

Big Data, Data Science, Statistics

Nancy Reid

21 July 2017



[Home](#) [News](#) [Quick Links](#) [Accommodation & Transportation](#) [Scientific Programme](#) [Social Programme & Tours](#) [Plan Your Trip](#) [Useful Information](#) [Registration](#)

Statistical Inference, Learning and Models in Big Data

**Beate Franke¹, Jean-François Plante², Ribana Roscher³,
En-Shiun Annie Lee⁴, Cathal Smyth⁵, Armin Hatefi⁵,
Fuqi Chen⁶, Einat Gil⁵, Alexander Schwing⁵,
Alessandro Selvitella⁸, Michael M. Hoffman⁵,
Roger Grosse⁵, Dieter Hendricks⁷ and Nancy Reid⁵**

¹*University College London, London, UK*

²*HEC Montréal, Montréal, Québec, Canada*

³*Freie Universität, Berlin, Germany*

⁴*University of Waterloo, Waterloo, Ontario, Canada*

⁵*University of Toronto, Toronto, Ontario, Canada*

E-mail: reid@utstat.utoronto.ca

⁶*Western University, London, Ontario, Canada*

⁷*University of Witswatersrand, Johannesburg, South Africa*

⁸*McMaster University, Hamilton, Ontario, Canada*



**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

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)

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)

For more information, allied activities off-site, and registration, please visit:

www.fields.utoronto.ca/programs/scientific/14-15/bigdata

Image Credits: Sheelagh Carpendale & InnoVis



**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 Norman

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

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 Vancouver and University of Toronto, and August.

ORGANIZING COMMITTEE

- Yoshua Bengio** (Montréal)
- Hugh Chipman** (Acadia)
- Sallie Keller** (Virginia Tech)
- Lisa Lix** (McGill)
- 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)

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)

For more information, allied activities off-site, and registration, please visit:

www.fields.utoronto.ca/programs/scientific/14-15/bigdata

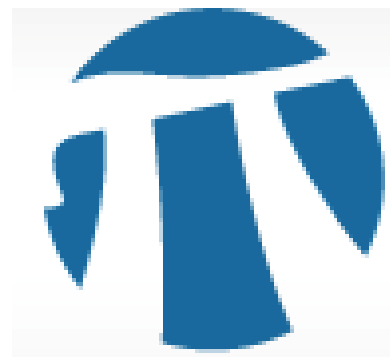
Image Credits: Sheelagh Carpendale & InnoVis



Canadian Institute for Statistical Sciences



Fields Institute
for Resesarch
in the
Mathematical
Sciences



Pacific Institute
for
Mathematical
Sciences



Centre de Recherches Mathématiques



NSERC
CRSNG



Ontario



**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

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)

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)

For more information, allied activities off-site, and registration, please visit:

www.fields.utoronto.ca/programs/scientific/14-15/bigdata

Image Credits: Sheelagh Carpendale & InnoVis

Workshops

- Opening Conference and Bootcamp
- Statistical Machine Learning
- Optimization and Matrix Methods
- Visualization: Strategies and Principles
- Big Data in Health Policy
- Big Data for Social Policy



Workshops

- Opening Conference and Bootcamp
- Statistical Machine Learning
- Optimization and Matrix Methods
- Visualization: Strategies and Principles
- Big Data in Health Policy
- Big Data for Social Policy



FieldsLive Video Archive

Workshops

- Opening Conference and Bootcamp
- Statistical Machine Learning
- Optimization and Matrix Methods
- Visualization: Strategies and Principles
- Big Data in Health Policy
- Big Data for Social Policy
- Networks, Web mining, and Cyber-security
- Statistical Theory for Large-scale Data
- Challenges in Environmental Science
- Complex Spatio-temporal Data
- Commercial and Retail Banking



FieldsLive Video Archive





FIELDS

THE FIELDS INSTITUTE



**THEMATIC PROGRAM ON
STATISTICAL INFERENCE,
LEARNING, AND MODELS FOR**

JANUARY - JUNE, 2015

PROGRAM

**BIG
DATA**

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

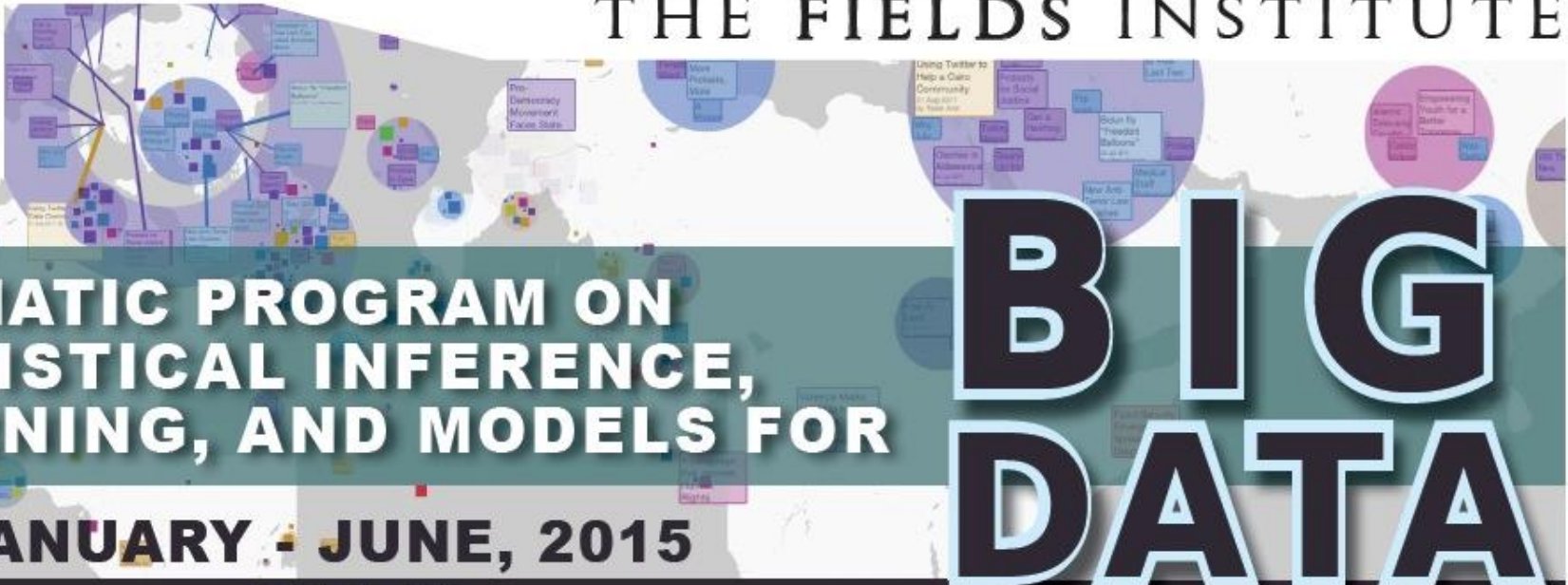
Workshop on Optimization and Matrix Methods in Big Data

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



FIELDS

THE FIELDS INSTITUTE



THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR

JANUARY - JUNE, 2015

BIG DATA

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

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

2. 'Big Data: it's not the Data'

Volume, Variety, Velocity, Veracity, Beyond the Vs

3. Strategies for Big Data Analysis

Data Wrangling, Visualisation, Reducing Dimensionality, Sparsity and Regularisation, Optimisation, Measuring Distance, Representation Learning, Sequential Learning, Multi-Disciplinarity

4. Illustrations

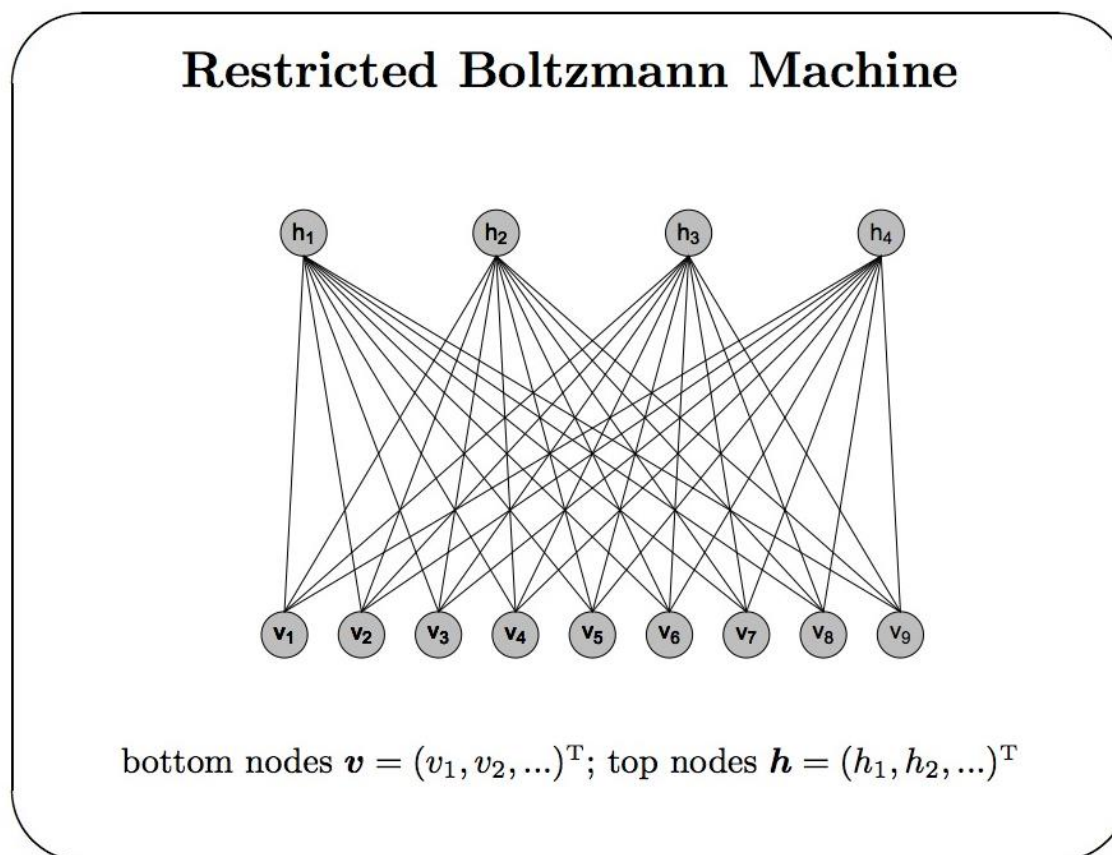
Public Health, Health Policy, Law and Order, Environmental Sciences, Education, Mobile Application Security, Image Recognition and Labelling, Digital Humanities, Materials Science

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

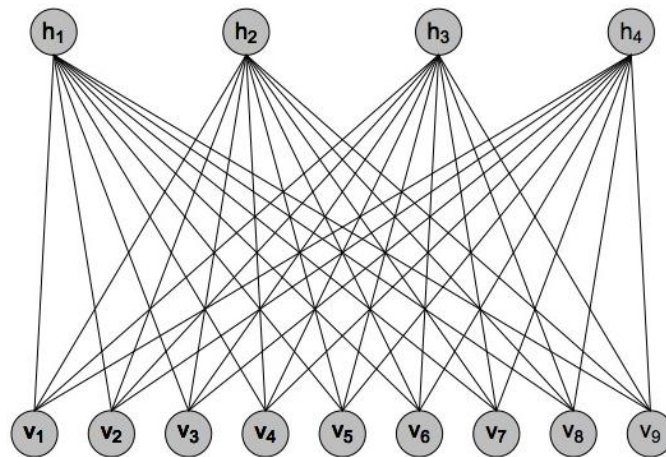
- Statistical Machine Learning



Statistical Machine Learning

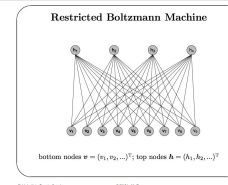
$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\} \quad \eta = (a, b, W)$$

Restricted Boltzmann Machine



bottom nodes $\mathbf{v} = (v_1, v_2, \dots)^T$; top nodes $\mathbf{h} = (h_1, h_2, \dots)^T$

Restricted Boltzmann machine

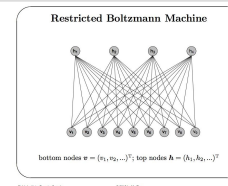


$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- natural gradient ascent

$$\eta \longleftarrow \eta + \epsilon i(\eta)^{-1} \nabla_{\eta} \ell(\eta; v, h) \quad \ell = \log f$$

Restricted Boltzmann machine



$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- natural gradient ascent

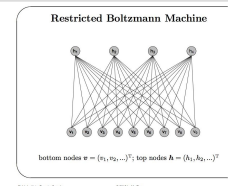
$$\eta \longleftarrow \eta + \epsilon i(\eta)^{-1} \nabla_{\eta} \ell(\eta; v, h) \quad \ell = \log f$$

- uses Fisher information as metric tensor

$$i = \mathbb{E}(-\ell'')$$

Girolami and Calderhead (2011); Amari (1987); Rao (1945)

Restricted Boltzmann machine



$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- natural gradient ascent

$$\eta \longleftarrow \eta + \epsilon i(\eta)^{-1} \nabla_{\eta} \ell(\eta; v, h) \quad \ell = \log f$$

- uses Fisher information as metric tensor

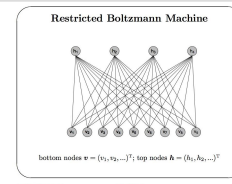
$$i = \mathbb{E}(-\ell'')$$

Girolami and Calderhead (2011); Amari (1987); Rao (1945)

- Gaussian graphical model approximation to force sparse inverse

Grosse and Salakhutdinov (2016) 32nd Internat. Conf. on Machine Learning

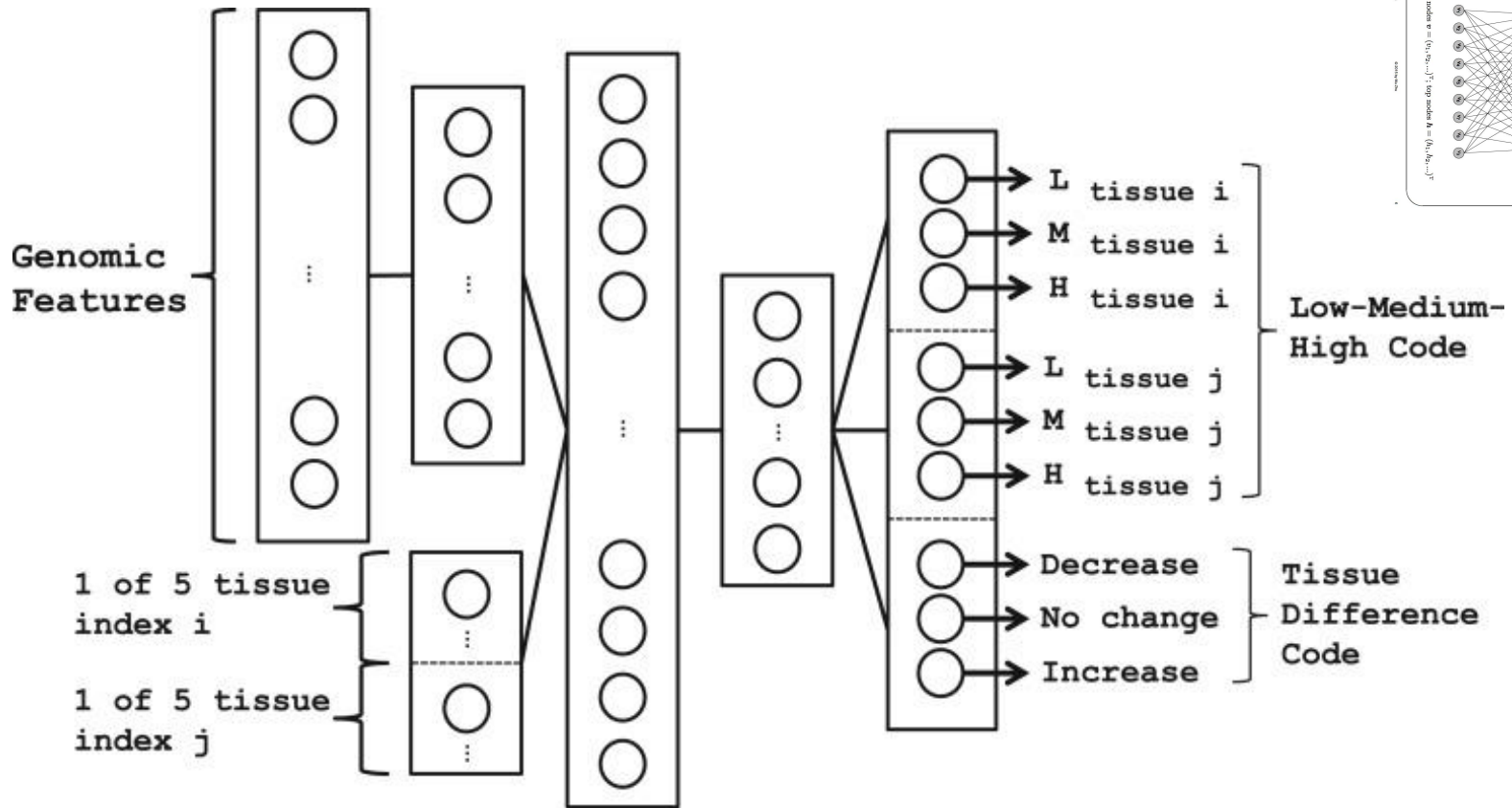
Restricted Boltzmann machine



$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- model for $h \mid \underline{v}$ is a logistic regression
- with odds ratio depending only on \underline{v}
- deep learning has ~ 10 layers, with millions of units
in each layer
- estimating parameters is an **optimization** problem

Restricted Boltzmann machine



Brendan Frey, Infinite Genomes Project

FieldsLive January 27 2015

Leung et al Bioinformatics 2014

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

- Optimization

$$\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log f(y_i | x_i; \theta) - \mathcal{P}_{\lambda}(\theta) \right\}$$

Optimization

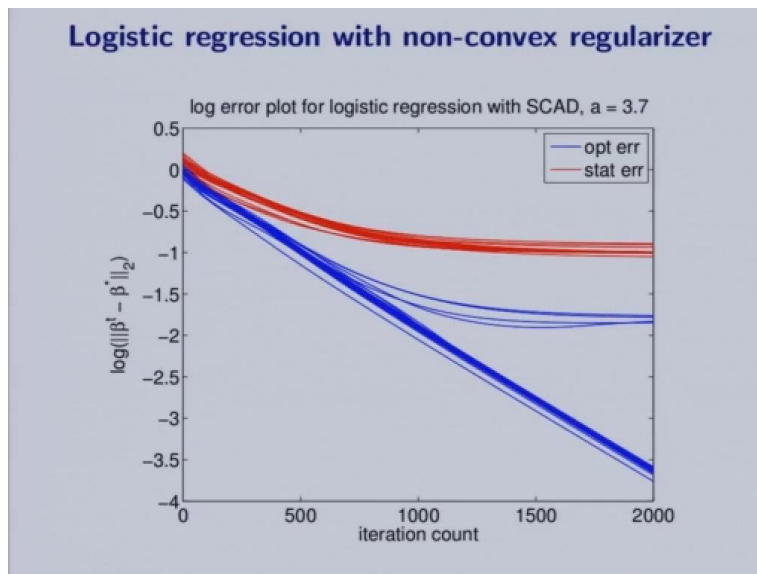
$$\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log f(y_i | x_i; \theta) - \mathcal{P}_{\lambda}(\theta) \right\}$$

- lasso penalty $\mathcal{P}_{\lambda}(\theta) = \lambda \|\theta\|_1 = \lambda \sum |\theta_j|$
- $\|\theta\|_1$ is convex relaxation of $\|\theta\|_0$
- many interesting penalties are non-convex
- optimization routines may not find global optimum

Optimization

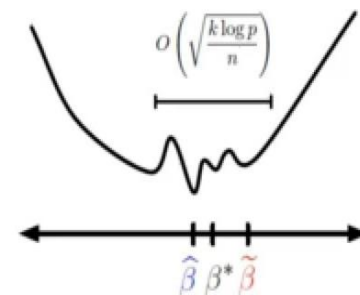
$$\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log f(y_i | x_i; \theta) - \mathcal{P}_{\lambda}(\theta) \right\}$$

- **statistical error** $\hat{\theta} - \theta^*$ neighbourhood of true value
- **approximation error** $\theta_t - \hat{\theta}$ iterating over t



Wainwright FieldsLive Jan 16 2015

Loh and Wainwright *JMLR* 2015

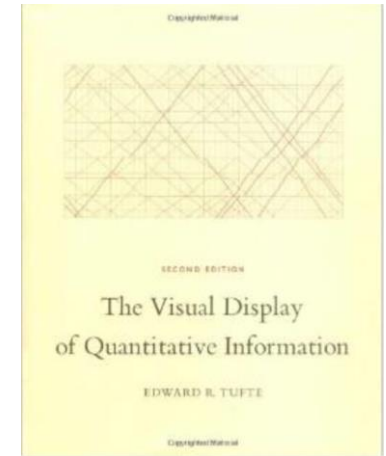


Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

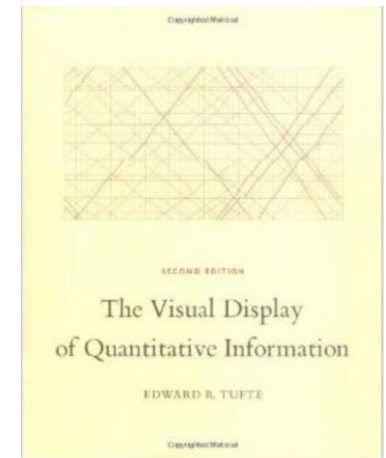
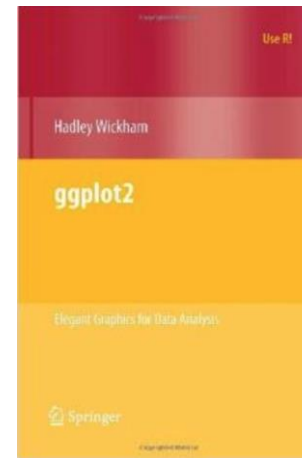
Visualization

- statistical graphics
 - data representation
 - data exploration
 - filtering, sampling aggregation



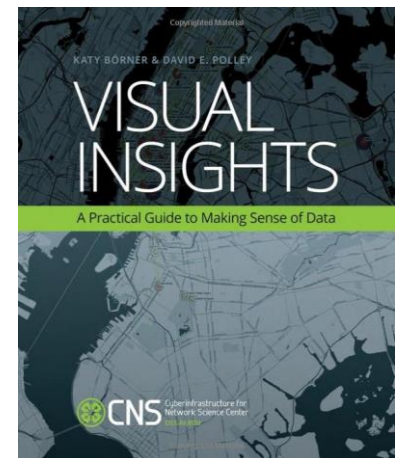
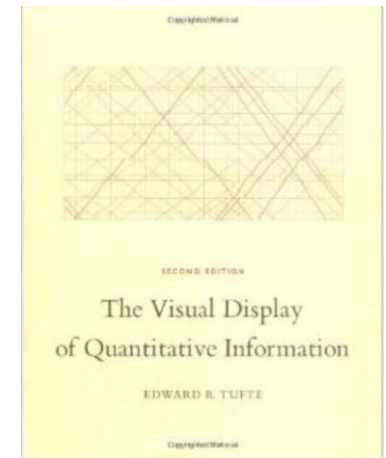
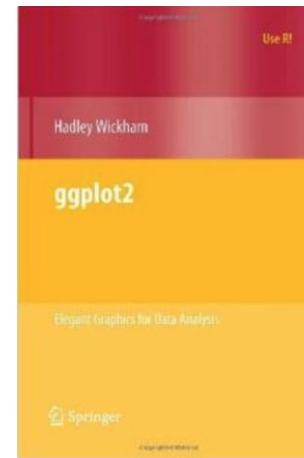
Visualization

- statistical graphics
 - data representation
 - data exploration
 - filtering, sampling aggregation
- information visualization
- scientific visualization
- cognitive science and design



Visualization

- statistical graphics
 - data representation
 - data exploration
 - filtering, sampling aggregation
- information visualization
- scientific visualization
- cognitive science and design



**Global health
innovation - global
development
professionals
network**

How the world got fat: a visualisation of global obesity over 40 years

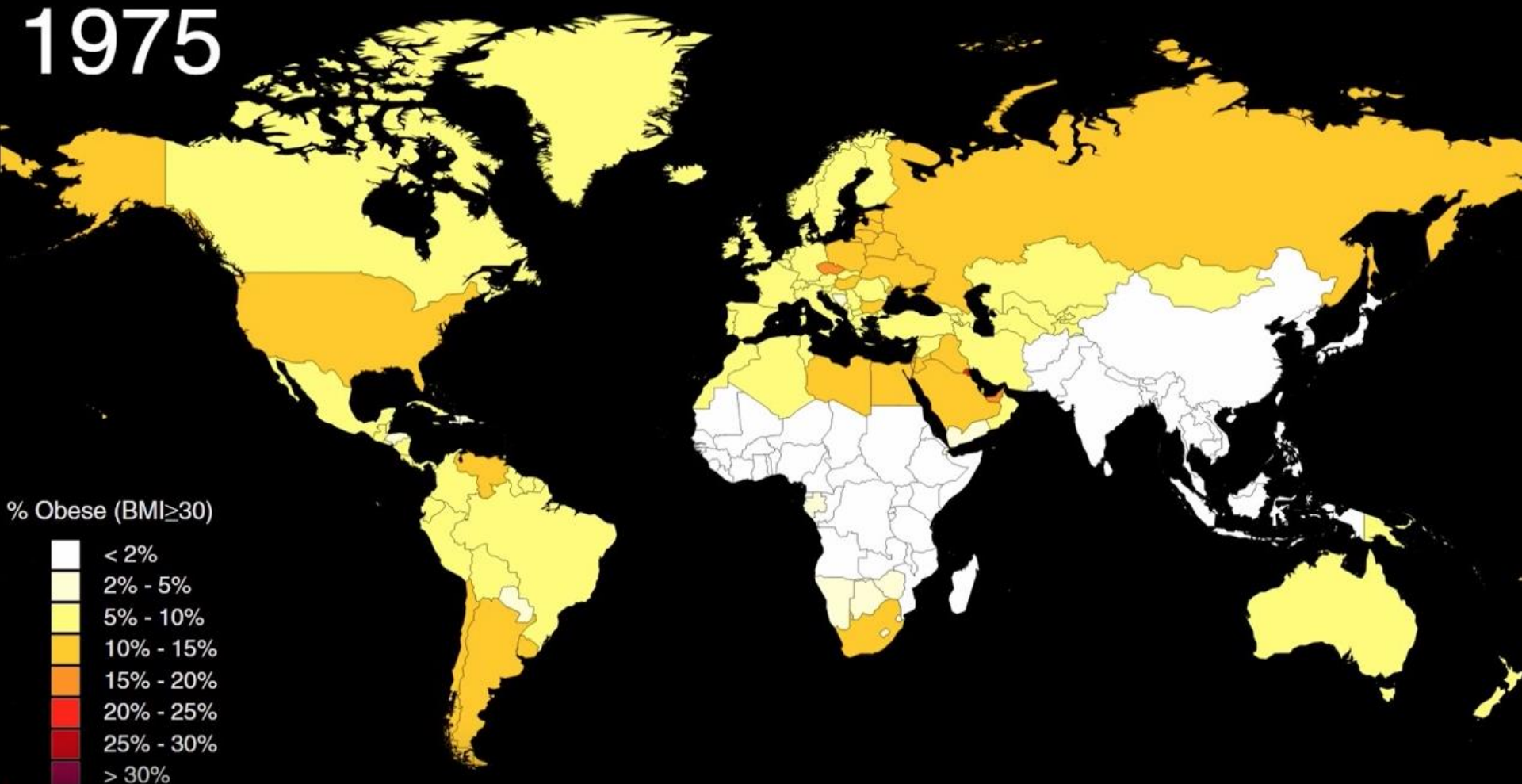
Max Galka

January 3 2017

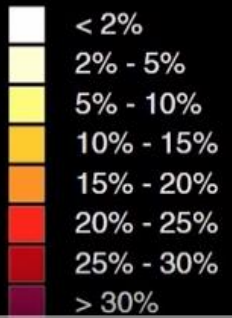
Global health
innovation - global

How the world got fat: a visualisation of

1975



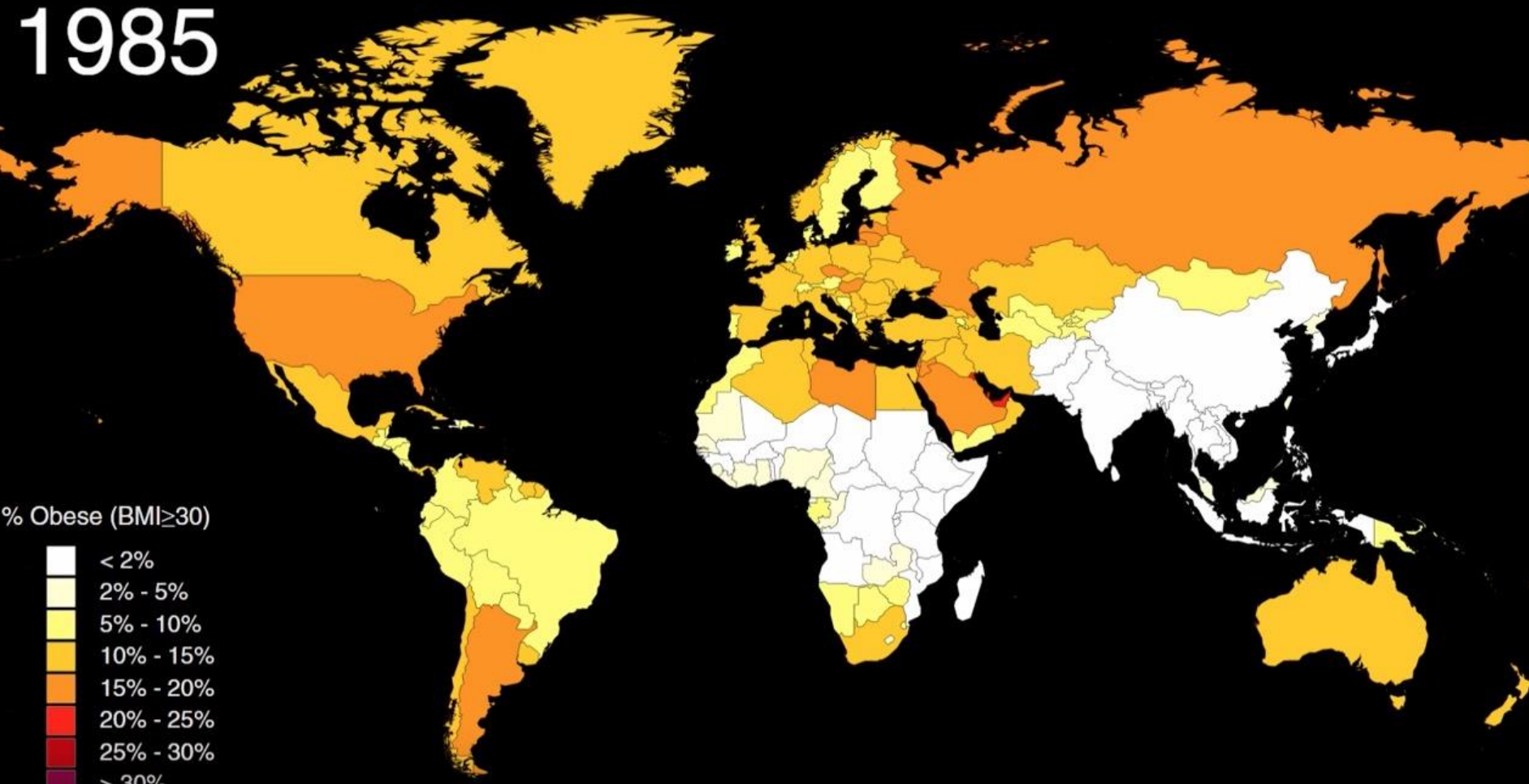
% Obese (BMI ≥ 30)



Global health
innovation - global

How the world got fat: a visualisation of

1985



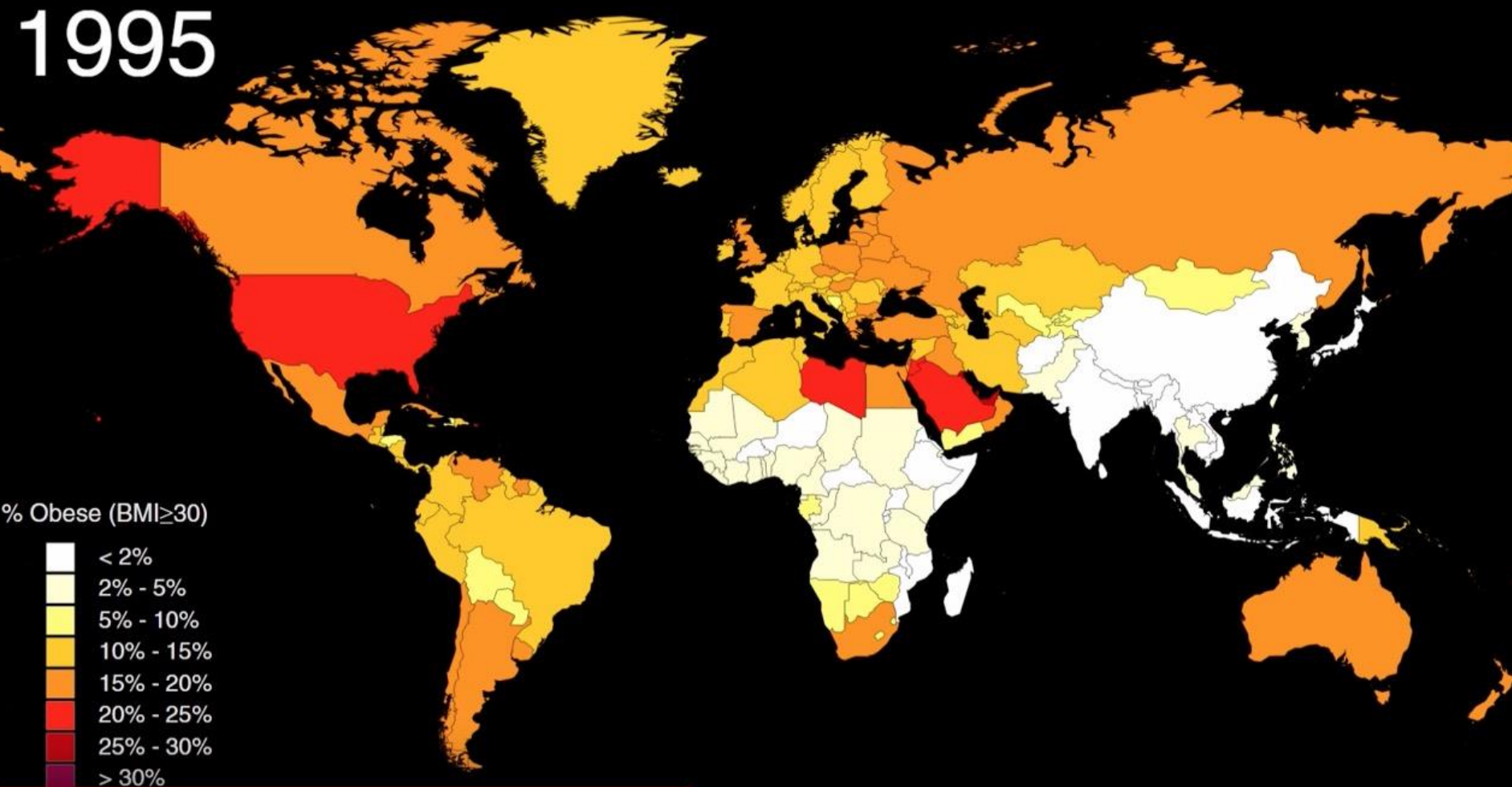
% Obese (BMI ≥ 30)

- < 2%
- 2% - 5%
- 5% - 10%
- 10% - 15%
- 15% - 20%
- 20% - 25%
- 25% - 30%
- > 30%

Global health innovation - global

How the world got fat: a visualisation of

1995



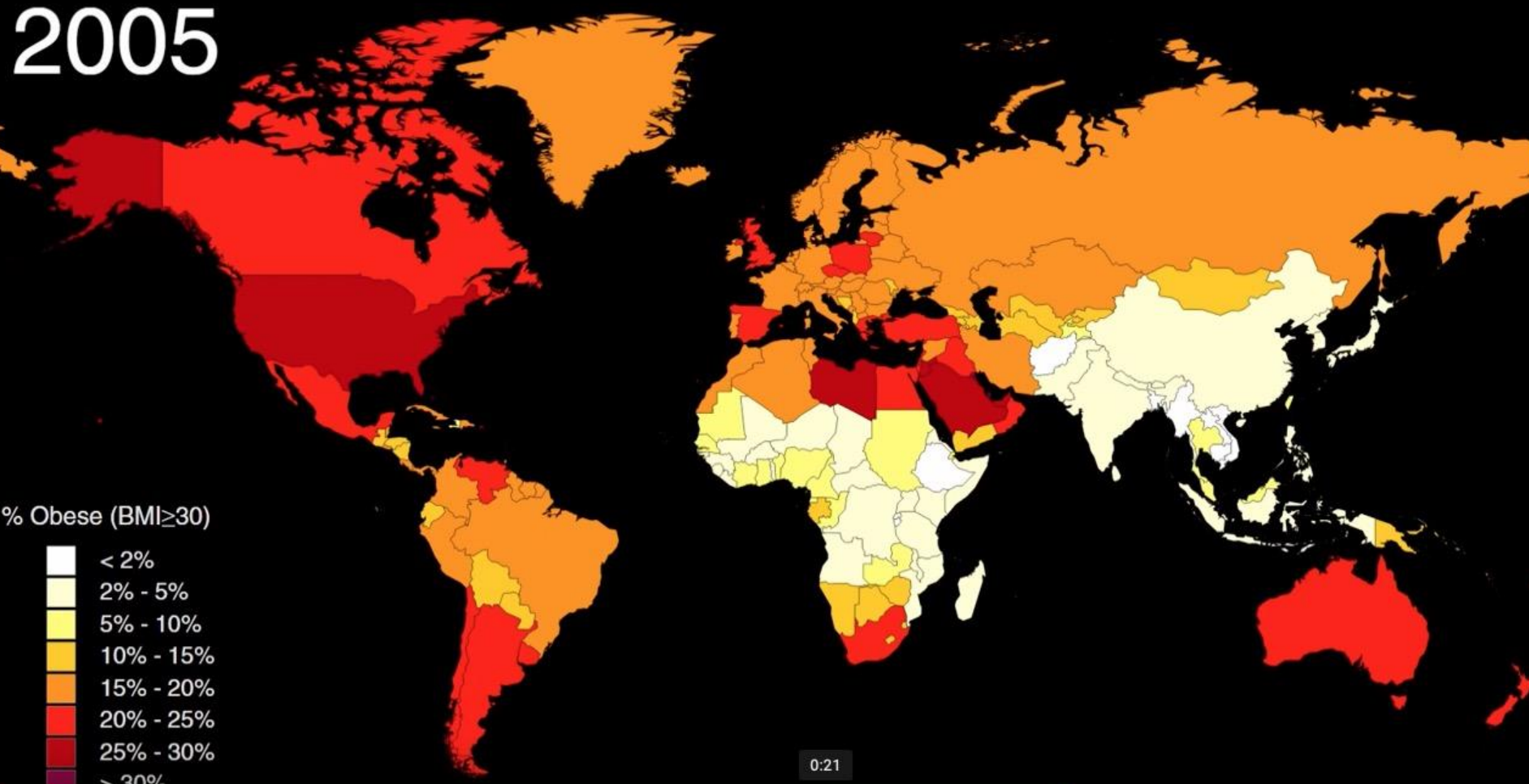
% Obese (BMI ≥ 30)

- < 2%
- 2% - 5%
- 5% - 10%
- 10% - 15%
- 15% - 20%
- 20% - 25%
- 25% - 30%
- > 30%

Global health
innovation - global

How the world got fat: a visualisation of

2005

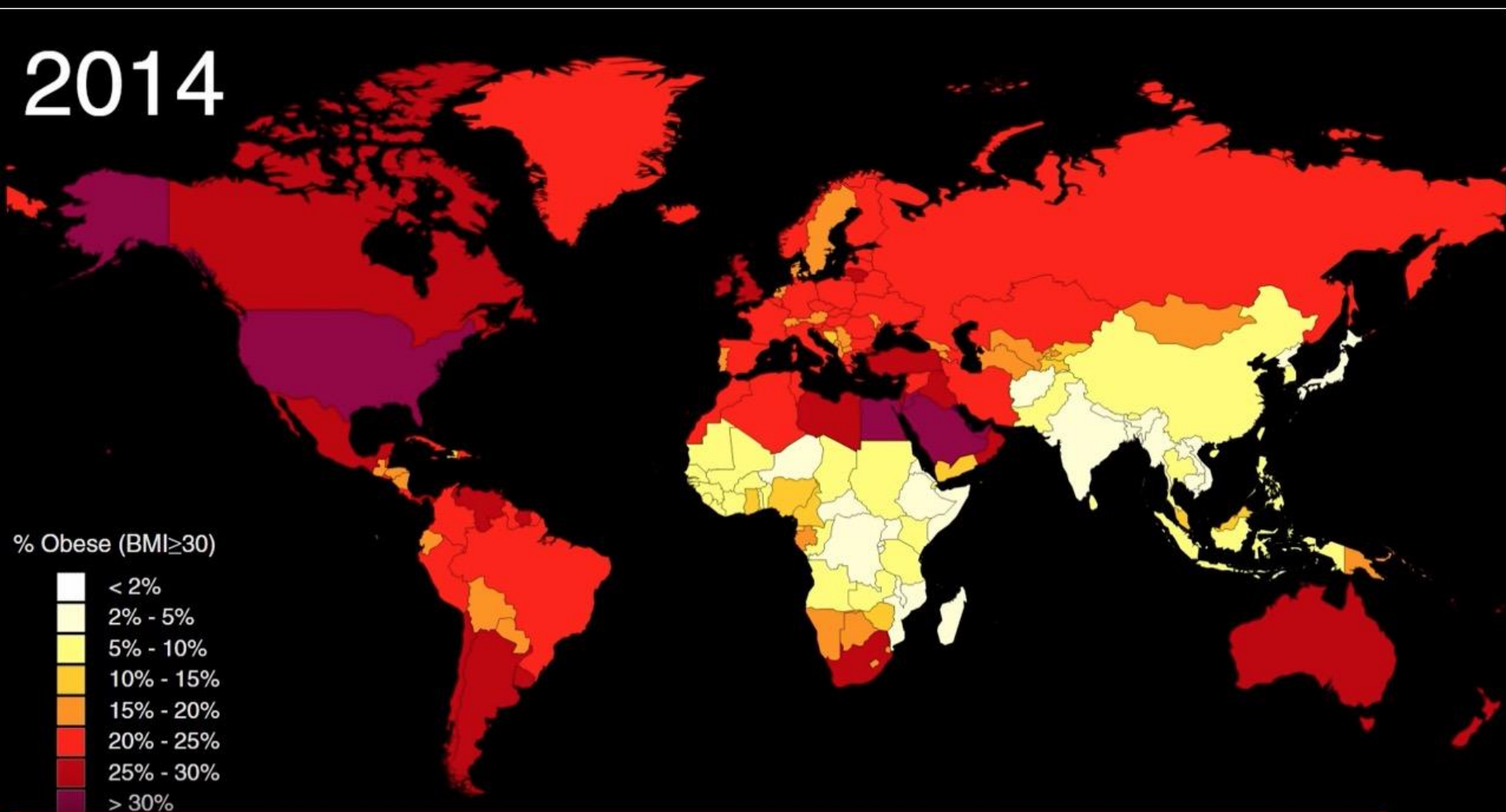


0:21

Global health
innovation - global

How the world got fat: a visualisation of

2014

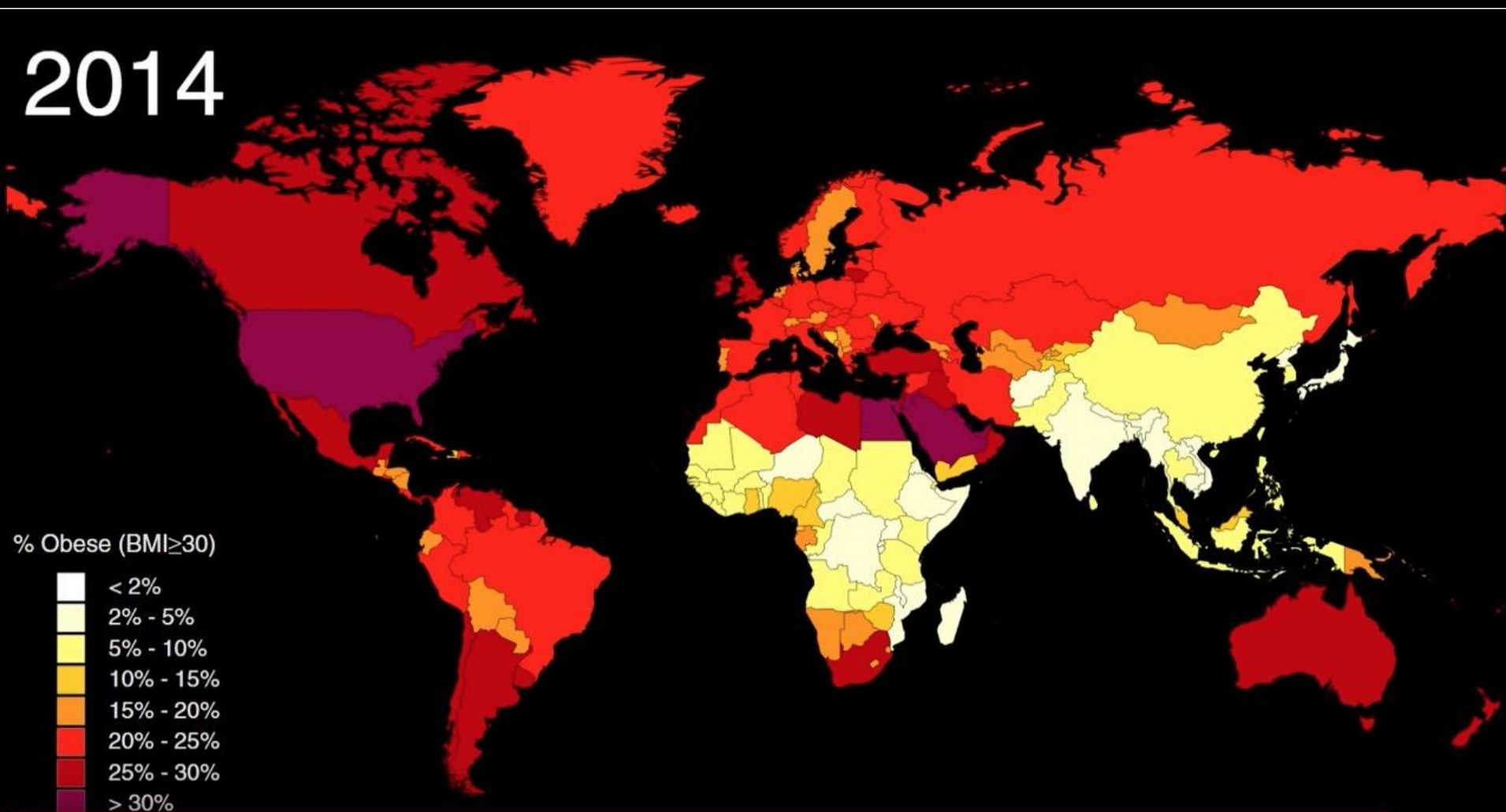


% Obese (BMI ≥ 30)

Global health
innovation - global

How the world got fat: a visualisation of

2014

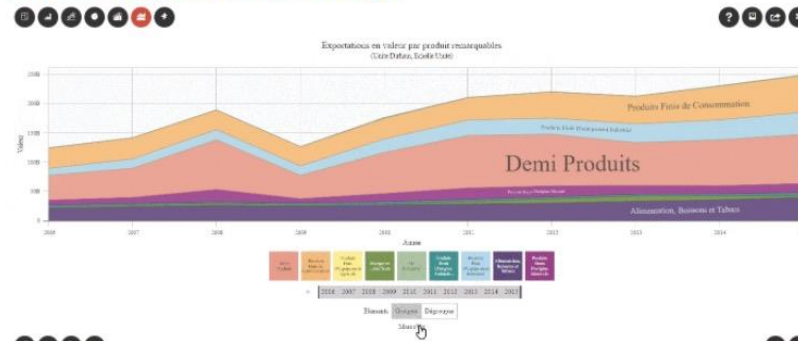


Morocco opendata visualization engine

9 JANUARY 2017

Open Data enables governments, businesses and entrepreneurs around the world to act as a catalyst and tool for social and economic change in diverse sectors. In Morocco in order to realize the full potential of Open Data, a Civic Tech organization built [Marocviz](#), a visualization platform for Morocco's public data. The platform, makes information accessible and digestible with [narrative visualizations](#). The platform is currently one of 70 proposals that have been shortlisted out of the 736 applications for a grant from the [innovateAFRICA](#) fund.

marocviz_visualizations.gif





BIENVENUE SUR MAROCVIZ

Le Moteur de visualisation de données ouvertes convertie pour vous les données publiques en formes narratives de visualisation.

EXPLORER



MAROCVIZ L'OPENDATA VISUALISATION ENGINE

La plate-forme convertie pour vous les données publiques en formes narratives de visualisation

Rapide et facile:

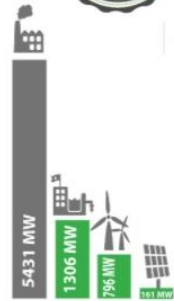
Trouver et visualiser les données dont vous avez besoin en quelques secondes. La plate-forme vous guide avec ses multiples options de recherche et navigation afin de répondre à vos questions ou simplement découvrir et explorer les données qui vous intéressent.



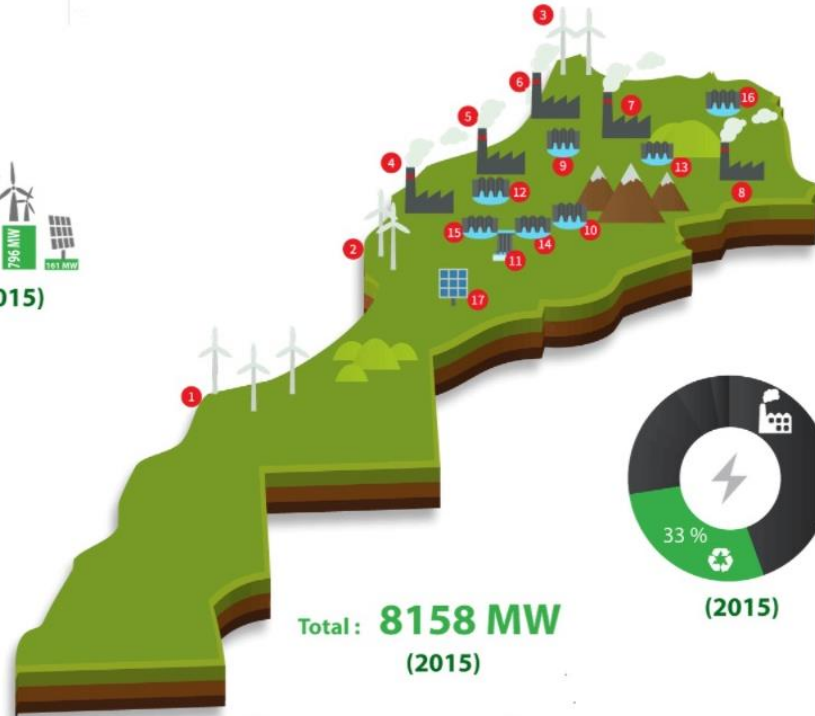


Marocviz.ma

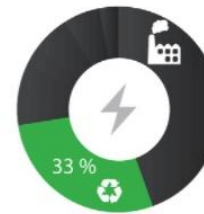
Les Centrales et leur Puissance



(2015)



Total: **8158 MW**
(2015)



(2015)

Thermal Power Station

- Jorf Lasfar : 1356 MW
- Mohammedia : 600 MW
- Thahaddart : 384 MW
- Al wahda : 800 MW
- Ain Beni Mathar : 470 MW

Wind Power Station

- 1- Akhfenir & Tarfaya : 501 MW
- 2-Cape Sim : 60 MW
- 3-Tangier : 140 MW

Hydropower Station

- 9-Al Wahda Dam : 240 MW
- 10-El Borj & Tanafnit : 40 MW
- 11-Afourer : 465 MW
- 12- Al massira : 128 MW
- 13- Allal al fasi & Idriss1 : 280 MW
- 14- Bin el ouidane : 135 MW
- 15-Hassan I : 67 MW
- 16-MV : 23 MW

Solar power station

- 17 NOOR 1 : 160 MW

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

“Even ‘Safe’ Pollution Levels Can Be Deadly”

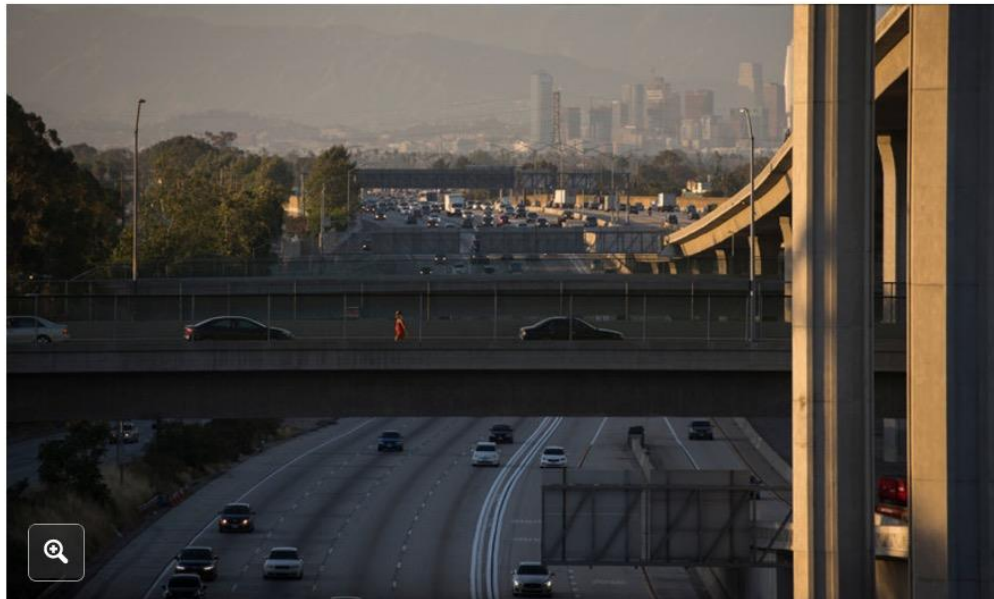
The New York Times

WELL

Even ‘Safe’ Pollution Levels Can Be Deadly

[Leer en español](#)

By NICHOLAS BAKALAR JUNE 28, 2017



Melissa Lyttle for The New York Times

“Even ‘Safe’ Pollution Levels Can Be Deadly”

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 29, 2017

VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

ABSTRACT

BACKGROUND

Studies have shown that long-term exposure to air pollution increases mortality. However, evidence is limited for air-pollution levels below the most recent National Ambient Air Quality Standards. Previous studies involved predominantly urban populations and did not have the statistical power to estimate the health effects in underrepresented groups.

METHODS

We constructed an open cohort of all Medicare beneficiaries (60,925,443 persons) in the continental United States from the years 2000 through 2012, with 460,310,521 person-years of follow-up. Annual averages of fine particulate matter (particles with a mass median aerodynamic diameter of less than 2.5 μm [$\text{PM}_{2.5}$]) and ozone were estimated according to the ZIP Code of residence for each enrollee with the use of previously validated prediction models. We estimated the risk of death associated with exposure to increases of 10 μg per cubic meter for $\text{PM}_{2.5}$ and 10 parts per billion (ppb) for ozone using a two-pollutant Cox proportional-hazards model that controlled for demographic characteristics, Medicaid eligibility, and area-level covariates.

From the Departments of Environmental Health (Q.D., Yan Wang, A.Z., P.K., J.D.S.) and Biostatistics (Yun Wang, C.C., F.D.), Harvard T.H. Chan School of Public Health, Boston. Address reprint requests to Dr. Dominici at Harvard T.H. Chan School of Public Health, Biostatistics Department, Bldg. 2, 4th Flr., 655 Huntington Ave., Boston, MA 02115, or at fdominic@hsph.harvard.edu.

N Engl J Med 2017;376:2513-22.

DOI: 10.1056/NEJMoa1702747

Copyright © 2017 Massachusetts Medical Society.

Health Policy

The NEW ENGLAND
JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 29, 2017

VOL. 376 NO. 26

Main conclusion

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

“In the entire Medicare population, there was significant evidence of adverse effects related to exposure to PM_{2.5} and ozone at concentrations below current national standards”

“Increases of 10 µg per cubic meter in PM_{2.5} ... associated with increase in all-cause mortality of 7.3% (7.1 to 7.5)”

Health Policy



Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

- **Mortality**

- beneficiaries of Medicare 2000 - 2012 (65+, US) – 61m persons
- age, sex, race, ZIP code, Medicaid status, date of death (censored)

Health Policy

The NEW ENGLAND
JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 29, 2017

VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

- **Mortality**

- beneficiaries of Medicare 2000 - 2012 (65+, US) – 61m persons
- age, sex, race, ZIP code, Medicaid status, date of death (censored)

- **Exposure**

- predicted annual average $PM_{2.5}$ for each ZIP code, using a neural network incorporating satellite, land-use, meteorological, simulation from chemical transport model

Health Policy

The NEW ENGLAND
JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 29, 2017

VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

- **Mortality**

- beneficiaries of Medicare 2000 - 2012 (65+, US) – 61m persons
- age, sex, race, ZIP code, Medicaid status, date of death (censored)

- **Exposure**

- predicted annual average $PM_{2.5}$ for each ZIP code, using a neural network incorporating satellite, land-use, meteorological, simulation from chemical transport model

- **Analysis**

- Cox-type regression analysis, with adjustment for spatial covariance

Lee et al 1992

- Cox mixed-effect analysis

random intercept location

Health Policy

The NEW ENGLAND
JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 29, 2017

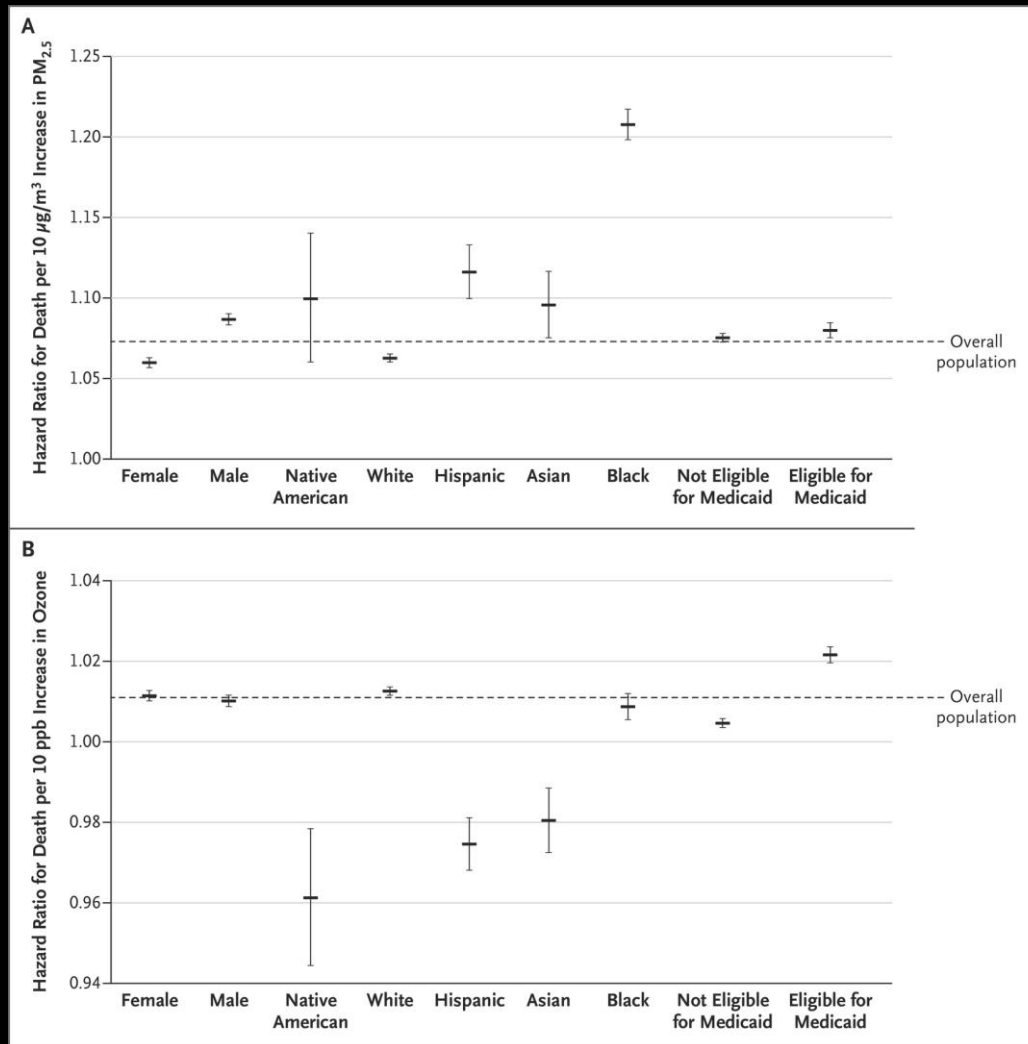
VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

- **Data Sources**
 - Behavioural Risk Factor Surveillance System -- confounders
 - US Census – zip code level
 - American Community Survey – zip code level
 - Dartmouth Atlas of Health Care – hospital level
 - Medicare Current Beneficiary Survey -- confounders
 - EPA Air Quality System – pollution
 - North American Regional Reanalysis -- temperature, humidity
- 22m deaths, 65m persons, 460m person-years
- “these data provided excellent power to estimate the risk of death at levels below the current [standards]” 12 μ g

Risk of Death Associated with an Increase of 10 μg per Cubic Meter in $\text{PM}_{2.5}$ Concentrations and an Increase of 10 ppb in Ozone Exposure, According to Study Subgroups.



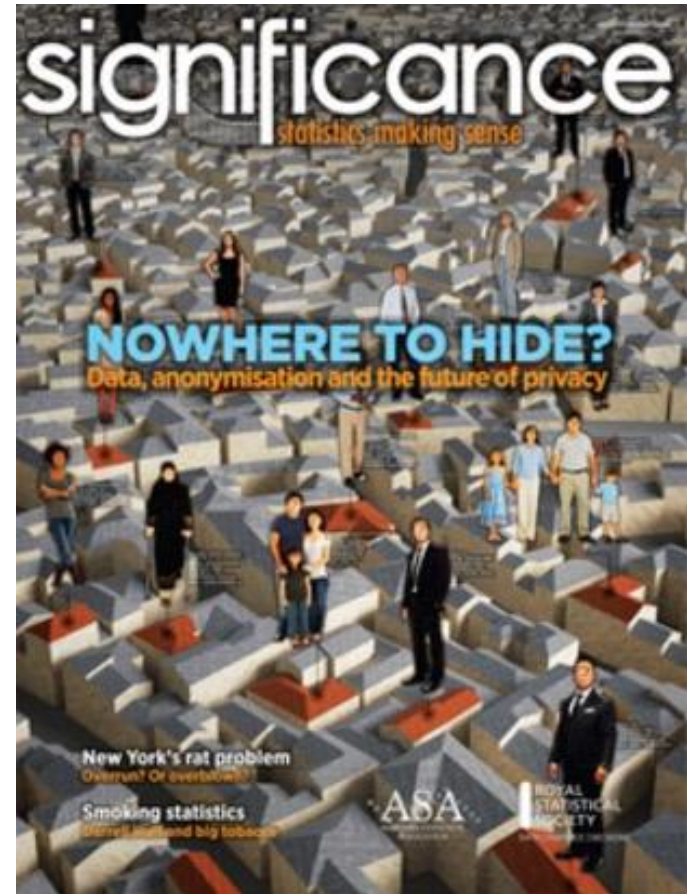
Di Q et al. N Engl J Med 2017;376:2513-2522



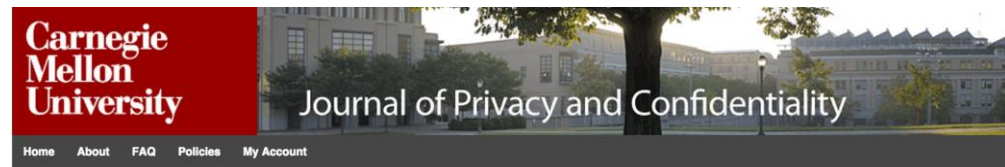
Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights



- Social Policy



Privacy

- “Big Data and Innovation, Setting the Record Straight: De-identification *Does Work*”

[Privacy Commissioner of Ontario, July 2014](#)

Privacy

- “Big Data and Innovation, Setting the Record Straight: De-identification *Does Work*”

[Privacy Commissioner of Ontario, July 2014](#)

- “No silver bullet: De-identification still doesn’t work”

[Narayan & Felten, July 2014](#)

Privacy

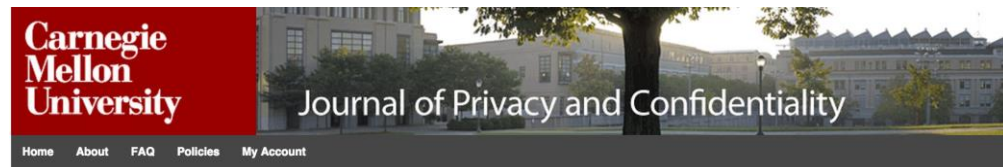
- “Big Data and Innovation, Setting the Record Straight: De-identification *Does Work*”

[Privacy Commissioner of Ontario, July 2014](#)

- “No silver bullet: De-identification still doesn’t work”

[Narayan & Felten, July 2014](#)

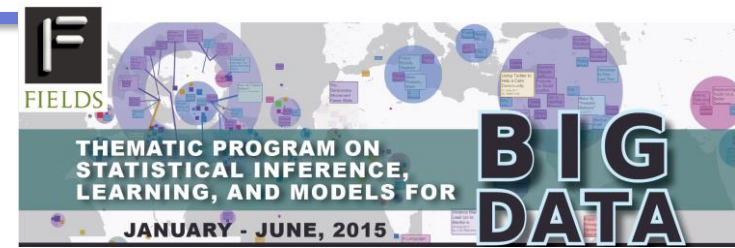
- **Statistical Disclosure Limitation**
- Differential Privacy
- Multi-party Communication



Some highlights

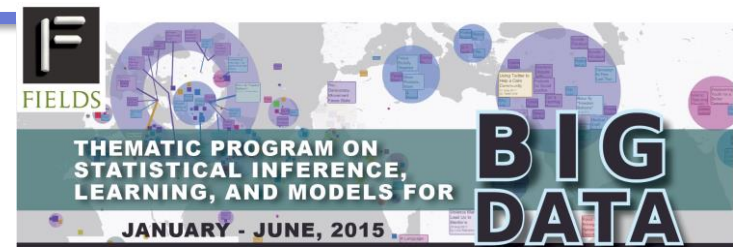
- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy
- inference, environmental science, networks, genomics, finance, physical sciences, software infrastructure, ...

What did we learn?



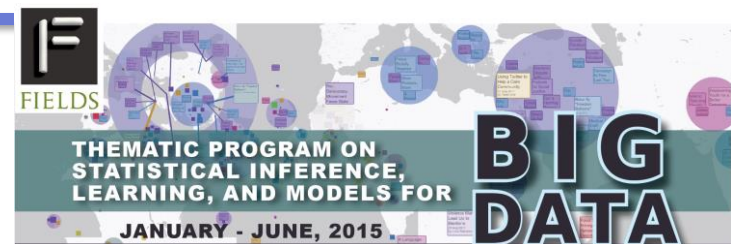
- Statistical models for big data are complex,
high-dimensional
 - inference is well-studied, but difficult

What did we learn?



- Statistical models for big data are complex, high-dimensional
 - inference is well-studied, but difficult
- Computational challenges include size and speed
 - ideas of statistical inference get lost in the machine

What did we learn?



- Statistical models for big data are complex, high-dimensional
 - inference is well-studied, but difficult
- Computational challenges include size and speed
 - ideas of statistical inference get lost in the machine
- Data owners understand 2., but not 1.
- Data modellers understand 1., but not 2.
- **Data science** may be the best way to combine these

That was yesterday

That was yesterday

Data science programs “springing up like mushrooms after rain”

HARVARDgazette

SCIENCE & HEALTH > ENGINEERING & TECHNOLOGY

Harvard launches data science initiative

Francesca Dominici and David Parkes named co-directors

March 28, 2017 | ✓ III







Berkeley, Hopkins, CMU, Washington, UBC, Yale, Toronto, ...

That was yesterday

Data science programs “springing up like mushrooms after rain”

68 Data Science & Big Data Master's degrees in United Kingdom

<p>M.B.A.</p> <h3>Master of Business Administration - Data</h3> <p>In this day and age, data is very easy to gather and store; it's knowing what to do with it that presents an obstacle. Studies...</p> <p> Nottingham Trent University ...</p>	<p>M.Sc.</p> <h3>Data Science for Business</h3> <p>This new Data Science for Business programme at the University of Stirling is the first in Scotland to be run in...</p> <p> University of Stirling Stirling Management School</p>	<p>M.Sc.</p> <h3>Big Data</h3> <p>This highly specialist program focuses on the growing importance of making use of Big Data technologies within...</p> <p> University of the West of Scotl... School of Engineering and Comp...</p>	<p>M.Sc.</p> <h3>Big Data and Text Analytics</h3> <p>The sheer volume of this information means that traditional stand-alone applications are no longer...</p> <p> University of Essex School of Computer Science and...</p>
--	--	--	--

Load more

That was yesterday

Data science programs “springing up like mushrooms after rain”

Data Science Africa 2017

Arusha, Tanzania

The last few years have witnessed an explosion in the quantity and variety of data available in Africa, produced either as a by-product of digital services, from sensors or measuring devices, satellites and from many other sources. A number of practical fields have been transformed by the ability to collect large volumes of data: for example, bioinformatics with the development of high throughput sequencing technology capable of measuring gene expression in cells, or agriculture with the widespread availability of high quality remote sensing data. For other data sources – such as mobile phone usage records from telecoms operators, which can be used to measure population movement and economic activity – we are just beginning to understand the practical possibilities.

Data science seeks to exploit advances in machine learning and statistics to make sense of the growing amounts of data available from various sources. In Africa, a number of problems in areas such as healthcare, agriculture, disaster response and wildlife conservation would benefit greatly if domain experts were exposed to data science techniques. These skills would allow practitioners to extract useful information from these abundant sources of raw data

Summer School on Machine Learning and Data Science

Dates: 17 July - 19 July 2017

Venue: Nelson Mandela African Institute of Science and Technology, Tanzania

In the tradition of previous Africa Data Science workshops, a summer school on machine learning and data science will be held prior to the main workshop. This summer school will target graduate students, researchers and professionals working with huge amounts of data or unique datasets.

That was yesterday

Data science programs “springing up like mushrooms after rain”

Data Science Africa 2017

Arusha, Tanzania

The last few years have witnessed an explosion in the quantity and variety of data available in Africa, produced either as a by-product of digital services, from sensors or measuring devices, satellites and from many other sources. A number of practical fields have been transformed by the ability to collect large volumes of data: for example, bioinformatics with the development of high throughput sequencing technology capable of measuring gene expression in cells, or agriculture with the widespread availability of high quality remote sensing data. For other data sources – such as mobile phone usage records from telecoms operators, which can be used to measure population movement and economic activity – we are just beginning to understand the practical possibilities.

Data science seeks to exploit advances in machine learning and statistics to make sense of the growing amounts of data available from various sources. In Africa, a number of problems in areas such as healthcare, agriculture, disaster response and wildlife conservation would benefit greatly if domain experts were exposed to data science techniques. These skills would allow practitioners to extract useful information from these abundant sources of raw data

Summer School on Machine Learning and Data Science

Dates: 17 July - 19 July 2017

Venue: Nelson Mandela African Institute of Science and Technology, Tanzania

In the tradition of previous Africa Data Science workshops, a summer school on machine learning and data science will be held prior to the main workshop. This summer school will target graduate students, researchers and professionals working with huge amounts of data or unique datasets.

That was yesterday

Data science programs “springing up like mushrooms after rain”



What is data science?

- a course?
- a set of courses?

[Data 8](#) [Weekly Schedule](#) [Course Info](#) [Connector Courses](#) [Staff](#) [Python Help](#) ▾

Data 8: Foundations of Data Science

Fall 2016

Instructor: Ani Adhikari

University of Toronto **New Undergraduate Program Proposal**

(This template has been developed in line with the University of Toronto's Quality Assurance Process.)

What is data science?

- a course?
- a set of courses?
- a job?
- a technology?

[Data 8](#) [Weekly Schedule](#) [Course Info](#) [Connector Courses](#) [Staff](#) [Python Help](#) ▾

Data 8: Foundations of Data Science

Fall 2016
Instructor: Ani Adhikari

University of Toronto New Undergraduate Program Proposal

(This template has been developed in line with the University of Toronto's Quality Assurance Process.)

LEARN DATA SCIENCE IN YOUR BROWSER

What is data science?

- a course?
- a set of courses?
- a job?
- a technology?
- a new field of research?
- a collaboration?

Data 8: Foundations of Data Science

Fall 2016
Instructor: Ani Adhikari

University of Toronto New Undergraduate Program Proposal

(This template has been developed in line with the University of Toronto's Quality Assurance Process.)

LEARN DATA SCIENCE IN YOUR BROWSER

Data Science Institute

What is data science?

- a course?
- a set of courses?
- a job?
- a technology?
- a new field of research?
- a collaboration?

Data 8: Foundations of Data Science

Fall 2016
Instructor: Ani Adhikari

University of Toronto New Undergraduate Program Proposal

(This template has been developed in line with the University of Toronto's Quality Assurance Process.)

LEARN DATA SCIENCE IN YOUR BROWSER

Français | English



Paris-Saclay
Center for Data Science

What is data science?

- a course?
- a set of courses?
- a job?
- a technology?
- a new field of research?
- a collaboration?

Data 8 Weekly Schedule Course Info Connector Courses Staff Python Help ▾

Data 8: Foundations of Data Science

Fall 2016
Instructor: Ani Adhikari

University of Toronto New Undergraduate Program Proposal

(This template has been developed in line with the University of Toronto's Quality Assurance Process.)

LEARN DATA SCIENCE IN YOUR BROWSER

THE ALAN
TURING
INSTITUTE

Data Science Program(s)

JHU DSS

- mathematical reasoning
- statistical theory
- statistical and machine learning methods
- programming and software development
- algorithms and data structure
- communication results and limitations

Good Enough Practices in Scientific Computing

Greg Wilson^{1,‡*}, Jennifer Bryan^{2,‡}, Karen Cranston^{3,‡}, Justin Kitzes^{4,‡},
Lex Nederbragt^{5,‡}, Tracy K. Teal^{6,‡}

1 Software Carpentry Foundation / gwwilson@software-carpentry.org

2 University of British Columbia / jenny@stat.ubc.ca

3 Duke University / karen.cranston@duke.edu

4 University of California, Berkeley / jkitzes@berkeley.edu

5 University of Oslo / lex.nederbragt@ibv.uio.no

6 Data Carpentry / tkteal@datacarpentry.org

‡ These authors contributed equally to this work.

* E-mail: Corresponding gwwilson@software-carpentry.org

Data Science Research



- data collection and data quality
- large N , small p
 - computational strategies, e.g. Spark, Hadoop
 - divide and conquer
- small n , large p
 - inferential and computational strategies
 - dimension reduction
 - post-selection inference
 - inference for extremes
- ‘new’ types of data: networks, graphs, text, images, ...
 - “alternative sources”

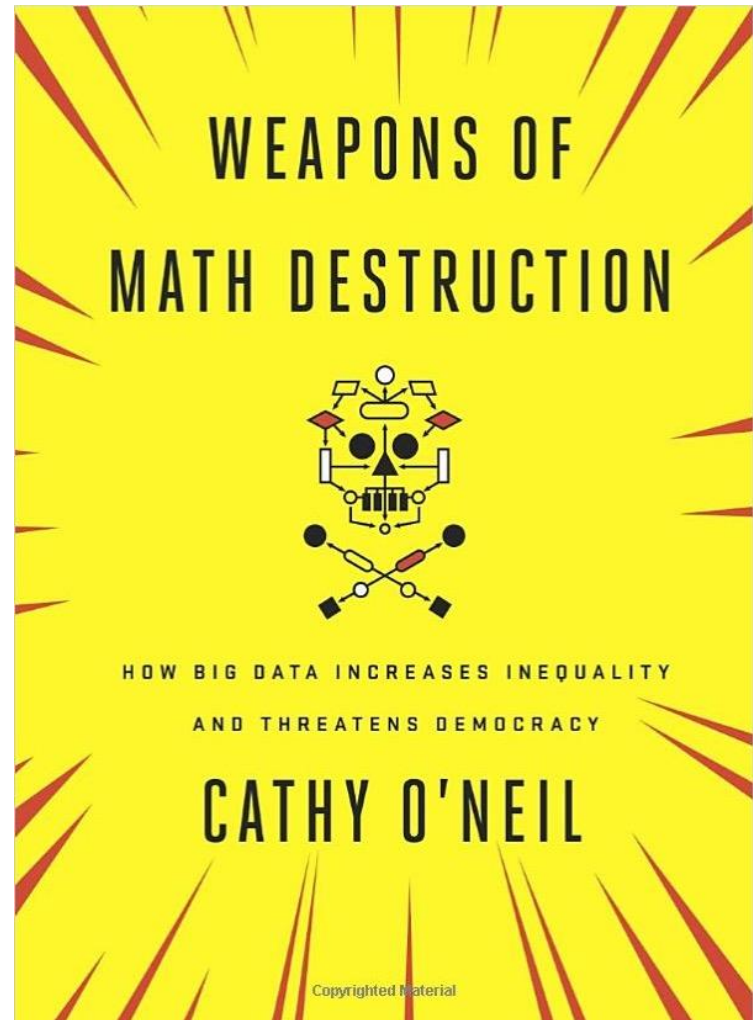
... Data Science Research

Leek 2017

- collaboration and communication
- data wrangling, database development, record linkage, privacy
- replicability, reproducibility, new workflows
- visualization
- outside the ivory tower -- industry, government, media, public

<https://simplystatistics.org/2017/07/19/my-unfunded-hhmi-teaching-professors-proposal/>

The push back



How big data threatens democracy and increases inequality

The push back

Big data
The Guardian's
Science Weekly

🔊 Weapons of math destruction: how big data and algorithms affect our lives - podcast

WS More or Less: Algorithms, Crime and Punishment

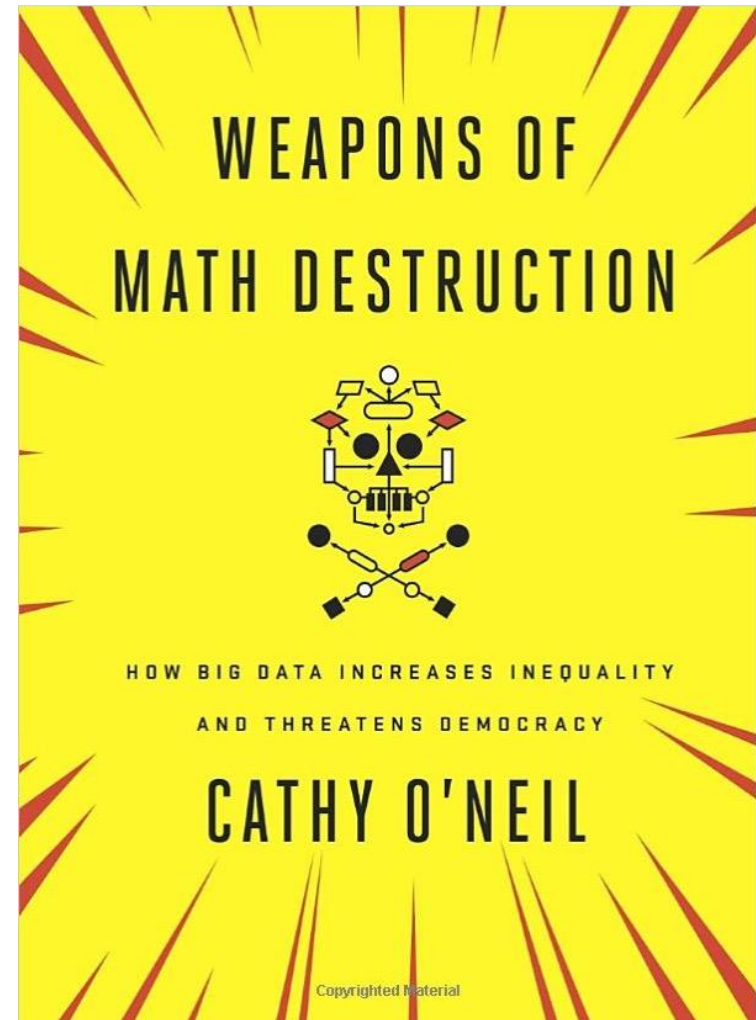
When maths can get you locked up.

Available now

🕒 9 minutes



Download MP3



How big data threatens democracy and increases inequality

The push back

Big data
The Guardian's
Science Weekly

🔊 Weapons of math destruction: how big data and algorithms affect our lives - podcast

WS More or Less: Algorithms, Crime and Punishment

When maths can get you locked up.

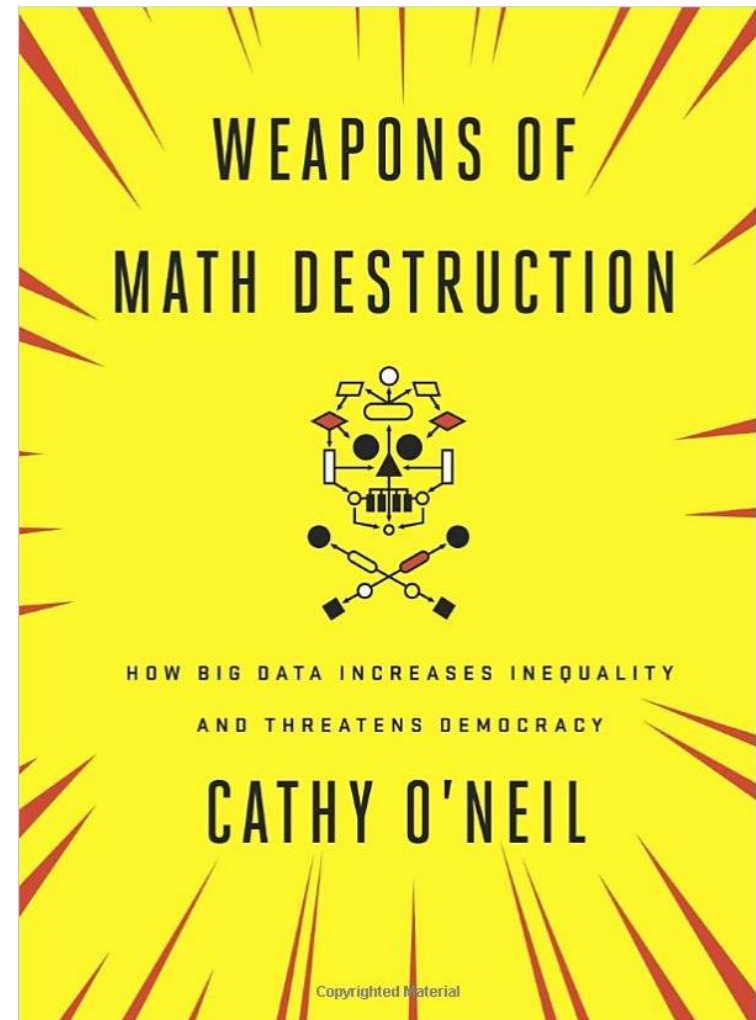
Available now

🕒 9 minutes



Download MP3

“if the assessment never asks about race, how could the algorithm throw up racially biased results?”



How big data threatens democracy and increases inequality

The push back

Big data

The Guardian's
Science Weekly

🔊 Weapons of math destruction: how big data and algorithms affect our lives - podcast

WS More or Less: Algorithms, Crime and Punishment

When maths can get you locked up.

Available now

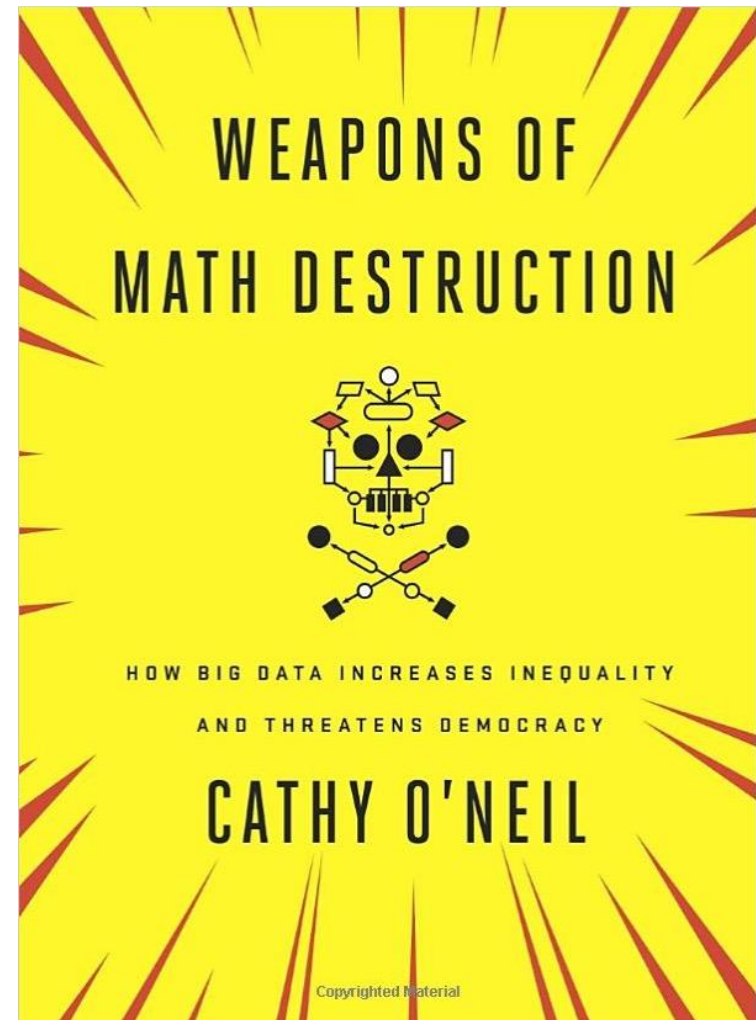
🕒 9 minutes



Download MP3

“if the assessment never asks about race, how could the algorithm throw up racially biased results?”

“Credit scores are used by nearly half of American employers to screen potential employees”



How big data threatens democracy and increases inequality

The push back

Big data in social sciences: a promise betrayed ?

Posted on March 22, 2017

In just 5 years, the mood at conferences on social science and big data has shifted, at least in France. Back in the early 2010s, these venues were buzzing with exchanges about the characteristics of the “revolution” ([the 4Vs](#)) with participants marveling at the research insights afforded by the use of tweets, website ratings, Facebook likes, Ebay prices or

“Big data is neither easier nor faster nor cheaper”

“Building a database doesn’t create its own uses”

Privacy



NATURE | COLUMN: WORLD VIEW



Cliver Sherlock

The DeepMind debacle demands dialogue on data

Mishandling of patient information shows how governments and companies must become more worthy of trust, says [Hetan Shah](#).

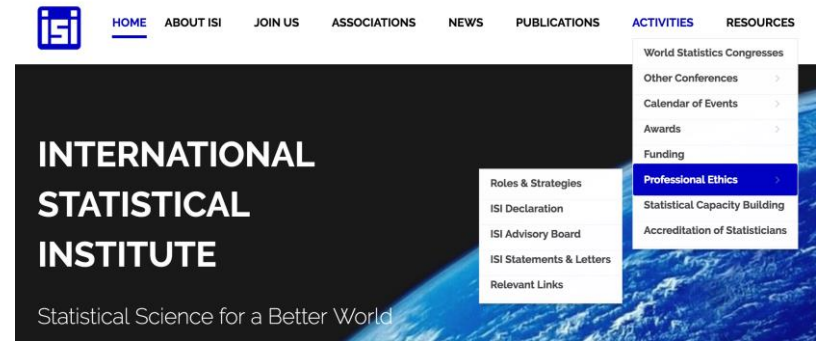
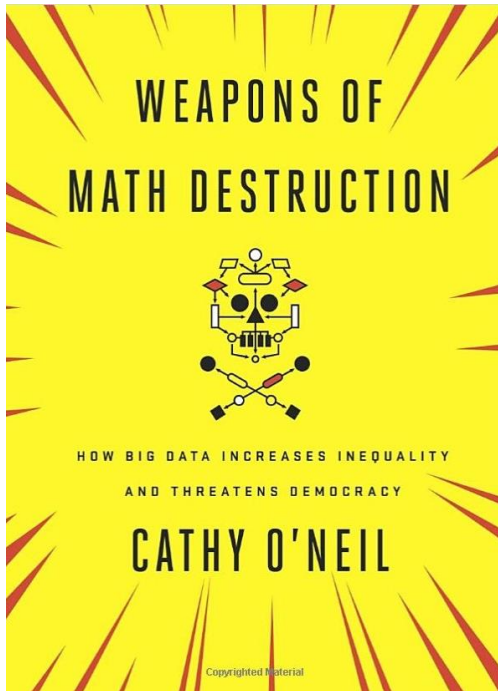
19 July 2017

Privacy



- DeepMind was to develop an app to check test results for signs of acute kidney injuries
- the arrangement failed to consider how patients expect their data to be used, and by whom
- had the project proceeded under open contracting, it would have been subject to public scrutiny
- it is unclear why an app for kidney injury requires the identifiable records of every patient seen by three hospitals over a five year period

Caution can be a good thing



“Digital Hippocratic Oath”

Caution can be a good thing

Guardian 2 July 2016

“...from data we will get the cure for cancer as well as better hospitals;

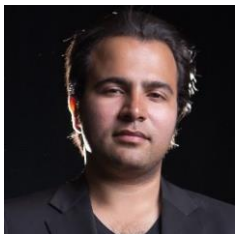
schools that adapt to children’s needs making them happier and smarter;

better policing and safer homes;

and of course jobs.”

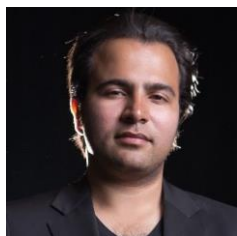
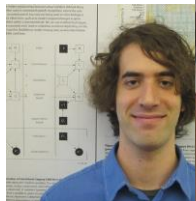


Thank You!





Thank You!



Intern

St
M





Thank You!



isi International Statistical Review

International Statistical Review (2016), 84, 3, 371–389 doi:10.1111/insr.12176

Statistical Inference, Learning and Models in Big Data