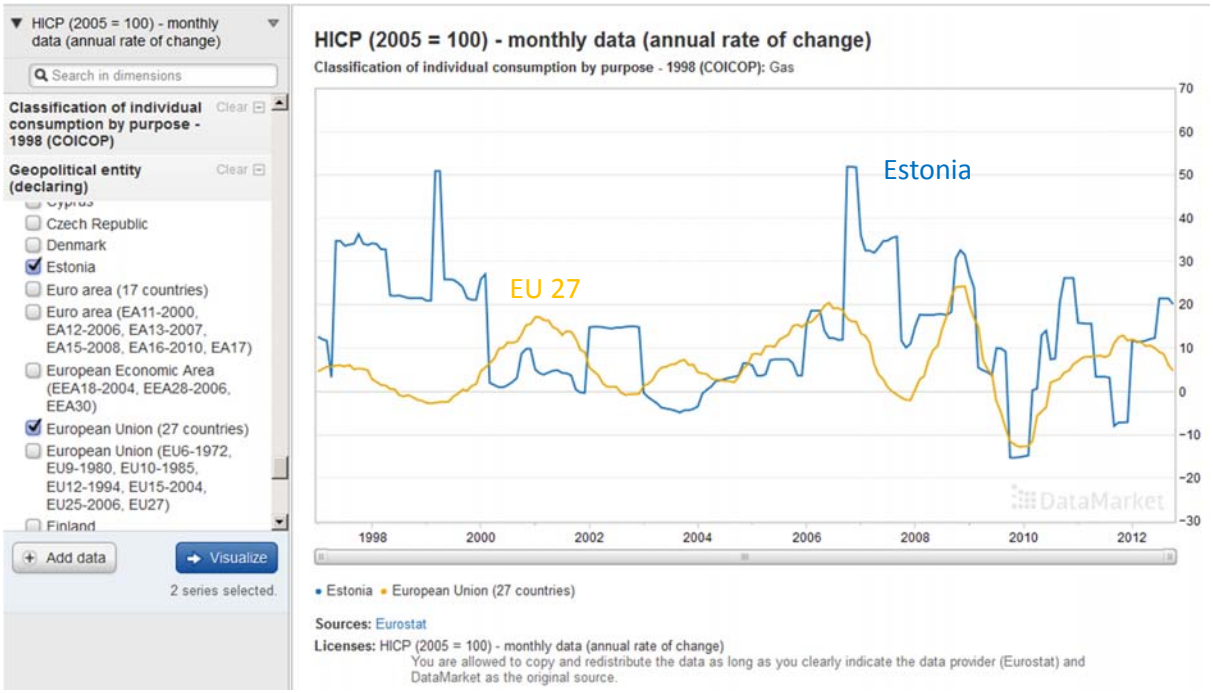
DataMarket: Monthly reported number of chickenpox, New York City, **Feb. 1963** -1972
http://datamarket.com/data/set/22v7/monthly-reported-number-of-chickenpox-new-york-city-1931-1972#!display=line&ds=22v7=provider:tsdl



DataMarket: Gas Prices from Jan. 1997 to Oct. 2012
http://datamarket.com/en/data/set/1a6e/#!ds=1a6e!qvc=4b:qvd=m.e&display=line



DataMarket: Gas Prices from Jan. 1997 to Oct. 2012
http://datamarket.com/en/data/set/1a6e/#!ds=1a6e!qvc=4b:qvd=m.e&display=line



DataMarket: Gas Prices from Jan. 1997 to Oct. 2012
http://datamarket.com/en/data/set/1a6e/#!ds=1a6e!qvc=4b:qvd=m.e&display=line

Is there a correlation? Germany and Estonia are part of EU27.

	A	B	C	D
1	HICP (2005 = 100) - monthly data (annual rate of change)			
2	Exported from datamarket.com			
3	Date exported	2012-11-30 18:35		
4	On DataMarket	http://datamarket.com/data/set/1a6e/#!display=line&height=370&width=720&hidden=null&ds=1a6e!qvc=4b:qvd=10.m.e		
5	License	You are allowed to copy and redistribute the data as long as you clearly indicate the data provider (Eurostat) and DataMarket as the original source.		
6	Provider	Eurostat		
7	Source URL	http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_hicp_manr&lang=en		
8	Units			
9				
10				
11				
12	Geopolitical en	Estonia	European	Germany (including former GDR from 1991)
13	Month			
14	1997-01	12.6	4.6	1.9
15	1997-02	12	5.2	2.9
16	1997-03	11.7	5.7	3
17	1997-04	3.3	5.7	3.7
18	1997-05	34.6	6	4
19	1997-06	34.6	6	4.5
20	1997-07	33.5	5.8	4.5
21	1997-08	33.8	6.1	4.6

Also relevant for calculating if:

- Twitter activity correlated to stock market behavior.
- Paper download counts correlated to citation counts.

Correlation

Use MS Excel or other statistics program to calculate

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$, where $\bar{x}$ and $\bar{y}$ are the means.

E15 =CORREL(C14:C203,D14:D203)

	A	B	C	D	E	F	G
11							
12	Geopolitical en	Estonia	European	Germany (including former GDR from 1991)			
13	Month						
14	1997-01	12.6	4.6	1.9			
15	1997-02	12	5.2	2.9	0.905826		
16	1997-03	11.7	5.7	3	0.049288		
17	1997-04	3.3	5.7	3.7			

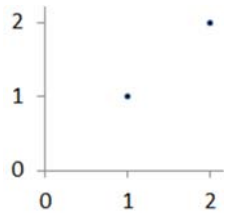
Note: Sequence of data pairs does not impact correlation result.

Correlation

Values run from -1 (no correlation) to 1 (identical).

D1 =CORREL(A1:A7,B1:B7)

	A	B	C	D	E	F
1	2	2	1	1		
2	1	1	2	-1		
3	2	2	1			
4	1	1	2			
5	2	2	1			
6	1	1	2			
7	2	2	1			



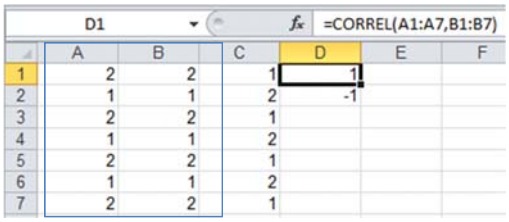
Positive association:
Upward trend (positive slope).

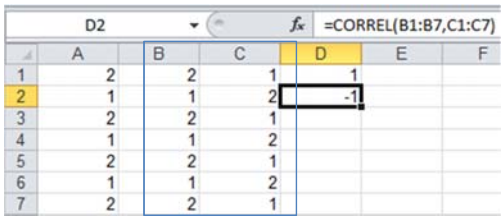
Negative association:
Downward trend (negative slope).

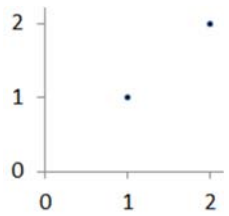
Non-associated: No trends.

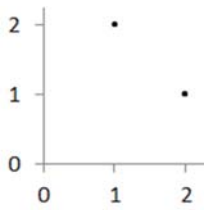
Correlation

Values run from -1 (no correlation) to 1 (identical).





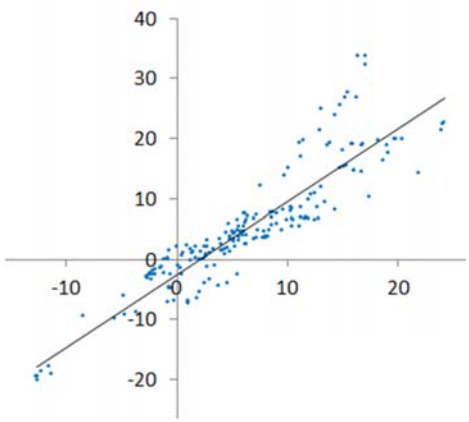




DataMarket: Gas Prices from Jan. 1997 to Oct. 2012
http://datamarket.com/en/data/set/1a6e/#!ds=1a6e!qvc=4b:qvd=m.e&display=line

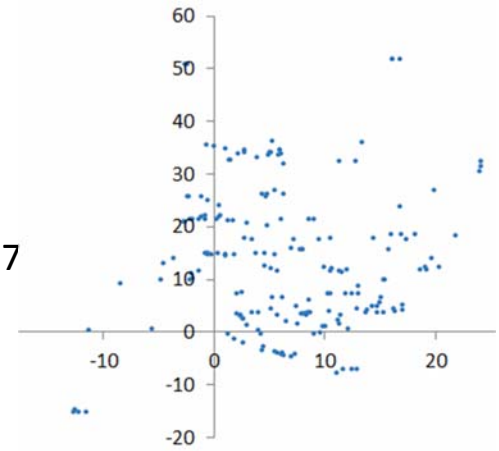
Is there a correlation? Germany and Estonia are part of EU27.
Visualization via scatterplot:

Germany (0.9058)



EU27

Estonia (0.04929)



EU27

Top-250 Movies from IMDb

Copy from web page at http://www.imdb.com/chart/top (on Nov 15, 2012), save as text file, open in Excel or other table editing program:

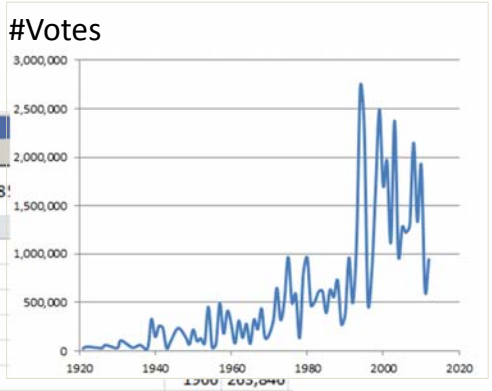
Rank	Rating	Title	Votes
1.	9.2	The Shawshank Redemption (1994)	857,280
2.	9.2	The Godfather (1972)	625,241
3.	9.0	T	
4.	8.9	P	
5.	8.9	T	
6.	8.9	1	
7.	8.9	S	
8.	8.9	T	

imdb-data.txt - Notepad

File Edit Format View Help

```
Rank    Rating  Title   Votes
1.      9.2     The Shawshank Redemption (1994) 8
2.      9.2     The Godfather (1972)    625,241
3.
4.
5.
6.
7.
8.
```

	A	B	C		
1	Rank	Rating	Title		
2	1	9.2	The Shawshank Redemption		
3	2	9.2	The Godfather		
4	3	9	The Godfather: Part II		
5	4	8.9	Pulp Fiction		
6	5	8.9	The Good, the Bad and the Ugly		
7	6	8.9	12 Angry Men	1957	211,144
8	7	8.9	Schindler's List	1993	444,891
9	8	8.9	The Dark Knight	2008	837,033

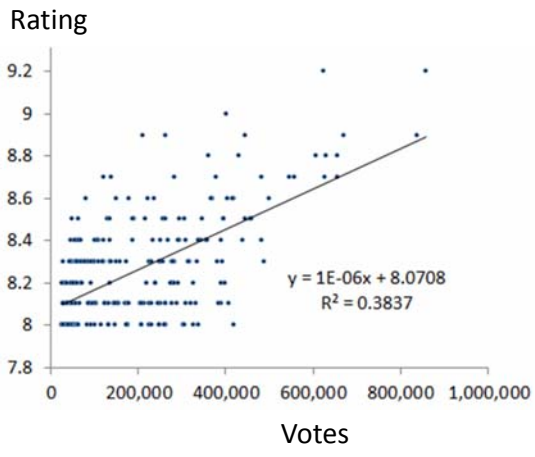


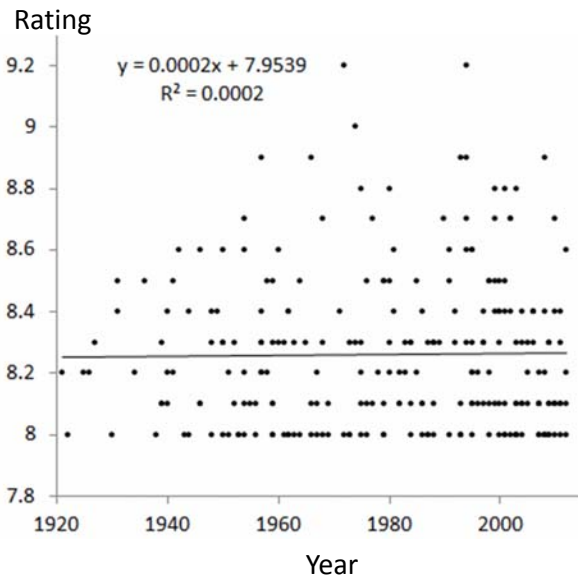
Is there a correlation between Rating and #Votes?

Any trends over time? Recent movies more popular?

Top-250 Movies from IMDb

	A	B	C	D	E
1	Rank	Rating	Title	Year	Votes
2	128	8.2	The Kid	1921	27,356
3	234	8	Nosferatu	1922	44,388
4	136	8.2	The Gold Rush	1925	33,973





Is there a correlation between Rating and #Votes?

Any trends over time? Recent movies more popular?

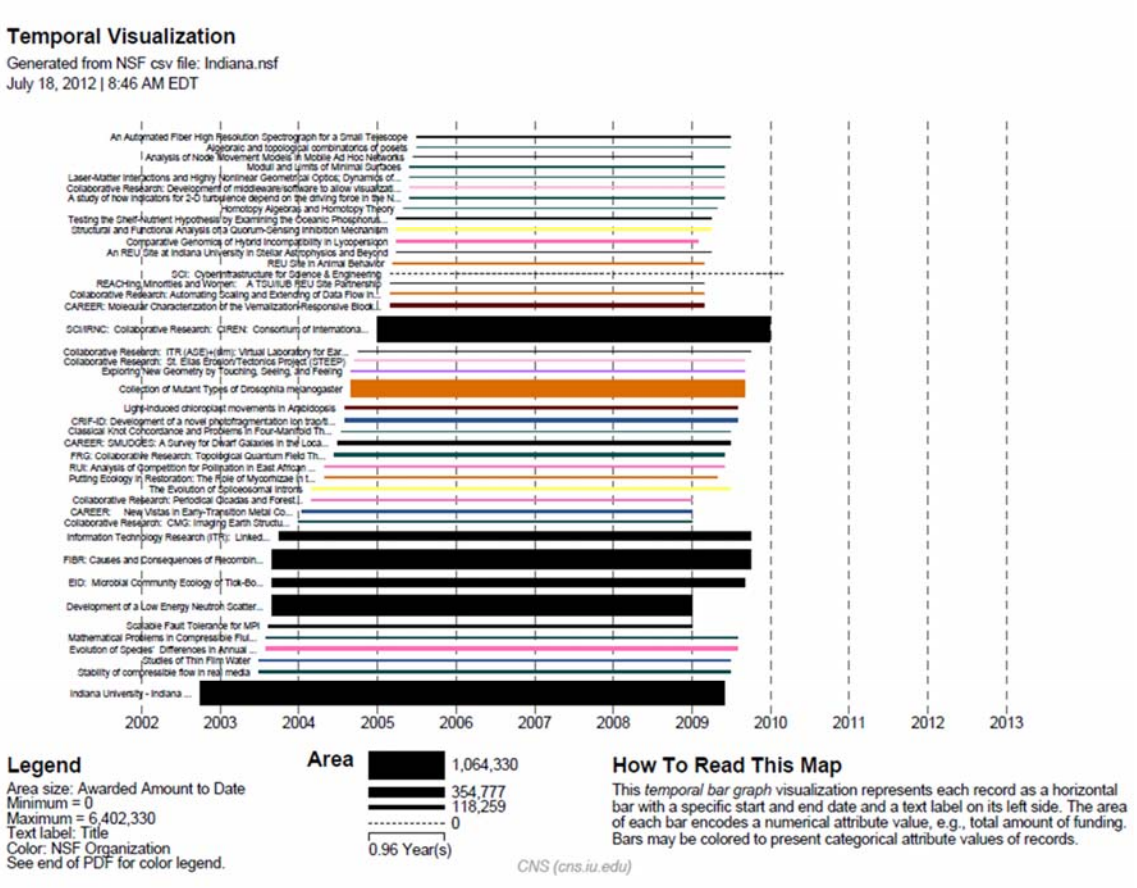
Visualize Data

Spreadsheet programs such as MS Excel or the free Apache Open Office (http://www.openoffice.org) support the easy generation of diverse charts and graphs.

Temporal bar graphs, see next slide and Unit 2: Hands-on.

Time slice data to generate, e.g.,

- Evolving geomaps, see Unit 3.
- Evolving networks, see Unit 7: Hands-on.



Relevant Tools

- TimeSearcher from HCIL supports the visual exploration of time-series data http://www.cs.umd.edu/hcil/timesearcher/
- Tableau, http://www.tableausoftware.com

Please post your favorite to Twitter or Flickr using tags "#ivmooc" and "#timetools."