

# Automated Web Data Collection with R

Theresa Gessler | Hauke Licht  
University of Zurich

# Introduction

## Plan of the session

- What: overview of the course
- Who: Introduction
- Why: Web Data
- HTML

# Introduction

## How this course works

- **learning by doing**
  - morning: mostly lecture to introduce techniques, guided exercise
  - afternoon: mostly independent exercises
  - (some deviations on some days)

## Rules

- **ask and answer**
  - discussions are better for every question you ask and every idea you share
- **respect and patience**
  - digital teaching and learning is difficult for all of us
- **please turn your camera on:** feedback is crucial to teaching
  - feel free to blur your background
  - use a virtual background
  - ...but please do turn your camera on

# Introduction

## Course content

**scraping:** /'skreɪpɪŋ/, *to remove (an outer layer, for example) from a surface by forceful strokes of an edged or rough instrument*

**web scraping:** to collect data from the web by removing the unnecessary parts (sometimes with a rough instrument)

→ family of different techniques

# Introduction

## Day 1

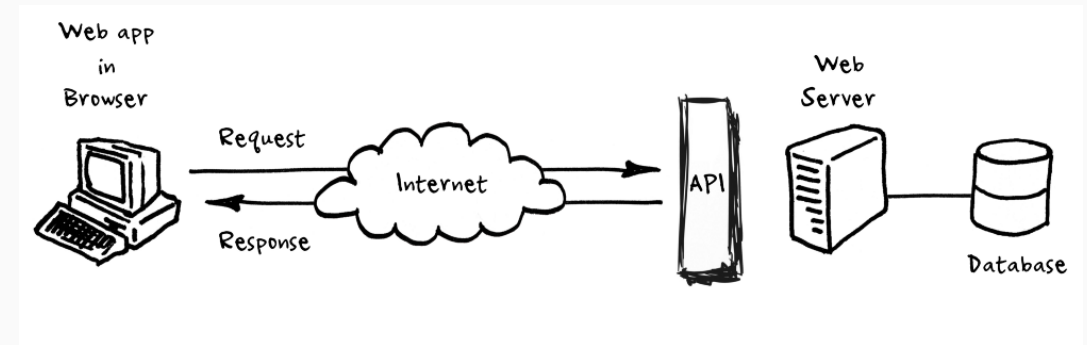
- course intro (morning)
- introduction to HTML (morning)
- downloading files (afternoon)
- some R practice (afternoon)

```
<div class="navbar navbar-default navbar-fixed-top" ro
<div class="container">
  <div class="navbar-header">
    <button type="button" class="navbar-toggle collap
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
    </button>
    <a class="navbar-brand" href="index.html">EUI Com
  </div>
  <div id="navbar" class="navbar-collapse collapse">
    <ul class="nav navbar-nav">
      <li>
        <a href="index.html">Overview</a>
      </li>
      <li>
        <a href="readings.html">Readings</a>
      </li>
      <li>
        <a href="code.html">Code & Slides</a>
      </li>
```

# Introduction

## Day 2

- Introduction to APIs (morning)
  - data provided by companies and organizations
  - interface built for data collection
- obtaining data from APIs with R (morning)
- social media research (afternoon)
- exercises: Wikipedia data (morning), Twitter (afternoon)



# Introduction

## Day 3

- Introduction to HTML
- using rvest to scrape HTML pages
- selecting parts of pages: CSS selectors
- exercise: presidential speeches

```
<div class="navbar navbar-default navbar-fixed-top" ro
<div class="container">
  <div class="navbar-header">
    <button type="button" class="navbar-toggle collap
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
    </button>
    <a class="navbar-brand" href="index.html">EUI Com
  </div>
  <div id="navbar" class="navbar-collapse collapse">
    <ul class="nav navbar-nav">
      <li>
        <a href="index.html">Overview</a>
      </li>
      <li>
        <a href="readings.html">Readings</a>
      </li>
      <li>
        <a href="code.html">Code & Slides</a>
      </li>
```

# Introduction

## Day 4

- scraping dynamic pages with RSelenium
  - simulating browsing behavior through code
- studying personalization with digital data
  - 'algorithmic auditing'



Image: Algoright: Auditing algorithms for bias



# Introduction

## Day 5

- how to scale up scraping projects
  - useful tools (e.g. parsing)
  - task scheduling
- ethics
- practice: working on your own projects
  - do think about them during the week!

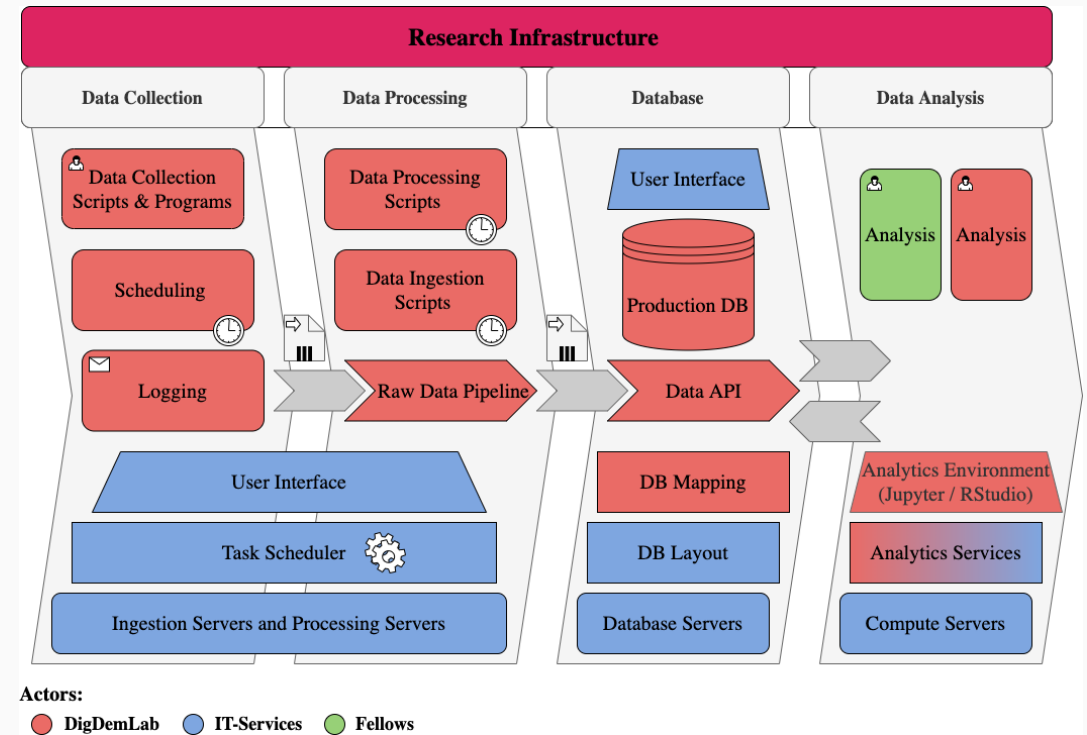


Image: DigDemLab

# Who

# Who

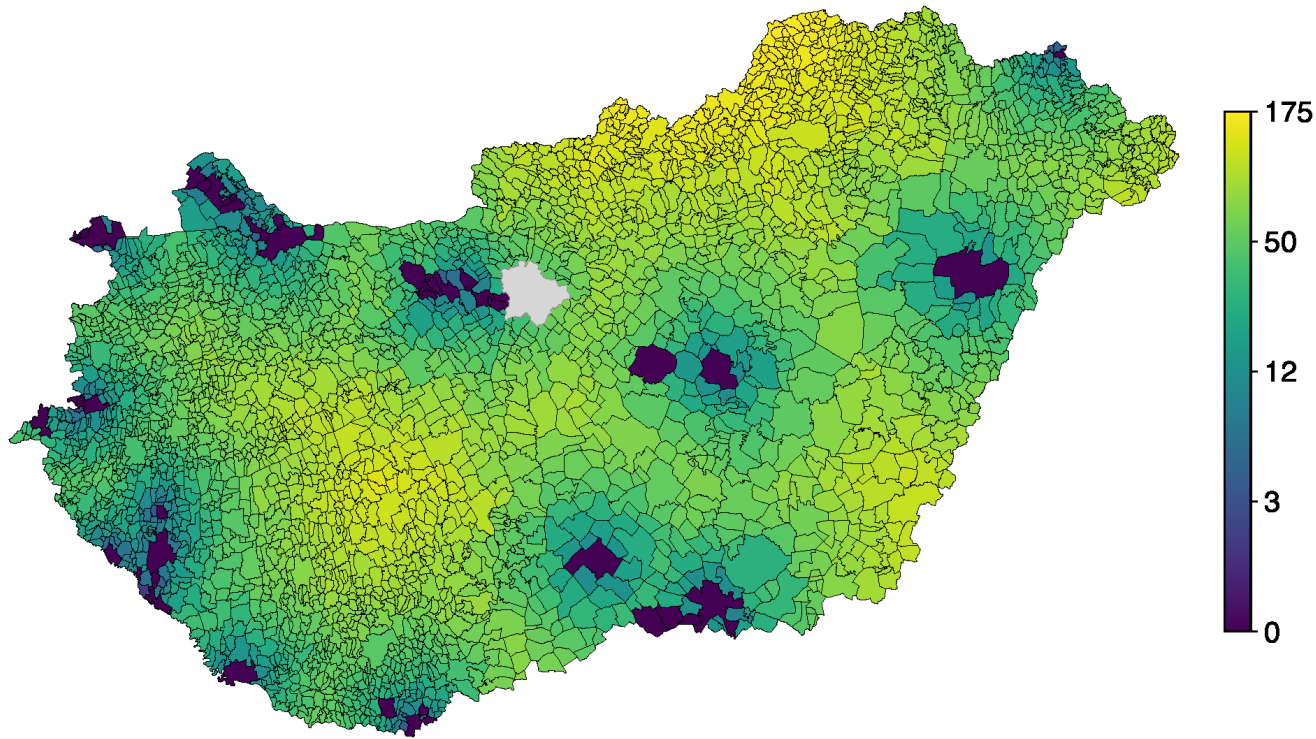
## Dr. Theresa Gessler



- **Postdoc** at the [Digital Democracy Lab](#) / Department of Political Science of the University of Zurich
- **reach me**
  - [gessler@ipz.uzh.ch](mailto:gessler@ipz.uzh.ch)
  - [www.theresagessler.eu](http://www.theresagessler.eu) | [@th\\_ges](#)
- I teach web data, text analysis, data journalism and substantive courses
- my research: **immigration** | **political parties** | **(digital) democracy** | **gender**

## Research: Immigration

Kilometers to Nearest Refugee Contact Settlement



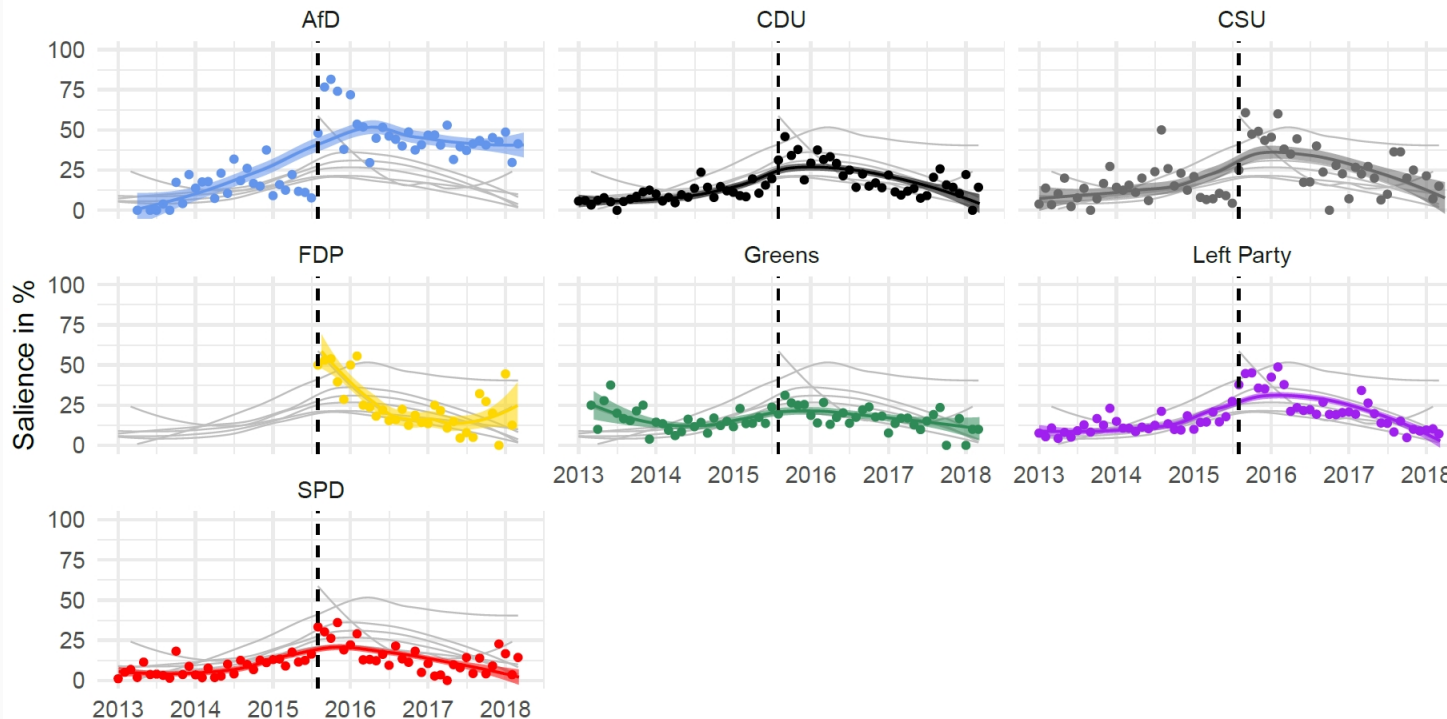
- identifying routes of refugees to measure political behavior

Gessler, T., Tóth, G., & Wachs, J. (2021). No country for asylum seekers? How short-term exposure to refugees influences attitudes and voting behavior in Hungary. *Political Behavior*.

# Who

## Research: Political Parties

### B Salience of immigration in Germany




- collecting press releases to study the salience of immigration

Gessler T., Hunger S. Nothing attracts a crowd as quickly as a crisis? How the refugee crisis and radical right parties shape party competition on immigration. Forthcoming.

## Research: Digital democracy / Gender

### Angela Merkel

 *Merkel ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter [Merkel \(Begriffsklärung\)](#) aufgeführt.*

**Angela**<sup>[1]</sup> **Dorothea Merkel** (\* 17. Juli 1954 in Hamburg als *Angela Dorothea Käsner*) ist eine deutsche Politikerin (CDU). Sie ist seit dem 22. November 2005 Bundeskanzlerin der Bundesrepublik Deutschland. Vom 10. April 2000 bis zum 7. Dezember 2018 war sie CDU-Bundesvorsitzende. Im Oktober 2018 erklärte sie, sich spätestens mit Ablauf der Legislaturperiode 2021 aus der Politik zurückzuziehen.

Merkel wuchs in der DDR auf und war dort als Physikerin am Zentralinstitut für Physikalische Chemie tätig. Bei der Bundestagswahl am 2. Dezember 1990 errang sie erstmals ein Bundestagsmandat. Bei den folgenden sieben Bundestagswahlen wurde sie in ihrem Wahlkreis in Vorpommern direkt gewählt.<sup>[2]</sup> Von 1991 bis 1994 war Merkel Bundesministerin für Frauen und Jugend im Kabinett Kohl IV und von 1994 bis 1998 Bundesministerin für Umwelt, Naturschutz und Reaktorsicherheit im Kabinett Kohl V. 1998 bis zu ihrer Wahl zur Bundesvorsitzenden der Partei amtierte sie als Generalsekretärin der CDU.

Nach dem knappen Sieg der Unionsparteien bei der vorgezogenen Bundestagswahl 2005 löste Merkel Gerhard Schröder als Bundeskanzler ab und führte zunächst eine große Koalition mit der SPD bis 2009 (Kabinett Merkel I). Nach der Bundestagswahl 2009 ging sie mit der FDP eine schwarz-gelbe Koalition ein (Kabinett Merkel II), der 2013 eine erneute große Koalition folgte, die auch nach der Bundestagswahl 2017 fortgesetzt wird (Kabinett Merkel III und IV).

#### Inhaltsverzeichnis [Verbergen]

##### 1 Leben

- 1.1 Elternhaus und frühe Kindheit (1954–1960)
- 1.2 Schulzeit und Studium (1961–1978)
- 1.3 Akademie der Wissenschaften in Ost-Berlin (1978–1989)
- 1.4 Familie
- 1.5 Freizeit

##### 2 Politische Laufbahn

- 2.1 Demokratischer Aufbruch (1989–1990)
- 2.2 Allianz für Deutschland (1990)
- 2.3 Beitritt zur CDU (1990)
- 2.4 Bundesministerin für Frauen und Jugend (1991–1994)
- 2.5 Bundesumweltministerin (1994–1998)
- 2.6 CDU-Generalsekretärin (1998–2000)
- 2.7 CDU-Vorsitzende (2000 bis 2018)
- 2.8 Oppositionsführerin (2002–2005)
  - 2.8.1 2002
  - 2.8.2 2003
  - 2.8.3 2004
  - 2.8.4 Vorgezogene Bundestagswahl 2005
- 2.9 Bundeskanzlerin (seit 2005)
  - 2.9.1 Große Koalition 2005 bis 2009
    - 2.9.1.1 Koalitionsverhandlungen



Angela Merkel (2019)

- collecting clickstream data and Wikipedia page content to study information-seeking

Gessler T. But is she married? Gender Bias and Users' Gendered Interest in Politicians on Wikipedia. Manuscript.

# Who

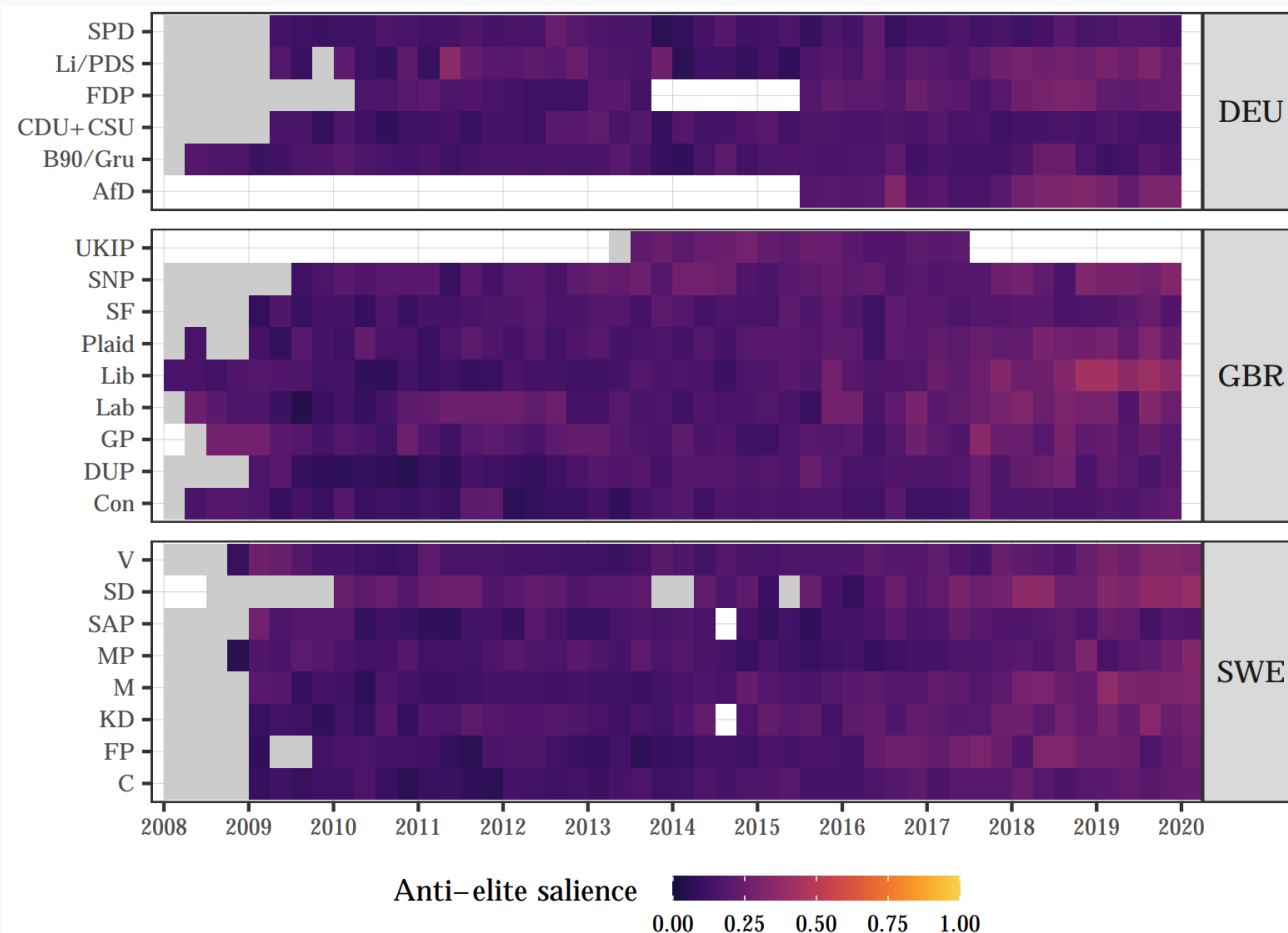
## Hauke Licht



- **PhD student** at the Department of Political Science of the University of Zurich
- **reach me**
  - [hauke.licht@uzh.ch](mailto:hauke.licht@uzh.ch)
  - [hauke\\_licht](#)
- my research: **text-as-data methodology** | **party competition** | **political representation**

# Who

## Research: Anti-elite rhetoric in multiparty competition



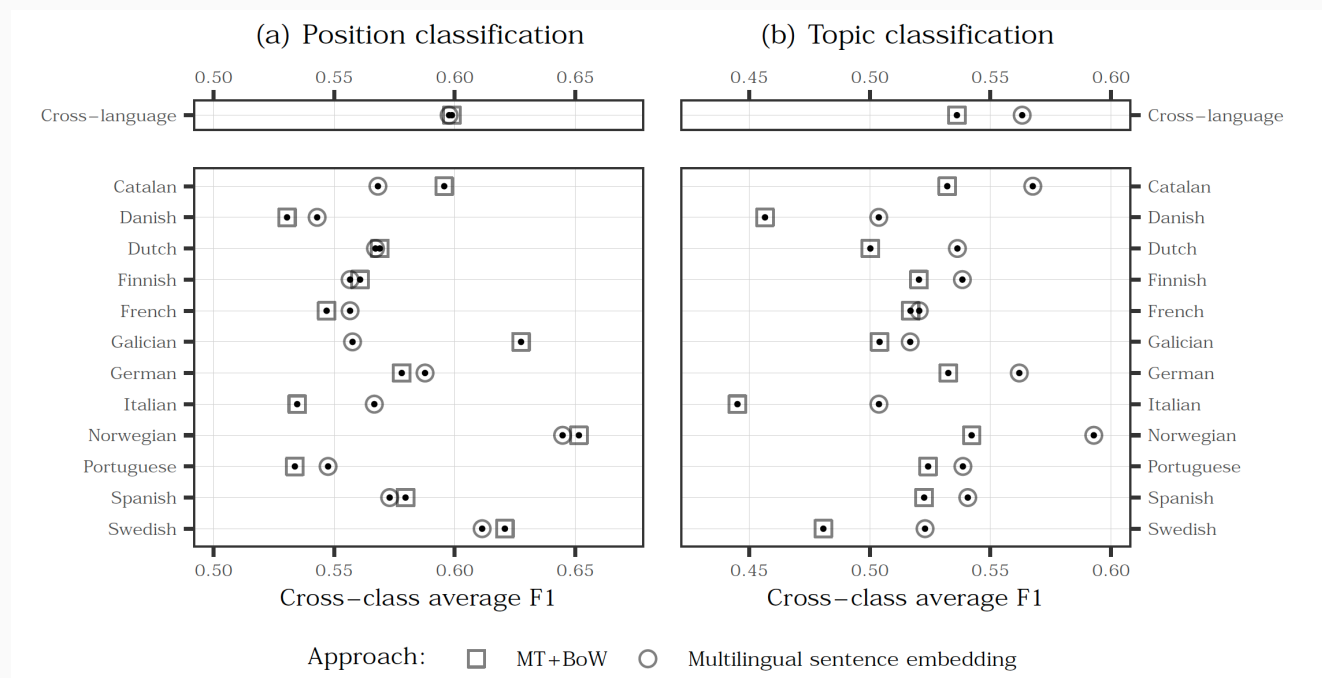
Quantifying how prevalent anti-elite rhetoric is in the electoral strategies of European parties by applying crowd coding and supervised machine learning to their Tweets

**"Measuring political rhetoric at scale using social media text"** (manuscript) with Tarik Abou-Chadi, Pablo Barberá and Whitney Hua



# Who

## Research: Cross-lingual text classification using multilingual text embeddings



Evaluate multilingual sentence embeddings methods for political text classification

**"Cross-lingual classification of political texts using multilingual sentence embeddings"** (under review)

# Who

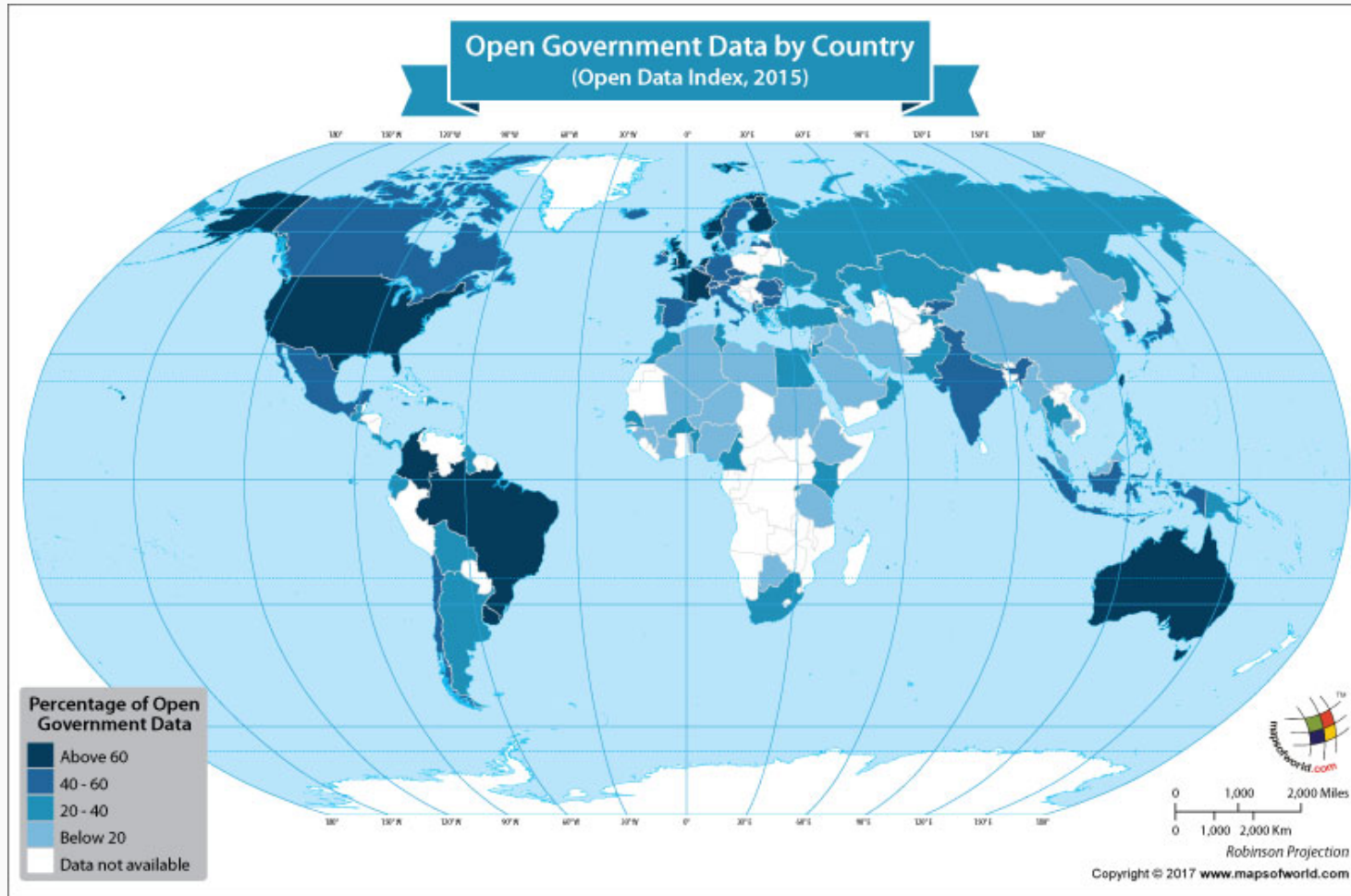
## Your turn!



- in 5 minute breakout rooms
  - why are you taking this course?
  - what research topics are you generally interested in?
  - what could you help others with?
- introduce each other afterwards

# Why Use Web Data?

# Why Use Web Data?



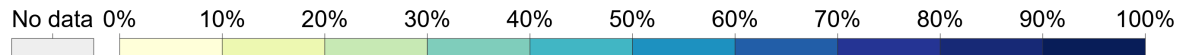
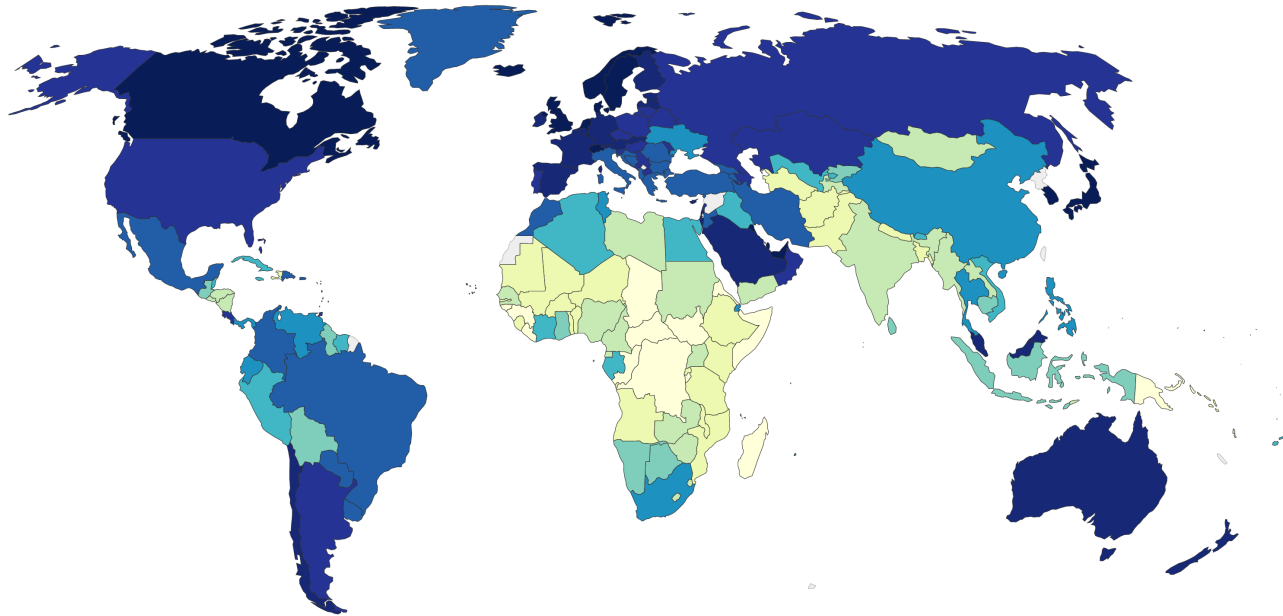
- increasing amount of public data online ('open government')

# Why Use Web Data?

## Share of the population using the Internet, 2017

All individuals who have used the Internet in the last 3 months are counted as Internet users. The Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc.

Our World  
in Data



Source: World Bank

[OurWorldInData.org/technology-adoption/](https://OurWorldInData.org/technology-adoption/) • CC BY

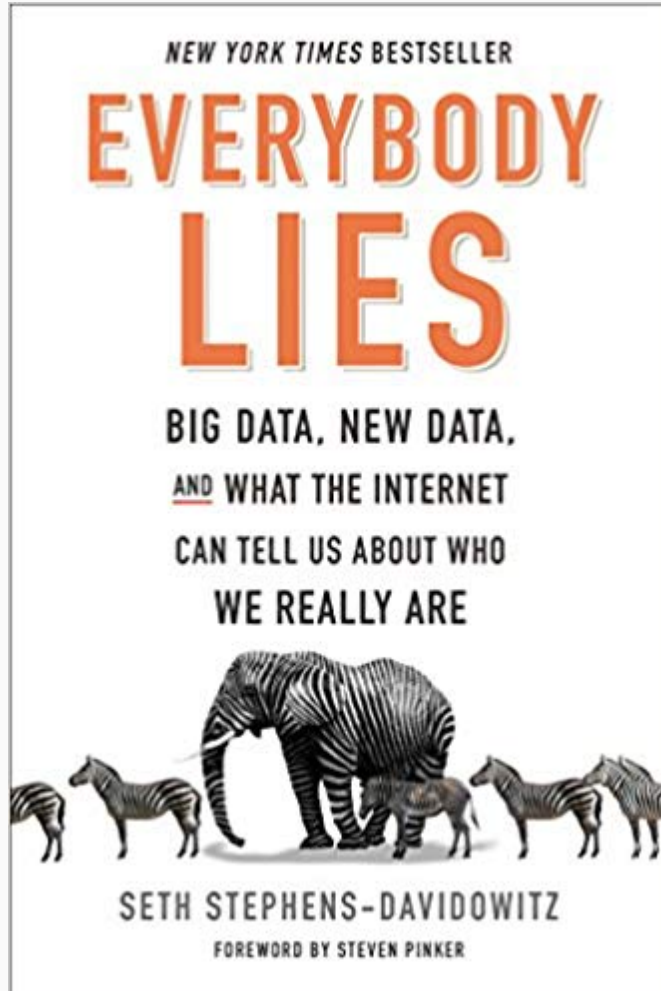
- increasing amount of people use the internet

# Why Use Web Data?



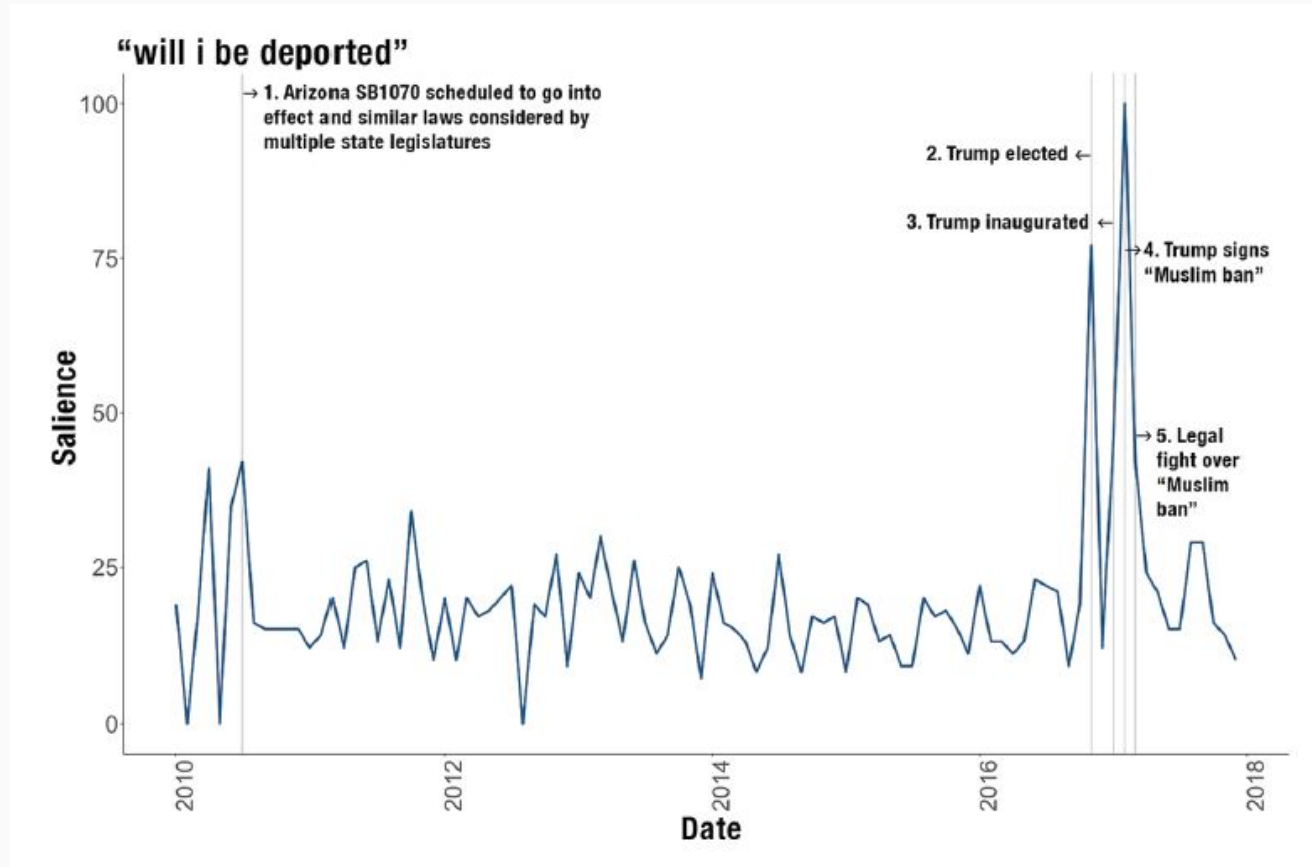
- increasing amount of politics happens online

# Why Use Web Data?



- we share everything online

# Why Use Web Data?



- that makes real world phenomena more visible online

Image: Chykina, Volha, and Charles Crabtree. "Using Google Trends to Measure Issue Salience for Hard-to-Survey Populations." *Socius* 4 (January 1, 2018):

2378023118760414. <https://doi.org/10.1177/2378023118760414>.



# What makes Web Data?

# What makes Web Data?

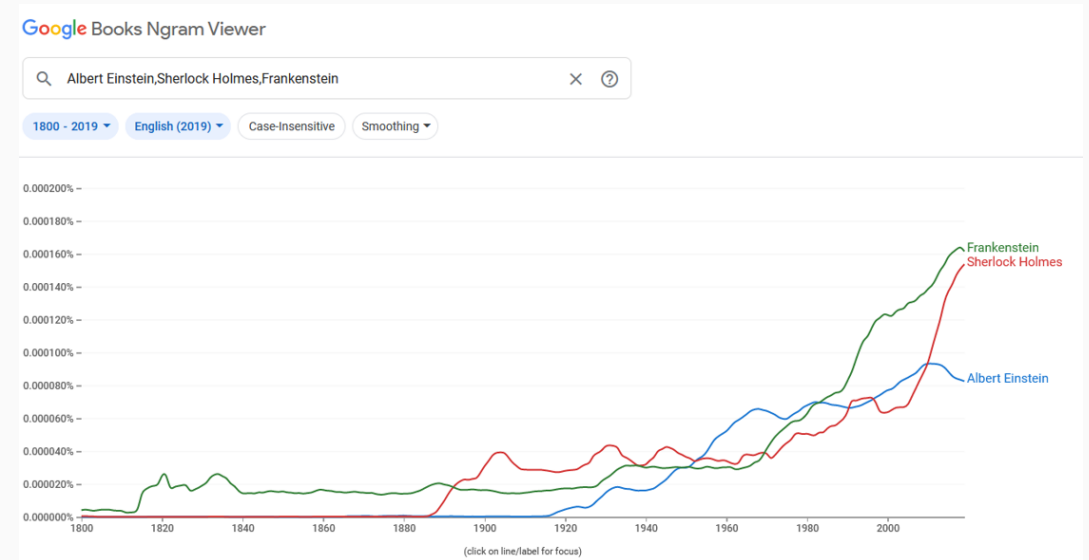
- always: **data collection from online sources**
- sometimes: different characteristics
  - traditional data made available online
  - 'digital trace data'

→ **10 characteristics of big data**, following Salganik (2018): 17-41

→ some characteristics are helpful whereas others are harmful for analysis

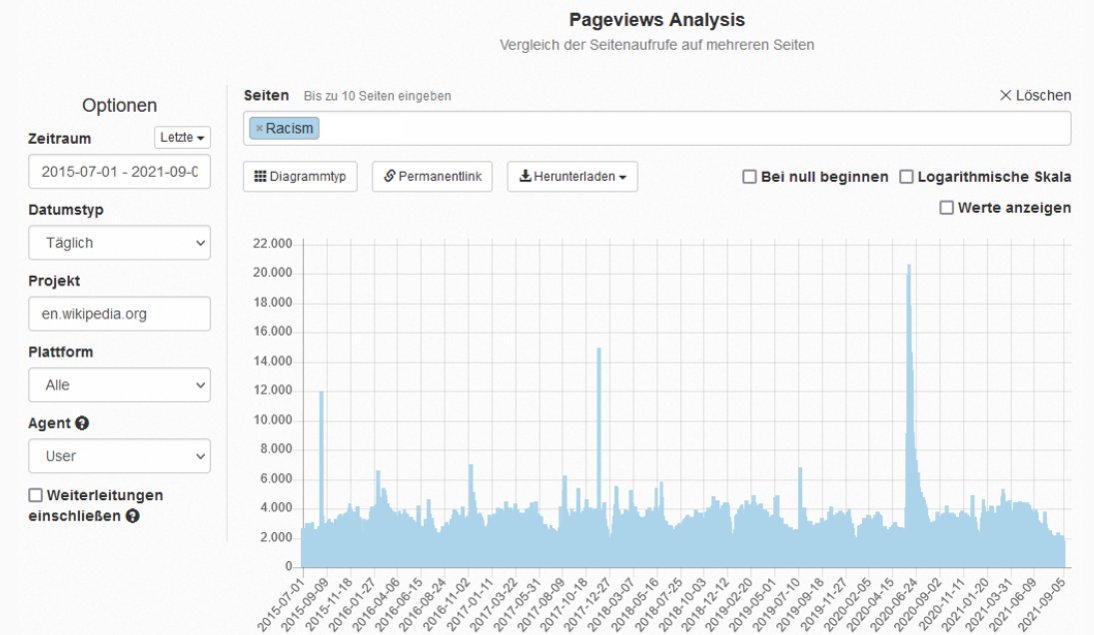
# What makes Web Data?

- Big



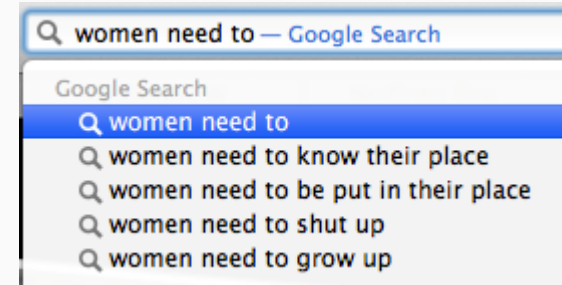
# What makes Web Data?

- Big
- Always-on



# What makes Web Data?

- Big
- Always-on
- Nonreactive



# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete

**Face editing: Japanese biker tricks internet into thinking he is a young woman**

Story: <https://www.bbc.com/news/world-asia-56447357>

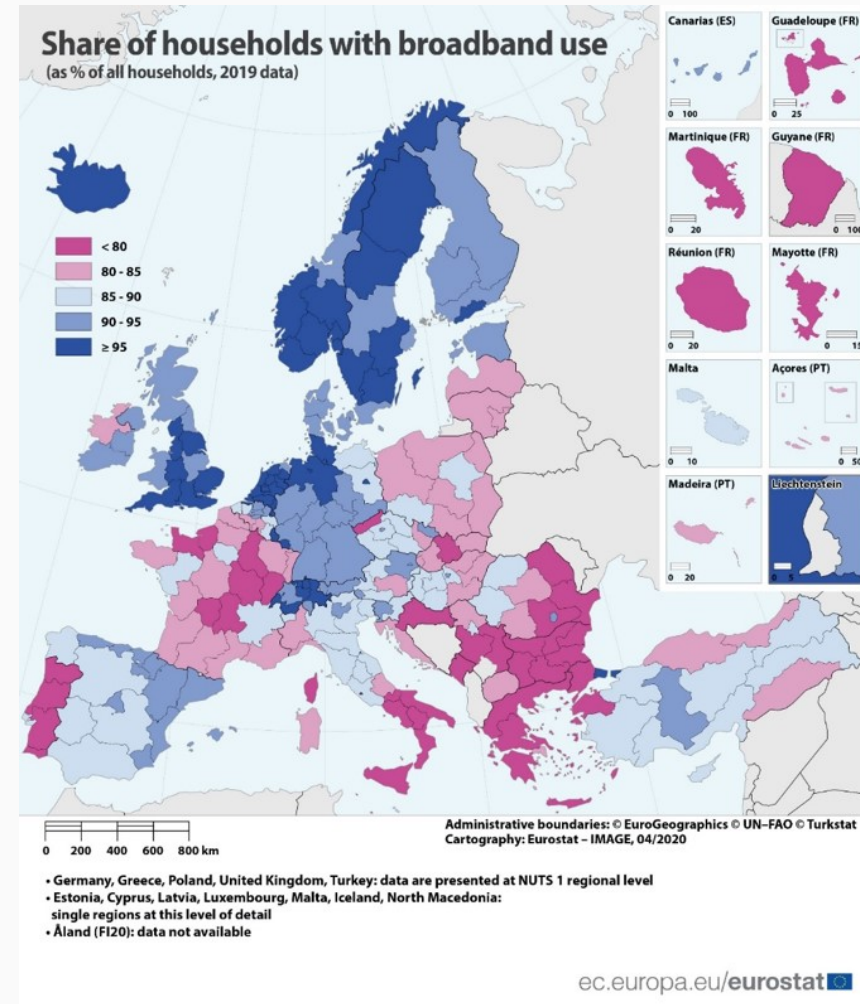
# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible



# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative





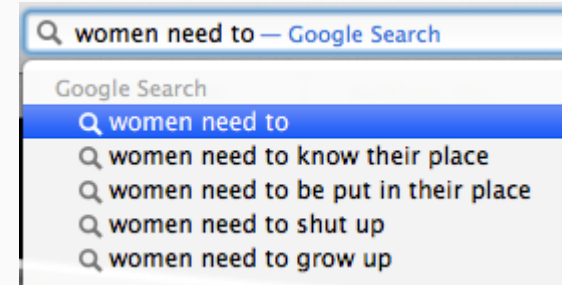
# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting



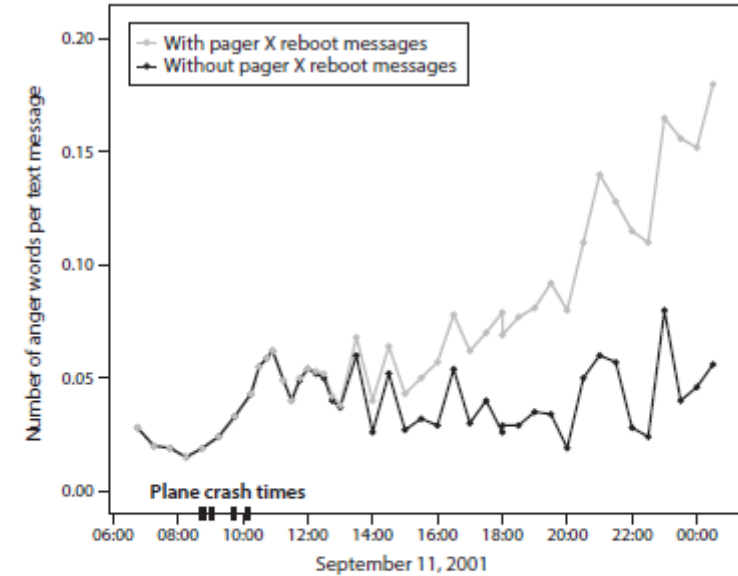
# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically Confounded



# What makes Web Data?

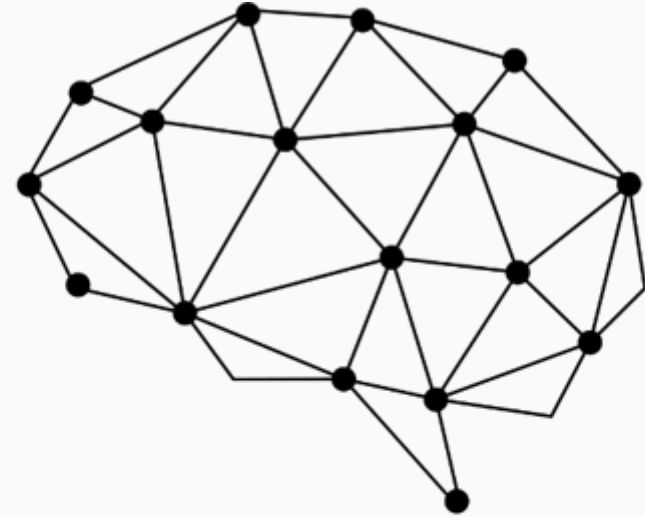
- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically Confounded
- Dirty



**Figure 2.4:** Estimated trends in anger over the course of September 11, 2001 based on 85,000 American pagers (Back, Küfner, and Egloff 2010; Pury 2011; Back, Küfner, and Egloff 2011). Originally, Back, Küfner, and Egloff (2010) reported a pattern of increasing anger throughout the day. However, most of these apparently angry messages were generated by a single pager that repeatedly sent out the following message: "Reboot NT machine [name] in cabinet [name] at [location]:CRITICAL:[date and time]". With this message removed, the apparent increase in anger disappears (Pury 2011; Back, Küfner, and Egloff 2011). Adapted from Pury (2011), figure 1b.

# What makes Web Data?

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically Confounded
- Dirty
- Sensitive



# Cambridge Analytica

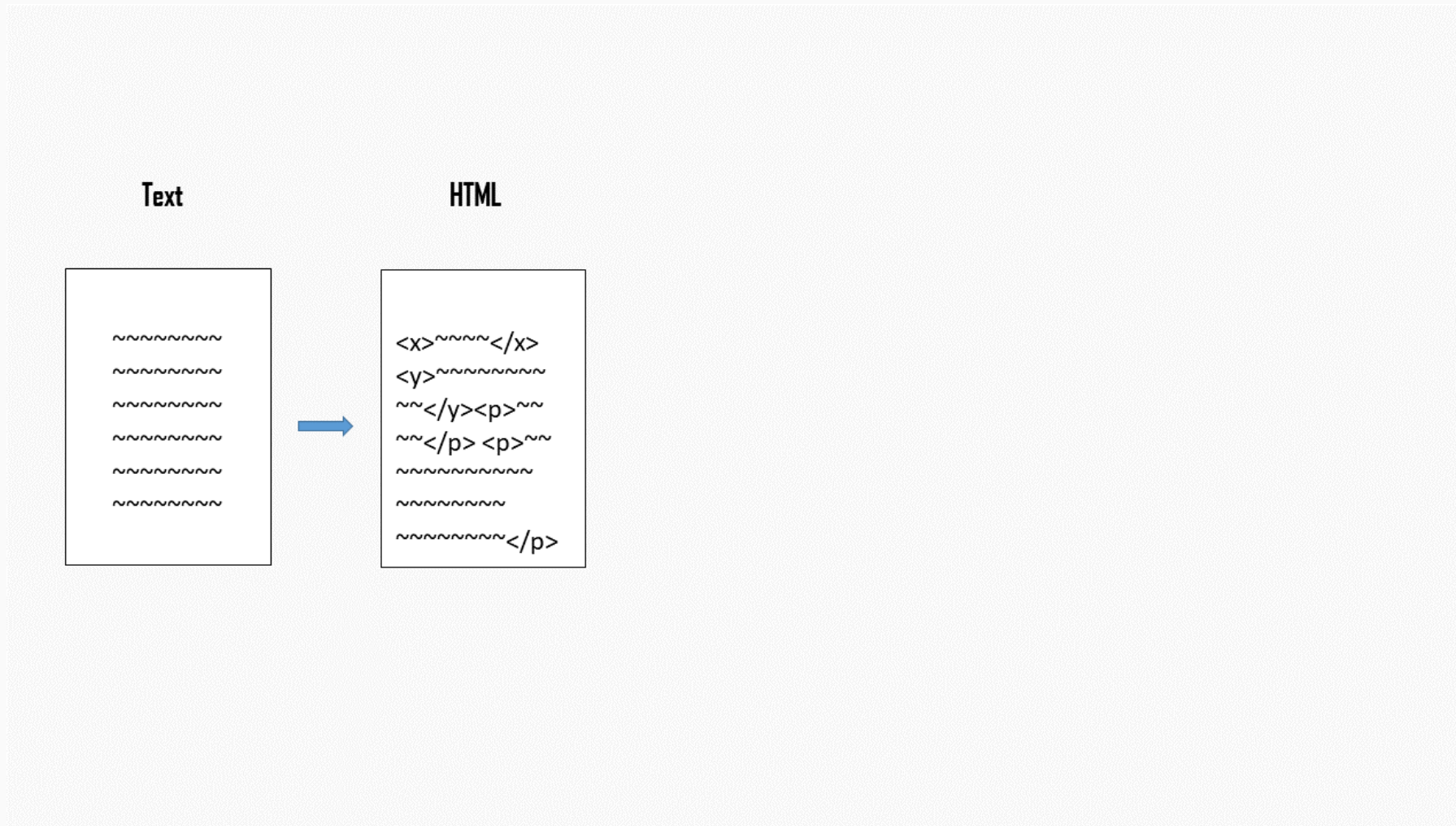
# HTML

# HTML

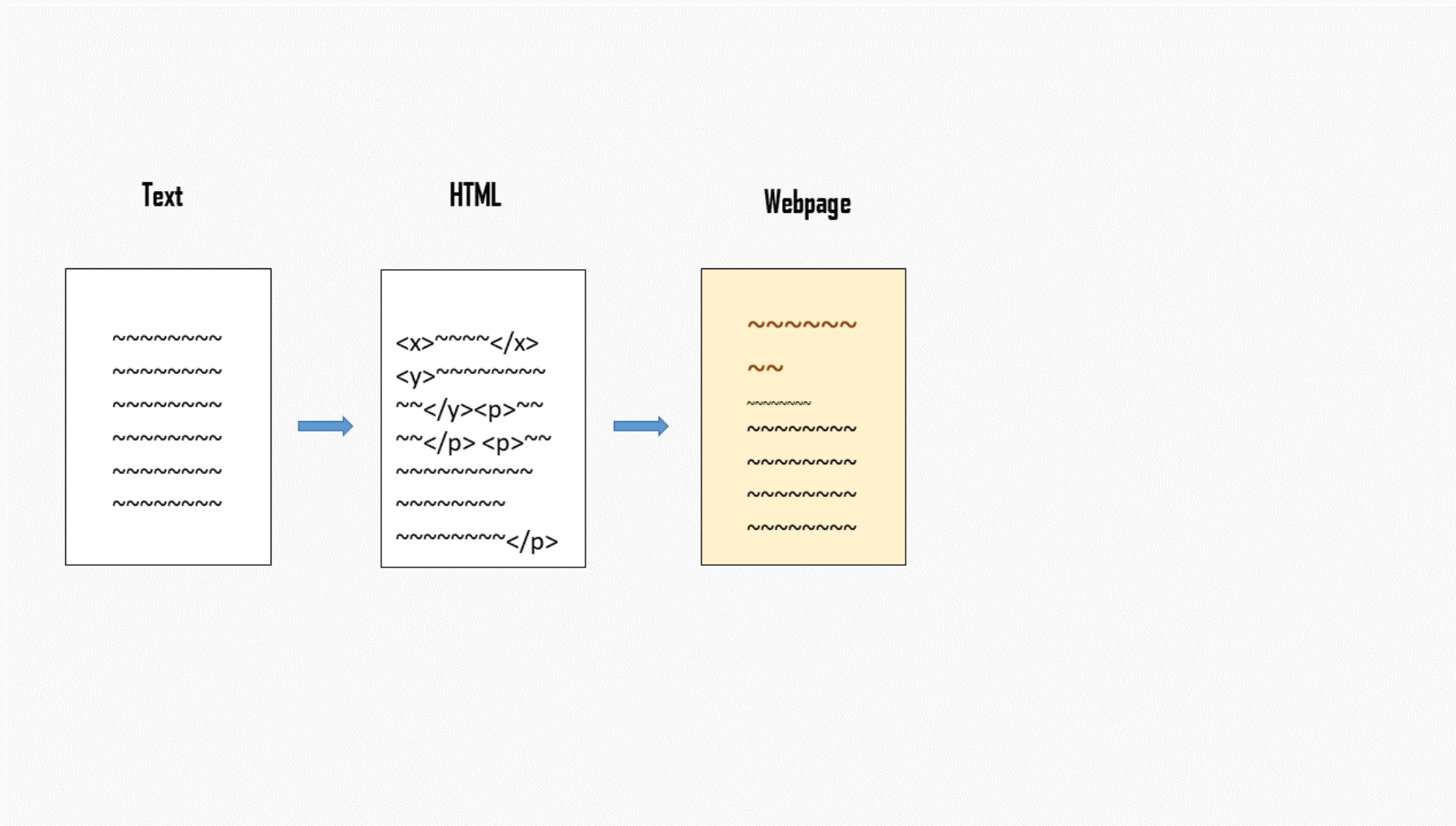
- **H**yper **T**ext **M**arkup **L**anguage
  - *markup*: additional description of formatting beyond the content of the text
- language consists of **HTML tags** to specify character / behaviour of text
- HTML tags typically consist of a starting and an end tag (exceptions: images, line breaks etc.)
  - many exceptions where it is not 'markup'
- they surround the text they are formatting

```
<tagname>Content goes here ... </tagname>
```

# HTML

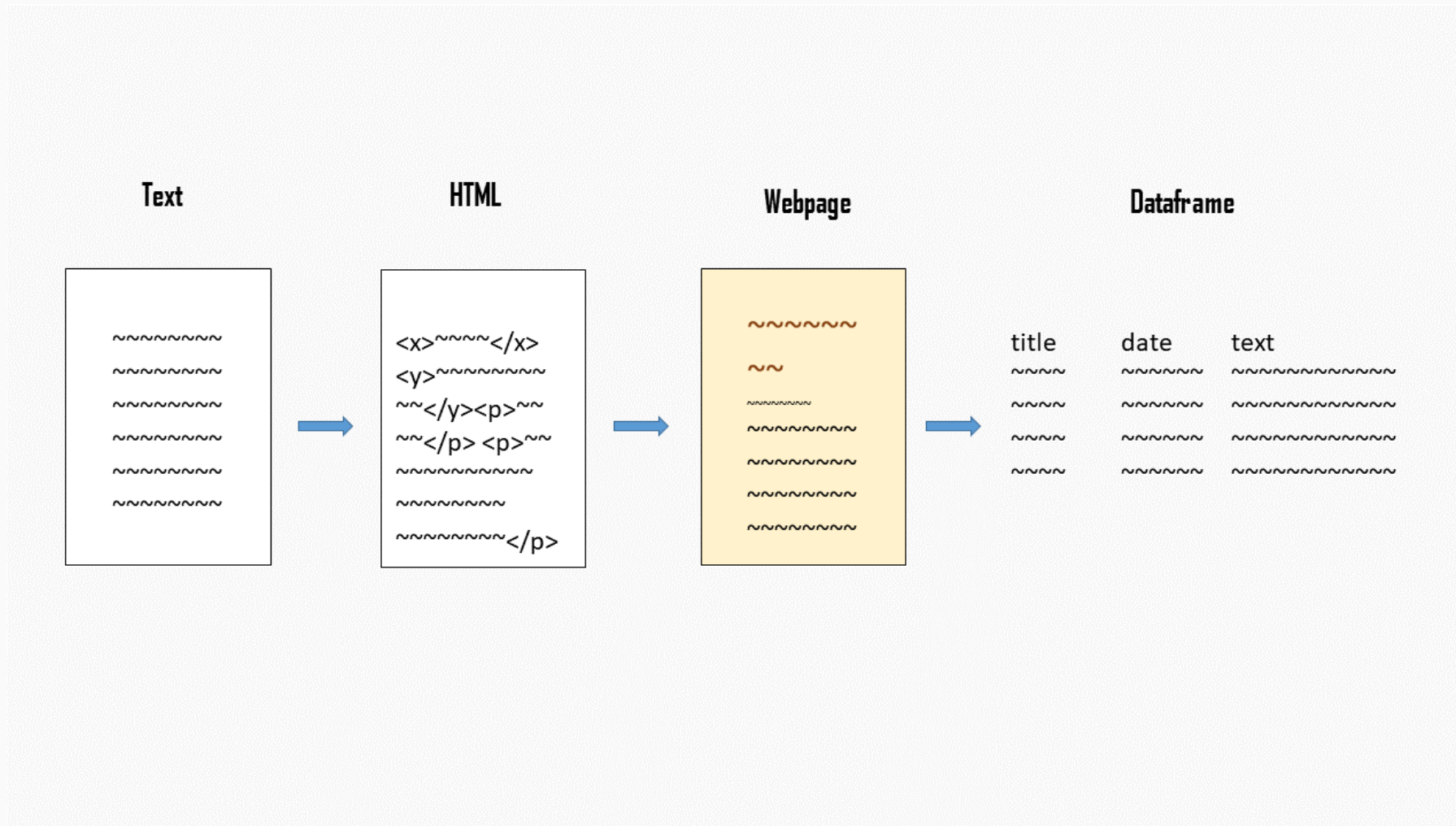


# HTML

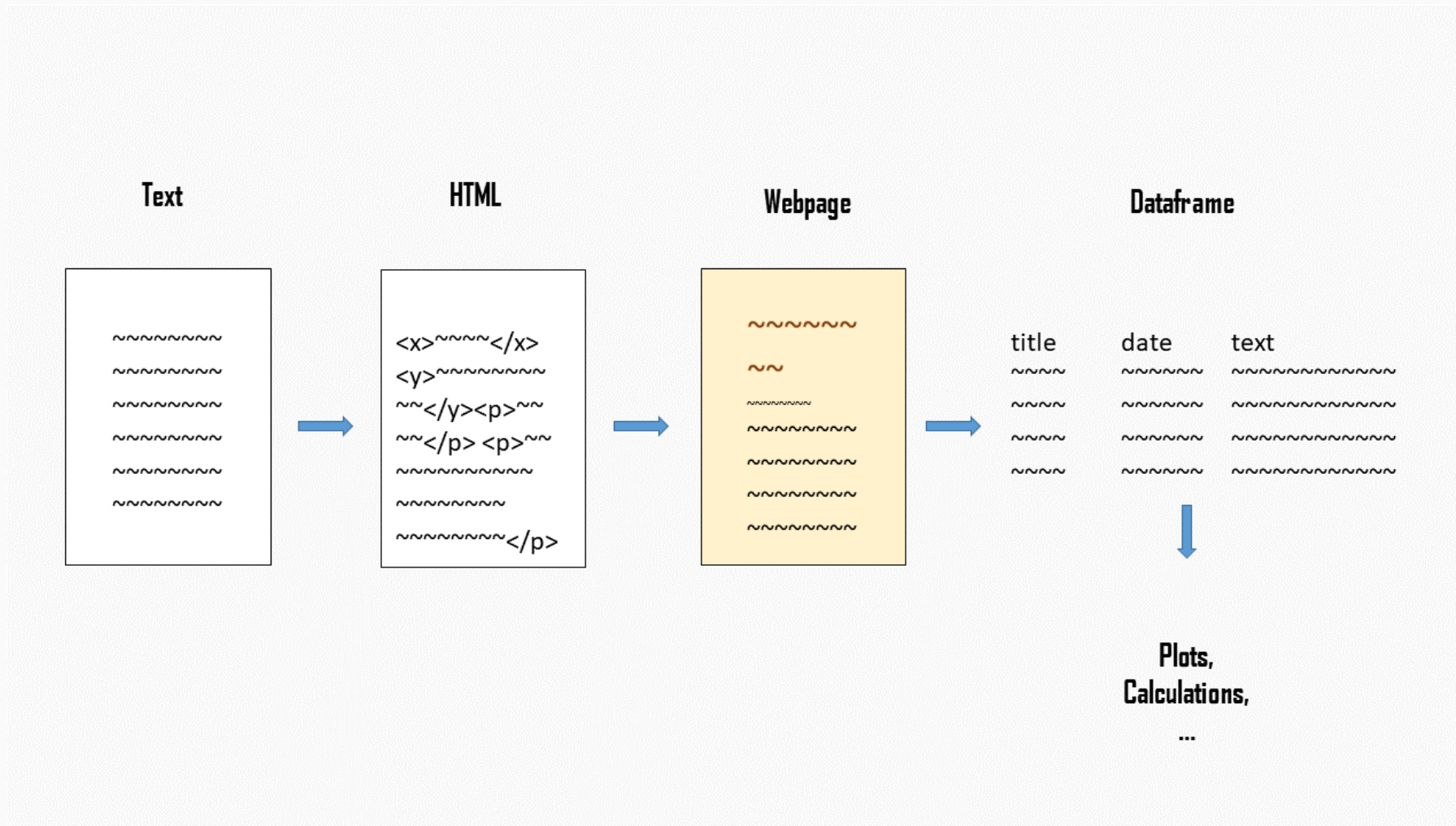




# HTML



# HTML



## Webscraping HTML Pages

- collecting data from HTML pages means removing the formatting but keeping any information it contains
  - 'parsing' of page structure
  - 'selecting' of parts of pages
- example webpage: <https://quotes.toscrape.com/>

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>Quotes to Scrape</title>
  <link rel="stylesheet" href="/static/bootstrap.min.css">
  <link rel="stylesheet" href="/static/main.css">
</head>
<body>
  <div class="container">
    <div class="row header-box">
      <div class="col-md-8">
        <h1>
          <a href="/" style="text-decoration: none">Quotes to Scrape</a>
        </h1>
      </div>
      <div class="col-md-4">
        <p>
          <a href="/login">Login</a>
        </p>
      </div>
    </div>
  </div>

  <div class="row">
    <div class="col-md-8">
      <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
        <span class="text" itemprop="text">The world as we have created it is a process of our thinking. It cannot be changed without
changing our thinking.</span>
        <span>by <small class="author" itemprop="author">Albert Einstein</small>
        <a href="/author/Albert-Einstein">(about)</a>
        </span>
      </div>
      <div class="tags">
        Tags:
        <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" / >

        <a class="tag" href="/tag/change/page/1/">change</a>

        <a class="tag" href="/tag/deep-thoughts/page/1/">deep-thoughts</a>

        <a class="tag" href="/tag/thinking/page/1/">thinking</a>

        <a class="tag" href="/tag/world/page/1/">world</a>
      </div>
    </div>
  </div>
```

## Example HTML Code

## Basic HTML tags

```
<html>
  <head>
    <title>Title of your web page</title>
  </head>
  <body>
    HTML web page content
  </body>
</html>
```

- we are mostly interested in what is inside the **body**, that is, the content of a webpage
- **head** gives meta information, often used by search engines
- tags can be **nested**

## Basic HTML Tags: Headings

**Headings** are defined by numbered h tags. Examples (with code and outcome):

```
<h1> your heading</h1>
```

```
<h2> a smaller heading</h2>
```

### a smaller heading

```
<h3> an even smaller heading</h3>
```

#### an even smaller heading

```
<h4> an even smaller heading</h4>
```

##### an even smaller heading

```
<h5> an even smaller heading</h5>
```

**an even smaller heading**

## Basic HTML Tags: Paragraphs

**Paragraphs** are defined by `div` or `p` tags.

Examples:

```
<p>this is a paragraph.</p><p>and this is the next.</p>
```

this is a paragraph.

and this is the next.

```
<div>this is a paragraph.</div><div>and this is the next.</div>
```

this is a paragraph.

and this is the next.

## Basic HTML Tags: Attributes

- All HTML elements can have attributes
- Attributes provide additional information about an element
  - they are included inside the tag

## Usage

- they are always specified in the starting tag
  - e.g. `<title attribute="x"> Title </title>`
- Attributes usually come in name and value pairs
  - e.g. `attributename="attributevalue"`



## Basic HTML Tags: Attributes - Links

- Most common case of attributes: **links**
  - text or images turned into a link by surrounding `<a>` tag (*anchor*)
  - link address specified as href attribute (*hyperreference*)

Example:

```
This is text <a href="http://quotes.toscrape.com/">with a link</a>.
```

This is text [with a link](http://quotes.toscrape.com/).

## Basic HTML Tags: Attributes

- other examples of attributes
  - src: location of an image
  - styles: formatting

Examples:

```

```

```
<p style="color:red">This is a paragraph.</p>
```

This is a paragraph.

## Styling with Classes

Webpages like blogs often define **Styles** and apply them to classes across the whole webpage. This use of classes is very common because it reduces the risk of accidentally formatting one instance of a repeated element differently.

```
<style>
p.error {
  color: red;   border: 1px solid red;
}
</style>

<p class="error">Red highlight</p>
```

Red highlight

# HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>Quotes to Scrape</title>
  <link rel="stylesheet" href="/static/bootstrap.min.css">
  <link rel="stylesheet" href="/static/main.css">
</head>
<body>
  <div class="container">
    <div class="row header-box">
      <div class="col-md-8">
        <h1>
          <a href="/" style="text-decoration: none">Quotes to Scrape</a>
        </h1>
      </div>
      <div class="col-md-4">
        <p>
          <a href="/login">Login</a>
        </p>
      </div>
    </div>
  </div>

  <div class="row">
    <div class="col-md-8">
      <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
        <span class="text" itemprop="text">The world as we have created it is a process of our thinking. It cannot be changed without
changing our thinking.</span>
        <span>by <small class="author" itemprop="author">Albert Einstein</small>
        <a href="/author/Albert-Einstein">(about)</a>
        </span>
      </div>
      <div class="tags">
        Tags:
        <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" / >

        <a class="tag" href="/tag/change/page/1/">change</a>

        <a class="tag" href="/tag/deep-thoughts/page/1/">deep-thoughts</a>

        <a class="tag" href="/tag/thinking/page/1/">thinking</a>

        <a class="tag" href="/tag/world/page/1/">world</a>
      </div>
    </div>
  </div>
```

Have another look at the webpage -  
do you understand more now?

Thanks - any questions?