

**Big Data: making
sense of information
and
analytics in oil and gas**

**How did we
get here?**

**What can we
do now?**

**What's in our
future?**

18,000BC

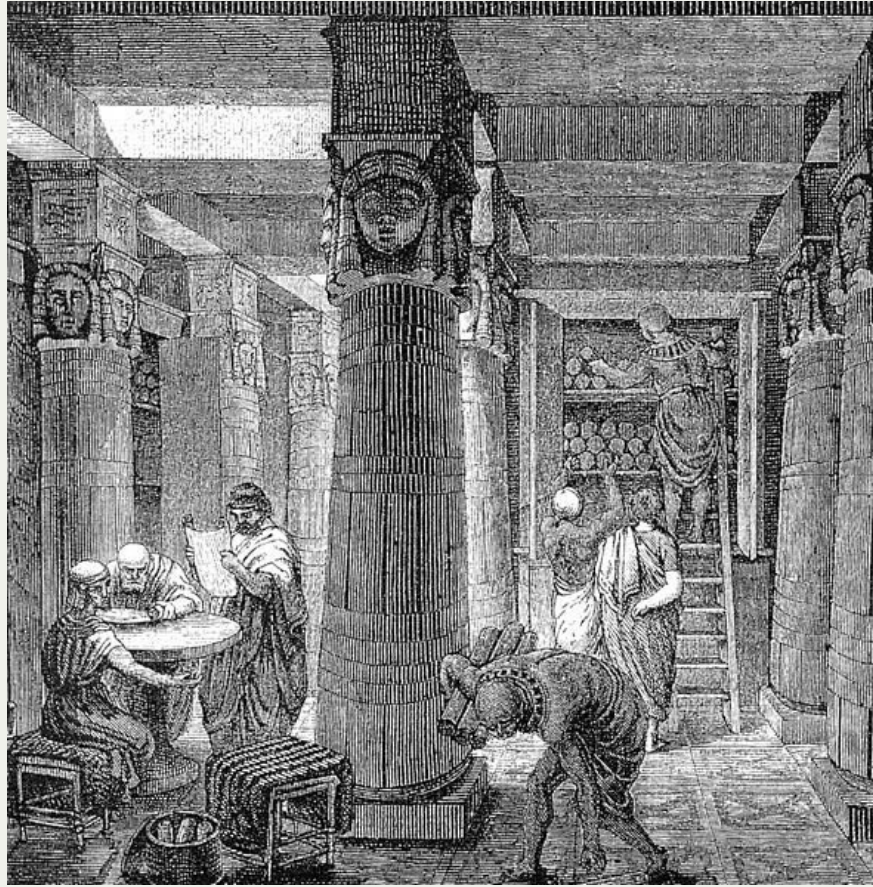
Ishango bone





300 BC – 48 AD

Library of Alexandria





1663

John Graunt

*Natural and Political
Observations Made upon the
Bills of Mortality*







1880





1890 US Census



1890 US Census

Conducted every 10 years



1890 US Census

Conducted every 10 years

Data predicted to take 10 years
to compile

Herman Hollerith

20 year old engineer





Tabulating machine



Tabulating machine

Used "punched cards"



Tabulating machine

Used "punched cards"

Reduced 10 years work down to just 3 months



Tabulating Machine Company

Tabulating Machine Company

Later merged with a number of other companies.

Tabulating Machine Company

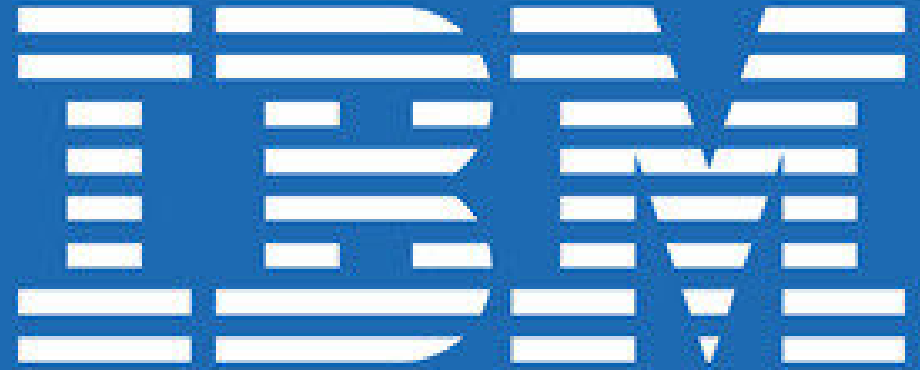
Later merged with a number of other companies.

To form International Business Machines

Tabulating Machine Company

Later merged with a number of
other companies.

To form International Business
Machines



1929

FIG. 1.

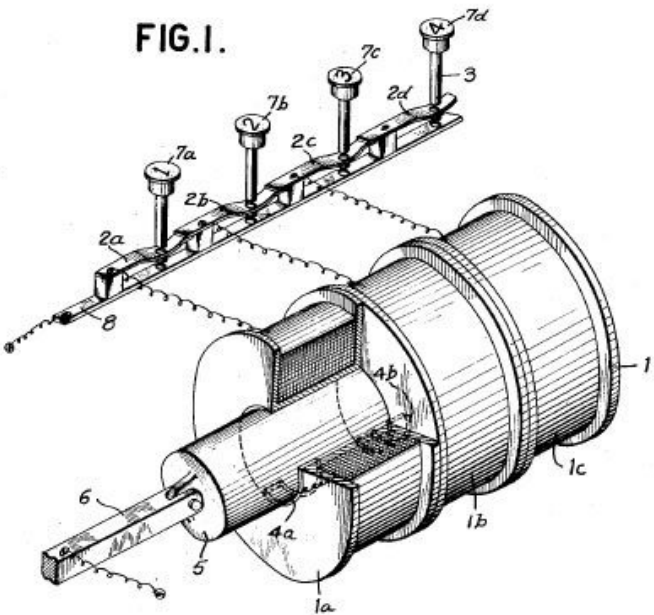
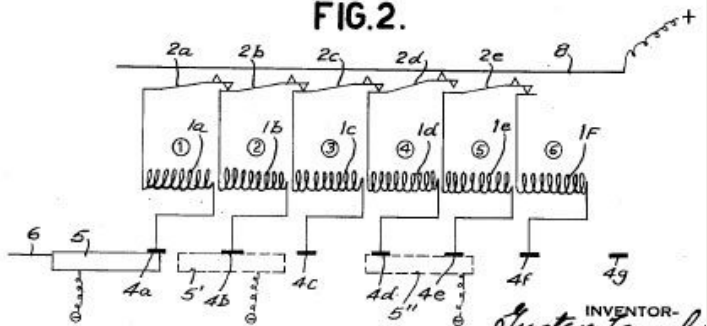


FIG. 2.



Magnetic drum memory

Invented by Gustav Tauschek in Austria

Used a ferromagnetic recording material

Could store 62.5 KB



Working for Rheinmetall



Working for Rheinmetall

Never used his invention



Working for Rheinmetall

Never used his invention

Worked on punch card
accounting systems



Working for Rheinmetall

Never used his invention

Worked on punch card
accounting systems

His subsidiary was acquired
and he sold 169 patents

IBM

1950



UNIVAC 1103

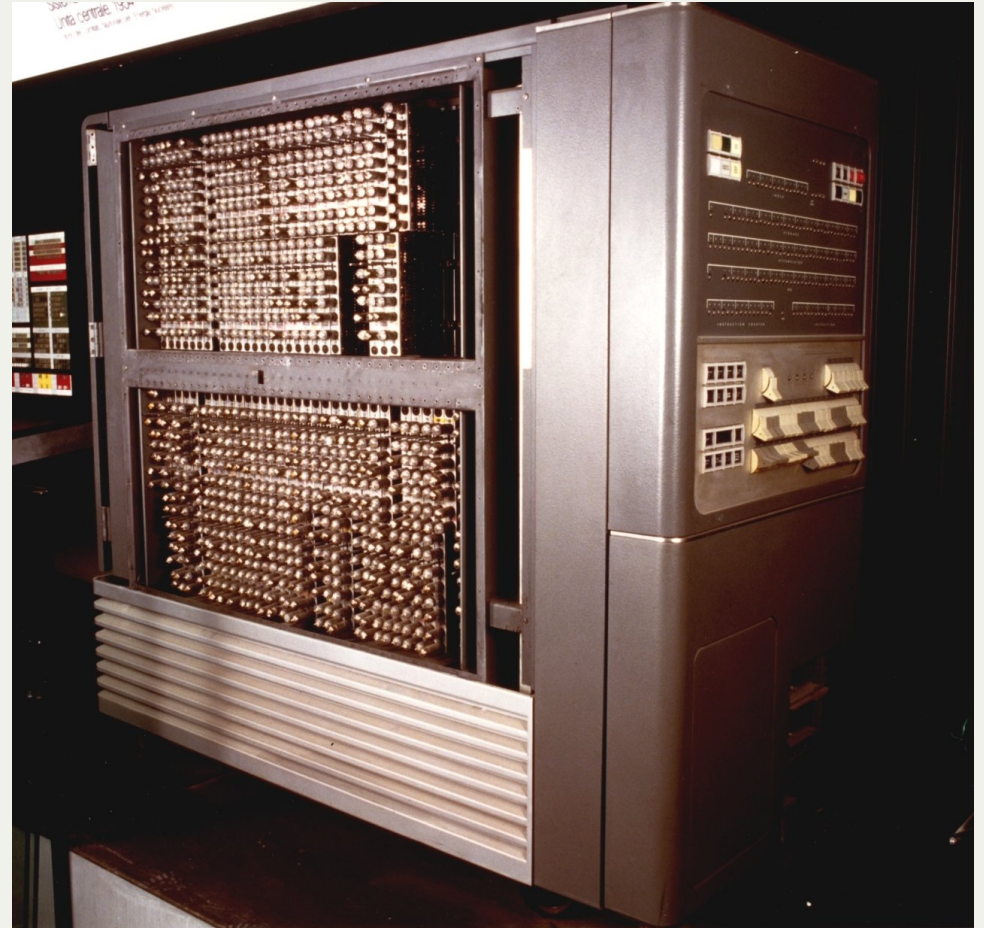
Weighed 17.5 tons

1954

IBM

IBM 704

The first mass-produced computer with floating-point arithmetic hardware.



1962

IBM



The IBM 1311 Disk Storage Drive

Each disk weighed 4.5kg

Could store 1.5MB

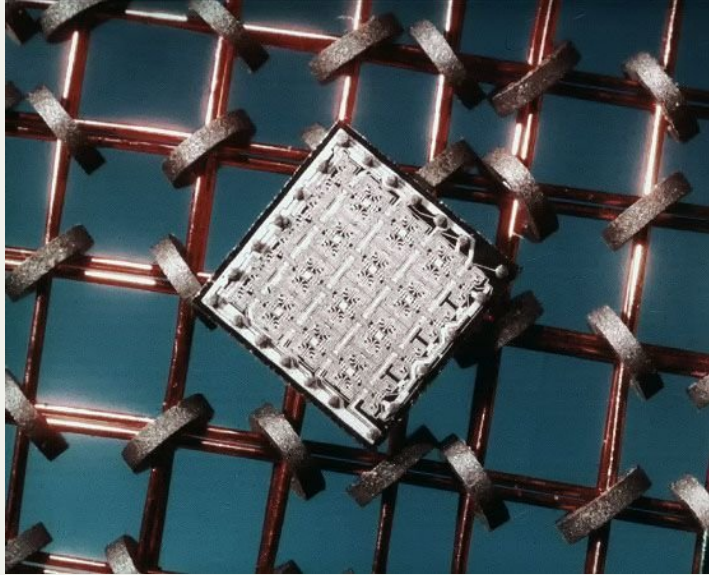
Shoebox Machine



Dr. E. A. Quade

IBM

1970



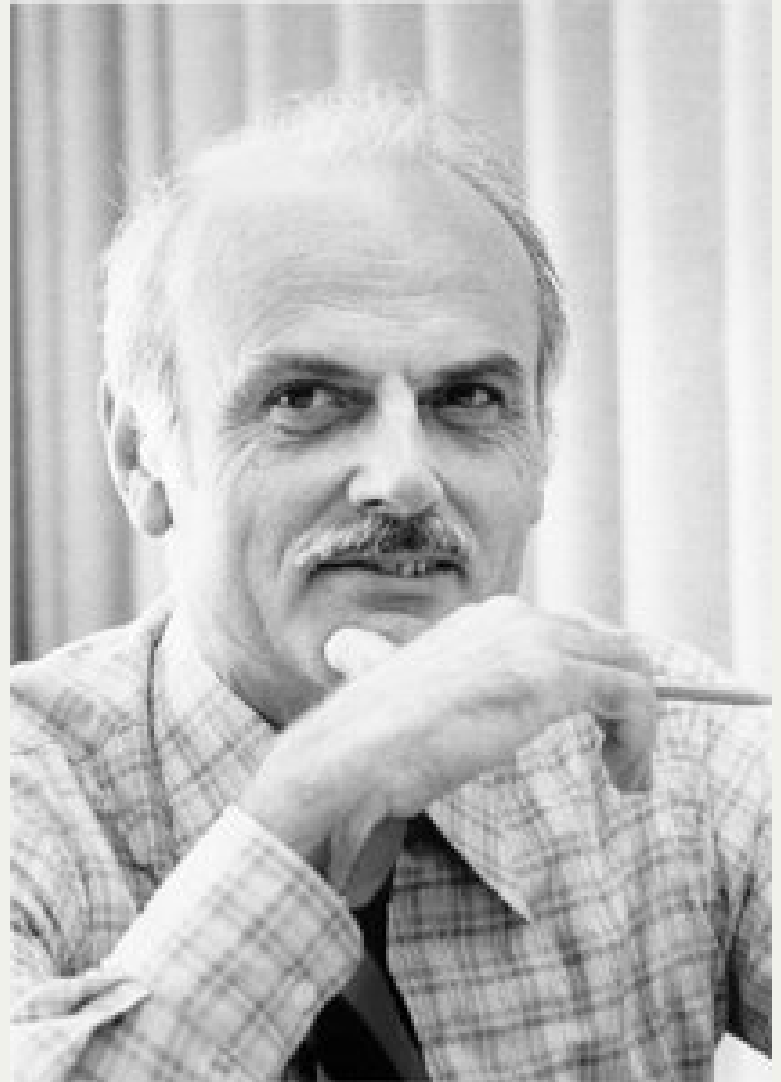
Semiconductor memory

Half the space for the same data

Edgar Frank "Ted" Codd

An Englishman working in the
US.

*"A Relational Model of Data for
Large Shared Data Banks"*



IBM

Relational Databases

Would become incredibly popular but at the time IBM refused to use Ted's ideas.



1979

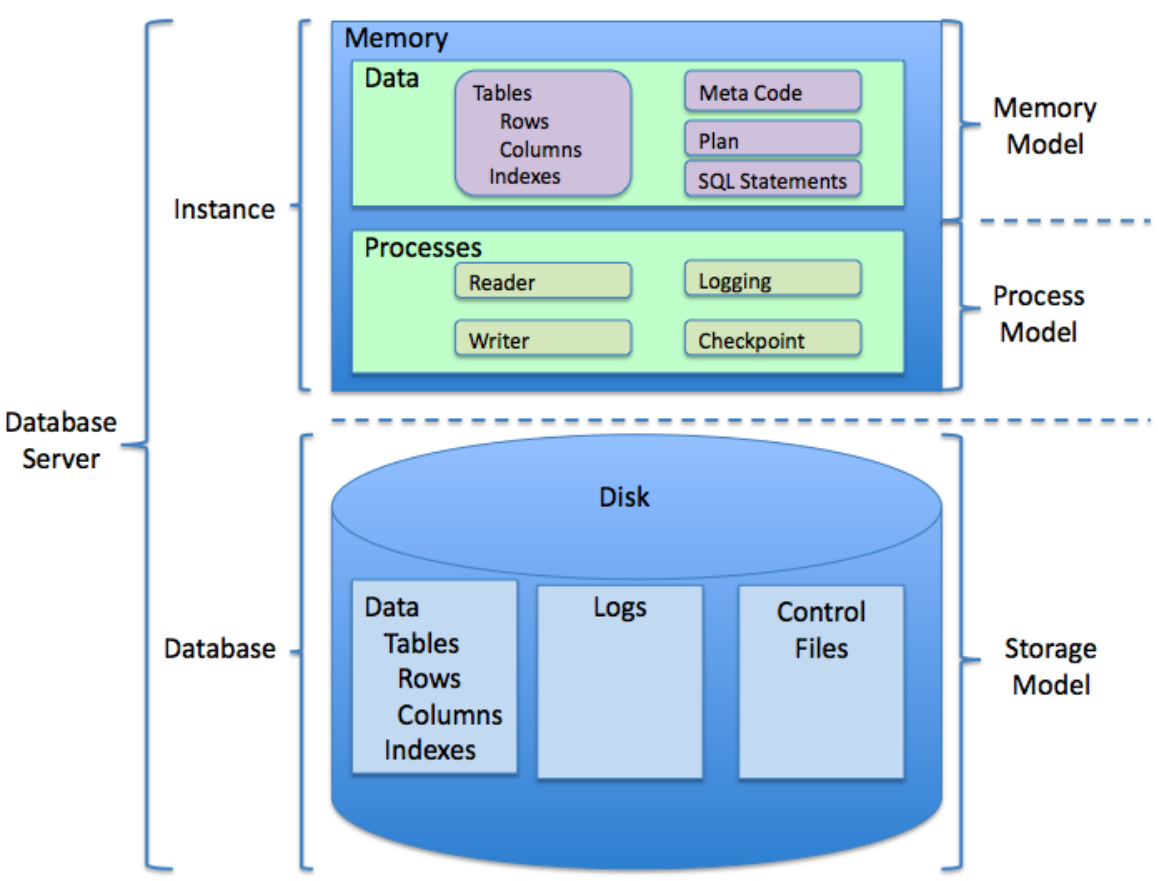
Larry Ellison



Oracle v2

**First commercially available
SQL-based RDBMS**

Relational Database Management System



1980

Commodore VIC-20



1 million sales

1MHz CPU

5KB RAM (expandable to 64KB)

20KB ROM



1981

IBM

MICROSOFT

CONSUMER PRODUCTS

A Division of Microsoft, Inc.

10700 Northup Way • Bellevue, WA 98004

CIRCLE 203 ON READER SERVICE CARD

Starting MS-DOS...

C:\> _

1984



INTERNATIONAL TELECOMMUNICATION UNION

CCITT

THE INTERNATIONAL
TELEGRAPH AND TELEPHONE
CONSULTATIVE COMMITTEE

BLUE BOOK

VOLUME VIII – FASCICLE VIII.7

**DATA COMMUNICATION NETWORKS
MESSAGE HANDLING SYSTEMS**

RECOMMENDATIONS X.400-X.420



IXTH PLENARY ASSEMBLY
MELBOURNE, 14-25 NOVEMBER 1988

Geneva 1989

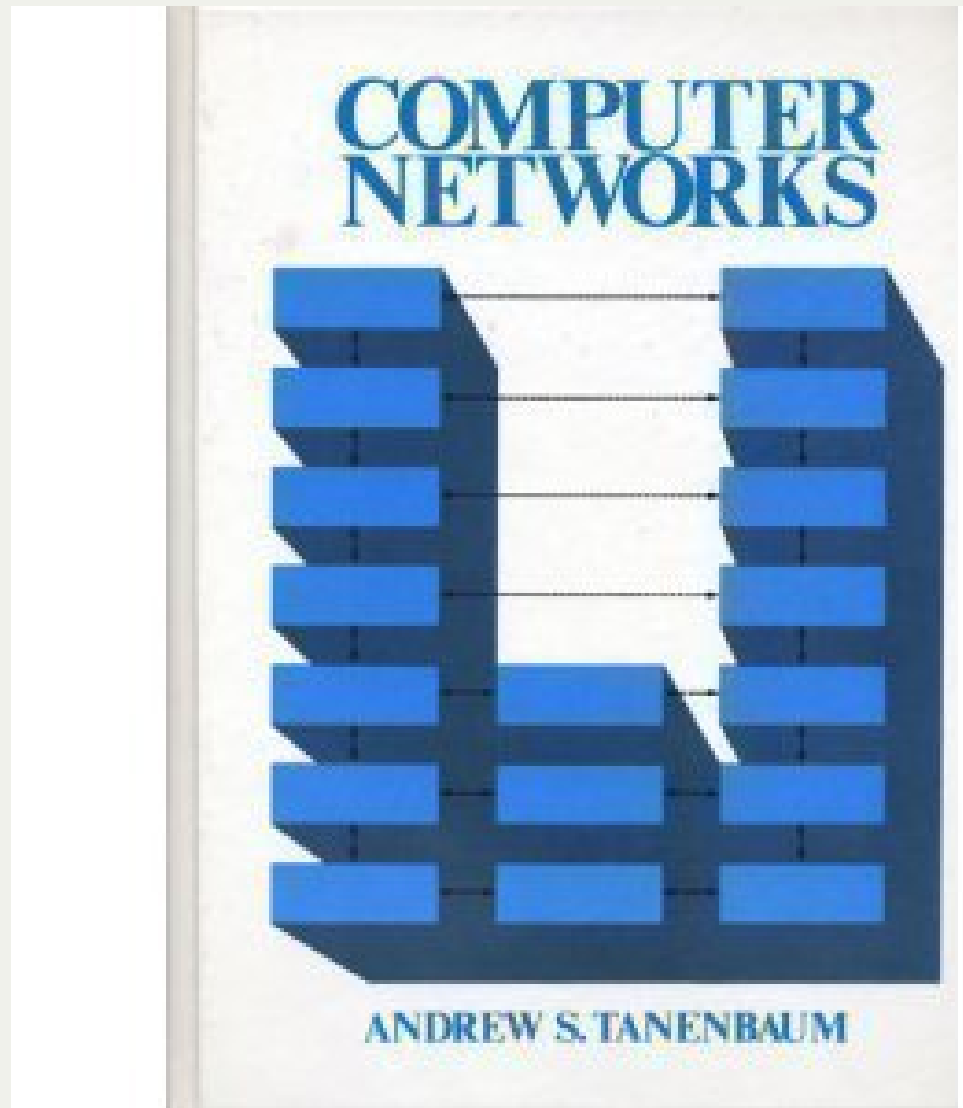


Open Systems Interconnection

The first standard for an "internet" is
published

Andrew S. Tanenbaum

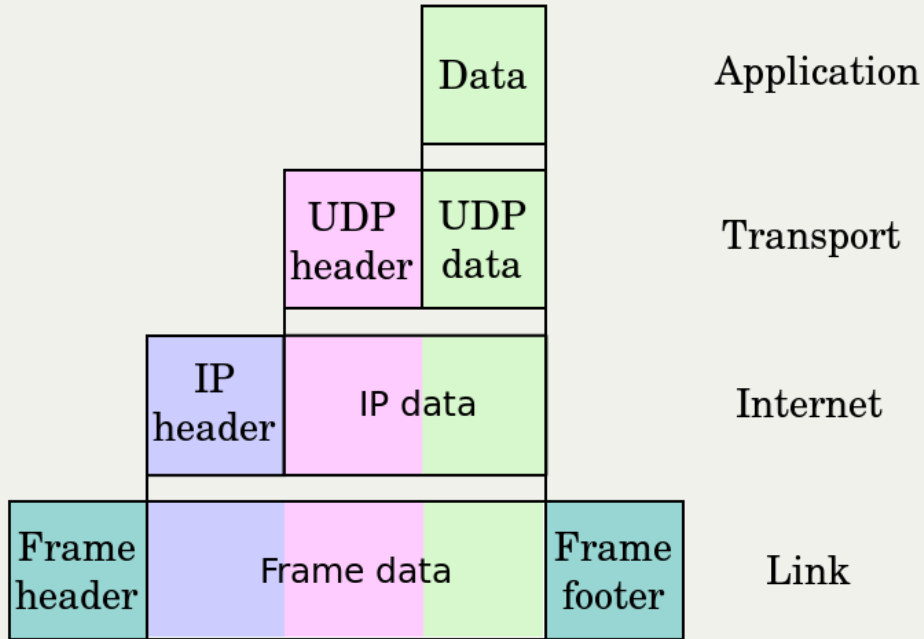
Criticised the OSI



Internet protocol suite

Known as TCP/IP

Developed at DARPA



1987



GSM Technical Specification

"standard for digital mobile telephony, including text messaging"



1990

Tim Berners-Lee

Designs a "web browser"

Called "WorldWideWeb"



1994

Amazon.com

Started as an online
bookstore



\$1 trillion

Market Summary > Amazon.com, Inc.

NASDAQ: AMZN

[+ Follow](#)

2,039.51 USD **+26.80 (1.33%)** ↑

Closed: 4 Sep, 19:49 GMT-4 · Disclaimer

After hours 2,037.00 **-2.51 (0.12%)**

1 day

5 days

1 month

6 months

YTD

1 year

5 years

Max



Amazon Web Services

AMAZON.COM, INC.
Segment Information
(in millions)

	Three Months Ended		Twelve Months Ended	
	December 31,		December 31,	
	2016	2017	2016	2017
	(unaudited)			
North America				
Net sales	\$26,240	\$37,302	\$ 79,785	\$ 106,110
Operating expenses	25,424	35,610	77,424	103,273
Operating income	<u>\$ 816</u>	<u>\$ 1,692</u>	<u>\$ 2,361</u>	<u>\$ 2,837</u>
International				
Net sales	\$13,965	\$18,038	\$ 43,983	\$ 54,297
Operating expenses	14,452	18,957	45,266	57,359
Operating income (loss)	<u>\$ (487)</u>	<u>\$ (919)</u>	<u>\$ (1,283)</u>	<u>\$ (3,062)</u>
AWS				
Net sales	\$ 3,536	\$ 5,113	\$ 12,219	\$ 17,459
Operating expenses	2,610	3,759	9,111	13,128
Operating income	<u>\$ 926</u>	<u>\$ 1,354</u>	<u>\$ 3,108</u>	<u>\$ 4,331</u>



Cloud and data processing

1997

Larry Page



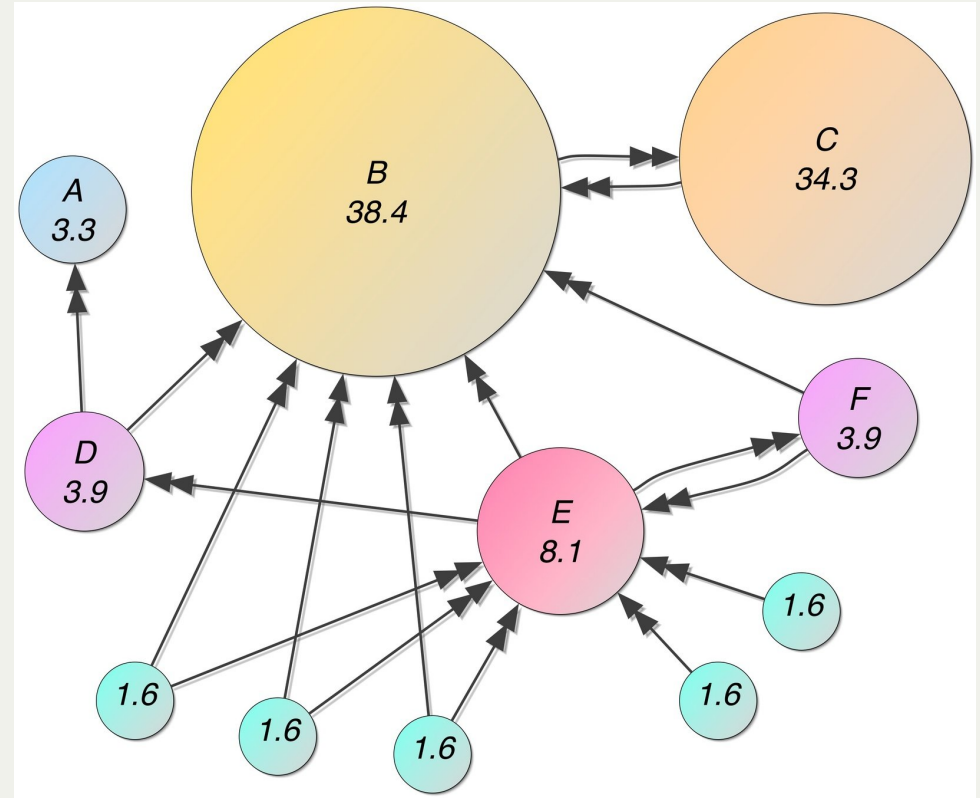
Sergey Brin



Backrub!

PageRank

Used "back links" between websites to rank the relevance of search terms.



Google

2000

Peter Lyman



Hal Varian



***“ How much information
was produced in 1999?”***

***“ If all printed material
published in the world each year
were expressed in ASCII, it could
be stored in less than 5 terabytes.***

“ The world's total production of information amounts to about 250 megabytes for each man, woman, and child on earth

285 terabytes/year



You are in: [Entertainment](#)

Front Page Tuesday, 4 January, 2000, 08:38 GMT

Westlife top millennium chart

World
UK

UK Politics

Business

Sci/Tech

Health

Education

Entertainment

Showbiz

Music

Film

Arts

TV and Radio

New Media

Reviews

Talking Point

In Depth

AudioVideo



Westlife: Top at end of 1999 and beginning of 2000

Pop band Westlife have beaten off all competition to take the first UK number one single of the new millennium.

The Irish boy band held on to the top spot after obtaining the last number one single of the century in last week's charts, with I Have A Dream/Seasons In The Sun.

See also:

- ▶ 27 Dec 99 | Entertainment Westlife top century's last chart
- ▶ 12 Dec 99 | Entertainment Westlife win song award

Internet links:

- ▶ BBC Radio 1
- ▶ The Worldwide Westlife Web Site (fan site)
- ▶ BBC Radio 2

The BBC is not responsible for the content of external internet sites

Top Entertainment stories now:

- ▶ Channel 4 boss warns of cuts
- ▶ Record numbers watch Big Brother
- ▶ Inquest opens into Entwistle's death
- ▶ Hindi soap set to storm US
- ▶ Gallagher attacks 'liar'

COMMONWEALTH GAMES

BBC SPORT

BBC Weather

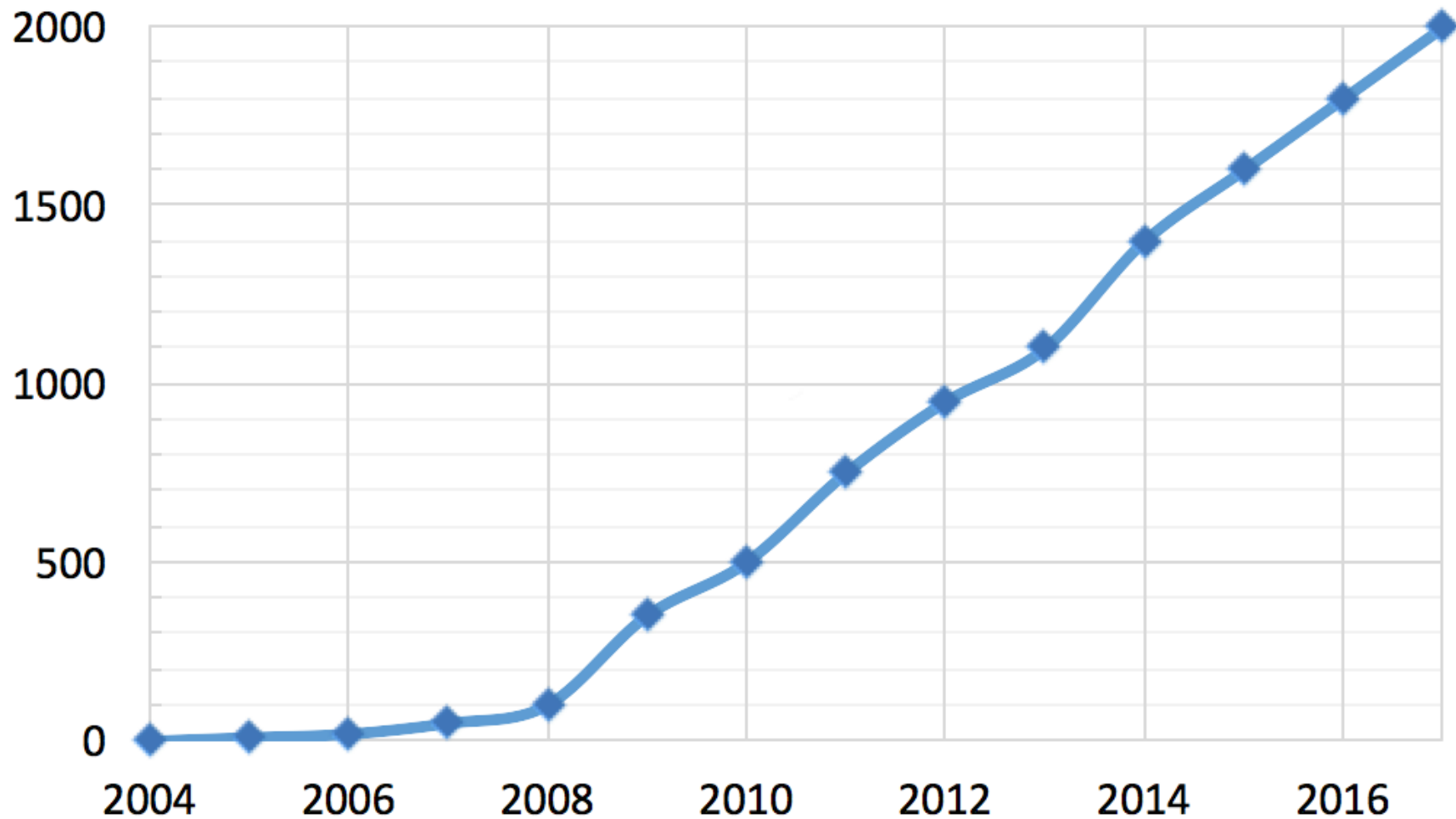
2004

Mark Zuckerberg



1 facebook 1 0

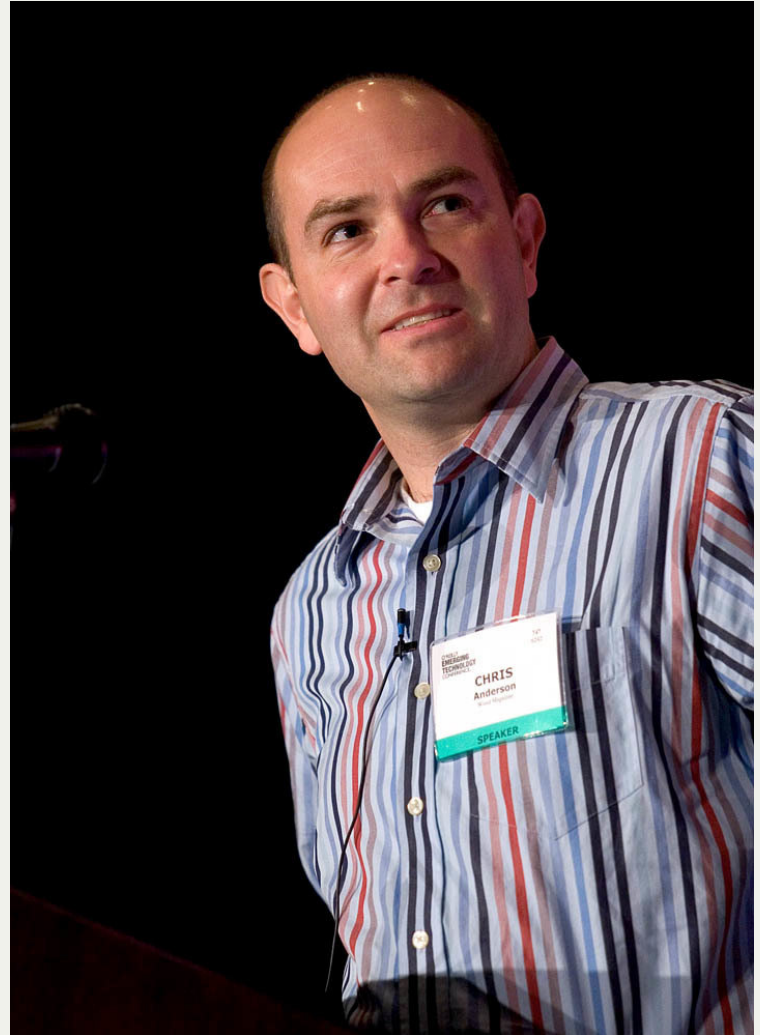
MONTHLY USERS ON FACEBOOK 2004-2017



2008

Chris Anderson

*THE END OF THEORY: THE DATA
DELUGE MAKES THE SCIENTIFIC
METHOD OBSOLETE*



“ At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics.

“faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.

2010

Eric Schmidt

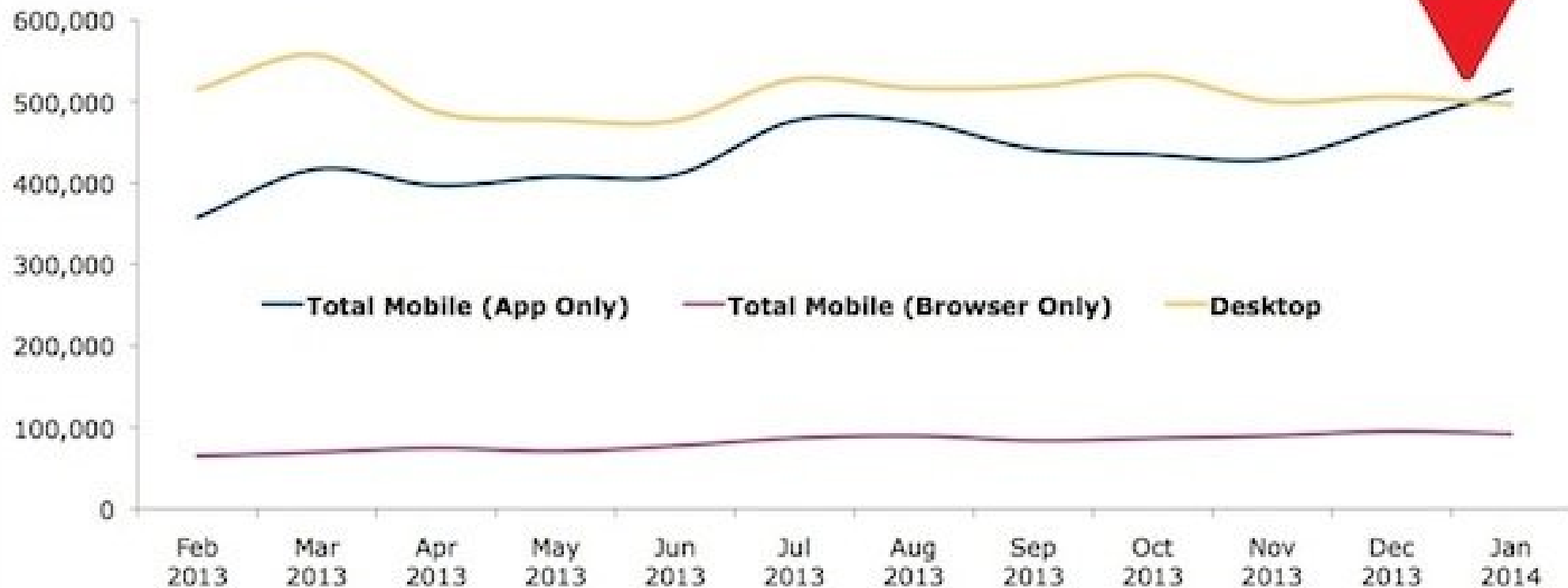
// as much data is now being created every two days, as was created from the beginning of human civilization to the year 2003



2014

Time Spent With the Internet, by Device, in the US

total minutes (mm) per month
February 2013 - January 2014





2016

Internet traffic.

In one year.

**1,000,000,000,000,000,
000,000 bytes**

10²¹ bytes

1 Zettabyte.

**We've come a long
way.**

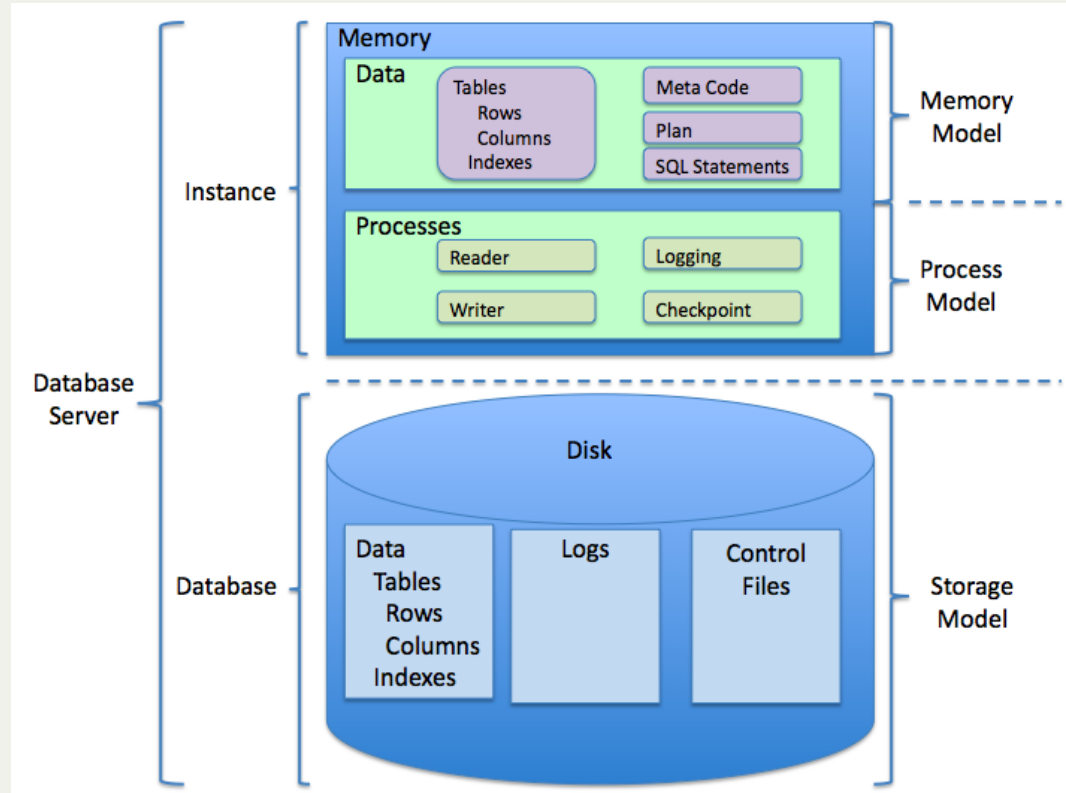
**... but what can we do with all
this data?**

Store it for analysis

Remember RDBMS?

That works fine for a while...

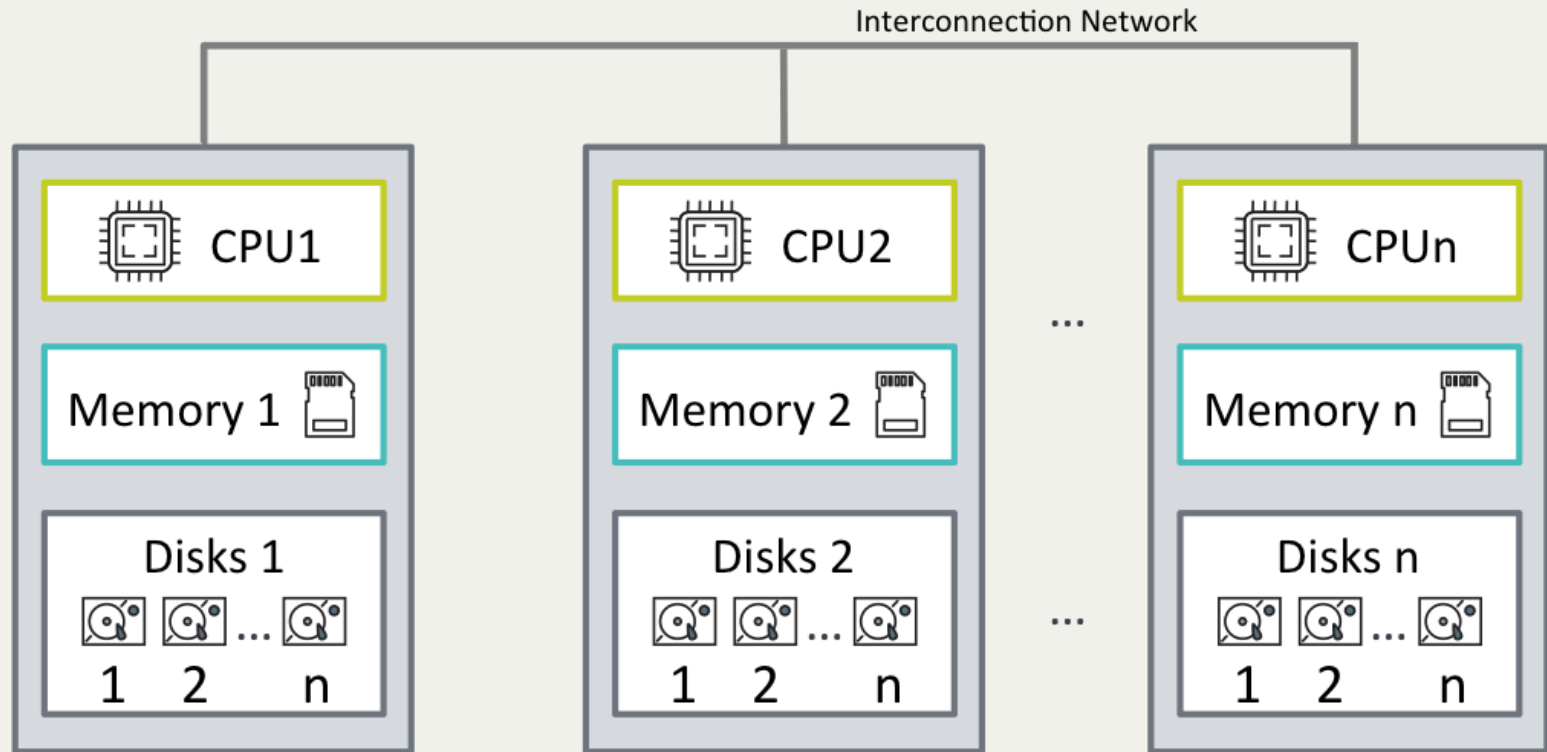
But eventually you need a
single big machine...







The answer is to "scale out"



High-availability of data

- Sharding
- Replication
- Availability zones

Tolerate failures

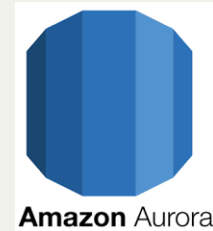
Windows

A fatal exception 0E has occurred at 0028:C0011E36 in UXD UHM(01) + 00010E36. The current application will be terminated.

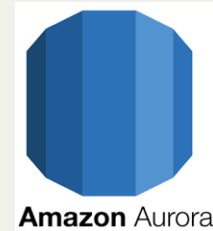
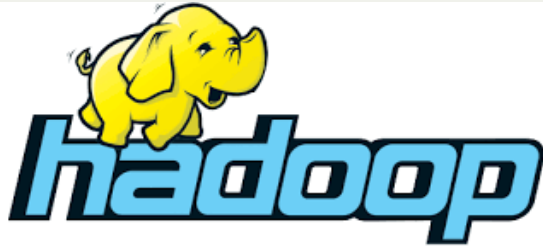
- * Press any key to terminate the current application.
- * Press CTRL+ALT+DEL again to restart your computer. You will lose any unsaved information in all applications.

Press any key to continue

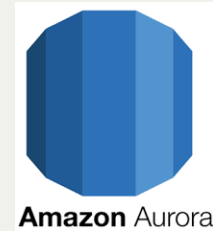
Many options



Many options



Many options



**What can we do with
the data?**



Machine learning!

Theory

Given N : $\{(x_1, y_1), \dots, (x_N, y_N)\}$

we need a function: $g : X \rightarrow Y$

scoring: $f : X \times Y \rightarrow \mathbb{R}$

estimate risk: $R_{emp}(g) = \frac{1}{N} \sum_i L(y_i, g(x_i))$

JOKES



**Use data to train a
model.**

**Use the model to
estimate something
new with partial data.**

Classification.



Regression.



Clustering.

The Netflix logo is centered in the lower half of the image. It consists of the word "NETFLIX" in a bold, white, sans-serif font. Each letter has a black drop shadow that gives it a 3D effect. The text is set against a solid red rectangular background.

NETFLIX

Classification.

Regression.

Clustering.

Let's get some data!

Open datasets

The image shows the Kaggle logo, which consists of the word "kaggle" in a lowercase, blue, sans-serif font. A small "TM" trademark symbol is located to the upper right of the letter "e". The logo is centered on a white rectangular background.

kaggle™

US Pipeline Accidents

Reviewed Dataset

Oil Pipeline Accidents, 2010-Present

Causes, injuries/fatalities, and costs of pipeline leaks and spills

Department of Transportation · last updated 2 years ago (Version 1)

36 voters

[Data](#) [Overview](#) [Kernels](#) [Discussion](#) [Activity](#) [Download \(217 KB\)](#) [New Kernel](#)

Data (217 KB) [API](#) `kaggle datasets download -d usdot/pipeline-accid...` [Download All](#)

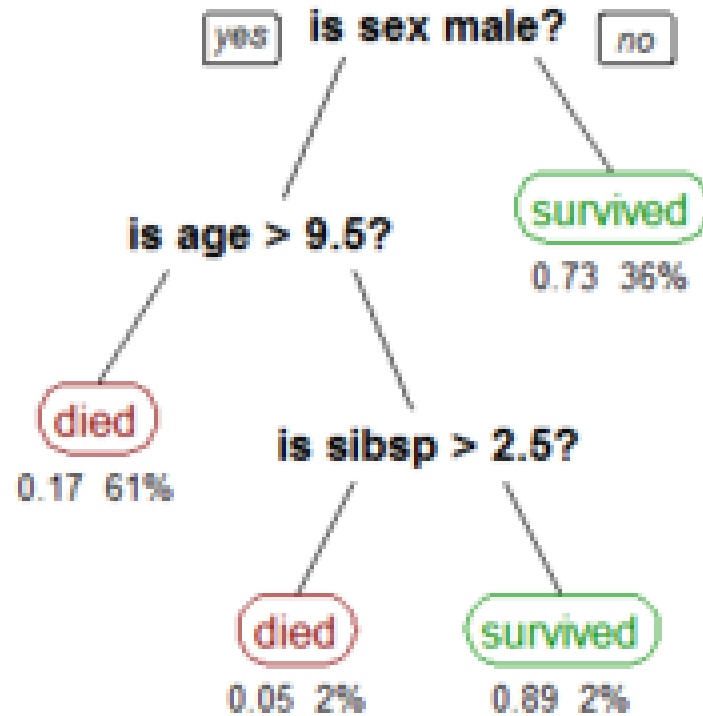
Data Sources	About this file Edit	Columns Edit
<ul style="list-style-type: none">database.csv 2795 x 48	Accident reports for oil pipelines	<ul style="list-style-type: none"># Report Number# Supplemental Number# Accident Year📅 Accident Date/Time# Operator IDA Operator NameA Pipeline/Facility NameA Pipeline LocationA Pipeline Type

Report Number	Accident Latitude	
Supplemental Number	Accident Longitude	
Accident Year	Cause Category	
Accident Date/Time	Cause Subcategory	Public Fatalities
Operator ID	Unintentional Release (Barrels)	All Fatalities
Operator Name	Intentional Release (Barrels)	Property Damage Costs
Pipeline/Facility Name	Liquid Recovery (Barrels)	Lost Commodity Costs
Pipeline Location	Net Loss (Barrels)	Public/Private Property Damage Costs
Pipeline Type	Liquid Ignition	Emergency Response Costs
Liquid Type	Liquid Explosion	Environmental Remediation Costs
Liquid Subtype	Pipeline Shutdown	Other Costs
Liquid Name	Shutdown Date/Time	All Costs
Accident City	Restart Date/Time	
Accident County		
Accident State		

Accident Date/Time	Pipeline Type	Liquid Type	Accident State	Cause Category	All Costs
-------------------------------	----------------------	--------------------	-----------------------	---------------------------	------------------

**I know where and
when I have a pipeline
accident.**

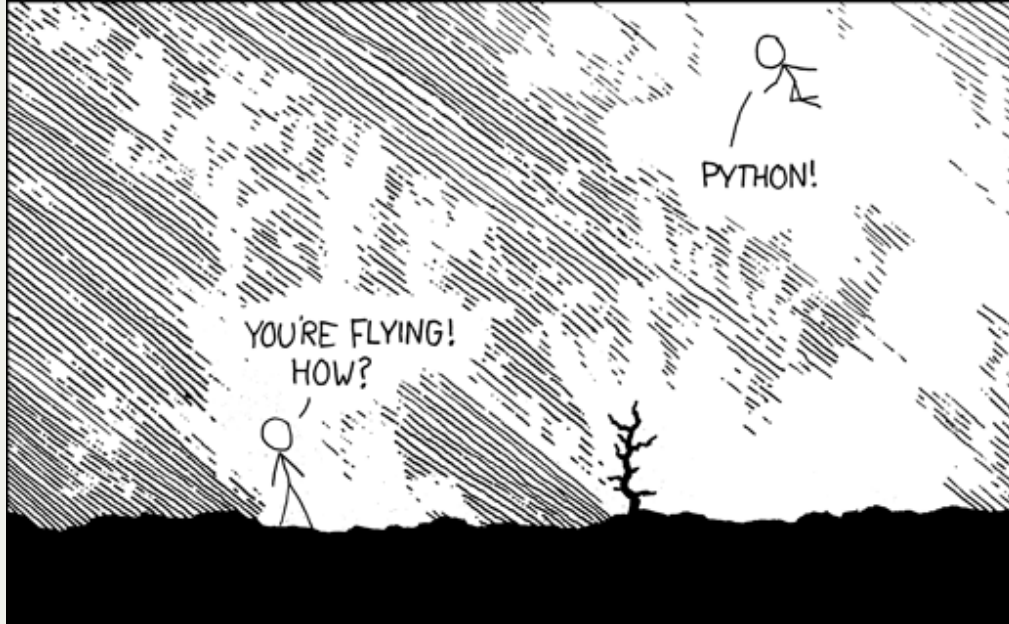
**What might the cause
be?**



Decision tree learning



pythonTM



I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"

I DUNNO...
DYNAMIC TYPING?
WHITESPACE?


COME JOIN US!
PROGRAMMING IS FUN AGAIN!
IT'S A WHOLE NEW WORLD UP HERE!



BUT HOW ARE YOU FLYING?

I JUST TYPED
import antigravity

THAT'S IT?



... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

Toolkit



```
import numpy as np
import pandas as pd
from sklearn import tree
from sklearn.base import TransformerMixin
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn_pandas import DataFrameMapper, CategoricalImputer

df = pd.read_csv("oil-pipelines-database.csv", index_col="Report Number")
df['Accident Date/Time'] = pd.to_datetime(df['Accident Date/Time'])
df = df[df["Pipeline Location"] == "ONSHORE"]
```

```
class DateEncoder(TransformerMixin):
    def fit(self, X, y=None):
        return self

    def transform(self, X):
        dt = X.dt
        return pd.concat([dt.year, dt.month, dt.day], axis=1)
```

```
cols = [
    ('Accident Date/Time', DateEncoder()), # 0, 1, 2
    (['Pipeline Type'], LabelEncoder()), # 3
    (['Liquid Type'], LabelEncoder()), # 4
    (['Accident State'], LabelEncoder()), # 5
    (['Cause Category'], LabelEncoder()), # 6
    (['All Costs'], StandardScaler()), # 7
]
```

```
clf = tree.DecisionTreeClassifier()
clf.fit(X, y)

def predict_cause(day_of_month, month, pipeline_type, liquid, state):
    value = clf.predict(np.array([
        [
            month,
            day_of_month,
            label_lookup("Pipeline Type")[pipeline_type],
            label_lookup("Liquid Type")[liquid],
            label_lookup('Accident State')[state]
        ]
    ]))
    return label_lookup("Cause Category", inverse=True)[value[0]]
```

```
predict_cause(15, 5, "UNDERGROUND", "CRUDE OIL", "OIL")
```

```
OUT[1]: CORROSION
```

```
predict_cause(1, 1, "UNDERGROUND", "CRUDE OIL", "C
```

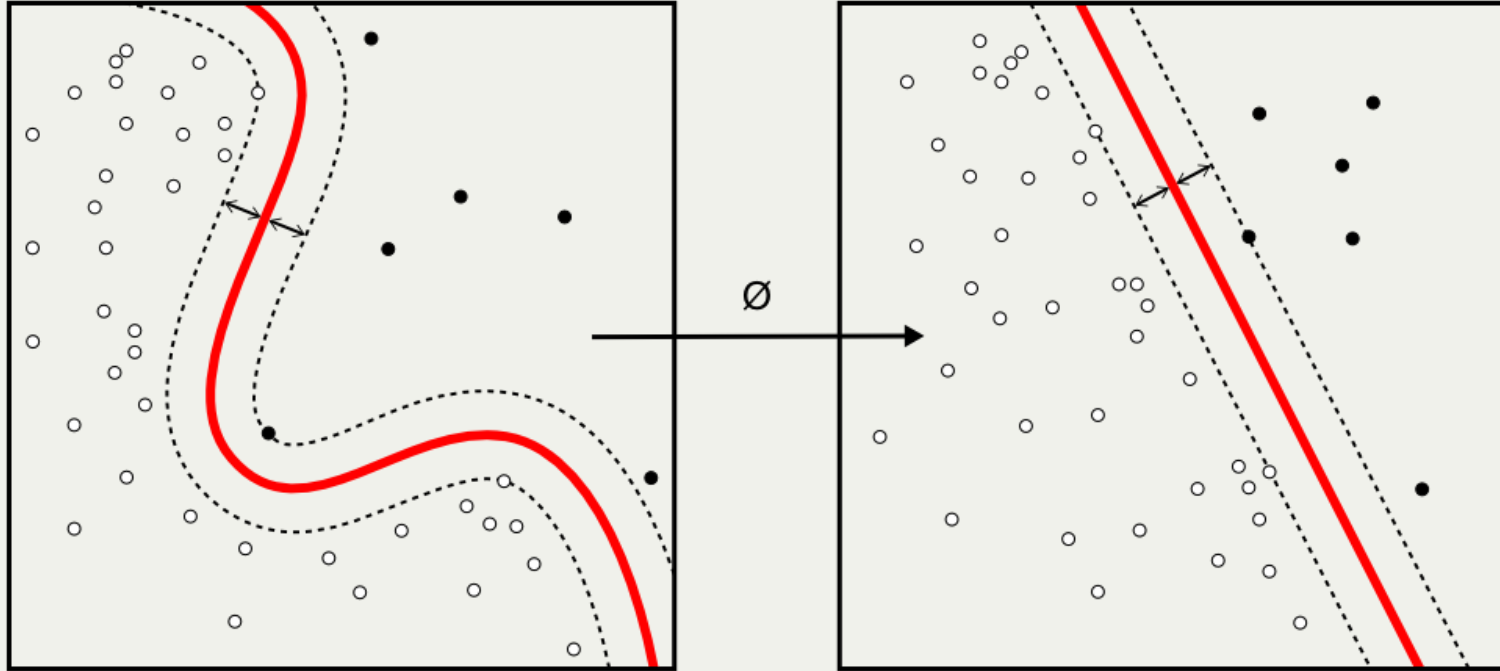
```
OUT[2]: MATERIAL/WELD/EQUIP FAILURE
```

**I now know what the
cause might be.**



**How much might it
cost us?**

Support Vector Regression



```
X_money = X_cause = tf_arr[:, 1: 7]
y_money = tf_arr[:, [7]]
svr_rbf = SVR(kernel='rbf', C=1e3, gamma=0.1)
y_rbf = svr_rbf.fit(X_money, y_money)

def predict_cost(day_of_month, month, pipeline_type, liquid, state, cause):
    value = y_rbf.predict(np.array([
        [
            month,
            day_of_month,
            label_lookup("Pipeline Type")[pipeline_type],
            label_lookup("Liquid Type")[liquid],
            label_lookup("Accident State")[state],
            label_lookup("Cause Category")[cause],
        ]
    ]))
    scale = StandardScaler()
    xp = scale.fit_transform(df['All Costs'].values.reshape(-1, 1)).flatten()
    fp = df['All Costs'].values.reshape(-1, 1).flatten()
    xp.sort()
    fp.sort()
    return np.interp(value[0], xp, fp)
```

```
In[1]: predict(15, 5, "UNDERGROUND", "CRUDE OIL", "TX")
```

```
Out[1]: "An UNDERGROUND CRUDE OIL pipeline accident in TX on day 15 of month 5  
was probably caused by CORROSION and will cost $309,853"
```

```
In[2]: predict(1, 1, "UNDERGROUND", "CRUDE OIL", "CO")
```

```
Out[2]: "An UNDERGROUND CRUDE OIL pipeline accident in CO on day 1 of month 1,  
was probably caused by MATERIAL/WELD/EQUIP FAILURE and will cost $2,18"
```



It's just that simple!

Not really this simple



Time consuming stuff:

- Feature selection & extraction
- Cross validation
- Hyper parameter tuning
- Model evaluation



scikit-learn user guide

Release 0.19.2

Is this AI?

Deep learning

**The machine that can teach
itself.**

 TensorFlow

 mxnet



Caffe

 torch

PYTORCH

theano

 Caffe2

A hand is formed by a dense collection of small, glowing green dots. The dots are arranged to create the shape of a hand with fingers spread, set against a light gray background. The text "What's next?" is written in a bold, black, sans-serif font across the center of the hand.

What's next?

A scenic landscape of a turquoise lake surrounded by mountains and forests, with the text "Data lakes." overlaid. The image shows a wide, calm lake reflecting the surrounding environment. The water is a vibrant turquoise color, and the sky is filled with soft, white clouds. The mountains are rugged and rocky, with patches of snow and green coniferous forests at their base. The overall atmosphere is serene and majestic.

Data lakes.



Predictive maintenance.

A bicycle is leaning against a metal railing. The bicycle is green and white. The railing is silver and has vertical bars. The background is a blurred outdoor scene with grass and a fence. The text "Anomaly detection." is overlaid in the center of the image in a bold, black, sans-serif font.

Anomaly detection.

Subsea UK Small Company of the Year 2018





I believe the children are our future

— *Whitney Houston* —

AZ QUOTES

**The changes will effect
young professionals
most.**

**Young professionals
have the most to gain
by understanding the
tools.**

**What steps can you
take?**

1. Look close to home.

Go and find out how your organisation is storing and analysing the data it has available.

2. Get some practical experience.

Work through online examples and experiment with open data sets.

3. Network and share your experiences.

Attend events and ask questions, meet with peers to compare notes.

**Thank you for
listening!**



Steven Rossiter
AgileTek Engineering
steve@agiletek.co.uk