Information Analytics

1 November 2016

Robin Boast and Rens Bod

Department of Media Studies

Faculty of Humanities

University of Amsterdam

Goals of this course

- To understand the process of signification and data creation;
- To understand the basics of statistical inference and probability.
- To critically apply elementary data-analysis methods and analyse meaningful patterns in the data; and
- To appoint and recognise the possibilities and limitations of information analysis

Schedule

1 Nov	Introduction to Analysis	Boast/Bod
8 Nov	Curated Content and Representation	Boast
15 Nov	Information and Data in the Humanities	Bod
22 Nov	Statistics and Statistical Inference	Boast
29 Nov	Analysing the Humanities	Bod
6 Dec	Bayesian Analysis": "and Analytical Questions in the Digital Humanities	Bod
13 Dec	Algorithms and Computational Logic	Boast
Week 19 Dec	Exam	
	No Lessons	
9	Project presentation and first work, and Development	Boast
10	Proposed methods (Assignment 1)	TBA
11	Discussion of the analyses (Assignment 2)	Boast
12	End Presentations	Boast/Bod

What is Analysis?

- **Analysis** is the process of breaking a complex substance into smaller parts in order to gain a better understanding of it
- In our situation, the substance is: DATA
- So what is our method?

Why Use a Method Anyway?

- There is only one reason:
  - because it helps you answer your research questions
- You have to start with a research question
  - then look for suitable methods and tools
  - pick the one that delivers the most interesting data
  - or use several to explore different angles/viewpoints
    - There no such thing as "*The* Method"

Differences Between Humanities & Sciences

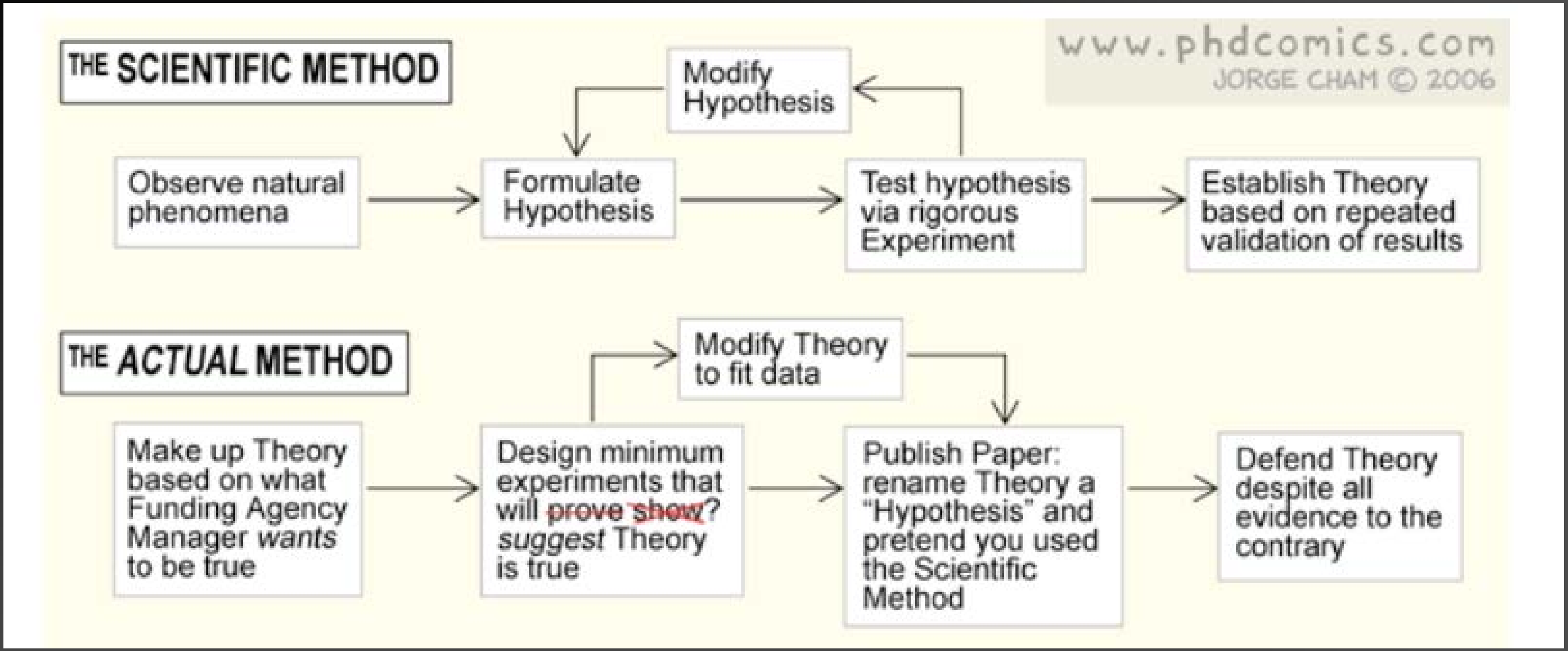
- Although could study the same objects
- Common wisdom:
  - sciences: truth-finding paradigm (exact science)
  - humanities: interpretative paradigm, many parallel ‘realities’

- The two paradigms are categorically different
  - at least, that’s what armchair philosophers claim
  - In concrete cases, they can fruitfully interact with and contribute to each other

Empirical Research

- Goal is to do measurable, repeatable research
  - using scientific method
  - used to distinguish science from non-science
- • Clearly laid out in sequence of steps:
  - 1. Hypothesis
  - 2. Method
  - 3. Results
  - 4. Conclusion

Scientific or Actual Method?


www.phdcomics.com
JORGE CHAM © 2006
THE SCIENTIFIC METHOD
Observe natural phenomena
Formulate Hypothesis
Modify Hypothesis
Test hypothesis via rigorous Experiment
Establish Theory based on repeated validation of results
THE ACTUAL METHOD
Make up Theory based on what Funding Agency Manager *wants* to be true
Design minimum experiments that will ~~prove~~ ~~show~~? *suggest* Theory is true
Modify Theory to fit data
Publish Paper: rename Theory a "Hypothesis" and pretend you used the Scientific Method
Defend Theory despite all evidence to the contrary

Qualitative and Quantitative Methods

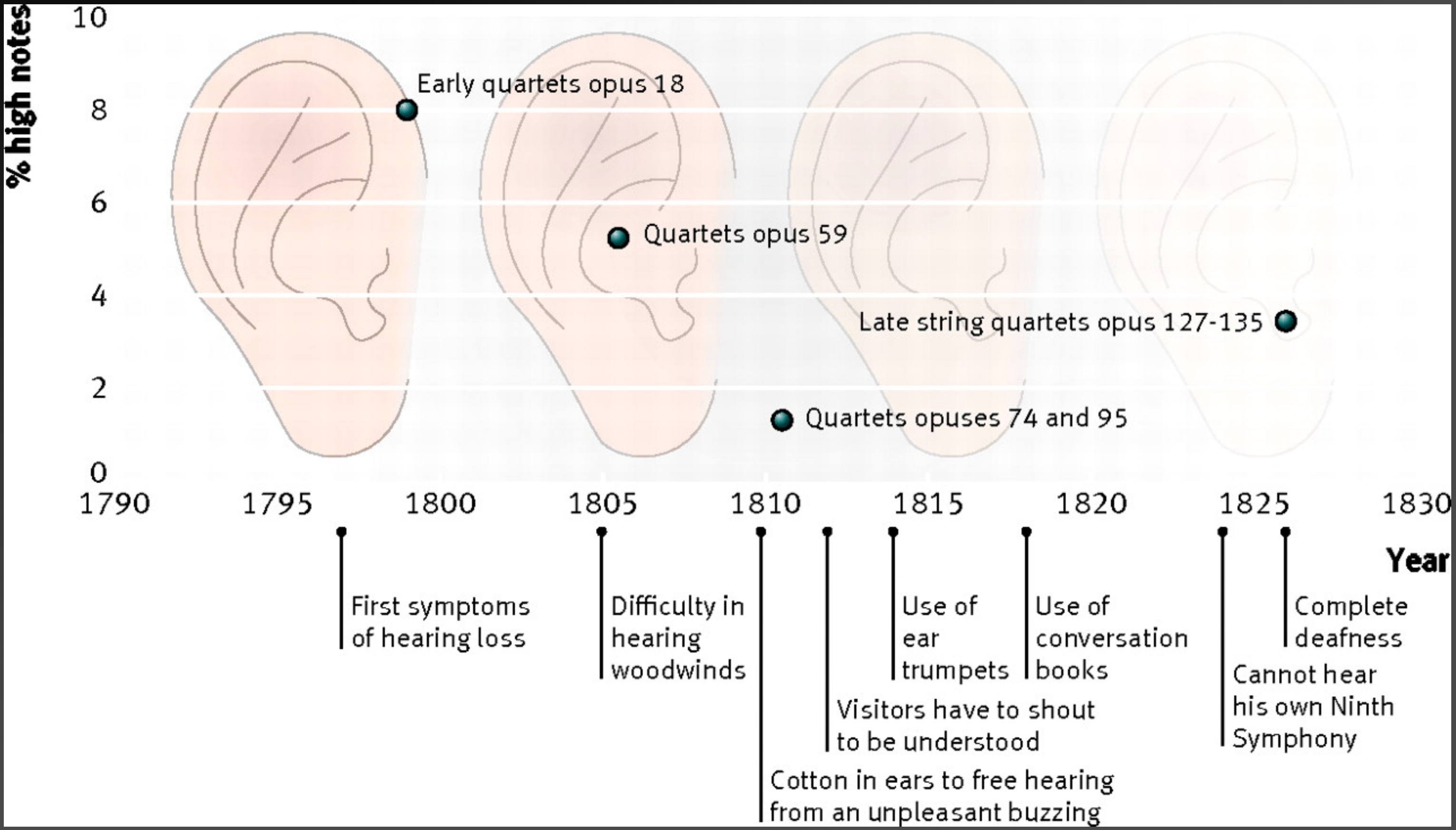
1. Get to know your data
2. Categorize the information
   - Identify themes or pattern
   - Organize them into coherent categories
3. Identify patterns and connections within and between categories
4. Interpretation – bring it all together

But what are patterns and interpretations?

- What is the role of patterns in the humanities?

- Patterns are ubiquitous in the humanities!
  - And they form a focus of study in Information Analytics
- Let me give some examples

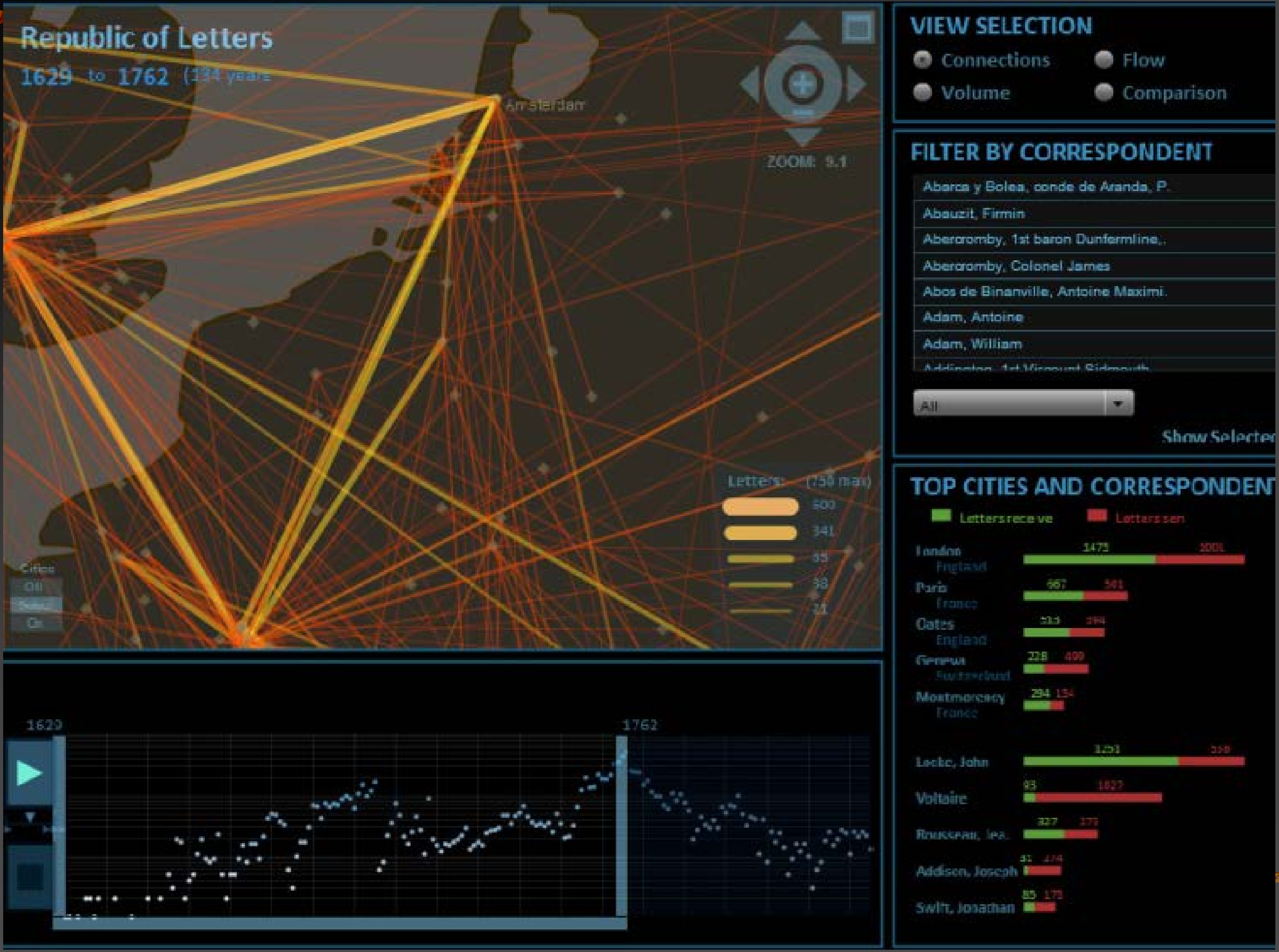
Division of Beethoven compositions into 3 style periods (Adler, 1911)



The way in which painters have represented the wind in a person's hair (A. Warburg, 1924)

Thus patterns are not always quantitative!

The network of the Republic of Letters (project at Stanford U.)



Republic of Letters

1629 to 1762 (134 years)

VIEW SELECTION

● Connections ○ Flow
○ Volume ○ Comparison

FILTER BY CORRESPONDENT

- Abarca y Bolea, conde de Aranda, P.
- Abauzit, Firmin
- Abercromby, 1st baron Dunfermline,.
- Abercromby, Colonel James
- Abos de Binanville, Antoine Maximi.
- Adam, Antoine
- Adam, William
- Addington, 1st Viscount Sidmouth

All ▾

Show Selected

TOP CITIES AND CORRESPONDENT

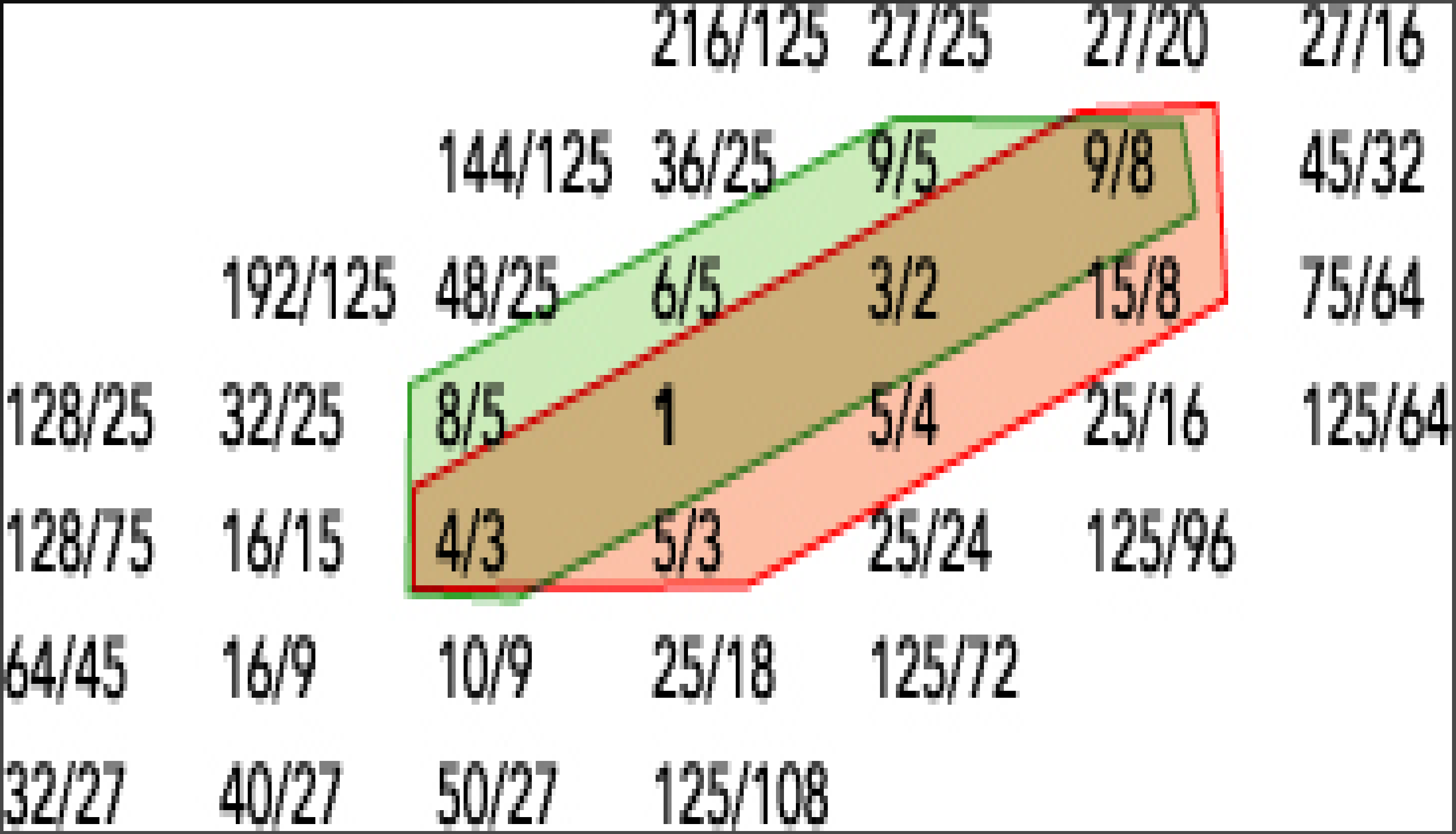
🟩 Letters receive 🟥 Letters sen

	🟩 Letters receive	🟥 Letters sen
London England	1475	1001
Paris France	967	541
Oates England	533	394
Geneva Switzerland	228	499
Montmorency France	294	154
Locke, John	1251	558
Voltaire	93	1627
Rousseau, Jea.	327	173
Addison, Joseph	31	174
Swift, Jonathan	85	173

The shift from voiceless to voiced consonants in language change (J. Grimm, 1822)

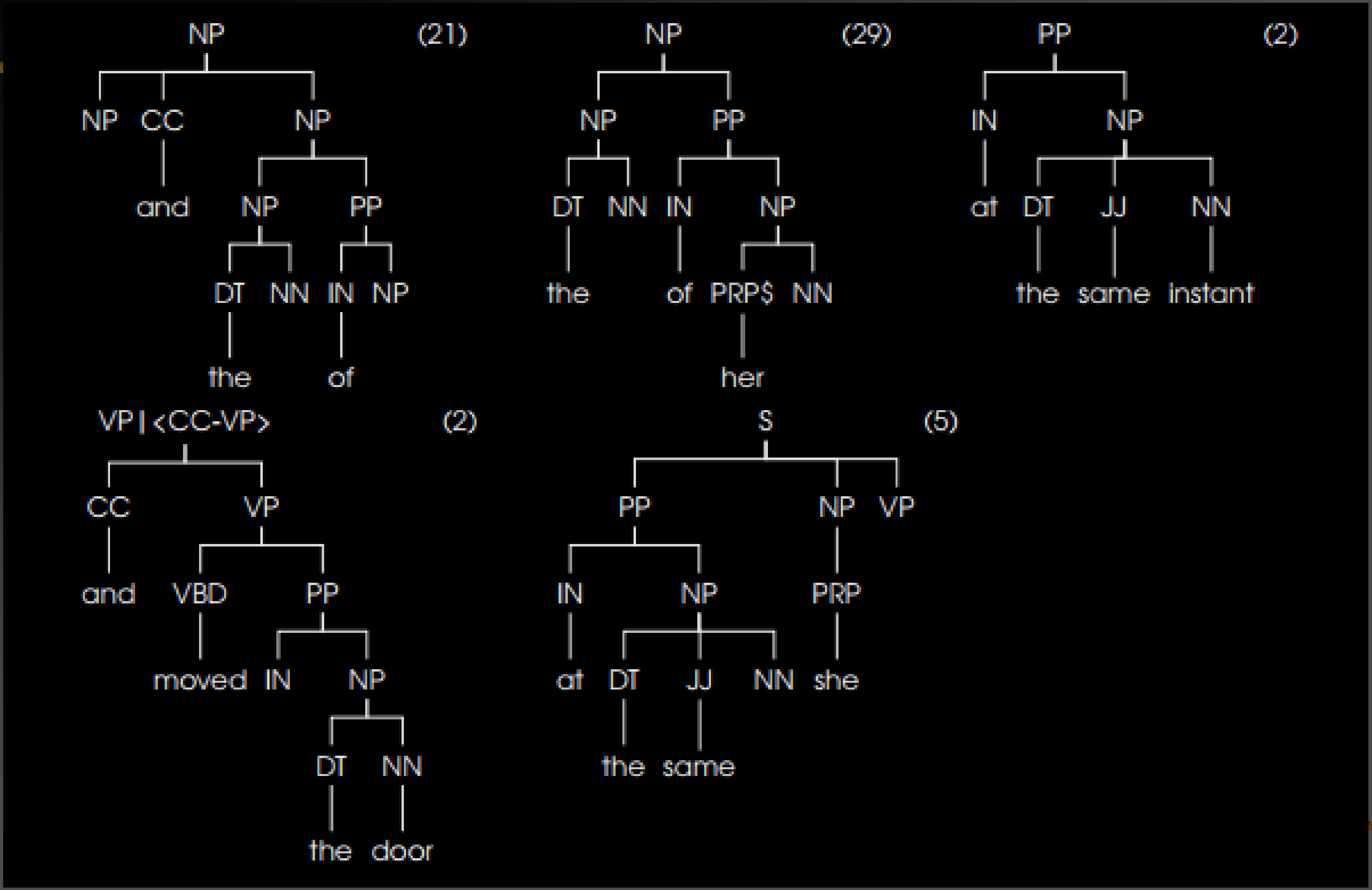
p b f	*t d th*	*k g x*
↓	↓	↓
f b p	*th t d*	*x k g*
↓	↓	↓
b f p	*d z t*	*g x k*

The convex structure of musical scales (A. Honingh 2006)



			216/125	27/25	27/20	27/16
		144/125	36/25	9/5	9/8	45/32
	192/125	48/25	6/5	3/2	15/8	75/64
128/25	32/25	8/5	1	5/4	25/16	125/64
128/75	16/15	4/3	5/3	25/24	125/96	
64/45	16/9	10/9	25/18	125/72		
32/27	40/27	50/27	125/108			

Recurring syntactic constructions in literature (V. Cranenburgh 2012)



So what is a pattern?

- From R. Bod (*A New History of the Humanities*, 2013):

“A pattern is a trend or a tendency that can range from the local to the global. It can consist of a regularity (often with exceptions) but also of a grammatical rule, or a historical trend. Some patterns may be similar to ‘laws’ such as the sound shift laws in linguistics or the laws of harmony in music, other patterns are entirely variable. The notion of ‘pattern’ is thus an umbrella term that covers everything that can be found between inexact trends and exact laws.”

Patterns are ubiquitous

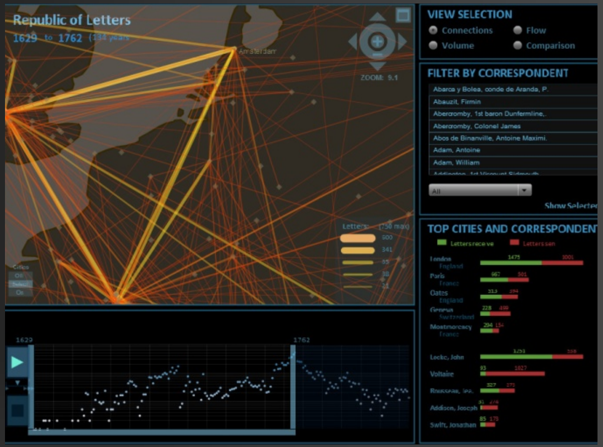
- Quest for patterns is found in:
  - all disciplines - from linguistics to historiography
  - all periods - from Antiquity up to the present day
  - all regions - from China to India to Africa to Europe

(R. Bod, 2013. *A New History of the Humanities*, OUP)

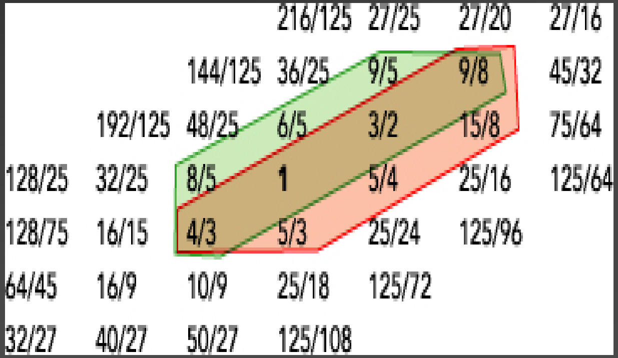
- Identification of patterns is enormously aided by digital techniques
  - Inducing patterns from massive amounts of data – but only if these data are in searchable format

There are even patterns that can *not* be identified without digital means:

Network of Republic of Letters?



Convexity of musical scales



			216/125	27/25	27/20	27/16
		144/125	36/25	9/5	9/8	45/32
	192/125	48/25	6/5	3/2	15/8	75/64
128/25	32/25	8/5	1	5/4	25/16	125/64
128/75	16/15	4/3	5/3	25/24	125/96	
64/45	16/9	10/9	25/18	125/72		
32/27	40/27	50/27	125/108			

Distinction between Deep and Shallow patterns

- In texts,
  - Deep: syntactic structures → hierarchical
  - Shallow: word sequences → linear
- In paintings
  - Deep: composition → hierarchical
  - Shallow: lines, colours → linear

- Both deep and shallow patterns are useful, and there is a continuity

Some patterns are local, others are (claimed to be) universal

LOCAL

- Division of Beethoven’s compositions into three style periods
- Way in which painters have represented the wind in a girl’s hair
- Shift from voiceless to voiced consonants in language change

UNIVERSAL

- Convex structure of traditional musical scales
- Basic word order of the world’s languages
- Use of recurrent phrases, themes and episodes in (oral) literature

Both local and universal, deep and shallow patterns exist – but need to be interpreted

- humanities 1.0 : the hermeneutic (interpretative) and critical tradition as it was developed during the 19th and early 20th century

- humanities 2.0: refers to the use of digital tools to humanities material: search for patterns

- humanities 3.0: refers to the hermeneutic and critical tradition applied to (deep and shallow) patterns found

→ humanities 3.0 is both pattern-oriented and hermeneutic

(Bod 2012)

Preview: Analyzing textual patterns

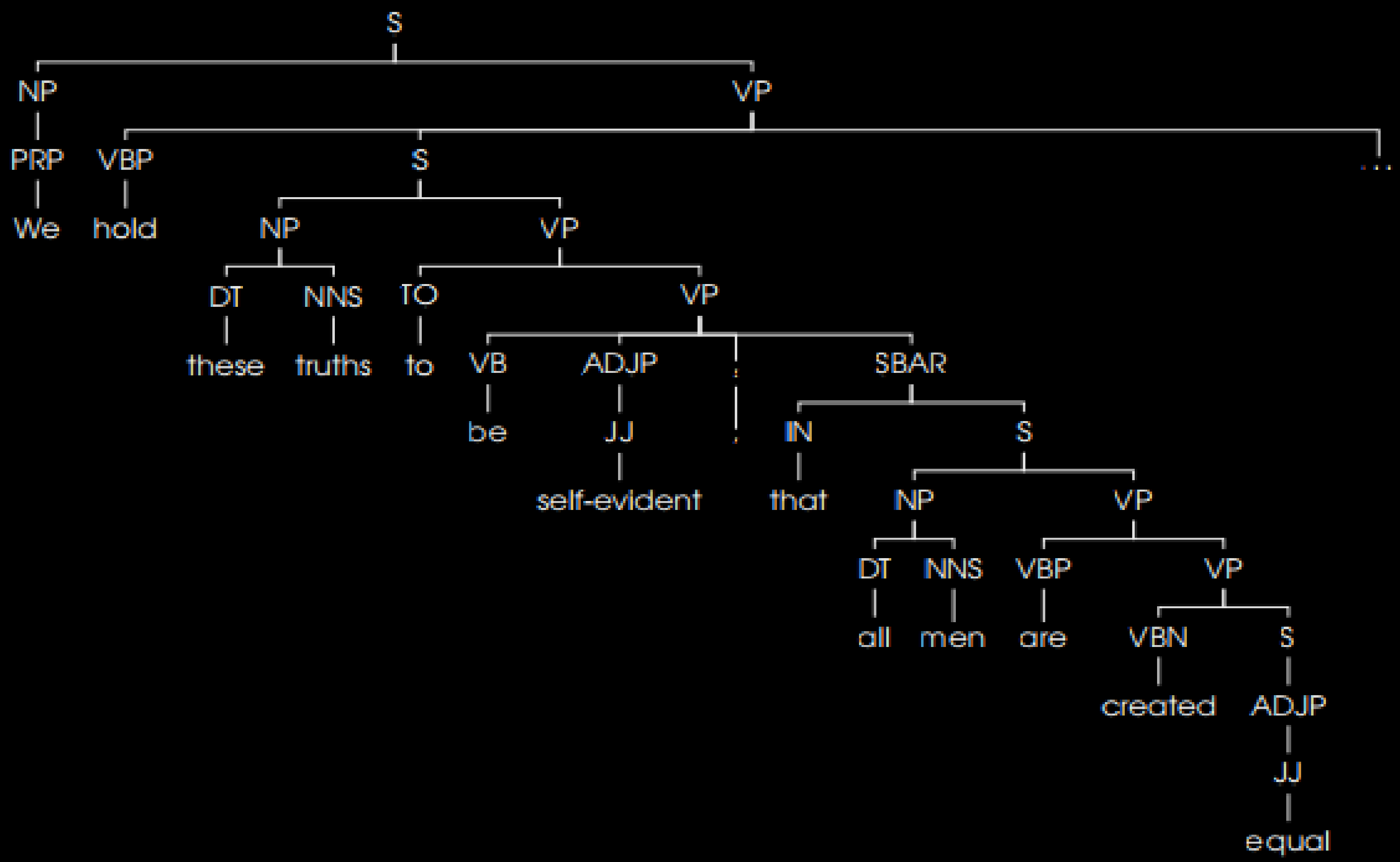
- Two kinds of textual patterns:
  - Words and word frequencies: n-grams (think of Google ngram viewer):
    - “flat”, linear, no hierarchical structure
  - Syntactic tree-patterns based on parsed texts:
    - “deep”, hierarchical, non-linear, discontinuous structure

Textual repositories

- From Literary domain, e.g. Project Gutenberg: 45.000 books: www.gutenberg.org …
- … to Journalist domain, e.g. KB Historical Newspapers, 9 million pages: kranten.kb.nl
  - And there is much and much more…
- We lose valuable information if we do not deal with the linguistic complexity of these textual sources

A sentence and its syntactic structure

"We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." — Declaration Of Independence, 1776, Thomas Jefferson



Textual similarity

- Question: where did Thomas Jefferson get his information from?
  - Which (philosophical) texts are similar to his text? Descartes? Locke? Burlamaqui? Others?
- For such questions it is not enough to look at words only, but need look at syntactic (and semantic) constructions
  - Easily obtainable
  - Current syntactic parsers for English, German, Dutch reach ~90-95% accuracy
  - Good enough for finding syntactic patterns

To be continued (15 November)!