Big Data & Automated Content Analysis
Week 1 – Wednesday: »Introduction«

Damian Trilling

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

3 February 2021

Afdeling Communicatiewetenschap
Universiteit van Amsterdam

Today

Getting to know each other

Setting the stage

- Defining "Big Data"
- Defining Computational (Social|Communication) Science

The toolbox

- The role of software in CSS
- Python: A language, not a program

Damian

dr. Damian Trilling

Universitair Hoofddocent (Associate Professor)

Communication in the Digital Society

- studied Communication Science in Münster and at the VU 2003–2009
- PhD candidate @ ASCoR 2009–2012
- political communication and journalism in a changing media environment
- computational research methods

@damian0604 d.c.trilling@uva.nl

REC-C 8th floor www.damiantrilling.net

Vlad

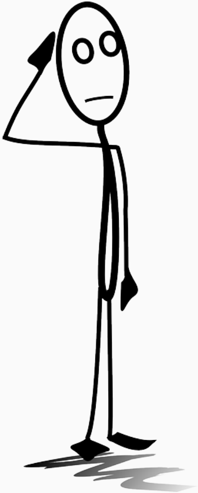
Vladislav Petkevic

Junior Lecturer Communication Science

- MSc in Communication Research (2020)
- Interested in politicial communication, especially election campaigns
- Even more interested in applying computational research methods (e.g. NLP, maschine vision) to studying social phenomena

v.petkevich@uva.nl

You



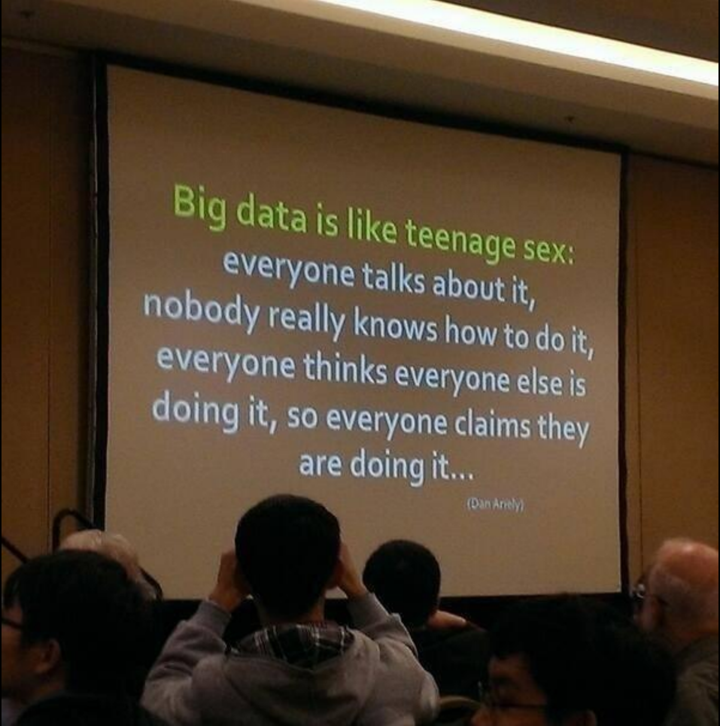
Your name?

Your background?

Your reason to follow this course?

Setting the stage

Defining “Big Data”


Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...
(Dan Ariely)

The “pragmatic” definition

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

The "commercial" definition

Gartner (n.d.)

"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

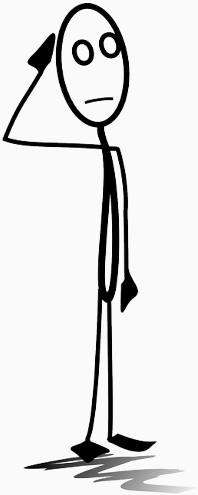
The "critical" definition

Boyd and Crawford (2012)

"

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.

"



Do you think we are doing Big Data analysis?

Setting the stage

Defining Computational (Social|Communication) Science

A very young field

Lazer et al. (2009)

"The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven 'computational social science' has been much slower."

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

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

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

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

CCS as a subset of CSS

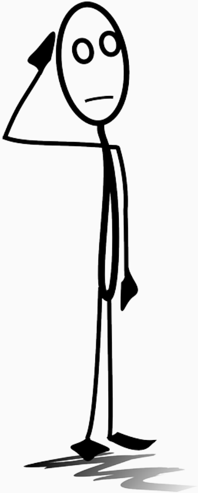
Hilbert et al. (2019)

"…our definition of computational communication science as an application of computational science to questions of human and social communication. As such, it is a natural subfield of computational social science" (followed by references to CSS definitions)

Data, analysis, theory

van Atteveldt and Peng (2018)

"...computational communication science studies generally involve: (1) large and complex data sets; (2) consisting of digital traces and other "naturally occurring" data; (3) requiring algorithmic solutions to analyze; and (4) allowing the study of human communication by applying and testing communication theory."



1. *What do you think? What is the essence of Big Data/CSS/CCS?*
2. *How will what we do here relate to theories and methods from other courses?*

The toolbox

The role of software in CSS

Why program your own tool?

Vis (2013)

"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."

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! (van Atteveldt et al., 2019)

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! (van Atteveldt et al., 2019)

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. $\Rightarrow$ ongoing open-science debate! (van Atteveldt et al., 2019)

Why program your own tool?

Vis (2013)

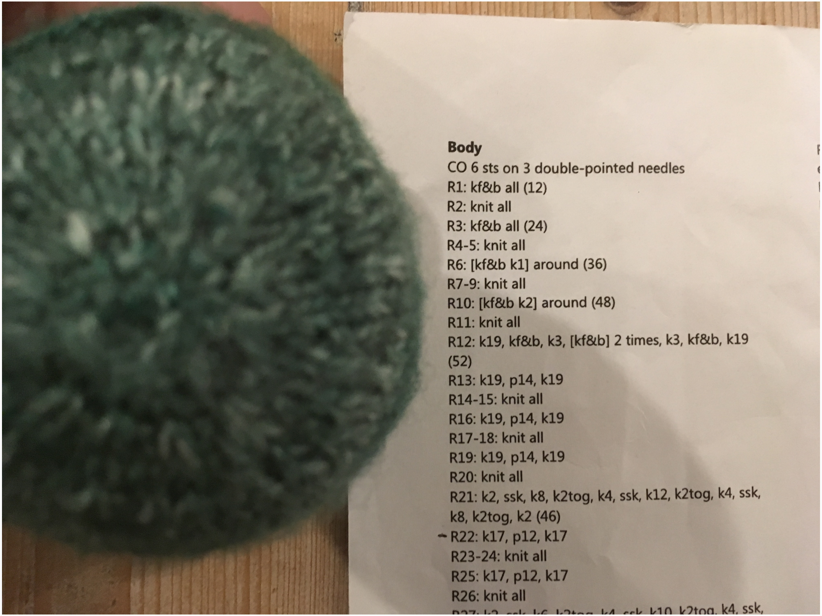
"[. . .] these [commercial] tools are often unsuitable for academic purposes because of their cost, along with the problematic 'black box' nature of many of these tools."

Mahrt and Scharkow (2013)

"[. . .] we should resist the temptation to let the opportunities and constraints of an application or platform determine the research question [. . .]"

The toolbox

Python: A language, not a program



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

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.

Which version?

We use Python 3.

http://www.google.com or http://www.stackexchange.com still may show you some Python2-code, but that can easily be adapted. Most notable difference: In Python 2, you write `print "Hi"`, this has changed to `print ("Hi")`.

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.

Which version?

We use Python 3.

http://www.google.com or http://www.stackexchange.com still may show you some Python2-code, but that can easily be adapted. Most notable difference: In Python 2, you write `print "Hi"`, this has changed to `print ("Hi")`.

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.

Which version?

We use Python 3.

http://www.google.com or http://www.stackexchange.com still may show you some Python2-code, but that can easily be adapted. Most notable difference: In Python 2, you write `print "Hi"`, this has changed to `print ("Hi")`.

Make sure you can run Python code and install packages. Otherwise, you won't be able to follow along on Friday. (See instructions you got. Use Vlad's office hours if you cannot figure it out.))

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Gartner. (n.d.). Big data. https://www.gartner.com/en/information-technology/glossary/big-data

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., González-Bailón, S., Lamberso, P., Pan, J., Peng, T.-Q., Shen, C., Smaldino, P. E., Van Atteveldt, W., Waldherr, A., Zhang, J., & Zhu, J. J. H. (2019). Computational Communication Science : A Methodological Catalyzer for a Maturing Discipline. *International Journal of Communication*, *13*, 3912–3934.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & van Alstyne, M. (2009). Computational social science. *Science*, *323*, 721–723. https://doi.org/10.1126/science.1167742

Mahrt, M., & Scharkow, M. (2013). The value of Big Data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33. https://doi.org/10.1080/08838151.2012.761700

van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Toward Open Computational Communication Science : A Practical Road Map for Reusable Data and Code University of Amsterdam , the Netherlands. *International Journal of Communication*, *13*, 3935–3954.

van Atteveldt, W., & Peng, T. Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10), 1–16. https://doi.org/10.5210/fm.v18i10.4878