M1a: In the beginning

DSE 10200: Introduction to Data Science

Instructor: Michael Grossberg

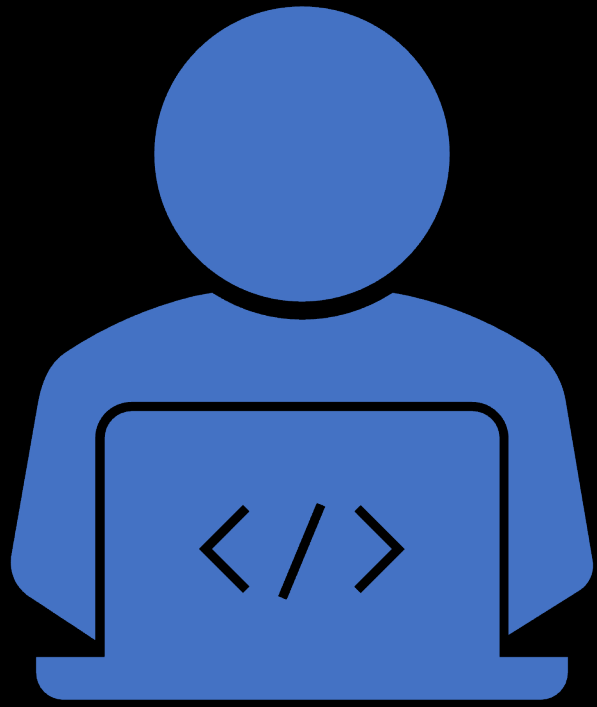
What’s up with “Data Science”?

Isn't it just statistics

computer science

ai

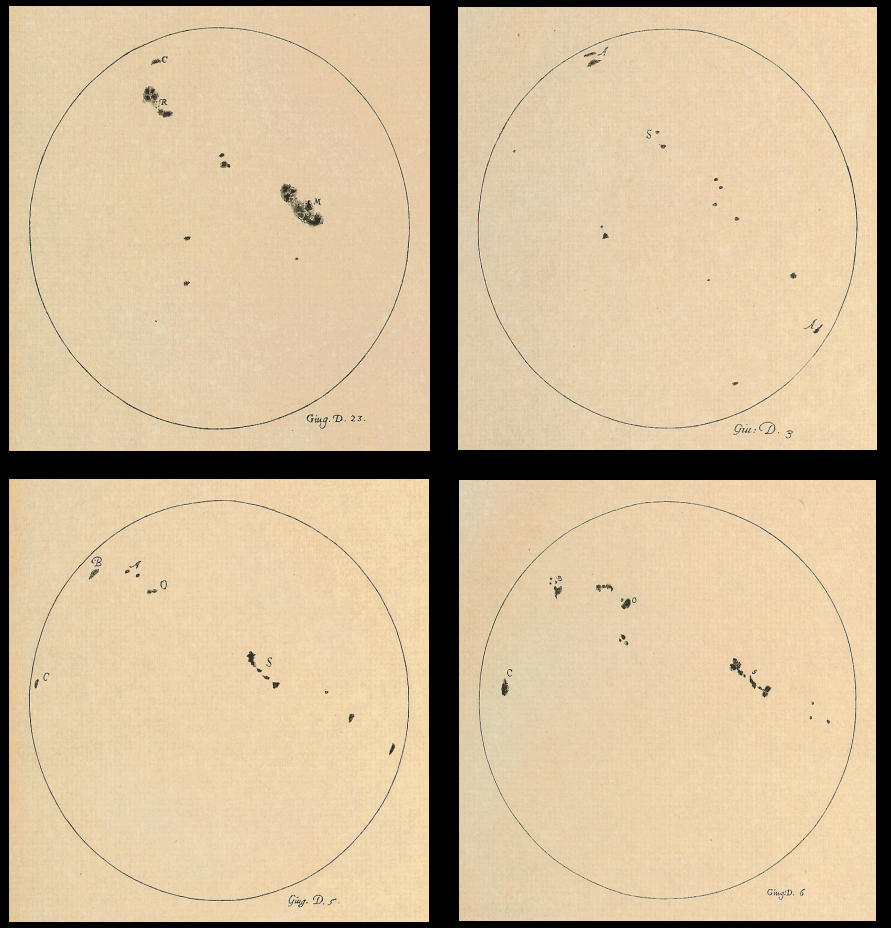
science



Throughout Most of Human History

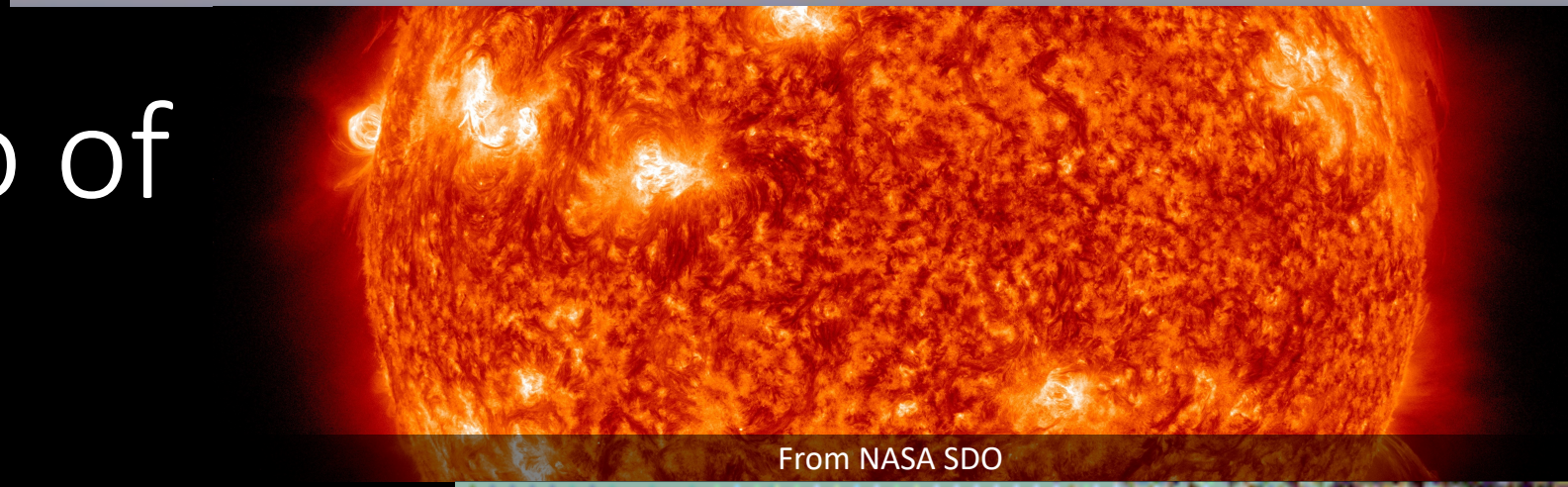
Data difficult to collect and store

Galileo's Hand-drawn Sunspots

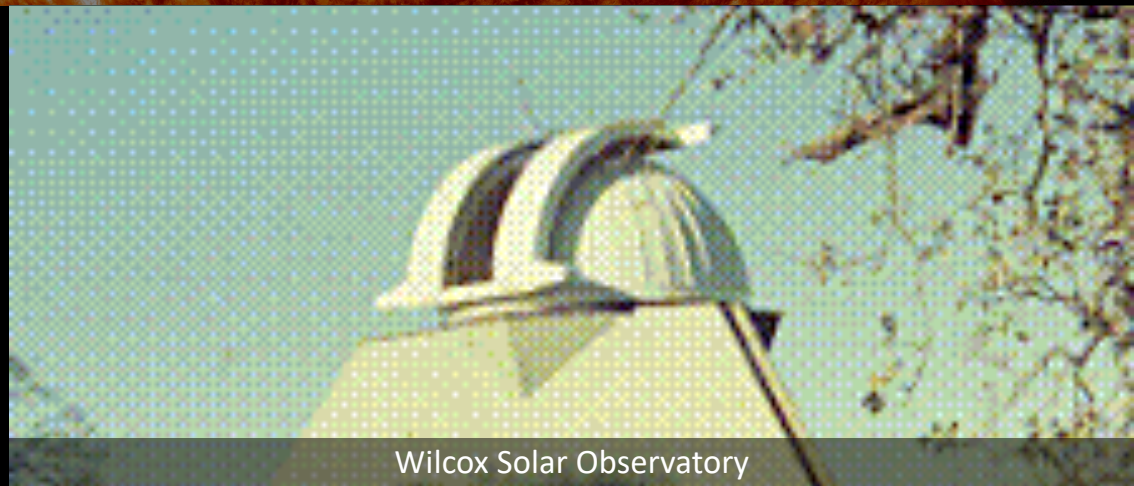


vs.

UltraHD video of the Sun 24/7

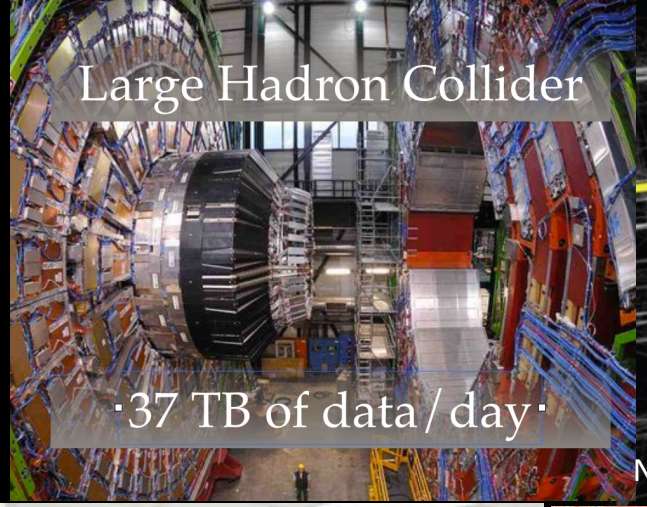


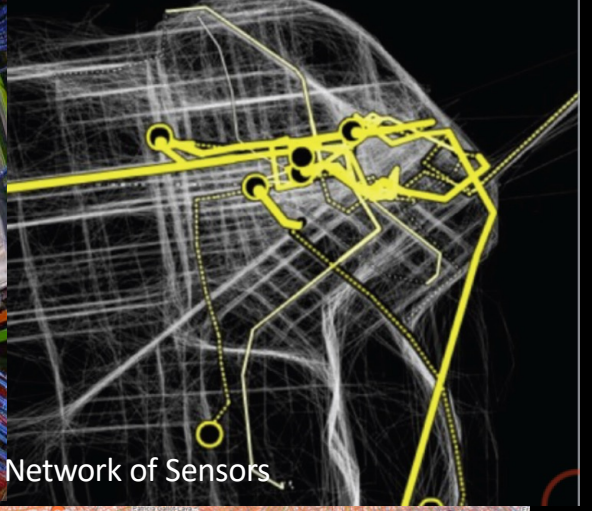
From NASA SDO

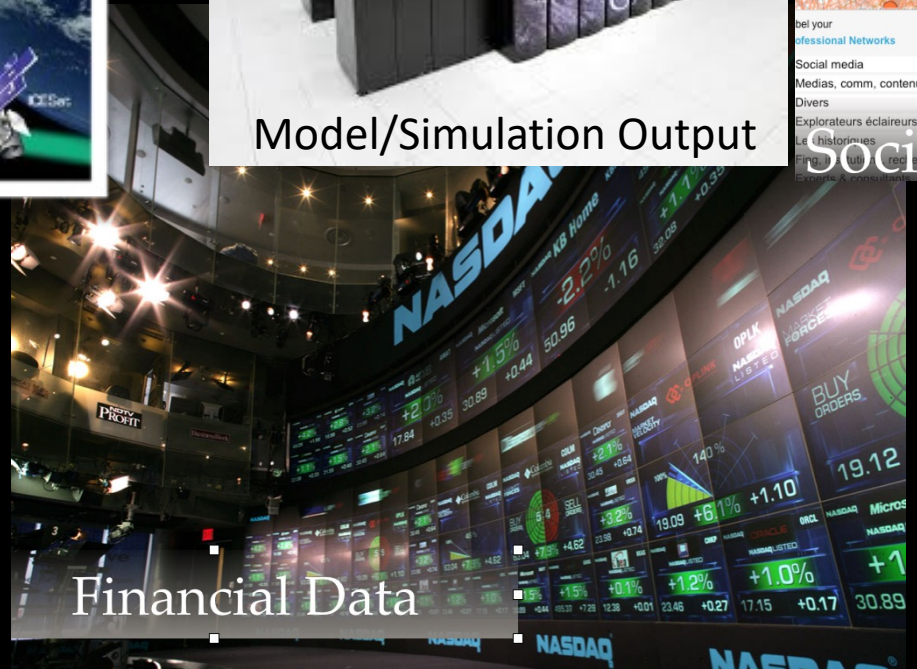


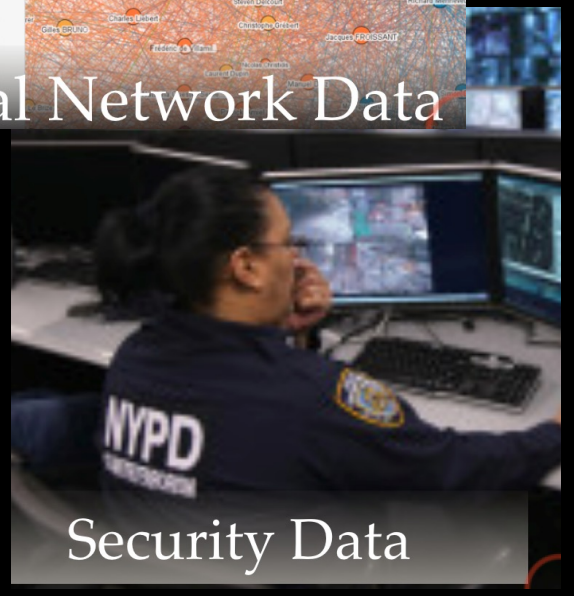
Wilcox Solar Observatory

Flood of Data









Every two days now we create as much information as we did from the dawn of civilization up until 2003

- Eric Schmidt, former CEO Google, 2010

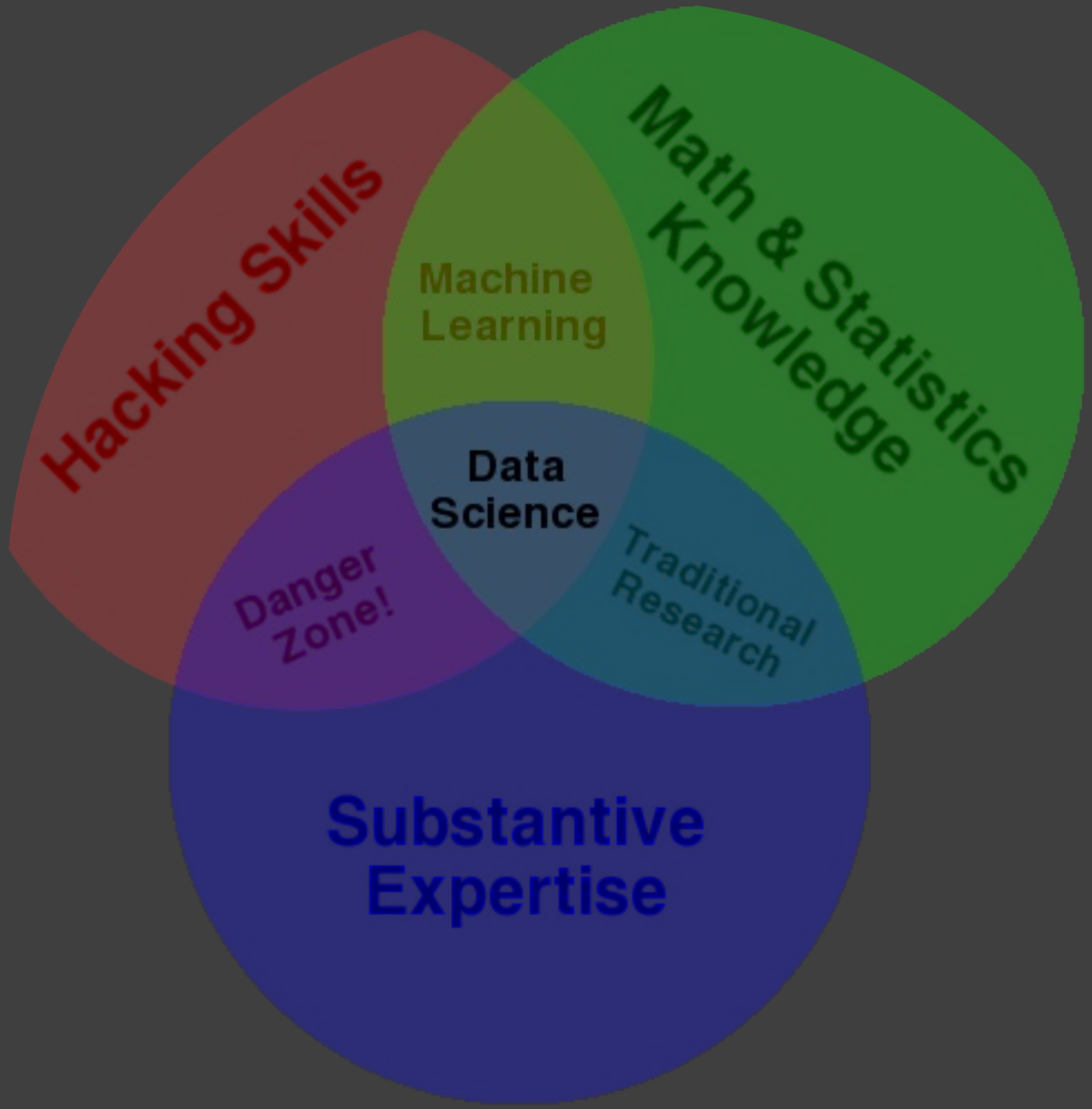
Data + Computation

+

Is Data Science a new thing?

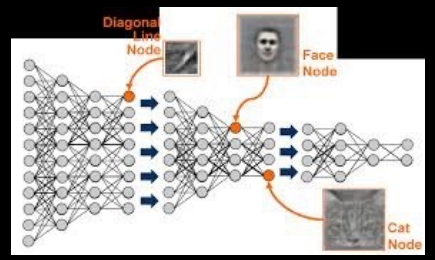
If things can combine to make something new

What is data science?



Coping with Data

Algorithms

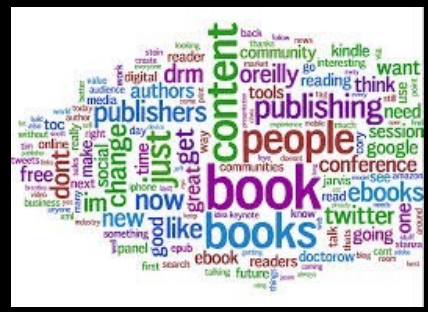


+

Computation

+

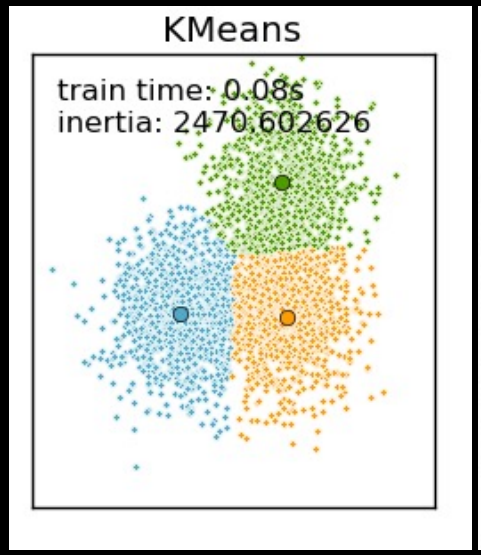
Visualization



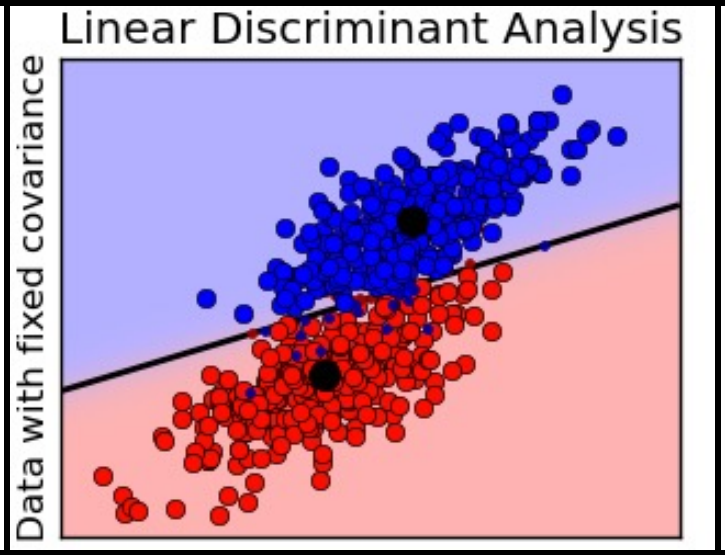
Core Subject Mater

grossberg@cs.ccny.cuny.edu

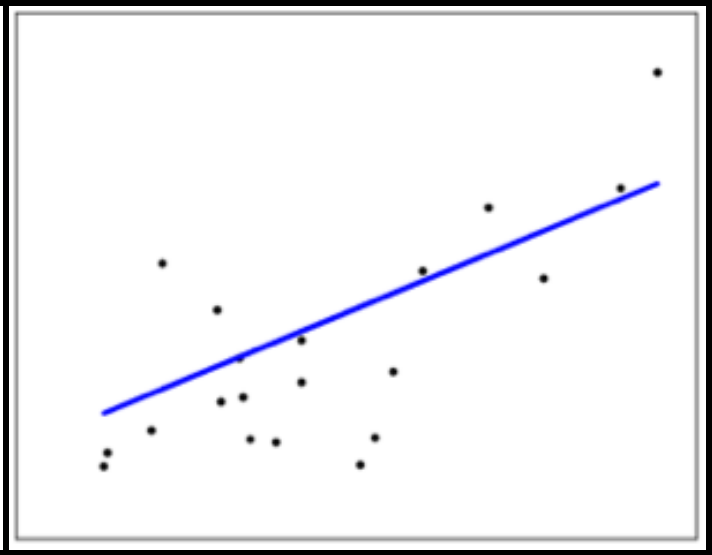
Machine Learning



Clustering Problem



Classification Problem



Regression Problem

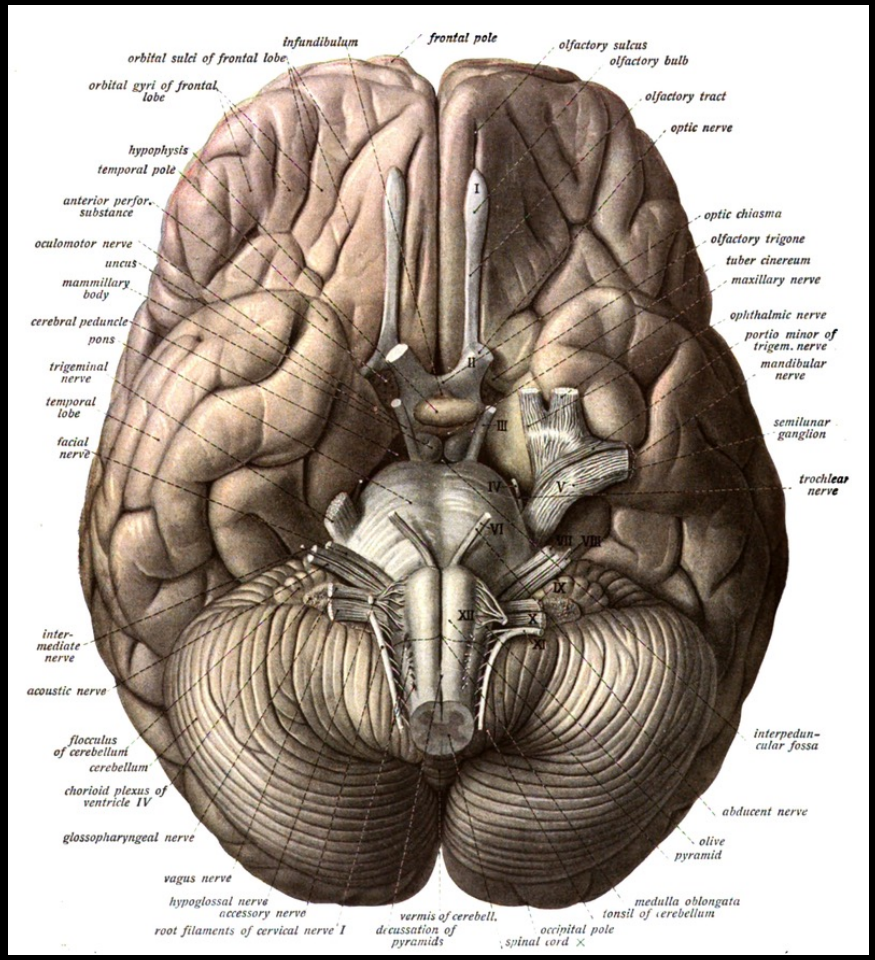
Goal: Explore structure of data and make predictions

Machine Learning in Engineering

Data Visualization

grossberg@cs.ccny.cuny.edu

Information

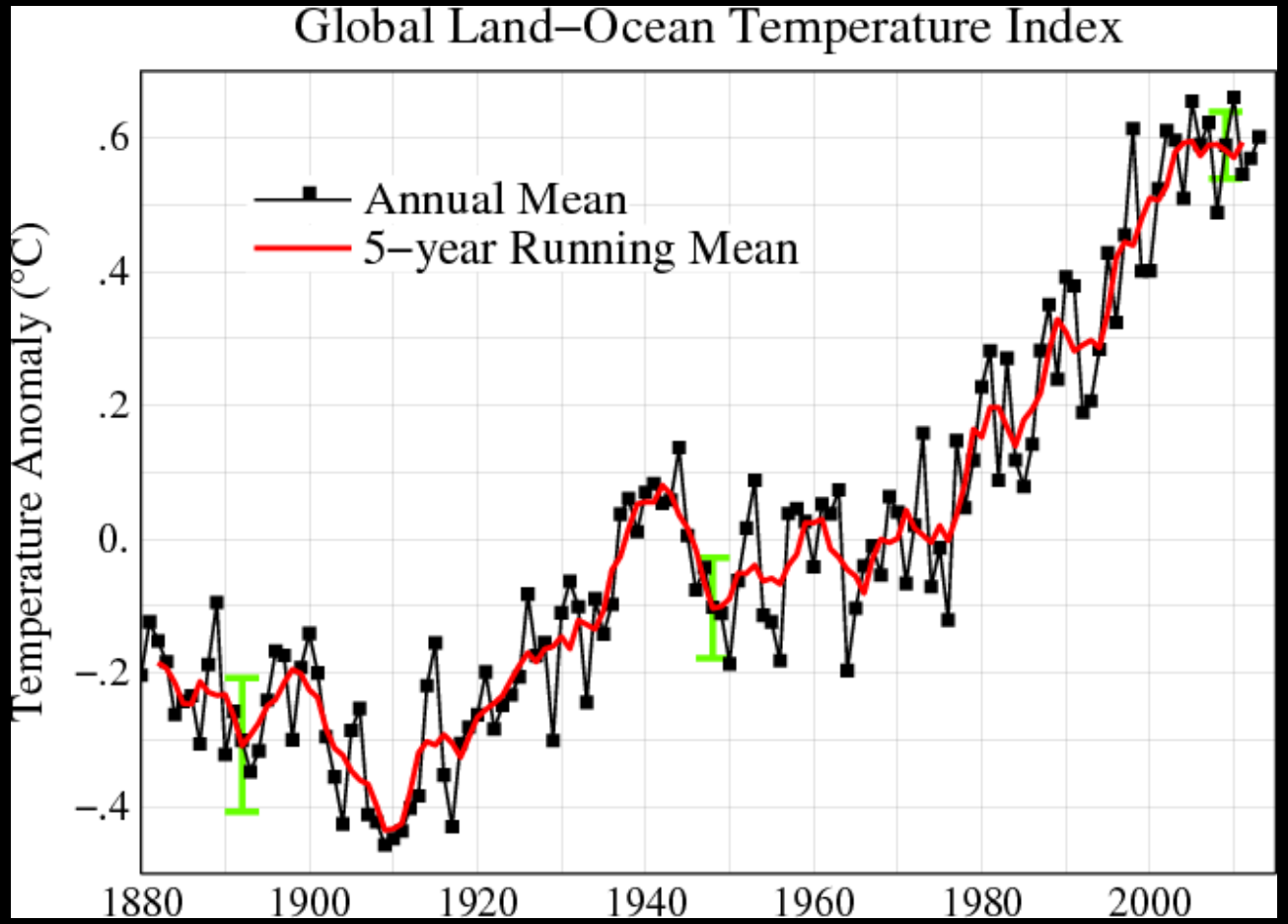


Understanding

Data Visualization Global Temp

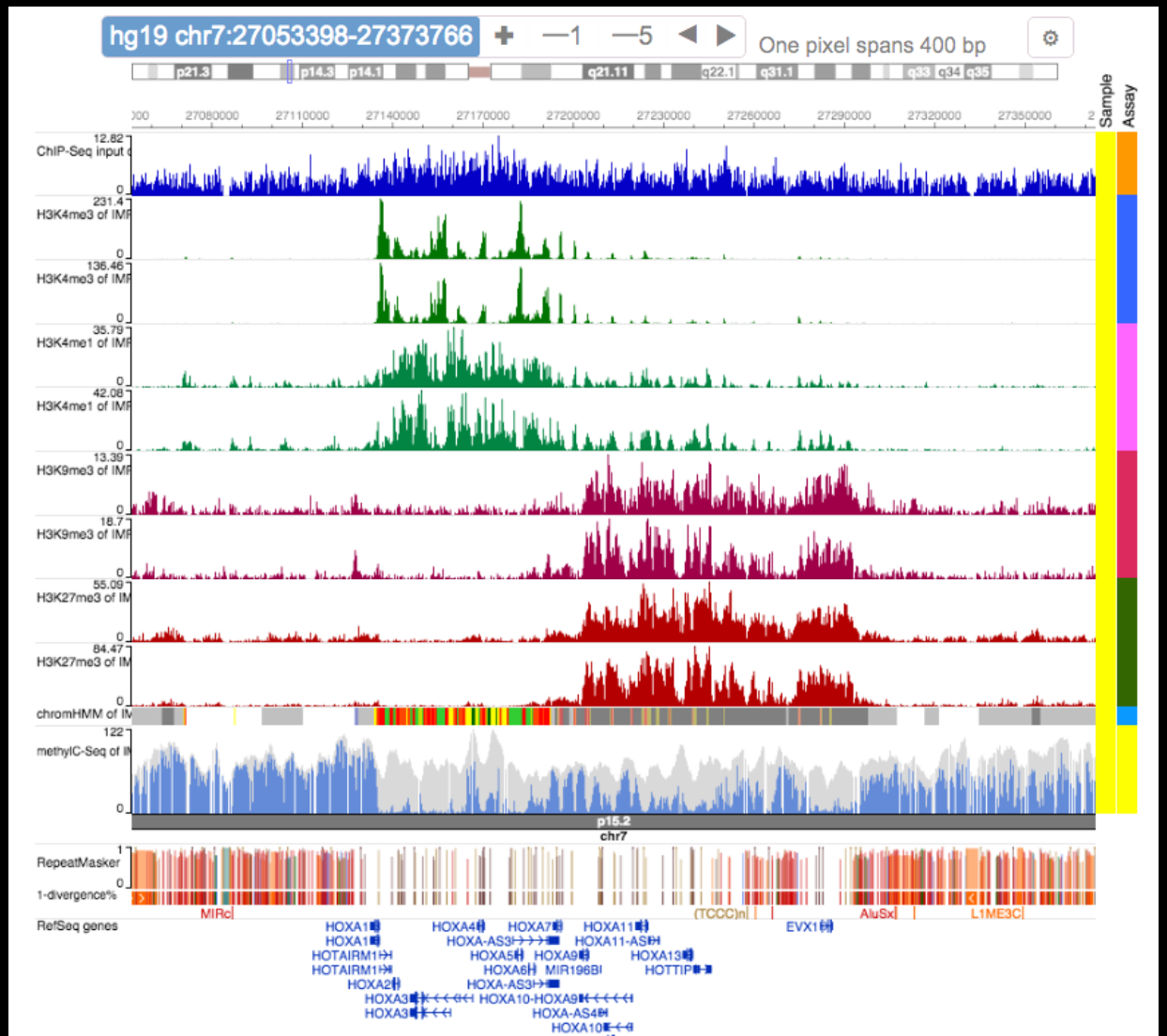
Year	Annual_Mean	5-year_Mean
1880	-0.20	*
1881	-0.12	*
1882	-0.15	-0.19
1883	-0.18	-0.19
1884	-0.26	-0.22
1885	-0.24	-0.25
1886	-0.23	-0.25
1887	-0.31	-0.21
1888	-0.19	-0.23
1889	-0.09	-0.23
1890	-0.32	-0.23
1891	-0.26	-0.26
1892	-0.30	-0.31
1893	-0.35	-0.29
1894	-0.32	-0.27
1895	-0.24	-0.25
1896	-0.17	-0.24
1897	-0.17	-0.21
1898	-0.30	-0.19
1899	-0.19	-0.20
1900	-0.14	-0.23
1901	-0.20	-0.24
1902	-0.30	-0.28
1903	-0.36	-0.31
1904	-0.43	-0.32
1905	-0.29	-0.35
1906	-0.25	-0.36
1907	-0.41	-0.37
1908	-0.42	-0.40
1909	-0.46	-0.44
1910	-0.45	-0.43
1911	-0.44	-0.43
1912	-0.40	-0.38
1913	-0.38	-0.32
1914	-0.22	-0.30
1915	-0.16	-0.31
1916	-0.35	-0.29
1917	-0.43	-0.31
1918	-0.31	-0.33
1919	-0.28	-0.30
1920	-0.26	-0.27
1921	-0.20	-0.26
1922	-0.28	-0.25
1923	-0.25	-0.23
1924	-0.23	-0.21
1925	-0.21	-0.19
1926	-0.08	-0.17
1927	-0.17	-0.18
1928	-0.16	-0.16
1929	-0.30	-0.16
1930	-0.11	-0.15
1931	-0.06	-0.16
1932	-0.10	-0.12
1933	-0.24	-0.13
1934	-0.09	-0.14
1935	-0.14	-0.11
1936	-0.10	-0.05
1937	0.04	-0.03
1938	0.06	0.02
1939	0.01	0.05
1940	0.07	0.06
1941	0.08	0.06
1942	0.05	0.08
1943	0.06	0.07
1944	0.14	0.04
1945	0.01	0.02
1946	-0.08	-0.02
1947	-0.04	-0.07
1948	-0.10	-0.10
1949	-0.11	-0.10
1950	-0.19	-0.09
1951	-0.06	-0.05
1952	0.02	-0.05
1953	0.09	-0.04
1954	-0.11	-0.06
1955	-0.12	-0.06
1956	-0.18	-0.07
1957	0.04	-0.04
1958	0.05	-0.02
1959	0.03	0.02
1960	-0.04	0.02
1961	0.05	0.03
1962	0.04	-0.01
1963	0.07	-0.03
1964	-0.20	-0.05
1965	-0.10	-0.06
1966	-0.04	-0.08
1967	-0.01	-0.03
1968	-0.05	-0.00
1969	0.06	-0.01
1970	0.04	0.00
1971	-0.07	0.04
1972	0.02	0.02
1973	0.16	0.01
1974	-0.07	-0.01
1975	-0.01	0.02
1976	-0.12	-0.00
1977	0.15	0.04
1978	0.05	0.08
1979	0.12	0.16
1980	0.23	0.15
1981	0.28	0.20
1982	0.09	0.20
1983	0.27	0.17
1984	0.12	0.14
1985	0.08	0.18
1986	0.14	0.19
1987	0.28	0.22
1988	0.35	0.28
1989	0.24	0.33
1990	0.39	0.31
1991	0.38	0.28
1992	0.19	0.29
1993	0.21	0.30
1994	0.28	0.29
1995	0.43	0.34
1996	0.32	0.42
1997	0.45	0.44
1998	0.61	0.44
1999	0.40	0.48
2000	0.40	0.51
2001	0.52	0.51
2002	0.61	0.53
2003	0.60	0.58
2004	0.51	0.59
2005	0.65	0.59
2006	0.59	0.57
2007	0.62	0.59
2008	0.49	0.59
2009	0.59	0.58
2010	0.66	0.57
2011	0.55	0.59
2012	0.57	*
2013	0.60	*
2014	*	*

Global Means Temp as Graph

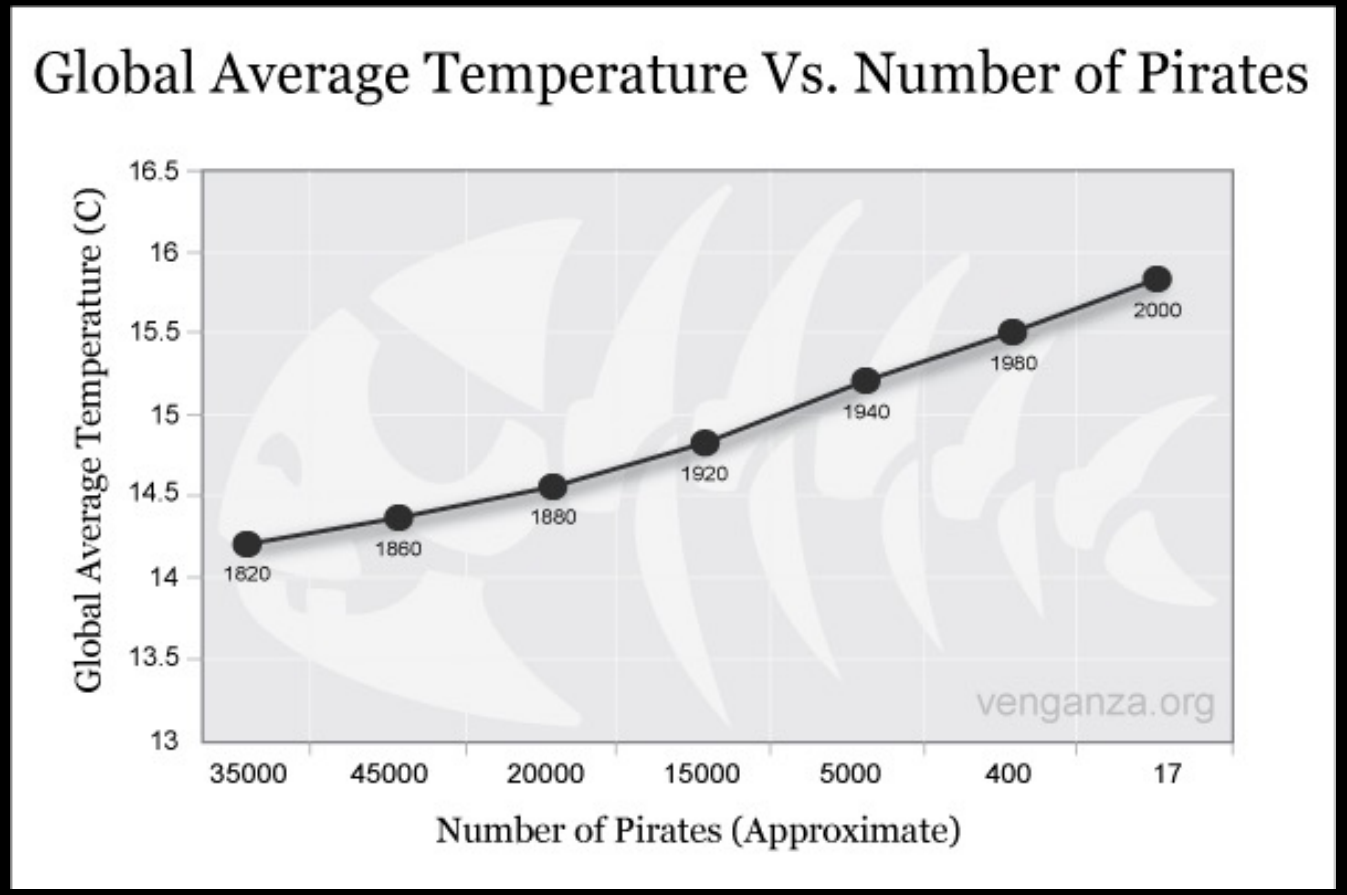


Hansen et al. (2006), NASA GISS

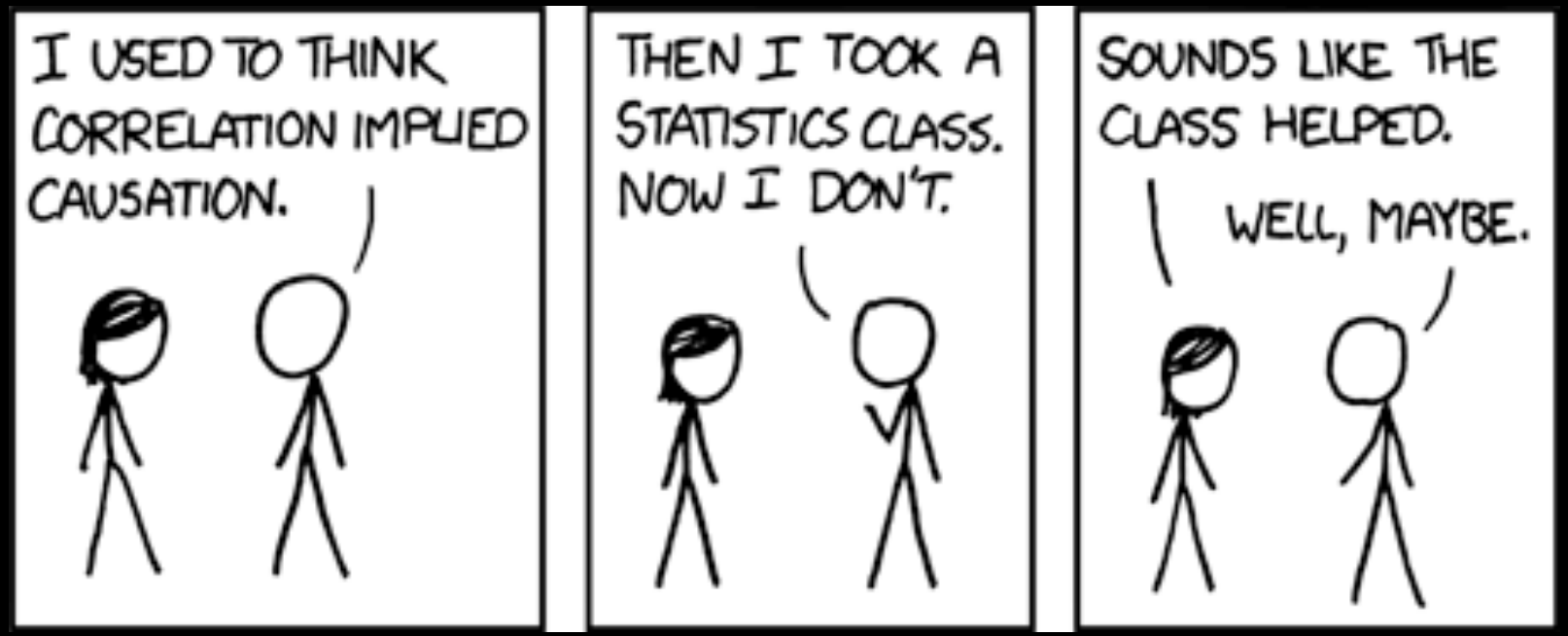
Viz Critical for Data Exploration



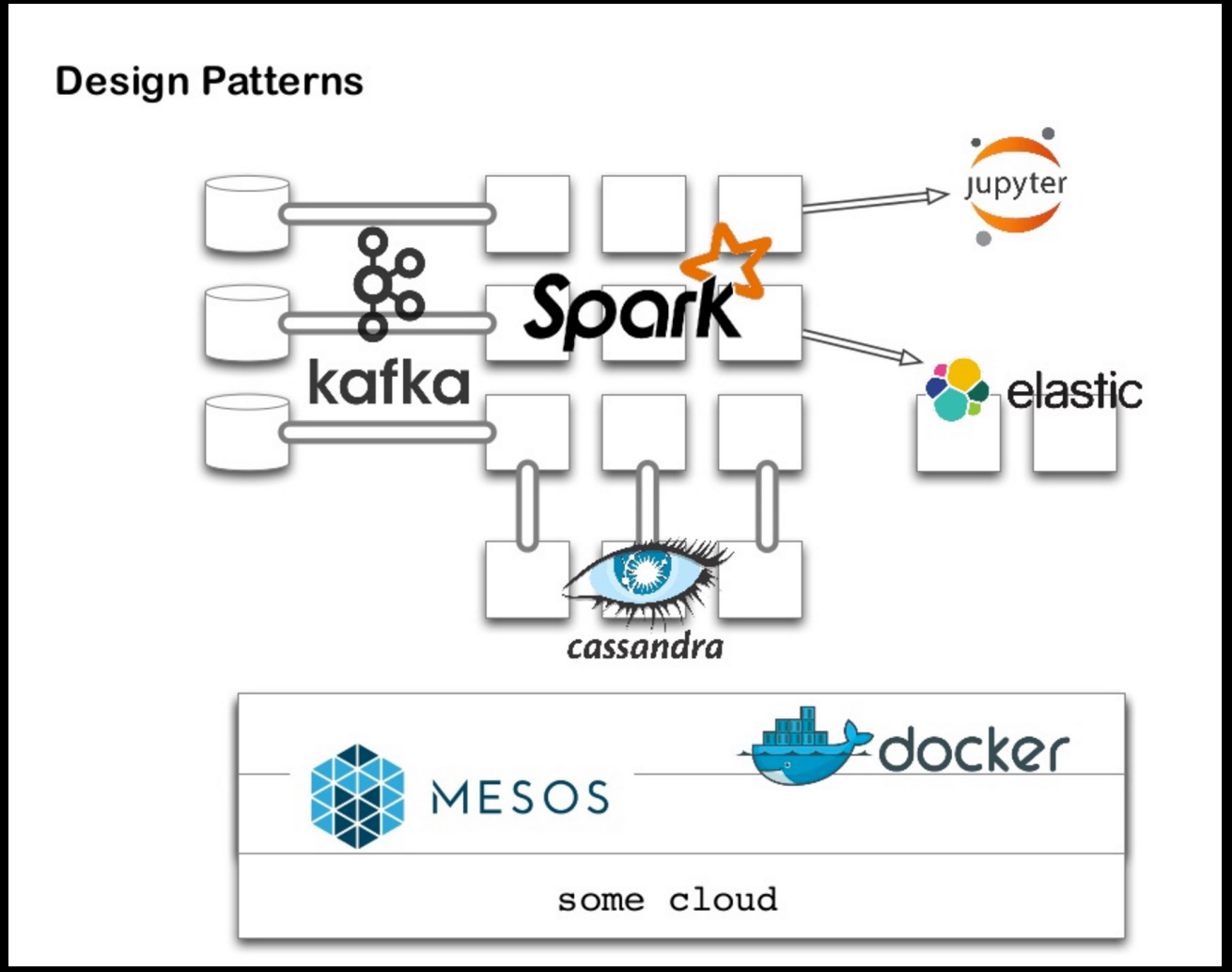
Statistics


Global Average Temperature Vs. Number of Pirates
Global Average Temperature (C)
16.5
16
15.5
15
14.5
14
13.5
13
1620
1860
1880
1920
1940
1980
2000
35000
45000
20000
15000
5000
400
17
Number of Pirates (Approximate)
venganza.org

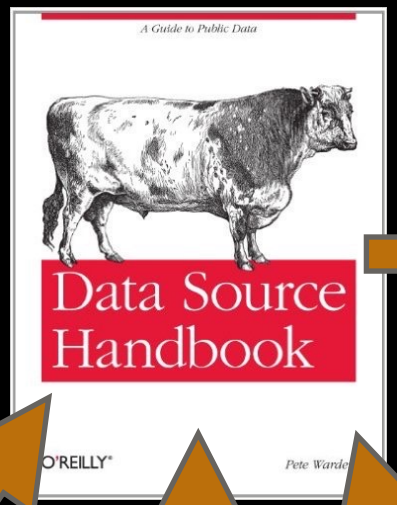
Statistics: Nonsense Protection

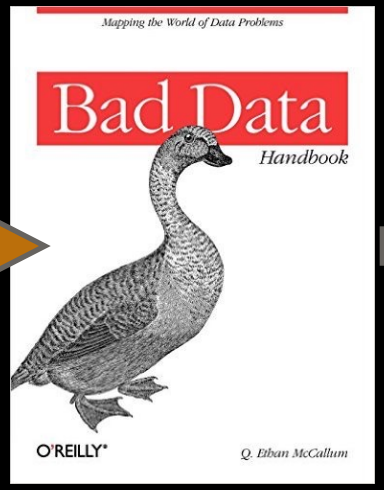


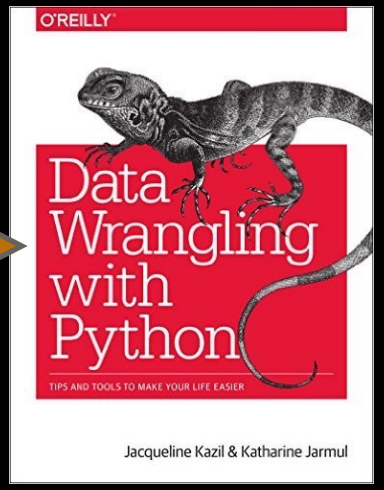
Handling Big Data

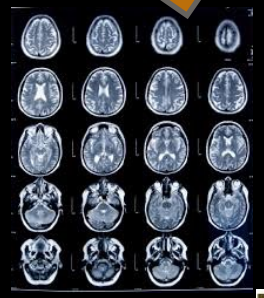


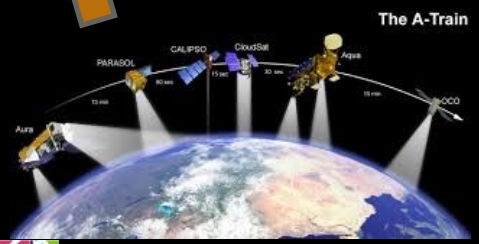
Data Engineering











Data Cleaning

Data Transformation
Signal/Image Processing


big
data
information
petabytes
storage
search
sets
large
management
analysis
government
internet
volume
technology
technologies
software
research
development
science
future
new
process
san
nas
million
exabytes
world
learning
size
billion
databases
business
mapreduce
applications
processing
based
amount
approaches
initiative
challenges
parallel
social
increasing
implementation
factor
structure
sources
state
terms
drive
flow
per
statistics
distributed
nearly
processes
second
scientific
companies
times
people
various
limits
high
private
simulations
many
online
hand
institute
paradigm
international
trends
algorithms
market
analytics
include
integration
needed
funding
economic
day
across
well
shared
time
relevant
intelligence
records
tools
used
critiques
program
year
queries
sensor
traffic
determine
between
even
networks
terabytes
results
variety
ability
privacy
higher
effective
rate
real
systems
out
hypothesis
huge
every
critique
years
now
visualization
complex
insight
massive
manage
center
infrastructure
uses
end
making
past
countries
related
national
database
framework
users
cost
most
handle
definition
worldwide
successful
work
one
replication
use
set
using
project
less
would