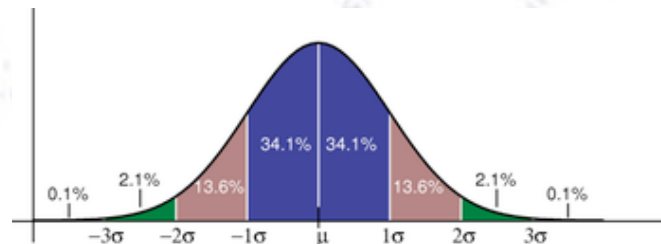


Applied Statistics

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.

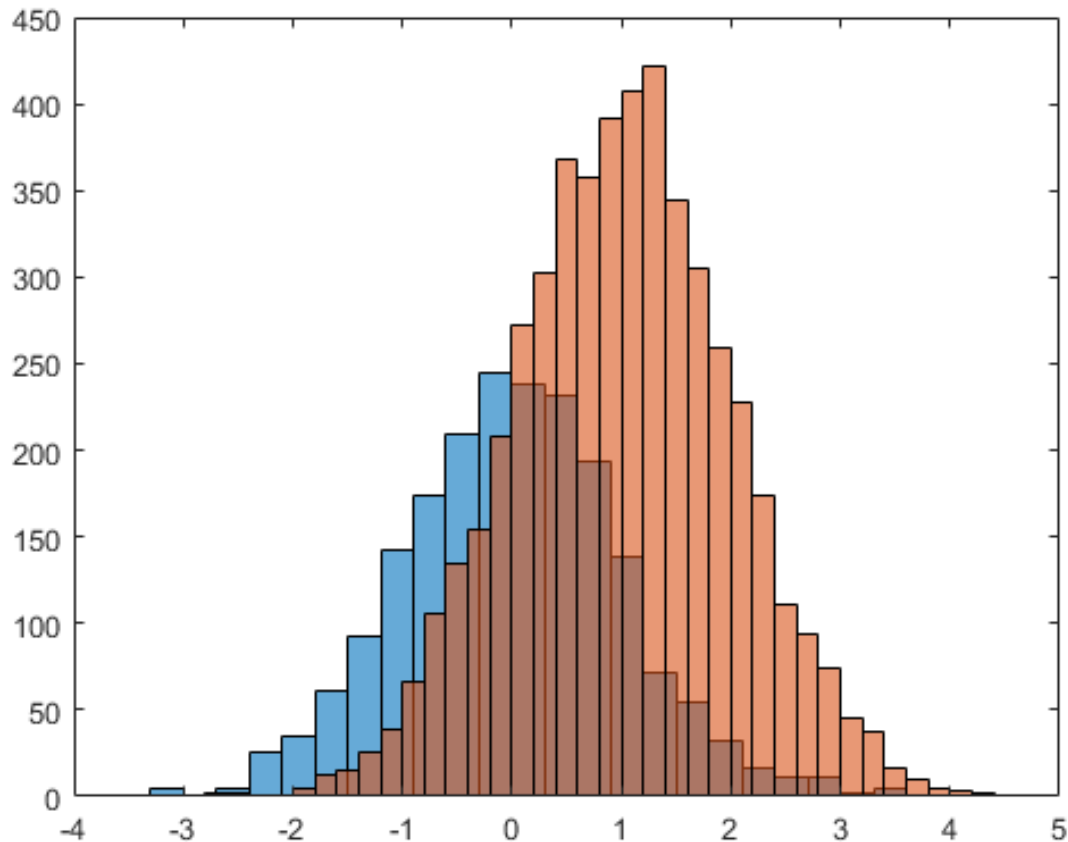
A pitfall is that continuous data is not always continuous, but may seem so!!! The problem arise, if plotting in a histogram with binning comparable (and possibly prime) to steps.

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.

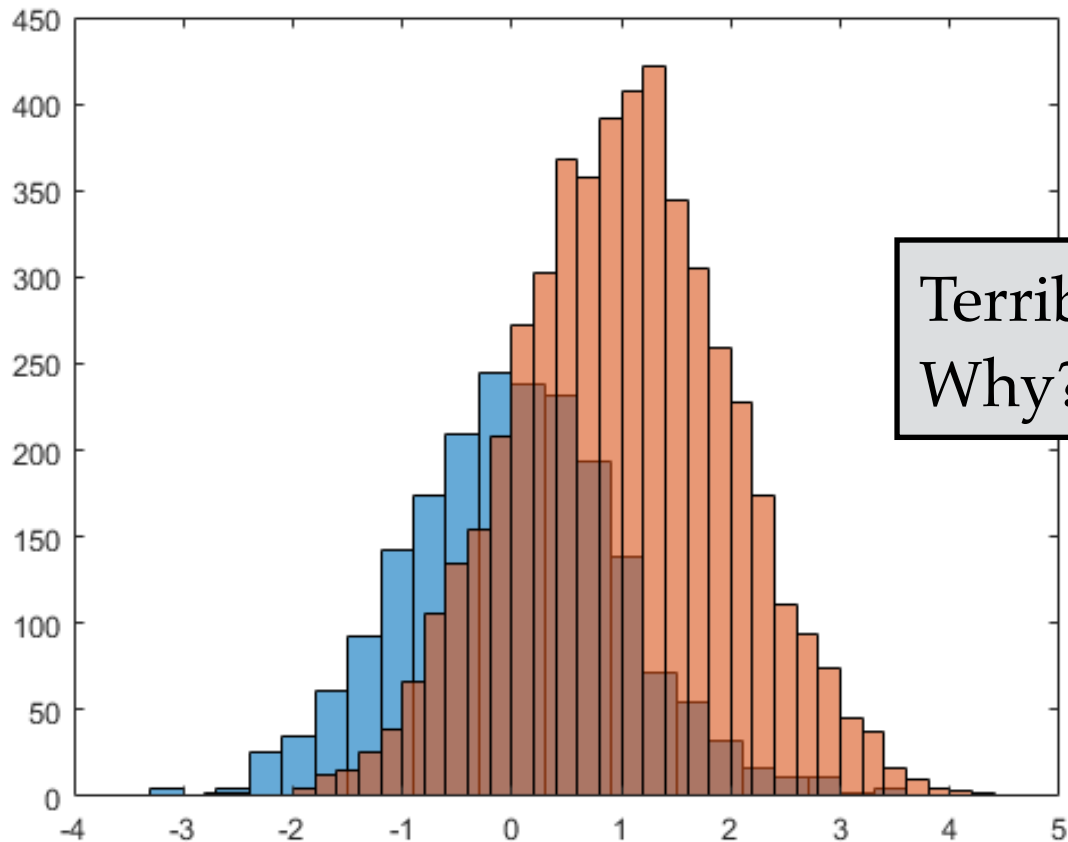
Ways of displaying data

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

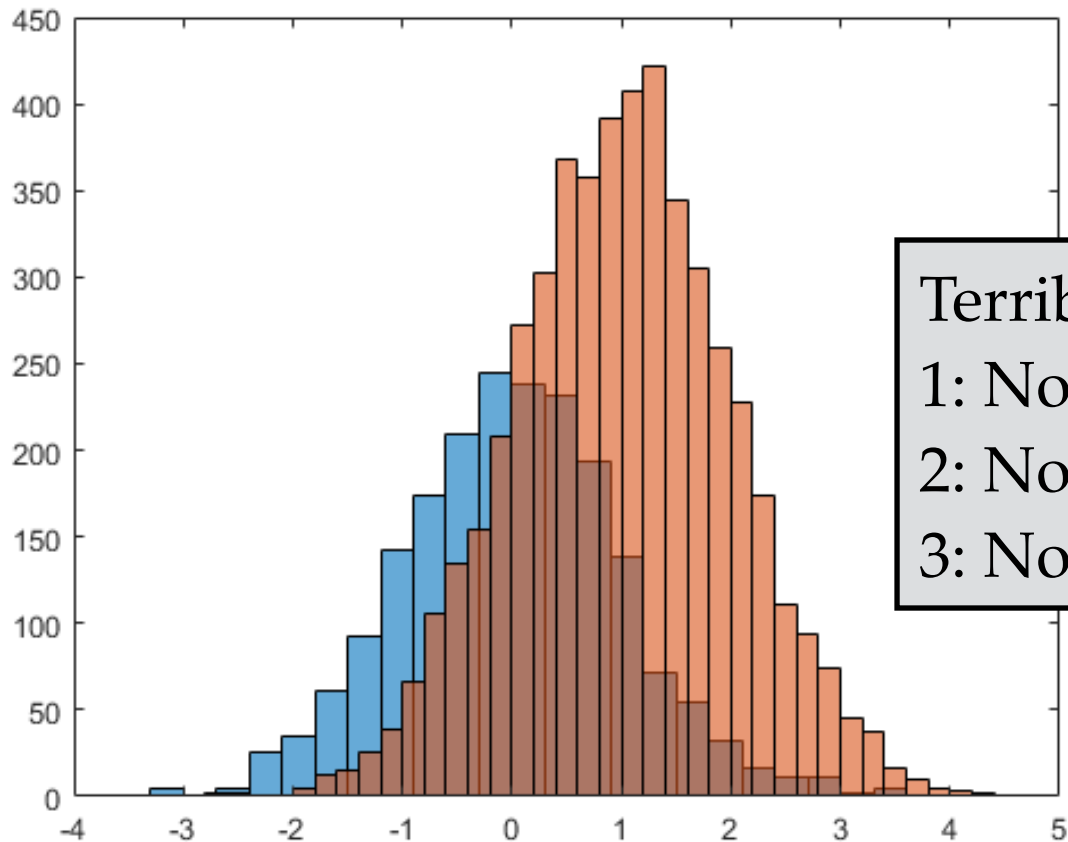
Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Terrible plot!!!
Why?

Ways of displaying data

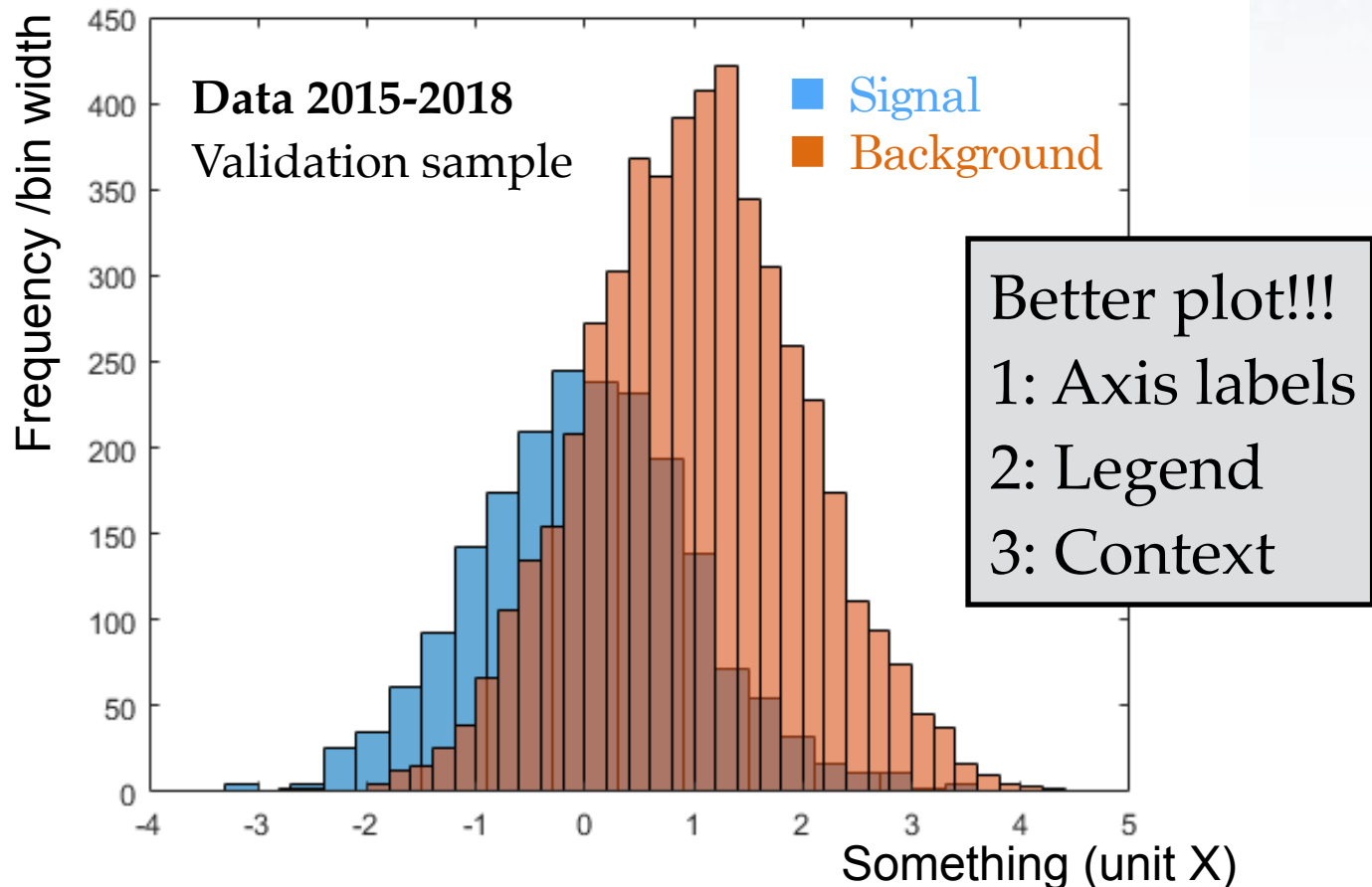
Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



Terrible plot!!!
1: No axis labels
2: No legend
3: No context

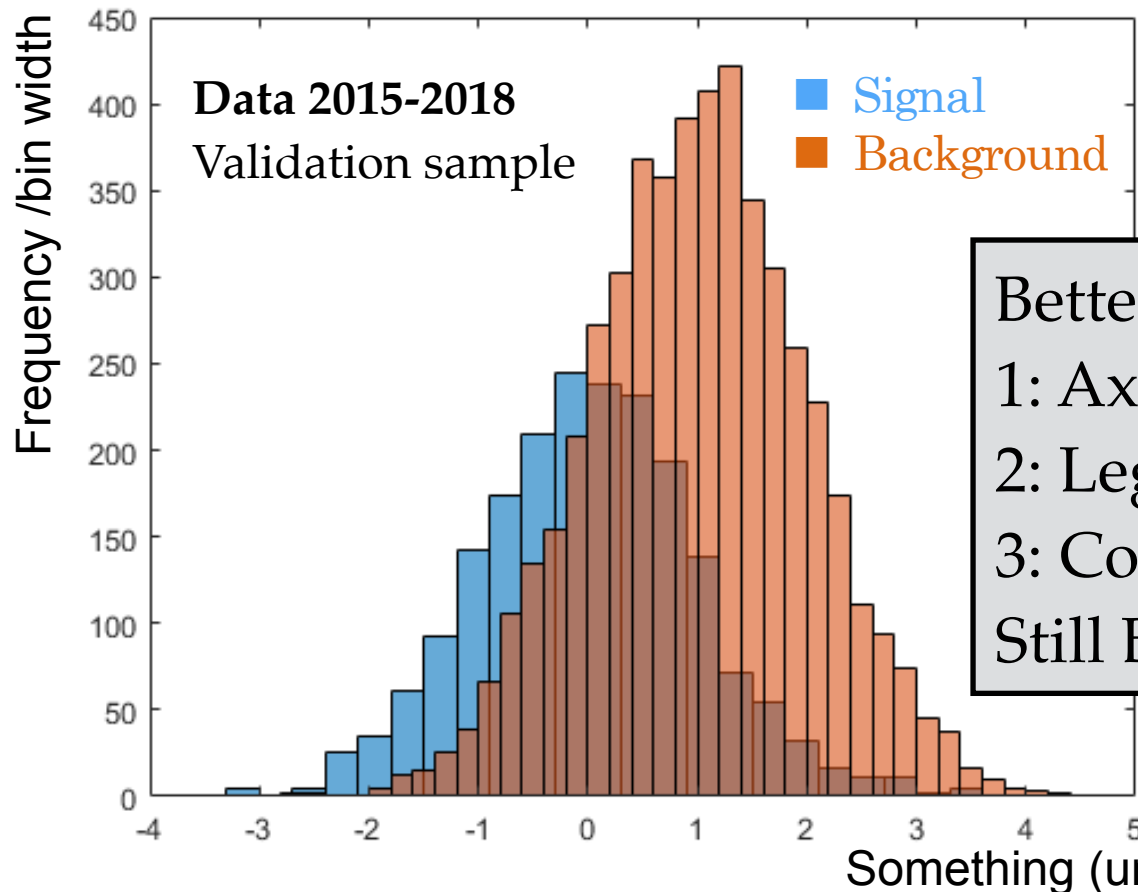
Ways of displaying data

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.



Ways of displaying data

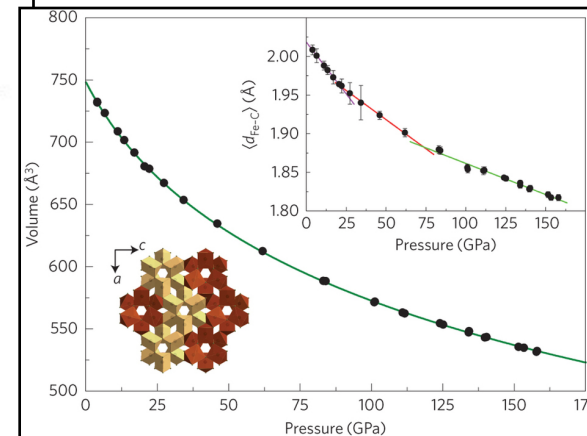
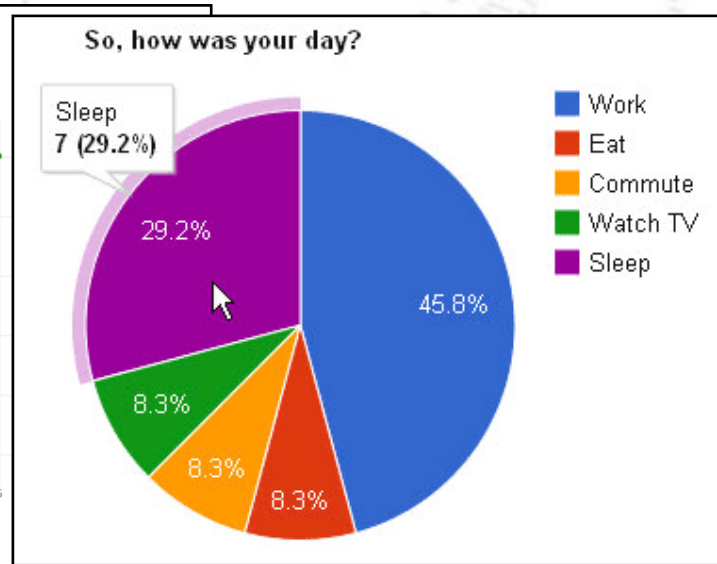
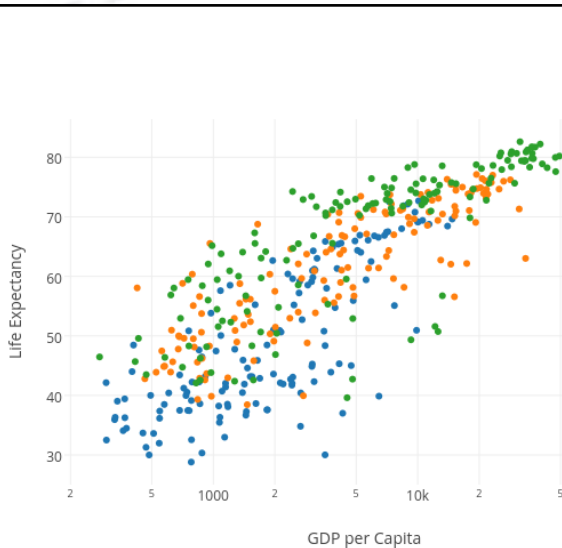
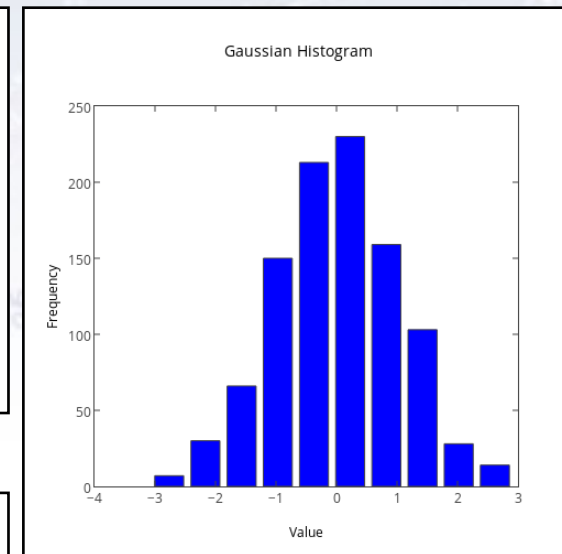
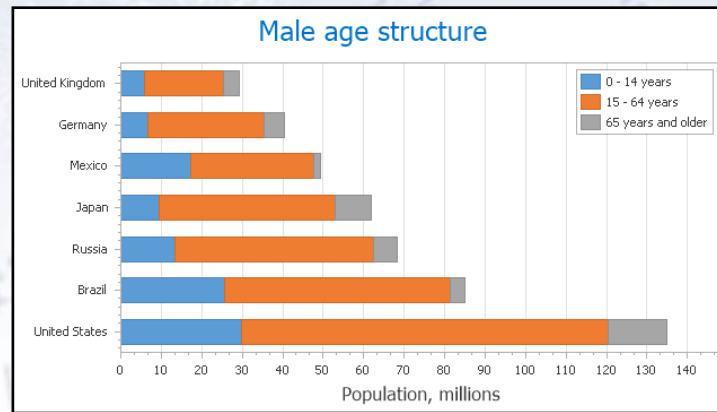
Given repeated measurement of a quantity, the most common way of displaying it, is with a 1D histogram. It is simple and easy to understand, but of course doesn't include more complex information.



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...

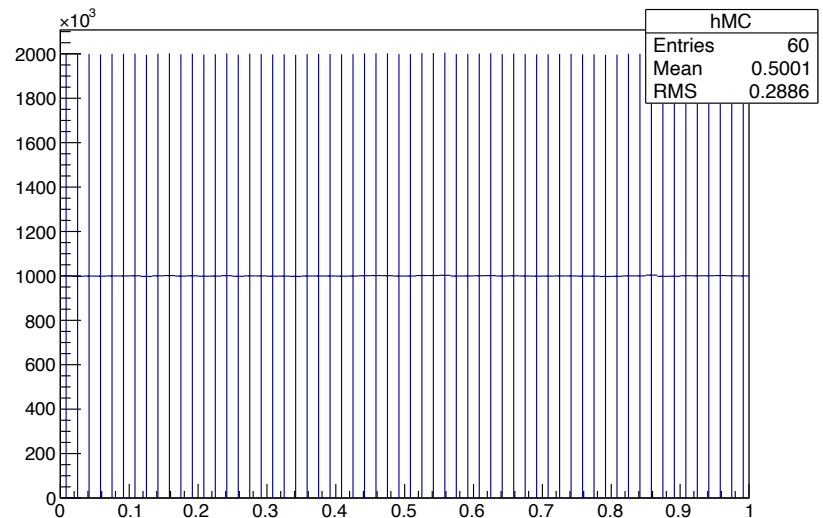
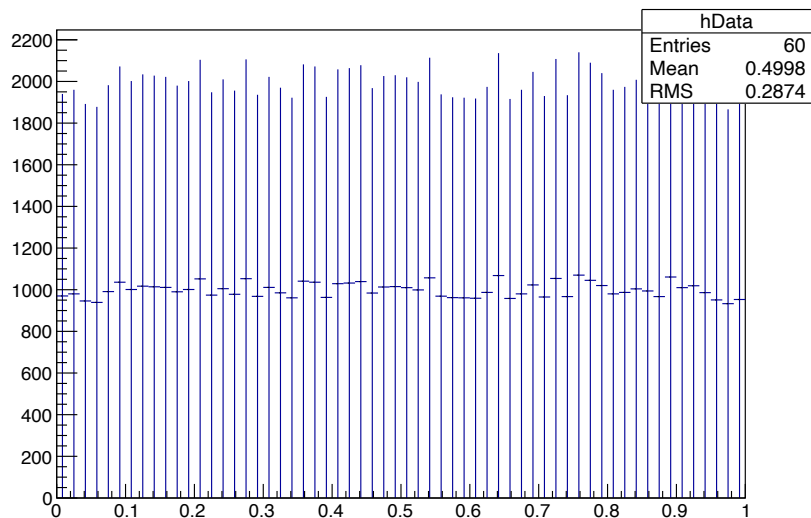
A note on plotting

Always plot your data!!!

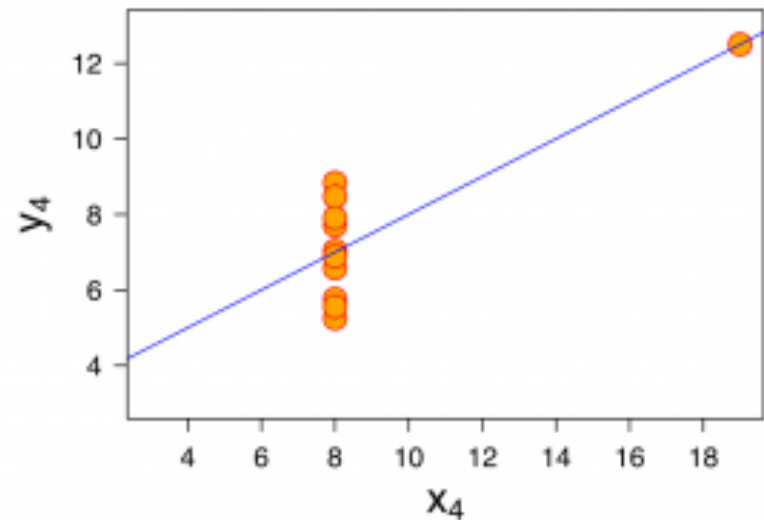
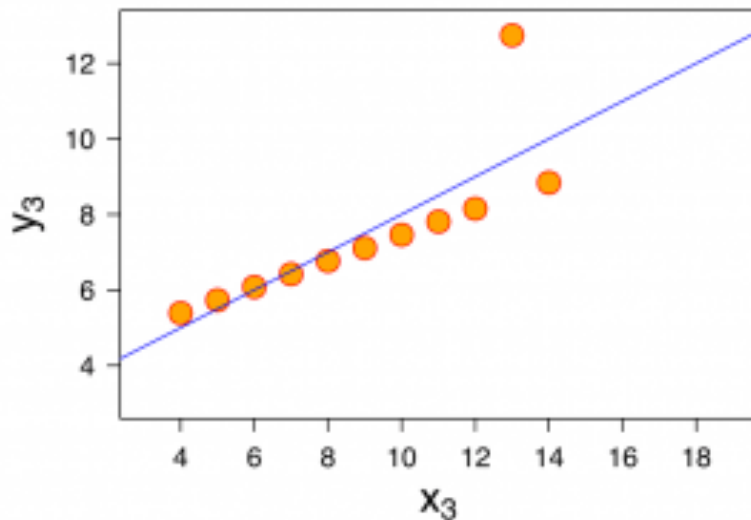
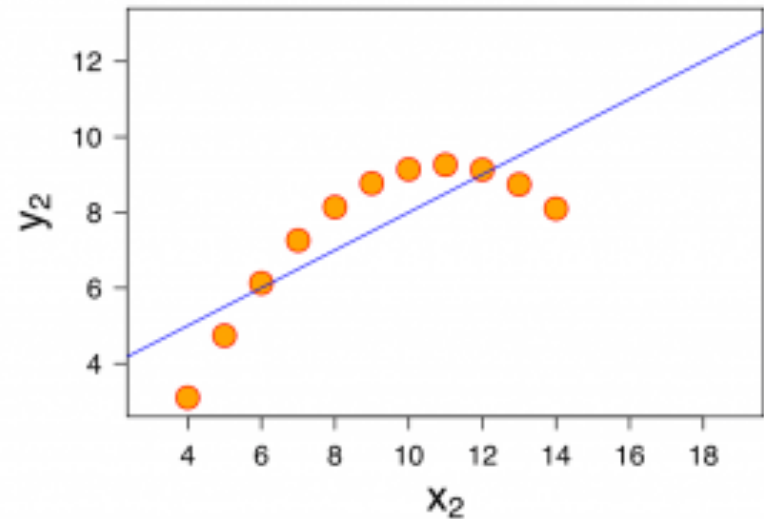
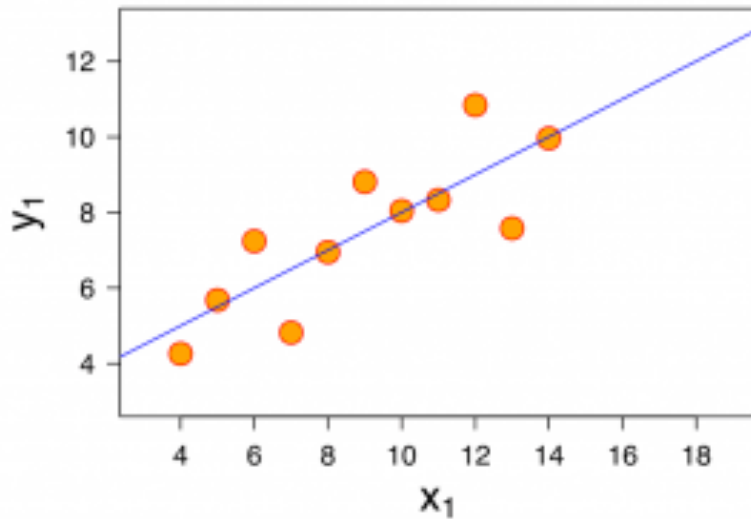
You never really know, what goes on in data, until you have SEEN it!

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...



Anscombe's Quartet





**Examples of
Poor Plots &
Illustrations**

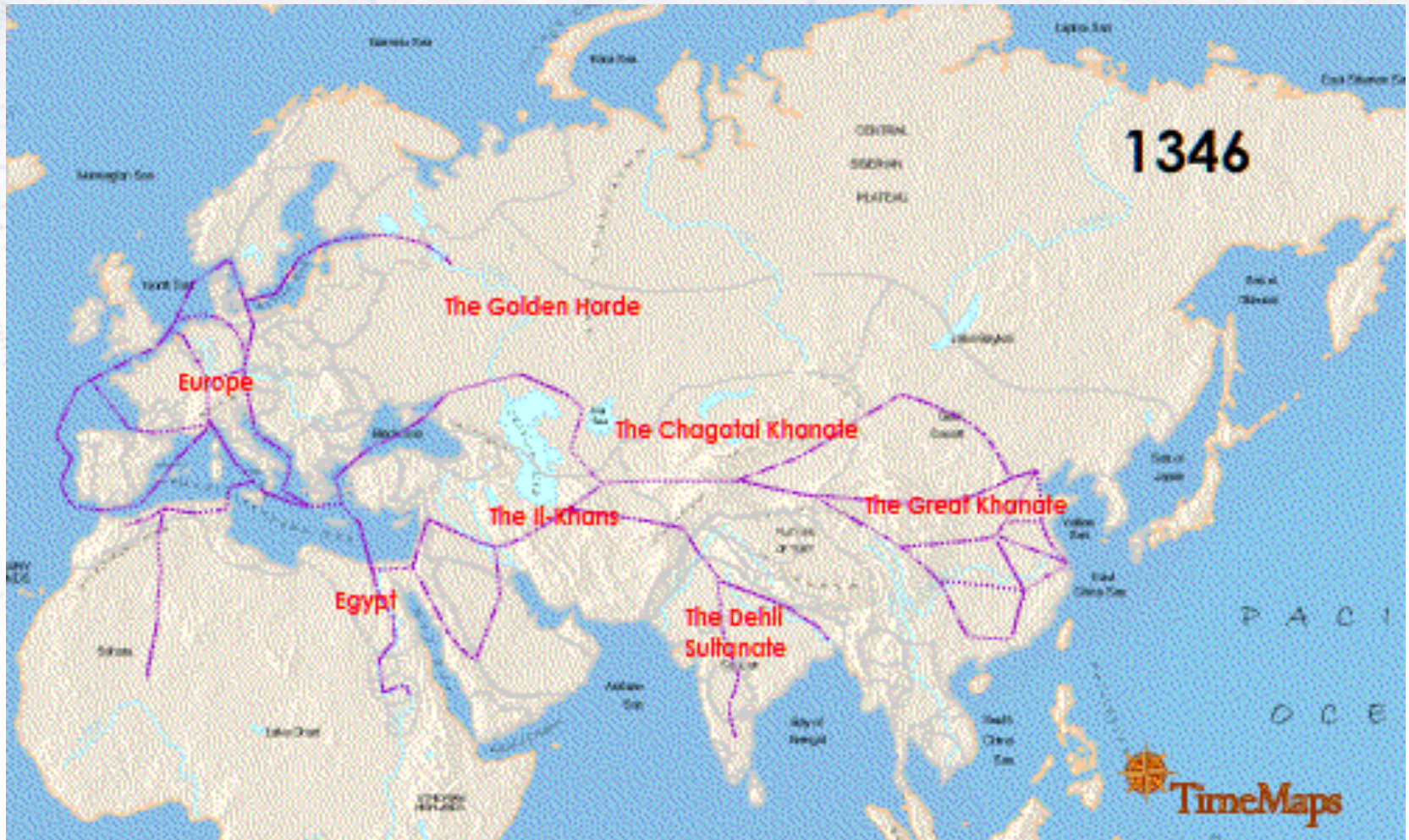
Black Death... and bad plots!

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?



AMERICANS WHO HAVE TRIED MARIJUANA

CBS NEWS POLL

51%
TODAY

43%
LAST YEAR

34%
1997



Source: MOE +/- 4%

HIGH SUPPORT FOR LEGALIZING MARIJUANA
MORE THAN HALF OF AMERICANS SAY THEY'VE TRIED POT



LIVE

CBSN

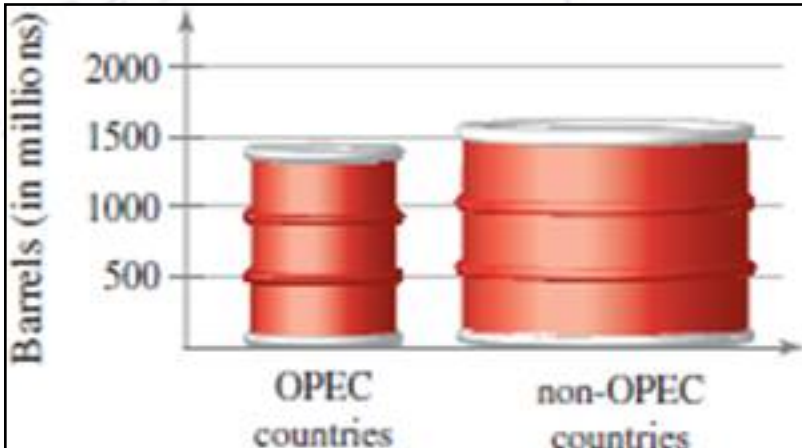
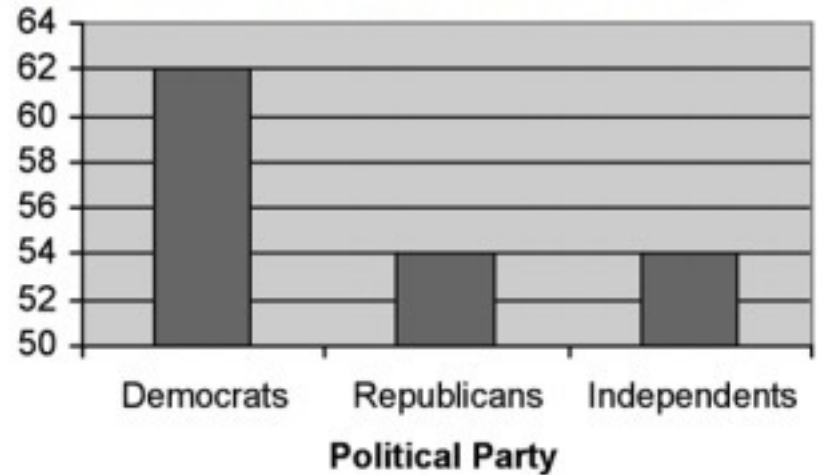
A mix of poor plots

Top 10 States in Ease of Doing Business

	Score	2015 Rank
1 Andhra Pradesh	98.78	2
2 Telangana	98.78	13
3 Gujarat	98.21	1
4 Chattisgarh	97.32	4
5 Madhya Pradesh	97.01	5
6 Haryana	96.95	14
7 Jharkhand	96.57	3
8 Rajasthan	96.43	6
9 Uttarakhand	96.13	23
10 Maharashtra	92.86	8

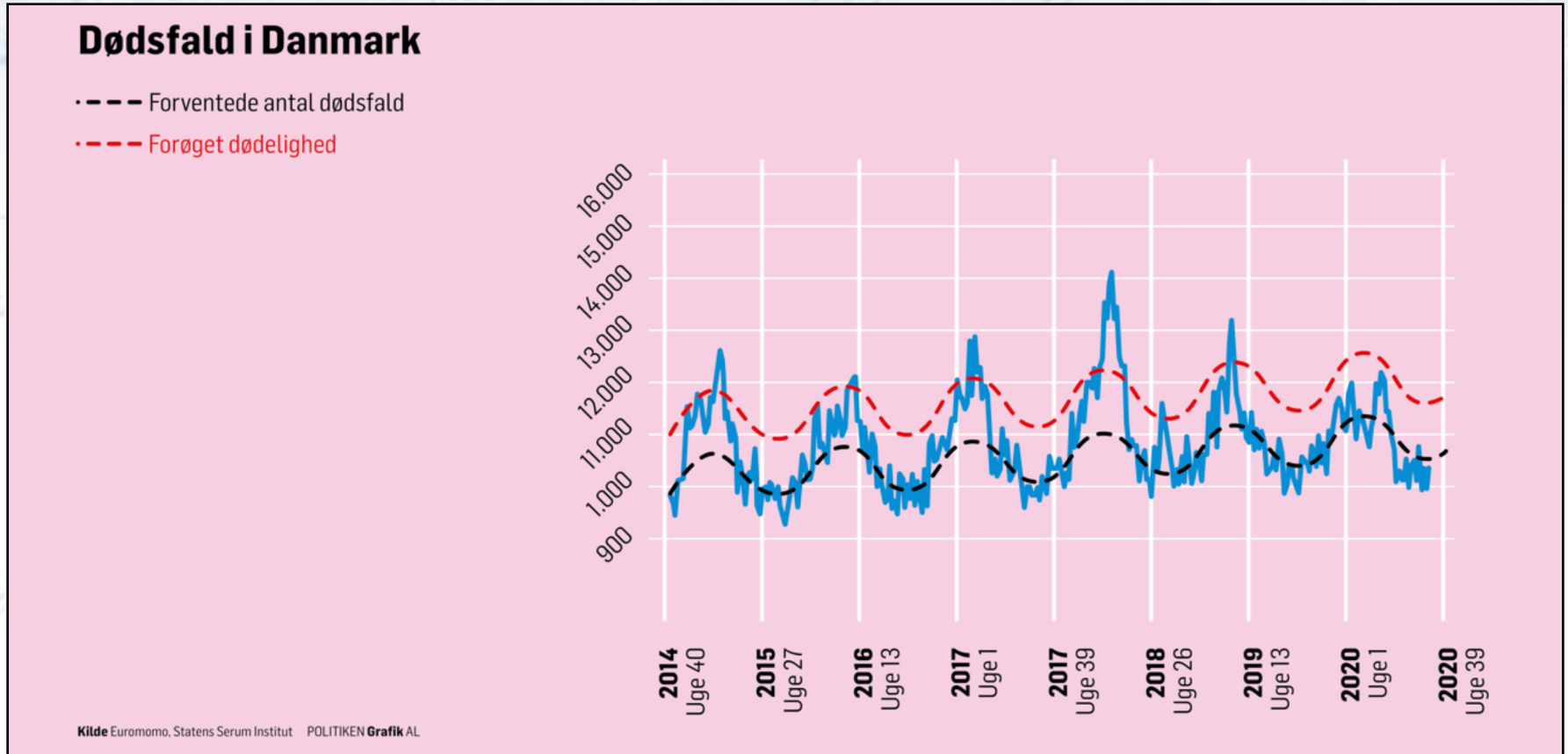
PTI GRAPHICS

Percent Who Agreed With Court



A new poor plot

Have a look at this plot (from Politiken, 11th of October 2020, sorry for the Danish), and see if you can see errors and suggest improvements.

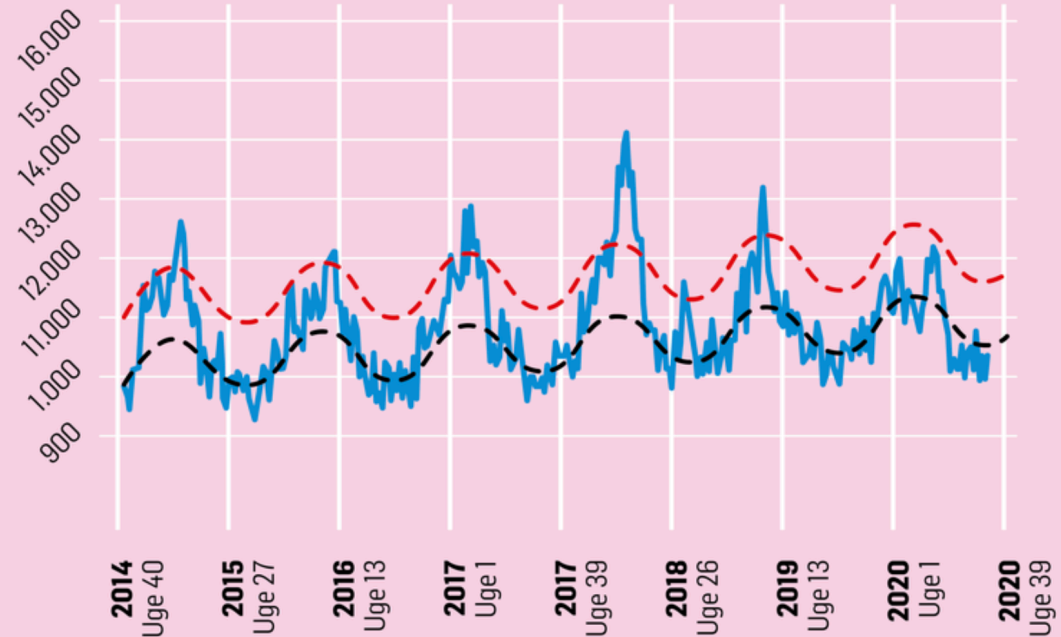


A new poor plot

Have a look at this plot (from Politiken, 11th of October 2020, sorry for the Danish), and see if you can see errors and suggest improvements.

Dødsfald i Danmark

- Forventede antal dødsfald
- Forøget dødelighed



My list:

- 1: Y-axis wrong!
- 2: X-axis very strange.
- 3: No legend for blue curve.
- 4: Models are terrible.

Kilde Euromomo, Statens Serum Institut POLITIKEN Grafik AL



**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.

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
 Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en caractères des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de L'Égur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

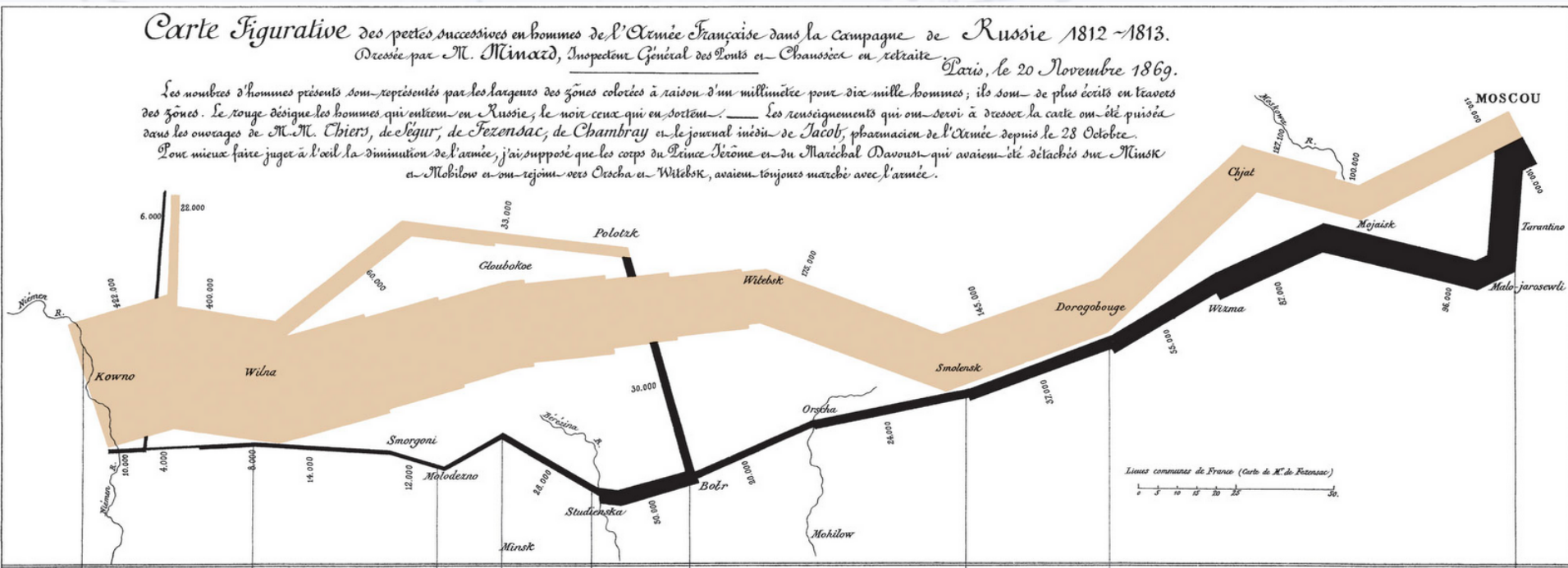
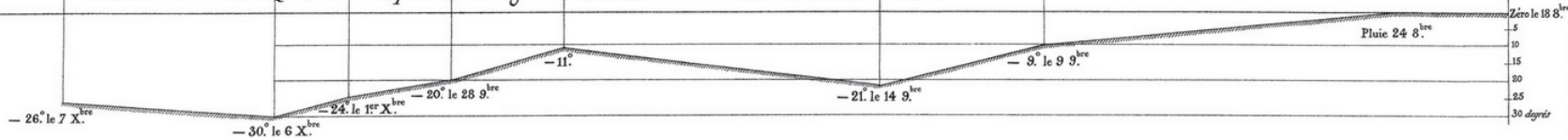


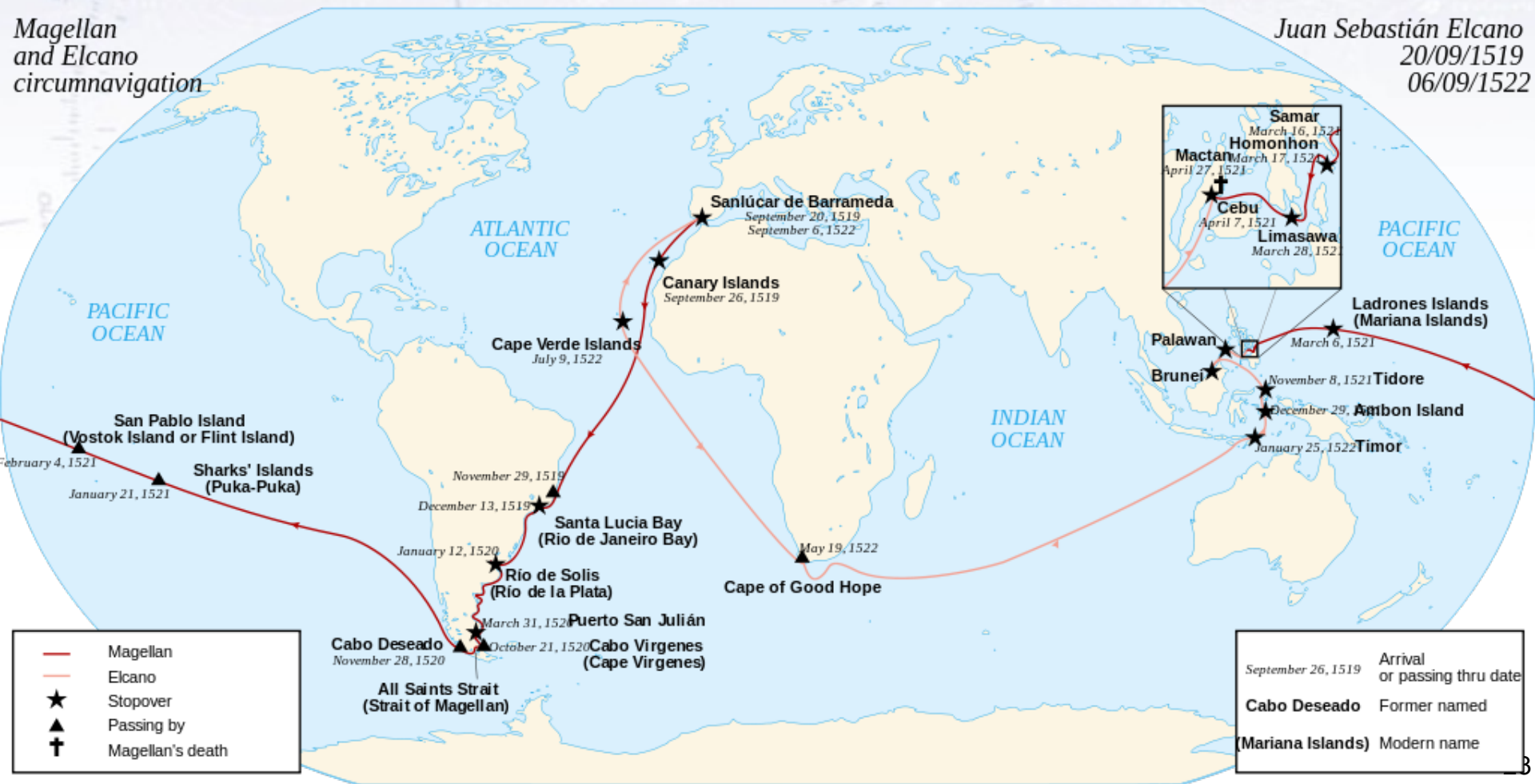
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Les Cosaques passent au galop le Niémen gelé.

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...

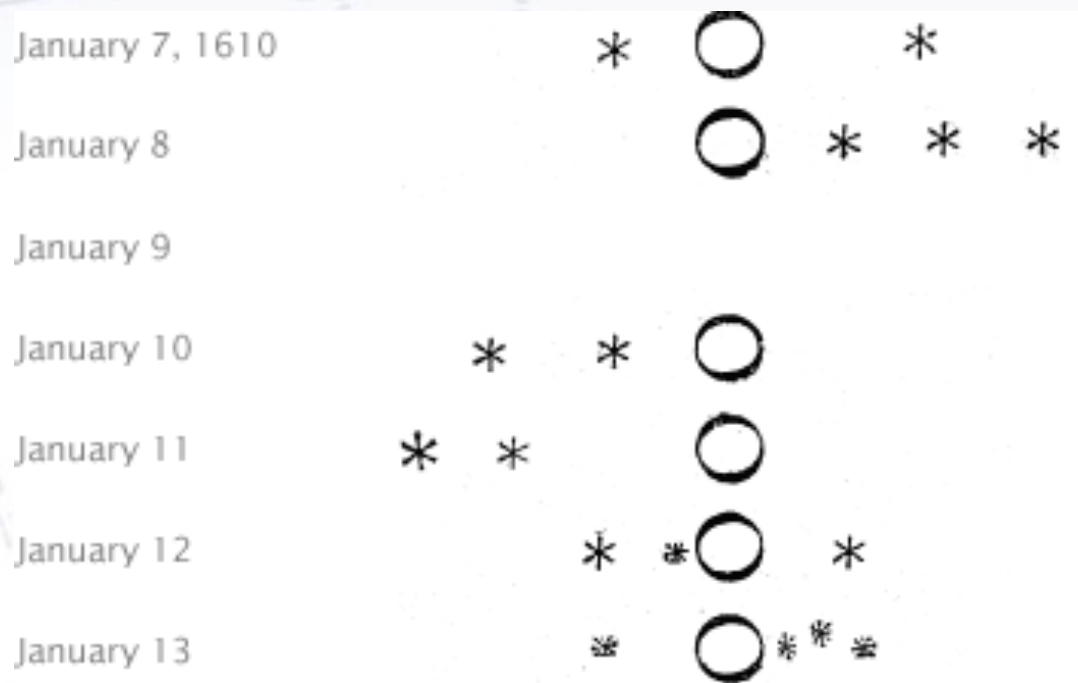


Galileo on Jupiter's moons

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 labelling and lack of extemporaneous axes.

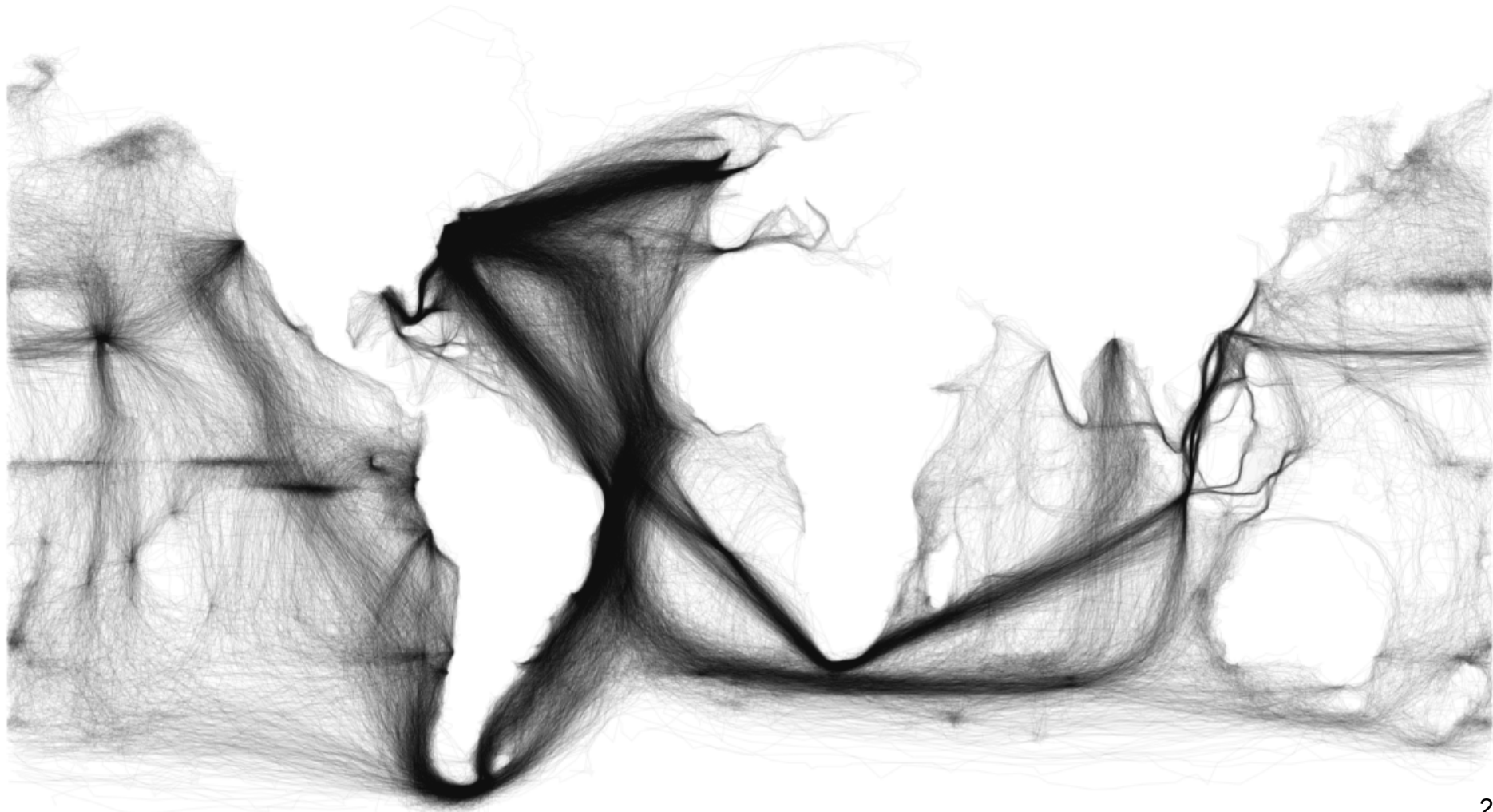
When you align each date's observations with Jupiter, as above, the helix pattern the moons trace as they orbit nearly jumps out!





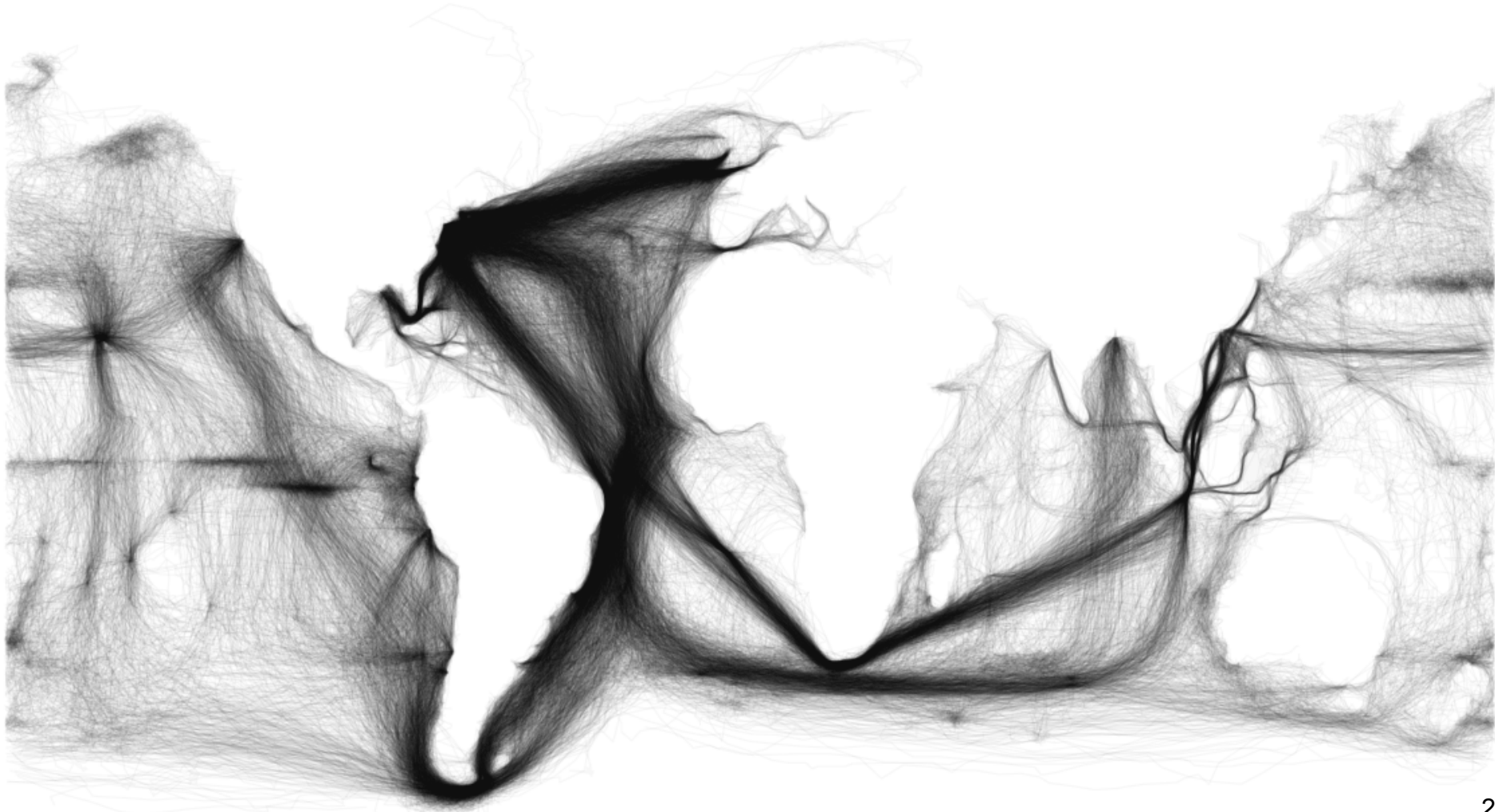
January 7, 1610			*	○	*
January 8				○	* * *
January 9					
January 10		*	*	○	
January 11	*	*		○	
January 12			*	○	*
January 13			*	○	* * *

Guess what this is...

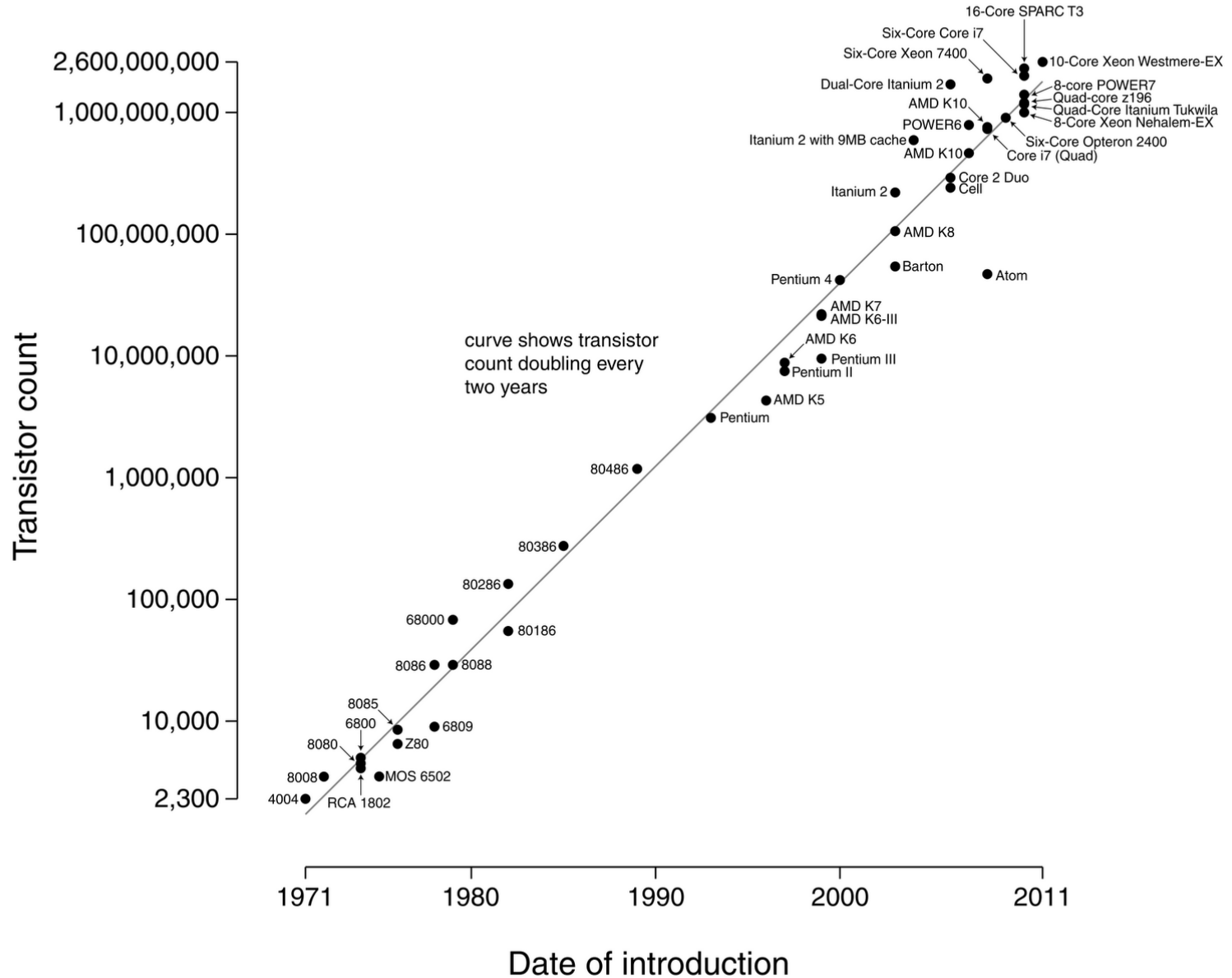


Guess what this is...

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.

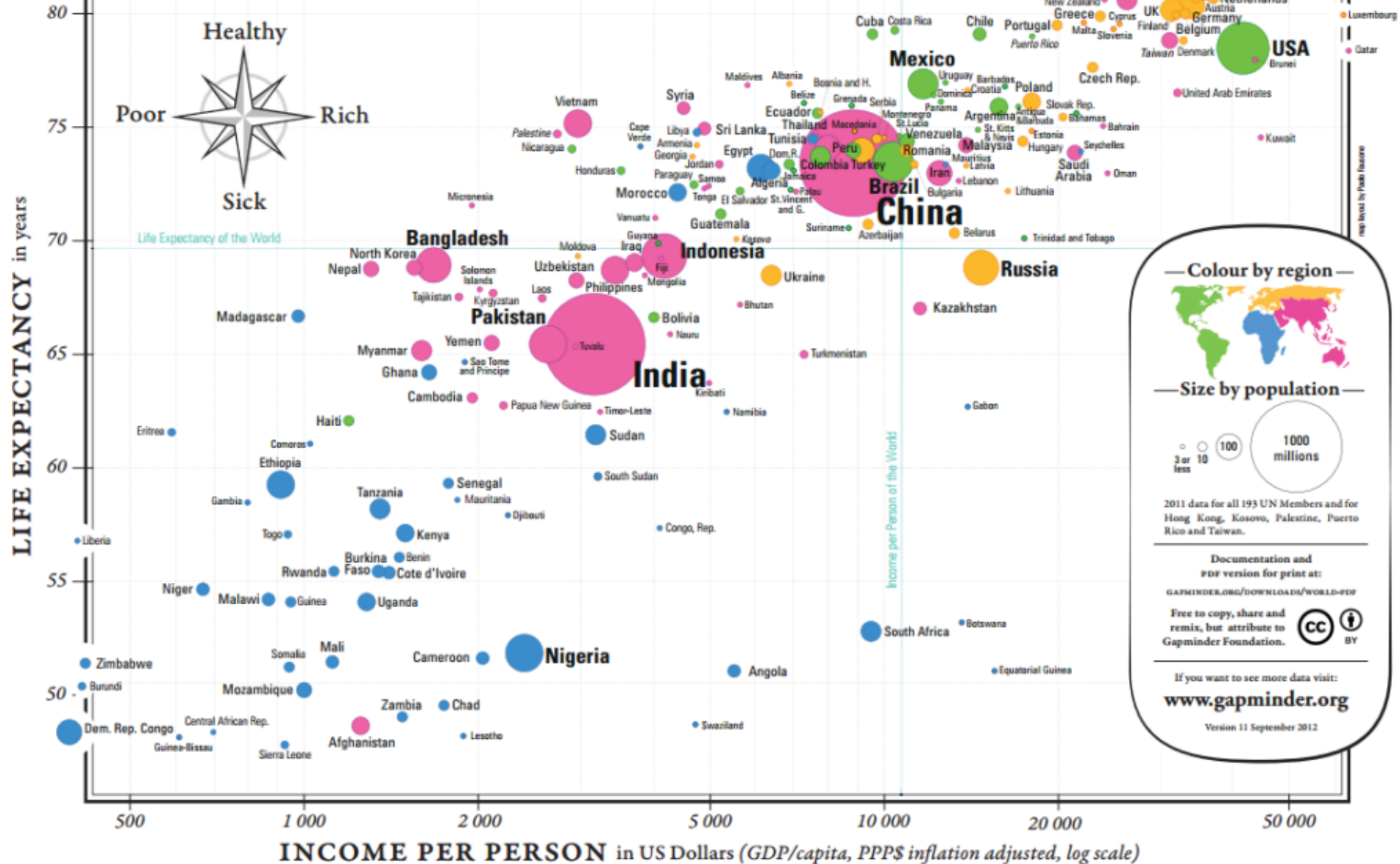


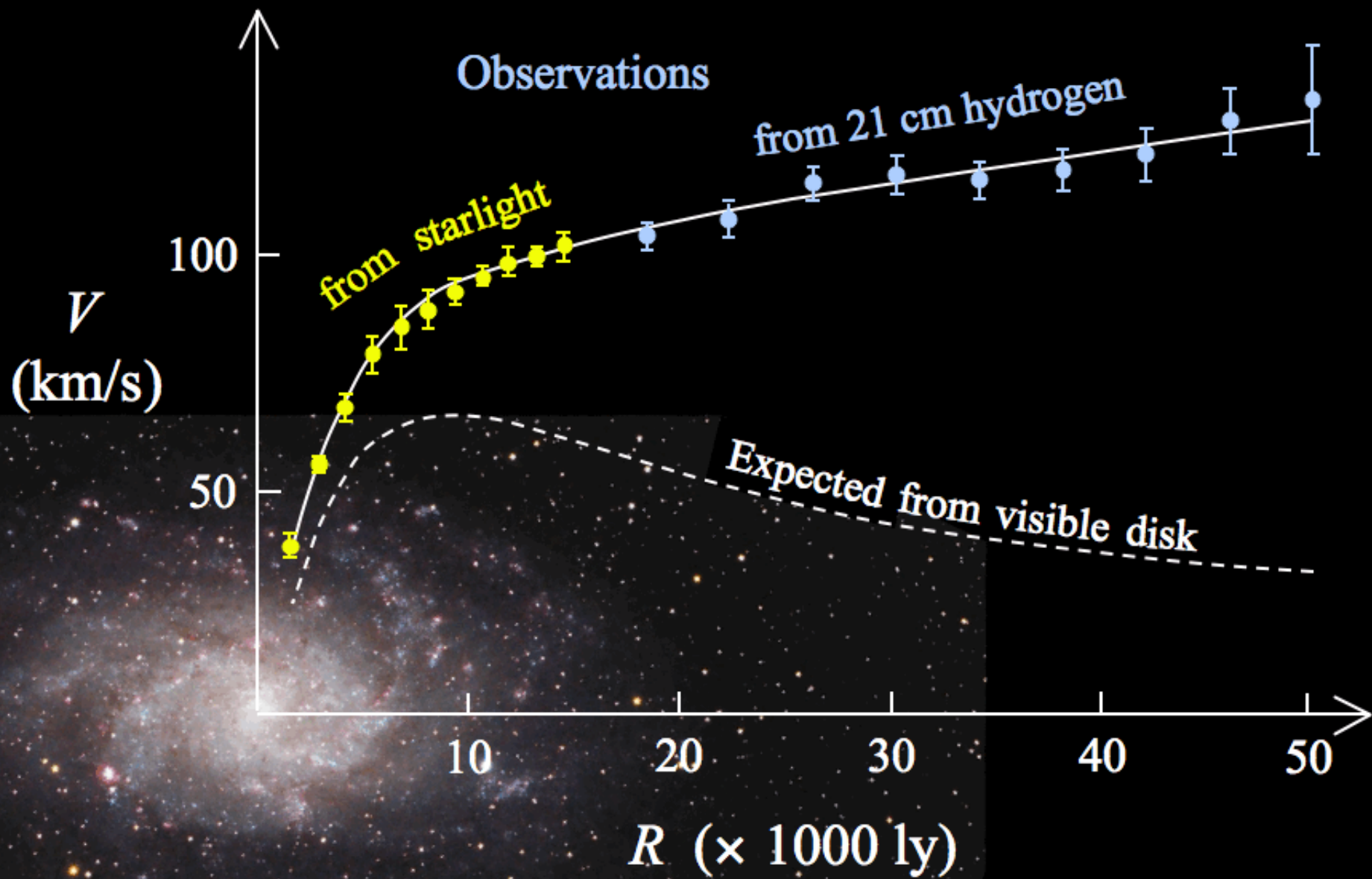
Microprocessor Transistor Counts 1971-2011 & Moore's Law



GAPMINDER WORLD 2012

Mapping the Wealth and Health of Nations



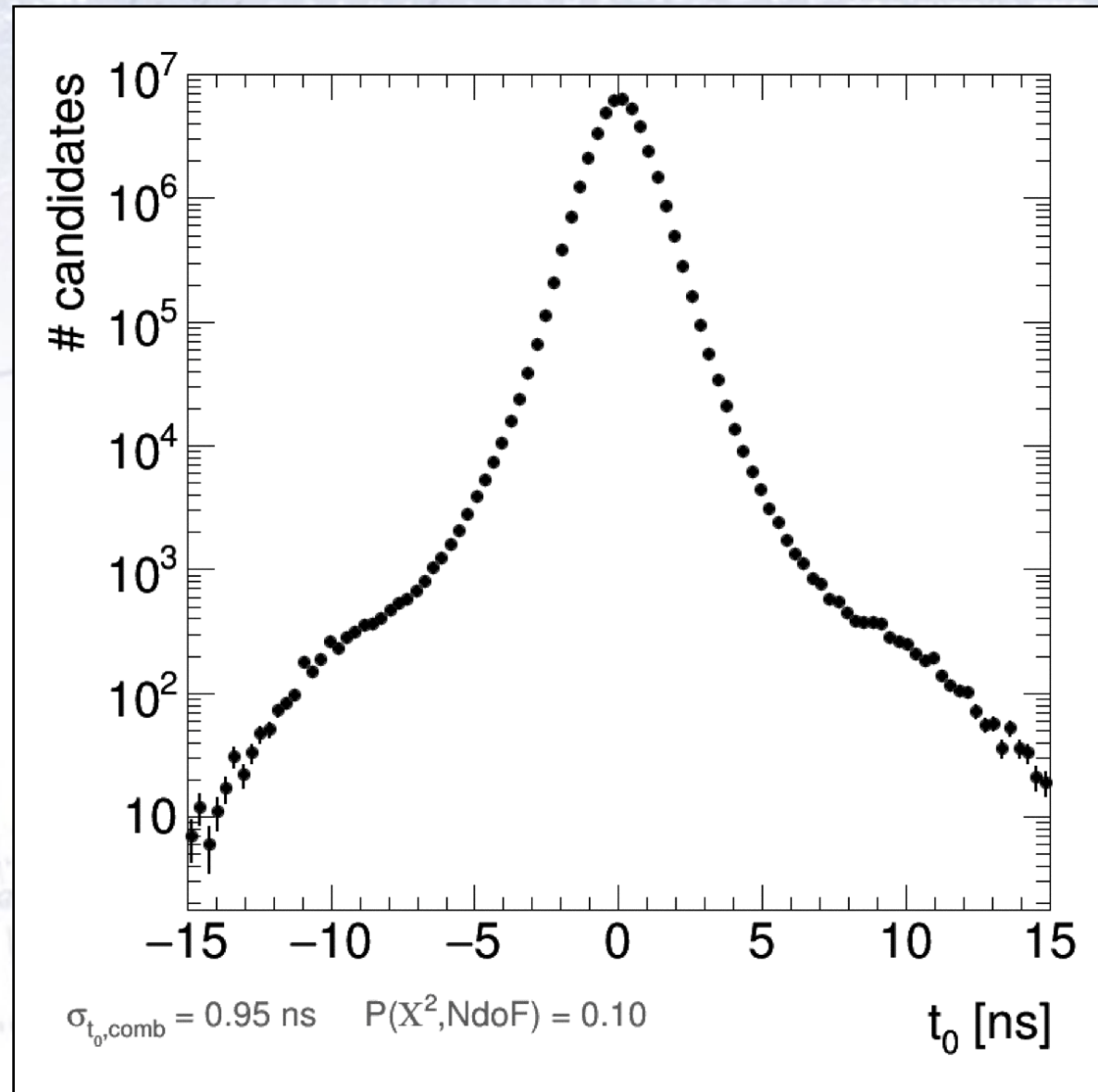


Animations of plots

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.

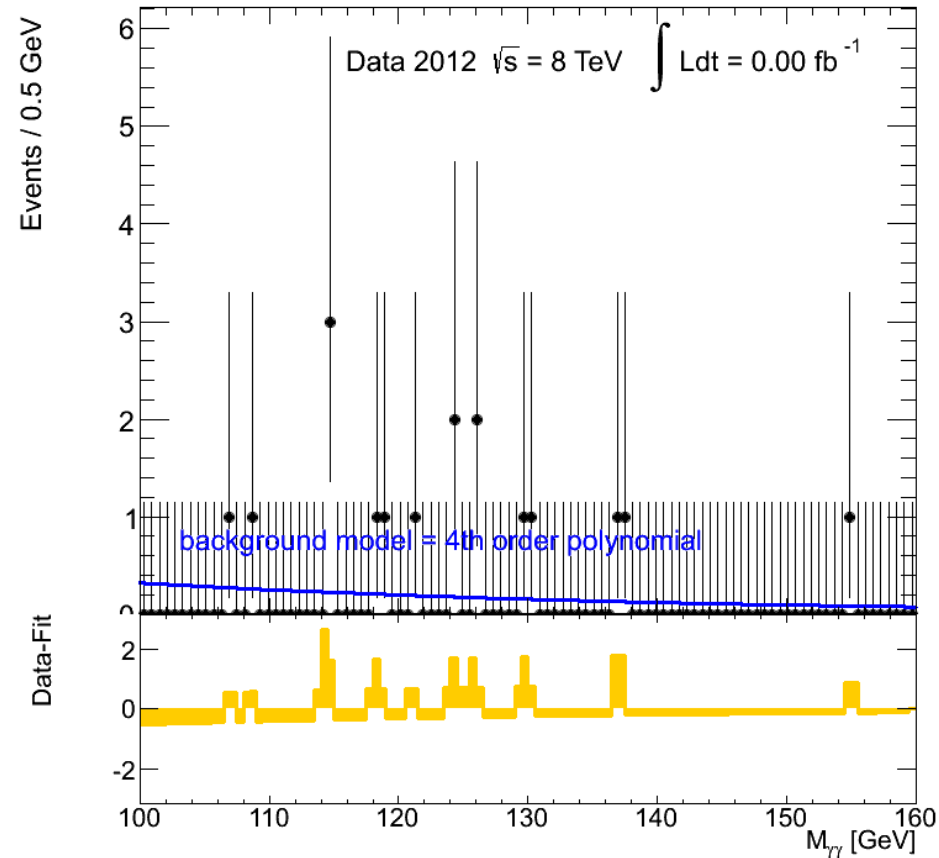
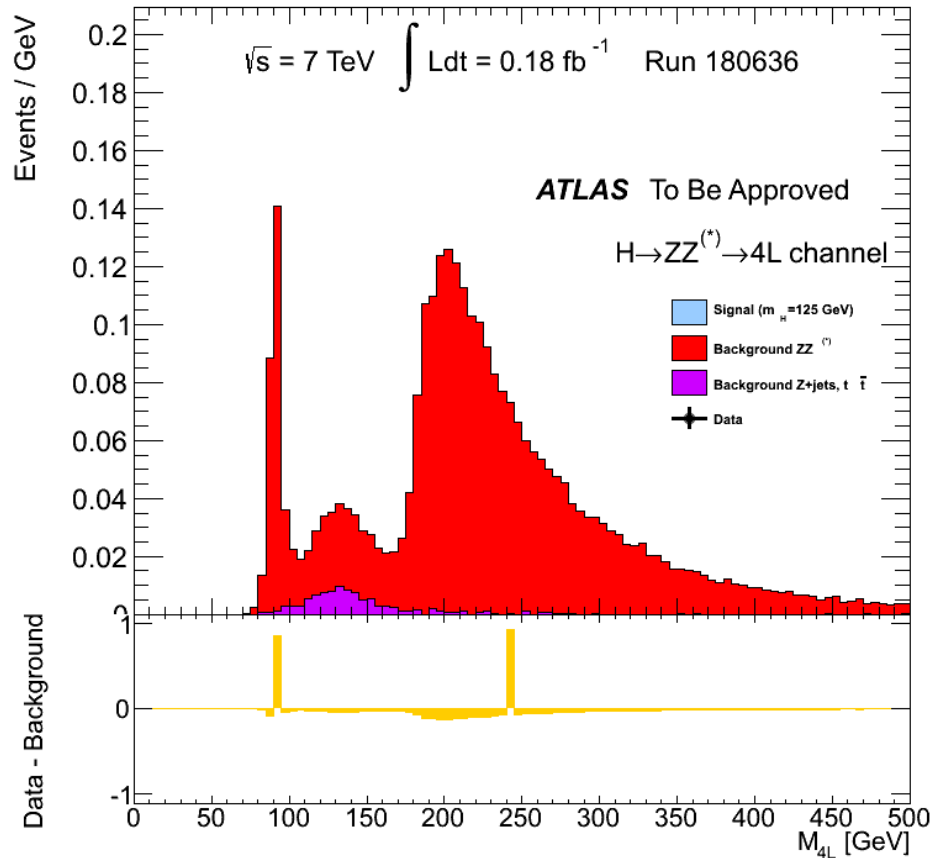
This is a very powerful way to make you points come across, typically well remembered by your audience.

Here is shown the timing of particles in ATLAS, and evidence for observing the small $\pm 5\text{ns}$ satellite bunches.



Animations of plots

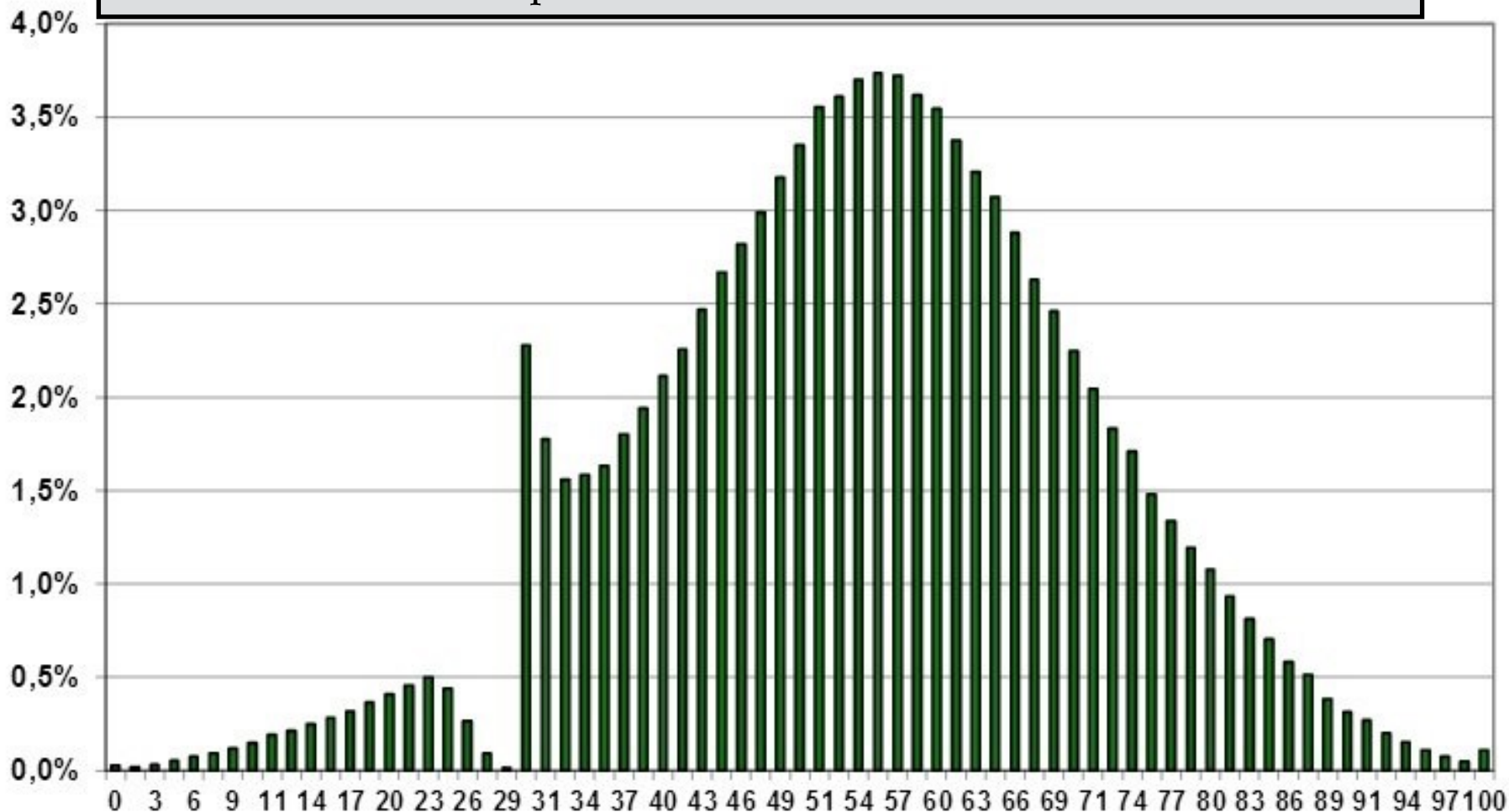
Following the Higgs discovery, the ATLAS collaboration produced the following two animations, which show the buildup of Higgs signal with time / data.



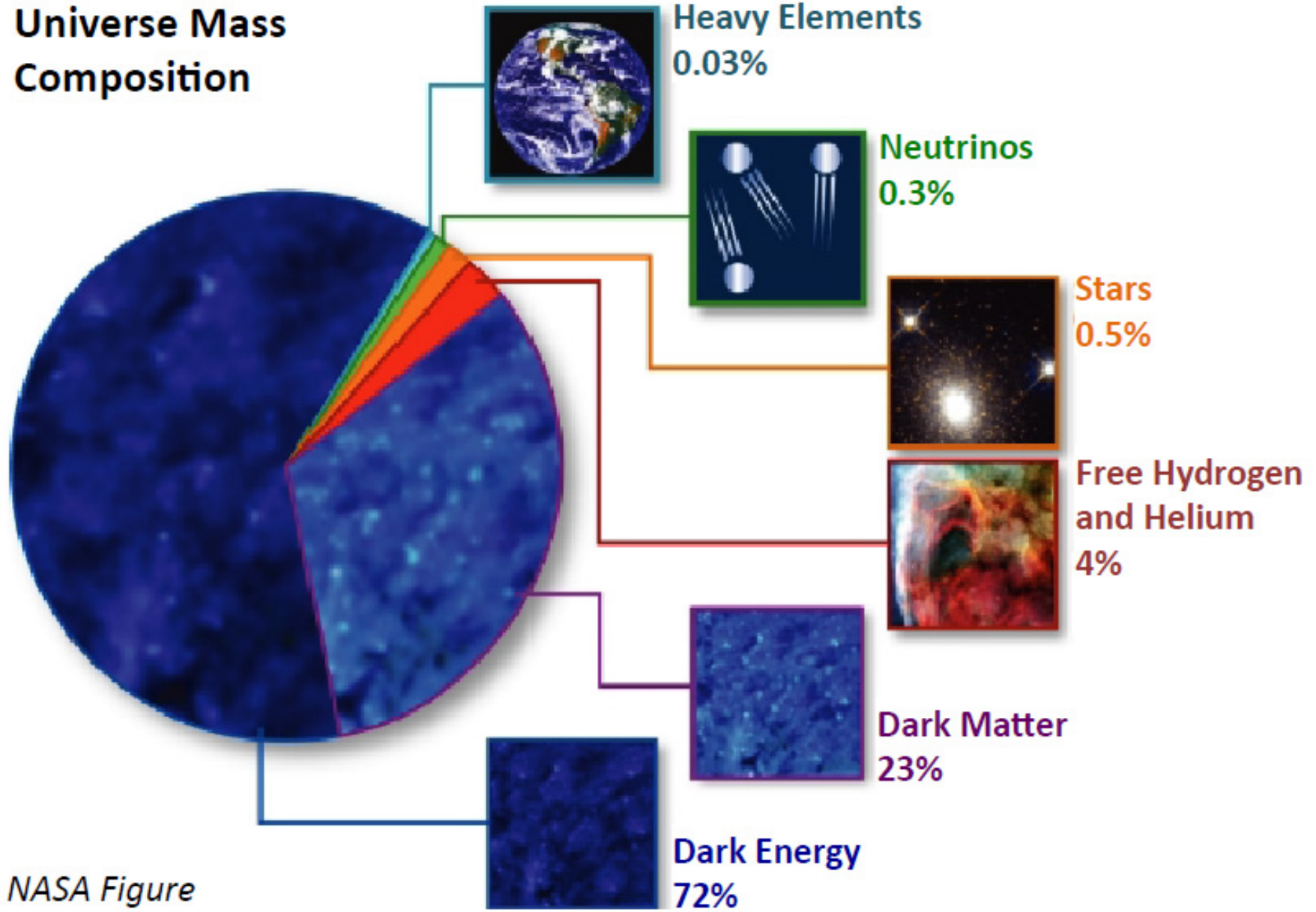
Distributions and social effects

2.1. Poziom podstawowy

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



Universe Mass Composition



NASA Figure

Overview

ATLAS Searches* - 95% CL Lower Limits (EPS-HEP 2011)

SUSY

- MSUGRA/CMSSM : 0-lep + $E_{T,miss}$
- Simplified model (light $\tilde{\chi}_1^0$) : 0-lep + $E_{T,miss}$
- Simplified model (light $\tilde{\chi}_2^0$) : 0-lep + $E_{T,miss}$
- Simplified model (light $\tilde{\chi}_1^{\pm}$) : 0-lep + $E_{T,miss}$
- Simplified model : 0-lep + b-jets + $E_{T,miss}$
- Pheno-MSSM (light $\tilde{\chi}_1^0$) : 2-lep SS + $E_{T,miss}$
- Pheno-MSSM (light $\tilde{\chi}_1^{\pm}$) : 2-lep OS^{SF} + $E_{T,miss}$
- GMSB (GGM) + Simpl. model : $\tilde{\gamma}\tilde{\gamma}$ + $E_{T,miss}$
- GMSB : stable $\tilde{\tau}$
- Stable massive particles : R-hadrons
- Stable massive particles : R-hadrons
- Stable massive particles : R-hadrons
- RPV ($\lambda'_{311}=0.01, \lambda_{312}=0.01$) : high-mass $e\mu$

Extra dimensions

- Large ED (ADD) : monojet UED : $\tilde{\gamma}\tilde{\gamma}$ + $E_{T,miss}$
- RS with $k/M_{Pl} = 0.1$: $m_{\tilde{\gamma}\tilde{\gamma}}$
- RS with $k/M_{Pl} = 0.1$: $m_{ee/\mu\mu}$
- RS with top couplings $g_L=1.0, g_R=4.0$: m_{tt}
- Quantum black hole (QBH) : $m_{dijet}, F(\chi)$
- QBH : High-mass $\sigma_{t+\chi}$
- ADD BH ($M_{th}/M_D=3$) : multijet $\Sigma p_T, N_{jets}$
- ADD BH ($M_{th}/M_D=3$) : SS dimuon $N_{ch. part.}$

LQ / W' / Ct. I.

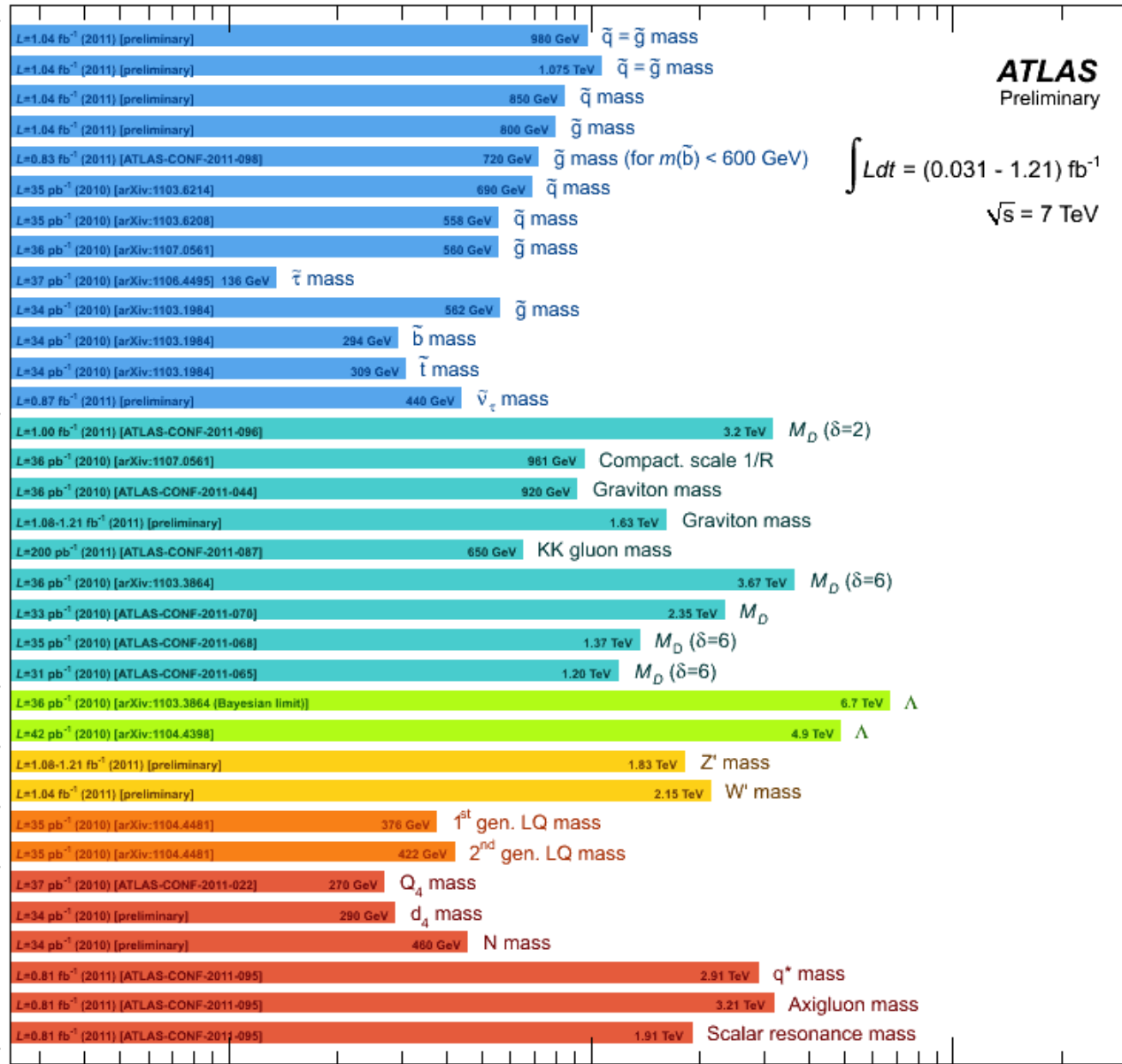
- qqqq contact interaction : $F_{\chi}(m_{dijet})$
- qq $\mu\mu$ contact interaction : $m_{\mu\mu}$
- SSM : $m_{ee/\mu\mu}$
- SSM : $m_{T,e/\mu}$

LQ

- Scalar LQ pairs ($\beta=1$) : kin. vars. in eejj, evjj
- Scalar LQ pairs ($\beta=1$) : kin. vars. in $\mu\mu jj, \mu\nu jj$
- 4th family : coll. mass in $Q_4\bar{Q}_4 \rightarrow WqWq$
- 4th family : $d_4\bar{d}_4 \rightarrow WtWt$ (SS dilepton)
- Major. neutr. ($V_{4-ferm.}, \Lambda=1$ TeV) : SS dilepton

Other

- Excited quarks : m_{dijet}
- Axigluons : m_{dijet}
- Color octet scalar : m_{dijet}



ATLAS Preliminary

$\int Ldt = (0.031 - 1.21) \text{ fb}^{-1}$
 $\sqrt{s} = 7 \text{ TeV}$

10⁻¹ 1 10 Mass scale [TeV]

*Only a selection of the available results shown