Please scan this code to register your presence on Ilias
for contact-tracing purposes. Your Ilias identity will be used.
– Only do this if you are *physically present*! –

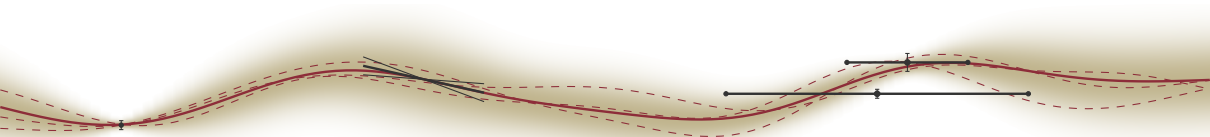
Data Literacy
Lecture 10
Sustainability and Data/AI

Philipp Hennig

10 January 2022

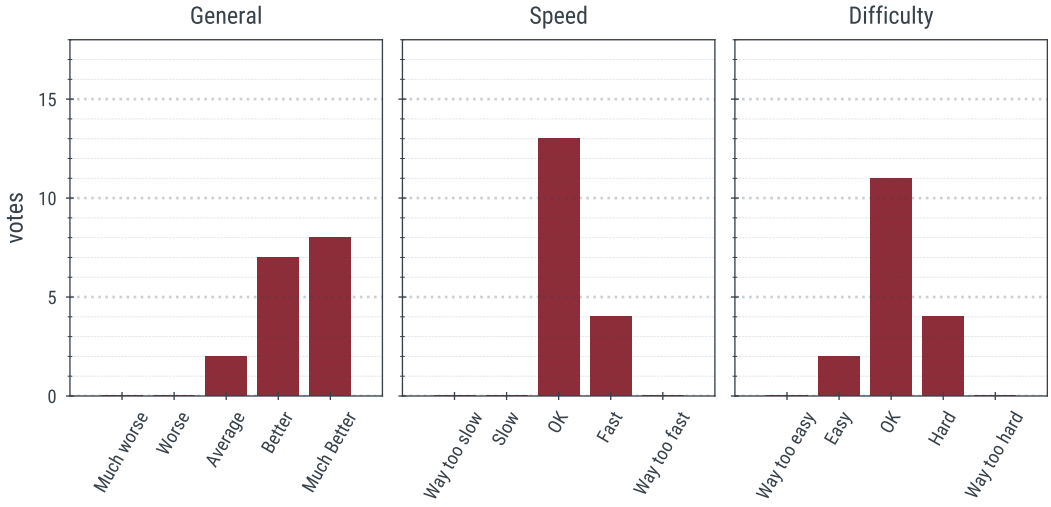
Eberhard Karls
Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



Feedback

quantitative



Detailed Feedback

your answers

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you *didn't* like

- there could have been more information about the data gender gap and more examples that show where data can be biased
- I really enjoyed today's lecture
- last two slides too fast

Things you liked

- The PCA example (6x)
- the political view and showing how politics treat the topic

Things you didn't understand

- The last parts on fairness and classification.
- Different definitions of fairness
- The ROC plots.

First Exam: 15 February 2022, 11:00–13:00
N5 + N3 + N4 (if necessary)

Please sign up for the exam on Ilias.

Second Exam: 12 April 2022, 09:30-11:30
B9N22

Reminder: You do not have to take the first exam to take the second one. But if you fail the first, you can take the second one.

Official evaluation available on Ilias. Main takeaways:

- people like:
  - the course overall
  - "engaging atmosphere"
  - examples
  - well-designed exercises
  - the "high-level" nature of the course
- people dislike:
  - mathematical topics are too tough / to fast
  - exercises are too easy
  - exercises are too hard
  - it is unclear which topics are important / what's needed for the exam
- proposed changes:
  - more introduction to exercise topics
  - remove the coding exercise, I don't think they are useful
  - spend more time on maths.
  - provide "*real* data literacy content"
- 11% of respondents (4/35) do not know there is a zoom call / recordings

We have to talk about Climate

especially in a course on data literacy

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

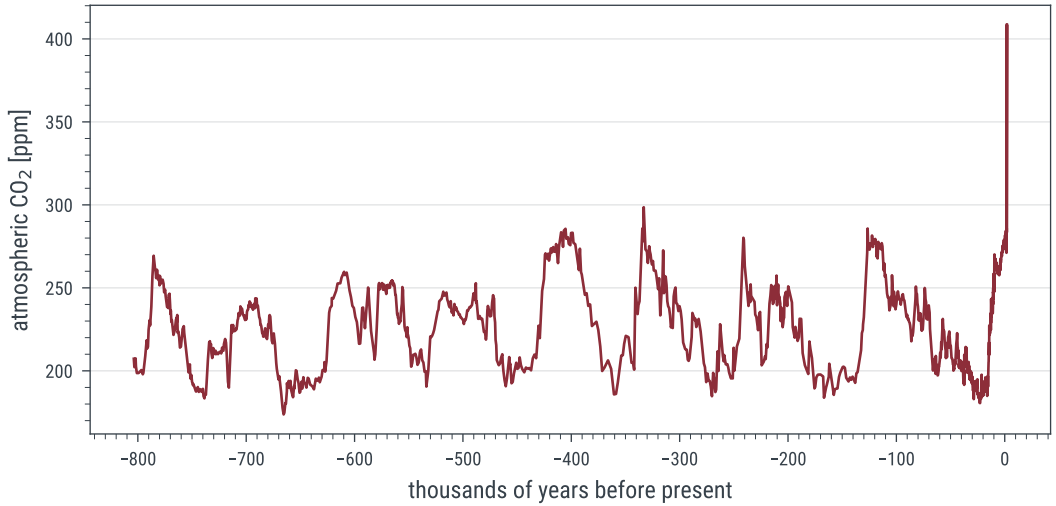
1. The climate desaster is the single-most existential crisis ever faced by mankind. Hence it *must* be addressed by all of us, especially those that want to be *experts in data*
2. Data is and will be crucial to allow us to understand and address the challenge of sustainability
3. Digital technology in general, and AI/ML in particular, are increasingly scrutinized as a major source of emissions. I will argue that
   - 3.1 since it is "naturally electric", relative to other sectors, digital technology is comparably easy to make sustainable, and has, relatively speaking, a low energy footprint. It should not be wrongly vilified.
   - 3.2 there *are* significant and urgent opportunities for efficiency gain in AI/ML in particular. Realizing them will make AI a better technology.
   - 3.3 AI/ML can be an enabling technology for a sustainable society.

As future experts on data science an ML, you should feel a responsibility to help society make sense of data, in particular on climate, and to help build technology the enables a sustainable future.

A Topic Nobody Can Avoid

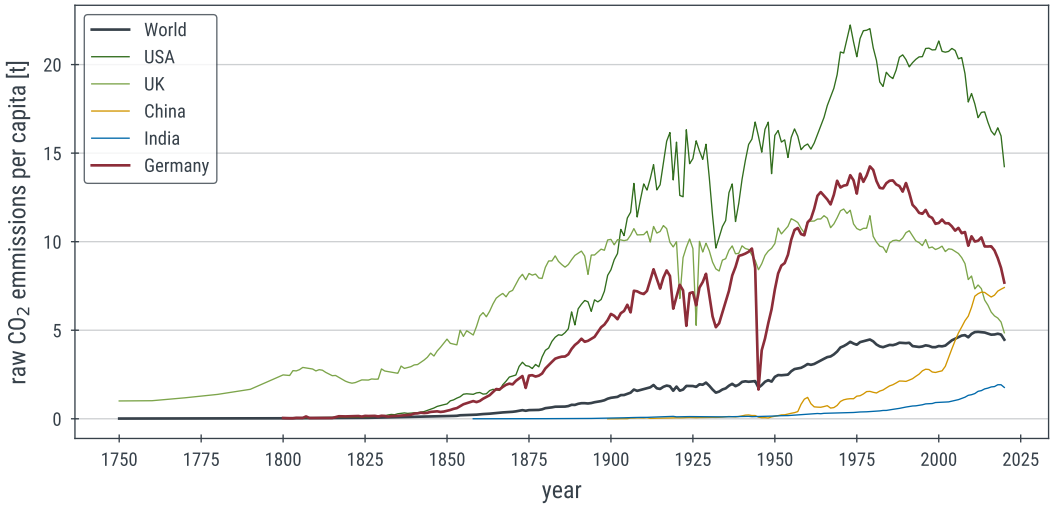
anthropogenic CO2 emmission is unpredented on geological scales

sources: NOAA Scripps and others, collected at our world in data



CO_2 was a byproduct of industrialization …

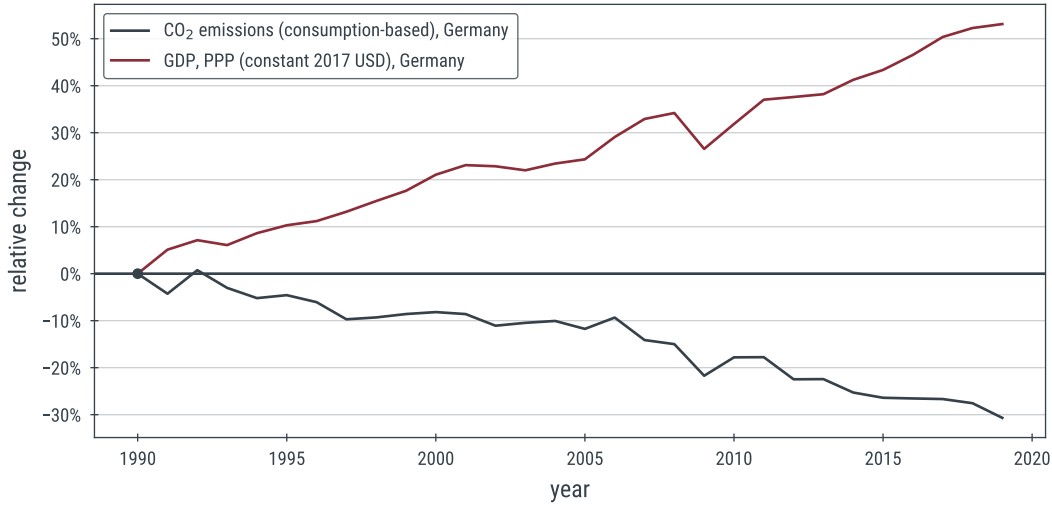
source: our world in data



…but has recently decoupled from economic growth

technological advancement and sustainability are not antagonists

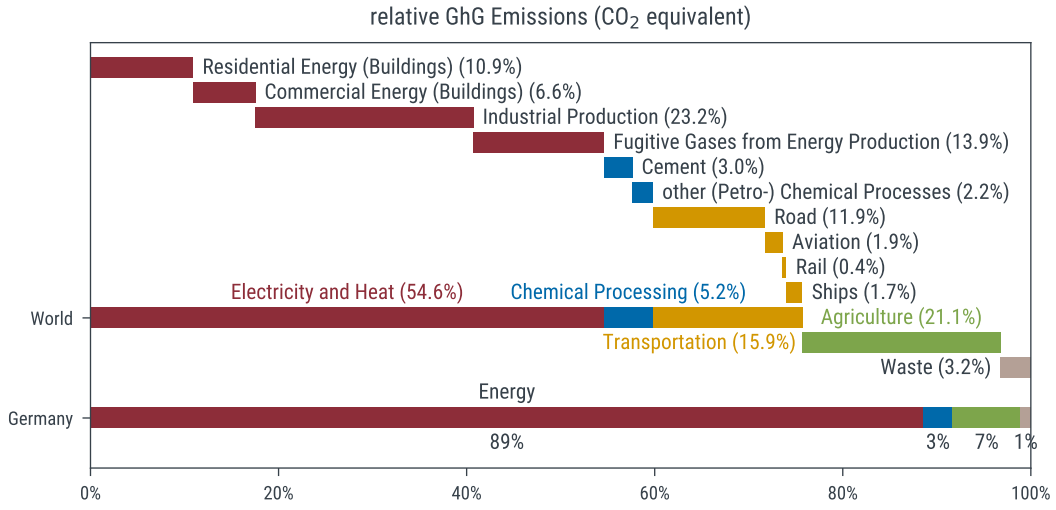
source: Global Carbon Project



Three sources make up 90% of our emissions.

thankfully, the largest sources are the easiest to change

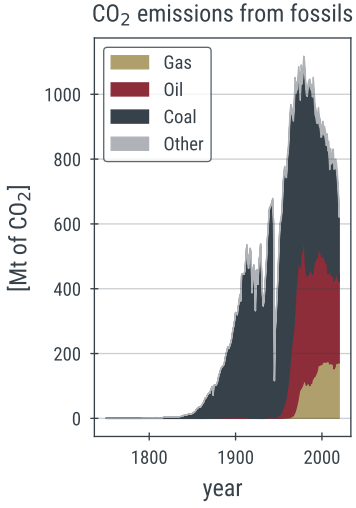
source: CAIT Climate Data Explorer, BMU

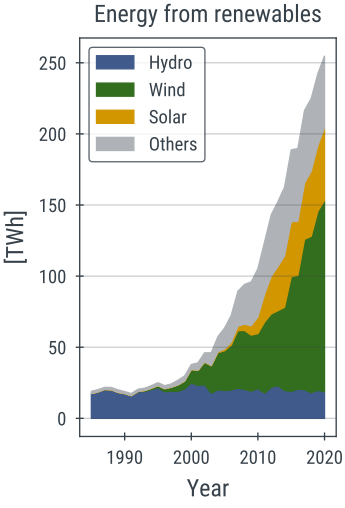


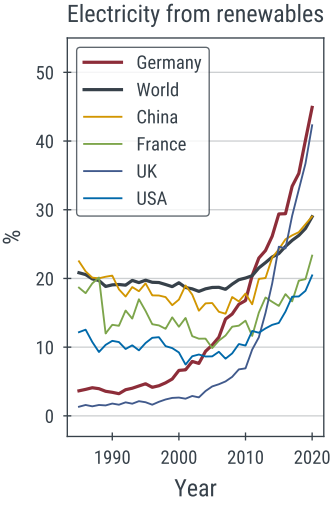
Electricity might rapidly be renewable

and with it all electric consumption, assuming it stays steady

source: our world in data

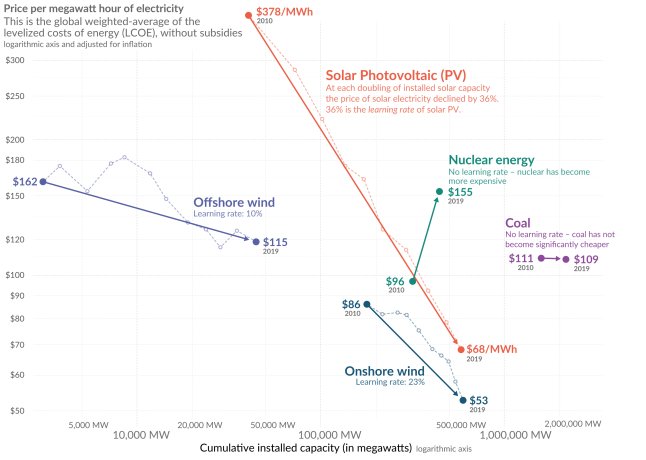






Renewables have already won

installing the remaining capacity needed will be cheaper than what we already have



Source: IRENA 2020 for all data on renewable sources; Lazard for the price of electricity from nuclear and coal – IAEA for nuclear capacity and Global Energy Monitor for coal capacity. Gas is not shown because the price between gas peaker and combined cycles differs significantly, and global data on the capacity of each of these sources is not available. The price of electricity from gas has fallen over this decade, but over the longer run it is not following a learning curve.

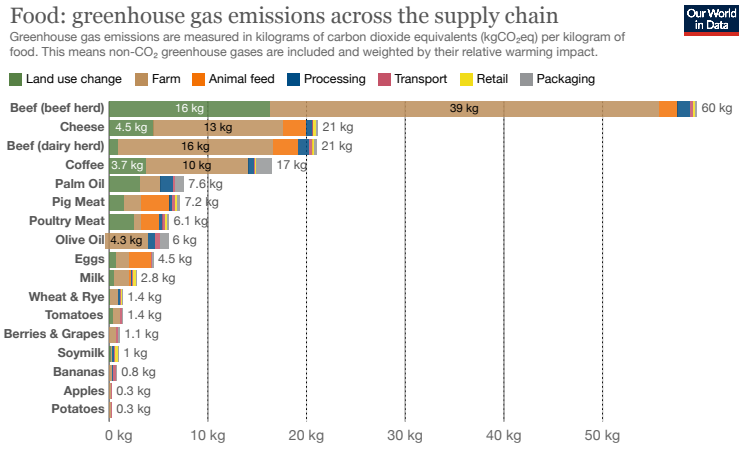
OurWorldinData.org – Research and data to make progress against the world's largest problems.

But other sectors pose tougher challenges

It is hard to regulate what people eat

source: owid

- Agricultural emissions are dominated by the farm itself, not transport
- Food produced by or from ruminating animals (cows, sheep, goats) is particularly bad because of their methane production.
- Achieving drastic reductions in GhG emissions requires essentially a vegan lifestyle, which is tough to regulate



Source: Poore, J., & Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. Science.
Note: Data represents the global median greenhouse gas emissions of food products based on a large meta-analysis of food production covering 38,700 commercially viable farms in 119 countries.
OurWorldInData.org/environmental-impacts-of-food • CC BY

Some takeaways so far:

- Our GhG emissions largely stem from
  - Electricity & Heat (domestic & industrial) (55%)
  - Agriculture (21%)
  - Transport (mostly Cars & Trucks) (16%)
  - Chemical Processes (5%)
  - Waste (3%)
- *All* these sources must be reduced. There is no silver bullet. But each sector requires *qualitatively different* actions. Consider two examples:

  Energy We must roughly quadruple our renewable electricity production, using wind and solar (and build or buy storage capacity). Doing so will be cheaper than what we have already invested, because wind and solar prizes have dropped exponentially with time. As much energy (and heat) as possible must be delivered as (or moved by) electricity or otherwise sustainably (electric cars, heatpumps, solar heating!).

  Agriculture A plant-based diet must be encouraged. This will be politically very hard.
- Everything that already runs on electricity will almost automatically become more sustainable over time (unless it grows faster than sustainable electricity!). Where technological solutions exist (cars!), they must be mandated. On the other end, air/sea transport and agriculture require much harder actions.

So what is the role of AI/ML and CS in all of this?

Die französische Non-Profit-Organisation The Shift Project schätzt, dass der gesamte Bereich der Informations- und Kommunikationstechnik (IKT) etwa 3,7 Prozent aller Treibhausgasemissionen weltweit verursacht und damit mehr als doppelt so viel wie die zivile Luftfahrt. **47 Milliarden Kilowattstunden Strom verbrauchen inzwischen allein in Deutschland Computer, elektronische Geräte wie Mobiltelefone, Tablets, Fernseher sowie die für den Einzelnen kaum sichtbaren Kommunikationsnetze und Rechenzentren. Auf sie entfielen 2017 bundesweit rund 13,2 Milliarden Kilowattstunden – damit verbrauchten sie ähnlich viel wie die Stadt Berlin.**

[…]

Allein schon beim Training einer Künstlichen Intelligenz (KI) zur Spracherkennung fällt fünfmal so viel CO2 an, wie ein Auto während seiner gesamten Lebensdauer ausstößt.

Der Tagesspiegel, 6.11.2019

80 Prozent des weltweiten Datenverkehrs im Internet fallen laut The Shift Project auf das Streaming von Videos. 2018 seien dadurch ungefähr so viel CO2-Emissionen ausgestoßen worden wie in ganz Spanien. 45 Prozent verursachten dabei die Produzenten, 55 Prozent die Verbraucher. **Wer ein nur zehnminütiges Youtube-Video schaut, verbrauche ähnlich viel Energie, als würde man fünf Minuten lang einen elektrischen 2000-Watt-Ofen im Hochbetrieb laufen lassen** […] Dabei verbrauchen sie umso mehr Daten, je höher die Videos aufgelöst sind. Dieter Janecek von den Grünen sieht politischen Handlungsbedarf: **„Anbieter wie Youtube sollten dazu verpflichtet werden, ihre Filme nicht in der höchstmöglichen, sondern einer niedrigeren Auflösung zu zeigen."**

Der Tagesspiegel, 6.11.2019

Is AI a climate killer?

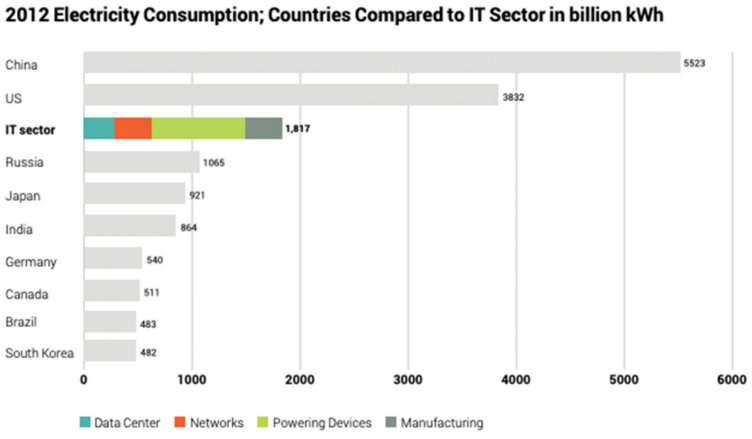
a true challenge for data literacy

> Etwa 33 Millionen Tonnen CO2-Emissionen im Jahr werden durch den Betrieb des Internets und internet-fähiger Geräte in Deutschland verursacht – so viel wie durch den innerdeutschen Flugverkehr.

"Bits und Bäume", Sabine Langkau & Sven Hilbig

So many big numbers!

Some real data literacy required



Source: Emerging Trends in Electricity Consumption for Consumer ICT, Peter Corcoran and Andres Andrae (2013) and CIA World Factbook. China/Russia/Canada figures are from 2014.

So many big numbers!

Some real data literacy required

source: Höfner & Frick. Bits und Bäume, 2019

The Politicians are Acting

Koalitionsvertrag 2021 – Emphasis mine

Nachhaltigkeit in der Digitalisierung

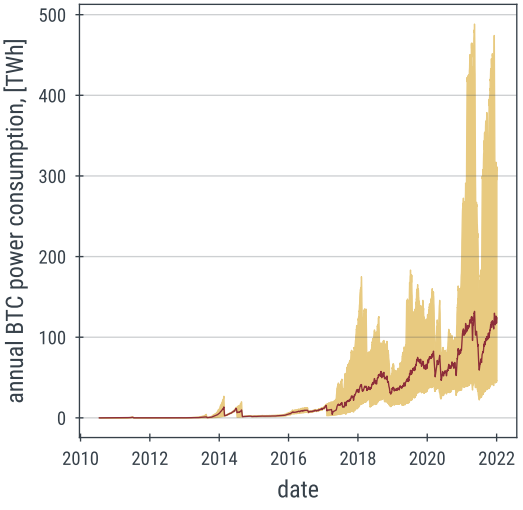
Wir wollen die Potentiale der Digitalisierung für mehr Nachhaltigkeit nutzen. Durch die Förderung digitaler Zwillinge (z. B. die Arbeit an einem virtuellen Modell eines analogen Produktes) helfen wir den Verbrauch an Ressourcen zu reduzieren. Wir werden Rechenzentren in Deutschland auf ökologische Nachhaltigkeit und Klimaschutz ausrichten, u. a. durch Nutzung der Abwärme. Neue Rechenzentren sind ab 2027 klimaneutral zu betreiben. Öffentliche Rechenzentren führen bis 2025 ein Umweltmanagementsystem nach EMAS (Eco Management and Audit Scheme) ein. Für IT- Beschaffungen des Bundes werden Zertifizierungen wie z. B. der Blaue Engel Standard. Ersatzteile und Softwareupdates für IT-Geräte müssen für die übliche Nutzungsdauer verpflichtend verfügbar sein. Dies ist den Nutzerinnen und Nutzern transparent zu machen.

There *are* power-hungry uses of compute

The power consumption of bitcoin

source: Cambridge Bitcoin Electricity Consumption Index

- The bitcoin network uses about 90 TWh per annum
- This is 7× Google's annual consumption (13TWh, source: NYT)



Time for some Numbers

It's true! Electronic devices do use a nontrivial amount of *electric* energy.

The network:

- T-Com self-reports a power consumption of 120 kWh / TB for data transfer to the consumer.
- *total* power consumption of their network has stayed relatively constant over 2016–2019

source: `https://www.cr-report.telekom.com`

The consumer:

- Mobile Phones (iphone 13 Pro Max): 4.3Ah @ 3.9 V = 16.8Wh battery, "lasts a day": average power consumption < 1W
- Laptop: MacBook Pro 16" 2021: About 9Ah @ 11.45 V = 100Wh battery: average power consumption $\sim$ 4W. Peak (benchmarks) up to 110W

source: `https://www.mtech.news/`

- NVidia Tesla V100 (just the GPU): 250 W
- 8 GPU workstation at peak: 2kW

Context:

- German *electricity* consumption 2020: 573.6 TWh
  i.e. average electricity draw per capita: ca. 800W

source: statista

Example I: virtual meetings

even local digital meetings can quickly be a net carbon reduction

- A zoom group call at 1080p resolution uses about 2.4 GB/hr, *for each participant*. Let's assume everyone's local setup draws 100W to manage the call.

$$0.12\text{kWh/GB} \cdot 2.4\text{GB/h/participant} + 0.100\text{kW/participant} = 0.4\text{kWh/h/participant}$$

- Both cars and planes use about `withouthotair.com`

$$5\text{l/100km/passenger} = 50\text{KWh/100km/passenger} = 11.5\text{kg}CO_2\text{/100km/passenger}.$$

(Basically all hydrocarbon fuels are about 10 kWh/l. Gasoline burns to 2.3kg CO_2 per litre. Very efficient electric cars with regenerative breaking may reach 25 KWh/100km/passenger)

- (Full) Trains are extremely energy efficient and use about 2 kWh/ 100km/passenger

Example I: virtual meetings

even local digital meetings can quickly be a net carbon reduction

- A zoom group call at 1080p resolution uses about 2.4 GB/hr, *for each participant*. Let's assume everyone's local setup draws 100W to manage the call.

$$0.12\text{kWh/GB} \cdot 2.4\text{GB/h/participant} + 0.100\text{kW/participant} = 0.4\text{kWh/h/participant}$$

- Both cars and planes use about `withouthotair.com`

$$5\text{l/100km/passenger} = 50\text{KWh/100km/passenger} = 11.5\text{kg}CO_2\text{/100km/passenger}.$$

(Basically all hydrocarbon fuels are about 10 kWh/l. Gasoline burns to 2.3kg CO_2 per litre. Very efficient electric cars with regenerative breaking may reach 25 KWh/100km/passenger)

- (Full) Trains are extremely energy efficient and use about 2 kWh/ 100km/passenger

$$\frac{0.4\text{ kWh / h / participant}}{50\text{ KWh/100car km/participant}} = 8\text{car km/h} \qquad \frac{0.4\text{ kWh / h / participant}}{2\text{ KWh/100km/participant}} = 20\text{ train km/h}$$

Example I: virtual meetings

even local digital meetings can quickly be a net carbon reduction

- A zoom group call at 1080p resolution uses about 2.4 GB/hr, *for each participant*. Let's assume everyone's local setup draws 100W to manage the call.

$$0.12\text{kWh/GB} \cdot 2.4\text{GB/h/participant} + 0.100\text{kW/participant} = 0.4\text{kWh/h/participant}$$

- Both cars and planes use about `withouthotair.com`

$$5\text{l/100km/passenger} = 50\text{KWh/100km/passenger} = 11.5\text{kg}CO_2\text{/100km/passenger}.$$

(Basically all hydrocarbon fuels are about 10 kWh/l. Gasoline burns to 2.3kg CO_2 per litre. Very efficient electric cars with regenerative breaking may reach 25 KWh/100km/passenger)

- (Full) Trains are extremely energy efficient and use about 2 kWh/ 100km/passenger

$$\frac{0.4\text{ kWh / h / participant}}{50\text{ KWh/100car km/participant}} = 8\text{car km/h} \qquad \frac{0.4\text{ kWh / h / participant}}{2\text{ KWh/100km/participant}} = 20\text{ train km/h}$$

For every hour you expect to be in the call, you can drive 8km by car. Or 20km by train.
If your car burns fuel, and your office uses sustainable electricity, take your bike! Don't bring your laptop, though!

Example II: AI Research

source: MIT Tech Review, 6 June 2019 / Strubell, Ganesh, McCallum, ACL 2019

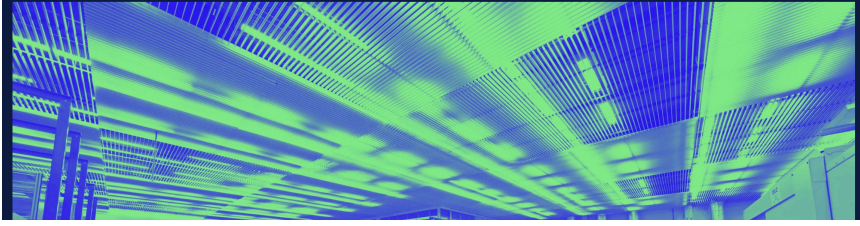
ARTIFICIAL INTELLIGENCE

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

By Karen Hao

June 6, 2019



Example II: AI Research

source: MIT Tech Review, 6 June 2019 / Strubell, Ganesh, McCallum, ACL 2019

In a new paper, researchers at the University of Massachusetts, Amherst, performed a life cycle assessment for training several common large AI models. They found that the process can emit more than 626,000 pounds of carbon dioxide equivalent—nearly five times the lifetime emissions of the average American car (and that includes manufacture of the car itself).

Common carbon footprint benchmarks

in lbs of CO2 equivalent

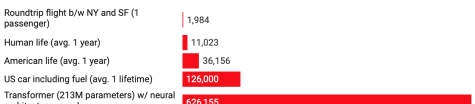


Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

The estimated costs of training a model once

In practice, models are usually trained many times during research and development.

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Example II: AI Research

Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)

Example II: AI Research

Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)
- Net CO_2 emissions: 0 kg (Uni uses renewable energy)

Example II: AI Research

Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)
- Net CO_2 emissions: 0 kg (Uni uses renewable energy)
- German energy mix in 2019:
  400 g/kWh $\times$ 5 MWh = **2000kg CO_2** (2020: 370g/kWh)

Example II: AI Research

Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)
- Net CO_2 emissions: 0 kg (Uni uses renewable energy)
- German energy mix in 2019:
  400 g/kWh $\times$ 5 MWh = **2000kg CO_2** (2020: 370g/kWh)
- flights from and to conference (FRA $\leftrightarrow$ YVR): **5000 kg CO_2** per Person

Example II: AI Research

Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)
- Net CO_2 emissions: 0 kg (Uni uses renewable energy)
- German energy mix in 2019:
  400 g/kWh $\times$ 5 MWh = **2000kg CO_2** (2020: 370g/kWh)
- flights from and to conference (FRA $\leftrightarrow$ YVR): **5000 kg CO_2** per Person
- videostreaming of the conference: 0.4 kWh/h $\cdot$ 0.4 kgCO_2 / kWh = 160g/h.
  2 people, 1 week: $<$ **20kg** total

Example II: AI Research

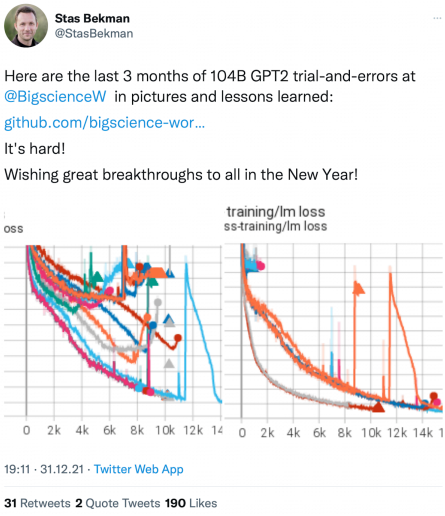
Even resource-hungry AI has a low carbon footprint compared to the human activities associated with it

- Recent large-scale paper: 5MWh (250 W/GPU $\times$ 20 000 GPU h)
- Net CO_2 emissions: 0 kg (Uni uses renewable energy)
- German energy mix in 2019:
  400 g/kWh $\times$ 5 MWh = **2000kg CO_2** (2020: 370g/kWh)
- flights from and to conference (FRA $\leftrightarrow$ YVR): **5000 kg CO_2** per Person
- videostreaming of the conference: 0.4 kWh/h $\cdot$ 0.4 kgCO_2 / kWh = 160g/h.
  2 people, 1 week: $<$ **20kg** total

If you are working with large-scale AI systems, your work is probably associated with significant *electric* energy demand (probably comparable to that of your entire private life).

We still have to do our homework!

Contemporary ML is extremely inefficient with human *and* natural resources



- ML continues to require trial and error. Algorithms that eliminate this process potentially reduce energy demand by $>90\%$.
- The goals of science (understanding what's going on) and sustainability (stop try and error) are aligned
- Most of the problem is with *algorithms*. Optimization, integration, linear algebra, simulation. Interested?

Numerics of Machine Learning
Winter Term 2022/23
In ML, methods research is sustainability work.

But AI *can* be a force for good

it's up to you how you want to use your education

If *you* want to contribute to sustainability through AI, consider

- improving data access in the public sector, so we can all understand better where energy is used
- build better control algorithms for the electricity grid and smart buildings
- advance virtual reality so people want to travel less
- help automate material science to find cheaper, lighter batteries
- build autonomous systems for precision agriculture/forestry to improve crop yield
- join research on controlling fusion reactors
- advance simulation methods to find stable, lightweight constructions for smart architecture
- …

…or join foundational methods research to make AI more resource efficient.

Summary:

- Humanity will *not* return to pre-industrial living. We need solutions that make our current lifestyle sustainable.
- Electricity and Heating are currently the primary source of emissions. Fortunately, they are also the easiest to make sustainable.
- Digital Technology is a nontrivial part of overall energy demand. But it is naturally electric. The primary goal must be to keep energy demand constant while technology grows.
- By replacing physical actions with digital/virtual ones; by improving the efficiency of physical processes; and by improving transparency of the energy cycle, ML and Data Science can contribute crucially to the sustainable society. But we also actually have to do it, not just say that it'll happen.

Please provide feedback:

If *you* want to help

1. pressure society & politics to scale up renewable energy sources, *fast*.
2. invest your technical skills in technology for sustainability, and sustainable technology.

Your Projects

Your homework for the rest of term

- From today, there will be no exercise sheets
- instead, in groups of ≤ 2:
  - find a dataset (examples on next slide)
  - pre-register your analysis by **this Friday**, on Ilias, see upcoming mail
  - write a short report (4 pages, NeurIPS style) detailing
    - the question being answered
    - explain your data collection
    - **make a visualization and/or do *some* statistical analysis** (you can use methodology not covered in the lecture, *if you can properly support/cite/explain it*.)
    - make decent plots (lecture 12)
    - discuss limitations/confidences/problems
    - support your argument with citations
  - provide a git repo holding your analysis (lecture 11)
- **Submit your report & repo by 7 February**

The project will be graded. The grade of the project makes up 20% of the overall grade (the other 80% arise from the exam)

Some Data Sources

feel free to find your own

clickable links!

- Energy Data for Germany by the BMWI
- Agri-environmental indicators by country
- All results of the 2021 Federal Election (currently only to voting district level, from "January 2022" also to polling station level).
- Any Covid Data. For an (unreasonably good) example, consider the Economist's excess death tracker with open data on github (thanks to Dmitry Kobak)
- All speeches ever given in the Bundestag.
- A collection of remote work salaries, and of anonymously collected AI/ML salaries (make a choropleth! But don't stop with it!)
- Subtitles for cinema and television (e.g. for visualizations like this)
- Someone collected about 200k illegal drug listings on the darknet, with prices.
- Spotify features for about 1.5M songs (or collect your own from the API)
- Make a map like this, for Tübingen
- use the python APIs for wikipedia, arxiv, twitter, google bigquery, etc.
- or just to go to https://www.reddit.com/r/datasets/

What are we looking for?

in your projects

- Project teams will be assigned to a tutor after this week's pre-registration
- This is the first time we do such projects. Feel invited to experiment. If unsure, contact your tutor.
- Projects can contain both visualization and analysis. If you decide to only do one, make sure you do it well. *Some* graphical content (plots) is needed in either case.
- The main goal of this part of the course is for you to collect experience working with raw data. That is, in particular
  - accessing data procedurally (i.e. using code, not a spreadsheet)
  - structuring a data-centric project in code / repo
  - dealing with messy real-world data
  - identifying things to talk about, and supporting them meaningfully with data, citations, analyses
  - presenting your analysis well, graphically (make 1–3 nice plots)
  - writing a (short) academic text, in decent English, with proper layout, and clean citations.
- The following weeks will cover some of the above, hopefully just in time for you to use it.