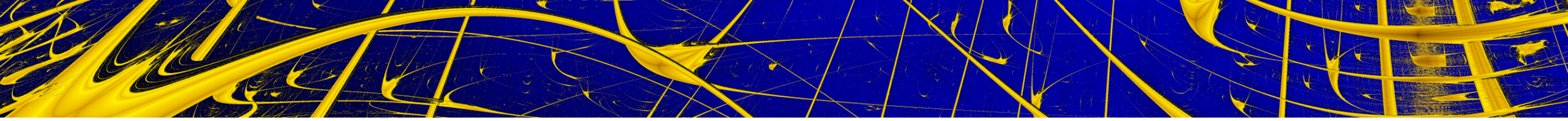


Introduction to Data Science

ESC 403

Lecture 1

Prof. Dr. Robert Feldmann
Institute for Computational Science
robert.feldmann@uzh.ch

www.ics.uzh.ch/~feldmann

Learning goals of this course

- Understand principles behind many Data Science, Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence (AI) methods
- Collect and analyze data with the help of these methods
- Apply to real-world data science problem
- Develop awareness of related ethical questions

Lectures

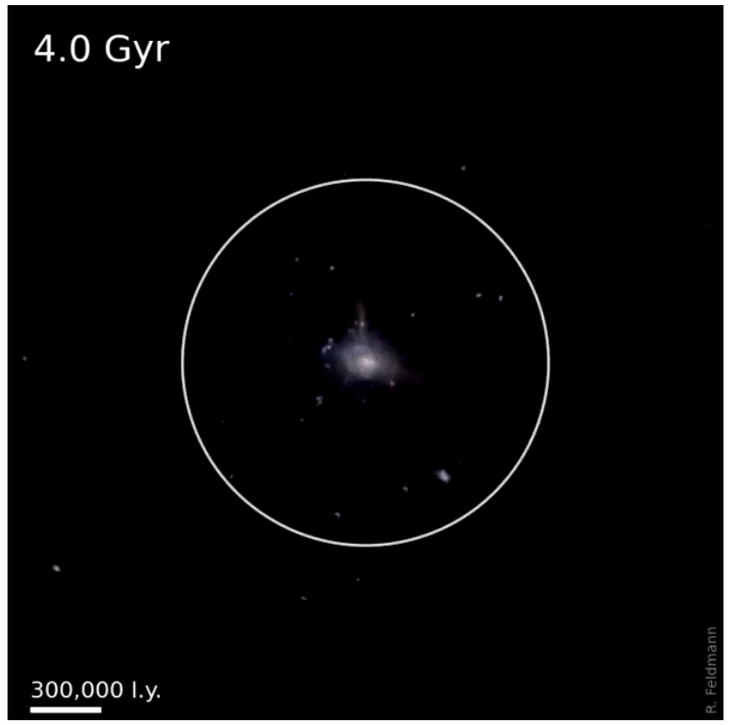
Exercises

Data Science Project

About myself

- Department of Astrophysics
- Computational Astrophysics & Data Science
  - How do galaxies form and evolve? What can we learn about the physics in galaxies from observational data?
  - Massively parallel simulations, Bayesian modeling, machine learning applied to astrophysical questions

www.ics.uzh.ch/~feldmann



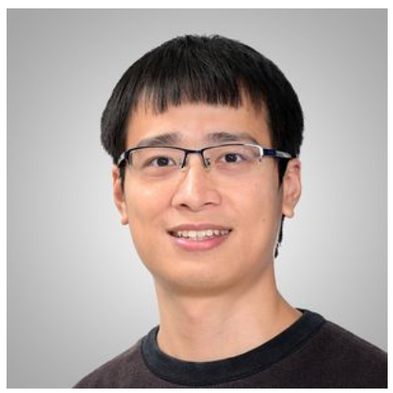
- If interested in a Bachelor's / Master's project in my group

=> contact me at: robert.feldmann@uzh.ch

Meet the team

Dr Darren Reed
(SIT / Astro)

Peng Yan
(ZHAW,
School of Engineering)



Luohong Wu
(Balgrist Universitätsklinik)

Mauro Bernardini
(Astro)

Logistics

Course website: OLAT

lms.uzh.ch olat.uzh.ch

- Wiki page
- Forum
- Lecture slides, Exercise sheets
- Dropbox to upload exercises & project proposals

Grading

- Data Science Project (**65% of grade**)
- Exam (**35% of grade**)
- Exercises (not graded but 50% of points needed)

Lecture notes

- uploaded to OLAT just before the lecture

Details

Data Science project

- team project (groups of 3-4): first step - find your team mates!
- submit a project proposal by March 28 (via OLAT)
- submit your finished project by May 17 (incl. slides, code)
- present/defend your project in the final weeks of this course

Examples of successful projects on the course website!

Exercise Sessions:

- Thursdays 1 pm Room Y03-G-91 (on site)
- streamed (zoom) but NOT recorded
- exercises handed out Tuesday, to be handed in following **Tuesday**
- Session: questions for current sheet, discussion of past sheet,
  extra topics/tutorials

Bring your laptop to the *exercise* sessions!

AI and Data Science is everywhere!

Entertainment

- personalized playlists and recommender systems (Youtube, Netflix, Spotify, Amazon)

Social Media & Communication

- facial recognition (Facebook), autocompletion (Gmail), facial filters (Snapchat)

E-commerce & Shopping

- product recommendations, targeted ads (Amazon), identification of counterfeits (eBay)

Finance & Banking

- personalized tax advice (TurboTax), fraud detection (PayPal), investment (JPMorgan Chase)

Healthcare & Medicine

- medical imaging & prognosis (Google DeepMind Health, Zebra Medical Vision)

Productivity & Design

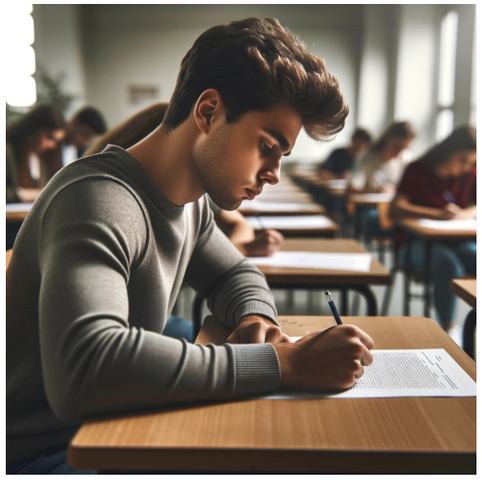
- coding (Copilot), grammar & text (Grammarly), design templates (Canva), internet search (Bing), LLM application building (Langchain)

What does this mean for us (people, society, economy)?

Examples of very recent technology

AI image generators

- Stable Diffusion (Stability AI, CompVis group)
  https://stability.ai/blog/stable-diffusion-public-release
- DALL-E (now v3) (OpenAI)
  https://openai.com/dall-e-3
- Midjourney (now v6)
  https://midjourney.com



w/ DALLE 3

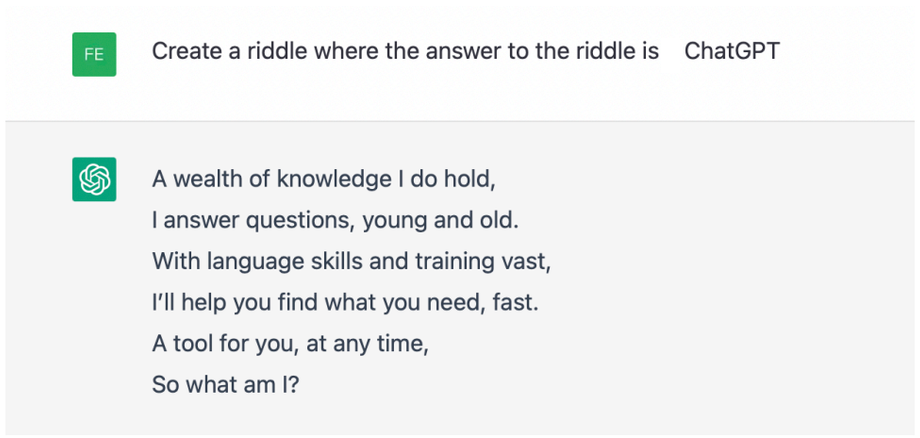
AI video generators

- Make-a-video (Meta AI)
  https://makeavideo.studio
- SORA (OpenAI)
  https://cdn.openai.com/sora

w/ SORA

Large Language Models (LLMs)

- ChatGPT-4 (OpenAI) https://openai.com/chatgpt
- Gemini (Google) https://gemini.google.com
- Claude-2 (Anthropic) https://claude.ai



w/ ChatGPT (GPT-3.5)

Technology is evolving quickly

Text prompt: *"A gorilla in an orange bathing suite jumping onto a tram in the city center of Zurich"*

DALLE 2 (**2023**)

DALLE 3 (**2024**)

Text prompt: *"A teddy bear painting a portrait"*

Make-a-video (**2023**)

Text prompt: *"A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in."*

SORA (**2024**)

AI and Data Science have a big impact on society!

COVID-19 response:

- predictive modeling, data mining, machine learning to analyze spread of the virus, identify hotspots, forecast number of infections, hospitalizations, and fatalities
- allowed public health officials and policy makers to make more informed decisions to regarding implementing measures to contain the virus

Development of autonomous driving / driver assist systems:

- often combination of AI and classical ML methods; for navigation, obstacle detection, and decision-making
- Will potentially make driving safer and more efficient (e.g., fewer accidents, traffic jams)

Natural Science:

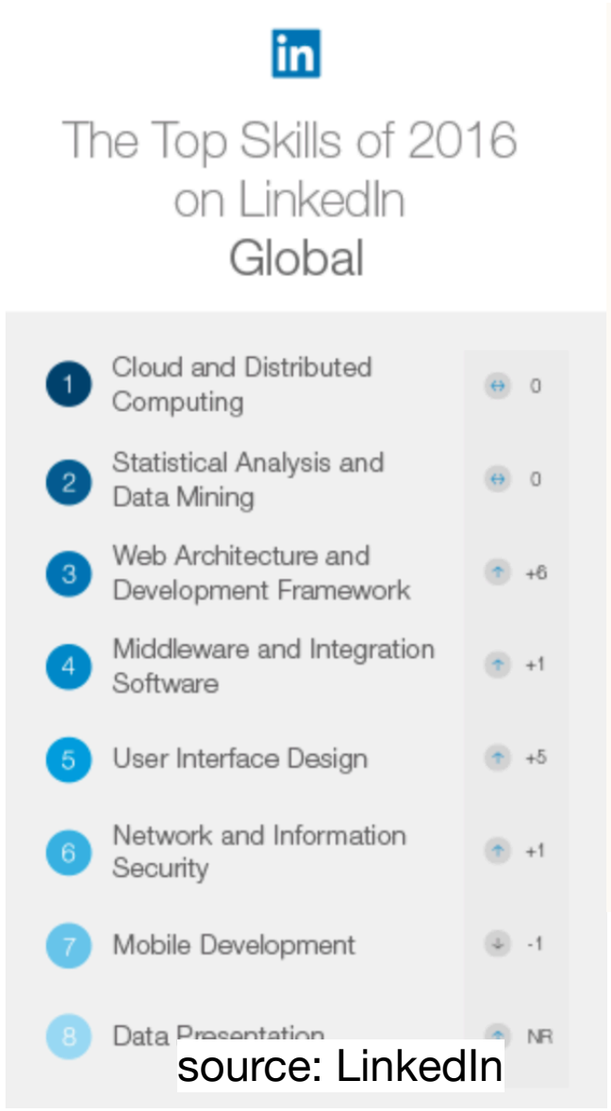
- particle physics: identify collision processes
- astronomy: classify galaxies
- cosmology: infer cosmological parameters

Paradigm shift: Instead of hypothesis ➜ data, we now have data ➜ hypothesis

Jobs in Data Science / ML / Big Data / AI

2015

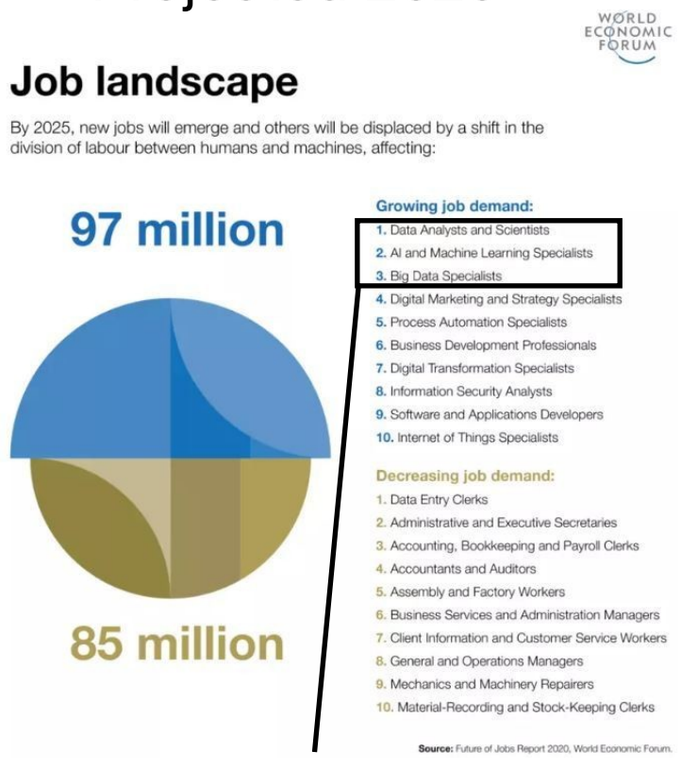
2016



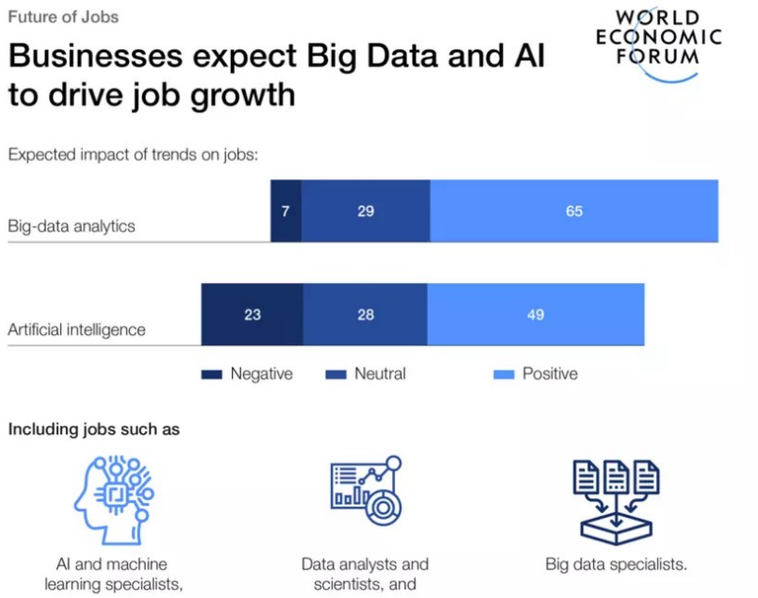
source: LinkedIn

2020

Projected 2025



1. Data Analysts and Scientists
2. AI and Machine Learning Specialists
3. Big Data Specialists



"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades."

Google's Chief Economist
Dr. Hal R. Varianin 2009

What we want to learn today

What is Data Science, AI, ML?

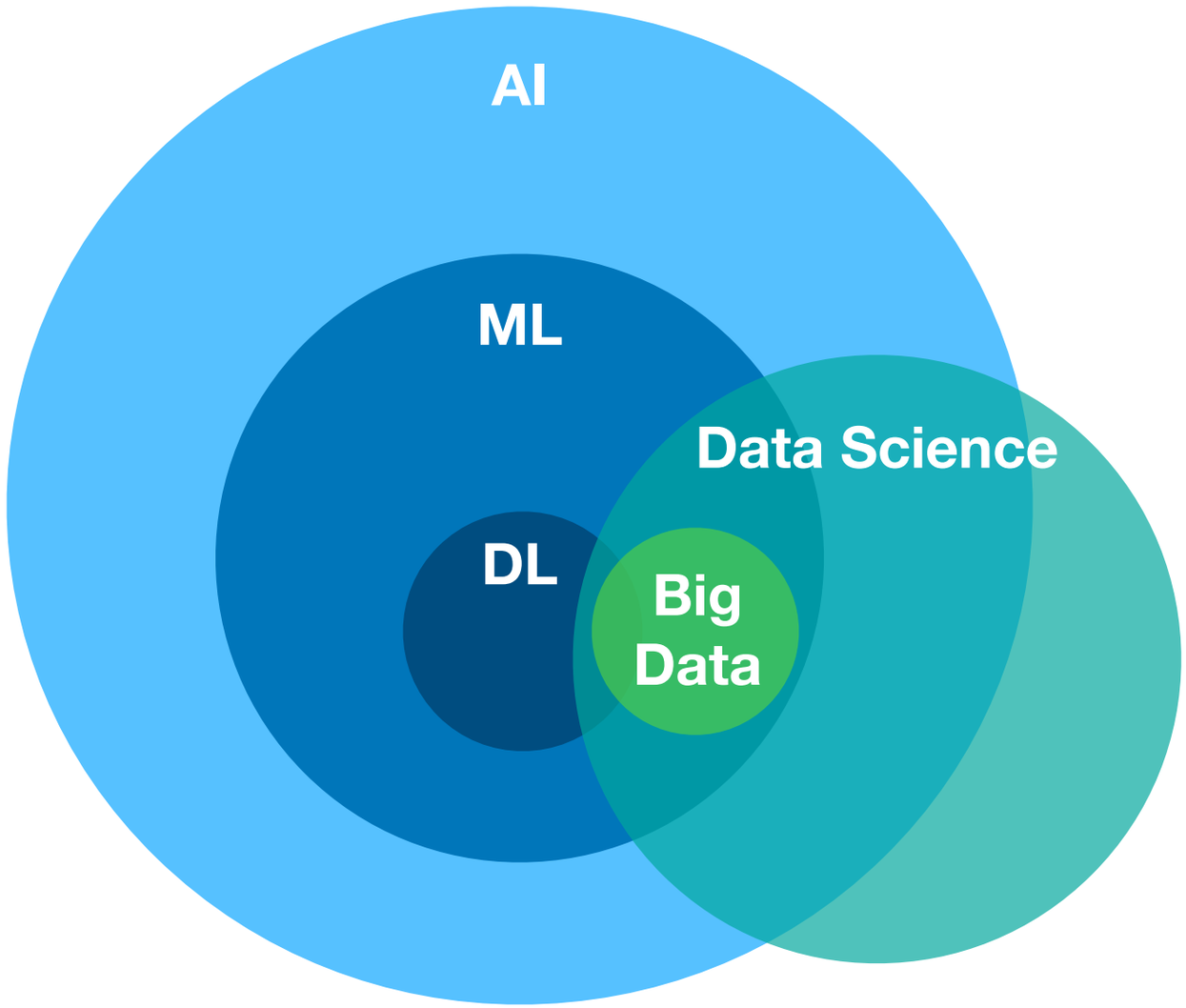
Why is this a thing now?

Which tools are important for a Data Scientist?

What are pitfalls, dangers, and problems?

What are Data Science, ML, DL, AI, Big Data?

- interrelated fields / terms, each with its unique focus and applications.
- some discussion on how to define / relate them



What is Artificial Intelligence?

Artificial Intelligence (AI) describes the ability of machines to perform tasks that would normally require human intelligence, such as:

- visual perception
- speech recognition
- decision-making
- language translating
- image generation

AI involves the development of computer programs that can process and learn from data and make predictions or propose actions based on that learning.

Ranges from simple, rule-based systems to complex, deep neural networks

Categories of AI

- Artificial Narrow Intelligence (ANI) weak AI
  ability to complete a specific task (e.g., NLP, computer vision)

- Artificial General Intelligence (AGI) strong AI
  ability to incorporate human behavior (e.g., interpret emotion), performs on par with humans

- Artificial Super Intelligence (ASI) strong AI
  surpasses human intelligence and ability

What is Machine Learning / Deep Learning?

Machine Learning (ML) is a subset of AI focusing on the development of algorithms and statistical models that enable computers to perform specific tasks without using explicit instructions.

ML is about making predictions or inferences from data, and it is a core tool within data science for building models from input data.

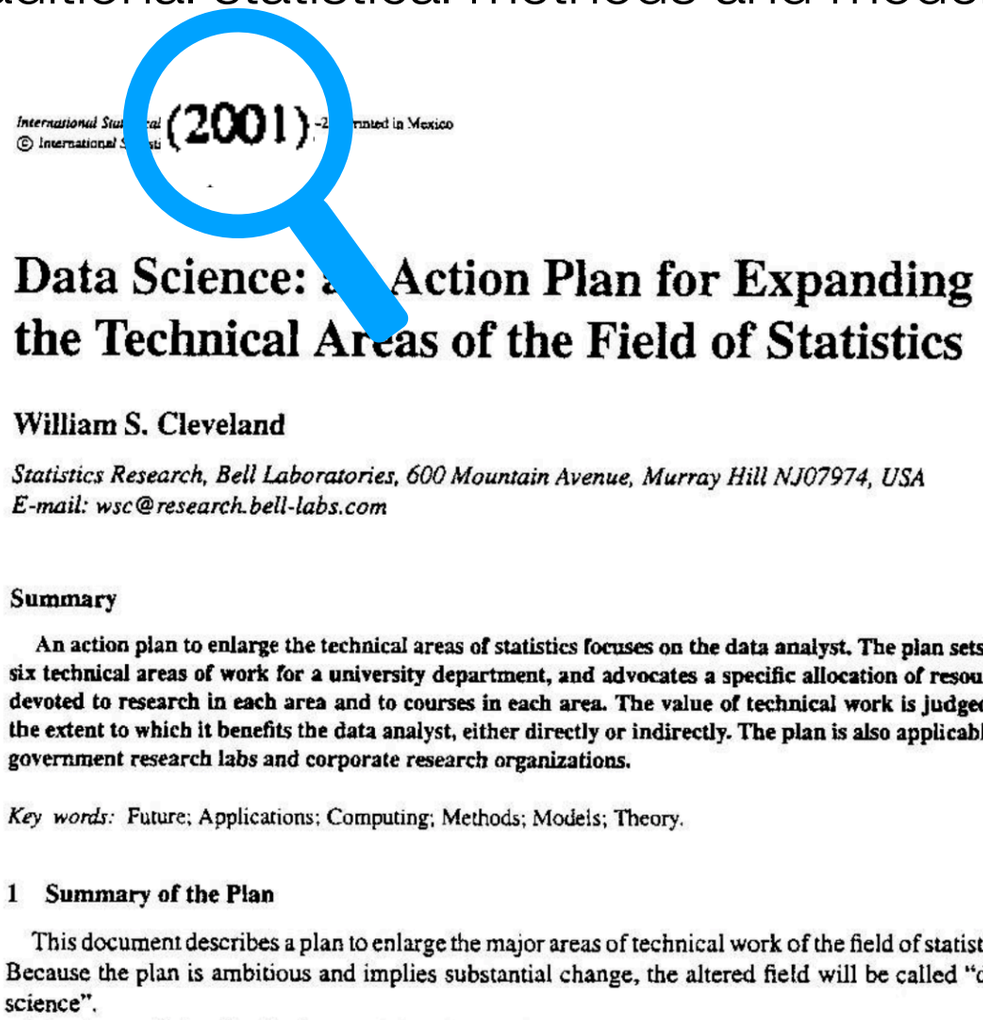
Categories of ML

- Classical Machine Learning
  - broad set of methods
  - backbone of many data analyses
  - require human intervention (e.g., for feature extraction)
- Deep Learning (DL)
  - Class of ML methods that uses deep neural networks
  - can perform many tasks without human intervention
- Reinforcement Learning
  - Learning via feedback (rewards/penalties) from actions (e.g., by interacting with environment)

What is Data Science

Data Science is an interdisciplinary field, combining aspects of statistics, mathematics, computer science, and domain expertise, that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Data science encompasses a broad range of techniques for data analysis, including both traditional statistical methods and modern machine learning techniques.



International Statistical (2001), -2 Printed in Mexico
© International Statistical

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA
E-mail: wsc@research.bell-labs.com

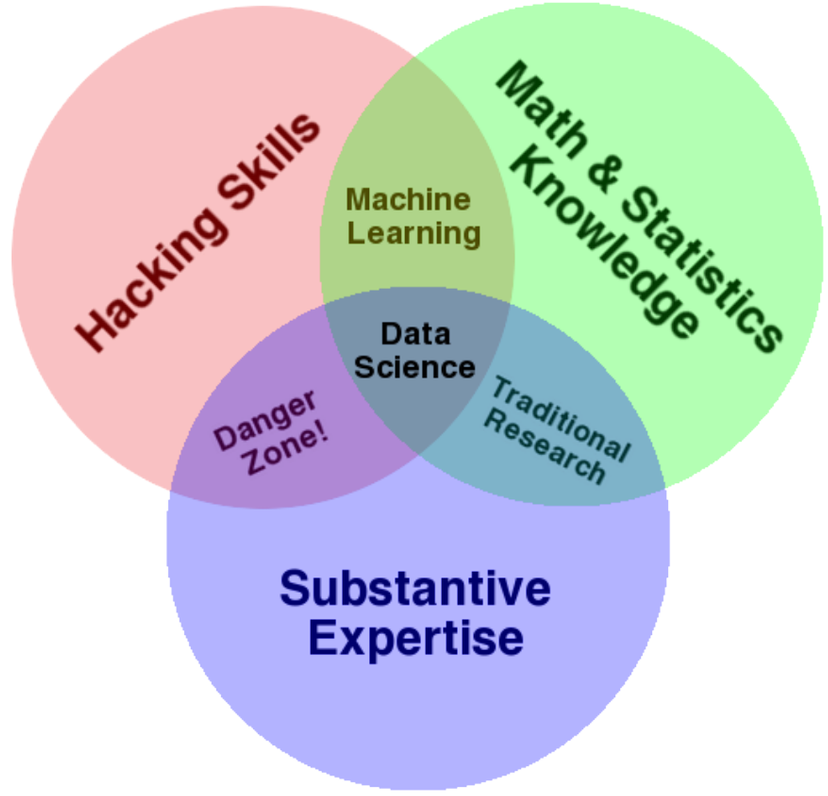
Summary

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department, and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Key words: Future; Applications; Computing; Methods; Models; Theory.

1 Summary of the Plan

This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science".



Data Science: The science of learning from data, with all that this entails

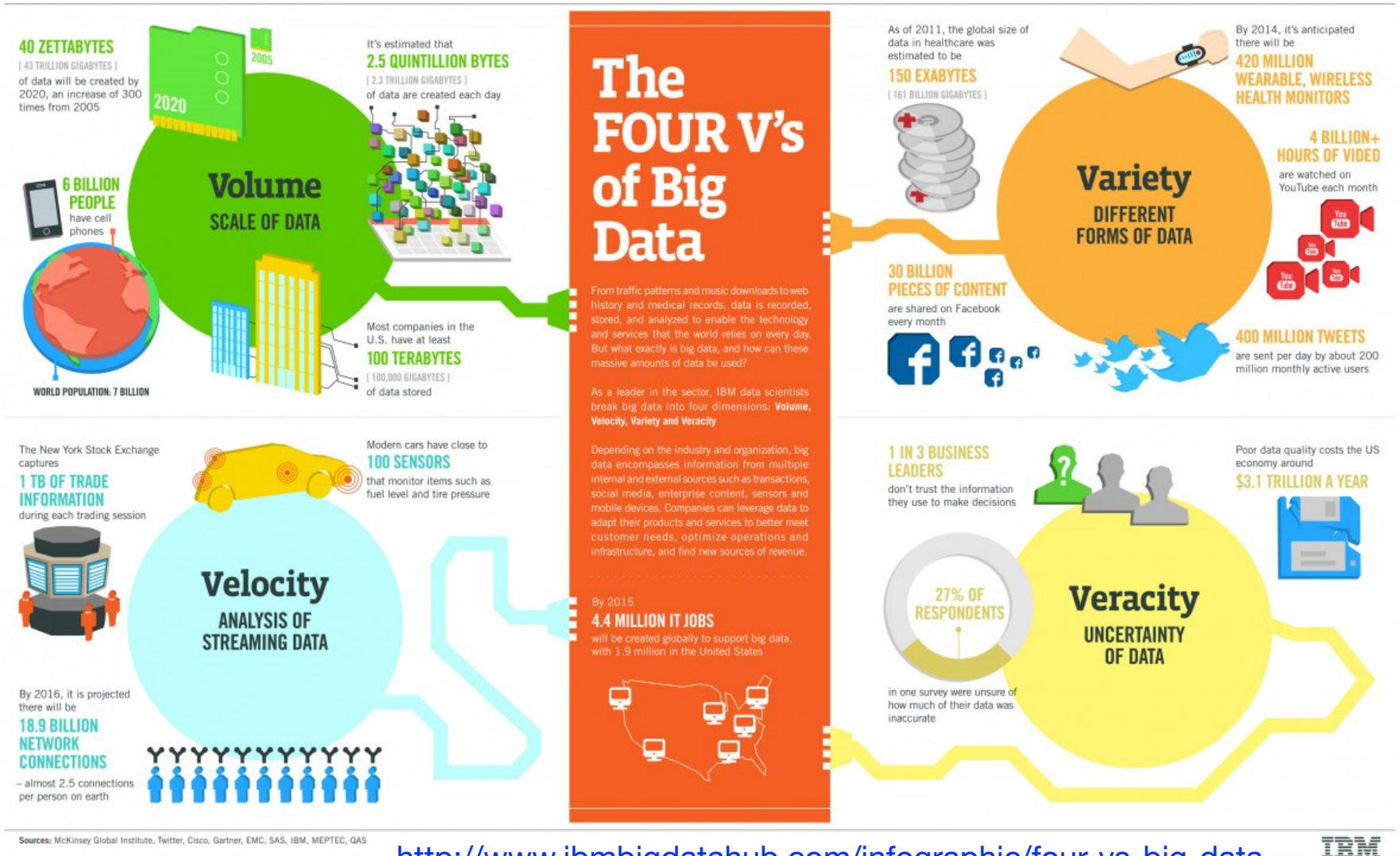
David Donoho 2015

What is Big Data?

'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.

Gartner, Inc (research and advisory company)

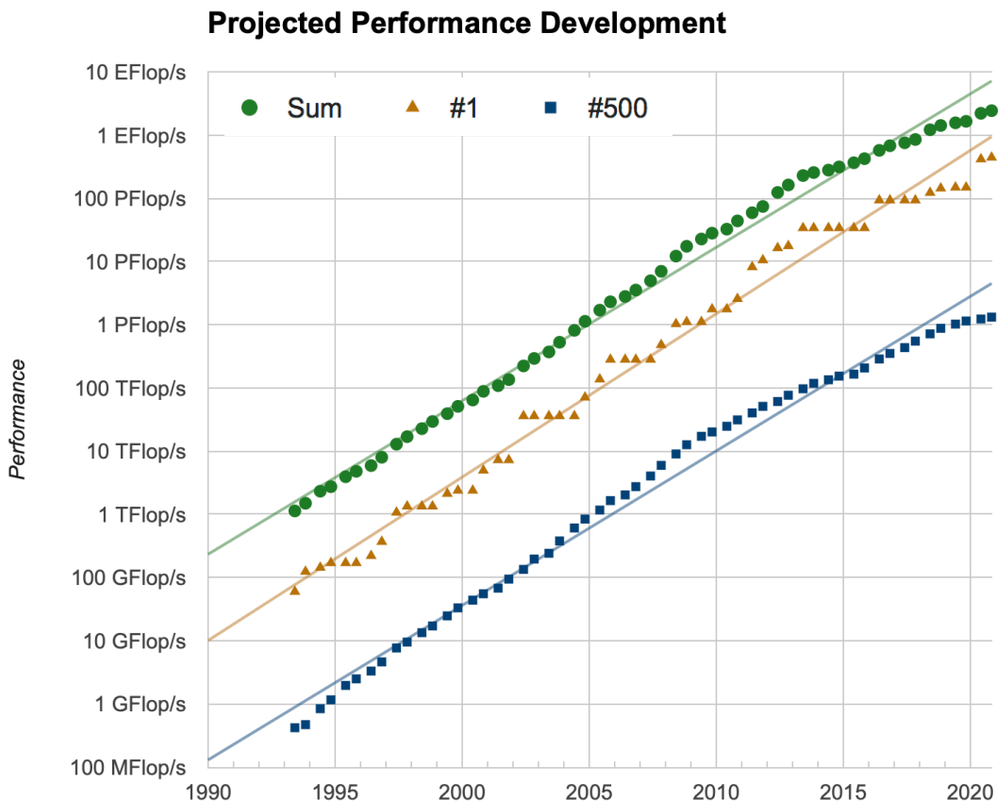
The **four V's** characterize Big Data and the challenges associated with managing and analyzing it



http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Goal of Big Data Analysis: Generate **Value**

Why is Data Science, Big Data, and AI a thing now?



www.top500.org/statistics/perfdevel/

Increase in computing power (Moore's law)

- multicore CPU, GPUs, TPUs
  => train larger models (e.g., deep NNs)
- cheaper memory
  => can store more data, models

Availability of larger data sets

- e.g., ImageNet, COCO, GPT-3

https://image-net.org
https://cocodataset.org
https://github.com/openai/gpt-3

- human classified data (e.g., Galaxy Zoo)

www.zooniverse.org/projects/zookeeper/galaxy-zoo

Cloud computing

- e.g., AWS (Amazon), Azure (Microsoft), GCP (Google)
  => access to compute power & scaling

Public / Open Source software

- e.g., TensorFlow, PyTorch
  => simplifies development & deployment of ML models

Algorithmic development

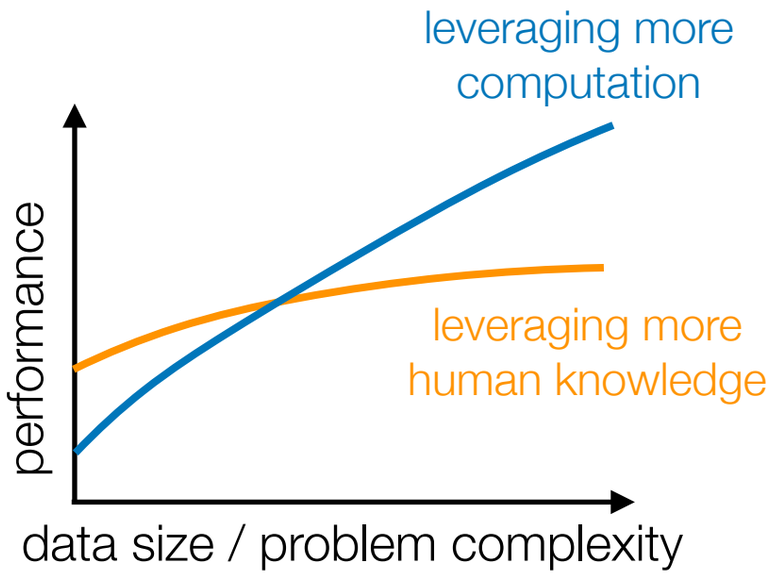
- significant advances in deep learning (e.g., CNN, transformers)
  => more efficient, easier to train models

Why is Data Science, Big Data, and AI a thing now?

"The Bitter Lesson" (Sutton 2019)

General methods that leverage computation (search & learning) are ultimately much more effective than methods that are based on how humans think.

- Reason is Moore's law
- Leveraging human knowledge leads to short term gain but makes it harder to exploit computation
- Lesson has been (re-)learned repeatedly



Examples

- Computer chess (Deep Blue vs Kasparov 1996/97): Massive, deep search much more effective than programs based on intuitive rules that humans use to play
- Similary in Go (AlphaGo vs Lee Sedol 2016): Search combined with heuristic based on AI/ML; even stronger successor AlphaGo Zero entirely self-taught without human games
- Language processing, computer vision, …

Where does data come from?

Everywhere!

- e.g., government, health, businesses, science, biometrics

Government data:

- data.gov (US)
- www.census.gov/data.html (US Census Bureau)
- http://data.europa.eu/euodp/en/data (Europe)
- data.gov.uk (UK)
- opengovernmentdata.org (various)

Health:

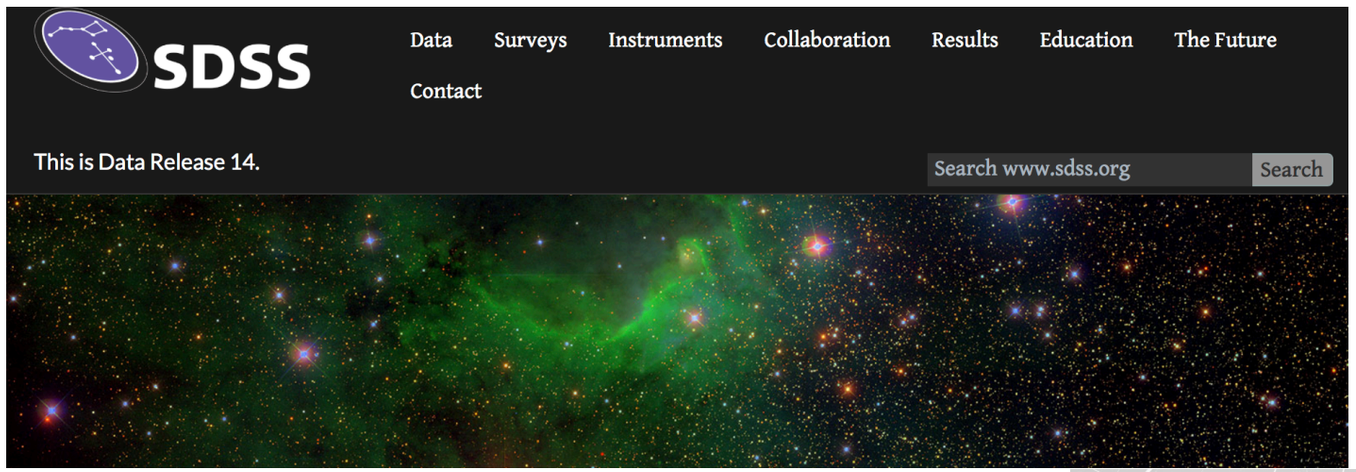
- www.healthdata.gov
- http://www.who.int/en/

Businesses:

- Facebook: https://developers.facebook.com/docs/graph-api
- The New York Times archive (search back to 1851) http://developer.nytimes.com

Science:

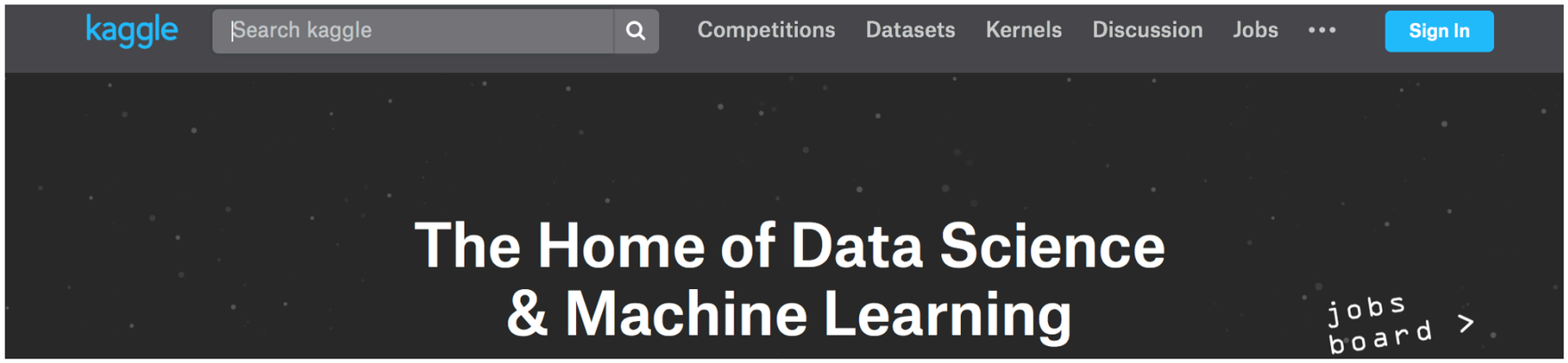
- push for open data and open access
- most data in astronomy & astrophysics is public



Where does data come from?

- Many sources (government, NPOs, science) provide the data for free
- Many data-specific companies that sell data

www.kaggle.com



Important components of Data Science

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation

Data Science Components: Data Exploration and Preparation

Preparation / Data Cleaning

- datasets contain artifacts, missing columns etc
- identify & address such issues: reformatting, recoding values, grouping, smoothing, etc

Exploratory Data Analysis (EDA):

- sanity-check its most basic properties
- expose unexpected features

Coale & Stephan 1962: “Teenage Widowers”

“An examination of tables from the 1950 U. S. Census of Population, and of the basic Persons punch card, shows that a few of the cards were punched one column to the right of the proper position in at least some columns. The result is that numbers reported in certain rare categories very young widowers [..] were greatly exaggerated. These errors occurred in spite of a careful checking program, and illustrate the ***necessity for users to view data concerning rare categories with special caution****.”*

Farman, Gardinar and Shanklin 1985: Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction

- wake-up call regarding Chlorofluorocarbons: Antartica’s ozone level had dropped by 10% below normal January levels
- authors were hesitant about publishing because previous satellite data showed no ‘ozon hole’
- but apparently, previous analysis dismissed large drops as errors / unreliable

Data Science Components: Data Representation and Transformation

- Data from multitude of sources (data bases, web, image libraries etc.)
- Many different representations of data:
  - e.g., text files, time series data, sound files, images, network data
  - graphs, trees, matrices
- Often advantageous to transform data prior to analysis, depending on data type

Variable Types

Numerical (quantitative)

- Continuous (any numerical value)
- Discrete ('interval')

Categorical (qualitative)

- Nominal (no particular order)
- Dichotomous (2 categories)
- Ordinal (categories with order)

Data Science Components: Computing with Data

- Knowledge of various computing languages:
  - Python, R, perhaps C/C++/Fortran, Shell scripts etc

- Knowledge of Computing Environments
  - IDEs, Jupyter Lab/Notebook, …
  - libraries & tools, version management

- Being able to make use of parallel & cloud computing
  - Run large number of jobs / cores on supercomputers, GPUs
  - Computing on distributed data (e.g., MapReduce, Spark)

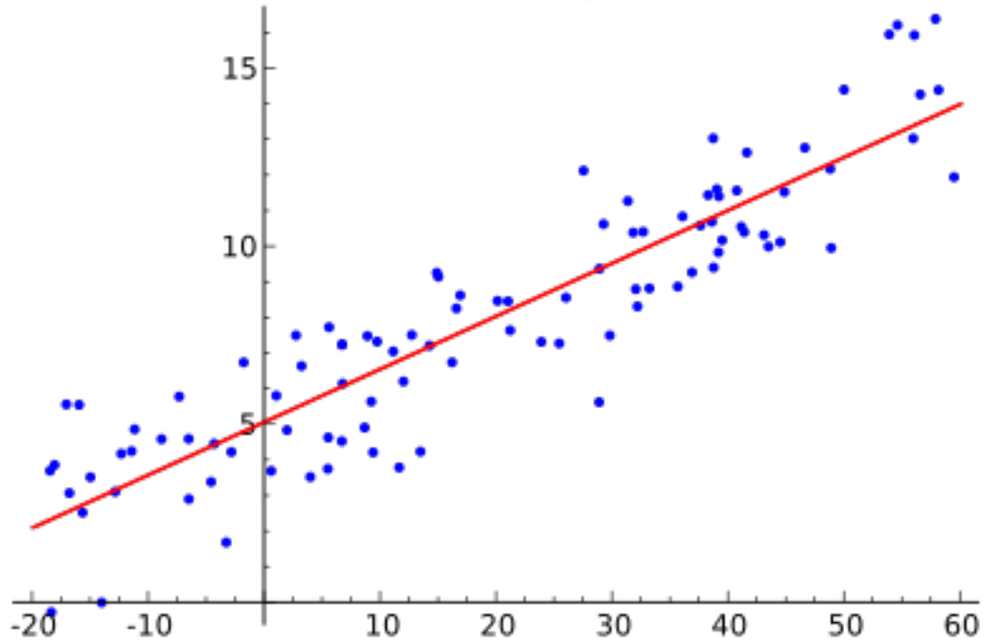
- Knowledge of ML frameworks and their APIs (e.g., TensorFlow, PyTorch)

Data Science Components: Data Modeling

- **Inferential** modeling
  - develop stochastic model that describes the data
  - infer properties of the model

  Interested in the model parameters!

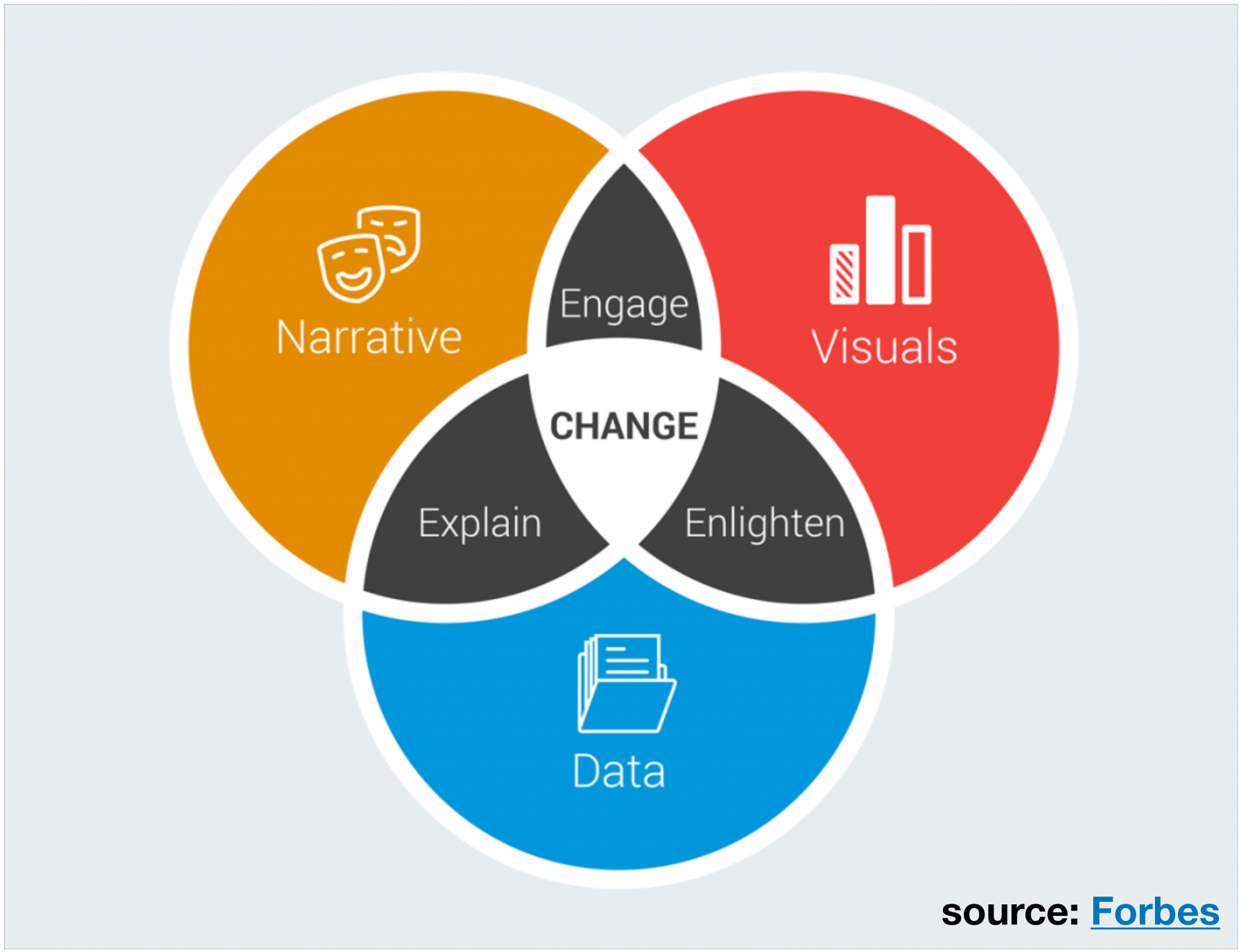
linear regression



- **Predictive** modeling
  - train models based on data
  - apply models to predict outcome over some data Universe

  Interested in the accuracy of the model predictions!

Data Science Components: Data Visualization & Presentation



source: Forbes

Danger ahead

- Key concerns:
  - Are the algorithms, data sources, and processing methods well documented?
  - Are the results reproducible by other researchers?
  - Are the results statistically sound?
  - Are the results numerically reliable?
  - Have the results been validated?

Danger ahead

Example: Reproducibility in biomedicine

Numerous cases where pharma products look good based on clinical trials but later disappoint in real-world usage, or results cannot be reproduced in separate studies.

- In 2004, GlaxoSmithKline acknowledged that while some trial of Paxil found it effective for depression in children, other unpublished studies show no benefit
- In 2011, Bayer researchers reported they were able to reproduce the results of only 17 of 67 published studies they examined
- In 2012, Amgen researchers reported they were able to reproduce the results of only 6 of 53 published cancer studies.
- In 2014, a review of Tamiflu found that while it makes symptoms to disappear a bit sooner, it did not stop serious complications or kept people out of the hospital.

Exposed a fundamental flaw in methodology:

Only publicizing the results of successful trials introduces a significant bias into the results.

Better, publish all trials! http://www.alltrials.net

Danger ahead

"I remember my friend John von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Enrico Fermi

- Backtesting / Retrodiction:
  - test predictive models on historical data
  - (in-)famous in finance, but used in many fields incl. climatology, cosmology,…

http://www.financialtrading.com/issues-related-to-back-testing

- Overfitting:
  - fit model with higher level of complexity than the data
  - noise and artifacts are fitted along true trends
  - overfitted models often fail when applied to future data

Ways to deal with it, and we will talk about it later in this course.

Assuming the data speaks for itself

"As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned?'"

David Bollier 2010

"People have this notion that you can have an agnostic method of running over data, but the truth is that the moment you touch the data, you've spoiled it. For any operation, you have destroyed that objective basis for it."

Jesper Andersen

The truth

- Researchers are always also interpreters of data
- how data is acquired ('veracity') and cleaned is subjective
- often data from different sources (w/ errors etc), complicating interpretation & analysis
- understanding the sources and biases of the data now more important than ever, otherwise 'more is less' often better

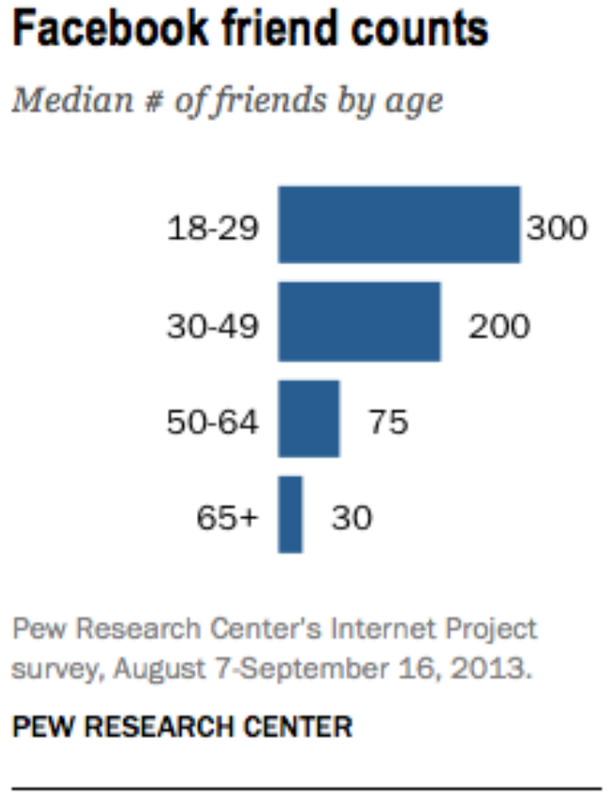
Common Pitfalls

Forgetting about context and sample origin

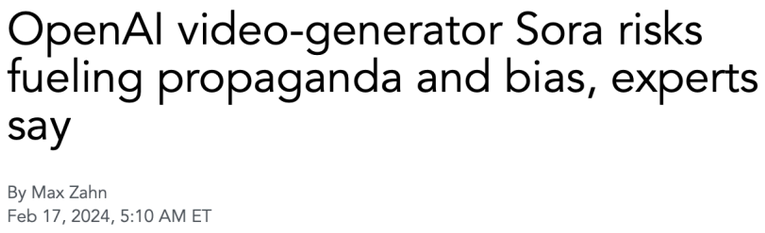
Example:

- researcher wants to study personal networks
- use ‘friends’ lists on social media
- maximum on Facebook is 5000 ‘friends’

- but these two measures are likely not equivalent
- maintaining context big challenge if data from different sources



Critical questions for Data Science and AI



Kranzberg's First Law *"Technology is neither good nor bad; nor is it neutral."*

[…] technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves

[…] the same technology can have quite different results when introduced into different contexts or under different circumstances"

Melvin Kranzberg 1986

Critical questions for Data Science and AI

The data-centric revolution triggers both utopian and dystopian visions

Will large scale search data help to create better tools, services, and public good?
Or will it usher in a new wave of privacy incursions and invasive marketing?

Will data analytics help us understand online communities and political movements?
Or will analytics be used to track protesters and suppress speech?

Will AI solve various problems in our society?
Or will they lead to more disinformation, filter bubbles, and societal degradation?

Critical questions for Data Science and AI

Clearly important to ask:

- Who gets access to what data?
- How is data collected (data source!) & analyzed?
- To what ends is the data analyzed?
- How are the data and models trained with it used?

Requires your commitment as a data scientist

Critical questions for Data Science and AI

Example #1 <u>When is it ethical to use accessible data?</u>

- in 2006, Harvard based research group collected public Facebook profiles of 1700 college students & studied how their interests and friendships changed over time (Lewis et al. 2008)
- supposedly anonymous research data was made public
- but possible to de-anonymize parts of the data set (Zimmer 2008)

Can ‘public’ data simply be used without requesting permission?

What if data (e.g., from a blog) is analyzed out of context?

Who is responsible for making sure that individuals and communities are not hurt by the research?

- Some help: Institutional Review Boards (research ethics committees)
- lot of responsibility by the individual researcher

Example #2

When is it ethical to use AI generated content in science, education, or business?

- Large Language models and AI image generators are trained on large data sets often scraped from the web

What about the copyright(s) & privacy rights of the original authors/creators of those sources?

What about the loss of income/traffic to the original authors/websites in AI based search?

Is it plagiarism to include AI generated content in scientific papers?

Science (journal)

Authors who use AI-assisted technologies as components of their research study or as aids in the writing or presentation of the manuscript should note this in the cover letter and in the acknowledgments section of the manuscript. **Detailed information should be provided in the methods section:** The full prompt used in the production of the work, as well as the AI tool and its version, should be disclosed. Authors are accountable for the accuracy of the work and for ensuring that there is no plagiarism.

Nature (journal)

Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our authorship criteria. Notably an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. **Use of an LLM should be properly documented in the Methods** section (and if a Methods section is not available, in a suitable alternative part) of the manuscript.

What about using AI tools as part of this course?

Guidelines on A.I. tools for lecturers and students at the Faculty of Science

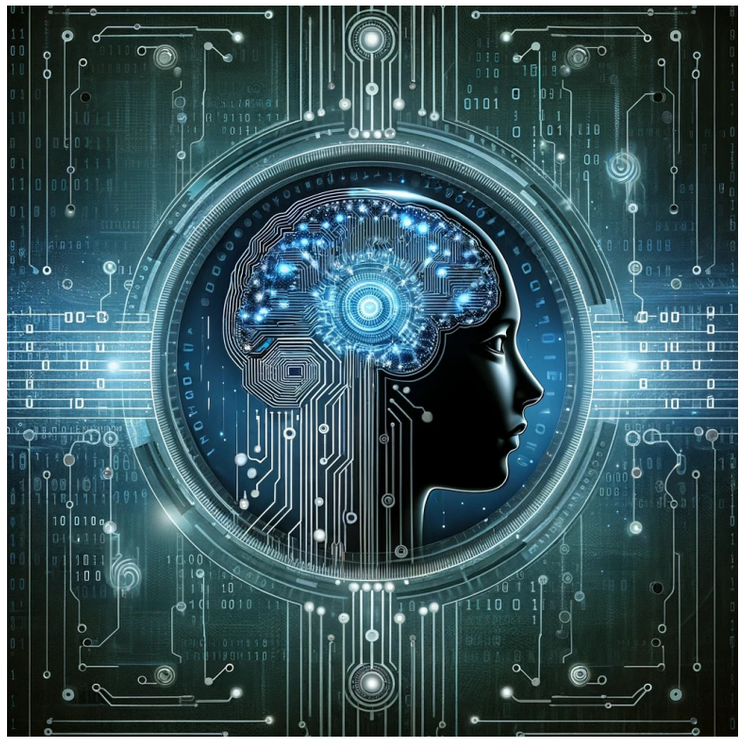
1. Lecturers and students are encouraged to experiment with, apply, and critically evaluate A.I. tools within the context of teaching and learning activities. The decision to use A.I. tools should be driven by their potential to enhance learning outcomes and should serve as a supplementary component to existing didactic strategies, rather than being adopted purely for their innovative appeal.
2. Lecturers and students are expected to critically evaluate the outputs generated by A.I. tools and determine the veracity and suitability of such results. The responsibility for the validity and authenticity of their work rests entirely with the user.
3. Lecturers and students should bear in mind that A.I. tools can amplify human biases, leading to potentially unfair or discriminatory outcomes.
4. The Faculty of Science strongly advocates for the development of students' critical and ethical thinking abilities, particularly in the context of using A.I. tools.
5. Lecturers grading remote learning and online examinations should take in account the accessibility of A.I. tools to students to ensure fairness and accuracy in evaluation.
6. In terms of academic work and publications, the focus should be directed towards the quality of data and their analyses. As A.I. tools become more widely used, grading of language quality should become less important.
7. Awareness of legal limitations related to the use of A.I. tools is crucial, especially concerning Data Protection laws in Switzerland and the General Data Protection Regulation (GDPR) in Europe. Any data uploaded to A.I. tools may be utilized for further A.I. training and could potentially be accessed by third-party applications, beyond the control of the users.

see: https://www.mnf.uzh.ch/en/studium/rund-ums-studium/Studium-und-KI.html

What about using AI tools as part of this course?

Rules for this course

- You may use AI tools during your project and exercises (but not during the exam)
- You must document any such use (which tool/version, what prompt etc.) in your homework / project documentation
- You are taking full responsibility for anything you hand in / present whether or not you make use of AI tools



Learning to use the various tools is a valuable skill, and saves time, but does not replace learning how to solve basic data science problems on your own!

Tools

In this course we will be using mainly Python

Why Python?

- Free, available for UNIX, OS X, WIN: e.g., https://www.anaconda.com
- Easy to bring up and use;
- General purpose scripting language, many specialized modules for
  - Data Access
  - Data Cleaning
  - Analysis
  - Visualization
  - Reporting
- Good development environments available (e.g. spyder, jupyter) https://jupyter.org
- Huge developer community

https://www.anaconda.com

6M+
Users

1,000+
Data Science Packages

150+
Enterprise Customers

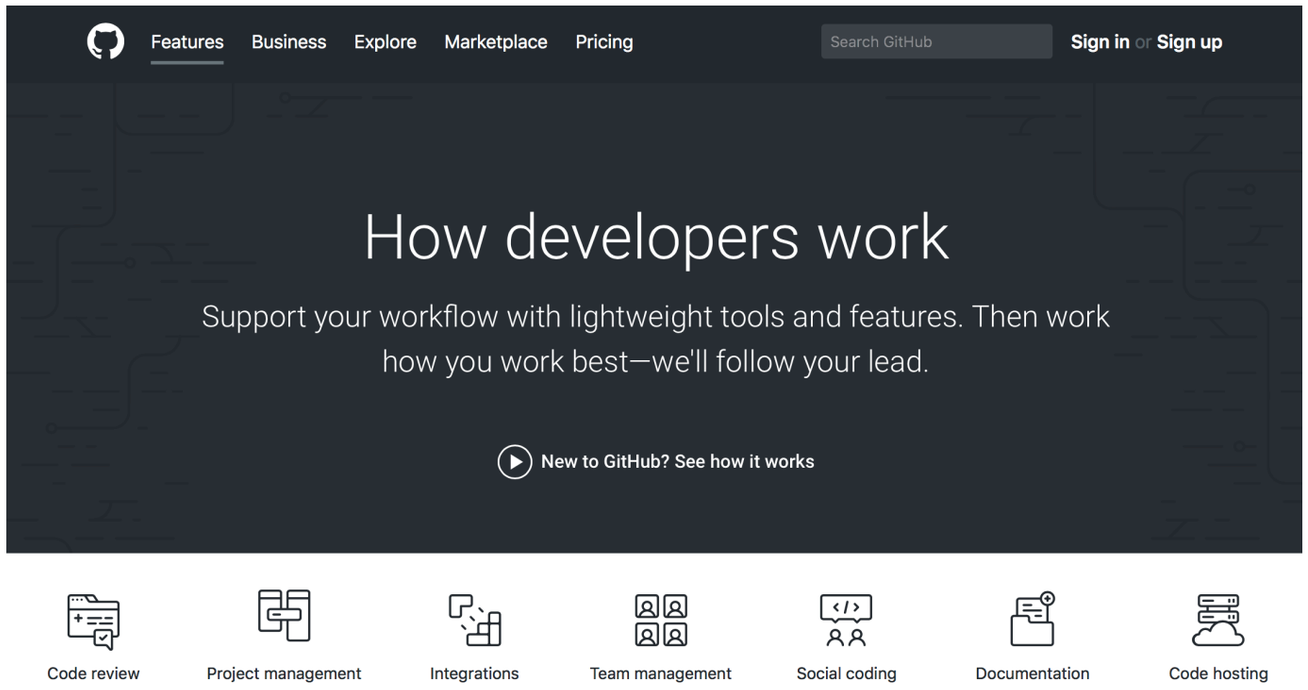
Anaconda Distribution

With over 6 million users, the open source Anaconda Distribution is the easiest way to do Python data science and machine learning. It includes hundreds of popular data science packages and the *conda* package and virtual environment manager for Windows, Linux, and MacOS. Conda makes it quick and easy to install, run, and upgrade complex data science and machine learning environments like Scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Services' *Machine Learning AMIs* and *Anaconda for Microsoft* on Azure and Windows.

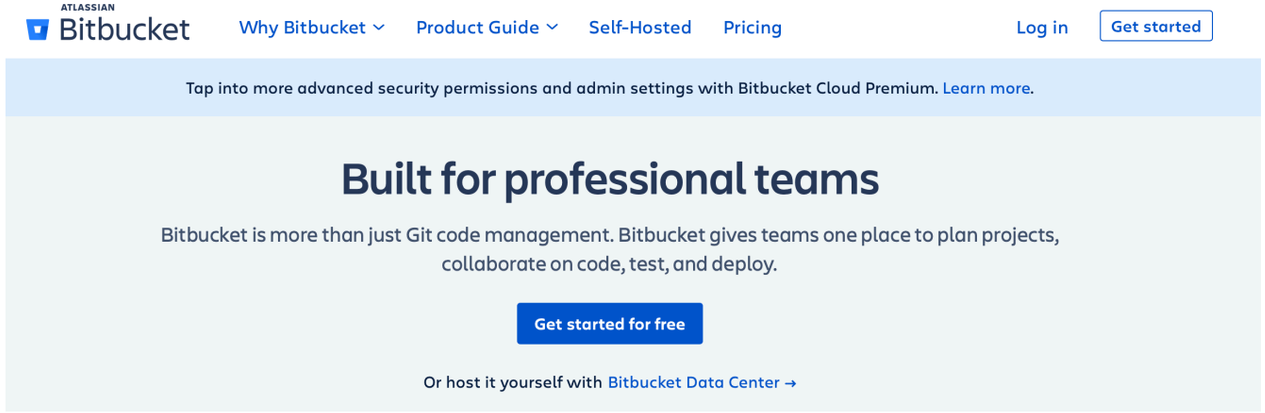
How to develop your code & share it?

Source-code repositories

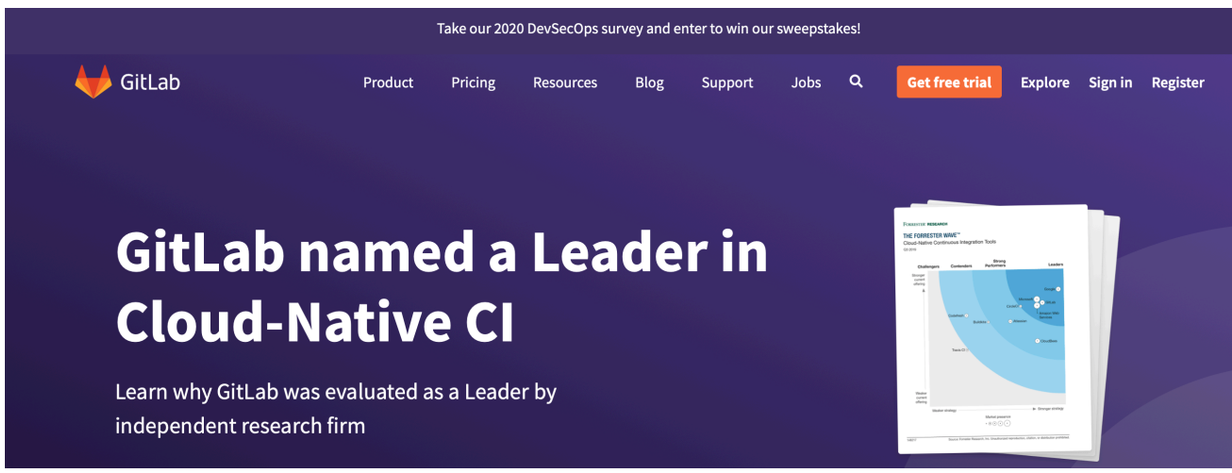
https://github.com



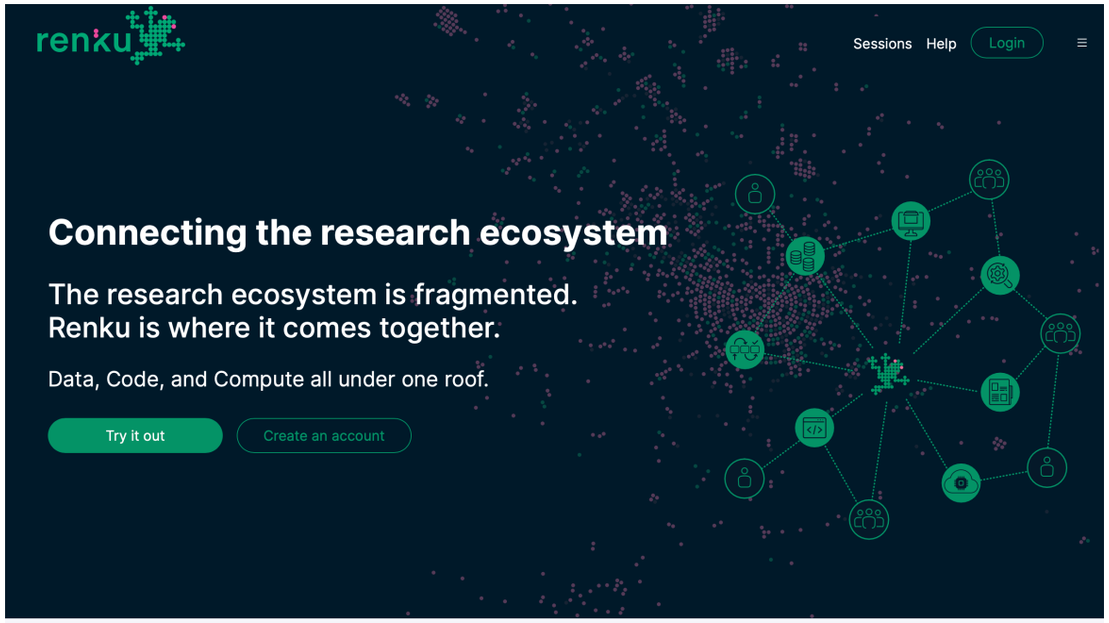
https://bitbucket.org



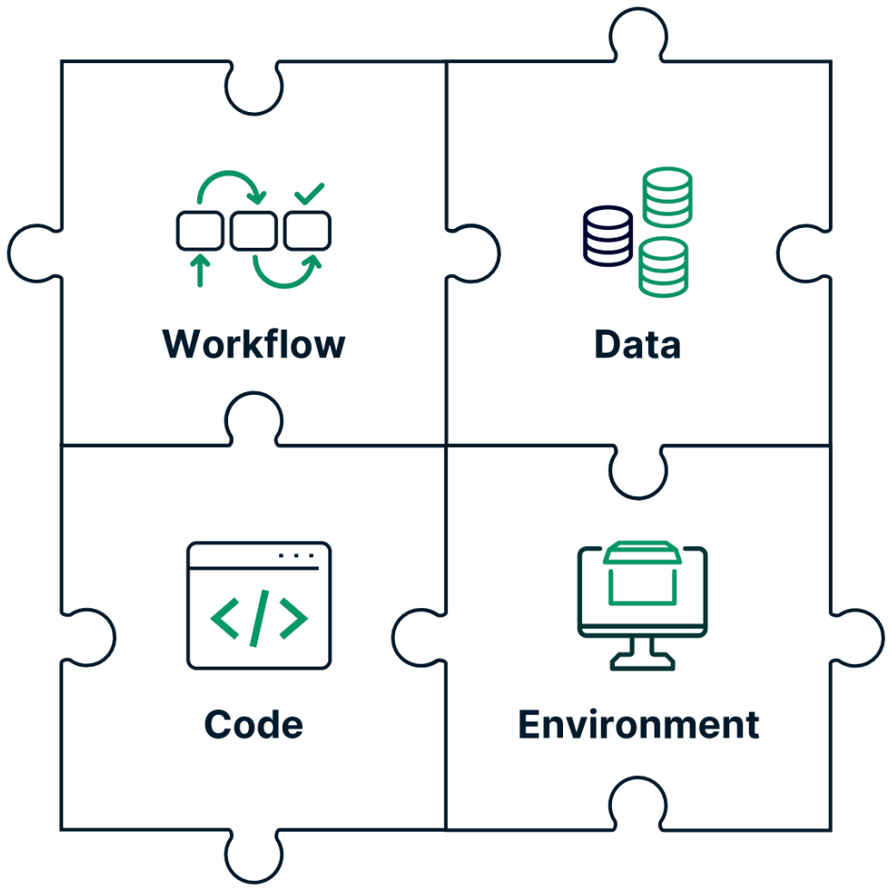
https://about.gitlab.com



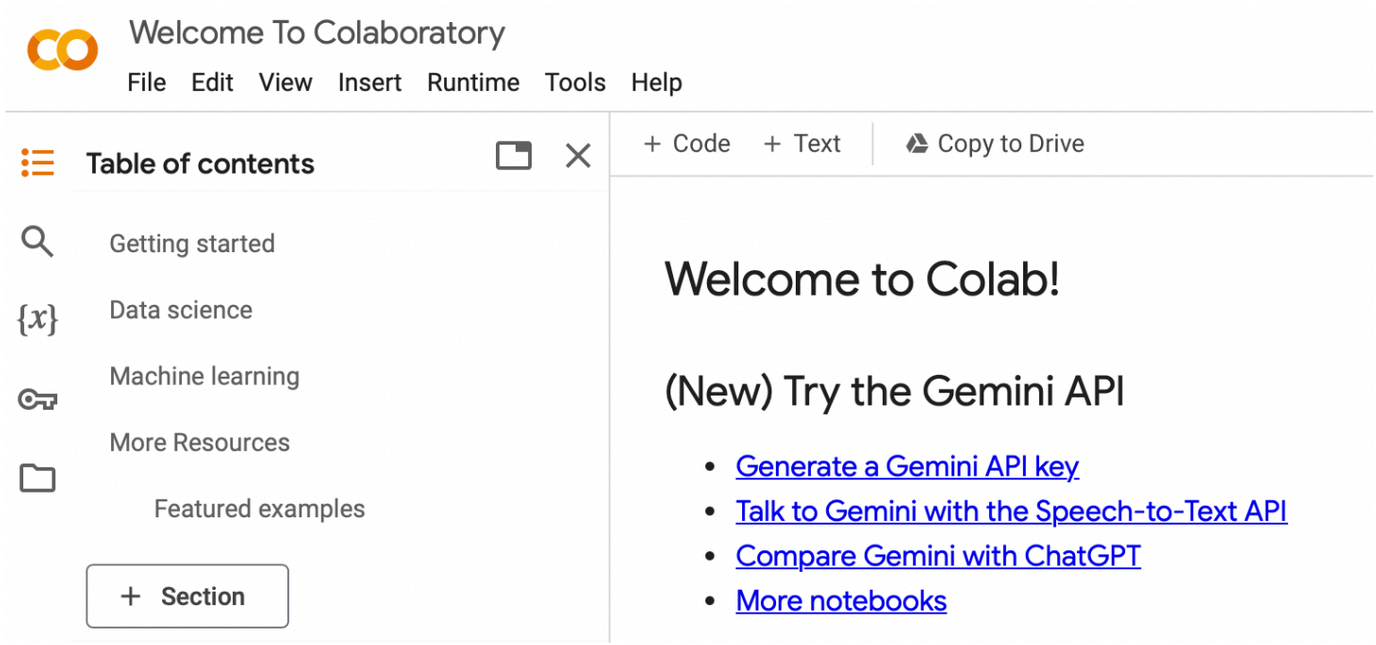
How to develop your code & share it?



Renku is open source and free to use!



https://renkulab.io/



https://colab.research.google.com/

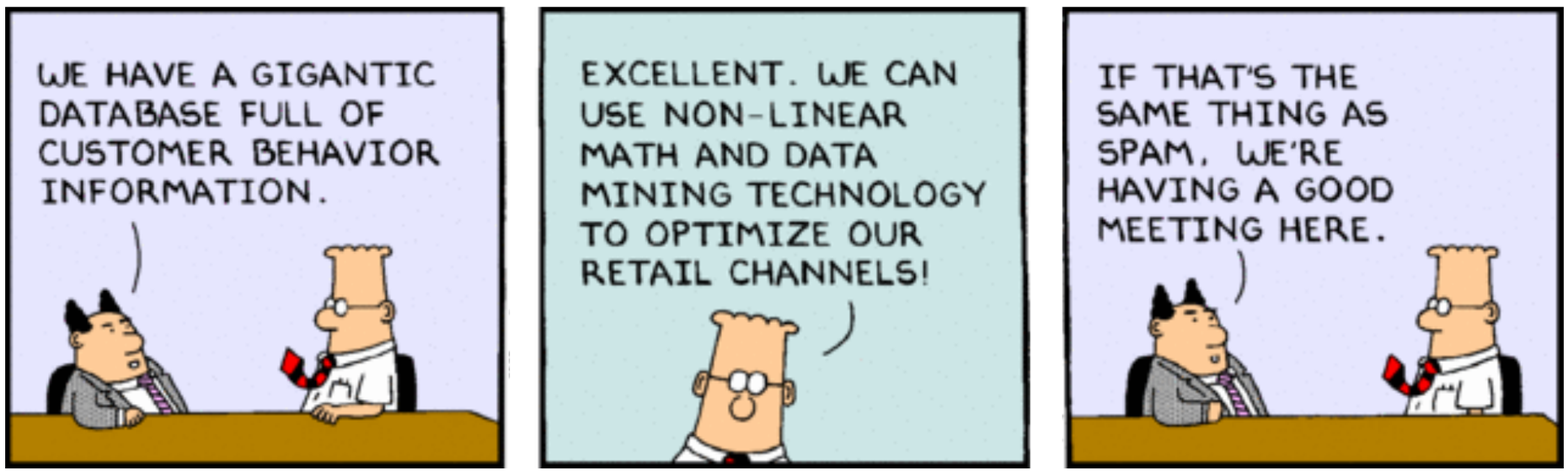
What will happen in this course?

Next lecture

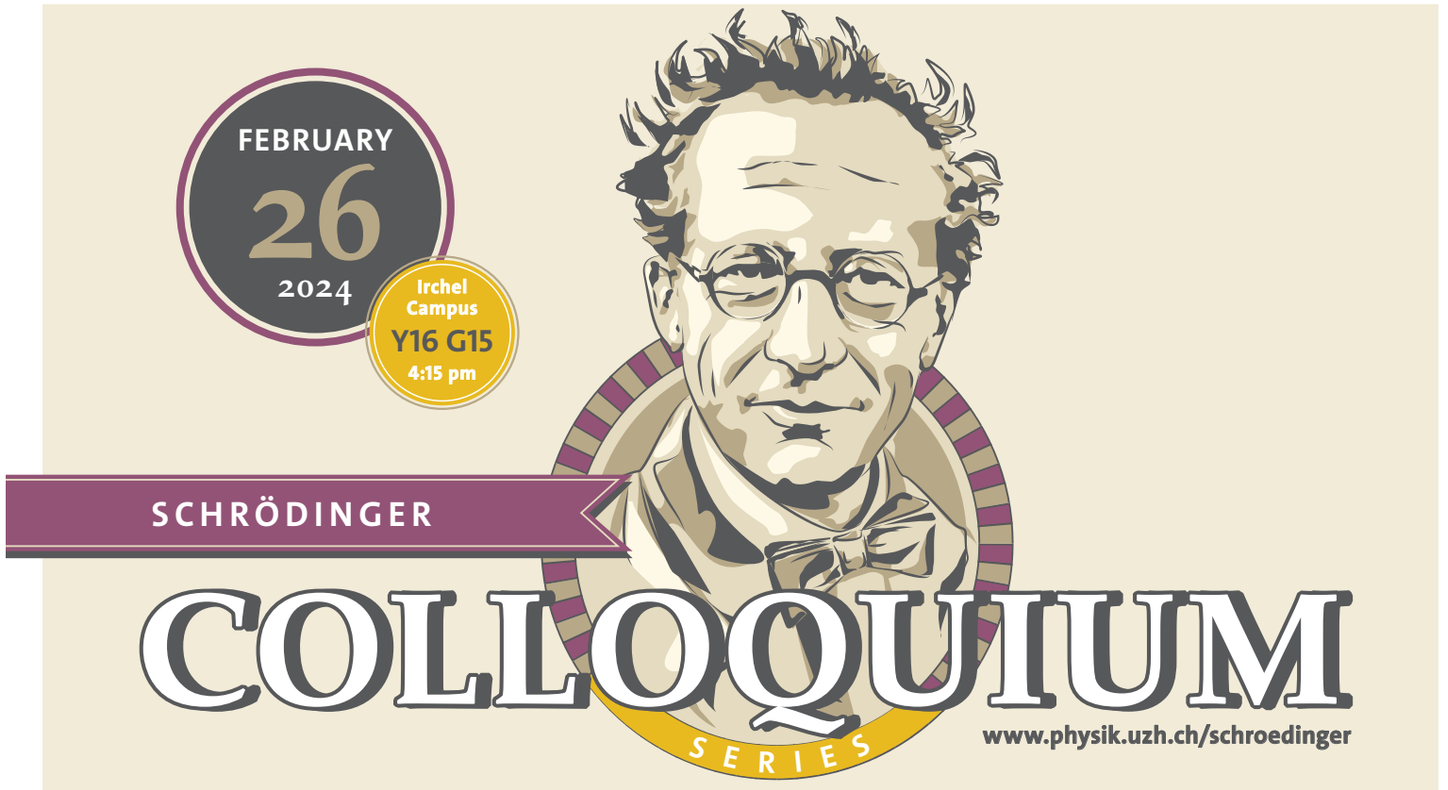
- cook-book for a data scientist: what to do? which order? what to avoid?

Subsequent lectures

- Introduction to statistical learning: regression, classification, validation, …
- Neural Networks / Deep Learning
- Bayesian Methods
- parallel & distributed computing
- Your presentation!



Faculty of Science



PROF. ROBERTO TROTTA SISSA & Imperial College London

The Promise of Machine Learning for Cosmology and Astroparticle Physics

Cosmological and astrophysical data are becoming sufficiently large and complex to soon prove intractable by traditional statistical techniques. Unravelling the twin mysteries of dark energy and dark matter will require new data analysis methods capable of extracting knowledge from upcoming data streams, including the Euclid satellite, the Nancy Grace Roman space telescope and LSST/Vera Rubin observatory. The recent rise of machine learning to prominence in the physical sciences promises to deliver the necessary tools — but questions remain about the scalability, interpretability and trustworthiness of the approach.

After an accessible introduction to machine learning in cosmology, I will present recent advances in simulation-based inference techniques for fast, scalable inference with guaranteed coverage. I will discuss a general, statistically principled solution to the ubiquitous problem of covariate shift in supervised learning, which achieves gold standard performance in a variety of settings, including supernova classification, photometric redshift estimation and weak lensing calibration. I will then give example applications in supernova type Ia cosmology, dark matter direct detection and gravitational waves astronomy.

A series of special physics colloquia in honor of Erwin Schrödinger, who was a professor at UZH from 1921 – 1927.
Lectures are intended for a broad audience from the Faculty of Science, aiming at experts and non-experts.