

The Evolution of Human-Data Interaction



Amit Somech

November 2018



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.

		Numeral 1			Bread
		Numeral 10			Wool
		Numeral 600			Sheep



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 every 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”



The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663	1664	1665	1666	1667	1668	1669	1670	1671	1672	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682	1683	1684	1685	1686	1687	1688	1689	1690	1691	1692	1693	1694	1695	1696	1697	1698	1699	1700	1701	1702	1703	1704	1705	1706	1707	1708	1709	1710	1711	1712	1713	1714	1715	1716	1717	1718	1719	1720	1721	1722	1723	1724	1725	1726	1727	1728	1729	1730	1731	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	1751	1752	1753	1754	1755	1756	1757	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771	1772	1773	1774	1775	1776	1777	1778	1779	1780	1781	1782	1783	1784	1785	1786	1787	1788	1789	1790	1791	1792	1793	1794	1795	1796	1797	1798	1799	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835	1836	1837	1838	1839	1840	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858	1859	1860	1861	1862	1863	1864	1865	1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	20100	20101	20102	20103	20104	20105	20106	20107	20108	20109	20110	20111	20112	20113	20114	20115	20116	20117	20118	20119	20120	20121	20122	20123	20124	20125	20126	20127	20128	20129	20130	20131	20132	20133	20134	20135	20136	20137	20138	20139	20140	20141	20142	20143	20144	20145	20146	20147	20148	20149	20150	20151	20152	20153	20154	20155	20156	20157	20158	20159	20160	20161	20162	20163	20164	20165	20166	20167	20168	20169	20170	20171	20172	20173	20174	20175	20176	20177	20178	20179	20180	20181	20182	20183	20184	20185	20186	20187	20188	20189	20190	20191	20192	20193	20194	20195	20196	20197	20198	20199	20200	20201	20202	20203	20204	20205	20206	20207	20208	20209	20210	20211	20212	20213	20214	20215	20216	20217	20218	20219	20220	20221	20222	20223	20224	20225	20226	20227	20228	20229	20230	20231	20232	20233	20234	20235	20236	20237	20238	20239	20240	20241	20242	20243	20244	20245	20246	20247	20248	20249	20250	20251	20252	20253	20254	20255	20256	20257	20258	20259	20260	20261	20262	20263	20264	20265	20266	20267	20268	20269	20270	20271	20272	20273	20274	20275	20276	20277	20278	20279	20280	20281	20282	20283	20284	20285	20286	20287	20288	20289	20290	20291	20292	20293	20294	20295	20296	20297	20298	20299	20300	20301	20302	20303	20304	20305	20306	20307	20308	20309	20310	20311	20312	20313	20314	20315	20316	20317	20318	20319	20320	20321	20322	20323	20324	20325	20326	20327	20328	20329	20330	20331	20332	20333	20334	20335	20336	20337	20338	20339	20340	20341	20342	20343	20344	20345	20346	20347	20348	20349	20350	20351	20352	20353	20354	20355	20356	20357	20358	20359	20360	20361	20362	20363	20364	20365	20366	20367	20368	20369	20370	20371	20372	20373	20374	20375	20376	20377	20378	20379	20380	20381	20382	20383	20384	20385	20386	20387	20388	20389	20390	20391	20392	20393	20394	20395	20396	20397	20398	20399	20400	20401	20402	20403	20404	20405	20406	20407	20408	20409	20410	20411	20412	20413	20414	20415	20416	20417	20418	20419	20420	20421	20422	20423	20424	20425	20426	20427	20428	20429	20430	20431	20432	20433	20434	20435	20436	20437	20438	20439	20440	20441	20442	20443	20444	20445	20446	20447	20448	20449	20450	20451	20452	20453	20454	20455	20456	20457	20458	20459	20460	20461	20462	20463	20464	20465	20466	20467	20468	20469	20470	20471	20472	20473	20474	20475	20476	20477	20478	20479	20480	20481	20482	20483	20484	20485	20486	20487	20488	20489	20490	20491	20492	20493	20494	20495	20496	20497	20498	20499	20500	20501	20502	20503	20504	20505	20506	20507	20508	20509	20510	20511	20512	20513	20514	20515	20516	20517	20518	20519	20520	20521</th

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—I TO IO PERSONS.

Supervisor's District No.	[7-556 b.]	Eleventh Census of the United States.				
Enumeration District No.		SCHEDULE No. 1.				
POPULATION AND SOCIAL STATISTICS.						
Name of city, town, township, precinct, district, boat, or other minor civil division.		County :	State :			
Street and No.:		Ward :	Name of Institution :	1890		
Enumerated by me on the _____ day of June, 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.						
2 Surname.						
3 Whether a soldier, sailor, or marine during the civil war (U. S. or Conf.), or widow of such person.						
4 Relationship to head of family.						
5 Whether white, black, mulatto, quadroon, octoroon, Chinese, Japanese, or Indian.						
6 Sex.						
7 Age at nearest birthday. If under one year, give age in months.						

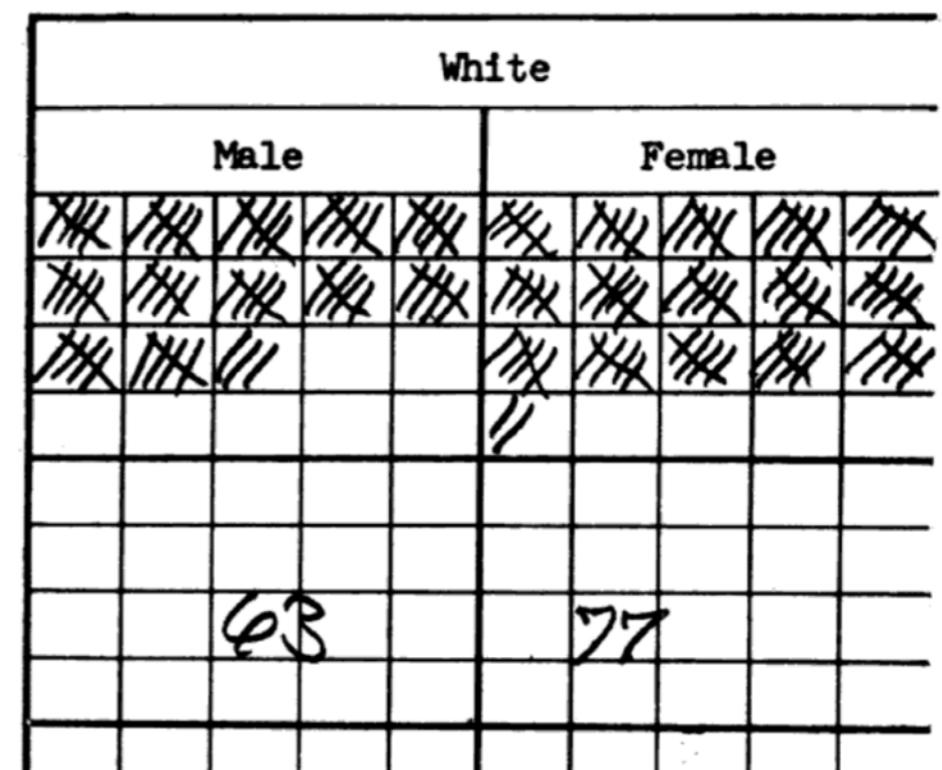


FIGURE 3.—Section of tally sheet, full size, with tally marks

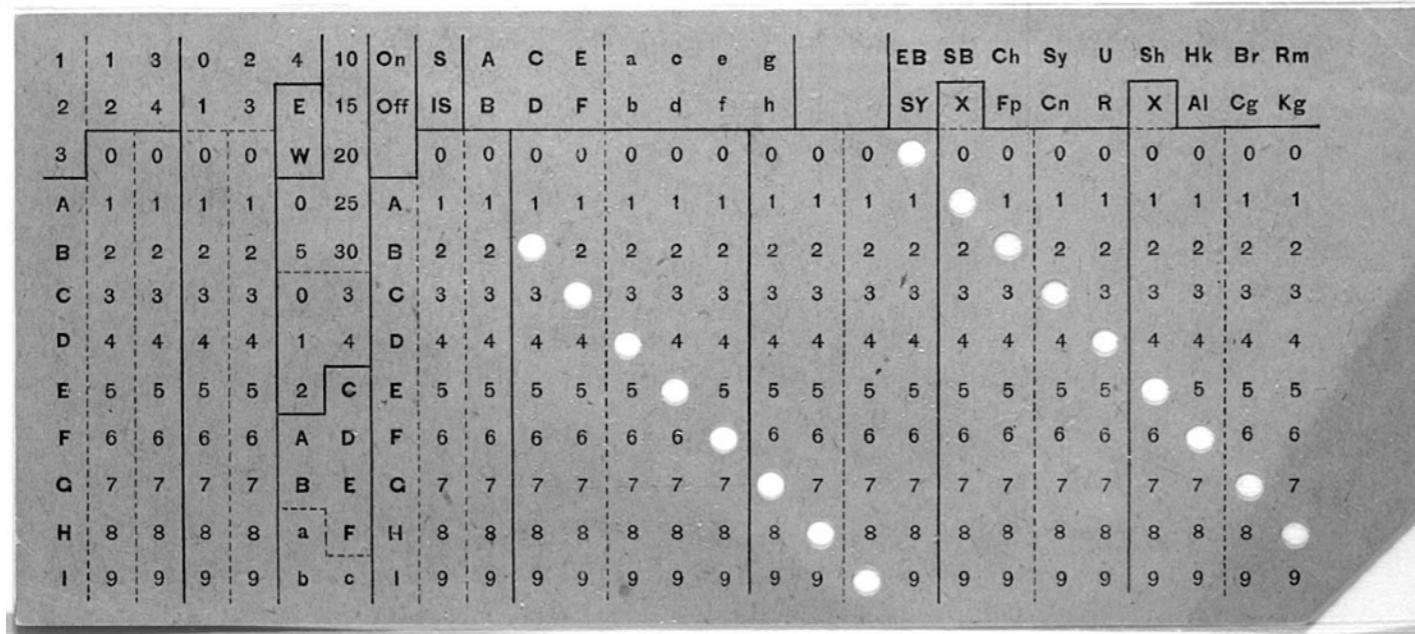
US Census form (for one family)

Manual “Count”

Big Data in the 19th century

1889: Herman Hollerith invents the “Tabulating machine”.

He suggests using *machine-readable* “punched cards”

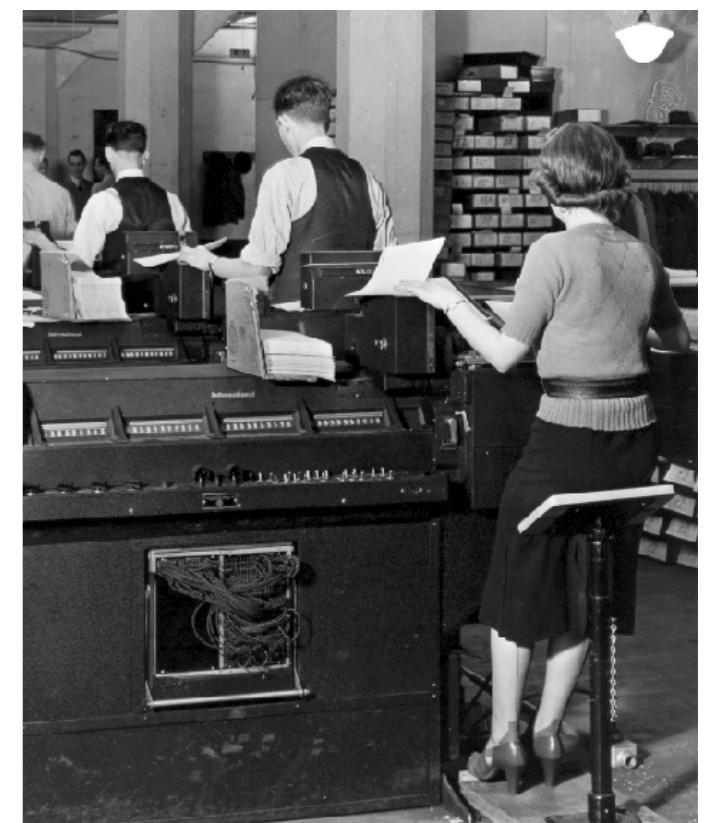
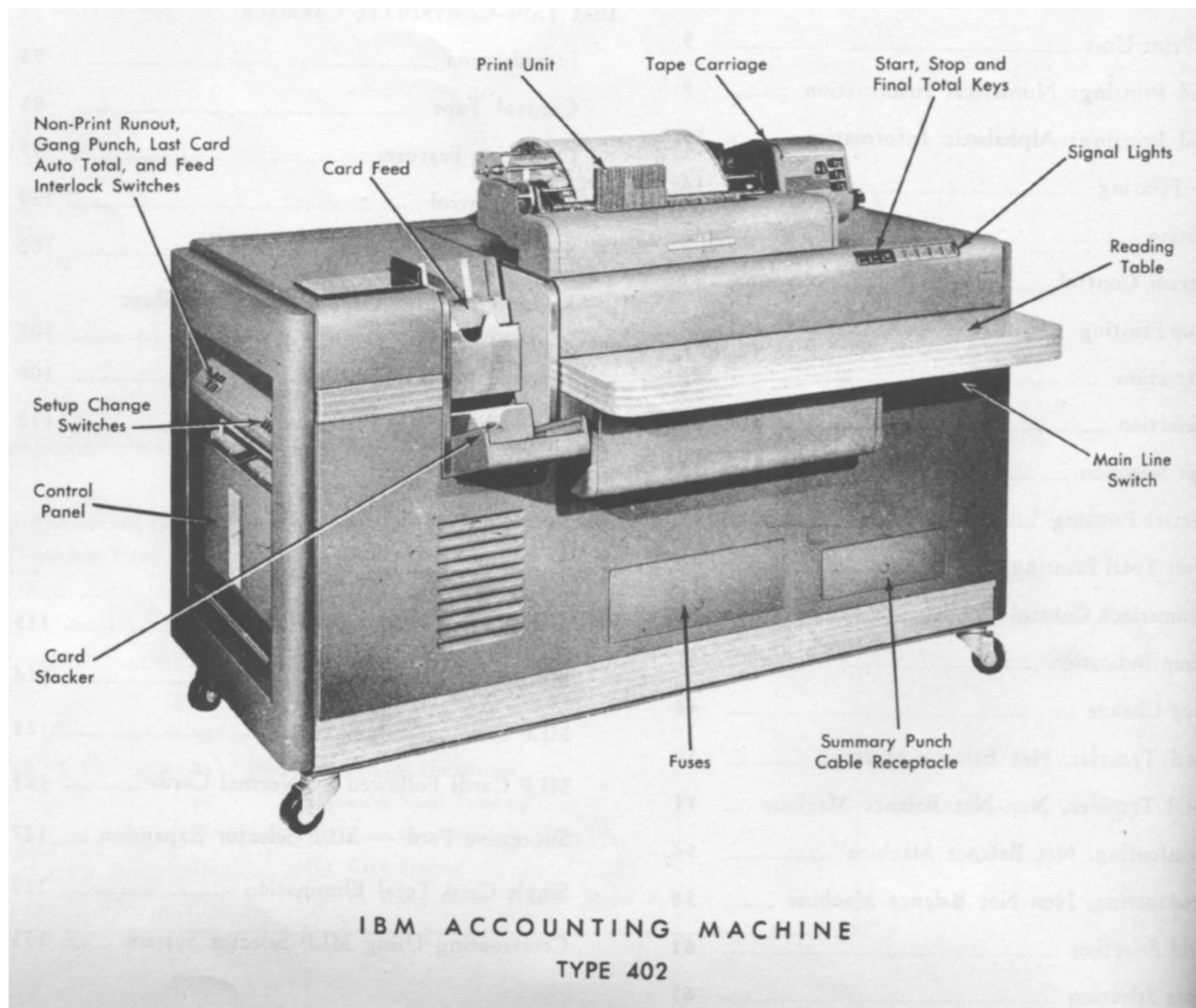


punched card representing one form



Big Data in the 19th century

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



IBM Type 285

Magnetic Storage

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



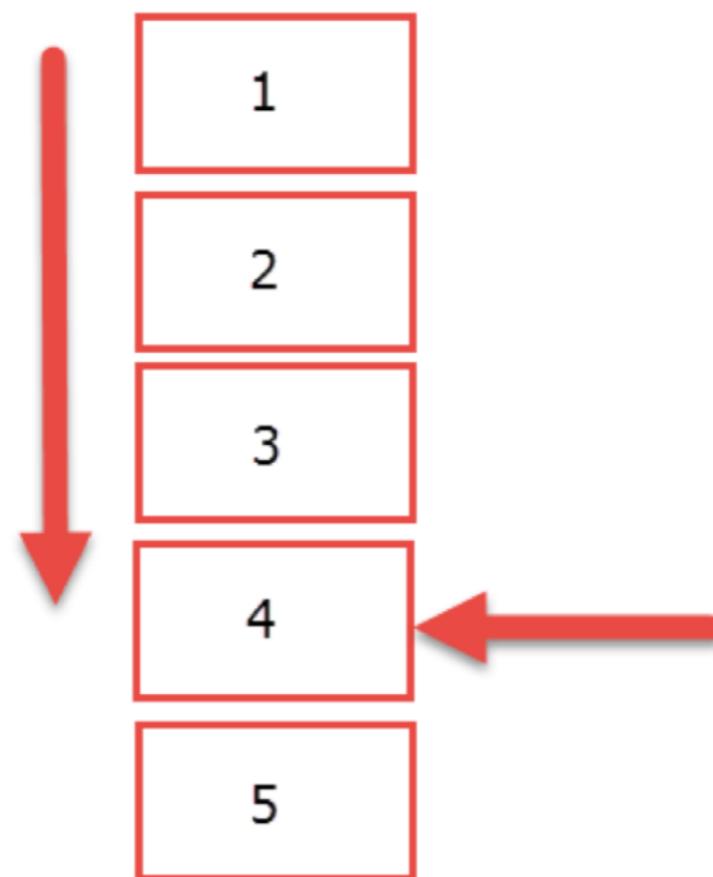
IBM 704 Mainframe



Magnetic Storage

Sequential Access Vs. Direct Access:

With **sequential access**, elements #1, 2, 3 must be processed before element #4 can be processed.



With **direct access**, element #4 in the list can be accessed without having to process the elements before it.

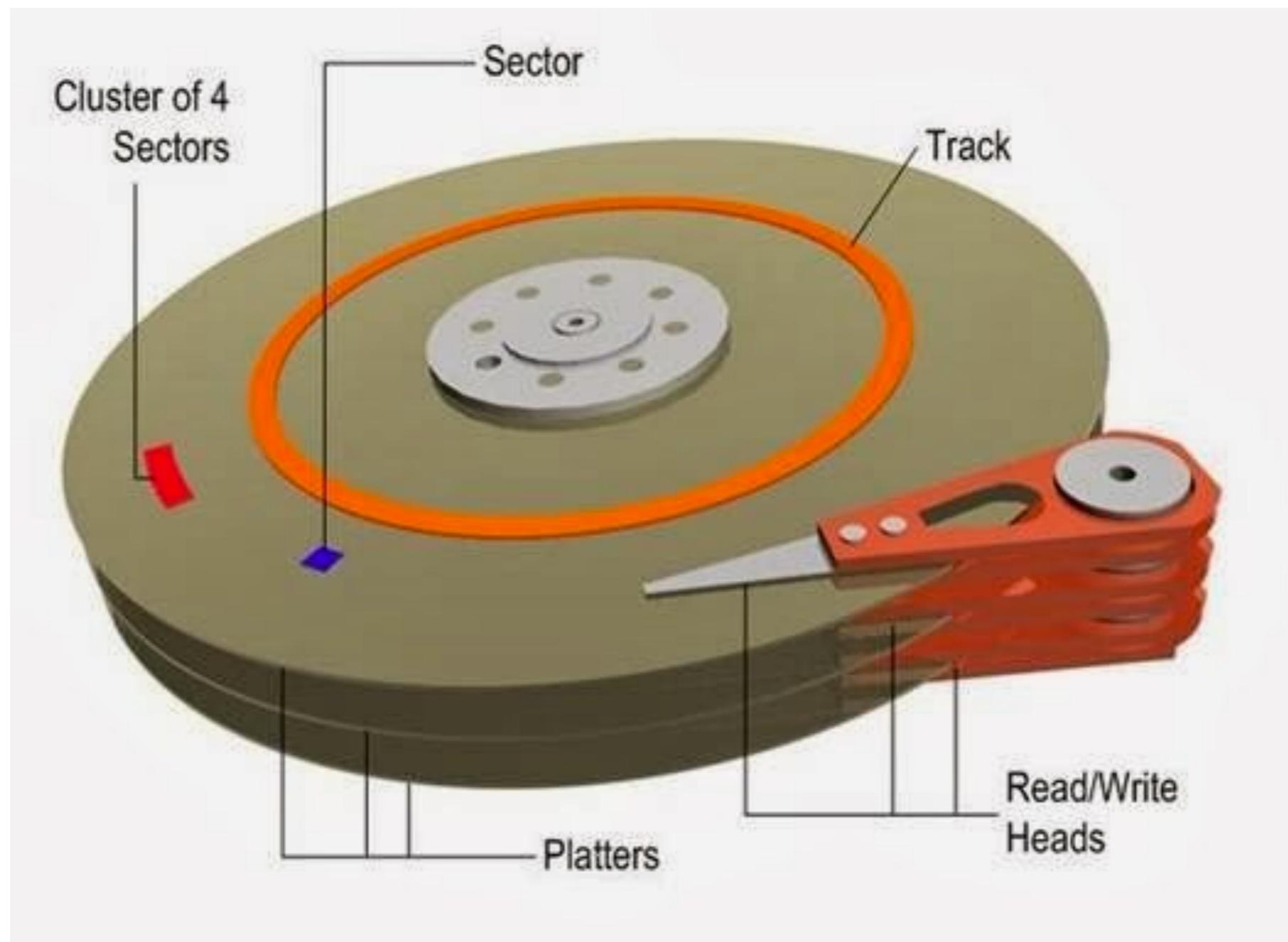
Magnetic Storage

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



Magnetic Storage

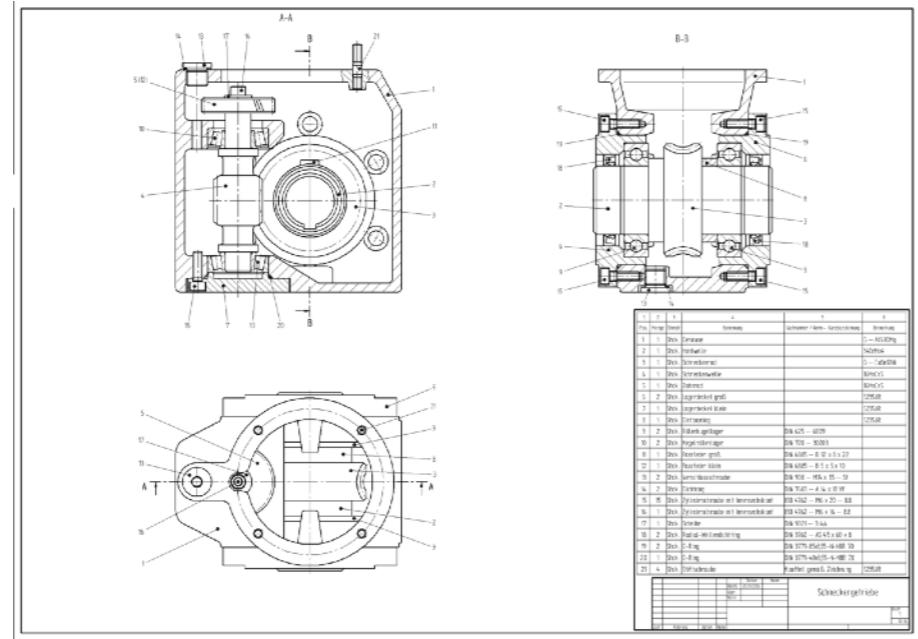
Hard Disk Schema:



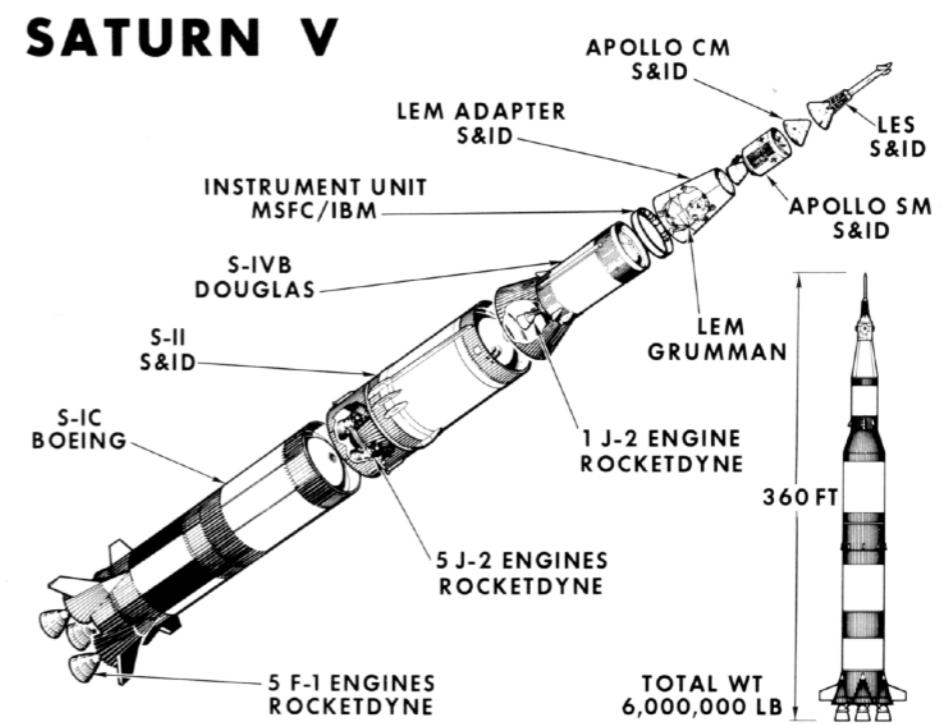
Databases and the Moon

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.

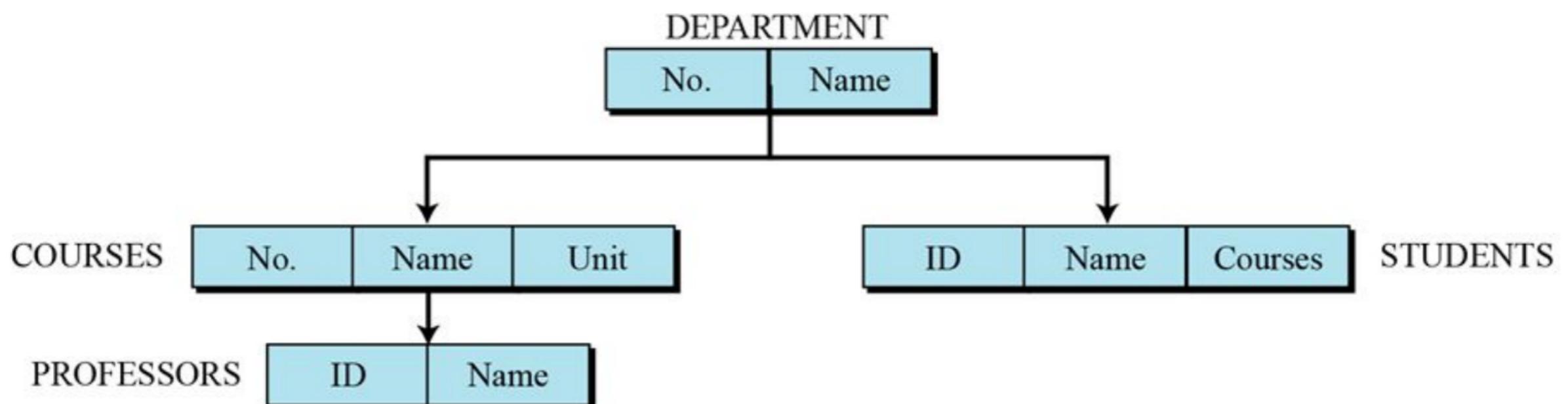


BOM Example



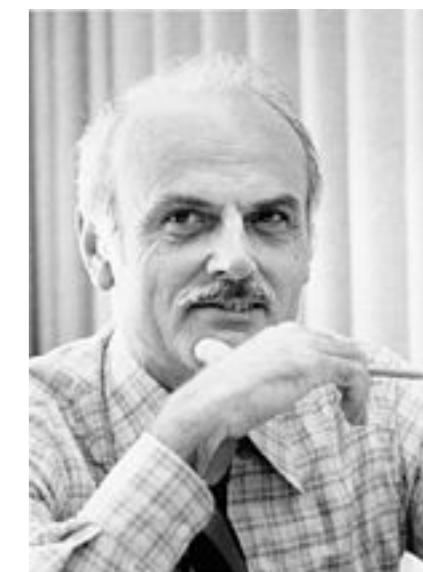
Databases and the Moon

IMS is a “Hierarchical Database”:



General Purpose Databases

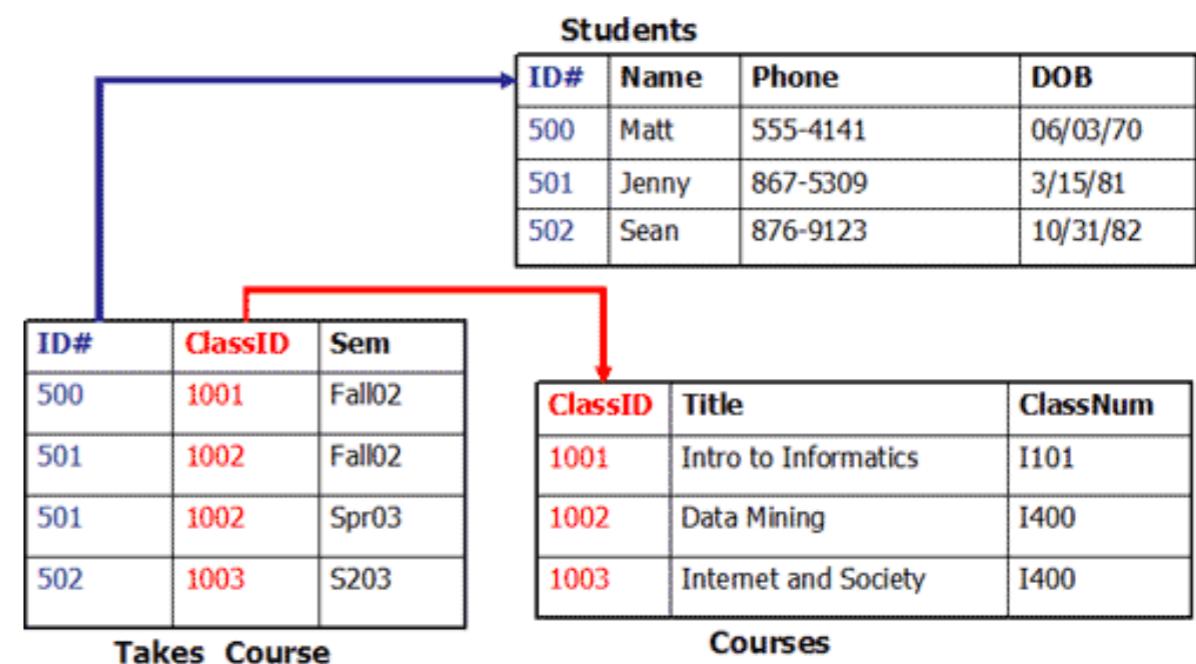
1970: “Ted” Codd (IBM) invents the Relational Database.



1971: IBM researchers develops the SQL “query language”.

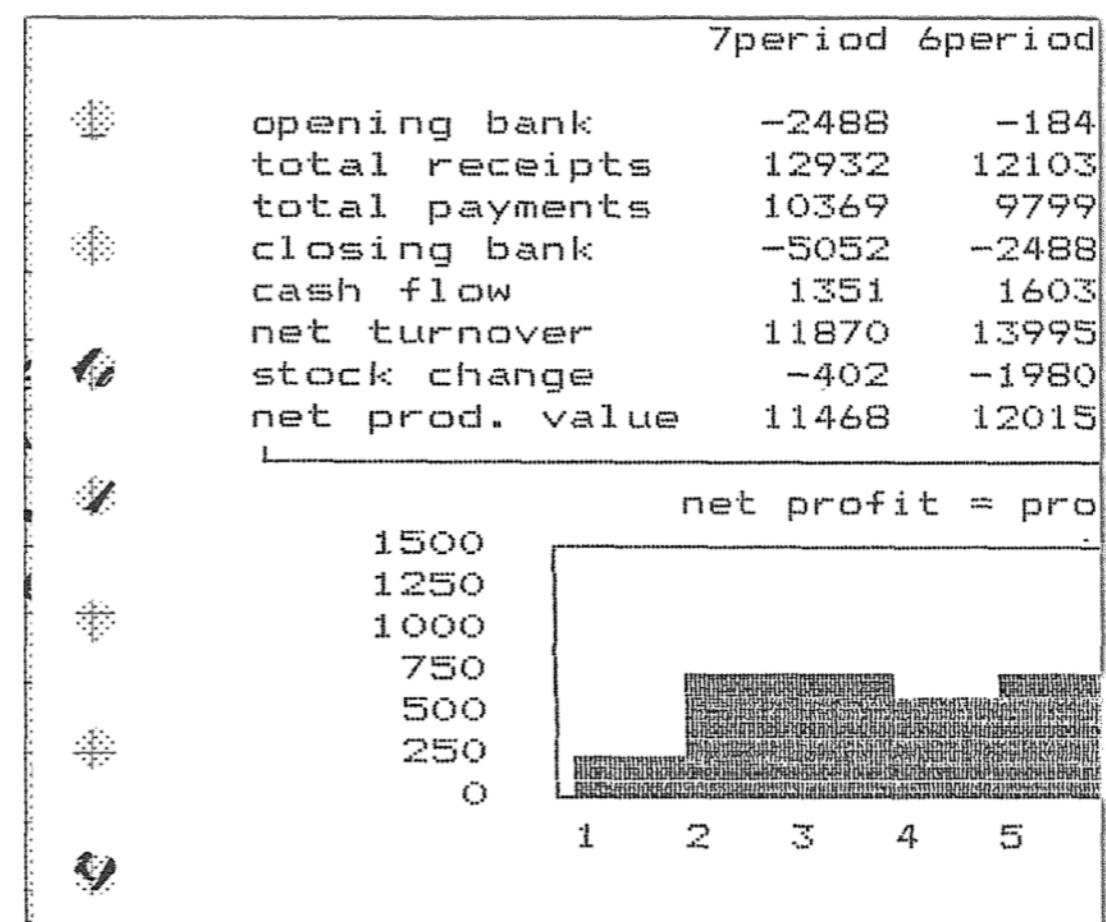
```
SELECT phone  
FROM students  
WHERE dob > "1993-01-01"
```

SQL Query



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					
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



Lotus 3.0 for MS DOS

The Internet

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



YOUR PC

OPEN FILE
RETRIEVE DOCUMENT
PRINT DOCUMENT
CLOSE FILE

YOUR PC WITH A MOTOROLA MODEM

OPEN FILE
PLAY ELECTRONIC GOLF
CHANGE JOE'S SCORE
READ COMICS IN LA TIMES
FINISH CROSSWORD PUZZLE IN NY TIMES
ORDER CASE OF CHIA PETS
E-MAIL EMILY
TRANSFER FUNDS
BUY STOCKS
ATTEMPT TO CORNER SILVER MARKET
RENEW LIBRARY BOOKS
PULL UP PROPOSAL FROM OFFICE

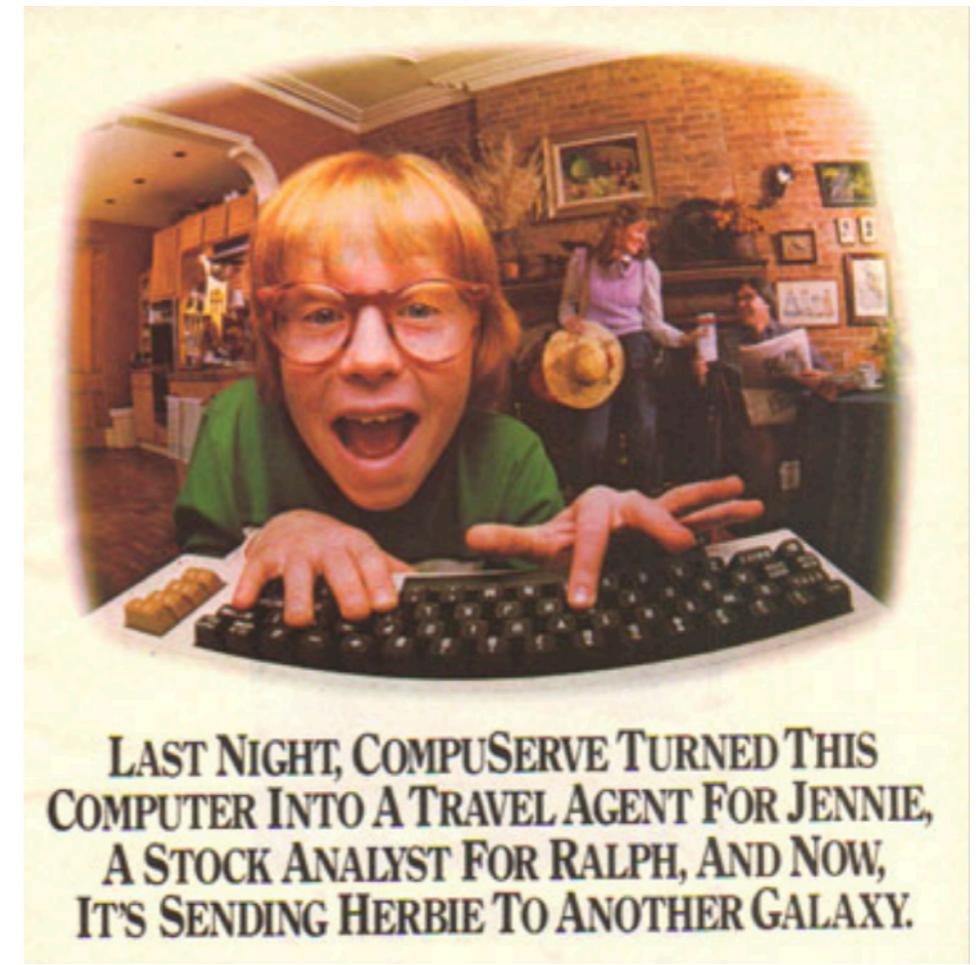
A collage of three images: a dark computer setup, a light-colored computer setup, and a colorful abstract painting.

YOUR PC

YOUR PC WITH A MOTOROLA MODEM

OPEN FILE
RETRIEVE DOCUMENT
PRINT DOCUMENT
CLOSE FILE

OPEN FILE
PLAY ELECTRONIC GOLF
CHANGE JOE'S SCORE
READ COMICS IN LA TIMES
FINISH CROSSWORD PUZZLE IN NY TIMES
ORDER CASE OF CHIA PETS
E-MAIL EMILY
TRANSFER FUNDS
BUY STOCKS
ATTEMPT TO CORNER SILVER MARKET
RENEW LIBRARY BOOKS
PULL UP PROPOSAL FROM OFFICE



Google

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



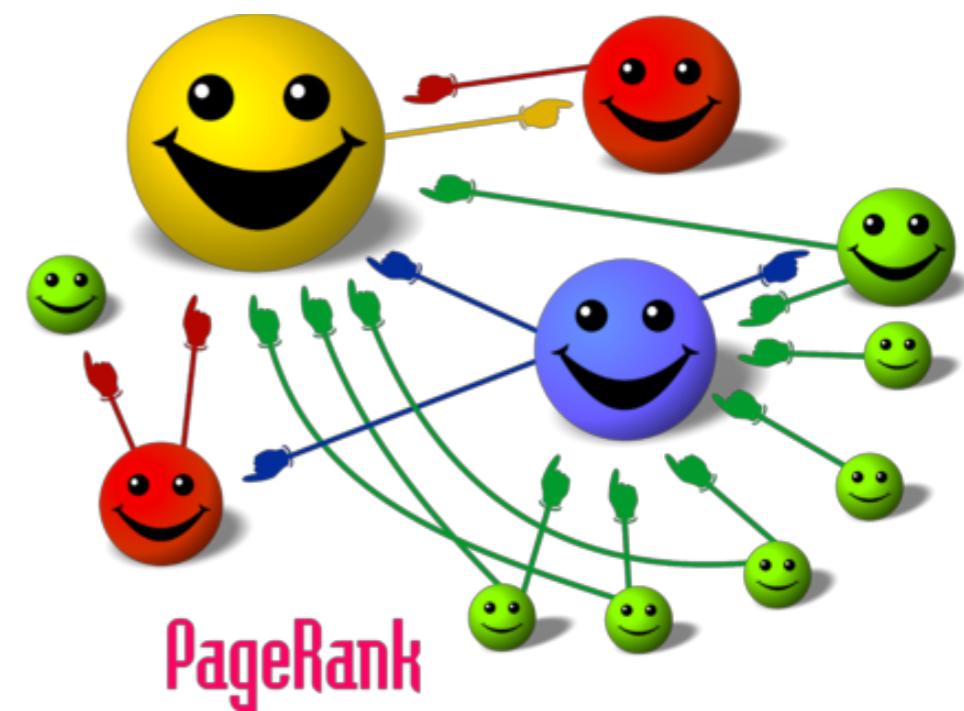
Copyright ©1998 Google Inc.

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:



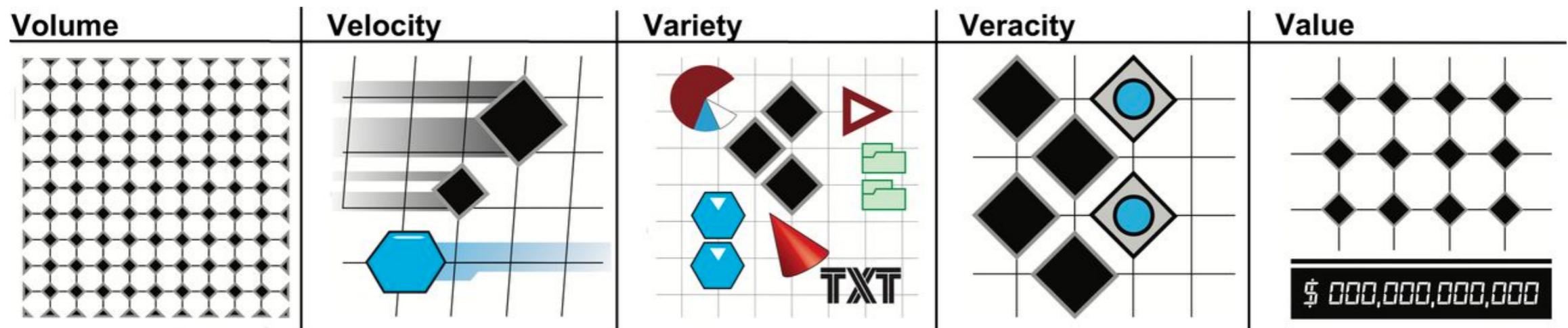
olympics	
olympics 2008	85,400,000 results
olympics schedule	880,000 results
olympics tv schedule	21,400,000 results
olympics 2012	738,000 results
olympics live	38,000,000 results
olympics medal count	312,000 results
olympics 2016	1,080,000 results
olympics basketball	3,540,000 results
olympics gymnastics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

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.

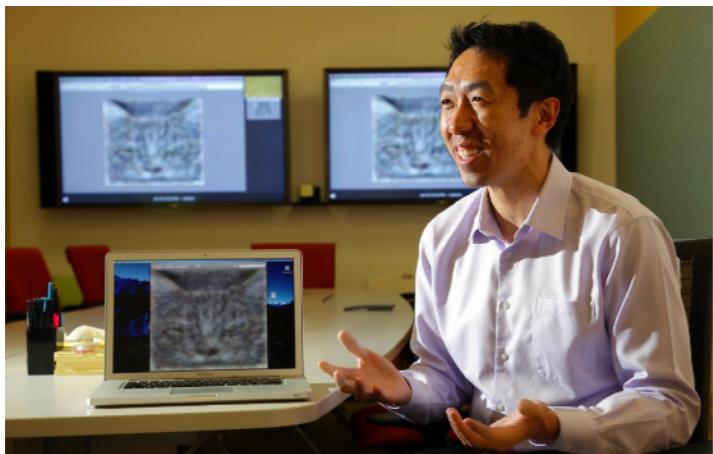
Also, “**NoSQL**” databases are becoming popular, e.g. graph DB, document DB.



Deep-Learning Revolution

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

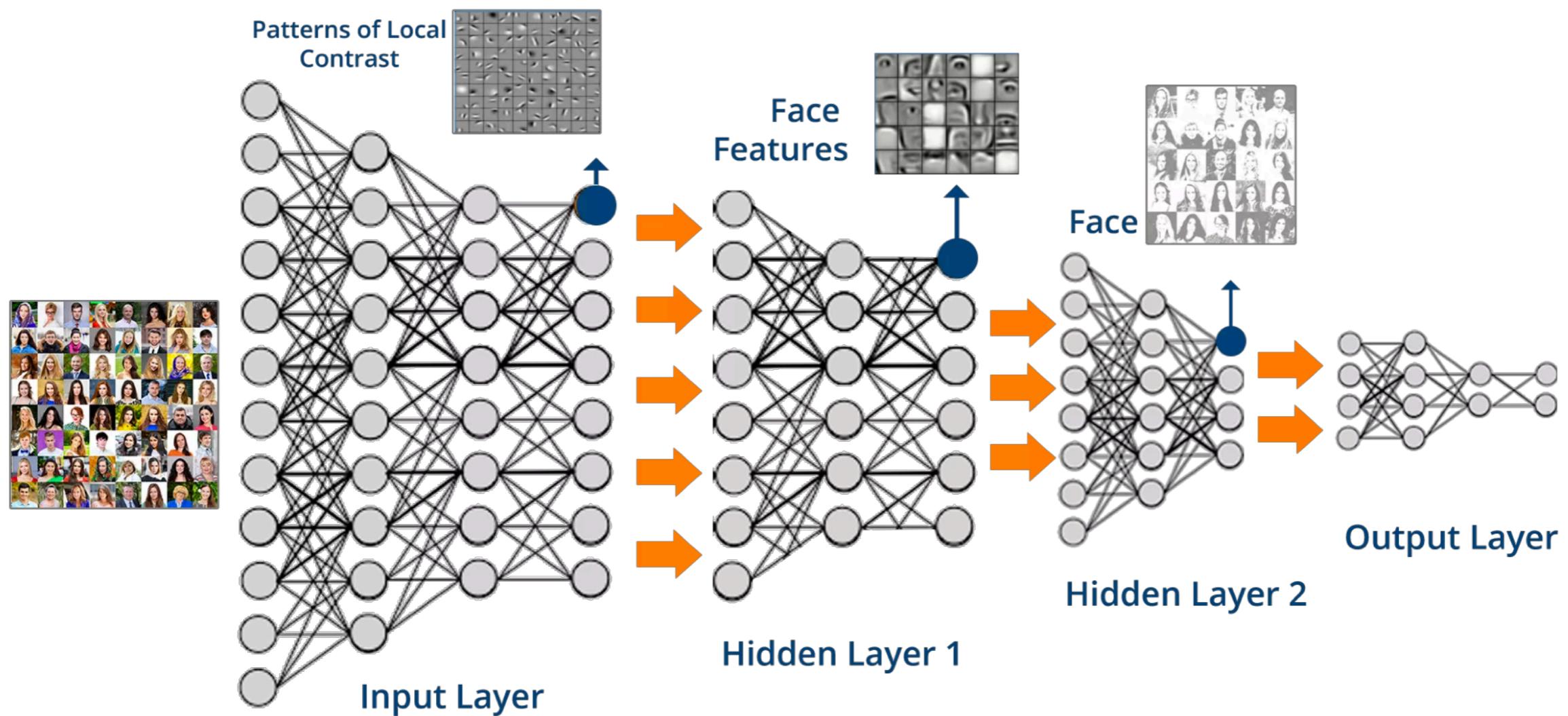
Their brain-like software successfully identified cats (without knowing what a cat is).



"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

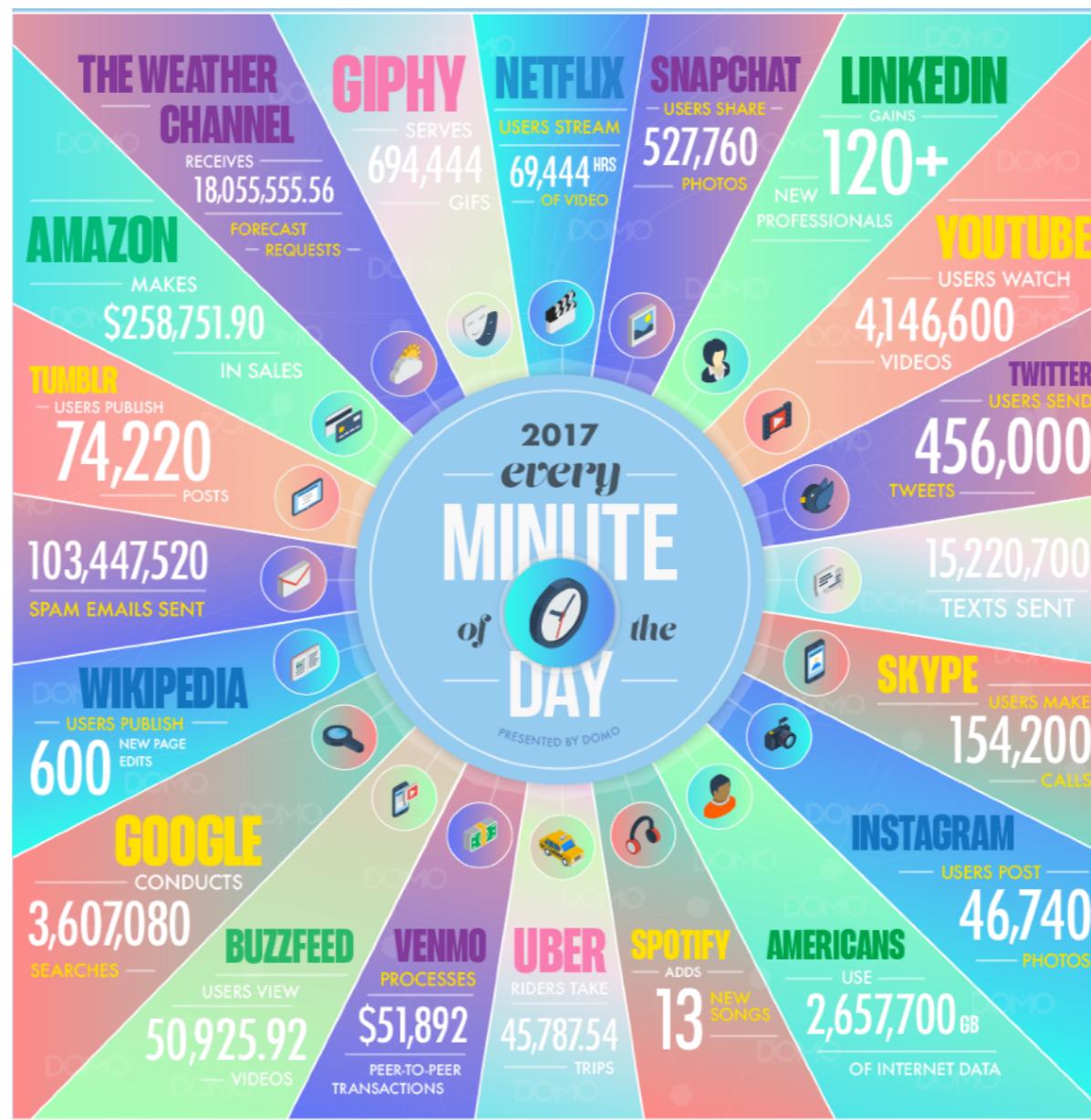
Deep-Learning Revolution

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

\$110,000
Median Base Salary

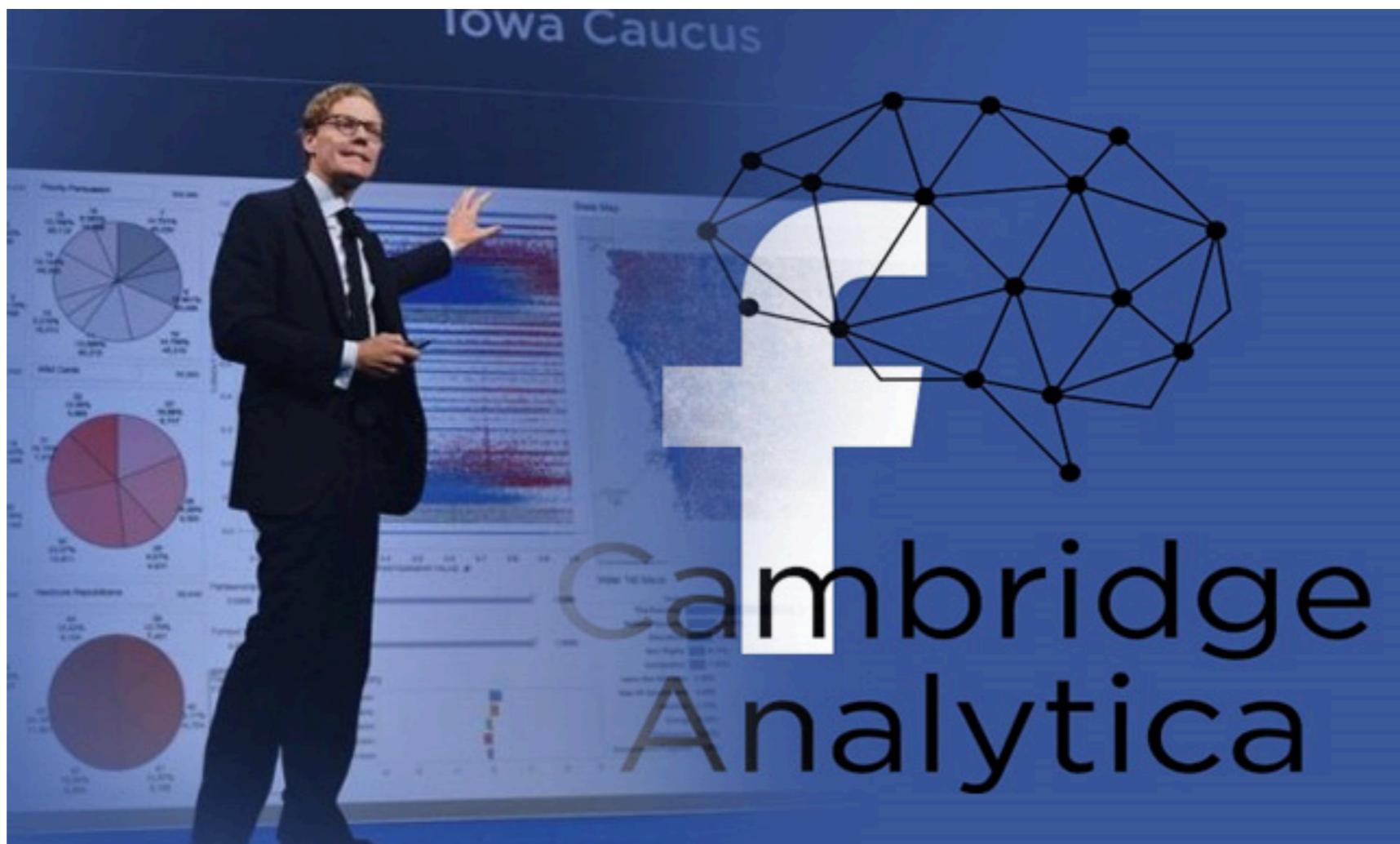
4.2 / 5
Job Satisfaction

4,524
Job Openings

[View Jobs](#)

The Dark(?) Side of Data

Privacy: US Presidential Election, 2016



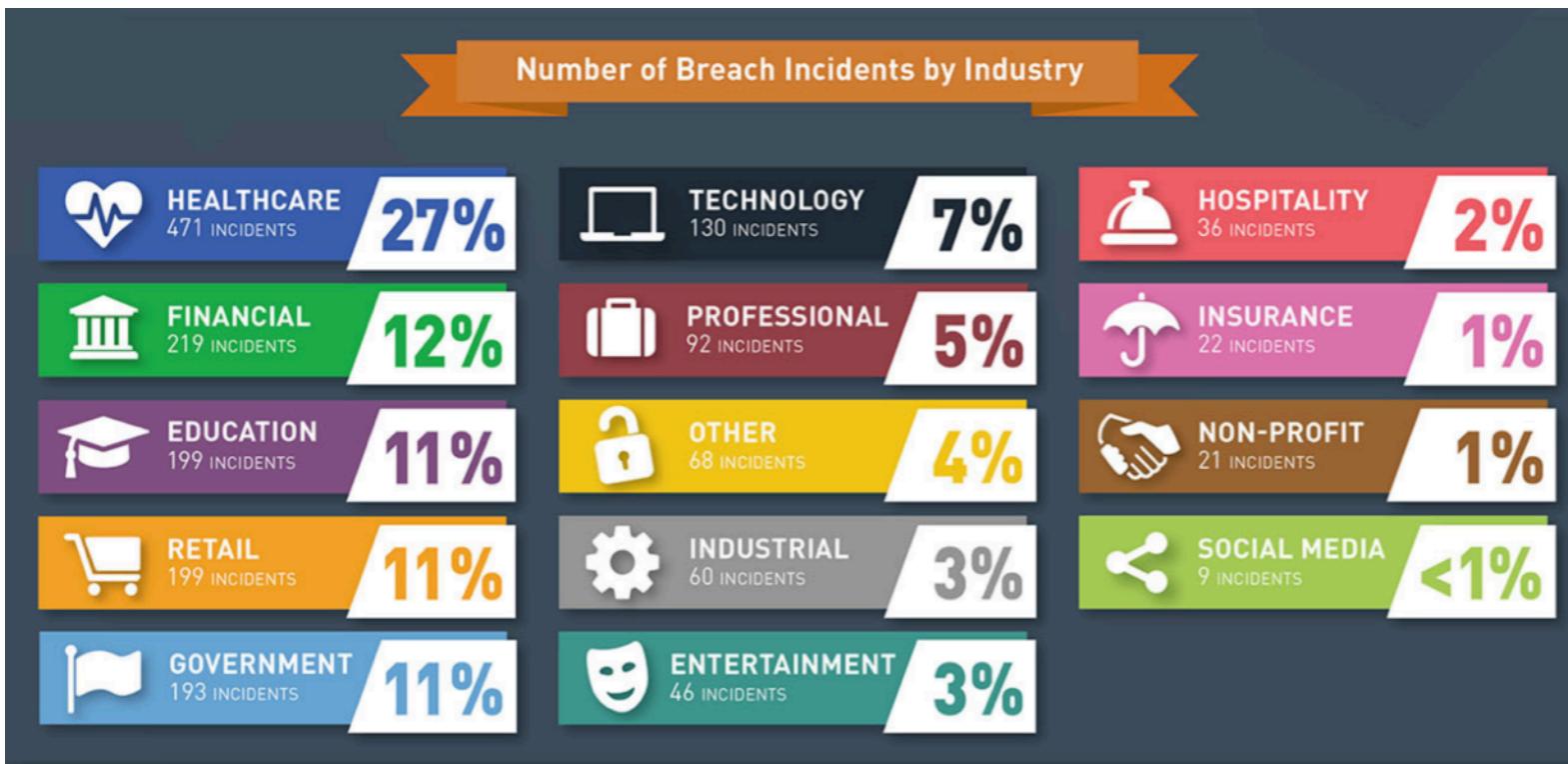
The Dark(?) Side of Data

Algorithmic “bias”: LinkedIn, 2016



The Dark(?) Side of Data

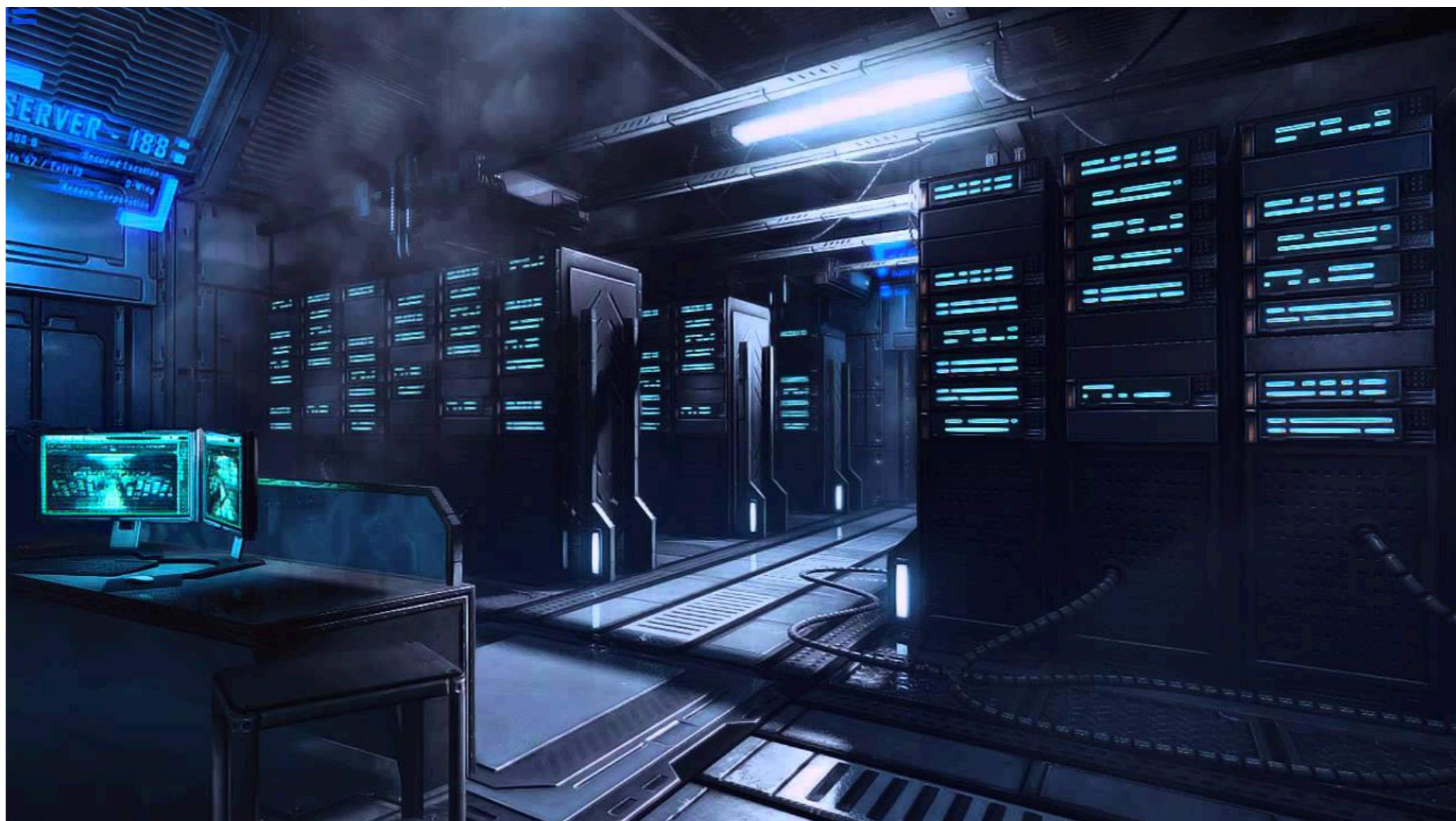
Data Breaches: 2.6 Billion data records were stolen in 2017.



U B E R

What's Next?

How can we handle the overflow of data?



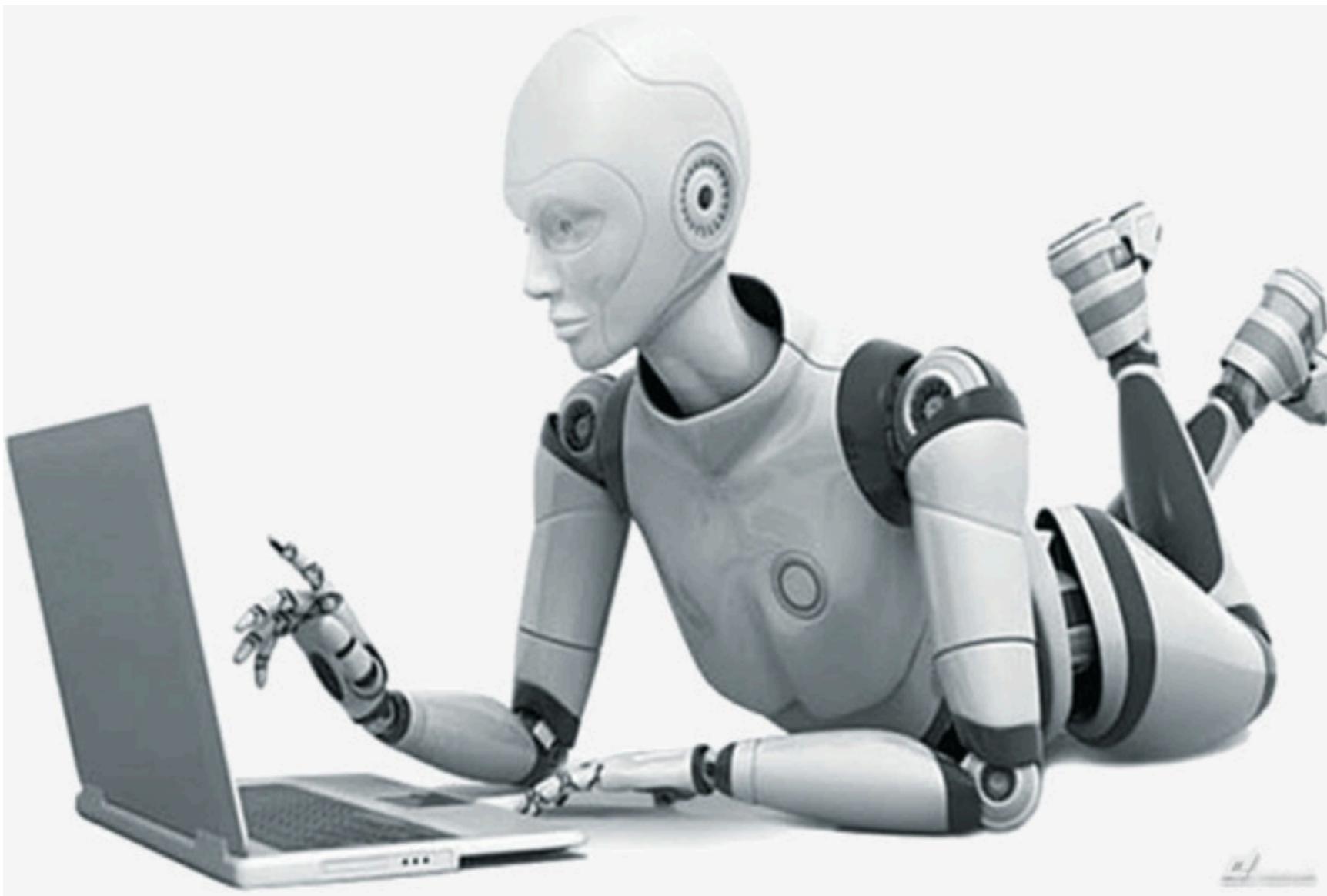
What's Next?

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

