In Praise of Small Data

Statistical Science and Data Science

Nancy Reid
University of Toronto

Department of Mathematics
Imperial College
20 March 2019

Outline

Statistics at a Crossroads

Examples: Statistics in the news

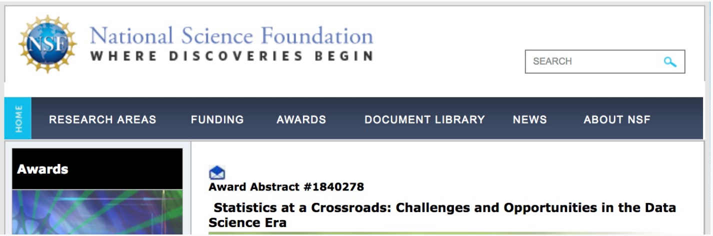
Statistical theory

Statistics and data science

Statistics at a Crossroads

Statistics at a crossroads

NSF award



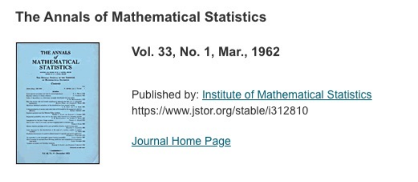
Award Abstract #1840278

Statistics at a Crossroads: Challenges and Opportunities in the Data Science Era

- NSF workshop and report
- "... we are at a crossroads with an unprecedented opportunity to modernize ... to become the major player in data science, but also with a non-ignorable risk to make ourselves obsolete in the broad community of data science."
- "... critical question, where do we go from here?"

... crossroads

Tukey 1962



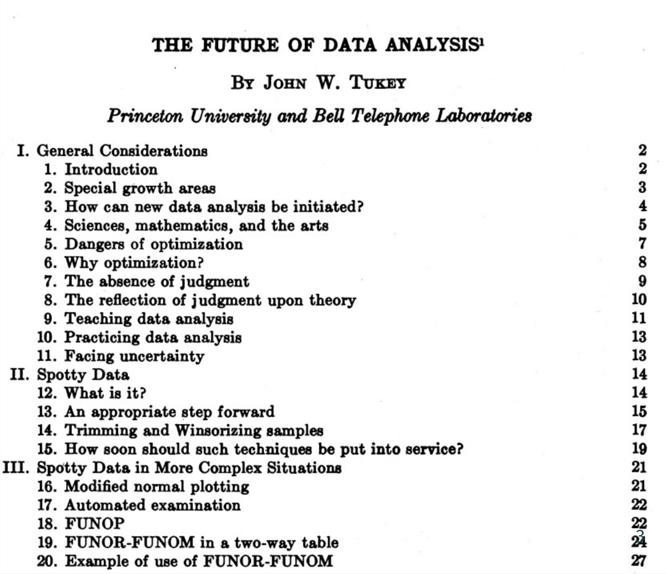
The Annals of Mathematical Statistics

Vol. 33, No. 1, Mar., 1962

Published by: Institute of Mathematical Statistics
https://www.jstor.org/stable/i312810

Journal Home Page

"The future of data analysis can ... lead to the provision of a great service to all fields of science and technology. Will it? That remains to ... our willingness to take up the rocky road of real problems in preferences to the smooth road of unreal assumptions ... Who is for the challenge?"



THE FUTURE OF DATA ANALYSIS[1]

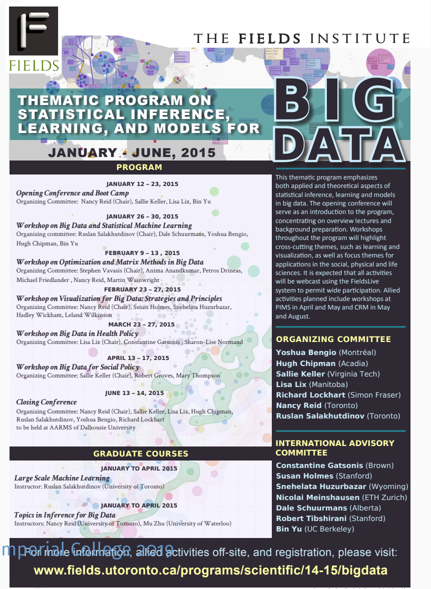
By John W. Tukey

Princeton University and Bell Telephone Laboratories

I. General Considerations 2
1. Introduction 2
2. Special growth areas 3
3. How can new data analysis be initiated? 4
4. Sciences, mathematics, and the arts 5
5. Dangers of optimization 7
6. Why optimization? 8
7. The absence of judgment 9
8. The reflection of judgment upon theory 10
9. Teaching data analysis 11
10. Practicing data analysis 13
11. Facing uncertainty 13
II. Spotty Data 14
12. What is it? 14
13. An appropriate step forward 15
14. Trimming and Winsorizing samples 17
15. How soon should such techniques be put into service? 19
III. Spotty Data in More Complex Situations 21
16. Modified normal plotting 21
17. Automated examination 22
18. FUNOP 22
19. FUNOR-FUNOM in a two-way table 24
20. Example of use of FUNOR-FUNOM 27

Big Data

2014



FIELDS

THE FIELDS INSTITUTE

THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR

BIG DATA

JANUARY - JUNE, 2015

PROGRAM

JANUARY 12 – 23, 2015

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

JANUARY 26 – 30, 2015

Workshop on Big Data and Statistical Machine Learning

Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

FEBRUARY 9 – 13 , 2015

Workshop on Optimization and Matrix Methods in Big Data

Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander , Nancy Reid, Martin Wainwright

FEBRUARY 23 – 27, 2015

Workshop on Visualization for Big Data: Strategies and Principles

Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson

MARCH 23 – 27, 2015

Workshop on Big Data in Health Policy

Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis , Sharon-Lise Normand

APRIL 13 – 17, 2015

Workshop on Big Data for Social Policy

Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson

JUNE 13 – 14, 2015

Closing Conference

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University

GRADUATE COURSES

JANUARY TO APRIL 2015

Large Scale Machine Learning

Instructor: Ruslan Salakhutdinov (University of Toronto)

JANUARY TO APRIL 2015

Topics in Inference for Big Data

Instructors: Nancy Reid (University of Toronto), Mu Zhu (University of Waterloo)

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops throughout the program will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life sciences. It is expected that all activities will be webcast using the FieldsLive system to permit wide participation. Allied activities planned include workshops at PIMS in April and May and CRM in May and August.

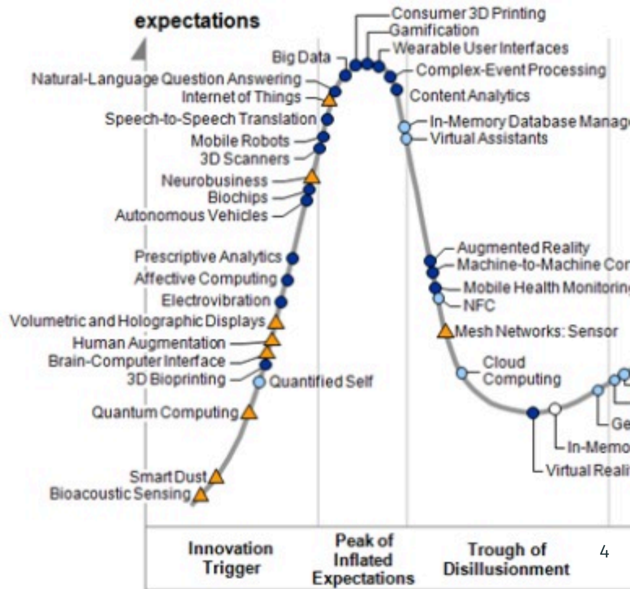
ORGANIZING COMMITTEE

Yoshua Bengio (Montréal)
Hugh Chipman (Acadia)
Sallie Keller (Virginia Tech)
Lisa Lix (Manitoba)
Richard Lockhart (Simon Fraser)
Nancy Reid (Toronto)
Ruslan Salakhutdinov (Toronto)

INTERNATIONAL ADVISORY COMMITTEE

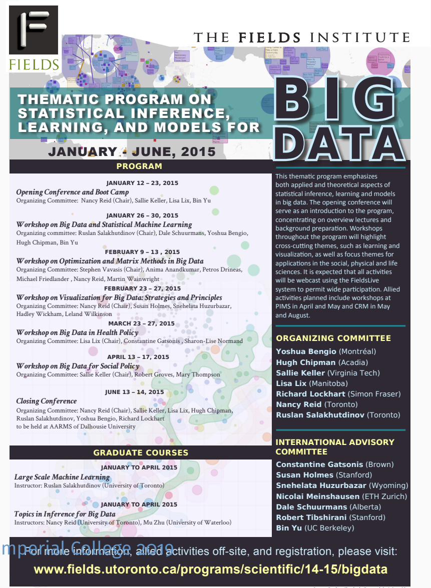
Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehelata Huzurbazar (Wyoming)
Nicolai Meinshausen (ETH Zurich)
Dale Schuurmans (Alberta)
Robert Tibshirani (Stanford)
Bin Yu (UC Berkeley)

For more information, allied activities off-site, and registration, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/bigdata



Big Data

2014



FIELDS

THE FIELDS INSTITUTE

THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR

BIG DATA

JANUARY - JUNE, 2015

PROGRAM

JANUARY 12 – 23, 2015

Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

JANUARY 26 – 30, 2015

Workshop on Big Data and Statistical Machine Learning
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

FEBRUARY 9 – 13 , 2015

Workshop on Optimization and Matrix Methods in Big Data
Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander , Nancy Reid, Martin Wainwright

FEBRUARY 23 – 27, 2015

Workshop on Visualization for Big Data: Strategies and Principles
Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson

MARCH 23 – 27, 2015

Workshop on Big Data in Health Policy
Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis , Sharon-Lise Normand

APRIL 13 – 17, 2015

Workshop on Big Data for Social Policy
Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson

JUNE 13 – 14, 2015

Closing Conference
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University

GRADUATE COURSES

JANUARY TO APRIL 2015

Large Scale Machine Learning
Instructor: Ruslan Salakhutdinov (University of Toronto)

JANUARY TO APRIL 2015

Topics in Inference for Big Data
Instructors: Nancy Reid (University of Toronto), Mu Zhu (University of Waterloo)

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops throughout the program will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life sciences. It is expected that all activities will be webcast using the FieldsLive system to permit wide participation. Allied activities planned include workshops at PIMS in April and May and CRM in May and August.

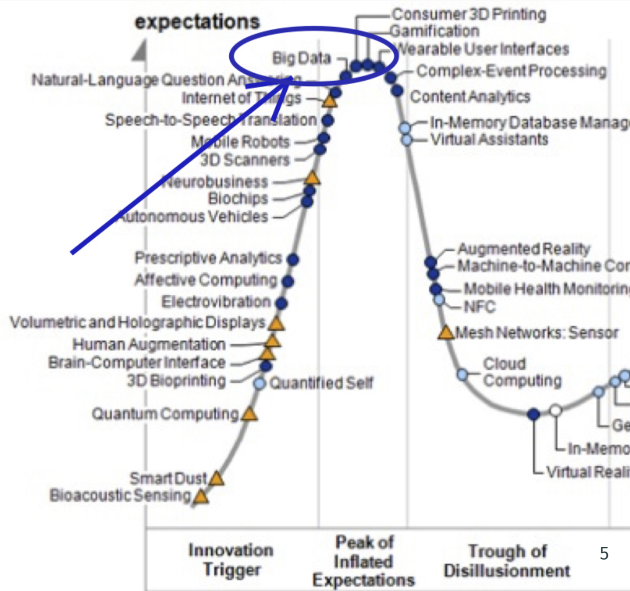
ORGANIZING COMMITTEE

Yoshua Bengio (Montréal)
Hugh Chipman (Acadia)
Sallie Keller (Virginia Tech)
Lisa Lix (Manitoba)
Richard Lockhart (Simon Fraser)
Nancy Reid (Toronto)
Ruslan Salakhutdinov (Toronto)

INTERNATIONAL ADVISORY COMMITTEE

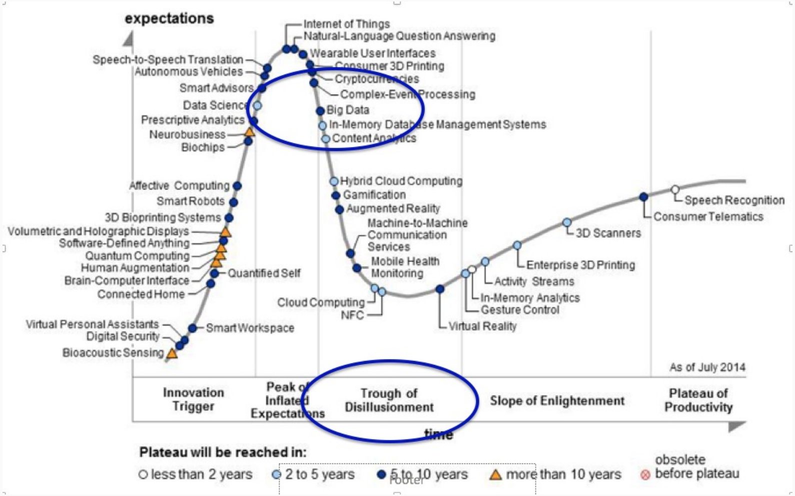
Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehelata Huzurbazar (Wyoming)
Nicolai Meinshausen (ETH Zurich)
Dale Schuurmans (Alberta)
Robert Tibshirani (Stanford)
Bin Yu (UC Berkeley)

For more information, allied activities off-site, and registration, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/bigdata



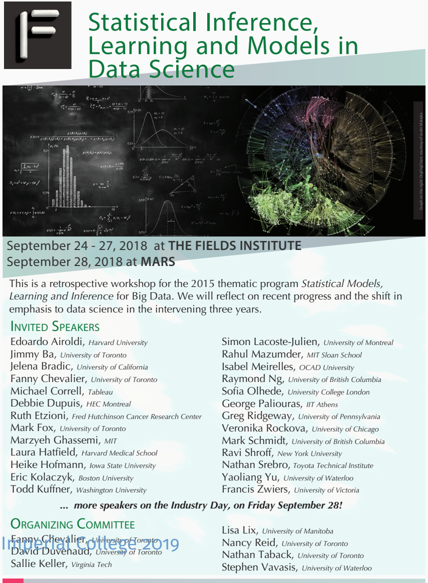
Big Data

2015



Data Science

2018



Statistical Inference, Learning and Models in Data Science

September 24 - 27, 2018 at **THE FIELDS INSTITUTE**
September 28, 2018 at **MARS**

This is a retrospective workshop for the 2015 thematic program *Statistical Models, Learning and Inference* for Big Data. We will reflect on recent progress and the shift in emphasis to data science in the intervening three years.

Invited Speakers

Edoardo Airoldi, *Harvard University*
Jimmy Ba, *University of Toronto*
Jelena Bradic, *University of California*
Fanny Chevalier, *University of Toronto*
Michael Correll, *Tableau*
Debbie Dupuis, *HEC Montreal*
Ruth Etzioni, *Fred Hutchinson Cancer Research Center*
Mark Fox, *University of Toronto*
Marzyeh Ghassemi, *MIT*
Laura Hatfield, *Harvard Medical School*
Heike Hofmann, *Iowa State University*
Eric Kolaczyk, *Boston University*
Todd Kuffner, *Washington University*

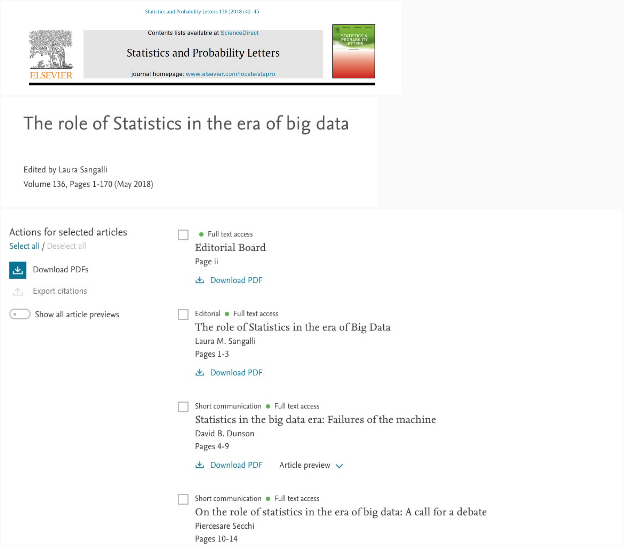
Simon Lacoste-Julien, *University of Montreal*
Rahul Mazumder, *MIT Sloan School*
Isabel Meirelles, *OCAD University*
Raymond Ng, *University of British Columbia*
Sofia Olhede, *University College London*
George Paliouras, *IIT Athens*
Greg Ridgeway, *University of Pennsylvania*
Veronika Rockova, *University of Chicago*
Mark Schmidt, *University of British Columbia*
Ravi Shroff, *New York University*
Nathan Srebro, *Toyota Technical Institute*
Yaoliang Yu, *University of Waterloo*
Francis Zwiers, *University of Victoria*

... more speakers on the Industry Day, on Friday September 28!

Organizing Committee

Fanny Chevalier, *University of Toronto*
David Duvenaud, *University of Toronto*
Sallie Keller, *Virginia Tech*

Lisa Lix, *University of Manitoba*
Nancy Reid, *University of Toronto*
Nathan Taback, *University of Toronto*
Stephen Vavasis, *University of Waterloo*



Statistics and Probability Letters 136 (2018) 42–45

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The role of Statistics in the era of big data

Edited by Laura Sangalli
Volume 136, Pages 1-170 (May 2018)

Actions for selected articles
Select all / Deselect all

Download PDFs
Export citations
Show all article previews

□ ● Full text access
Editorial Board
Page ii
Download PDF

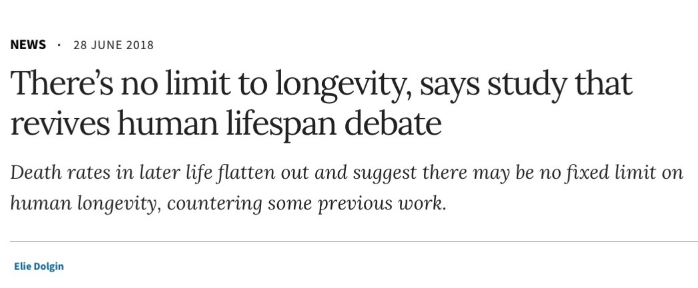
□ Editorial ● Full text access
The role of Statistics in the era of Big Data
Laura M. Sangalli
Pages 1-3
Download PDF

□ Short communication ● Full text access
Statistics in the big data era: Failures of the machine
David B. Dunson
Pages 4-9
Download PDF   Article preview

□ Short communication ● Full text access
On the role of statistics in the era of big data: A call for a debate
Piercesare Secchi
Pages 10-14

Examples: Statistics in the news

Example 1: human longevity



NEWS · 28 JUNE 2018

There’s no limit to longevity, says study that revives human lifespan debate

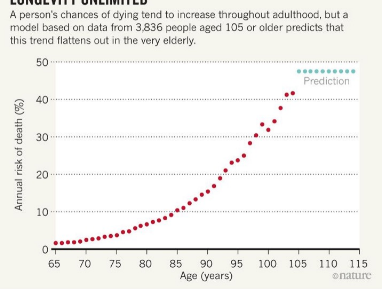
Death rates in later life flatten out and suggest there may be no fixed limit on human longevity, countering some previous work.

Elie Dolgin

Nature News
June 28 2018

“the study included fewer than 100 people who lived to 110 or beyond”
“even small inaccuracies in the Italian longevity records could lead to a spurious conclusion”

E. Dolgin, Nature

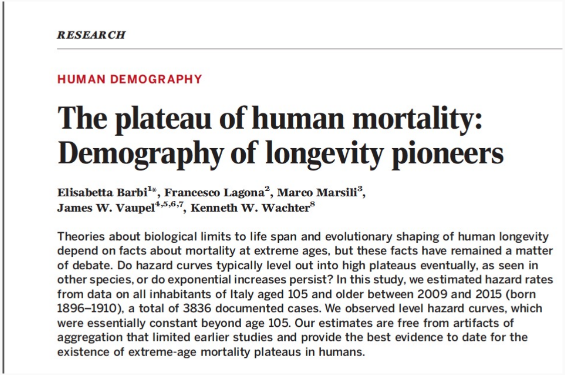


LONGEVITY UNLIMITED

A person’s chances of dying tend to increase throughout adulthood, but a model based on data from 3,836 people aged 105 or older predicts that this trend flattens out in the very elderly.

… is there a limit to human longevity?

Example 1



RESEARCH

HUMAN DEMOGRAPHY

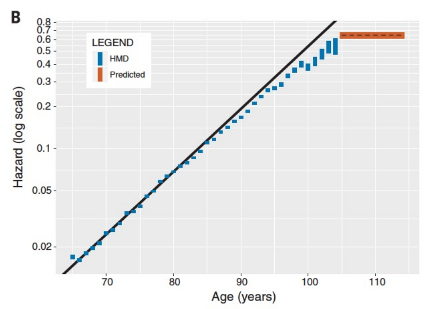
The plateau of human mortality: Demography of longevity pioneers

Elisabetta Barbi[1*], Francesco Lagona[2], Marco Marsili[3], James W. Vaupel[4,5,6,7], Kenneth W. Wachter[8]

Theories about biological limits to life span and evolutionary shaping of human longevity depend on facts about mortality at extreme ages, but these facts have remained a matter of debate. Do hazard curves typically level out into high plateaus eventually, as seen in other species, or do exponential increases persist? In this study, we estimated hazard rates from data on all inhabitants of Italy aged 105 and older between 2009 and 2015 (born 1896–1910), a total of 3836 documented cases. We observed level hazard curves, which were essentially constant beyond age 105. Our estimates are free from artifacts of aggregation that limited earlier studies and provide the best evidence to date for the existence of extreme-age mortality plateaus in humans.

Science

June 29 2018



“We observed level hazard curves, which were essentially constant beyond age 105”

“… provide the best evidence to date for the existence of extreme-age mortality plateaus”

… is there a limit to human longevity?

Example 1

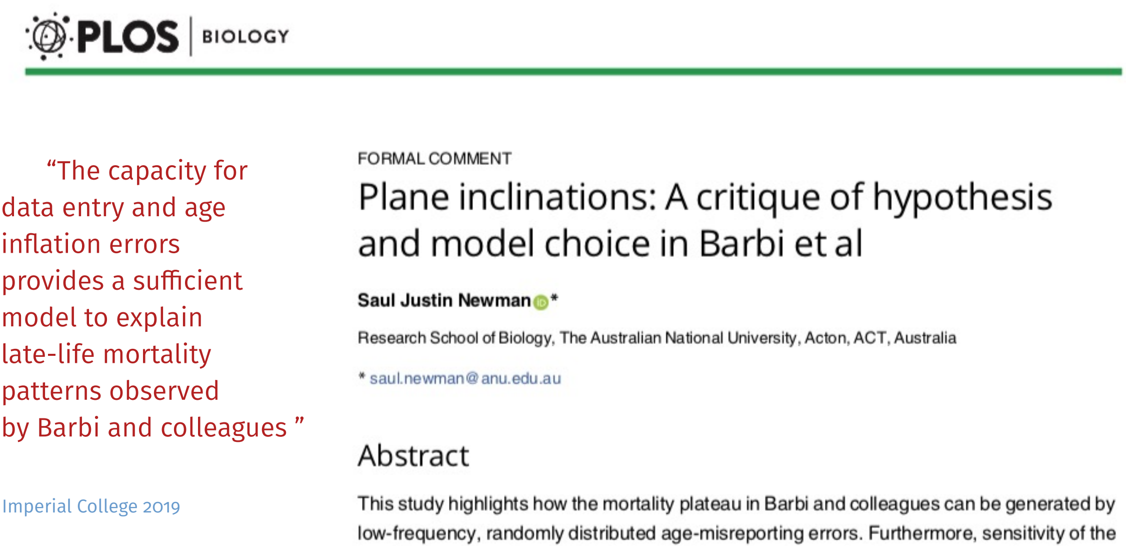
“This study is unlikely to be the last word on the age-limit dispute, says Haim Cohen, a molecular biologist at Bar-Ilan University in Ramut-Gan Israel ‘I’m sure that the debate is going to continue’.”

Dolgin, Nature, June 2018

… is there a limit to human longevity?

Example 1

“The capacity for data entry and age inflation errors provides a sufficient model to explain late-life mortality patterns observed by Barbi and colleagues ”



PLOS | BIOLOGY

FORMAL COMMENT

Plane inclinations: A critique of hypothesis and model choice in Barbi et al

Saul Justin Newman*

Research School of Biology, The Australian National University, Acton, ACT, Australia

* saul.newman@anu.edu.au

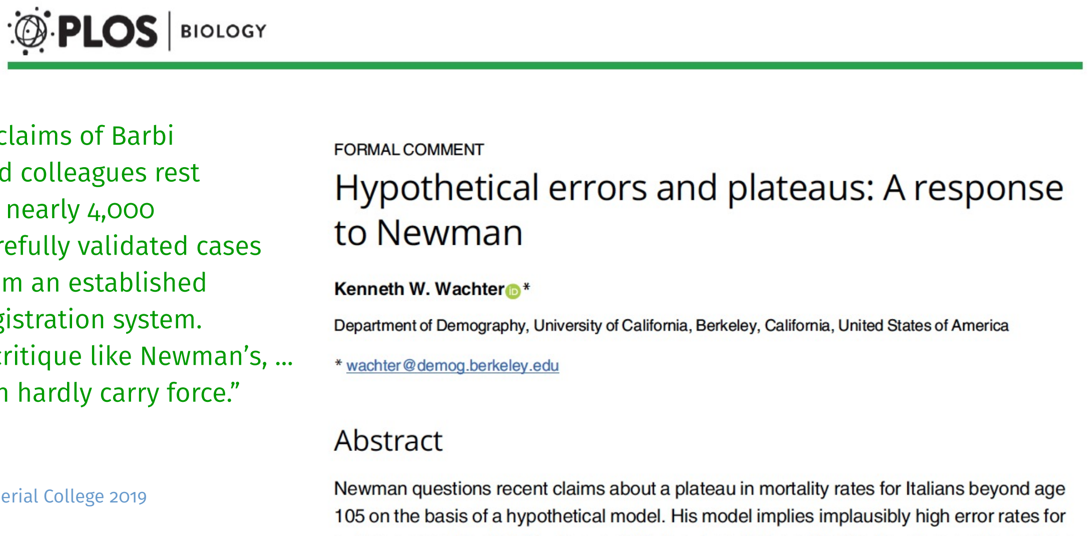
Abstract

This study highlights how the mortality plateau in Barbi and colleagues can be generated by low-frequency, randomly distributed age-misreporting errors. Furthermore, sensitivity of the

… is there a limit to human longevity?

Example 1

"… claims of Barbi and colleagues rest on nearly 4,000 carefully validated cases from an established registration system. A critique like Newman's, … can hardly carry force."



PLOS | BIOLOGY

FORMAL COMMENT

Hypothetical errors and plateaus: A response to Newman

Kenneth W. Wachter *

Department of Demography, University of California, Berkeley, California, United States of America

* wachter@demog.berkeley.edu

Abstract

Newman questions recent claims about a plateau in mortality rates for Italians beyond age 105 on the basis of a hypothetical model. His model implies implausibly high error rates for

… is there a limit to human longevity?

Newman 2018

- claims that age-misreporting can generate spurious late life plateaus
- Barbi et al (2018) fit a parametric model and used likelihood ratio test to compare to a constant hazard

  for age > 105
- Newman argued that a modelling choice they made influenced their results
- "of the 861 … combinations tested, the model selected by Barbi et al generated the single largest late-life mortality plateau"

- statistics: Gompertz model, LRT, power analysis

  $h(x) = ae^{bx}e^{\beta_1 C + \beta_2 M}$
- data science: 861 such fits, plus simulated errors

  with probabilities ranging from 10^{-3} to 10^{-6}
- domain: all inhabitants of Italy aged $\geq$ 105 years 2009–2015 (3836 cases) + Human Mortality Database

… is there a limit to human longevity?

Barbi et al 2018

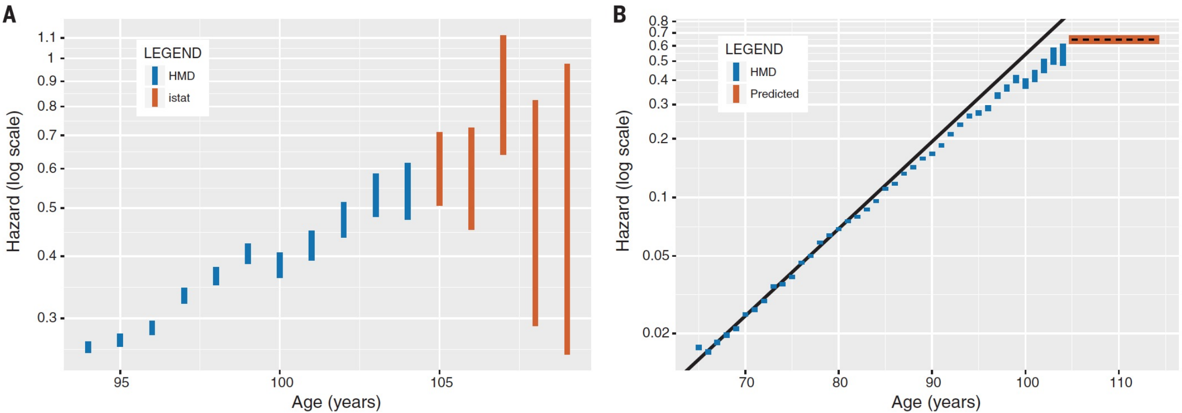
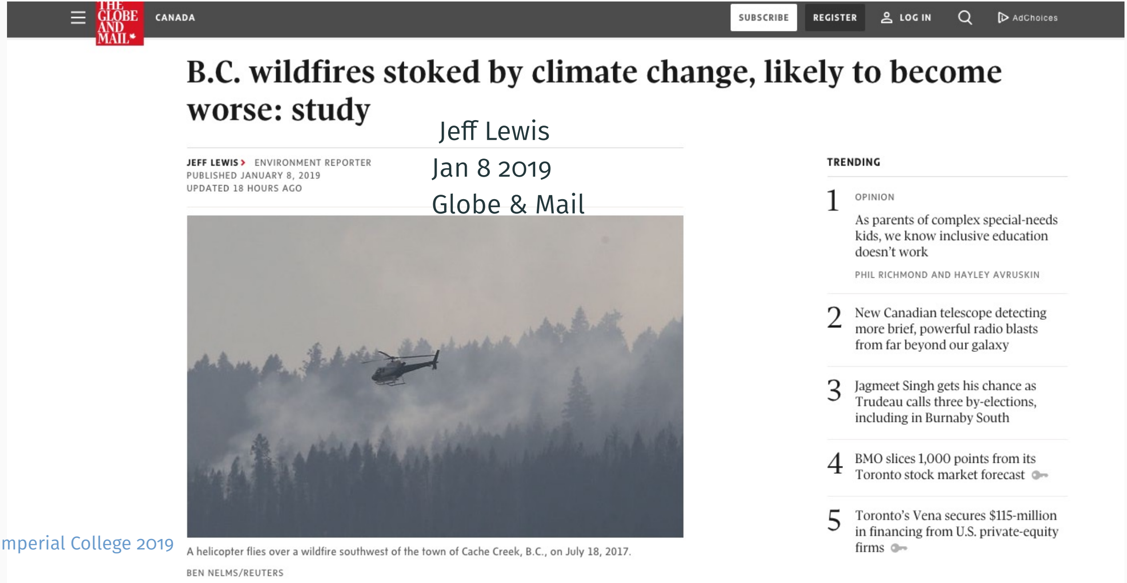


Fig. 1. Yearly hazards on a logarithmic scale for the cohort of Italian women born in 1904. Confidence intervals were derived from Human Mortality Database (HMD) data for ages up to 105 and from ISTAT data beyond age 105. (**A**) Closeup with 95% confidence intervals based solely on single-cohort data. (**B**) Broad view with estimated plateau beyond age 105 (black dashed line) and 95% confidence bands (orange) predicted from the model parameters based on the full ISTAT database, along with a straight-line prediction (black) from fitting a Gompertz model to ages 65 to 80.

Barbi *et al.*, *Science* **360**, 1459–1461 (2018)   29 June 2018   **1 of 3**

Example 2: wildfire



B.C. wildfires stoked by climate change, likely to become worse: study

JEFF LEWIS › ENVIRONMENT REPORTER
PUBLISHED JANUARY 8, 2019
UPDATED 18 HOURS AGO

A helicopter flies over a wildfire southwest of the town of Cache Creek, B.C., on July 18, 2017.

BEN NELMS/REUTERS

TRENDING

1 OPINION
As parents of complex special-needs kids, we know inclusive education doesn't work
PHIL RICHMOND AND HAYLEY AVRUSKIN

2 New Canadian telescope detecting more brief, powerful radio blasts from far beyond our galaxy

3 Jagmeet Singh gets his chance as Trudeau calls three by-elections, including in Burnaby South

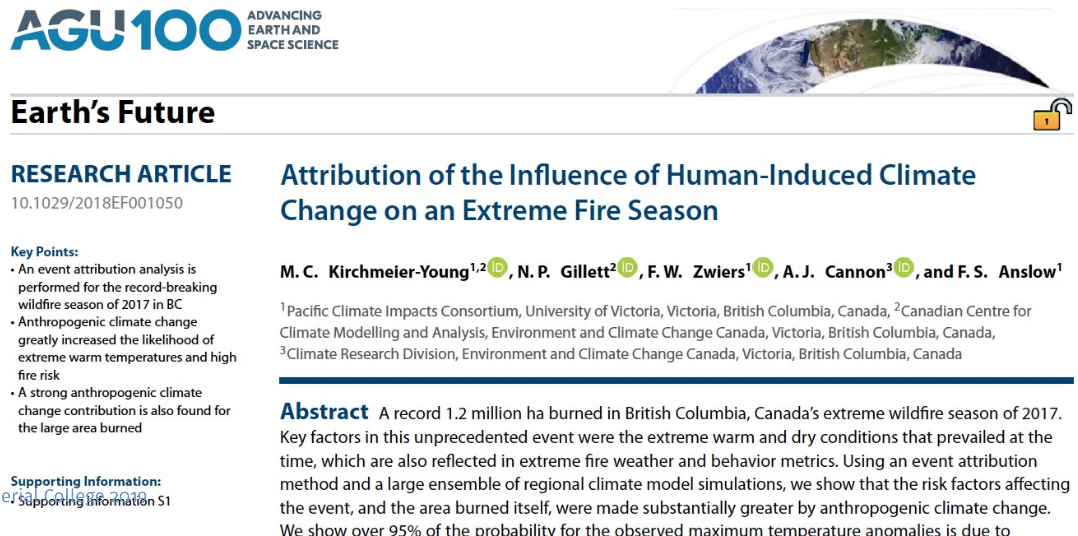
4 BMO slices 1,000 points from its Toronto stock market forecast

5 Toronto's Vena secures $115-million in financing from U.S. private-equity firms

Jeff Lewis
Jan 8 2019
Globe & Mail

Event attribution

Example 2



AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth's Future

RESEARCH ARTICLE

10.1029/2018EF001050

Attribution of the Influence of Human-Induced Climate Change on an Extreme Fire Season

M. C. Kirchmeier-Young[1,2], N. P. Gillett[2], F. W. Zwiers[1], A. J. Cannon[3], and F. S. Anslow[1]

[1] Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada, [2] Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada, [3] Climate Research Division, Environment and Climate Change Canada, Victoria, British Columbia, Canada

Key Points:

- An event attribution analysis is performed for the record-breaking wildfire season of 2017 in BC
- Anthropogenic climate change greatly increased the likelihood of extreme warm temperatures and high fire risk
- A strong anthropogenic climate change contribution is also found for the large area burned

Supporting Information:

- Supporting Information S1

Abstract A record 1.2 million ha burned in British Columbia, Canada's extreme wildfire season of 2017. Key factors in this unprecedented event were the extreme warm and dry conditions that prevailed at the time, which are also reflected in extreme fire weather and behavior metrics. Using an event attribution method and a large ensemble of regional climate model simulations, we show that the risk factors affecting the event, and the area burned itself, were made substantially greater by anthropogenic climate change. We show over 95% of the probability for the observed maximum temperature anomalies is due to

… event attribution

Kirchmeier-Young et al 2018

- “anthropogenic climate change increased the area burned by a factor of 7 - 11”
- “We use a large ensemble of CanRCM4 … consisting of 50 realizations on a 50-km grid. Each realization is driven by a member of the CanESM2 … large ensemble … We utilize data from 1961 to 2020.”
- “A data set of gridded maximum (and minimum) temperature and precipitation anomalies was created by interpolating monthly values calculated from surface station observations relative to a 30-year climatology. Observational data was acquired from numerous sources and interpolated using a thin plate spline methodology.”
- “… values for each year and large ensemble realization were pooled together for two time periods: 1961-1970 and 2011-2020, resulting in 500 values for each decade (10 years x 50 realizations). ”

… event attribution

Kirchmeier-Young et al 2018

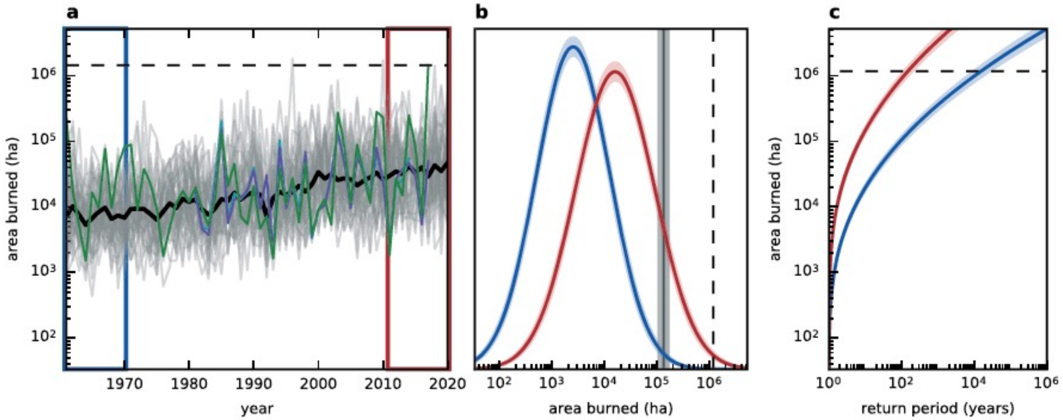


Figure 5. Time series (a, log scale) of regression-predicted annual burned area in the BC Southern Cordillera for bias-corrected CanRCM4 realizations (gray) and ensemble mean (bold), reanalysis (turquoise/purple), and observations (green). The dashed line marks the observed 2017 value. Probability distributions (b) for area burned amounts (log scale) from decades outlined in corresponding colors in (a). The gray bar indicates the area burned amount in the distribution with reduced anthropogenic influence (blue) of a corresponding percentile to the 2017 amount (dashed line) in the

… event attribution

Kirchmeier-Young et al 2018

- complex computer simulation of global climate

mathematics
numerical analysis

- creation of regional climate scenarios

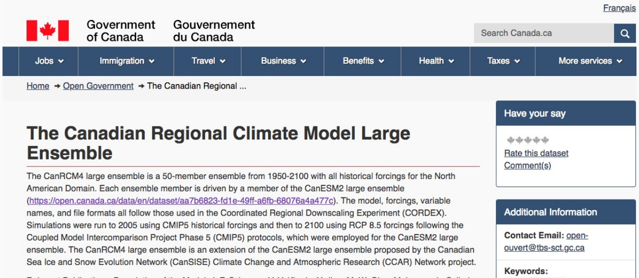
mathematics, statistics

- combined with available observational data

statistics, data science

- modelled with regression and kernel density estimation

statistics
mathematics



The Canadian Regional Climate Model Large Ensemble

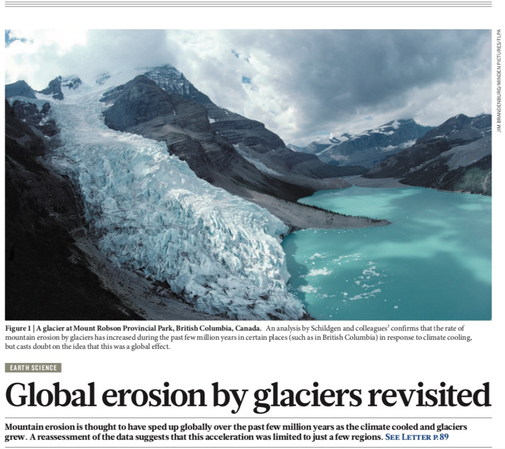
The CanRCM4 large ensemble is a 50-member ensemble from 1950-2100 with all historical forcings for the North American Domain. Each ensemble member is driven by a member of the CanESM2 large ensemble (https://open.canada.ca/data/en/dataset/aa7b6823-fd1e-49ff-a6fb-68076a4a477c). The model, forcings, variable names, and file formats all follow those used in the Coordinated Regional Downscaling Experiment (CORDEX). Simulations were run to 2005 using CMIP5 historical forcings and then to 2100 using RCP 8.5 forcings following the Coupled Model Intercomparison Project Phase 5 (CMIP5) protocols, which were employed for the CanESM2 large ensemble. The CanRCM4 large ensemble is an extension of the CanESM2 large ensemble proposed by the Canadian Sea Ice and Snow Evolution Network (CanSISE) Climate Change and Atmospheric Research (CCAR) Network project.

Additional Information

Contact Email: open-ouvert@tbs-sct.gc.ca

Keywords:

Example 3: Glaciers



JIM BRANDENBURG/MINDEN PICTURES/FLPA

Figure 1 | A glacier at Mount Robson Provincial Park, British Columbia, Canada. An analysis by Schildgen and colleagues[3] confirms that the rate of mountain erosion by glaciers has increased during the past few million years in certain places (such as in British Columbia) in response to climate cooling, but casts doubt on the idea that this was a global effect.

EARTH SCIENCE

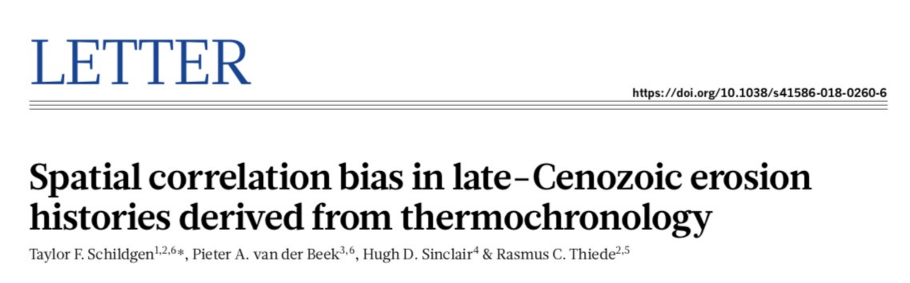
Global erosion by glaciers revisited

Mountain erosion is thought to have sped up globally over the past few million years as the climate cooled and glaciers grew. A reassessment of the data suggests that this acceleration was limited to just a few regions. SEE LETTER P. 89

"Mountain erosion is thought to have sped up globally … A reassessment of the data suggests that this acceleration was limited…"

Kirby, Nature July 2018

… Global erosion by glaciers revisited



LETTER

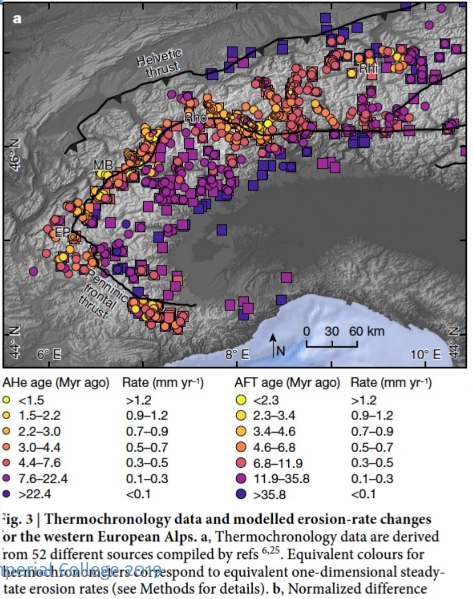
https://doi.org/10.1038/s41586-018-0260-6

Spatial correlation bias in late-Cenozoic erosion histories derived from thermochronology

Taylor F. Schildgen[1,2,6]*, Pieter A. van der Beek[3,6], Hugh D. Sinclair[4] & Rasmus C. Thiede[2,5]

… Global erosion by glaciers revisited

Schildgen et al 2018



AHe age (Myr ago)	Rate (mm yr^{-1})	AFT age (Myr ago)	Rate (mm yr^{-1})
● <1.5	>1.2	● <2.3	>1.2
● 1.5–2.2	0.9–1.2	● 2.3–3.4	0.9–1.2
● 2.2–3.0	0.7–0.9	● 3.4–4.6	0.7–0.9
● 3.0–4.4	0.5–0.7	● 4.6–6.8	0.5–0.7
● 4.4–7.6	0.3–0.5	● 6.8–11.9	0.3–0.5
● 7.6–22.4	0.1–0.3	● 11.9–35.8	0.1–0.3
● >22.4	<0.1	● >35.8	<0.1

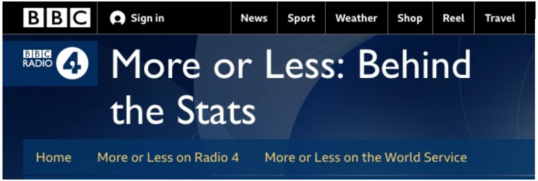
ig. 3 | Thermochronology data and modelled erosion-rate changes or the western European Alps. a, Thermochronology data are derived rom 52 different sources compiled by refs [6,25]. Equivalent colours for thermochronometers correspond to equivalent one-dimensional steady-tate erosion rates (see Methods for details). **b,** Normalized difference

> "… as we use increasingly sophisticated analyses of **'big data'** to gain insight into global trends in geology, we must not lose sight of the physical processes that operate locally"

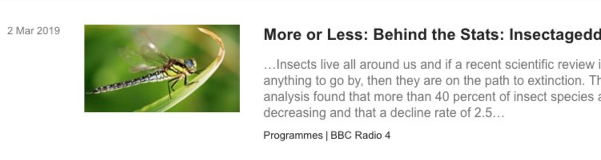
Kirby, Nature July 2018

Example 4: Insectageddon

More or Less, BBC



More or Less: Behind the Stats


2 Mar 2019

More or Less: Behind the Stats: Insectagedd

…Insects live all around us and if a recent scientific review i anything to go by, then they are on the path to extinction. Th analysis found that more than 40 percent of insect species a decreasing and that a decline rate of 2.5…

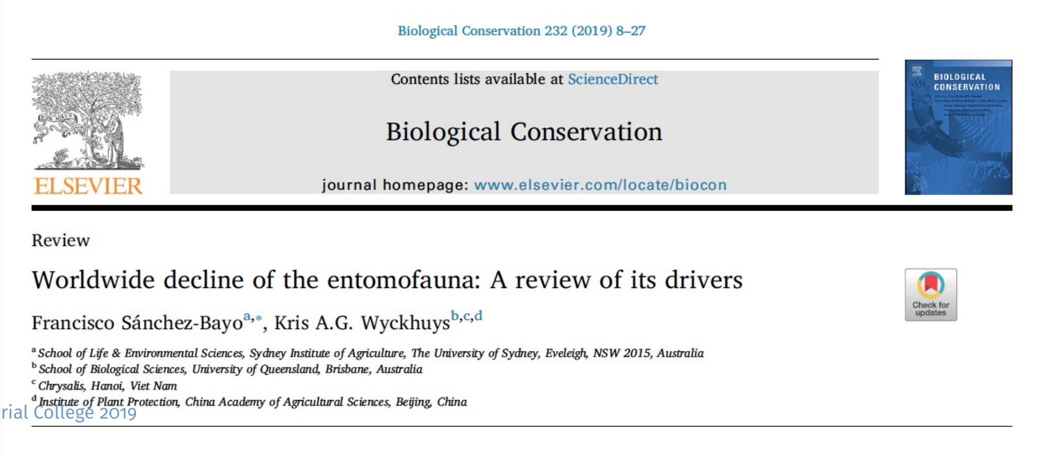
Programmes | BBC Radio 4

Plummeting insect numbers 'threaten collapse of nature'

… insects

Example 4

“if a recent scientific review is anything to go by, more than more than 40 percent of insect species are decreasing and that a decline rate of 2.5 percent per year suggests they could disappear in one hundred years”



Biological Conservation 232 (2019) 8–27

ELSEVIER

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon

BIOLOGICAL CONSERVATION

Review

Worldwide decline of the entomofauna: A review of its drivers

Check for updates

Francisco Sánchez-Bayo[a,*], Kris A.G. Wyckhuys[b,c,d]

[a] *School of Life & Environmental Sciences, Sydney Institute of Agriculture, The University of Sydney, Eveleigh, NSW 2015, Australia*
[b] *School of Biological Sciences, University of Queensland, Brisbane, Australia*
[c] *Chrysalis, Hanoi, Viet Nam*
[d] *Institute of Plant Protection, China Academy of Agricultural Sciences, Beijing, China*

… insects

- “we aimed at compiling all long-term insect surveys… past 40 years… peer-reviewed literature databases”
- “search on the online Web of Science database … 653 publications”
- “Reports that … were excluded”
- “Only surveys that reported changes over time were considered”
- “this review covers 73 reports on entomofauna declines and examines their likely causes”
- (on biomass) “we have 3 studies that we added to the survey”

$$653 \longrightarrow 73 \longrightarrow 0 \longrightarrow 3 \xrightarrow{\hspace{8cm}}$$

Plummeting insect numbers 'threaten collapse of nature'

… insects

Example 4

More or Less: Behind the Stats

2 Mar 2019

More or Less: Behind the Stats: Insectageddon

…Insects live all around us and if a recent scientific review is anything to go by, then they are on the path to extinction. The analysis found that more than 40 percent of insect species are decreasing and that a decline rate of 2.5…

Programmes | BBC Radio 4

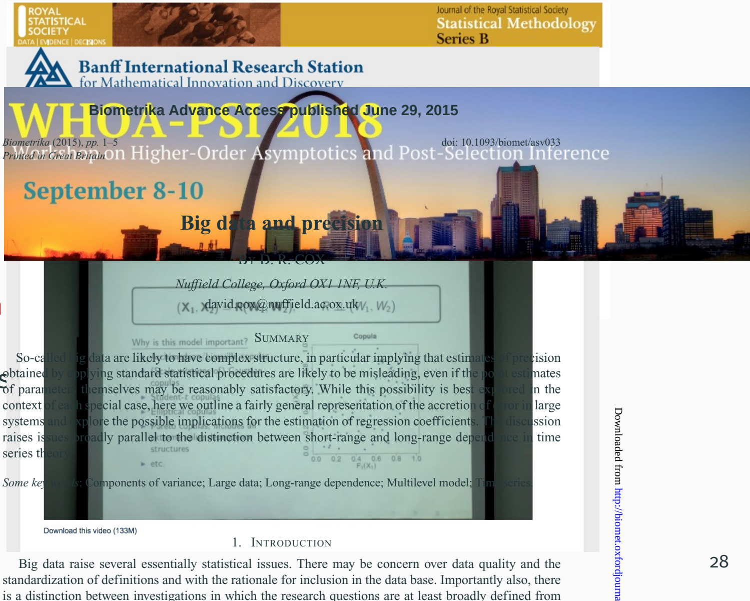
- entomologist: “we simply don’t have the data”
- statistician: “they simply don’t have the proof”
- author: “even if we don’t have enough data to prove it statistically or whatever, we know that this is happening.

  So it’s better to do it now, than 10 years later when we have more serious problems”

Statistical theory

Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- high-dimensional inference
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- foundations
- ...


ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistical Methodology
Series B
Banff International Research Station
for Mathematical Innovation and Discovery
WHOA-PSI 2018
Workshop on Higher-Order Asymptotics and Post-Selection Inference
September 8-10
(X1, X2) ~ K(W1, W2)
Why is this model important?
Copula
copulas
Student-t copulas
Elliptical copulas
Pareto copulas
structures
etc.
0.0 0.2 0.4 0.6 0.8 1.0
F1(X1)
Download this video (133M)

Biometrika Advance Access published June 29, 2015

Biometrika (2015), pp. 1–5
Printed in Great Britain
doi: 10.1093/biomet/asv033

Big data and precision

BY D. R. COX
Nuffield College, Oxford OX1 1NF, U.K.
david.cox@nuffield.ox.ac.uk

SUMMARY

So-called big data are likely to have complex structure, in particular implying that estimates of precision obtained by applying standard statistical procedures are likely to be misleading, even if the point estimates of parameters themselves may be reasonably satisfactory. While this possibility is best explored in the context of each special case, here we outline a fairly general representation of the accretion of error in large systems and explore the possible implications for the estimation of regression coefficients. The discussion raises issues broadly parallel to the distinction between short-range and long-range dependence in time series theory.

Some key words: Components of variance; Large data; Long-range dependence; Multilevel model; Time series.

1. INTRODUCTION

Big data raise several essentially statistical issues. There may be concern over data quality and the standardization of definitions and with the rationale for inclusion in the data base. Importantly also, there is a distinction between investigations in which the research questions are at least broadly defined from

Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- **high-dimensional inference**
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- **foundations**
- ...

High-dimensional asymptotic theory

- $f(y;\theta)$, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$

$y_1, \ldots, y_n$ independent

- classical: p fixed, $n \to \infty$

$\sqrt{n}(\hat{\theta} - \theta)V^{-1/2} \xrightarrow{d} N_p(0, I)$

- semi-classical: $p_n/n \to 0$, or $p_n^{3/2}/n \to 0$

Huber, Portnoy; Sartori, Lunardon, ...

- moderate dimension $p_n/n \to \kappa \in (0,1)$

Candes, Lei/Bickel/El Karoui, ...

- high dimension $p_n \sim n^{\alpha}$

$p > n$

- ultra-high dimension $p_n \sim e^n$

Moderate dimension

Lei, El-Karoui, Bickel 18

- $\hat{\beta} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i^{\mathrm{T}} \beta)$ *M*-estimator
- coordinate-wise asymptotic normality

$$\max_j d_{TV} \left\{ \mathcal{L} \left(\frac{\hat{\beta}_j - \mathrm{E}(\hat{\beta}_j)}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} \right), N(0,1) \right\} = o(1)$$

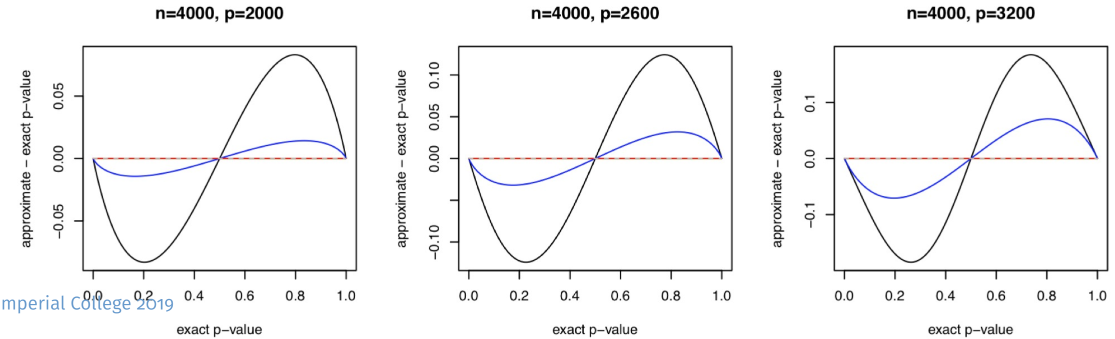
- “For instance for least-squares, standard degrees of freedom adjustments effectively take care of many dimensionality-related problems”

- in least squares, ‘standard degrees of freedom adjustments’ can be derived using higher order asymptotics for p fixed
- e.g. $n = 50, p = 30$ or $n = 500, p = 300$ moderate or classical?

… moderate dimension

- normal theory linear regression $y = X\beta + \sigma\epsilon,\ \epsilon \sim N(0,1)$
- exact test available based on t-statistic $t = (\hat{\beta}_j - \beta_j)/v_{jj}$ least squares = mle
- all likelihood quantities are functions of t, n and p
- modified log-likelihood root, derived from higher order asymptotics depends only on $t, n - p$ r^* depends on $t, n, n-p$

Plot of differences between approximate and exact p-values for one-sided alternative against the true p-value:



… moderate dimension

Sur, Candes, '18

logistic regression

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta, \quad y_i \sim \mathrm{Bernoulli}(p_i)$$

- if the MLE exists, then

$p/n \to \kappa \in (0,1)$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta_j) \longrightarrow 0; \qquad \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta_j)^2 \longrightarrow \sigma_*^2$$

- Likelihood Ratio Test for $H:\beta_j = 0$ has scaled χ^2

$$w(\beta_j) = 2\{\ell(\hat{\beta}) - \ell(\tilde{\beta}_{(j)})\} \xrightarrow{d} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2$$

- $(a_*, \sigma_*, \lambda_*)$ characterized as the solution of three equations
- e.g. $n = 50, p = 30$ or $n = 500, p = 300$

moderate or classical?

… moderate dimension

Sur, Candes, '18

logistic regression

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}} \beta, \quad y_i \sim \text{Bernoulli}(p_i)$$

- if the MLE exists, then

$p/n \to \kappa \in (0,1)$

$$\frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j - a_* \beta) \longrightarrow 0; \qquad \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j - a_* \beta)^2 \longrightarrow \sigma_*^2$$

- in logistic regression, change the score equation a little maximum likelihood estimate always exists

Firth 93; Kosmidis/Firth 09

- usual limit theory seems to be fine with large p

Sartori; Lunardon 18

Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- **high-dimensional inference**
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- **foundations**
- ...

Foundations

- how to get from data to conclusions
- with generalizable strategies
- what principles do we use to develop these strategies
- how are these strategies to be evaluated — efficiency, precision
- probability to describe physical haphazard variability — frequentist

  subject to empirical validation
- probability to describe the uncertainty of knowledge — Bayesian

  degree of belief

samsi
NSF Duke NCSU UNC
JOBS AT SAMSI
FORMS & RESOURCES
WORKING GROUPS
CO
PROGRAMS & ACTIVITIES
PARTICIPATING
PEOPLE
NEWS A
Home > Programs and Activities > Year-long Research Programs > Model Uncertainty: Mathematical and Statistical (MUMS) > Bayesian, Fiducial, and Frequentist (B
APPLY

Bayesian, Fiducial, and Frequentist (BFF) Conferences

***** Deadline for applications for this workshop is March 20, 2019 *****

Applications received after March 20th are subject to availability.

Location

This workshop will be held at Penn Pavilion on the campus of Duke University.

Role of Probability

R & Cox 2013

- probability to describe physical haphazard variability — frequentist
    - probabilities represent features of the "real" world in somewhat idealized form
    - subject to empirical test and improvement
- probability to describe the uncertainty of knowledge — Bayesian
    - measures rational or "impersonal" degree of belief,

      or — Jeffreys, 1939,1961
    - measures a particular person's degree of belief

      F.P. Ramsey, 1926
    - linked to personal decision making

… role of probability

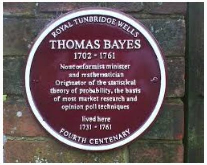
R & Cox 2013

- confidence intervals or *p*-values refer to empirical probabilities

“[7 – 11]”

- inference is assessed by behaviour of the procedure under hypothetical repetition

- the Bayesian approach to inference describes uncertainty of knowledge
- this can be interpreted empirically by appeal to a notion of calibration



Statistics and data science

Statistical science

Cox & Donnelly 2011

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

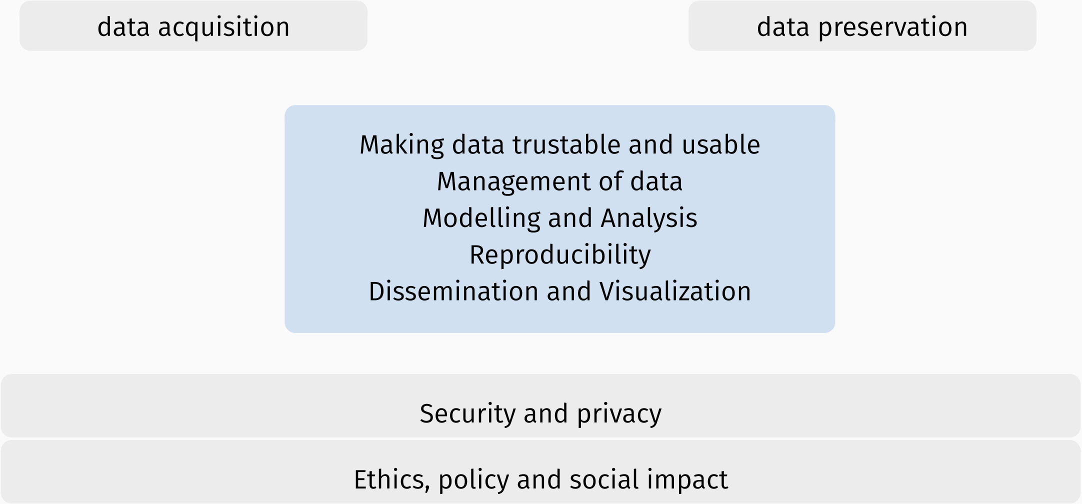
- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately
- visualization strategies, conveyance of uncertainties

but not pessimistically

Data science

“workflow”



data acquisition

data preservation

Making data trustable and usable
Management of data
Modelling and Analysis
Reproducibility
Dissemination and Visualization

Security and privacy

Ethics, policy and social impact

… data science workflow

Making data trustable and usable
Management of data

provenance, sampling, cleaning, digitizing
size, speed, accessibility

Modelling and Analysis
Reproducibility
Dissemination and Visualization

interpretable vs predictive methods
accessibility and impact
data, code, output

mathematics statistics computer science domain expertise

Security and privacy

disclosure limitation, anonymization, encryption

Ethics, policy and social impact

fairness and transparency

In praise of small data

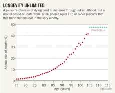
”you don’t have as much data as you thought”

- correlation/dependence/heterogeneity/multiple scales

Cox 15, Meng

- rare events

CERN, extremes



LONGEVITY UNLIMITED

A person's chances of dying tend to increase throughout adulthood, but a model based on data from 3,836 people aged 105 or older predicts that this trend flattens out in the very elderly.

- subgroup analyses/‘data slices’

wildfires, social media

Plummeting insect numbers 'threaten collapse of nature'

The rate of insect extinction is eight times faster than that of mammals, birds and reptiles. Photograph: Courtesy of Entomologischer Verein Krefeld

- complex models/many parameters/high-dimensional inference

sparsity, new asymptotics

- ...

Thank you!

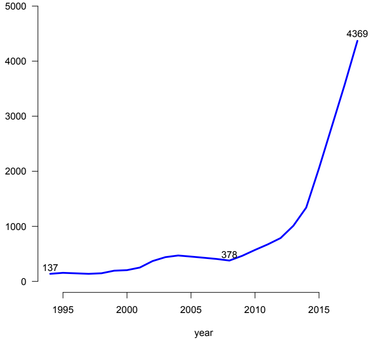
Imperial College London
Royal Albert Hall
Kensington Gore
Museums
South Kensington

Collaborations

CANSSI
INCASS

- Canadian Statistical Sciences Institute
- launched in 2012
- funded 2014–2021 by Natural Sciences and Engineering Research Council
- national scope, virtual institute
- **Collaborative Research Teams**
  multidisciplinary, multi-institution, statistical leadership, scientific engagement

… collaborations



Statistical Sciences

UNIVERSITY OF TORONTO

- statistical genetics
- spatial modelling
- machine learning (with CS)
- visualization (with CS)
- demography (with Sociology)
- astrostatistics (with A and A)
- cognitive neuroscience (with Pyschology)
- data science (with iSchool)
- financial insurance
- actuarial science
- teaching stream

In praise of small data

counterpoint