Mathematical Use and Abuse of Big Data in Biology and Medicine

Giuseppe Longo

Centre Cavaillès, CNRS et ENS, Paris
and Biology Departement, Tufts Univ., Boston
ww.di.ens.fr/users/longo

Calude C., Longo G. *The Deluge of Spurious Correlations in Big Data*, **Found. Sci.,** 2016
(http://www.di.ens.fr/users/longo/download.html)

Big Data

IBM [18 : "What is Big Data ?"] estimates that

« Every day, we create **2.5 quintillion bytes of data –** so much that 90% of the data in the world today has been created in the last two years alone. »

Fantastic tool for knowledge and science!

After Greek **observation** and **speculation :**

Experimental method (Galileo),
Mathematics (Descartes, Newton)
Immense Databases (if soundly used …)

E.g. *statistics* or extensive *use* of the immense data bases on,

Big Data

IBM [18 : "What is Big Data ?"] estimates that

« Every day, we create **2.5 quintillion bytes of data –** so much that 90% of the data in the world today has been created in the last two years alone. »

Fantastic tool for knowledge and science!

After Greek **observation** and **speculation :**

Experimental method (Galileo),
Mathematics (Descartes, Newton)
Immense Databases (if soundly used …)

E.g. *statistics* or extensive *use* of the immense data bases on, e.g. Biogical Rhythms (cardiac, metabolic …) in

Longo G., Montévil M., **Perspectives on Organisms: Biological Time, Symmetries and Singularities**, *Springer*, Berlin, 2014.

Big Data analysis as « The End of Science »

Big Data analysis as « The End of Science »

C. Anderson, 2008: « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete »

« Correlation is enough . . . . We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms **find patterns** where science cannot ».

« with enough data, the numbers speak for themselves ... **Correlation supersedes causation**, and science can advance even without coherent models, unified theories. »

The largest the best … *Independently* of any analysis of the "meaning" or "content", prediction and **rules for action** are **provided by the data mining** (NSF project).

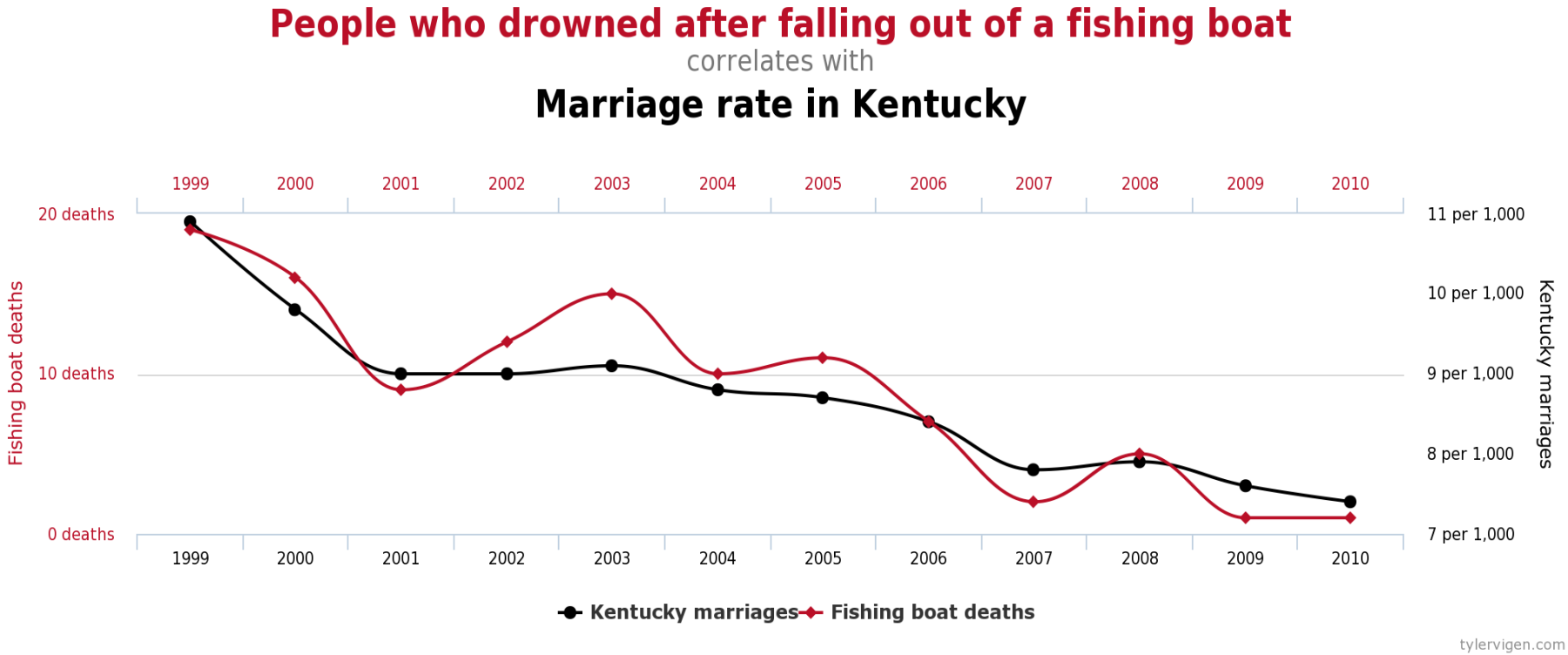
References are in:
Calude C., Longo G. The Deluge of Spurious Correlations in Big Data, 2016 (http://www.di.ens.fr/users/longo/download.html).

Just looking at Big Data

An **empirical** response: A large collection of spurious correlations : http://www.tylervigen.com/spurious-correlations, 2015.

Just looking at Big Data

An **empirical** response: A large collection of spurious correlations : http://www.tylervigen.com/spurious-correlations, 2015.



Use Mathematics to fight the Big Data Folies

PART I: By some use of “Ramsey theory” (born in the 1920's), prove:
all sufficiently large set of numbers contains Spurious Correlations

PART II: Since this large enough **data set is arbitrary**, it may have been obtained by a **random** generator of digits or numbers (series of dice throws or quantum spins measurements).

PART III: My motivations from Cancer Biology (Soto Lab., Boston)

Use Mathematics to fight the Big Data Folies

PART I: By some use of “Ramsey theory” (born in the 1920's), prove:

Informal: "Given **any arbitrary correlation** on sets of data, **there exists** a large enough number (size) such that **any data set of that size or more**, realises that type of correlation."

Calude C., Longo G. The Deluge of Spurious Correlations in Big Data, 2016
(http://www.di.ens.fr/users/longo/download.html)

Use Mathematics to fight the Big Data Folies

PART I: By some use of “Ramsey theory” (born in the 1920's), prove:

Informal: "Given **any arbitrary correlation** on sets of data, **there exists** a large enough number (size) such that **any data set of that size or more**, realises that type of correlation."

PART II: Since this large enough **data set is arbitrary**, it may have been obtained by a **random** generator of digits or numbers (series of dice throws or quantum spins measurements).

Note: it is exactly the **size of the data** that allows our result: the more data, the more arbitrary, meaningless and useless (for future action) correlations will be found in them. *How large?*

Calude C., Longo G. The Deluge of Spurious Correlations in Big Data, 2016
(http://www.di.ens.fr/users/longo/download.html)

"Colored" Van der Waerden and Ramsey Theorems

"Colored" Van der Waerden

Finite **Van der Waerden theorem** (for sequences of digits or colors):

For any positive integers c and k there is a positive integer γ such that every string, made out of c digits or colours, of length more than γ contains an arithmetic progression with k occurrences of the same digit or colour, i.e. a monochromatic arithmetic progression of length k.

"Colored" Van der Waerden

Finite **Van der Waerden theorem** (for sequences of digits or colors):

For any positive integers c and k there is a positive integer γ such that every string, made out of c digits or colours, of length more than γ contains an arithmetic progression with k occurrences of the same digit or colour, i.e. a monochromatic arithmetic progression of length k.

c = 2 (**B**lack/**R**ed)**,** k = 3

B R R **B** **B** R R **B**

"Colored" Van der Waerden

Finite **Van der Waerden theorem** (for sequences of digits or colors):

For any positive integers c and k there is a positive integer γ such that every string, made out of c digits or colours, of length more than γ contains an arithmetic progression with k occurrences of the same digit or colour, i.e. a monochromatic arithmetic progression of length k.

c = 2 (**B**lack/**R**ed), k = 3

B R R **B** **B** R R **B** R (3 R at distance 3)
B R R **B** **B** R R **B** **B** (3 **B** at distance 4)

W(2,3) = 9 (the Van der Waerden number: any length 9 sequence ...)

"Colored" Ramsey Theorems

Finite **Ramsey theorem** (for n-ary relations **[A]n** on a set A):

For all positive integers s, n, c there is a positive integer γ such that for every finite set A containing more than γ elements and for every partition $P : [A]n \rightarrow \{1, 2, \ldots, c\}$ there exists a subset B of A containing s elements whose n-sets are monochromatic, i.e. P(x) has the same value (colour) for every x in [B]n .

"Colored" Ramsey Theorems

Finite **Ramsey theorem** (for n-ary relations **[A]n** on a set A):

For all positive integers s, n, c there is a positive integer γ such that for every finite set A containing more than γ elements and for every partition $P : [A]n \rightarrow \{1, 2, \ldots, c\}$ there exists a subset B of A containing s elements whose n-sets are monochromatic, i.e. P(x) has the same value (colour) for every x in [B]n .

Very vague and informal intuition:
Take two gases of different colors, mixing up in a box of size 10; suppose that one can see the different colors of the molecules (!); **Given any** length *l* below 10 (our *s*,*n*,*c*), these theorems compute the amount of time (*γ)* one has to wait to “see” a regularity (a straight line, say) of length at least *l* ….

Hints to applications of Ramsey Theorems

Let ***A*** be a relational database. Fix s, n, c,
a **correlation of variables** in ***A***
is a set B of size s (e.g. the number of years and quantities)
whose n-ary relations (a1, a2 … an)
form the correlation (for the *given criteria* or colors c)

When the **correlation applies**, all elements are given **the same color**, out of c (are monochromatic).

Then **by Ramsey theorem** one has that:
given any "correlation", i.e. any s, n and c, there always exists a large enough number γ such that any set A of size greater than γ, in any way "P" it colored, contains a set B of size s whose subsets of n elements (n-ary relations) are all correlated – that is, monochromatic.

Since A is arbitrary, it may be generated by a **random** process ...

NEXT

Ramsey: How large is γ ?

Let c = 2 and $\gamma = R(s,n)$ the Ramsey number of 2, s and n
(i.e. given r and n, any set ***A*** of cardinality $R(r,n)$ contains a subset **B** of cardinality s, with [**B**]n monochromatic)

Immensly large if ***card(B) > min(B)*** [Paris-Harrington, 1978; Longo, 1981]

Upper and lower bounds have been computed for R(s,s): these are exponentials compatible with today's size of Big Data
[Erdos, Szkeres, 1947; Szemered, 1980; Conlon, 2009]

For n = s : A (corrected) exponential upperbound:

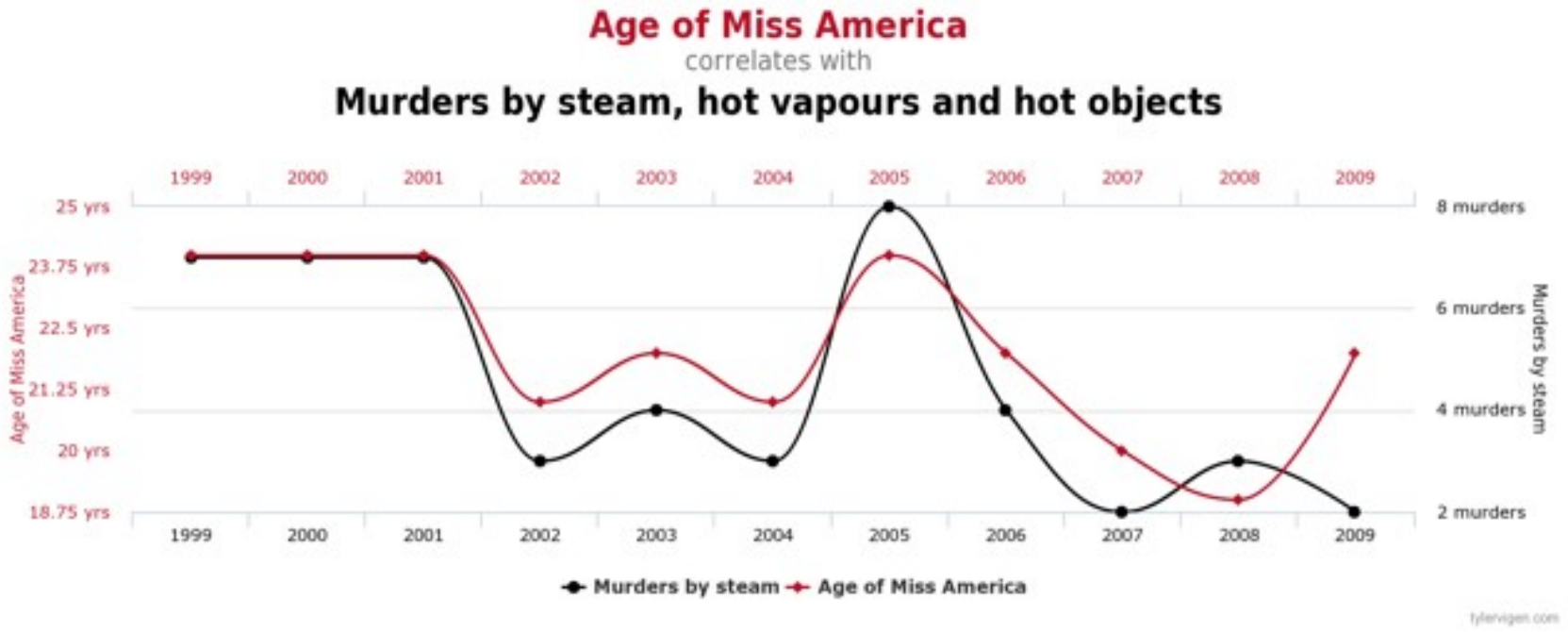
$$R(s,s) \leq [1+o(1)]\frac{4^{s-1}}{\sqrt{\pi s}}.$$

An exponential lower bound,

$$R(s,s) \geq [1+o(1)]\frac{s}{\sqrt{2}e}2^{s/2},$$

What is a ''spurious'' correlation ?

Theory dependent *definition …*



How many "spurious" correlations ?

''Spurious'' a relative notion

Define: a correlation is "**spurious**" when it belongs to a *randomly generated set A* set

A very weak definition ….

But then, what "random" means ?

Randomness for sequences of numbers

Randomness for sequences of numbers

Algorithmic Information Theory

(Kolmogorof, 1960; Martin-Löf, 1965; Chaitin, 1970; Calude, 2002) :

Analysis of **randomness** for *sequences of numbers*

Easy extension to n-ary relations, i.e. valid for both VdW and Ramsey frames:

- any finite set is computationally isomorphic to a sequence with a "low cost" of coding

How many ''random'' A ?

Random finite sequence, an approximation: **incompressible**

In **practice** (compression algorithms) : The best average rate of algorithmic compressibility is of about 86.4%

Example:
A string x is compressed into a file Zip(x) which has about 86,4% of the length of x.

The **probability that a binary string x** of (relatively short) length 2048 is reduced by 13.6%

is smaller than 10expo(-82)

(10expo(82) is the number of hydrogen atoms in the Universe)

In other words, for large n, very few strings of length n are compressible, *that is surely* **not random**.

Warning: sound Statistical Analyses

Current Statistical Analysis are **hypothesis** (thus, **theory**) **driven**:
- ***research*** *hypotheses* (with alternatives),
- ***null*** *hypothesis* (*no sense* relationship between two data sets)
- give ***probability's*** *thresholds*

Typically:

- A comparison is **statistically significant** if the relationship between the data sets would be an unlikely realization of the *null hypothesis* given a **threshold probability—**the significance level.
- The process of distinguishing between the *null hypothesis* and the **alternative hypothesis** is aided by identifying two conceptual types of errors (type 1 & type 2), and by specifying parametric limits on e.g. how much type 1 error will be permitted.

Note : **type I** error is the incorrect rejection of a true null hypothesis (a "**false positive**"), while a **type II** error is the failure to reject a false null hypothesis (a "**false negative**").

Part III: Motivations for Big Data analysis, The case of Cancer

Collaboration, since 2008 with

C. **Sonnenschein**, A. **Soto**
Department of Integrative Physiology and Pathobiology
Tufts University School of Medicine, Boston

Tissue Organization Field Theory (**TOFT**)

M. **Montévil**,
PhD student, then joint post-doc U. Boston and ENS, Paris

http://sackler.tufts.edu/Faculty-and-Research/Faculty-Research-Pages/Ana-Soto-and-Carolos-Sonnenschein

DNA decoding, 2000-01 (fantastic tehcnological achievement)

Robert A. Weinberg (Biology, MIT),
a major promotor of the *Somatic Mutation Theory* (SMT) of cancer :

Co-author of a "classic" synthesis:

Hanahan D and Weinberg RA. *The hallmarks of cancer*. **Cell,** 100, 57–70, **2000**. ***(20,000 citations** by 2010)*

Year 2000 (-01): the "decoding of DNA"

« ... **cancer biology and treatment … will become a science** with a conceptual structure and logical coherence that rivals that of chemistry or physics »

Within the frame of SMT: Nowell, P. C. *The clonal evolution of tumor cell populations*. **194** , 123-128 (1976)

DNA decoding, 2000-01 (fantastic tehcnological achievement)

F. Collins, 2001: « we have grasped the **code written by God** »

C. Venter, 2001: the "decoder" of the human genome

A. von Eschenbach, director Nat. Cancer Inst. 2003: "to eliminate the suffering and death from **cancer**, and to do so by 2015"
Diagnosis and prognosis withn two or three years … *NO WAY !*

DNA decoding, 2000-01 (fantastic tehcnological achievement)

F. Collins, 2001: « we have grasped the **code written by God** »

C. Venter, 2001: the "decoder" of the human genome

A. von Eschenbach, director Nat. Cancer Inst. 2003: "to eliminate the suffering and death from **cancer**, and to do so by 2015"
Diagnosis and prognosis withn two or three years … *NO WAY !*

C. Venter, interview for the *Spiegel*, *July 29, 2010*:
Title: « **We have learned nothing from the genome** »
« ... phonies … the ill-founded belief that those who know the DNA sequence also know every aspect of life. This nonsense ... »

Yet, **we did learn a lot**: *the case of cancer* ..

Cancer and the DNA decoding

From the massive DNA decoding of cells in cancer tissues:

1 - Gene-expression signatures for benign and malignant cancer may coexist in the same tumor.
2 - Genetic analyses do not allow to discriminate between a tumor that (has or) will metastasize(d) from another that (has or) will not.
3 – DNA sequencing does not help in distinguishing a primary from a metastatic cancer.

(Maffini et al. 2005; Hendrix et al. 2007; Bussard et al. 2010 ; Imielinski et al., 2012; Gerlinger et al. *Engl J Med* 366;10, 2012)

References in:

http://www.di.ens.fr/users/longo/files/BiologicalConseq-ofCompute.pdf

''Coming Full Circle – form endless comlexity to simplicity and back again'' by R. A. **Weinberg**, MIT Center for Molecular Oncology,
Cell 157, March 27, *2014*

2014 : Somatic Mutation Theory (SMT): Capitulation

''Coming Full Circle – form endless comlexity to simplicity and back again'' by R. A. **Weinberg**, MIT Center for Molecular Oncology,
Cell 157, March 27, *2014*

2014 : Somatic Mutation Theory (SMT): Capitulation

« **Half a century** of cancer research had generated an enormous body of observations about the behavior of the disease, but there were **essentially no insights** into how the disease *begins and progresses* to its life-threatening conclusions. »

« … essentially incoherent phenomena that constituted "cancer research [at the molecular level]" … **one should never, ever confuse cancer research with science** »

DISTINCTION : BIOLOGICAL VS CLINICAL RESEARCH

Cancer and Big Data

Since « … one should never, ever confuse cancer research with science » … « myriads of unexpected mutations » (Weinberg, 2014)

Let's then predict and act on the grounds of Dig Data !

Cancer and Big Data

Since « … one should never, ever confuse cancer research with science » … « myriads of unexpected mutations » (Weinberg, 2014)

Let's then predict and act on the grounds of Dig Data !

Purely **Big Data** Driven cancer research: ***predict*** *and* ***act***:

Cancer Institute, Oregon Health & Science Univ. & Intel, 2016 : http://www.informationweek.com/big-data/big-data-analytics/can-big-data-help-cure-cancer-/d/d-id/1326295

Many Biology University Labs & IBM, 2016: http://www.businessinsider.in/IBMs-Watson-can-now-do-in-minutes-what-takes-cancer-doctors-weeks/articleshow/47168413.cms

Some references (*downloadable:* Google: Giuseppe Longo)

Bravi B., Longo G. Biology, from Noise to Functional Randomness. *in* Springer LNCS 9252, pp 3-34, 2015

Calude C., Longo G. The Deluge of Spurious Correlations in Big Data, *in* Foundations of Science, 1-18, March, 2016

Calude C., Longo G. Classical, Quantum and Biological Randomness as Relative Unpredictability. Spec issue, **Natural Computing**, Springer, 2016.

Soto A., Longo G, Noble D. (eds.) **From the century of the genome to the century of the organism: New theoretical approaches**, *Special issue of* **Progress in Biophysics and Molecular Biology,** Vol. 122, 1, Elsevier, 2016.

Longo G. The Biological Consequences of the Computational World: Mathematical Reflections on Cancer Biology, ***submitted,*** 2017.