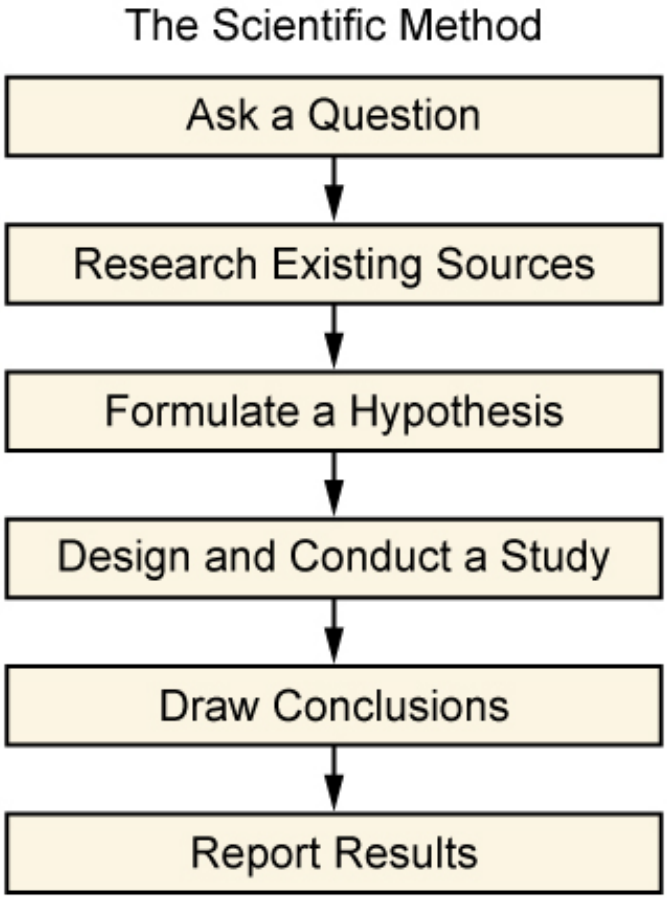
Data Preparation

TM the Great Unread

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

July 26, 2016

Research design process



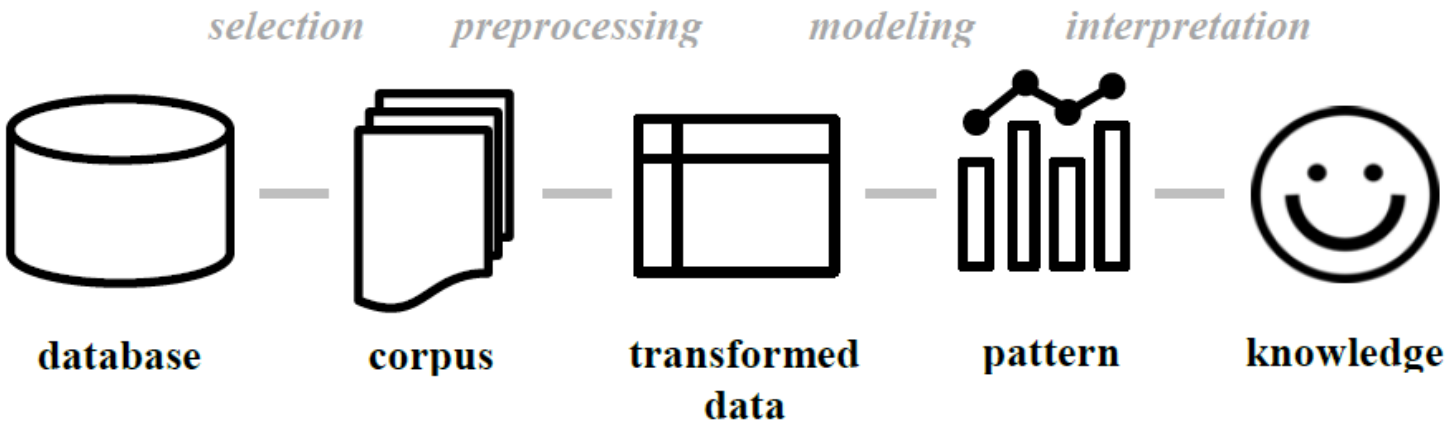
Design types:

- Exploratory
- Descriptive
- Causal

Independent of type, the actual process is always iterative

$\Rightarrow$ 'Mixed model' of research design

TM workflow



General multistep process of knowledge discovery

Constructed out of a need for handling ‘extraction of useful information (knowledge) from rapidly growing volumes of digital data’

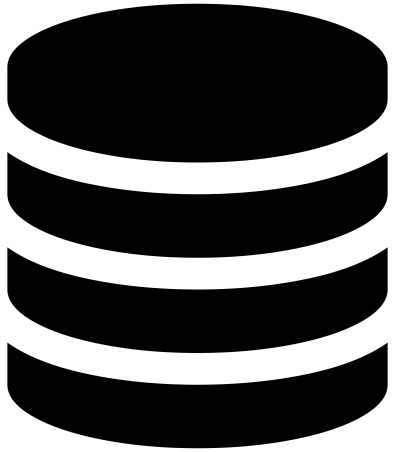
For each project you develop a pipeline within this workflow

Accessing data

Select specific documents (target data/corpus) relevant to your research question from the database.

Online databases and research libraries are excellent resources

- Beware of legal issues, always ask your provider/supervisor
- Historical sources can be a problem



We will focus on documents stored locally in a plain text format (R is very accommodating though).

```
1 library(rvest)
2 walkingdead.l <- html("http://www.imdb.com/title/tt1520211/")
3 cast.v <- html_text(html_nodes(walkingdead.l,
4                "#titleCast .itemprop span"))
```

Packages

"You gotta know when to be lazy. Done correctly, it's an art form that benefits everyone." (Nicholas Sparks, *The Choice*)

Package (or library): Collection of resources (code and data) and associated documentation [1]

Download and install packages from *The Comprehensive R Archive Network* (CRAN) repositories [2] or from local files.

```
1 install.packages("tm")
2 library(tm)
3 library() # available packages
4 sessionInfo() # attached packages
```

[1] http://www.inside-r.org/category/packagetags/naturallanguageprocessing

[2] or CRAN-like

Preprocessing

To prepare a document ('a string of characters') we need to split or *tokenize* it at the relevant level(s).
Unstructured data are very noisy. To increase the signal we therefore remove irrelevant data through preprocessing.

Use a range of text normalization techniques to preprocess the data:

- Casefolding
- Removal of non-alphanumeric characters (punctuation, blanks)
- Removal of numeral and stopwords
- Reduction of inflectional forms through stemming and lemmatization
- Synonym substitution

Remember that one man's rubbish may be another's treasure