The Evolution of Human-Data Interaction

Amit Somech

November 2018

TEL AVIV UNIVERSITY
אוניברסיטת תל אביב

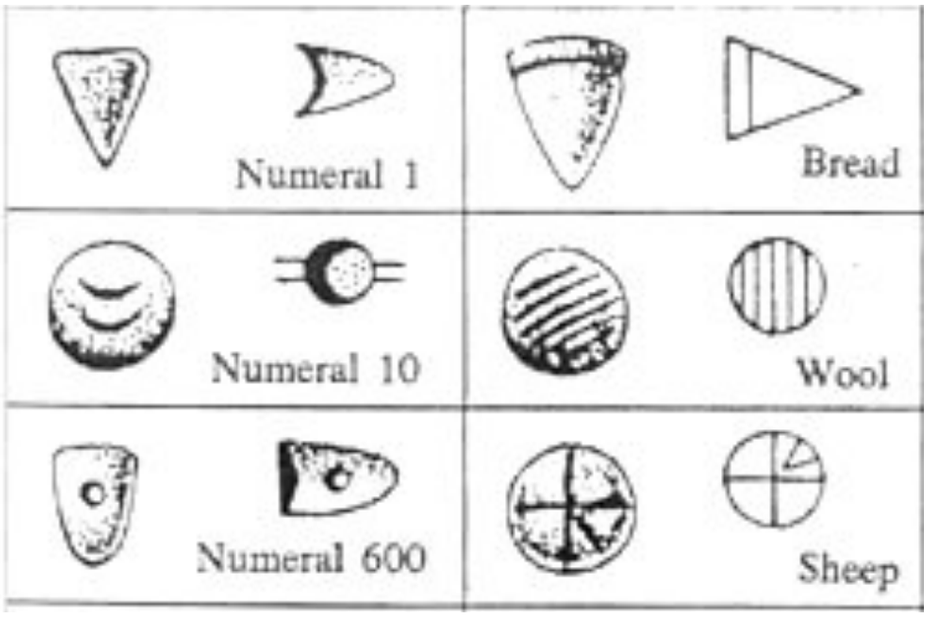
Palaeolithic Data Storage

Palaeolithic Data Storage

20000 B.C: Tribespeople would mark notches into sticks or bones, to keep track of trading activity or supplies.

Early Accounting

5000 B.C: Accounting developed using a token system: The token shapes represented various kinds of merchandise.



Early Accounting

3000 B.C: Reducing tokens dimensionality (from 3d to 2d) resulted in “Cuneiform script” (כתב יתדות).

1500 B.C: Changing from ***logograms*** (one sign = 1 item) to ***syllabograms*** (one sign = one syllable).

A clay “tablet”

Roman (distributed) Census

600 A.D: Roman empire held a census very 5 years.

A record for each citizen: name, age, family, list of property: (land, # of slaves, cattle, etc.).

Censors in each region of the empire prepared ***lists*** and sent them to Rome, for an overall ***count***.

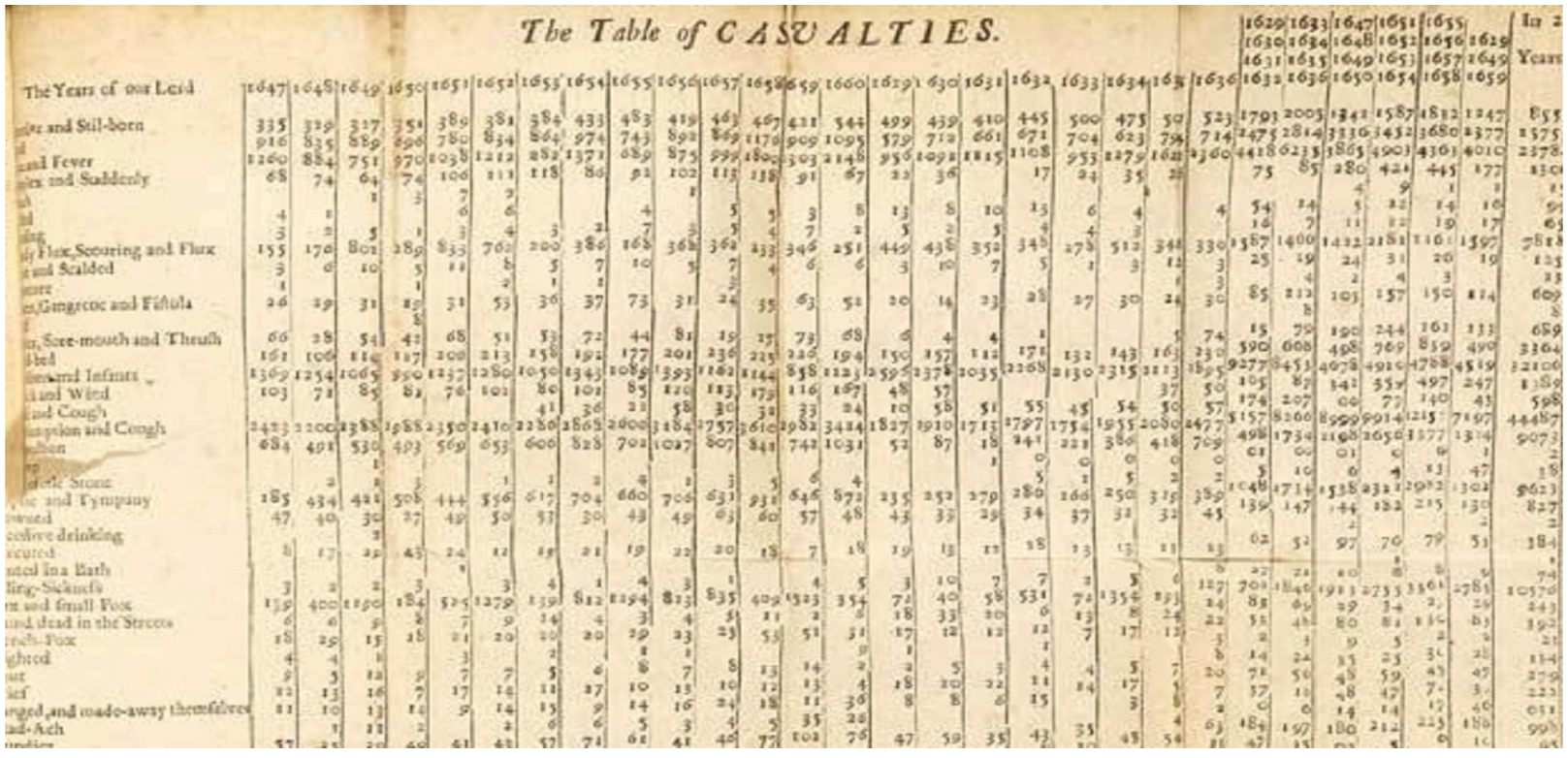
Roman "Data Center" →

Beginning of Statistics

1660 A.D: John Graunt, a shop keeper from London analyzed the weekly “bills of mortality”.

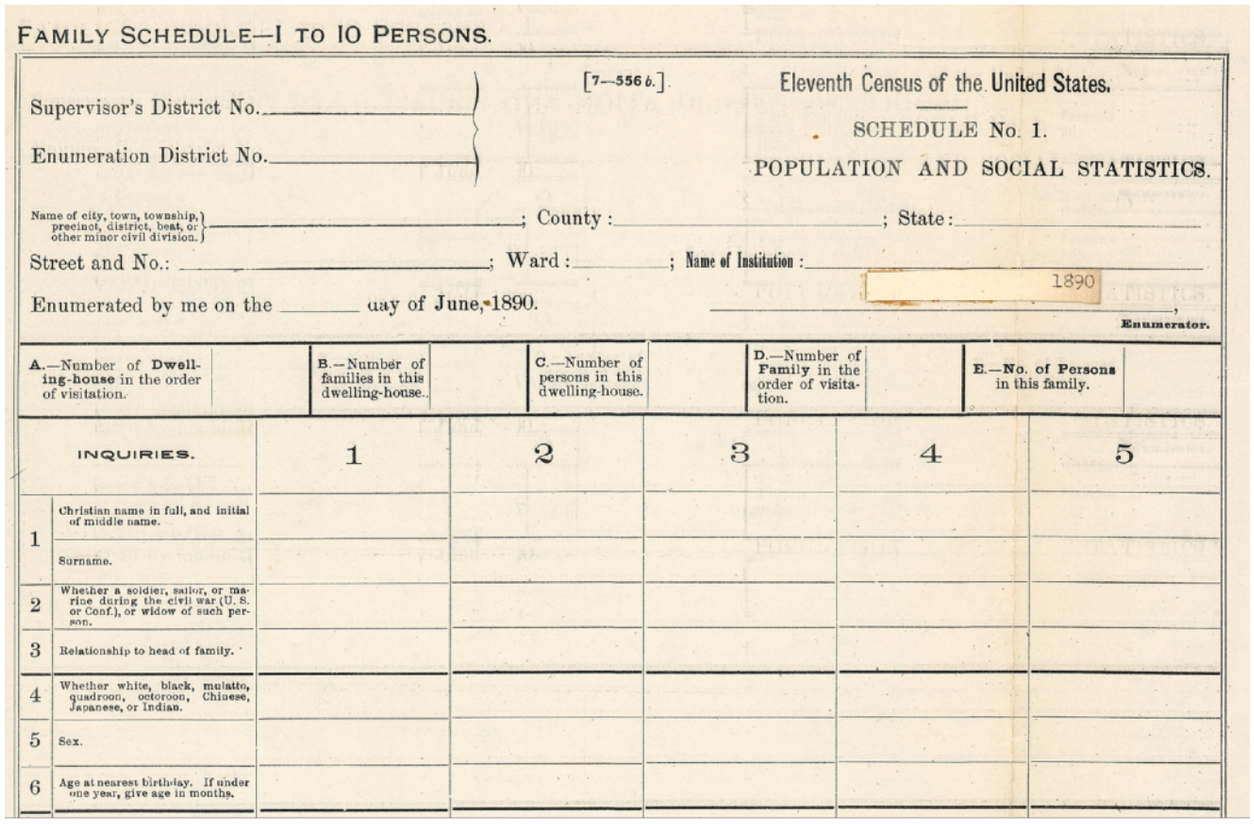
Created the first “life table” - listing the probability of dying in each age group.

1665 style “Spread sheet”



Big Data in the 19th century

1880: USA Census has a problem: It takes 8 years to "tabulate" and analyze the raw forms (50M records)



FAMILY SCHEDULE—1 TO 10 PERSONS.

[7—556 b.] Eleventh Census of the United States.

Supervisor's District No._____

Enumeration District No._____

SCHEDULE No. 1.

POPULATION AND SOCIAL STATISTICS.

Name of city, town, township, precinct, district, beat, or other minor civil division._____; County:_____; State:_____

Street and No.:_____; Ward:_____; Name of Institution:_____

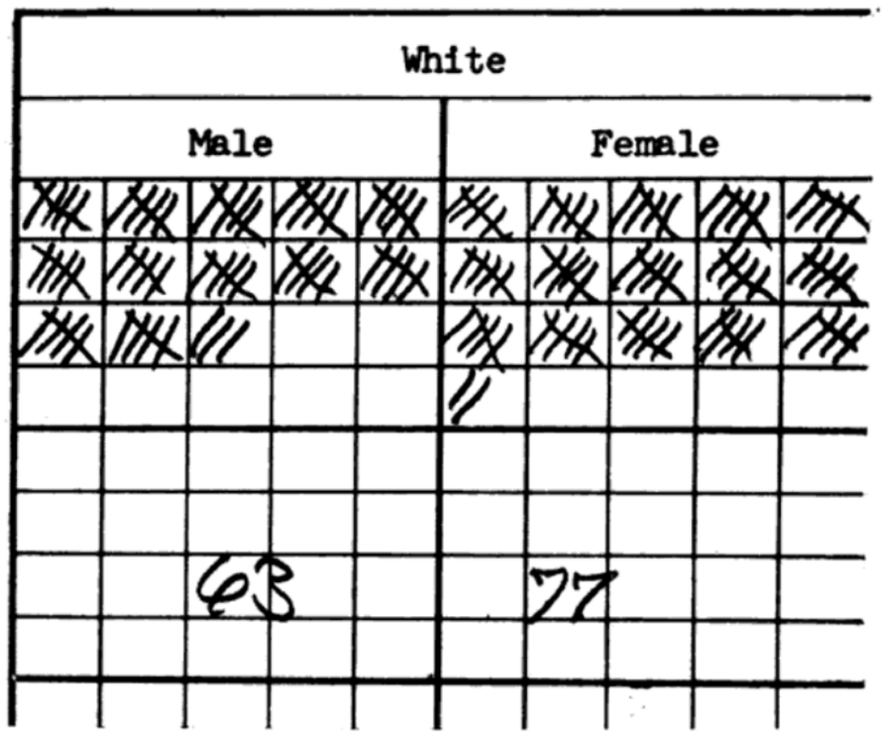
Enumerated by me on the _____ day of June, 1890. 1890

_____, Enumerator.

A.—Number of Dwelling-house in the order of visitation.	B.—Number of families in this dwelling-house.	C.—Number of persons in this dwelling-house.	D.—Number of Family in the order of visitation.	E.—No. of Persons in this family.

	INQUIRIES.	1	2	3	4	5
1	Christian name in full, and initial of middle name.					
	Surname.					
2	Whether a soldier, sailor, or marine during the civil war (U. S. or Conf.), or widow of such person.					
3	Relationship to head of family.					
4	Whether white, black, mulatto, quadroon, octoroon, Chinese, Japanese, or Indian.					
5	Sex.					
6	Age at nearest birthday. If under one year, give age in months.					

US Census form (for one family)



White

Male	Female
63	77

FIGURE 3.—Section of tally sheet, full size, with tally marks

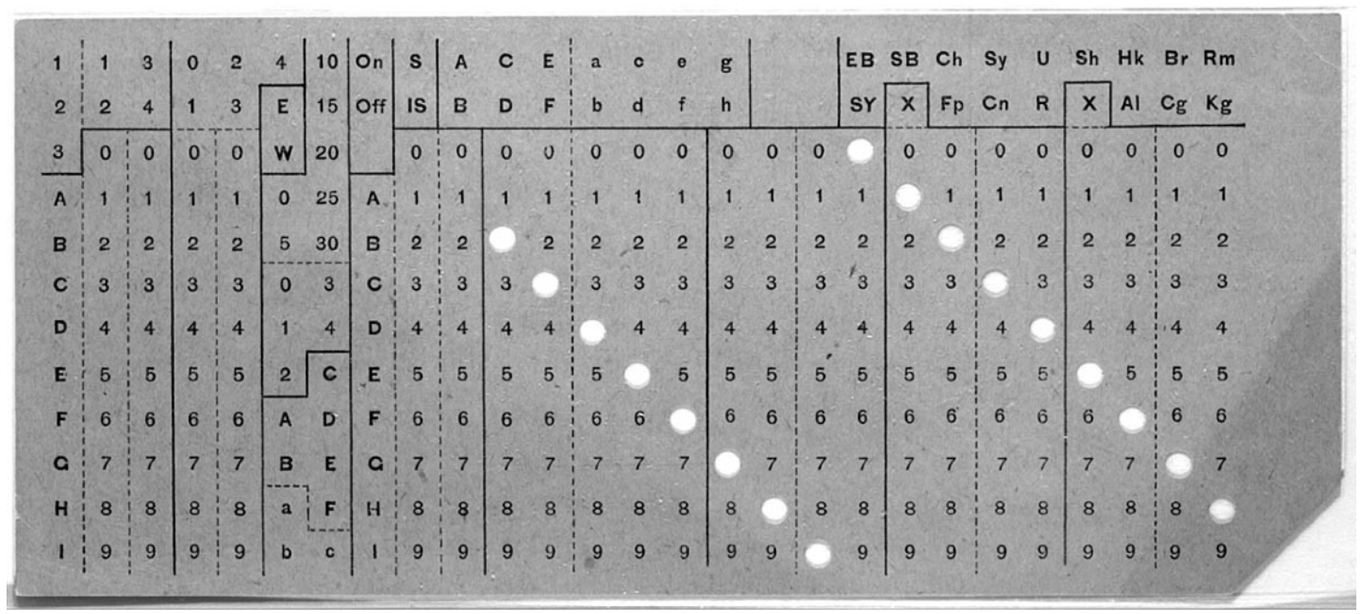
Manual "Count"

Big Data in the 19th century

1889: Herman Hollerith invents the “Tabulating machine”.

He suggests using *machine-readable*

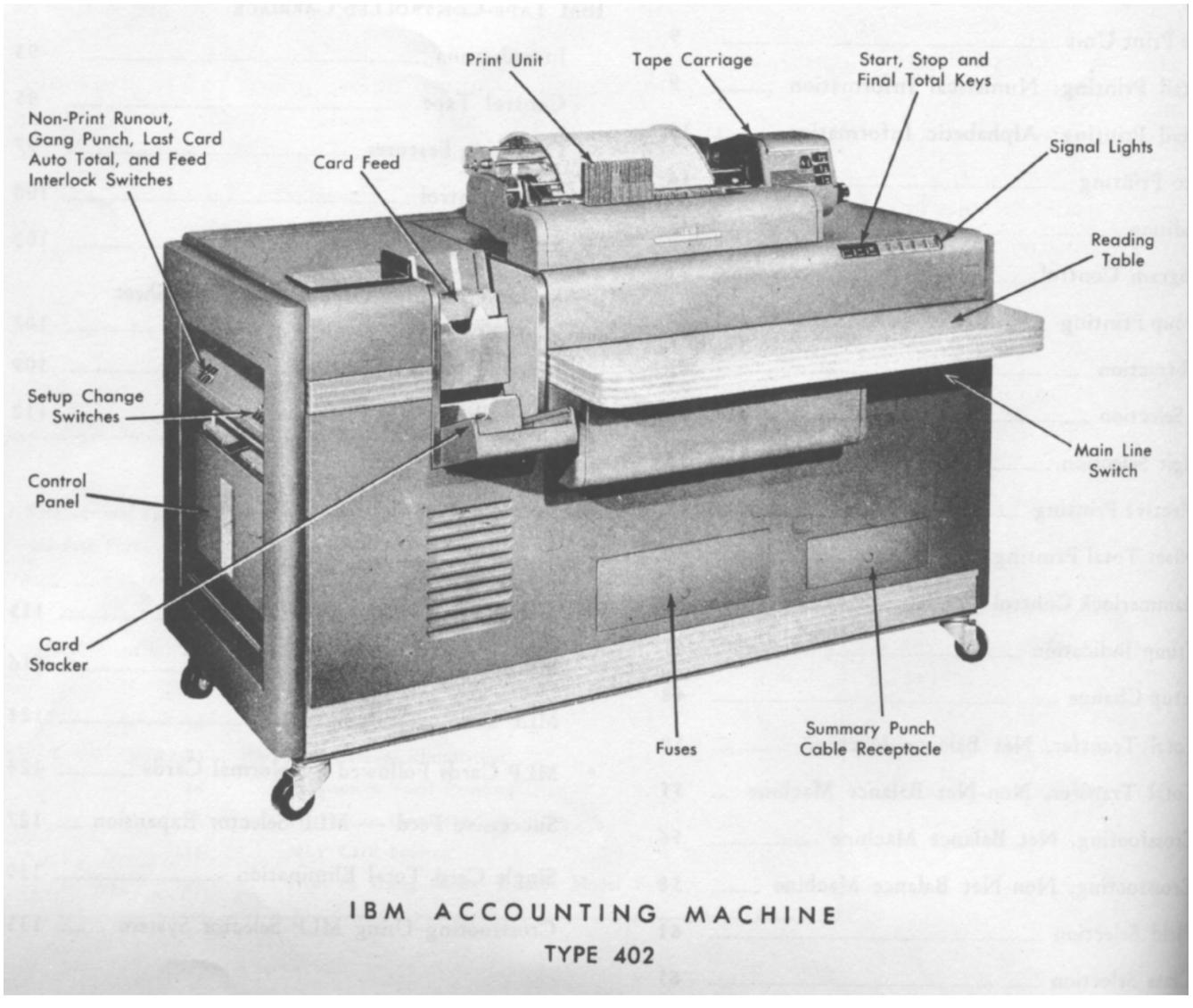
“punched cards”



punched card representing one form

Big Data in the 19th century

1911: Hollerith's (very successful) company joins the *Computing-Tabulating-Recording Company* (CTR).



IBM ACCOUNTING MACHINE
TYPE 402

IBM Type 285

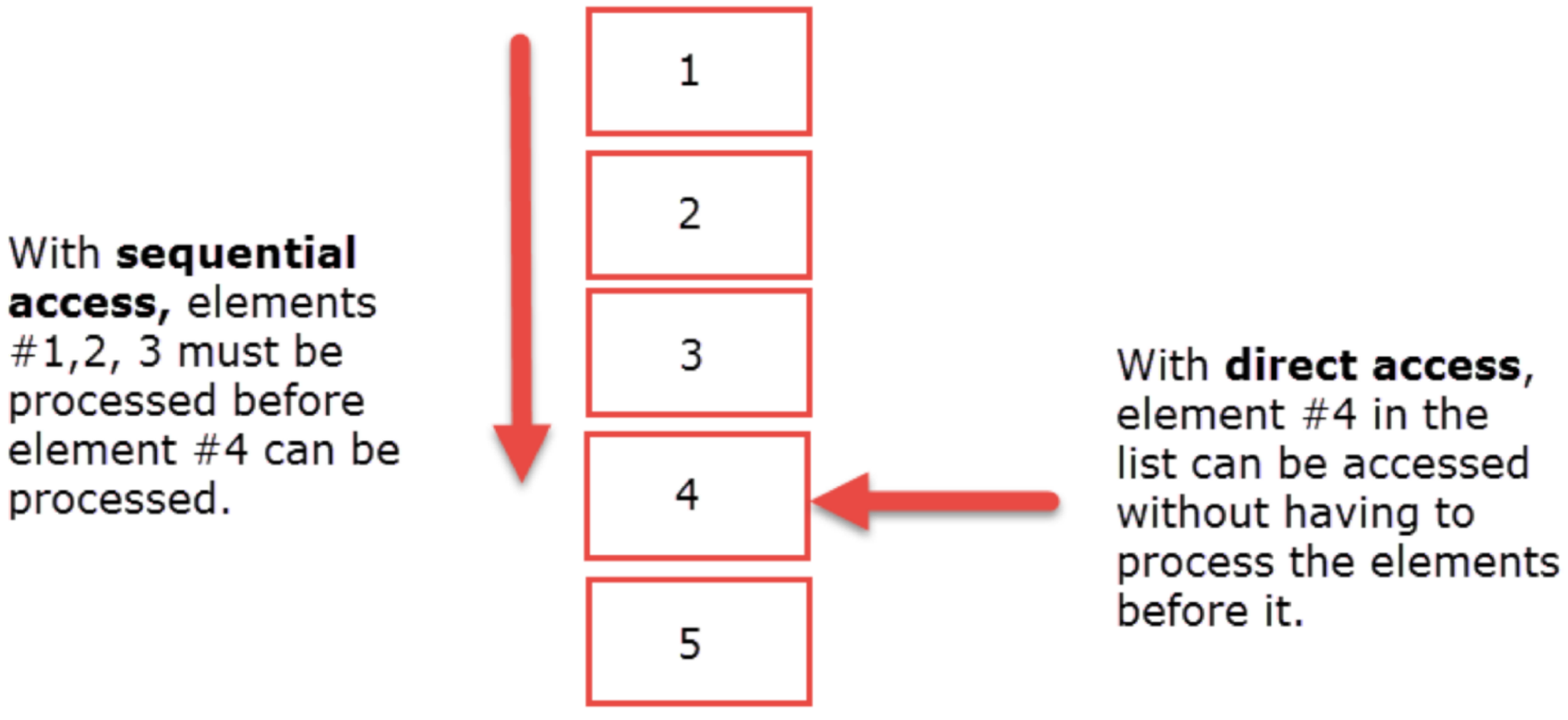
Magnetic Storage

1952s: IBM computers are now using magnetic tapes as storage! 500M long tape could hold about 3.5MB.

IBM 704 Mainframe

Magnetic Storage

Sequential Access Vs. Direct Access:

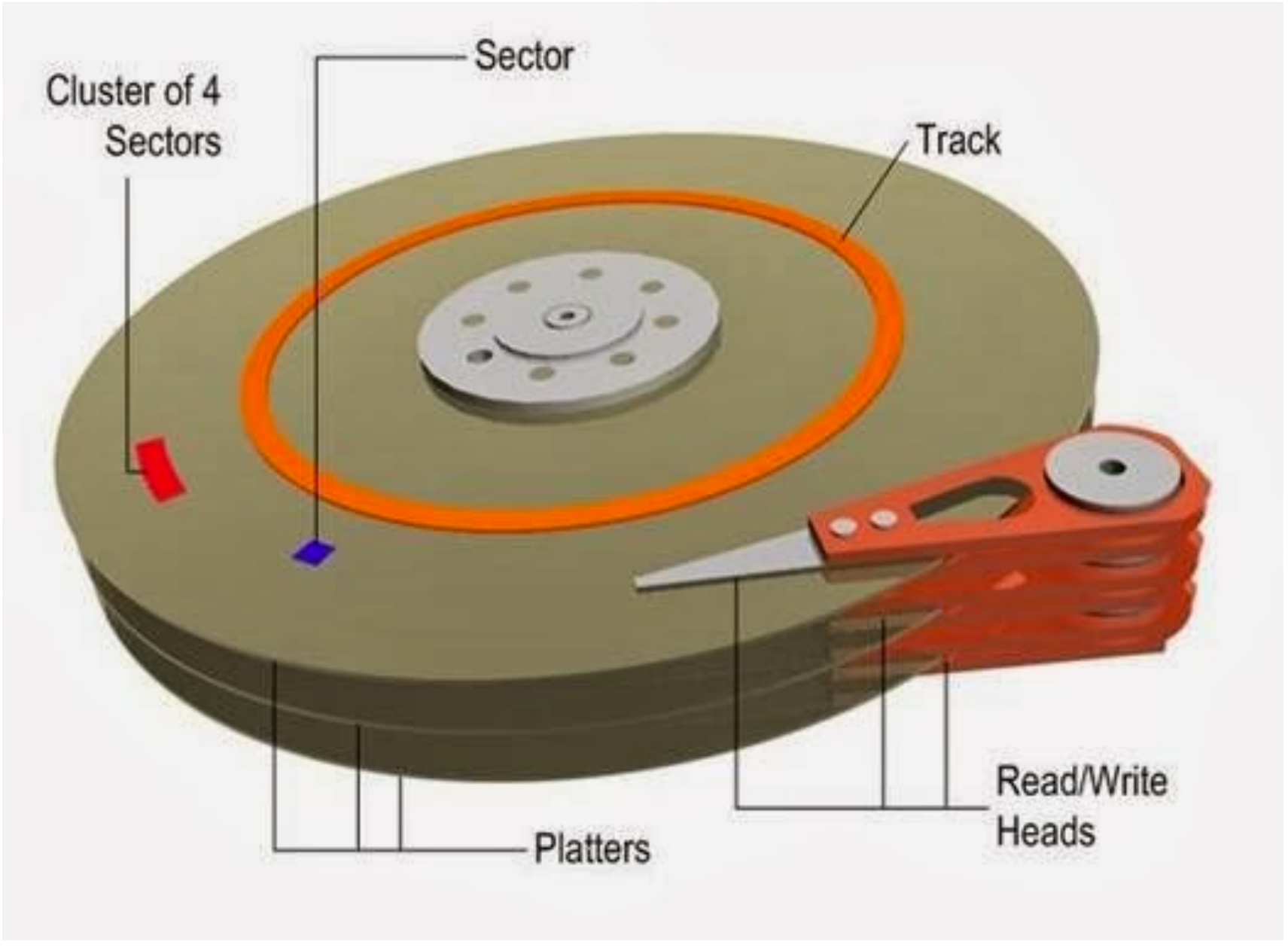


Magnetic Storage

1957: *IBM 350 is the first commercial magnetic "hard" drive. (Capacity: 3.5MB, Access time: 6 seconds).*

Magnetic Storage

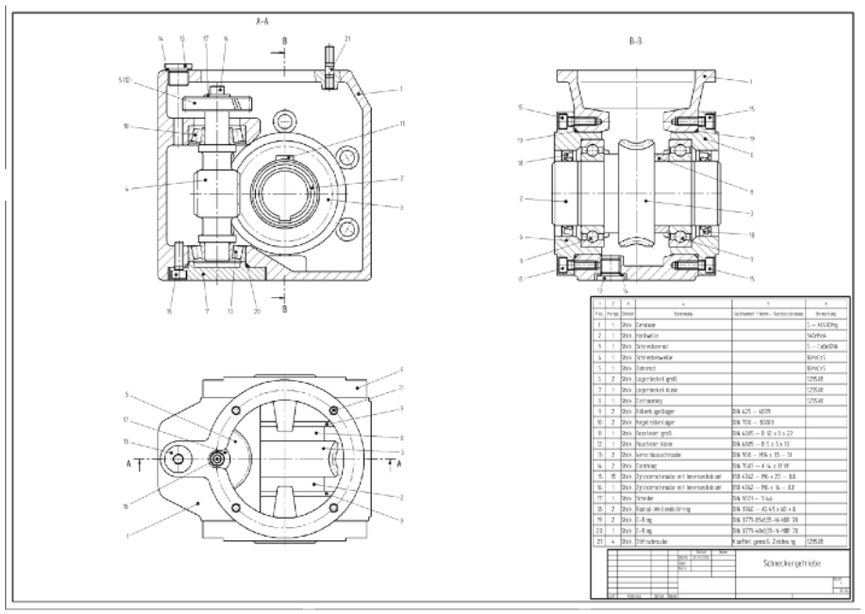
Hard Disk Schema:



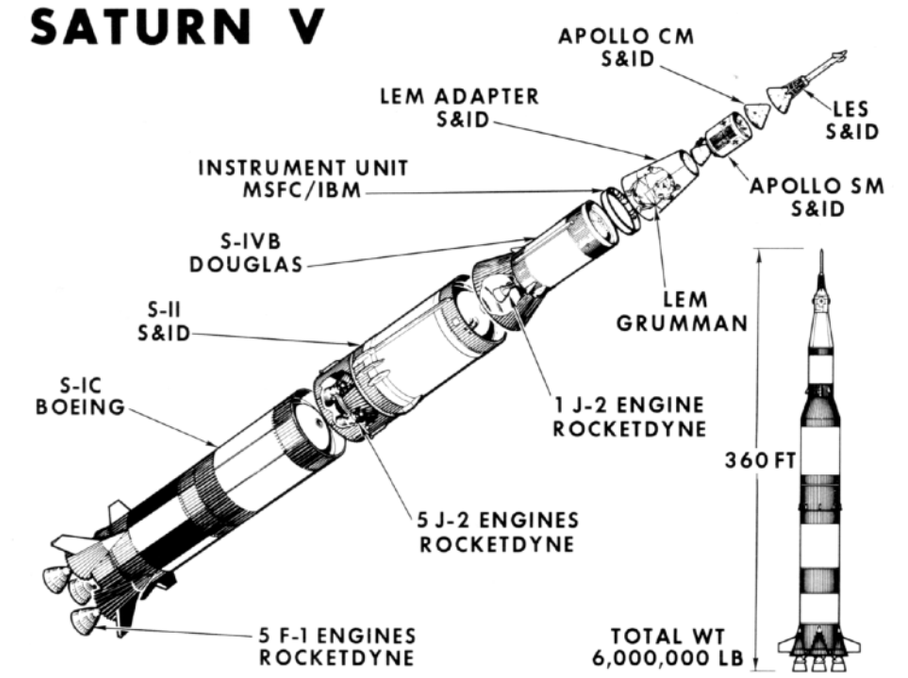
Databases and the Moon

1961: John F. Kennedy challenged American industry to send an American man to the moon. IBM is included.

1966: IBM develops the Information Management System (IMS) to store a massive Bill of Materials (BOM) for the development of Saturn V.

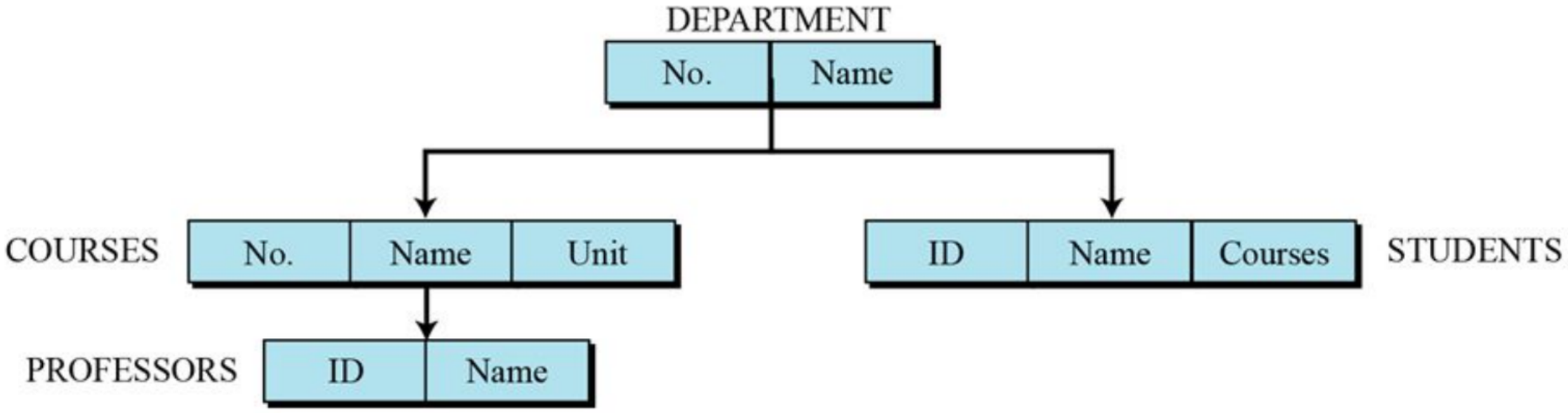


BOM Example

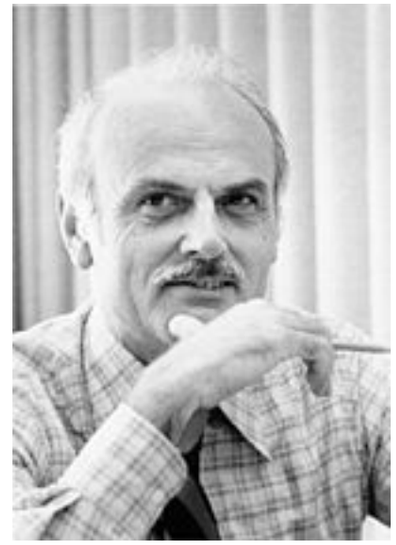


Databases and the Moon

IMS is a “Hierarchical Database”:



General Purpose Databases



1970: *"Ted" Codd (IBM) invents the* Relational Database.

1971: IBM researchers develops the SQL "query language".

```
SELECT phone
FROM   students
WHERE  dob > "1993-01-01"
```

SQL Query

Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

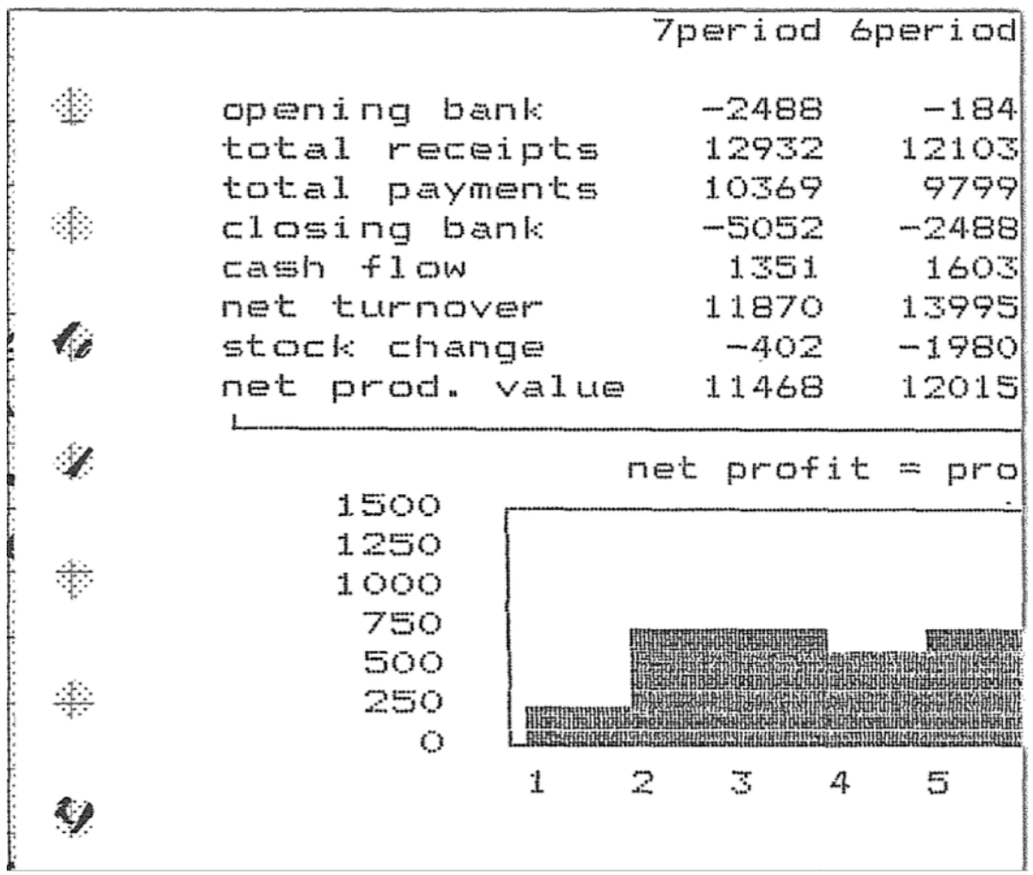
Takes_Course

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Courses

General Purpose Databases

1976-1982: Ideas of “business intelligence” and “Decision Support Systems” are now implemented in pioneer large companies.



Personal Databases

1981: IBM released the PC. One of the first programs was “Lotus-1-2-3”

A:A1: 'EMP MENU

Worksheet Range Copy Move File Print Graph Data System Quit

Global Insert Delete Column Erase Titles Window Status Page Hide

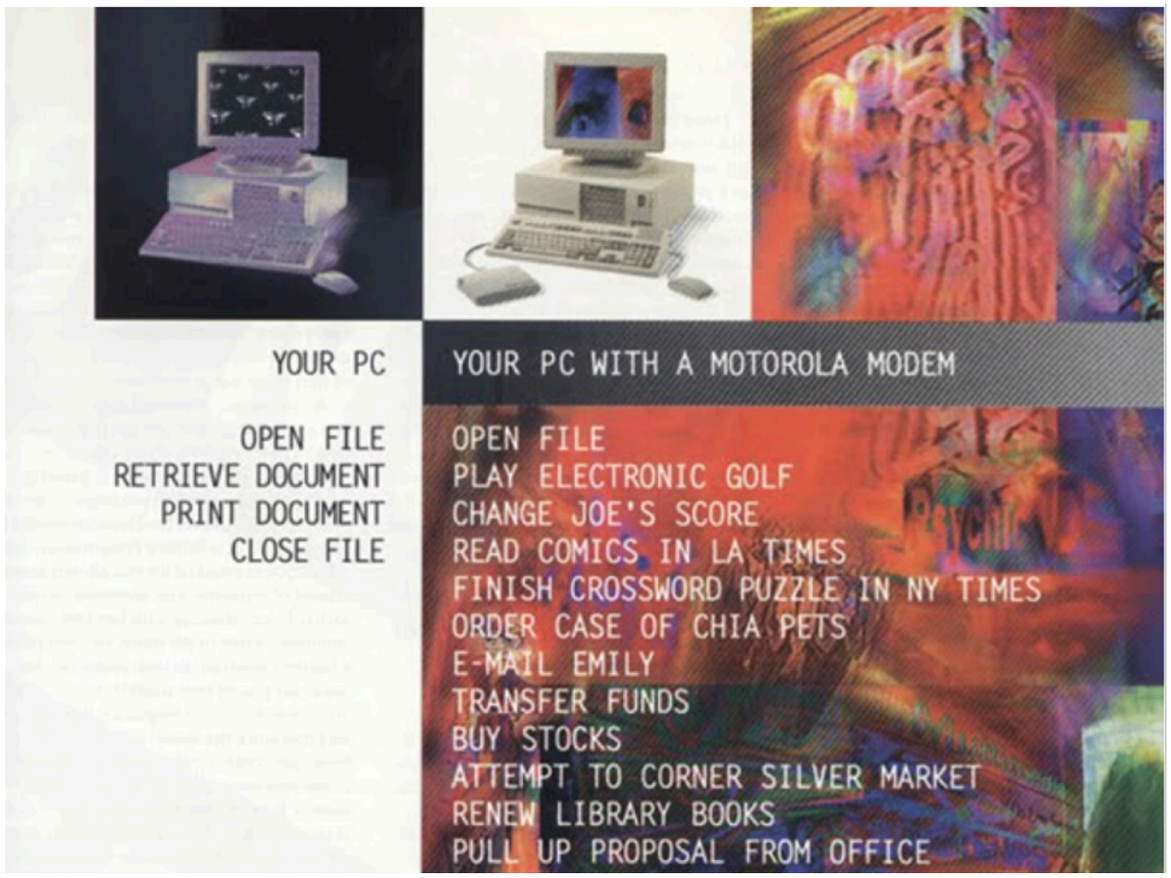
	A	B	C	D	E	F	G
1	EMP	EMP_NAME	DEPTNO	JOB	YEARS	SALARY	BONUS
2	1777	Azibad	4000	Sales	2	40000	10000
3	81964	Brown	6000	Sales	3	45000	10000
4	40370	Burns	6000	Mgr	4	75000	25000
5	50706	Caeser	7000	Mgr	3	65000	25000
6	49692	Curly	3000	Mgr	5	65000	20000
7	34791	Dabarrett	7000	Sales	2	45000	10000
8	84984	Daniels	1000	President	8	150000	100000
9	59937	Dempsey	3000	Sales	3	40000	10000
10	51515	Donovan	3000	Sales	2	30000	5000
11	48338	Fields	4000	Mgr	5	70000	25000
12	91574	Fiklore	1000	Admin	8	35000	---
13	64596	Fine	5000	Mgr	3	75000	25000
14	13729	Green	1000	Mgr	5	90000	25000
15	55957	Hermann	4000	Sales	4	50000	10000
16	31619	Hodgedon	5000	Sales	2	40000	10000
17	1773	Howard	2000	Mgr	3	80000	25000
18	2165	Hugh	1000	Admin	5	30000	---
19	23907	Johnson	1000	VP	1	100000	50000
20	7166	Laflare	2000	Sales	2	35000	5000

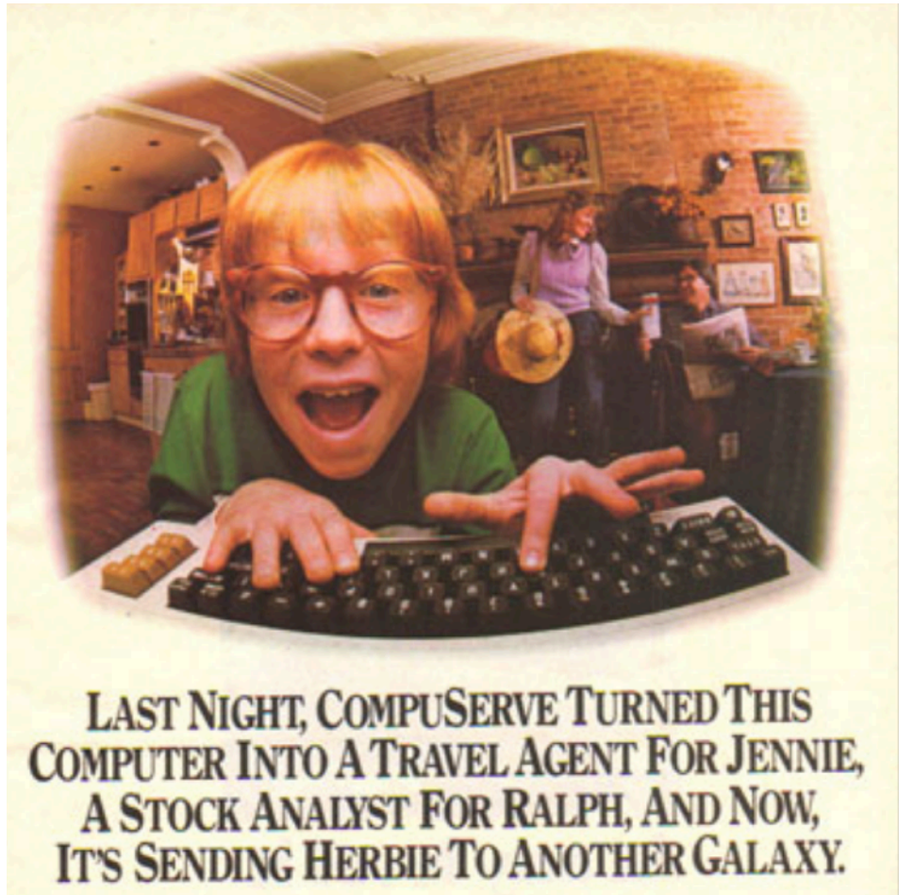
DATA.WK3

Lotus 3.0 for MS DOS

The Internet

1991: Internet Service Providers are now selling modems and internet access to the general public.





Google

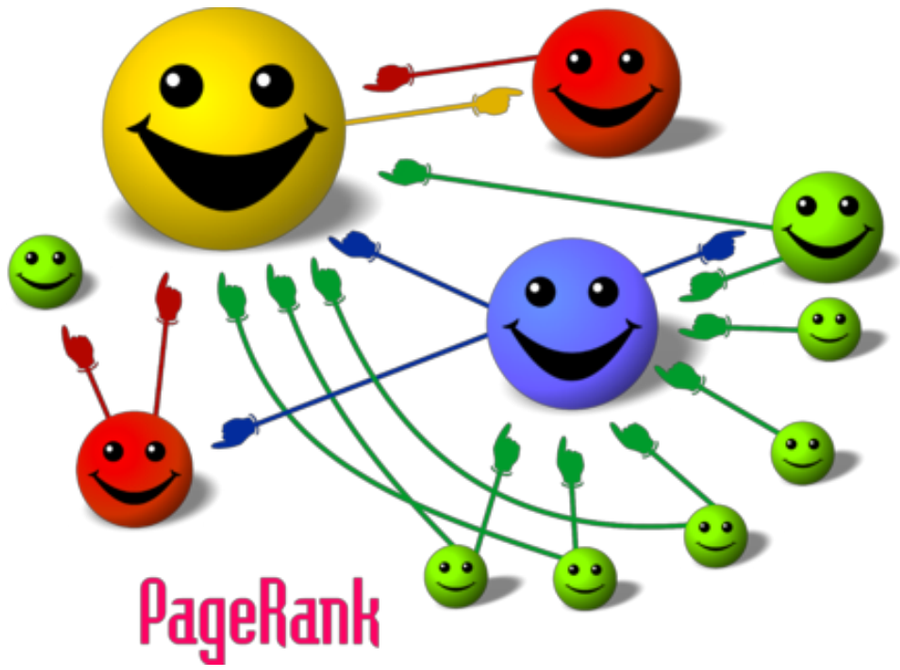
1997: The domain google.com is registered by two phD students: Larry Page and Sergey Brin

Google

2000: Google.com becomes the second most popular search engine. Why?

1. **It is the most accurate.** PageRank - redefining search “relevance”
2. **It is the largest** (in terms of indexed pages)

PageRank: ***The score of a website is proportional to the total scores of other websites which are linking to it***



Web 2.0: User Generated Data

2004: The term WEB 2.0 is now a popular term for websites facilitating **user generated data** (first coined in 1999 by Darcy Dinucci):

"The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens"

2005: facebook.com launches

2005: **youtube.com** launches

2007: Apple releases iPhone 1

The Netflix Prize

2006-2009: Netflix conducts an open competition for 1M$ for outperforming their movie **recommendations** algorithm. More than 30,000 participated.

2008: google.com now uses the “Suggest” feature:

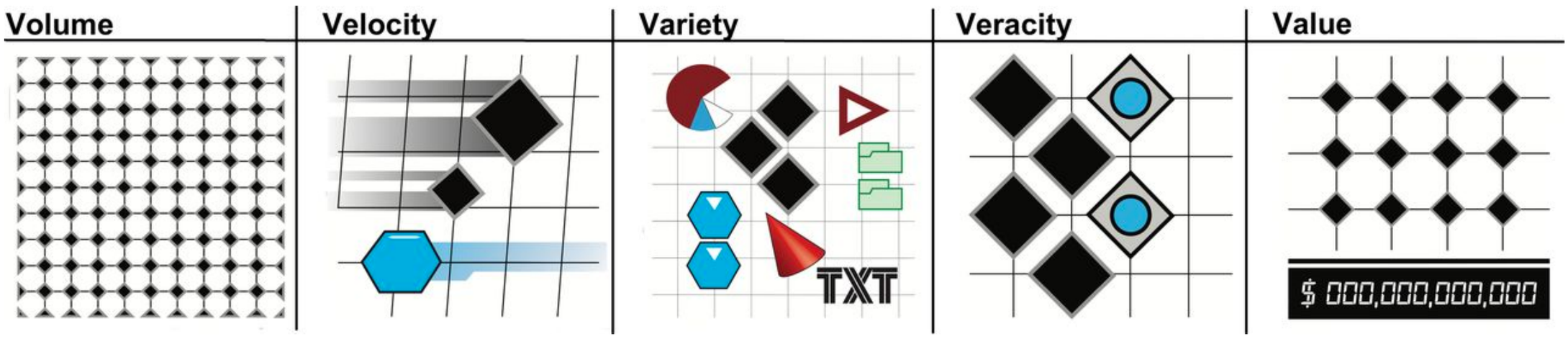
Google Suggest LABS

olympics	
olympics 2008	85,400,000 results
olympics schedule	880,000 results
olympics tv schedule	21,400,000 results
olympics 2012	738,000 results
olympics live	38,000,000 results
olympics medal count	312,000 results
olympics 2016	1,080,000 results
olympics basketball	3,540,000 results
olympics gymnastics	950,000 results
olympics online	18,200,000 results
	close

The “Big Data” Era

“As much data is now being created every two days, as was created from the beginning of human civilization to the year 2003” E. Schmidt, 2010

The main challenges of Big data:



Databases 2.0

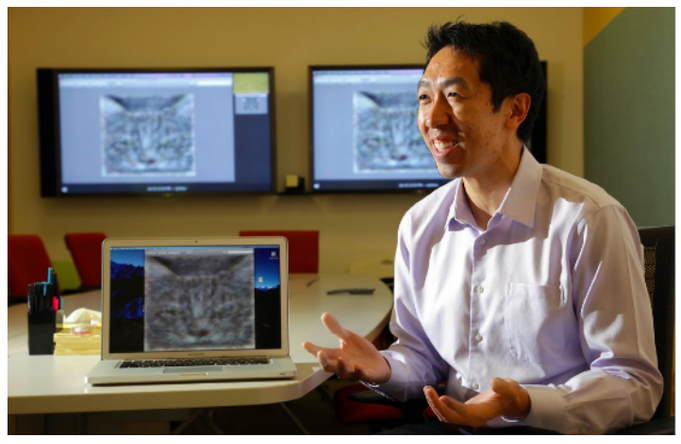
2012: Apache **Spark** is released, an open-source distributed in-memory database.

Also, “**NoSQL”** databases are becoming popular, e.g. graph DB, document DB.

Deep-Learning Revolution

2012: A mysterious “X lab” at Google publishes results of the Cat Experiment:

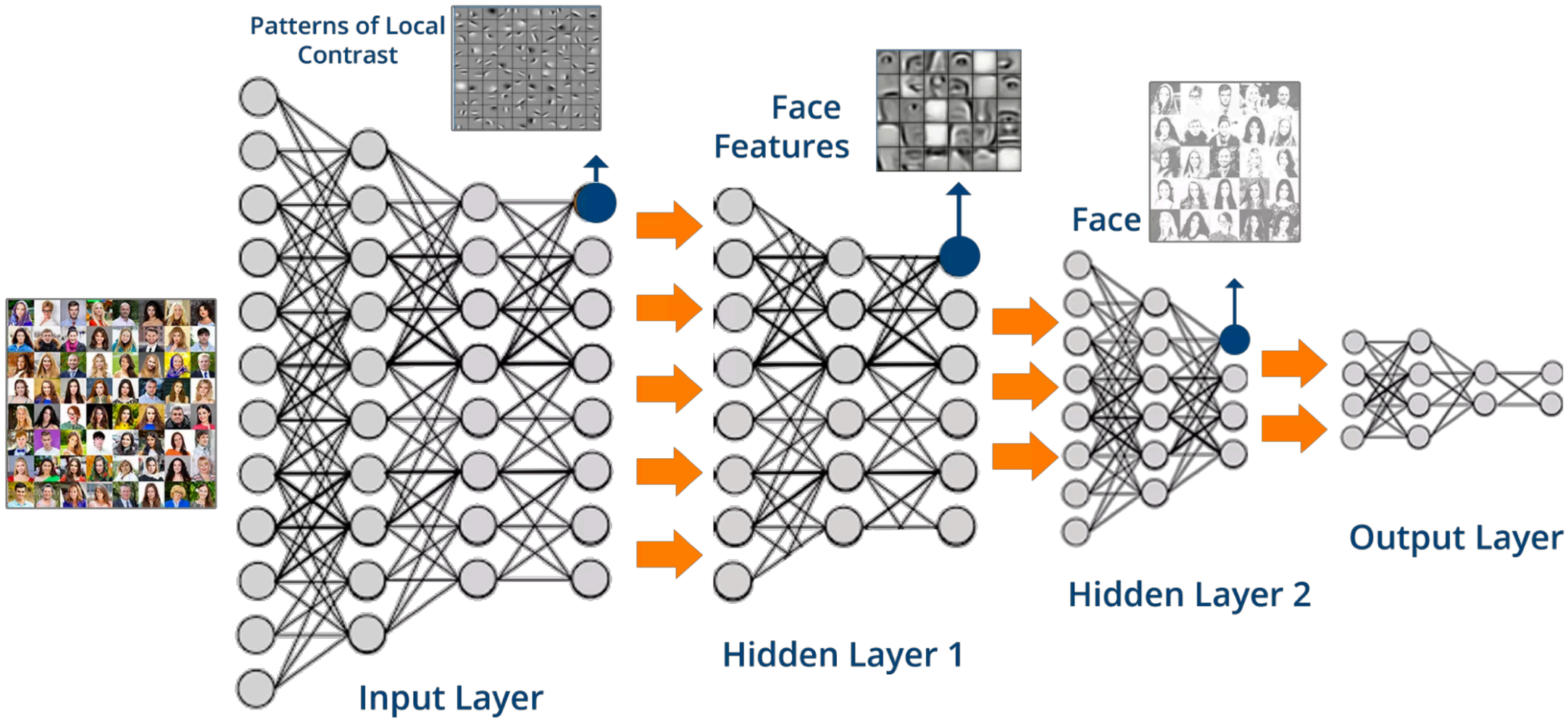
Their brain-like software successfully identified cats (without knowing what a cat is).



"Instead of having researchers trying to find out how to find edges, you instead throw a ton of data at the algorithm and you have the software automatically learn from the data” **Andrew NG,2012**

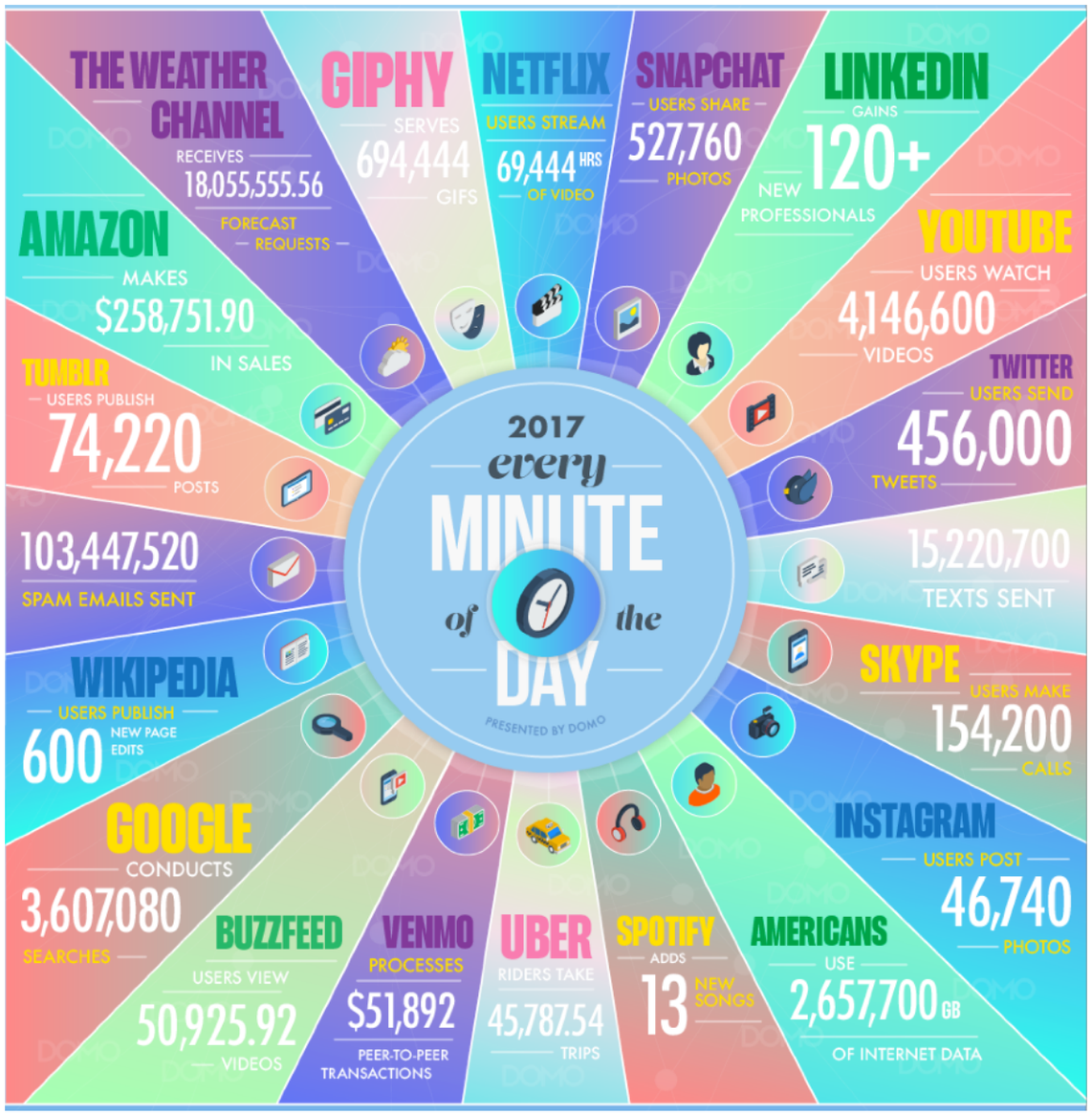
Deep-Learning Revolution

Deep-neural Architecture:



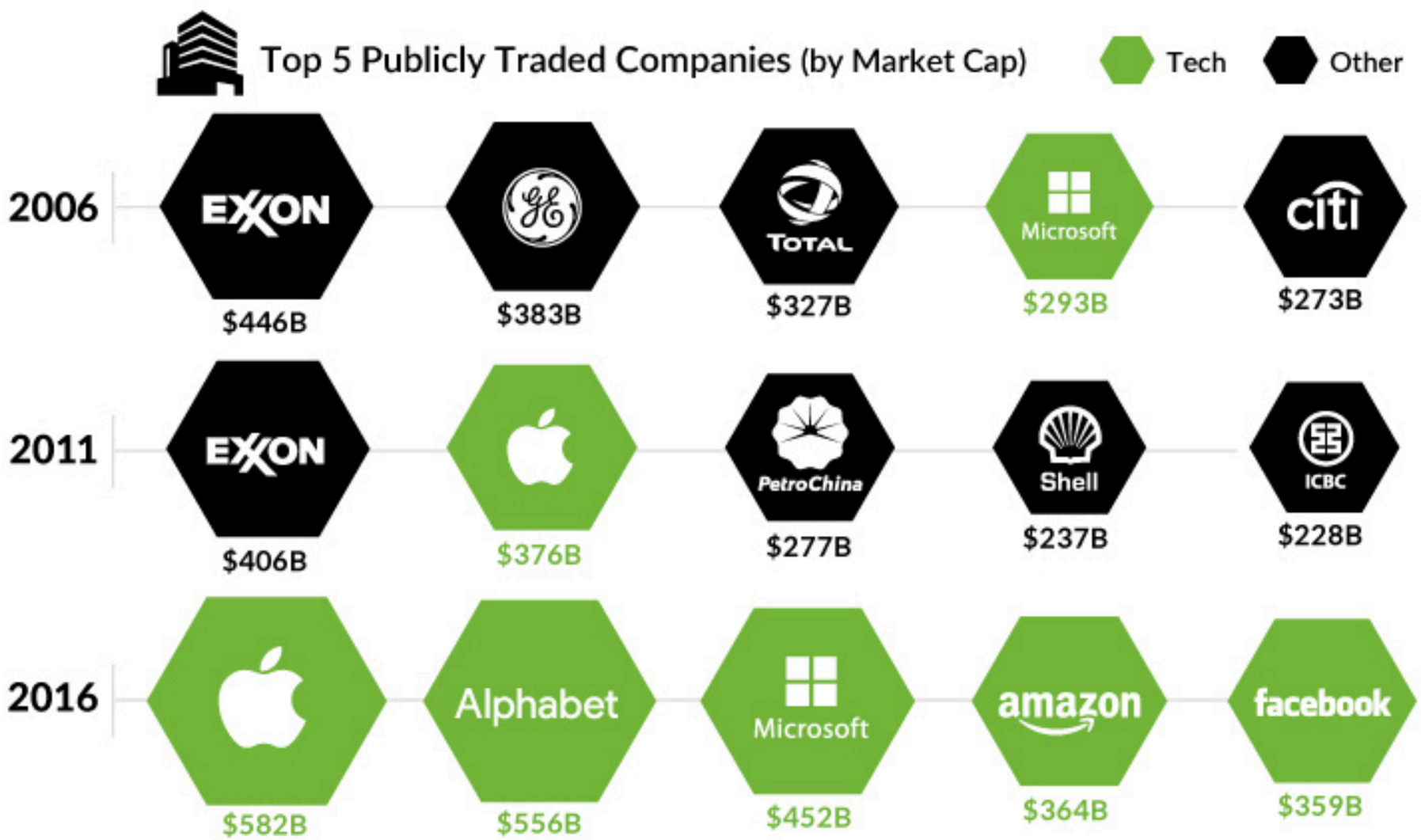
What Happens Now

2017: Every day 2.5 quintillion (10^{18}) Bytes are generated. Every minute:



What Happens Now

The top most valued corporations in the world



What Happens Now

2018: According to GlassDoor.com, the #1 best job in America is:

1 Data Scientist

4.8 / 5
Job Score

4.2 / 5
Job Satisfaction

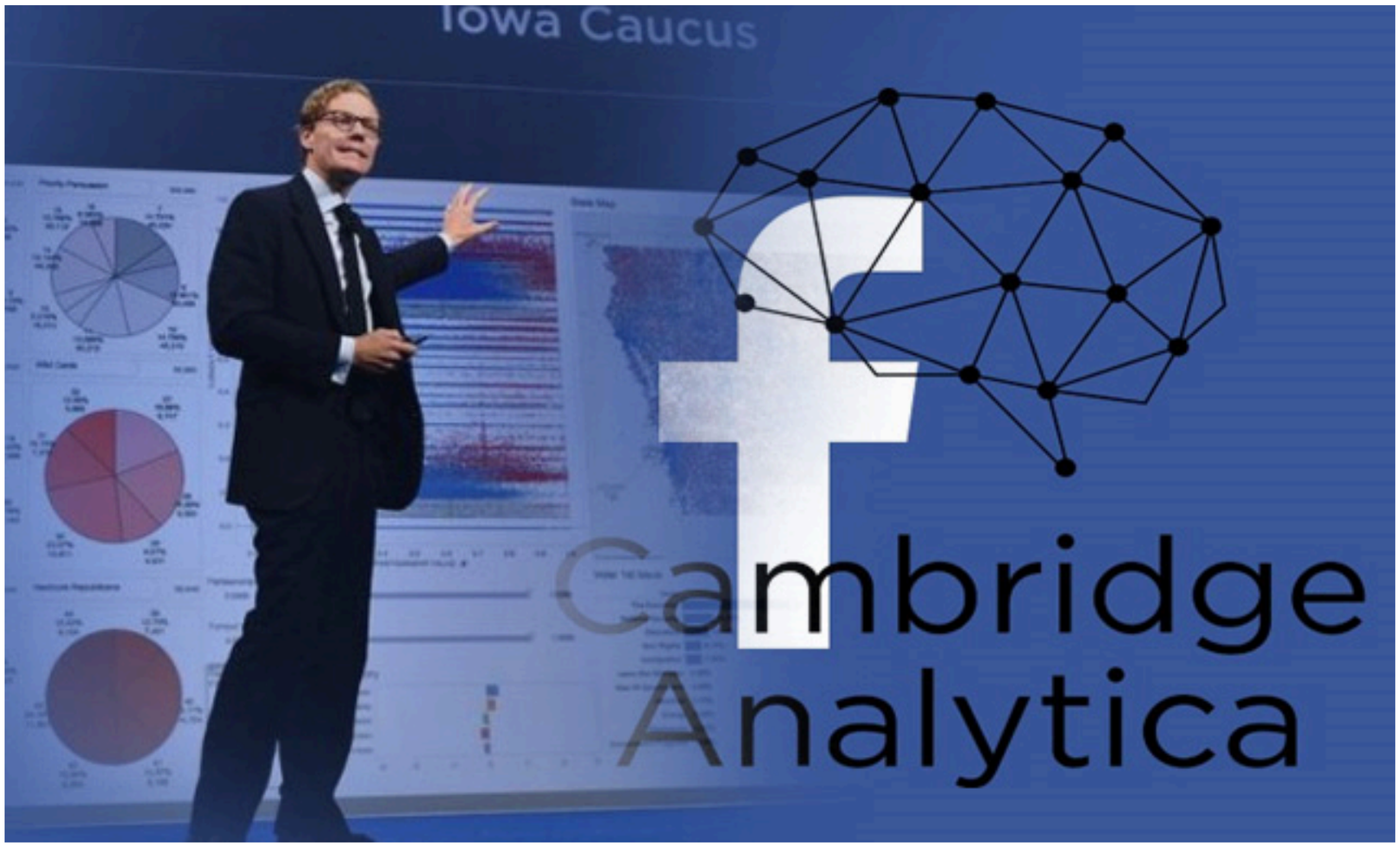
$110,000
Median Base Salary

4,524
Job Openings

View Jobs

The Dark(?) Side of Data

Privacy: US Presidential Election, 2016

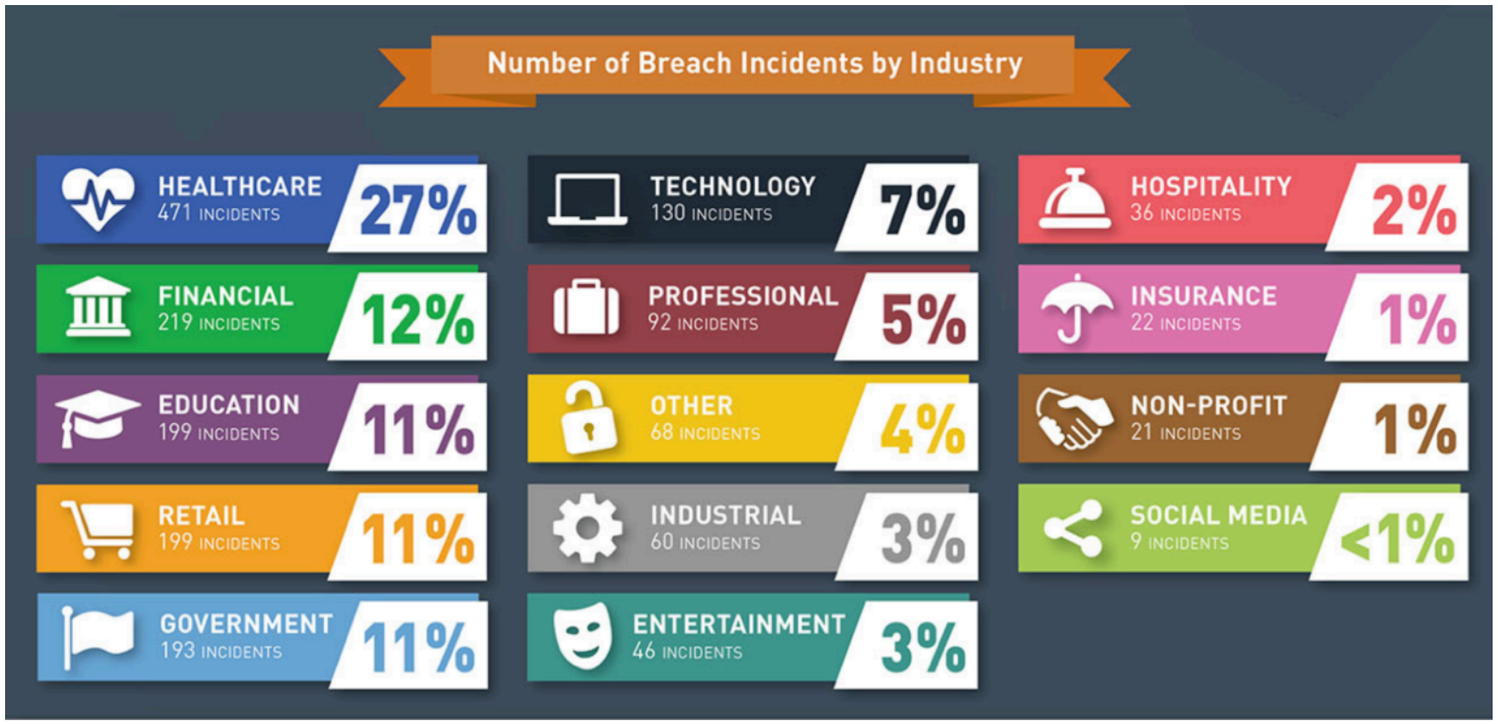


The Dark(?) Side of Data

Algorithmic "bias": LinkedIn, 2016

The Dark(?) Side of Data

Data Breaches: 2.6 Billion data records were stolen in 2017.



EQUIFAX®

Verizon

UBER

What’s Next?

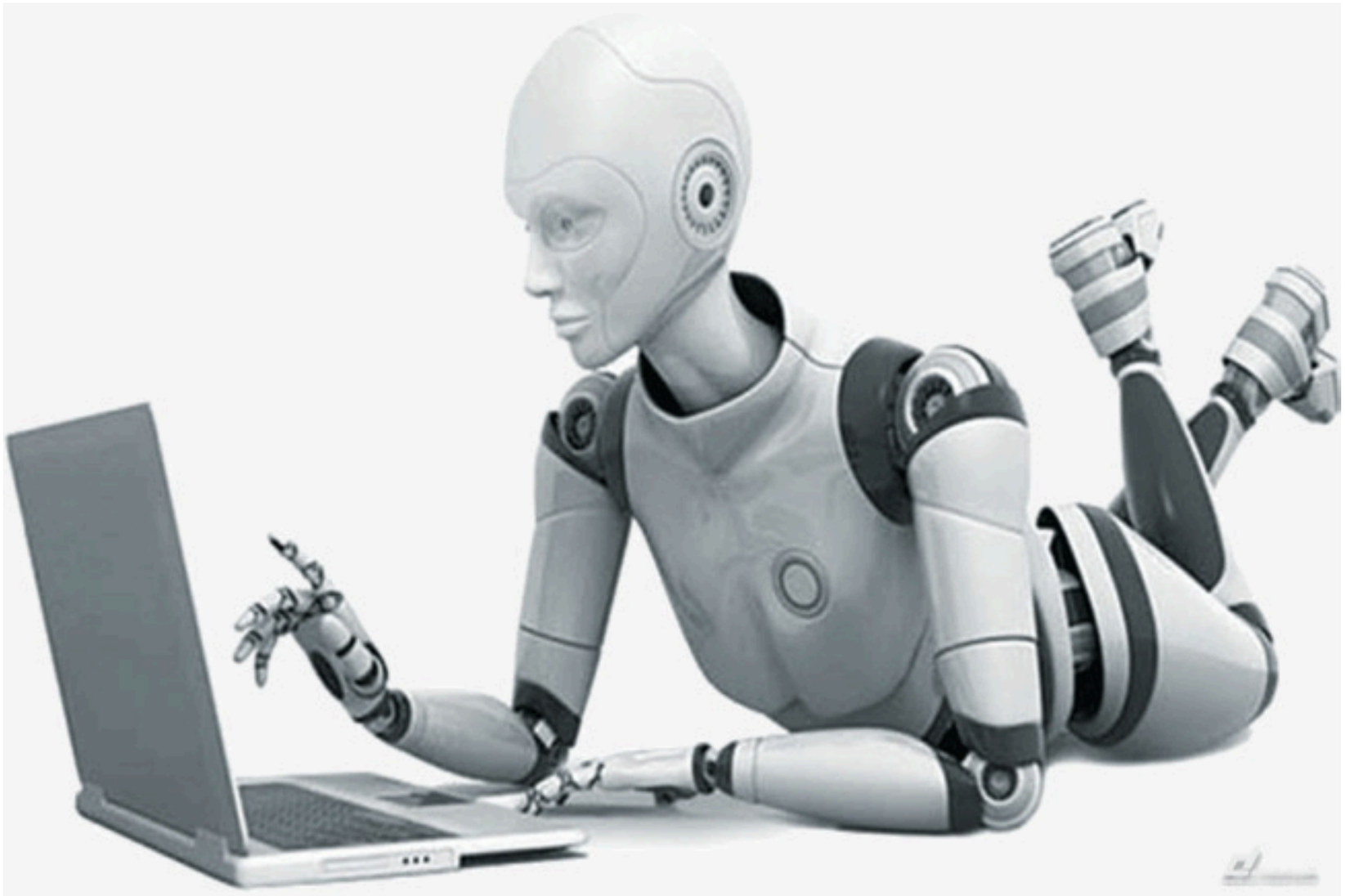
How can we handle the overflow of data?

What’s Next?

Can we “solve” data analysis like Google solved the web search problem?

What’s Next?

Can we teach a machine to analyze and understand data?



Questions