Big Data and Automated Content Analysis I+II

Week 1 – Wednesday
»Introduction«

Damian Trilling

d.c.trilling@uva.nl
@damian0604
www.damiantrilling.net

Afdeling Communicatiewetenschap
Universiteit van Amsterdam

5 February 2019

Today

Introducing…	What is Big Data?	Methods	What have others done?	What can we do?	The schedule
●○○○	○○○○○○○○	○○○○○○○○○○○○○	○○○○○○○○○	○○○○○	○○○○

… the people

Introducing…
        … the people

Introducing...	What is Big Data?	Methods	What have others done?	What can we do?	The schedule
○●○○	○○○○○○○○	○○○○○○○○○○○○○	○○○○○○○○○	○○○○○	○○○○

...the people

Introducing...
Damian

dr. Damian Trilling
Assistant Professor Political Communication &
Journalism

- studied Communication Science in Münster and at the VU 2003–2009

- PhD candidate @ ASCoR 2009–2012

- interested in political communication and journalism in a changing media environment and in innovative (digital, large-scale, computational) research methods

@damian0604   d.c.trilling@uva.nl
REC-C 8th floor   www.damiantrilling.net
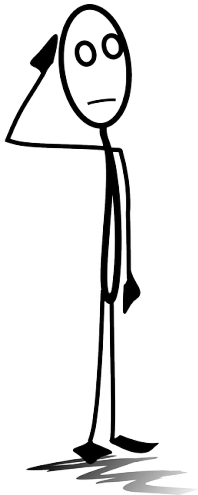
Introducing...
Anne

dr. Anne Kroon
Assistant Professor Corporate Communication

- Stereotypes in the media

- ACA and word embeddings

@annekroon   a.c.kroon@uva.nl

Introducing...
You

Your name?
Your background?
Your reason to follow this course?

Introducing. . .
oooo

What is Big Data?
●ooooooo

Methods
oooooooooooo

What have others done?
ooooooooo

What can we do?
ooooo

The schedule
oooo

Definitions

What is Big Data?

Introducing... | **What is Big Data?** | Methods | What have others done? | What can we do? | The schedule
oooo | oooooooo | oooooooooooo | ooooooooo | ooooo | oooo
Definitions

What is Big Data?

Introducing... | **What is Big Data?** | Methods | What have others done? | What can we do? | The schedule
oooo | oo●ooooo | oooooooooooooo | ooooooooo | ooooo | oooo

Definitions

What is Big Data?

A simple technical definition could be:

Everything that needs so much computational power and/or storage that you cannot do it on a regular computer.

Introducing. . .   **What is Big Data?**   Methods   What have others done?   What can we do?   The schedule
oooo   oooo●oooo   oooooooooooooo   ooooooooo   ooooo   oooo
Definitions

What is Big Data?

What is Big Data?

Vis, 2013

- "commercial" definition (Gartner): "'Big data' is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"

What is Big Data?

Vis, 2013

- boyd & Crawford definition:
  1. Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
  2. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
  3. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

What is Big Data?

Vis, 2013

- "commercial" definition (Gartner): "'Big data' is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"
- boyd & Crawford definition:
  1. Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
  2. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
  3. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

Introducing... | **What is Big Data?** | Methods | What have others done? | What can we do? | The schedule
0000 | 00000●000 | 0000000000000 | 000000000 | 00000 | 0000

Definitions

Implications & criticism

Introducing. . .   **What is Big Data?**   Methods   What have others done?   What can we do?   The schedule
0000   00000●000   0000000000000   000000000   00000   0000

Definitions

Implications & criticism

boyd & Crawford, 2012

❶ Big Data changes the definition of knowledge

❷ Claims to objectivity and accuracy are misleading

❸ Bigger data are not always better data

❹ Taken out of context, Big Data loses its meaning

❺ Just because it is accessible does not make it ethical

❻ Limited access to Big Data creates new digital divides

Introducing... | **What is Big Data?** | Methods | What have others done? | What can we do? | The schedule
oooo | oooooo○oo | oooooooooooooo | ooooooooo | ooooo | oooo

Definitions

APIs, researchers and tools *make* Big Data

Introducing. . . **What is Big Data?** Methods What have others done? What can we do? The schedule

Definitions

APIs, researchers and tools *make* Big Data

Vis, 2013

Inevitable influences of:

- APIs
- filtering, search strings, . . .
- changing services over time
- organizations that provide the data

Epistemologies and paradigm shifts

Kitchin, 2014

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough
- Data-driven science: knowledge discovery guided by theory

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough
- Data-driven science: knowledge discovery guided by theory
- Computational social science and digital humanities: employ Big Data research within existing epistemologies
  - DH: descriptive statistics, visualizations
  - CSS: prediction and simulation

Introducing. . .   **What is Big Data?**   Methods   What have others done?   What can we do?   The schedule
0000   0000000●   000000000000   000000000   00000   0000
Are we doing Big Data research?

Are we doing Big Data research in this course?

Introducing. . .   What is Big Data?   Methods   What have others done?   What can we do?   The schedule
0000   0000000●   000000000000000   000000000   00000   0000
Are we doing Big Data research?

Are we doing Big Data research in this course?

Depends on the definition

- Not if we take a definition that *only* focuses on computing power and the amount of data

Introducing. . .   **What is Big Data?**   Methods   What have others done?   What can we do?   The schedule
0000   0000000●   000000000000   000000000   00000   0000
Are we doing Big Data research?

Are we doing Big Data research in this course?

Depends on the definition

- Not if we take a definition that *only* focuses on computing power and the amount of data
- **But:** We are using the same techniques. And they *scale* well.

Introducing... | **What is Big Data?** | Methods | What have others done? | What can we do? | The schedule
0000 | 0000000● | 000000000000 | 000000000 | 00000 | 0000
Are we doing Big Data research?

Are we doing Big Data research in this course?

Depends on the definition

- Not if we take a definition that *only* focuses on computing power and the amount of data
- **But:** We are using the same techniques. And they *scale* well.
- Oh, and about that high-performance computing in the cloud: We actually *do* have access to that, so if someone has a really great idea...

Methods

What we will learn the next weeks

What we will learn the next weeks

1. How to collect data

APIs, scrapers and crawlers, feeds, databases, . . .
Storage in different file formats

Introducing...	What is Big Data?	**Methods**	What have others done?	What can we do?	The schedule
○○○○	○○○○○○○○	○●○○○○○○○○○○○○	○○○○○○○○○	○○○○○	○○○○

Which techniques?

What we will learn the next weeks

1. How to collect data

APIs, scrapers and crawlers, feeds, databases, . . .
Storage in different file formats

2. How to analyze data

Sentiment analysis, automated content analysis, regular
expressions, natural language processing, cluster analysis, machine
learning, network analysis

| Introducing. . . | What is Big Data? | **Methods** | What have others done? | What can we do? | The schedule |
| 0000 | 00000000 | 00●0000000000 | 000000000 | 00000 | 0000 |

Which techniques?

The ACA toolbox

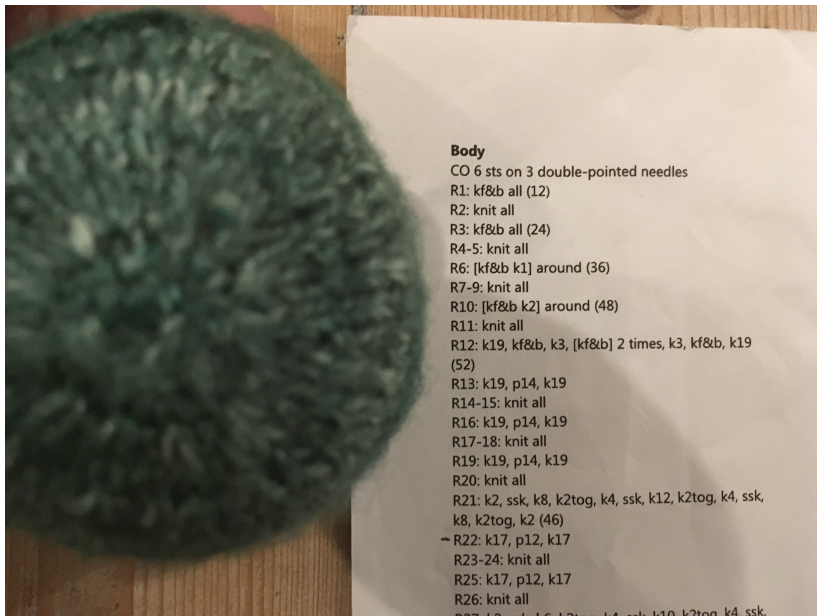|	**Methodological approach**		
	Counting and Dictionary	*Supervised Machine Learning*	*Unsupervised Machine Learning*
Typical research interests and content features	visibility analysis sentiment analysis subjectivity analysis	frames topics gender bias	frames topics
Common statistical procedures	string comparisons counting	support vector machines naive Bayes	principal component analysis cluster analysis latent dirichlet allocation semantic network analysis

deductive ⟶ **inductive**

Boumans, J.W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*, 1. 8–23.

Introducing...   What is Big Data?   **Methods**   What have others done?   What can we do?   The schedule

Which tools?

The methods

The tools we use for this

The programming language Python (and the huge amount of Python modules others already wrote).

Introducing. . .   What is Big Data?   **Methods**   What have others done?   What can we do?   The schedule
oooo   oooooooo   oooo○●oooooooo   ooooooooo   oooo○   ooooo
Which tools?

A language, not a program:
Python

Body
CO 6 sts on 3 double-pointed needles
R1: kf&b all (12)
R2: knit all
R3: kf&b all (24)
R4-5: knit all
R6: [kf&b k1] around (36)
R7-9: knit all
R10: [kf&b k2] around (48)
R11: knit all
R12: k19, kf&b, k3, [kf&b] 2 times, k3, kf&b, k19 (52)
R13: k19, p14, k19
R14-15: knit all
R16: k19, p14, k19
R17-18: knit all
R19: k19, p14, k19
R20: knit all
R21: k2, ssk, k8, k2tog, k4, ssk, k12, k2tog, k4, ssk, k8, k2tog, k2 (46)
➤ R22: k17, p12, k17
R23-24: knit all
R25: k17, p12, k17
R26: knit all
R27: k2, ssk, k6, k2tog, k4, ssk, k10, k2tog, k4, ssk,

An algorithm in a language that's a bit harder (I think) than Python

Introducing...   What is Big Data?   **Methods**   What have others done?   What can we do?   The schedule
oooo   oooooooo   ooooooo●oooooo   ooooooooo   ooooo   oooo
Which tools?

Python

What?

- A language, not a specific program
- Huge advantage: flexibility, portability
- One of *the* languages for data analysis. (The other one is R.)

Python

What?

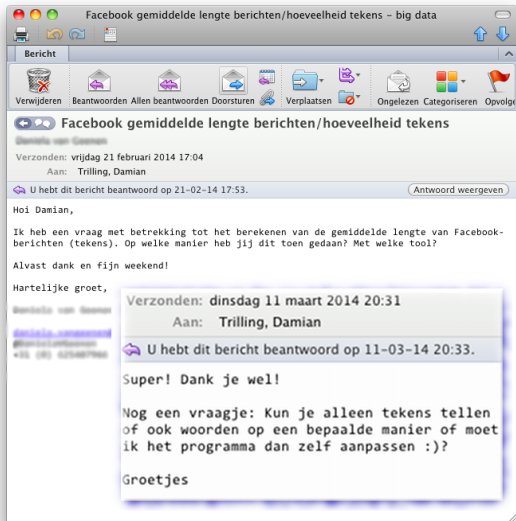- A language, not a specific program

- Huge advantage: flexibility, portability

- One of *the* languages for data analysis. (The other one is R.)
  But Python is more flexible—the original version of Dropbox was written in Python. Some people say: R for numbers, Python for text and messy stuff.

Introducing...	What is Big Data?	**Methods**	What have others done?	What can we do?	The schedule
○○○○	○○○○○○○○	○○○○○○○●○○○○○○	○○○○○○○○○	○○○○○	○○○○

Which tools?

Python

What?

- A language, not a specific program
- Huge advantage: flexibility, portability
- One of *the* languages for data analysis. (The other one is R.)

Which version?

We use Python 3.
`http://www.google.com` or `http://www.stackexchange.com` still offer a lot of Python2-code, but that can easily be adapted. Most notable difference: In Python 2, you write `print "Hi"`, this has changed to `print ("Hi")`

Why program your own tool?

Introducing...	What is Big Data?	Methods	What have others done?	What can we do?	The schedule
0000	00000000	000000000●0000	000000000	00000	0000

Which tools?

Why program your own tool?

If the task would have been done with a (commercial) tool, we can only research what the tool allows us to do ($\Rightarrow$ our discussion from some minutes ago).

| Introducing... | What is Big Data? | **Methods** | What have others done? | What can we do? | The schedule |
| 0000 | 00000000 | 0000000000●0000 | 000000000 | 00000 | 0000 |

Which tools?

Why program your own tool?

If the task would have been done with a (commercial) tool, we can only research what the tool allows us to do ($\Rightarrow$ our discussion from some minutes ago).

Luckily, the problem is easily solved

The task was done with a self-written Python program. We change the line

```
lengte_list.append(len(row[textcolumn]))
```

to

```
lengte_list.append(len(row[textcolumn].split()))
```

Why program your own tool?

```
Moreover, the tools we use can limit the range of
questions that might be imagined, simply because they
do not fit the affordances of the tool.  Not many
researchers themselves have the ability or access to
other researchers who can build the required tools in
line with any preferred enquiry.  This then
introduces serious limitations in terms of the scope
of research that can be done. Vis, 2013
```

Introducing...	What is Big Data?	Methods	What have others done?	What can we do?	The schedule
○○○○	○○○○○○○○	○○○○○○○○○○○●○○	○○○○○○○○○	○○○○○	○○○○

Which tools?

Why program your own tool?

Some considerations regarding the use of software in science

Assuming that science should be *transparent* and *reproducible by anyone*

Some considerations regarding the use of software in science

Assuming that science should be *transparent* and *reproducible by anyone*, we should

use tools that are

- platform-independent
- free (as in beer and as in speech, gratis and libre)
- which implies: open source

Introducing...	What is Big Data?	**Methods**	What have others done?	What can we do?	The schedule
0000	00000000	000000000000●0	000000000	00000	0000

Which tools?

Some considerations regarding the use of software in science

Assuming that science should be *transparent* and *reproducible by anyone*, we should

use tools that are

- platform-independent
- free (as in beer and as in speech, gratis and libre)
- which implies: open source

This ensures it can our research (a) can be reproduced by anyone, and that there is (b) no black box that no one can look inside. ⇒ ongoing open-science debate!

Introducing...   What is Big Data?   **Methods**   What have others done?   What can we do?   The schedule

Which tools?

Why program your own tool?

[...] these [commercial] tools are often unsuitable
for academic purposes because of their cost, along
with the problematic 'black box' nature of many of
these tools. Vis, 2013

[...] we should resist the temptation to let the
opportunities and constraints of an application or
platform determine the research question [...] Mahrt &
Scharkow, 2013, p. 30

What have others done?

Introducing. . .   What is Big Data?   Methods   **What have others done?**   What can we do?   The schedule
0000   00000000   0000000000000   0●0000000   00000   0000
Online news sharing

Online news sharing

Castillo, El-Haddad, Pfeffer, & Stempeck, 2013

Online news sharing

The question

"We describe the interplay between website visitation patterns and social media reactions to news content. [. . .] We also show that social media reactions can help predict future visitation patterns early and accurately." (p.1)

Castillo, El-Haddad, Pfeffer, & Stempeck, 2013

Introducing... | What is Big Data? | Methods | **What have others done?** | What can we do? | The schedule
0000 | 00000000 | 0000000000000 | 00●000000 | 00000 | 0000

Online news sharing

Online news sharing

The method

- data set 1: log files provided by Al Jazeera
- data set 2: Facebook and Twitter API
- analysis: link 1 and 2 and estimate the relationships

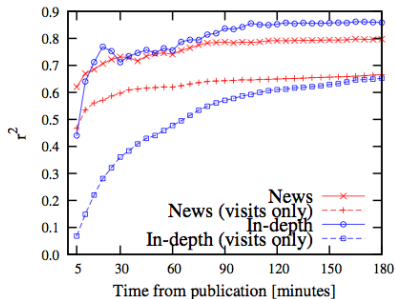Castillo, El-Haddad, Pfeffer, & Stempeck, 2013

Figure 7. Proportion of explained variance (r^2) for the prediction of total volume of visits, for News and In-depth articles.

Figure 8. Distribution of effective shelf-life.

It takes about 3 hours to be able to explain > 0.6 of the variance for In-Depth articles, and the additional variables are profitable from the first minutes. After 10-20 minutes we observe the largest difference in our regression models ($+0.5$ in terms of r^2).

We take a closer look at the model variables after 20 minutes to identify the sources of this improvement. For this purpose we stepwise fit the model variables by AIC (Akaike information criterion) as implemented in

Table 5. Modeling effective shelf-life: Significance levels for regression models after 20 minutes.

Variable	In-depth		News	
Visits R^2	0.0005		0.0921	
Social media R^2	0.4457		0.2193	
Social media R^2 adjusted	0.2274		0.1505	
Twitter tweets	0.0138	*	0.0061	**
Twitter entropy	0.0027	**	0.0024	**
Twitter avg. followers			0.0001	***
Volume of unique tweets	0.0026	**		
Unique tweets %	0.0190	*	0.0445	*
Corporate retweets	0.0001	***		
Traffic from e-mail/IM	0.0482	*		

Introducing... | What is Big Data? | Methods | What have others done? | What can we do? | The schedule
0000 | 00000000 | 0000000000000 | 0000●0000 | 00000 | 0000
Online news sharing

Online news sharing

The issues

- You need that guy at Al Jazeera
- You need the infrastructure to cope with the data
- Very much tailored to one outlet

Castillo, El-Haddad, Pfeffer, & Stempeck, 2013

Introducing. . .   What is Big Data?   Methods   What have others done?   What can we do?   The schedule
oooo   ooooooooo   oooooooooooo ooooo●ooo   ooooo   oooo
Partisan asymmetries

Partisan asymmetries

Conover, Gonçalves, Flammini, & Menczer, 2012

Partisan asymmetries

The question

How does the Twitter behavior differ between right-wing and left-wing users?

Conover, Gonçalves, Flammini, & Menczer, 2012

Partisan asymmetries

The method

Starting with two hashtags (one used by progressives, one used by conservatives), 55 co-occuring hashtags were identified. Identification of follower networks, retweet networks, mention networks within tweets with these hashtags.
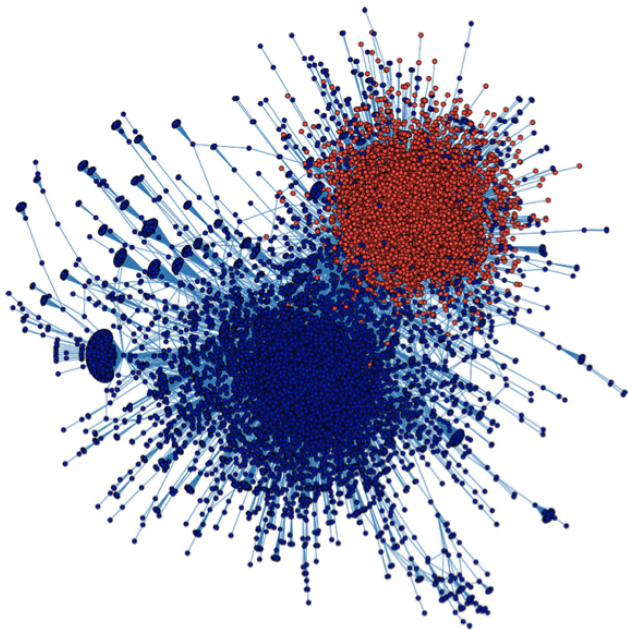
Conover, Gonçalves, Flammini, & Menczer, 2012

Figure 3 The network of political retweets, laid out using a force-directed algorithm. Node colors reflect cluster assignments, which correspond to politically homogeneous communities of left- and right-leaning users with 87% accuracy. (See Section 3.3.)

Partisan asymmetries

The issues

- The two seeds #tcot and #p2 oversample (extremly) partisan content, inevitably leading to the structure in Figure 3.

- But maybe no problem, as the rest of the paper aims to *compare* these groups.

- Not an empirical problem, but still: What do we learn *exactly* from this study apart from the – interesting – case?

Conover, Gonçalves, Flammini, & Menczer, 2012

Introducing. . .   What is Big Data?   Methods   What have others done?   What can we do?   The schedule
oooo   oooooooo   oooooooooooooo   ●oooo   oooo
Considerations regarding feasibility

To which extent could *we* conduct such studies?

Introducing. . .   What is Big Data?   Methods   What have others done?   **What can we do?**   The schedule
0000   00000000   0000000000000 000000000   0●000   0000
Considerations regarding feasibility

What can we do?

Cool research, sure, but what can *we* do?

- Dependency from third parties:
  scraping < API < server-side implementation
- Restrictions (e.g., Twitter: sprinkler, garden hose, fire hose)
  ⇒ Vis, 2013: Data are made!
- We can't just trust the numbers. Some tasks require human
  coders – or a qualitative approach, at least as a pre-study.

Introducing... | What is Big Data? | Methods | What have others done? | **What can we do?** | The schedule
0000 | 00000000 | 000000000000 | 000000000 | 00●00 | 0000
Considerations regarding feasibility

What can we do?

pros, cons, and feasibility

- APIs (Twitter, Facebook, ...)
- server-side implementations ($\Rightarrow$ Al Jazeera-example)
- scraping
- client-side log files
- "traditional" methods (surveys etc.)

Introducing...   What is Big Data?   Methods   What have others done?   What can we do?   The schedule
0000   00000000   000000000000000   000000000   000●0   0000
Considerations regarding feasibility

What can we do?

What helps us answer *our* questions?

1. Draft a RQ.
2. Think of *different* ways to approach it
3. Think of *different* data sources.
4. Think of *different* analyses.

Introducing. . .   What is Big Data?   Methods   What have others done?   **What can we do?**   The schedule
0000   00000000   000000000000000   000000000   000●0   0000
Considerations regarding feasibility

What can we do?

What helps us answer *our* questions?

❶ Draft a RQ.

❷ Think of *different* ways to approach it

❸ Think of *different* data sources.

❹ Think of *different* analyses.

And then:

❶ Check what the technical possibilities are. (e.g., is there an API? How can we get the data?)

❷ Re-evaluate all steps.

Introducing. . .   What is Big Data?   Methods   What have others done?   **What can we do?**   The schedule
0000   00000000   0000000000000   000000000   0000●   0000
Examples from last year

Some final projects from previous courses

Questions

- How can house prices on funda.nl be predicted?

- Where are those who edit Wikipedia-entries about companies geographically located, related to the company's HQ?

- Can we predict ratings on Hostelworld?

- What do people write in their Tinder profiles – and how do men and women differ?

- How international is the ICA, given all presentations given at all conferences in the last decade?

The schedule

The schedule

Each week

In general: A lecture (Wednesday) and a lab session (Friday).
Each week one method.

Examinations

A mid-term take-home exam in week 5 and an individual research
project on which you work during the whole course.

The schedule

Each week

In general: A lecture (Wednesday) and a lab session (Friday).
Each week one method.

Examinations

A mid-term take-home exam in week 5 and an individual research
project on which you work during the whole course.

Self-study

Play around! You really have to *do* it to learn programming. See it
as your weekly assignment ;-)

Introducing... | What is Big Data? | Methods | What have others done? | What can we do? | The schedule
0000 | 00000000 | 0000000000000 | 000000000 | 00000 | 00●0

Next meetings

Next meetings

Week 1: Introduction

Friday, 7–4

No meeting due to *Etmaal van de Communicatiewetenschap.*
Your task: Read chapters 2 and 3!
Test everything in your VM!
And, have a look at the alternative (new) chapter on Canvas
(Computational Analysis of Communication).