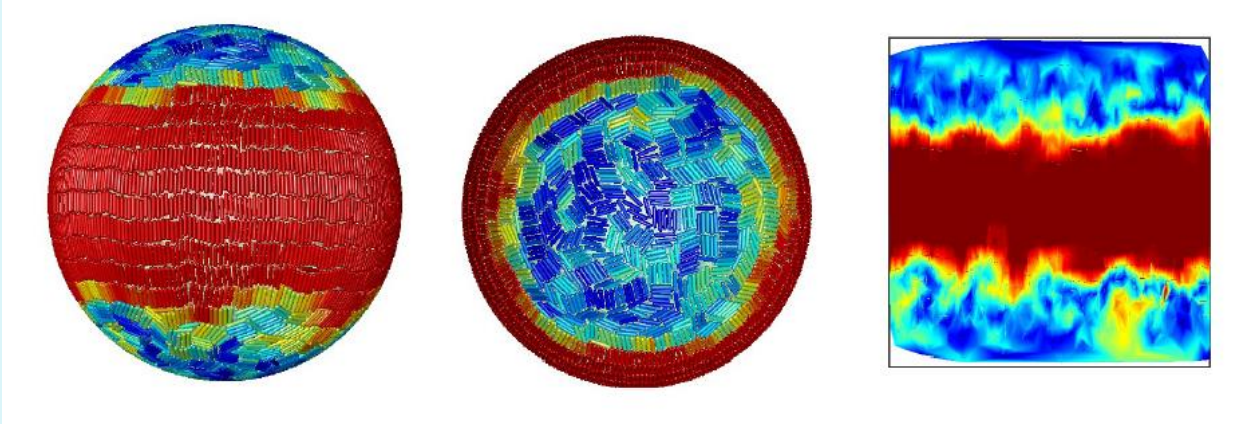
The Philosophy of Big Data: from hard sciences to soft sciences

Elshad Allahyarov

1) Physics Department, Case Western Reserve University, Cleveland OH, USA
2) Theoretical Department, OIVT Russian Academy of Sciences, Moscow, Russia
3) Institute for Theoretical Physics, HHU Düsseldorf / Theoretical Chemistry UDE Essen, Germany



Motivation

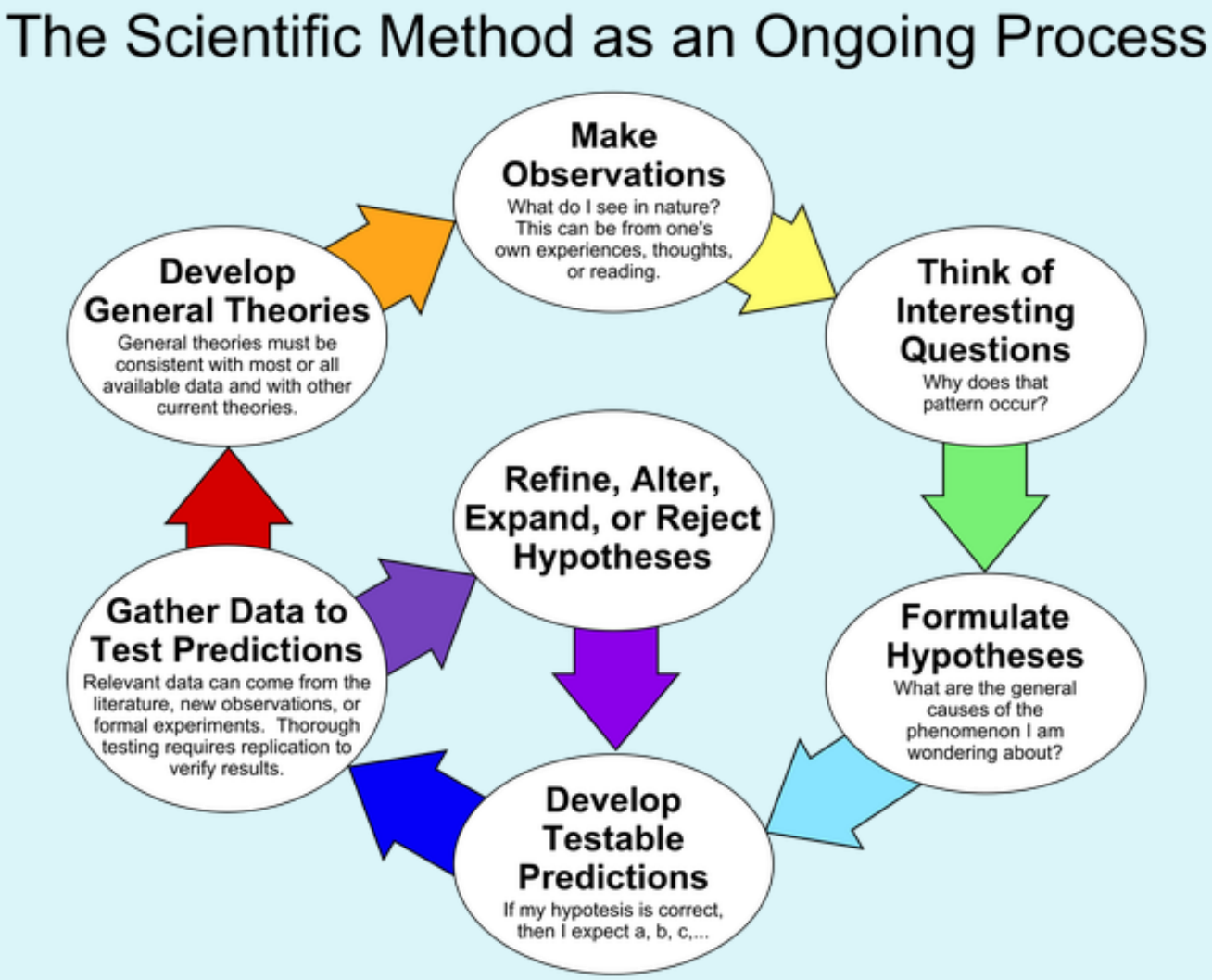
- **What is the difference between big data and hard science?**
- **Information sharing- how safe is it nowadays?**
- **Scientific databases build on scientific results.**
- **Scientific databases operated by Big Data methods.**
- **Big Data benefits from advances in Hard Sciences.**
- **A cooperation between Big Data and Hard Science is needed to get deeper insight into the dynamics and stability of the collected data.**

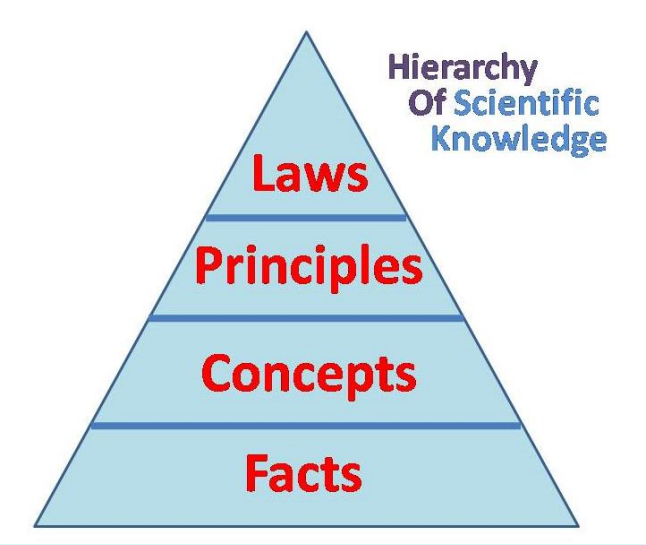
Main difference between hard and soft sciences

Hard science and soft science are colloquial terms used to compare scientific fields on the basis of perceived methodological rigor, exactitude, and objectivity.

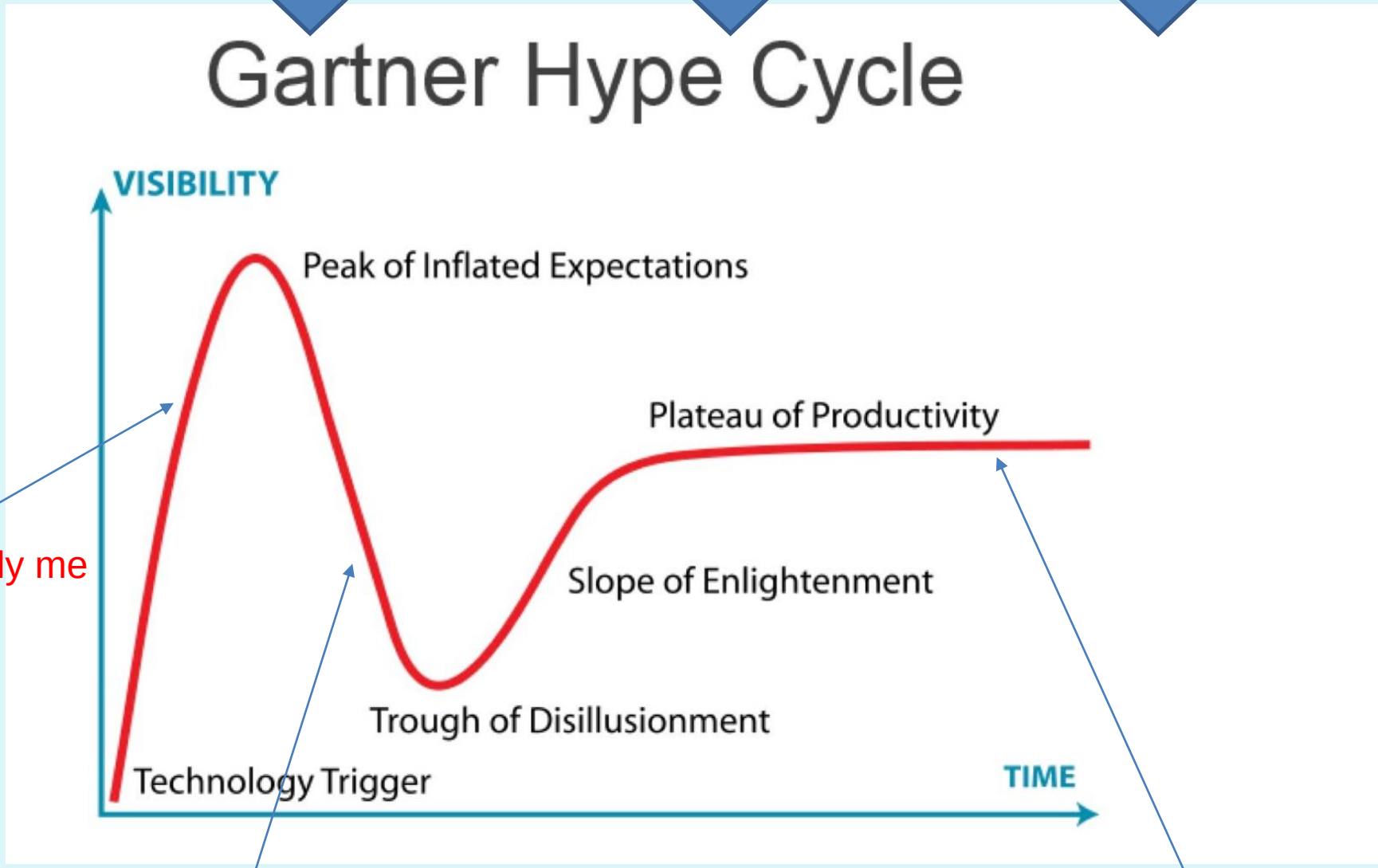
Soft sciences- physiology, history, sociology, weather monitoring, meteorology, psychology, linguistics, language studies, economy, financing, banking, media, communication, public health, health care, insurance.

Hard Sciences- physics, chemistry, biology.





Published every year for practically all emerging technologies

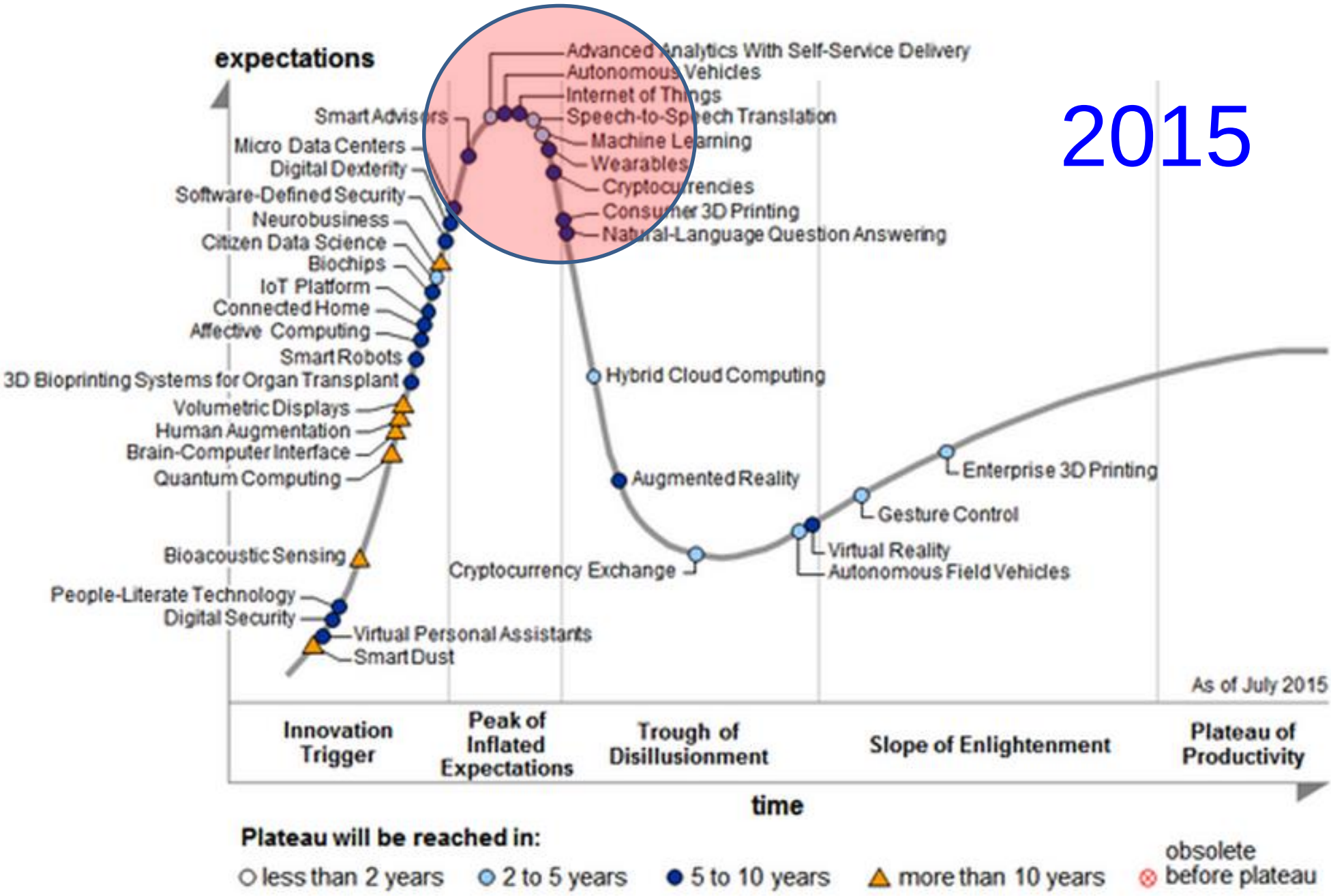


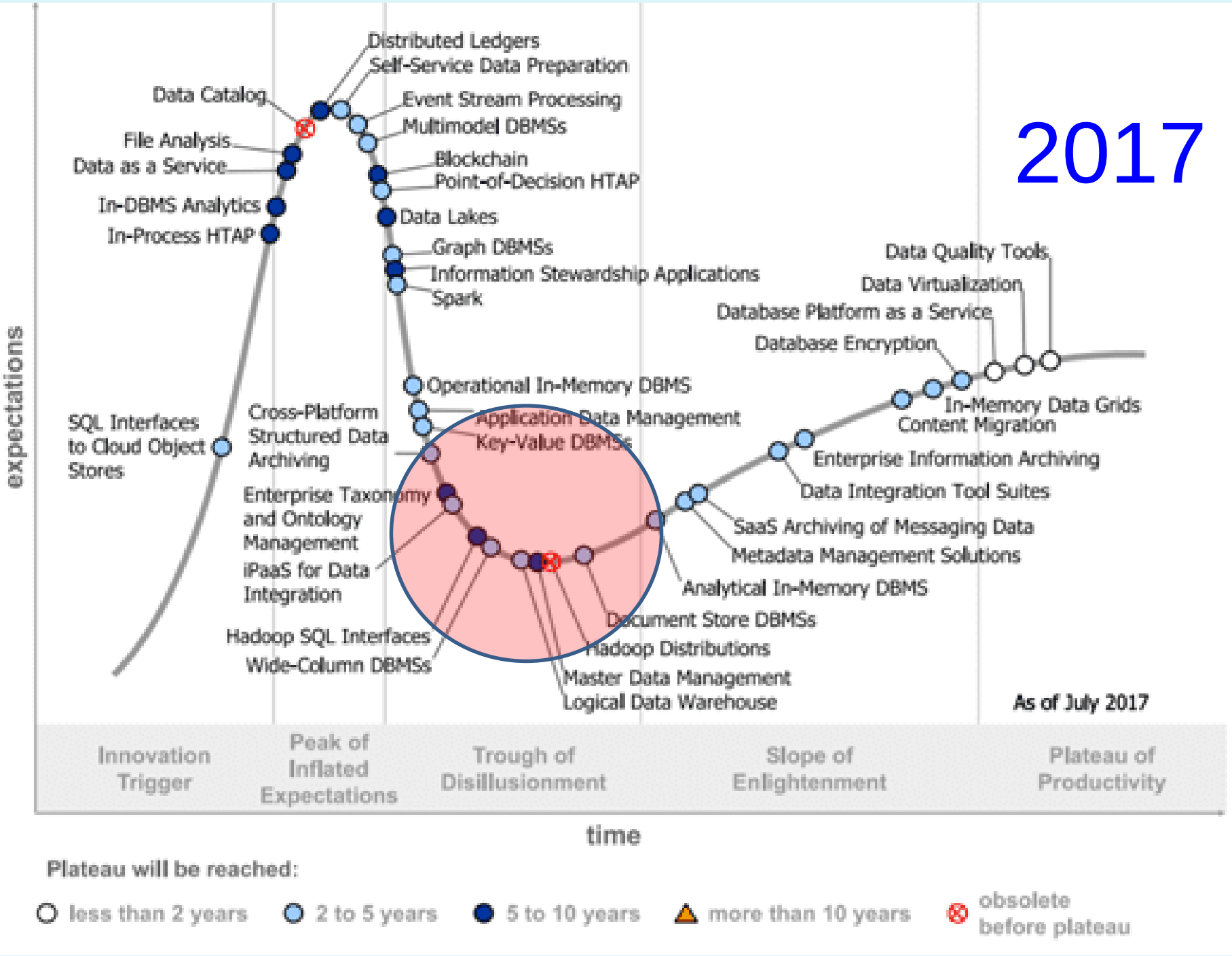
Me and only me

Me and Mozart

Only Mozart

Figure 1. Hype Cycle for Emerging Technologies, 2015





Big Data is Falling into the Trough of Disillusionment

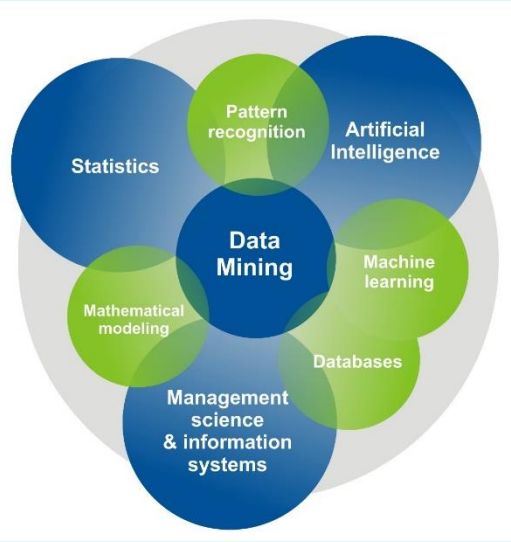
The purpose of big data (sources: IoT, b-b, b-g, b-c, social networks,…)

- to store,
- to manage,
- **to utilize data**
- **to cipher available information**

} technical stuff

- to serve some specific purposes

} potential scientific applications

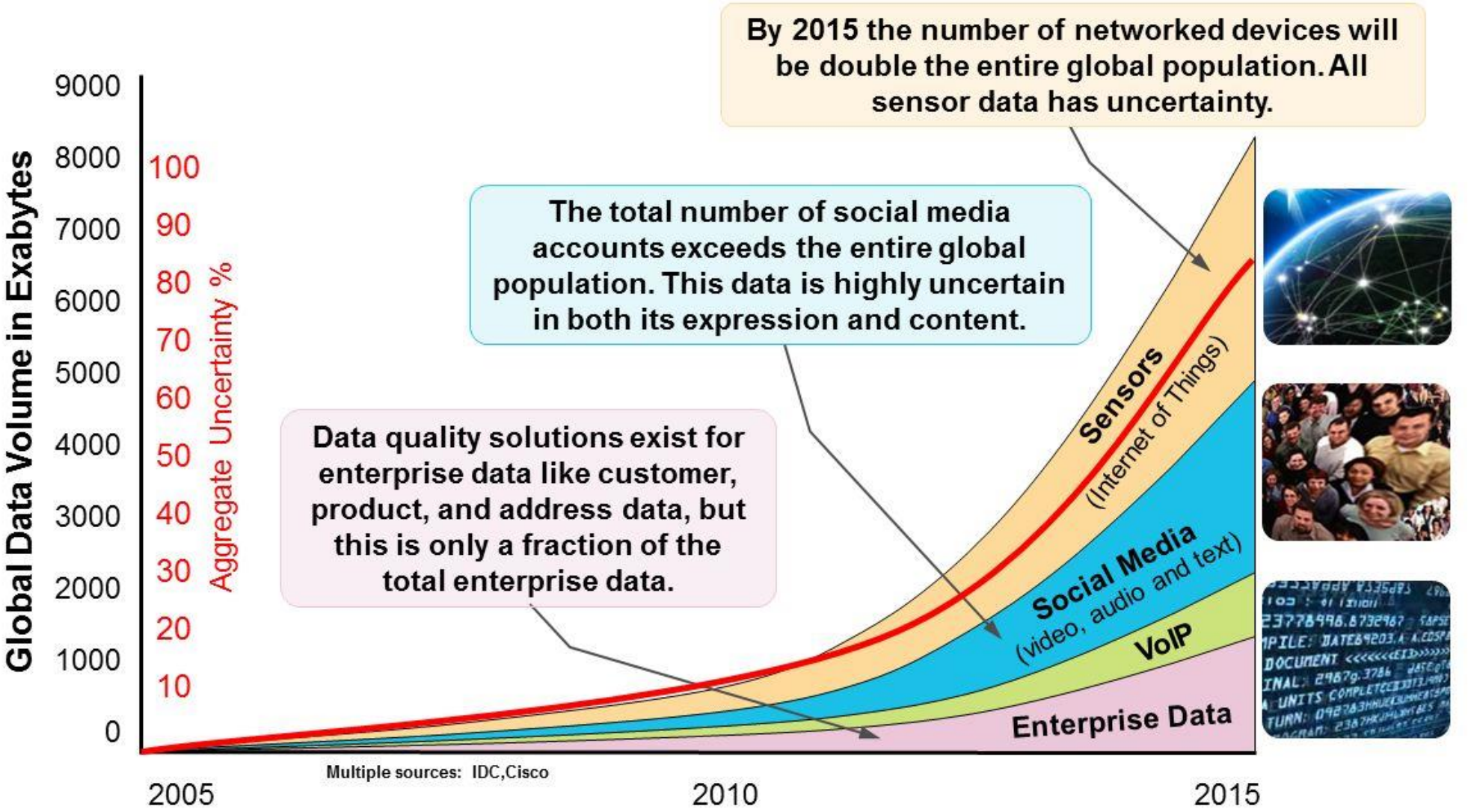


basics of Big Data courses taught to students

For the utilization you need additional things:

- data managment platforms (oracle, ibm),
- architectures,
- analytical methods,
- software tools, Hadoop, Pig, Phythoon, R,
- parallel computing, cloud computing,
- statistical approaches,
- vizualization techniques.
- Frameworks: cloud, hadoop, spark, mapreduce,
- Microsoft's Big Data Service, HDInsight Cluster
- Azure Storage Account,
- deep dive, cognitivity analyses,
- to get new relationships among the data,

Today, exploratory (thus horizontal, 2D) analysis of Big Data is fast, large-scale, data-driven and involves extensive use of advanced statistical methods and visualization techniques.



Whereas more data will be generated each year, its availability and sharing will face growing limitations because of emerging scandals which generate stronger government regulations.

Recent Scandals that mark a shakeup of the Big Data Landscape.

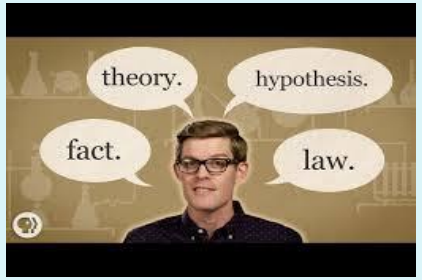
1. Uber:
It holds personal information containing your addresses, credit cards, driving license numbers, email addresses, phone numbers, detailed data on your movements and travel history. In 2016 hackers stole personal information from 57 million Uber users around the world, including names and driver's license numbers of around 600,000 drivers in the U.S

2. Anthem: Health insurer in the US. In 2015 there was a breach affecting medical records and personal information of 79 million people. The company settled litigation for a record $115m.

3. The case of hiQ Labs, Inc. vs LinkedIn Corp. In 2017 the U.S. District Court in California granted hiQ Labs, Inc. a preliminary injunction against LinkedIn Corp which prohibits LinkedIn from preventing hiQ's access to LinkedIn users' public profile data. So, even without your consent, your data can be used by third companies.

4. **Cambridge Analytica**. In 2016 this data analytics firm illicitly procured the data of 75 million Facebook users — without their knowledge or consent — and then enlisted that to inform voter-targeting strategies for Donald Trump's presidential campaign.

The main difference between
Big Data Technology and Hard Science
is similar to the difference between
Exploration vs. Discovery.



An Old World Example

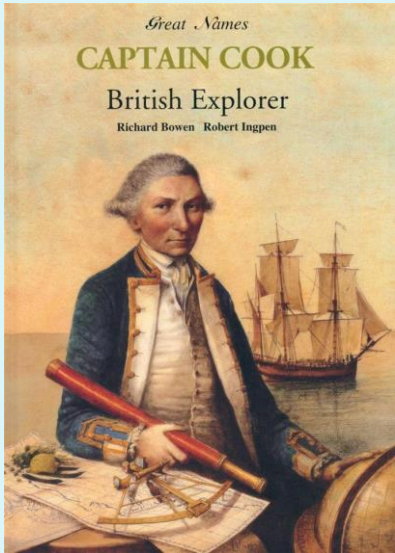
In the 15th through 18th centuries, there were many voyages of exploration and discovery. Some you could characterize as exploration and some as discovery. Let's look at two examples that show the difference.

Christopher Columbus was on a voyage of discovery.

He knew exactly what question he wanted to answer – I want to get to the East Indies – and knew what direction or area to look – sailing directly west. Now, he did find a different answer, discovering the Americas, but his mission was one of discovery.

first voyage in 1492

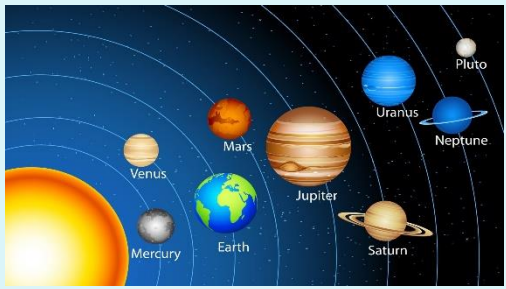
Captain James Cook set out on a different mission – to explore the Pacific. He was trying to explore new areas to find answers to a broad suite of questions. As he explored, he would identify specific areas that showed promise. Then, he would transition into discovery mode to answer specific questions relevant to that area.



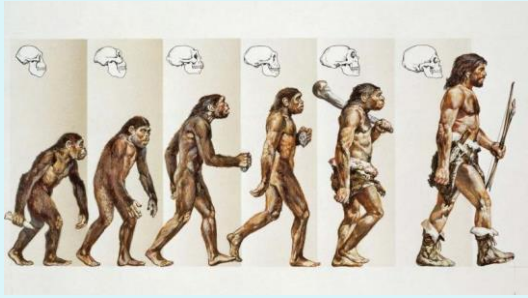
first voyage in 1768

Examples of scientific discoveries utilizing big data collected by scientific methods

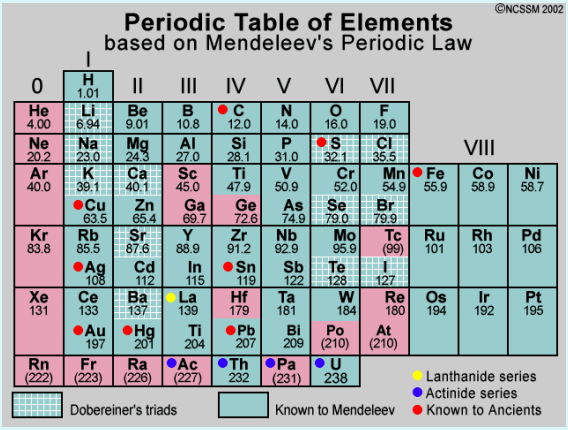
First –the question “why?”, then gathering facts, then analytics and discovery



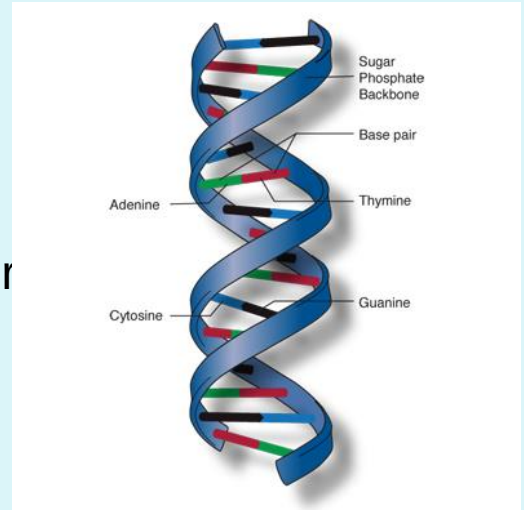
Nicolaus Copernicus, 1543, the Sun is the center of the Universe and it made the planets move around it in perfect circles



Charles Darwin, 1858, On the Origin of Species by Natural Selection (before him Thomas Malthus, in cooperation with Alfred R. Wallace)



Dmitri Mendeleev, 1869, periodic table of elements, developed mainly to illustrate periodic trends of the then-known elements



James Watson and Francis Crick, 1953, the discovery of the double helix, the twisted-ladder structure of deoxyribonucleic acid (DNA).

Scientific ‘Big Data’ databases

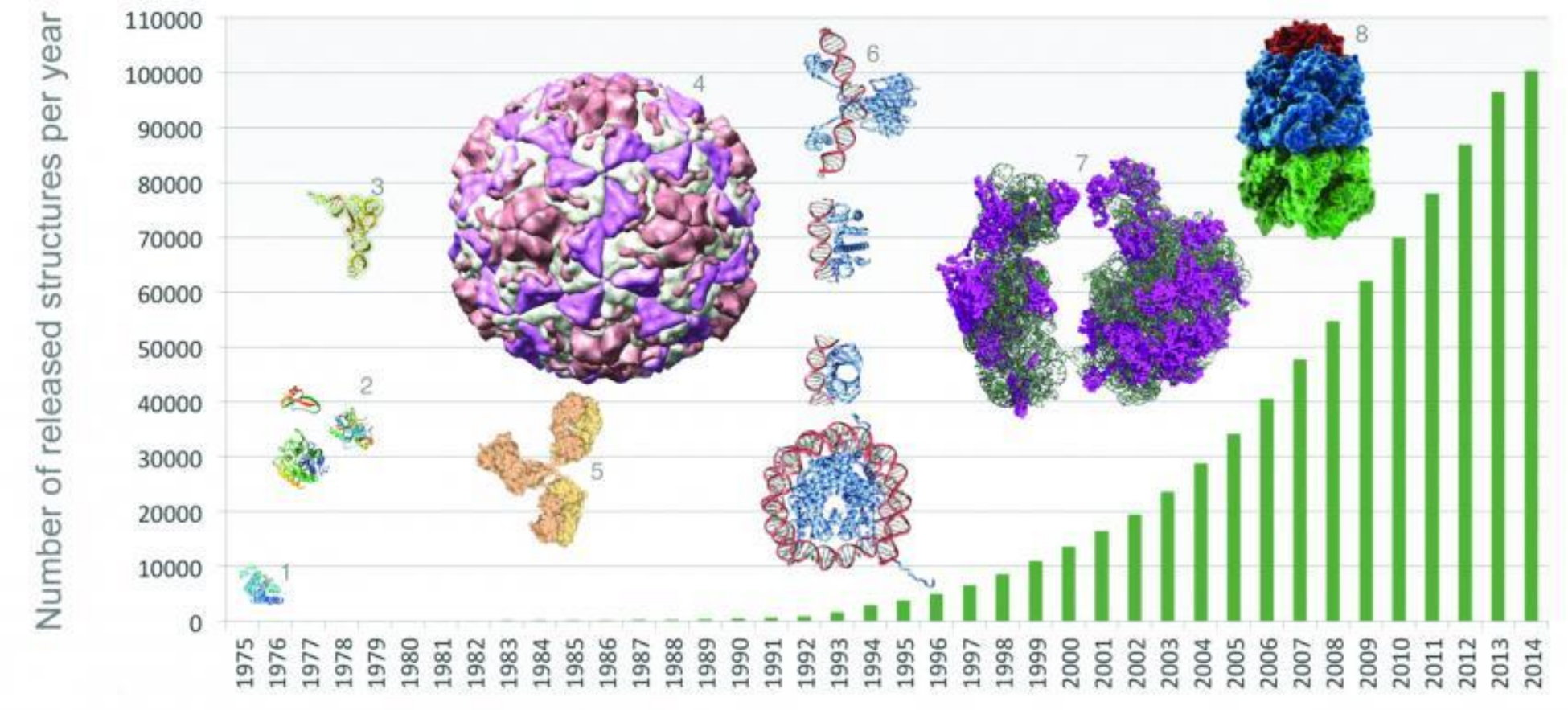
During the progress of Science, huge data was collected and verified which led to the formation of different databases: the first Big Data collections. Scientists now have access to numerous large data sets of relevance to multiple scientific domains.

1. Protein Data Bank (started in 1971).

WORLDWIDE PDB
PROTEIN DATA BANK

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

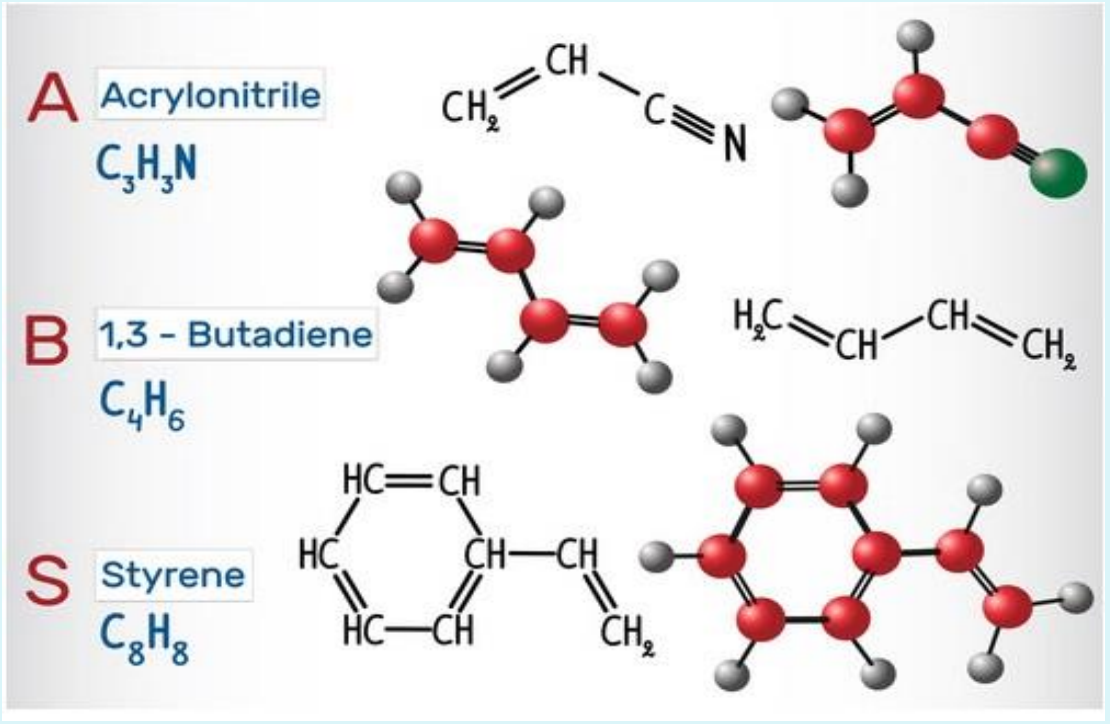
The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.



Scientific ‘Big Data’ databases

2. Polymer Database (started in 2003).

Polymer Database "PoLyInfo" systematically provides various data required for polymeric material design. The main data source is academic literature on polymers. Information on polymers including properties, chemical structures, IUPAC names, processing methods of measured samples, measurement conditions, used monomers and polymerization methods are stored in a object database



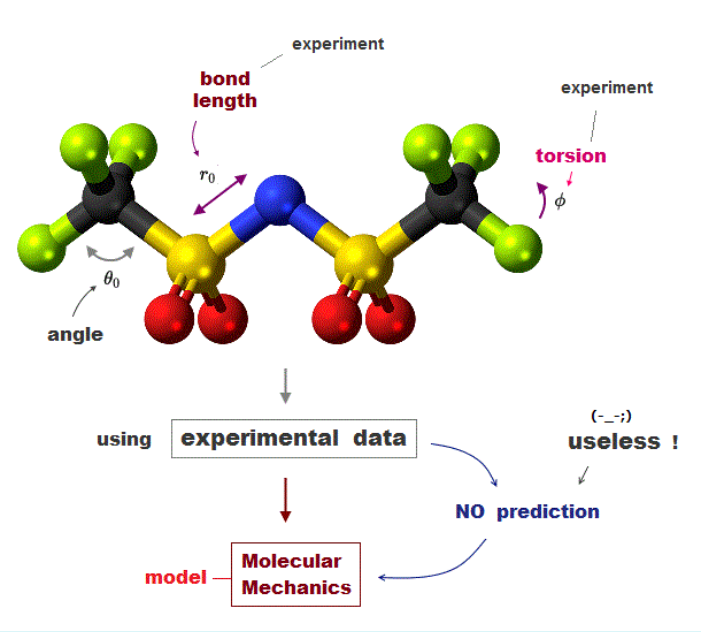
Number of open data (Sep. 27, 2017)	
Homopolymers	14,798
Copolymers	5,068
Polymer Blends	1,795
Composites	2,043
Monomers	17,104
Property points	294,856
Literature data	15,647

Scientific ‘Big Data’ databases

3. Force-Field Database (started in 1990).

In the context of molecular modeling, a force field is developed to fit energy functions or interatomic potentials. Different force fields are designed for different purposes. All are implemented in various computer software.

Popular Force-Fields developed for molecular dynamics of macromolecules are:
-- AMBER (Assisted Model Building and Energy Refinement),
-- CHARMM (Chemistry at HARvard Molecular Mechanics),
-- GROMOS (GROningen MOlecular Simulation),
-- MMFF (Merck Molecular Force Field)
--…… (50-60 more)



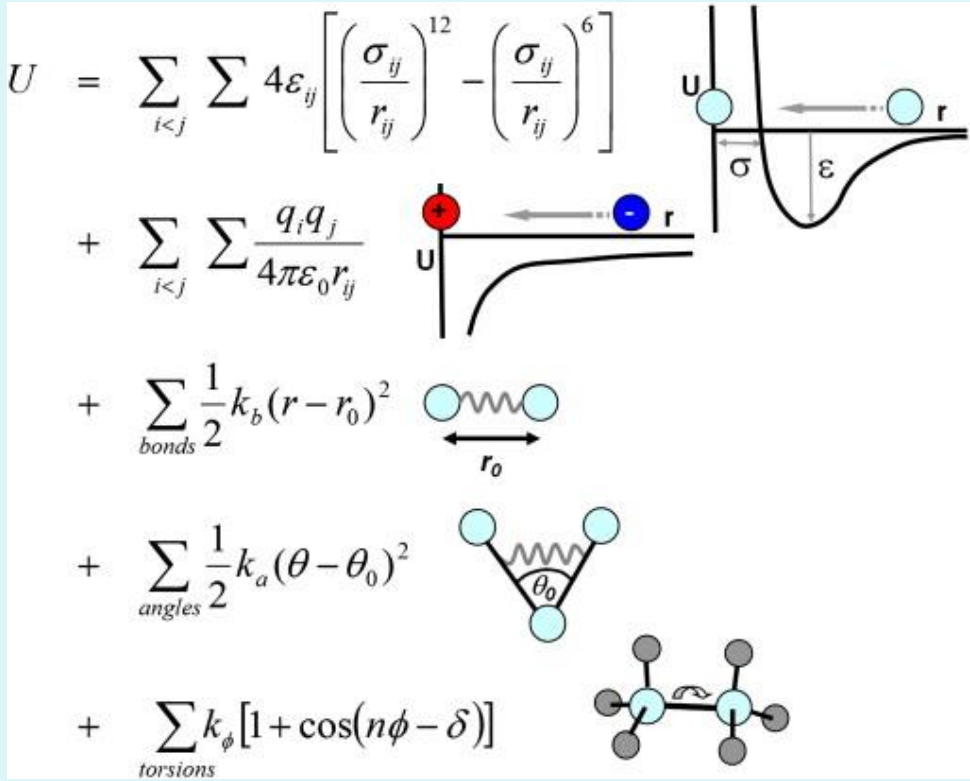
$$U = \sum_{i<j}\sum 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

$$+ \sum_{i<j}\sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

$$+ \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2$$

$$+ \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$

$$+ \sum_{torsions} k_\phi [1 + \cos(n\phi - \delta)]$$

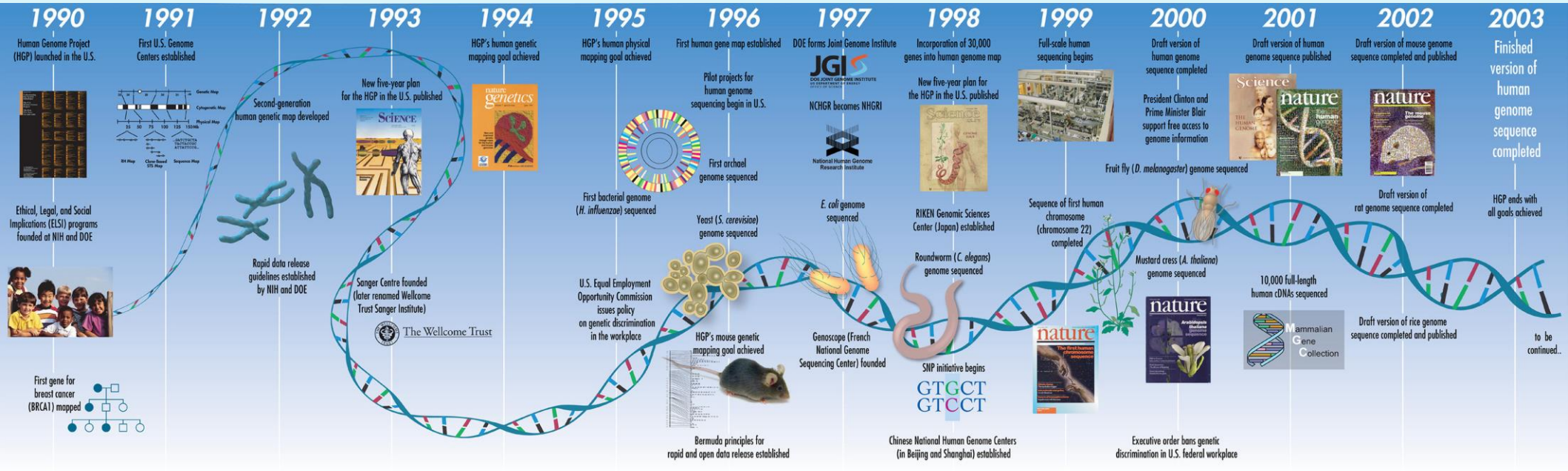


Scientific ‘Big Data’ databases

4. The Human Genome Project (HGP) (started in 1990).

The Human Genome was an international scientific research project with the goal of determining the sequence of nucleotide base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint

DNA Sequencing Technologies were Key to the Human Genome Project



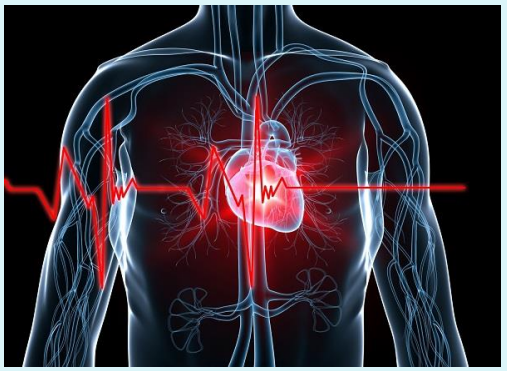
Scientific ‘Big Data’ databases

5. National Oceanic and Atmospheric Administration (NOAA) database (started in 1970).

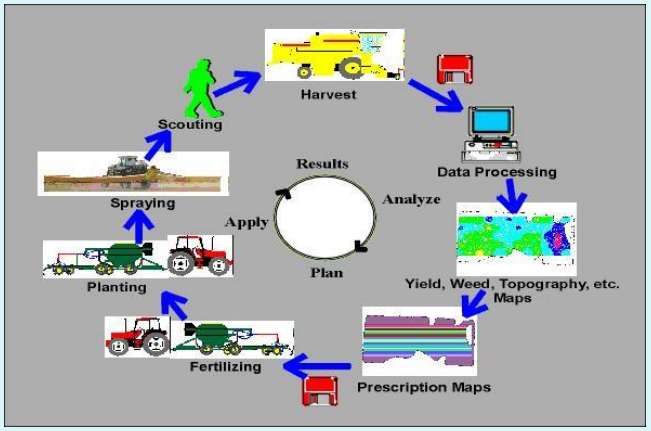
NOAA maintains several Databases containing data on climate patterns, earthquakes, ozone levels, and ocean temperatures. These data are useful to scientists in many fields, including environmental science, energy, public health, and medicine.

Examples how Science has benefited from Big Data developed approaches.

There is growing tendency to use Big Data methods in contemporary Sceince

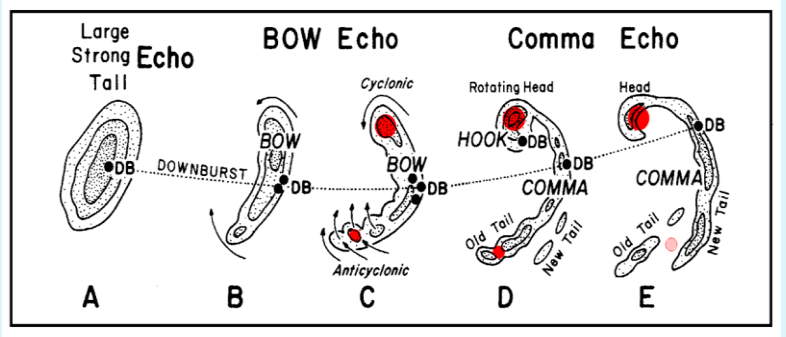


1. Cardiologists and data specialists at Stanford University, and University of California San Francisco have developed a data science algorithm that uses patient electronic health records, especially heart beat records and evaluates other risk factor records, to predict second heart attack for a patient. This is advantageous as the doctors don't need to physically analyze and evaluate the patient hands on for identifying elevated risk of heart attack in patients, allowing hospitals to save valuable time and resources for other aspects of patient recovery.



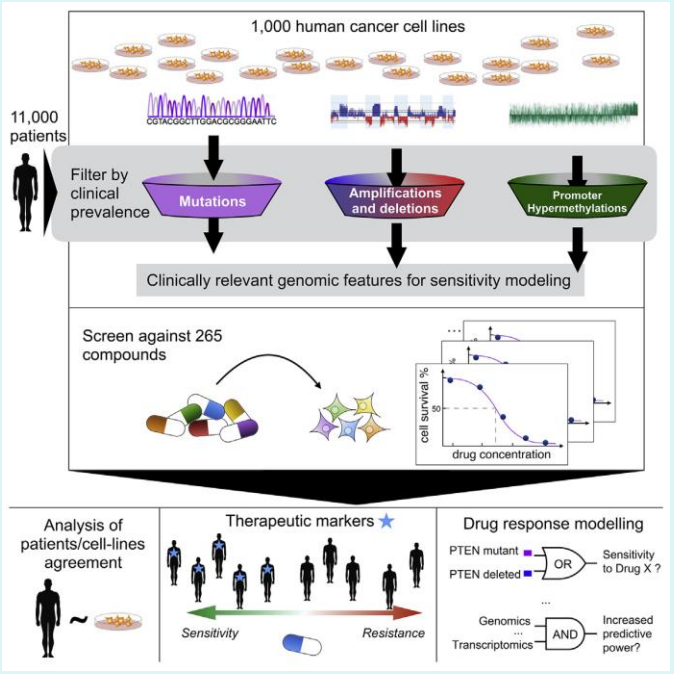
2. Precision Farming: This new farming approach is based on collecting farming information on the plant seeds over many farms. Then, instead of using biological research in greenhouses and fields or months and years, now the research starts at the computational level (in-silico) where data can be analyzed, experiments planned, and hypotheses developed. From here, a much smaller number of plants needs to be validated in the field for performance across a wide range of environments, when a breeder can then determine which exact hybrid is best for a particular area.

3. Prediction of severe weather in the global climate system. Penn State's IT College, and Accuweather Inc published in 2017 a pioneering work that utilized the power of big data and data science. The researchers utilized a 'bow echo' signature signal, which is caught in the radar before a severe thunderstorm, hurricane or tornado develops. Though the bow echo signal is easily missed by human eyes, catching it early can help predict severe weather. By harnessing the vast data collected by the National Oceanic and Atmosphere Administration (NOAA), the researchers used machine learning to accurately and efficiently detect bow echoes and automatically predict severe thunderstorms, tornadoes and hurricanes.



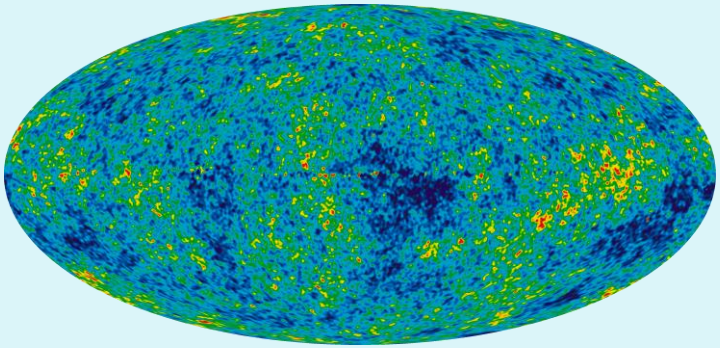
Examples how Science has benefited from Big Data developed approaches.

4. Drug Discovery (DD) is now extensively based on Big Data and Machine Learning for “right” drug molecules. Computer-aided drug discovery (CADD) approaches using pharmacophores and molecular modeling to conduct so-called “virtual” screens of compound libraries. Precision medicine is a modern approach to treatment, where doctors select the best course of treatment for the patient based on the patients personalized genetic information.

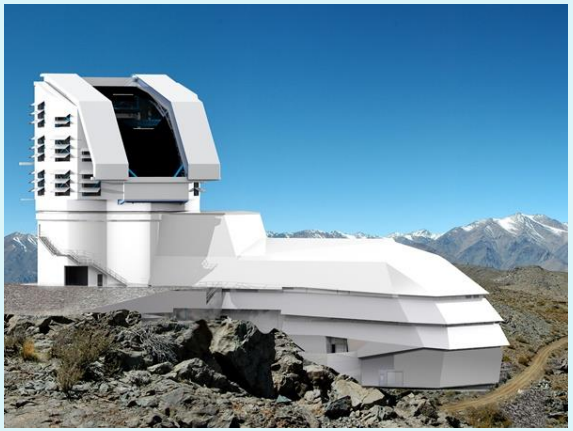


5. Large Hadron Collider (LHC). The collider experiments in high-energy physics. The LHC generates up to 600 million collisions per second and produces 15 petabytes (15 million gigabytes) of data per year. Big Data, distributed computing and sophisticated data analysis all played a crucial role in the discovery of the Higgs boson. But the discovery of the Higgs boson was not data-driven.

6. NASA's Kepler telescope (NKT). In 2017 Google developed and applied data science algorithms on data or signals collected by NKT to identify a Solar System like our own called Kepler-90 star system elsewhere in the universe.



Examples how Science has benefited from Big Data developed approaches.



7. Large Synoptic Survey Telescope (LSST). It is being built in Chili. Starting in 2022, the LSST will capture images of the entire night sky every three days over a 10-year period,

8. The sanctification of Big Data by Science:

In the US the science already started to look for a joining points with the Big Data. In 2018 National Science Foundation (NSF) and National Institutes of Health (NIH) joined forces "to develop new methods to derive knowledge from data; construct new infrastructure to manage, curate and serve data to communities; and forge new approaches for associated education and training,"

The "program aims to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting information from large, diverse, distributed, and heterogeneous data sets in order to accelerate progress in science and engineering research."

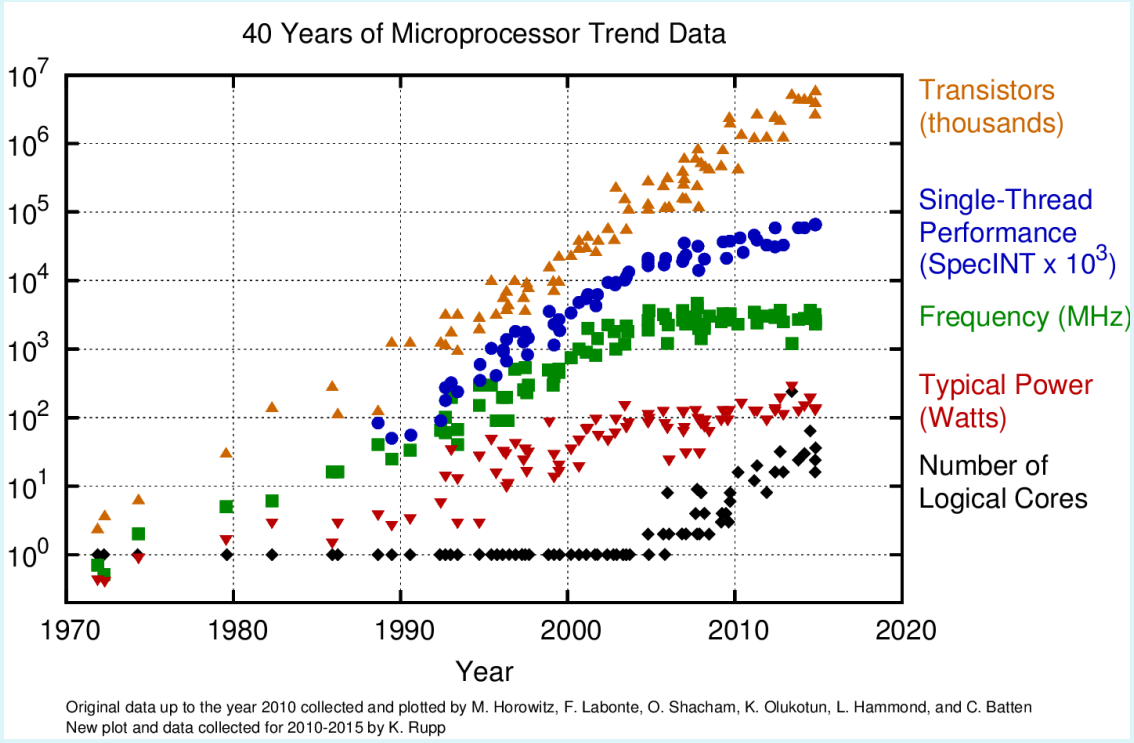
Aim: a progress in the use of Big Data to improve our understanding of ourselves and the world

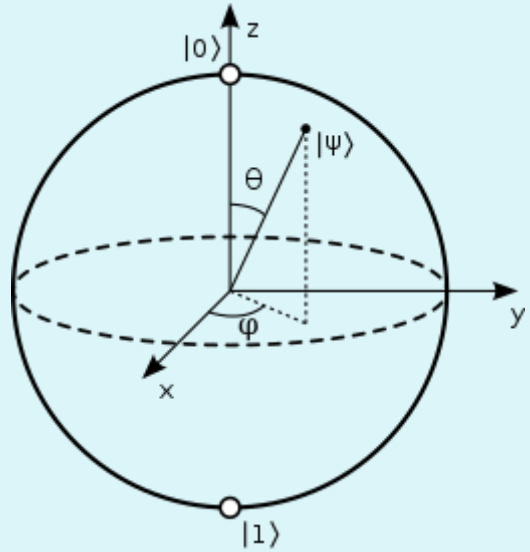
Examples how BIG DATA applications will benefit from the advances in HARD SCIENCE.

Big Data starts to play role in the progress of fundamental SCIENCE. However, the science, especially physics, material science and chemistry never stopped and still are continuously and successfully solving many of Big Data problems.

New developments in quantum computing will tremendously increase the speed of the processors and the efficiency of parallel computation.
Today, Google has a quantum 72-qubit computer they claim is 100 million times faster ***(a factor 10^9)*** than any of today's systems.

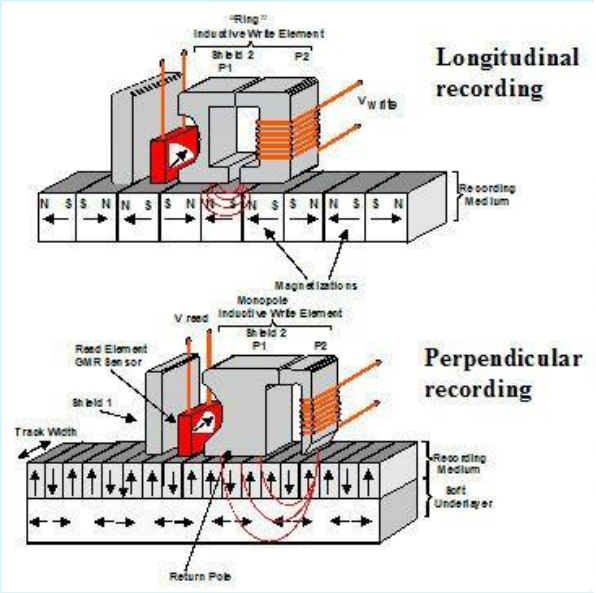
Processor frequency is stalled because of Heating problems, field-generated noise.

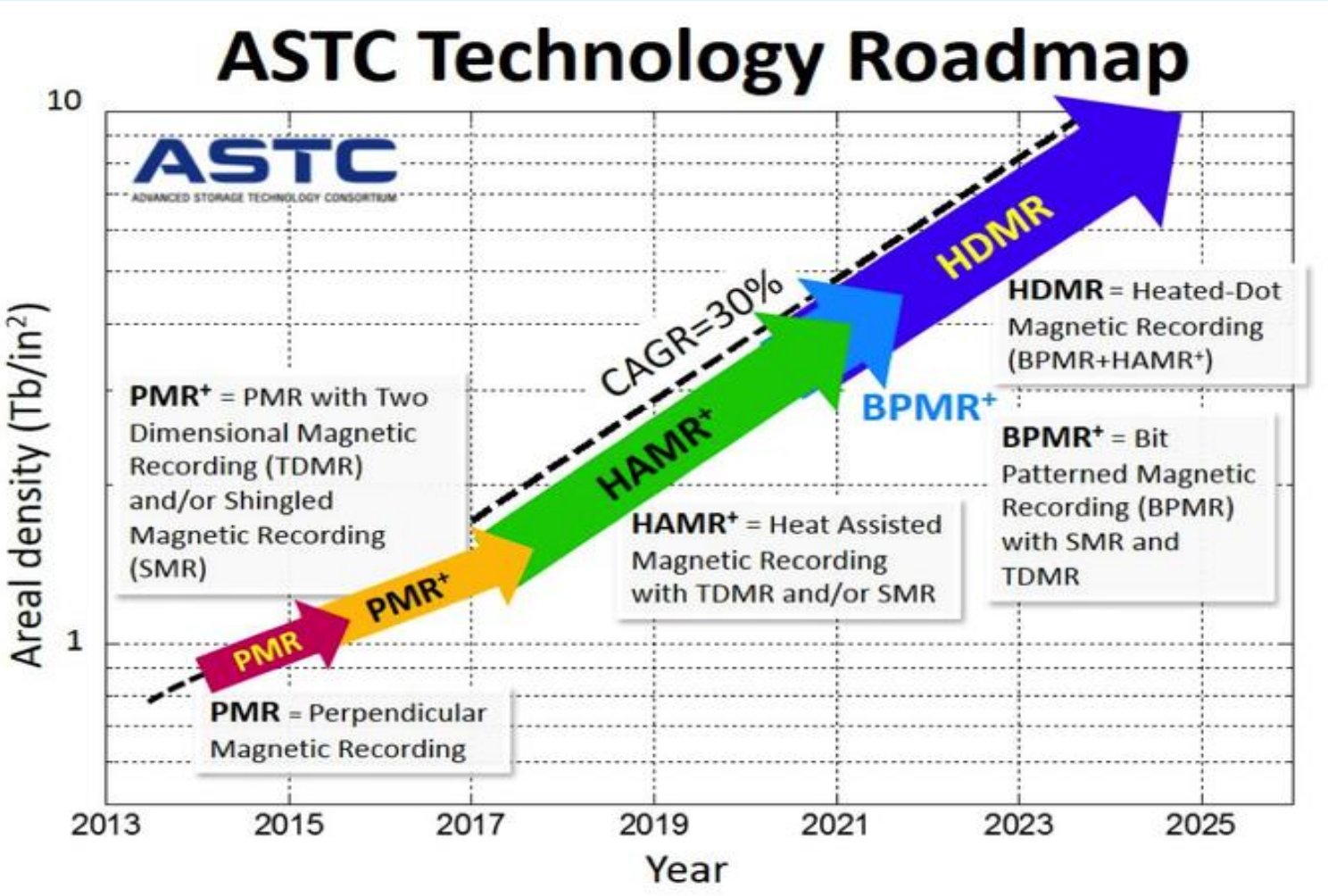


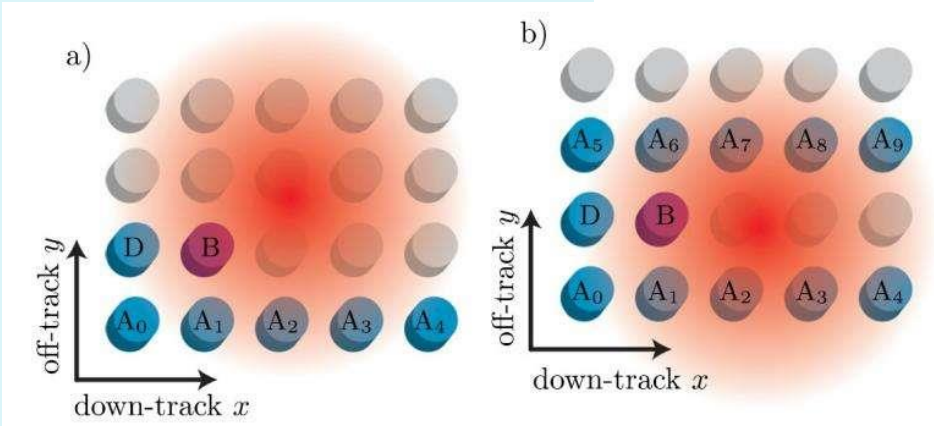


The Bloch sphere is a representation of a qubit, the fundamental building block of quantum computers.

New technology has been developed to increase the in storage capacitance.

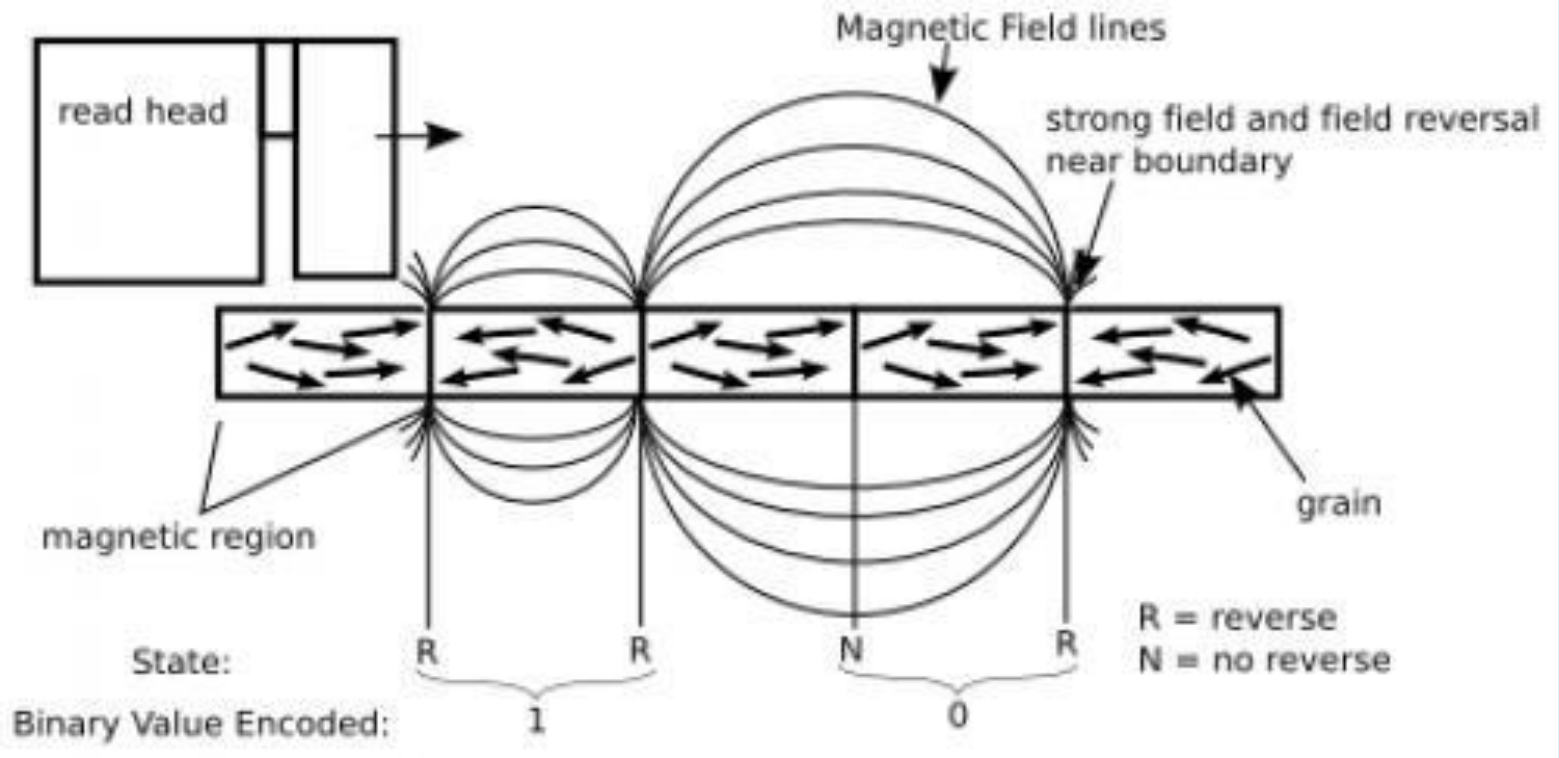


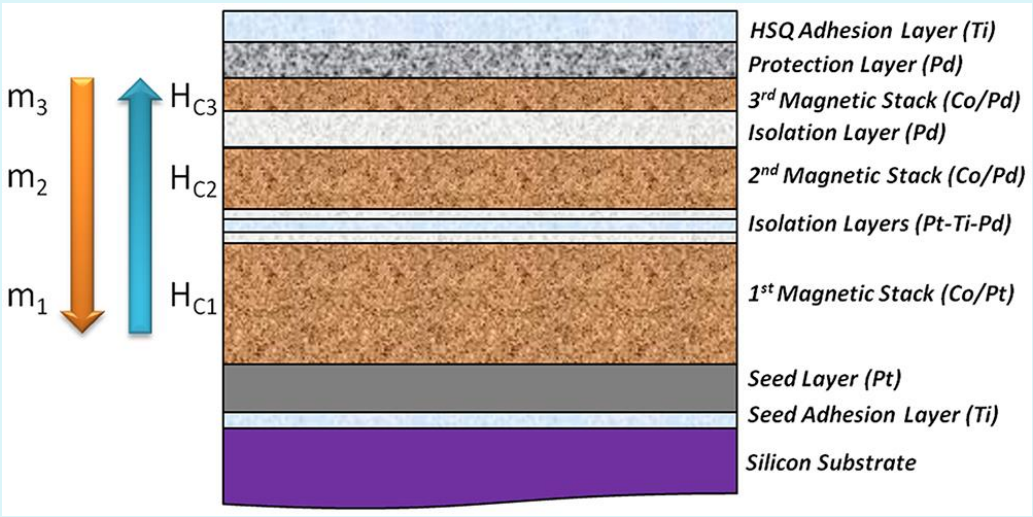


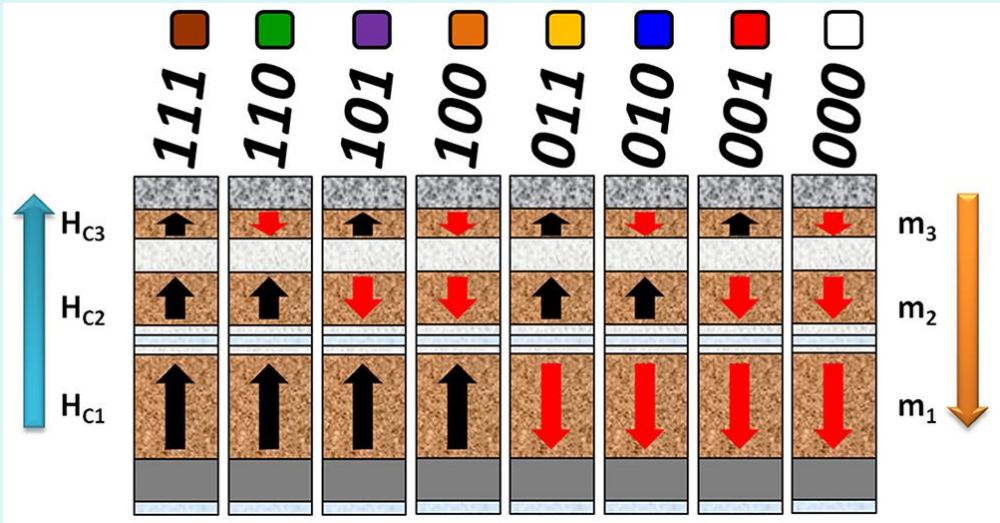


Emerging Technologies For Capacity Growth

3D magnetic storage breakthrough enables 100TB+ hard drives

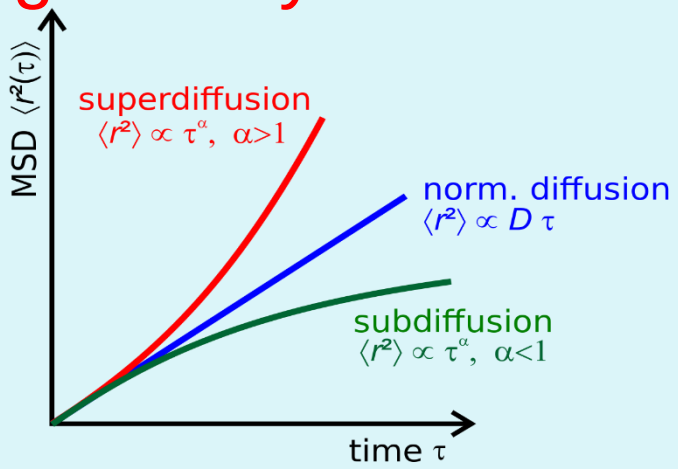




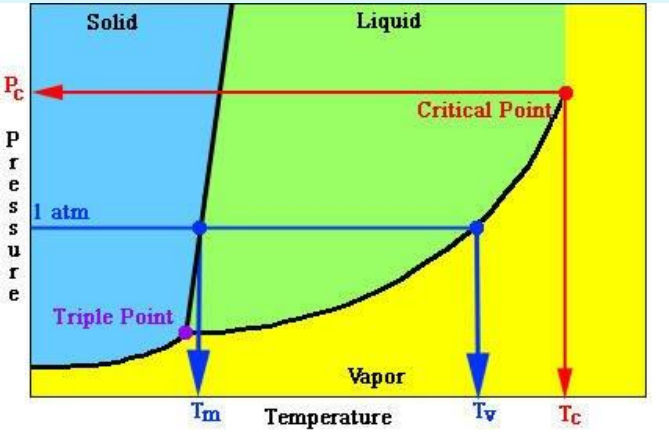


New scientific methods to describe big data systems

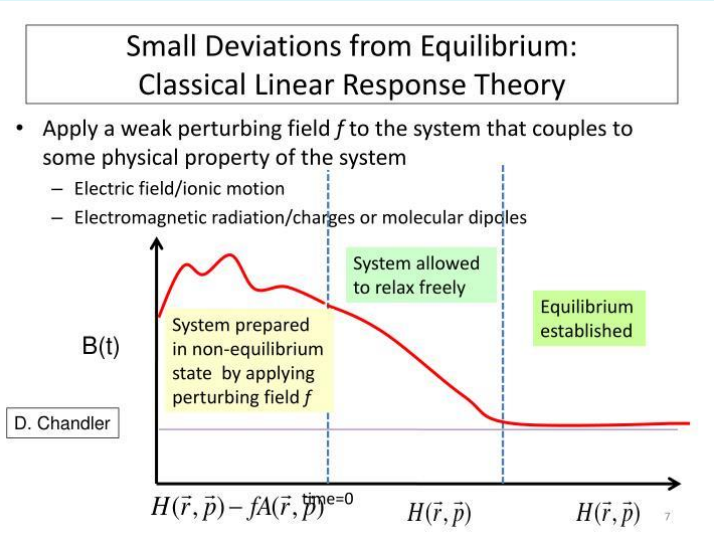
1. Mean square displacement (MSD) analyzes of dynamical properties in collected data. This will enable to segregate processes on different time scales, to detect fast and slow processes, to define saturation tendencies.



2. Defining Phase Diagrams for different states of the data. For example, it is possible to produce a general law which describes how the democracy in any society depends on the oil produced per capita.



3. Assessing the Stability of Data through its linear/nonlinear response to external loads. For example, how trustable are observed correlations between processes A and B in the normal state of the economy if the processes decorrelate under financial crisis conditions.



Other approaches: higher order correlations q4-q6 analyses, DFT method for the ground state search., etc.

CONCLUSIONS

Conclusion 1: Whereas Big Data detects patterns and correlations, Hard sciences focus on 'what causes these correlations, how these correlations are related to other correlations in other systems, and what is the impact of initial conditions and other system parameters?

Conclusion 2: The successful application of Big Data methods in soft science and data-driven science can change data sharing in hard science (scientists withhold information, claim ownership, credits for publication, very tight competition, less sharing with other labs), and make it easy for younger scientists to freely access experimental data and simulation codes for his projects.
In other words, an interaction with big data hopefully will contribute to the data sharing in hard science.

Conclusion 3. Rapid evolution of Big data is a driving force for hard science progress. Quantum Computer Qubits, magnetic dot storages are bright examples.

Conclusion 4. The hard science can contribute to the
a) dynamical analyses of big data applications ,
b) parametrizing the big data findings to create theoretical models and mathematical apparatus for it,
c) while we understand the present and the past with the big data, hard science methods will give us ability to predict future and make decisions about future.

Conclusion 5. The hierarchy in the Big Data can be built using hard science methods, which can turn the Big Data into a solid scientific discipline.

Final Conclusion. BIG DATA needs BIG THEORY.

Thank you for your attention and the opportunity for being at BDDB-2018