ETH zürich

Social Data Science

Session 01: What is Social Data Science?

Dr. David Garcia

ETH zürich

Who am I?

- Dr. David Garcia
- Website: dgarcia.eu
- Twitter: @dgarcia_eu
- Github: dgarcia-eu
- Email: dgarcia@ethz.ch

- Group leader at the Medical University of Vienna, Section for Science of Complex Systems
- Faculty member of the Complexity Science Hub Vienna
- Privatdozent at ETH Zurich

For questions, suggestions, etc, go to the course Moodle:

https://moodle-app2.let.ethz.ch/course/view.php?id=4091

Notes:

- Learning goals of Session 01:
  - To know the principles of Social Data Science and its relationship to other disciplines
  - To be aware of the new oportunities of digital trace data
  - To have an overview of the course and the exercise setup
- Remember the course Moodle: https://moodle-app2.let.ethz.ch/course/view.php?id=4091

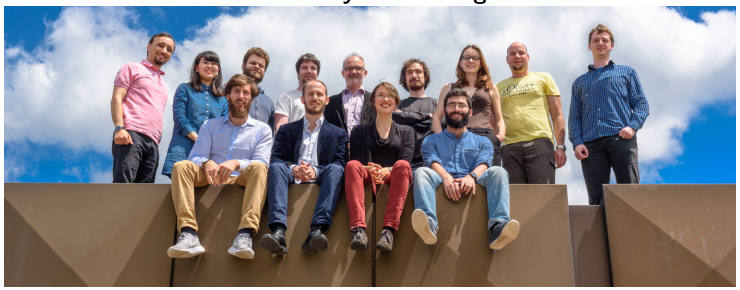
Notes:

Where do I work?

COMPLEXITY SCIENCE HUB VIENNA

Who contributed to this course?

Chair of Systems Design



With the former help of:

- Qixuan (Kevin) Zhang
- Marc Egger
- Corneel van der Pol
- Alex Huang and Tamás Kriváchi

Notes:

- Website of the Complexity Science Hub: https://www.csh.ac.at

Complexity science links state-of-the-art mathematics, modelling, data and computer science with fundamental questions posed from various disciplines, such as medicine, economics, ecology or social sciences, and opens new paths to a deeper understanding of systemic risks, resilience, efficiency, and the requirements for sustainable innovation and creativity.

Notes:

- Website of the Chair of Systems Design: https://www.sg.ethz.ch/teaching/sds/
- Important predeceding courses:

1. Collective Dynamics of Firms
2. Complex Adaptive Systems
3. Agent-Based Modelling of Social Systems

ETH zürich

Who are you?

- **And more important: What do you expect from this course?**
- Students from various levels: MSc, PhD, MAS, external...
- Students from many departments and disciplines

1. MTEC, GESS
2. Computer Science, Engineering studies (ITET, MAVT)
3. Mathematics, Statistics

ETH zürich

Overview of Session 01

1. What is Social Data Science?
2. New data for Social Science
3. Course overview and administrative details
4. Exercise 01: R Crash Course

Notes:

Notes:

What do we want with this course?

The Quantitative Understanding of Human Behavior

- **Quantitative**
  As opposed to qualitative or descriptive, we aim for robust findings grounded in strong evidence
- **Understanding**
  Not just predicting, we want to be able to generalize and combine knowledge
- **Human**
  We will not study particles or objects. Measurement and ethics will be a challenge
- **Behavior**
  Observable changes, structures, dynamics, and patterns; not just stories or theories

How are we going to do it?

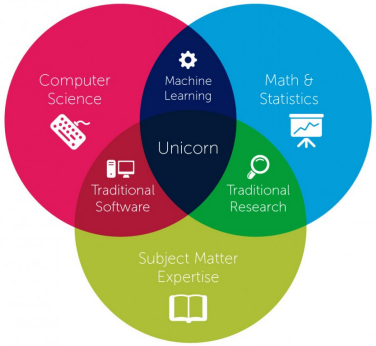
Retrieving, processing, analyzing, and interpreting Digital Traces

Notes:

Notes:

Visualizing Friendships. Paul Butler, Facebook, 2010
https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

What is Data Science?



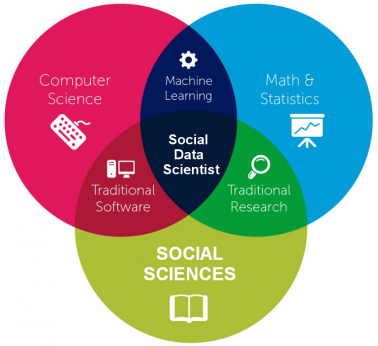
Data: plural of *datum*, "given (things)"

Facts in the form of stored and transmittable information

- Data is given to us, it is not fabricated nor simulated
- Application of methods from Computer Science and Statistics to empirical questions and practical problems
- Ability to **combine** methods is more important than excellence in individual disciplines
- Extreme importance of **communication** and **synthesis**

"Data Scientist: The Sexiest Job of the 21st Century"

What is Social Data Science?



- The application of Data Science to study human behavior and social interaction
- Gathering digital trace data with Computer Science techniques
- Statistical analysis of **"messy" Social Data**
- Interpreting results in a wider theoretical scope in: Sociology, Political Science, Psychology, and even Anthropology and Economics

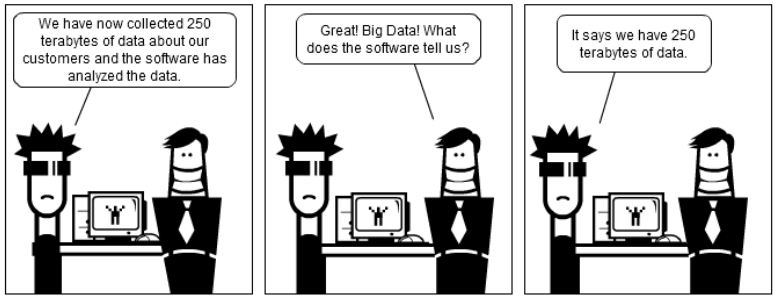
The main principle: **Social Data Science is question-driven**

Notes:

- The original diagram can be attributed to Drew Conway: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- "Data Scientist: The Sexiest Job of the 21st Century". Thomas Davenport, Harvard Business review: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Notes:

Why not Big Data or Machine Learning?



The key word in "Data Science" is not Data, it is Science

"You may have 100 GB and only 3 KB are useful for answering the real question you care about."

Why not Big Data or Machine Learning?

- **Similarity:** gathering and processing of large datasets
- **Difference:** Social Data Science aims at understanding, not prediction
  - Predictions can be very accurate, but might not give any understanding

Social Data Science is **question-driven**, not data-driven

- **What you won't learn in this course:**
  MongoDB, Cassandra, Hadoop, Spark, Artificial Neural Networks, Deep Learning, etc
- **What you will learn:**
  To retrieve and analyze datasets of human behaviour at least **one order of magnitude larger** than traditional methods in the Social Sciences

Notes:

Social Media Cartoon: SocMedSean:
http://www.socmedsean.com/
comic-the-critical-element-of-a-successful-big-data-strategy/
Credit to Michal Kosinski for finding this cartoon.
Quote from Rafa Irizarry, Roger Peng, and Jeff Leek:
http://simplystatistics.org/2013/12/12/
the-key-word-in-data-science-is-not-data-it-is-science/

Notes:

Why not Statistics?

Statistics

"Statistics investigates and develops specific methods for evaluating hypotheses in the light of empirical facts (data)."

- **Similarity:** testing hypotheses against data, generalizing
- **Difference:** statistics develops methods, Social Data Science applies them when needed

How to do research in Statistics:

1. Here is a method to test hypotheses or fit models
2. Here are lots of proofs evidencing its robustness and validity
3. Here is an application of the method to a conviniently clean dataset for a very particular question

The Data Science approach to analysis

How to do research in Social Data Science:

1. Here is a question
2. Here is a big and messy dataset extracted directly from empirical evidence
3. Here are the steps to clean and process the dataset
4. Here is an analysis (and a visualization) that addressess the question

Social Data Science is **question-driven**, not method-driven

- **What you won't learn in this course:**
  Formal statistical methods, state-of-the art models, proofs of efficiency or validity
- **What you will learn in this course:**
  Intuitive statistical principles of **interdisciplinary communicability**

Notes:

Statistics definition from: Philosophy of Statistics, 2014
https://plato.stanford.edu/entries/statistics/
"For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt... I have come to feel that my central interest is in data analysis." - John Tukey, 1962 "The Future of Data Analysis"

Notes:

A Guide to Teaching Data Science Stephanie C. Hicks, Rafael A. Irizarry
https://arxiv.org/abs/1612.07140

Why not Sociology or Psychology?

- **Similarity:** Social Data Science addresses the same fundamental questions as the Social Sciences
- **Difference:** Social Data Science is empirical and complementary to the Social Sciences

"The institutional and cultural orientation of social-science disciplines [...] have historically emphasized the advancement of particular theories over the solution of practical problems."
Duncan J. Watts

Social Data Science is **question-driven**, not theory-driven

- **What you won't learn in this course:**
  To critically argue about Social Science theories and their consequences
- **What you will learn in this course:**
  To **test propositions** from relevant Social Science theories against empirical data

Complementary approaches to human behavior

Internal validity:
The extent to which the design of a study can identify causal mechanisms
External validity:
The extent to which research results can be generalized and applied to the phenomenon they study.

Traditional Social Science methods: Experiments, surveys, interviews...

- **Strengths:** High internal validity, stardardized methods
- **Weaknesses:** Low external validity, small sample sizes

Social Data Science methods: Observational studies, digital trace analysis...

- **Strengths:** High external validity, large sample size
- **Weaknesses:** Low internal validity, unclear standards

Social Data Science is complementary to the Social Sciences, not a replacement

Notes:

Quote from "Should social science be more solution-oriented?" Duncan J. Watts.
Nature Human Behaviour 1 (2017)
http://www.nature.com/articles/s41562-016-0015

Notes:

Internal and External Validity, General Issues
W. Huitt, J. Hummel, D. Kaeck, 1999
www.edpsycinteractive.org/topics/intro/valdgn.html

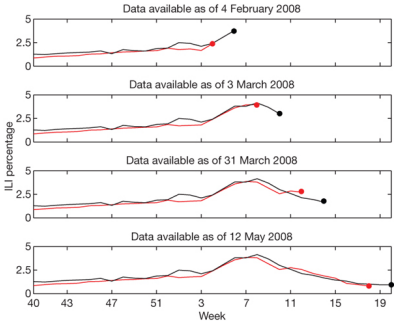
The importance of questions in Social Data Science

Deep Thought, the supercomputer to find the answer to the Ultimate Question of Life, the Universe, and Everything

Understanding our questions is a prerequisite to understanding their answers

Social Data Science Stories: Google Flu Trends

- **Nowcasting:** *predicting the present*
  - Estimation of the value of a quantity based on signals that appear at the same time
  - Nowcasting is valuable when data reporting is delayed
- **Google Flu Trends:** nowcasting influenza-related physician visits based on Google search volumes
  - CDC reports have a two week delay
  - Used 45 search terms from 50 million candidate terms
  - High weekly accuracy between 2003 and 2008
  - Applied to various regions in the US



Notes:

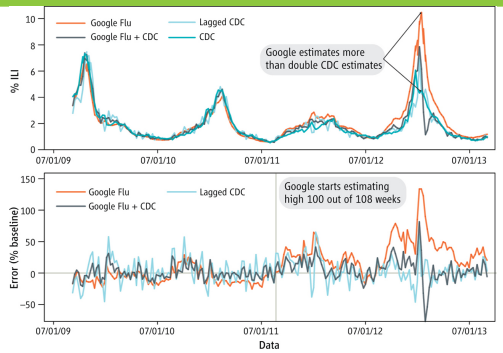
The Hitchhiker's Guide to the Galaxy (2005)
http://www.imdb.com/title/tt0371724/
https://en.wikipedia.org/wiki/42_(number)

Notes:

Detecting influenza epidemics using search engine query data. J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski and L. Brilliant. Nature, 2009
http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

When Google Flu Trends stopped working

- In February 2013, Google Flu Trends started overestimating
- Just a lagged CDC prediction with 2 week-old data outperformed Google Flu Trends
- Impossible to know the source of the error outside Google. Candidates:
  - News-related searches
  - Seasonal effects
  - Changes in Google's algorithm and interface



Big Data Hubris: "The often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" (Lazer et al, Science, 2014)

Take home message: All data is better than Big data

Beware of making a Data Piñata

- Also known as kitchen-sink regression, garbage-in garbage-out, and post hoc storytelling
- Some patterns will always come out with sufficient data, what matters is making sense of them

Notes:

http://www.nature.com/news/when-google-got-flu-wrong-1.12413
http://science.sciencemag.org/content/343/6176/1203
A predictive term for flu was "high school basketball", because the flu season and the high school basketbal season tend to coincide every year.

Notes:

- http://www.urbandictionary.com/define.php?term=data%20pi%C3%B1ata
- https://en.wikipedia.org/wiki/Kitchen_sink_regression
- Credit to Daniel Gayo-Avello for coming up with the term before me

New data for Social Science

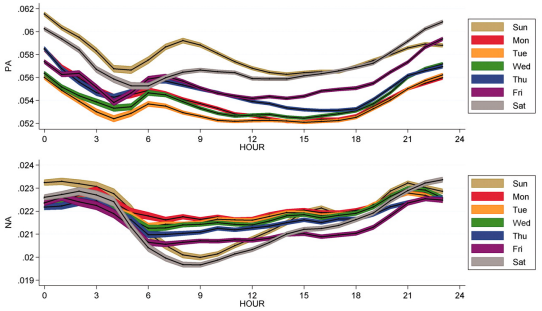
New data for Social Science

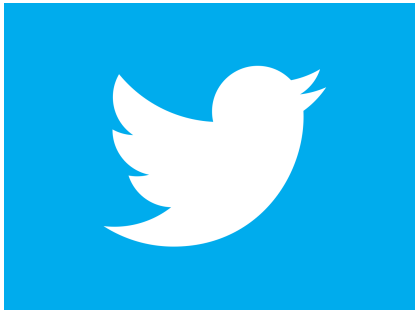
- **Big Data**
  - Observing large amounts of humans across demographics
- **Fast Data**
  - Quantifying aspects of human behavior in real time
- **Long Data**
  - Retrieving longitudinal data and at various timescales
- **Deep Data**
  - Gathering persistent information on individuals
- **Mixed Data**
  - Combining heterogeneous datasources and unstructured data
- **Weird Data**
  - Locating subcommunities and behavior unobservable in experiments or surveys

Notes:

Notes:

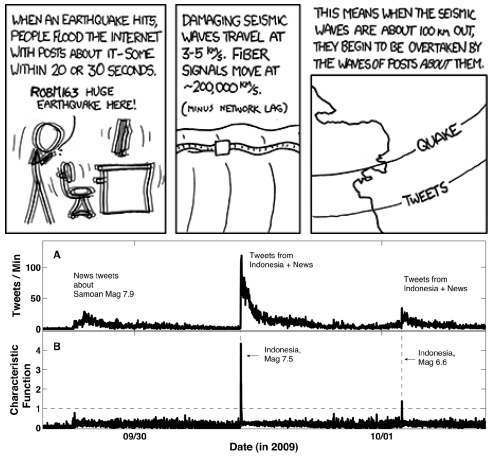
Big Data: Mood oscillations in Twitter

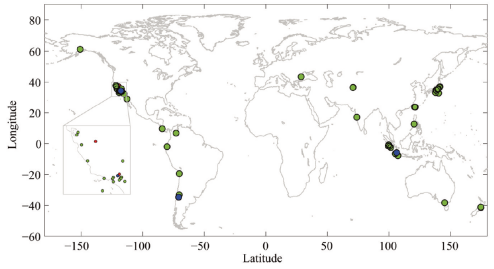




- Text analysis of expression of affect in 500 Million tweets worldwide
- Weekly and hourly oscillations of mood
- Possible thanks to **Big Data**
- Seasonal changes in mood linked to changes daylength, not absolute hours of light

Fast Data: Earthquake detection





- Twitter activity reacts to earthquakes within 2 minutes of its origin
- Considerably faster than seismographic detections in poorly instrumented regions thanks to **Fast Data**

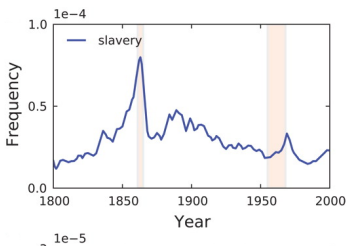
Notes:

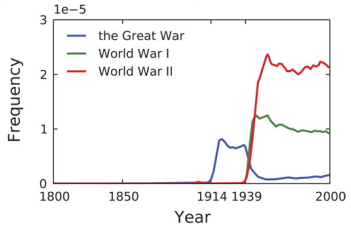
Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. S. Golder, M. Macy. Science, 2011.
http://science.sciencemag.org/content/333/6051/1878.full

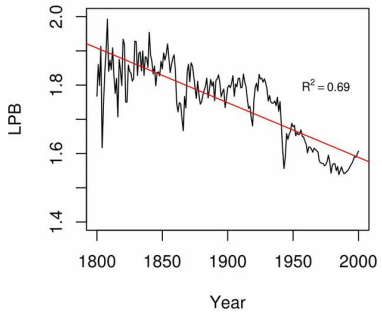
Notes:

- XKCD: Seismic Waves: https://xkcd.com/723/
- Twitter earthquake detection: earthquake monitoring in a social world. P. Earle, D. Bowden, M. Guy. Annals of Geophysics, 2011
  http://www.annalsofgeophysics.eu/index.php/annals/article/view/5364

Long Data: Culturomics in Google Books

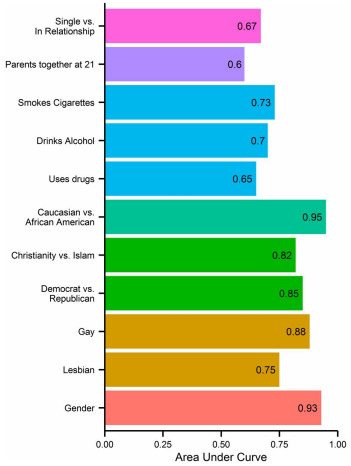


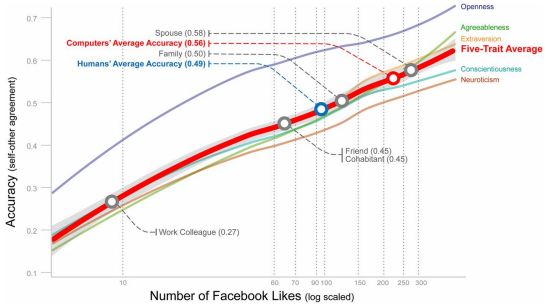




- Google scanned millions of historical books
- Culturomics: quantitative and digitized analysis of culture
- Cultural trends like linguistic biases and social topics are captured in books thanks to **Long Data**

Deep Data: Personality on Facebook





- Likes on Facebook can be used to predict personality traits and private information
- Personality prediction better than friends for > 100 likes
- The list of likes of a user is **Deep Data**

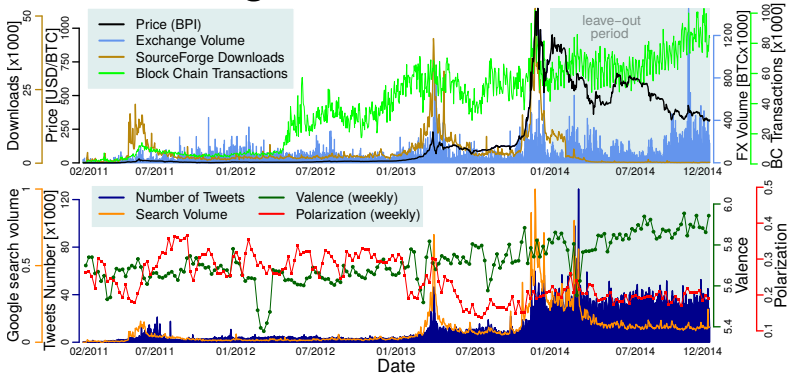
Notes:

- Quantitative Analysis of Culture Using Millions of Digitized Books. J. Michel et. al. Science, 2011. http://science.sciencemag.org/content/331/6014/176.full
- Linguistic positivity in historical texts reflects dynamic environmental and psychological factors R. Ilieva, J. Hoover, M. Dehghani, and R. Axelrod. PNAS, 2016. http://www.pnas.org/content/113/49/E7871.full

Notes:

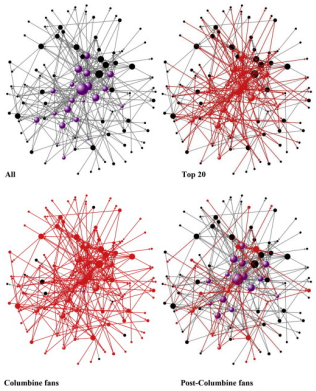
- Private traits and attributes are predictable from digital records of human behavior, M. Kosinski, D. Stillwell, T. Graepel. PNAS, 2013
  www.pnas.org/content/110/15/5802.full
- Computer-based personality judgments are more accurate than those made by humans W. Youyou, M. Kosinski, D. Stillwell. PNAS, 2015
  http://www.pnas.org/content/112/4/1036.full

Mixed Data: Bitcoin Signals



- Social and economic signals of Bitcoin can be aligned, related, and converted into **Mixed Data**
- Combining signals shows how Word-of-Mouth drives bubbles and how search precedes crashes

Weird Data: Mass Shooting fans and anonymity



Examples of **Weird Data**:

1. Pro-school shooting fans can be found through Youtube. They form a tightly-connected network
2. Anti-social norms emerge in online communities that are almost entirely anonymous

Notes:

Social signals and algorithmic trading of Bitcoin. David Garcia, Frank Schweitzer. Royal Society Open Science

http://rsos.royalsocietypublishing.org/content/2/9/150288

Notes:

- Glamorizing rampage online: School shooting fan communities on YouTube, A. Oksanen, J. Hawdon, P. Räsänen. Technology in Society, 2014
  http://www.sciencedirect.com/science/article/pii/S0160791X1400044X
- 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community, M. S. Bernstein, A. Monroy-Hernandez, D. Harry, P. Andre, K. Panovich, G. Vargas. ICWSM, 2011
  http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2873/4398

Course overview and administrative details

Course Topics

- **Day 1: Introduction to Social Data Science**
  - Session 01: What is Social Data Science?
  - Session 02: Temporal Orientation
- **Day 2: Social Dynamics**
  - Session 03: The Simmel Effect
  - Session 04: Social Impact Theory
- **Day 3. Affect**
  - Session 05: Sentiment Analysis
  - Session 06: Emotions
- **Day 4. Social Networks**
  - Session 07: Social Network Analysis
  - Session 08: Social Network Phenomena
- **Day 5. Datathon**

Notes:

Notes:

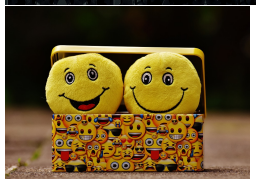
ETH zürich

Day 2: Social Dynamics

4. **The Simmel Effect**
   Trends in baby names
5. **Social Impact Theory**
   Measuring influence in Twitter

ETH zürich

Day 3: Affect



6. **Sentiment Analysis**
   Evaluating sentiment analysis tools
7. **Emotions**
   Facebook sharing of emotional content

Notes:

A great course to complement this block is 363-0541-00L Systems Dynamics and Complexity by F. Schweitzer (Autumn Semester)

https://www.sg.ethz.ch/teaching/systems-dynamics-and-complexity/

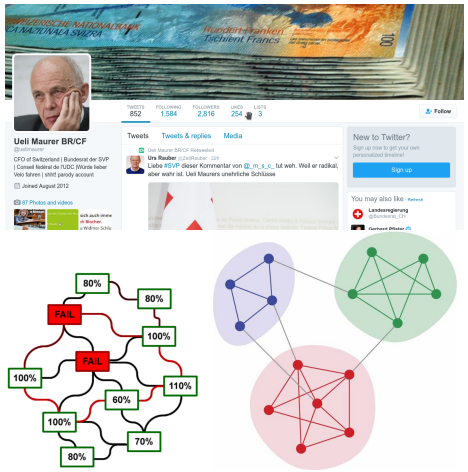
Notes:

A great course to complement this block is 252-3005-00L Natural Language Understanding (Spring Semester)

http://www.vvz.ethz.ch/Vorlesungsverzeichnis/lerneinheitPre.do?lerneinheitId=111765&semkez=2017S&lang=en

ETH zürich

Day 4: Social Networks



8. **Social Network Analysis**
   The friendship paradox in Twitter
9. **Social Network Phenomena**
   The network of Swiss politicians in Twitter

ETH zürich

Day 5: Datathon

10. **Project topic examples**
    Followed by time and feedback set up projects
11. **Short presentations (5min) of projects**
    Very quick explanation by each student of their project idea

The course will be graded in two parts:

- 50% Your work during the datathon and your presentation
- 50% Your project report. Conditions:
  - 3 pages max (11pt min font)
  - Along with codes
  - Deadline: TBD

Notes:

- A great course to complement this block is 363-0588-00L Complex Networks by I. Scholtes (Spring Semester)
  https://www.sg.ethz.ch/teaching/cn/
- Cascading failure, Wikipedia, 2008
  https://en.wikipedia.org/wiki/Cascading_failure
- Community structure, Graphstream Project, 2017
  https://data.graphstream-project.org/talks/CSSS2012/gs-communities.html#/step-2

Notes:

Project Structure

1. **Motivation**
   What question do you seek to answer and why?
2. **Project plan**
   Identify the tasks to be done to answer the question
3. **Data retrieval**
   Use interfaces and resources to collect all data necessary for the project
4. **Data processing**
   Filter data, normalize values, and compute additional variables
5. **Analysis**
   Perform statistical analyses and visualizations that assess the question
6. **Conclusion**
   Evaluate answers to the question and their reliability
7. **Critique**
   Identify limitations and alternative explanations

Exercise Materials

- Exercises sheets
  - exercise sheets in R markdown (+pdf)
  - contain exercise instructions and hints
  - fill in the missing code to solve exercise
- Exercises sessions
  - **you** participate in solving during the session
  - different students may get different solutions
  - emphasis on interpretation and critique

Google Trends

Set up gtrendsR

To get access to Google Trends data, you must provide Google account login credentials. We provided one for you that you can use, but feel free to use your own instead.

Extract data from Google Trends.

For each country we need the FOI, which is the ratio between the volume of searches for "2015" and "2013", in the year 2014

Note that with Google Trends we can query a maximum of 5 countries at a time, so we won't get all the data in one go. Rather it is worth making a for loop that goes through all the country codes. Google Trends doens't give absolute volumes, but relative ones. It always sets the largest data to 100 and scales the rest accordingly. However if you search for two things at the same time, or two countries at the same time, the results will have the correct proportion to each other. This means that to get the correct FOI, for each country code you need to extract data for "2015" and "2013" in the same search! In other words you should only have one call of the gtrends function for each country.

```
# Example: gtrends(c("2013","2015"),geo="CH")

#your code goes here
```

Regress GDP per capita PPP on FOI and plot

Now that you have the FOI index and GPD per capita, PPP value for each country, you can make a regession and plot the result. Please intepret the result and give a simple explanation.

```
#your code goes here
```

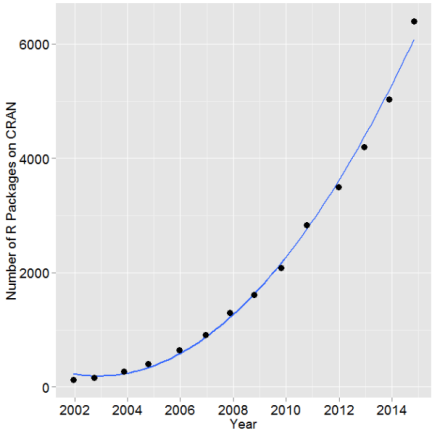
Practice is essential to become a Social Data Scientist

Notes:

Notes:

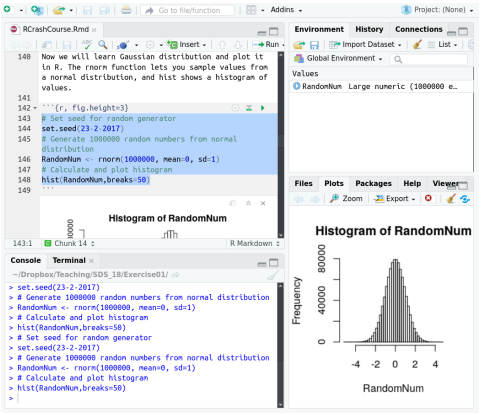
What is R?

- An open source *programming language and software environment* for statistical computing and graphics
- Supported by the R Foundation for Statistical Computing
- Widely used among statisticians and data miners for developing statistical software and data analysis
- You can code R scripts (programs) or run commands in the interactive terminal



Installing R Studio

- Follow the instructions from the links provided below to download and install R and R Studio on your computer:
  - R: https://www.r-project.org
  - R Studio: https://www.rstudio.com
  - R Markdown: http://rmarkdown.rstudio.com
- The precise steps will depend on your system
- Once you have successfully installed R and R Studio, you can join the R crash course



Notes:

Figure references:

https://www.r-bloggers.com/rs-growth-continues-to-accelerate/

Notes:

ETH zürich

R Crash Course

1. What is Social Data Science?
2. New data for Social Science
3. Course overview and administrative details
4. Exercise 01: R Crash Course

Notes: