Scientific Data Format

The Incredible Life of Scientific Data

Graziano Giuliani
ggiulian@ictp.it

STIMULATE – Ferrara 2019

BLOOM

one year in a Virginia garden

INFO WE TRUST

Information is in Data

- Information is the form we give to our ideas
  - From Latin : give form to an idea
- Any information comes from data we collect
  - We live in an out of equilibrium universe which changes in time
  - Any transition between different states generates data we can collect
    - Our minds are built in such a way that we put reason and meaning to change
    - History is sum of our trusted views to which we give significant meaning
- Science objective is to find the best rules that can explain reality behind the endless data input we are subject to

Where do data come from?

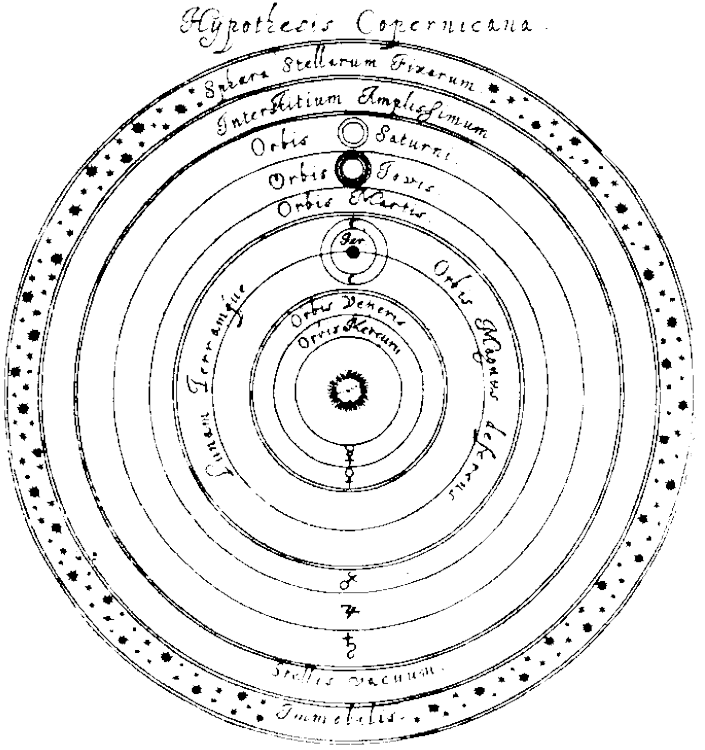
- Any change that leaves a record somewhere is a datum
  - In “nature” any event generates a new state
    - Records fade in time
    - Multiple event signatures are mixed
  - Human mind is the only known apparatus which is able to think on how to think
  - Feeding information to human mind is the objective of data collection

Scientific data

- From Galilei onward, the scientific method is established as the most successful process to interpret reality by analyzing data and find mathematical abstract relationships between measurable quantities. The Scientific process is the only human mind construct which by itself has impacted the very way we perceive reality

How it works

- We detect something
- We try to explain what we have seen and build an hypothesis
- We test our new knowledge against nature in experiments
- If hypothesis survives the test, we call it a scientific theory and we use it until we have a new one which better fits reality or new evidence undermines it requesting the process to start over
- Nature's reality is always the final judge of a scientific theory.



Working with data

- We need data
  - We need to understand data
    - To get someone trust on our model of reality
    - New established views
  - New questions rise from knowledge
- And we need more data….

Collect data

- Data collection requires a clean process
  - Understand every measuring device, identify errors, missing values, and corrupt records
  - Throw away, replace, and/or fill missing values/errors
  - Encoding removes details to concentrate our understanding of the data

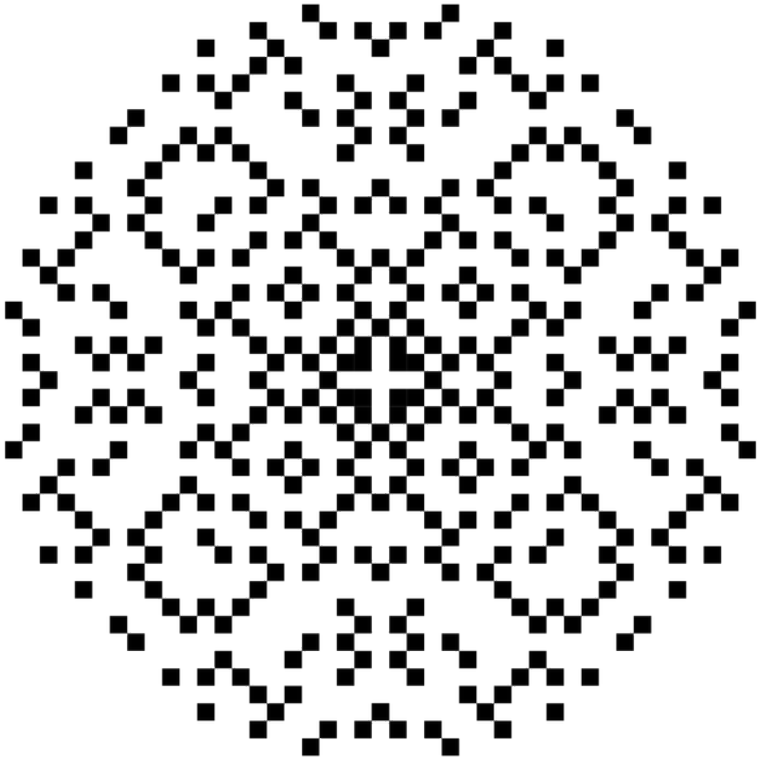
Too much data!

- Big jump in technology has generated multiple information inflation events
  - Constant move towards more durable, portable, compact
    - Written language
    - Libraries
    - Movable printing types
    - Digital encoding and storage
  - Authority crises: which information is trustworthy?

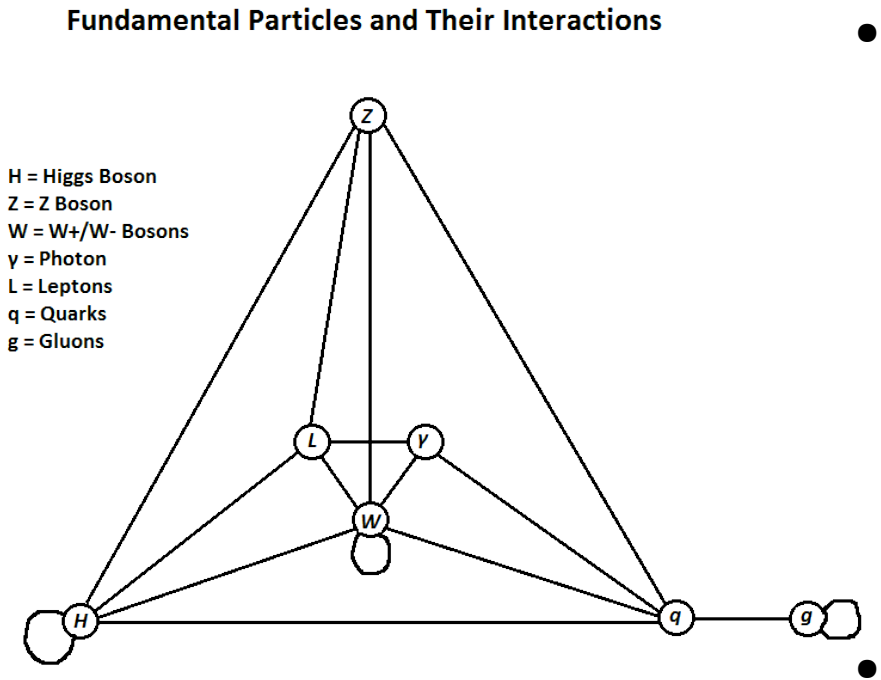
信任

Data Analysis

- We must now understand what patterns and values our data has to back up our findings.
  - Different types of visualizations and statistical testings
  - Derive hidden meanings behind data through various graphs and analysis.



Model



- Find a Mathematical model of the reality
  - Predictive Algorithms
  - Test predictions against measurements
  - Iterate until satisfaction
  - Side effect: paper publication
- Model results are new data in themselves

Data Visualization

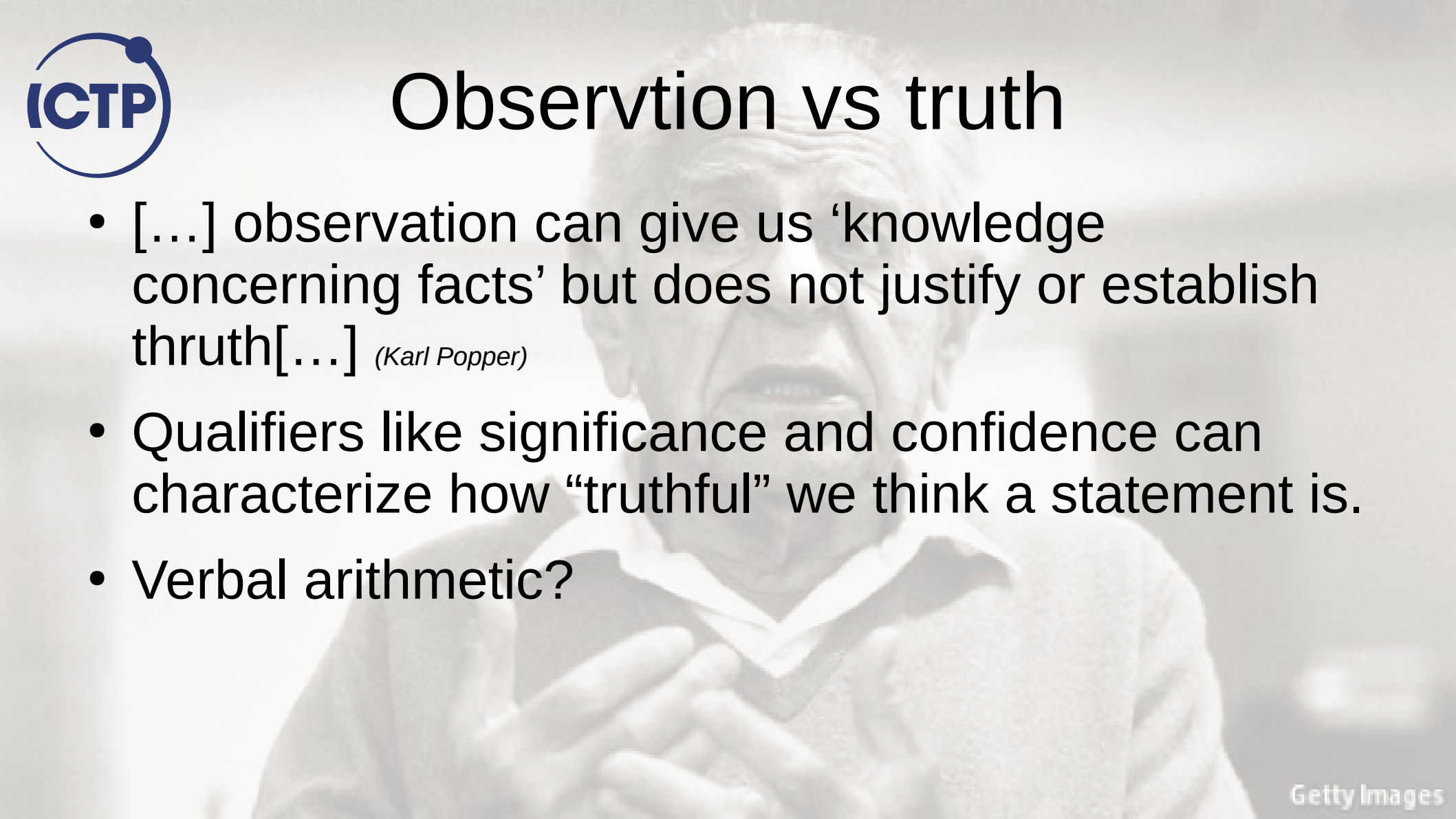
- Scientific results must be presented to humans
  - Humans have limited capabilities
    - Best sensors are the eyes
      - Good images are worth thousand pages of text
      - Our minds are capable of seeing limited patterns and understand multiple variations only up to small number of dimensions
    - Images and words together have greater impact
- Artists have unconventional ways to produce big effects on human minds
  - Not always a scientist is a good graphical artist

Human Mind

- What is human mind better at
  - Comparison-making machine
    - Deviations capture our attention if on single image
      - Better at add/subtract than multiply/divide
      - Good at catching behavior or pattern
    - Better with images than words
  - Certain. Probable. Confident. Reliable. Meaningful. Significant. How trustworthy a statement is?

Observtion vs truth

- […] observation can give us ‘knowledge concerning facts’ but does not justify or establish thruth[…] *(Karl Popper)*
- Qualifiers like significance and confidence can characterize how “truthful” we think a statement is.
- Verbal arithmetic?



Confidence

- Confidence is expressed qualitatively and tells us how certain we are that scientific findings are valid. The level of confidence is determined by the type, amount, quality and consistency of evidence. A “very high confidence” means that there is at least a 9 in 10 chance of a finding being correct.

Confidence Terminology	Degree of confidence in being correct
Very high confidence	At least 9 out of 10 chance
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very low confidence	Less than 1 out of 10 chance

Certainty

- The certainty of scientific findings is then described using likelihoods. Findings are assessed probabilistically using observations, modeling results or expert judgment.

Likelihood Terminology	Likelihood of the occurrence/ outcome
Virtually certain	> 99% probability
Extremely likely	> 95% probability
Very likely	> 90% probability
Likely	> 66% probability
More likely than not	> 50% probability
About as likely as not	33 to 66% probability
Unlikely	< 33% probability
Very unlikely	< 10% probability
Extremely unlikely	< 5% probability
Exceptionally unlikely	< 1% probability

Storytelling

Single stream and different timelines

Forward, backward or circular streams

Repetition and multi-causal networks

Offer supposedly privileged point of views

Pose questions and then callback to give answers

Build meaning along with answer to questions

Provide answers thwarting expectations

Keeping data

University of Washington Libraries, Special Collections Division

- The experiment is difficult or impossible to repeat
- The phenomenon is rare
- The knowledge comes out by statistical analysis of the data
- A changing process is in act and we are interested in the change effects
- The cost of the measure is high
- The measure has economical impact
- New theory can falsify the established by better explanations

Data archive

- Almost no scientist can use just paper and pencil to record reality. And Darwin did all his work with simple notebooks...
- Technology has changed the way we keep data
- Records in a digital world are coded on magnetic supports to have high information density
  - Will this magnetic supports last for centuries?
  - Will readers for this supports be available indefinitely in the future?
  - Will the coding used be crackable in the future?

Data Shepherd

- Data used for science must have a shepherd in charge of their maintenance
  - Data must be readable from their support
  - Metadata must be accessible and support queries
  - Metadata semantics must be sensible and accurate
  - User access to data is granted with clear procedures
  - Data authoring and citation is granted

Data embedding?

- Why not have data and data analysis workflow neatly included in the paper writing process?
- Enter Stephen Wolfram
  - Mathematica® was launched in 1988
  - “A New Kind of Science” of 2002
  - Wolfram’s proprietary notebook showcased innovative technology, but decades after its introduction, still has few users.
- Where is the “good idea”?
  - **THE NOTEBOOK**

Mathematica

- Mathematica exemplify the wonderful product of a single mind:
  - Monolithic
  - Consistent
  - Integrated
  - Closed
- Proprietary business model
  - Capital but not mind driven
  - Corporate and not Science driven
  - Increase inequalities among rich and poor scientists

Jupyter

- Jupyter ports Notebook technology to the masses
- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Format

- Jupyter Notebooks are an open document format based on JSON. They contain a complete record of the user's sessions and include code, narrative text, equations and rich output.
- JavaScript Object Notation (JSON) is an open-standard file format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types (or any other serializable value). It is a very common data format used for asynchronous browser–server communication, including as a replacement for XML in some AJAX-style systems

Computing

- Computing part is performed by sending messages to a kernel
- Kernels are processes that run interactive code in a particular programming language and return output to the user. Kernels also respond to tab completion and introspection requests.
- The Notebook communicates with computational Kernels using the Interactive Computing Protocol, an open network protocol based on JSON data over ZMQ (high-performance asynchronous messaging library) and WebSockets (standard for full-duplex communication channels over a single TCP connection)

What it looks like?

- https://nbviewer.jupyter.org

- https://github.com/graziano-giuliani/STIMULATE/blob/master/stimulate.ipynb