BIG DATA

Machiel Jansen

SURF SARA

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.

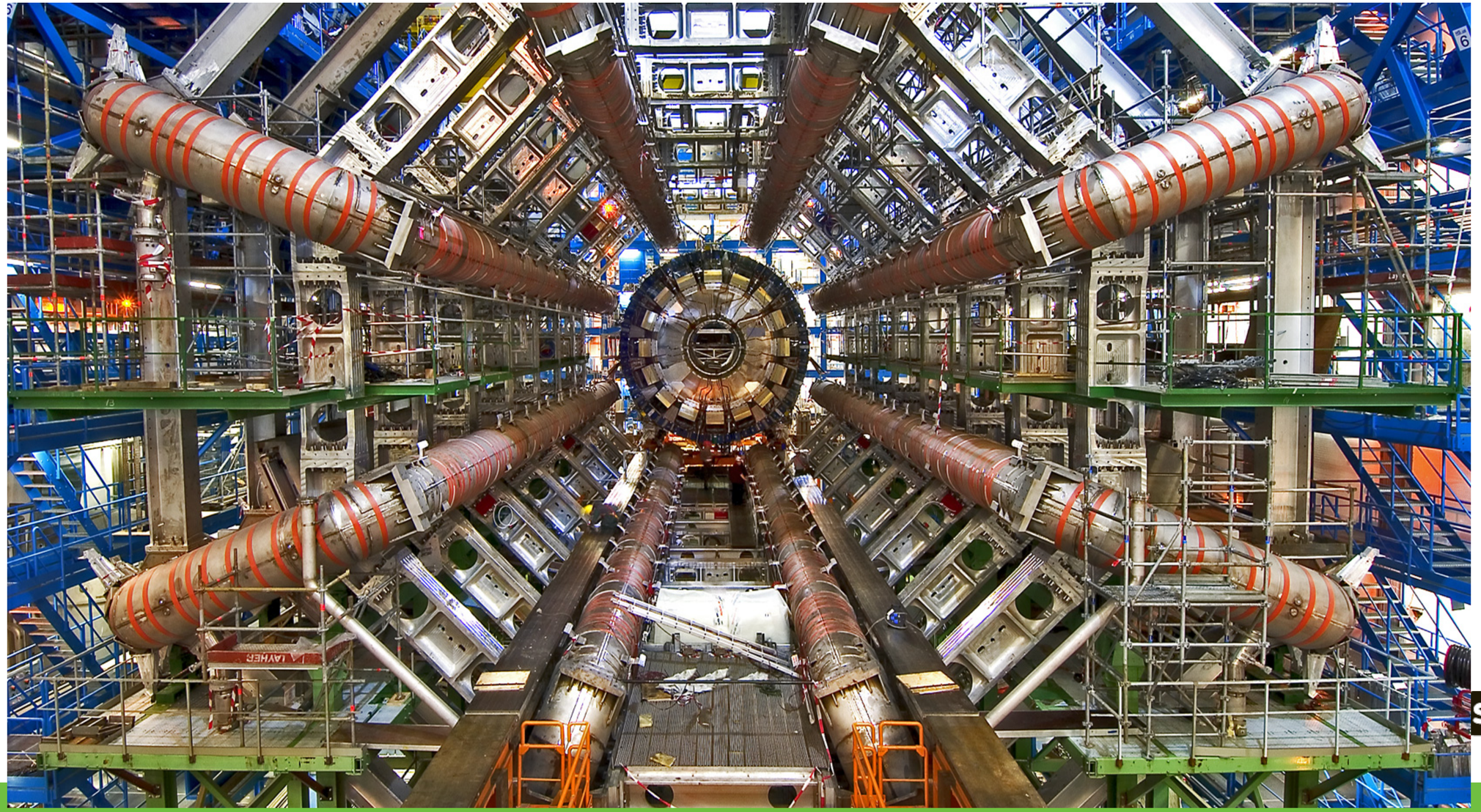
SURFmarket en SURFspot

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.

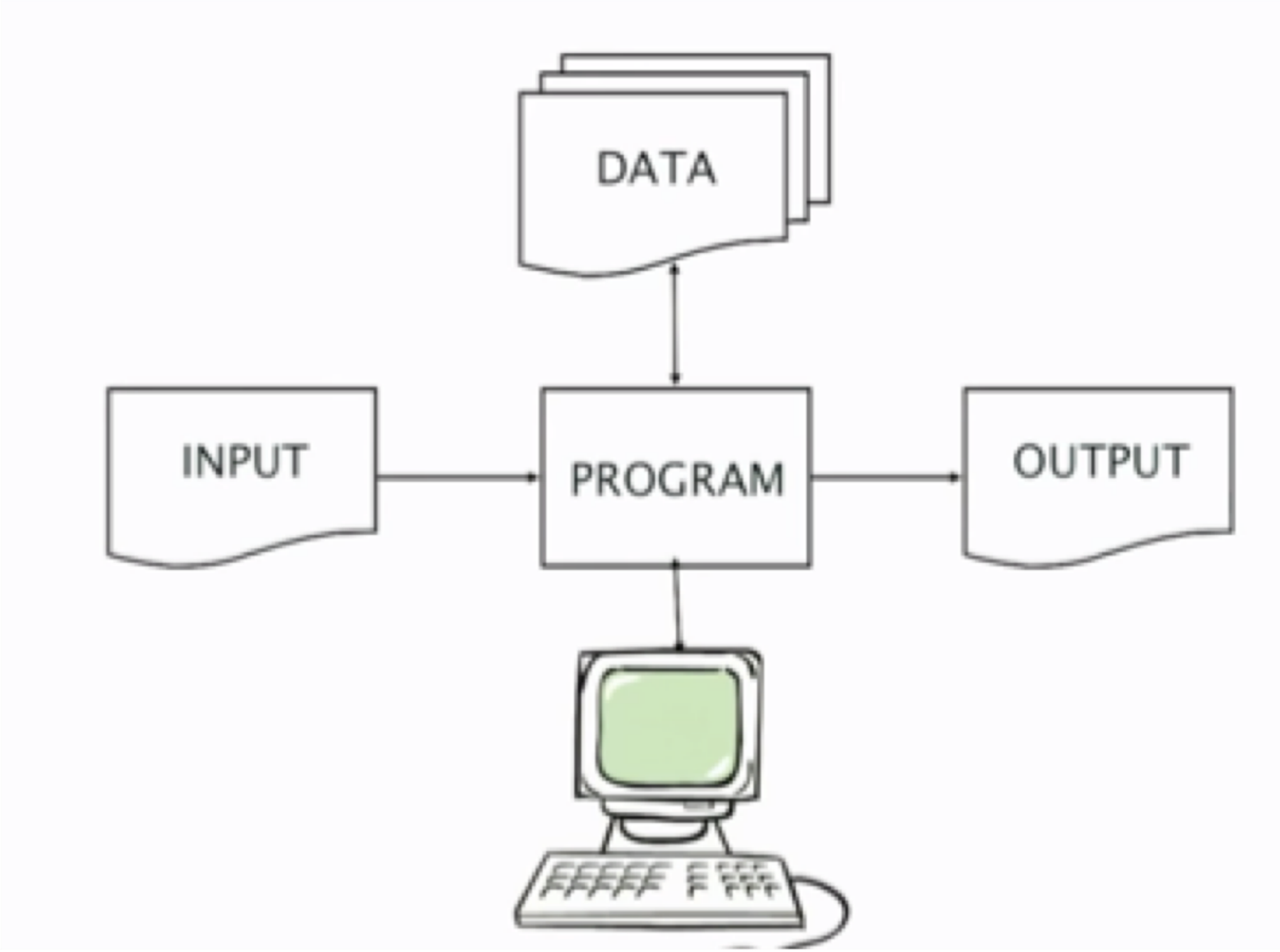
SURFsara

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.

Large Hadron Collider



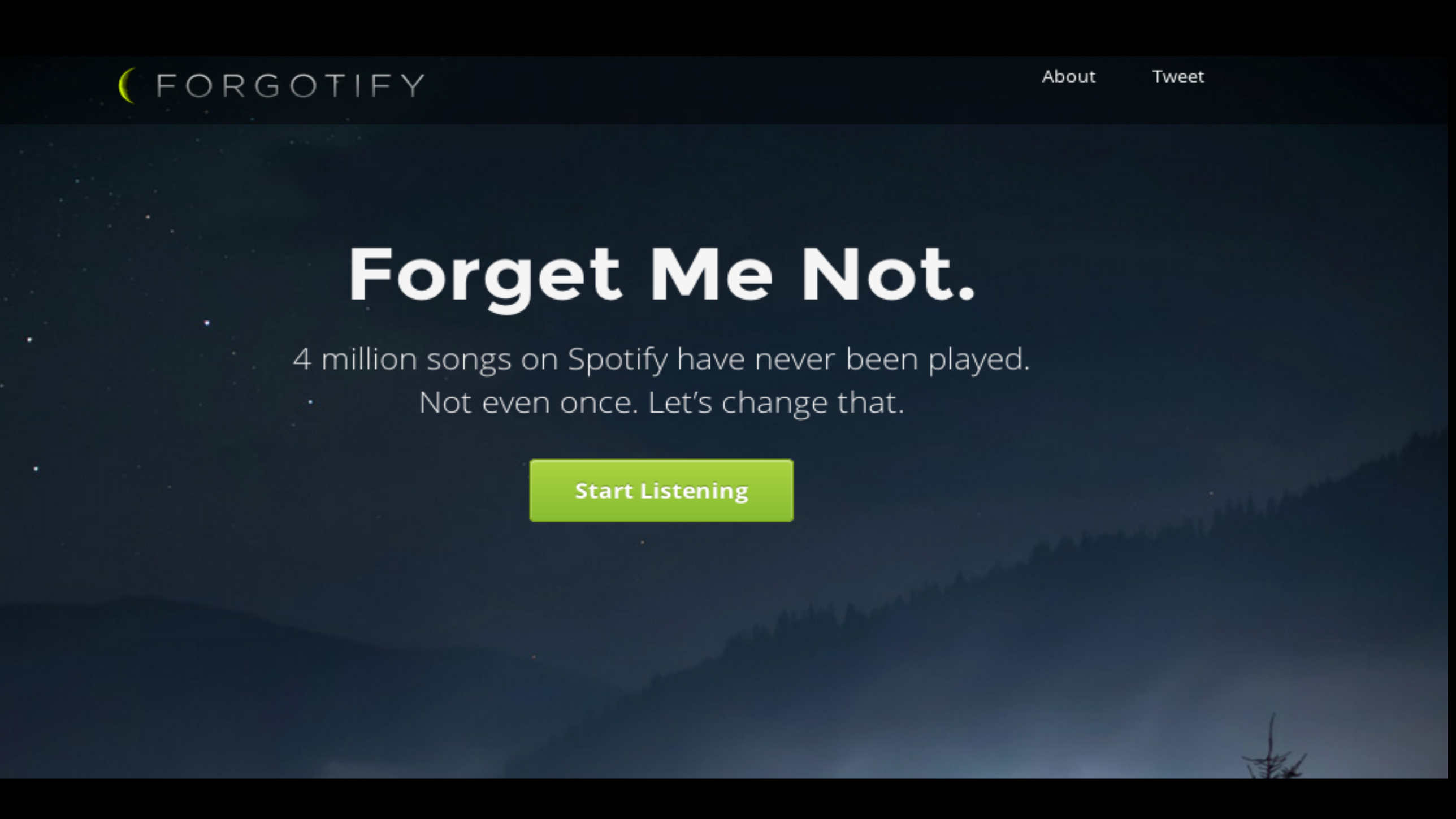
SURF SARA

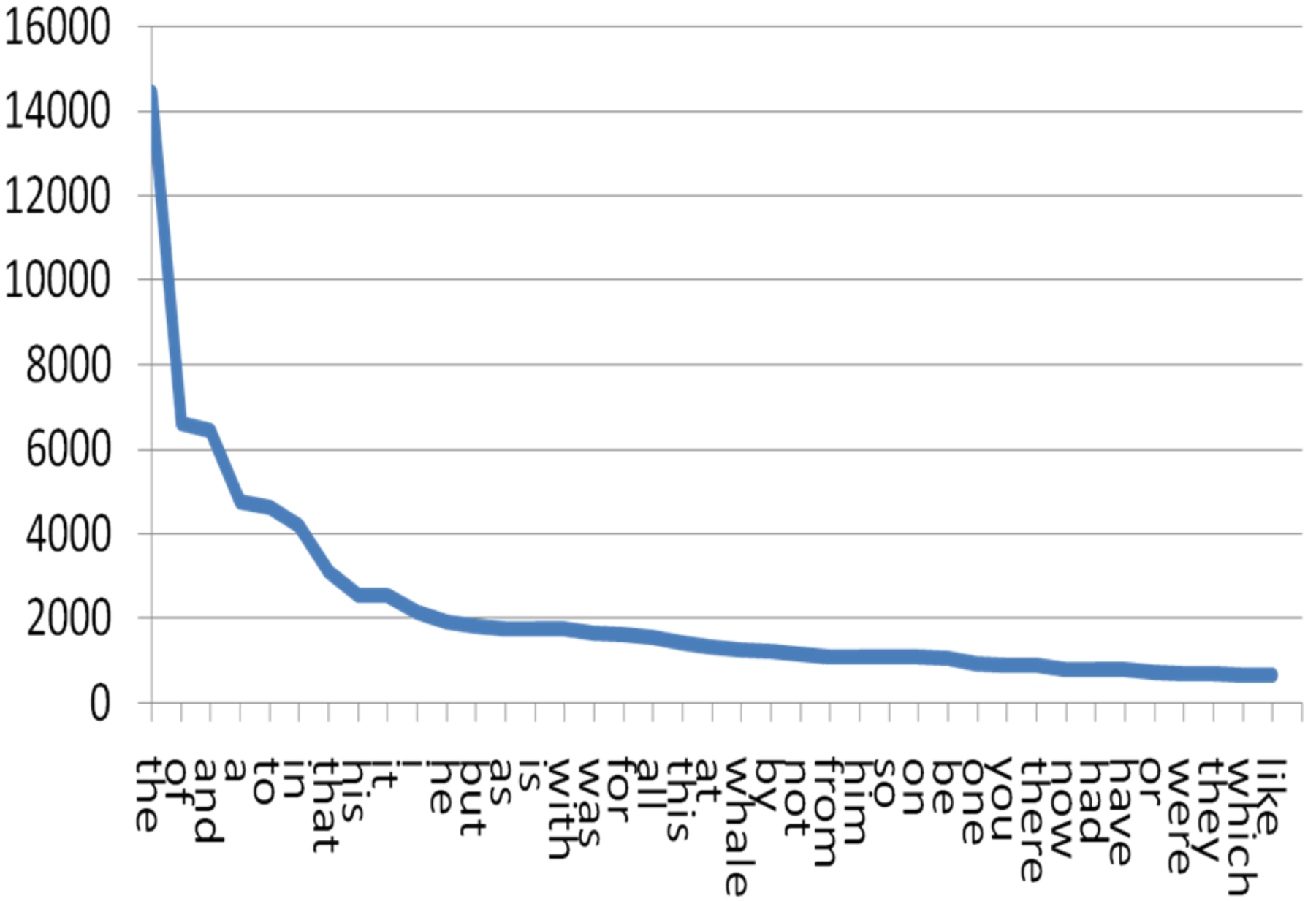

DATA
INPUT
PROGRAM
OUTPUT

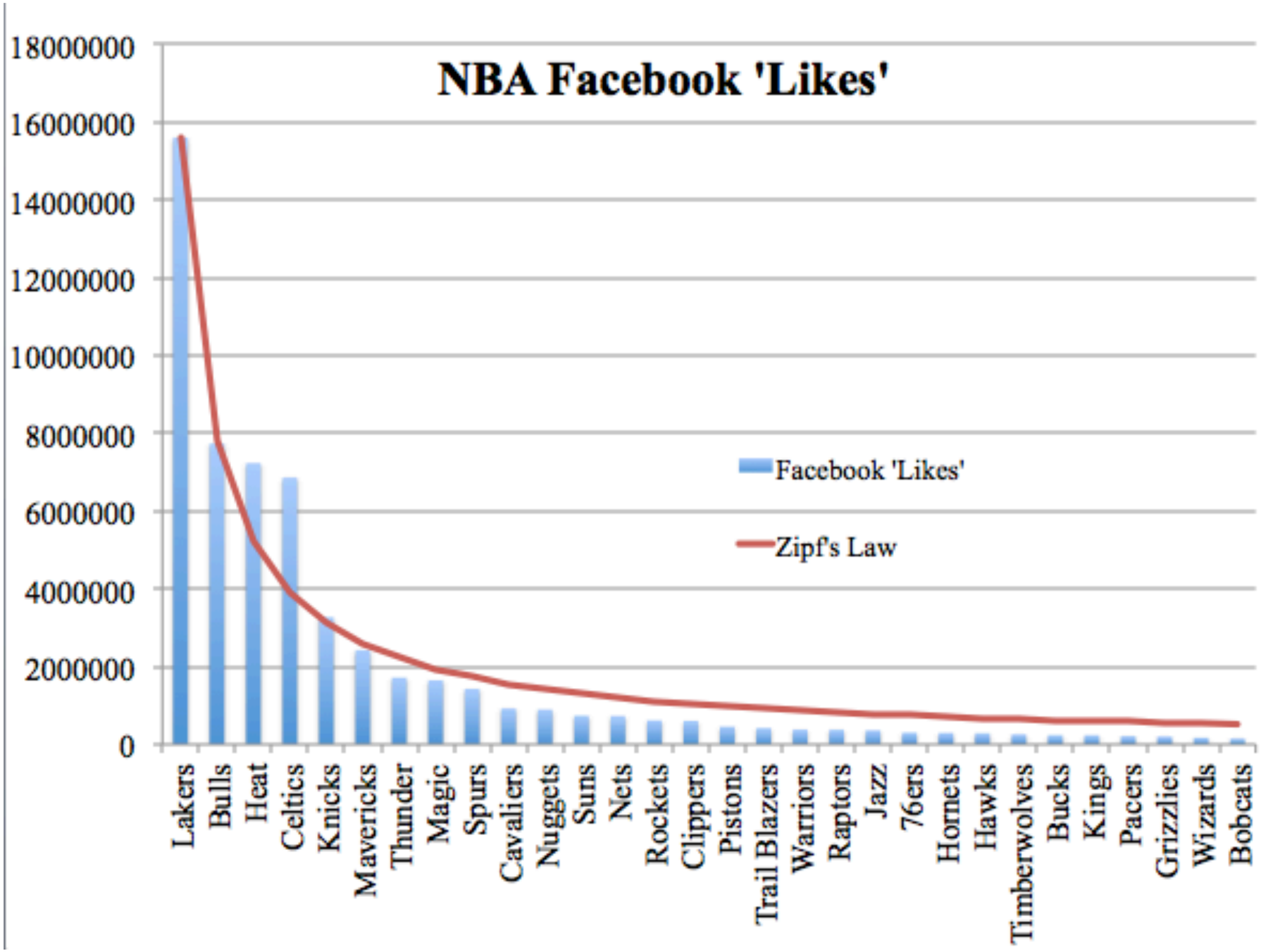
ANATOMY OF THE LONG TAIL

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

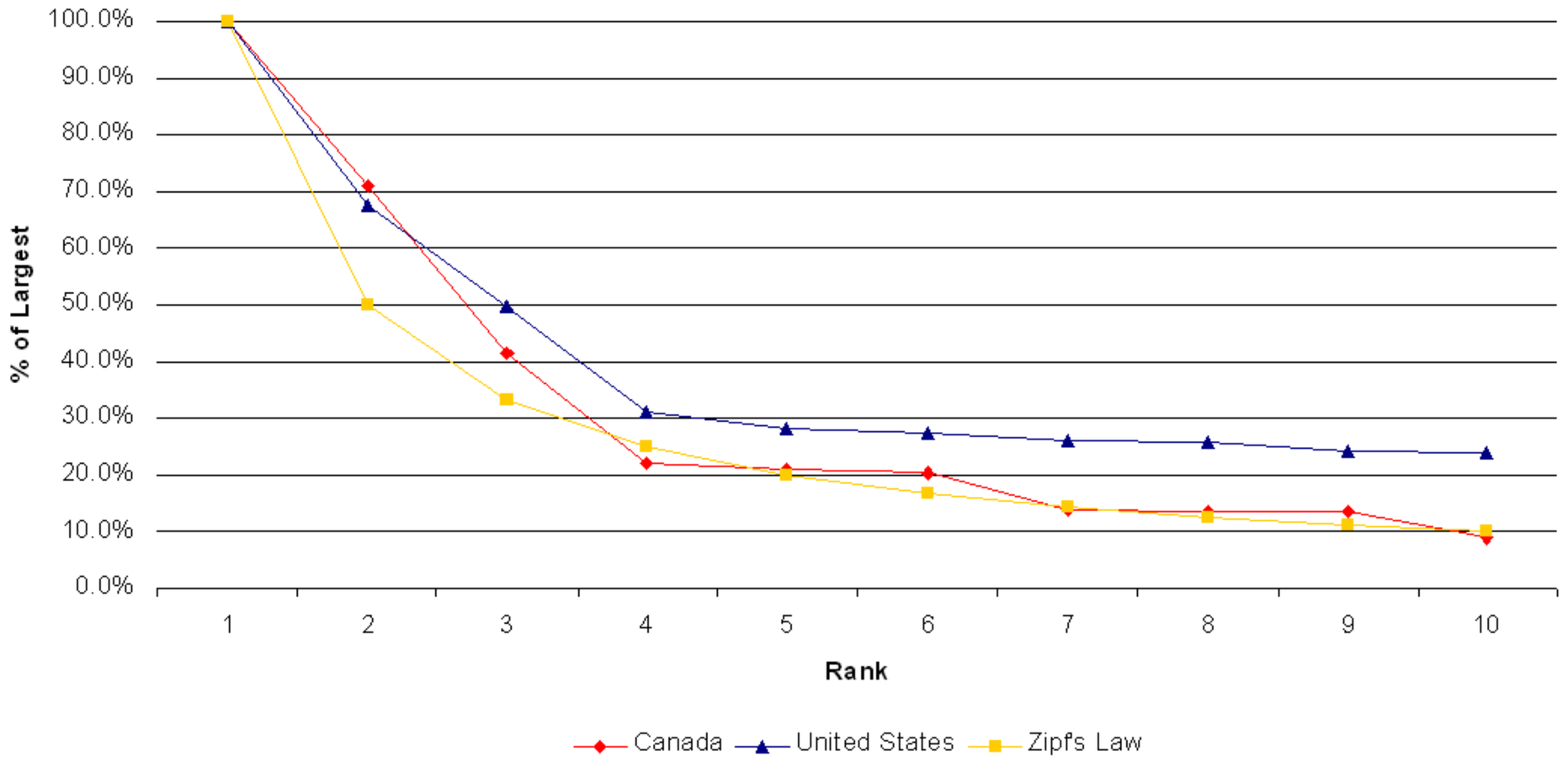

FORGOTIFY
About
Tweet
Forget Me Not.
4 million songs on Spotify have never been played.
Not even once. Let's change that.
Start Listening


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


NBA Facebook 'Likes'
18000000
16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
Facebook 'Likes'
Zipf's Law
Lakers
Bulls
Heat
Celtics
Knicks
Mavericks
Thunder
Magic
Spurs
Cavaliers
Nuggets
Suns
Nets
Rockets
Clippers
Pistons
Trail Blazers
Warriors
Raptors
Jazz
76ers
Hornets
Hawks
Timberwolves
Bucks
Kings
Pacers
Grizzlies
Wizards
Bobcats

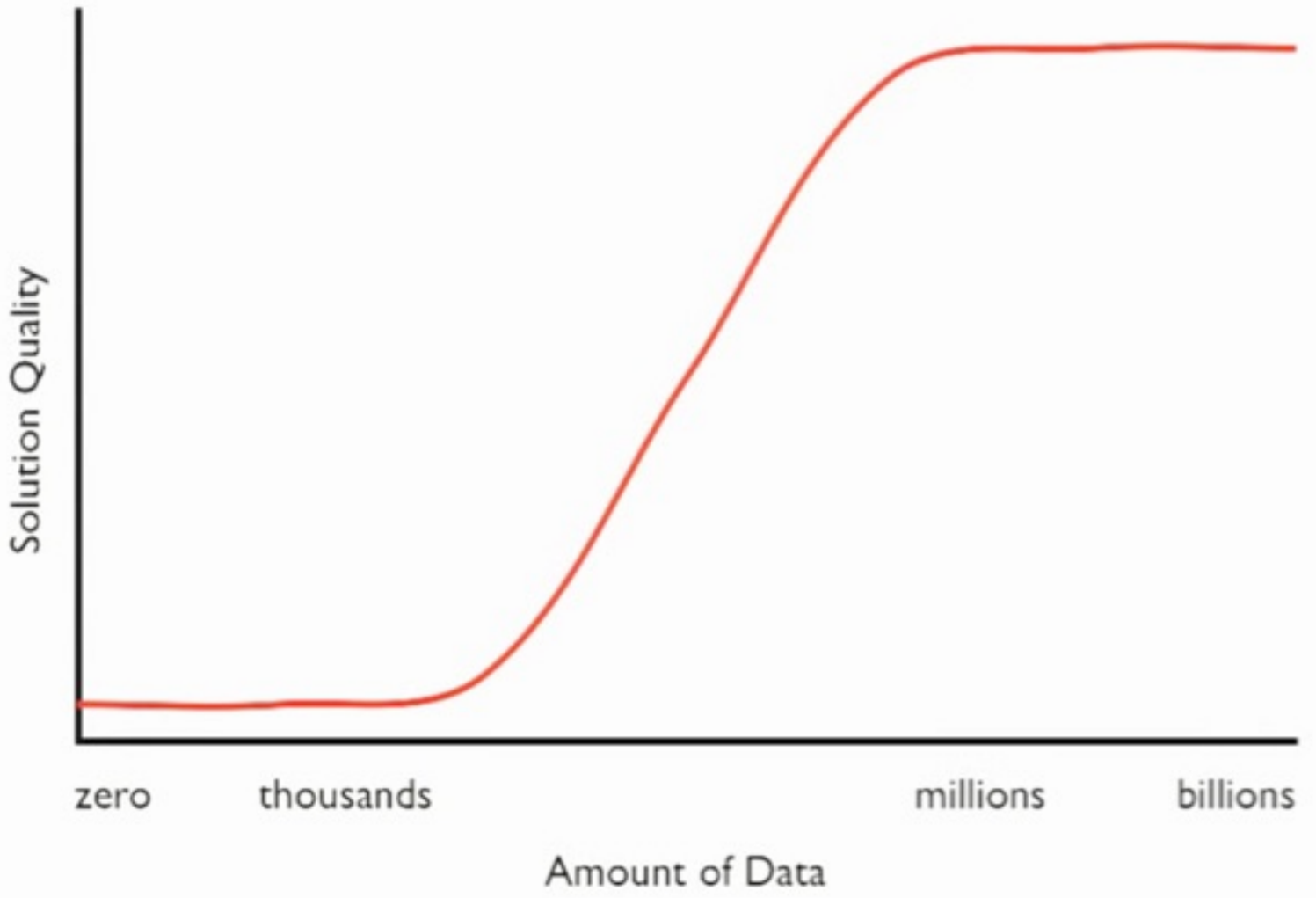
SURF SARA

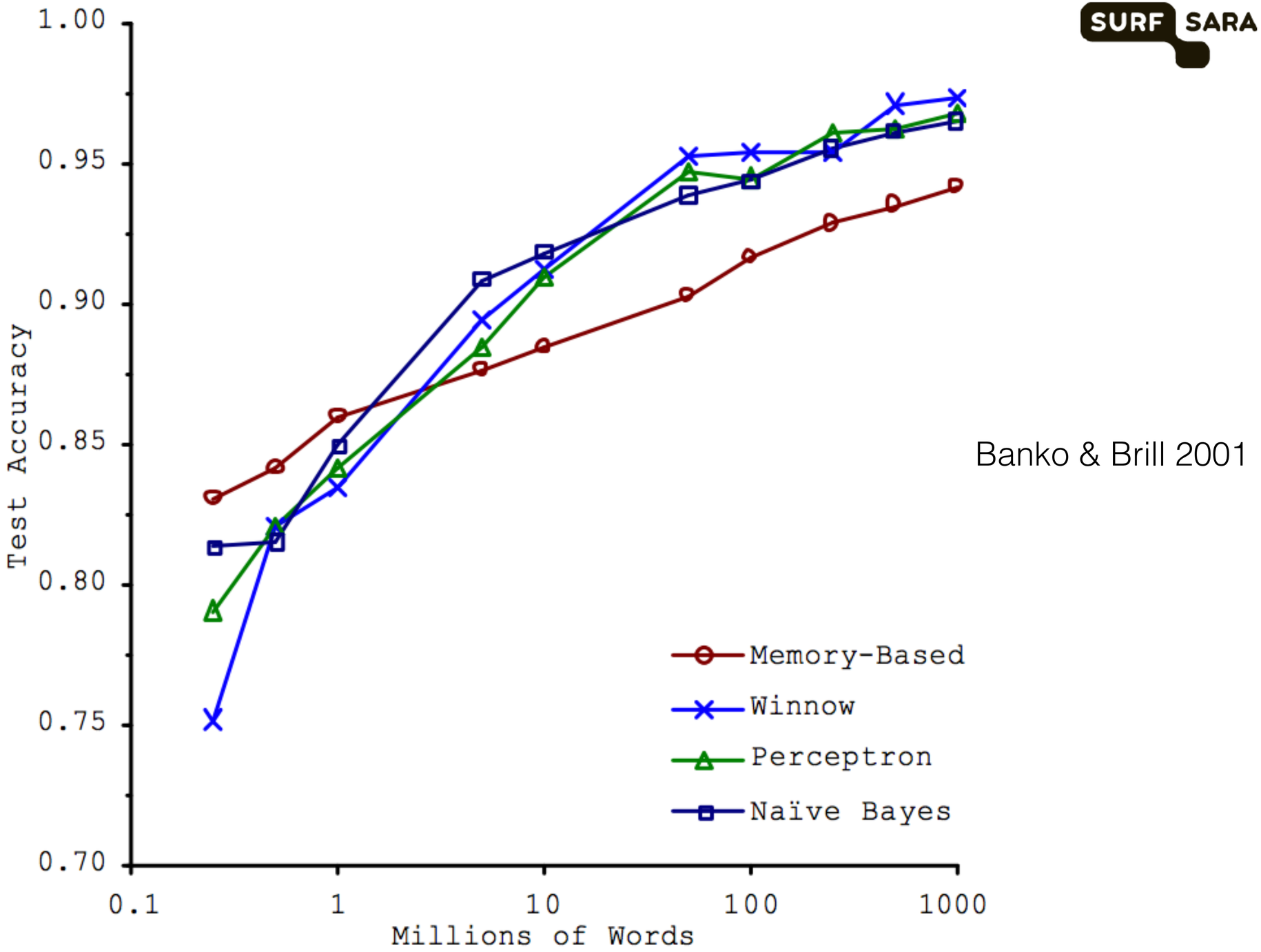
Zipfs law and city populations



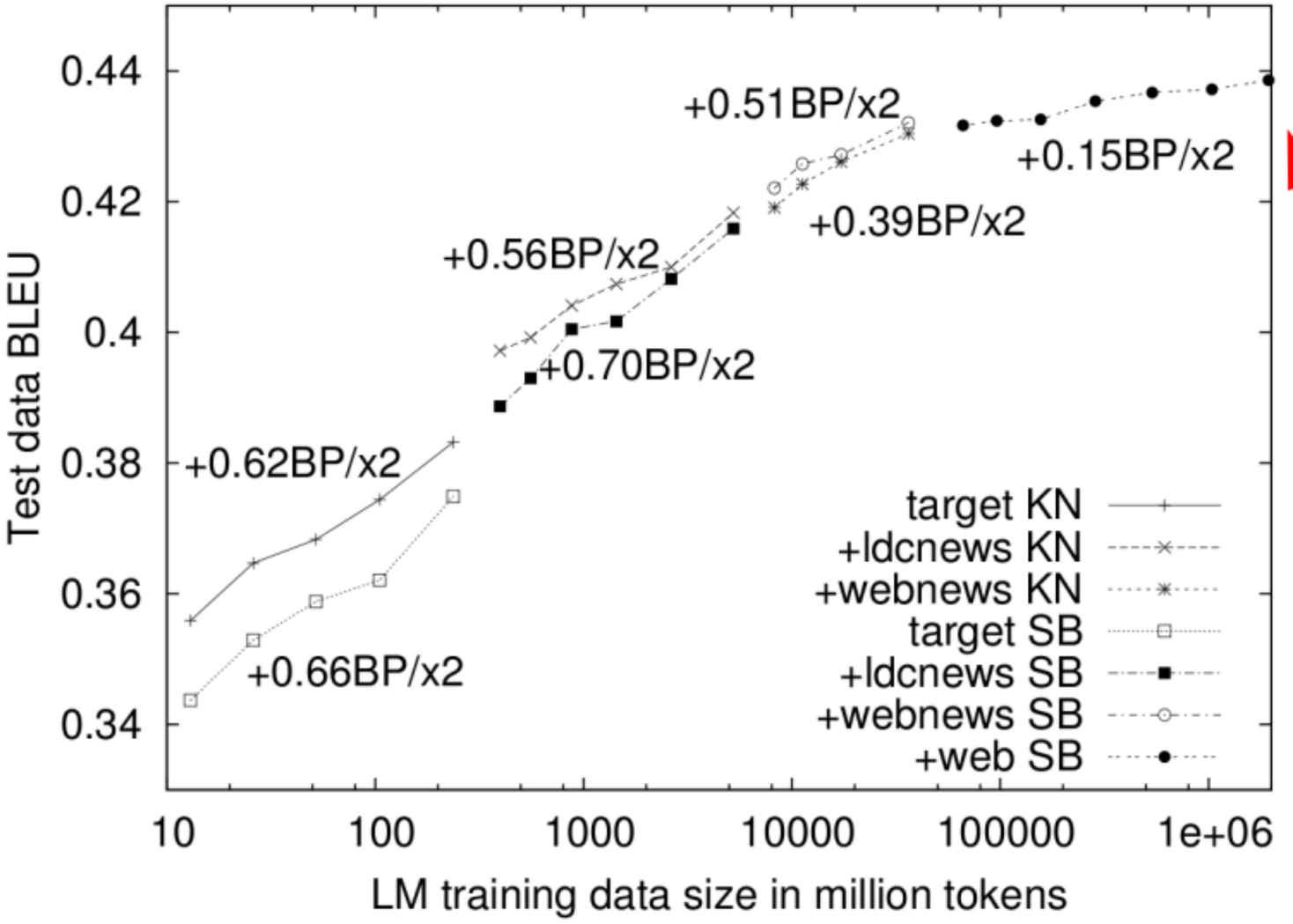

Original
Hays and Efros - 2007

Data Threshold





Banko & Brill 2001



Thorsten Brants, Ashok Popat, Peng Xu, Franz Och, Jeffrey Dean. Large Language Models in Machine Translation. In: Proceedings of EMNLP, 2007

How to get here if you are (not) Google?

Open source available

widely adopted by internet and big data companies

SURF SARA

Big Data Technology as strategy

UBER

ebay

airbnb

SURF SARA

What does this technology offer?

- Frameworks for developing scalable applications for data analytics, machine learning, streaming and more.
- noSQL (non relational) Databases for storing

SURF SARA

What does this technology offer?

Advantages:

Scalability to thousands of machines without changing the code

Easy development, reduced complexity

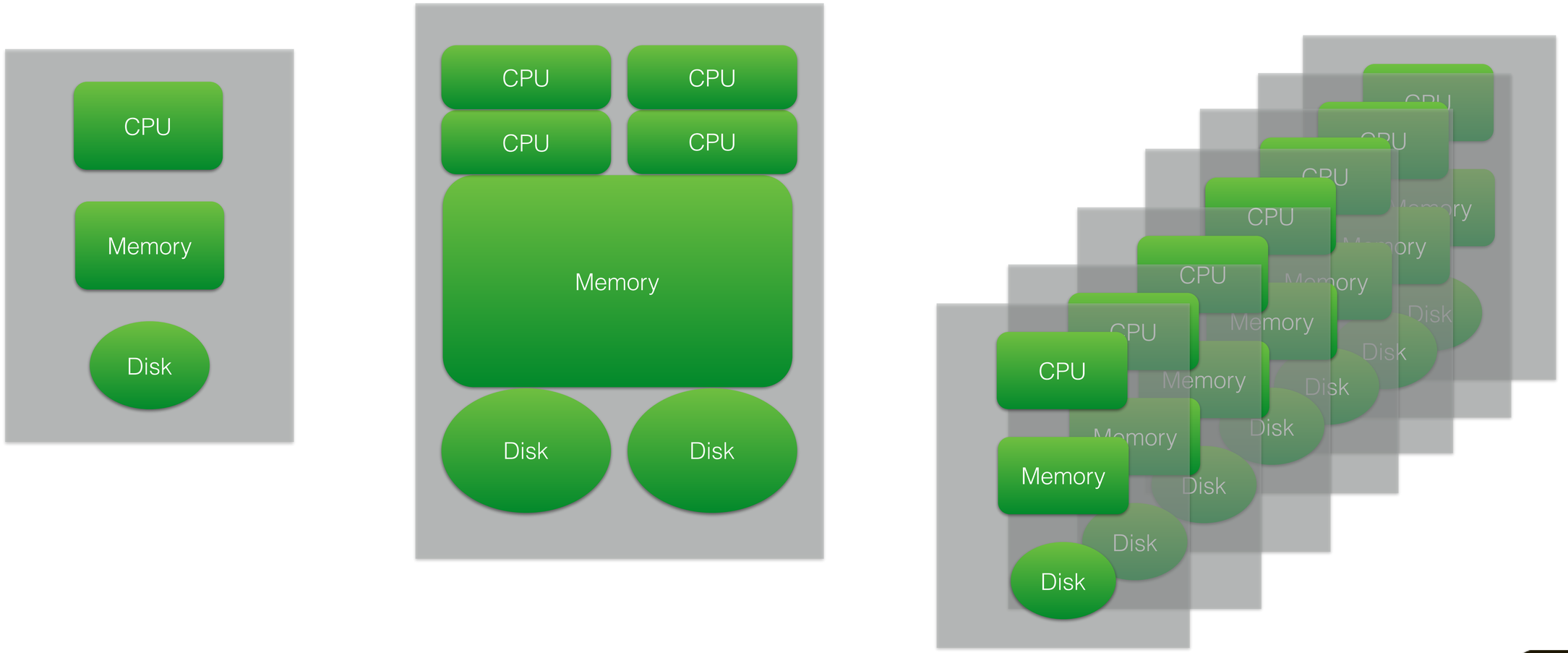
Support for unstructured and semistructured data formats

Disadvantages:

You have to write code in these frameworks

Not all tasks (e.g. simulations, transactions) can be done in these frameworks/databases

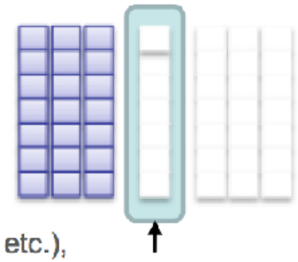
Scalability



Traditionally Parallel programming is hard

Fundamental issues

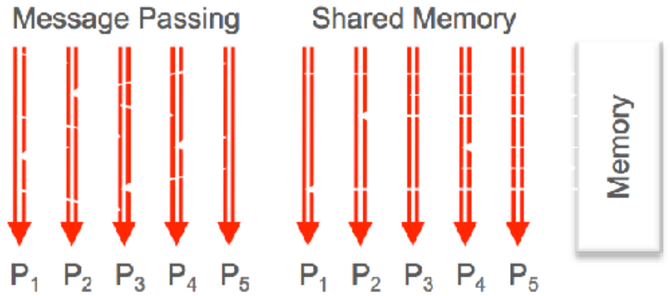
scheduling, data distribution, synchronization, inter-process communication, robustness, fault tolerance, …



Architectural issues

Flynn's taxonomy (SIMD, MIMD, etc.), network typology, bisection bandwidth
UMA vs. NUMA, cache coherence

Different programming models

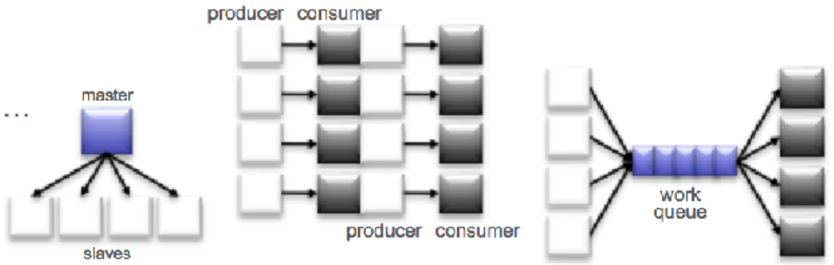


Common problems

livelock, deadlock, data starvation, priority inversion…
dining philosophers, sleeping barbers, cigarette smokers, …

Different programming constructs

mutexes, conditional variables, barriers, …
masters/slaves, producers/consumers, workå queues, …



The reality: programmer shoulders the burden of managing concurrency…

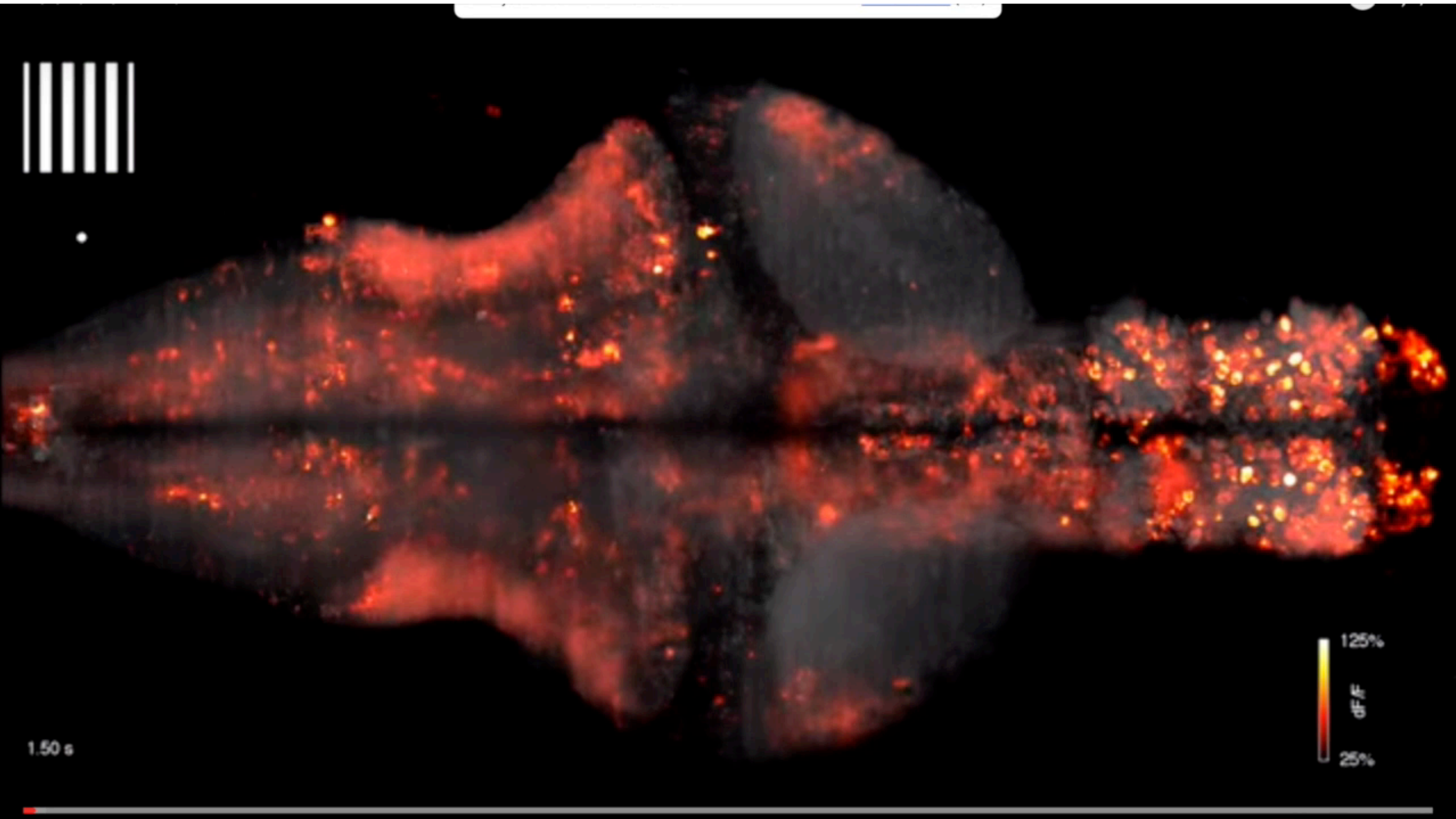
Slide: Jimmy Lin

SURF SARA

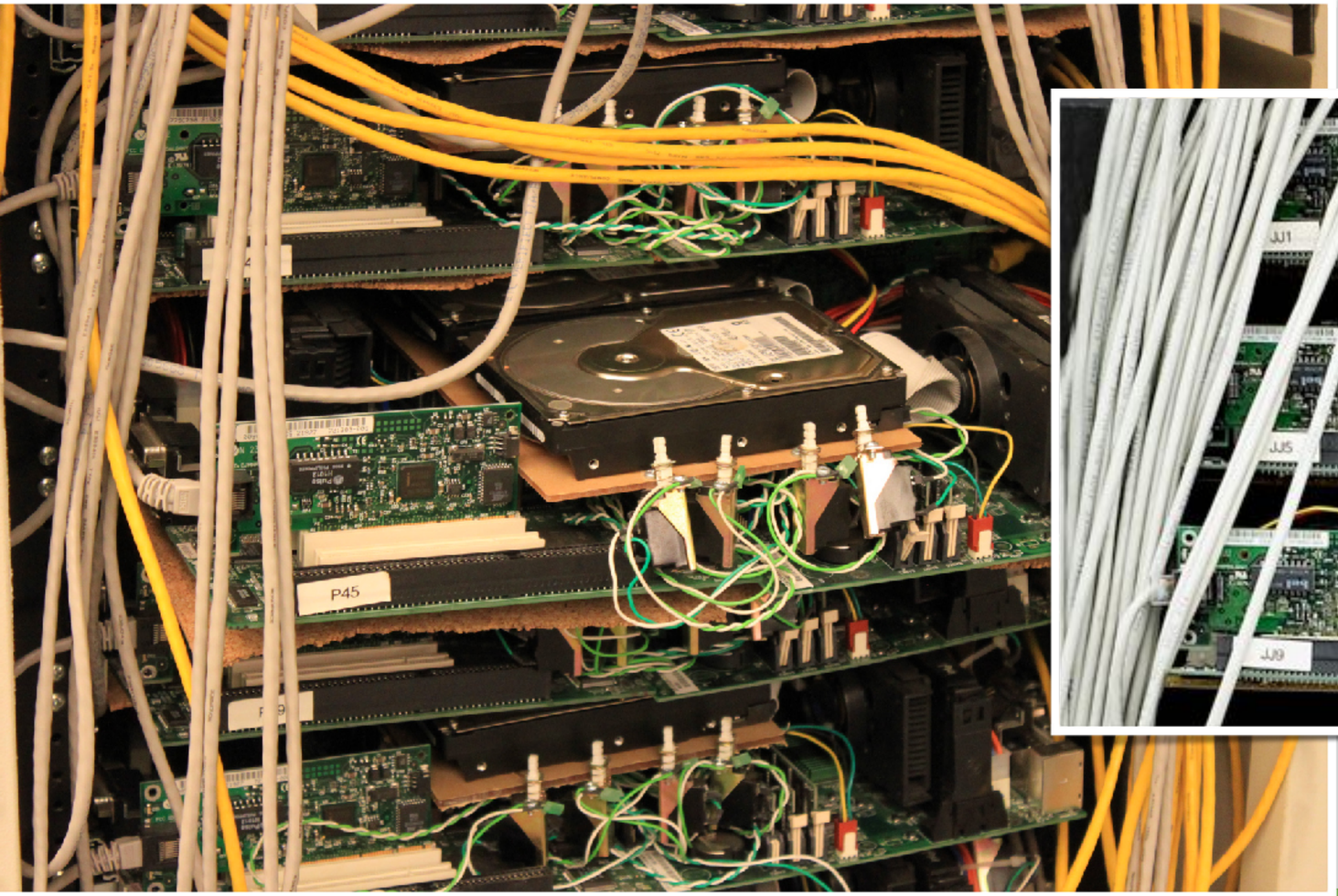
Reducing complexity

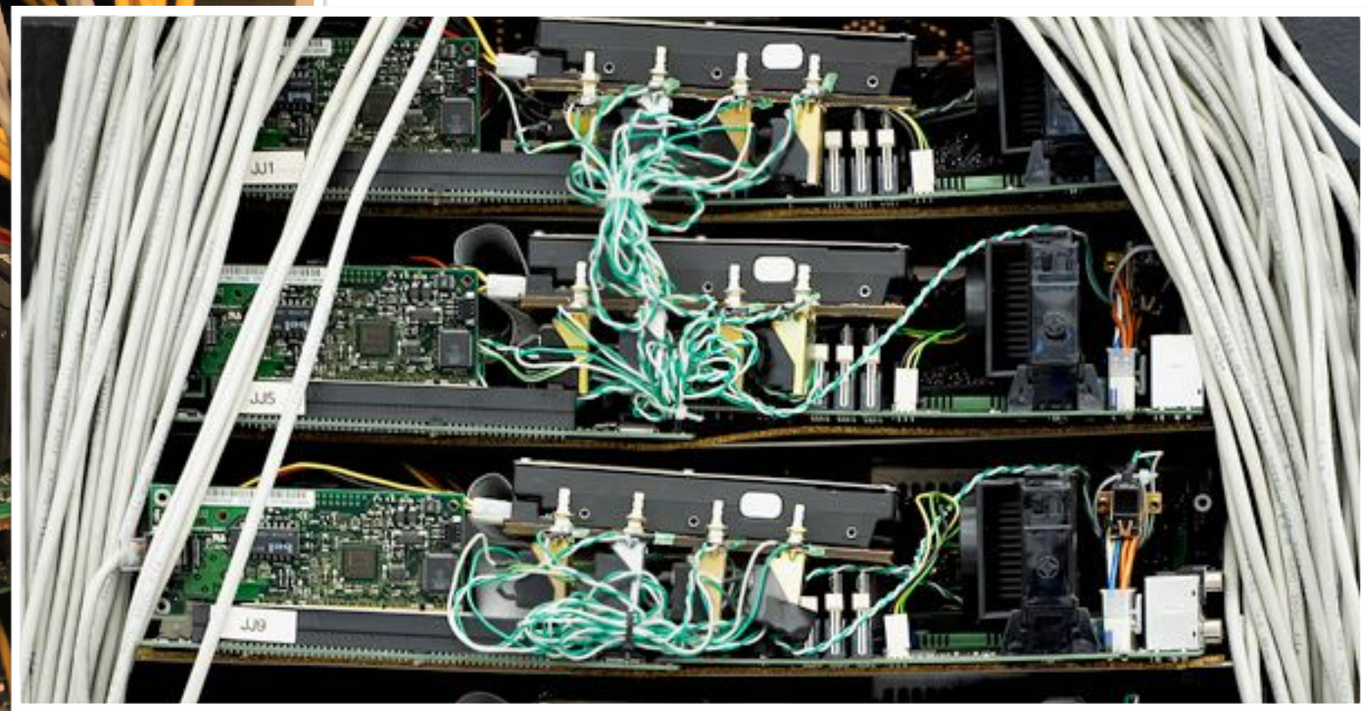
- Simple API's to build applications that scale
- Shared mutable state is avoided
- (Human) fault tolerance is important
- Emphasis on SQL
- Web based Notebooks for development

SURF SARA


125%
Ux
25%
1.50 s

Googles approach





MapReduce: Simplified Data Processing on Large Clusters

Dean and Sanjay Ghemawat

google.com

day, etc. Most such computations are conceptu-
straightforward. However, the input data is usually
e and the computations have to be distributed across
ndreds or thousands of machines in order to finish in
reasonable amount of time. The issues of how to par-
llelize the computation, distribute the data, and handle
ailures conspire to obscure the original simple compu-
tation with large amounts of complex code to deal with
these issues.

As a reaction to this complexity, we designed a new
abstraction that allows us to express the simple computa-
tions we were trying to perform but hides the messy de-
tails of parallelization, fault-tolerance, data distribution

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*

ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points.

The file system has successfully met our storage needs. It is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as research and development efforts that require

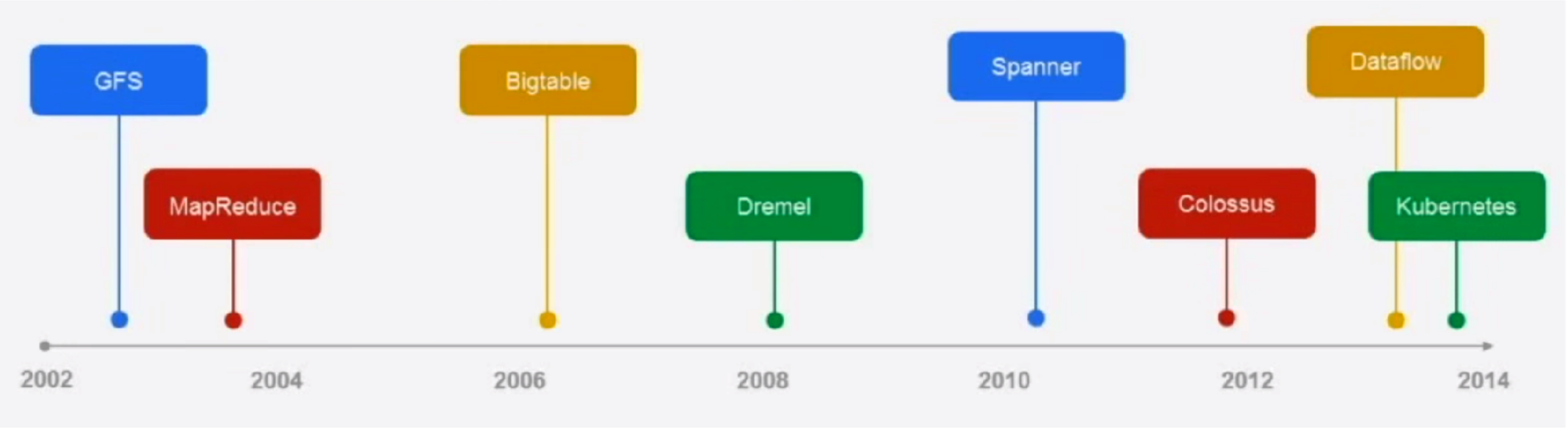
1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

First, component failures are the norm rather than the exception. The file system consists of hundreds or even thousands of storage machines built from inexpensive commodity parts and is accessed by a comparable number of client machines. The quantity and quality of the compo-

modify

Google Innovations in Software



Rethinking old ideas

The Pragmatic Programmers

Seven Concurrency Models in Seven Weeks

When Threads Unravel

Paul Butcher

Series editor: *Bruce A. Tate*
Development editor: *Jacquelyn Carter*

The Pragmatic Programmers

Seven Databases in Seven Weeks

A Guide to Modern Databases and the NoSQL Movement

Eric Redmond and Jim R. Wilson

Series editor: *Bruce A. Tate*
Development editor: *Jacquelyn Carter*

SURF SARA

Big Data technology in science

Adoption is slow.

Some researchers take the utility view on ICT (it should work!)

Others stick to traditional/proven technology or software

Scalability is underestimated or postponed

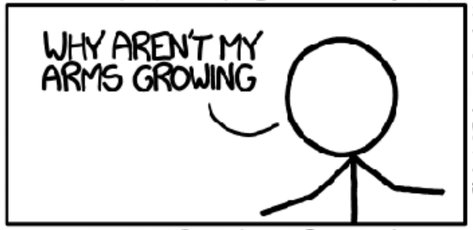
Some are critical : we have been doing this for years

SURF SARA

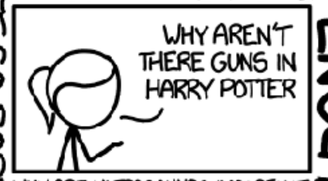
QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS

WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT


WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER,
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN


WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47S SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE


WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG


WHY DO IGUANAS DIE
WHY ARE TREES TALL
WHY ARE THERE DINOSAUR GHOSTS
WHY AREN'T THERE
WHY IS THERE LAVA
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LIFE SO BORING
WHY IS YKK ON ALL ZIPPERS
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Preferential Attachment

VARIETY

VELOCITY