STATISTICAL MODELING VS. MACHINE LEARNING – SIMILARITIES AND DIFFERENCES

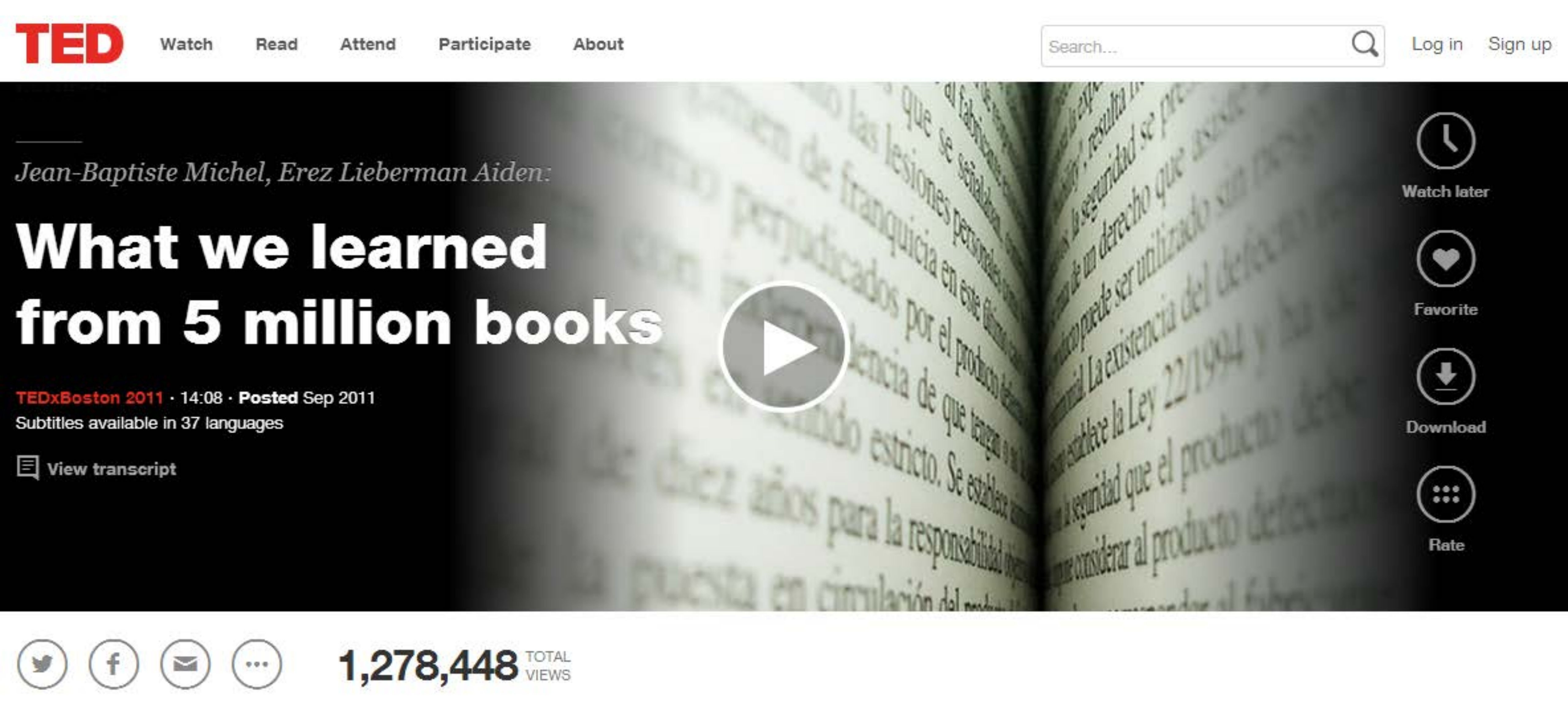
WK RECH 2019, Frankfurt School of Finance & Management

I. Illustrative Examples: Google Books and Google Flu Trends
II. Two Paradigms: Statistical Modeling (SM) vs. Machine Learning (ML)
III. Bankruptcy Prediction: The SM vs. ML Way
IV. Reflections

ILLUSTRATIVE EXAMPLES: GOOGLE BOOKS AND GOOGLE FLU TRENDS

GOOGLE BOOKS STUDY

PADERBORN UNIVERSITY



GOOGLE BOOKS STUDY

Paper

Published in final edited form as:
Science. 2011 January 14; 331(6014): 176–182. doi:10.1126/science.1199644.

Quantitative analysis of culture using millions of digitized books

Jean-Baptiste Michel[1,2,3,4,*,†], **Yuan Kui Shen**[5], **Aviva P. Aiden**[6], **Adrian Veres**[7], **Matthew K. Gray**[8], **The Google Books Team**[8], **Joseph P. Pickett**[9], **Dale Hoiberg**[10], **Dan Clancy**[8], **Peter Norvig**[8], **Jon Orwant**[8], **Steven Pinker**[4], **Martin A. Nowak**[1,11,12], and **Erez Lieberman Aiden**[1,12,13,14,15,16,*,†]

[1] Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts 02138, USA.

[2] Institute for Quantitative Social Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

[3] Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA.

[4] Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[5] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA.

[6] Harvard Medical School, Boston, Massachusetts, 02115, USA.

[7] Harvard College, Cambridge, Massachusetts 02138, USA.

[8] Google, Inc., Mountain View, California, 94043, USA.

[9] Houghton Mifflin Harcourt, Boston, Massachusetts 02116, USA.

[10] Encyclopaedia Britannica, Inc., Chicago, Illinois 60654, USA.

[11] Dept of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

[12] Dept of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.

[13] Broad Institute of Harvard and MIT, Harvard University, Cambridge, Massachusetts 02138, USA.

[14] School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

[15] Harvard Society of Fellows, Harvard University, Cambridge, Massachusetts 02138, USA.

[16] Laboratory-at-Large, Harvard University, Cambridge, Massachusetts 02138, USA.

Abstract

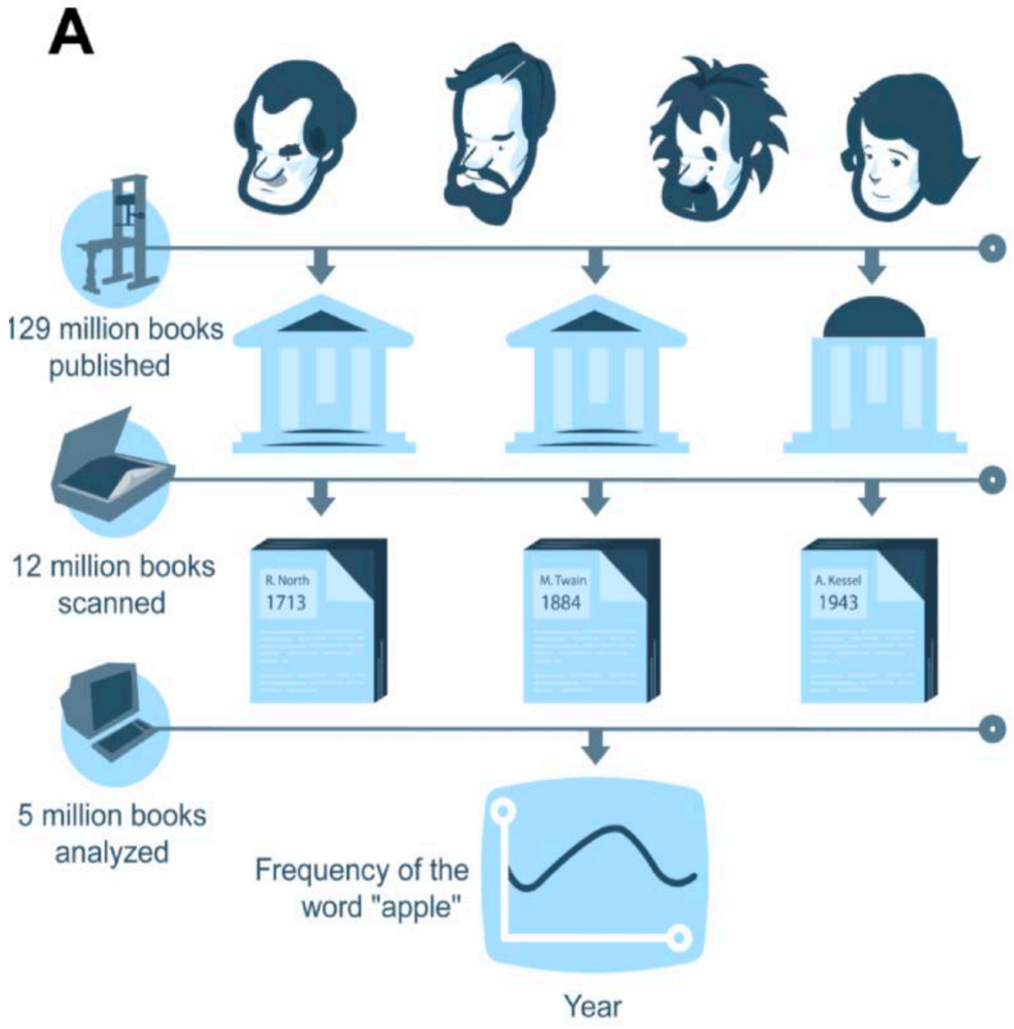
We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics', focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. 'Culturomics' extends the boundaries

† To whom correspondence should be addressed. jb.michel@gmail.com (JBM); erez@erez.com (EA)..
* these authors contributed equally

Supporting Online Material www.sciencemag.org/cgi/content/full/??? Materials and Methods Figs. S1 to S19

PADERBORN UNIVERSITY

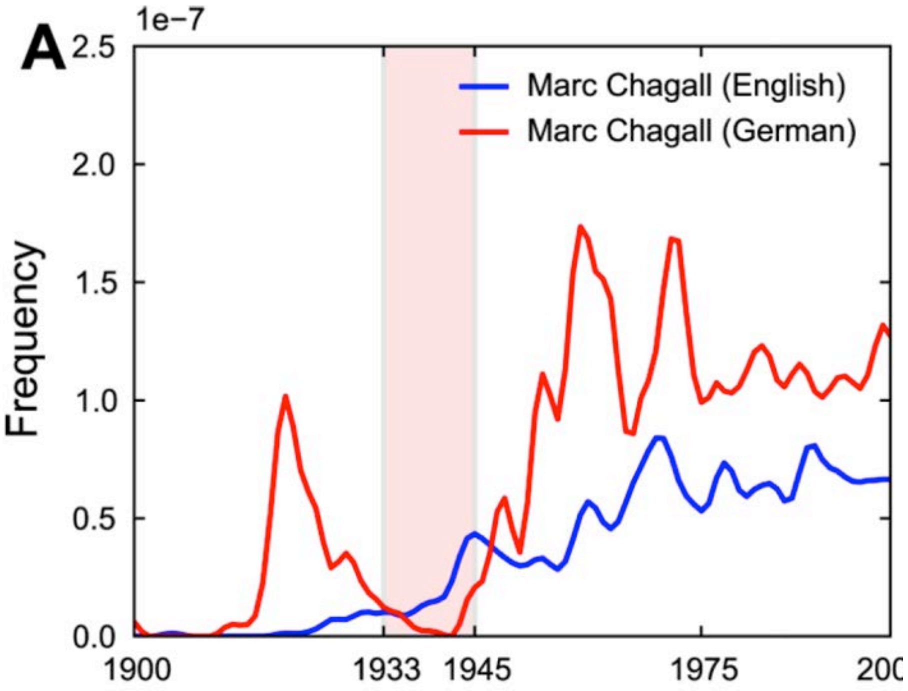
Data and Methods

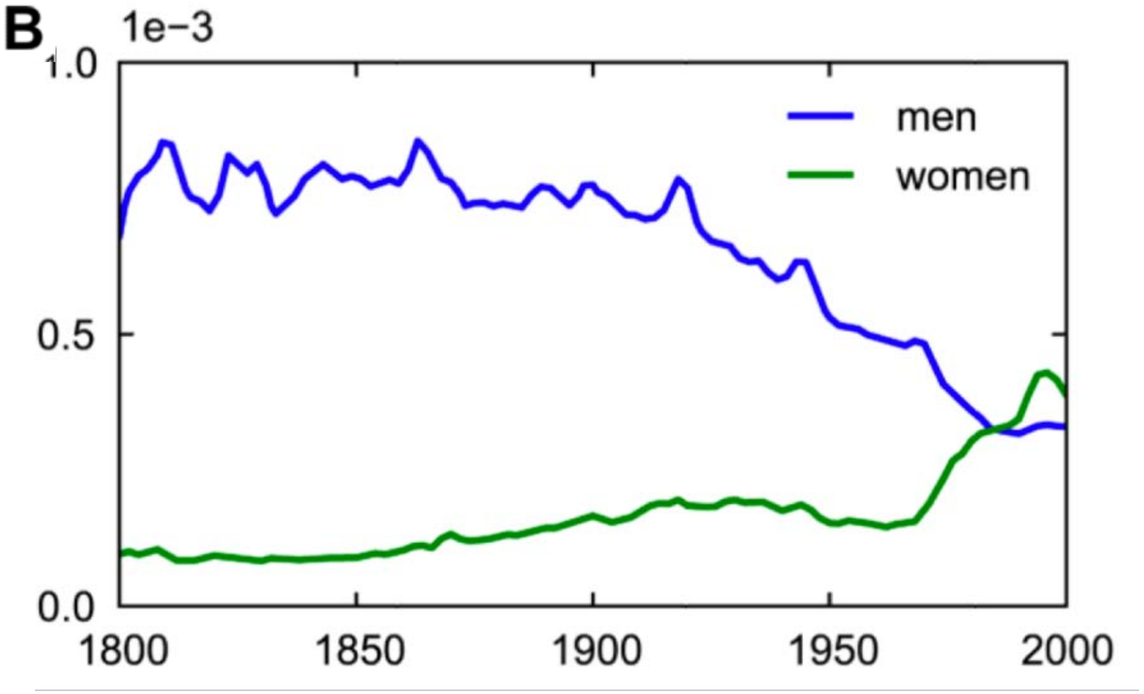


Source: Michel et al. (2011)

GOOGLE BOOKS STUDY

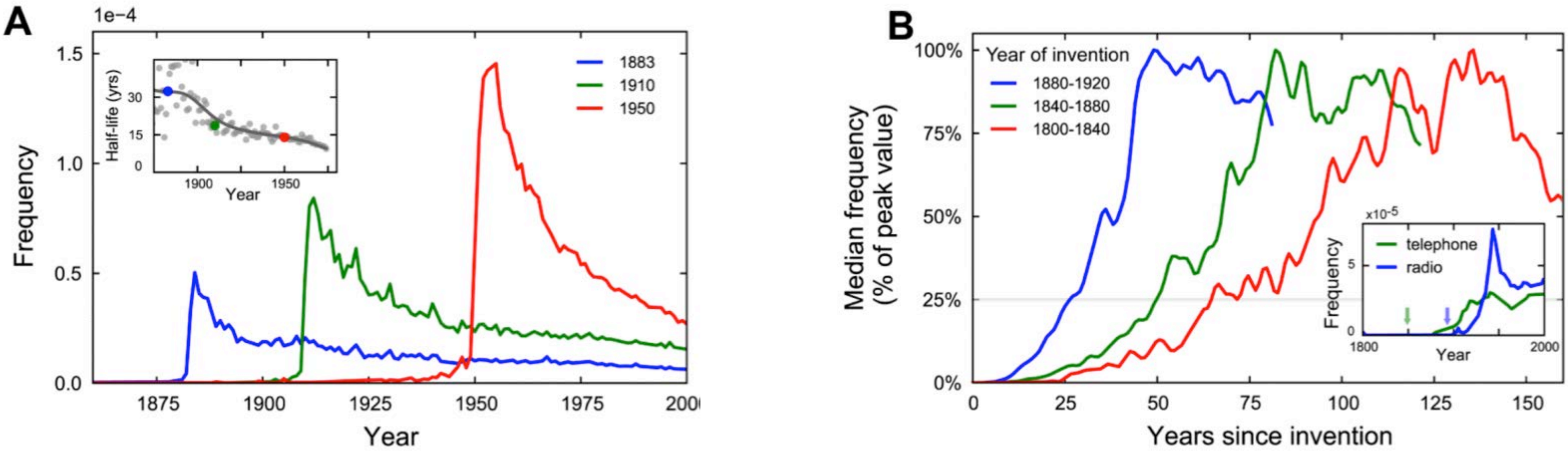
Selected Findings





Source: Michel et al. (2011)

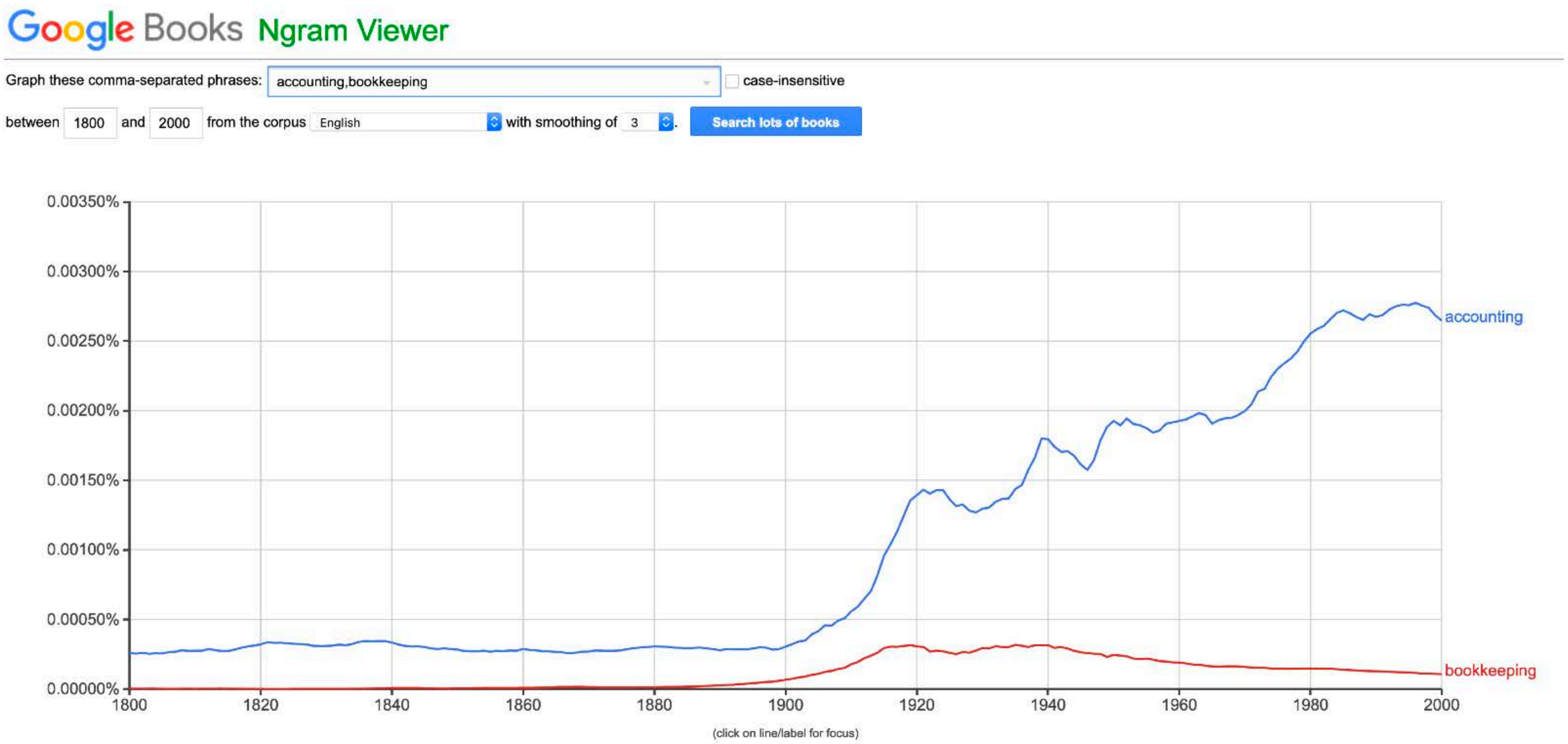
Selected Findings



Source: Michel et al. (2011)

PADERBORN UNIVERSITY

Try It Yourself at https://books.google.com/ngrams



GOOGLE FLU TRENDS STUDY

Google Search Terms

GOOGLE FLU TRENDS STUDY

Original Paper

nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virological and clinical data, including influenza-like illness (ILI) physician visits. The CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1–2-week reporting lag.

In an attempt to provide faster detection, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines[5] and over-the-counter drug sales[6]. About 90 million American adults are believed to search online for information about specific diseases or medical problems each year[7], making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted search queries submitted to a Swedish medical website (A. Hulth, G. Rydevik and A. Linde, manuscript in preparation), visitors to certain pages on a US health website[8], and user clicks on a search keyword advertisement in Canada[9]. A set of Yahoo search queries containing the words 'flu' or 'influenza' were found to correlate with virological and mortality surveillance data over multiple years[10].

Our proposed system builds on this earlier work by using an automated method of discovering influenza-related search queries. By processing hundreds of billions of individual searches from 5 years of Google web search logs, our system generates more comprehensive models for use in influenza surveillance, with regional and state-level estimates of ILI activity in the United States. Widespread global usage of online search engines may eventually enable models to be developed in international settings.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha\,\text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (http://www.cdc.gov/flu/weekly) was used to help build our models. For each of the nine surveillance regions of the United States, the CDC reported the average percentage of all outpatient visits to sentinel providers that were ILI-related on a weekly basis. No data were provided for weeks outside of the annual influenza season, and we excluded such dates from model fitting, although our model was used to generate unvalidated ILI estimates for these weeks.

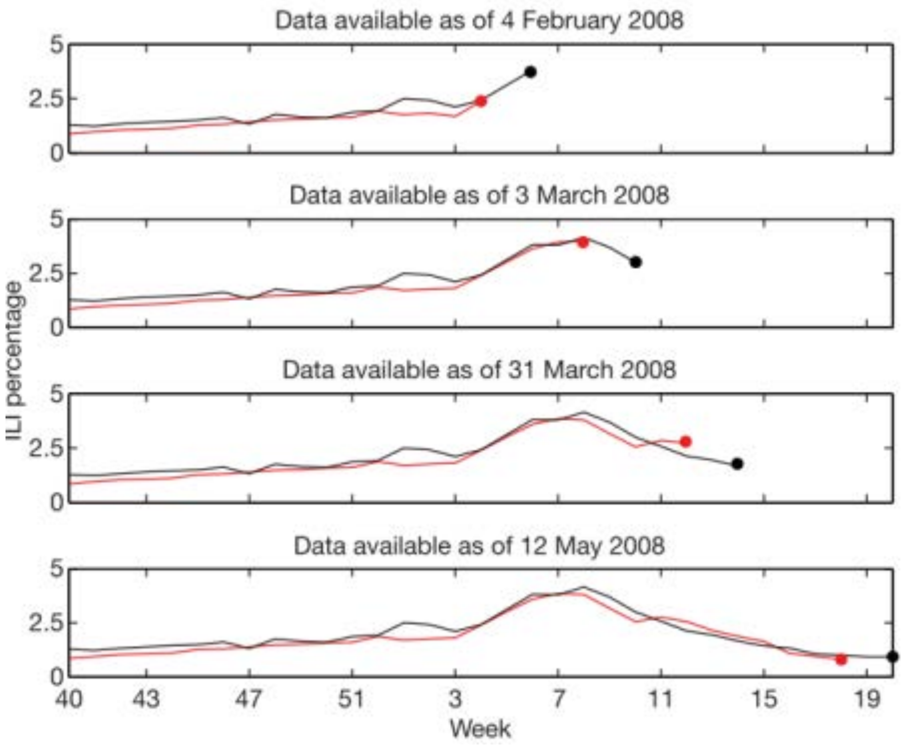
We designed an automated method of selecting ILI-related search queries, requiring no previous knowledge about influenza. We measured how effectively our model would fit the CDC ILI data in each region if we used only a single query as the explanatory variable, $Q(t)$. Each of the 50 million candidate queries in our database was separately tested in this manner, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. Our approach rewarded queries that showed regional variations similar to the regional variations in CDC ILI data: the chance that a random search query can fit the ILI percentage in all nine regions is considerably less than the chance that a random search query can fit a single location (Supplementary Fig. 2).

The automated query selection process produced a list of the highest scoring search queries, sorted by mean Z-transformed correlation across the nine regions. To decide which queries would be included in the ILI-related query fraction, $Q(t)$, we considered different sets of n top-scoring queries. We measured the performance of these models based on the sum of the queries in each set, and picked n such that we obtained the best fit against out-of-sample ILI data across the nine regions (Fig. 1).

[1]Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. [2]Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA.

GOOGLE FLU TRENDS STUDY

Google Flu Study

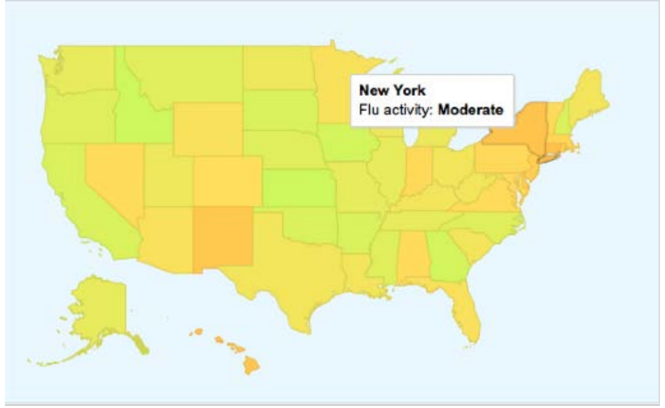


“The final model was validated on 42 points per region of previously untested data from 2007 to 2008, which were excluded from all previous steps. Estimates generated for these 42 points obtained a **mean correlation of 0.97** (min: 0.92, max: 0.99, n: 9 regions) with the CDC-observed ILI percentages.

Harnessing the collective intelligence of millions of users, **Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today**.

Source: Ginsberg et al. (2009)

GOOGLE FLU TRENDS STUDY



GOOGLE FLU TRENDS STUDY

PADERBORN UNIVERSITY

How to Forecast the Flu with Google Search Terms?

X Y

	Date	Location	“coughing”	„soar throat“	“cold”	…	ILI level
Observation 1	01.01.2019	Frankfurt	2321	3441	5513	…	0.020
Observation 2	01.01.2019	Berlin	1968	3201	4236	…	0.008
Observation 3	02.01.2019	Frankfurt	2331	3446	5657	…	0.021
…	…	…	…	…	…	…	

$$Y = f(X) + \varepsilon$$

GOOGLE FLU TRENDS STUDY

Response

FINAL FINAL

POLICYFORUM

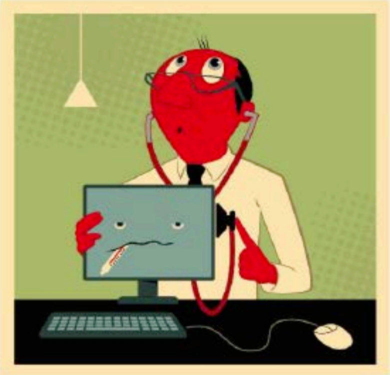
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?



The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy.

*Corresponding author. E-mail: d.lazer@neu.edu.

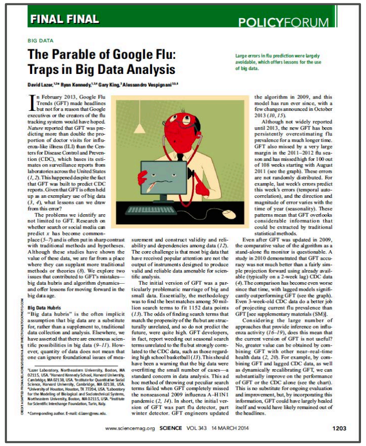
CREDIT: ADAPTED FROM AXEL KORES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014 1203

GOOGLE FLU TRENDS STUDY

PADERBORN UNIVERSITY

Traps in Big Data Analysis



FINAL FINAL POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

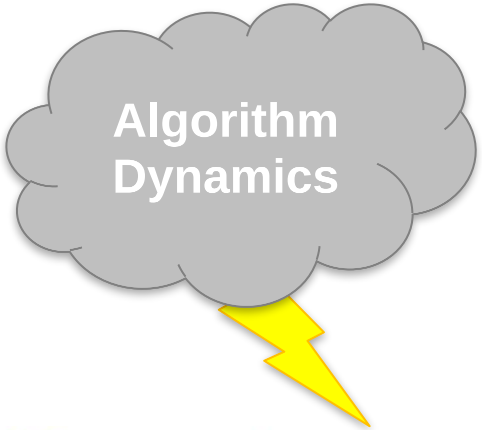
Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

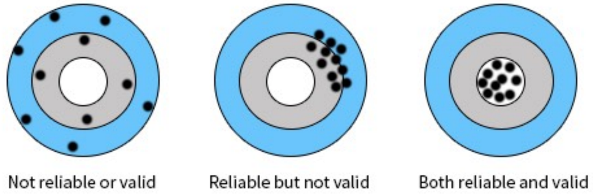
David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

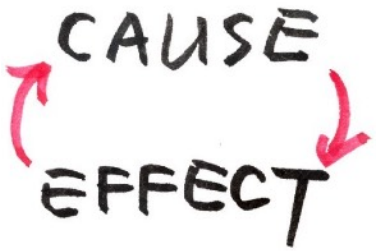
In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from …

… the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August …



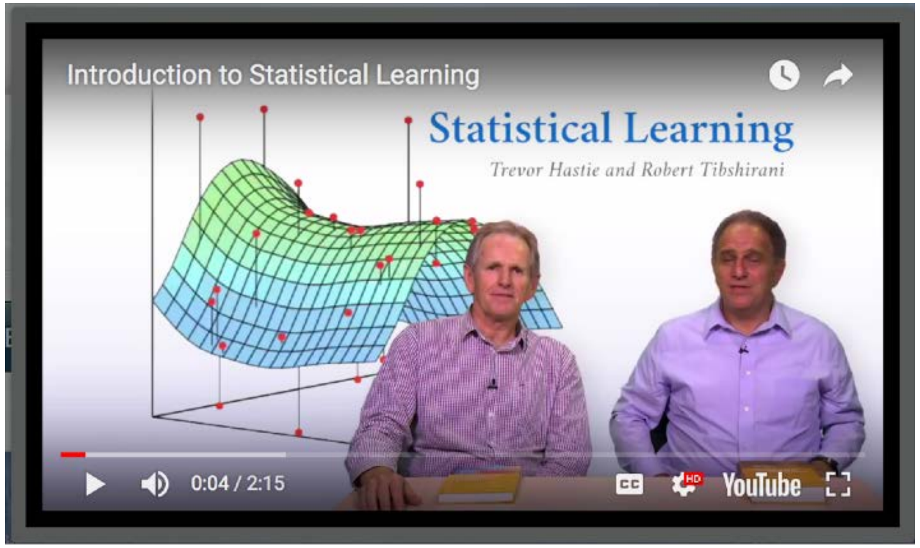




TWO PARADIGMS: STATISTICAL MODELING VS. MACHINE LEARNING

PADERBORN UNIVERSITY

What is Statistics?



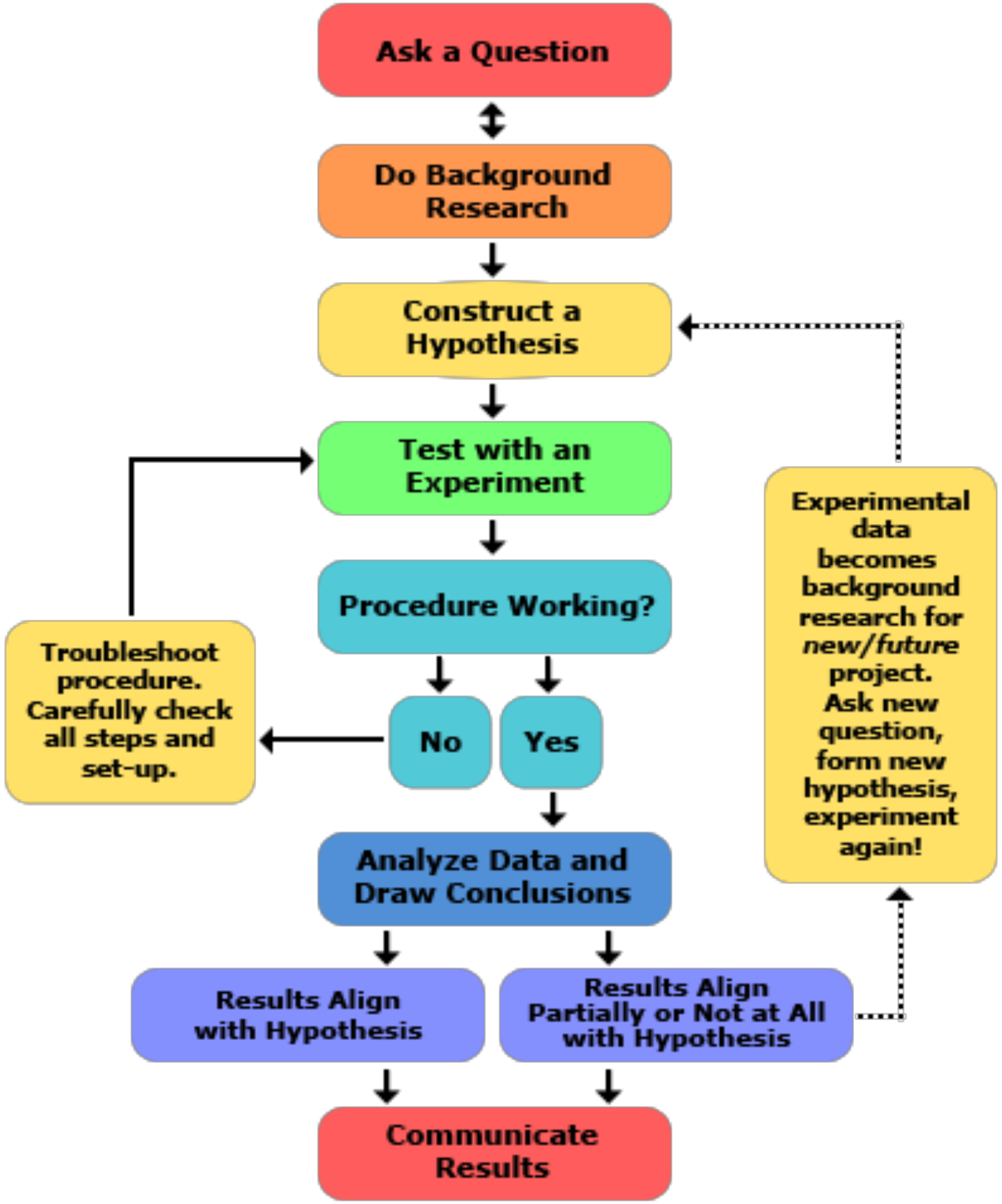
Source: https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about

> "Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and to understand "what the data says"." (Friedman et al., 2001)

> "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data" (Merriam-Webster)

THE STATISTICAL MODELING PARADIGM

The Scientific Method: Hypothetico-deductive



Source: https://www.sciencebuddies.org

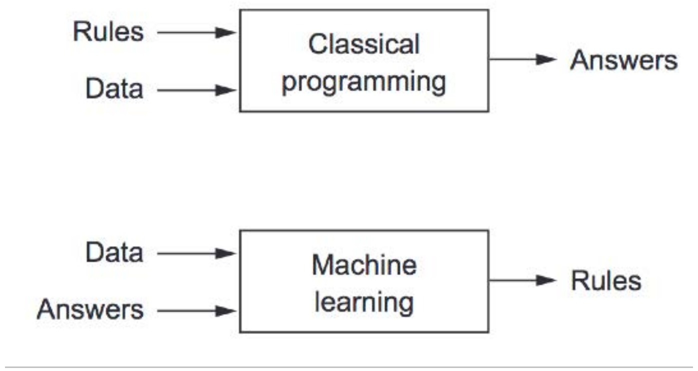
Example: Lady Tasting Tea

Source: https://www.youtube.com/watch?v=lgs7d5saFFc

THE MACHINE LEARNING PARADIGM

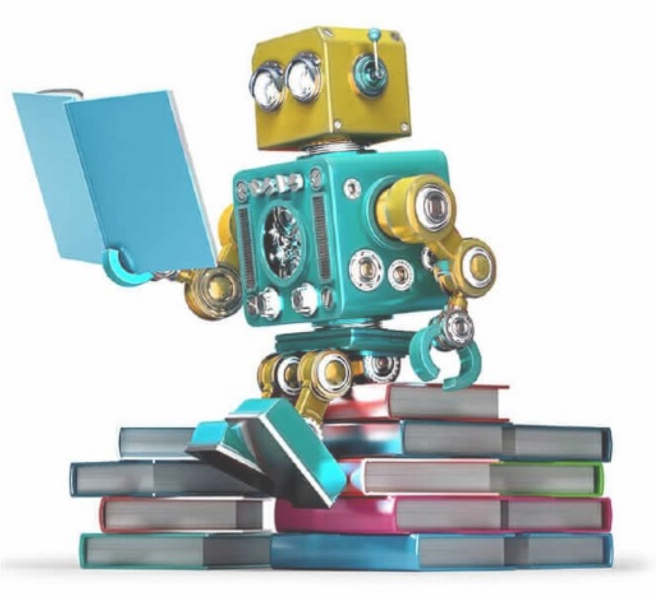
What is Machine Learning?

"The field of study that gives computers the ability to learn without being explicitly programmed." (Samuel, 1959)



Source: Chollet (2018)

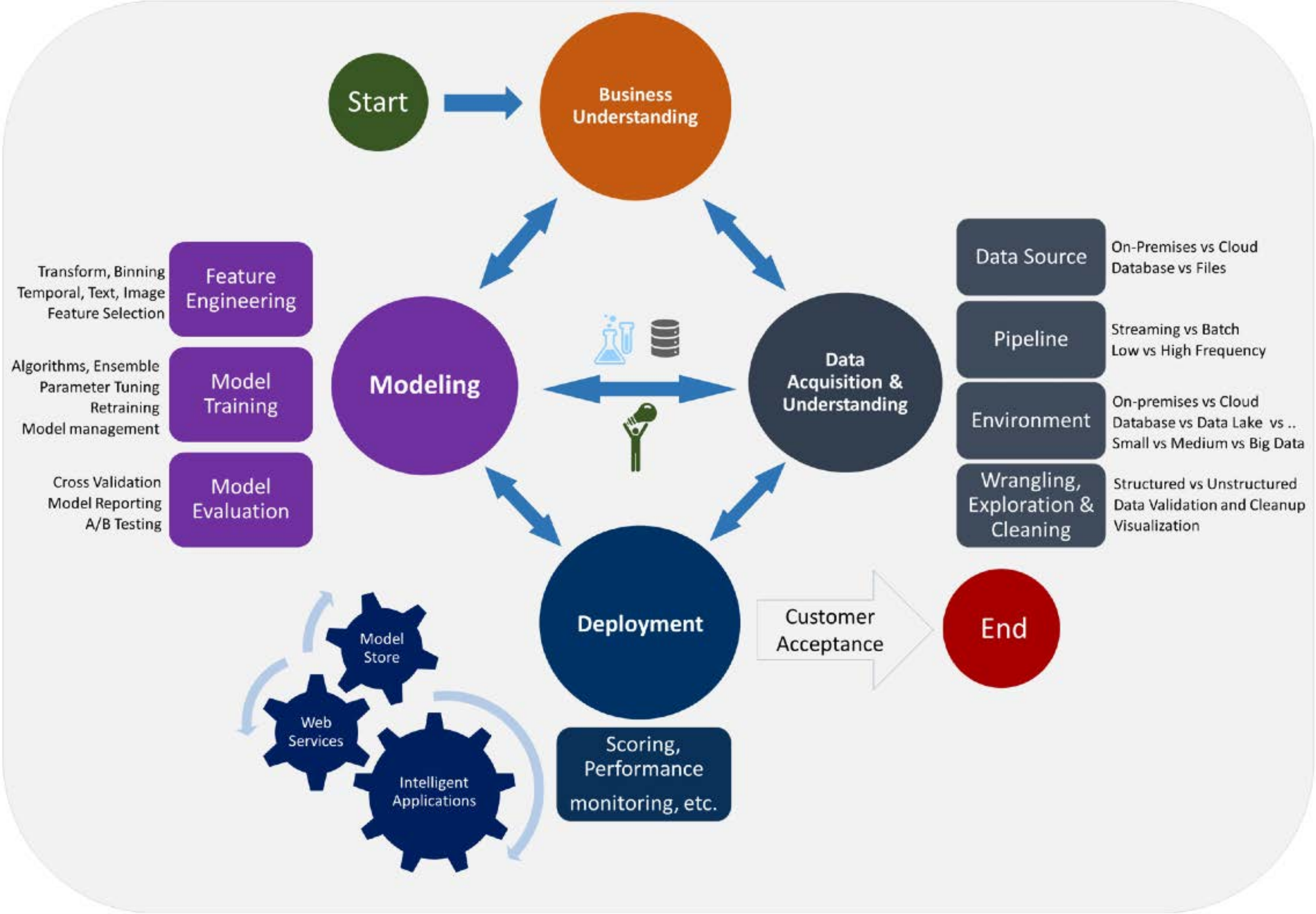
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997)



THE MACHINE LEARNING PARADIGM

PADERBORN UNIVERSITY

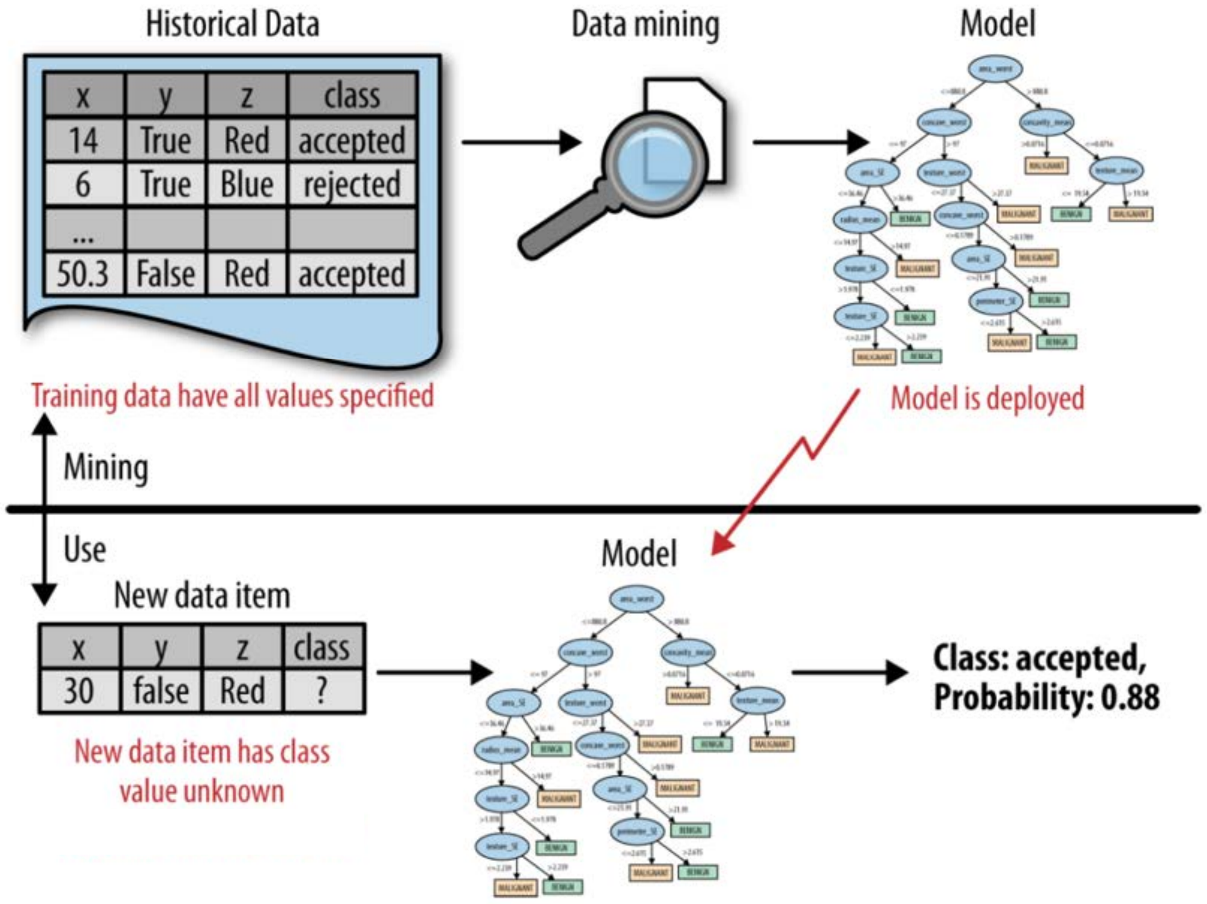
The Data Science Lifecycle: Data-driven, Inductive, Iterative



Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

PADERBORN UNIVERSITY

Example: Supervised Machine Learning for Credit Risk Scoring



Source: Provost & Fawcett (2013)

BANKRUPTCY PREDICTION:
THE SM VS. ML WAY

BANKRUPTCY PREDICTION

Governing Machine Learning in Governments

DANISH BUSINESS AUTHORITY

SKAT

Per Rådberg Nagbøl
Phd Student
ITU Copenhagen

EARLY WARNING EUROPE

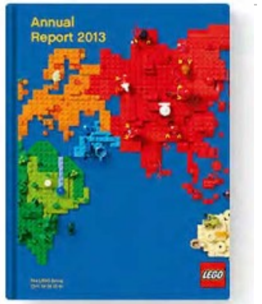
Early Warning Europe provides free, impartial and confidential counselling to companies in distress

BANKRUPTCY PREDICTION – THE SM WAY

Research Design

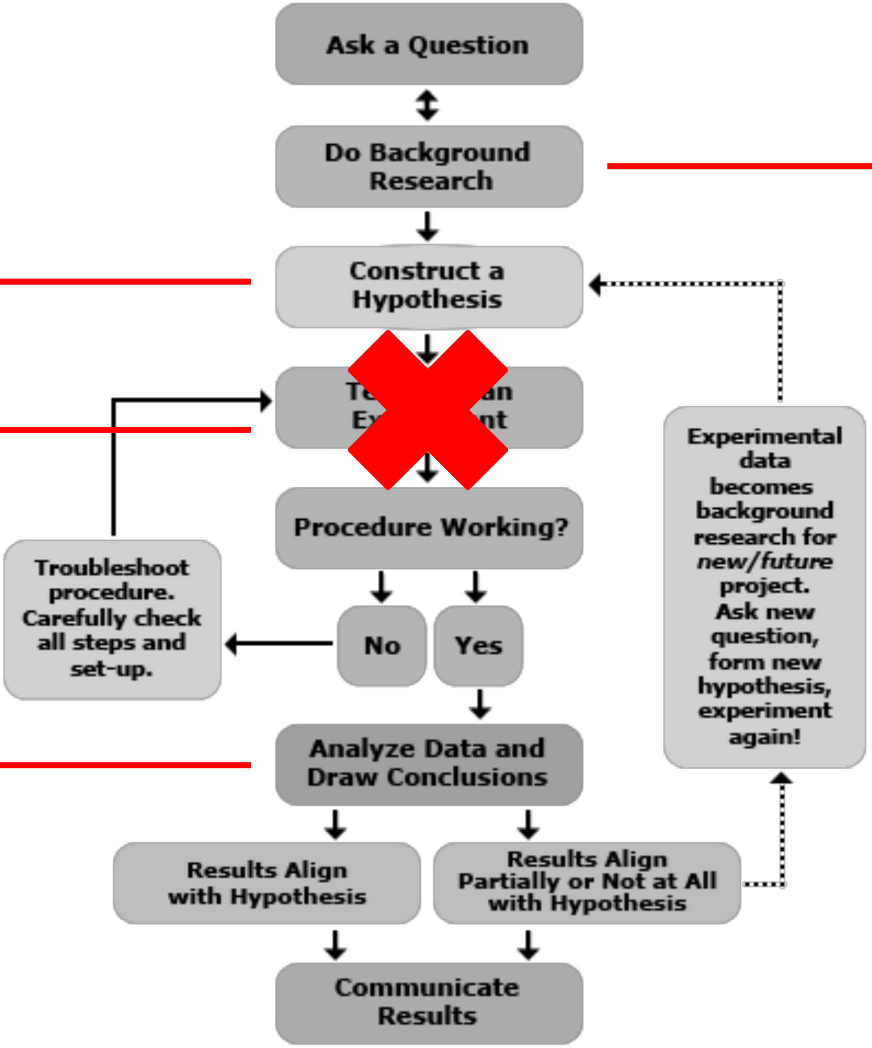
$$\text{(I)} \quad Z = .012X_1 + .014X_2 + .033X_3 + .006X_4 + .999X_5$$

where X_1 = Working capital/Total assets
X_2 = Retained Earnings/Total assets
X_3 = Earnings before interest and taxes/Total assets
X_4 = Market value equity/Book value of total debt
X_5 = Sales/Total assets
Z = Overall Index



STATSTIDENDE

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$



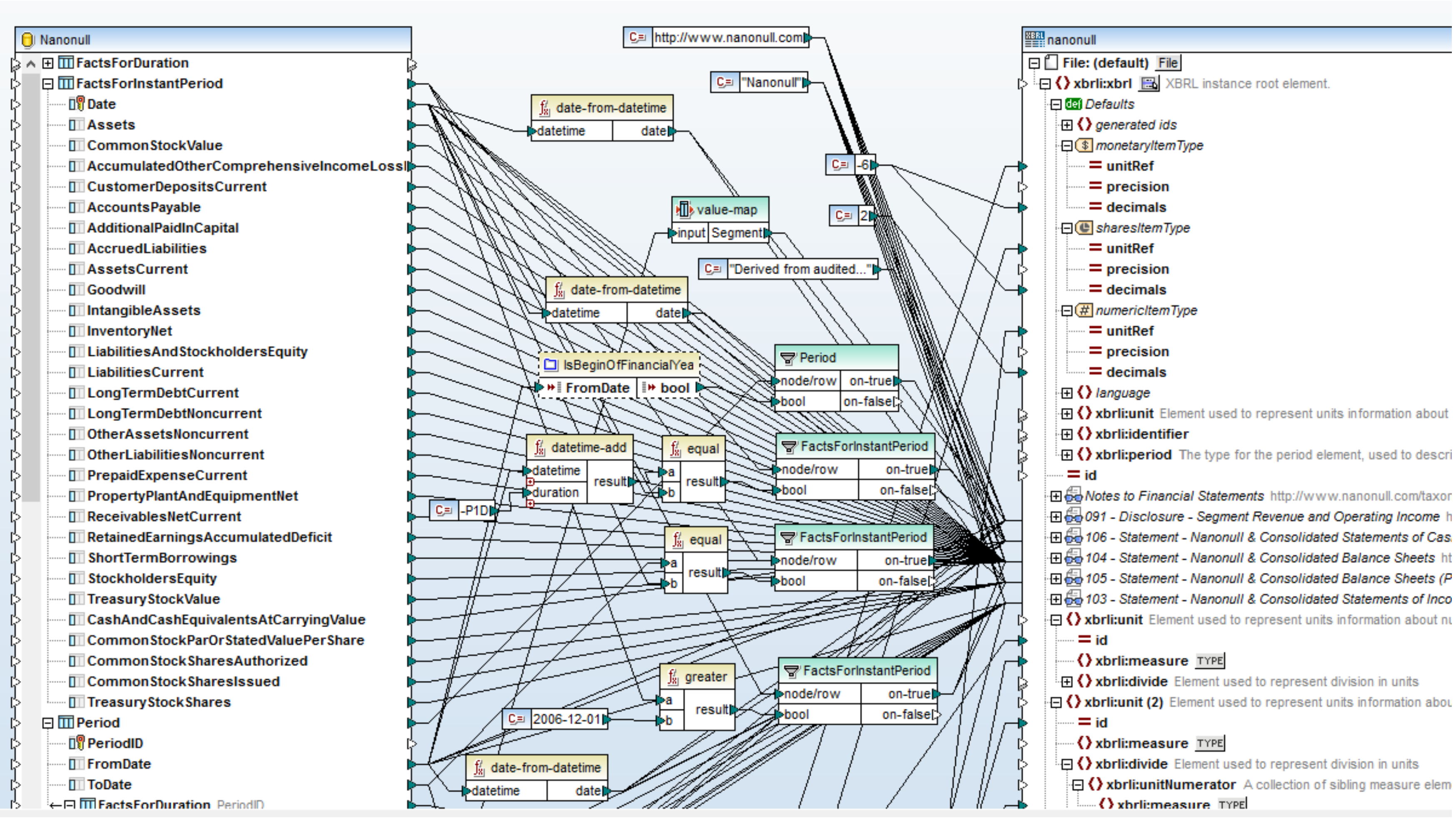
The Journal of FINANCE

FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY

Edward I. Altman*

BANKRUPTCY PREDICTION – THE SM WAY

Dataset: 62.000+ Annual Reports in XBRL from 2014



PADERBORN UNIVERSITY

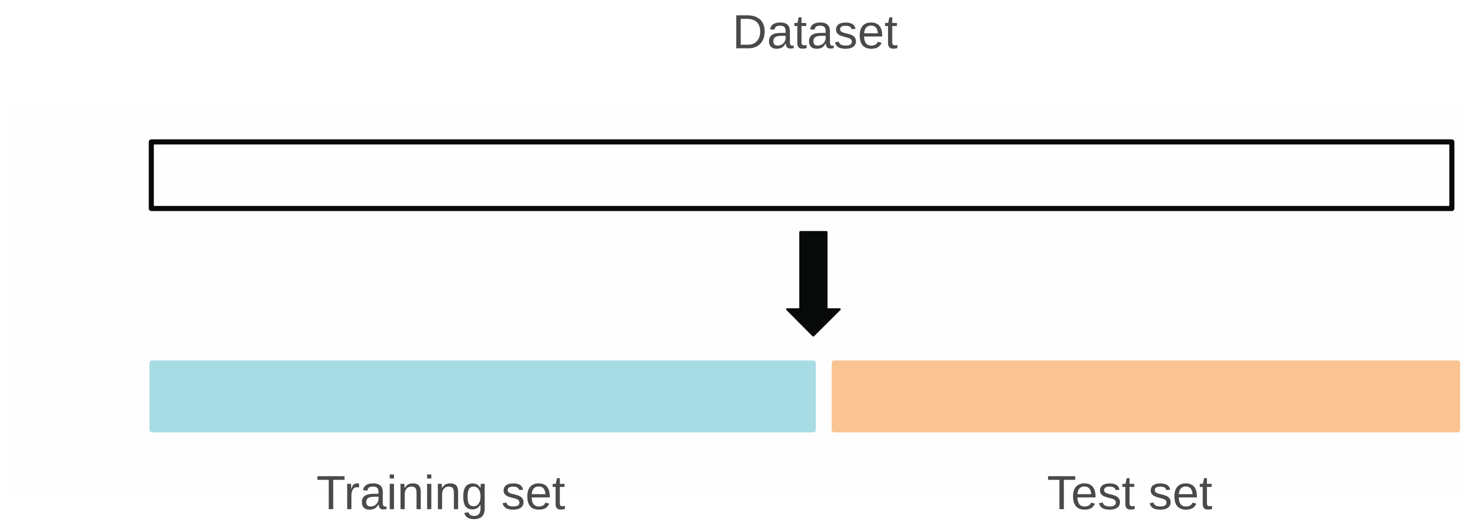
Results of Logistic Regression

```
==============================================
                       Dependent variable:
                   ---------------------------
                           bankruptcy
----------------------------------------------
Constant                -4.027*** (0.036)
X1                      -1.501*** (0.067)
X2                     -0.00000 (0.00004)
X3                      0.00001 (0.00005)
X4                     -0.00000 (0.00000)
----------------------------------------------
Observations                 50,377
Log Likelihood             -4,432.134
Akaike Inf. Crit.           8,874.268
==============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

X_1 = Working capital/Total assets
X_2 = Retained Earnings/Total assets
X_3 = Earnings before interest and taxes/Total assets
X_4 = Market value equity/Book value of total debt
X_5 = Sales/Total assets

PADERBORN UNIVERSITY

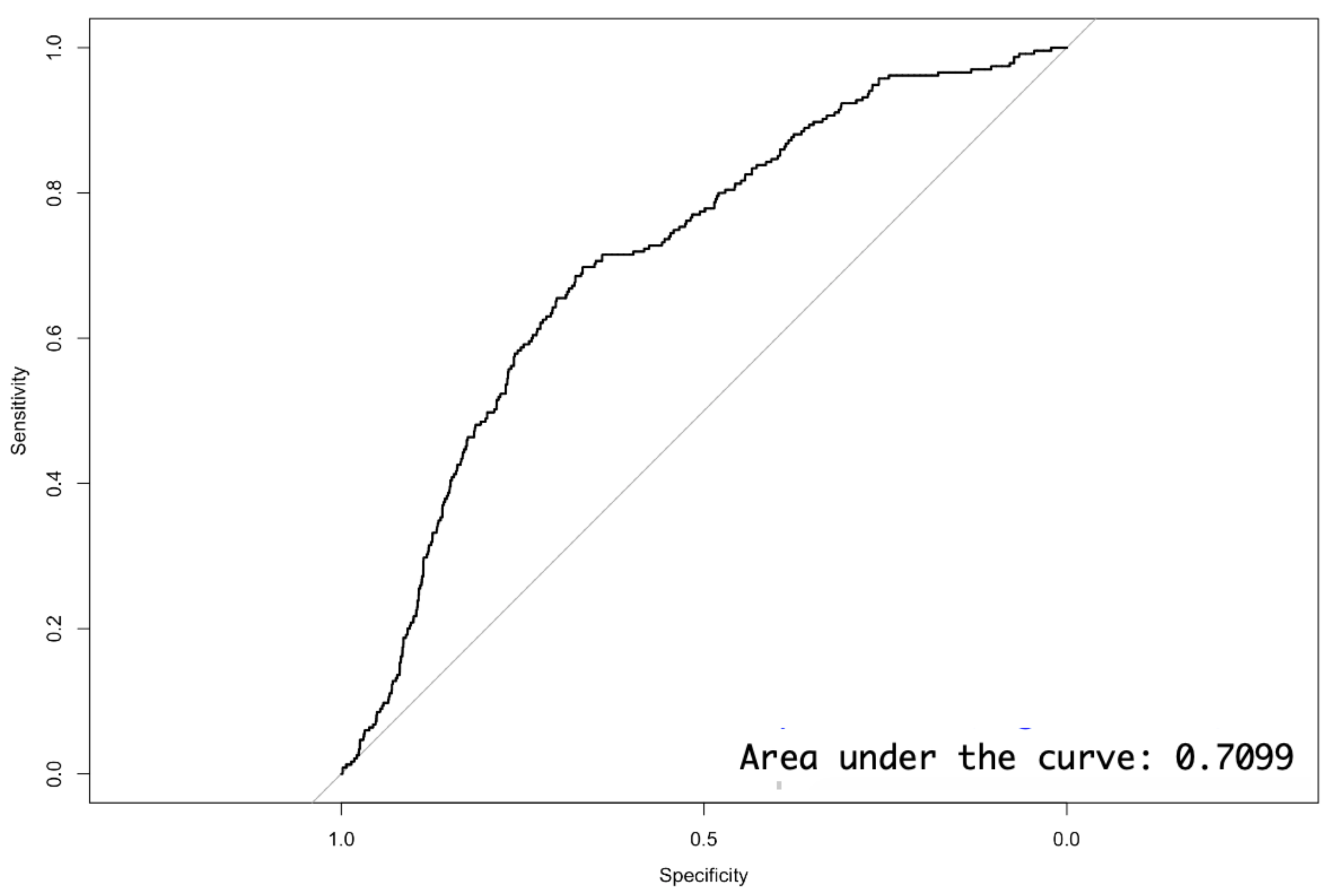
Training and Test Sets



Source: James et al. (2013)

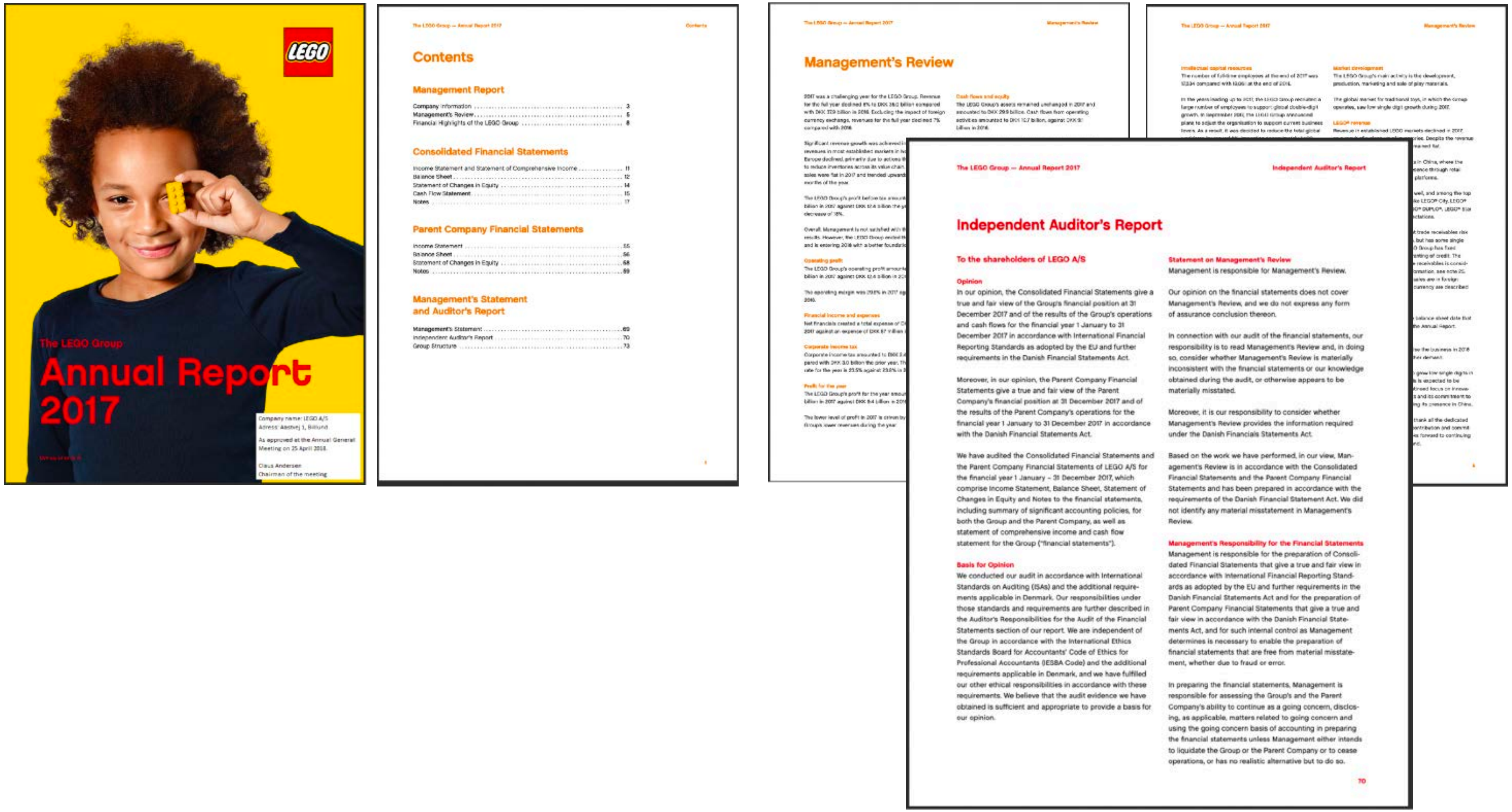
PADERBORN UNIVERSITY

Predictive Accuracy of Logistic Regression on Test Set



BANKRUPTCY PREDICTION – THE ML WAY

Going Beyond Financial Ratios (i.e., reading 62.000 annual reports)



The Bag-Of-Words (BOW) Model

Lorem Ipsum

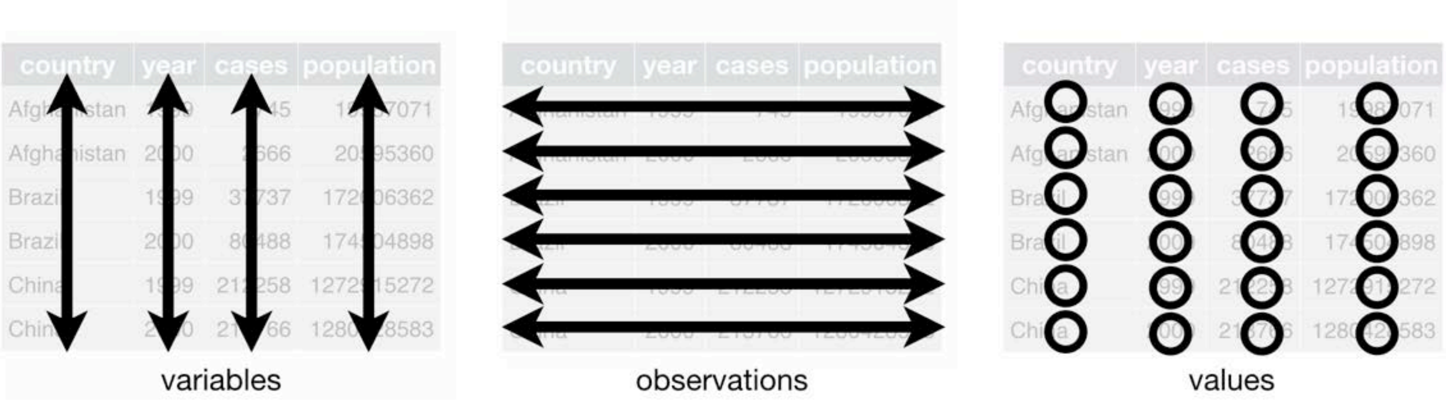
"Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit..."
"There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain..."

What is Lorem Ipsum?

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Why do we use it?

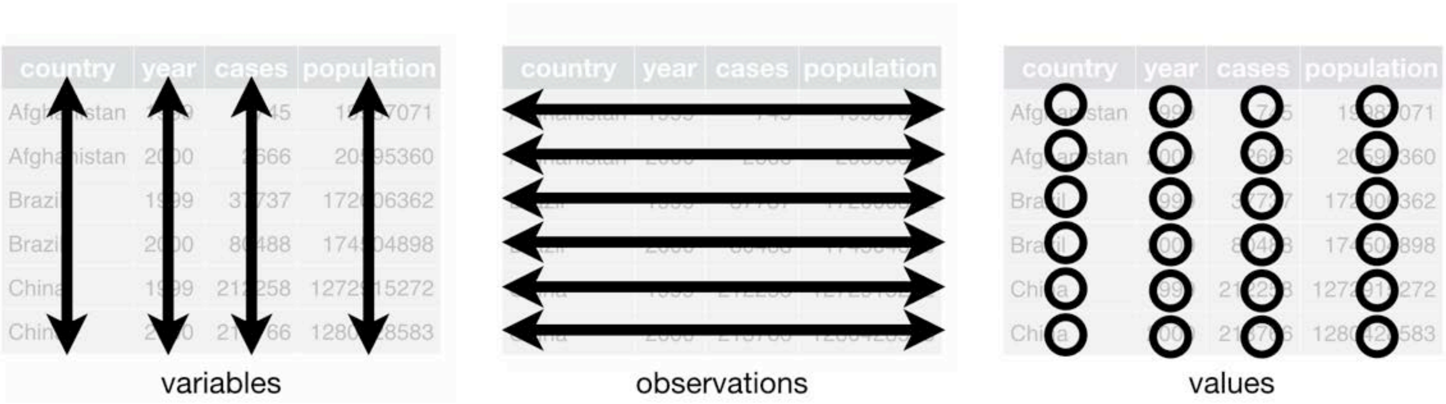
It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).



The Bag-Of-Words (BOW) Model

- Treat every document as a unordered set of words
- Ignore word order, sentence structure, and punctuation

- Tidy data frame:
  - Every document is an observation (row)
  - Every word is a variable (column)
  - The presence of a word in a document (aka. token) is represented by the cell values



BANKRUPTCY PREDICTION – THE ML WAY

The Bag-Of-Words (BOW) Model

$$\overbrace{\phantom{\text{word 1 word 2 word 3 word 4 word 5 ...}}}^{X} \qquad Y$$

	"word 1"	"word 2"	"word 3"	"word 4"	"word 5"	…	Bankruptcy?
Lego	0	0	0	2	1	…	**No**
Maersk	0	0	0	2	0	…	**No**
Jysk Fragt	0	1	1	1	0	…	**Yes**
…	…	…	…	…	…	…	

$$Y = f(X) + \varepsilon$$

Tree-based Classification Algorithms

BANKRUPTCY PREDICTION – THE ML WAY

Example: Loan Default

BANKRUPTCY PREDICTION – THE ML WAY

Example: Loan Default

Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

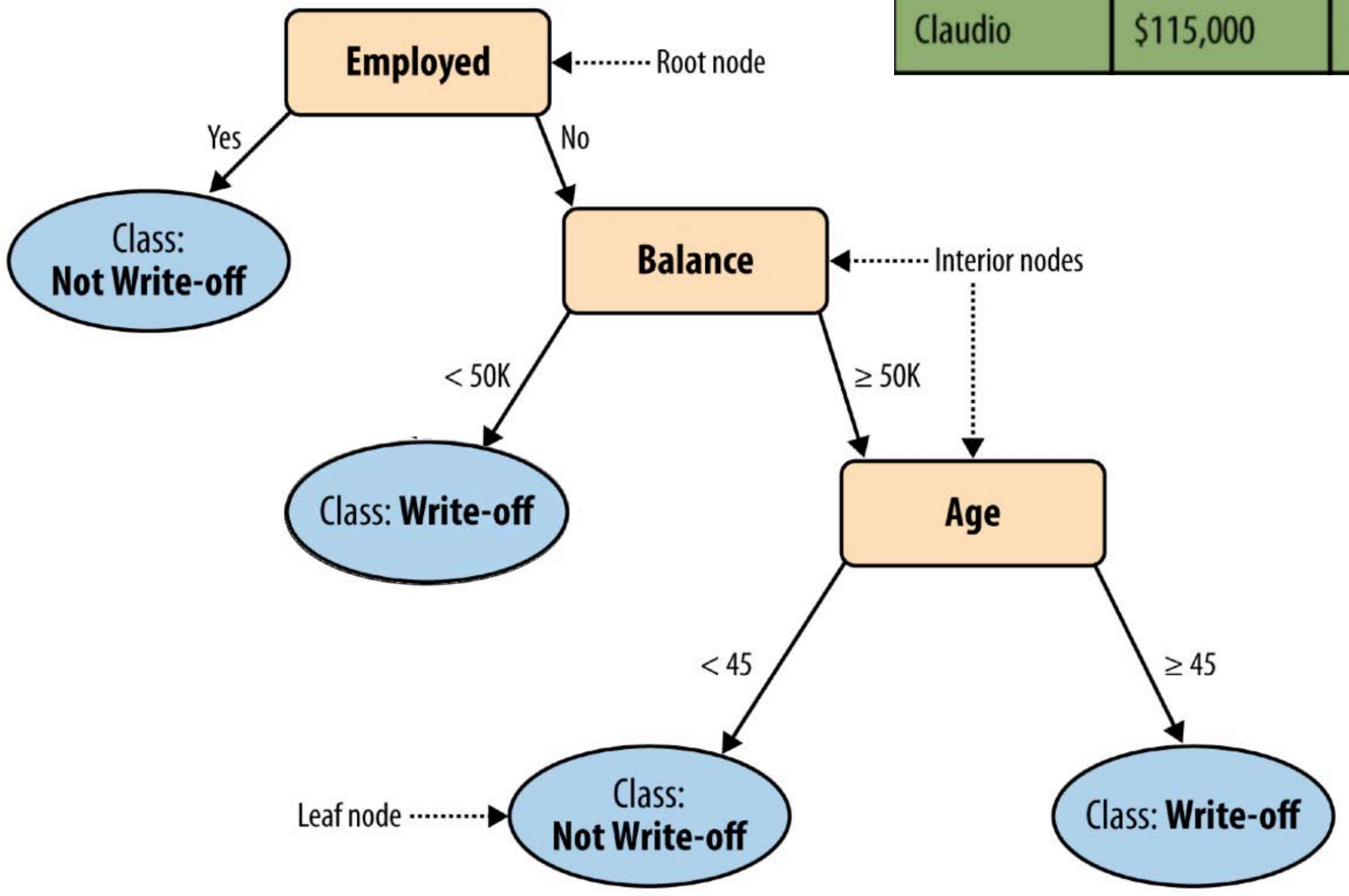
This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

Source: Provost & Fawcett (2013)

BANKRUPTCY PREDICTION – THE ML WAY

Example: Loan Default

Claudio	$115,000	40	no



Source: Provost & Fawcett (2013)

PADERBORN UNIVERSITY

The CART Algorithm: Top-down, Greedy Search

- It is computationally infeasible to consider all possible sequences and combinations of splits.

- Instead, do **recursive binary partitioning**
  - **Top-down**: Start with zero splits and successively partition the feature space into two parts.
  - **Greedy**: At each step, make the best possible split at that particular step (i.e., the split with the highest information gain, i.e., reduction in entropy).
  - Stop when some condition (e.g., minimal number of observations in one leaf) is met.

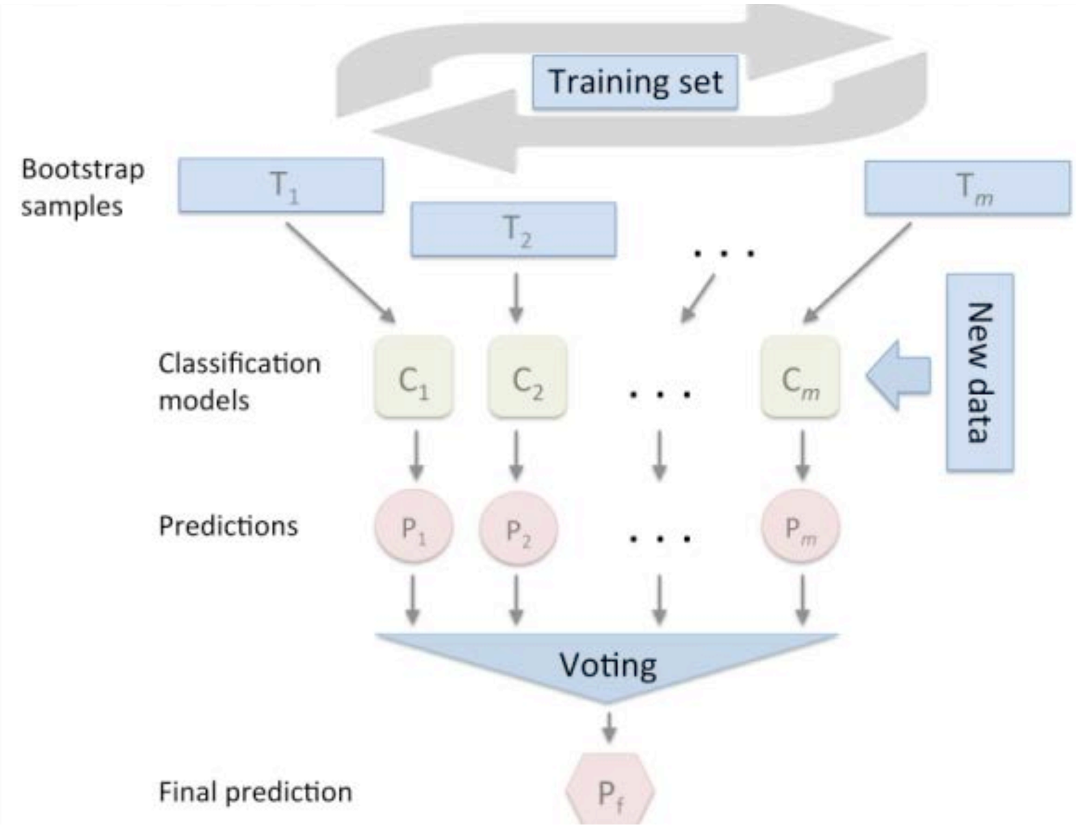
- That is, we consider all predictors $X_1, . . .,X_p$, and all possible split points s for each of the predictors, and then choose the predictor and split point with the highest information gain at each step.

Source: Provost & Fawcett (2013)

Random Forests

Bootstrap Aggregation (Bagging)

- A way to reduce overfitting of a machine learning algorithm is to take many training sets from the population, build a separate model on each training set, and average the resulting predictions.



Source: James et al. (2013)

From Bagging to Random Forest

- **Following the idea of bagging**, we draw multiple random samples (bootstrap samples) from the training data and create a decision tree on each sample
  - Typically, 2/3 of the rows in the training set

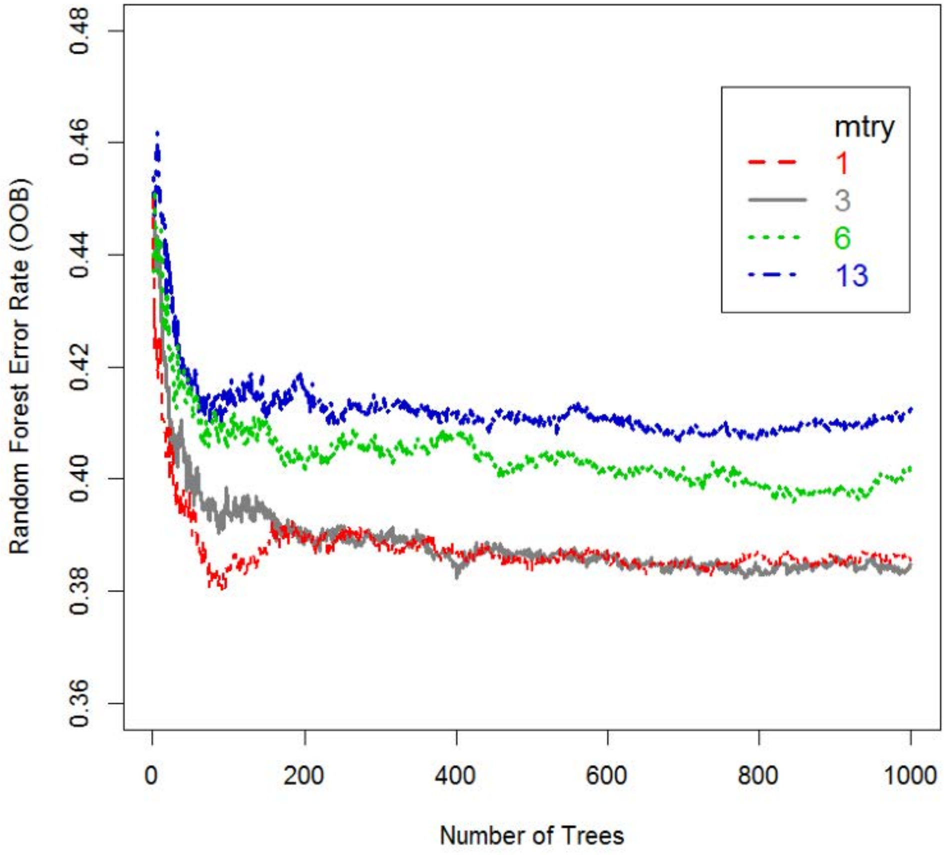
- However, in Random Forests we **allow only a random subset (*m*) of all the predictors (*p*) to be used at each split of the decision tree**
  - Typically, *m = SQRT(p)*

- **Why** does this work?
  - In bagging, if there is one strong predictor, all the trees will use this predictor in the top split
    ➔ all of the trees will look quite similar to each other
    ➔ their predictions will be highly correlated
    ➔ only a little bit of variance will be removed

Source: James et al. (2013)

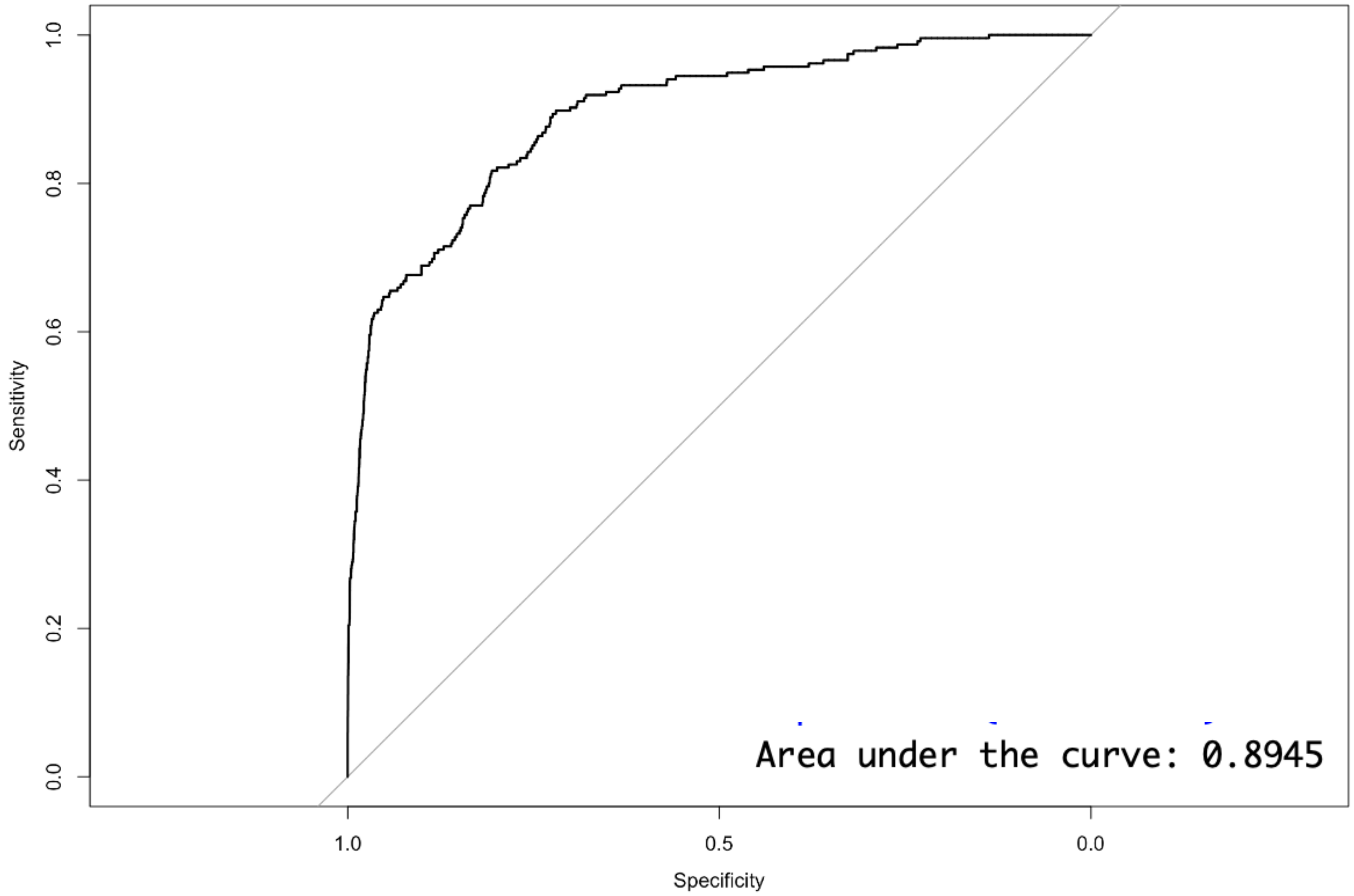
PADERBORN UNIVERSITY

A Random Forest has many Trees



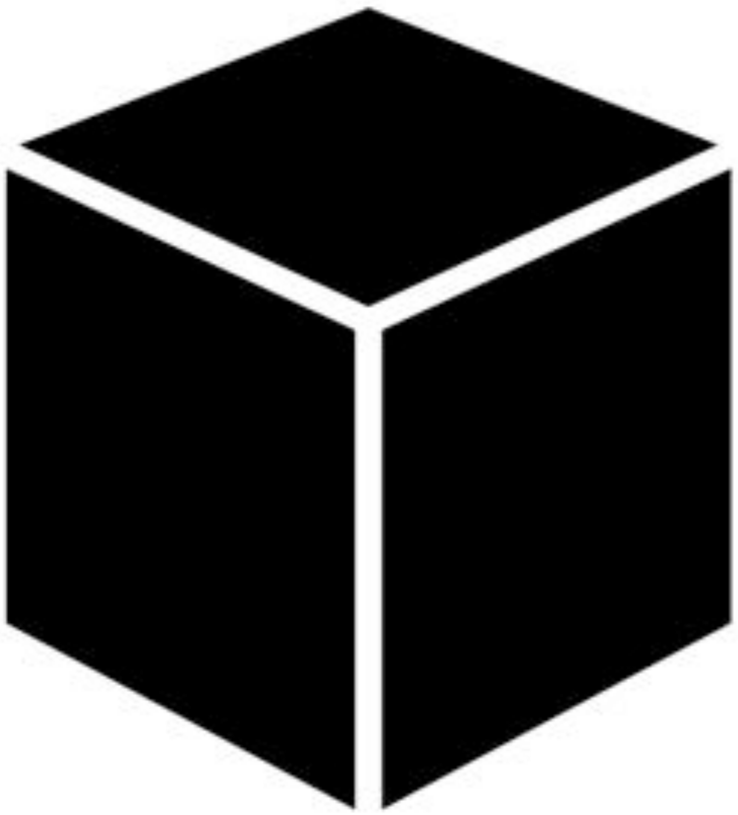
BANKRUPTCY PREDICTION – THE ML WAY

Predictive Accuracy of Random Forest with BOW on Test Set



PADERBORN UNIVERSITY

How Does the Model Look Like?



Black Box

PADERBORN UNIVERSITY

Next Steps

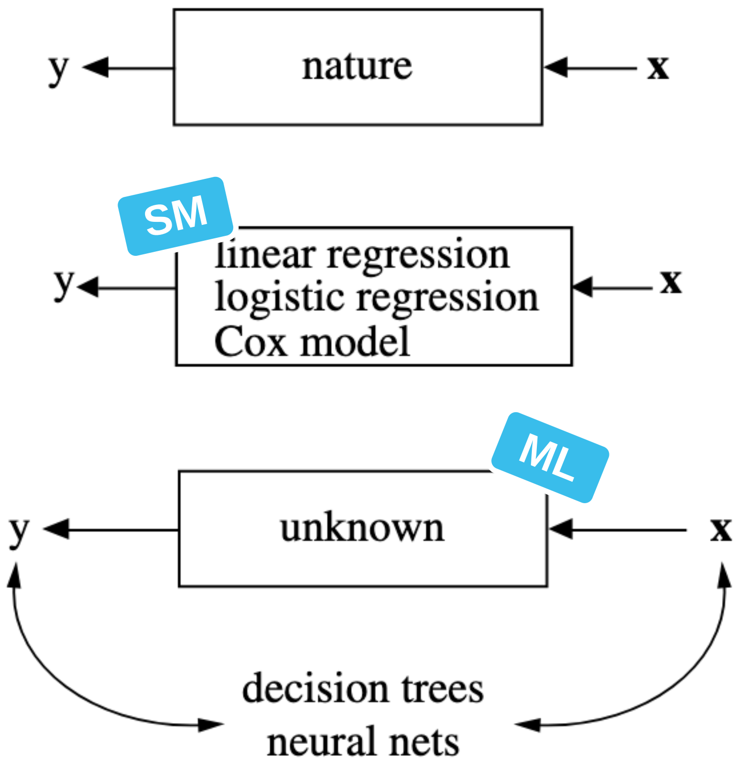
- Include more text from annual reports (e.g., management review, management’s statement, CSR)
- Quantify the informativeness of different sections of annual reports
- Use artificial neural networks to better capture syntax and semantics of text
- Try to open the black box of machine learning algorithms

REFLECTIONS

PADERBORN UNIVERSITY

Similarities and Differences of Statistical Modeling and Machine Learning

- **Data**
  - Both work with almost the same data structures
  - ML wrangles messy data until it fits into rows and columns
- **Methods**
  - Both use regression and classification techniques
  - SM applies mainly additive linear models
  - ML uses on non-linear methods that work on high-dimensional data
  - ML makes use of unsupervised techniques for data preparation
- **Process**
  - SM is theory/hypothesis-driven (no fishing for correlations!)
  - ML is mainly data-driven
- **Outputs**
  - SM: focus on causal explanations
  - ML: focus on predictive accuracy (on unseen test data!)



Source: Breiman (2001)

REFLECTIONS

PADERBORN UNIVERSITY

Paper

European Journal of Information Systems (2016), 1–14

www.palgrave-journals.com/ejis/

ISSUES AND OPINION

Utilizing big data analytics for information systems research: challenges, promises and guidelines

Oliver Müller[1], Iris Junglas[2], Jan vom Brocke[1] and Stefan Debortoli[1]

[1]*Institute of Information Systems, University of Liechtenstein, Vaduz, Liechtenstein;* [2] *College of Business, Florida State University, Tallahassee, U.S.A.*

Correspondence: Oliver Müller, Institute of Information Systems, University of Liechtenstein, Fürst-Franz-Josef-Strasse, Vaduz 9490, Liechtenstein
Tel: +423 265 13 00;
Email: oliver.mueller@uni.li

Abstract
This essay discusses the use of big data analytics (BDA) as a strategy of enquiry for advancing information systems (IS) research. In broad terms, we understand BDA as the statistical modelling of large, diverse, and dynamic data sets of user-generated content and digital traces. BDA, as a new paradigm for utilising big data sources and advanced analytics, has already found its way into some social science disciplines. Sociology and economics are two examples that have successfully harnessed BDA for scientific enquiry. Often, BDA draws on methodologies and tools that are unfamiliar for some IS researchers (e.g., predictive modelling, natural language processing). Following the phases of a typical research process, this article is set out to dissect BDA's challenges and promises for IS research, and illustrates them by means of an exemplary study about predicting the helpfulness of 1.3 million online customer reviews. In order to assist IS researchers in planning, executing, and interpreting their own studies, and evaluating the studies of others, we propose an initial set of guidelines for conducting rigorous BDA studies in IS.
European Journal of Information Systems advance online publication, 9 February 2016; doi:10.1057/ejis.2016.2

Keywords: big data; analytics; data source; methodology; information systems research

Why worry about big data analytics (BDA)?

The proliferation of the web, social media, mobile devices, and sensor networks, along with the falling costs for storage and computing resources, has led to a near ubiquitous and ever-increasing digital record of computer-mediated actions and communications – a record that has been termed 'big data'. Studies agree (e.g., Hilbert & López, 2011; IDC, 2011, 2014) that the volume of data being generated and stored today is growing exponentially (Kitchin, 2014) – but, big data is not just about volume (Constantiou & Kallinikos, 2015; Yoo, 2015). Accompanying the increased size of data sets is a growing variety of data sources and formats. In fact, IDC (2011) claims that more than 80% of all digital content is unstructured and that two-third is generated by individuals, rather than enterprises. The velocity of data has increased too, resulting in reduced latency between the occurrence of a real-world event and its mirroring digital footprint (vom Brocke *et al*, 2014).

As volume, variety, and velocity (Laney, 2001) increase, the veracity of data is drawn into question (Bendler *et al*, 2014). Unlike research data collected with a specific research question in mind and measured using validated instruments, big data often just 'happens'. Private and government organisations increasingly collect big data without a concrete purpose

Received: 18 November 2014
Revised: 7 December 2015
Accepted: 1 January 2016

THE END

Prof. Dr. Oliver Müller
Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics
Universität Paderborn
Warburger Str. 100, 33098 Paderborn
R: Q2.457
E: oliver.mueller@uni-paderborn.de
T: +49-5251-605245
W: https://wiwi.uni-paderborn.de/dep3/mueller/

- Michel, J. B., Shen, Y. K., Aiden, A. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176-182.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229.
- Mitchell, T. M. (1997). Machine learning. McGraw Hill.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001).The elements of statistical learning. Springer.
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly Media.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
- Müller, O., Junglas, I., Brocke, J. V., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, 25(4), 289-302.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

- Icons: Depb Dew, Jemis Mali from the Noun Project