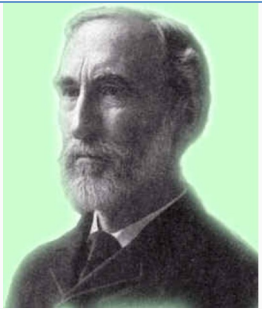
In Praise of Small Data

Statistical Science and Data Science

Nancy Reid
University of Toronto

January 15 2020



Outline

The Gibbs sampler

The Gibbs distribution

- probability distribution, or measure, for complex system

  also called Boltzmann distribution; Stigler

- density function

$$f(x) = \frac{1}{Z} \exp\{-\beta E(x)\}$$

  $E(\cdot)$ is the energy function

  x is a state of the system

  β is called (inverse) temperature

- partition function

$$Z = \Sigma_x \exp\{-\beta E(x)\}$$

  $\int_x \exp\{-\beta E(x)\} dx$

- used in statistical physics, quantum mechanics, probability theory, statistical modelling, machine learning, ...

  $E(x; \theta), Z = Z(\beta, \theta)$

The Gibbs Sampler

Geman & Geman, 1984

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-6, NO. 6, NOVEMBER 1984 721

Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images

STUART GEMAN AND DONALD GEMAN

"we introduce a stochastic model for the original image, based on the Gibbs distribution, and a new restoration algorithm, based on stochastic relaxation and annealing"

"the computational problem is overcome ...
with a sampling method that we call the Gibbs Sampler.

... Gibbs Sampler

Geman & Geman, 1984

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-6, NO. 6, NOVEMBER 1984 721

Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images

STUART GEMAN AND DONALD GEMAN

" When Don and I were growing up it was a standard go-to for mother's day, valentine's day, or just an old-fashioned "I'm sorry." We were sitting together writing our paper on the Bayesian approach to image processing and wondering what to call that particular version of stochastic relaxation. Don mentioned the Whitman's Sampler. It was the perfect metaphor."

... Gibbs Sampler

Geman & Geman, 1984

- the Gibbs sampler is one of a wide range of Markov chain Monte Carlo algorithms

- from 1990 onward these revolutionized statistical inference
- replacing difficult integrals with finite sums over computer-generated points

- led to an explosion of applications of Bayesian inference in complex problems

Gelfand & Smith, 1990; Casella & George, 1992

- and to a wealth of interesting mathematical, statistical, and probabilistic questions which themselves lead to new applications

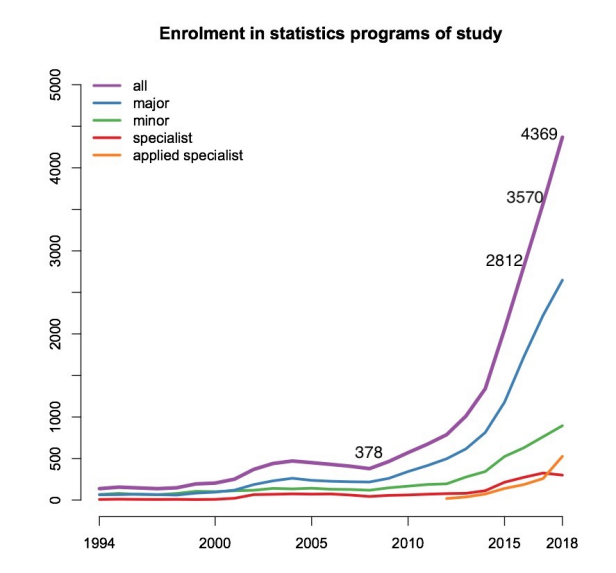
- and to a new generation of computational approaches to statistical science

“What’s new in statistics?”

Enrollments are sky-rocketing

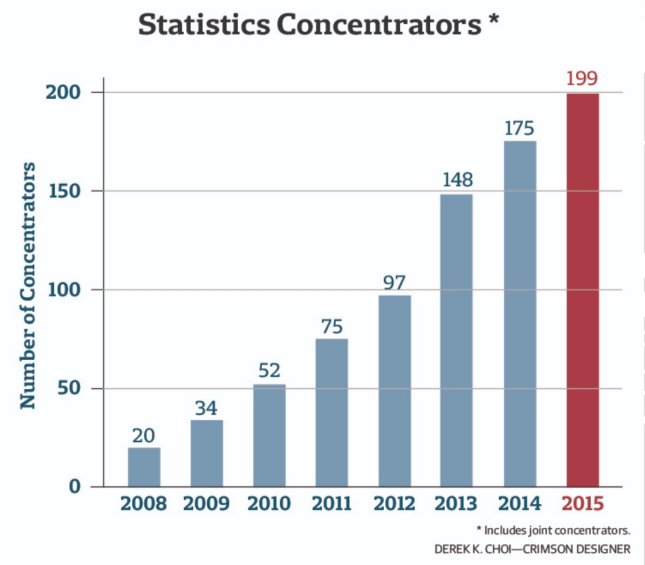
Statistical Sciences

UNIVERSITY OF TORONTO



Enrollments are sky-rocketing

HARVARD
Faculty of Arts and Sciences
DEPARTMENT OF STATISTICS

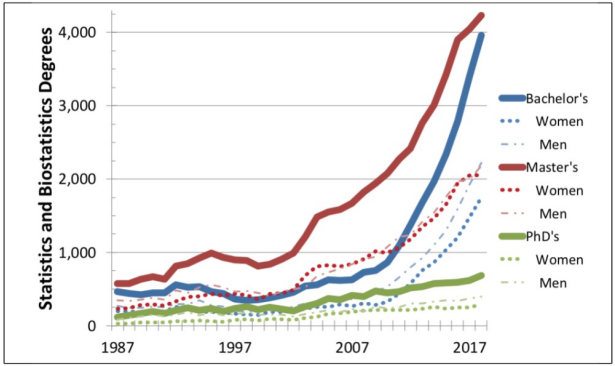


Enrollments are sky-rocketing

Am. Stat. News 2019

(Bio)Statistics Bachelor's Degrees Nearly Quintuple This Decade

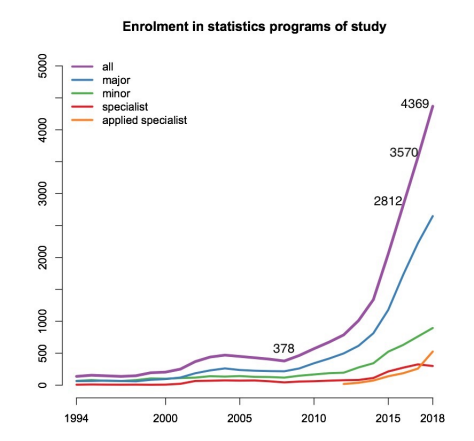
Highlights from 2018 Degree Release



Statistical science is looking outwards

Statistical Sciences

UNIVERSITY OF TORONTO



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

New York Times, August 2009

For Today's Graduate, Just One Word: Statistics

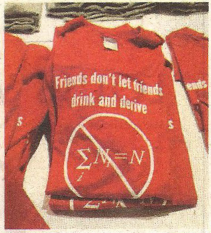
By STEVE LOHR

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.



DANIEL ROSENBAUM FOR THE NEW YORK TIMES

A question: would 9 out of 10 statisticians wear this shirt?

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a by-product of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by 2012, according to a projection by IDC, a research firm.

Yet data is merely the raw material of knowledge. "We're rapidly entering a world where everything can be monitored and measured," said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology's Center for Digital Business. "But the big problem is going to be the ability of humans to use, analyze and make sense of the data."

The new breed of statisticians tackle that problem. They use powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data.

Continued on Page A3

FIELDS

THE FIELDS INSTITUTE

THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR BIG DATA

JANUARY - JUNE, 2015

PROGRAM

JANUARY 12 – 23, 2015

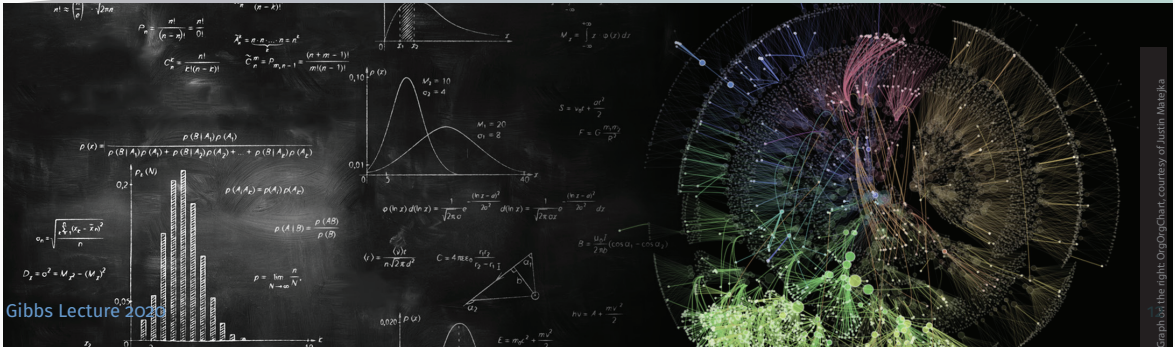
Opening Conference and Boot Camp

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

JANUARY 26 – 30, 2015

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,

Statistical Inference, Learning and Models in Data Science



Graph on the right: OrgOrgChart, courtesy of Justin Matejka

Gibbs Lecture 2020

Big Data 2015

THE FIELDS INSTITUTE

FIELDS

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, Snehalata 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 Fieldscive 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)

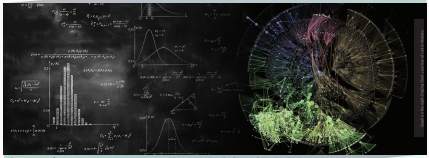
INTERNATIONAL ADVISORY COMMITTEE

Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehalata 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

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*

HDSR
SCIENCE REVIEW

The Alan Turing Institute

BERKELEY
Institute for Data Science

DATA SCIENCE DEGREE PROGRAMS
TOP MASTER'S DEGREES

ta Analytics Bootcamps in Denve
iding a Path to Powerful Skills

Paris-Saclay Center for Data Science
universite PARIS-SACLAY

Institute of Data

LEARN DATA SCIENCE IN YOUR BROWSER

HDSR

HARVARD DATA SCIENCE REVIEW

A Microscopic, Telescopic, and Kaleidoscopic View of Data Science

HOME ABOUT MISSION AND SCOPE ISSUE 1.1

Current Issue · 1.2

FROM THE EDITOR-IN-CHIEF

Xiao-Li Meng

Five Immersive 3D Surroundings of Data Science

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

but not pessimistically

- visualization strategies, conveyance of uncertainties

Data science

"workflow"

data acquisition

data preservation

Making data trustable and usable
Management of data

Modelling and Analysis

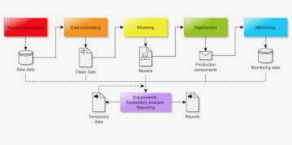
Reproducibility
Dissemination and Visualization

Security and privacy

Ethics, policy and social impact

Data science

“workflow”

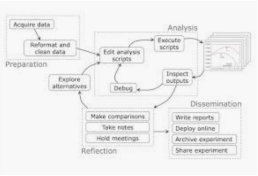


The Data Science Workflow - Towards ...
towardsdatascience.com

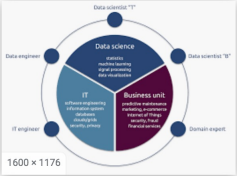
Teaching the Data Science Process
kdnuggets.com

Data Science Workflow - The Process for ...
business-science.io

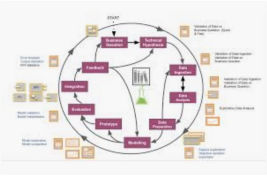
What is Data Science? – D...
dataquest.io



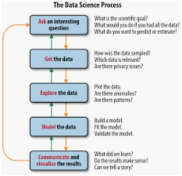
Data Science Workflow: Overview and ...
m-cacm.acm.org



Teaching the data science process ...
towardsdatascience.com



Accelerating Data Science Workflows
bbvadata.com



Development Workflows for Data ...
resources.github.com

https://www.google.com/imgres?imgurl=https%3A%2F%2Fmico.medium

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	IS, CS, Stat
Modelling and Analysis	interpretable vs predictive methods	Math, Stat, CS
Reproducibility Dissemination and Visualization	accessibility and impact data, code, output	IS, DS

mathematics statistics computer science **domain expertise**

Security and privacy	disclosure limitation, anonymization, encryption	CS, Stat
Ethics, policy and social impact	fairness and transparency	SS, Hum, DS

Examples: Statistics in the news

Examples: Statistics in the news

Example 1: Wildfire

Example 1: wildfire

"Toronto's National Newspaper"

THE GLOBE AND MAIL CANADA SUBSCRIBE REGISTER LOG IN AdChoices

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

Jeff Lewis

Jan 8 2019

Globe & Mail

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

Event attribution

Example 1

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth's Future

RESEARCH ARTICLE
10.1029/2018EF001050

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

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

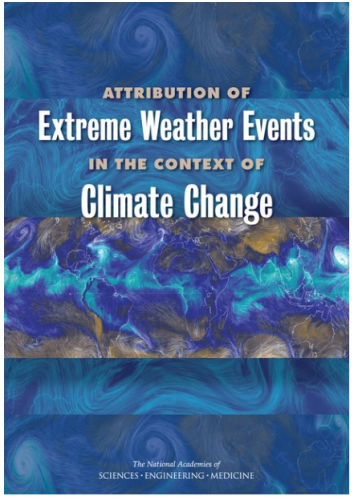
M. C. Kirchmeier-Young[1,2], N. P. Gillett[2], F. W. Zwiers[1], A. J. Cannon[3], and F. S. Anslow[1]

[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

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

Example 1



The relatively young science of **extreme event attribution** seeks to tease out the influence of human-caused climate change from other factors, such as natural sources of variability like El Niño, as contributors to individual extreme events.

Consensus Report

National Academy of Sciences Engineering and Medicine

The data

Kirchmeier-Young et al 2019

"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** … ."

Français

Government of Canada | Gouvernement du Canada

Search Canada.ca

Jobs | Immigration | Travel | Business | Benefits | Health | Taxes | More services

Home → Open Government → The Canadian Regional ...

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-a8fb-68076a4e477c). 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.

Relevant Publications: Description of the Model: J. F. Scinocca, V. V. Kharin, Y. Jiao, M. W. Qian, M. Lazare, L. Solheim, G. M. Flato, S. Biner, M. Desgagne, B. Dugas, Coordinated global and regional climate modeling. J. Clim. 29, 17–35 (2016). https://doi.org/10.1175/JCLI-D-15-0161.1 Examples of applications of the large ensemble: Fyfe, J.C., C. Derksen, L. Mudryk, G.M. Flato, B.D. Santer, N.C. Swart, N.P. Molotch, X. Zhang, H. Wan, V.K. Arora, J. Scinocca, 2017: Large near-term projected snowpack loss over the western United States, Nature Comm., 8:14996, https://doi.org/10.1038/ncomms14996 Kirchmeier-Young, M. C., N. P. Gillett, F. W. Zwiers, A. J. Cannon, F. S. Anslow, 2018: Influence of human-induced climate change on British Columbia's extreme 2017 fire season.

Publisher - Current Organization Name: Environment and Climate Change Canada

Licence: Open Government Licence - Canada

Resources

Resource Name	Resource Type	Format	Language	Links
CanRCM4 Large Ensemble Output	Dataset	NetCDF	English French	Access
Data Dictionary	Guide	PDF	English	Access
Data Dictionary	Guide	PDF	French	Access

Have your say

Rate this dataset
Comment(s)

Additional Information

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

Keywords: large ensembles, regional climate model, climate

Subject: Nature and Environment

Maintenance and Update Frequency: As Needed

Date Published: 2018-09-26

Temporal Coverage: 1950-01-01 to 2100-12-31

Openness Rating:

About this Record

Record Released: 2018-10-12

Simulation of global climate

physics, mathematics

numerical analysis

Creation of regional climate scenarios

mathematics, statistics

The data

Kirchmeier-Young et al 2019

"… values were pooled together for two time periods: **1961-1970 and 2011-2020**, resulting in **500 values** for each decade (10 years x 50 realizations). ""

Français

Government of Canada | Gouvernement du Canada

Search Canada.ca

Jobs | Immigration | Travel | Business | Benefits | Health | Taxes | More services

Home → Open Government → The Canadian Regional ...

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-a8fb-68076a4e477c). 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.

Relevant Publications: Description of the Model: J. F. Scinocca, V. V. Kharin, Y. Jiao, M. W. Qian, M. Lazare, L. Solheim, G. M. Flato, S. Biner, M. Desgagne, B. Dugas, Coordinated global and regional climate modeling. J. Clim. 29, 17–35 (2016). https://doi.org/10.1175/JCLI-D-15-0161.1 Examples of applications of the large ensemble: Fyfe, J.C., C. Derksen, L. Mudryk, G.M. Flato, B.D. Santer, N.C. Swart, N.P. Molotch, X. Zhang, H. Wan, V.K. Arora, J. Scinocca, 2017: Large near-term projected snowpack loss over the western United States, Nature Comm., 8:14996, https://doi.org/10.1038/ncomms14996 Kirchmeier-Young, M. C., N. P. Gillett, F. W. Zwiers, A. J. Cannon, F. S. Anslow, 2018: Influence of human-induced climate change on British Columbia's extreme 2017 fire season.

Publisher - Current Organization Name: Environment and Climate Change Canada

Licence: Open Government Licence - Canada

Resources

Resource Name	Resource Type	Format	Language	Links
CanRCM4 Large Ensemble Output	Dataset	NetCDF	English French	Access
Data Dictionary	Guide	PDF	English	Access
Data Dictionary	Guide	PDF	French	Access

Have your say

Rate this dataset

Comment(s)

Additional Information

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

Keywords: large ensembles, regional climate model, climate

Subject: Nature and Environment

Maintenance and Update Frequency: As Needed

Date Published: 2018-09-26

Temporal Coverage: 1950-01-01 to 2100-12-31

Openness Rating:

About this Record

Record Released: 2018-10-12

Simulation of global climate

physics, mathematics

numerical analysis

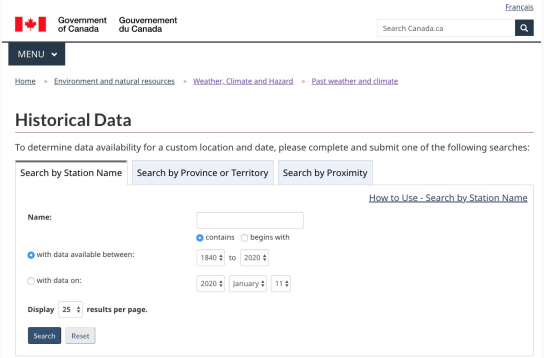
Creation of regional climate scenarios

mathematics, statistics

The data

Kirchmeier-Young et al 2019

"A data set of **temperature and precipitation anomalies** was created … from surface station observations "



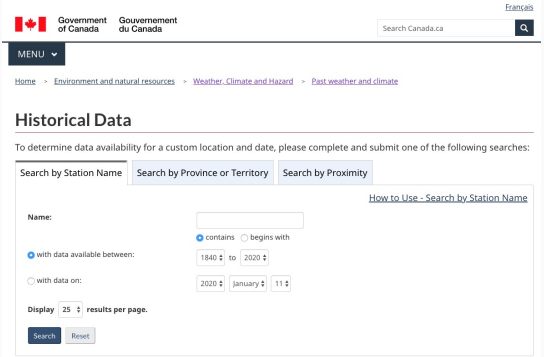
Observational data

statistics, data science

The data

Kirchmeier-Young et al 2019

"Observational data was acquired from numerous sources and interpolated using a **thin plate spline methodology** "



Observational data

statistics, data science

- Fire weather indices

National Aeronautics and Space Administration
Goddard Institute for Space Studies

Goddard Space Flight Center
Sciences and Exploration Directorate
Earth Sciences Division

GISS Home | Global Fire WEather Database (GFWED)

- Precipitation

MSWEP Retrospective

Latest version: V1.2 (released March 2, 2018)

- Fire locations and perimeters

BRITISH COLUMBIA Data Catalogue

What is DataBC? | Dataset Usage | Geographic Services | Blog | Developers | Contact

Datasets | Organizations | Groups | Stay Up To Date | About

- Mean air temperature anomalies

UEA University of East Anglia

Climatic Research Unit

CRU

Home | About CRU | Data | Academic Programmes | Research | Staff and Students | Information Sheets | Publications | Media | News/Events

Environment and Climate Change Canada

BC Hydro
Power smart

RioTinto

Event attribution

Kirchmeier-Young et al 2019

- “anthropogenic climate change increased the **area burned** by a factor of 7 - 11”
- “86 - 91 percent of the area burned can be attributed to anthropogenic climate change”
- “anthropogenic factors increased the likelihood of the extreme warm temperature by over 20 times ”
- “anthropogenic factors increased the likelihood of extreme fire weather indices by 2-4 times”

The analysis

Kirchmeier-Young et al 2019

"anthropogenic climate change increased the area burned by a factor of 7 - 11"

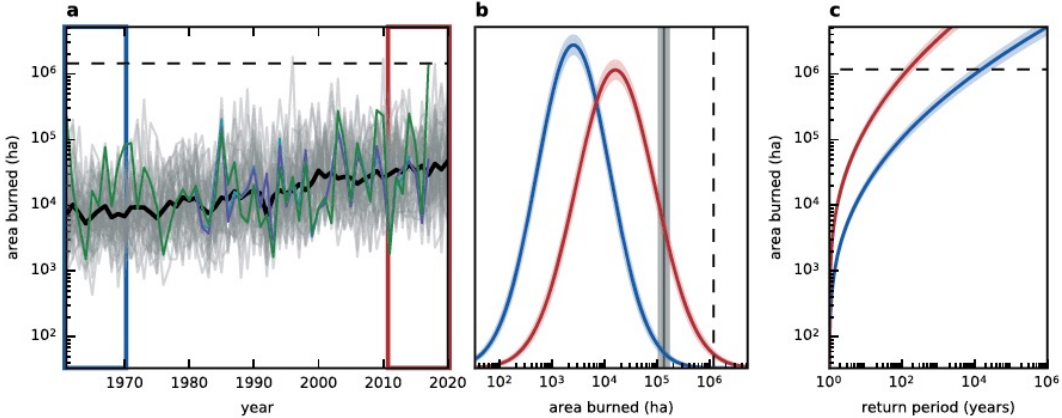


Figure 5. Time series (a, log scale) of regression-predicted annual burned area in the BC Southern Cordillera for bias-corrected CanRCM4 realizations (gray) and ensemble mean (bold), reanalysis (turquoise/purple), and observations (green). The dashed line marks the observed 2017 value. Probability distributions (b) for area burned amounts (log scale)

Event attribution

Kirchmeier-Young et al 2019

- “anthropogenic climate change increased the **area burned** by a factor of 7 - 11”
- “86 - 91 percent of the area burned can be attributed to anthropogenic climate change”
- “anthropogenic factors increased the likelihood of the extreme warm temperature by over 20 times ”
- “anthropogenic factors increased the likelihood of extreme fire weather indices by 2-4 times”

Event attribution

Is climate change to blame for Australia's bushfires?

11 November 2019

Share

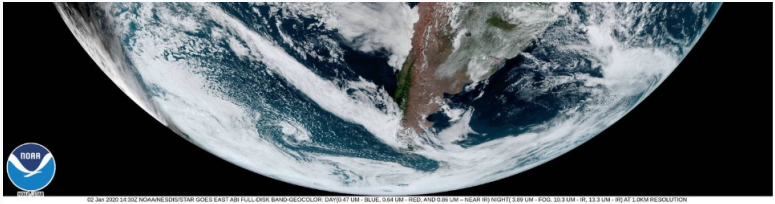
Australia fires

"The science around climate change is complex – it's not the cause of bushfires but scientists have long warned that a hotter, drier climate would contribute to Australia's fires becoming more frequent and more intense."

"We find it very difficult in general to attribute climate change impacts to a specific event, particularly while the event is running, said Dr Richard Thornton, chief executive of the Bushfires & Natural Hazards Co-operative Research Centre."

Climate and weather

The Washington Post
Democracy Dies in Darkness



Satellite image showing weather on Jan. 2, 2019. (NOAA)

By **Andrew Freedman**
January 2

"For the first time, scientists have detected the "fingerprint" of human-induced climate change on daily weather patterns at the global scale"

"If verified by subsequent work, the findings … would upend the long-established narrative"

"The new study … uses statistical techniques and climate model simulations"

Sippel et al., 2020

Examples: Statistics in the news

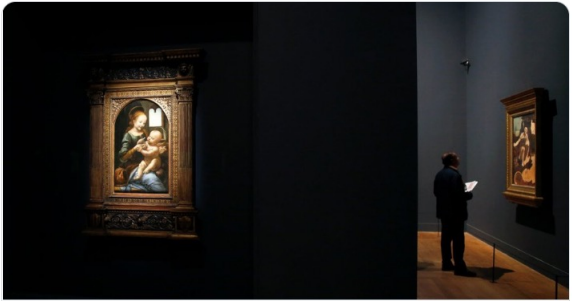
Example 2: Art and Life

Art and Life

NY Times Twitter

NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



Another Benefit to Going to Museums? You May Live Longer
Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.
nytimes.com

Example: Art and Life

Fancourt & Steptoe, 2019

Calling Bullshit @callin_bull · Dec 23, 2019

Want to live longer? Try driving a BMW 7 series or a Mercedes S class.

> **NYT Health** @NYTHealth · Dec 22, 2019
>
> Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't. nyti.ms/2Q9AmZV

8 49 178

Show this thread

Calling Bull with R, Carrie Diaz Eaton, Thursday 6 pm

... art and life

amolpatil1 @amolpatil1 · Dec 22, 2019
Replying to @NYTHealth
In other news, People who eat brunches in absurdly priced Museum cafeterias, live longer
8 28 2.9K

ThankGodWeLiveInTheseTimes @ThankTimes · Dec 23, 2019
Replying to @NYTHealth
Rich people live longer.

WhO kNeW?!
17

CinqL @ashestodust80 · Dec 23, 2019
Replying to @NYTHealth
Seriously you get paid for this??? Oh my god.
18

Q: Did you see that going to the opera makes you live longer?

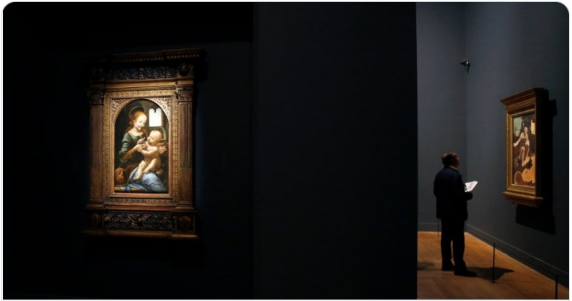
A: No, it just makes it feel longer.

Thomas Lumley, Stats Chat

… art and life

NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



Another Benefit to Going to Museums? You May Live Longer
Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.
nytimes.com

… art and life

NY Times, Dec 22

The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

- … evidence that simply being exposed to the arts may help people live longer
- researchers in London … followed thousands of people 50 and older
- study controlled for socioeconomic factors like … income, education level and mobility
- … researchers collected data from 6,710 people who responded to questionnaires about how often they went to concerts, museums, galleries, the theater or the opera
- the researchers combed through the data they had collected to search for patterns

thebmj Research Education News & Views Campaigns

Archive For a

Research » Christmas 2019: Express Yourself

The art of life and death: 14 year follow-up analyses of associations between arts engagement and mortality in the English Longitudinal Study of Ageing

BMJ 2019 ; 367 doi: https://doi.org/10.1136/bmj.l6377 (Published 18 December 2019)

Cite this as: *BMJ* 2019;367:l6377

RESEARCH

The art of life and death: 14 year follow-up analyses of associations between arts engagement and mortality in the English Longitudinal Study of Ageing

Daisy Fancourt,[1] Andrew Steptoe[1]

[1]Department of Behavioural Science and Health, University College London, London WC1E 7HB, UK

Correspondence to: D Fancourt d.fancourt@ucl.ac.uk (or @Daisy_Fancourt on Twitter; ORCID 0000-0002-6952-334X)

Cite this as: *BMJ* 2019;367:l6377
http://dx.doi.org/10.1136/bmj.l6377

Accepted: 24 September 2019

ABSTRACT

OBJECTIVE

To explore associations between different frequencies of arts engagement and mortality over a 14 year follow-up period.

DESIGN

Prospective cohort study.

PARTICIPANTS

English Longitudinal Study of Ageing cohort of 6710 community dwelling adults aged 50 years and older (53.6% women, average age 65.9 years, standard

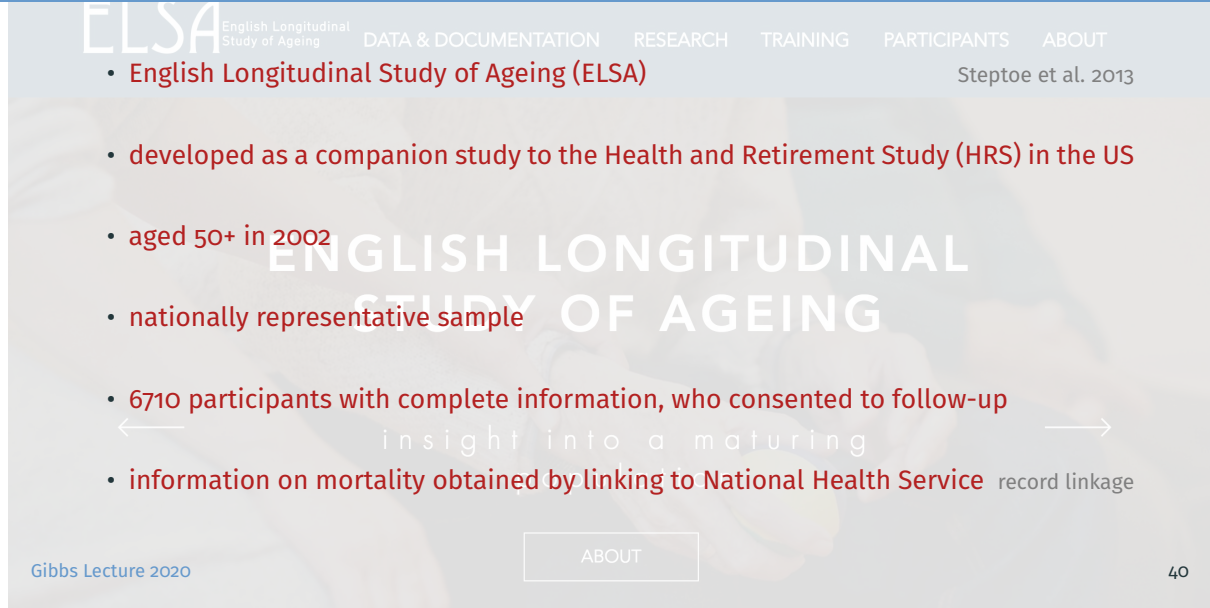
of demographic, socioeconomic, health related, behavioural, and social factors. Results were robust to a range of sensitivity analyses with no evidence of moderation by sex, socioeconomic status, or social factors. This study was observational and so causality cannot be assumed.

CONCLUSIONS

Receptive arts engagement could have a protective association with longevity in older adults. This association might be partly explained by differences in cognition, mental health, and physical activity among those who do and do not engage in the arts.


Search...
ELSA
English Longitudinal Study of Ageing
DATA & DOCUMENTATION
RESEARCH
TRAINING
PARTICIPANTS
ABOUT
ENGLISH LONGITUDINAL STUDY OF AGEING
insight into a maturing population
ABOUT

The people



- English Longitudinal Study of Ageing (ELSA) Steptoe et al. 2013
- developed as a companion study to the Health and Retirement Study (HRS) in the US
- aged 50+ in 2002
- nationally representative sample
- 6710 participants with complete information, who consented to follow-up
- information on mortality obtained by linking to National Health Service record linkage

The “Raw” Data

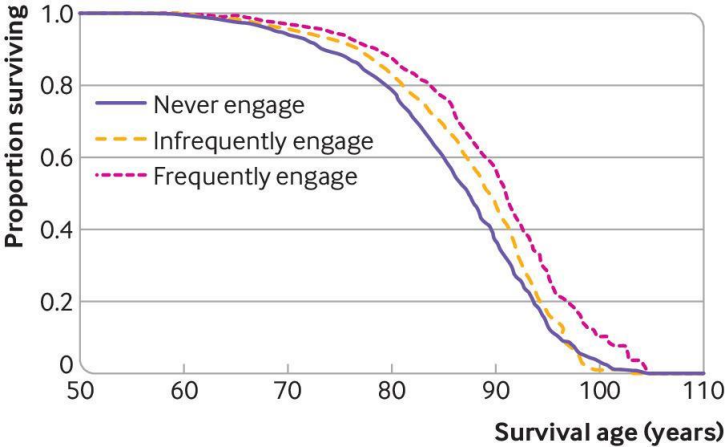
Table 1. Fancourt & Steptoe

Receptive Arts Engagement

	Never	Infrequently	Frequently
Died	837 (47.5%)	809 (26.6%)	355 (18.6%)
Survived	925	2233	1551
Total	1762	3042	1906

The Data

adjusted for confounders



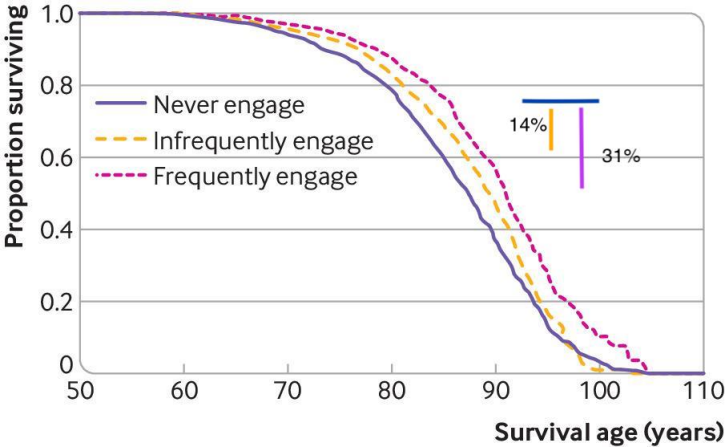
The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

The Data

adjusted for confounders



The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Confounders

pace Twitter

Table 2 | Cox proportional hazards models showing associations between receptive arts engagement and 14 year mortality by calculating the percentage of protective association explained by specific confounding factors

Explanatory factors	Adjusted hazard ratio (95% CI)	P	PPAE (%)
Basic model (age)	0.67 (0.63 to 0.71)	<0.001	—
+sex	0.67 (0.63 to 0.72)	<0.001	0
+education, occupational status, and employment status	0.67 (0.63 to 0.72)	<0.001	0
+wealth	0.70 (0.65 to 0.75)	<0.001	9.1
+cancer, lung disease, cardiovascular disease, or other long term condition	0.67 (0.62 to 0.71)	<0.001	0
+mobility and disability	0.71 (0.66 to 0.75)	<0.001	12.1
+depressive symptoms and psychiatric conditions	0.68 (0.64 to 0.72)	<0.001	3.0
+cognition	0.72 (0.67 to 0.76)	<0.001	15.2
+sensory impairment (hearing and eyesight)	0.67 (0.63 to 0.72)	<0.001	0
+sedentary behaviours	0.69 (0.65 to 0.74)	<0.001	6.1
+other health behaviours (drinking and smoking)	0.70 (0.65 to 0.74)	<0.001	9.1
+loneliness, living status, and marital status	0.69 (0.64 to 0.73)	<0.001	6.1
+social, civic, and hobby engagement	0.71 (0.67 to 0.76)	<0.001	12.1
=all	0.80 (0.75 to 0.87)	<0.001	41.9

PPAE=percentage of protective association explained.
Analysed using receptive arts engagement as a continuous variable. Each line of the table shows an explanatory factor or set of explanatory factors added to the basic model. The final line shows all of these factors entered simultaneously.

Confounders

pace Twitter

Table 2 | Cox proportional hazards models showing associations between receptive arts engagement and 14 year mortality by calculating the percentage of protective association explained by specific confounding factors

Explanatory factors	Adjusted hazard ratio (95% CI)	P	PPAE (%)
Basic model (age)	0.67 (0.63 to 0.71)	<0.001	—
+sex	0.67 (0.63 to 0.72)	<0.001	0
+education, occupational status, and employment status	0.67 (0.63 to 0.72)	<0.001	0
+wealth	0.70 (0.65 to 0.75)	<0.001	9.1
+cancer, lung disease, cardiovascular disease, or other long term condition	0.67 (0.62 to 0.71)	<0.001	0
+mobility and disability	0.71 (0.66 to 0.75)	<0.001	12.1
+depressive symptoms and psychiatric conditions	0.68 (0.64 to 0.72)	<0.001	3.0
+cognition	0.72 (0.67 to 0.76)	<0.001	15.2
+sensory impairment (hearing and eyesight)	0.67 (0.63 to 0.72)	<0.001	0
+sedentary behaviours	0.69 (0.65 to 0.74)	<0.001	6.1
+other health behaviours (drinking and smoking)	0.70 (0.65 to 0.74)	<0.001	9.1
+loneliness, living status, and marital status	0.69 (0.64 to 0.73)	<0.001	6.1
+social, civic, and hobby engagement	0.71 (0.67 to 0.76)	<0.001	12.1
=all	0.80 (0.75 to 0.87)	<0.001	41.9

PPAE=percentage of protective association explained.
Analysed using receptive arts engagement as a continuous variable. Each line of the table shows an explanatory factor or set of explanatory factors added to the basic model. The final line shows all of these factors entered simultaneously.

Arts Engagement and Mortality

Supplementary Analyses

Supplementary Table 1: Cox proportional hazards regression models showing associations between cultural engagement and 14-year mortality split by gender

	Men (n=3,115)		Women (n=3,595)	
	HR	95% CI	HR	95% CI
USING CULTURE AS A CONTINUOUS EXPOSURE				
Cultural engagement	0.84	0.76-0.93	0.82	0.74-0.91
USING CULTURE AS A CATEGORICAL EXPOSURE				

Statistical Analysis

- key analysis was based on a regression model for survival data
  proportional hazards regression
- checked the proportional hazards assumption
  using residuals
- weighted analysis to accommodate non-response
  which was relatively minimal
- three sets of sensitivity analyses compared to initial analysis
  - several subgroup analyses
    age, sex, SES, etc.
  - finer adjustment for confounders
  - further testing of model assumptions
    including reverse causality

The Conclusions

- results are broadly consistent with related literature
- study found a dose-response effect
- "this study suggests that receptive arts engagement **could** have independent longitudinal protective associations with longevity"

  my emphasis
- "this study did not compare the relative effect size of arts and other known predictors of mortality, but other factors undoubtedly have a larger bearing on mortality risk"
- "A causal relationship cannot be assumed, and unmeasured confounding factors might be responsible for the association"

The Conclusions

"As we always say, **correlation doesn't imply causation**—but it doesn't sell newspapers either"

Calling Bullshit

The Stats Chat chocolate rule: "if you're going to a concert or visiting a museum primarily for the health effects, you're doing it wrong"

Lumley

Daisy Fancourt
@Daisy_Fancourt

Today my paper w/ @andrewp_steptoe is published in the @bmj_latest showing arts engagement is associated with longevity in older adults. Confounders obviously a big challenge but results consistent in well adjusted models & multiple sensitivity analyses.

Reproducibility and Visualization

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

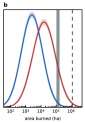
Ethics, policy and social impact

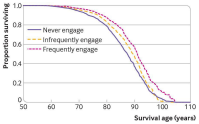
Reproducibility

- there are some really good statistical analyses
- paired with really good science Kirchmeier et al
- paired with really good social science

  Fancourt & Steptoe
- there are some really bad ones too

- reproducibility of science is harmed by rote use of any tool
- *p*-values are one of those tools that are misunderstood and mis-used
- let your alarm bells ring when you hear "small, but statistically significant"
- it's probably more complicated than that most science is





Dementia

Claims about a treatment for Alzheimer's should be met with caution

More trials would be a good idea

SOCIETING • FACEBOOK

How Quitting Facebook Could Change Your Life

By Erin Corbett January 31, 2019

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

Visualization

Australia fires: Misleading maps and pictures go viral

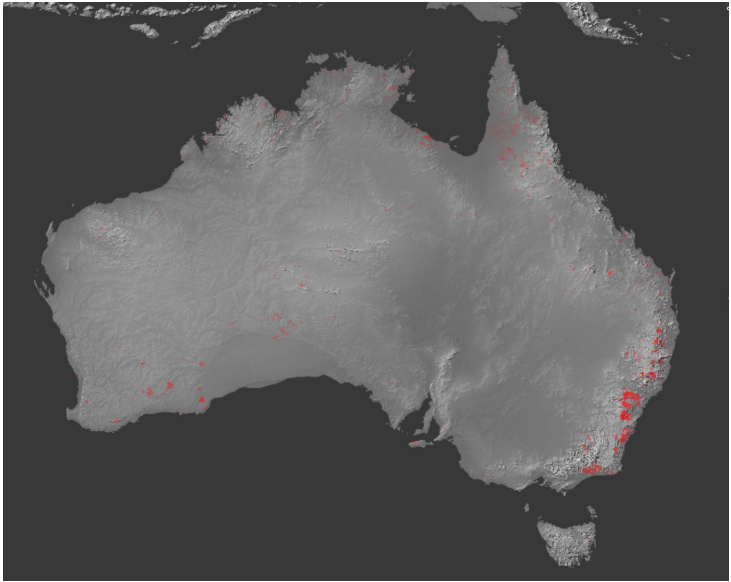
By Georgina Rannard
BBC News

7 January 2020

Share

Visualization

statschat.org.nz; Guardian Nick Evershed



In praise of small data

"you don't have as much data as you thought"

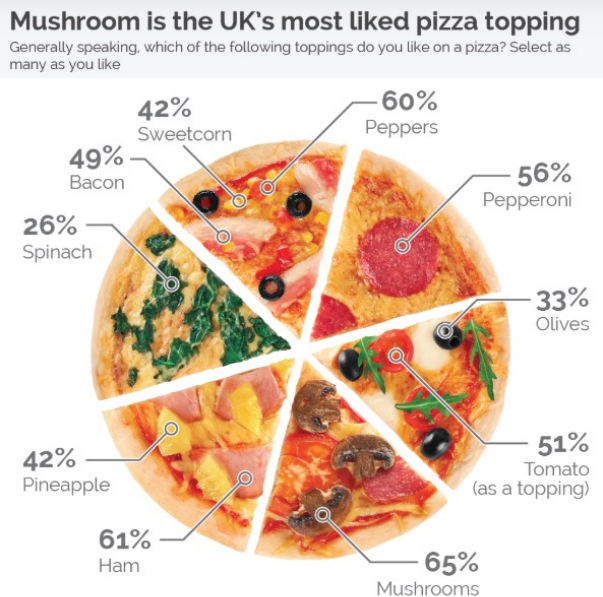
- Wildfires and climate change
- Going to the museum
- "We have a billion observations"
- "We have 1000 observations in every county on every day"
- Lots of data needs complex modelling

- 50 climate simulations, times 10 years
- 6710 people with complete information
- "but we're looking for extremely rare events" Higgs boson
- but what about correlation in time and space
- new statistical theory for high-dimensions, complex dependence, extreme values

quality is much more important than quantity

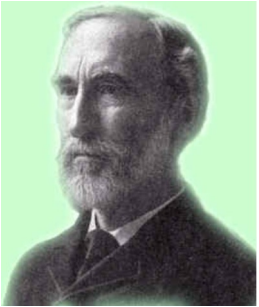
Pie Charts!!

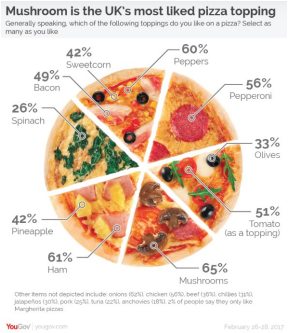
YouGov



In praise of small data

Thank you!





References i

Slide 2. Gibbs (1902). *Elementary Principles in Statistical Mechanics.* https://books.google.com/books?hl=en&lr=&id=2oc-AAAAIAAJ&oi=fnd&pg=PA32&dq=J+Gibbs&ots=QnQo7iS-TG&sig=Lz5FJ9sKyWSI09ldnrkBhmvu8eE#v=onepage&q&f=false

Slide 3. Geman and Geman (1984). *IEEE Transactions on Pattern Analysis and Machine Intelligence.* PAMI 6, 721–741.

Slide 5. Gelfand and Smith (1990). *Journal of the American Statistical Association* 85, 398–409. Casella and George (1992). *The American Statistician* 46, 167–174.

Slide 8. AmStat News, Dec 2019, 6–9. https://magazine.amstat.org/wp-content/uploads/2019/11/DECEMBER2019.pdf

Slide 11. Talks from the Big Data program available at https://www.fields.utoronto.ca/video-archive/2014 and from the Data Science workshop at https:/www.fields.utoronto.ca/video-archive/event/2553

Slide 14. Meng (2019). Harvard Data Science Review 2. https://review.datascience.harvard.edu/

Slide 15. Cox and Donnelly (2011). *Principles of Applied Statistics.* Cambridge University Press.

References ii

Slide 21. Kirchmeier-Young et al. (2019). Earth's Future, 7. https://doi.org/10.1029/2018EF001050

Slide 22. *NASEM Consensus Report on Attribution of Extreme Weather Events.* https://www.nap.edu/catalog/21852/attribution-of-extreme-weather-events-in-the-context-of-climate-change

Slide 30. BBC News Australia, 11 Nov 2019. https://www.bbc.com/news/world-australia-50341210

Slide 31. Freedman, A. (2020) Washington Post, Jan 2. https://www.washingtonpost.com/weather/2020/01/02/signal-human-caused-climate-change-has-emerged-every-day-weather-study-finds/?et_cid=3149408&et_rid=49319077&utm_campaign=SoT-23734&utm_medium=email&utm_source=Science_on_Tap.
Sippel et al.(2020). *Nature Climate Change* 10, 35–41.

Slide 32. NY Times museum tweet https://twitter.com/NYTHealth/status/1208754010422939650

Slide 34. Lumley, T. Stats Chat (2020/01/07) https://www.statschat.org.nz/2020/01/07/something-to-do-on-your-holiday/

References iii

Slide 36. Cramer, M., NY Times (2019/12/22). https://www.nytimes.com/2019/12/22/us/arts-health-effects-ucl-study.html?smtyp=cur&smid=tw-nythealth

Slide 37. Fancourt and Steptoe (2019) *British Medical Journal* 367:16377. https://www.bmj.com/content/367/bmj.l6377

Slide 39. *English Longitudinal Study on Aging.* https://www.elsa-project.ac.uk/

Slide 40. Steptoe et al. (2013). *International Journal of Epidemiology* 42, 1640-8. doi:10.1093/ije/dys168.

Slide 51. Abbott, Nature News (2019) https://www.nature.com/articles/d41586-019-03261-5
Economist (2019/10/24) https://www.economist.com/science-and-technology/2019/10/24/claims-about-a-treatment-for-alzheimers-should-be-met-with-caution
Salzberg, Forbes (2019/01/06) https://www.forbes.com/sites/stevensalzberg/2020/01/06/can-intermittent-fasting-reset-your-immune-system/#63d2dfcc27ac
de Cabo and Mattson (2019). *New England Journal of Medicine* 381:2541–51. DOI:10.1056/NEJMra1905136
Corbett, Fortune (2019/01/31). https://fortune.com/2019/01/31/quitting-facebook-life-changing/

References iv

Allcott et al. (2019). National Bureau of Economic Research
https://www.nber.org/papers/w25514

Slide 54. Lumley, T. Stats Chat (2020/01/08).
https://www.statschat.org.nz/2020/01/08/misleading-with-maps/

Slide 56. Smith, M. YouGov (2019/03/06) https://yougov.co.uk/topics/politics/articles-reports/2017/03/06/does-pineapple-belong-pizza