Data Science for Digital Humanities

BY

HENDRIK HEUER

INSTITUTE FOR
INFORMATION MANAGEMENT
BREMEN GMBH

Universität Bremen

About me

- **Hendrik** Heuer, MSc.
- Doctoral Researcher at the **Institute of Information Management Bremen (ifib)** at the **University of Bremen**
- Mail: hheuer@uni-bremen.de
- Website: http://hen-drik.de

KTH VETENSKAP OCH KONST

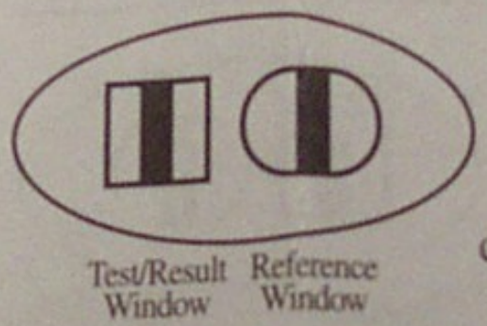
Aalto University

Universität Bremen

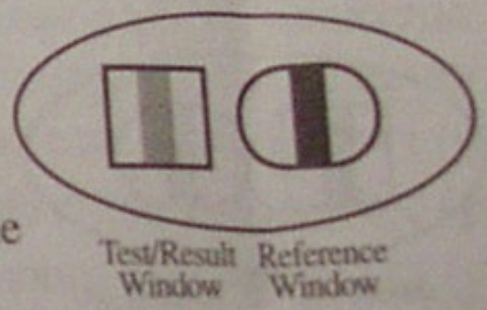
been applied. A positive result may be observed in as early as 1 minute. To verify a negative result, allow the test to run for the full 3 minutes.

DO NOT interpret your result after 10 minutes.

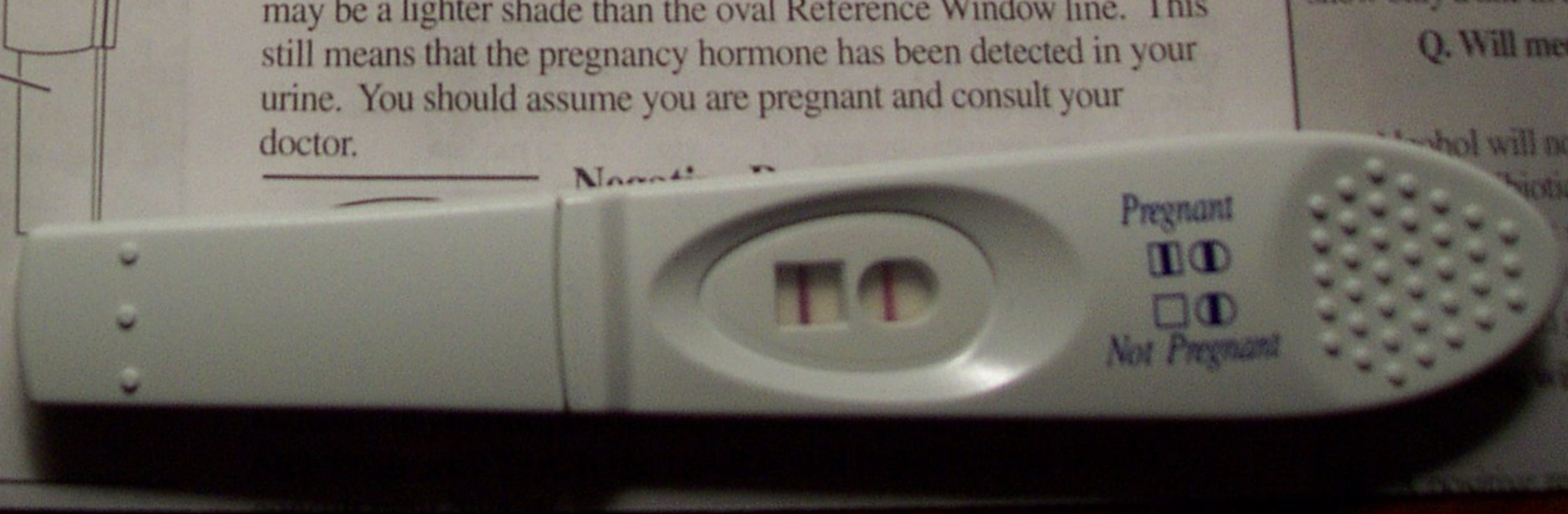
Positive Result



The Test Line can vary in shades of color compared to the Reference Line

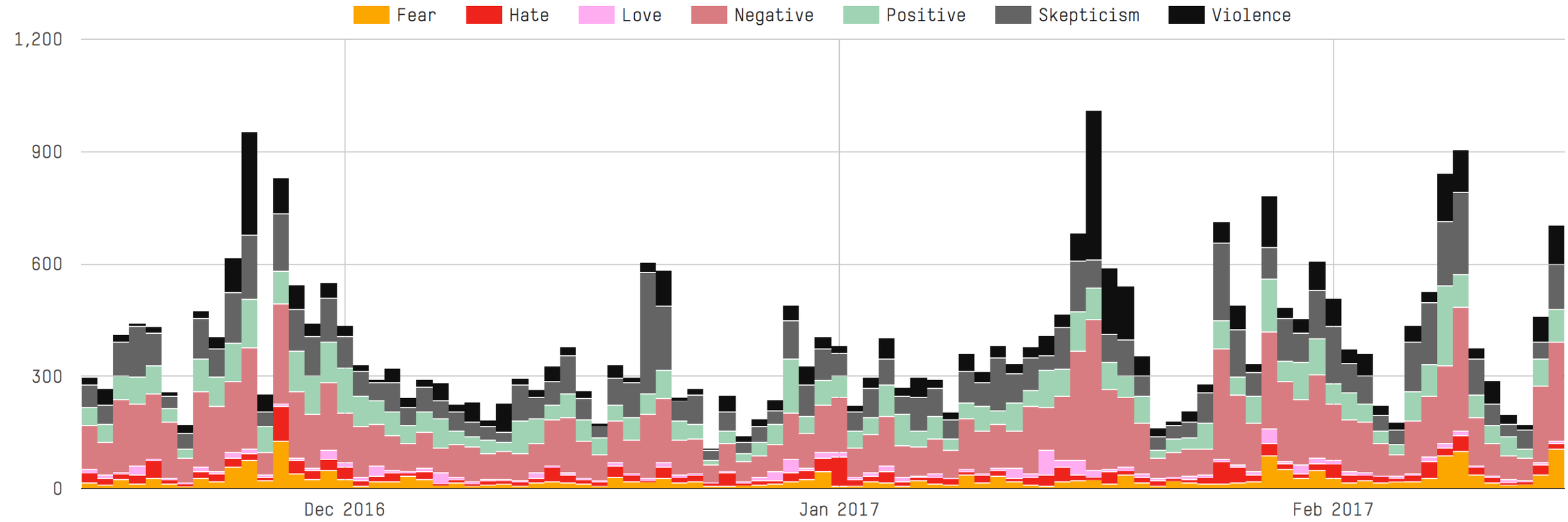


A positive result is observed when there is a purple-pink line in the oval Reference Window and the same color line in the square Test/Result Window. The line in the square Test/Result Window may be a lighter shade than the oval Reference Window line. This still means that the pregnancy hormone has been detected in your urine. You should assume you are pregnant and consult your doctor.



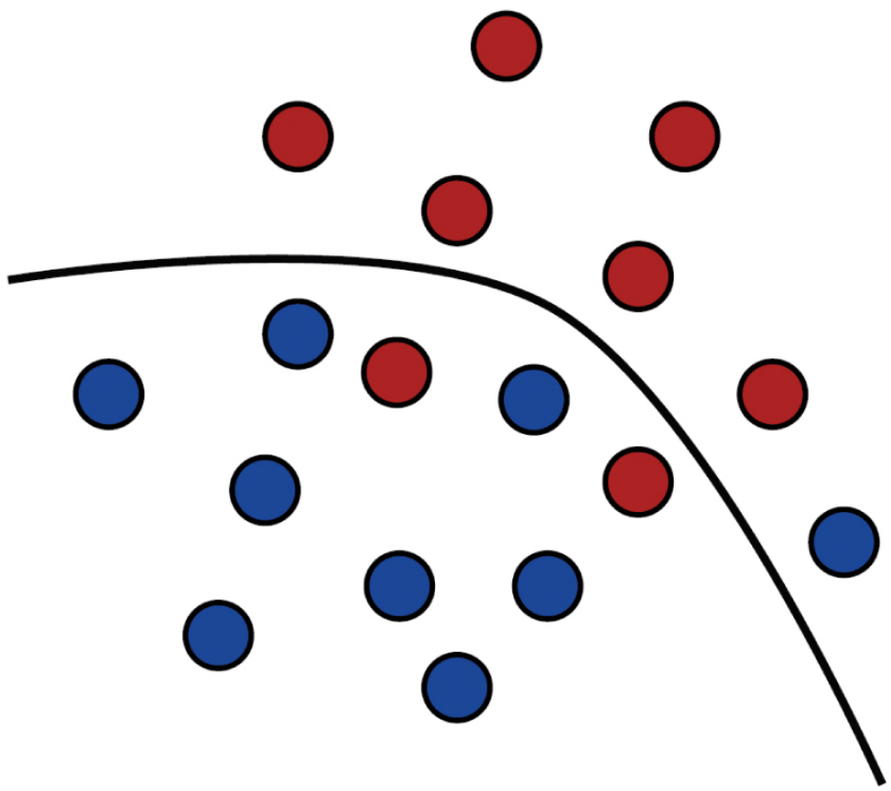
Sentiment Analysis

Gavagai



Analysing millions of images and enormous text sources **using machine learning** and deep learning techniques **is simple** & straightforward in the Python ecosystem

"Data Science is **statistics on a Mac**."

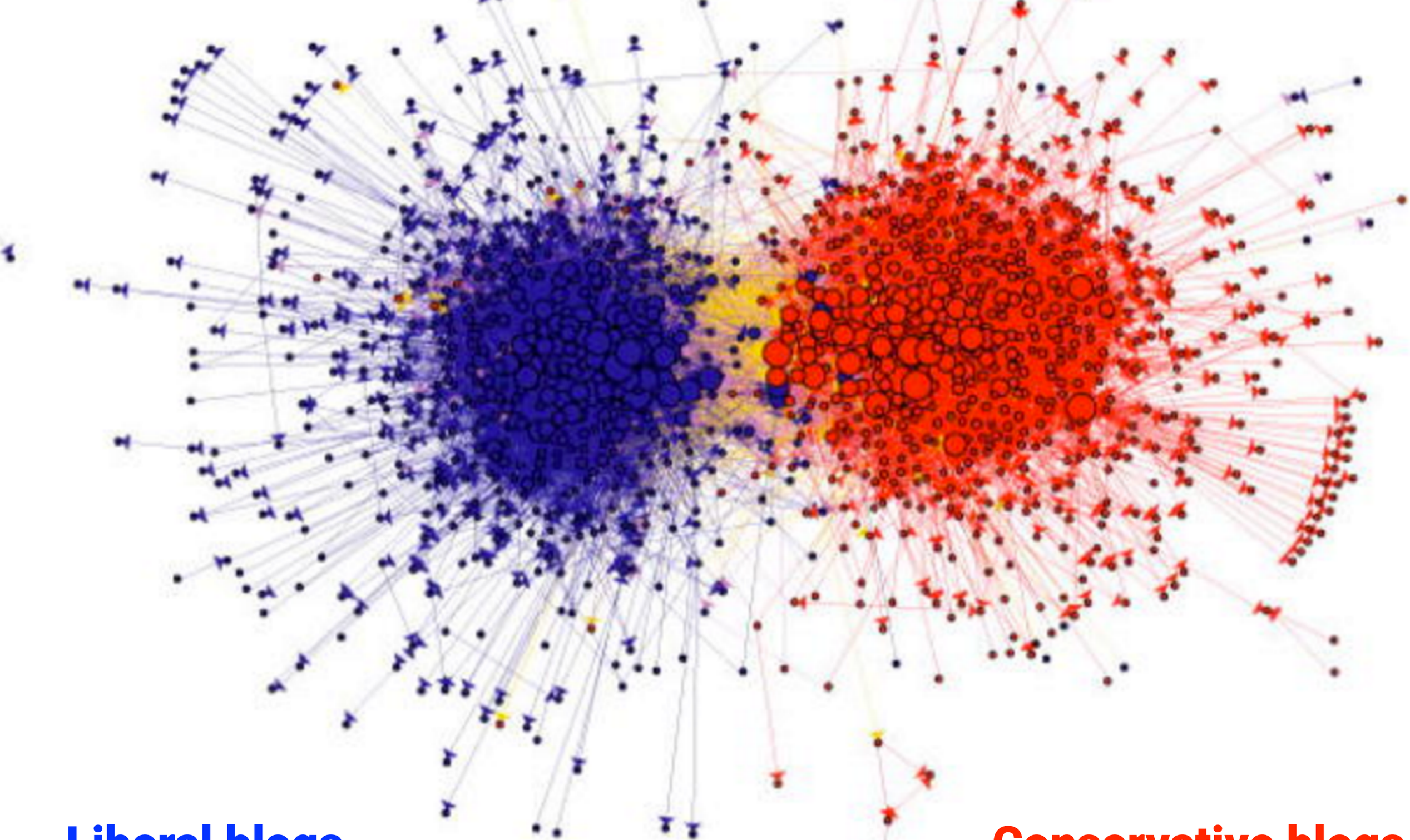


"A **data scientist** is someone who is **better at statistics than any software engineer** and **better at software engineering than any statistician**."

Bild: Foundations of Machine Learning
Zitate: https://datadistributist.wordpress.com/2014/02/04/the-truth-about-data-science/

Data Science

- introduced by **Cleveland** in 2001
- enlargement of **statistics**
- enabling analysts to **"learn from data"**
- **"Sexiest job of the 21st century"**
  - **Davenport & Patil, 2012**
    Harvard Business Review



Computational Social Science, Lazer, Pentland, Adamic, et al, Science 06 Feb 2009

$$informatics = mathematics + x$$

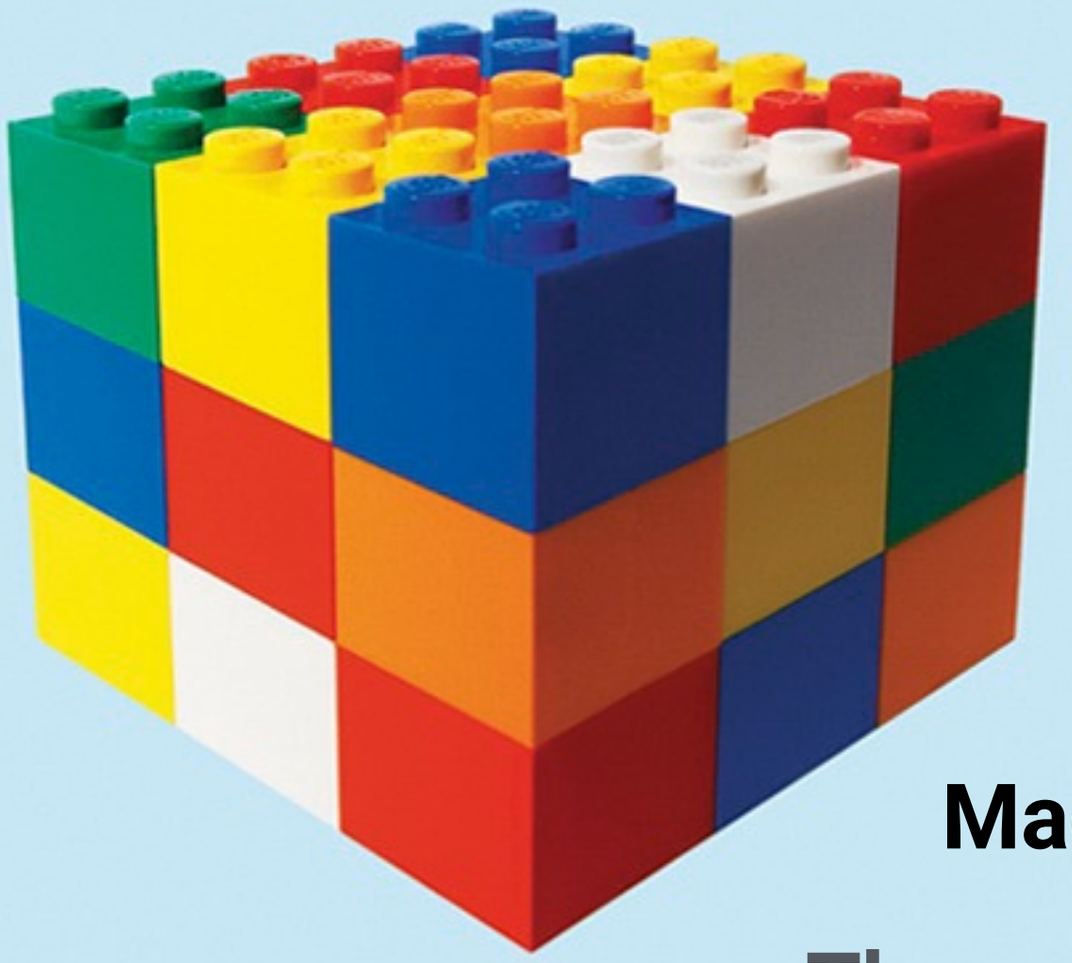
Text Processing

Natural Language Toolkit

word2vec spaCy

Topic Modeling

gensim



Visualization

d3.js Highcharts

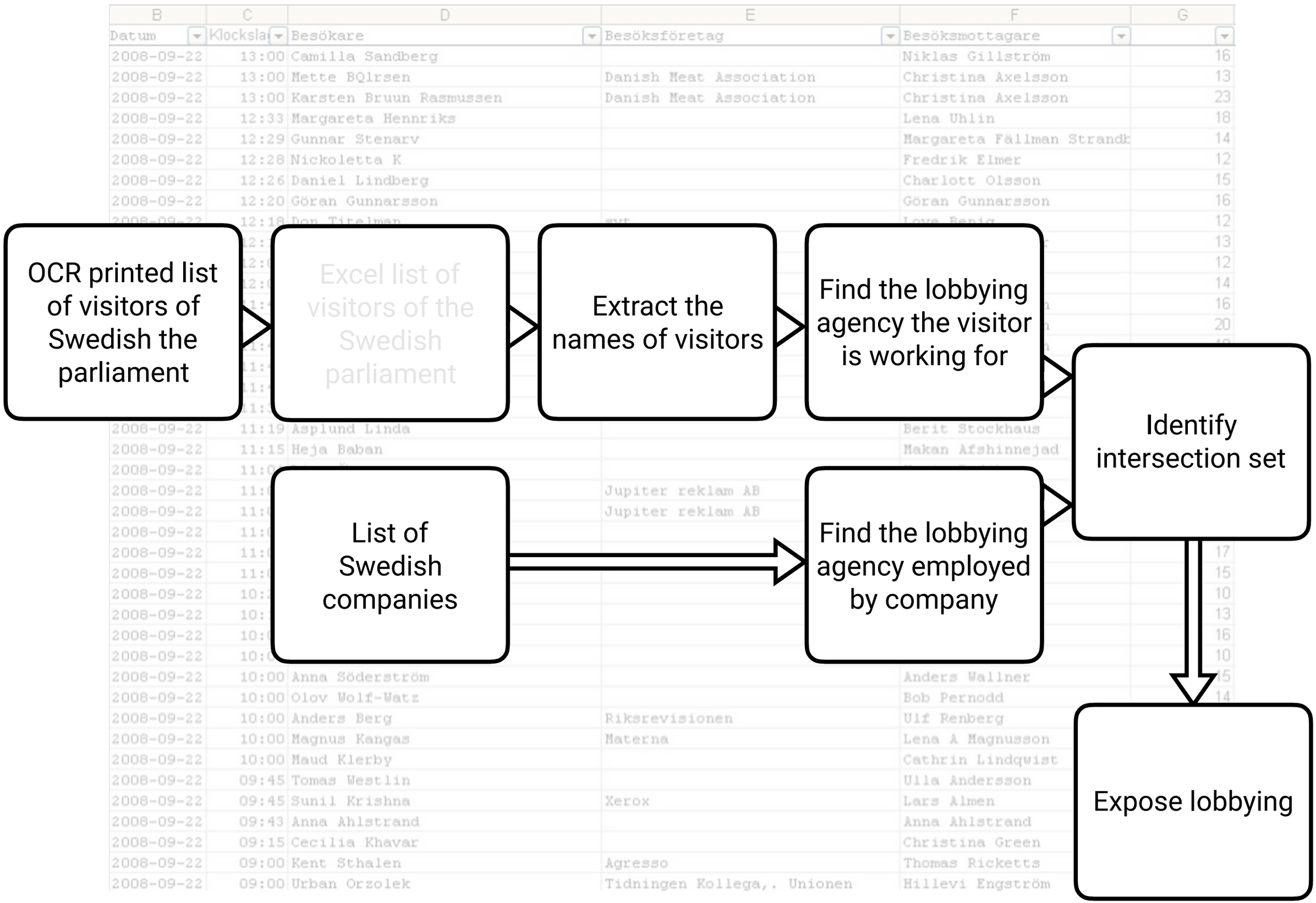
Google Chart API

Machine Learning

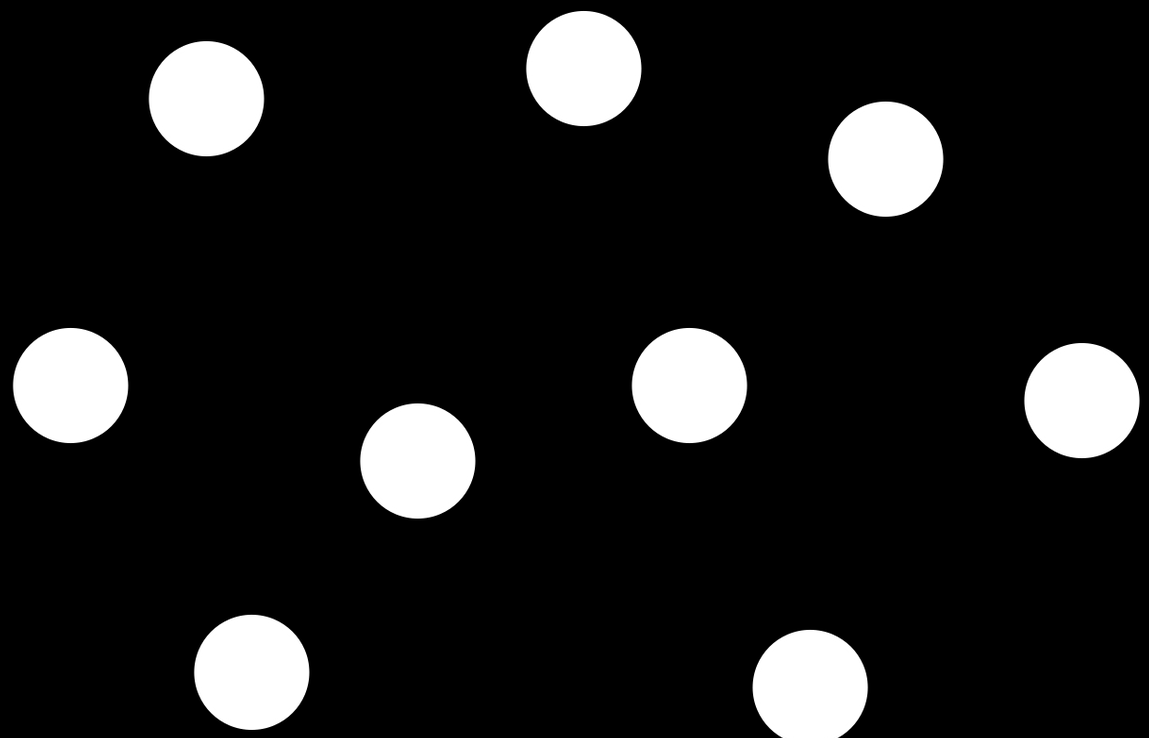
Theano scikit-learn

Caffe Keras Tensorflow

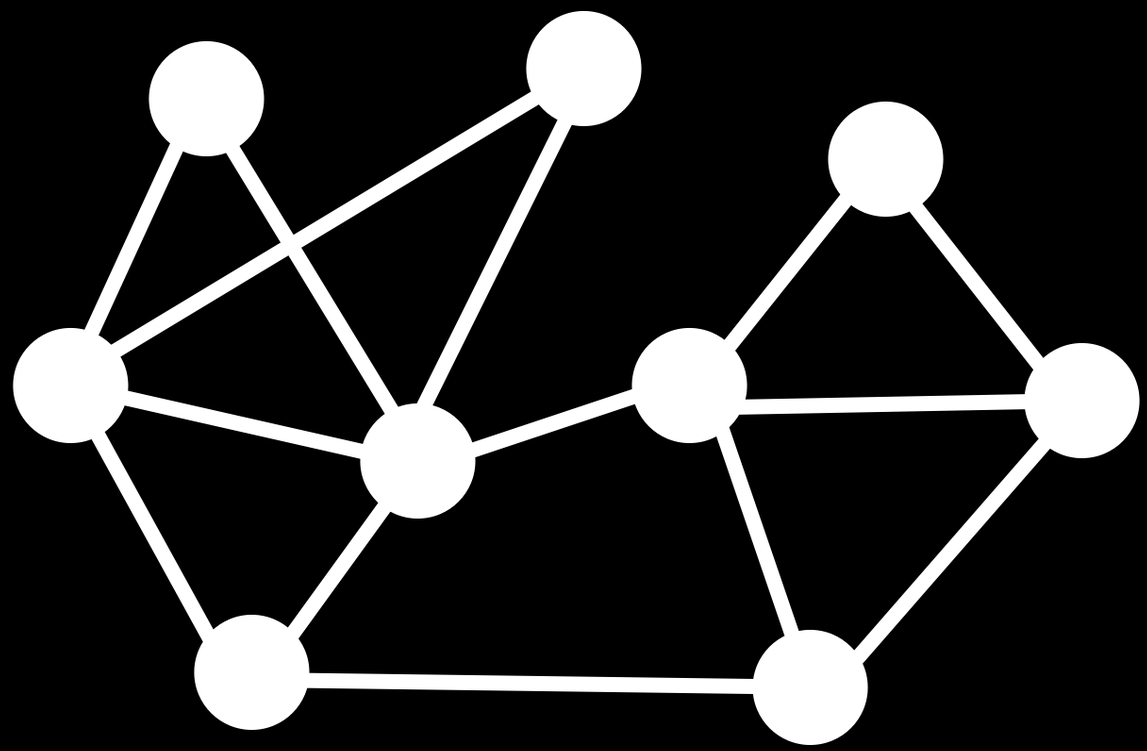
Imagine a journalist who wants to **cross-reference** the names on the guest list of a parliament with online information about **lobbyists** to identify which party meets which **company**



Lobbying in the Swedish parliament
Helena Bengtsson, Sveriges Television (SVT) / Guardian
PyCon Sweden 2014



Small-world experiments



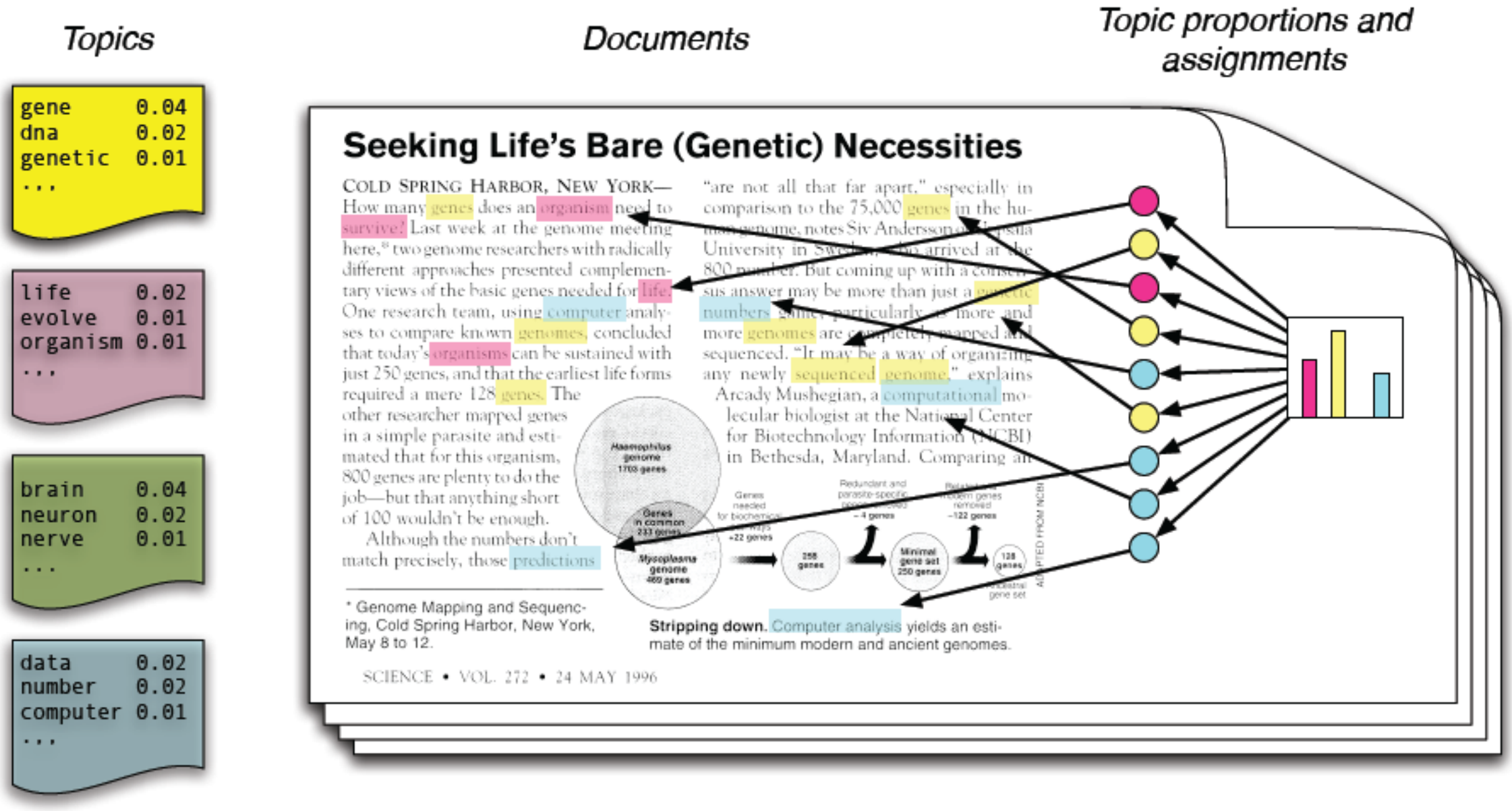
Small-world experiments

Karinthy (1929)
Milgram (1967)
Duncan Watt (2001),
Leskovec and Horvitz (2007)

Digital Humanities

- are the academic disciplines that study the **expressions of the human mind** (Rapport Duurzame Geesteswetenschappen, 2010)
- **intersection** of the **humanities** and **computer science** opens up **new research methods** and creates a new environment in which the humanities become subject to new approaches

Topic Modelling (LDA)



Blei, David M., Andrew Y. Ng, and Michael I. Jordan.
"Latent Dirichlet Allocation."
Journal of Machine Learning Research 3.Jan (2003): 993-1022.

Latent Dirichlet Allocation

- I ate a banana and spinach smoothie for breakfast
- I like to eat broccoli and bananas.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Edwin Chen, https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation

Latent Dirichlet Allocation

- I ate a banana and spinach smoothie for breakfast
- I like to eat broccoli and bananas.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

- Topic A: food
- Topic B: cute animals

Edwin Chen, https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation

Latent Dirichlet Allocation

- I ate a banana and spinach smoothie for breakfast
- I like to eat broccoli and bananas.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

- Topic A: food
- Topic B: cute animals
- Sentences 1 and 2: 100% Topic A
- Sentences 3 and 4: 100% Topic B
- Sentence 5: 60% Topic A, 40% Topic B

Edwin Chen, https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation

Latent Dirichlet Allocation

- I ate a banana and spinach smoothie for breakfast
- I like to eat broccoli and bananas.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

- Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching
- Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster

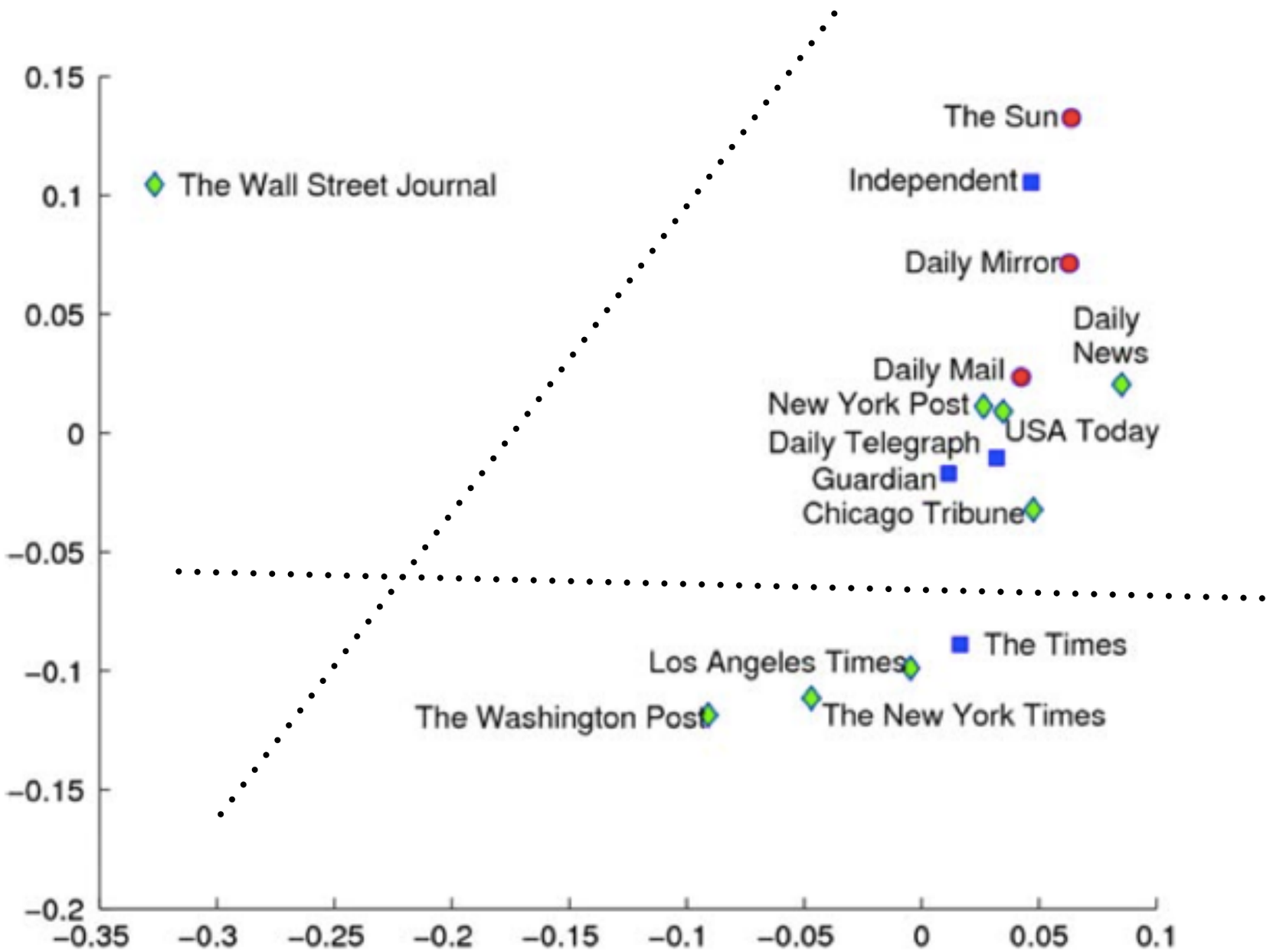
Edwin Chen, https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation

Massive-scale automated analysis of news-content

- 2.5 million articles from 498 different English-language news outlets
  - Reuters
  - New York Times Corpus
- automatically annotated into 15 topic areas
- the topics were compared in regards to readability, linguistic subjectivity and gender imbalances

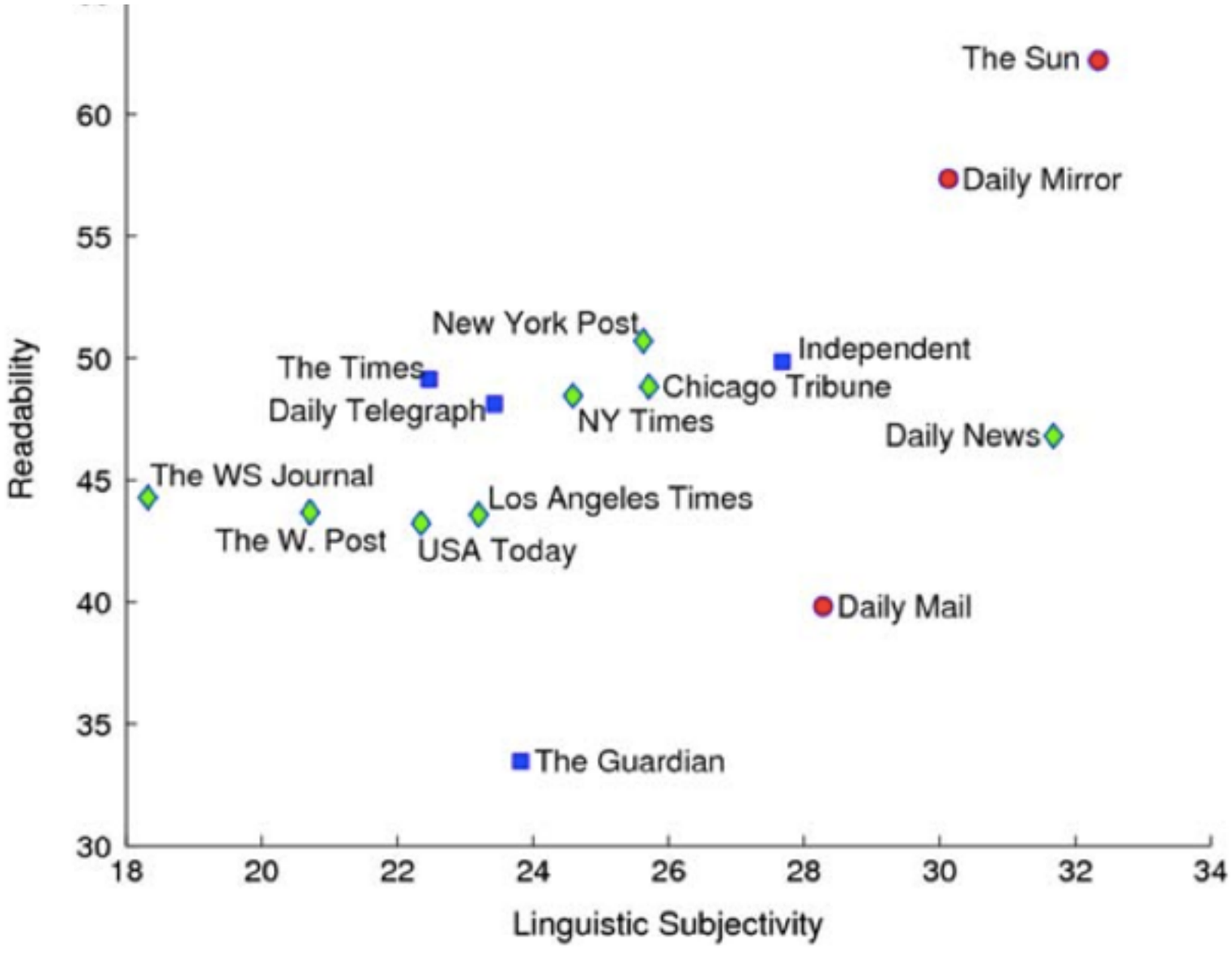
I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, ‘Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender’, Digital Journalism, vol. 1, no. 1, 2013. DOI:10.1080/21670811.2012.714928

Comparing Newspapers



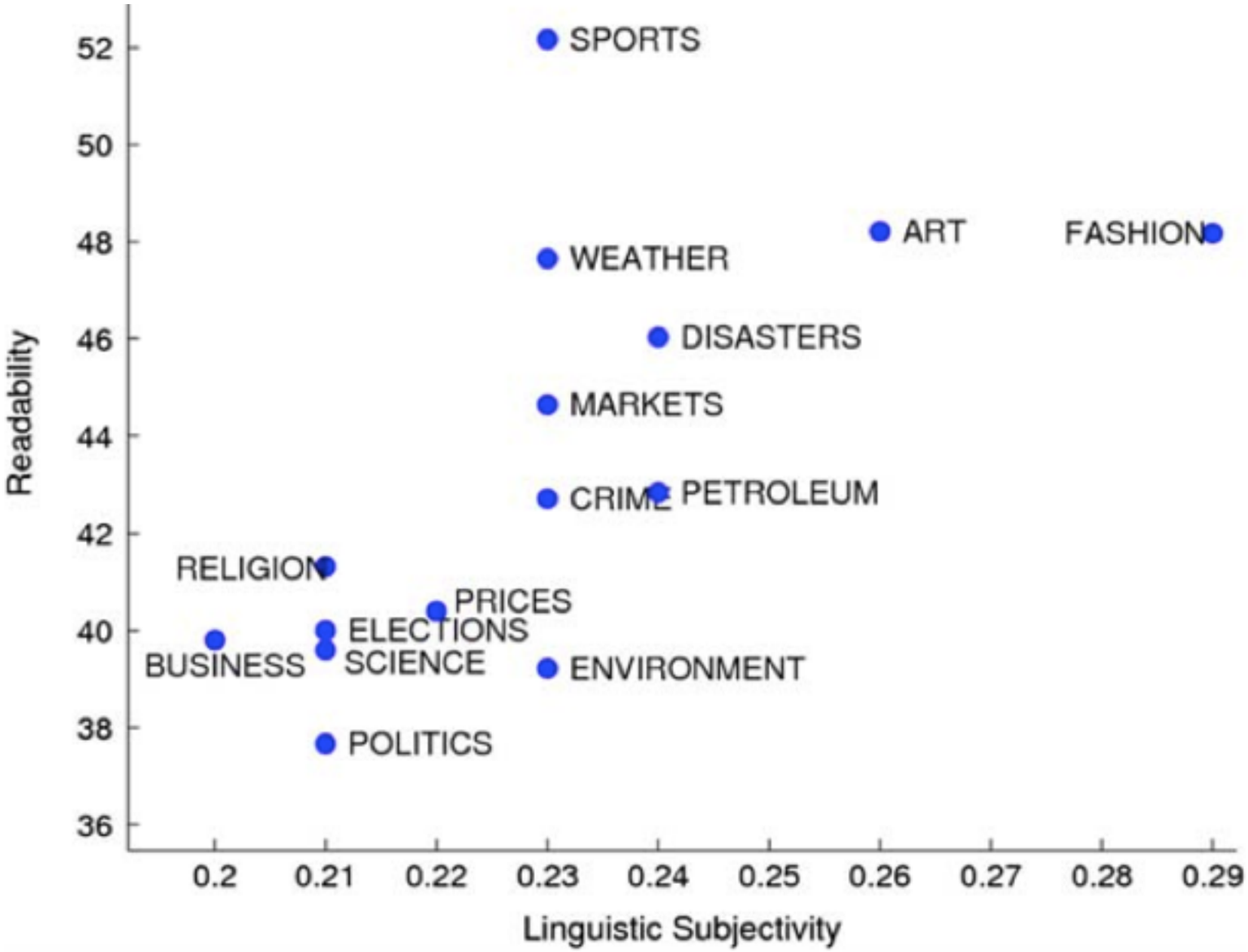
I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, 'Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender', Digital Journalism, vol. 1, no. 1, 2013. DOI:10.1080/21670811.2012.714928

Comparing Newspapers



I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, 'Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender', Digital Journalism, vol. 1, no. 1, 2013. DOI:10.1080/21670811.2012.714928

Comparing Topics



I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, 'Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender', Digital Journalism, vol. 1, no. 1, 2013. DOI:10.1080/21670811.2012.714928

Macroanalysis

- Jockers uses digital methods for **literary history**
- showing that factors such as author **gender**, author **nationality**, and **date of publication** affect the **choice of literary themes in novels**
- measurable, data-driven **proxy for literary themes**
- identified and extracted hundreds of topics from a corpus of **3346 works of 19th-century British, Irish, and American fiction**

Topic Modelling (LDA)

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman and Alexei A. Efros.
"Generative Visual Manipulation on the Natural Image Manifold",
in European Conference on Computer Vision (ECCV). 2016.

Image: https://en.wikipedia.org/wiki/Gezi_Park_protests

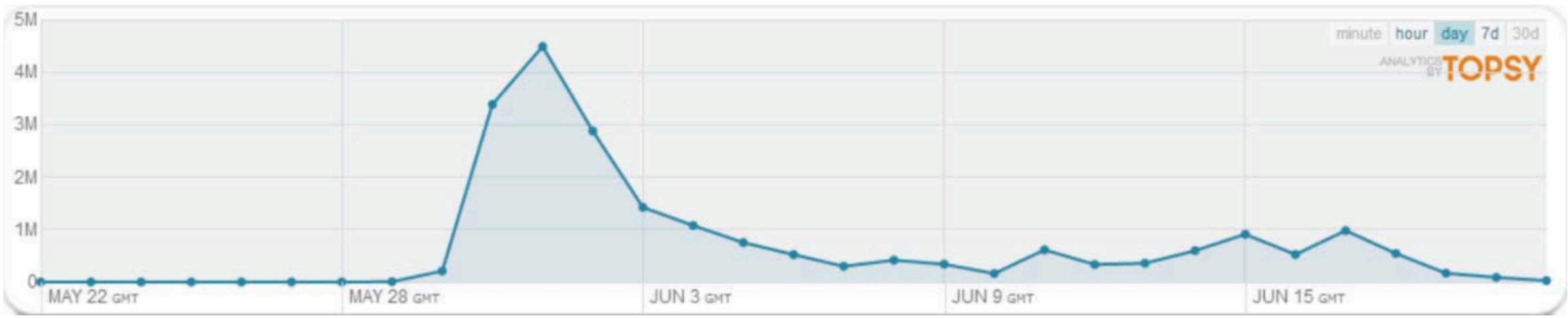


Figure 1: The frequency of top 20 hashtags associated with Gezi Protests. (Banko and Babacan, 2013)

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman and Alexei A. Efros.
"Generative Visual Manipulation on the Natural Image Manifold",
in European Conference on Computer Vision (ECCV). 2016.

Image: https://en.wikipedia.org/wiki/Gezi_Park_protests

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

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: $\mathrm{logit}(I(t)) = \alpha \mathrm{logit}(Q(t)) + \varepsilon$, 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. $\mathrm{logit}(p)$ 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 sepa-

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. surement and construct validity and reliability and dependencies among data (*12*).

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

Big Data Hubris

ILLUSTRATION: ISTOCKPHOTO.COM

Big Data Hubris

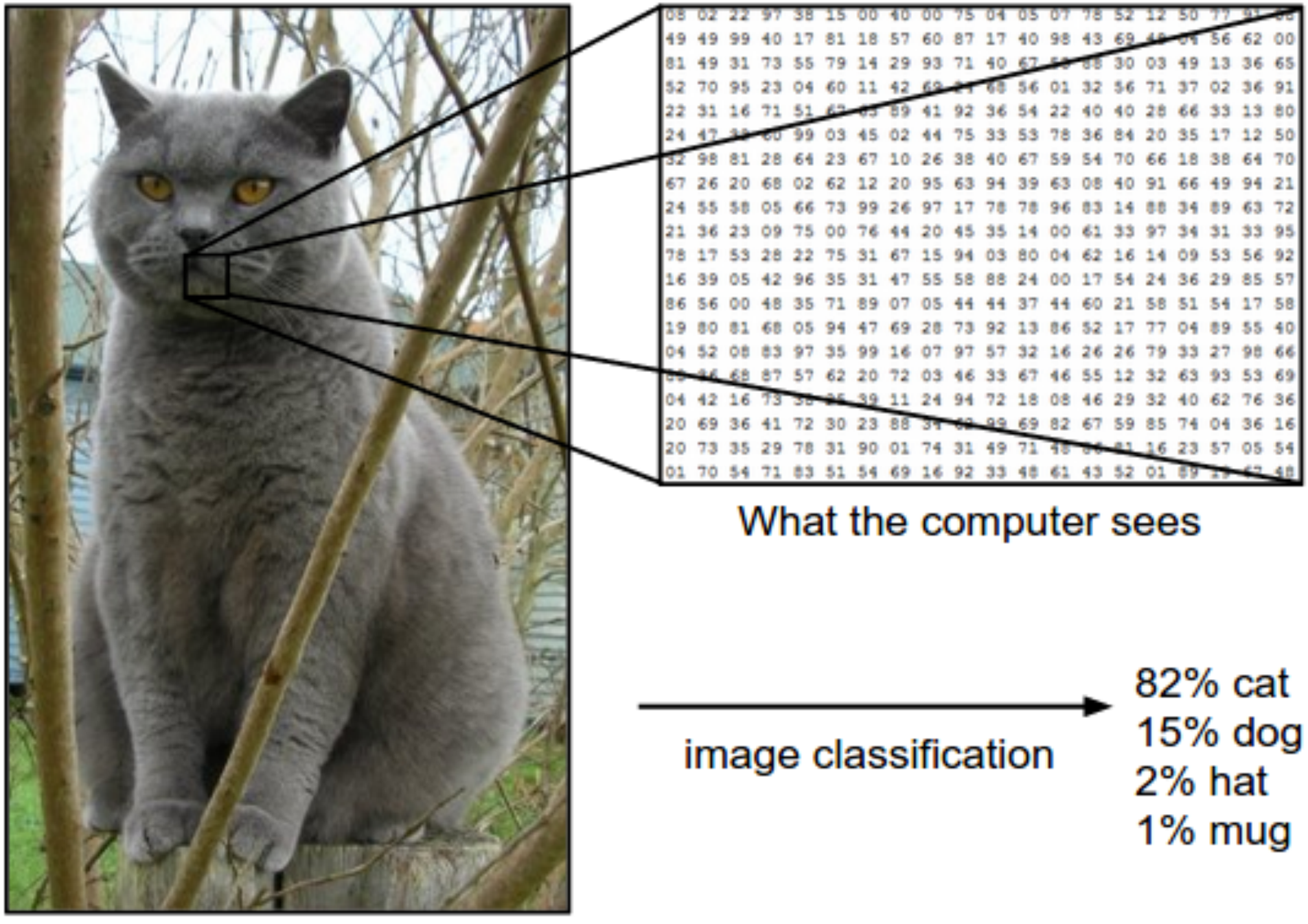
"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional

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

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 influ-

What the Computer sees



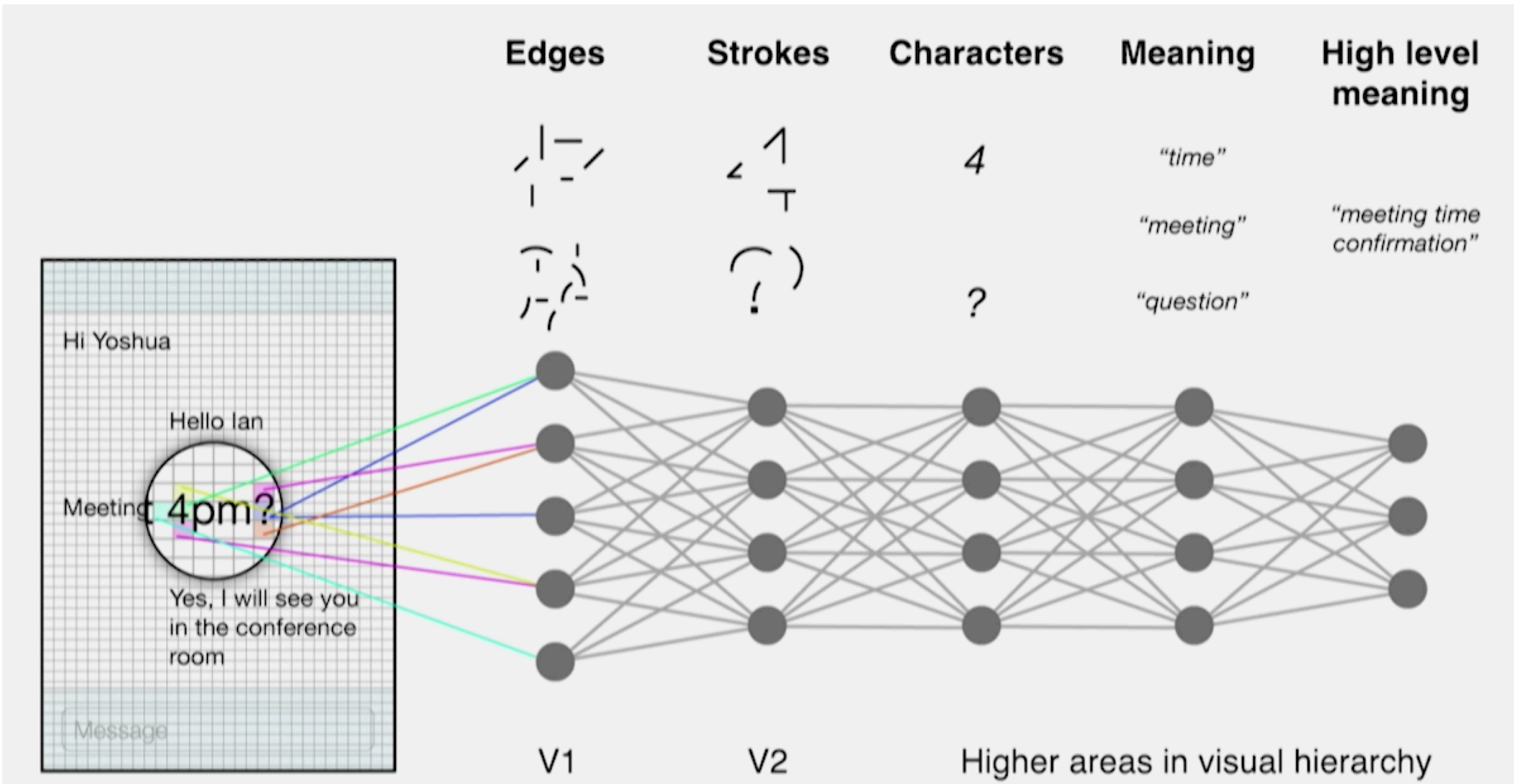
Computer Vision

Source: Fei-Fei Li & Justin Johnson & Serena Yeung CS231n, Stanford University

ImageNet
1000 object class (categories)

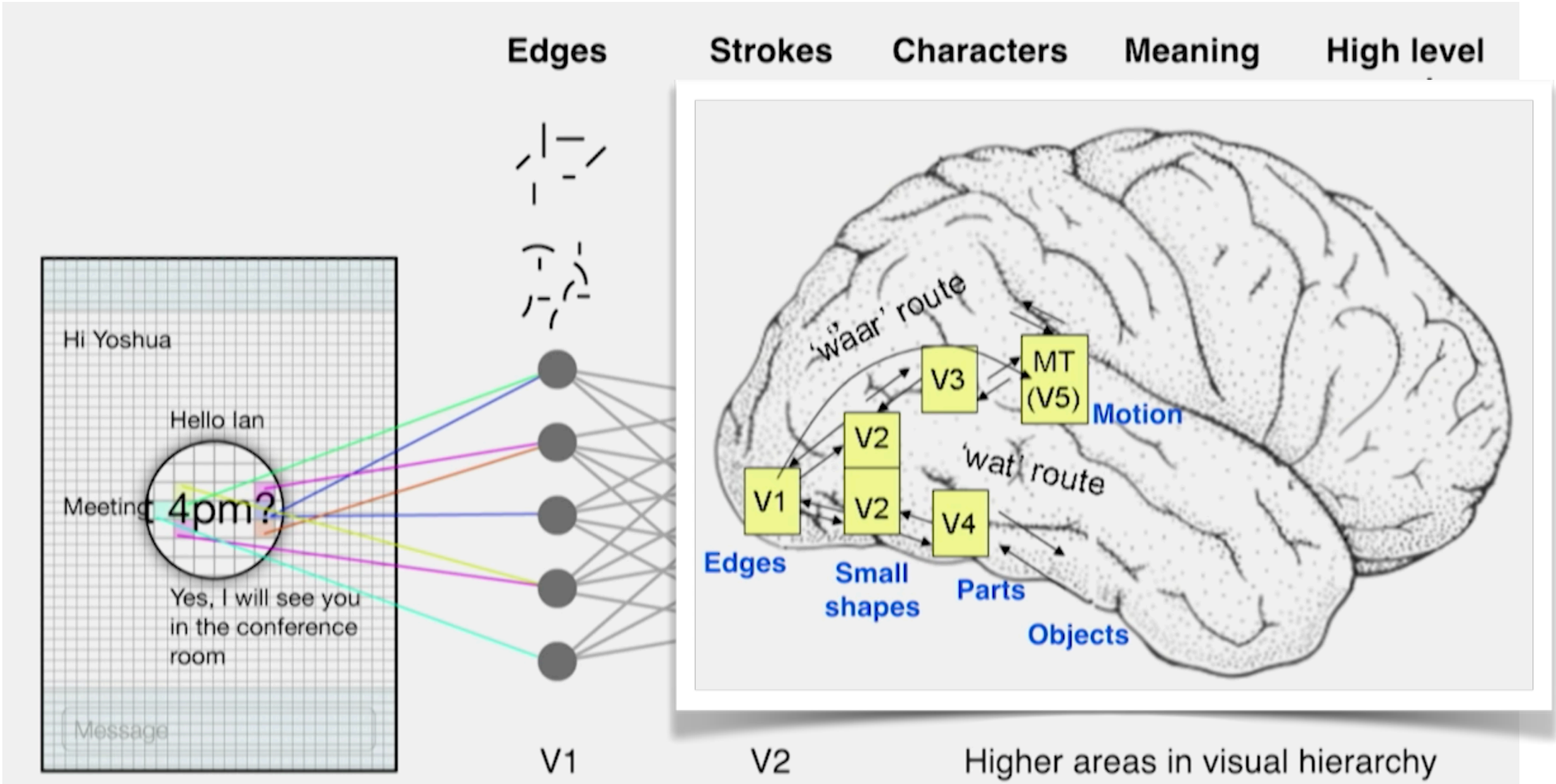
Images for training: 1,2 millions
Images for testing: 100k

Representation Learning

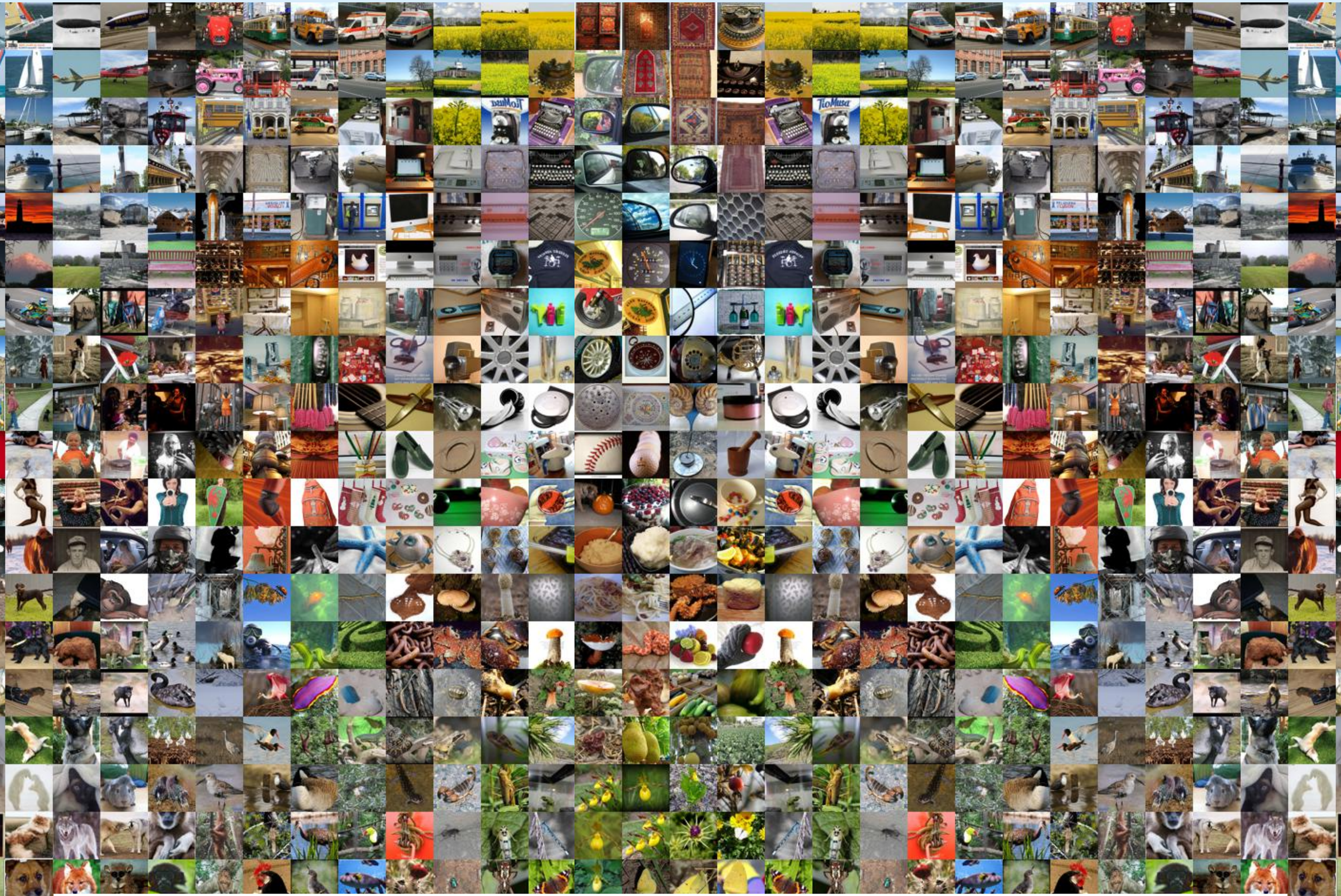


The Rise of Artificial Intelligence through Deep Learning | Yoshua Bengio | TEDxMontreal

Representation Learning

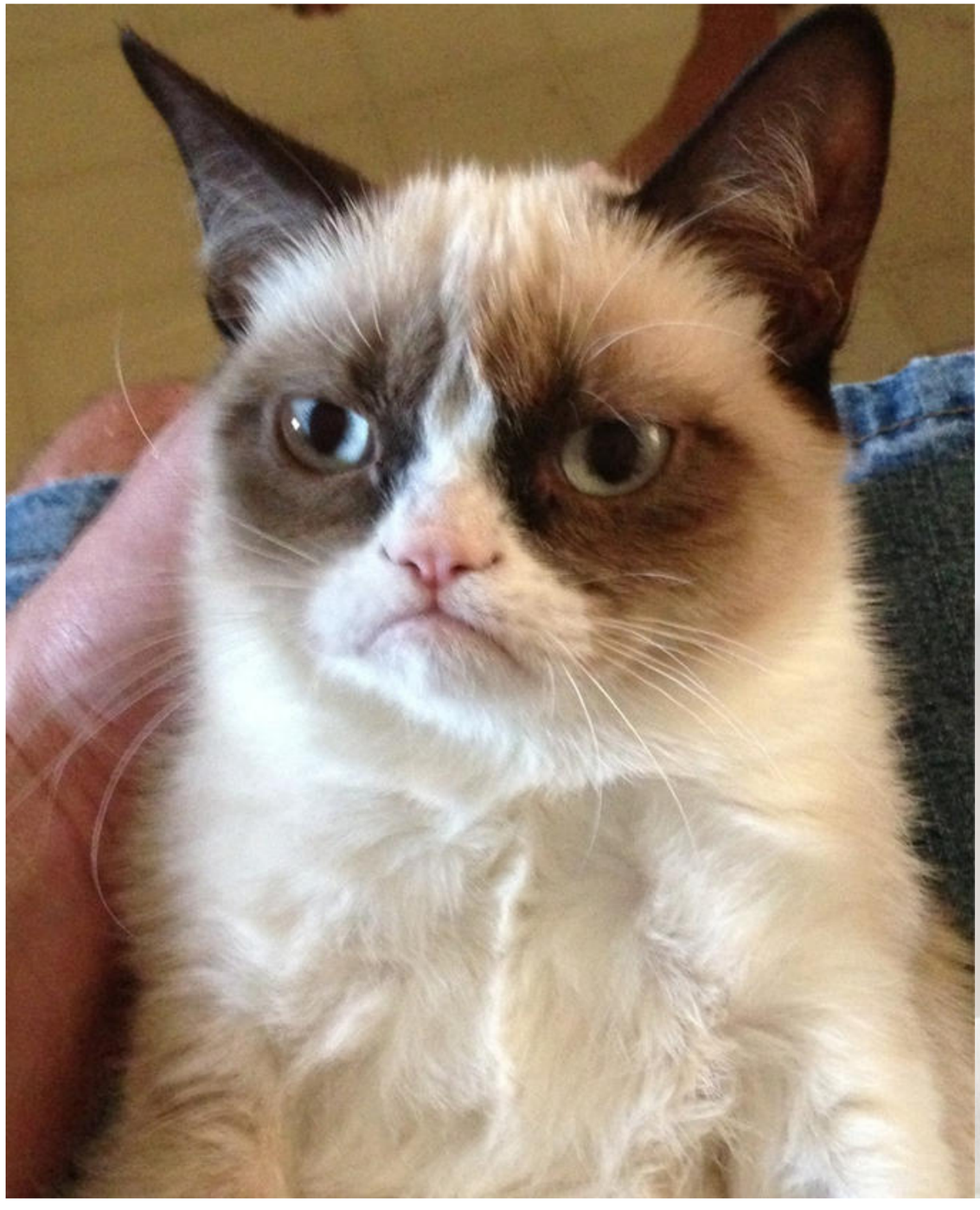


The Rise of Artificial Intelligence through Deep Learning | Yoshua Bengio | TEDxMontreal



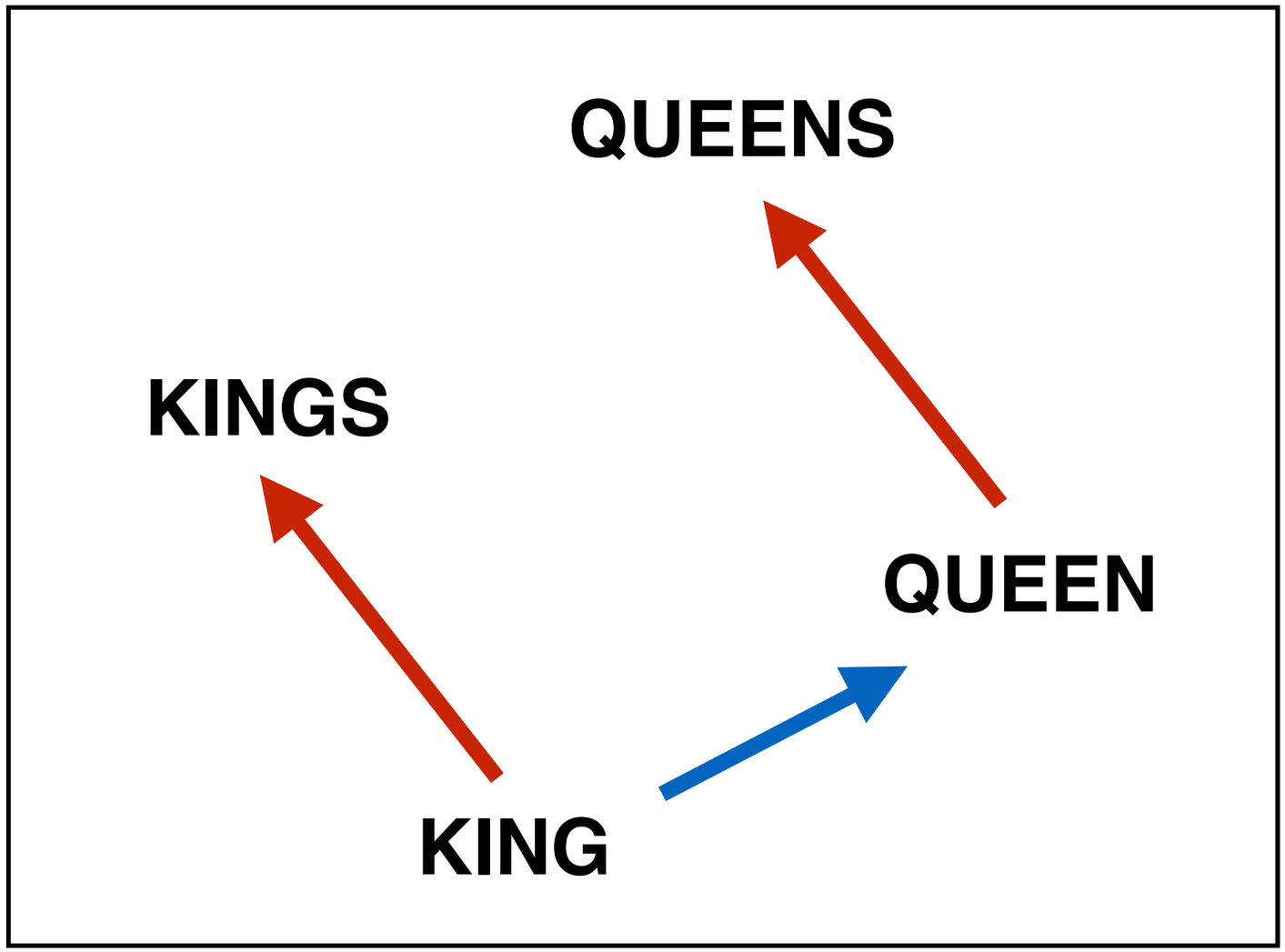
Source: Andrej Karpathy: What I learned from competing against a ConvNet on ImageNet

Transfer Learning



David Teniers the Younger: Archduke Leopold Wilhelm and the artist in the archducal picture gallery in Brussels

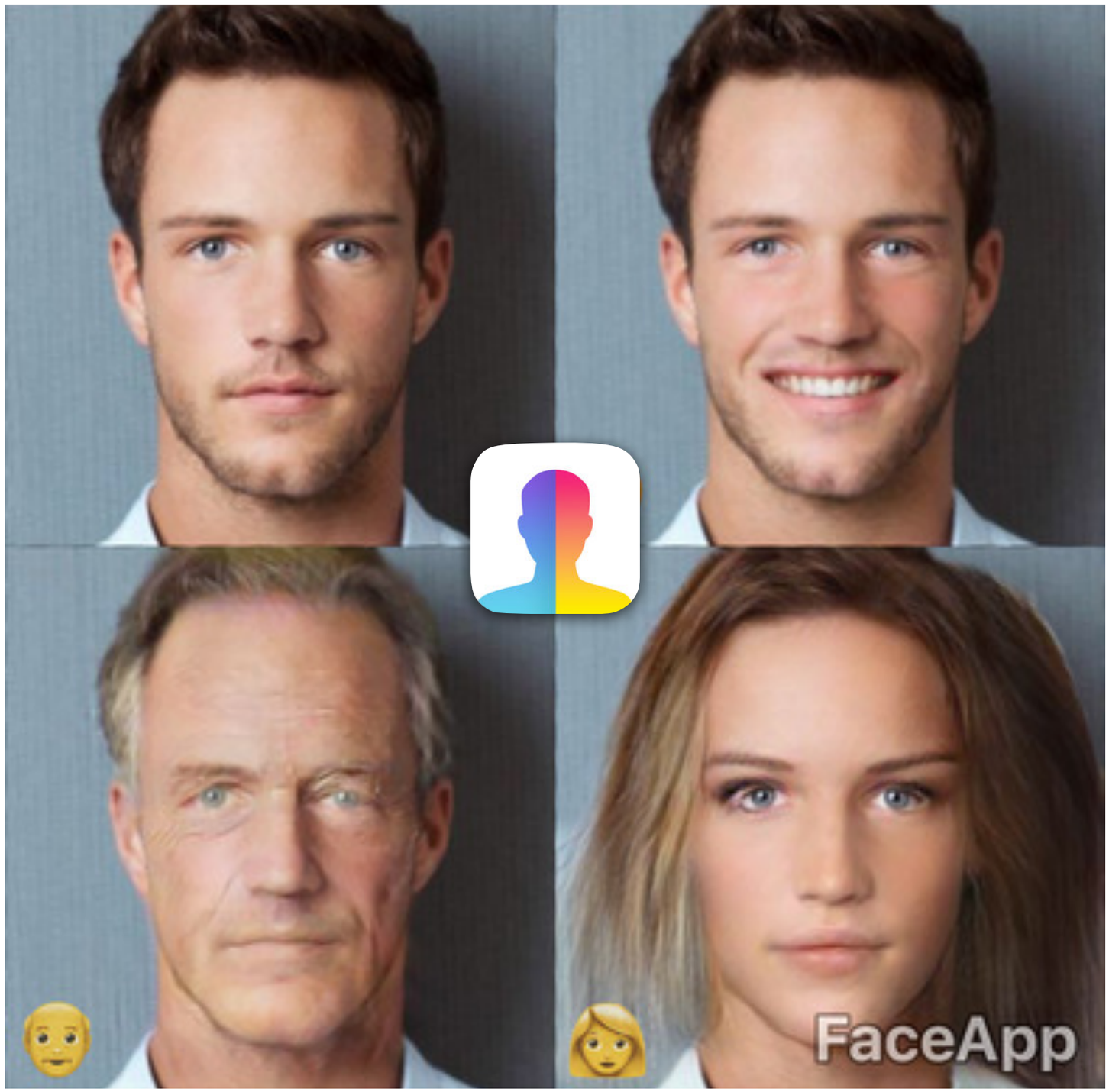
word2vec



Vectors can encode relationships

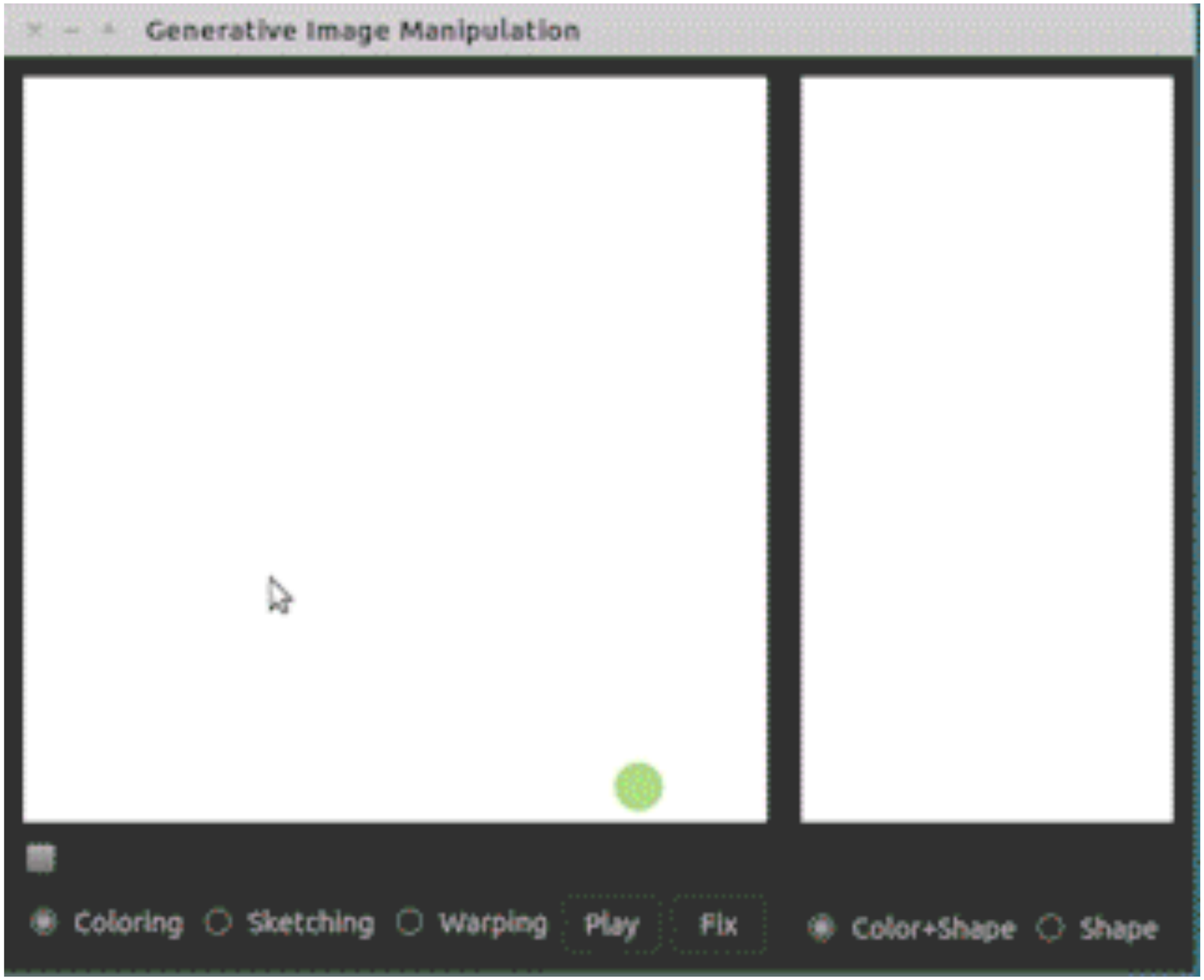
T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘Efficient Estimation of Word Representations in Vector Space’, CoRR, vol. abs/1301.3781, 2013 [Online]. Available: http://arxiv.org/abs/1301.3781

Generative Models



Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, Brendan, Frey. Adversarial Autoencoders. http://arxiv.org/abs/1511.05644. 2016.

Generative Models

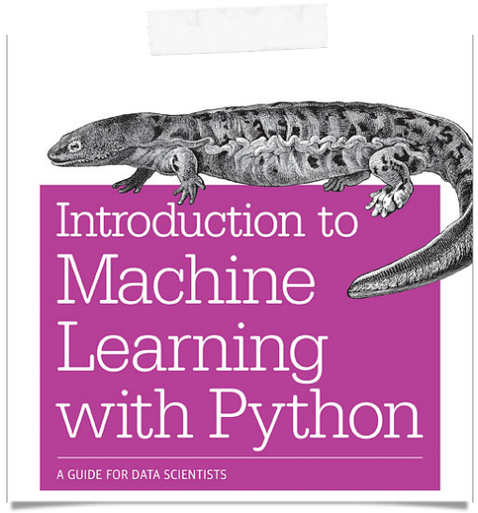


Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman and Alexei A. Efros.
"Generative Visual Manipulation on the Natural Image Manifold",
in European Conference on Computer Vision (ECCV). 2016.

Reading List

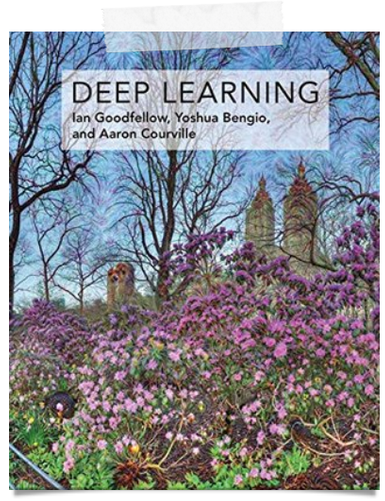
Doing Data Science

- Cathy O'Neil & Rachel Schutt
- one of the first books on Data Science
- Schutt was a data scientist with Google



Introduction to Machine Learning with Python

- Andreas C. Müller & Sarah Guido
- from the creator of scikit-learn
- practical introduction to Python and ML



Deep Learning

- Ian Goodfellow, Yoshua Bengio, & Aaron Courville
- best textbook on deep learning
- surprisingly readable

Data Science for Digital Humanities

BY

HENDRIK HEUER

INSTITUTE FOR
INFORMATION MANAGEMENT
BREMEN GMBH

Universität Bremen