Big Data
Map Reduce
Spark

Machiel Jansen

SURF SARA

Agenda

- Big Data
- HDFS – Map Reduce
- Pig + Oefening
- Lunch
- Python intro notebook
- Spark intro + oefening
- Spark log analyse oefening

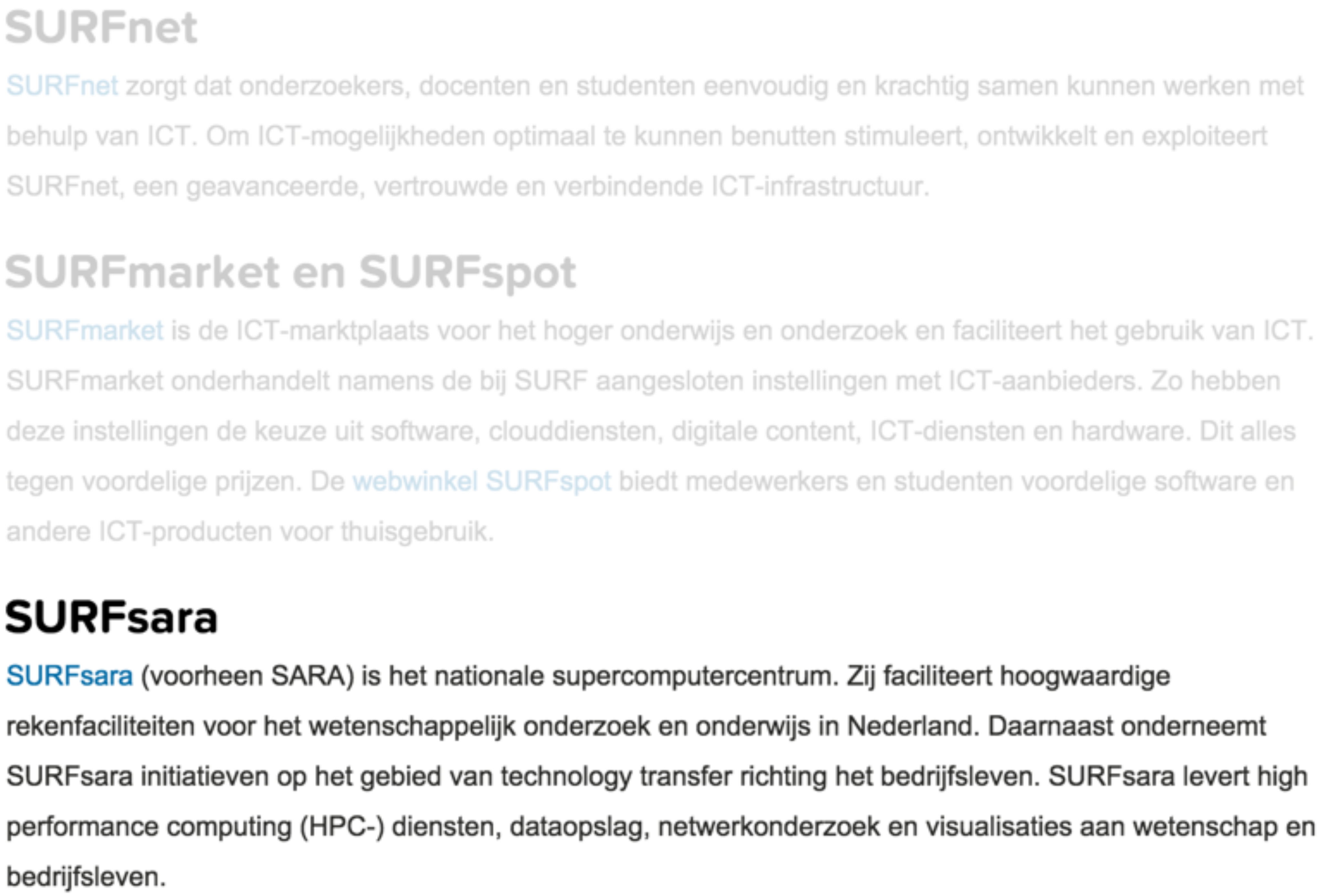
We zullen regelmatig pauzeren.

Stel vooral vragen!

What?

We provide large scale compute and data services for academic and research institutes

Who?



SURFnet

SURFnet zorgt dat onderzoekers, docenten en studenten eenvoudig en krachtig samen kunnen werken met behulp van ICT. Om ICT-mogelijkheden optimaal te kunnen benutten stimuleert, ontwikkelt en exploiteert SURFnet, een geavanceerde, vertrouwde en verbindende ICT-infrastructuur.

SURFmarket en SURFspot

SURFmarket is de ICT-marktplaats voor het hoger onderwijs en onderzoek en faciliteert het gebruik van ICT. SURFmarket onderhandelt namens de bij SURF aangesloten instellingen met ICT-aanbieders. Zo hebben deze instellingen de keuze uit software, clouddiensten, digitale content, ICT-diensten en hardware. Dit alles tegen voordelige prijzen. De webwinkel SURFspot biedt medewerkers en studenten voordelige software en andere ICT-producten voor thuisgebruik.

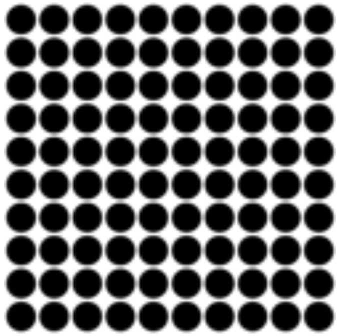
SURFsara

SURFsara (voorheen SARA) is het nationale supercomputercentrum. Zij faciliteert hoogwaardige rekenfaciliteiten voor het wetenschappelijk onderzoek en onderwijs in Nederland. Daarnaast onderneemt SURFsara initiatieven op het gebied van technology transfer richting het bedrijfsleven. SURFsara levert high performance computing (HPC-) diensten, dataopslag, netwerkonderzoek en visualisaties aan wetenschap en bedrijfsleven.

Cartesius

Lisa

Visualisation



Grid

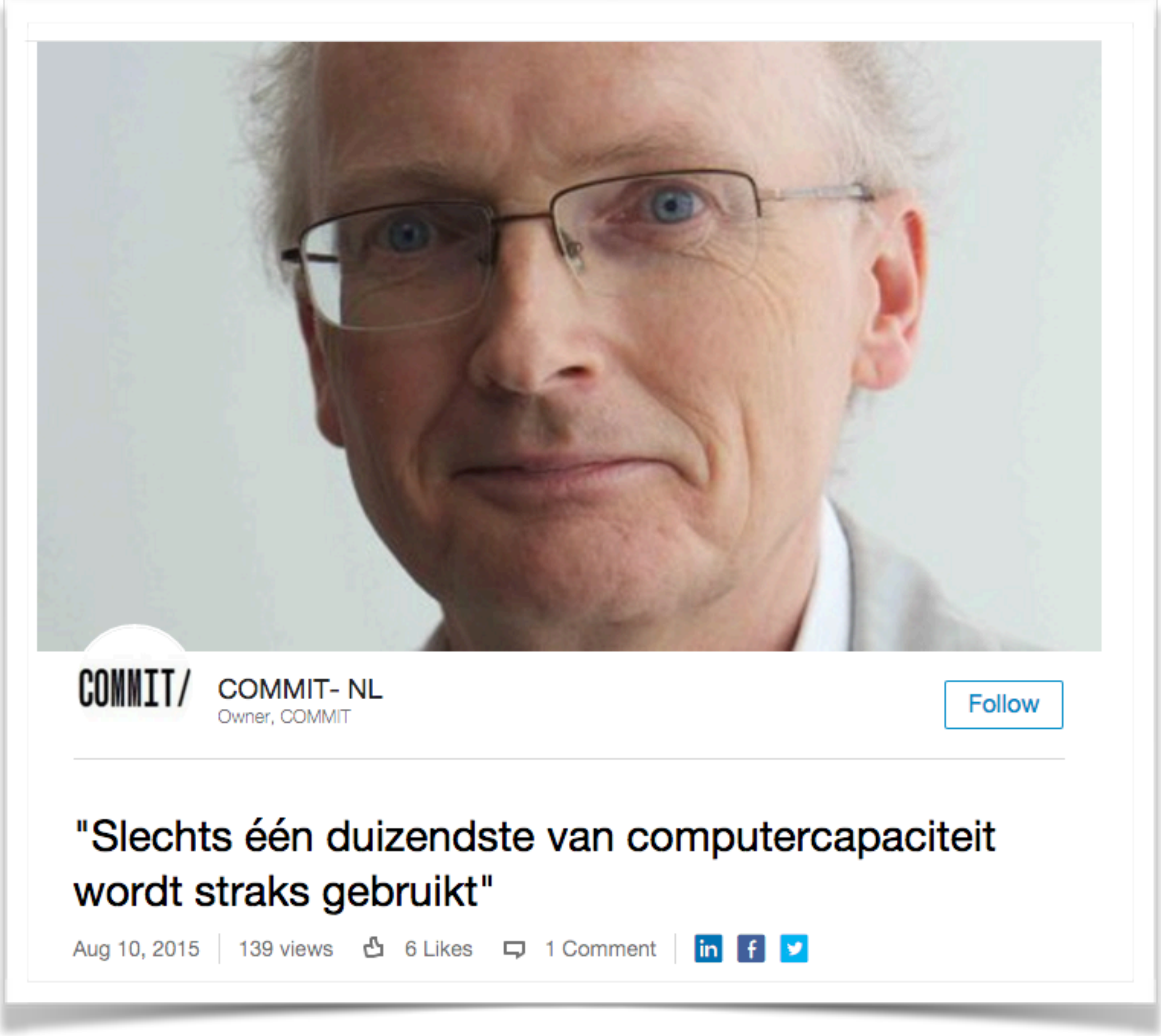
Cloud

Hadoop

Data Services

Gebrekkige kennis van parallel programmeren zorgt ervoor dat straks slechts één duizendste van de capaciteit van computers wordt gebruikt. Hierdoor zijn berekeningen onnodig langzaam en onnauwkeurig. Dat vertraagt de ontwikkeling van de Nederlandse kenniseconomie.

Henri Bal - VU Amsterdam



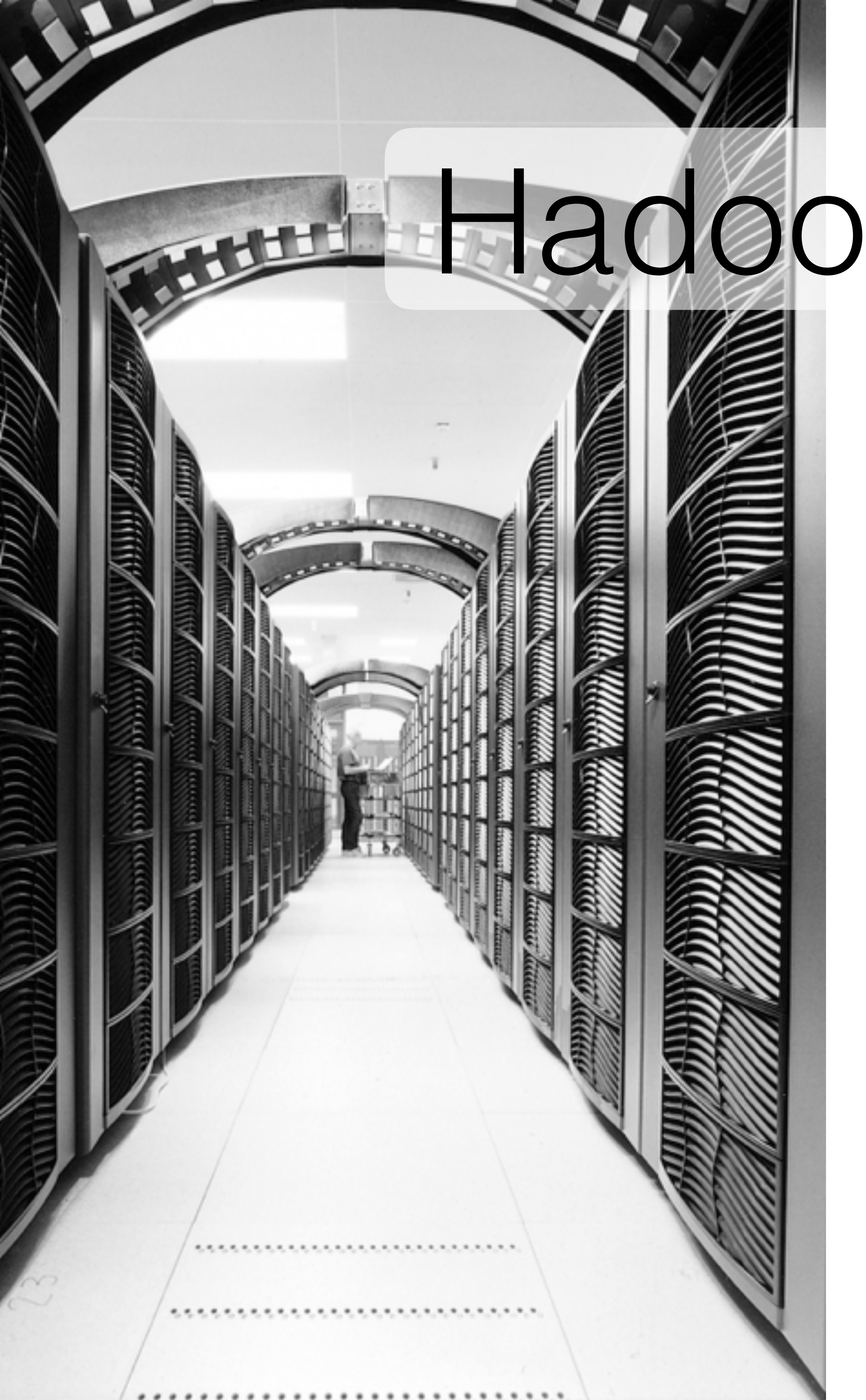
COMMIT/ COMMIT- NL
Owner, COMMIT

Follow

"Slechts één duizendste van computercapaciteit wordt straks gebruikt"

Aug 10, 2015 | 139 views 6 Likes 1 Comment

Hadoop cluster



197 machines –
  64 GB RAM – total 10TB RAM

1576 parallel jobs

4 x 2.4 TB Disk – in total 2.3 PB

Hortonworks HDP 2.2 (Hadoop 2.6)
Kerberos authentication

YARN for Apache Spark, Storm, Pig, Hive, HBase and many more

SURF SARA

hadoop

APACHE HBASE

MESOS

Tomas Barton (@barton_tomas)

Spark

APACHE GIRAPH

elasticsearch.

cascading

HIVE

kafka

APACHE STORM™
Distributed • Resilient • Real-time

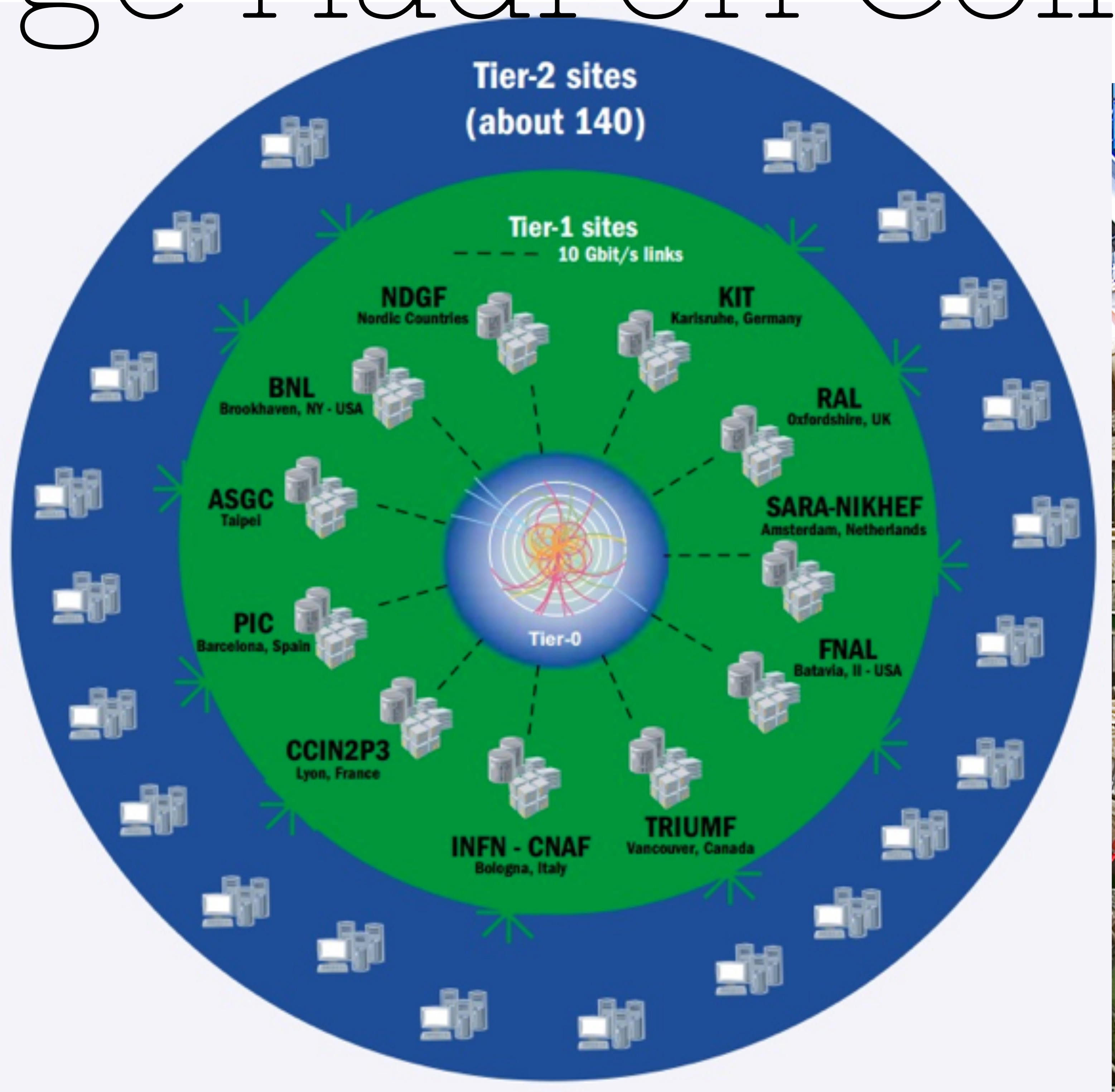
Big Data toolbox

Hadoop as Big Data's

Swiss Army Knife

Batch (MR, Pig)
Streaming (Storm, Spark)
In memory (Spark, Tez)
SQL (Hive
Database (HBase)

Large Hadron Collider



Big Data

There is a need for systems that
can work with different kinds of
data formats and sources
without requiring strict schema
definitions up front,
do it at scale
and cost-effective.

This presentation: Technology perspective

- Big Data is about scalability
- Doing things big, really big changes business, analytics, and technology
- Big Data technology is very much a programmers or hackers (in the good sense) world.
- The technology is rapidly changing and to wait for 'mature' products could mean missing out
- Understanding the fundamentals of the technology helps to understand opportunities for Big Data applications and use cases.

Doing Big Data science

Janet Echelman

'Big Data' in practice often means small datasets in relational databases on laptops or traditional clusters

Scalability is often ignored

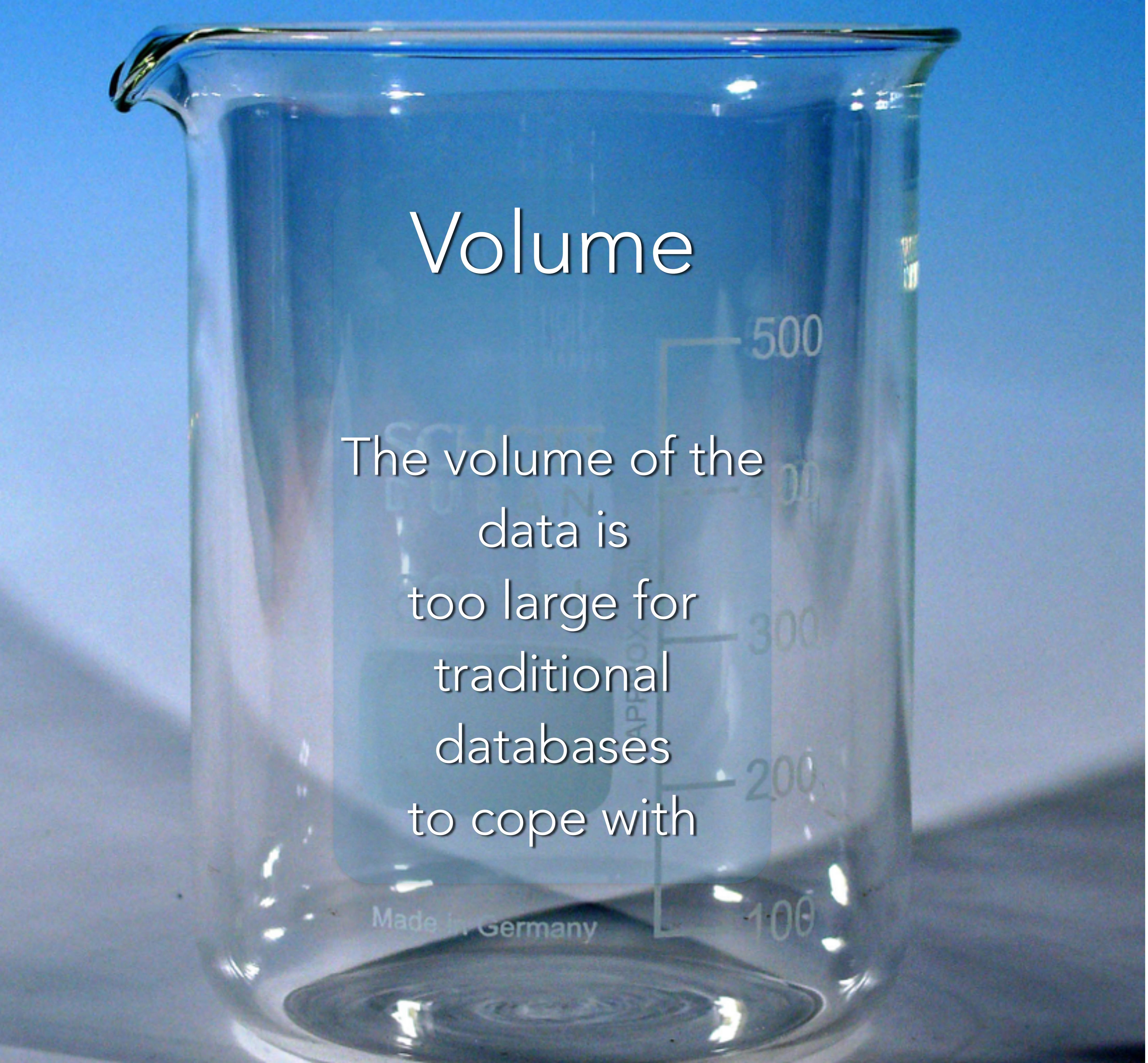
Knowledge of scalable solutions is scarce - also among data scientists

Making use of scale

Traditional models for HPC are difficult

Big Data programming models are much easier

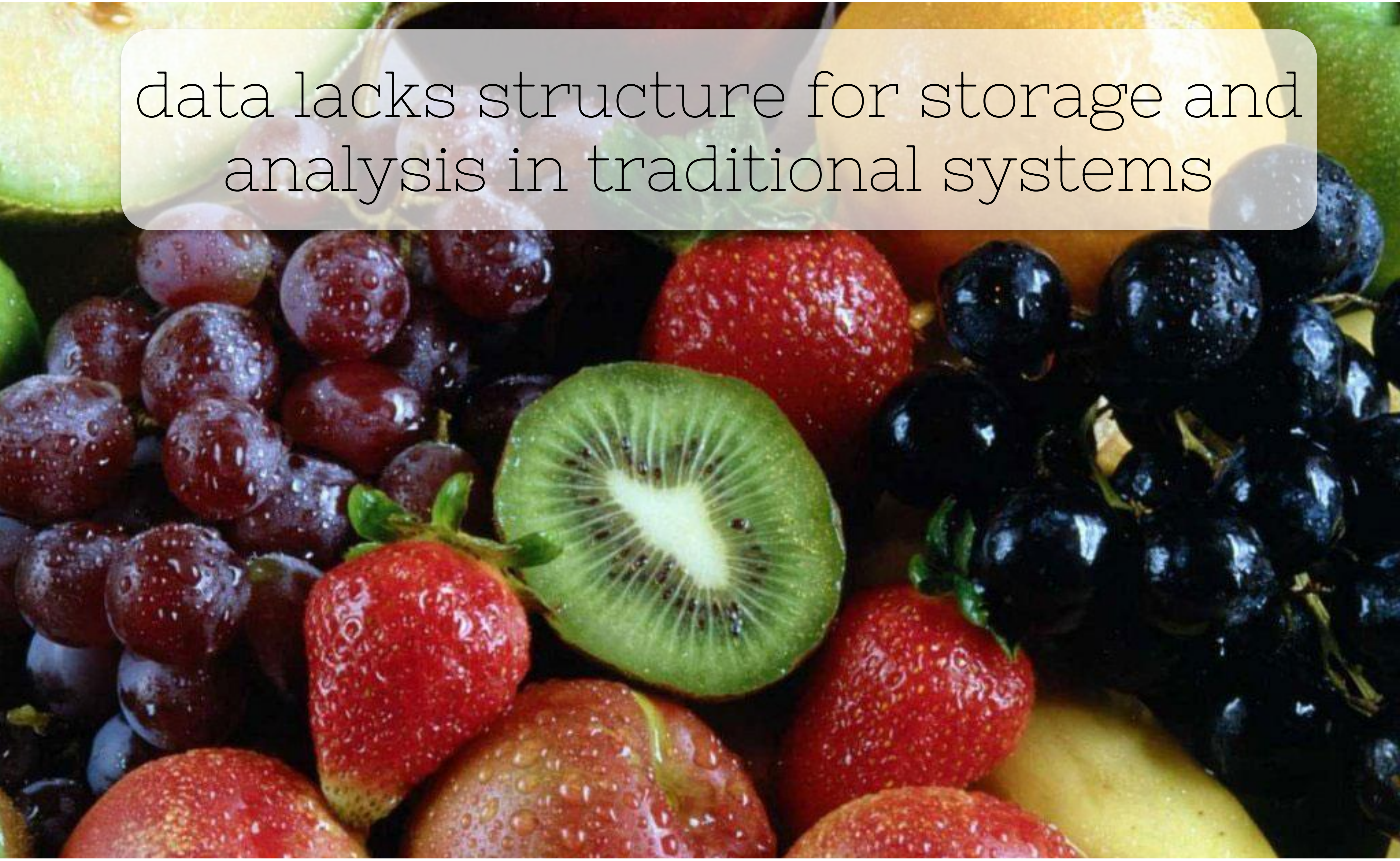
Big Data technology often isolates complexity



Volume

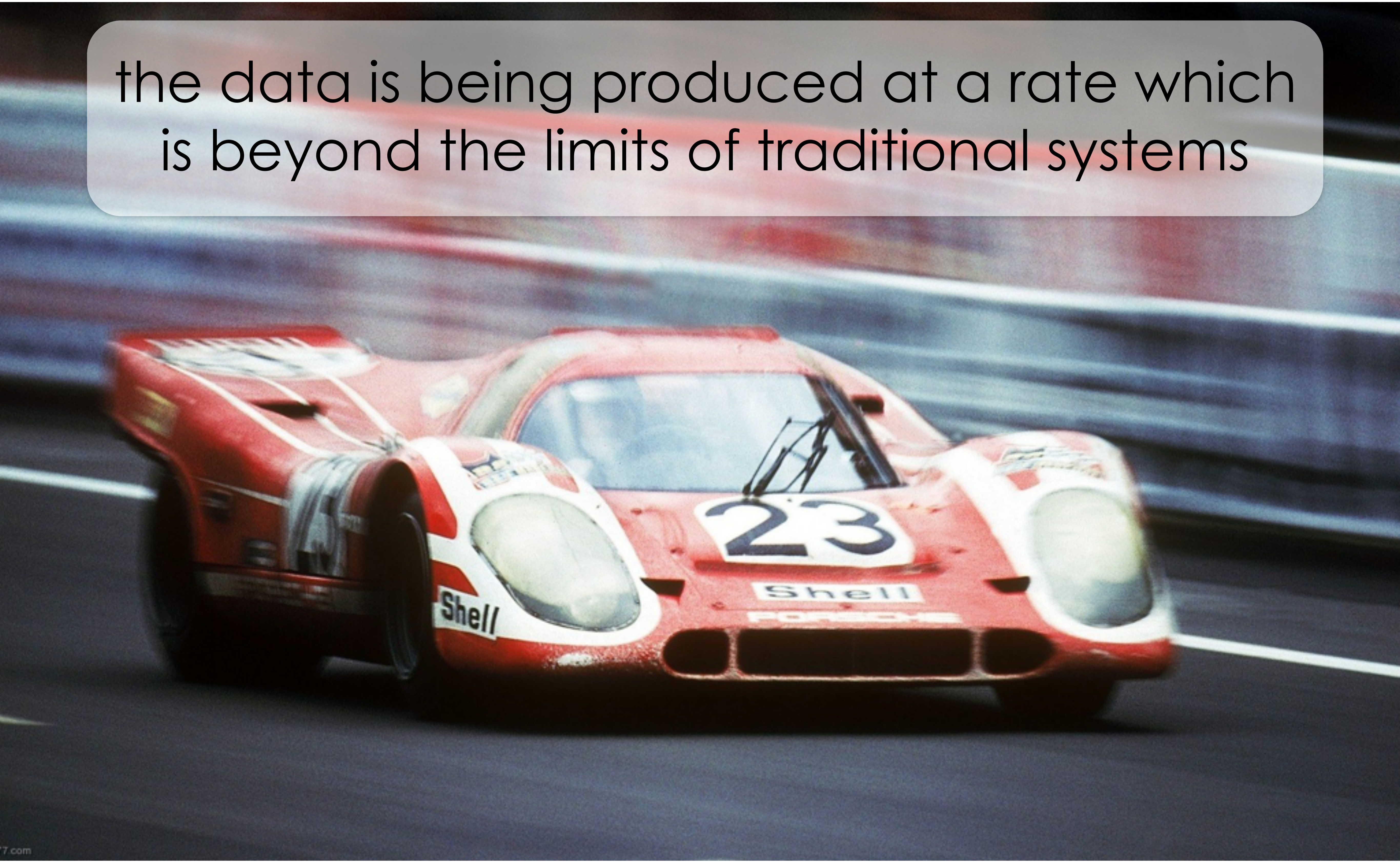
The volume of the data is too large for traditional databases to cope with

Variety



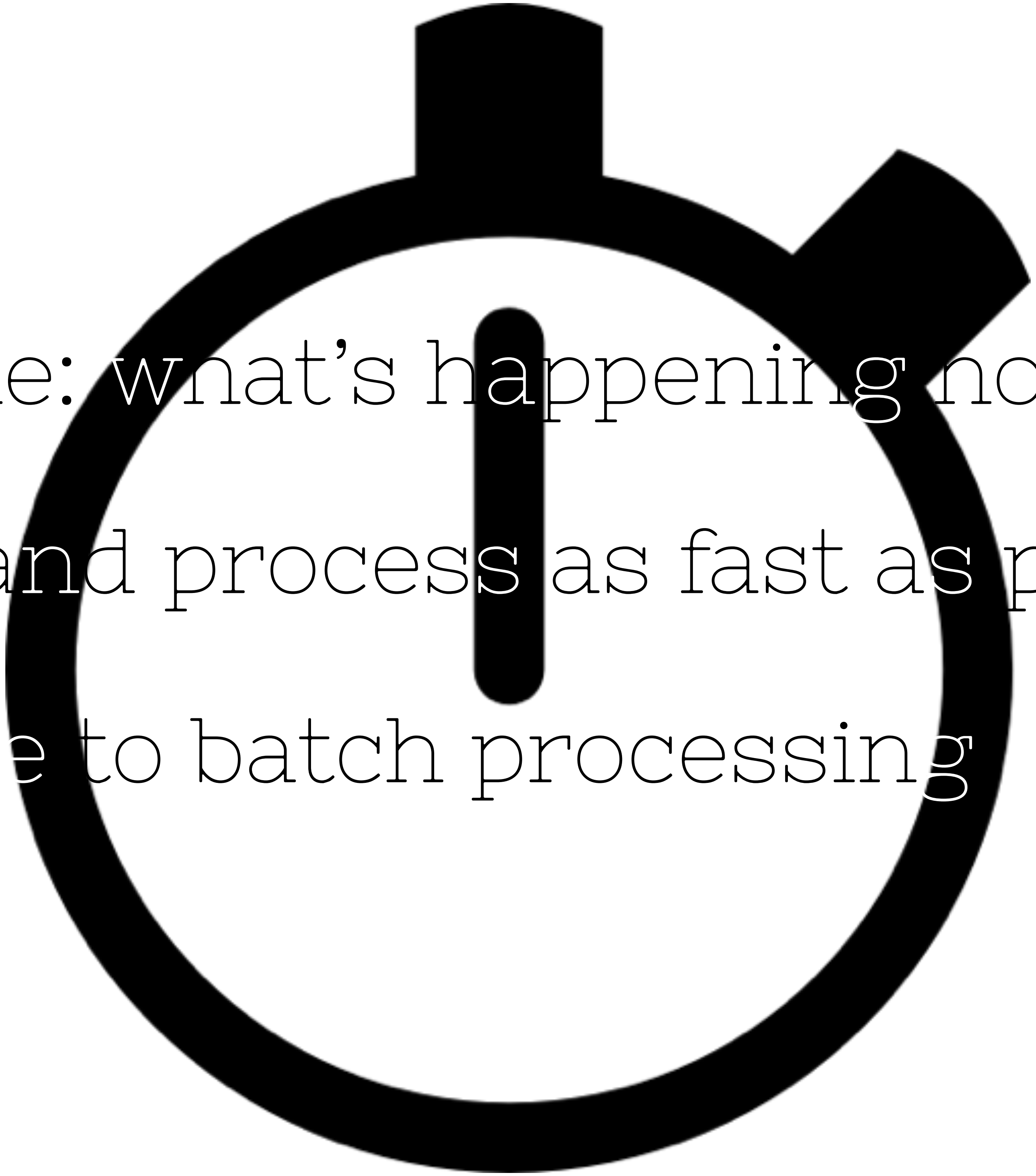
data lacks structure for storage and analysis in traditional systems

Velocity



the data is being produced at a rate which is beyond the limits of traditional systems

Real Time



Real time: what’s happening now?

Collect and process as fast as possible

Compare to batch processing

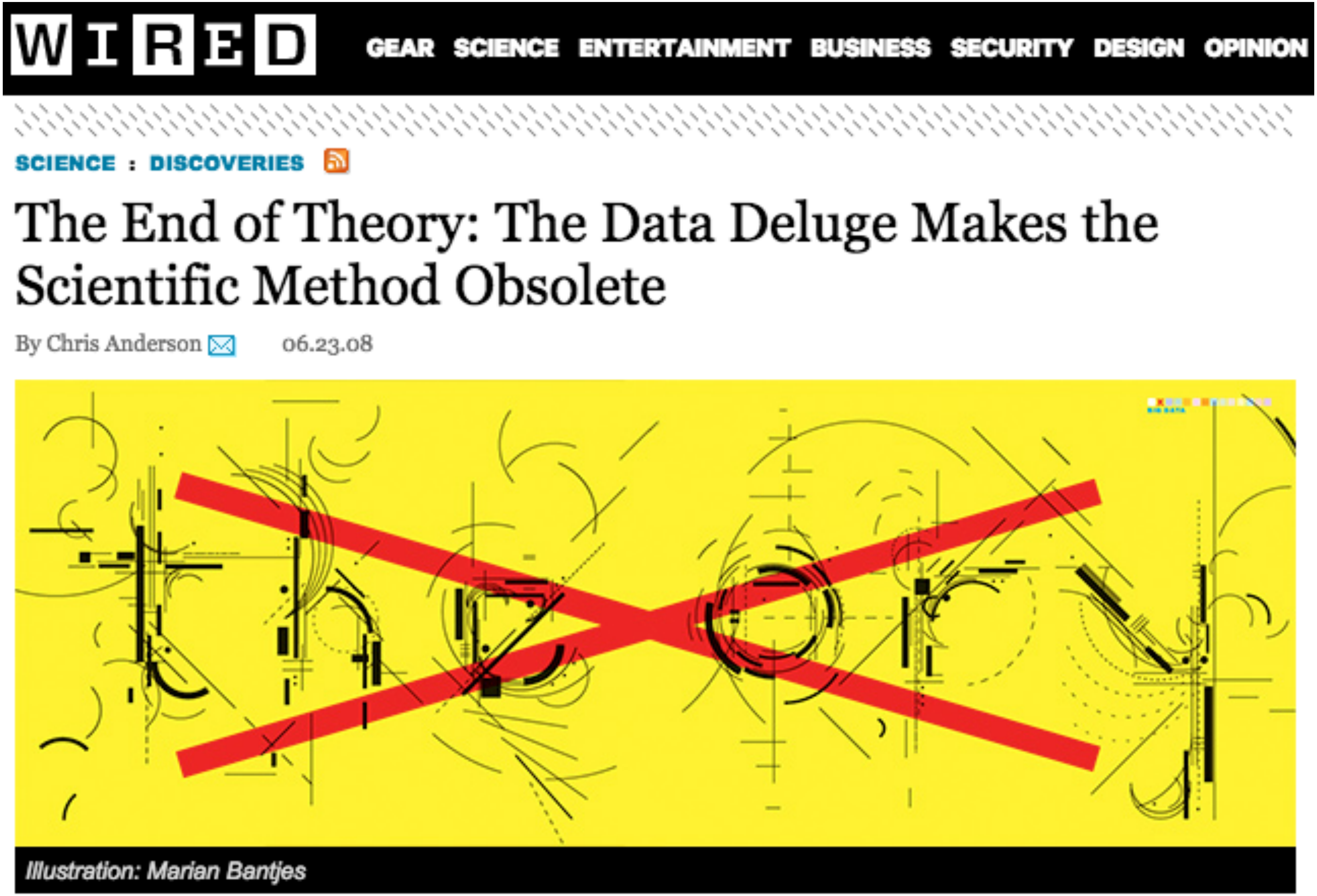
So what’s new

- Scaling is difficult - we want it to be easy and scale massively (if needed)
- Traditional databases want us to define schema’s and structure data BEFORE we store it - we want to store first and worry about schema’s later
- We want it cheap

No aselect samples but all (or almost all) data

Not only accurate but sloppy data – accept inaccuracy

Not causal models but rather correlations

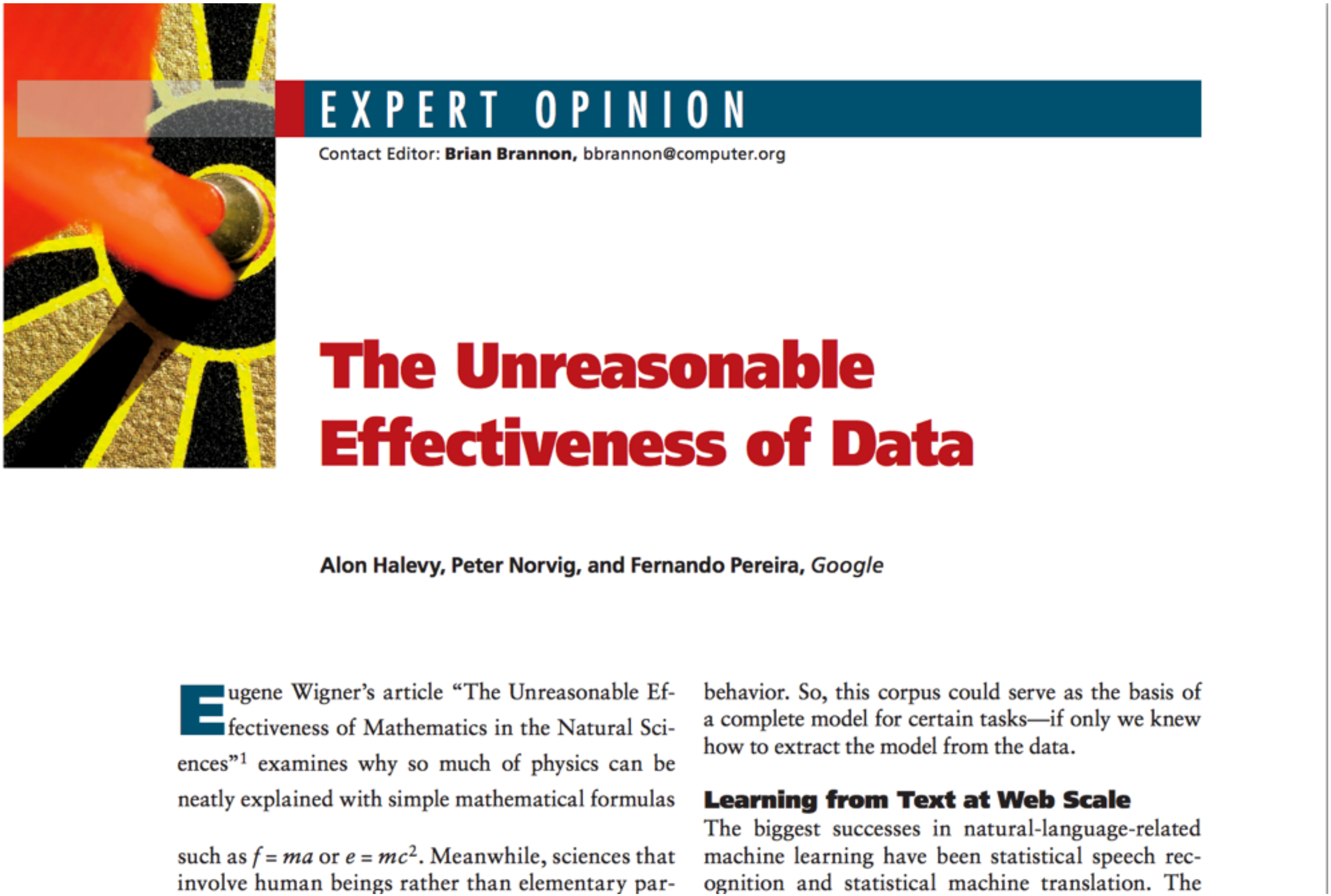


SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

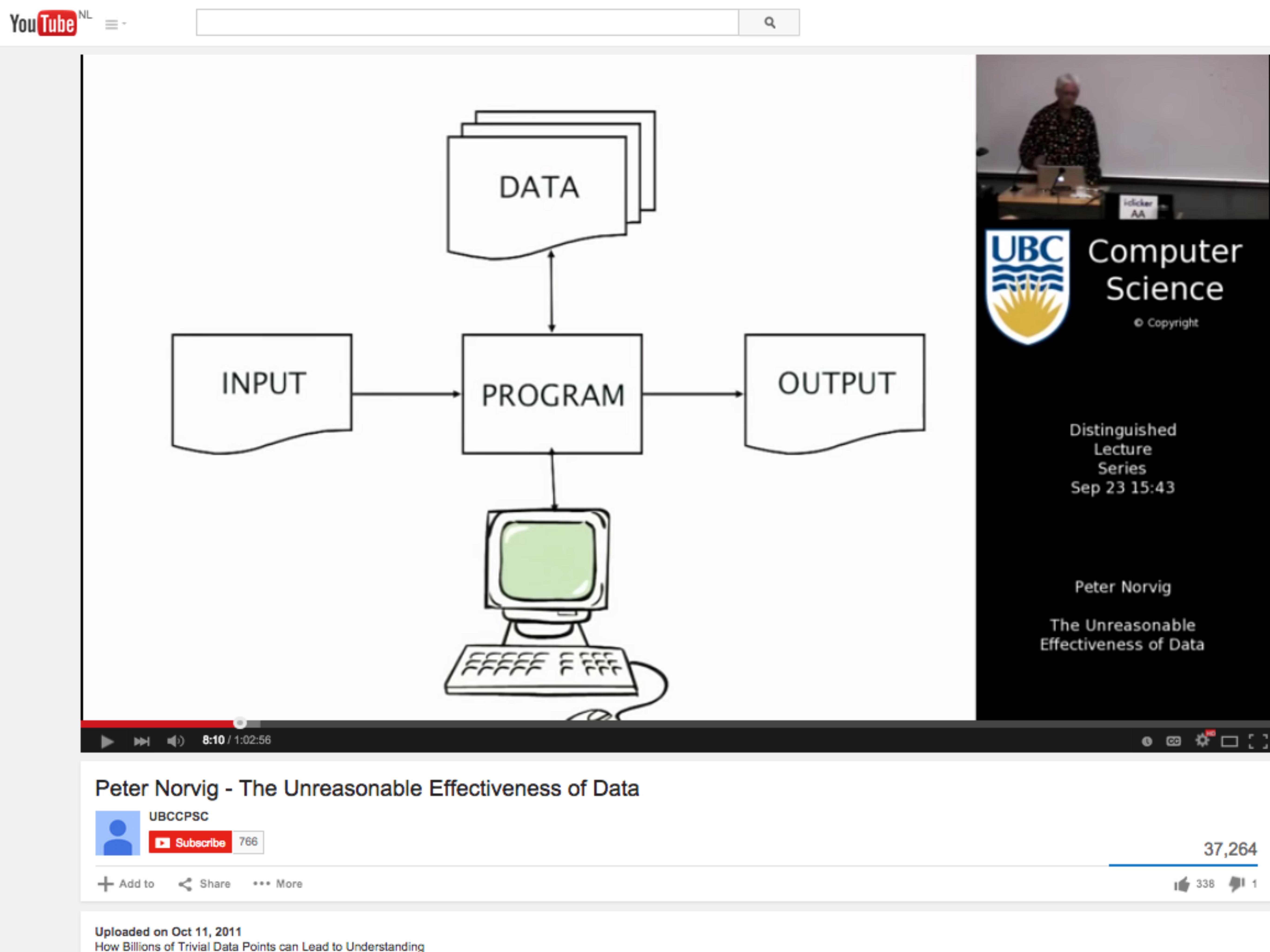
Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary par-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The



Computer Science

© Copyright

Distinguished Lecture Series Sep 23 15:43

Peter Norvig

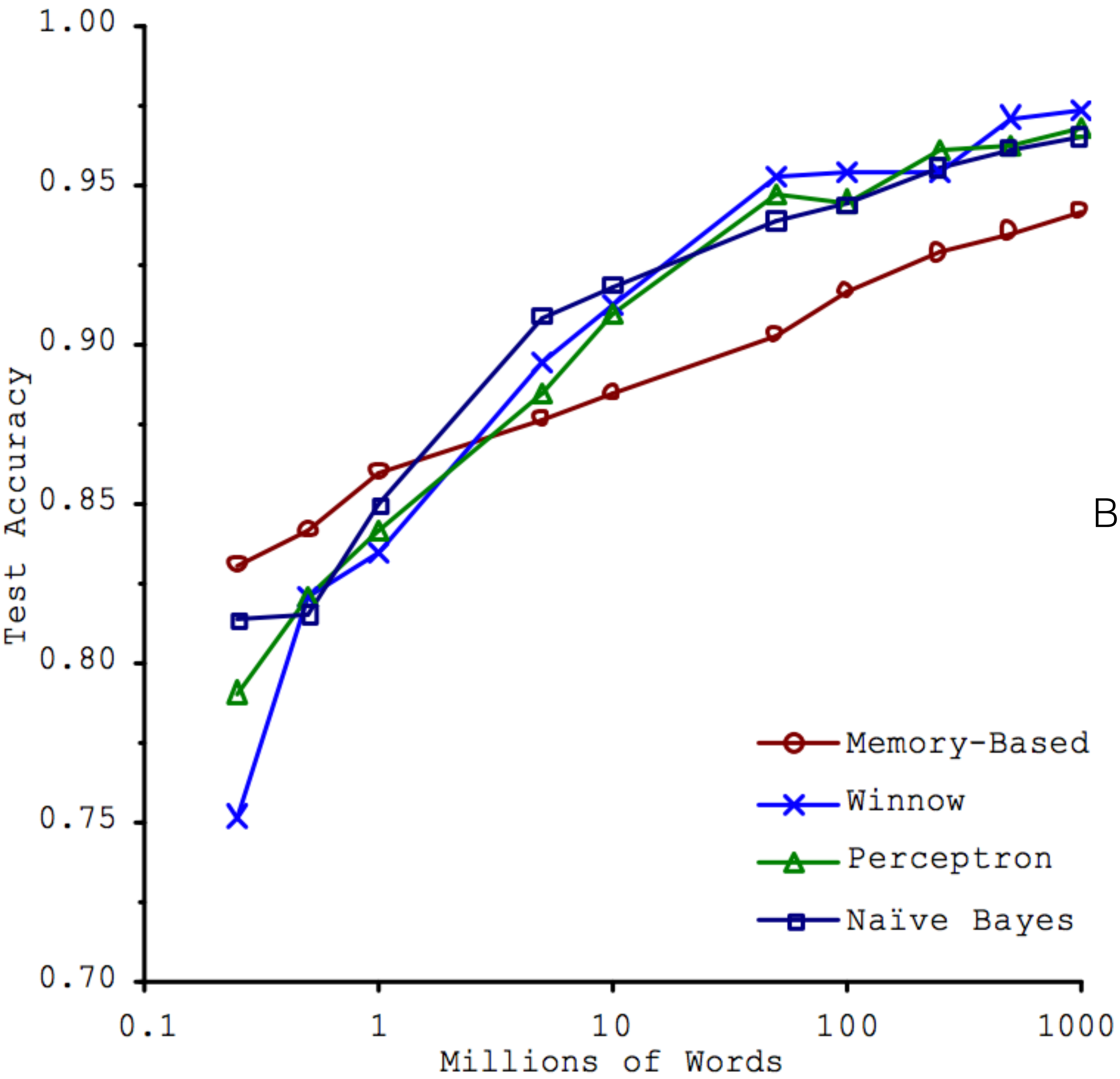
The Unreasonable Effectiveness of Data

Peter Norvig - The Unreasonable Effectiveness of Data

UBCCPSC

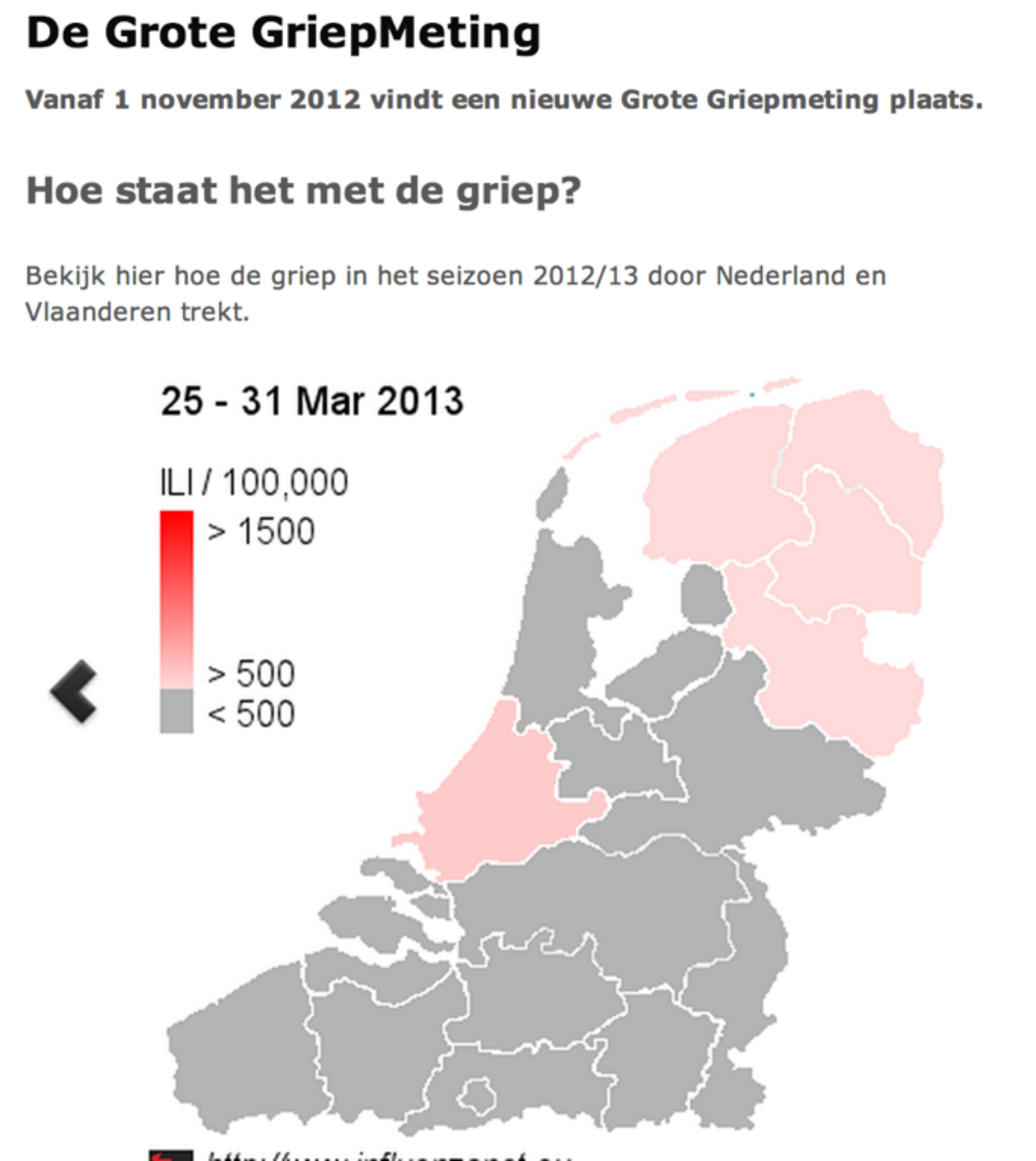
Uploaded on Oct 11, 2011

How Billions of Trivial Data Points can Lead to Understanding



Banko & Brill 2001

Google translate



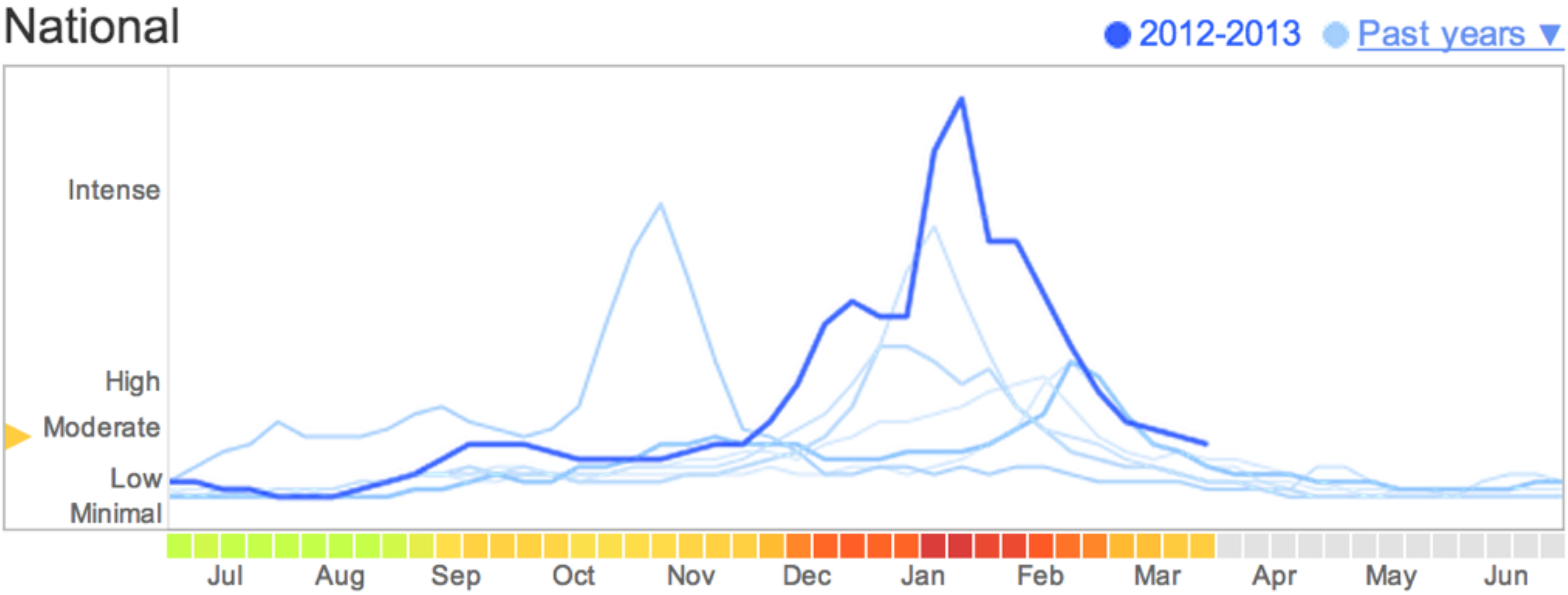
De Grote GriepMeting

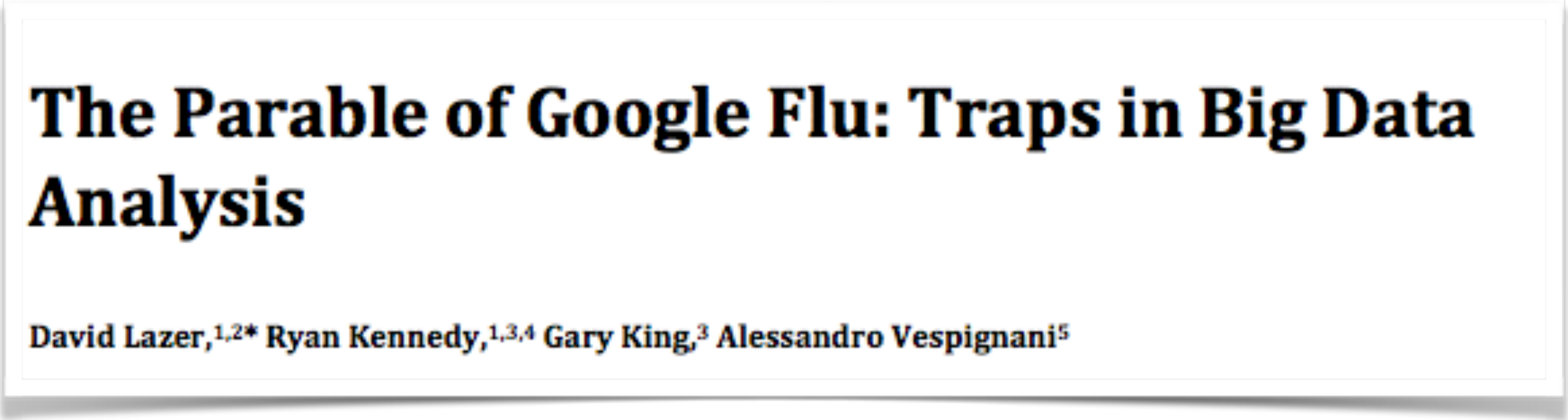
Vanaf 1 november 2012 vindt een nieuwe Grote Griepmeting plaats.

Hoe staat het met de griep?

Bekijk hier hoe de griep in het seizoen 2012/13 door Nederland en Vlaanderen trekt.

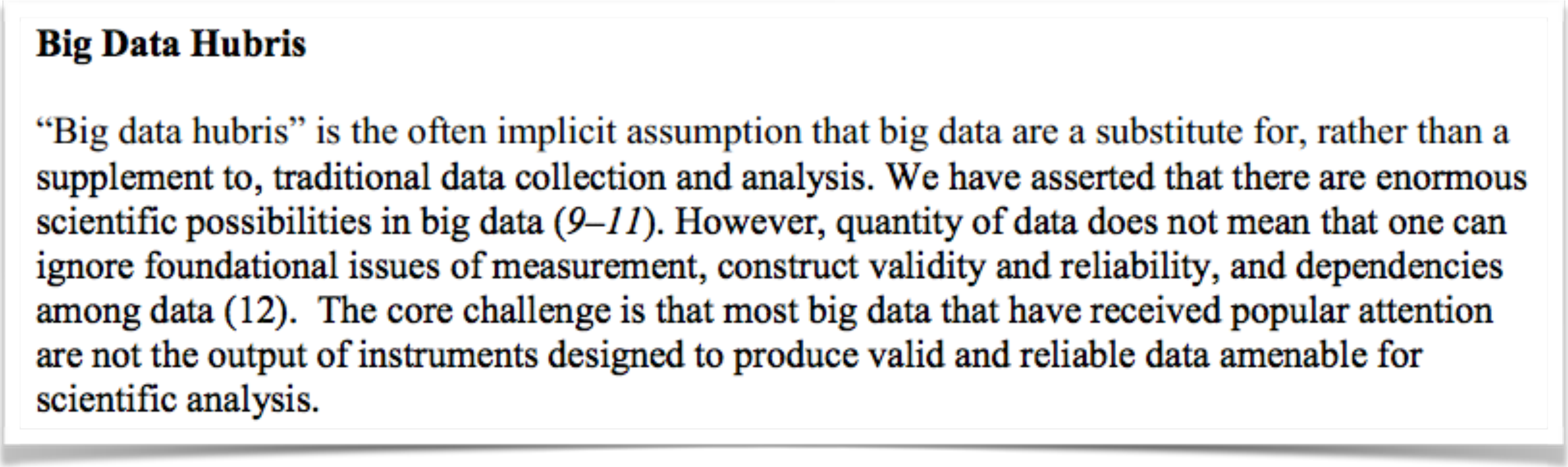
Google flu


National
● 2012-2013
● Past years ▼
Intense
High
Moderate
Low
Minimal
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun



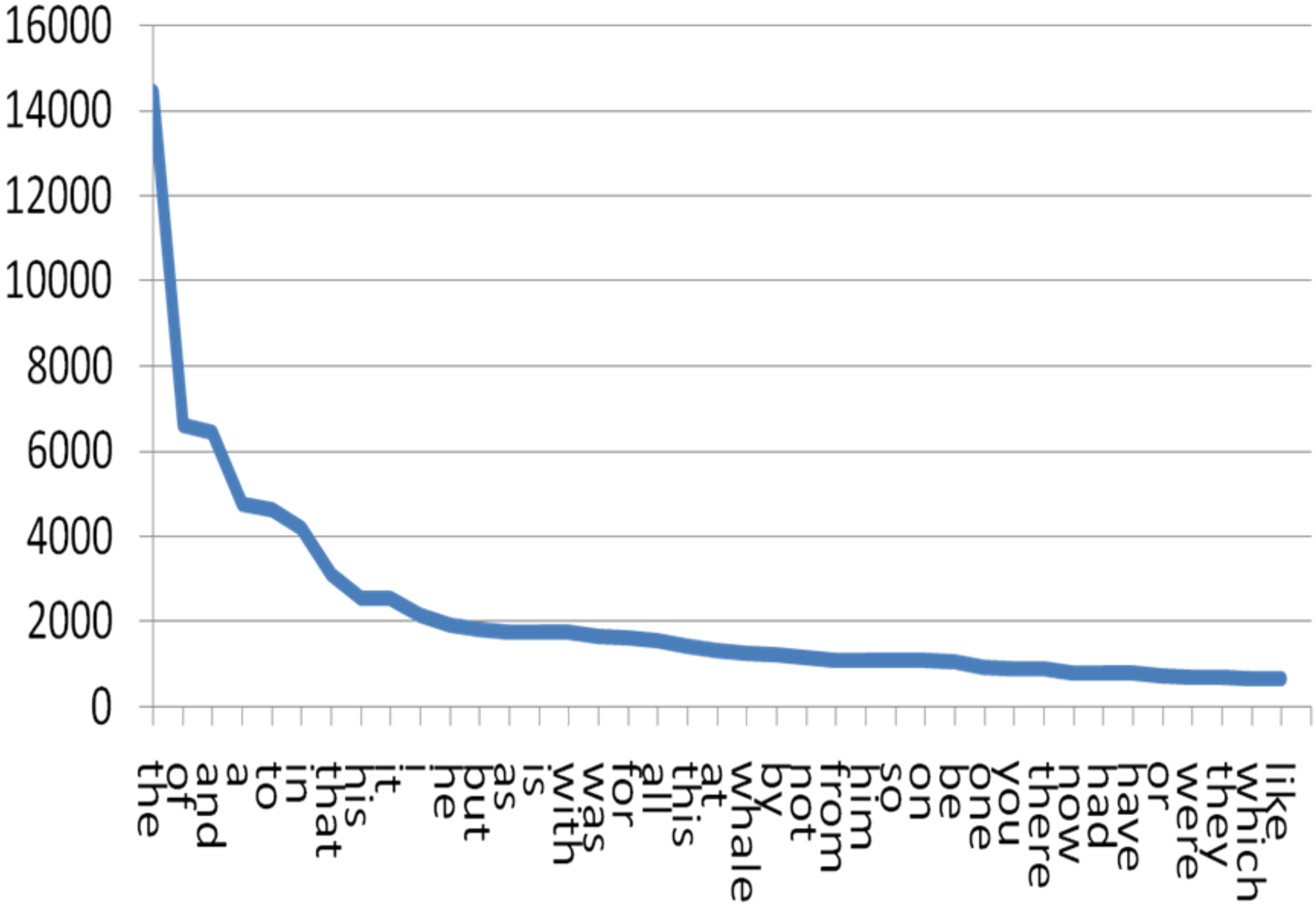
The Parable of Google Flu: Traps in Big Data Analysis

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



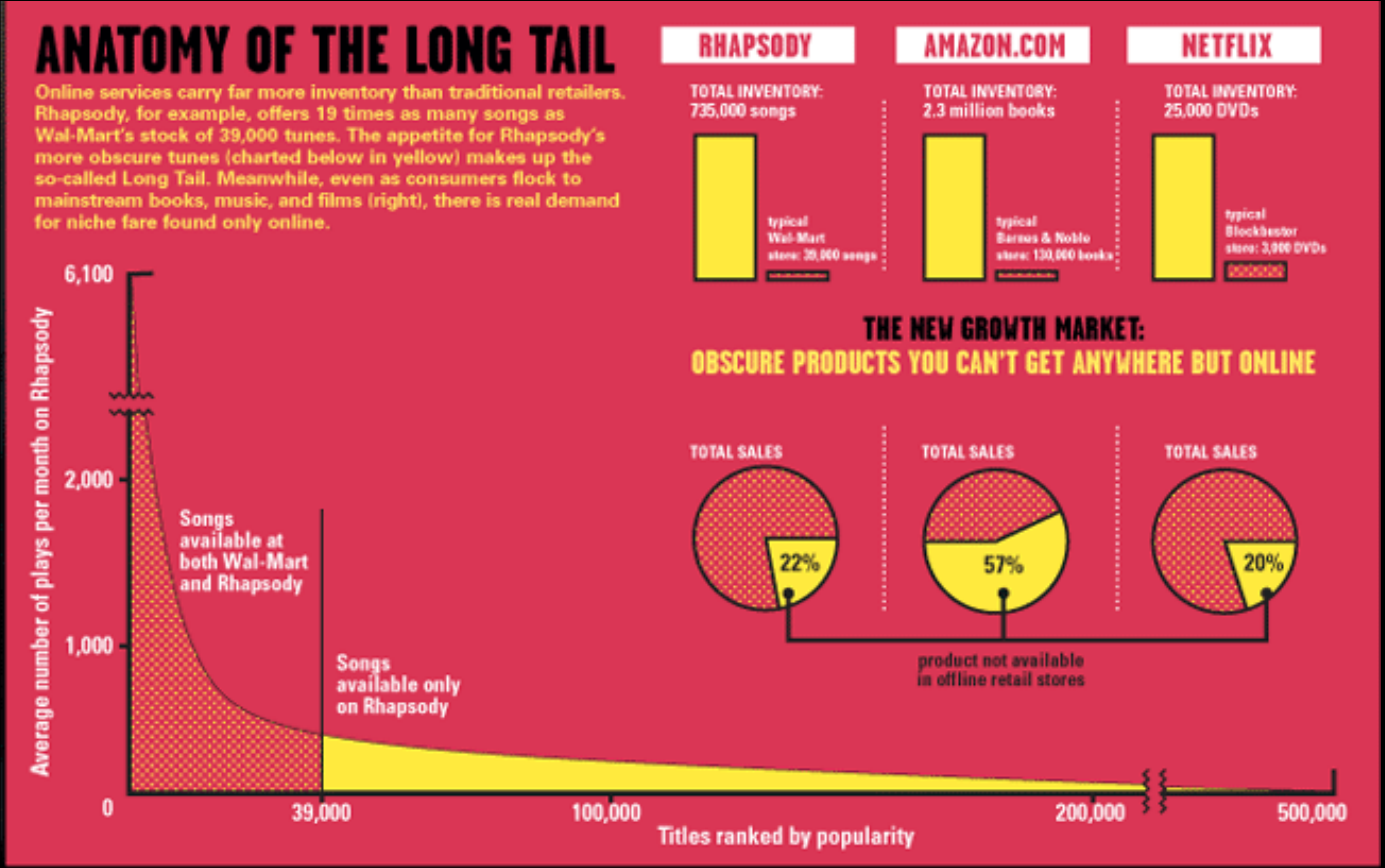
Big Data Hubris

“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. 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, 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.

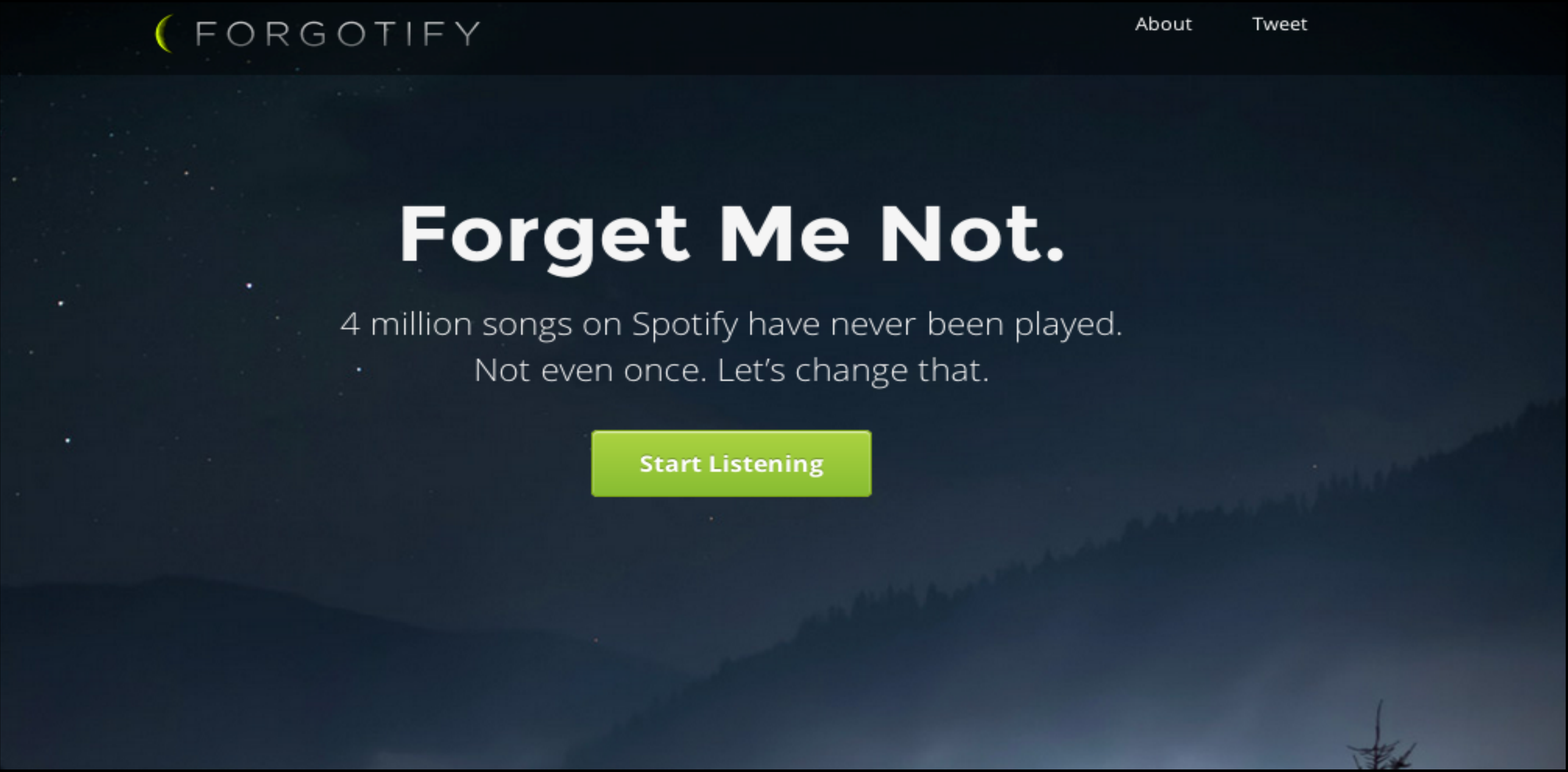

16000
14000
12000
10000
8000
6000
4000
2000
0
the
of
and
a
to
in
that
his
it
i
he
but
as
is
with
was
for
all
this
at
whale
by
not
from
him
so
on
be
one
you
there
now
had
have
or
were
they
which
like

ANATOMY OF THE LONG TAIL

Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.



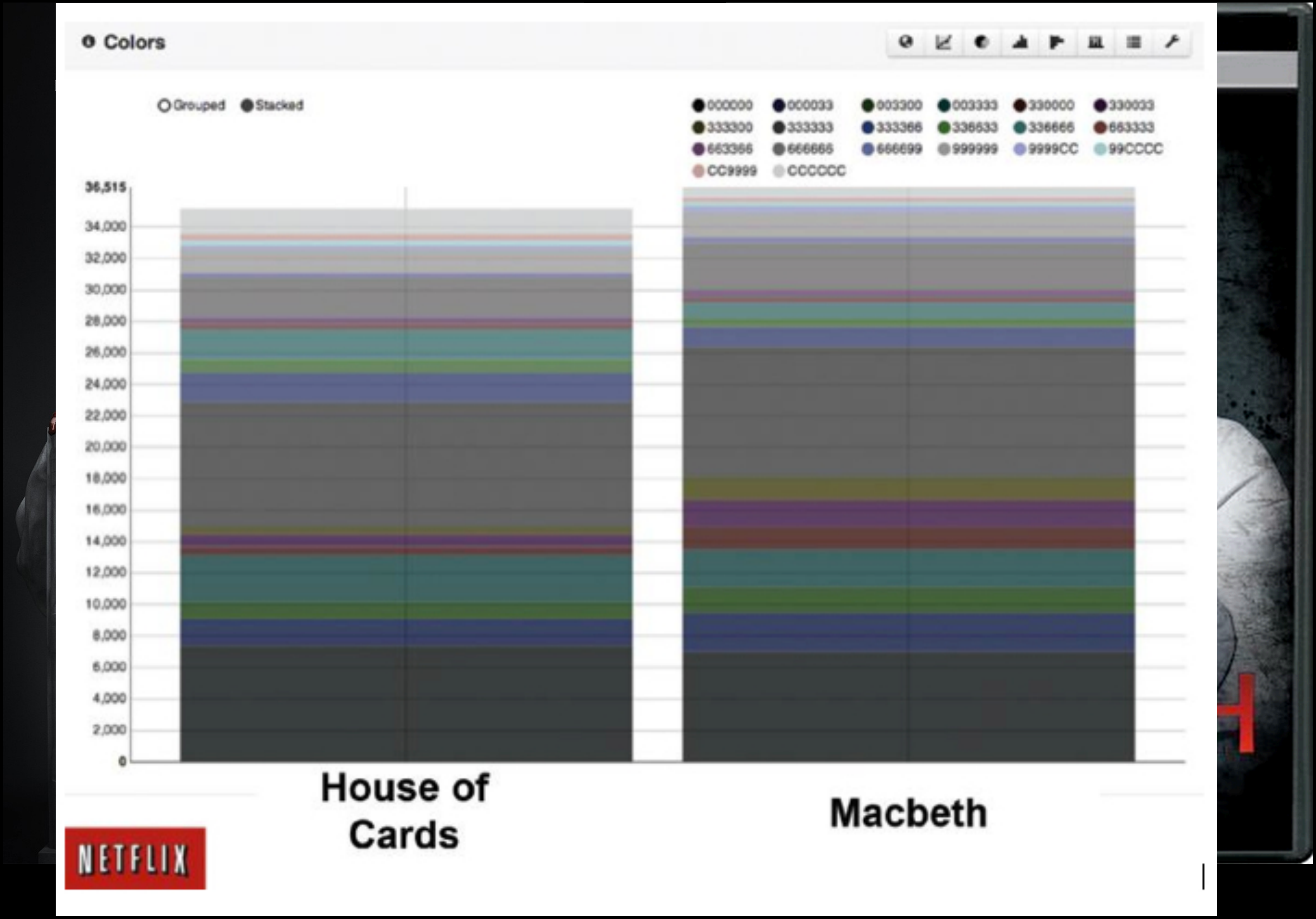
Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks



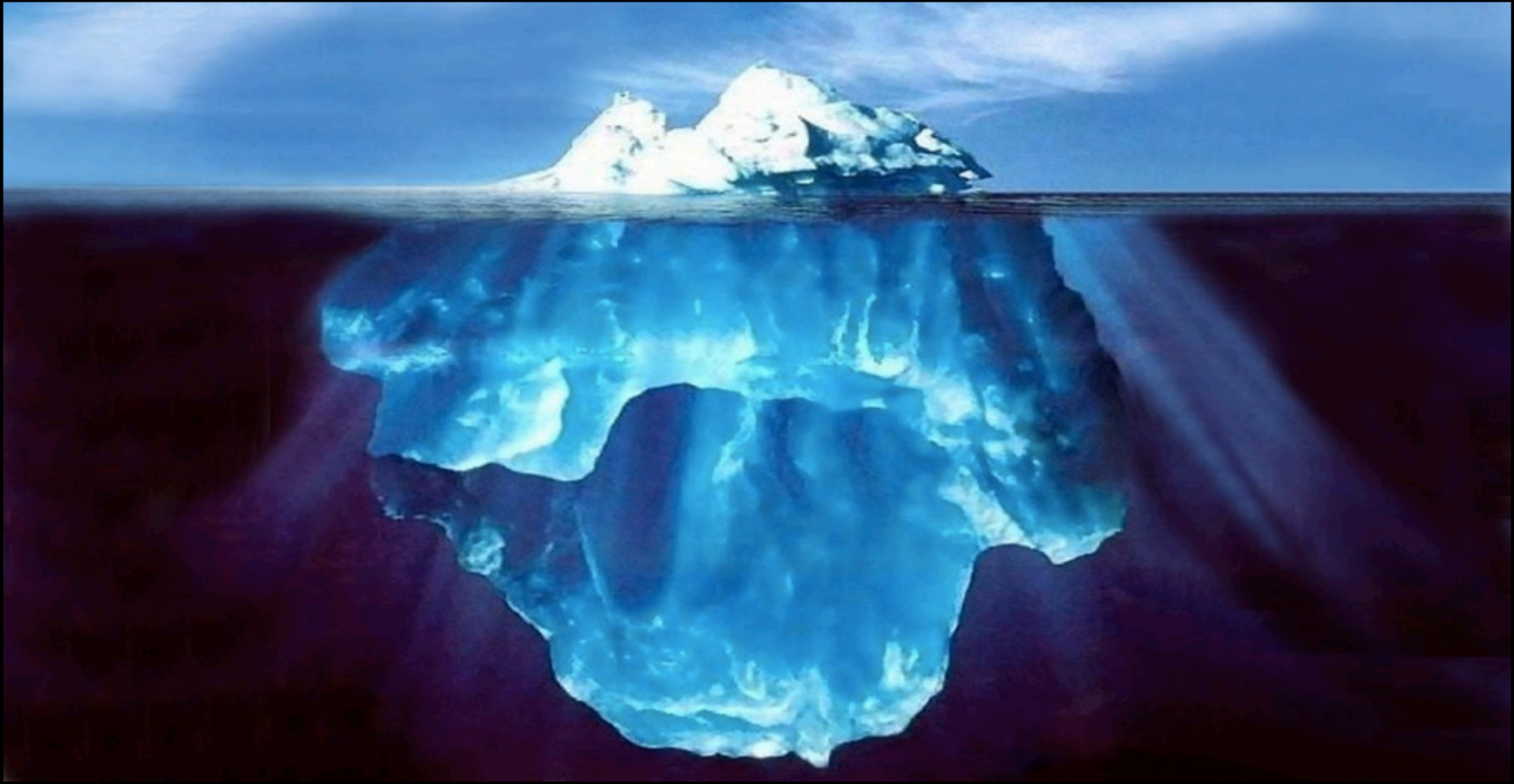
FORGOTIFY

Forget Me Not.

4 million songs on Spotify have never been played.
Not even once. Let's change that.



http://www.wired.com/insights/2014/03/big-data-lessons-netflix/

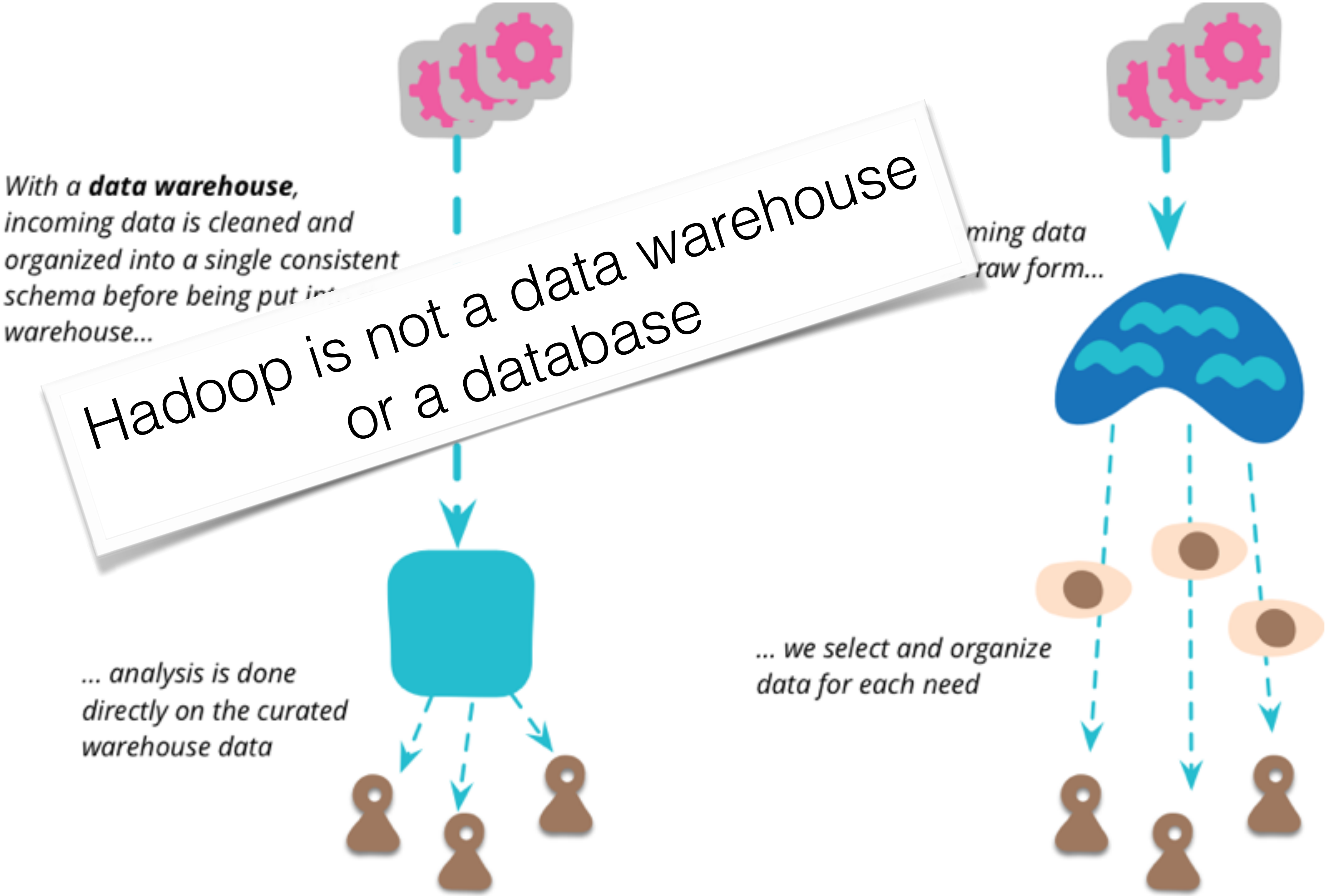


Immutability and data

The data lake

Never (no never) delete anything

Data lake

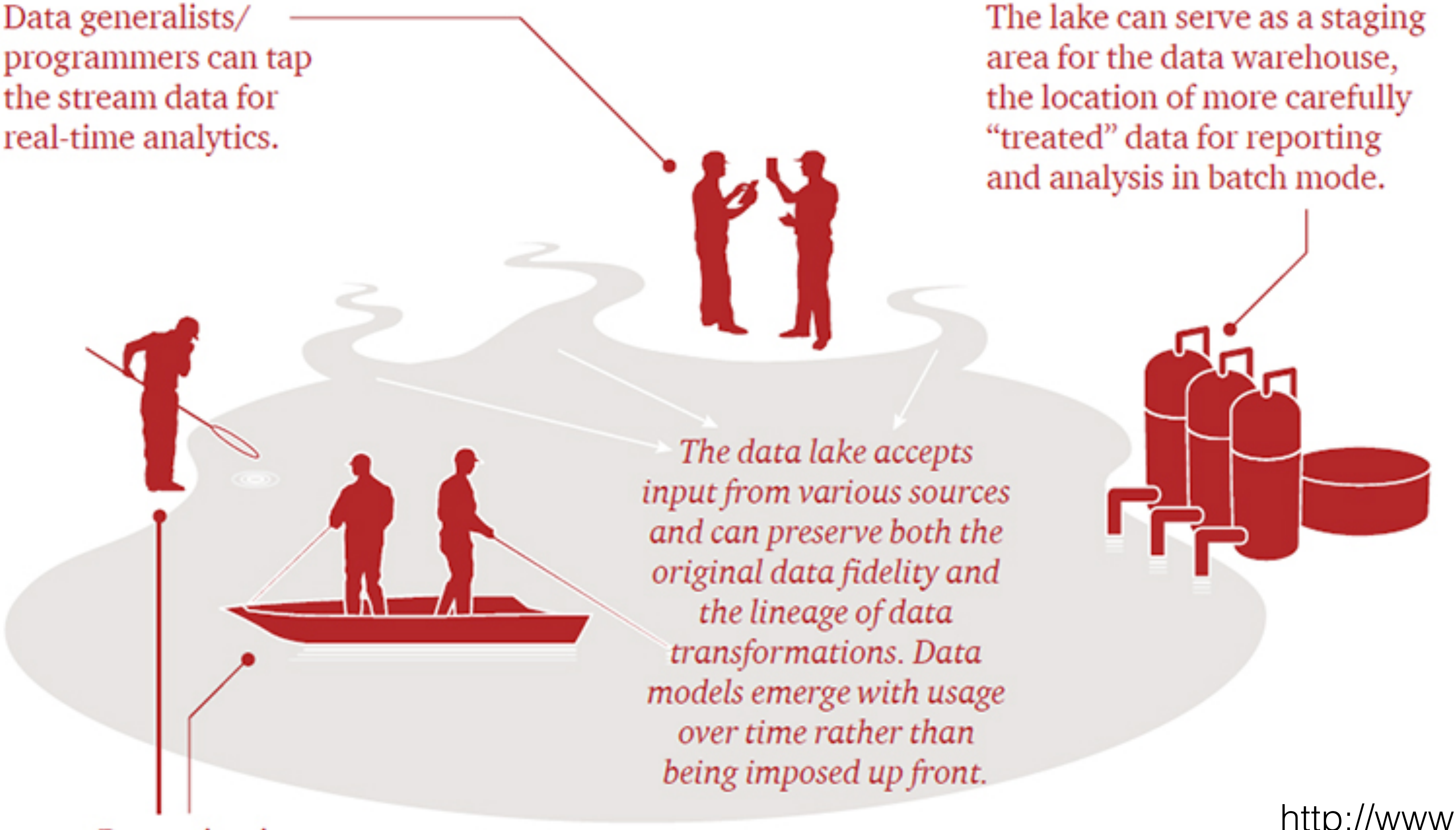

With a **data warehouse**, incoming data is cleaned and organized into a single consistent schema before being put in
warehouse...
ming data
raw form...
... analysis is done directly on the curated warehouse data
... we select and organize data for each need

Hadoop is not a data warehouse or a database

Taken from Martin Fowlers website

Data lake

A repository for large quantities and varieties of data, both structured and unstructured.



Data generalists/ programmers can tap the stream data for real-time analytics.

The lake can serve as a staging area for the data warehouse, the location of more carefully "treated" data for reporting and analysis in batch mode.

The data lake accepts input from various sources and can preserve both the original data fidelity and the lineage of data transformations. Data models emerge with usage over time rather than being imposed up front.

Data scientists use the lake for discovery and ideation.

Data lakes take advantage of commodity cluster computing techniques for massively scalable, low-cost storage of data files in any format.

http://www.pwc.com/