Big Data, organisation and analysis

A second view on Big Data

Steffen M. Noe, Spring 2025

The Big V's

Big data is categorised into the four (or five, or six) V's

Volume

- Quantity of the generated and stored data
- Size usually larger than Terabytes
- Determines potential value and insight of the data

Variety

- Type and nature of the data
- mostly unstructured and semi-structured data
- Still applies to very big structured data

Velocity

- Speed at which data is generated or processed
- often real-time availability
- Continuous data production

Veracity

- Truthfulness an reliability
- often seen as "data quality"
- affects analysis possibilities

Value

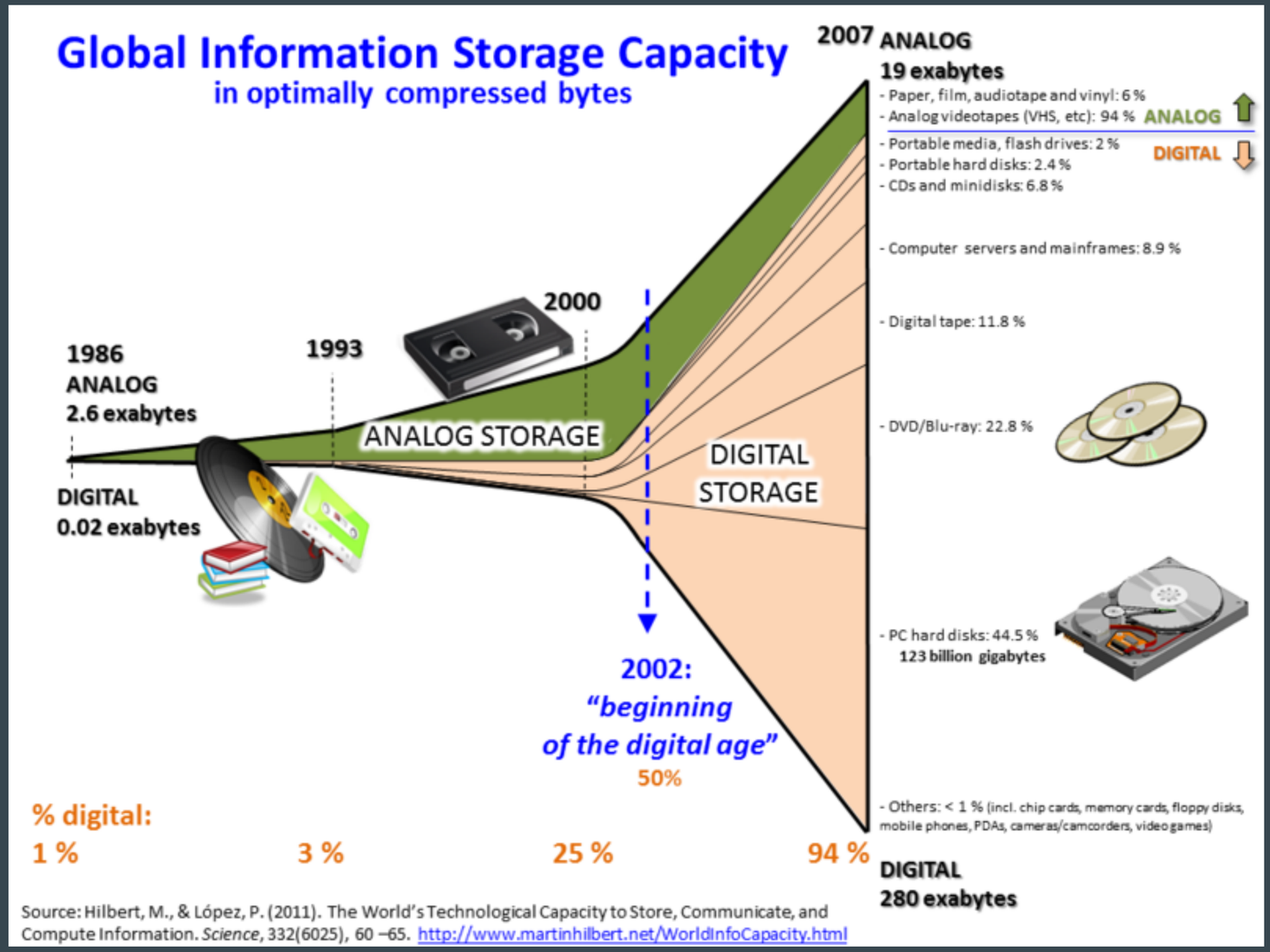
- What "worth" can be extracted?
- Depends on volume, variety and veracity

Variability

- Characterises change in formats and structures
- including different sources

Some history

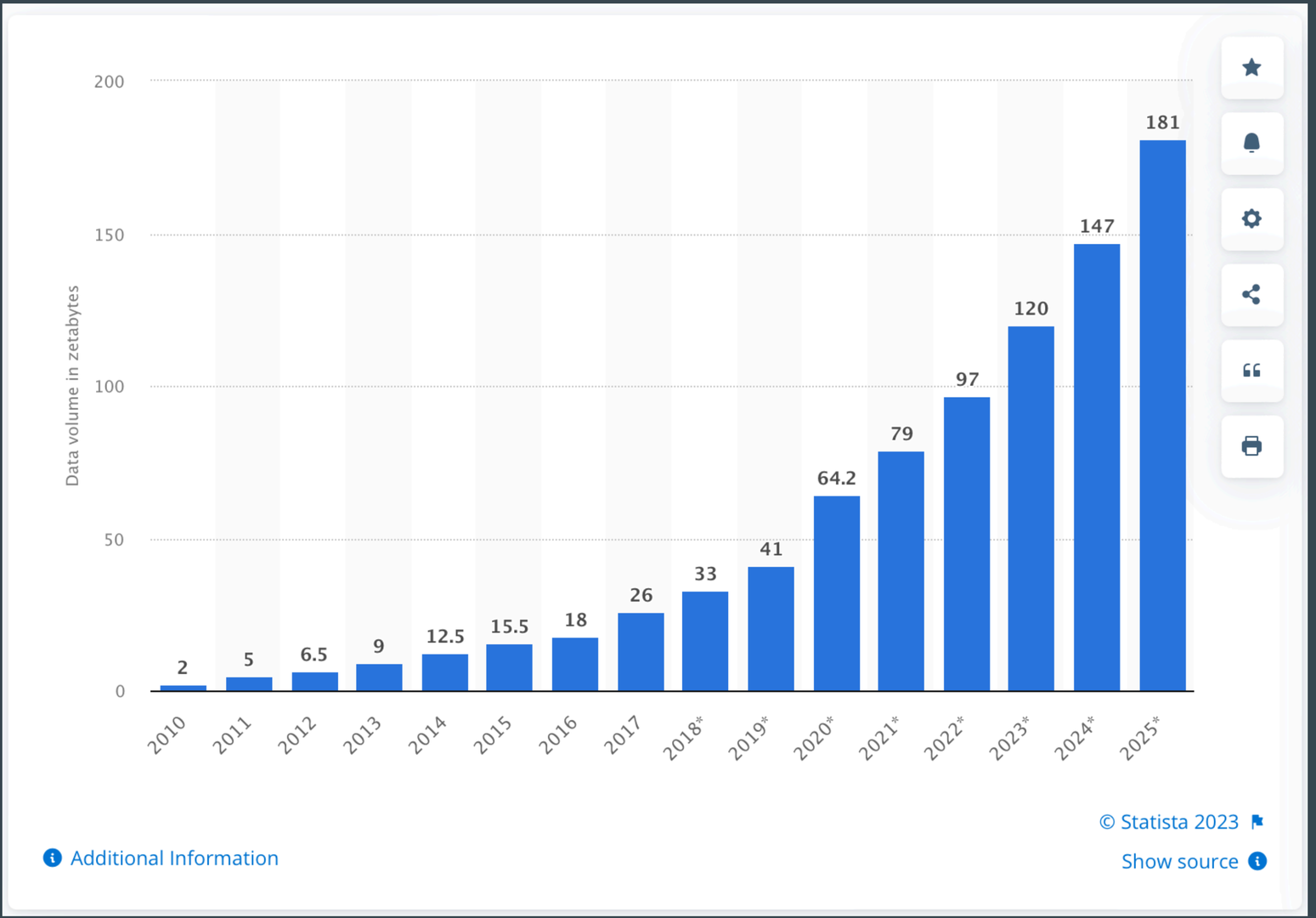
The transformation from analog to digital data



https://en.wikipedia.org/wiki/Big_data

A view to the future

Data is expected to grow exponentially further



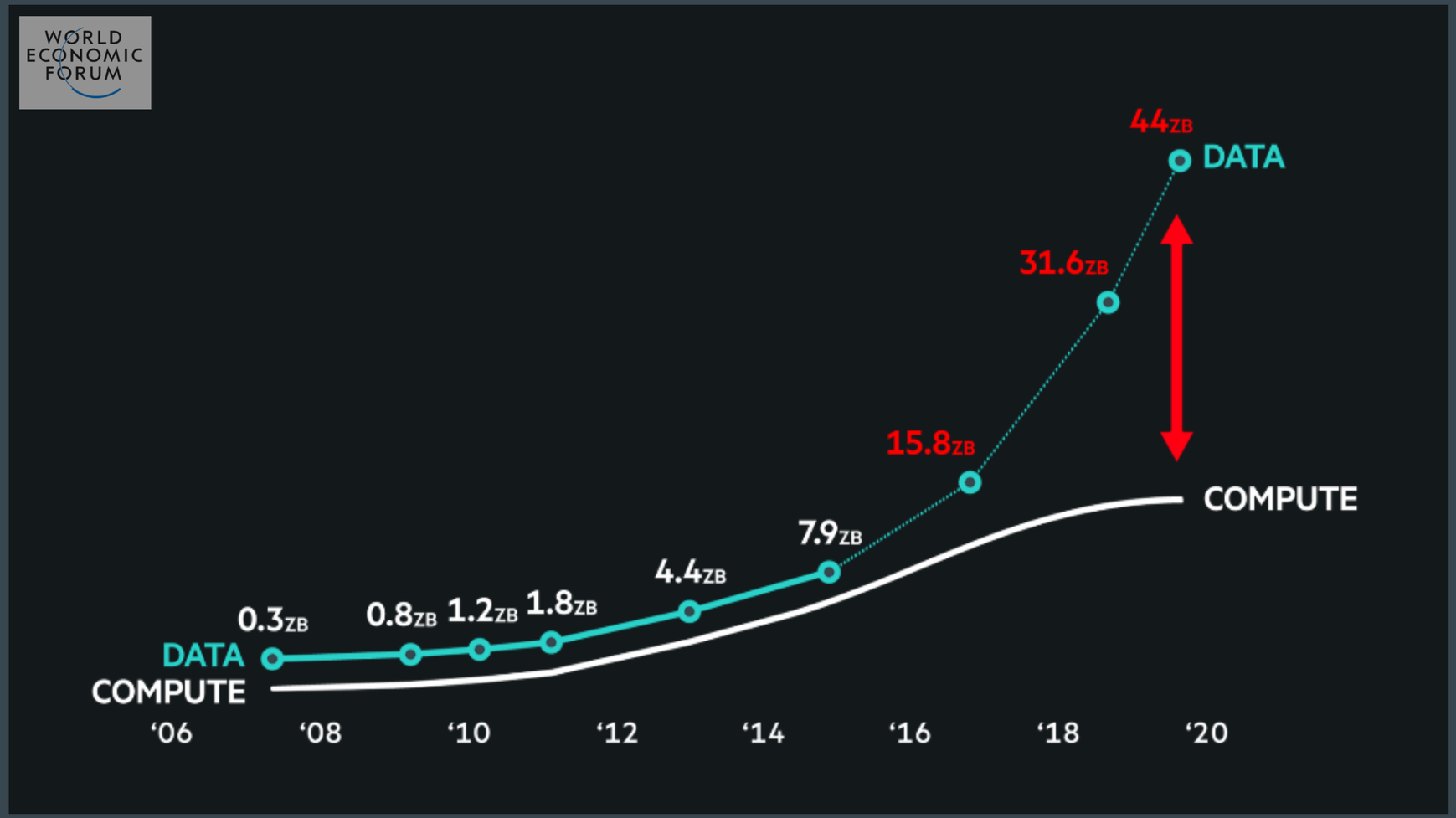
Will there be some limits?

- The most likely ones are of technical / physical nature
- The still ongoing sensor development lead to more and cheaper ones, a limit may be the speed of data transfer from sensor to storage
- Because Big Data also tend to combine different data sources there is growth of “secondary data” even without new data production
- For such derived data, produced by e.g., machine learning, statistical models, forecast process models the limits are the available computing power and time

The correlation between computing power and data

"Every two years, we create more data than we've created in all history" Kirk Bresniker, 2018

- Data grows exponential
- Moore's law slow's down.
- The computing power does not follow the exponential growth anymore



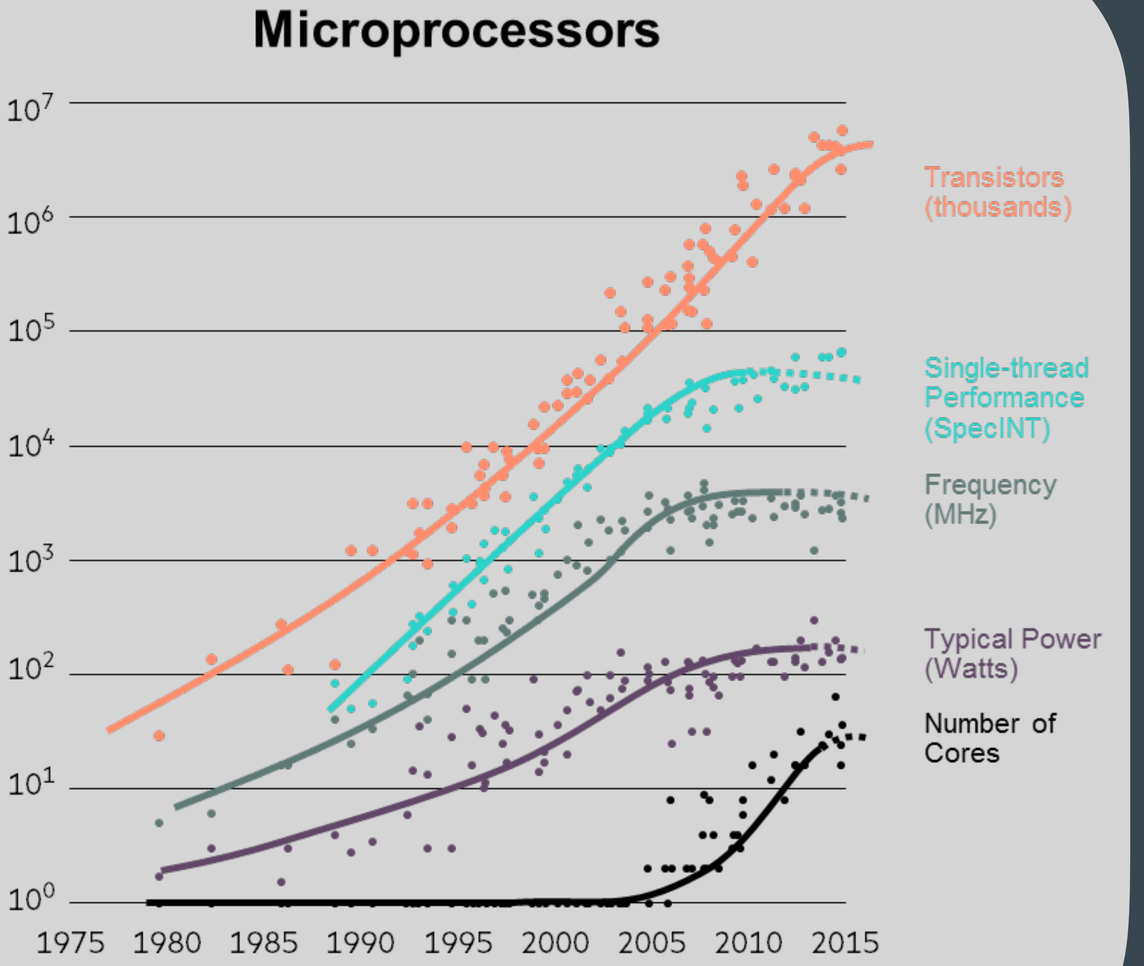
https://www.weforum.org/agenda/2018/09/end-of-an-era-what-computing-will-look-like-after-moores-law/

It seems we are reaching some physical limits

MOSFET or Dennant scaling

https://www.weforum.org/agenda/2018/09/end-of-an-era-what-computing-will-look-like-after-moores-law/

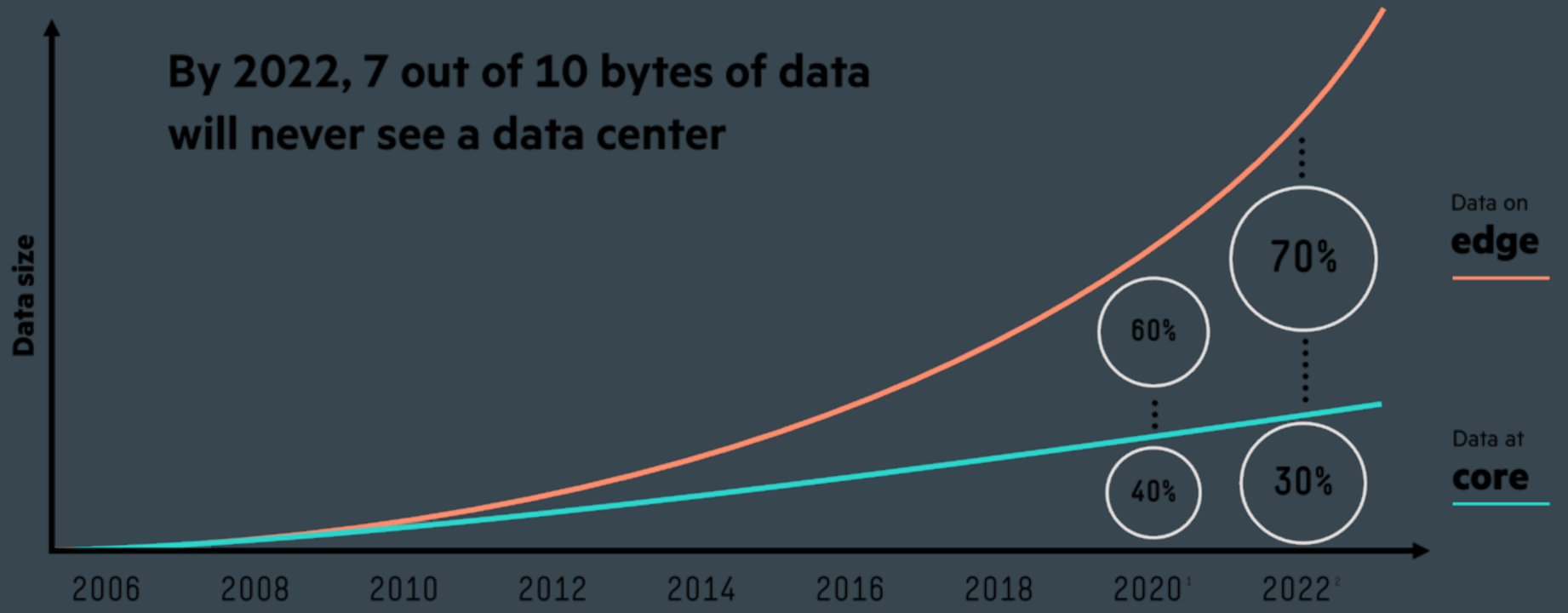
- tells how every generation of microelectronics can shrink in size
- there are physical limits
- since 2006 the increase in transistors per chip didn't increase the performance
- number of cores was increased then
- since 2016 thermal and electrical limits prevent performance growth



40 years of Microprocessor Trend Data Image: Karl Rupp

New paradigms

Shift of data from the center to the edge



1. International Data Corporation (IDC) https://www.idc.com/getfile.dyn?containerId=US41883016&attachmentId=47265871&id=null&bid=null&cid=null&patnerId=null
2. M2M Global Forecast & Analysis 2011-22

https://www.weforum.org/agenda/2018/09/end-of-an-era-what-computing-will-look-like-after-moores-law/

Where does the data come from?

- Machine or sensor data
- Social data
- Transactional data

Machine or sensor data

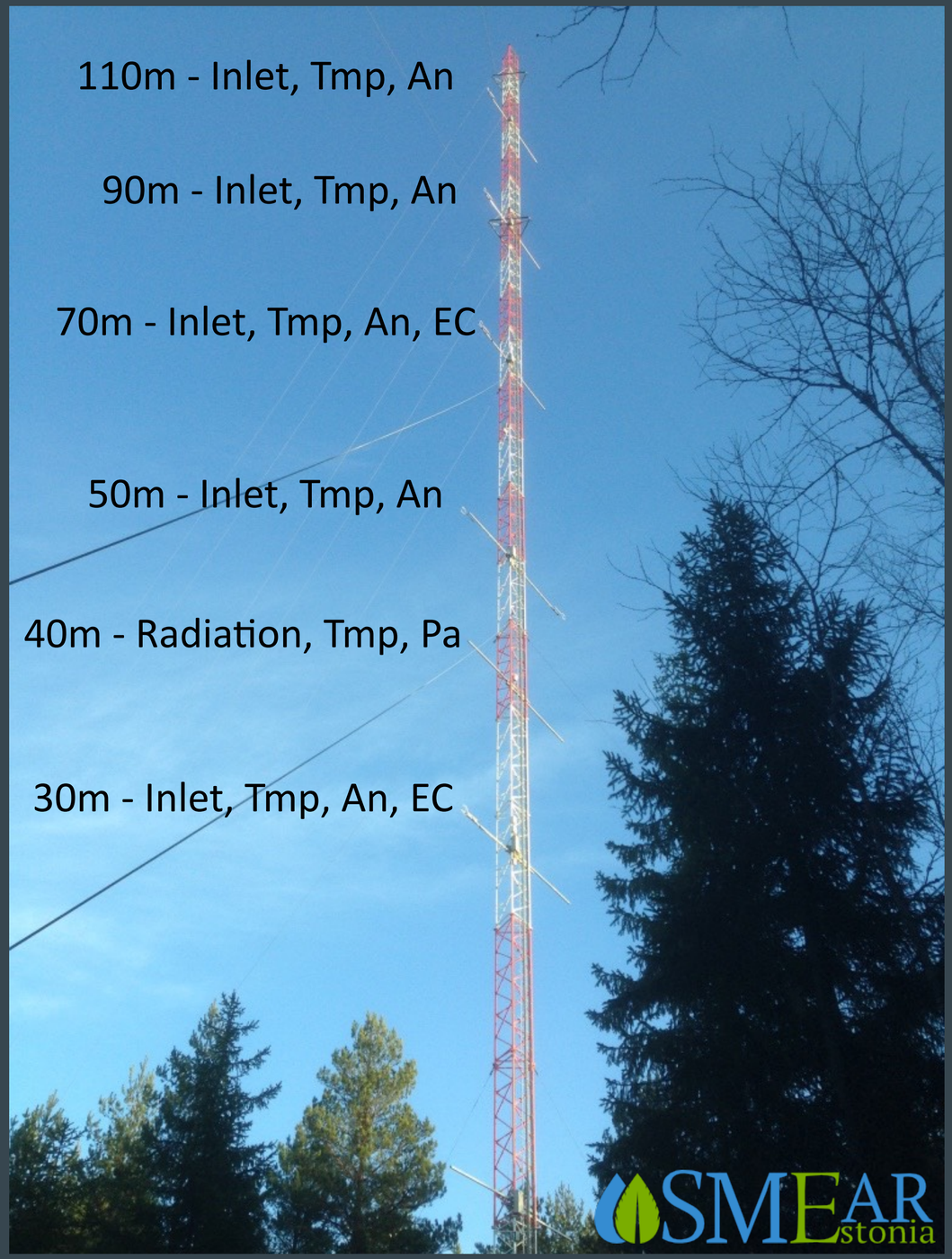
including “measurement data”

- These data appear in any part of our society
- Before the digital data processing became available, such data were “expensive” but very valuable. (Census, Inventories, Cadasters)
- Data are produced automatically in relation to an event or after a fixed time schedule
- This part acts like a feed-forward loop, e.g., we track the data storage and create new data from it.
- Large scale sources are Healthcare, Science, Economy, Industry, ...

Machine data example

SMEAR Estonia

- 130m high atmospheric tower
- Sensor network for:
  - CO2/CH4/H2O/CO
  - Temperature, 3D windspeed (u-v-z) and horizontal wind direction, atm. pressure
- 5 sensor heights
- data production speed 10 Hz
- daily data production 10 million rows of raw data



Social data

- Mostly referred to as data from social media platforms
- These are produced constantly when people interact
- Large scale sources are Twitter, Facebook, Instagram, Youtube, TikTok, and many others similar platforms
- These data contain information in the communication by text, pictures, and videos
- There is as well data in the connection, who is active and where?
- Telecommunication data also belongs to that segment

Social data

- Social data are of large value
- Most “platform businesses” use social data for their business model
- Companies
  - Alphabet (Google, GoogleCloud, Colab, Deep Mind, Calico, Waymo, Bard, PaLM2, ...)
  - Amazon (diverse online markets, AWS, Twitch, Ring, A9.com, ...)
  - Apple (Software, Hardware, iCloud, Drive.ai, Apple Studios, InVisage, ...)
  - ByteDance (TikTok)
  - Meta (Facebook, Instagram, Messenger, WhatsApp, Hardware, Llama, ...)
  - Microsoft (Software, Hardware, GitHub, Bing, LinkedIn, Skype, Nuance communications, Azure, Copilot, ...)
  - Tencent (WeChat, payment, BYD cars, Gaming platforms,...)
  - Tesla (cars, automotive, AI)
  - Twitter, X (messaging, news services, Vine, Periscope, xAI, ...)
  - OpenAI (ChatGPT differen flavors)
  - Anthropic (Claude differen flavors)
  - Perplexity (AI powered search engine)

- advertising
- online shopping
- streamlined services
- web search
- information retrieval
- Cloud computing services
- ML and AI development by utilizing the data streams their platforms generate
- Automation technology, self-driving systems
- Tooling (web based tools)

Transactional data

The technical side

- Linked with the “hardware” of the communication systems
- Internet backbones
- Telecommunication via cable (DSL, IP telephony)
- Mobile networks (GSM, G4/5/6)
- WLAN/LAN (Routers, Switches)
- Satellite communication
- Geopositioning (GPS)

Transactional data

The social-economic side

- Bank money transfers
- Stock exchange
- automated high speed stock trading
- Buy/sell platforms
- Distributed ledger systems (Blockchain)
- Electronic wallets
- Fungible tokens
- Telecom connections

- Health care (e-recipes, patient data)
- local governance (cadastre, ID card services)
- legal transactions (commodity and change thereof)
- Search engine data (incl. AI)
- AI chat data

Challenge

- Try to make a list what services you use that generate big data
- Please try also to assess the percentage of use between the services you name