Một số nền tảng của khoa học dữ liệu

$$data\ science = \int_{\Omega} data \cdot (math \wedge stat \wedge cs \wedge ...)?$$

Nguyễn Xuân Long

Department of Statistics and Dept of EECS
Michigan Institute for Data Science
University of Michigan, Ann Arbor

Vietnam Institute for Advanced Study in Mathematics
Hanoi and HCMC, 5/2017

Outline

Summer School 2012: Modern statistical methods in machine learning

VIASM School on Statistical machine learning (2015)

- Picture taken with students from HCMC
- Lectures on Bayesian nonparametrics

Chủ đề năm nay là Data Science.
Điều gì mới?

(phạm vi, xu thế)

(vai trò của công nghệ và toán học)

(những bất biến)

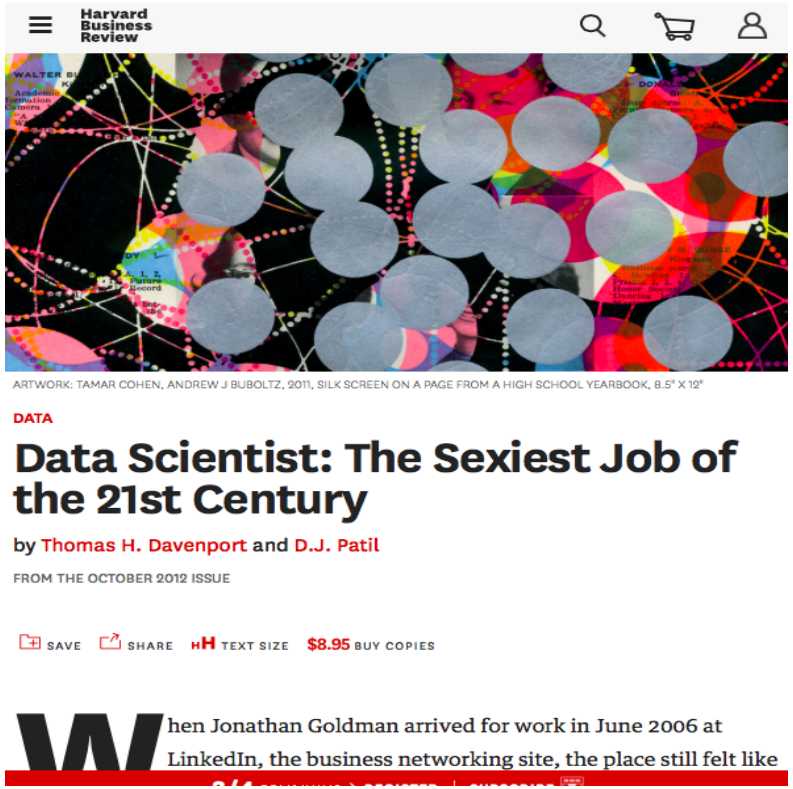
(Berkeley, 2003)

Trong khi đó ...

2009: Hal Varian, chief economist at Google, famously said: "I keep saying that the sexy job in the next 10 years will be statisticians" – New York Times

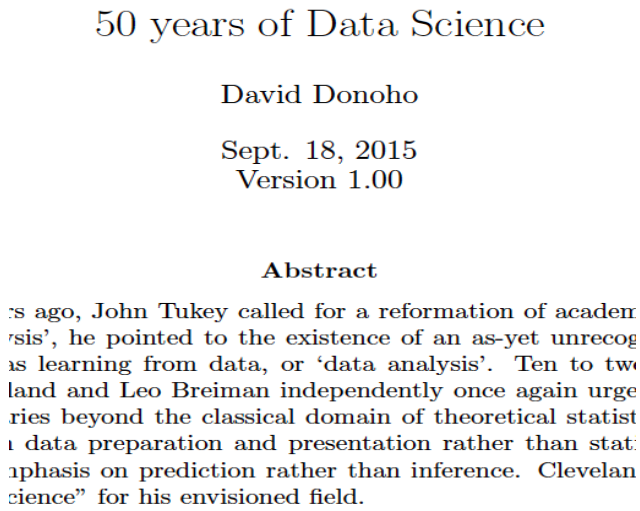
2012: a Harvard Business Review's article thinks of data scientists

2017: continued rosy job forecast for experts in data science, statistics, machine learning, artificial intelligence



Institutional changes

David Donoho là một nhà thống kê nổi tiếng của ĐH Stanford (wavelet, compressed sensing). Ông ấy đã viết về "Data Science Moment", trong một bài báo gây xôn xao dư luận hơn một năm trước đây.



50 years of Data Science

David Donoho

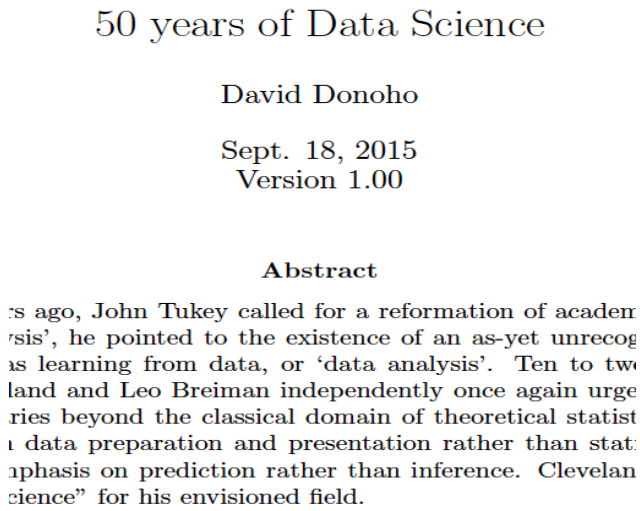
Sept. 18, 2015
Version 1.00

Abstract

:s ago, John Tukey called for a reformation of academ
'sis', he pointed to the existence of an as-yet unrecog
as learning from data, or 'data analysis'. Ten to tw
land and Leo Breiman independently once again urge
ries beyond the classical domain of theoretical statist
data preparation and presentation rather than stat
nphasis on prediction rather than inference. Clevelan
cience" for his envisioned field.

Institutional changes

David Donoho là một nhà thống kê nổi tiếng của ĐH Stanford (wavelet, compressed sensing). Ông ấy đã viết về "Data Science Moment", trong một bài báo gây xôn xao dư luận hơn một năm trước đây.



50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

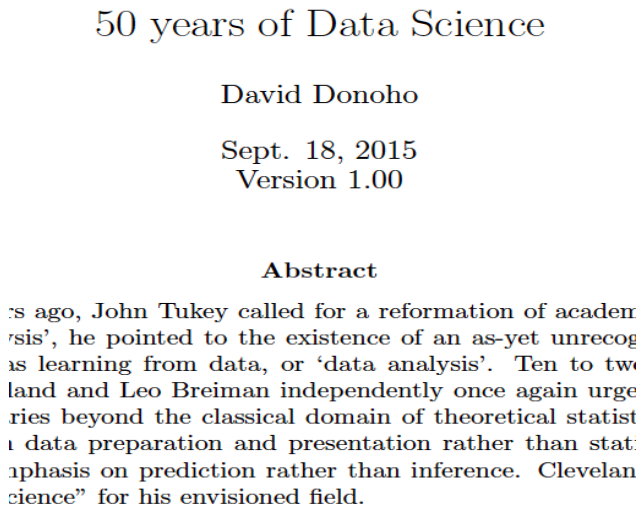
Abstract

:s ago, John Tukey called for a reformation of academ
'sis', he pointed to the existence of an as-yet unrecog
as learning from data, or 'data analysis'. Ten to tw
land and Leo Breiman independently once again urge
ries beyond the classical domain of theoretical statist
data preparation and presentation rather than stat
nphasis on prediction rather than inference. Clevelan
cience" for his envisioned field.

Khoảnh khắc gì vậy?

Institutional changes

David Donoho là một nhà thống kê nổi tiếng của ĐH Stanford (wavelet, compressed sensing). Ông ấy đã viết về "Data Science Moment", trong một bài báo gây xôn xao dư luận hơn một năm trước đây.



50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

Abstract

:s ago, John Tukey called for a reformation of academ
'sis', he pointed to the existence of an as-yet unrecog
as learning from data, or 'data analysis'. Ten to tw
land and Leo Breiman independently once again urge
ries beyond the classical domain of theoretical statist
data preparation and presentation rather than stat
phasis on prediction rather than inference. Clevelan
cience" for his envisioned field.

Khoảnh khắc gì vậy?

"... Tháng 9 năm 2015, ĐH Michigan vừa đưa ra thông báo về một chương trình "Data Science Initiative" (DSI) trị giá 100 triệu USD:

"Data science has become a fourth approach to scientific discovery, in addition to experimentation, modeling, and computation"
— Provost Martha Pollack.

"...Trang web của Chương trình DSI cho ta (GS Donoho) biết một bức tranh của KHDL:

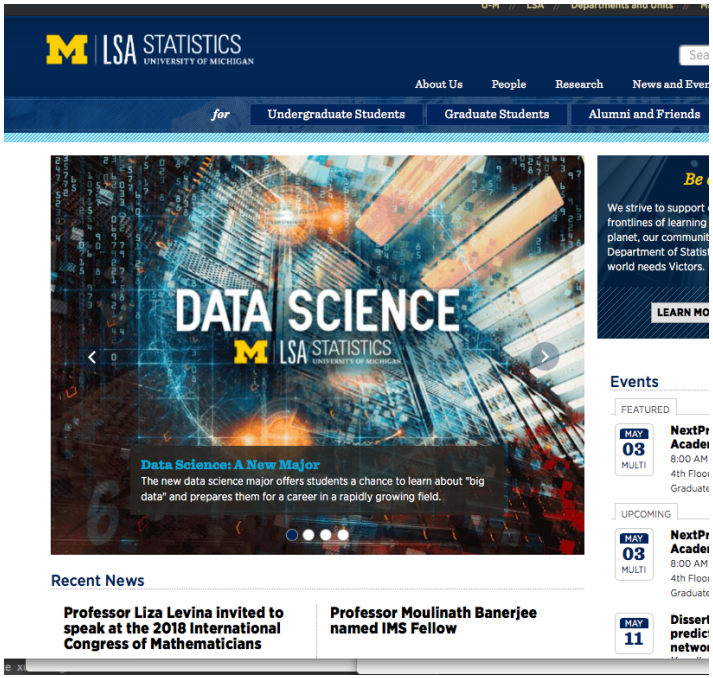
This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.

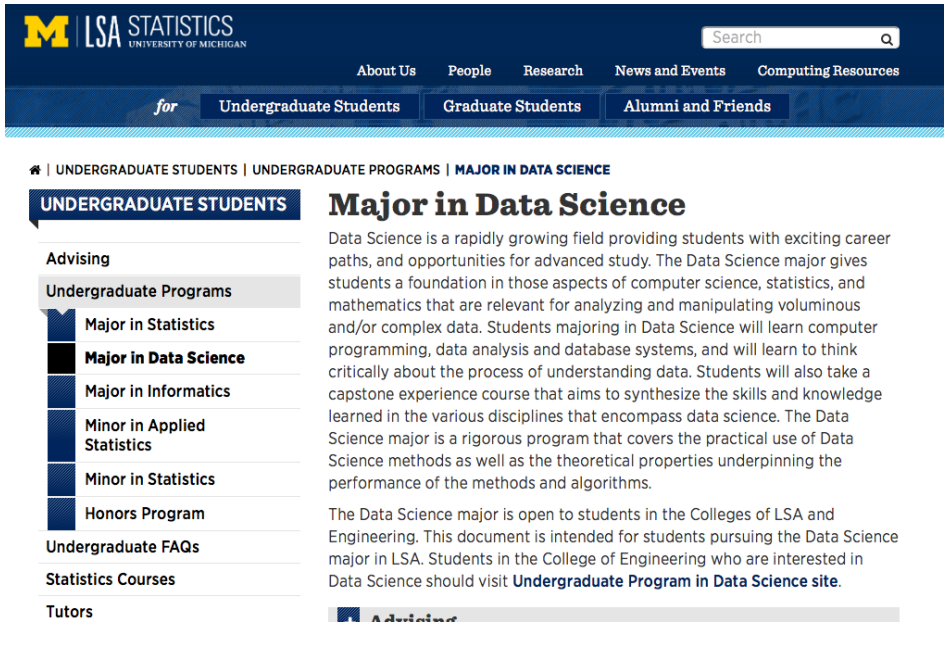
More university initiatives and degrees in US

- similar initiatives at NYU, Columbia, MIT, UC Berkeley,...

- new undergraduate majors and Master's program in Data Science mushrooming in major universities, including Columbia, Carnegie Mellon University, Cornell, New York University, Northwestern U, Stanford, UC Berkeley, U of Michigan, U of Illinois, U of Wisconsin,...

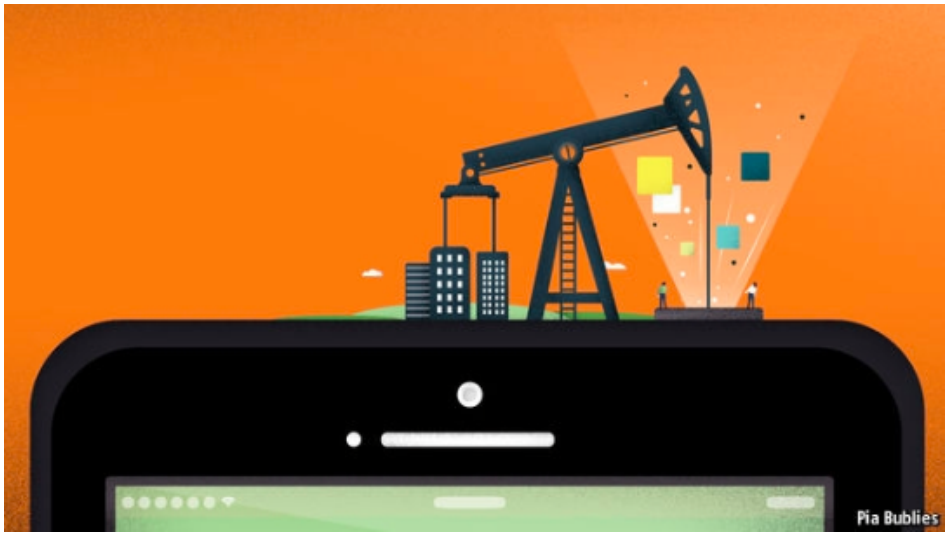
U Michigan's Statistics Department Website





"Data Science is a rapidly growing field providing students with exciting career paths, and opportunities for advanced study. The Data Science major gives students **a foundation in those aspects of computer science, statistics, and mathematics** that are relevant for analyzing and manipulating voluminous and/or complex data. Students majoring in Data Science will learn computer programming, data analysis and database systems, and will learn to think critically about the process of understanding data..."

data, data and data



- "data is fuel of the future" — Economist article, May 6, 2017
  - it is becoming so easy to collect and generate data, anytime, anywhere
- it is also becoming easy to process, even manipulate, data
- powerful computing helps handling bigger and more complex data
- more data, more needs and ambitions

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients
- astronomers gain new understanding of cosmos from signals collected by improved telescope and communication technologies

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients
- astronomers gain new understanding of cosmos from signals collected by improved telescope and communication technologies
- engineers make *automated car driving* possible by utilizing more sensory data types

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients
- astronomers gain new understanding of cosmos from signals collected by improved telescope and communication technologies
- engineers make *automated car driving* possible by utilizing more sensory data types
- bank analysts make more accurate prediction of loan performances, thereby earning more profits

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients
- astronomers gain new understanding of cosmos from signals collected by improved telescope and communication technologies
- engineers make *automated car driving* possible by utilizing more sensory data types
- bank analysts make more accurate prediction of loan performances, thereby earning more profits
- political operatives make more accurate voting pattern forecast and improve winning chance for candidate's campaign

data and ambitions

- molecular biologists gain better understanding of the relationship between gene expression levels and cancer tumors, leading to new treatments for cancer patients
- astronomers gain new understanding of cosmos from signals collected by improved telescope and communication technologies
- engineers make *automated car driving* possible by utilizing more sensory data types
- bank analysts make more accurate prediction of loan performances, thereby earning more profits
- political operatives make more accurate voting pattern forecast and improve winning chance for candidate's campaign
- top companies such as Google, Facebook, Amazon,... make money by learning about and exploiting our Internet browsing patterns

is data science for real?

- understood differently by different communities
  (business, industry, researchers, academic scientists)
- all agreed on its growing roles and vast impacts on all facets of society
- much hypes are commercially driven
- but an intellectual core of data science is also emerging

opportunities and risks

- trong thời đại "big data", lắm cơ hội song cũng nhiều cạm bẫy
  - dữ liệu có thể đánh lừa
  - dữ liệu thu thập được dễ dàng cũng dễ làm ta lười đi: lười trong suy nghĩ, hời hợt trong cách thức tổ chức và cẩu thả trong phân tích

opportunities and risks

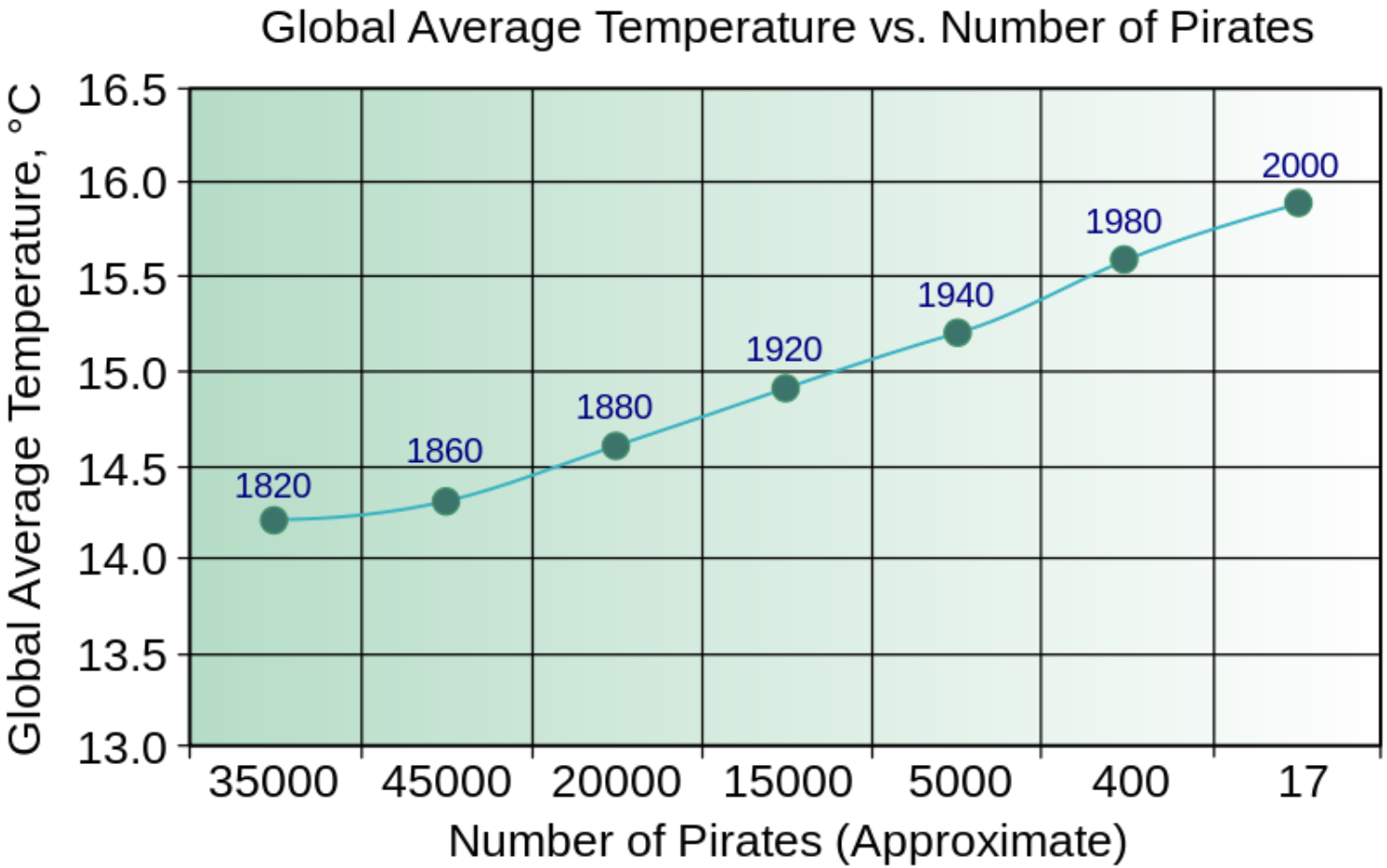
- trong thời đại "big data", lắm cơ hội song cũng nhiều cạm bẫy
  - dữ liệu có thể đánh lừa
  - dữ liệu thu thập được dễ dàng cũng dễ làm ta lười đi: lười trong suy nghĩ, hời hợt trong cách thức tổ chức và cẩu thả trong phân tích
- trong các địa hạt nghiên cứu: dễ bị loá mắt bởi những phương pháp nóng và đang hợp thời thượng
  - để thành công bền vững, phải đi vào những vấn đề nền tảng

opportunities and risks

- trong thời đại "big data", lắm cơ hội song cũng nhiều cạm bẫy
  - dữ liệu có thể đánh lừa
  - dữ liệu thu thập được dễ dàng cũng dễ làm ta lười đi: lười trong suy nghĩ, hời hợt trong cách thức tổ chức và cẩu thả trong phân tích
- trong các địa hạt nghiên cứu: dễ bị loá mắt bởi những phương pháp nóng và đang hợp thời thượng
  - để thành công bền vững, phải đi vào những vấn đề nền tảng
- Điều này có nghĩa: phải xây dựng trên nền tảng vững chắc của toán, thống kê, khoa học máy tính, nếu ta muốn phát triển KHDL một cách vững vàng

"lies, damned lies and statistics" – Mark Twain

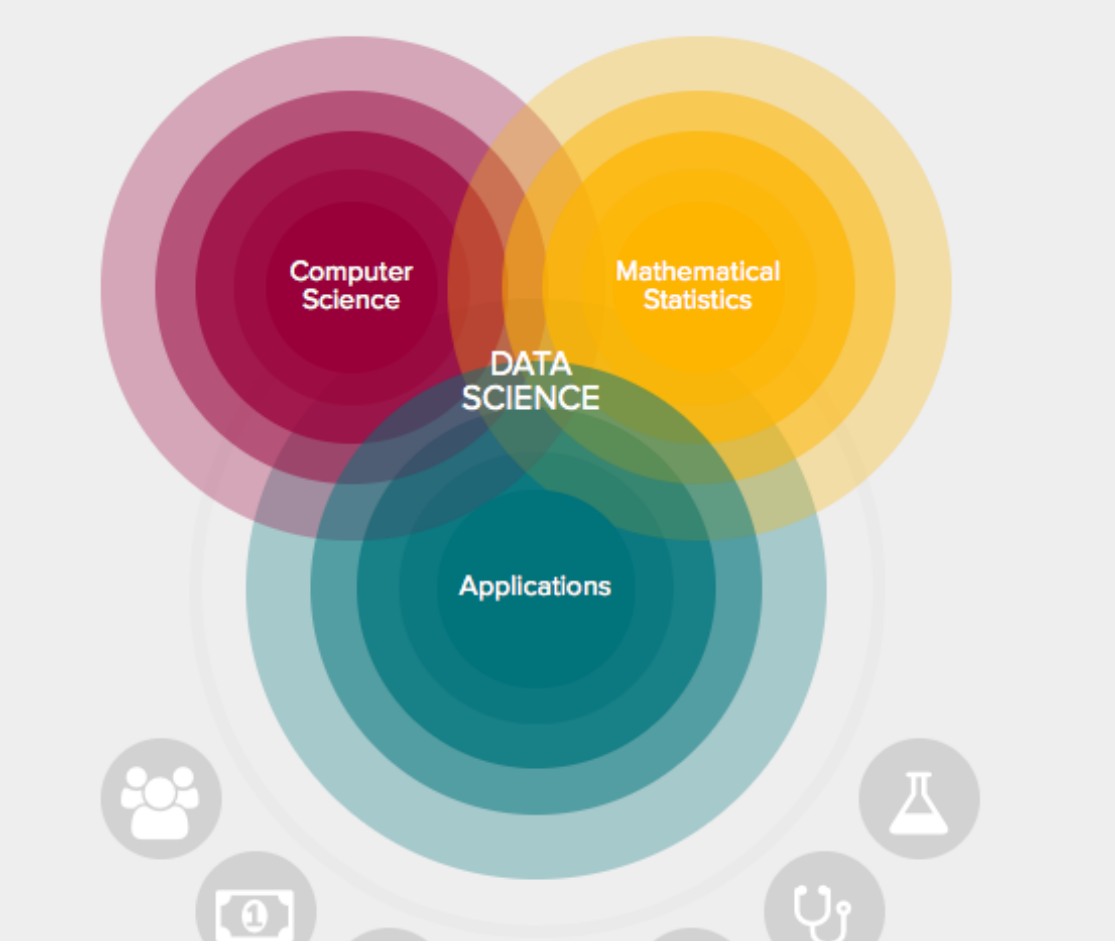
This small data set is scientific (not fake), but what can we make of it?



Big data give bigger room for propaganda and fake news

Outline

the major actors according to New York University's DSI Website



Skill set $\neq$ Core foundation

Skill set

A data scientist requires a large skill set that can be overwhelming to learn

- modern programming languages, from C/C++, R, Matlab, Python
- modern database techniques, from spreadsheets to SQL to ditributed databases and live data streams
- algorithms and models: from classical tools such as logistic regression, linear regression to support vector machines, graphical models, deep learning, Bayesian nonparametrics
- distributed architecture: Hadoop, Map/Reduce, SPARK, GPU

Skill set reflects Core foundation

Elements of the science

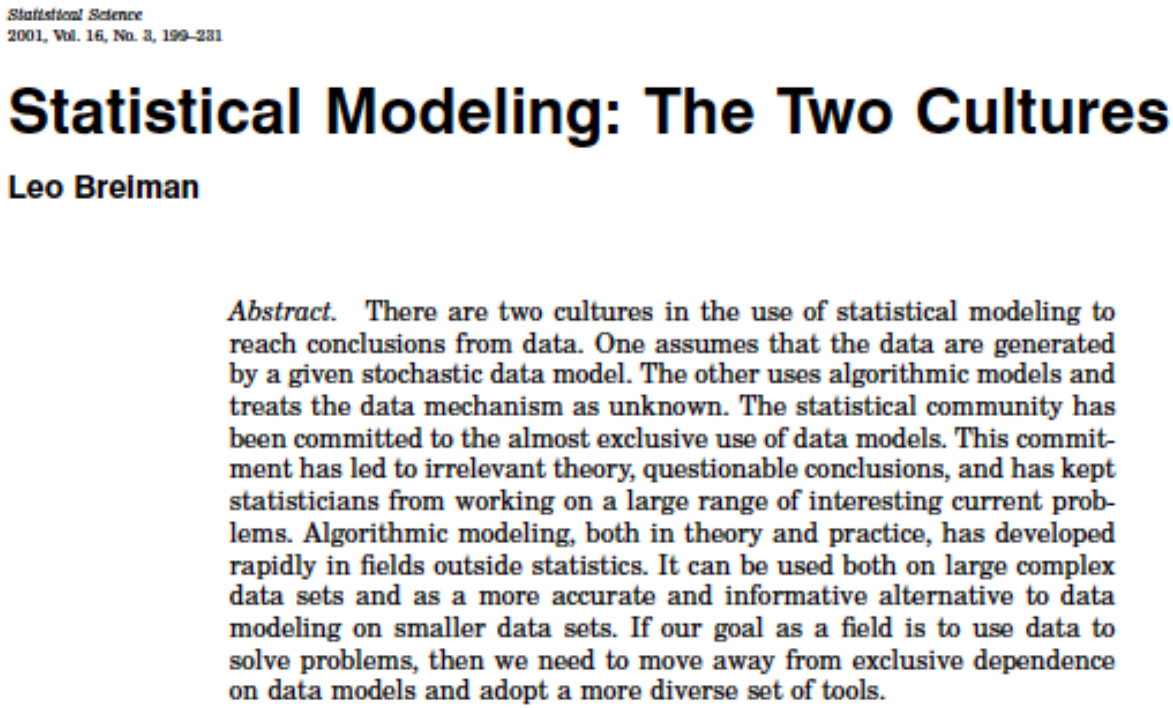
... the science of learning from data, with all that it entails

- elements of data representation
  - from data storage to abstract data structures and visualization
  - both in a mathematical sense and sense of a computer system
- elements of statistical learning and inference
  - from statistical modeling to prediction and inference
  - interplay of algorithmic and statistical efficiency
- elements offered by real-world data domains in sciences and industry
  - fresh challenges in real-world scale and complexity

Rest of lecture: Elements of statistical learning and inference

What do we want ~~to do with~~ *to draw from* data?

Leo Breiman là nhà thống kê/ học máy lỗi lạc của Berkeley, tác giả decision trees và random forests, viết về văn hoá trong mô hình suy diễn



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Có hai mục tiêu suy diễn/ học thống kê

- dự báo (to predict)
- thiết lập quy luật, hiểu biết về mối quan hệ (to infer/ understand)

Prediction

Mọi phương pháp dự báo đều bắt đầu từ dữ liệu X để ước lượng Y

$$X \longrightarrow \boxed{METHOD} \longrightarrow Y$$

Ví dụ: dữ liệu X về dung lượng giao thông của các địa điểm
Y là một giá trị nhị phân để dự báo đường có bị tắc hay không

Prediction

Mọi phương pháp dự báo đều bắt đầu từ dữ liệu X để ước lượng Y

$$X \longrightarrow \boxed{METHOD} \longrightarrow Y$$

Ví dụ: dữ liệu X về dung lượng giao thông của các địa điểm
Y là một giá trị nhị phân để dự báo đường có bị tắc hay không

Giả dụ: X và Y được liên hệ với nhau bởi một đại lượng chân lý nào đó kết nối X và Y, ta tạm gọi là θ

$$X \longleftrightarrow \theta \longleftrightarrow Y$$

Prediction

Mọi phương pháp dự báo đều bắt đầu từ dữ liệu X để ước lượng Y

$$X \longrightarrow \boxed{METHOD} \longrightarrow Y$$

Ví dụ: dữ liệu X về dung lượng giao thông của các địa điểm
Y là một giá trị nhị phân để dự báo đường có bị tắc hay không

Giả dụ: X và Y được liên hệ với nhau bởi một đại lượng chân lý nào đó kết nối X và Y, ta tạm gọi là θ

$$X \longleftrightarrow \theta \longleftrightarrow Y$$

Chúng ta không bao giờ biết được chính xác θ (chỉ có Trời mới biết!)

Prediction

Mọi phương pháp dự báo đều bắt đầu từ dữ liệu X để ước lượng Y

$$X \longrightarrow \boxed{METHOD} \longrightarrow Y$$

Ví dụ: dữ liệu X về dung lượng giao thông của các địa điểm
Y là một giá trị nhị phân để dự báo đường có bị tắc hay không

Giả dụ: X và Y được liên hệ với nhau bởi một đại lượng chân lý nào đó kết nối X và Y, ta tạm gọi là θ

$$X \longleftrightarrow \theta \longleftrightarrow Y$$

Chúng ta không bao giờ biết được chính xác θ (chỉ có Trời mới biết!)
Breiman nói đến hai cách tiếp cận khác nhau trong bài toán prediction

- Algorithmic Modeling
- Data Modeling

ALGORITHMIC MODELING không quan tâm lắm đến bản chất sự liên hệ giữa dung lượng giao thông với sự tắc đường mô tả bởi θ, nó chỉ muốn tìm ra một hàm dự báo $y = f(x)$

$$\{X, Y\} \Longrightarrow \boxed{\textit{ALGORITHM MODEL}} \Longrightarrow f$$

Chạy thuật toán $\boxed{\textit{blackbox}}$ như *decision tree, kernel machine, neural net* để học f, với mục tiêu sao cho $f(X)$ càng gần với Y càng tốt

Vậy f có liên quan gì đến θ?

DATA MODELING tìm cách thiết lập một chất keo toán học kết dính X với Y, thông qua một mô hình xác suất liên hệ biến ngẫu nhiên X và biến ngẫu nhiên Y

$$\{X, Y\} \longrightarrow \boxed{\textit{DATA MODEL}} \longrightarrow \mathbb{P}(Y|X, \hat{\theta})$$

$\hat{\theta}$ là một ước lượng của chân lý θ, thông qua đó ta thiết lập được liên hệ giữa Y và X thông qua phân bố điều kiện của Y khi cho X.

Phân bố này cho ta một hàm số dự báo gọi là optimal Bayes classifier

$$f(x) = 1 \text{ if } \mathbb{P}(Y|X, \hat{\theta}) > 1/2, \quad 0 \text{ otherwise}$$

Lý thuyết học thống kê đối với bài toán phân lớp cho biết, hai phương pháp chỉ là hai mặt của một vấn đề:

- DATA MODELING is good only if $\hat{\theta}$ approximates well the true θ
- ALGORITHMIC MODELING is good only if the outcome f approximates well the optimal Bayes classifier đối với true θ

Lý thuyết học thống kê đối với bài toán phân lớp cho biết, hai phương pháp chỉ là hai mặt của một vấn đề:

Data $\Rightarrow$ *DATA MODEL* $\Longleftrightarrow$ *ALGORITHM MODEL* $\Rightarrow$ Prediction function f

- DATA MODELING is good only if $\hat{\theta}$ approximates well the true θ
- ALGORITHMIC MODELING is good only if the outcome f approximates well the optimal Bayes classifier đối với true θ

- There is no silver bullet!
- Distinction between data model and algorithm model may disappear as both become richer, e.g.:
  - Bayesian nonparametric data model may be described algorithmically
  - Randomized algorithm entails a stochastic process (e.g., Markov chain)

Hai văn hoá

Machine Learning

- chú trọng đến mô hình thuật toán nhiều hơn mô hình dữ liệu
  (đặc biệt cho prediction/ supervised learning)
- cộng đồng học máy rất năng động: học một công cụ toán/ thống kê, rebranding, và sở hữu nó rất nhanh; chạy theo mốt với tốc độ chóng mặt

Hai văn hoá

Machine Learning

- chú trọng đến mô hình thuật toán nhiều hơn mô hình dữ liệu (đặc biệt cho prediction/ supervised learning)
- cộng đồng học máy rất năng động: học một công cụ toán/ thống kê, rebranding, và sở hữu nó rất nhanh; chạy theo mốt với tốc độ chóng mặt

Statistics

Statistics sâu về toán học và có bề dày về các lý thuyết suy diễn

- prediction, parameter estimation, hypothesis test
- không chỉ quan tâm đến kết quả ước lượng, mà cả mức độ bất định của ước lượng (confidence intervals, credible intervals,...)

Hai văn hoá

Machine Learning

- chú trọng đến mô hình thuật toán nhiều hơn mô hình dữ liệu (đặc biệt cho prediction/ supervised learning)
- cộng đồng học máy rất năng động: học một công cụ toán/ thống kê, rebranding, và sở hữu nó rất nhanh; chạy theo mốt với tốc độ chóng mặt

Statistics

Statistics sâu về toán học và có bề dày về các lý thuyết suy diễn

- prediction, parameter estimation, hypothesis test
- không chỉ quan tâm đến kết quả ước lượng, mà cả mức độ bất định của ước lượng (confidence intervals, credible intervals,...)

Machine Learning học hỏi nhiều từ nền tảng toán thống kê và mô hình dữ liệu. Ngược lại, sự trưởng thành của Machine Learning giúp các nhà thống kê bừng tỉnh, đánh giá đúng hơn vai trò của thuật toán và computing architecture.

Outline

Topic modeling

INPUT:

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

OUTPUT:

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Hai bài báo

D. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

- ứng dụng cho xử lý văn bản, xử lý ảnh trong trí tuệ nhân tạo, các ngành khoa học xã hội
- 19000 trích dẫn ở Google Scholar

J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

- ứng dụng trong sinh học phân tử, di truyền học
- 20000 trích dẫn ở Google Scholar

"Beauty and the Beast"

Dữ liệu thô:

$\mathbf{x} =$

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Chân lý:

$\boldsymbol{\theta} =$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

- $\mathbf{x}$ làm ký hiệu cho đầu vào của một thuật toán. I.e., $\mathbf{x}$ là biểu diễn của tập các văn bản

- $\boldsymbol{\theta}$ là đầu ra thuật toán; nó là biểu diễn cho các chủ đề chúng ta muốn chắt lọc từ nội dung của dữ liệu $\mathbf{x}$

- $\mathbf{x}$ lấy giá trị ở trong một không gian mà chúng ta sẽ định nghĩa một cách hình thức. Tương tự như vậy với $\boldsymbol{\theta}$

Để có thể suy diễn được chủ đề $\boldsymbol{\theta}$ từ dữ liệu thô $\mathbf{x}$, chúng ta cần phải thiết lập được một sự liên hệ giữa $\boldsymbol{\theta}$ và $\mathbf{x}$.

Mô hình xác suất $\mathbb{P}(\mathbf{x}|\boldsymbol{\theta})$

- dữ liệu được xem là hiện sinh (realization) của một biến ngẫu nhiên $\mathbf{x}$, còn chủ đề $\boldsymbol{\theta}$ là một tham số cho phân phối của dữ liệu

- dùng ngôn ngữ xác suất để xác lập mô hình toán học cho dữ liệu được xem là hết sức tự nhiên, vì dữ liệu thường có yếu tố bất định:
  - mỗi lần thu thập dữ liệu mới ta thường nhận một giá trị khác nhau, cho dù quy luật sinh ra dữ liệu có thể được xác định bởi cùng một phân bố "chân lý"
  - dẫu ta không bao giờ biết chắc chắn được phân bố chân lý, bằng những hiểu biết từ dữ liệu thực tế chúng ta có thể tìm một lớp phân bố xấp xỉ khả dĩ cho phân bố chân lý
  - lớp phân bố ấy sẽ được tham số hoá bởi $\boldsymbol{\theta}$

- bằng dữ liệu thực, chúng ta sẽ tìm giá trị tốt nhất có thể cho $\boldsymbol{\theta}$

Biểu diễn cho mô hình sinh dữ liệu $\mathbb{P}(\mathbf{x}|\boldsymbol{\theta})$

$$\mathbb{P}\left(\mathbf{x} = \right.$$

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

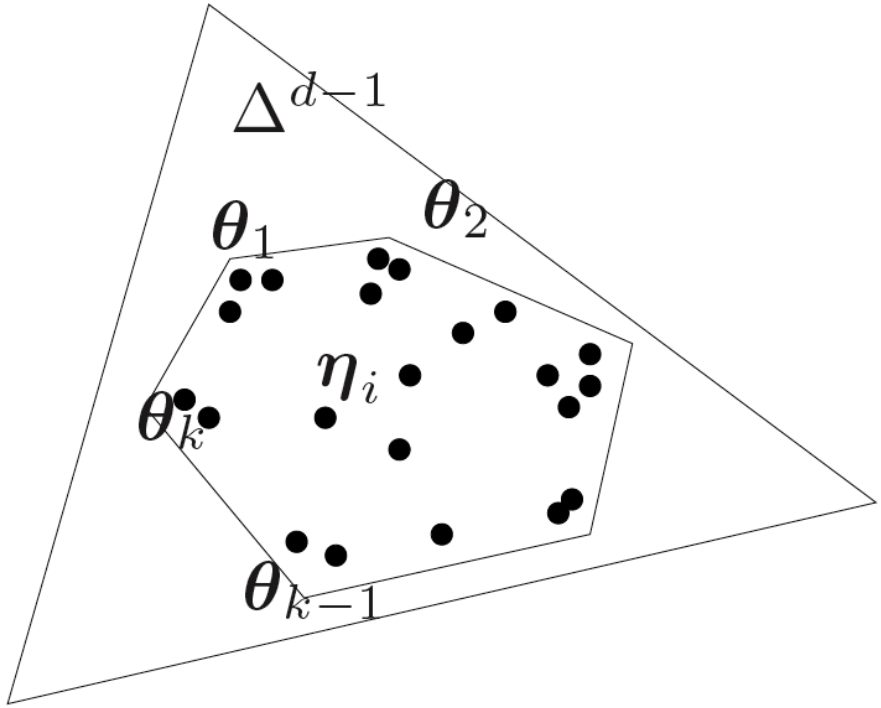
$$\left| \boldsymbol{\theta} = \right.$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

$$\left. \right)$$

- Cho V là bộ từ điển gồm có d phần tử (từ vựng) được đánh số bằng $V = \{1, \ldots, d\}$
  - với văn bản dài n từ, ta viết $\mathbf{x} = (x_1, \ldots, x_n) \in V^n$.
- Giả sử rằng có k chủ đề ẩn khác nhau $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, tương ứng với k phần tử nằm trong đơn hình xác suất

$$\Delta^{d-1} := \{u \in [0,1]^d | u_1 + \ldots + u_d = 1\}.$$



Mỗi chủ đề $\theta_i \in \Delta^{d-1}$ tương ứng với một tần suất xuất hiện nhất định của các từ trong V.

- Ví dụ: một chủ đề về "giáo dục" sẽ có nhiều từ với tần suất xuất hiện cao, như "trường", "lớp", "học sinh", "học phí", v.v.
- Một chủ đề về "đảng" sẽ có những từ khác với tần suất cao, như "nghị quyết", "quán triệt", "tiến lên", "to lớn", v.v.

- Thật khó tưởng tượng một văn bản chỉ đề cập đến một chủ đề duy nhất. Thực tế hơn, mỗi văn bản bao gồm sự pha trộn của nhiều chủ đề cùng một lúc, dẫu mức độ pha trộn có thể nhiều ít khác nhau.

- Thật khó tưởng tượng một văn bản chỉ đề cập đến một chủ đề duy nhất. Thực tế hơn, mỗi văn bản bao gồm sự pha trộn của nhiều chủ đề cùng một lúc, dẫu mức độ pha trộn có thể nhiều ít khác nhau.

- Còn ở ứng dụng trong sinh học phân tử, mỗi cá thể trong một cộng đồng có thể là kết quả của sự pha trộn các nguồn gốc khác nhau về mã di truyền, từ châu Phi, Á hay Âu.

- Trực quan trên gợi ra một khái niệm hình học rất tự nhiên: điểm ngẫu nhiên nằm trong tập bao lồi của các chủ đề:

$$G := \mathrm{conv}\{\theta_1, \ldots, \theta_k\} \subset \Delta^{d-1}.$$

Điểm ngẫu nhiên này, với ký hiệu $\boldsymbol{\eta}$, được xác lập bởi công thức

$$\boldsymbol{\eta} = \beta_1\theta_1 + \ldots + \beta_k\theta_k,$$

trong đó $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ là một biến ngẫu nhiên lấy giá trị trong đơn hình xác suất Δ^{k-1}.

- Trực quan trên gợi ra một khái niệm hình học rất tự nhiên: điểm ngẫu nhiên nằm trong tập bao lồi của các chủ đề:

$$G := \mathrm{conv}\{\theta_1, \ldots, \theta_k\} \subset \Delta^{d-1}.$$

Điểm ngẫu nhiên này, với ký hiệu $\boldsymbol{\eta}$, được xác lập bởi công thức

$$\boldsymbol{\eta} = \beta_1\theta_1 + \ldots + \beta_k\theta_k,$$

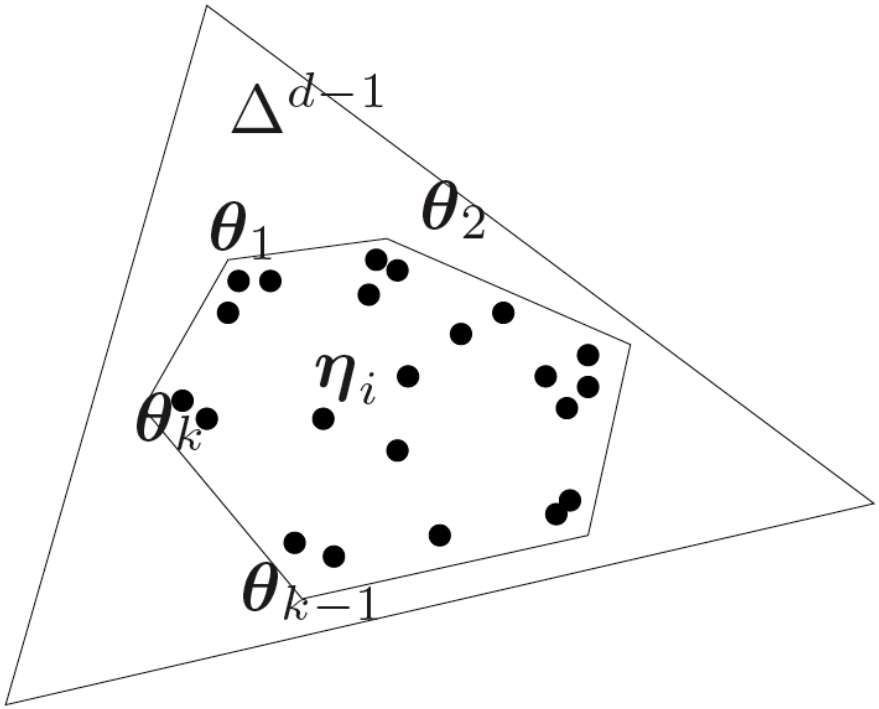
trong đó $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ là một biến ngẫu nhiên lấy giá trị trong đơn hình xác suất Δ^{k-1}.

- ta sẽ giả dụ rằng $\boldsymbol{\beta}$ tuân theo phân phối Dirichlet trong Δ^{k-1}

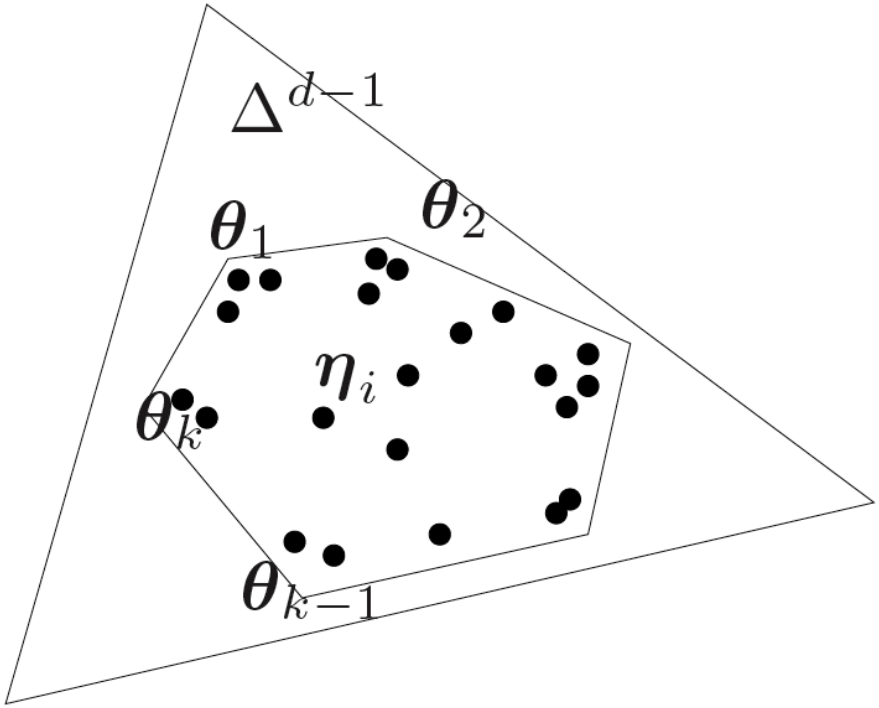
$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{\Gamma(\sum_{j=1}^k \gamma_j)}{\prod_{j=1}^k \Gamma(\gamma_j)} \prod_{j=1}^k \beta_j^{\gamma_j - 1}.$$

trên đây, Γ biểu thị hàm số Gamma, còn $\gamma_1, \ldots, \gamma_k$ là những tham số của mật độ Dirichlet.

Probability distribution on topic polytope G



- Định nghĩa trên đây cho (gián tiếp) một phân phối đối với biến ngẫu nhiên $\boldsymbol{\eta}$ lấy giá trị ở trong đa diện lồi G, ký hiệu bởi $\mathrm{d}P(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\gamma})$
  - có thể viết ra tường minh bằng công thức chuyển biến



Tham số $\gamma_1, \ldots, \gamma_k$ điều khiển mật độ của biến ngẫu nhiên $\boldsymbol{\eta} \in G$.

- nếu $\gamma_1 = \ldots = \gamma_k = 1$, $p(\cdot|\boldsymbol{\gamma})$ trở thành hằng số. Do đó khi $k < d$, nếu $\theta_1, \ldots, \theta_k$ là k đỉnh của đa diện lồi G, biến ngẫu nhiên $\boldsymbol{\eta}$ sẽ tuân theo phân bố *đồng nhất* trong lòng đa điện G.
- nếu $\gamma_1 = \ldots = \gamma_k \ll 1$, phân phối của $\boldsymbol{\eta}$ tập trung phần lớn khối (mass) ở sát các mặt biên của G.
- ngược lại nếu $\gamma_1 = \ldots = \gamma_k \gg 1$, phần lớn khối tập trung ở sâu bên trong G.

Chúng ta đã gần xong việc định nghĩa cho phân bố của dữ liệu $\mathbf{x}$:

- mỗi văn bản sẽ có tương ứng một tần suất $\boldsymbol{\eta}$ kể trên.
- khi biết $\boldsymbol{\eta}$, văn bản $\mathbf{x} = (x_1, \ldots, x_n) \in V^n$ được coi là một dãy các từ vựng trong V.
- mỗi x_i là một biến ngẫu nhiên độc lập tuân theo phân phối phân loại (categorical distribution): tung một con xúc xắc d mặt sao cho xác suất của x_i lấy giá trị mặt j sẽ là η_j, $j = 1, \ldots, d$.
- công thức hàm khối xác suất cho $\mathbf{x}$, khi biết $\boldsymbol{\eta}$:

$$p(\mathbf{x}|\boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{j=1}^{d} \eta_i^{1(x_i=j)}.$$

ở trên, $1(A) := 1$ nếu biểu hiện A đúng, và 0 nếu A sai.

Chúng ta đã gần xong việc định nghĩa cho phân bố của dữ liệu $\mathbf{x}$:

- mỗi văn bản sẽ có tương ứng một tần suất $\boldsymbol{\eta}$ kể trên.
- khi biết $\boldsymbol{\eta}$, văn bản $\mathbf{x} = (x_1, \ldots, x_n) \in V^n$ được coi là một dãy các từ vựng trong V.
- mỗi x_i là một biến ngẫu nhiên độc lập tuân theo phân phối phân loại (categorical distribution): tung một con xúc xắc d mặt sao cho xác suất của x_i lấy giá trị mặt j sẽ là η_j, $j = 1, \ldots, d$.
- công thức hàm khối xác suất cho $\mathbf{x}$, khi biết $\boldsymbol{\eta}$:

$$p(\mathbf{x}|\boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{j=1}^{d} \eta_i^{1(x_i=j)}.$$

  ở trên, $1(A) := 1$ nếu biểu hiện A đúng, và 0 nếu A sai.

Tóm lại, nếu đã biết giá trị các tham số $\theta_1, \ldots, \theta_k$, cũng như $\gamma_1, \ldots, \gamma_k$ được cho trước, mô hình sinh dữ liệu được định nghĩa bởi hàm mật độ sau

$$\boxed{p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\boldsymbol{\eta}) \, \mathrm{d}P(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\gamma}).}$$

Statistical inference of topic polytope

- Cho m văn bản $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, mỗi văn bản có đúng n từ.
  - như vậy, lượng dữ liệu thô là mn từ trong m văn bản

- Cho m văn bản $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, mỗi văn bản có đúng n từ.
  - như vậy, lượng dữ liệu thô là mn từ trong m văn bản

Bài toán học chủ đề

Cho trước số chủ đề k và các tham số $\gamma_1, \ldots, \gamma_k > 0$. Giả sử

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \sim \mathbb{P}(\mathbf{x}|\boldsymbol{\theta}).$$

Hãy tìm các chủ đề $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ để xác định phân bố sinh ra tập m văn bản.

Remarks

- Chúng ta không thể tìm ra chính xác các chủ đề θ, chỉ có thể ước lượng chúng.
- Nhưng ta kỳ vọng rằng nếu số lượng dữ liệu m, n càng lớn thì các ước lượng đối với θ càng tiến gần đến giá trị "chân lý" của chúng.
- Mặt khác, đây cũng không phải là vấn đề đơn giản về mặt tính toán, mà phải cần những hỗ trợ về thuật toán đủ mạnh
- Đó là các thuật toán "học máy" lấy đầu vào là tập m văn bản, và đầu ra chính là các ước lượng về "chủ đề"!

Convex geometry

Một bài toán kinh điển trong hình học lồi: *làm thế nào để ước lượng được các đỉnh của một đa diện lồi* G nếu như chúng ta quan sát được m hiện sinh độc lập rút ra theo phân bố đồng nhất từ trong lòng của G?

Convex geometry

Một bài toán kinh điển trong hình học lồi: *làm thế nào để ước lượng được các đỉnh của một đa diện lồi* G nếu như chúng ta quan sát được m hiện sinh độc lập rút ra theo phân bố đồng nhất từ trong lòng của G?

Khái quát hơn cho một phân bố Dirichlet trên một đa diện lồi:

Ước lượng đỉnh của một đa diện lồi

Cho trước $\gamma \in \mathbb{R}_+^k$. Giả sử $\eta^{(1)}, \ldots, \eta^{(m)}$ là m hiện sinh độc lập của một biến ngẫu nhiên $\boldsymbol{\eta}$ lấy giá trị trong G như đã định nghĩa ở mục trước. Hãy đưa ra một ước lượng tốt nhất cho đa diện G.

"Tốt nhất" được đo bằng một metric cụ thể, như Hausdorff: với hai đa diện G và G' bất kỳ trong một không gian chung $\mathbb{R}^d$

$$d_H(G, G') = \min\{\epsilon \geq 0 | G \subset G'_\epsilon, G' \subset G_\epsilon\}$$

trong đó $G_\epsilon = \{\theta + v | \theta \in G, v \in \mathbb{R}^d, \|v\|_2 \leq \epsilon\}$.

Luật số lớn trong xác suất

Có một ước lượng rất tự nhiên và kinh điển: lấy hình bao lồi của các hiện sinh $\eta^{(1)}, \ldots, \eta^{(m)}$,

$$\hat{G} := \mathrm{conv}(\eta^{(1)}, \ldots, \eta^{(m)}).$$

Theorem

If $m \to \infty$, then $d_H(G, \hat{G}) \to 0$ in probability.

- Một số kết quả tinh hơn về tốc độ hội tụ, và về phân bố tiệm cận (Reitzner, 2005; Bárány & Vu, 2007)
- Bài toán suy diễn chủ đề của ta phức tạp hơn bài toán kinh điển ở một số khía cạnh
  - Thứ nhất, ở Bài toán 1 ta không có trong tay các hiện sinh của vector tần suất $\eta^{(1)}, \ldots, \eta^{(m)}$, thay vào đó là các hiện sinh $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ ($\mathbf{x}$ là các văn bản (tập hợp các từ)
  - Thứ hai, vector tham số γ có thể bao gồm các giá trị khác với 1, và khác nhau (để phản ánh xác thực hơn dữ liệu thực tế). Nói cách khác, phân phối trong lòng đa diện lồi G thường không bao giờ là đồng nhất, tuy đó là một giả thuyết khá phổ biến trong hình học lồi cổ điển.

Maximum Likelihood Estimation
(Ước lượng bằng cực đại khả dĩ)

Trong thống kê toán có một phương pháp ước lượng rất mạnh và hữu dụng, dựa vào phép lấy cực đại của hàm khả dĩ của Fisher.

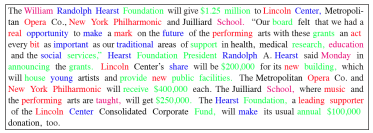
Hàm *khả dĩ* (likelihood function) là một hàm số đối với tham số $\boldsymbol{\theta}$ của (mật độ) phân bố $p(\mathbf{x}|\boldsymbol{\theta})$.

Ước lượng cực đại của hàm khả dĩ, viết tắt bằng MLE, khi đã được cho các dữ liệu $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ được định nghĩa như sau:

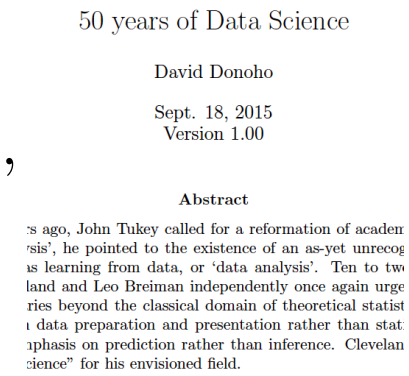
$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p(\mathbf{x}^{(i)}|\boldsymbol{\theta}).$$

Optimization

$$\max_{\boldsymbol{\theta}} \mathbb{P}\left(\mathbf{x} = \right.$$


The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

,


50 years of Data Science

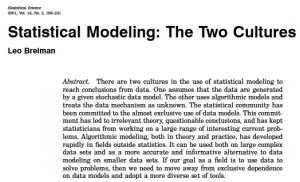
David Donoho

Sept. 18, 2015
Version 1.00

Abstract

's ago, John Tukey called for a reformation of academ
'sis', he pointed to the existence of an as-yet unrecog
as learning from data, or 'data analysis'. Ten to tw
land and Leo Breiman independently once again urge
ries beyond the classical domain of theoretical statist
l data preparation and presentation rather than stat
mphasis on prediction rather than inference. Clevelan
cience" for his envisioned field.

,


Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

$$, \cdots \Big| \boldsymbol{\theta}\right)$$

So we obtain

$$\boldsymbol{\theta} =$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Một bảo đảm quan trọng đối với ước lượng MLE cho lớp mô hình chủ đề:

Theorem (Nguyen, 2015)

Với $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ *là ước lượng MLE. Viết* $\hat{G} := conv(\hat{\theta}_1, \ldots, \hat{\theta}_k)$. *Nếu* m *và* n *cùng tiến tới vô cùng, trong đó* $\log\log m \leq \log n = o(m)$, *ta có*

$$d_H(\hat{G}, G) = O_P\left(\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}\right)^{1/(2(q+\alpha))}.$$

Trên đây, $q = \min\{(k-1), (d-1)\}$, α *là một hằng số dương phụ thuộc vào* γ, O_P *có nghĩa là "hằng số cận trên trong xác suất", được xác định đối với phân phối* $p(\mathbf{x}|\boldsymbol{\theta})$.

Nếu số lượng văn bản càng tăng, thì suy diễn của ta càng chính xác!

Màn dạo đầu:

Bayesian inference vs Frequentist inference

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$
- Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát.

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$
- Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát.
- Nhưng có một sự tranh cãi quyết liệt về ý nghĩa của khái niệm $\boldsymbol{\theta}$: nó có bất định hay không?

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$
- Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát.
- Nhưng có một sự tranh cãi quyết liệt về ý nghĩa của khái niệm $\boldsymbol{\theta}$: nó có bất định hay không?
- Có hai trường phái đối với câu hỏi này: Các nhà *thống kê tần suất* cho rằng $\boldsymbol{\theta}$ hoàn toàn có thể xác định được. Phương pháp ước lượng cực đại của hàm khả dĩ (MLE) chính là một sản phẩm của các nhà tần suất.

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$
- Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát.
- Nhưng có một sự tranh cãi quyết liệt về ý nghĩa của khái niệm $\boldsymbol{\theta}$: nó có bất định hay không?
- Có hai trường phái đối với câu hỏi này: Các nhà *thống kê tần suất* cho rằng $\boldsymbol{\theta}$ hoàn toàn có thể xác định được. Phương pháp ước lượng cực đại của hàm khả dĩ (MLE) chính là một sản phẩm của các nhà tần suất.
- Ngược lại, trường phái *Bayes* cho rằng tri thức về $\boldsymbol{\theta}$ cũng bất định, và do đó về mặt toán học, nó cũng phải được biển diễn bởi một biến ngẫu nhiên.

- Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\mathbf{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\mathbf{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$
- Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát.
- Nhưng có một sự tranh cãi quyết liệt về ý nghĩa của khái niệm $\boldsymbol{\theta}$: nó có bất định hay không?
- Có hai trường phái đối với câu hỏi này: Các nhà *thống kê tần suất* cho rằng $\boldsymbol{\theta}$ hoàn toàn có thể xác định được. Phương pháp ước lượng cực đại của hàm khả dĩ (MLE) chính là một sản phẩm của các nhà tần suất.
- Ngược lại, trường phái *Bayes* cho rằng tri thức về $\boldsymbol{\theta}$ cũng bất định, và do đó về mặt toán học, nó cũng phải được biển diễn bởi một biến ngẫu nhiên.
- Vì chúng ta không thể quan sát được khái niệm trừu tượng $\boldsymbol{\theta}$, chúng ta không thể xác quyết đúng sai trong tranh cãi này.

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.
- Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu.

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.
- Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu.
  - Phân phối này được gọi là *phân bố tiên nghiệm*, ký hiệu bởi $\Pi(\boldsymbol{\theta})$, nó là chắt lọc của tri thức mà nhà thống kê có được trước khi anh ta quan sát.

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.
- Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu.
  - Phân phối này được gọi là *phân bố tiên nghiệm*, ký hiệu bởi $\Pi(\boldsymbol{\theta})$, nó là chắt lọc của tri thức mà nhà thống kê có được trước khi anh ta quan sát.
  - Sau khi đã thu thập dữ liệu, m văn bản $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, kiến thức của nhà thống kê về $\boldsymbol{\theta}$ được tổng kết thành *phân bố hậu nghiệm*.

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.
- Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu.
  - Phân phối này được gọi là *phân bố tiên nghiệm*, ký hiệu bởi $\Pi(\boldsymbol{\theta})$, nó là chắt lọc của tri thức mà nhà thống kê có được trước khi anh ta quan sát.
  - Sau khi đã thu thập dữ liệu, m văn bản $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, kiến thức của nhà thống kê về $\boldsymbol{\theta}$ được tổng kết thành *phân bố hậu nghiệm*.
- Phân bố hậu nghiệm được xác định bằng công thức Bayes nổi tiếng:

$$\boxed{\mathrm{d}\Pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot \mathrm{d}\Pi(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}) \cdot \mathrm{d}\Pi(\boldsymbol{\theta})}.}$$

- Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được.
- Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu.
  - Phân phối này được gọi là *phân bố tiên nghiệm*, ký hiệu bởi $\Pi(\boldsymbol{\theta})$, nó là chắt lọc của tri thức mà nhà thống kê có được trước khi anh ta quan sát.
  - Sau khi đã thu thập dữ liệu, m văn bản $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, kiến thức của nhà thống kê về $\boldsymbol{\theta}$ được tổng kết thành *phân bố hậu nghiệm*.
- Phân bố hậu nghiệm được xác định bằng công thức Bayes nổi tiếng:

$$\boxed{\mathrm{d}\Pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot \mathrm{d}\Pi(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}) \cdot \mathrm{d}\Pi(\boldsymbol{\theta})}.}$$

- Có thể chứng minh được rằng khi số lượng dữ liệu càng nhiều thì phân bố hậu nghiệm tập trung càng gần về giá trị "chân lý" của $\boldsymbol{\theta}$ (Nguyen, 2015): niềm tin của các nhà thống kê Bayes cũng tiệm cận dần tới chân lý!

Differentiation vs Integration

- MLE, a popular estimation approach of the Frequentist school, requires taking differentiation of information given by data (via likelihood function)
- Bayesian inference requires integration with respect to prior and likelihood over the domain of parameter (possible truth)
- In this way, the two schools can be seen as complementary, but there are other distinctions that are technically fascinating and can sometimes be philosophically difficult to reconcile
  - topic of the second part of lecture

Thuật toán học chủ đề

- Thách thức cho Thống kê hiện đại và học máy: mục tiêu tính toán hiệu quả những đại lượng trọng yếu như ước lượng MLE, hay phân bố hậu nghiệm.
  - để có độ tin cậy cao các thuật toán phải quét qua một số lớn các văn bản có độ dài khác nhau, con số đó có thể đến hàng vạn, hàng triệu, thậm chí hàng tỷ
  - thời gian chạy thuật toán có thể mất hàng ngày hoặc hàng tuần để có được kết quả tương đối chính xác của bài toán tìm cực đại cho hàm khả dĩ.
- Thường chỉ cần tìm các giải pháp xấp xỉ
  - phép suy diễn biến phân (variational inference) (Blei et al, 2003).
  - phương pháp suy diễn Bayes xấp xỉ dựa vào Markov Chain Monte Carlo

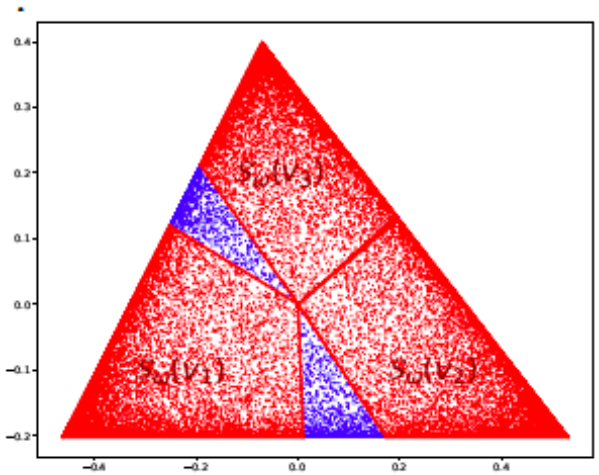
Thuật toán học chủ đề

- Thách thức cho Thống kê hiện đại và học máy: mục tiêu tính toán hiệu quả những đại lượng trọng yếu như ước lượng MLE, hay phân bố hậu nghiệm.
  - để có độ tin cậy cao các thuật toán phải quét qua một số lớn các văn bản có độ dài khác nhau, con số đó có thể đến hàng vạn, hàng triệu, thậm chí hàng tỷ
  - thời gian chạy thuật toán có thể mất hàng ngày hoặc hàng tuần để có được kết quả tương đối chính xác của bài toán tìm cực đại cho hàm khả dĩ.
- Thường chỉ cần tìm các giải pháp xấp xỉ
  - phép suy diễn biến phân (variational inference) (Blei et al, 2003).
  - phương pháp suy diễn Bayes xấp xỉ dựa vào Markov Chain Monte Carlo

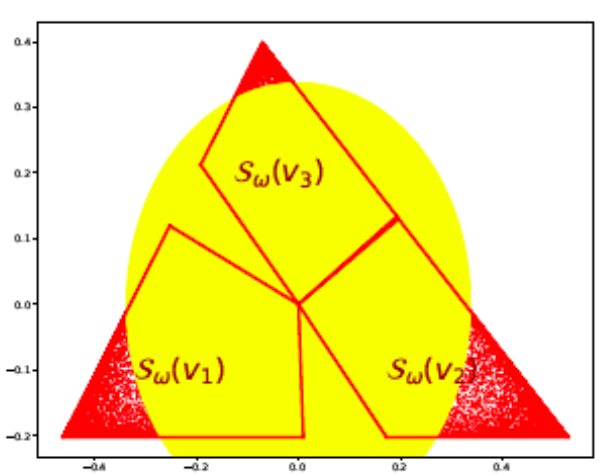
Có một thực trạng khó chịu: phép biến phân xấp xỉ cho ra kết quả rất nhanh, nhưng lại chưa trúng. Ngược lại, thuật toán mô phỏng chuỗi Markov đưa ra được ước lượng khá trúng, nhưng lại không nhanh.

Algorithms based on convex geometry

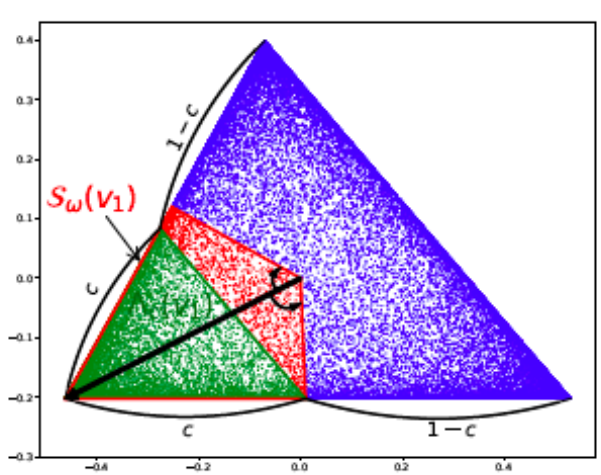
Convex geometry based algorithms can be both fast and accurate (Yurochkin & Nguyen, 2016; Yurochkin, Guha & Nguyen, 2017)



(a) An incomplete coverage



(b) Full coverage



(c) Cap $\Lambda_c(v_1)$ and cone $\mathcal{S}_\omega(v_1)$

For simulated data

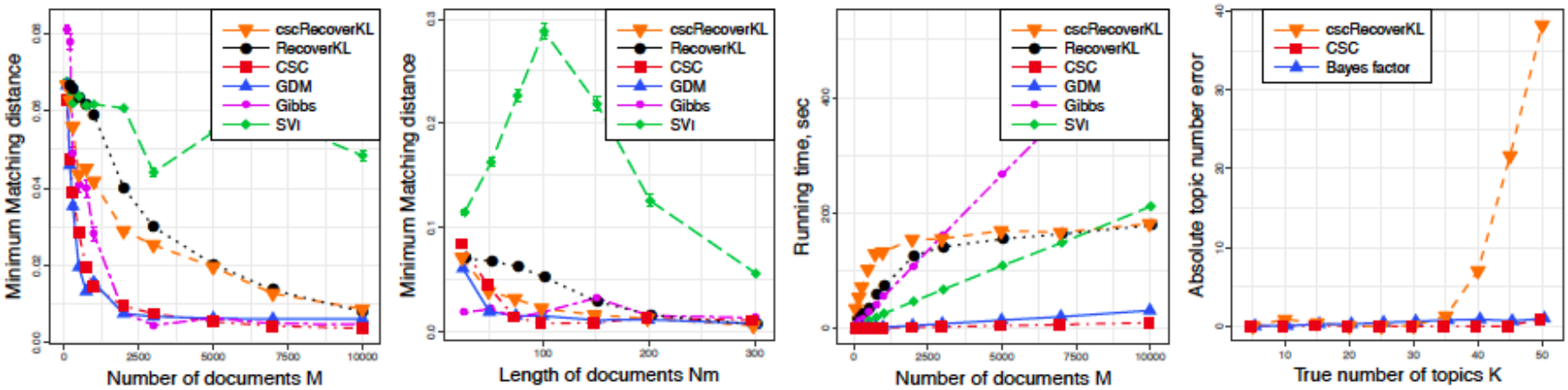


Figure 3: Minimum matching Euclidean distance for (a) varying corpora sizes, (b) varying length of documents. (c) Running times for varying corpora sizes. (d) Estimation of number of topics

Applications to analysis of New York Times articles

Table 1: Modeling topics of NYTimes articles

	K	Perplexity	Coherence	Time
cscRecoverKL	27	2603	-238	37 min
HDP Gibbs	229	1435	-433	3 days
LDA Gibbs	80	1522	-300	10 hours
CSC	159	1568	-322	19 min

What have we illustrated?

Elements of the science

... the science of learning from data, with all that it entails

- elements of data representation
  - from data storage to abstract data structures and visualization
  - both in a mathematical sense and sense of a computer system
- elements of statistical learning and inference
  - from statistical modeling to prediction and inference
  - interplay of algorithmic and statistical efficiency
- elements offered by real-world data domains in sciences and industry
  - fresh challenges in data scale and complexity

What kind of math?

For an undergraduate:

- probability and statistics (and real and functional analysis): how to think about uncertainty and inference
- linear algebra and matrix/tensor analysis: how to represent
- continuous and discrete optimization: how to optimize

More math in advanced data science research:

- combinatorics: random discrete structures
- differential geometry: for statistical theory and algorithm analysis
- topology: topological data analysis
- and so on

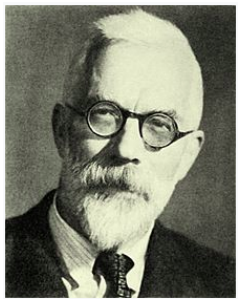
Kết luận: $\boxed{\textit{data science} \geq \int_{\Omega} \textit{data} \cdot (\textit{math} \wedge \textit{stat} \wedge \textit{cs})}$

- Công nghệ thông tin mang cho ta nguồn dữ liệu dồi dào, làm nảy sinh một nhu cầu cần suy diễn với những dữ liệu ấy.
- Thống kê và toán học cho ta một nền móng và công cụ để suy diễn một cách hợp lý.
  - Ngược lại sự phát triển của khoa học suy diễn từ dữ liệu, dữ liệu lớn, cũng đưa toán học tới gần với các ứng dụng của xã hội hiện đại hơn bao giờ hết. Vấn đề suy diễn với các cấu trúc phức tạp cũng có tác dụng thúc đẩy sự phát triển nội tại của toán học.
- Không chỉ đòi hỏi sự chuẩn bị tốt về toán học, một chuyên gia về KHDL phải có sự tò mò đối với thực tiễn, có khả năng diễn đạt thành thục bằng ngôn ngữ của toán, có sự liều lĩnh để sử dụng ngôn ngữ ấy vào dữ liệu thực.

Outline

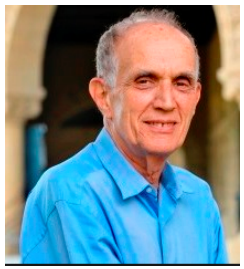
Frequentist

Bernoulli, Gauss, Kolmogorov
Pearson, Wald, Stein, Lehmann



Ronald Fisher (1890–1962)

Jerzy Neyman (1894–1981)

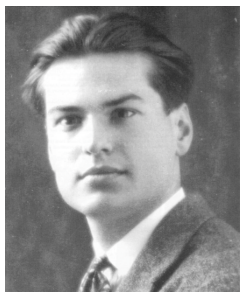


Brad Efron (1936–)

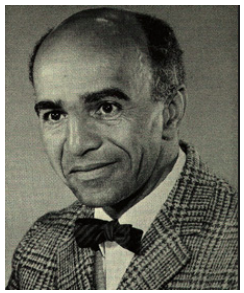
Bayesian

Bayes, Laplace, Borel
Good, Jeffreys, Savage, Lindley

Thomas Bayes (1702–1761)



Bruno de Finetti (1906–1985)



David Blackwell (1919–2010)

A simple question

Nếu trong năm vừa qua bạn tới trường 200 ngày, và tới trễ mất 102 ngày.
Hỏi: ngày mai xác suất bạn đi trễ là bao nhiêu?

A simple question

Nếu trong năm vừa qua bạn tới trường 200 ngày, và tới trễ mất 102 ngày. Hỏi: ngày mai xác suất bạn đi trễ là bao nhiêu?

"Quá dễ không cần nghĩ! Tôi không cần phải học thống kê để trả lời câu hỏi này. Khả năng trễ xấp xỉ bằng 1/2 (chính xác là 0.51)!".

Another question

8/11/2016 là ngày nước Mỹ đi bầu cử tổng thống.

Trên trang FiveThirtyEight.com Nate Silver (người đã dự báo chính xác kết quả bầu cử tổng thống vài lần trước) có ghi, lúc 9 giờ sáng:

"Xác suất Hillary Clinton thắng cử là 71%"

Tối hôm đó, Donald Trump được tuyên bố thắng cử trong sự kinh ngạc của rất nhiều người.

a familiar mathematician's question

11/23/2016 2:21PM

Just before the election day, I have been thinking quite hard what the prediction means exactly.

When you toss a coin with probability 1/2, that means something when you toss the coin 1000 times.

Elction happens only once, so what does the probability of winning mean?

11/23/2016 7:50PM

Right, you can't run elections 1000 times. Or, unless the universe we live in is just one of many creations. 1/2 is the frequentist probability, which makes sense only if you can run an experiment many times. The probability of winning an election should be interpreted as the Bayesian probability, which is what you believe in (the chance that something (will) happen). It's a well-defined notion: it's a conditional probability of an event — to be interpreted as the

In 1950, an economist preparing a report asked statistician David Blackwell (not yet a Bayesian) to estimate the probability of another world war in the next five years.

In 1950, an economist preparing a report asked statistician David Blackwell (not yet a Bayesian) to estimate the probability of another world war in the next five years.
Blackwell answered: "Oh, that question just doesn't make sense. Probability applies to a long sequence of repeatable events, and this is clearly a unique situation. The probability is either 0 or 1, but we won't know for five years."
The economist replied,
"I was afraid you were going to say that. I've spoken to several other statisticians, and they all told me the same thing."

Excerpt from "A theory that would not die" – Sharon McGrayne

We use probability to quantify uncertainty of our knowledge given data, in our life (often instinctively) and in our work

There are two *interpretations* of probabilities: frequentist probability and Bayesian probability

- the mathematical concept of probability is the same, it is how to use probability to quantify knowledge that is different!

We use probability to quantify uncertainty of our knowledge given data, in our life (often instinctively) and in our work

There are two *interpretations* of probabilities: frequentist probability and Bayesian probability

- the mathematical concept of probability is the same, it is how to use probability to quantify knowledge that is different!

Mathematical Statistics provides the mathematical foundation to resolving the question of (inductive) inference.

The two different interretations correspond to two distinct approaches to statistical inference

In a nutshell

Frequentist approach is natural when data can be obtained by repeated experiments

Bayesian approach is more appropriate when one needs to draw conclusion based on only the data available

Statistical Learning/Inference

Vague definition: statistical inference is about making sense of data.

Operational definition: a computational process of turning data to statistics, prediction and understanding.

Vague definition: statistical inference is about making sense of data.

Operational definition: a computational process of turning data to statistics, prediction and understanding.

Mathematical representation of data and knowledge

- the data are represented by a variable x taking values in $\mathcal{X}$
- the inference is made regarding some notion of true knowledge (truth) represented by parameter $\theta \in \Theta$.

x the records of your lateness in past year; θ the probability you will be late tomorrow.

x a collection of heights sampled from a population; θ the typical height of nation.

x is the blinking cursor on radar screen; $\theta \in \mathbb{R}^3$ the actual location of airplane.

x represents poll numbers; $\theta = 1$ or 0, whether you will become president or not.

x the weird dream you have tonight; θ the winning lottery number tomorrow.

x your birth date; θ your future life event.

A less obvious example, x is the collection of data pair of the form (y, z), where z is the binary class label representing the "class" of y. θ is a classification function

- θ maps y to prediction $\theta(y)$ of true label z

A clustering problem involves subdividing a collection of data points represented by x into "clusters", which can be represented by θ.

point of departure:

data is random

Statisticians universally agree to take x to be *random*.
Moreover, there has to be some sort of link between x and θ that allows us to draw inference about θ given the information we have based on x.

Statisticians universally agree to take x to be *random*.
Moreover, there has to be some sort of link between x and θ that allows us to draw inference about θ given the information we have based on x.

- using the language of probability, it is assumed that x is a random variable distributed according to a distribution function parameterized by θ. We write

$$x \sim f(x|\theta)$$

meaning of $f(x|\theta)$

$f(x|\theta)$ is called a probability model for data x.

It means: if I know the true θ, f gives me a probabilistic mechanism for obtaining the random x.

The statistical inference problem can be reduced to this: given the data (which is a realization of random variable x), what can we say about θ?

the distribution-free appearance

Most statisticians agree up to this point. Moreover, the distribution-teller f is assumed known, so the remaining issue left is with θ.

It is worth noting that there are many problems especially those tacked in machine learning, distribution-free approaches may be prefered. This means that the probability model f is implicit rather than being explicit. Nonetheless, it is still extremely useful to think that such an underlying model f exists, if one is to understand and analyze the properties of the inference approaches, even if such approaches do not overtly make use of any probability assumption.

the Great Debate:

unknown θ is random, or is this not?

Having said that data x is a realization of some probability model $f(x|\theta)$, statisticians may not agree about the nature of θ:

Bayesians say that θ is random; frequentists insist that it is not.
This leads to the famous frequentist-Bayesian divide.

in some settings, the distinction is quite superficial, but in some other settings, the difference is fundamental. In some other settings they are quite complementary.

it appears that the span of Bayes and frequentist axes do not cover all of statistical inference. With the increased interest in large scale inference, computational considerations represent another formidable dimension that promises to rewrite the entire landscape of modern statistics.

no longer does the frequentist vs Bayes debate excite as much passion as it did half a century ago.

pragmatic viewpoints seem to prevail these days: anything goes if it "solves" a problem.

an increasingly accepted view now is that we should continue to honor the distinction between the two approaches, while recognizing that with computational axis now becoming a major player, a complete story of statistical inference may yet to be written.

Bayesian inference

basic concepts

Since the Bayesian approach posits that parameter θ is random, it needs to be endowed with a probability distribution π on the parameter space Θ, namely $\pi(\theta)$.

- π shall be called "prior distribution"

The conditional distribution of θ given data x can now be obtained via the law of probability — this is famously called the Bayes' rule

$$\boxed{\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta)f(x|\theta)d\theta}}$$

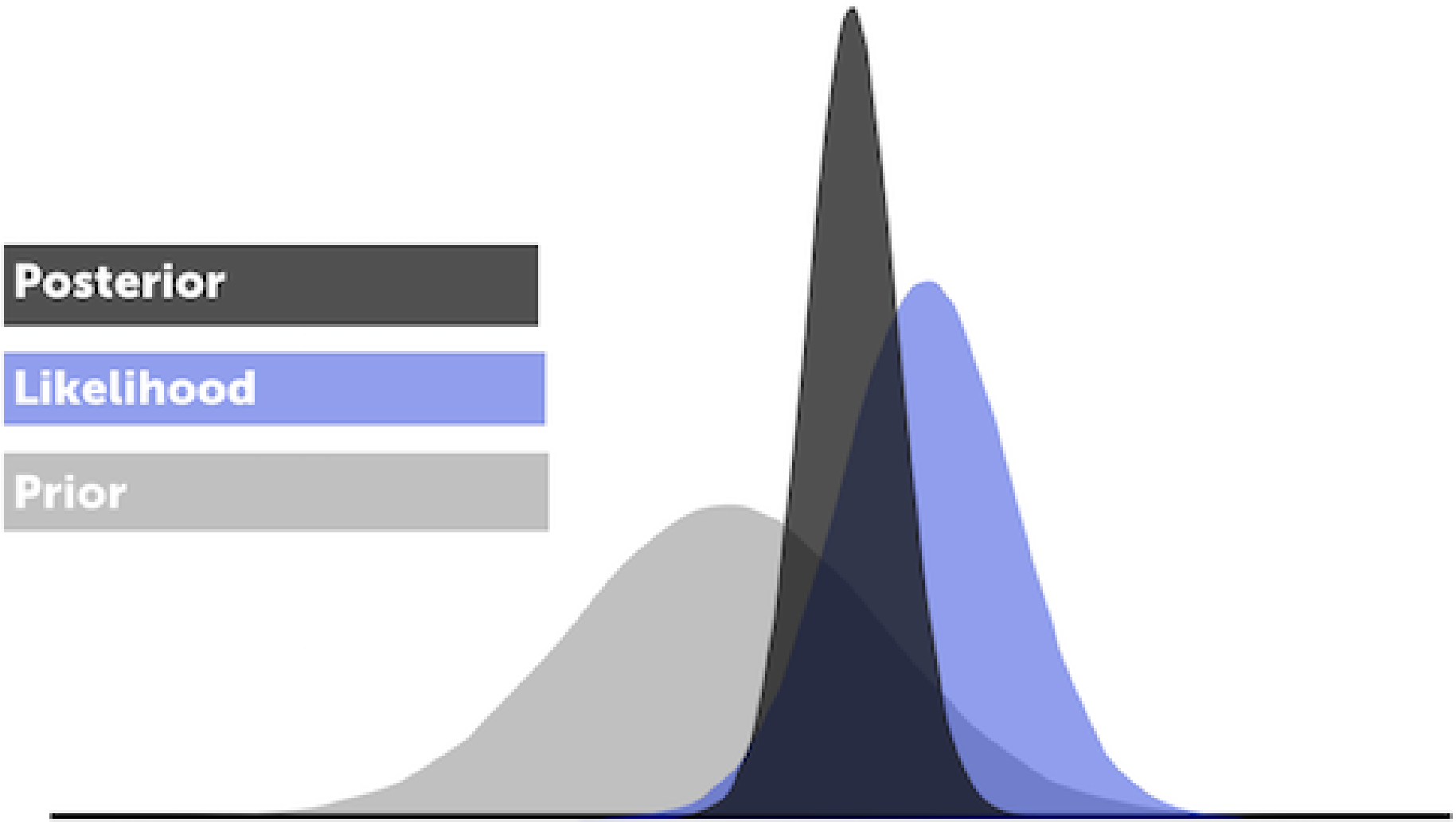
This conditional distribution is called *posterior distribution*.
Bayesian inference is drawn on the basis of the posterior distribution, conditionally on the data x.

Note that the denominator is equal to the marginal distribution of x which is induced by the joint distribution of θ and x. Because the marginal of x does not depend on θ, we may write that

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta),$$

or

$$\boxed{\text{posterior} \propto \text{prior} \times \text{likelihood}}$$



(Ilustration of Kim Larsen)

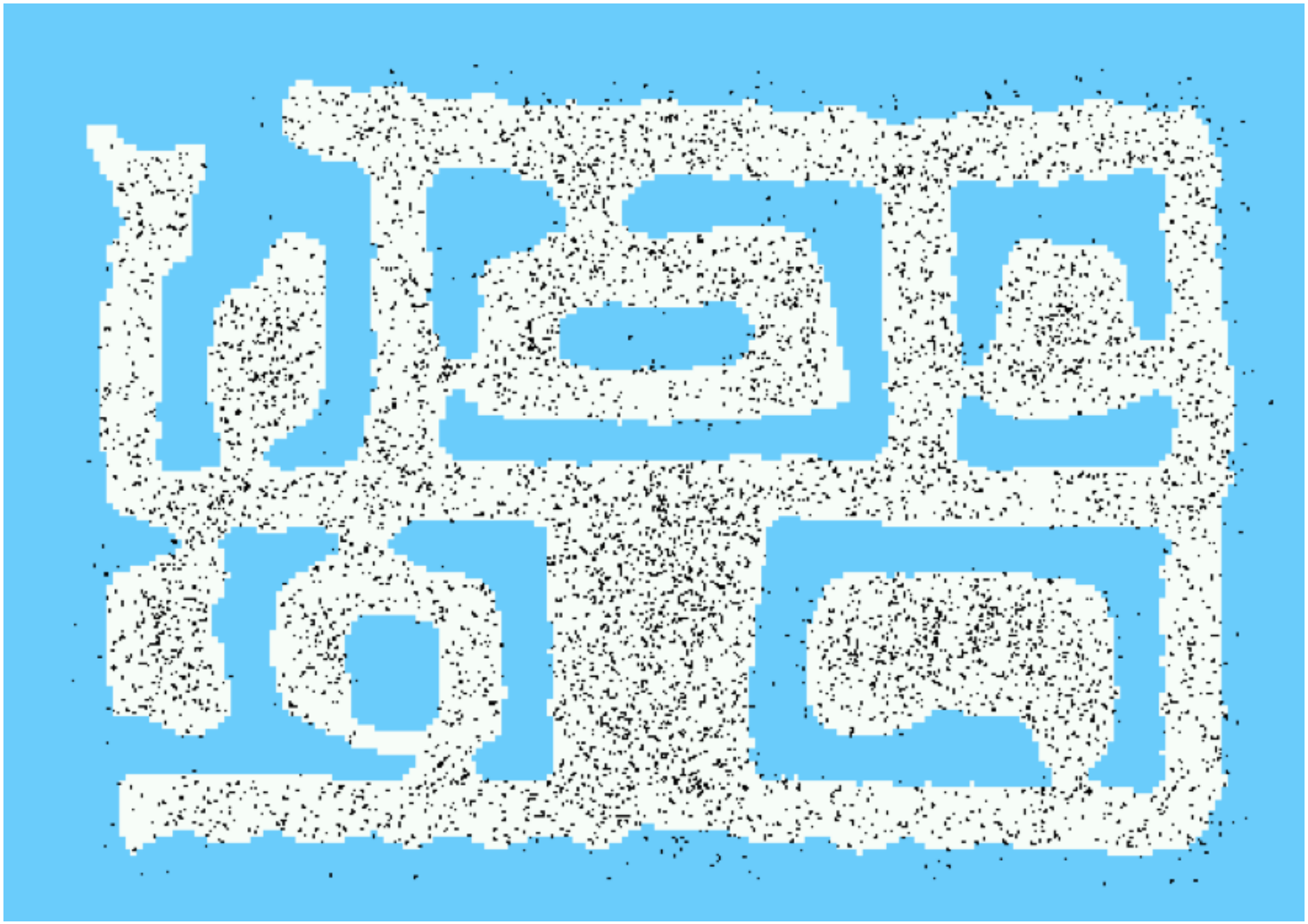
$$\boxed{\text{posterior} \propto \text{prior} \times \text{likelihood}}$$

This is one of the most influential equations in the history of development of human thoughts. The equation makes it clear that the inference about θ given observation x must come from a combination of both the model and the prior knowledge about the model.

What is more striking is that it puts both the causes (θ) and effects (x) on the same conceptual level, because both of them have probability distributions.

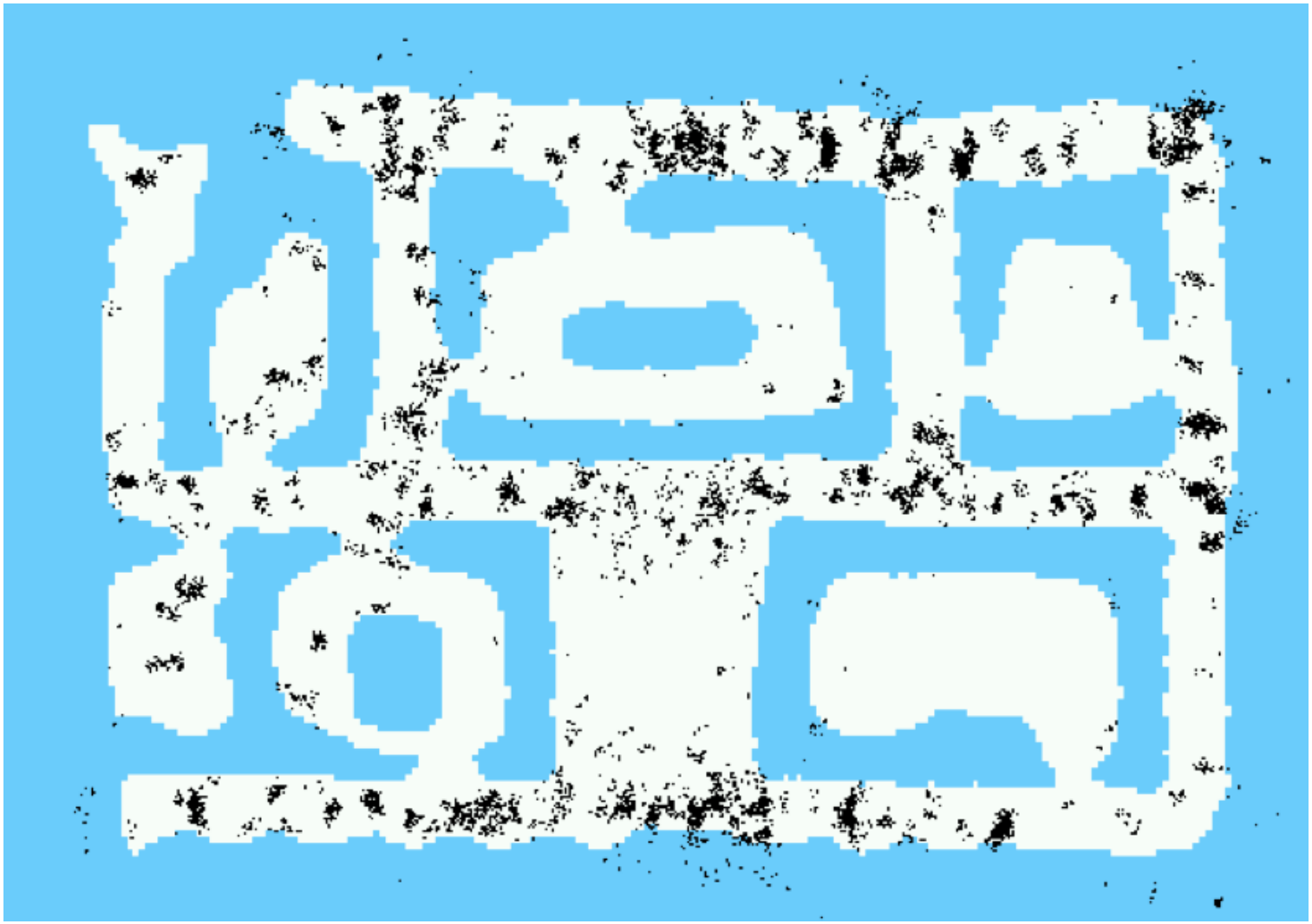
A robot's Bayesian analysis to self-localize

Sample of possible locations from prior distribution

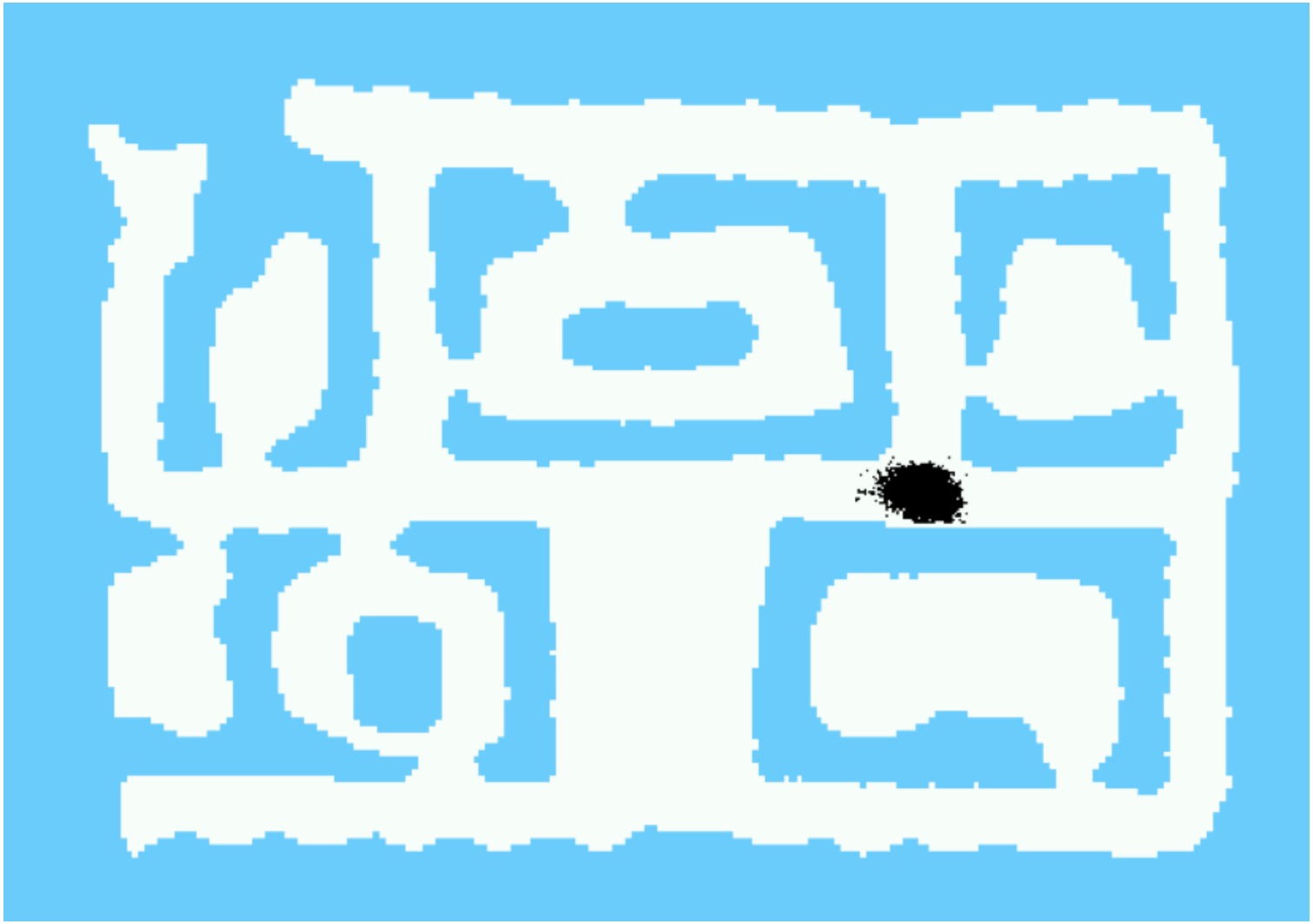


[Source: Choset et al, 2005]

Sample from posterior distribution after 10 data points (via sensor scans)



Sample from posterior distribution after 65 data points



criticism of Bayesian viewpoint

First, what if θ is deterministic?

Second, where does this prior come from? Requiring the priors means inference is no longer *objective*. Moreover, the business of putting priors on things seems treacherous.

Third, computationally expensive to calculate the posterior distribution. All these questions are legitimate and can be addressed.

advantages of Bayesian viewpoint

First, at the conceptual level, the Bayesian choice simplifies inference greatly, by equating statistical inference to the inverse problem in probability.

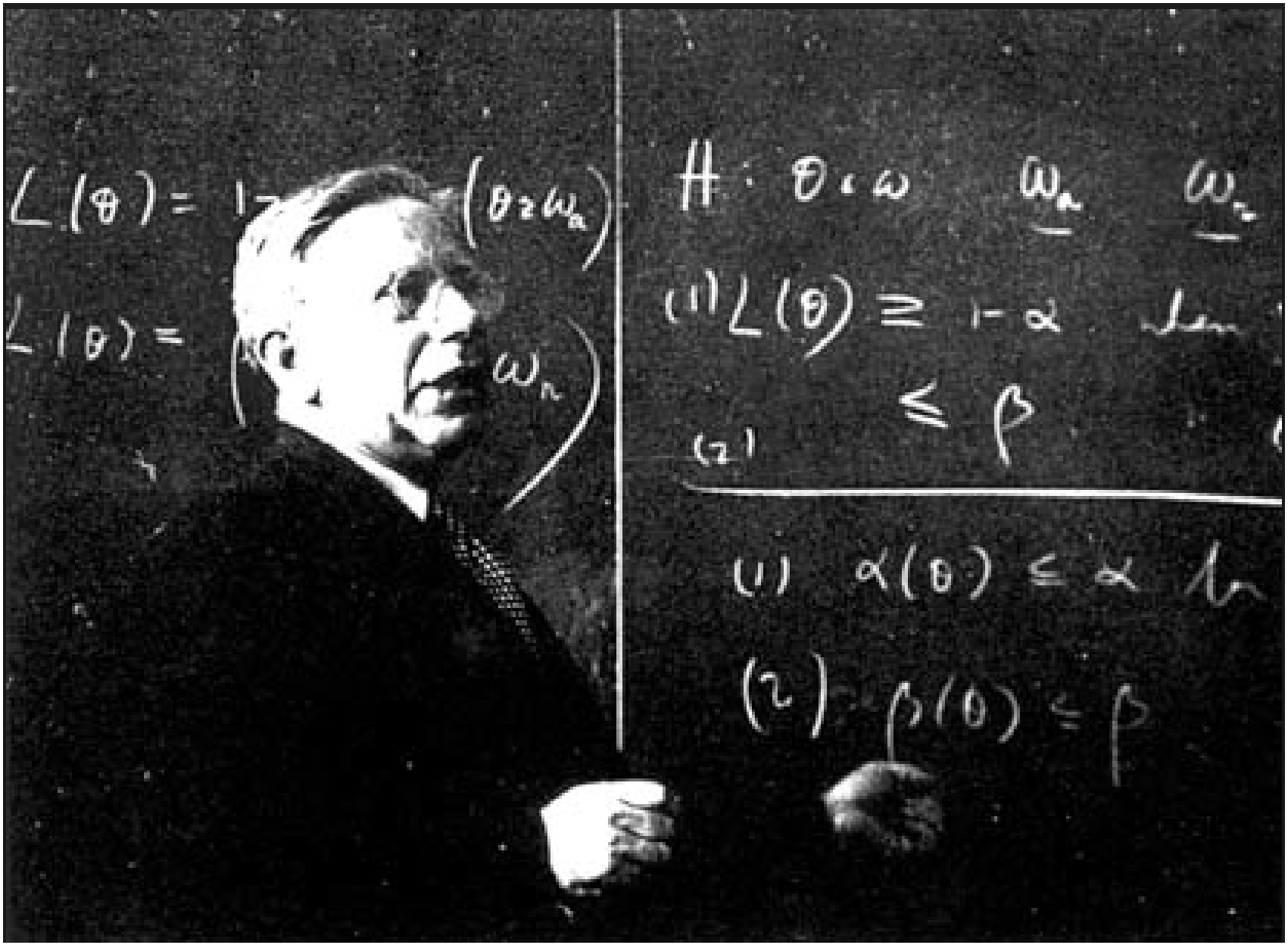
- in fact, at the time of Bayes and Laplace (all the way up to the advent of modern statistics in the 20th century), Statistics was often called inverse probability.

Second, at a more pragmatic level, by being explicit about the prior distribution as another source of inference, the Bayesian choice does not shy away from subjectivity.

Third, prior distribution is the best way of summarizing our available prior knowledge (or the lack of such knowledge) as well as the residual of uncertainty, thus allowing for the incorporation of this imperfect information about the decision making process.

Bayes vs Frequentist
via lense of decision-theoretic framework

Abraham Wald (1902–1950)



basic decision-theoretic concepts

To evaluate the quality of an inference procedure, in this framework we need a notion of loss, a function $\ell : \Theta \times \mathcal{D} \to [0, +\infty)$.

- Θ is the space of parameters representing the state of nature, $\mathcal{D}$ is called a decision space.

Thus the value $\ell(\theta, \delta)$ represents the quality of action δ regarding the state of nature θ.

parameter estimation

In parameter estimation, the goal is to simply estimate θ given data x. Then, δ is a function of data x, and $\mathcal{D} = \Theta$, while ℓ represents some sort of estimation error.

parameter estimation

In parameter estimation, the goal is to simply estimate θ given data x. Then, δ is a function of data x, and $\mathcal{D} = \Theta$, while ℓ represents some sort of estimation error.

Squared loss function is given by $\ell(\theta, \delta) := (\theta - \delta)^2$, for real-valued θ, δ.

parameter estimation

In parameter estimation, the goal is to simply estimate θ given data x. Then, δ is a function of data x, and $\mathcal{D} = \Theta$, while ℓ represents some sort of estimation error.

Squared loss function is given by $\ell(\theta, \delta) := (\theta - \delta)^2$, for real-valued θ, δ. For classification problem, a zero-one loss function is defined as $\ell(\theta, \delta) := \mathbb{I}(\theta \neq \delta)$.

frequentist risk

A frequentist is interested in the following notion, *frequentist risk*, the expected loss that one incurs by averaging over all data realizations

$$\boxed{R(\theta, \delta) = \mathbb{E}_\theta \ell(\theta, \delta(x))}$$

where θ is fixed, and the expectation is taken with respect to the distribution of x, namely, some $f(x|\theta)$.

Clearly one would prefer a procedure/decision δ which yields smallest risk. But the frequentist risk is a function of θ.

This raises some difficulty: how do we compare between different procedures?

Given two procedures δ_1 and δ_2, if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ, i.e., δ_1 dominates δ_2 everywhere, we say that δ_2 is *inadmissible* — we should always opt for δ_1 instead.

More commonly, there are ranges of the parameter space in which δ_1 dominates δ_2, and then ranges where δ_2 dominates δ_1. Moreover, we do not know which range does the true θ lie in.

unbiased procedures

A way to get around it, from a frequentist viewpoint, is to restrict the class of procedures to a subset with a sensible property, such as classes of unbiased procedures.

- that is, those δ such that $\mathbb{E}\delta(x) = \theta$.

Within the unbiased class one may look for the procedure with a minimum risk (e.g., minimum variance).

Example

For squared loss, we have $\ell(\theta, \delta(x)) = (\theta - \delta(x))^2$. Then,

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta(\theta - \delta(x))^2 \\ &= (\theta - \mathbb{E}_\theta \delta(x))^2 + \mathbb{E}_\theta(\delta(x) - \mathbb{E}_\theta \delta(x))^2 \\ &= \text{Bias}^2 + \text{Variance}. \end{aligned}$$

This motivated minimum variance unbiased (MVU) estimators.

Example

For squared loss, we have $\ell(\theta, \delta(x)) = (\theta - \delta(x))^2$. Then,

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta(\theta - \delta(x))^2 \\ &= (\theta - \mathbb{E}_\theta\delta(x))^2 + \mathbb{E}_\theta(\delta(x) - \mathbb{E}_\theta\delta(x))^2 \\ &= \text{Bias}^2 + \text{Variance}. \end{aligned}$$

This motivated minimum variance unbiased (MVU) estimators.

Critique: Bayesian statisticians do not find this criterion appealing, because it involves averaging over all possible data realizations x, leaving no particular attention to the actual data x available at hand. Finally, it must be noted that insisting on unbiasedness would leave out good procedures. In fact, Bayesian procedures are generally unbiased.

Example

We want to estimate the height of a daughter, denoted θ, given the height of the mother, denoted by x.
Assume that $(x, \theta) \sim N_2(\mu, \Sigma)$, where $\mu := (\mu_1, \mu_2) = (160, 160)^T$ (in centimeters). Σ has equal variances, $\Sigma_{11} = \Sigma_{22} = \sigma^2$, and correlation $\rho := \Sigma_{12}/\sigma^2 = 0.5$.
The Bayesian approach takes the conditional mean of θ given x:

$$\mathbb{E}(\theta|x) = \mu_2 + \rho(x - \mu_1) = 160 + \rho(x - 160).$$

Verify that this is unbiased. Given fixed θ, $\mathbb{E}(x|\theta) = \mu_1 + \rho(\theta - \mu_2)$, so

$$\begin{aligned}\mathbb{E}_\theta(\mu_2 + \rho(x - \mu_1)) &= 160 + \rho(\mathbb{E}_\theta x - 160) \\ &= 160 + \rho(160 + \rho(\theta - 160) - 160) \\ &= 160 + \rho^2(\theta - 160) \neq \theta.\end{aligned}$$

Example (continued...)

Consider instead a linear estimator of the form

$$\delta(x) := 160 + \eta(x - 160),$$

then $\mathbb{E}_\theta \delta(x) = 160 + \eta(160 + \rho(\theta - 160) - 160) = \theta$ iff we set $\eta = 2$.
Is this a good estimator?
Suppose that the mother is observed to measure at 170 cm, then the Bayesian estimate for the daughter would be 160.25, while the unbiased estimator gives 180cm. This contradicts our common sense, that exceptionally tall people are flunctuation, where most people would be concentrated near the mean.

Example (continued...)

Consider instead a linear estimator of the form

$$\delta(x) := 160 + \eta(x - 160),$$

then $\mathbb{E}_\theta \delta(x) = 160 + \eta(160 + \rho(\theta - 160) - 160) = \theta$ iff we set $\eta = 2$.
Is this a good estimator?
Suppose that the mother is observed to measure at 170 cm, then the Bayesian estimate for the daughter would be 160.25, while the unbiased estimator gives 180cm. This contradicts our common sense, that exceptionally tall people are flunctuation, where most people would be concentrated near the mean.

Mathematically, the Bayesian estimate naturally exercises a technique that is now known as "shrinkage", or regularization/penalization. Shrinkage estimators are now widely accepted as they provide superior performance both practically and theoretically.

Minimax optimality in frequentist inference

Another way is to compare the procedure based on the worse case scenario, $\sup_\theta R(\theta, \delta)$. The procedure that achieves the infimum of the worse case scenario is called the minimax optimal procedure.

- that is, δ^* such that

$$\inf_{\delta \in \mathcal{D}} \sup_\theta R(\theta, \delta) = \sup_\theta R(\theta, \delta^*).$$

This is an elegant mathematical solution to the problem of comparison.

But from an inferential standpoint, the minimax criterion picks procedure that pays attention only at the most difficult range of parameter θ, the range of values which may never actually happen.

It is possible that the true value of θ is not within such a range, and that for the typical true value of θ, there exists procedures δ' for which the risk $R(\theta, \delta)$ is far smaller than the risk incurred by the minimax optimal procedure:

$$R(\theta, \delta') \ll R(\theta, \delta^*).$$

The minimax criterion is most appropriate in situations when the state of nature θ acts in an adversarial manner.

- for instance, θ represents a mechanism for generating spam mails. Then one can expect the spammers to change θ from time to time to get around improved spam filters.

Moreover, the minimax risk is in some sense indicative of the difficulty of an inference problem.

Bayesians generally ignore the worst case scenarios, though not entirely. Instead, they look to introduce a "weight function" of θ to tell which part of the parameter deserves more attention, and seek for procedure that minimizes according to this weight function. This weight does not merely come from our prior knowledge — it is also driven by the data that we observe. As one may guess, this weight function will come from the posterior distribution of θ given the data.

Frequentist is pessimistic

By centering on frequentist risk

$$\boxed{R(\theta, \delta) = \mathbb{E}_\theta l(\theta, \delta(x))}$$

a frequentist approach

- is in favor of procedures that perform well on average among all possible data realizations, including the data that were not actually observed.
- some procedures are also evaluated by averaging over all possible experiments, including the ones that could have been performed, but were not.
- such is the spirit of a frequentist, always seeing the data as just an instance of a lot more to come, perennially worried about what may happen far into the future, instead of dwelling on the availability of the presence.

Bayesian decision theory

The Bayesian approach integrates on the space Θ since θ is unknown, instead of integrating on the space $\mathcal{X}$, as x is known.

It relies on the *posterior expected loss*, or posterior risk

$$\rho(x, \pi) = \int l(\theta, \delta(x))\pi(\theta|x)\ d\theta.$$

The integration is with respect to the posterior distribution $\pi(\theta|x)$

Definition

(Bayes action). The Bayes action, $\delta^*(x)$ is the value of $\delta(x)$ that minimizes the posterior risk.

If ℓ is the square loss, then $\delta^*(x) = \mathbb{E}(\theta|x)$, the posterior mean.

If ℓ is the absolute error loss, $\ell(\theta, \delta) = |\theta - d|$, given that $\Theta = \mathcal{D} = \mathbb{R}$, then $\delta^*(x)$ is the median of the posterior distribution $\pi(\theta|x)$.

The zero-one loss arises in the contexts of binary hypothesis test and classification problem. Here $\Theta = \mathcal{D} = \{0, 1\}$, and $\ell(\theta, d) = \mathbb{I}(\theta \neq d)$. The two hypotheses H_0 and H_1 are to be distinguished. Then, $\delta^*(x) = 1$ if $\pi(H_1|x) \geq \pi(H_0|x)$ and 0 otherwise.

Bayes and Frequentist

Bayes perspective: considers expected loss by averaging over parameter θ

Frequentist perspective: considers expected loss by averaging over data x.

These two concepts are linked by Fubini's theorem.

Definition

Let $R(\theta, \delta)$ be the frequentist risk, and π a distribution over θ. A Bayes rule is a decision δ which minimizes the expected risk, obtained by averaging the frequentist risk over θ:

$$r(\pi, \delta) = \int R(\theta, \delta) d\pi(\theta).$$

The minimum expected risk is called "Bayes risk".

By Fubini's theorem, we can write

$$r(\pi, \delta) = \int \int l(\theta, \delta(x) f(x|\theta) \pi(\theta) d\theta dx = \int \rho(x, \pi) p(x) dx.$$

It can be observed that the Bayes action achieves the Bayes risk.

While this sounds great for a Bayesian, a "hard-core" frequentist might simply shrug at this result, for she would not care about the Bayes risk in the first place!

Summary

BAYES

– parameter θ is random
– data x is fixed
– overtly subjective and optimistic
– suitable when data can't be replicated, and knowledge highly uncertain
– computationally more expensive

FREQUENTIST

– parameter θ is non-random;
– data x is one of more to come
– overtly pessimistic and aims to be objective
– suitable when data experiments replicable
– good for evaluation of procedures

Summary

BAYES	FREQUENTIST
– parameter θ is random	– parameter θ is non-random;
– data x is fixed	– data x is one of more to come
– overtly subjective and optimistic	– overtly pessimistic and aims to be objective
– suitable when data can't be replicated, and knowledge highly uncertain	– suitable when data experiments replicable
– computationally more expensive	– good for evaluation of procedures

— "Big data" does not always mean data experiments are easily replicable for the inference question of interest, partly because space of inference questions also grows quickly

— Sometimes it is not easy to know which setting to put ourselves in; this is a choice we must make. There is also a computational consideration that increasingly becomes another central pillar of inference (e.g., what if θ cannot be computed efficiently even if we know in which form it is supposed to be)

First detection of gravitational waves

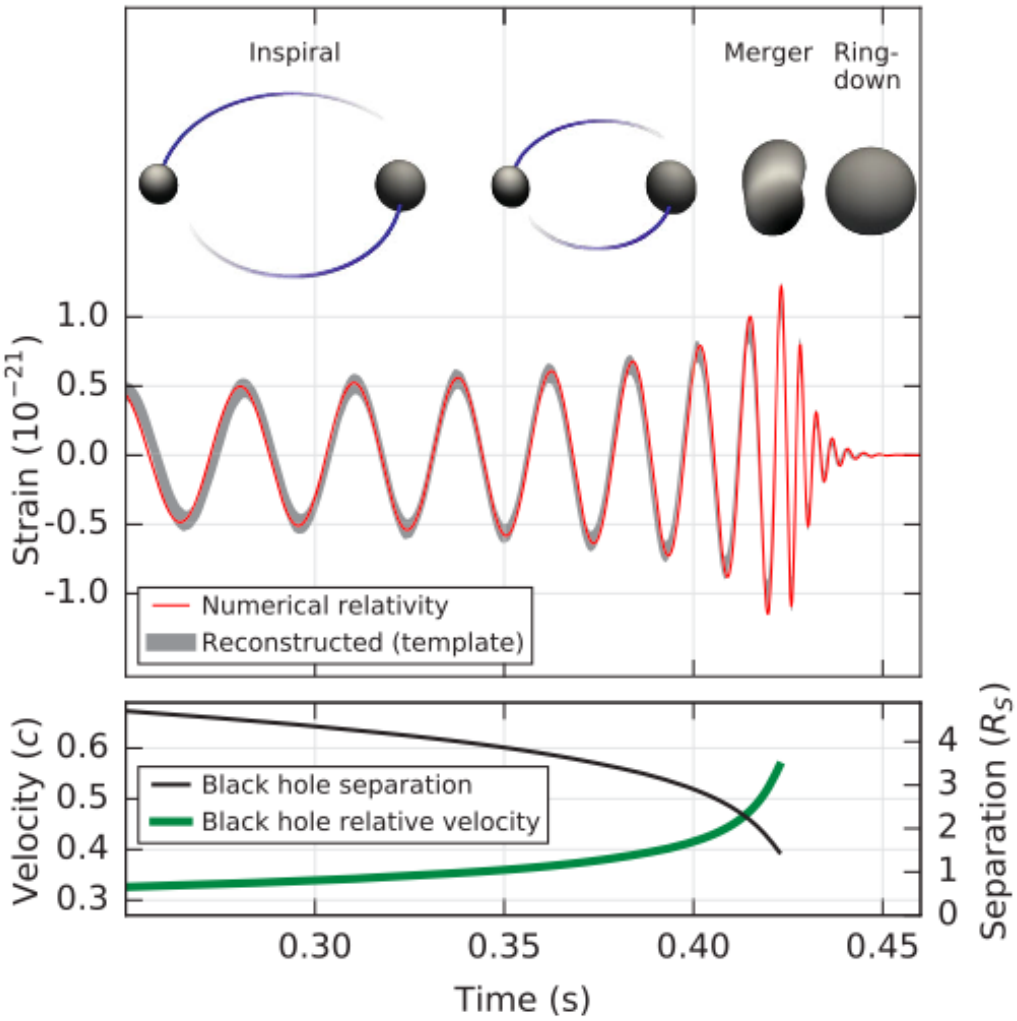
- predicted by Einstein 100 years ago
- gravitational waves estimated to have travelled for 1 billion years, due to the merger of two black holes
- energy released by the merger reached a level greater than combined power of light radiated by all the stars in the observable universe

PRL 116, 061102 (2016) Selected for a Viewpoint in *Physics* PHYSICAL REVIEW LETTERS week ending 12 FEBRUARY 2016

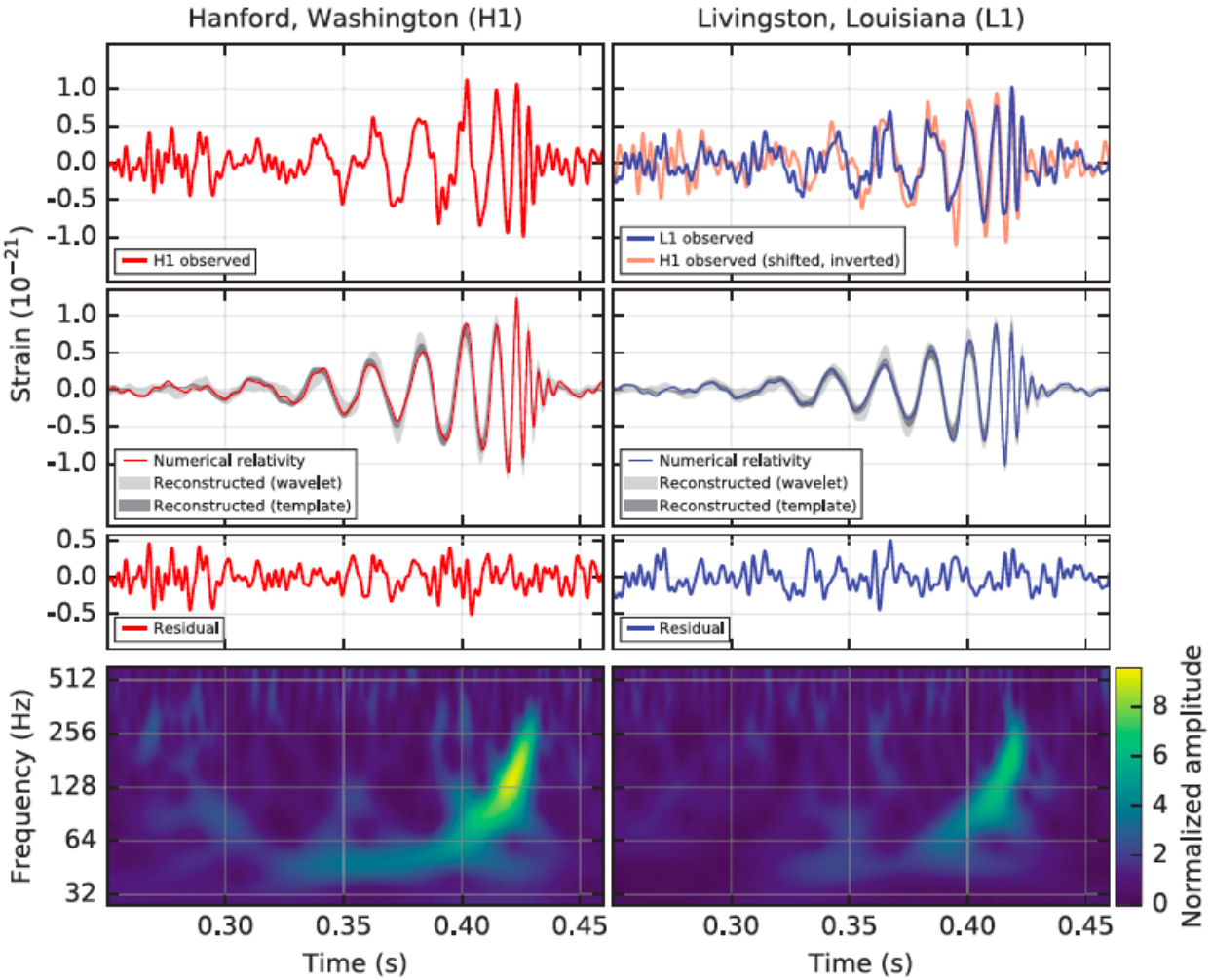
Observation of Gravitational Waves from a Binary Black Hole Merger

B. P. Abbott *et al.*[*]
(LIGO Scientific Collaboration and Virgo Collaboration)
(Received 21 January 2016; published 11 February 2016)

On September 14, 2015 at 09:50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal sweeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a matched-filter signal-to-noise ratio of 24 and a false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a significance greater than 5.1σ. The source lies at a luminosity distance of 410^{+160}_{-180} Mpc corresponding to a redshift $z = 0.09^{+0.03}_{-0.04}$. In the source frame, the initial black hole masses are $36^{+5}_{-4}M_{\odot}$ and $29^{+4}_{-4}M_{\odot}$, and the final black hole mass is $62^{+4}_{-4}M_{\odot}$, with $3.0^{+0.5}_{-0.5}M_{\odot}c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct



widely lauded to be the greatest scientific discovery of 21 century, and a great triump for experimental astrophysics



- much work go into statistical analysis to validate the theory
  - to make sure signals not due to noise or other sources of disturbance
  - to identify correctly the source of event consistent with physical theory
  - to provide accurate estimate of source parameters (about black holes)

PRL 116, 061102 (2016) Selected for a Viewpoint in *Physics* PHYSICAL REVIEW LETTERS week ending 12 FEBRUARY 2016

Observation of Gravitational Waves from a Binary Black Hole Merger

B. P. Abbott *et al.*[*]
(LIGO Scientific Collaboration and Virgo Collaboration)
(Received 21 January 2016; published 11 February 2016)

On September 14, 2015 at 09:50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal sweeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a matched-filter signal-to-noise ratio of 24 and a false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a significance greater than 5.1σ. The source lies at a luminosity distance of 410^{+160}_{-180} Mpc corresponding to a redshift $z = 0.09^{+0.03}_{-0.04}$. In the source frame, the initial black hole masses are $36^{+5}_{-4}M_{\odot}$ and $29^{+4}_{-4}M_{\odot}$, and the final black hole mass is $62^{+4}_{-4}M_{\odot}$, with $3.0^{+0.5}_{-0.5}M_{\odot}c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

DOI: 10.1103/PhysRevLett.116.061102

- first two tasks were achieved by sophisticated physics model based hypothesis tests (maximum likelihood based tests), i.e., frequentist approach
- the last task was achieved by a refined Bayesian analysis

Frequentist or Bayes, which side are you?

- understanding this foundation of inference can help us to avoid pitfalls and abusive practices (e.g., the tendency to draw conclusion you want to draw a priori, via misuses of p-value, the sensitivity of model choice)
- the Bayes vs frequentist divide will always exist due to the unsolvable nature of inductive inference and the nature of knowledge
- most of the time it appears that the two approaches actually complement other rather than contradict
- depending on data/problem, one may choose which approach to proceed
  - if this sounds a bit Bayesian, that is because the author's outlook actually is; although neither would a hard-core frequentist nor a hard-core Bayesian approve of this inconsitency!

Outline

References

For Section 1: Foundations of Data Science

D. Donoho. 50 years of Data Science, *John Tukey Centennial*, 2015.

L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231, 2001.

M. I. Jordan & T. Mitchell. Machine Learning: Trends, perspectives, and prospects. *Science*, 349 (6245), 255–260, 2015.

For Section 2: Topic Modeling, an illustration

D. Blei, A. Ng, & M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.

J. Pritchard, M. Stephen & P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics*, 2000.

X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21:618–646, 2015.

X. Nguyen. Tìm chân lý từ dữ liệu thô: mô hình chủ đề và hình học. *Pi Magazine*, 5, 2017.

M. Reitzner. Central limit theorems for random polytopes. *Probability Theory and Related Fields*, 133:483–507, 2005.

I. Bárány and V. Vu. Central limit theorems for Gaussian polytopes. *Annals of Probability*, 35:1593–1621, 2007.

For Section 3: Bayes and frequentist inference

J. Berger. Statistical decision theory and Bayesian analysis, Springer 1985.
P. Bickel & K. Doksum. Mathematical statistics: basic ideas and selected topics, vol. 1, Prentice Hall, 2000.
C. Bishop. Pattern recognition and machine learning, 2007.
G. Casella & R. Berger. Statistical inference, 2001.
T. Hastie, R. Tibshirani & J. Friedman. Elements of statistical learning, Springer, 2009.
O. Kallenberg. Foundations of modern probability. Springer, 2010.
M. I. Jordan, Introduction to probablistic graphical models. Unpublished text book.
C. Robert. The Bayesian choice: From decision-theoretic foundation to computational implementation, Springer, 2007.
A. van der Vaart. Asymptotic Statistics, Cambridge University Press, 2000.
L. Wasserman. All of Statistics: a concise course in statistical inference, Springer, 2004.