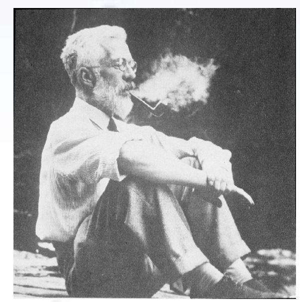
Applied Statistics

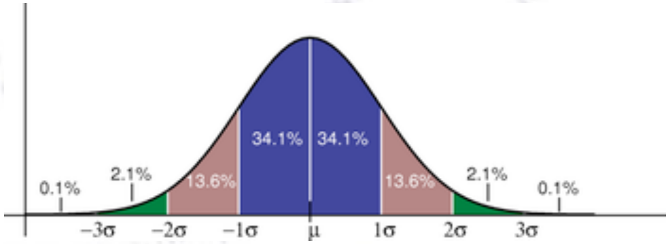
Types of data and ways of illustrating it








Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Types of data

To first order, data comes in two general types and then "the rest":
- **Discreet** (typically counting data, i.e. positive integers)
- **Continuous** (at least more or less)
- **The rest**, i.e. text, images - but typically convertible into two first.

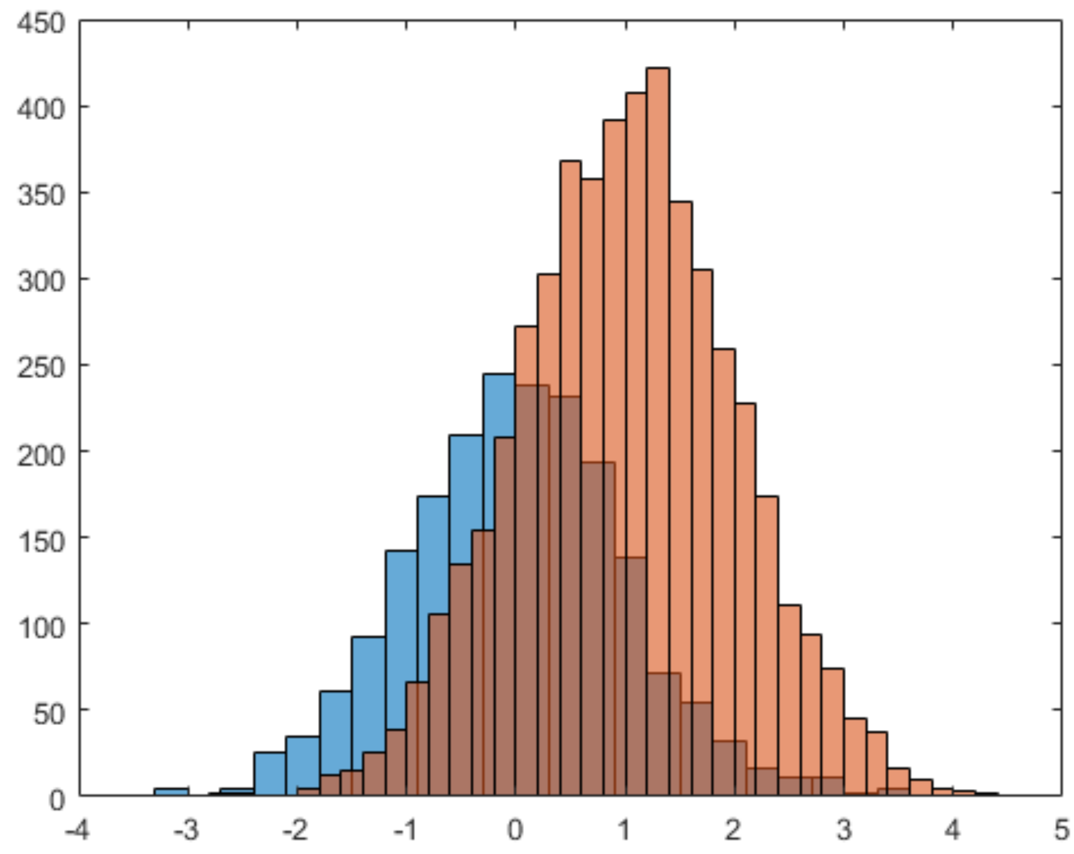
A pitfall is that continuous data is not always continuous, but may seem so!!! The problem arise, if plotting in a histogram with binning comparable (and possibly prime) to steps.

Most basically, one has repeated measurements of things (i.e. 1D distributions). However, often there are several dimensions in the data (possibly 1000s), leading to near-infinite complexity.

Data can be paired in different ways, and/or divisible into groups, experiments, periods, etc. This "meta-data" is important to keep track of.

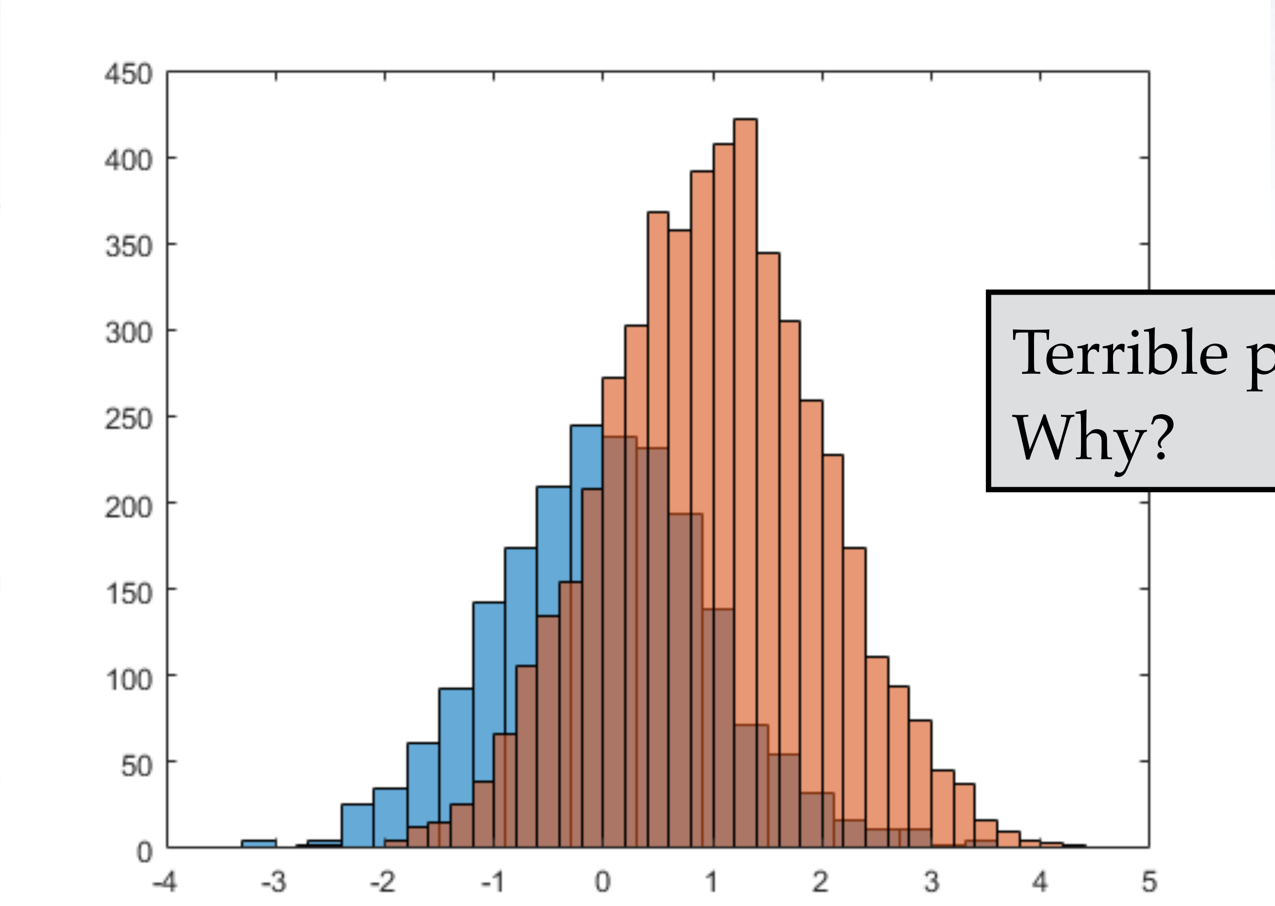
Ways of displaying data

Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Ways of displaying data

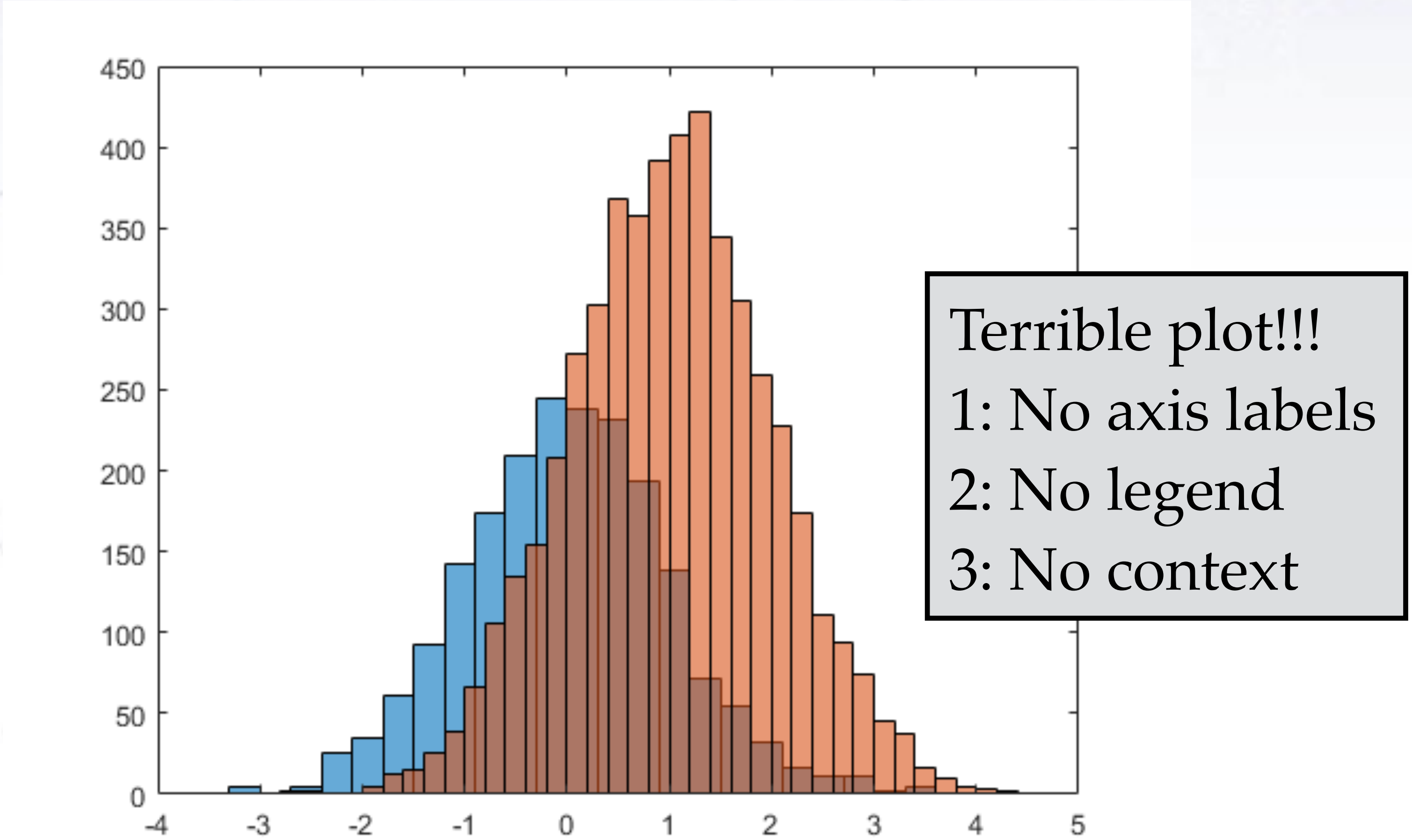
Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Terrible plot!!!
Why?

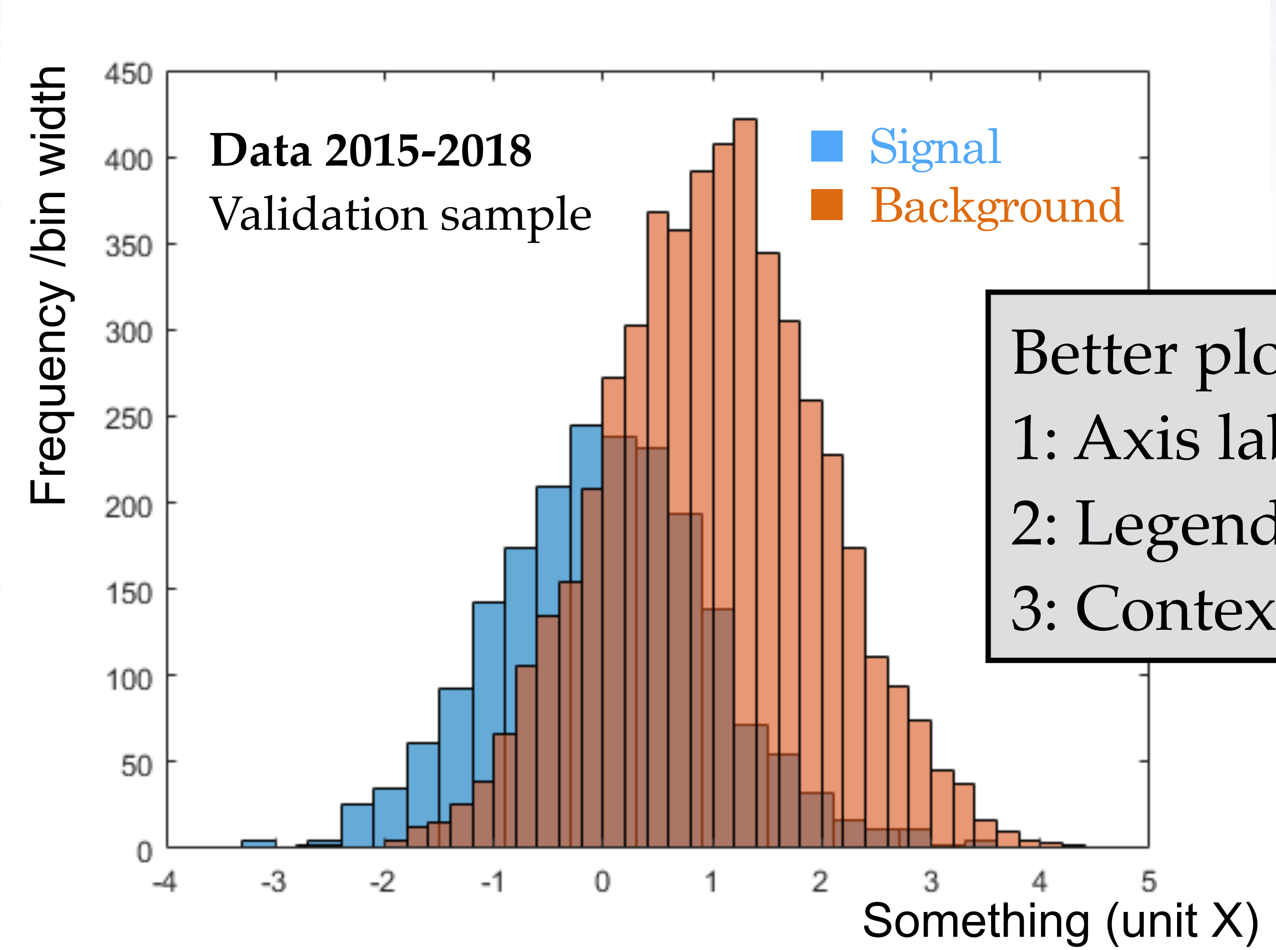
Ways of displaying data

Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Ways of displaying data

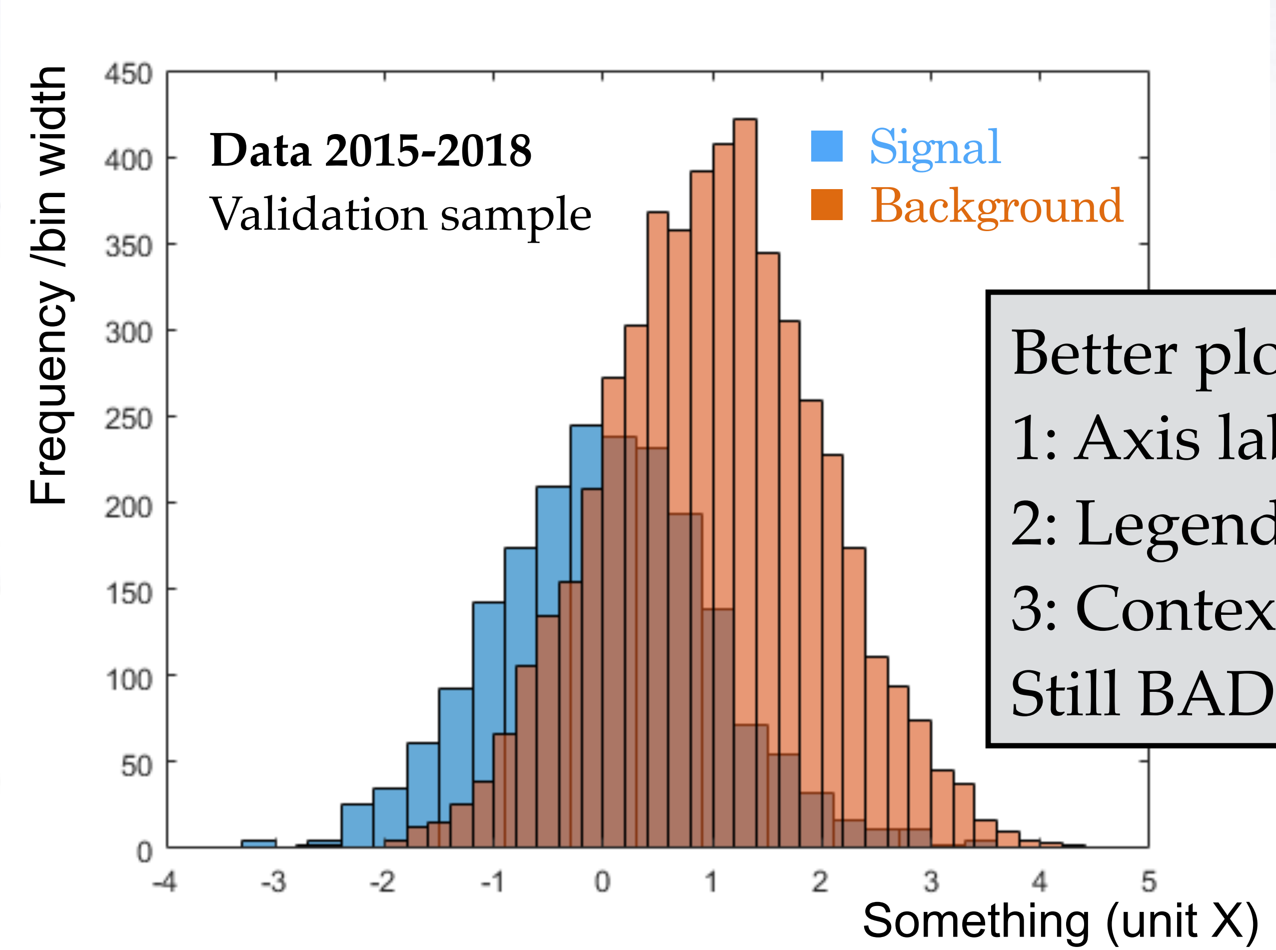
Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Better plot!!!
1: Axis labels
2: Legend
3: Context

Ways of displaying data

Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.

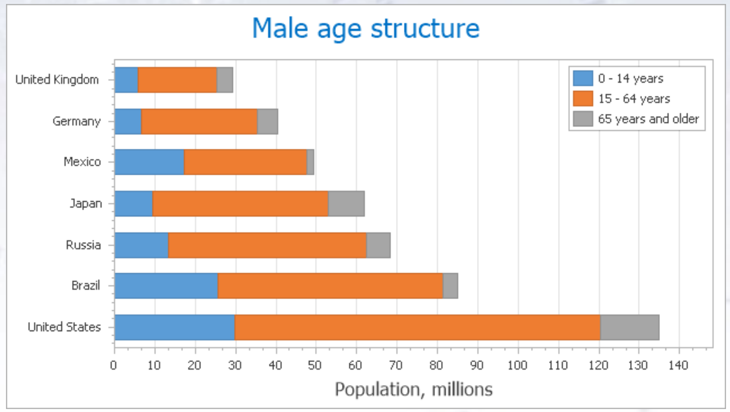


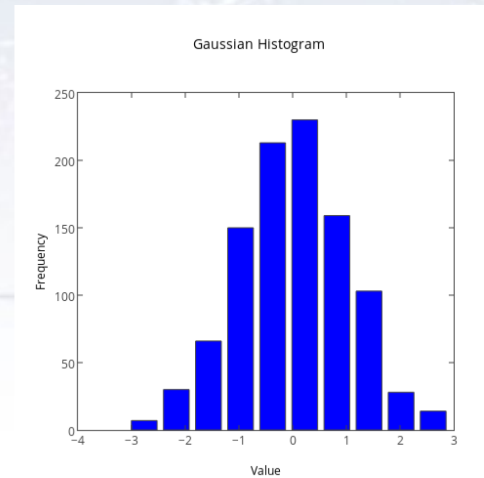
Better plot!!!
1: Axis labels
2: Legend
3: Context
Still BAD! Why?

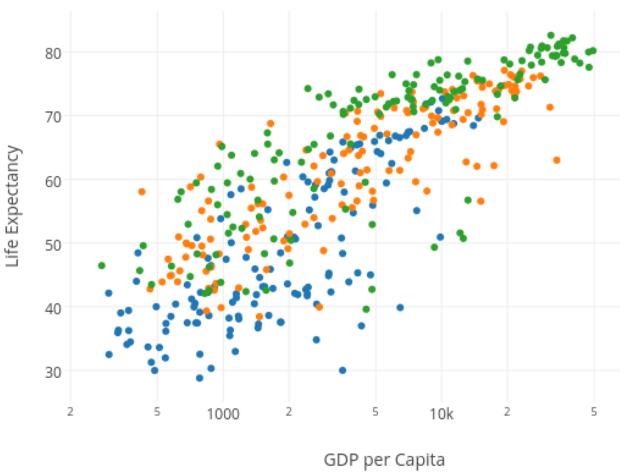
Displaying data

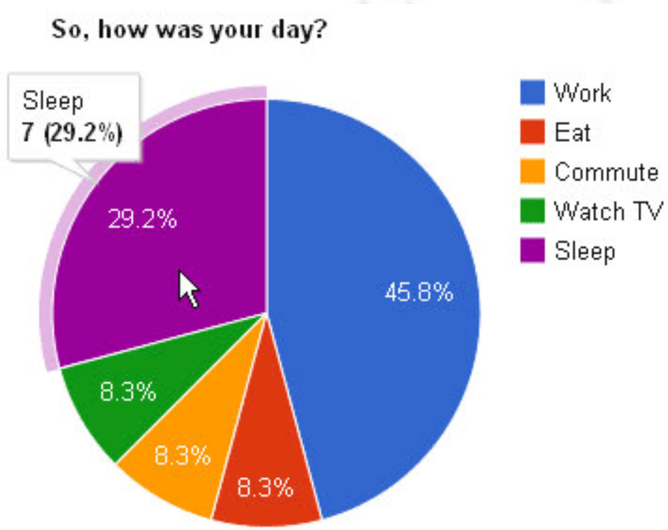
There are a multitude of ways to display data, some of which are:

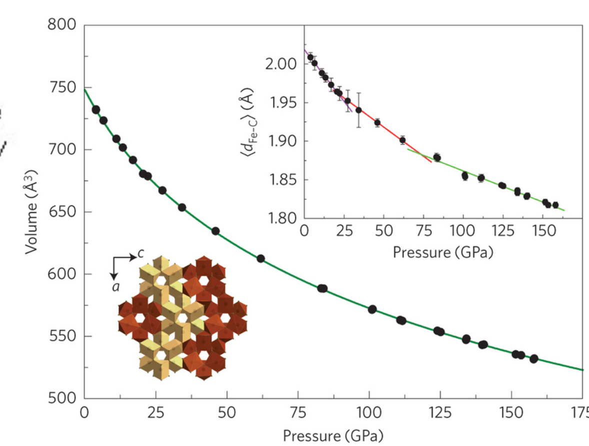
- Bar charts
- Histograms
- Scatter plots
- Pie charts
- Line / fits











A note on plots

Simple plots (for checks):

Most plots you produce is for yourself! Make sure they have labels on the axis, but otherwise don't put too much work into their style.

Time scale to produce: Minutes

Important plots (for showing):

Some plots are for others, and they should be clear cut and illustrative, or the message will be lost. Ask yourself (and then a fellow student), if they understand the plot, and what could be done to improve them.

Time scale to produce: Hours

Central (i.e. money) plots (for public circulation):

A few key plots will be shown elsewhere by others, but ONLY if they are of good quality and illustrate a relevant point well. For these few plots (2-10 in a thesis) you should invest some time in getting them right, as they hold the result of months/years work.

Time scale to produce: Days

A note on plotting

Always plot your data!!!

You never really know, what goes on in data, until you have SEEN it!

A note on plotting

Always plot your data!!!

You never really know, what goes on in data, until you have SEEN it!

A Ph.D. student a few months ago comes into my office, asking for help with statistics, as his likelihood fit gave good results, but his Chi2 not!

TP: Have you seen the histograms? **Ph.D.**: No, but they are so simple…

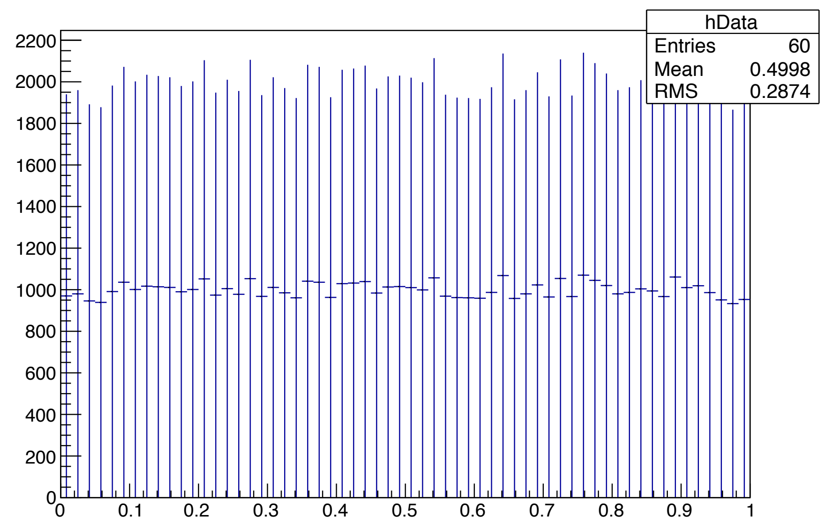
A note on plotting

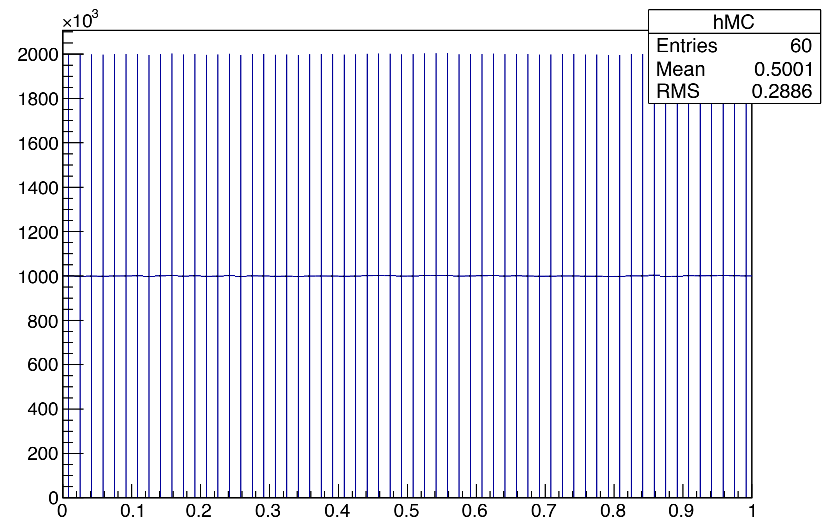
Always plot your data!!!

You never really know, what goes on in data, until you have SEEN it!

A Ph.D. student a few months ago comes into my office, asking for help with statistics, as his likelihood fit gave good results, but his Chi2 not!

TP: Have you seen the histograms? **Ph.D.**: No, but they are so simple…





Examples of

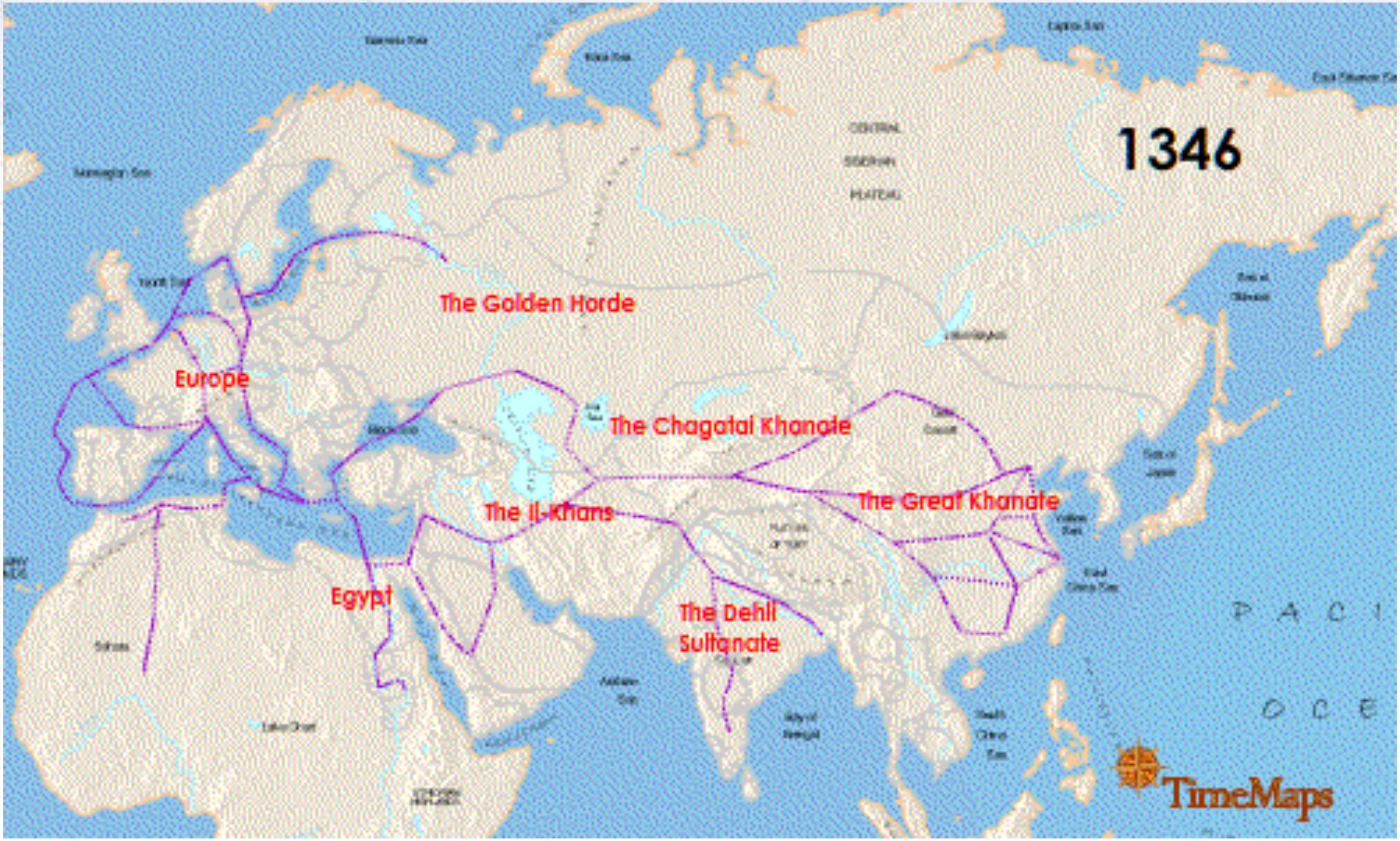
Poor Plots & Illustrations

Black Death… and bad plots!

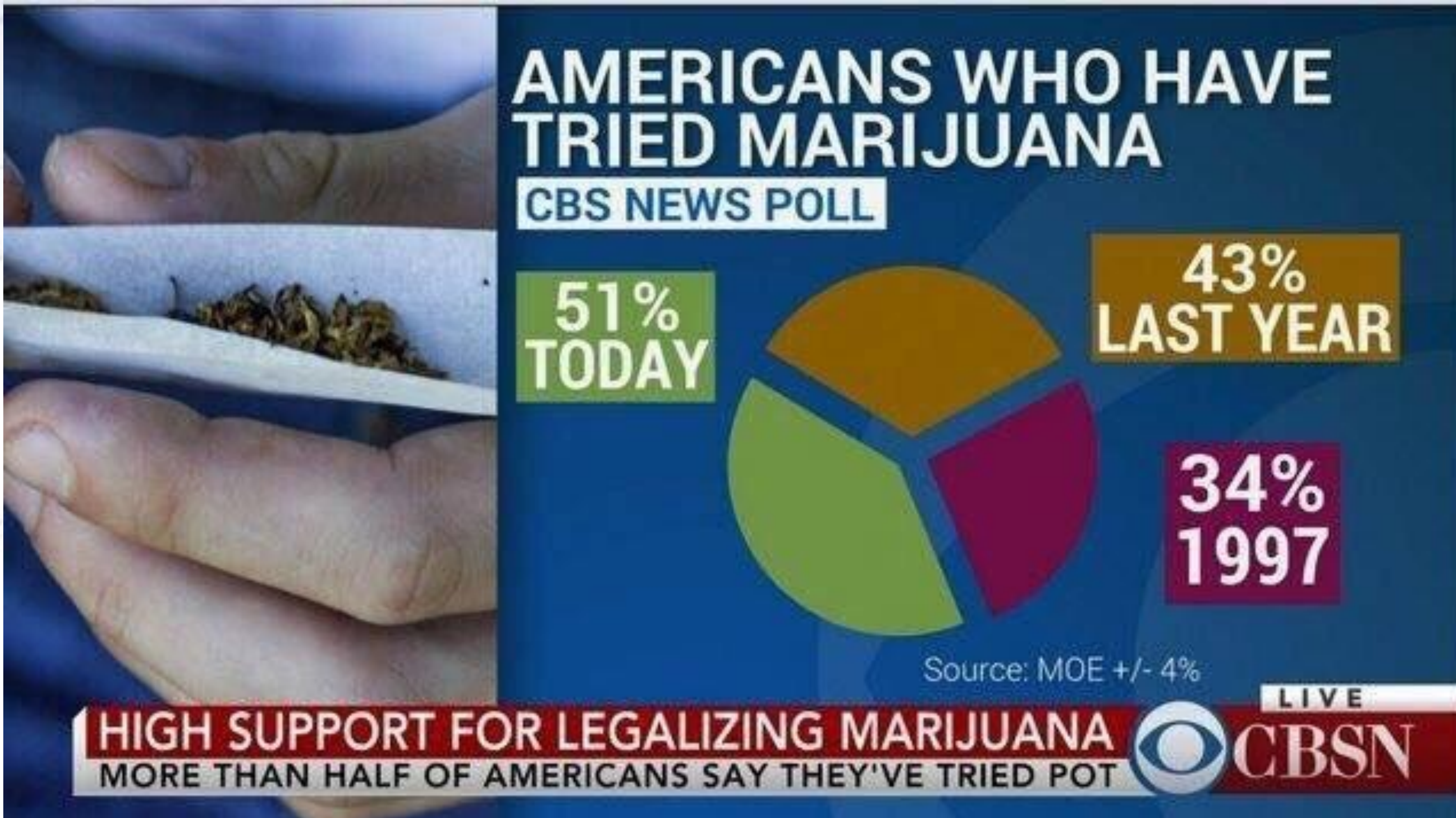
While at first the plot looks very cool, but the use of colors is often misinterpreted. The colors imply differing intensities or mortality rates, but the legend indicates they represent time. The arrows convey all the real information here…

Black Death… and bad plots!

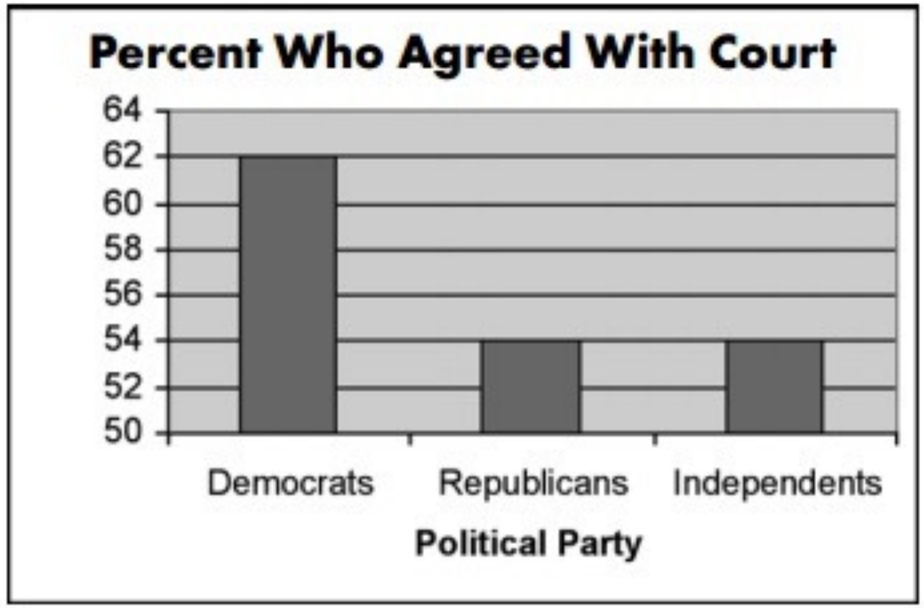
While at first the plot looks very cool, but the use of colors is often misinterpreted. The colors imply differing intensities or mortality rates, but the legend indicates they represent time. The arrows convey all the real information here... but now we can do better!

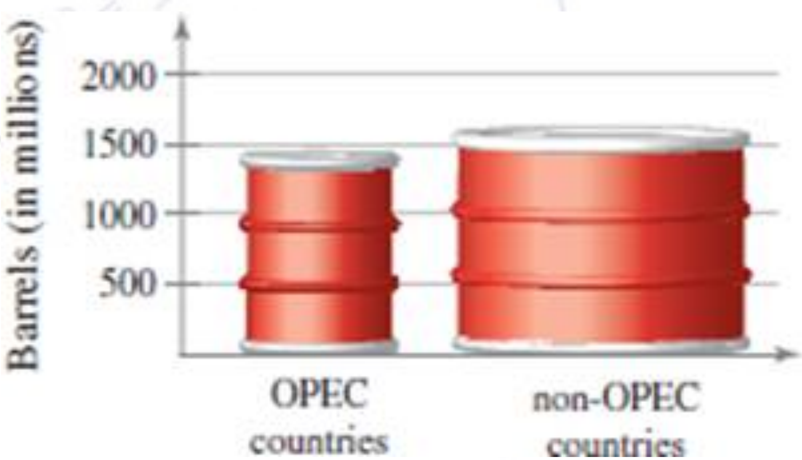


128% of Americans have tried marijuana?



A mix of poor plots

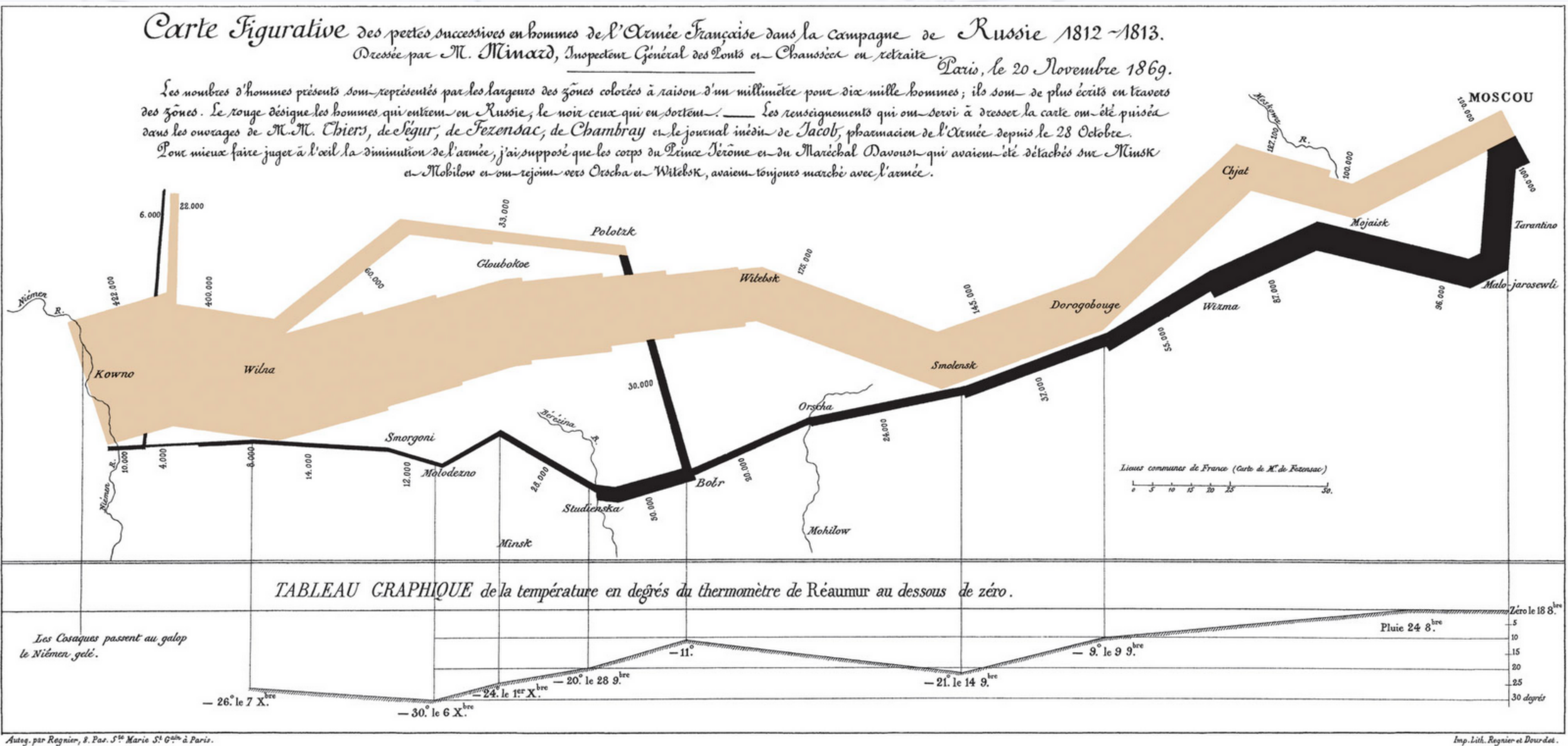




Examples of Great Plots & Illustrations

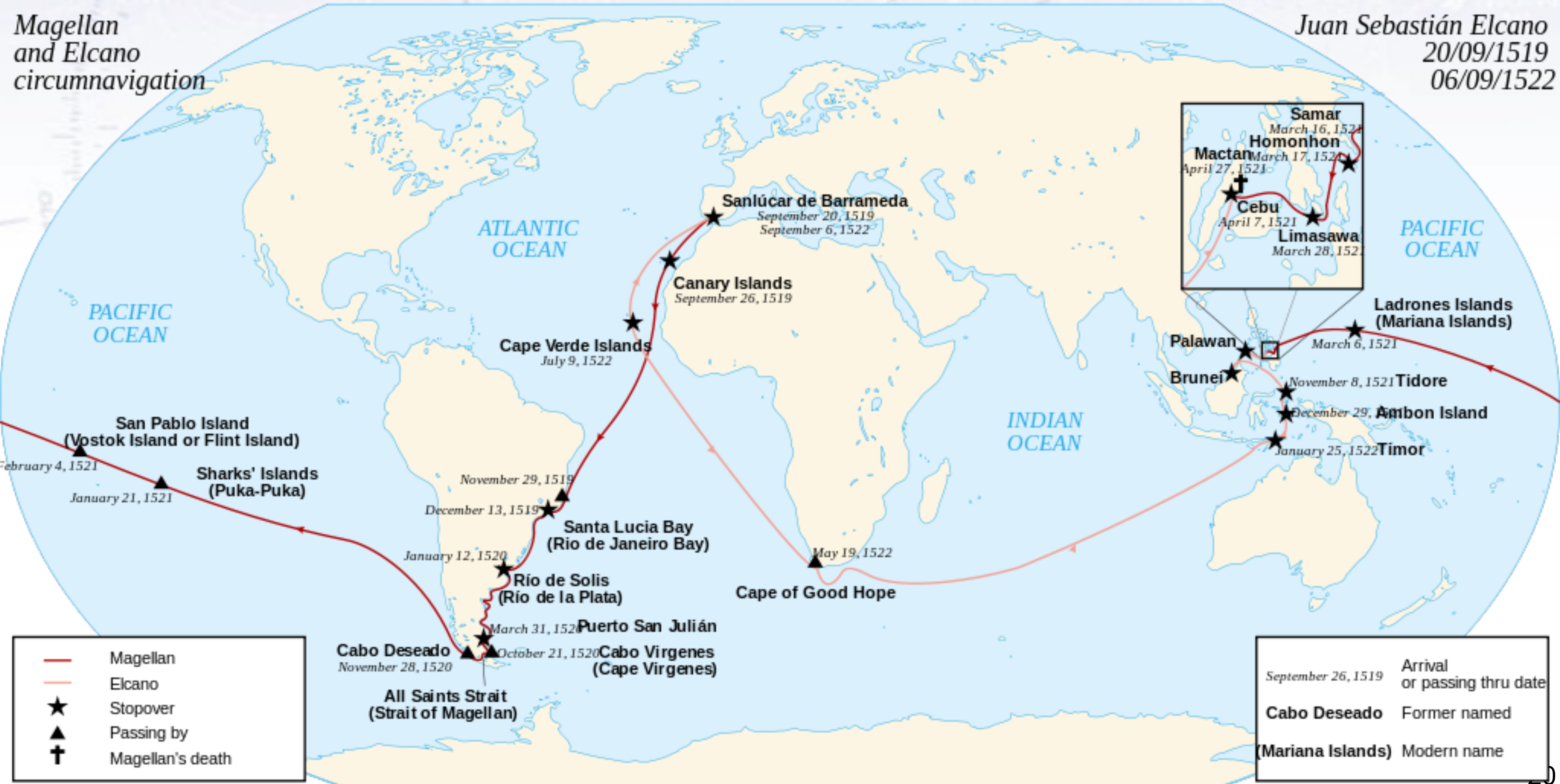
Napoleon’s march on Russia

This has been called by some the "greatest figure ever made". It illustratively tells the story of Napoleon's catastrophic march and retreat in Russia in 1812, losing 400.000 men. The graph contains a massive amount of data, showing landmarks and geographic course the army took, the size of the army over time, and the temperature of the bitter Russian winter. You can study this figure and gain insight as to why Napoleon lost.



Magellan’s circumnavigation

In addition to starting the globalisation age, the expedition (unwittingly) discovered the need for an **International Date Line**. Despite numerous deaths, an accurate ships log was kept for over 1000 days. When the surviving sailors returned, they realised their log was **off the local calendar by 1 day**. Cool, consider the death, drama, plight, and sheer insanity of the voyage…

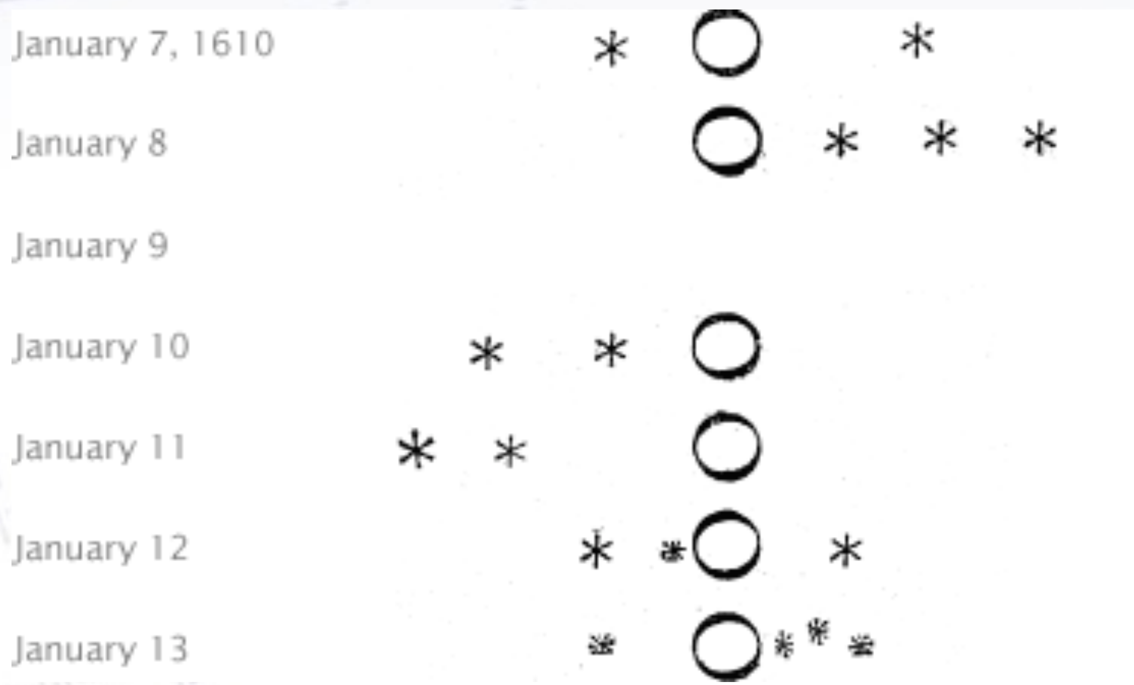


Galileo on Jupiter’s moons

This is a condensed version of the famous observations Galileo made of the Jovian moons. Jupiter is shown as the O, and the moons as *. Using these simple observations, Galileo deduced that each little * was actually orbiting Jupiter, which gave credence to the controversial Copernican theory that the Earth is not center of the Universe.

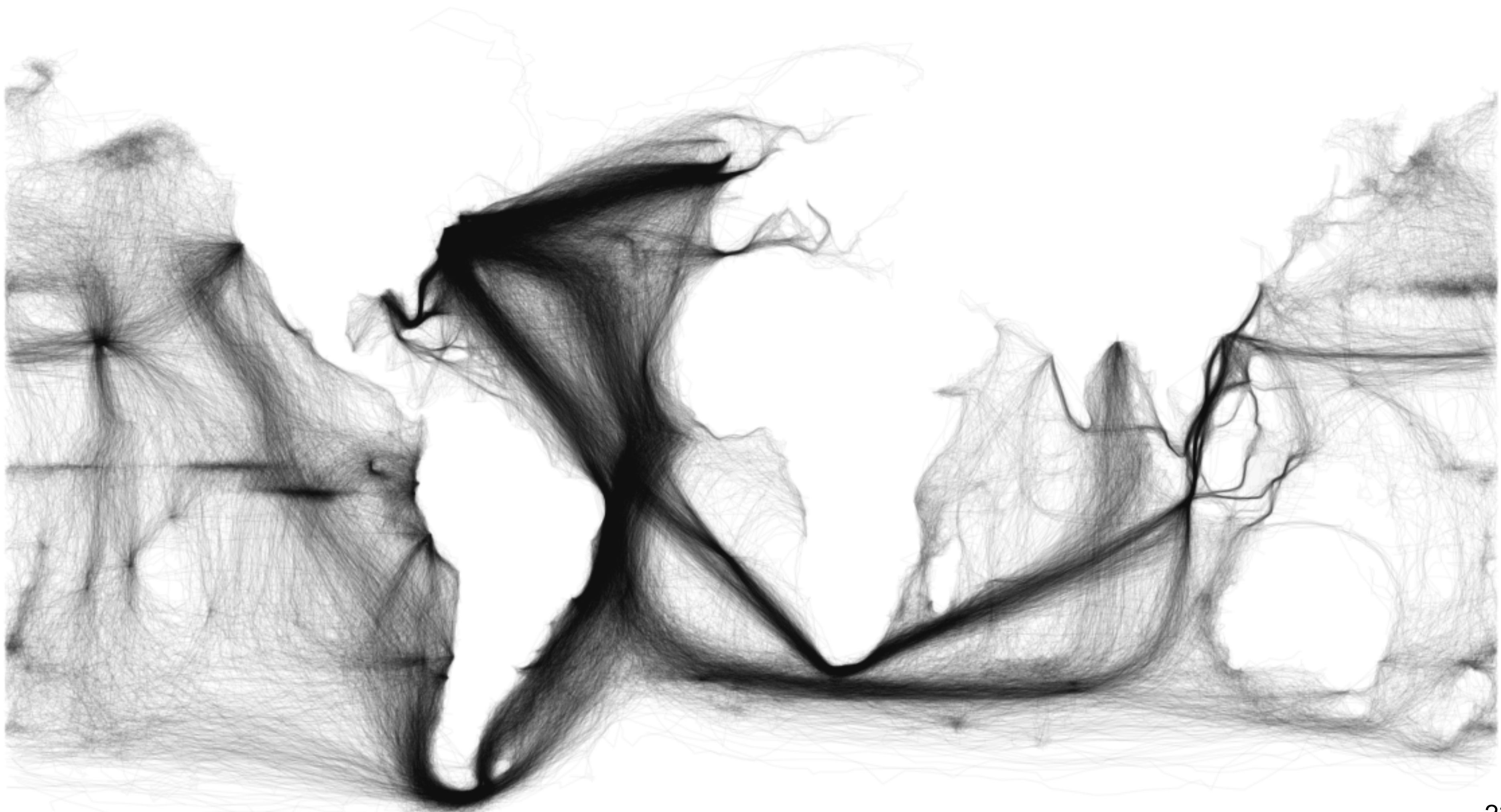
What is great about this figure is its simplicity, Tufte would approve the sparse labeling and lack of extemporaneous axes.

When you align each date's observations with Jupiter, as above, the helix pattern the moons trace as they orbit nearly jumps out!



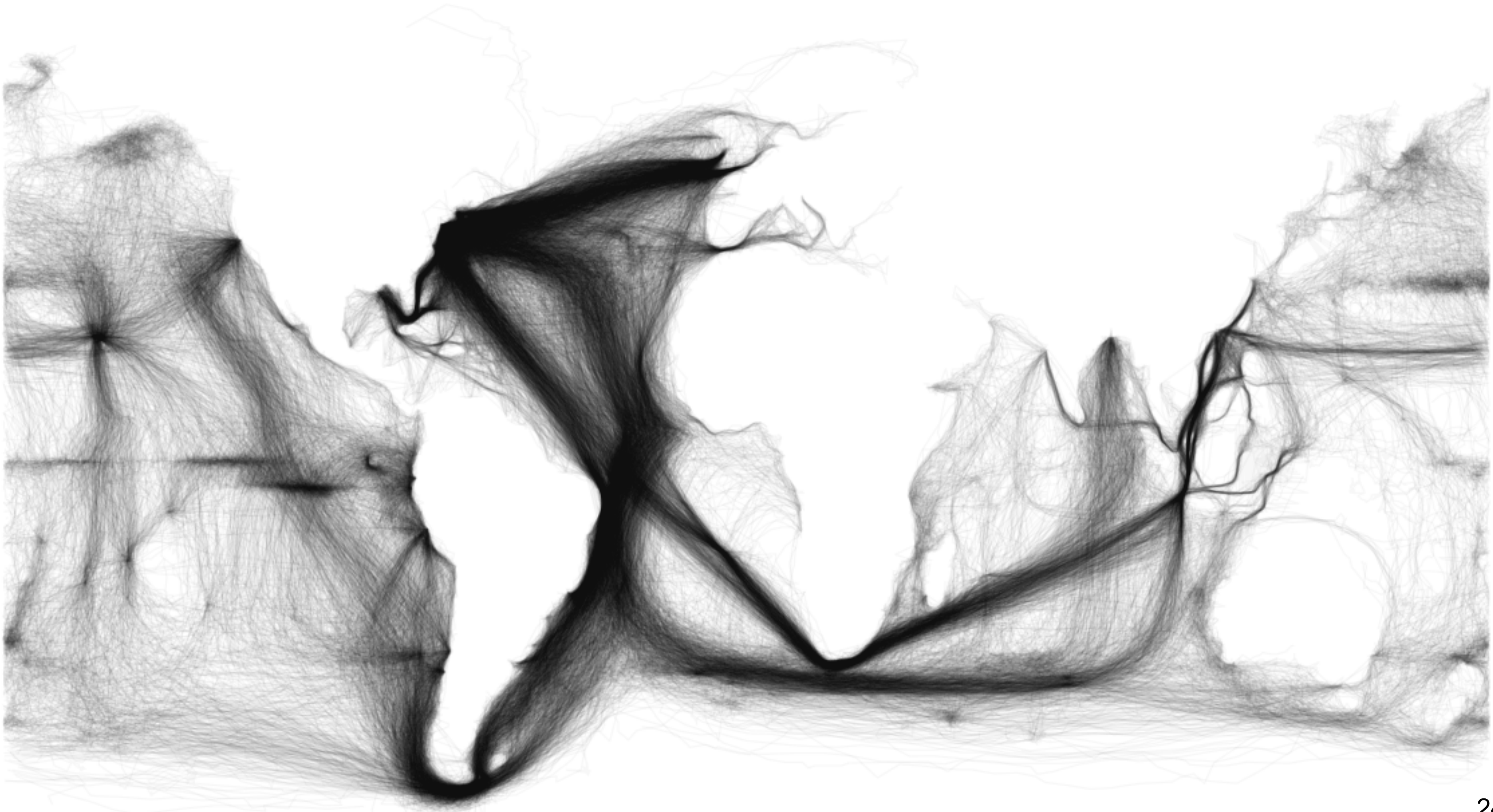

January 7, 1610
January 8
January 9
January 10
January 11
January 12
January 13

Guess what this is…

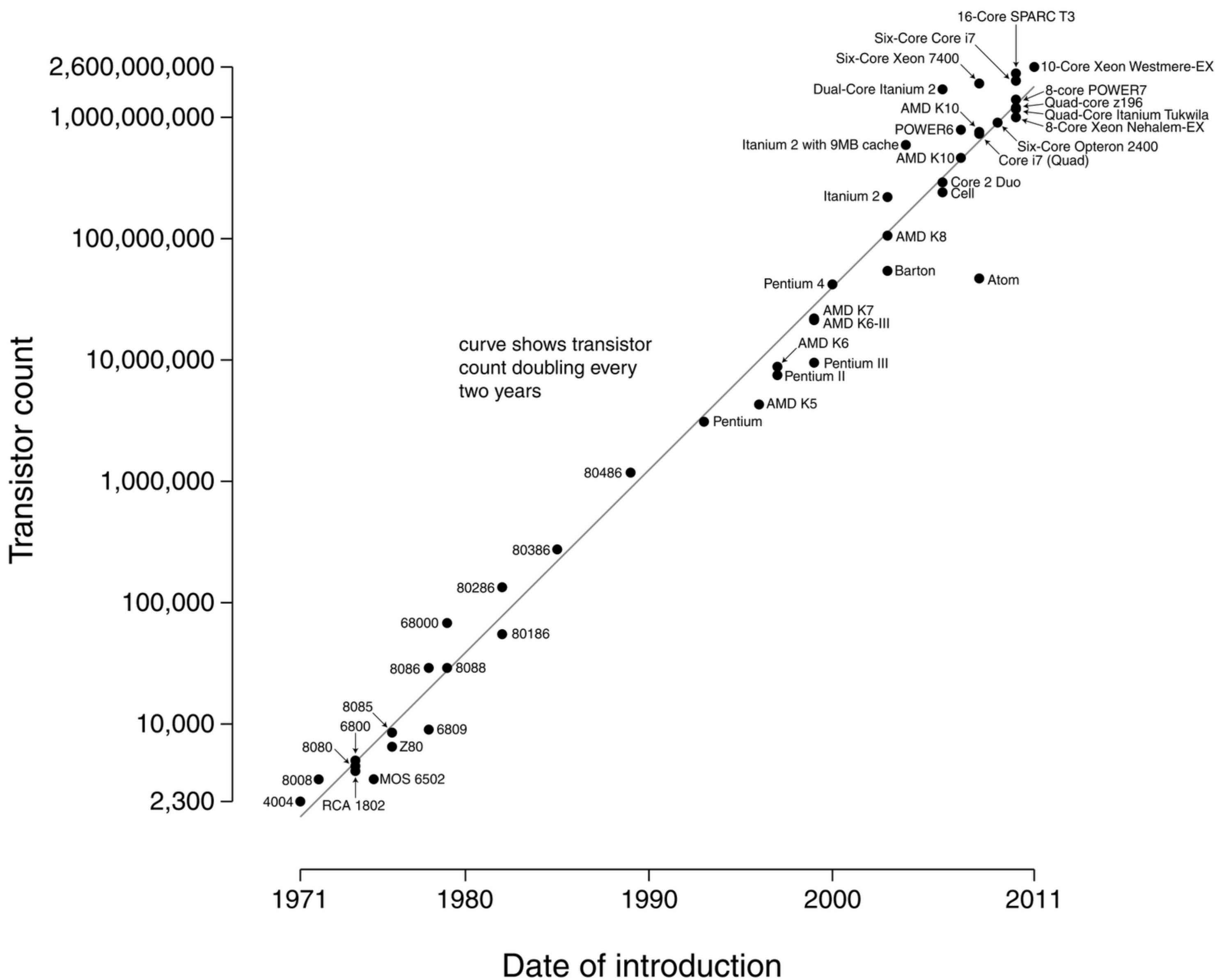


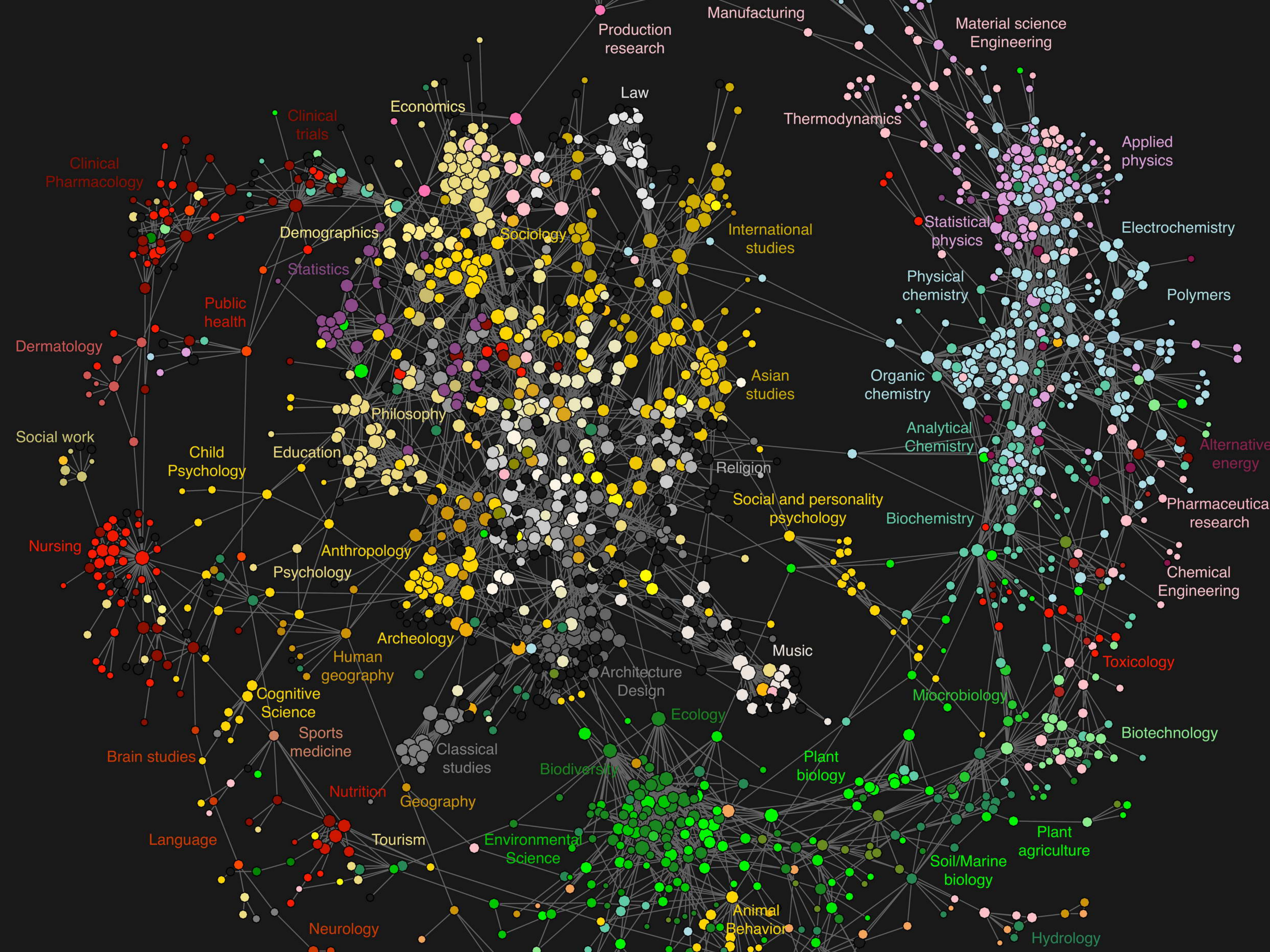
Guess what this is…

Plot millions of journal entries from 18th and 19th century ship logs, and you reveal a picture of ocean trade you've never seen before.



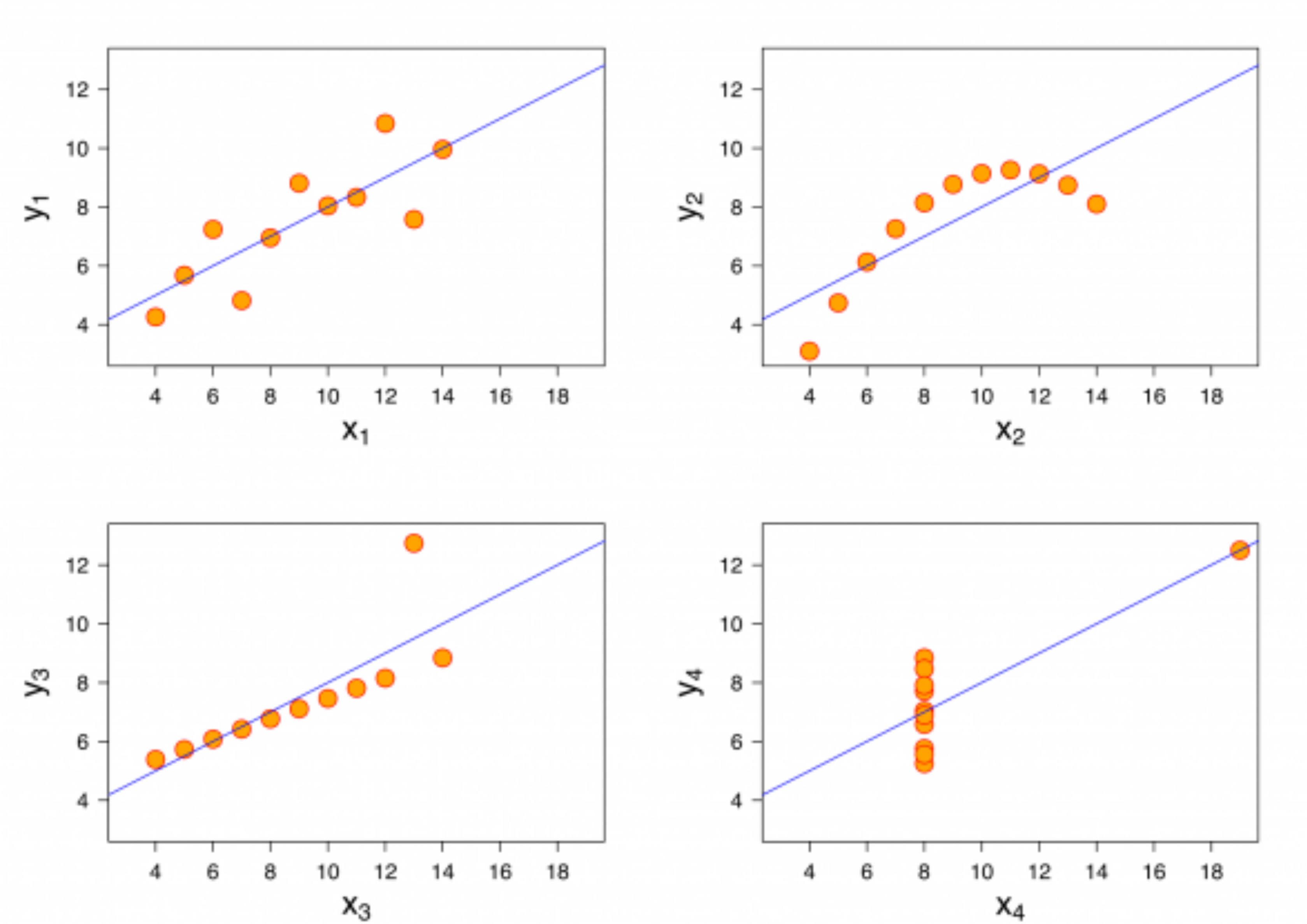
Microprocessor Transistor Counts 1971-2011 & Moore's Law

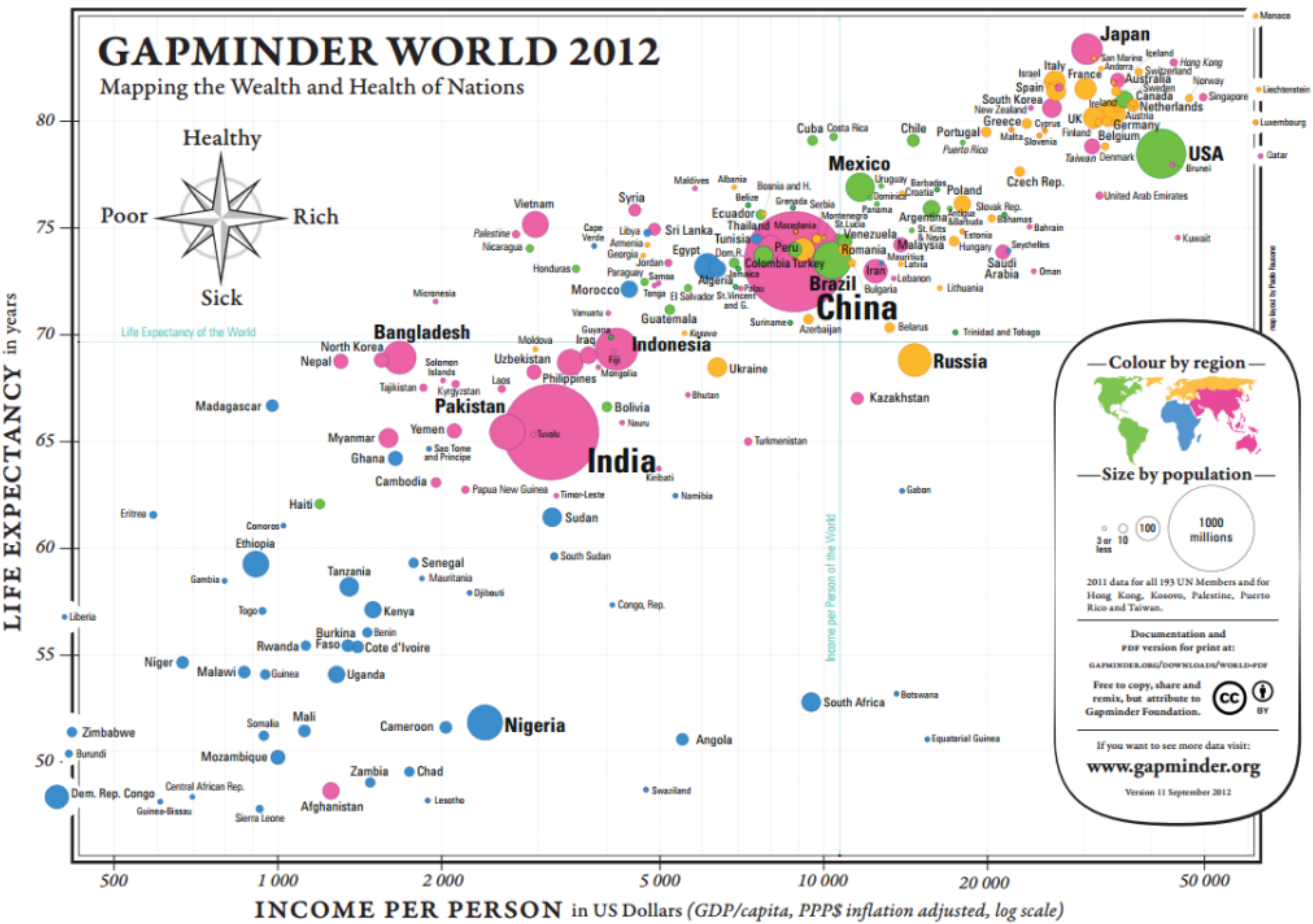


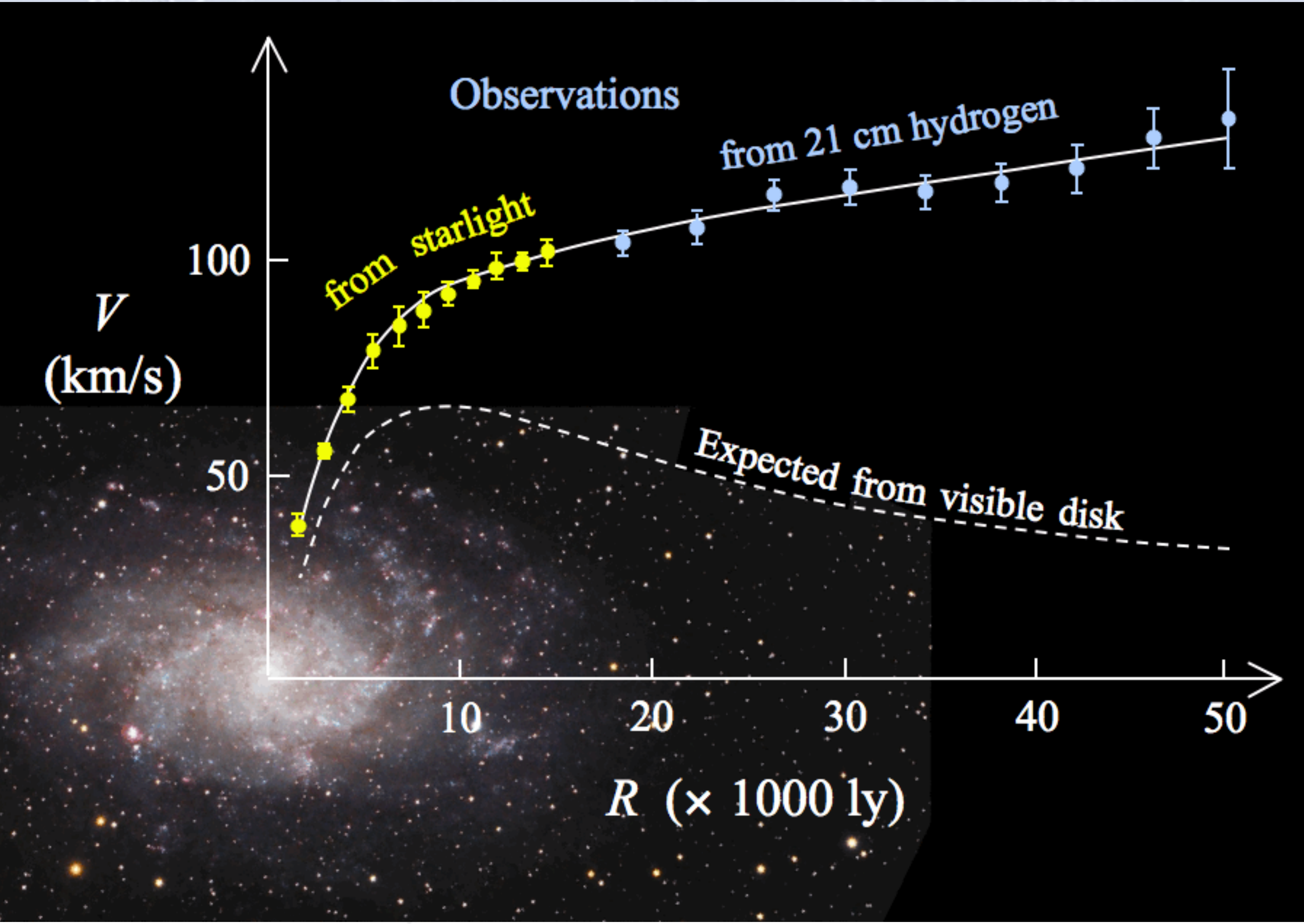


Production research
Manufacturing
Material science Engineering
Law
Economics
Clinical trials
Clinical Pharmacology
Thermodynamics
Applied physics
Demographics
Sociology
International studies
Statistical physics
Electrochemistry
Statistics
Public health
Physical chemistry
Polymers
Dermatology
Asian studies
Organic chemistry
Philosophy
Analytical Chemistry
Alternative energy
Social work
Child Psychology
Education
Religion
Social and personality psychology
Biochemistry
Pharmaceutical research
Nursing
Anthropology
Psychology
Chemical Engineering
Archeology
Human geography
Music
Toxicology
Architecture Design
Cognitive Science
Microbiology
Ecology
Sports medicine
Biotechnology
Classical studies
Plant biology
Brain studies
Biodiversity
Nutrition
Geography
Plant agriculture
Language
Tourism
Environmental Science
Soil/Marine biology
Animal Behavior
Neurology
Hydrology

Anscombe's Quartet




GAPMINDER WORLD 2012
Mapping the Wealth and Health of Nations
Healthy
Poor
Rich
Sick
LIFE EXPECTANCY in years
INCOME PER PERSON in US Dollars (GDP/capita, PPP$ inflation adjusted, log scale)
500
1 000
2 000
5 000
10 000
20 000
50 000
80
75
70
65
60
55
50
Life Expectancy of the World
Income per Person of the World
Japan
USA
Mexico
China
Brazil
Russia
Indonesia
Bangladesh
Pakistan
India
Nigeria
South Africa
Colour by region
Size by population
3 or less
10
100
1000 millions
2011 data for all 193 UN Members and for Hong Kong, Kosovo, Palestine, Puerto Rico and Taiwan.
Documentation and PDF version for print at:
GAPMINDER.ORG/DOWNLOADS/WORLD-PDF
Free to copy, share and remix, but attribute to Gapminder Foundation.
If you want to see more data visit:
www.gapminder.org
Version 11 September 2012



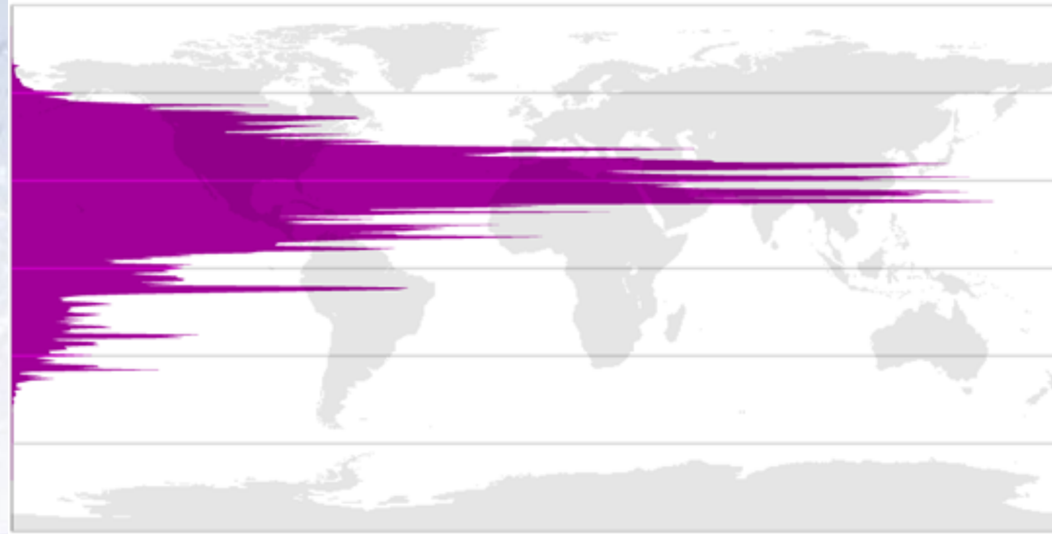
Observations
from 21 cm hydrogen
from starlight
100
50
V
(km/s)
Expected from visible disk
10
20
30
40
50
R (× 1000 ly)

Backgrounds

Putting a suitable background, which is simple, yet supports the graphs show, is a very effective way of producing great plots.

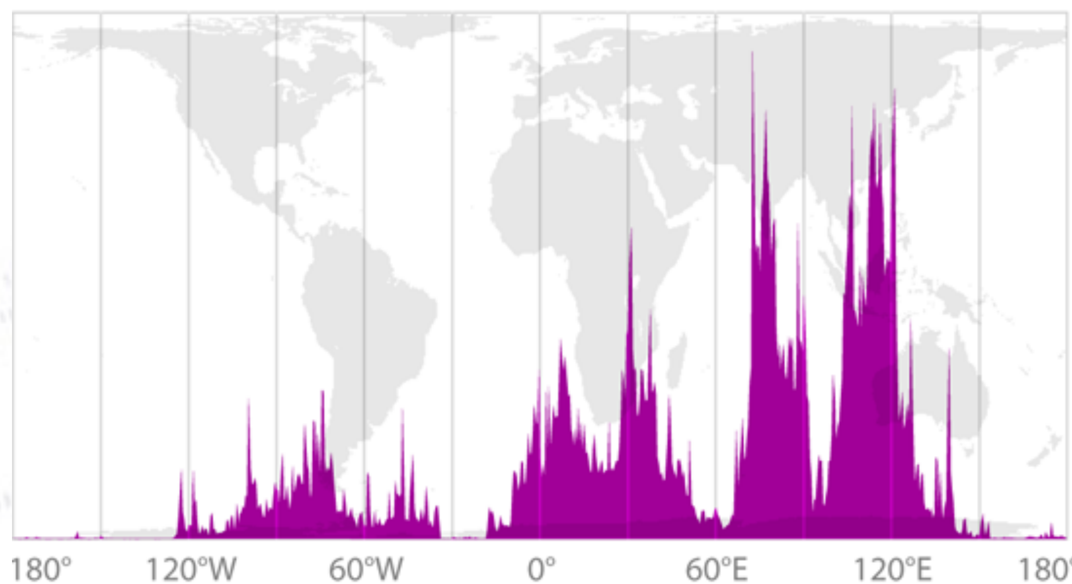
Just imagine these plots without the world map. Almost useless…

The World's Population in 2000, by Latitude



(horizontal axis shows the sum of all population at each degree of latitude)

The World's Population in 2000, by Longitude



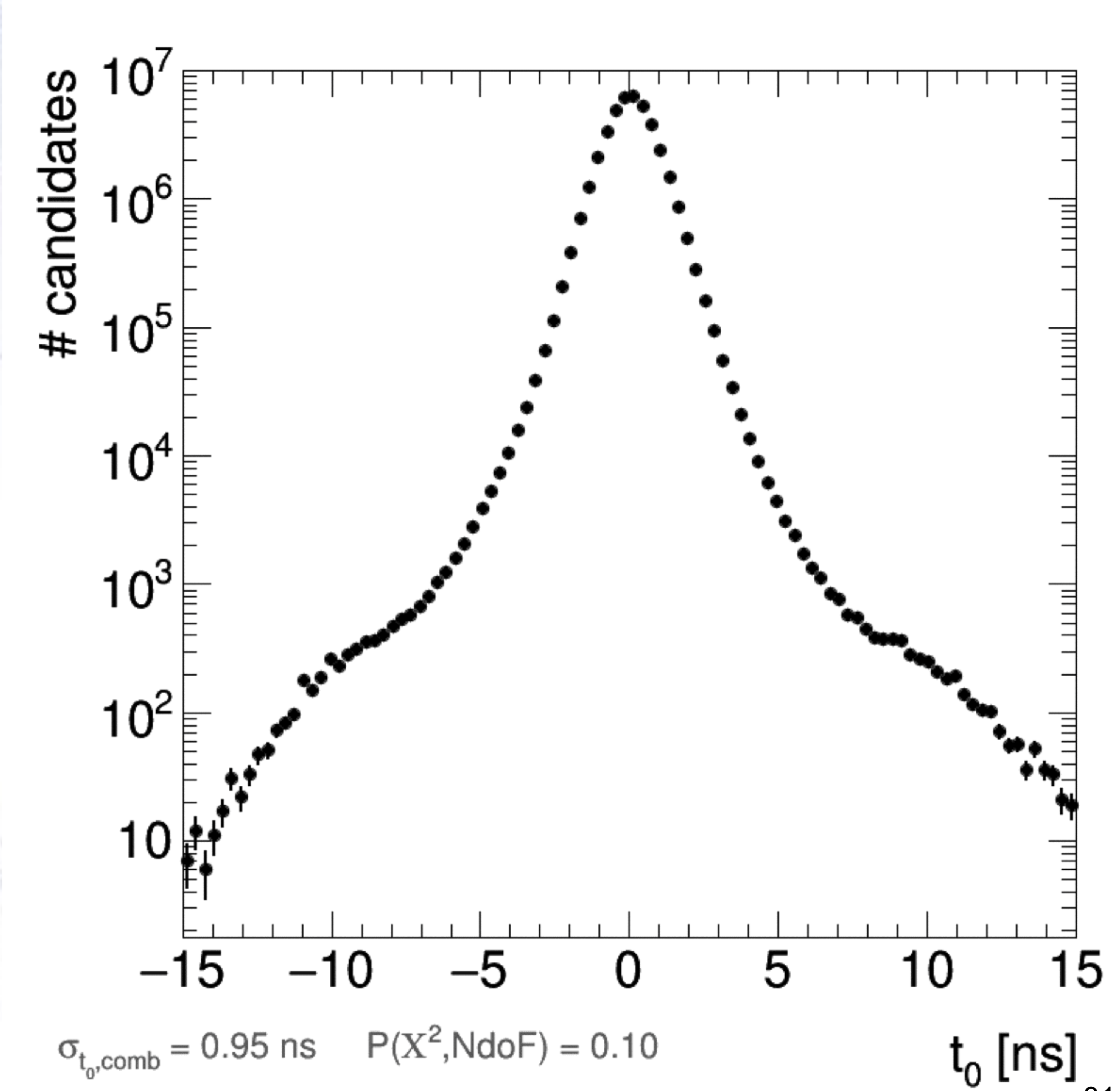
(vertical axis shows the sum of all population at each degree of longitude)

Animations of plots

Once you know how to generate plots en masse, it is surprisingly simple to make short animations illustrating the effect of the changes between plots.

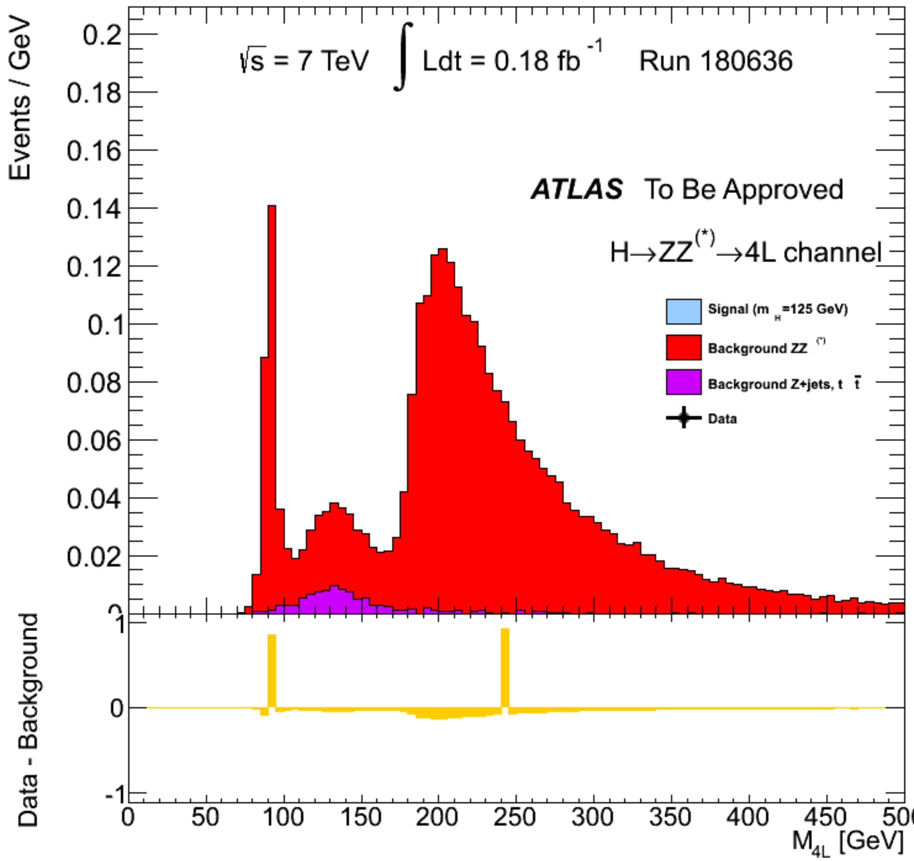
This is a very powerful way to make you points come across, typically well remembered by your audience.

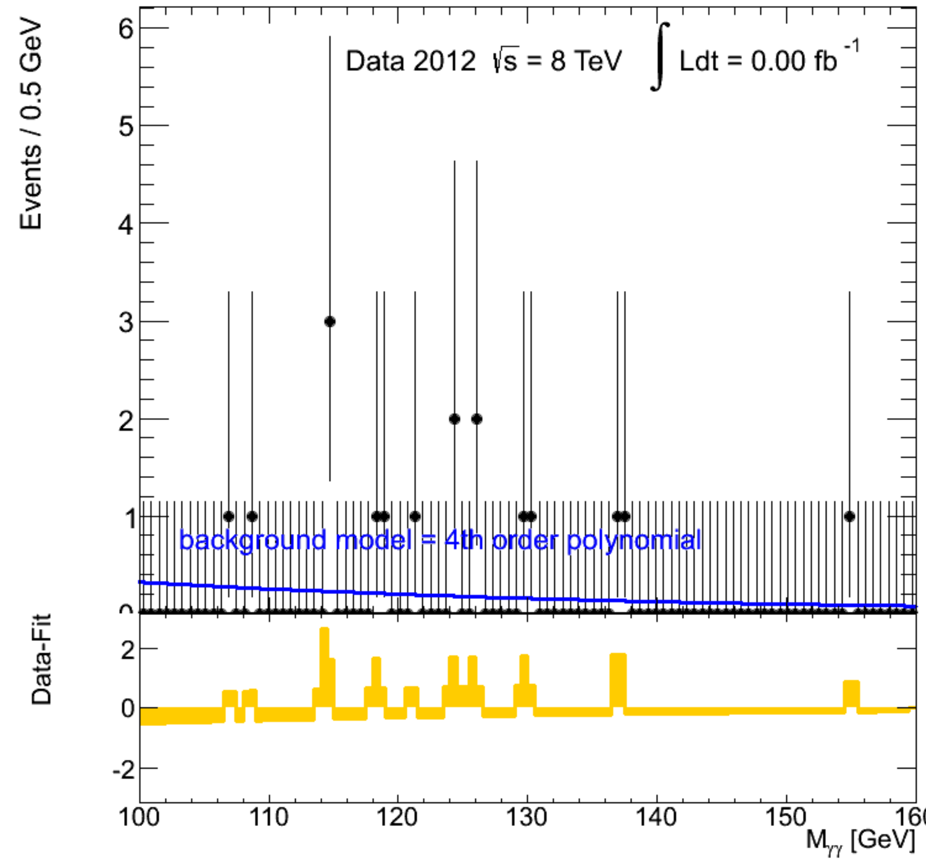
Here is shown the timing of particles in ATLAS, and evidence for observing the small ±5ns satellite bunches.



Animations of plots

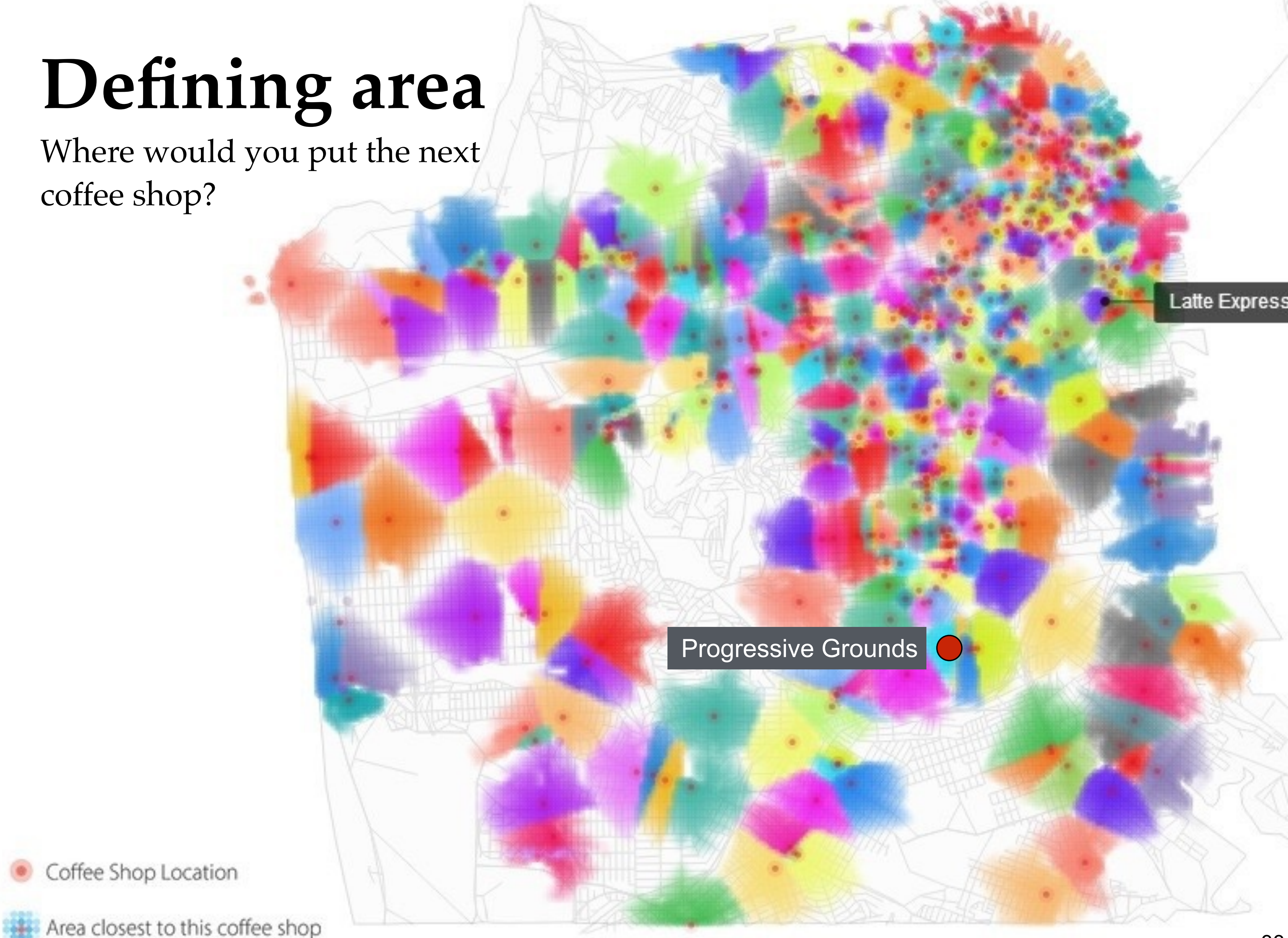
Following the Higgs discovery, the ATLAS collaboration produced the following two animations, which show the buildup of Higgs signal with time/data.





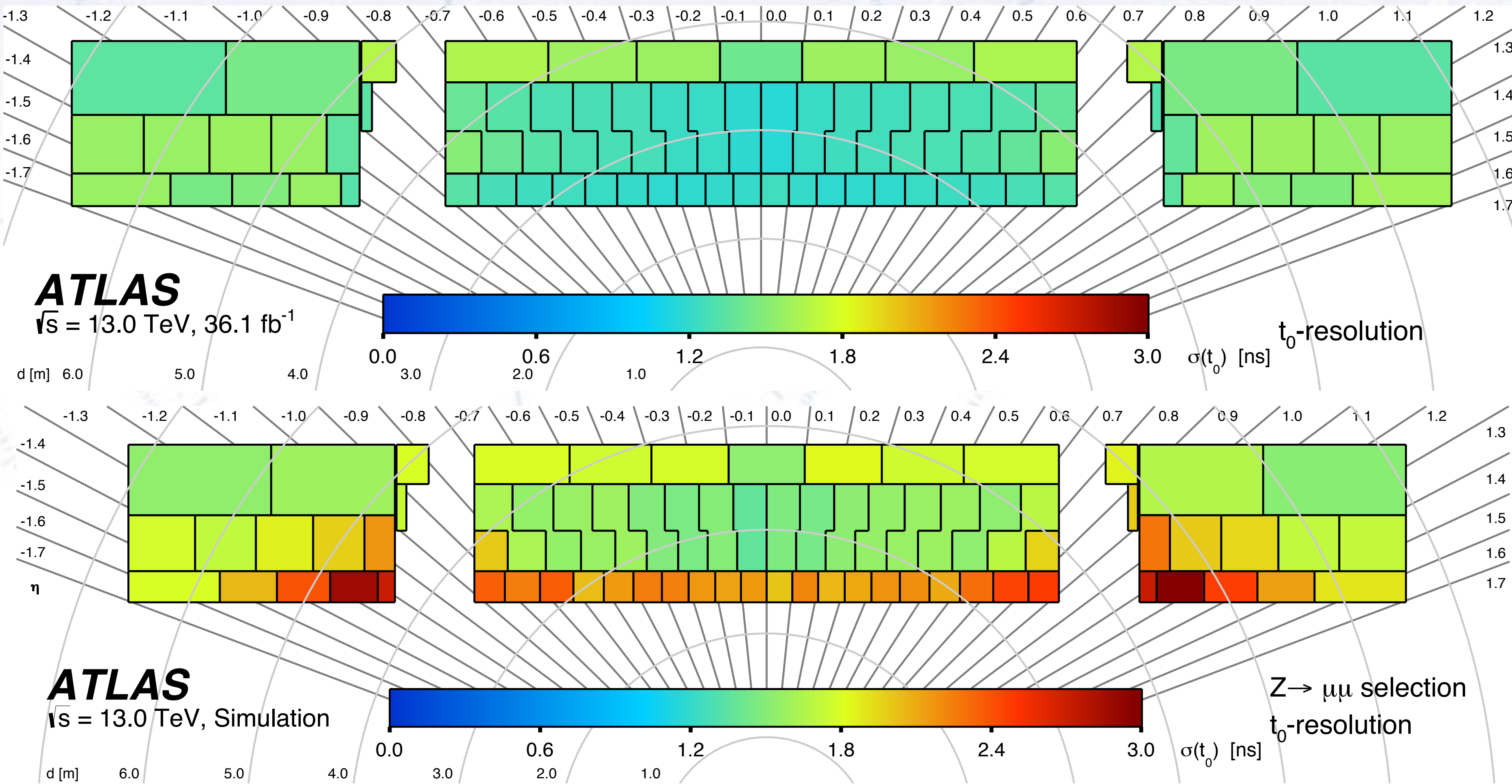
Defining area

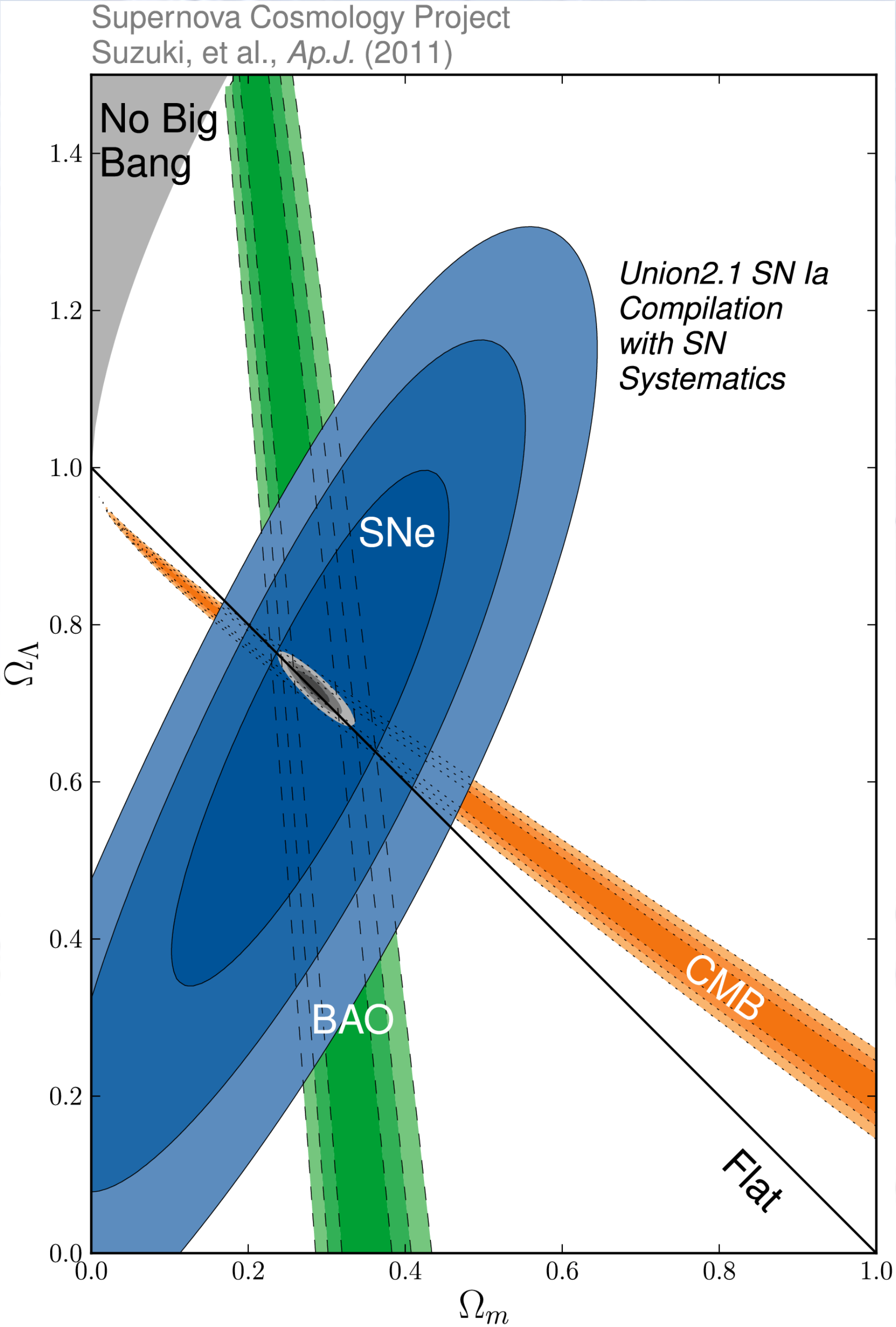
Where would you put the next coffee shop?

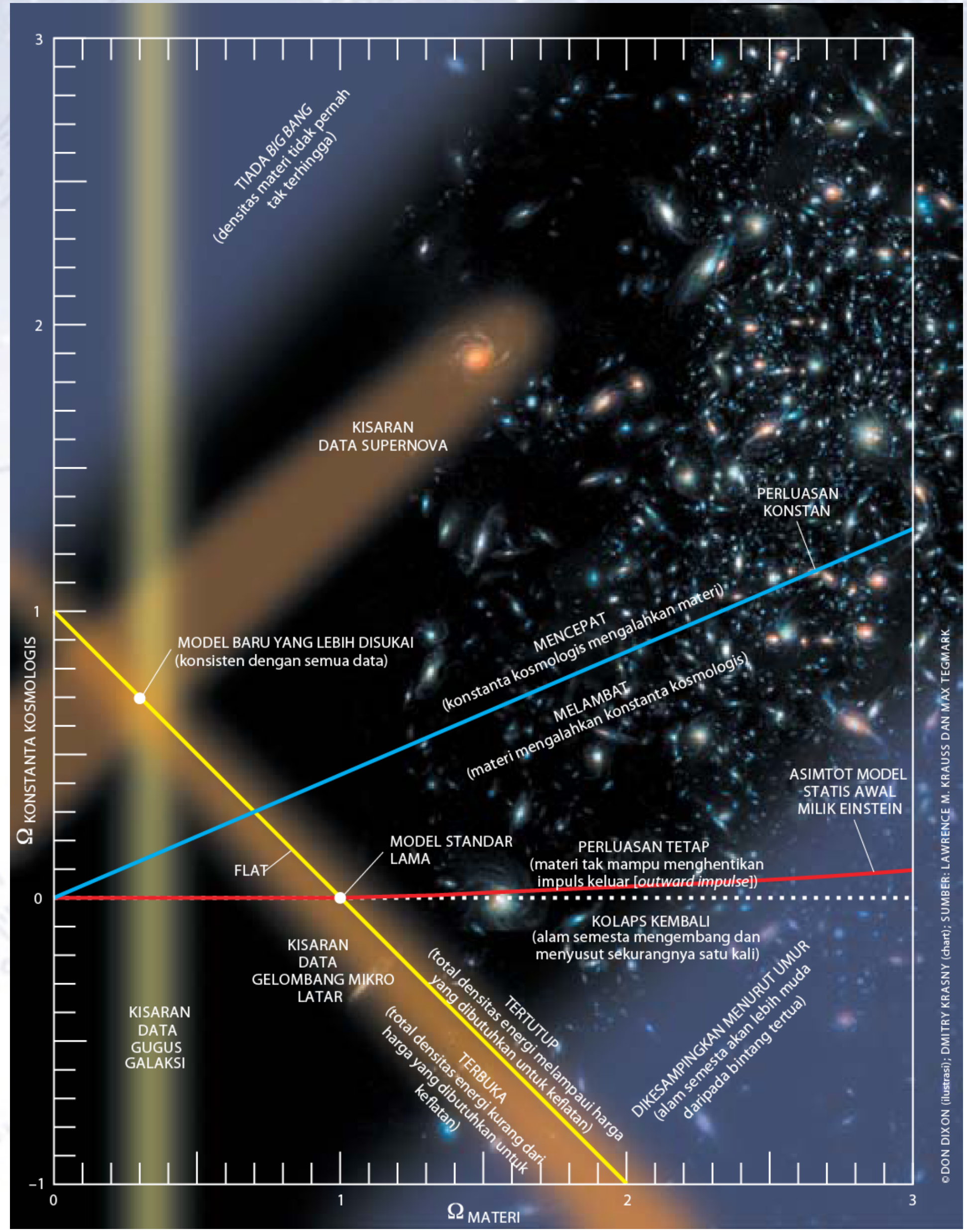


Showing detector performance

The following plot took a while to produce, but have since become defining for the hadronic calorimeter in ATLAS. It shows where the detector performs best, and also the differences between data (top) and simulation (bottom).



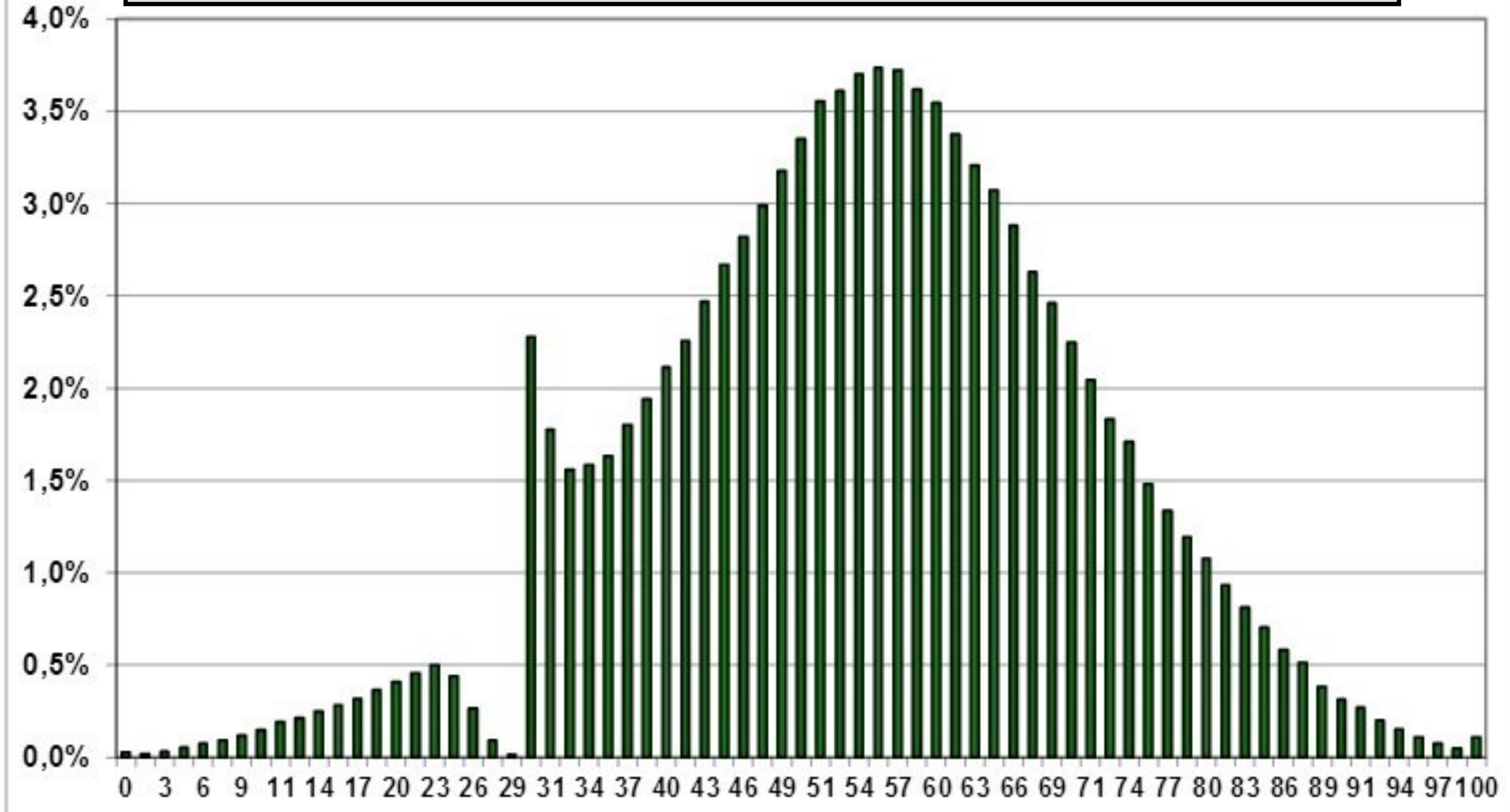

Supernova Cosmology Project
Suzuki, et al., *Ap.J.* (2011)
No Big Bang
Union2.1 SN Ia Compilation with SN Systematics
SNe
BAO
CMB
Flat
Ω_Λ
Ω_m
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

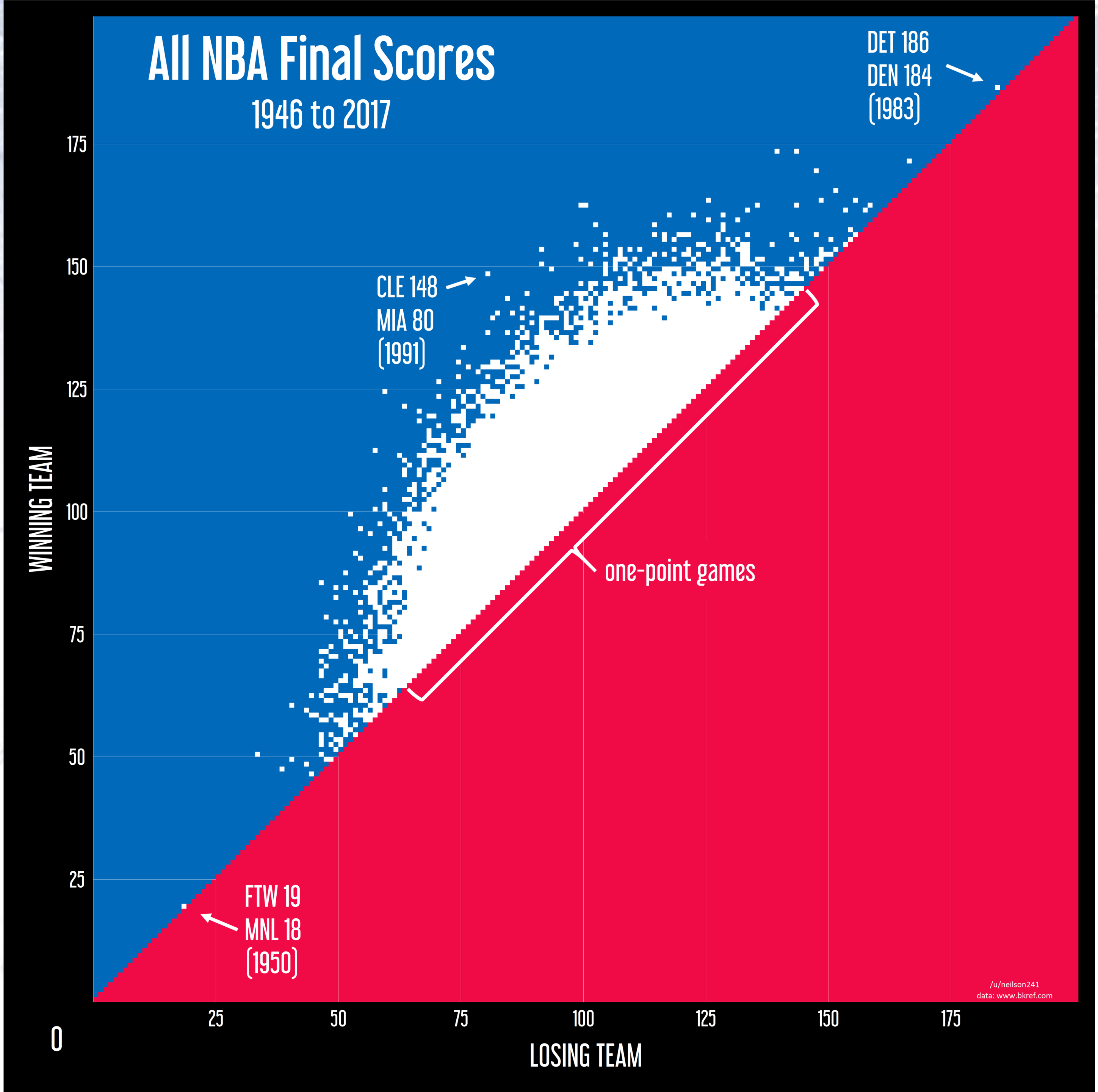

TIADA BIG BANG
(densitas materi tidak pernah tak terhingga)
KISARAN DATA SUPERNOVA
PERLUASAN KONSTAN
MODEL BARU YANG LEBIH DISUKAI
(konsisten dengan semua data)
MENCEPAT
(konstanta kosmologis mengalahkan materi)
MELAMBAT
(materi mengalahkan konstanta kosmologis)
ASIMTOT MODEL STATIS AWAL MILIK EINSTEIN
MODEL STANDAR LAMA
PERLUASAN TETAP
(materi tak mampu menghentikan impuls keluar [*outward impulse*])
FLAT
KOLAPS KEMBALI
(alam semesta mengembang dan menyusut sekurangnya satu kali)
KISARAN DATA GELOMBANG MIKRO LATAR
KISARAN DATA GUGUS GALAKSI
TERTUTUP
(total densitas energi melampaui harga yang dibutuhkan untuk keflatan)
TERBUKA
(total densitas energi kurang dari harga yang dibutuhkan untuk keflatan)
DIKESAMPINGKAN MENURUT UMUR
(alam semesta akan lebih muda daripada bintang tertua)
Ω KONSTANTA KOSMOLOGIS
Ω MATERI
-1
0
1
2
3
©DON DIXON (ilustrasi); DMITRY KRASNY (chart); SUMBER: LAWRENCE M. KRAUSS DAN MAX TEGMARK

Distributions and social effects

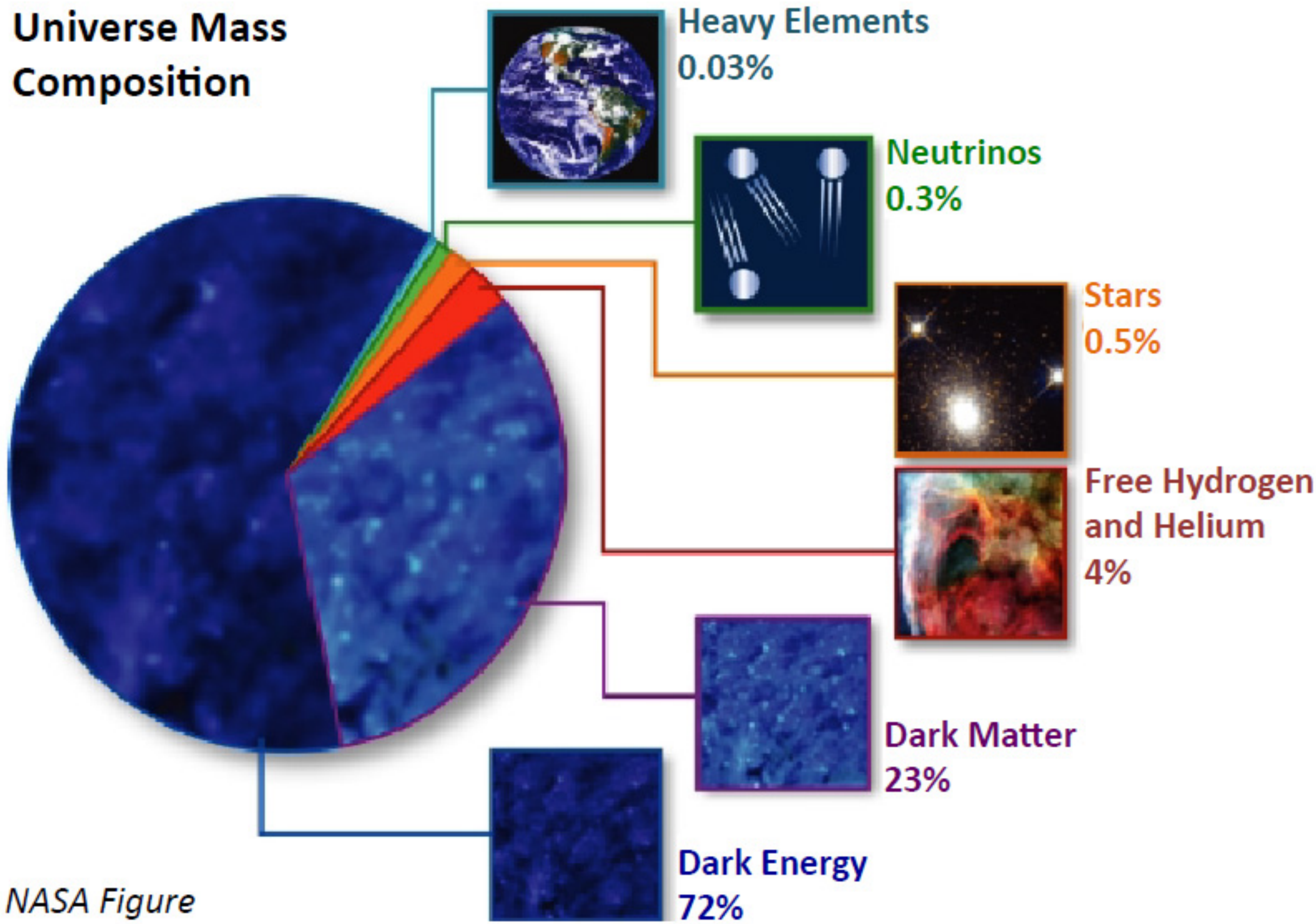
2.1. Poziom podstawowy

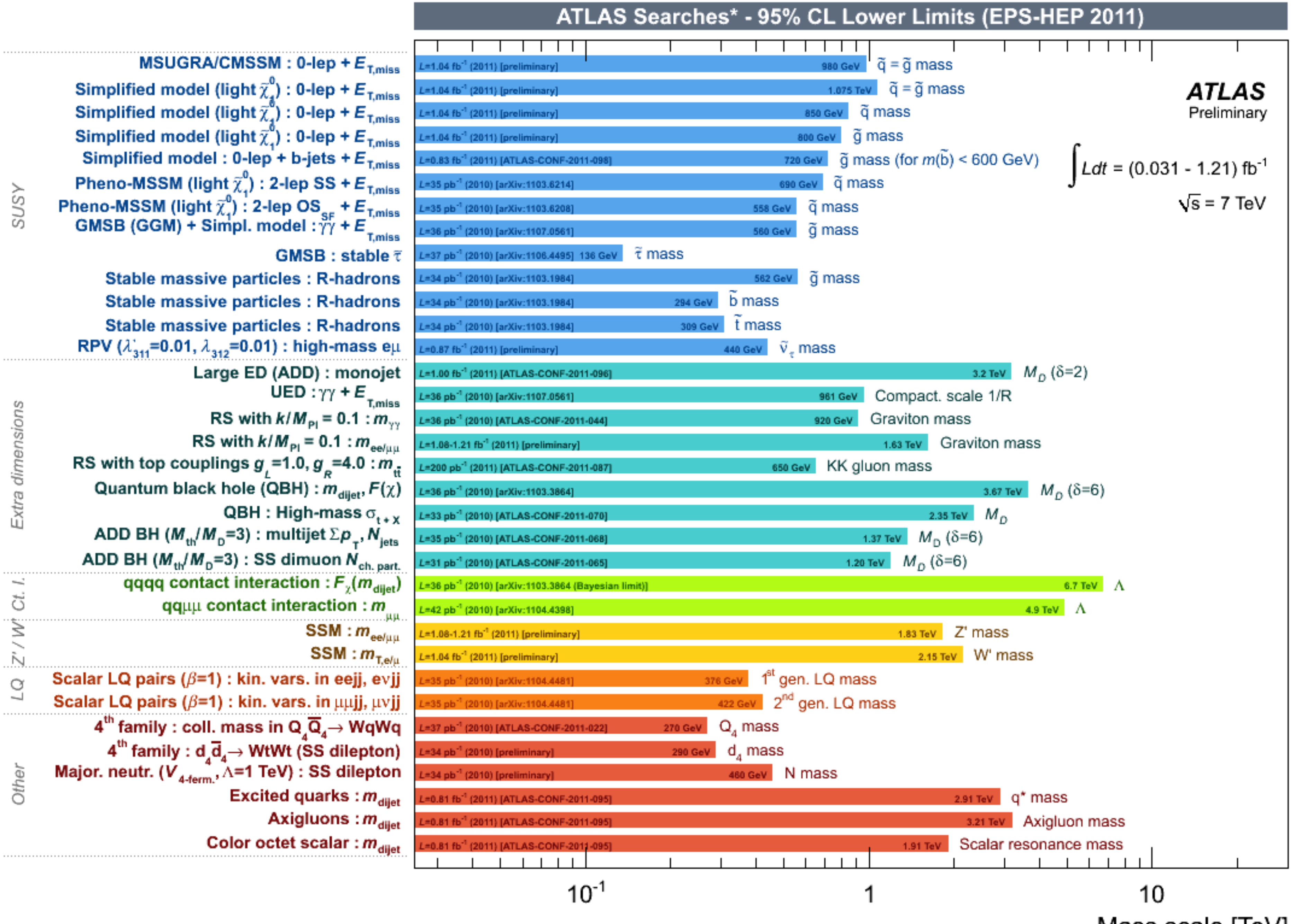
Distribution of Matura (high school exit exam) results in Poland in 2013.
The minimum score to pass is 30%.



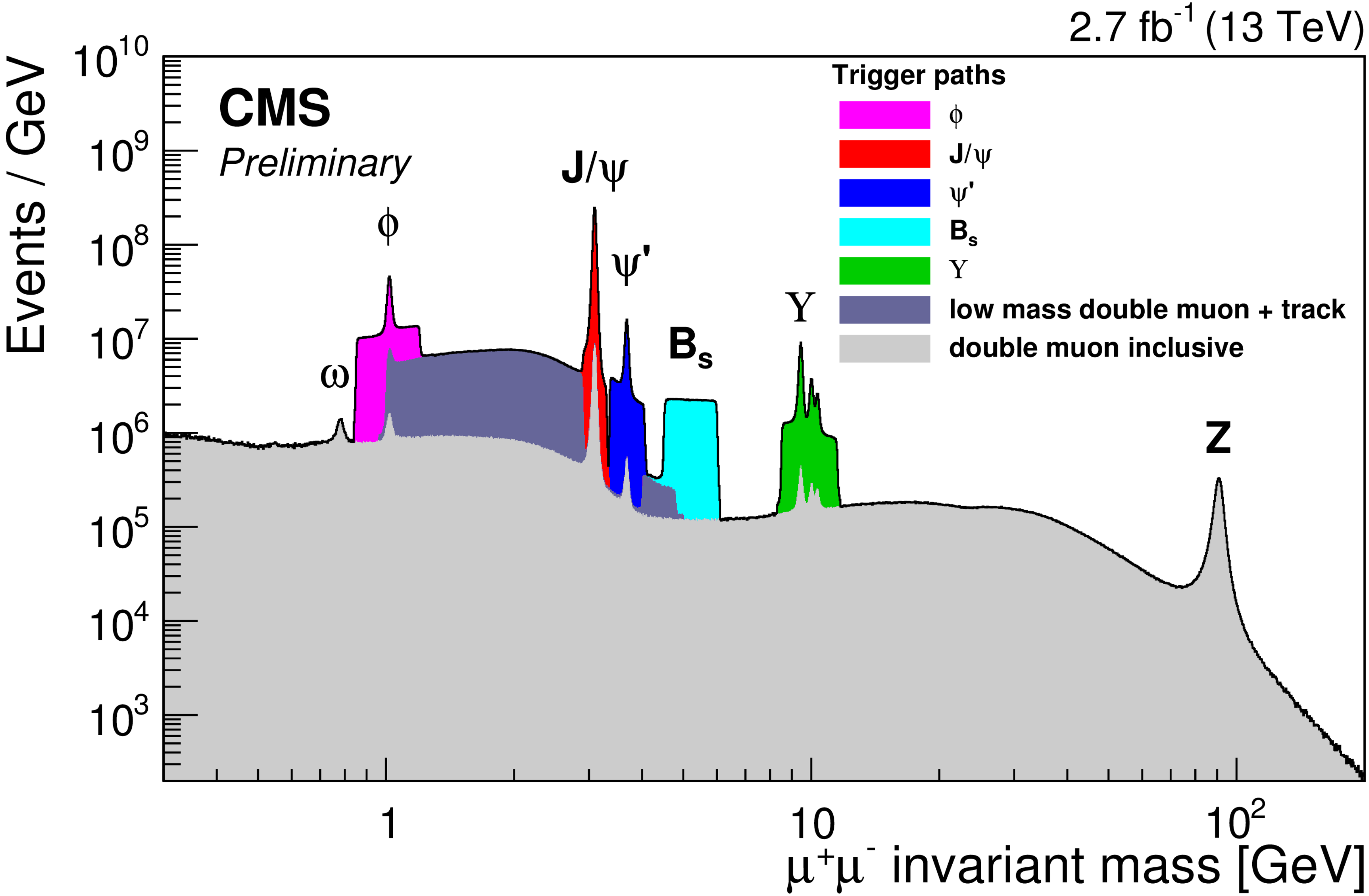


All NBA Final Scores
1946 to 2017
DET 186
DEN 184
(1983)
CLE 148
MIA 80
(1991)
one-point games
FTW 19
MNL 18
(1950)
WINNING TEAM
LOSING TEAM
175
150
125
100
75
50
25
0
25
50
75
100
125
150
175
/u/neilson241
data: www.bkref.com


Universe Mass
Composition
Heavy Elements
0.03%
Neutrinos
0.3%
Stars
0.5%
Free Hydrogen
and Helium
4%
Dark Matter
23%
Dark Energy
72%
NASA Figure



*Only a selection of the available results shown


2.7 fb^{-1} (13 TeV)
CMS
Preliminary
Events / GeV
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
ω
ϕ
J/ψ
ψ'
B$_s$
Υ
Z
Trigger paths
ϕ
J/ψ
ψ'
B$_s$
Υ
low mass double muon + track
double muon inclusive
1
10
10^2
$\mu^+\mu^-$ invariant mass [GeV]