Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#201 Data Science or Statistics ?

été 2020

Statistics ?

"All the News That's Fit to Print"

The New York Times

VOL. CLXIX .... No. 58,645 — FRIDAY, MARCH 27, 2020

JOB LOSSES SOAR; U.S. VIRUS CASES TOP WORLD

New Data Shows Staggering Toll of Outbreak

Since the coronavirus descended on Brooklyn Hospital Center three weeks ago, the staff has handled over 800 potential cases.

Unfilled Posts Force Scramble By Washington

Online Class With No Way To Get There

Courage at a Brooklyn Hospital, At the Front of an Invisible War

With Supplies Waning, Workers Soldier On

No Crowd, but I'll Take You Out to the Ballgame

A Fan Writes a Fantasy for Opening Day

Weekly unemployment claims

"All the News That's Fit to Print"

The New York Times

Late Edition

VOL. CLXIX ... No. 58,688 — NEW YORK, SATURDAY, MAY 9, 2020 — $3.00

U.S. UNEMPLOYMENT IS WORST SINCE DEPRESSION

Georgia Killing Puts Spotlight on a Police Force's Troubled History

Long Path to Arrests of Ex-Officer and Son in Black Man's Death

April's Rate of 14.7% Touches All Parts of Economy

In Flynn Case, Russia Inquiry Is Barr's Target

If West Wing Still Isn't Safe, Is Any Office?

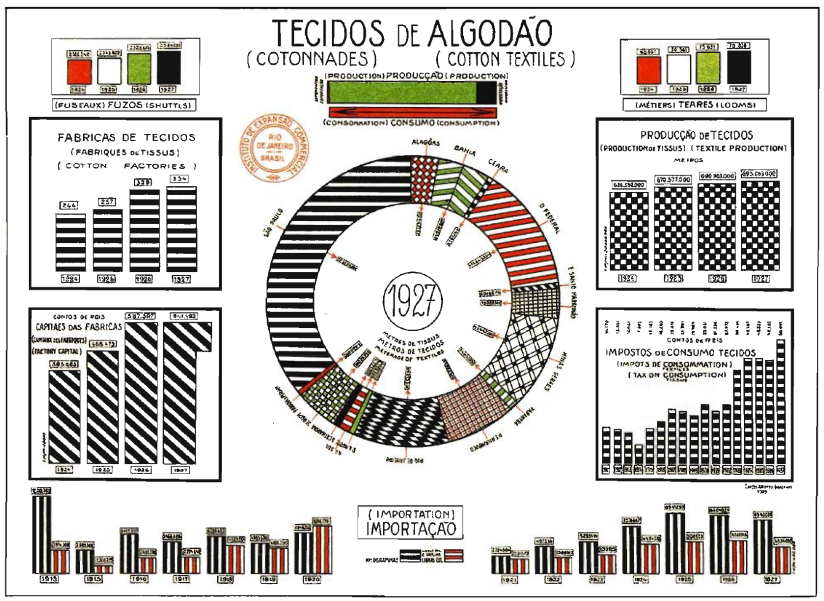
As Official Tally Ignores Reality, Mexico's Hospitals Are Overrun

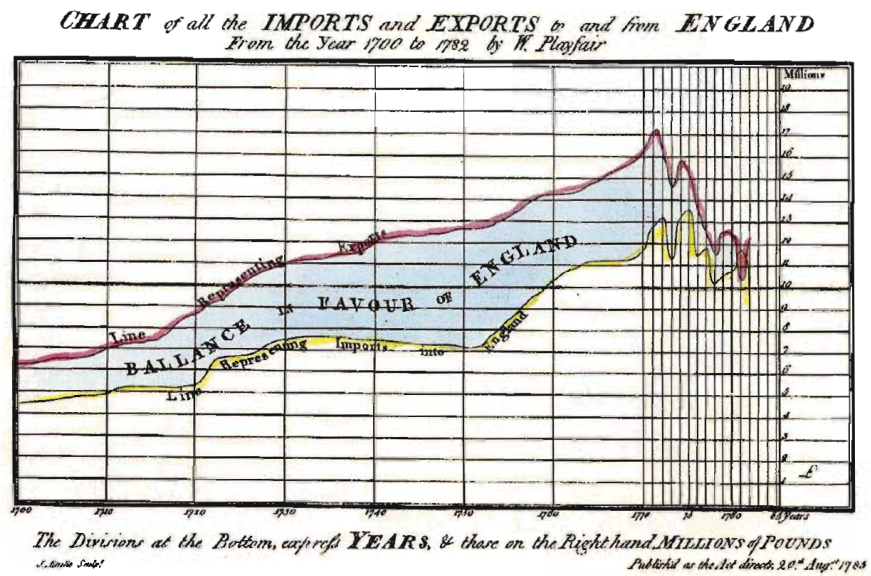
A Mexico City crematory. Regional virus deaths are officially low.

20,500,000 jobs lost in April

Statistics ?

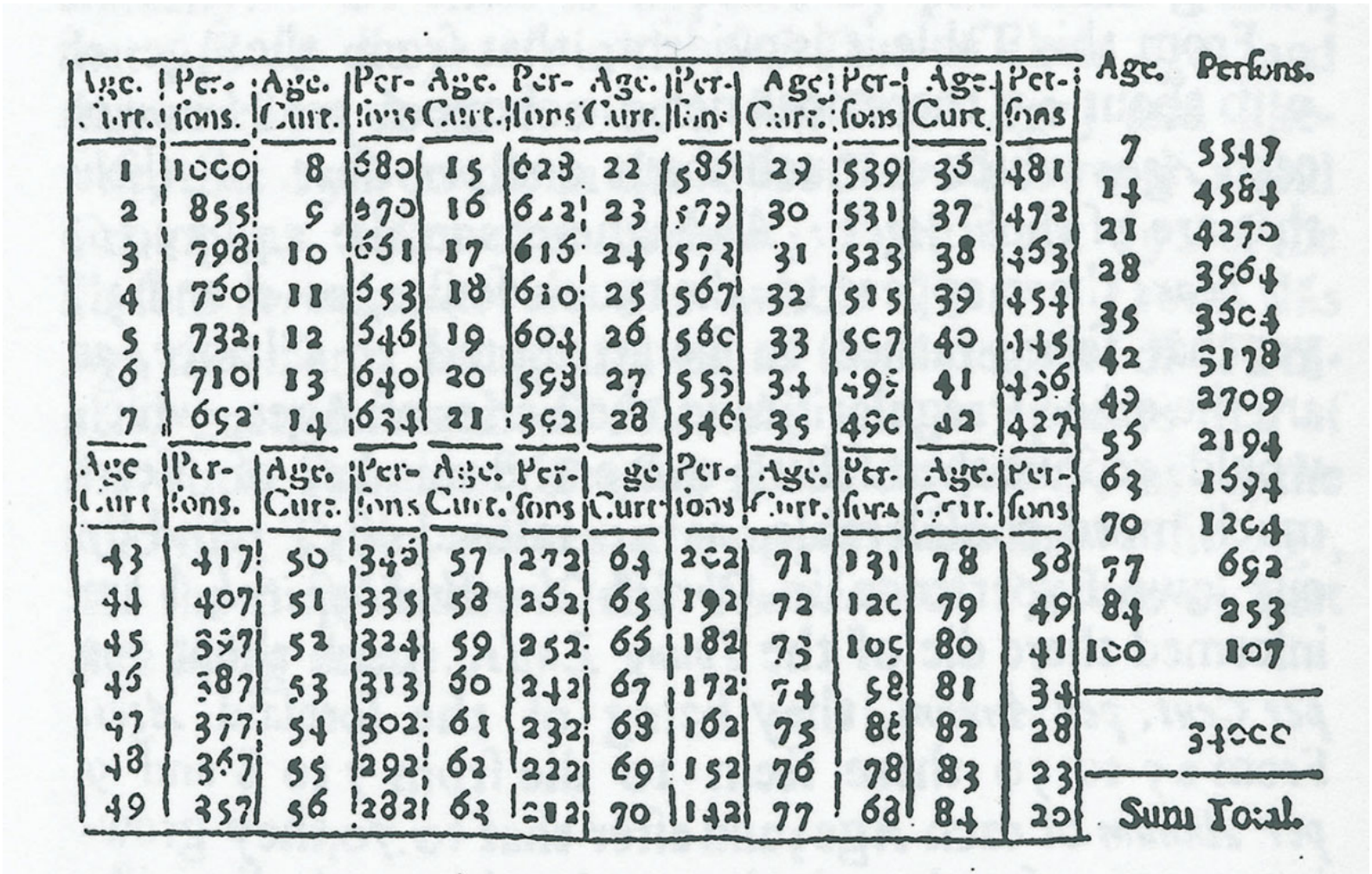
German Statistik, introduced by Gottfried Achenwall (1749), originally designated the analysis of data about the state, signifying the "science of state" (then called political arithmetic).





Statistik was data to be used by governmental and (often centralized) administrative bodies

Statistics ?



Age.	Persons.
7	5547
14	4584
21	4270
28	3964
35	3504
42	3178
49	2709
56	2194
63	1594
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

Graunt's' (1662) Observations on the Bills of Mortality

Statistics ?

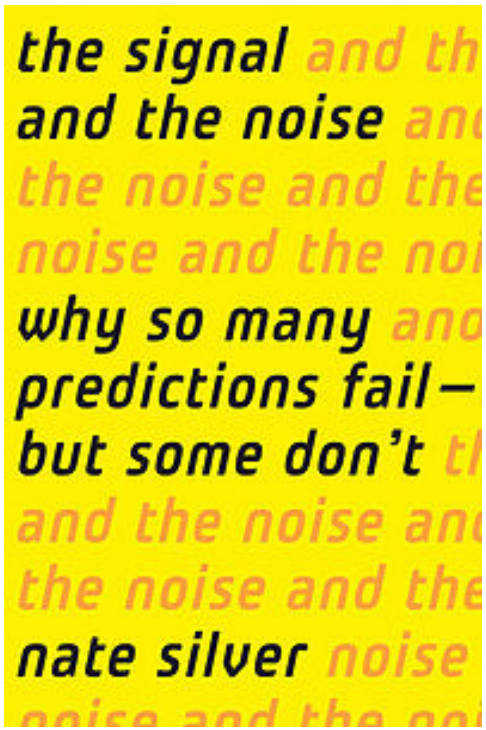
Graunt concluded that the plague was caused by person-to-person infection rather than the competing theory of "infectious air" based on the pattern of infections through time, "thought as we think today [...] they reasoned about their data" Where Shall the History of Statistics Begin?

"People use statistics as the drunken man uses lamp posts - for support rather than illumination",
Andrew Lang (or not)

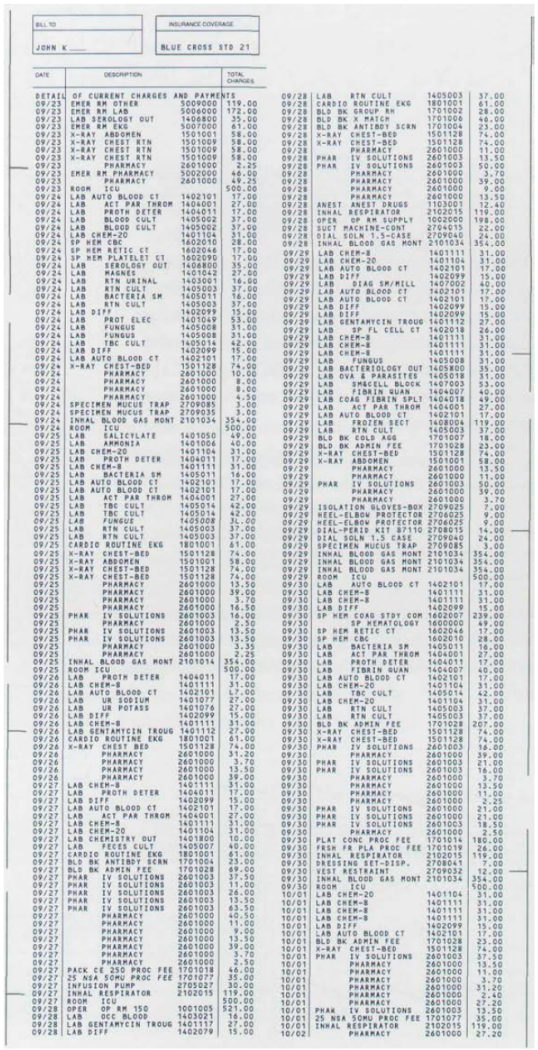
Statistics ? Data ?

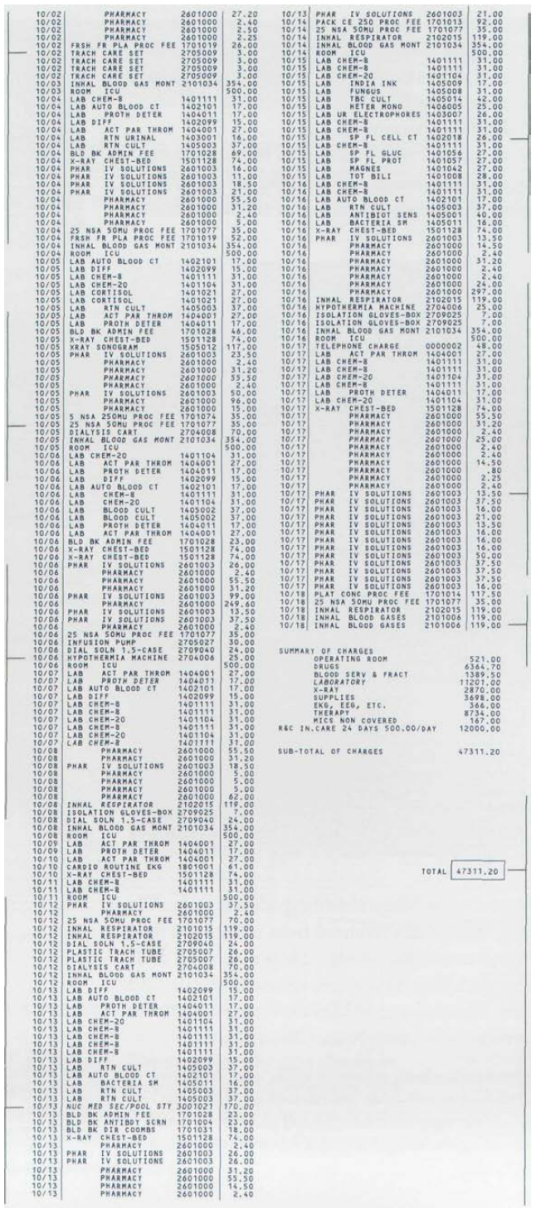
Big Data context

Data, data everywhere



(source Edward Tufte)





Agenda

Data are x_i's, or $\boldsymbol{x}_i = (x_{1,i}, \cdots, x_{k,i})$'s.

- functions, c.d.f. $F(x) = \mathbb{P}[X \leq x]$ and density (histogram)
- statistical indicators
  - central value(s): average(s), median
  - dispersion: variance, standard deviation, inequalities
  - approximations
  - quantiles
- inference: from frequentist to Bayesian
- testing, significance, p-value
- bivariate analysis (x_i, y_i)
- multivariate analysis $(\boldsymbol{x}_i = (x_{1,i}, \cdots, x_{k,i}))$
  - projection & dimension reduction
  - cluster analysis
- networks
- time series