Humankind 2.0: The Technologies of the Future
2. Big Data

Piero Scaruffi, 2016

See http://www.scaruffi.com/singular/human20.html for the full text of this discussion
Big Data

- Very soon Homo Sapiens will be producing more data every year than in the previous 200,000 years.
Big Data

• 2016:
  – The Internet has 32,585GB of Internet traffic every second
  – 2.4 million emails are sent every minute
  – Google processes 52,000 search queries per second
  – 5,000 tweets are posted on Twitter every second
  – YouTube delivers 115,846 videos per second
  – 50,000 Likes are generated daily on Facebook
  – 60,000 new photos are uploaded to Facebook every second

Sources: internetlivestats.com and onesecond.designly.com
From Silicon Valley to Data Valley

- More data will soon come from
  - Internet of Things
  - Wearables
  - Genetics
  - Nanorobots
  - Robots
data is the oil
of the 21st century

DATA IS THE NEW OIL

Once “refined” data yield...

Self-driving Car (and plane)
Drones
Ads on social media
Upselling
Wearables
Smart objects
Decision making
Big Data

- **Infrastructure**
  - Apache Spark: an engine for large-scale data processing (Matei Zaharia at UC Berkeley, 2009)
  - Nebula (NASA + Anso Labs + RackSpace) -> OpenStack (more than 500 companies in 2015)
Big Data

• Who has the biggest big data problem?
Big Data

- **Google and Facebook**
  - Google: Pregel (2010) -> Apache Giraph
Big Data

- Hadoop
Big Data

- Hive

Apache Hive + HBase Architecture
Big Data

- Big Query
Big Data

• Giraph
  – It’s all about graphs
  – Main success factor of Google's search engine: better ranking of search results
  – Google's ranking is based on PageRank, a graph algorithm
  – Facebook’s social graph has more than one billion users and more than 100 billion friendships
  – Twitter’s social graph has billion of follower relationships

\[ p_i = \sum_{j \in \{ (j,i) \}} \frac{p_j}{d_j} \]
Big Data

- Giraph

Facebook reveal trillion edge version of Apache Giraph in Graph Search
Big Data

- Big big-data users
Understanding Big Data

• The Data Loop
  – Most data are produced by machines
  – Increasingly data are read by machines
  – Most data are used to trigger events that produce more data
Big Data

• Visualizing big data
Hessian matrix from a quadratic programming problem
Linear programming problem
Computational fluid dynamics: shallow-water equations
Linear programming problem
Social network: people and the web pages they like
Big Data

• What those pictures are: solving a large system of linear equations with a large number (millions) of unknowns

Images by Margot Gerritsen (Stanford Univ), Tim Davis & Yifan Hu
http://www.cise.ufl.edu/research/sparse/matrices/
Big Data

• Beyond "Data Analytics", "Business Intelligence", etc
  – The platforms are available for free (open source)
  – Cloud storage is cheaper and cheaper
  – The math is widely available (eg, "Mining of Massive Datasets"): anybody can use those methods to analyze big data.
  – There is no "top secret" in Big Data
  – There is a huge number of big customers
Big Data

• Beyond "Data Analytics", "Business Intelligence", etc
  – But we still do old-fashioned "data analytics" (eg, Amazon's "recommendation engine")
  – Target figured out a teen girl was pregnant before her father did

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

FEB 16, 2012 @ 11:02 AM 2,927,596 VIEWS
Big Data

• Beyond "Data Analytics", "Business Intelligence", etc
  – There is still no killer application
  – Maybe that's why Google, Facebook, etc give their platforms for free to third-party developers
  – Understanding big data is one field that will require a shift from competition to cooperation.
Big Data

• Big data requires an interdisciplinary approach
  – Nuclear power, the Moon mission and the Internet are examples of big-data innovations driven by interdisciplinary teams
  – Solving problems in human society is not just a math problem
Big Data

• Big data requires an interdisciplinary approach
  – Harvard University's Institute for Quantitative Social Science
  – UC Berkeley's Institute for Data Science
  – US Government's "Big Data Research and Development Initiative"
Big Data’s Killer App

• Interpret data as people, not numbers
  – Capture all the data about my body, my routines, most recent medical knowledge, and conditions around me (epidemics, pollution, etc) and
    • 1. prevent health problems;
    • 2. alert me about health problems;
    • 3. suggest improvements to my lifestyle
  – Sloan Foundation’s Microbenet
Big Data’s Killer App

- Applications that guess the future (predictive applications)
  - Look for patterns and then build hypotheses
  - Stanford's "Big Data in Biomedicine" with the motto "Data science will shape human health for the 21st century"
Small Data

• Big data: a combination of structured and unstructured data that can reach exabytes
• Problem: in many cases big data is overkill
• Small data are around us and are most of the data that we need for our apps
• “Small data connects people with timely, meaningful insights, organized and packaged to be ... actionable for everyday tasks.” (Digital Clarity Group)
• Small data is the right data
• Big Data is good for centralized models (control)
• Small data is good for distributed models (crowd)
Quantified Self Movement

- Gary Wolf and Kevin Kelly (2007)
- You discover aspects of yourself that are obvious to all your friends but you never realized.
- Keeping a diary of your life, written by someone who follows you nonstop.
- The data will tell you who you really are.
- The data will help you improve yourself.
Big Data for Everybody

- Democratizing knowledge
  - 18th century:
    - The "Encyclopedie" to share all the world's knowledge with ordinary people
    - Prussia introduced compulsory primary education
  - 19th century:
    - Education not a privilege but a duty (mandatory for all children)
    - The PhD
Big Data for Everybody

• Democratizing knowledge
  – 20th century:
    • The World-wide Web and the smartphone ("prosthetic knowledge", Rich Oglesby)
  – 21st century:
    • Big data
Big Data

• Problems of democratizing knowledge
  – We don't even have access to the data that we generate
  – Ordinary people are the object, not the subject, of big data.
Bibliography
Contact

• www.scaruffi.com

See http://www.scaruffi.com/singular/human20.html for the full text of this discussion