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
Jean-Baptiste Michel, Erez Lieberman Aiden:

What we learned from 5 million books

TEDxBoston 2011 • 14:08 • Posted Sep 2011
Subtitles available in 37 languages

View transcript

1,278,448 TOTAL VIEWS
Quantitative analysis of culture using millions of digitized books

Jean-Baptiste Michel1,2,3,4,5, Yvan Kui Shen1, Anna P. Abiador6, Adrien Veres2, Matthew K. Gray2, The Google Books Team3, Joseph P. Pickhardt1, Dale Hoiberg2, Dan Clancy2, Peter Norvig1, Jon Orwant7, Steven Pinker8, Martin A. Nowak9,11,12, and Ernie Lieberman Aiden10, Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
2 Harvard Medical School, Boston, Massachusetts, 02115, USA.
3 Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138, USA.
4 Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA.
5 Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts 02138, USA.
6 Google, Inc., Mountain View, California, 94043, USA.
7 Houghton Mifflin Harcourt, Boston, Massachusetts 02115, USA.
8 Encyclopaedia Britannica, Inc., Chicago, Illinois 60654, USA.
9 Dept of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
10 Dept of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.
11 Dept of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.
12 Broad Institute of Harvard and MIT, Harvard University, Cambridge, Massachusetts 02138, USA.
13 School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.
14 Harvard Society of Fellows, Harvard University, Cambridge, Massachusetts 02138, USA.
15 Laboratory of Large, Harvard University, Cambridge, Massachusetts 02138, USA.

Abstract
We constructed a corpus of digital 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, consensus, and historical epidemiology. "Culturomics" extends the boundaries

1 For correspondence, please contact: jeh@helixstanford.com (JM), mnowak@hms.harvard.edu (MA).
GOOGLE BOOKS STUDY

Data and Methods

Source: Michel et al. (2011)
GOOGLE BOOKS STUDY

Selected Findings

Source: Michel et al. (2011)
Selected Findings

A

B

Source: Michel et al. (2011)
Try It Yourself at https://books.google.com/ngrams
Google Search Terms

Google FLU TRENDS STUDY
GOOGLE FLU TRENDS STUDY

Original Paper

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg, Matthew H. Mohr, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

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. In addition to seasonal influenza, a new strain of influenza virus against which we have no previous immunity exists and that demonstrates human-to-human transmission would result in a pandemic with millions of fatalities. A timely detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza.

In this study, we develop an approach to measure weekly influenza activity for the entire world each day. Here we present a method of analyzing large numbers of Google search queries to track influenza-like illnesses in 166 countries. Because the relative frequency of certain queries is highly associated with the percentage of physician visits to 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 an reporting lag of about one week. This approach may enable us to use search queries to detect influenza epidemics across a large population of web users.

Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance System (EIS), rely on both virologic and clinical data, including influenza-like illnesses (ILI) diagnoses. The CDC provides national and regional data from these surveillance systems on a weekly basis, typically with a 2-3 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 cell volume to telephone traffic analysis (1) and even the counter dog diet (2). About 90 million Americans use Google searches every day, and are believed to search online for information about specific diseases or medical problems each week, making web search terms a uniquely reliable source of information about health trends. Previous attempts at using online activity for influenza surveillance have involved search queries submitted to a flyweight malware (A. Hult, G. Freyhold and A. Lind, manuscript in preparation). We therefore consider certain pages on a U.S. health website (3), and user clicks on a search argument given a search query (4). A set of “Vinci” search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales (5).

Our proposed system builds on this earlier work by using an automated method of detecting influenza-related search queries. By processing hundreds of millions of individual searches from users of Google web search, we can estimate trends in self-reported influenza-like illnesses on a regional scale. We track the percentage of all Google searches containing the words “flu” or “influenza” that are made by users of the United States. We use this data to estimate the influenza season activity in the United States. We therefore present a tool to monitor influenza activity in the United States and worldwide, and we provide a method to detect an influenza epidemic or pandemic in real-time.

To verify the accuracy of the model, we compare the model’s predictions with the CDC’s ILI surveillance data (6). The model is able to accurately predict the onset of influenza activity in the United States and worldwide, with a high degree of sensitivity and specificity. The model is therefore a valuable tool for detecting influenza epidemics and pandemics in real-time.

Public health-related findings data from the CDC’s Influenza Surveillance Network (7) was used to help collect our results. For each of the states, we used the Mississippi region as the region of interest, although we model with the national estimate for ILI estimates for these regions.

We designed an automated method of selecting ILI-related search queries, requiring no previous knowledge about influenza. We trained and tested our model to detect the CDC’s ILI data in each region. If we used a single query for each region, our model would not be effective. Instead, we used a combination of multiple queries to detect the ILI data in each region. This method is highly effective and can detect influenza activity in real-time. The model is able to detect influenza activity in each region with high accuracy.

We suggest that this method can be used to detect influenza activity in real-time, and that it can be used to predict the onset of influenza activity in real-time. The model is therefore a valuable tool for detecting influenza epidemics and pandemics in real-time.

References:
1. A. Hult, G. Freyhold and A. Lind, manuscript in preparation.
2. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
3. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
4. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
5. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
6. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
7. J. Vinci, search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data on multiple scales.
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.

“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.

Source: Ginsberg et al. (2009)
How to Forecast the Flu with Google Search Terms?

### Google Flu Trends Study

<table>
<thead>
<tr>
<th>Observation</th>
<th>Date</th>
<th>Location</th>
<th>“coughing”</th>
<th>„soar throat“</th>
<th>“cold”</th>
<th>...</th>
<th>ILI level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01.01.2019</td>
<td>Frankfurt</td>
<td>2321</td>
<td>3441</td>
<td>5513</td>
<td>...</td>
<td>0.020</td>
</tr>
<tr>
<td>2</td>
<td>01.01.2019</td>
<td>Berlin</td>
<td>1968</td>
<td>3201</td>
<td>4236</td>
<td>...</td>
<td>0.008</td>
</tr>
<tr>
<td>3</td>
<td>02.01.2019</td>
<td>Frankfurt</td>
<td>2331</td>
<td>3446</td>
<td>5657</td>
<td>...</td>
<td>0.021</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

\[ Y = f(X) + \varepsilon \]
The Parable of Google Flu: Traps in Big Data Analysis

David Luu1,2,3, Ray Kennedy1,2,3, Gary King1,2,3 and Alessandro Vespignani1,2,3

In February 2013, Google Flu Trends (GFT) made headlines not for a reason that Google executives or the creators of the flu tracking system would have hoped. Yahoo 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 (I). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is held up as an exemplary use of big data (2, 4), what lessons can we draw from this case?

The problem is not limited to GFT. Research on whether search or social media can predict a flu season's comings and goings (3) and other areas of public health has shown that traditional methods and applications can produce inaccurate estimates. Although these studies have shown the value of these data, we are far from a place where they can supplement more traditional methods or replace them. We explore the issues that contribute to GFT's missteps—big data, data collection and analysis, and algorithm design—and offer lessons from ever-evolving big data.

Big Data Blues

"Big data works" 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 outlined the numerous scientific possibilities for big data (9–11). However, quantity of data does not mean that one can ignore foundational issues of research.

The paradox is that big data that have required regular attention are not the output of instruments designed to produce valid and reliable data suitable for scientific analysis. The initial vision of GFT was particularly problematic: merging big data with small data. Essentially, the methodology was to find the best matches among 30 million search terms for 112 data points (10). The idea is that the similarity of scientific papers could reveal the importance of a particular topic. We should be cautious about using this approach, and we do not predict the future, were quite high. GFT developers, in fact, reported using an unsupervised learning algorithm that could identify the most strongly connected to the CDC data, such as those regarding high school basketball (11). This should have been warning that the big data were confounding the small number of cases—a standard concept in data analysis. This ad hoc method for throwing out popular search terms failed when GFT completely missed the unprecedented 2009 influenza A H1N1 pandemic (2, 10). Instead, the initial version of GFT was part of a directed, part winter directive. GFT directed updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2012 (4, 5).

Although not widely reported until 2013, the new GFT has been particularly overestimating flu prevalence for a much longer time. GFT has missed a large surge in the 2011–2012 flu season and has undercounted the 900000 of 1100000 cases starting with June 2011 (6). These errors are not randomly distributed. For example, last spring's errors predict this winter's errors (temporal autocorrelation), and the detection and magnitude of errors are often biased toward regions of high flu activity (7). These patterns mean that GFT overestimates considerable information that could be estimated by traditional methods.

Response

Even after GFT was updated in 2009, the competitive value of the algorithm as a stand-alone predictor is questionable. A study in 2014 demonstrated that GFT accuracy was not much better than a fairly simple projection based on already available data (8). Typically, a 2-week lagged CDC data (9). The computational framework was improved since that time, with lagged models significantly outperforming GFT (10). Findings from 3-week-old CDC data do a better job of projecting current flu prevalence than GFT (10) (see supplementary materials (SM)).

Can solving the large number of approaches that pose life sciences or the accuracy (6–11), 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 largest CDC data, as well as dynamically recalibrating GFT, we can more accurately improve the performance of GFT or the CDC alone (see the chart). This is to say, for ongoing evaluation and improvement, both by incorporating the CDC estimates, GFT should have largely remained itself and would have likely remained out of the headlines.
Traps in Big Data Analysis

Big Data Hubris

Algorithm Dynamics

Not reliable or valid
Reliable but not valid
Both reliable and valid

CAUSE

EFFEECT

what helps against flu
what helps against flu
what helps against stomach flu
what helps protect the computer against power fluctuations

Google Search  I'm Feeling Lucky
TWO PARADIGMS: STATISTICAL MODELING VS. MACHINE LEARNING
What is Statistics?

“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
THE STATISTICAL MODELING PARADIGM

Example: Lady Tasting Tea

Source: https://www.youtube.com/watch?v=Igs7d5saFFc
What is Machine Learning?

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

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

Example: Supervised Machine Learning for Credit Risk Scoring

Source: Provost & Fawcett (2013)
BANKRUPTCY PREDICTION: THE SM VS. ML WAY
Governing Machine Learning in Governments

Per Rådberg Nagbøl
Phd Student
ITU Copenhagen
Research Design

\[ Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \]

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

\[
\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p
\]
BANKRUPTCY PREDICTION – THE SM WAY

Dataset: 62,000+ Annual Reports in XBRL from 2014
# Results of Logistic Regression

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>bankruptcy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.027*** (0.036)</td>
</tr>
<tr>
<td>X1</td>
<td>-1.501*** (0.067)</td>
</tr>
<tr>
<td>X2</td>
<td>0.00000 (0.00004)</td>
</tr>
<tr>
<td>X3</td>
<td>0.00001 (0.00005)</td>
</tr>
<tr>
<td>X4</td>
<td>0.00000 (0.00000)</td>
</tr>
</tbody>
</table>

| 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
Training and Test Sets

Dataset

Training set

Test set

Source: James et al. (2013)
Predictive Accuracy of Logistic Regression on Test Set

Area under the curve: 0.7099
Going Beyond Financial Ratios (i.e., reading 62,000 annual reports)
The Bag-Of-Words (BOW) Model

Lorem Ipsum

“Necque poro quisquem est qui dolore ipsum qui dolore amit, consectetur adipiscing...”

“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 popularized 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).

<table>
<thead>
<tr>
<th>country</th>
<th>year</th>
<th>cases</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>2019</td>
<td>266</td>
<td>201231360</td>
</tr>
<tr>
<td>Brazil</td>
<td>2019</td>
<td>3737</td>
<td>17238362</td>
</tr>
<tr>
<td>Brazil</td>
<td>2019</td>
<td>4488</td>
<td>174194896</td>
</tr>
<tr>
<td>China</td>
<td>2019</td>
<td>2766</td>
<td>128726272</td>
</tr>
<tr>
<td>China</td>
<td>2019</td>
<td>2866</td>
<td>128726883</td>
</tr>
</tbody>
</table>

variables observations values
The Bag-Of-Words (BOW) Model

- Treat every document as an 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
The Bag-Of-Words (BOW) Model

<table>
<thead>
<tr>
<th></th>
<th>“word 1”</th>
<th>“word 2”</th>
<th>“word 3”</th>
<th>“word 4”</th>
<th>“word 5”</th>
<th>…</th>
<th>Bankruptcy?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lego</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>…</td>
<td>No</td>
</tr>
<tr>
<td>Maersk</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>…</td>
<td>No</td>
</tr>
<tr>
<td>Jysk Frigt</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>…</td>
<td>Yes</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

$Y = f(X) + \varepsilon$
BANKRUPTCY PREDICTION – THE ML WAY

Tree-based Classification Algorithms
Example: Loan Default
Example: Loan Default

<table>
<thead>
<tr>
<th>Name</th>
<th>Balance</th>
<th>Age</th>
<th>Employed</th>
<th>Write-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>$200,000</td>
<td>42</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Mary</td>
<td>$35,000</td>
<td>33</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Claudio</td>
<td>$115,000</td>
<td>40</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Robert</td>
<td>$29,000</td>
<td>23</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Dora</td>
<td>$72,000</td>
<td>31</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

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

Source: Provost & Fawcett (2013)
Example: Loan Default

Claudio | $115,000 | 40 | no

Source: Provost & Fawcett (2013)
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 $X_1, \ldots, X_p$, 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)
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)
A Random Forest has many Trees
Predictive Accuracy of Random Forest with BOW on Test Set

Area under the curve: 0.8945
How Does the Model Look Like?

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

Oliver Müller1, Iris Jungles2, Jan vom Brocke1 and Stefan Debortol1

1Facultad de Informática, University of Castilla-La Mancha, Spain
2Universität der Bundeswehr München, Germany

Correspondence: Oliver Müller, Institute of Information Systems, University of Castilla-La Mancha, Plaza de las Monjas, 40071 Albacete, Spain.

ISSUES AND OPINION

Abstract
This paper reviews the use of big data analytics (BDA) as a strategy for success in the information systems (IS) research. In broad terms, we understand BDA as the statistical modeling of large, diverse, and dynamic databases of user-generated content and digital traces. BDA is a new paradigm for utilizing big data sources and advanced analytics that already had its way into some social science disciplines, but which have only started in IS research. In this context, we present the research efforts of the authors to discuss BDA challenges and opportunities for IS research, and find them by means of an end-user study about predicting the number of 1.3 million online consumer reviews. In order to assist IS researchers in planning, executing, and interpreting their own studies, we propose an initial set of guidelines for conducting BDA studies in IS.

Keywords: big data analytics; data sources; 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 технологические for storage and computing resources, has led to a new appetite for transforming digital data into a valuable asset. BDA, as a result, has become a hot topic in IS research. Studies, for example, from 2011, 2012, and 2013, have shown that the volume of data is increasing at an exponential rate. One study from 2011, by McKinsey, estimated that the volume of digital data is expected to reach 1.7 exabytes per year by 2014. In contrast, another study from 2013, by Gartner, predicted that the volume of data would exceed 16,000 exabytes by 2025.

As a result, the need for data analytics has increased, resulting in reduced latency between the collection of data and its analysis. In addition, digital footprint research has shown that the volume of data is growing at an exponential rate.

As volume, velocity, and variety (Vendrus et al., 2014) increase, the velocity of data is drawn into question (Sambamurthy et al., 2014). Unlike research data collected with a specific research question in mind and measured using validated instruments, big data often arise from unexpected, uncontrolled, and unpredictable processes. As a result, big data can be challenging to collect and analyze without a concrete purpose.
THE END
• Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. Springer.
• James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

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