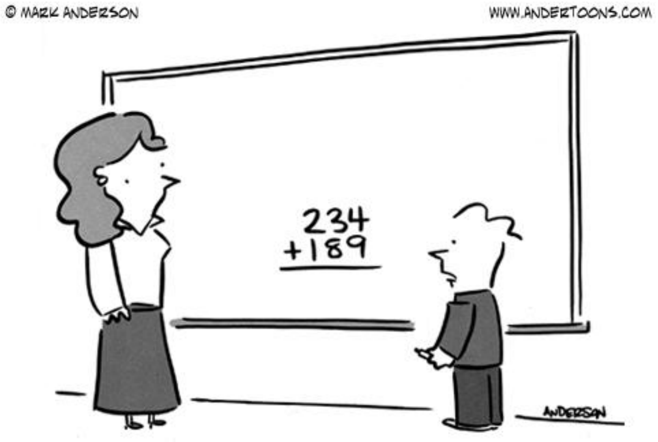
Multi-Core, Big Data processing, Heterogeneous Computing

Prof. S. Colafranceschi
(CERN, INFN, JMU)



"Does this count as big data?"

Intro/Disclaimer

- This presentation is crafted with cut and paste from my courses “Big Data, no problem” and ”Advanced Manufacturing”
  - Just intro, no technicalities, want to learn more: write: stefano.colafranceschi@cern.ch
  - browse: https://scolafranceschi.web.cern.ch/

Why do I teach tech related?

- A decade at CERN (world's largest and most prominent research organizations in the field of particle physics)
  - CERN's experiments generate data that is not only voluminous (>500Pb) but also incredibly complex
- Carried out Physics studies and simulations, took part in the Higgs Boson quest
- Wrote large firmware/software platforms for Supervisory Control and Data Acquisition systems (SCADA) for big data on HPC (clusters, decentralized, cloud)
- Founded earth observation high resolution startup

Industry 4.0 and Big Data

- Automation/Digital Manufacturing + Integration of various technologies: IoT, BigData, AI, CC
  - Rapid advances in digital sensors, networks, storage, and computation, along with their availability at low cost, are leading to the creation of huge collections of data
    - But how big is big? today's "big" may become tomorrow's "small."
    - 90% of the world's data generated has been generated in the last 2 years!

Big data objective

- To enables decision makers to make the right decisions based on predictions through the analysis of available data. Key aspects:
  - Data domain (searching for patterns)
  - Business intelligence domain (making predictions)
  - Statistical domain (making assumptions)

Big data analytics

- *Machine learning: the field of study that gives computers that ability to learn without being explicitly programmed* A. Samuel
  - Historically, many terms intended to describe the equivalent meaning of ML:
    - learning from data
    - pattern Recognition
    - data science
    - data mining, text mining
- Build systems that can perform/exceed human competence in handling many complex tasks

(My) 2 cent claim

The increasing amount of data generated is strongly correlated with our increasing quality of life and life expectation

- Long life to data!

Who started?

- The drive for generation and storage of data came from scientists
  - LHC generates 5 TB/s; “traditional data” stored in warehouses-like that rely in highly structured data

Who started?

Think about data, it's always been out there ("somewhere"):

- Maybe drawn on caverns
- Maybe just in our minds
- Maybe printed with old-style Kodak
- Maybe written by hand

- We did NOT invent data, we "implemented" a way to save some information into a digital (very handy) support.
  - It turns out that after 100k years (homo-sapiens) we have a way to keep collecting a lots of data → this huge basket poses a problem, nobody can look at the data, we design programs that do take a look to data, a program (supervised or not) would perform a data-analysis

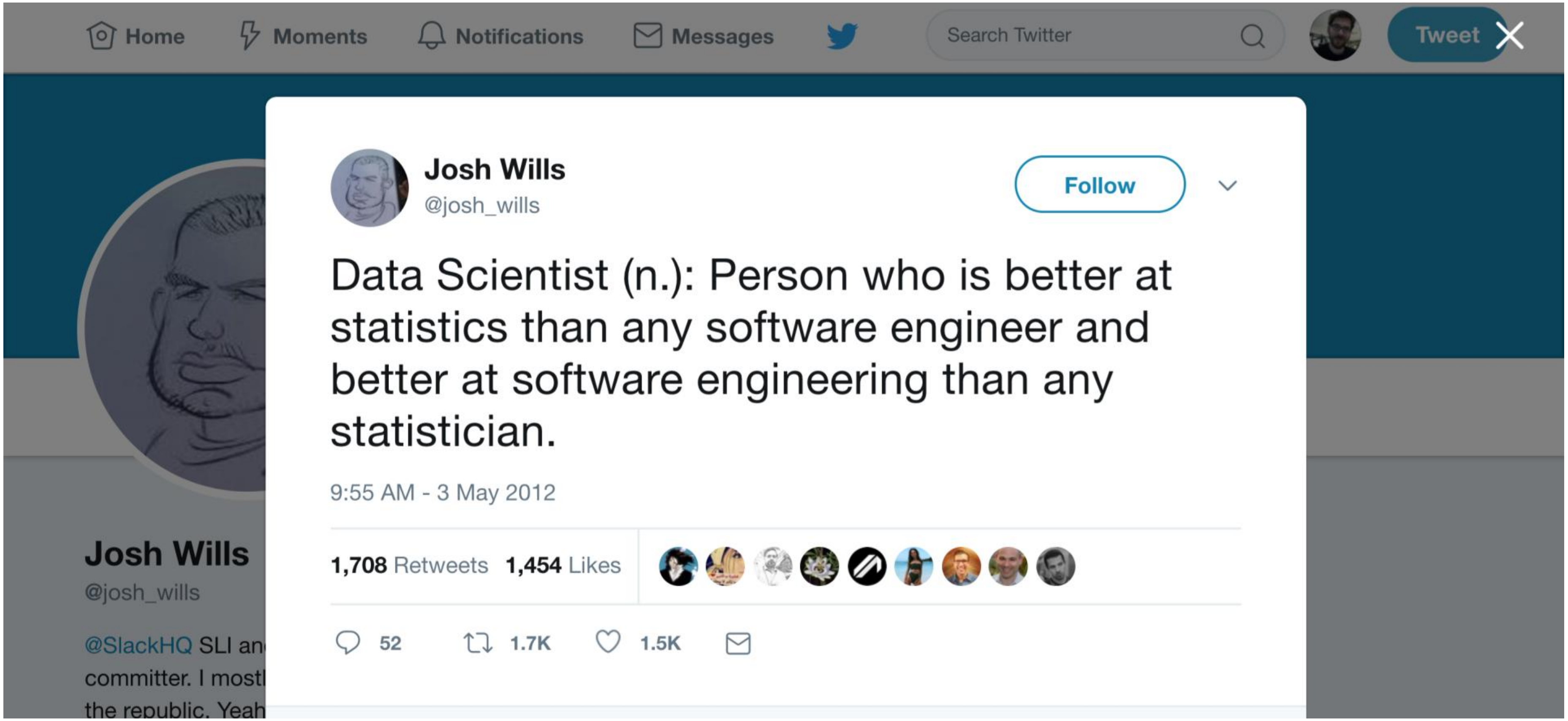
Where we are…

- SMAC (social, mobile, analytics, and cloud): a new paradigm for data to be generated, processed, and stored by enterprises
  - analyze any source of data, structured and stored in relational databases; semi-structured and emerging from sensors, machines, and applications; or unstructured obtained from social media and other human sources.
    - FB processes Petabytes every hour!

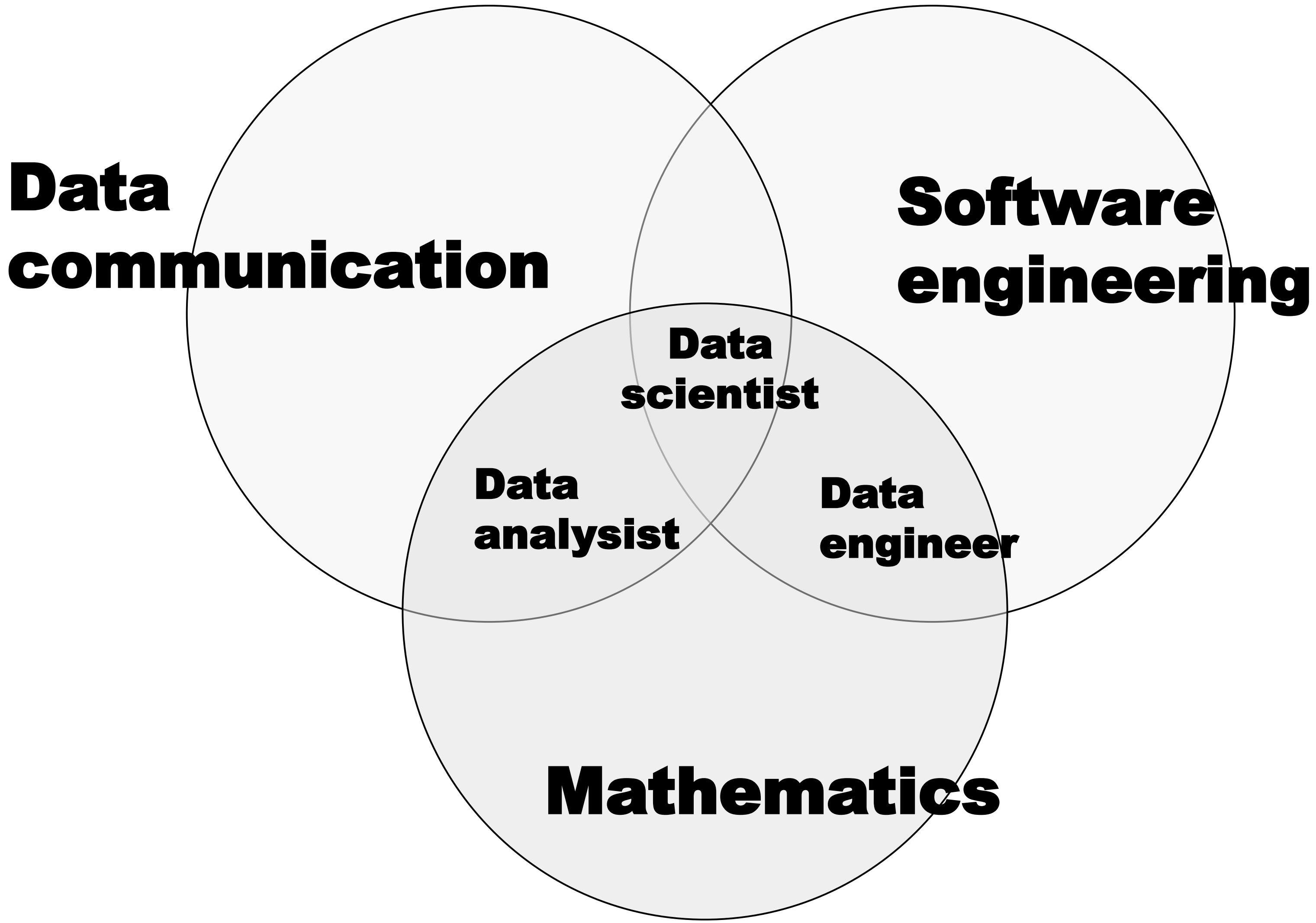
Where we are…

- By the way.. Google is a deliberate misspelling of the mathematical term googol: that’s 1 plus 100 zeros!

What Industry needs



What Industry needs



What Industry needs

- Data has the potential to enable new insights that can change the way business, science, and governments deliver services to their consumers and can impact society as a whole.
  - New algorithms, methods, infrastructures, and platforms are required
    - Big Data computing paradigm focuses on the sensing, collection, storage, management and analysis of data from variety of sources to enable new value and insights.

Challenges

- Life-cycle management of data
- Large-scale storage
- Flexible processing infrastructure
- Data modeling
- Scalable machine learning & data analysis algorithms
- Techniques for sampling and making trade-off between data processing time and accuracy
- Dealing with privacy & ethical issues involved in data sensing, storage, processing, and actions.

Big Data Science

- Very challenging when traditional methods lose their effectiveness due to **velocity**:
  - The velocity aspect of Big Data demands analytic algorithms that can operate data in motion,
    - algorithms that do not assume that all the data is available all the time for decision making, and decisions need to be made "on the go," (with summaries of past data?).
    - real-time processing systems (stream processing platforms)

Big Data Science

- Very challenging when traditional methods lose their effectiveness due to **large volume:**
  - The volume aspect of data demands that existing algorithms for different analytics data are adapted to take advantage of distributed systems where memory is NOT shared (different machines have only part of data to operate)
    - Deep learning NN layered with structures that hold different abstractions of the same data
      - natural language processing
      - text mining detection in the context of social media
    - Graph processing (powered by GPUs)

Big Data applications

- Capabilities offered by Big data produce "value"
  - Industries, Financial institutions, governments, entertainment, educational institutions, and researchers.. are applying Big Data analytics on a daily basis as part of their business as usual tasks.

Big Data and Industry

- Smart Manufacturing
  - Predictive Maintenance
    - By analyzing historical data and monitoring real-time data from machines, predictive maintenance models can be developed. These models can predict when equipment is likely to fail, allowing for proactive maintenance and reducing unplanned downtime.
  - Quality Control
    - Big data analytics can be used to monitor and control the quality of products throughout the manufacturing process. Any deviations from quality standards can be quickly detected and addressed, reducing waste and improving product consistency.

Big Data and Industry

- Supply Chain Optimization
  - Big data can also be used to optimize the entire supply chain, from raw material sourcing to distribution. This optimization can lead to cost savings, improved logistics, and better inventory management.
- Customization and Personalization
  - Industry 4.0 allows for greater customization and personalization of products. Big data helps in understanding customer preferences and tailoring products to meet individual needs.

Heterogeneous Computing

- Computing as a service/commodity
  - The ability of running on different hardware platform to exploit different technologies:
    - CPU, GPU, FPGA, custom hardware..

AI, computing and scale of economy

- But.. ideal experiment:
  - Technology creates a twin of you, that work for you (for nothing h24)
    - Employment will be higher, it’s just that it won’ be employment of humans
  - Automation can increase or decrease employment according to circumstances
    - Direct and indirect effect

House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Brush is 1micron wide
    - 10.000 persons/year to paint a house
      - no house painters are employed

House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Brush is 1micron wide
    - 10.000 persons/year to paint a house
      - no house painters are employed
  - Brush is 1mm wide
    - A few royal painter employed by the queen

House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Brush is 1micron wide
    - 10.000 persons/year to paint a house
      - no house painters are employed
  - Brush is 1mm wide
    - A few royal painter employed by the queen
  - Brush is 1cm
    - Hundreds of painters are employed

House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Brush is 1micron wide
    - 10.000 persons/year to paint a house
      - no house painters are employed
  - Brush is 1mm wide
    - A few royal painter employed by the queen
  - Brush is 1cm
    - Hundreds of painters are employed
  - Brush is 10cm
    - Realm of practicality, everyone has the home painted
    - Thousand of painters employed

House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Rollers and Spray Brush (~1m wide)
    - Demand of painters drop
      - The job is quick
      - No one needs the house repainted frequently

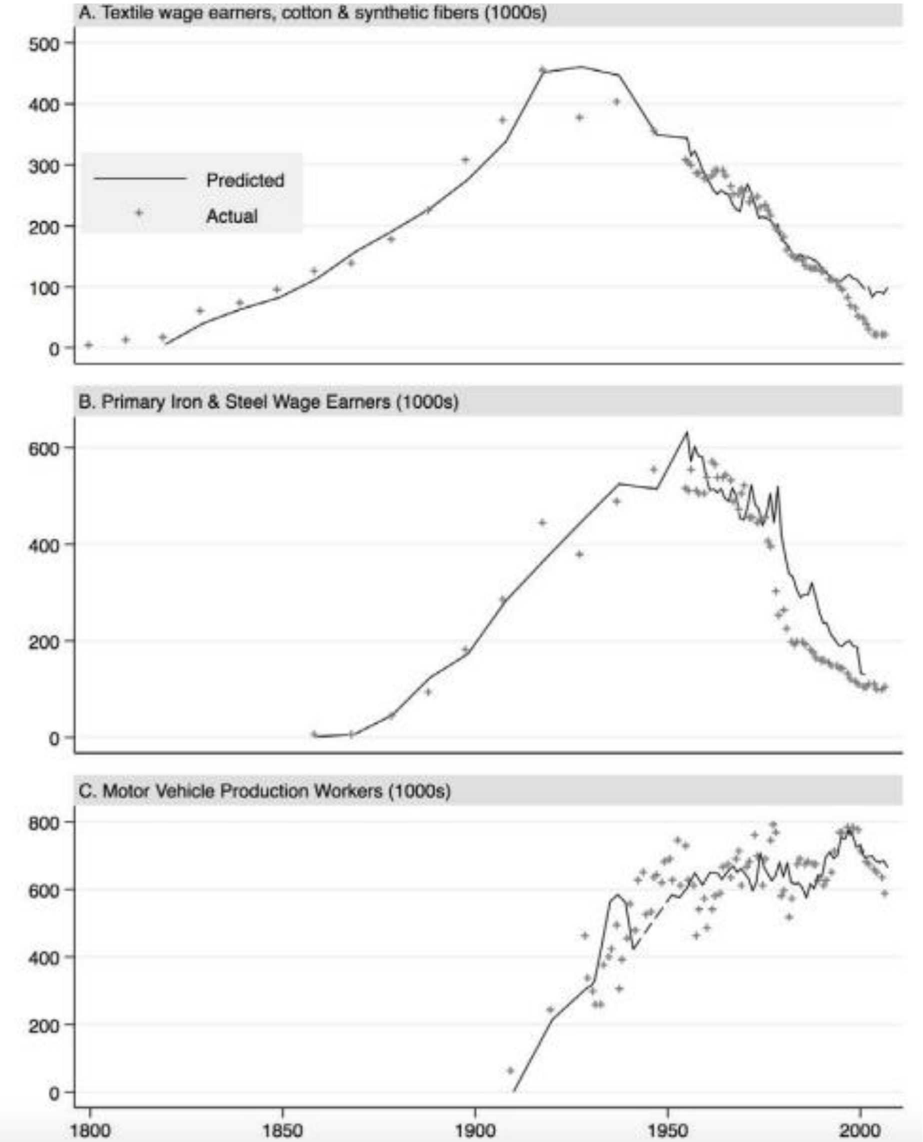
House painter example

- House painting as function of the technological level (width of the paintbrush)
  - Rollers and Spray Brush (~1m wide)
    - Demand of painters drop
      - The job is quick
      - No one needs the house repainted frequently
  - Self autonomous brush (~100m wide)
    - Paint the entire house in 30minutes
    - No more house painters existing

House painter example

- Technology can reduce cost and increase demand (“Inverted U curve”)
  - Making available new and more effective brushes, technology created a market
  - Further increase in technology lead to collapse the human (employment) in that market

James Bessen (National Bureau of Economic Research)



Inverted U curve

- The direct impact
  - technology depends on where you are on the curve
- The indirect impact
  - Some people manufacture brushes and robots
    - But we need much less people to make a robot if compared to the number of painters
      - Otherwise it will cost more to paint houses with robots and nobody will buy the robots
  - We spare money on painting with robot, we have more money for other things
    - This increase markets of other sectors

Summary

- Industry 4.0, powered by big data analytics, is transforming various industries by improving efficiency, reducing costs, enhancing product quality, and enabling innovative solutions.
  - The integration of data-driven technologies is a key driver of competitiveness and innovation in the modern business landscape.
- Industry 4.0 leverages big data as a fundamental component to enable the automation, optimization, and transformation of manufacturing and industrial processes.
- The integration of big data analytics into these processes is essential for achieving the goals of increased efficiency, flexibility, and competitiveness that Industry 4.0 aims to deliver.