GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

Teoria 6. Tractament de dades II

Departament de Matemàtiques

Data processing for visualization

- **Chapter 5 - Data processing for visualization (I)**
  - Uncertainty and error
  - Transformations and data massage (+ seminars & PRT1)
- **Chapter 6 (today) - Data processing for visualization (II)**
  - Dimensionality reduction
  - Computation and important metrics selection

6. Data processing for visualization (II). Contents:

1. Dimensionality reduction

1. Introduction
2. Correlograms
3. Feature projection – PCA
4. Discriminant analysis (linear) - LDA
5. T-Distributed stochastic neighbour embedding (t-SNE)
6. Tomography- Slice along a plane, 2D isosurfaces for a 3D field, isocontours

2.Computation and important metrics selection

1. Quality metrics: Noise reduction, clutter reduction, search outliers

6.1.1. Introduction: **Dimensionality reduction**

Dimensionality reduction (DR) means: the process of **transformation of data from high dimensional space to low dimensional space** while maintaining most of the meaningful insights from the original data.

The goal is to preserve the meaningful structure of a dataset while using fewer attributes to represent the items.

For example: We have a dataset contains hundred columns (i.e features) or it could be an array of points that make up a large sphere in the 3D space. DR?

DR entails lowering the number of columns to a smaller number, such as 2D.

6.1.1. Introduction: **Dimensionality reduction**

Dimensionality reduction (DR) has two primary use cases:

– data exploration

– machine learning.

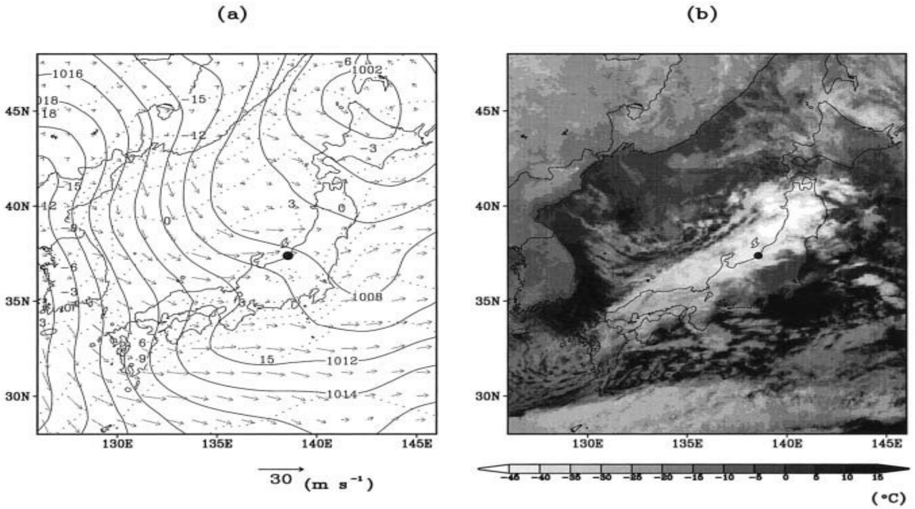
DR is a strategy for managing complexity in visualization:

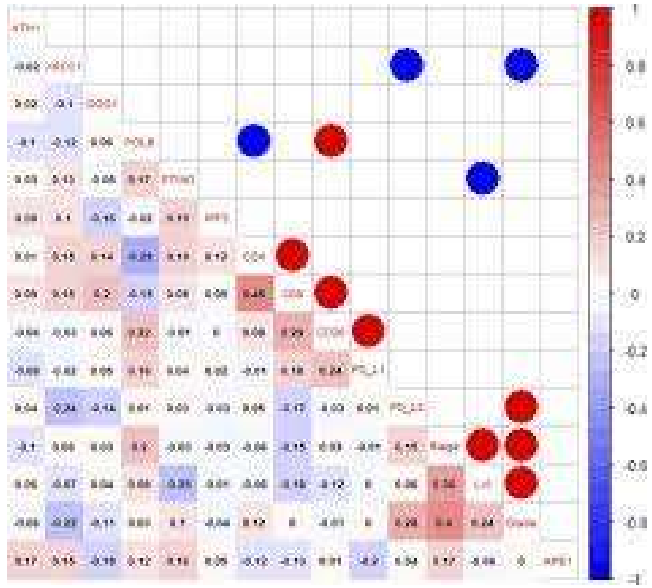
It is useful *for data exploration* because **dimensionality reduction to few dimensions** (e.g., 2D or 3D) **allows for visualizing the samples.**

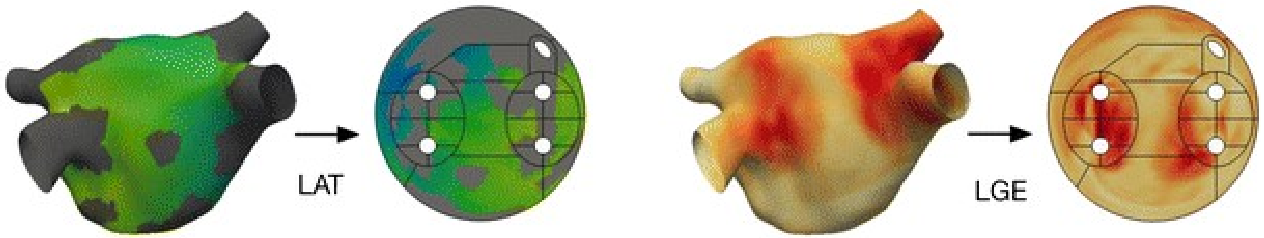
Such a visualization can then be used to obtain insights from the data *(e.g., detect clusters and identify outliers).*

6.1.1. Introduction: **Dimensionality reduction**

Examples of dimension reduction:

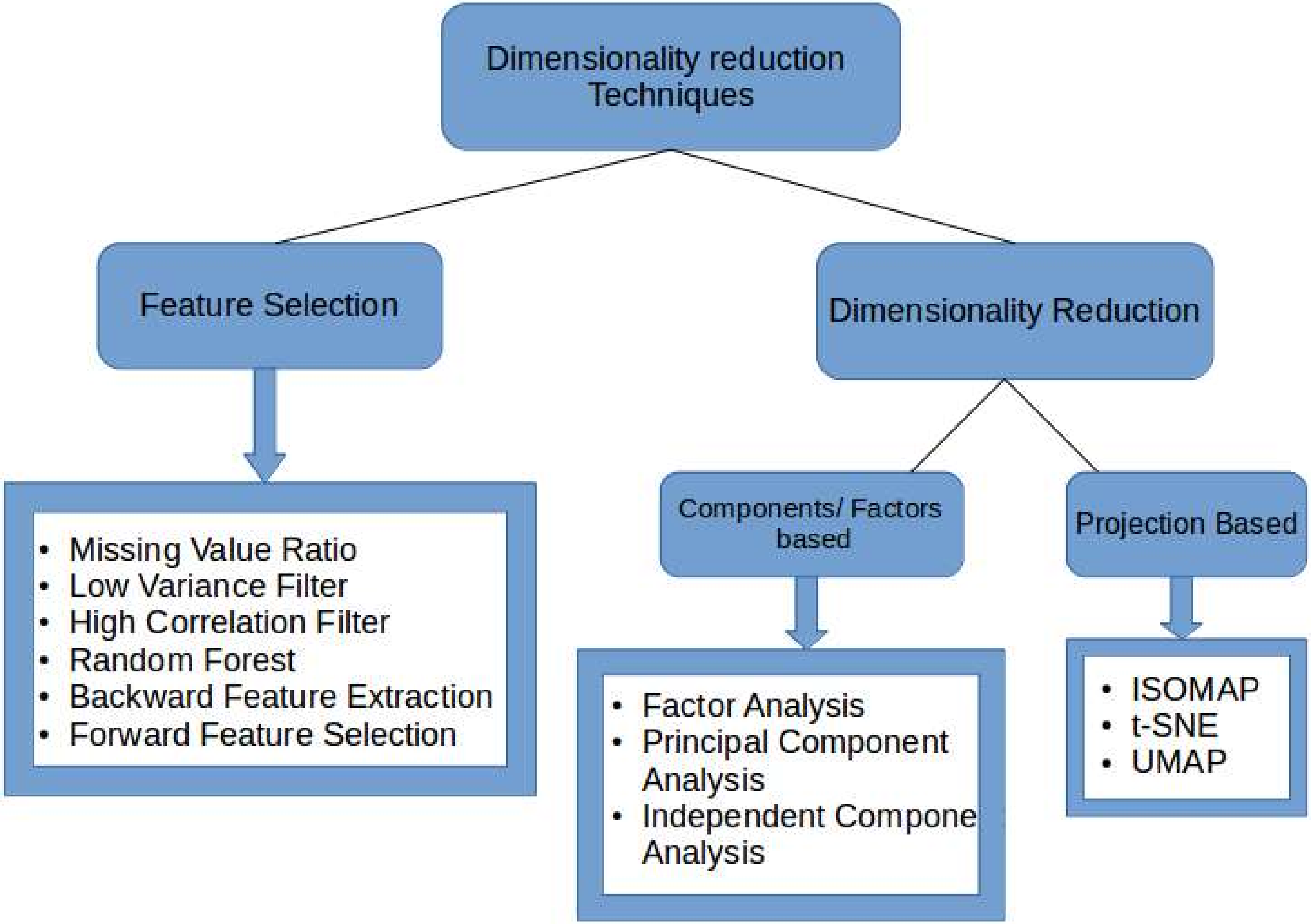






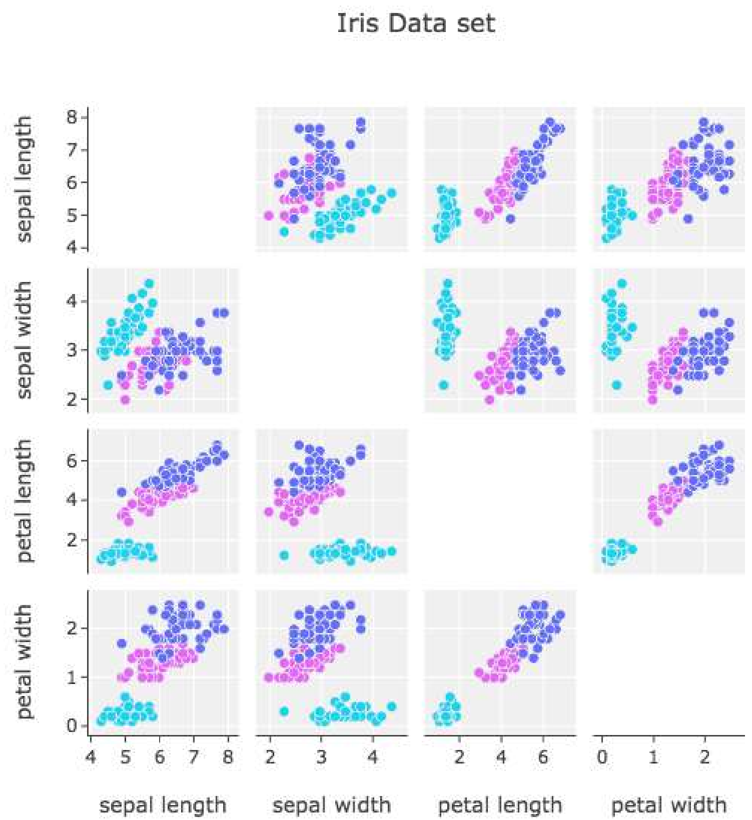
DOI: 10.1007/s10840-017-0281-3

6.1.1. Introduction: **Dimensionality reduction**



6.1.1. Remember: **Scatterplot matrix limitation**

We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots *to determine the correlation (if any) between a series of variables*.



!! When we have >3 or 4 quantitative variables – scatterplot matrices quickly become unwieldy

6.1.1. Correlation coefficients

We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots to determine *the correlation (if any) between a series of variables.*

!! When we have >3 or 4 quantitative variables – scatterplot matrices quickly become unwieldy

In this case, it is **more useful to quantify the amount of association between pairs of variables and visualize these quantities rather than the raw data**.
One common way to do this is to calculate **correlation coefficients.**

6.1.1. Correlation coefficient

- Having two sets of observations: x_i and y_i
- And: $\bar{x}$ and $\bar{y}$ the corresponding sample means

The correlation coefficient is:

$$R = \frac{\sum_i (x_i - \bar{x})(yi - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y - \bar{y})^2}}$$

The **correlation coefficient *R*** is **a number between -1 and 1 that measures to what extent two variables are correlated**

6.1.1. Correlation coefficients

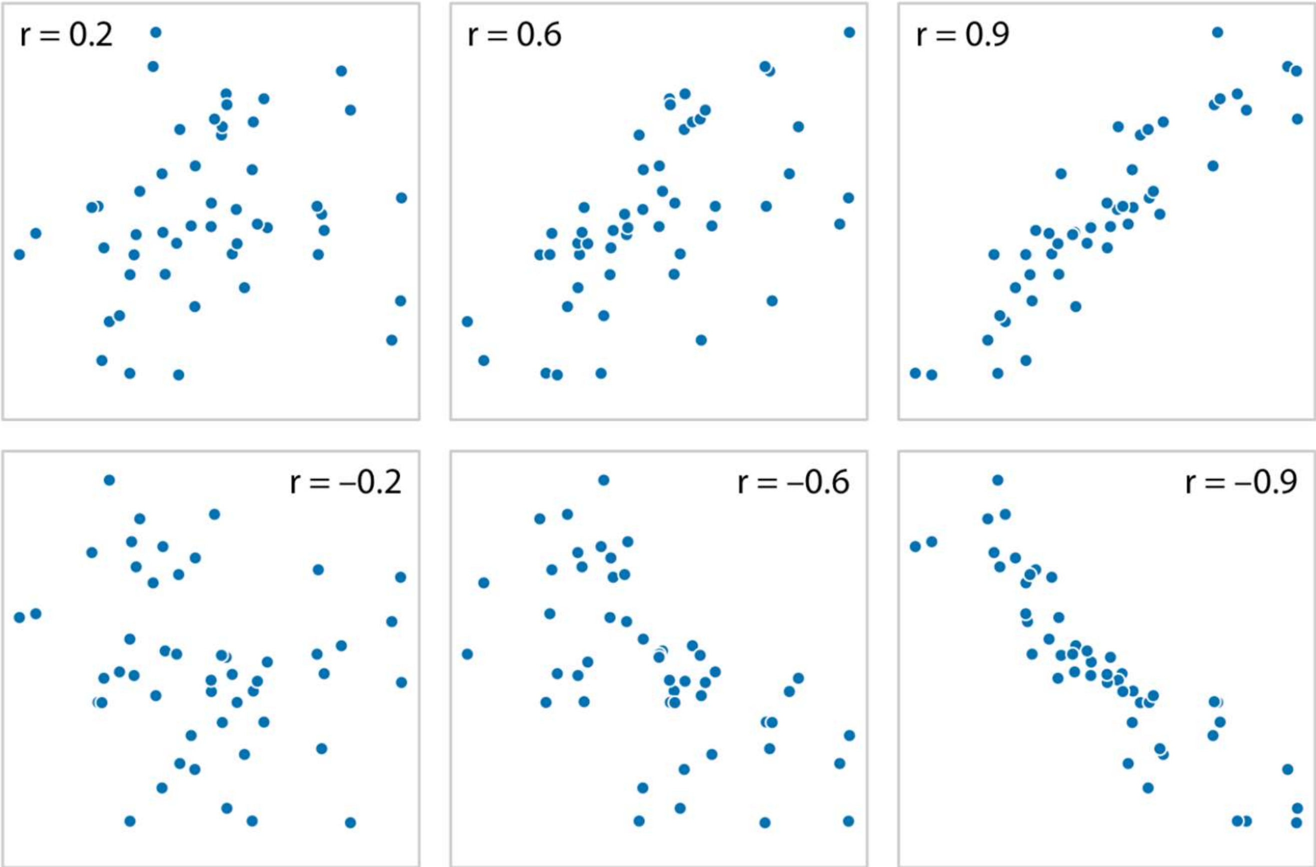
$$R = \frac{\sum_i (x_i - \bar{x})(yi - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y - \bar{y})^2}}, \ \textit{-1} < R < \textit{1}$$

- R **= *0*** means there is **no association** whatsoever
- R **= *1*** *or* ***-1*** indicates a **perfect association**

The sign of the correlation coefficient R indicates :

- R **>0**: variables are **correlated** (larger values in one variable coincide with larger values in the other)
- R **<0**: **anticorrelated** (larger values in one variable coincide with smaller values in the other)

6.1.1. Correlation coefficients



R>0

R<0

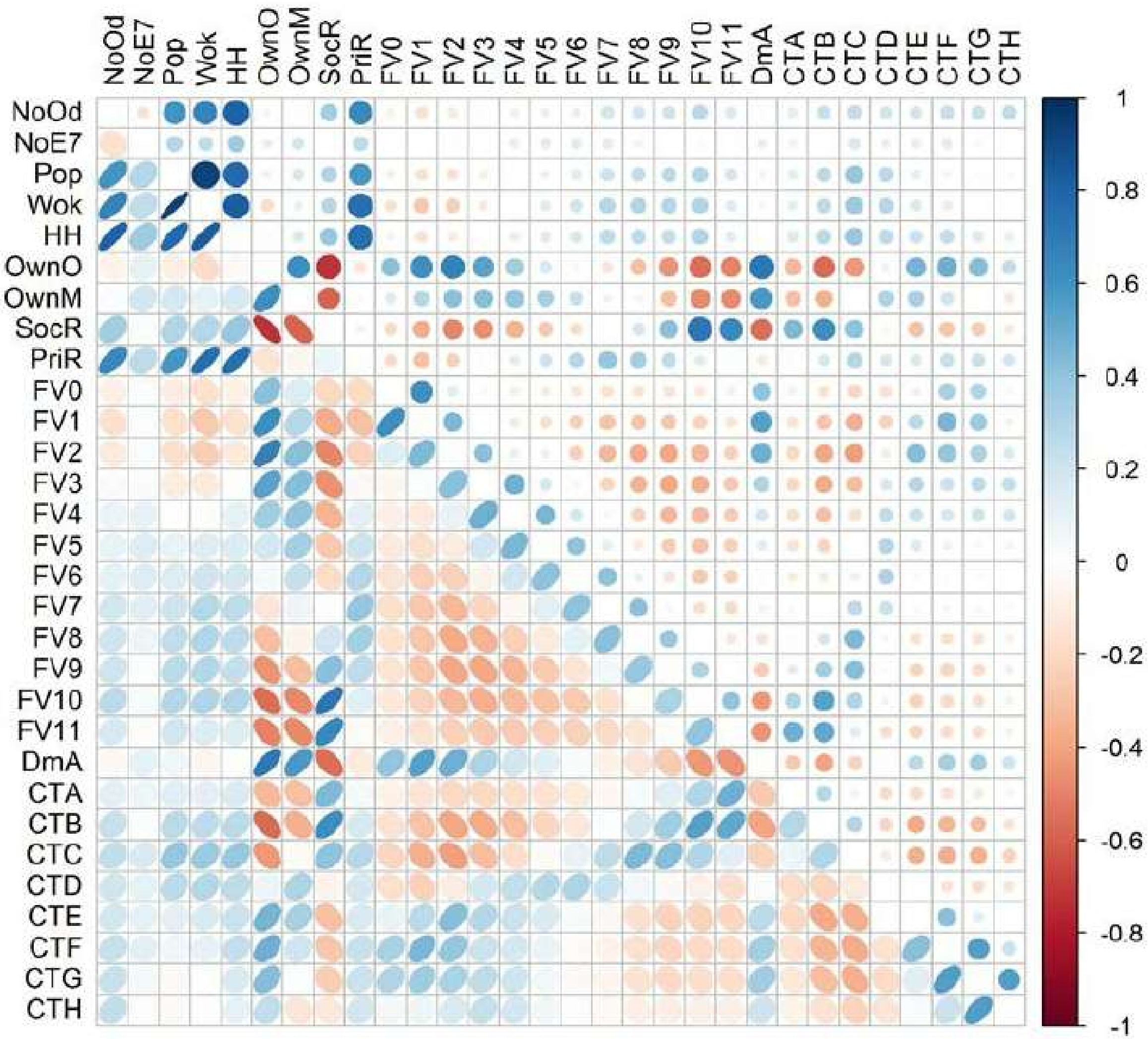
Wilke

Examples of correlations of different magnitude and direction, with associated correlation coefficient R

Escola d'Enginyeria UAB

UAB

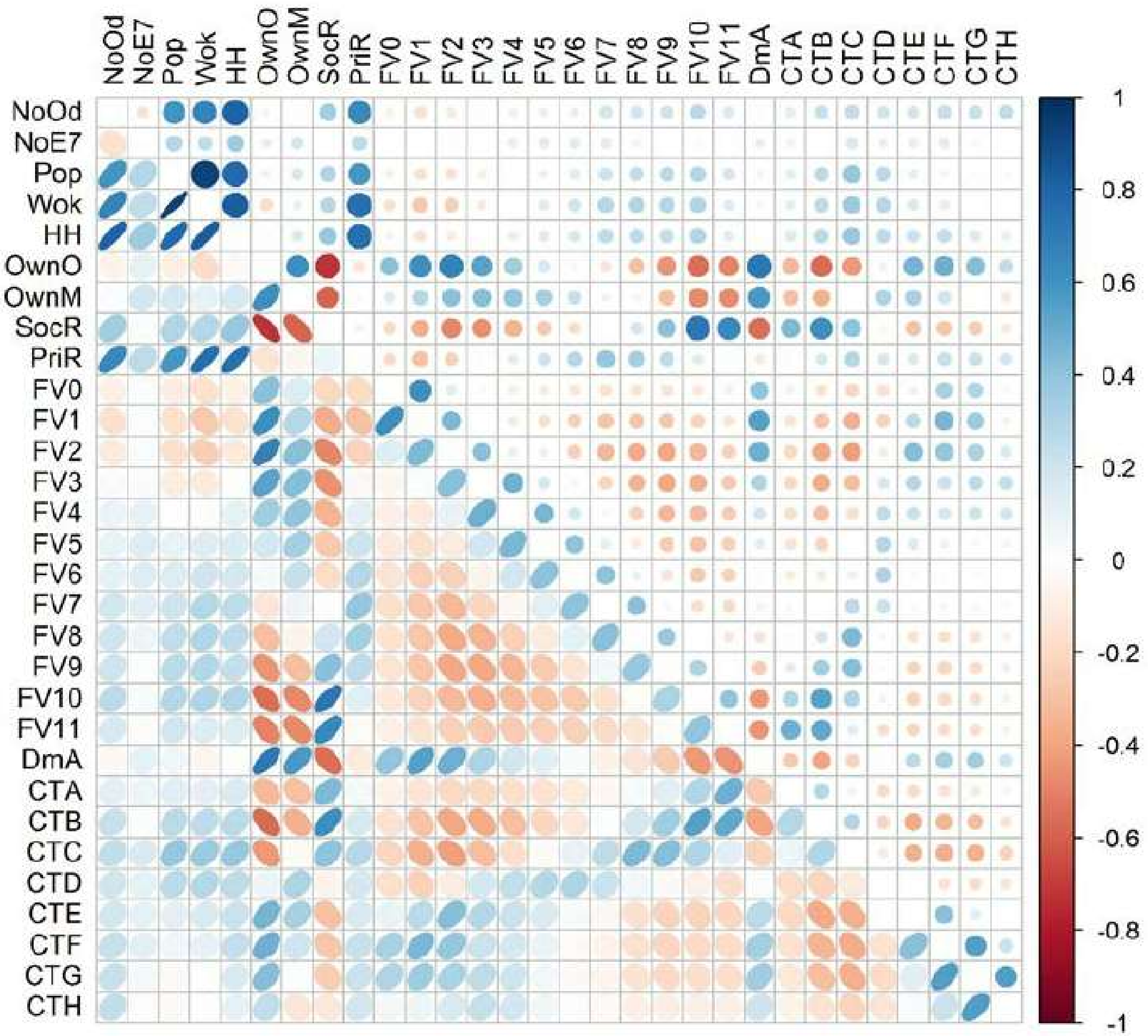
6.1.2. Correlogram



Correlation structure of the input variables. **Colour intensity and ellipse shape** are directly linked to correlation coefficient

DOI: 10.1016/j.proeng.2015.08.1069

6.1.2. Correlogram

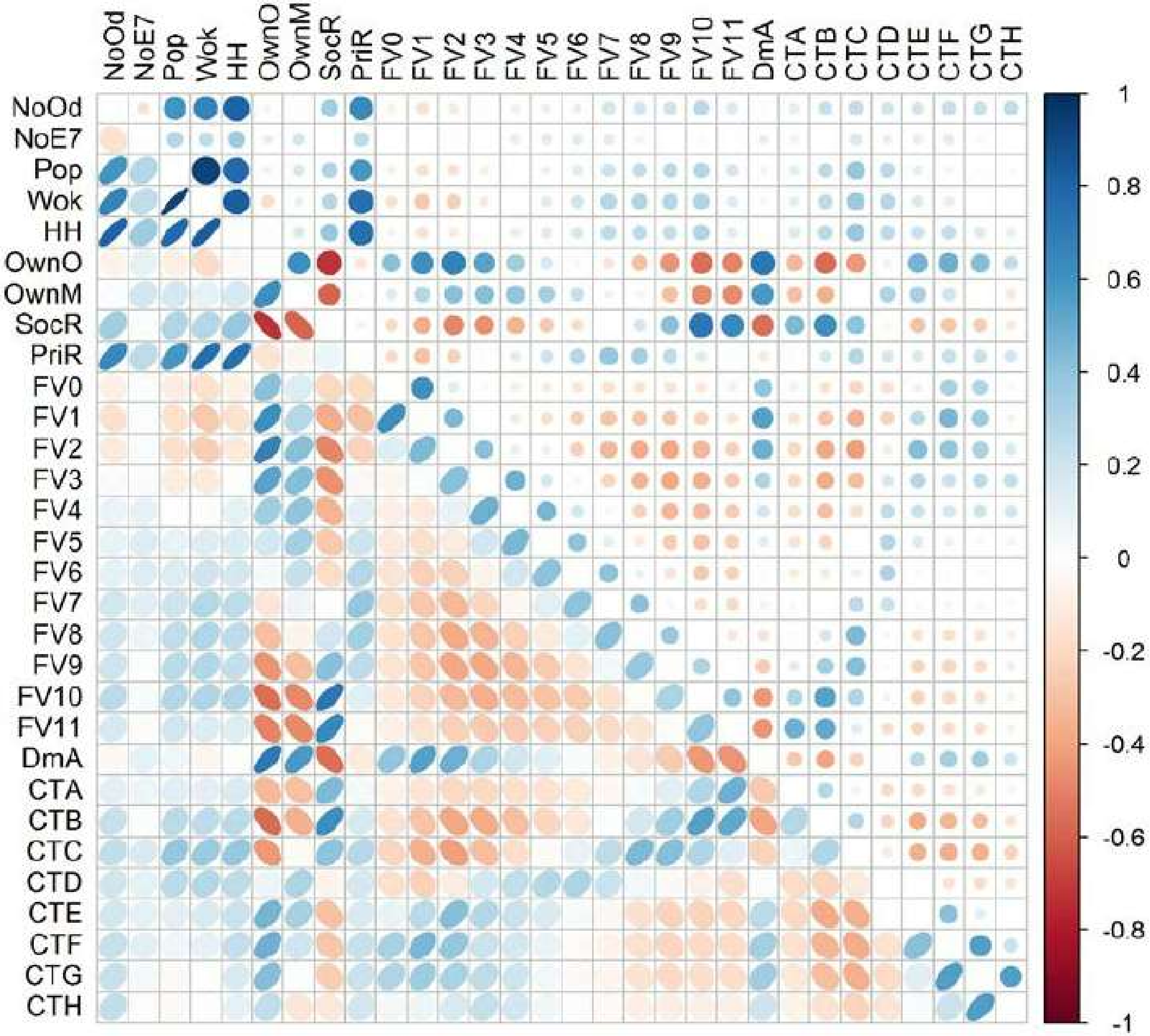


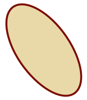
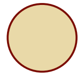
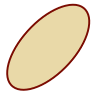
Correlation structure of the input variables. Colour intensity and ellipse shape are directly linked to correlation coefficient

R>0 is displayed in blue (correlated)

R<0 is shown in red (anticorrelated).

6.1.2 Correlogram



-  : there is a **higher negative correlation** coefficient between two variables.
-  : indicates a **weak correlation** for two factors
-  : there is a **higher positive correlation** coefficient between two variables

6.1. Correlation & Dimensionality reduction

DR relies on the key insight that **most high-dimensional datasets consist of multiple correlated variables** that convey overlapping information

Such datasets can be reduced to a smaller number of key dimensions without loss of much critical information.

DR can be achieved by:

- ***Feature elimination*** – we reduce the feature space by elimination feature
- ***Feature selection*** – process of selecting required features from all the features available in data. *Goal: to choose features that represent the dataset perfectly*
- ***Feature Engineering*** – process of **transforming raw data into feature**, which represent the dataset well

6.1. Correlation & Dimensionality reduction

DR relies on the key insight that **most high-dimensional datasets consist of multiple correlated variables that convey overlapping information**

Such datasets can be reduced to a smaller number of key dimensions without loss of much critical information.

There are many techniques for dimension reduction. We will see:

- Principal Components Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- T-Distributed Stochastic Neighbour Embedding (t-SNE) (next day)

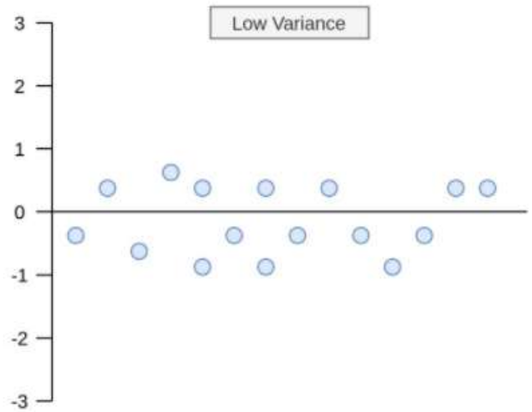
6.1. Remember: **Variance**

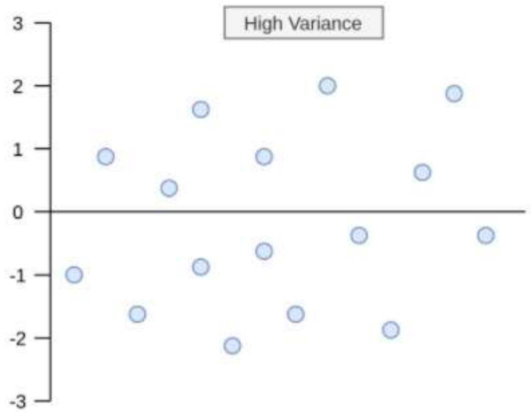
Before going further, **let's clarify some concepts:**

- We know that **variance** represents **the variation of values in a single variable.** *It depends on **how the values far from each other.***

Having a set of observations: x_i and $\bar{x}$ the corresponding sample mean:

$$var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$





Sergen Cansiz

6.1. Remember: **Covariance**

Before going further, let's clarify some concepts:

- Unlike the variance, **covariance** is calculated **between two different variables**. Its purpose is to find the value that indicates **how these two variables vary together**.

- Having two sets of observations: x_i and y_i
- And: $\bar{x}$ and $\bar{y}$ the corresponding sample means

$$cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Escola d'Enginyeria UAB

UAB

6.1. Remember: **Covariance vs correlation**

Before going further, let's clarify some concepts:

- **Covariance vs correlation**

- Having two sets of observations: x_i and y_i
- And: $\bar{x}$ and $\bar{y}$ the corresponding sample means

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$R(x,y) = \frac{\sum_i (x_i - \bar{x})(yi - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y - \bar{y})^2}}$$

6.1. Covariance vs correlation

Before going further, let's clarify some concepts:

- **Covariance vs correlation**

Covariance	Correlation
Covariance is a **measure** to indicate the extent to which **two random variables change *in tandem***	Correlation is a **measure** used to represent how strongly **two random variables are related *to each other***
Covariance **indicates the *direction* of the linear relationship** between variables	Correlation **measures *both the strength and direction* of the linear relationship** between variables
Covariance can **vary *between* $-\infty$ *and* $+\infty$**	Correlation **ranges *between -1 and 1***

6.1. Covariance vs correlation

Before going further, let's clarify some concepts:

- **Covariance vs correlation**

Covariance	Correlation
Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed	**Correlation is NOT influenced by the change in scale**
Covariance of two *dependent variables* measures **how much in real quantity** (i.e, e.g., cm, km, liters) **on average they *co-vary.***	Correlation of two *dependent variables* measures **the proportion of how much on average these variables *vary with respect to one another***
Covariance is **zero for independent variables**	Completely **independent variables have a zero correlation**

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**

Because covariance can only be calculated between two variables, **covariance matrices** stand for representing **covariance values of each pair of variables in multivariate data.** Also, the covariance between the same variables equals variance, so, **the diagonal shows the variance of each variable**

Symmetric matrix

$$\begin{array}{c c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix} \end{array} \qquad \begin{array}{c c} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

2 and 3- dimensional covariance matrices

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**

These *values* in the covariance matrix **show the distribution magnitude and direction** of multivariate data in multidimensional space.

By controlling these values we can have *information about how data spread among two dimensions.*

Symmetric matrix

$$\begin{array}{c|cc} & x & y \\ \hline x & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$}} \\ y & & \end{array}$$

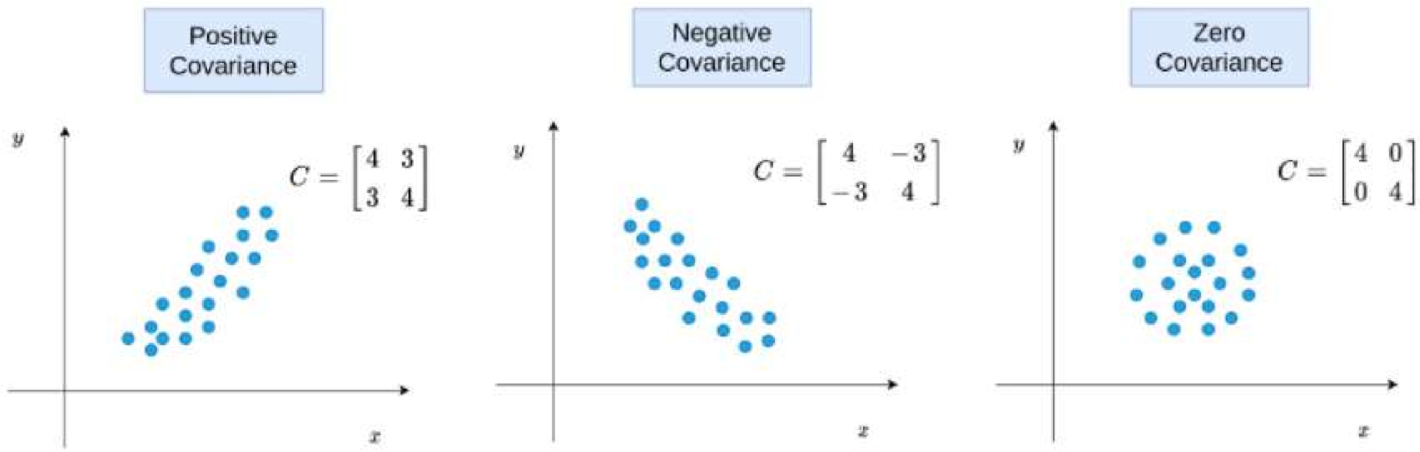
$$\begin{array}{ccc} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

2 and 3- dimensional covariance matrices

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**

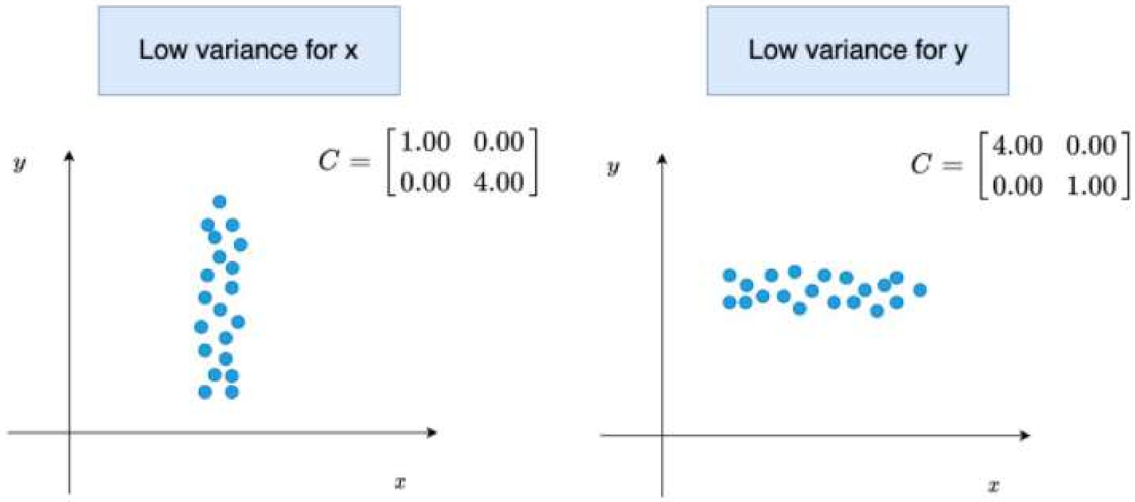


Sergen Cansiz

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**



Sergen Cansiz

Covariance near zero and different variances

6.1.3. Principal Components Analysis (PCA)

- When should I use PCA?

1. Do you want to reduce the number of variables, but **you *are not able to identify variables*** to completely remove from consideration?
2. Do you want ***to ensure your variables are independent*** of one another?
3. Are you ***comfortable making your independent variable less interpretable***?

6.1.3. Principal Components Analysis (PCA)

- PCA introduces a **new set of variables (smaller number of variables)**, called **principal components (PCs)**, by linear combination of the original variables in the data, standardized to zero mean and unit variance.
- **The axes or new variables are the PCs and are ordered by variance:**
  - The first component, *PC 1*, represents the *direction of the highest variance of the data*.
  - The direction of the *PC 2*, represents the *highest of the remaining variance* ***orthogonal*** *to the PC 1*.

This can be naturally **extended to obtain the required number of components, which together span a component space covering the desired amount of variance.**

6.1.3. Covariance matrix – relation with PCA

Before going further, let's clarify some concepts:

- **Eigenvalues and eigenvectors of covariance matrix:**

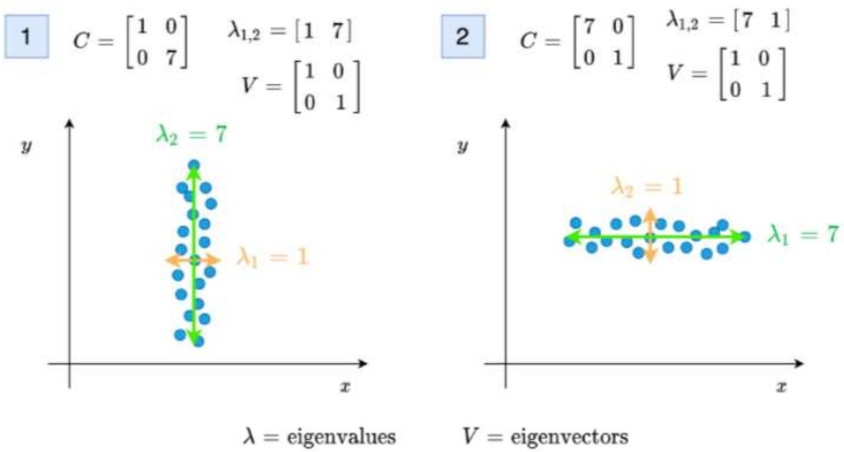
- The **eigenvalues represent the magnitude of the spread** in the direction of the principal components in PCA.

- The **eigenvectors show the direction**.

6.1.3 Covariance matrix – relation with PCA

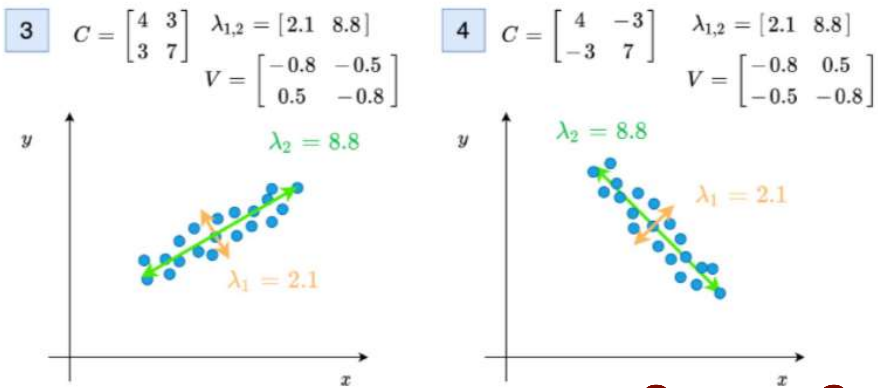
Before going further, let's clarify some concepts:

- **Eigenvalues and eigenvectors of covariance matrix**



- The first and second plots show the distribution of points when the covariance is near zero (independent variables).

Note: when the covariance is zero the eigenvalues=variance values



- The third and fourth plots represent the distribution of points when the covariance is different from zero.

Note: here we need to calculate the eigenvalues and eigenvectors

Sergen Cansiz

6.1.3 Principal Components Analysis (PCA)

PCA is maybe the most popular technique to examine high-dimensional data (unsupervised learning)

PCA computes a rotation matrix : $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{P}\times\boldsymbol{P}}$ **from the matrix of features** $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{P}}$

W can be understood as a **mapping function that transforms the observations in *X* to a rotated space**

The **coordinates of observations in *X* are transformed to** their new form, ***Z***, via: $Z = XW$

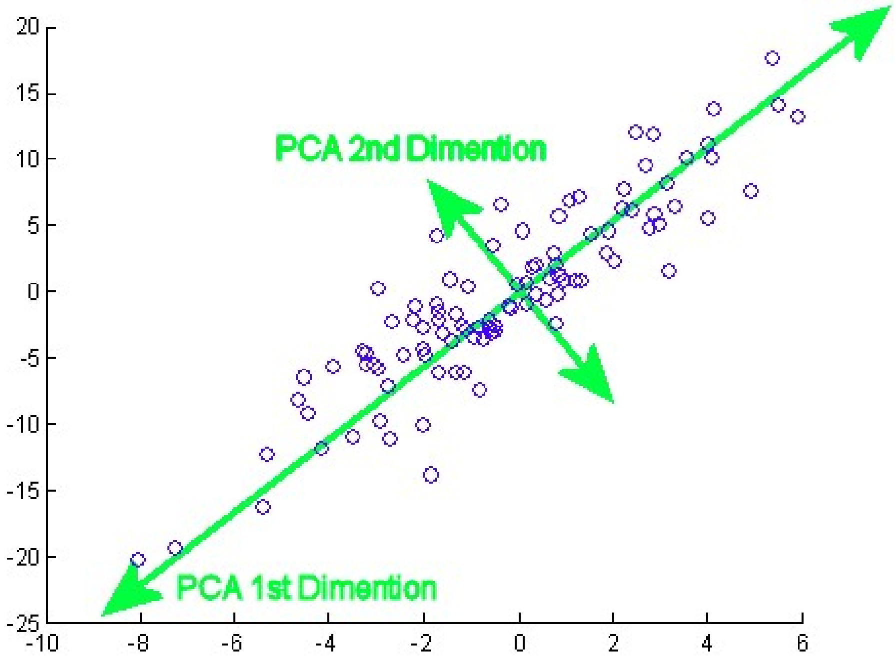
The rotation matrix, *W*, is constructed through orthogonal linear transformations. **Each of these transformations is performed in order to maximize the variance on the data**

6.1.3 Principal Components Analysis (PCA)

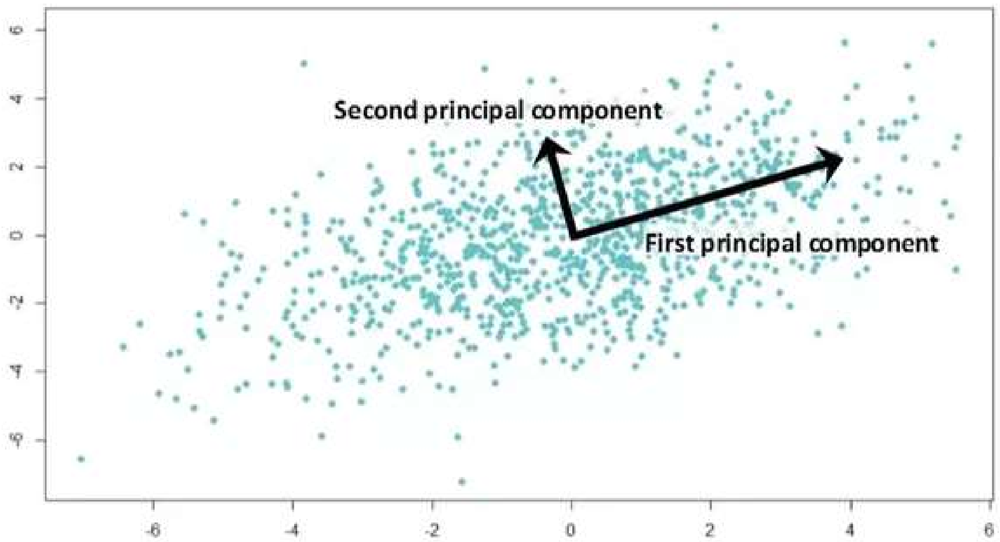
Steps:

1. Take the **matrix of features** $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{P}}$, *N > P*
2. Compute the **mean vector for each dimension**
3. Compute the **covariance matrix**
4. Compute the eigenvectors and corresponding eigenvalues for each dimension
5. Sort the eigenvectors by decreasing eigenvalues and choose ***P*** eigenvectors with the largest eigenvalues to form a new matrix $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{P} \times \boldsymbol{P}}$
6. Use this eigenvector matrix to transform the samples onto the new subspace: $Z = XW$

6.1.3 Principal Components Analysis (PCA)



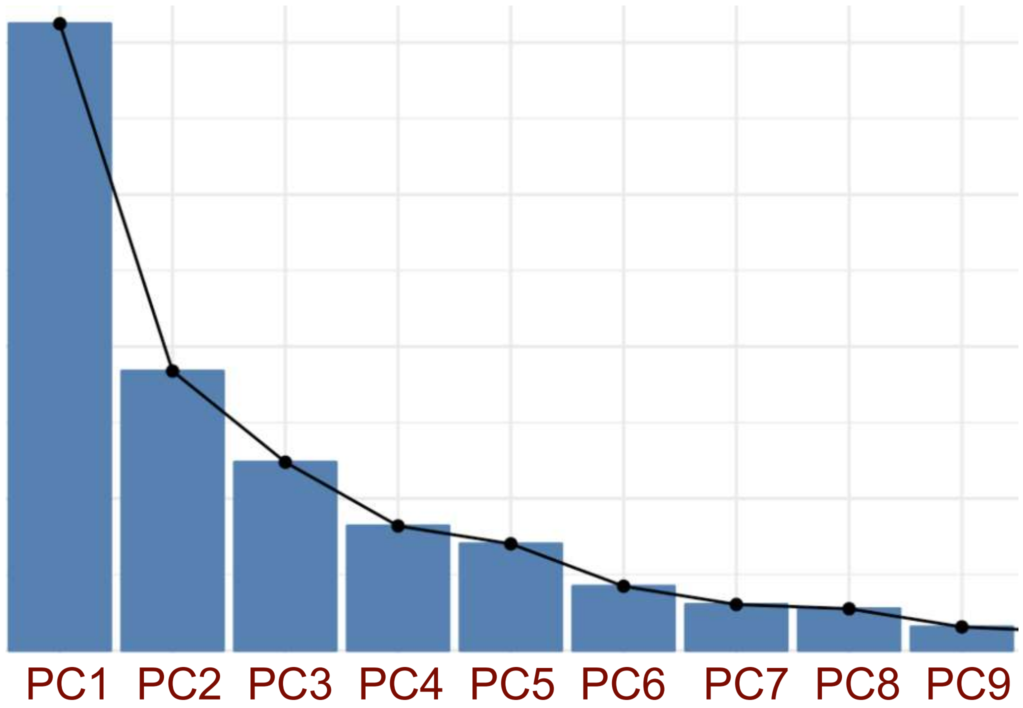
weigend.com



Wavy AI Research Foundation

6.1.3 Principal Components Analysis (PCA)

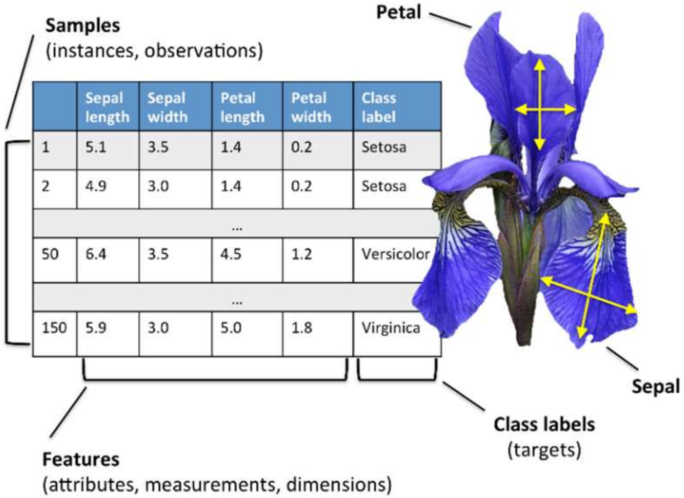
- In highly dimensional datasets, **the vast majority of the variance in the data is often captured by a small number of principal components.**
- A **plot of the distribution of the variance across principal components may look like this:**



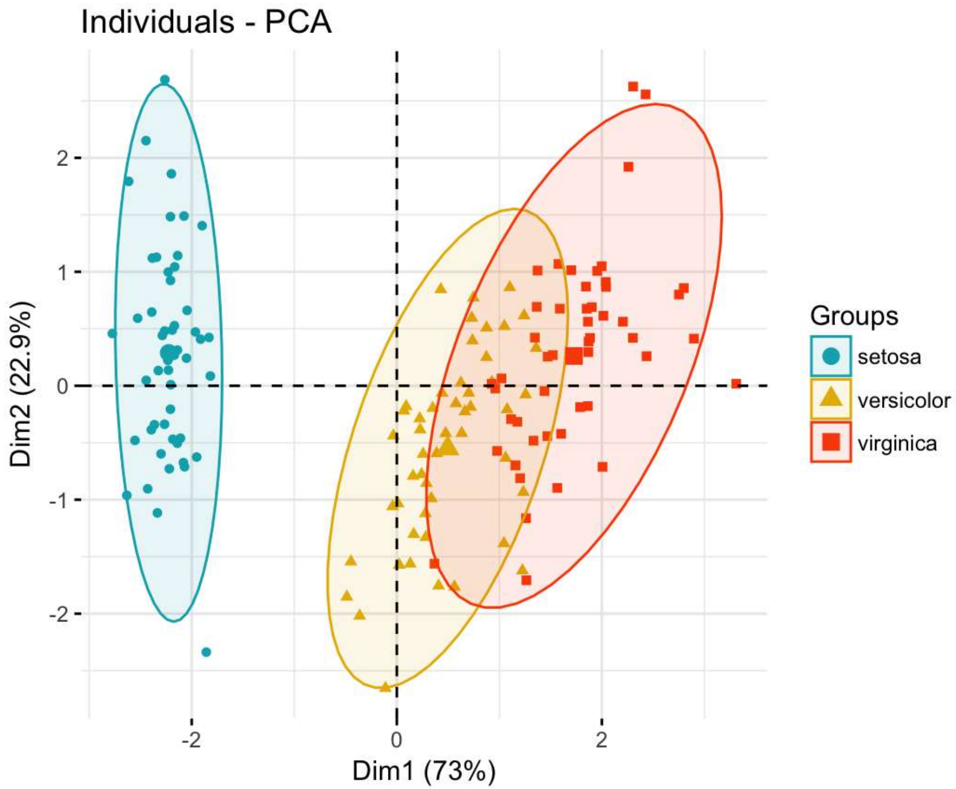
6.1.3 Principal Components Analysis (PCA)


```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width



PCA identifies the combination of attributes (PCs, or directions in the feature space) **that account for the most variance in the data.**



Here we plot the different samples on the 2 first PCs.

6.1.3 Principal Components Analysis (PCA)

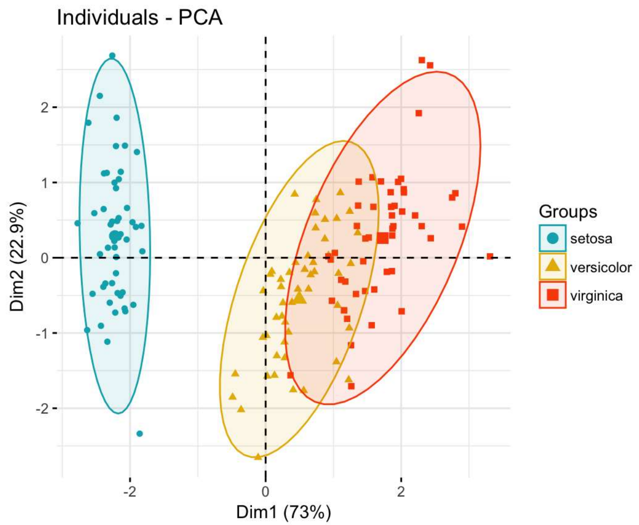
In the theoretical class, we saw:


```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the most variance in the data.

In practice:



How to do PCA Visualization using R:

The following functions, from factoextra package can be used:

`fviz_pca_ind()`: Graph of individuals

`fviz_pca_var()`: Graph of variables

`fviz_pca_biplot()` (or `fviz_pca()`): Biplot of individuals and variables

6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:


```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

First, we need to install the packages and load the libraries:

```
> install.packages("devtools")
> library("devtools")
> install.packages("factoextra")
> library("factoextra")
```

Afterwards, we prepare the dataframe if needed:

The variable Species (index = 5) is removed (not numerical)

We use ‘`center=TRUE`’ to center the variables to 0 and we scale them to have variance 1 by using ‘`scale.=TRUE`’

```
> iris_pca<-prcomp(iris[,-5], center=TRUE, scale.=TRUE)
```

6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:

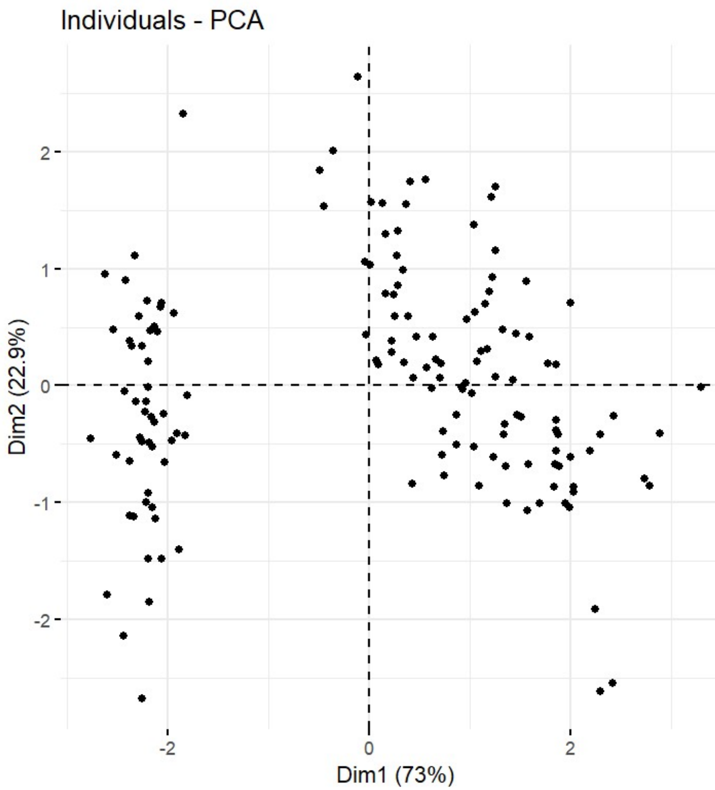
3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

How to do PCA Visualization using R:

```
fviz_pca_ind(iris_pca, geom="point")
```

Graph of individuals using only points



6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

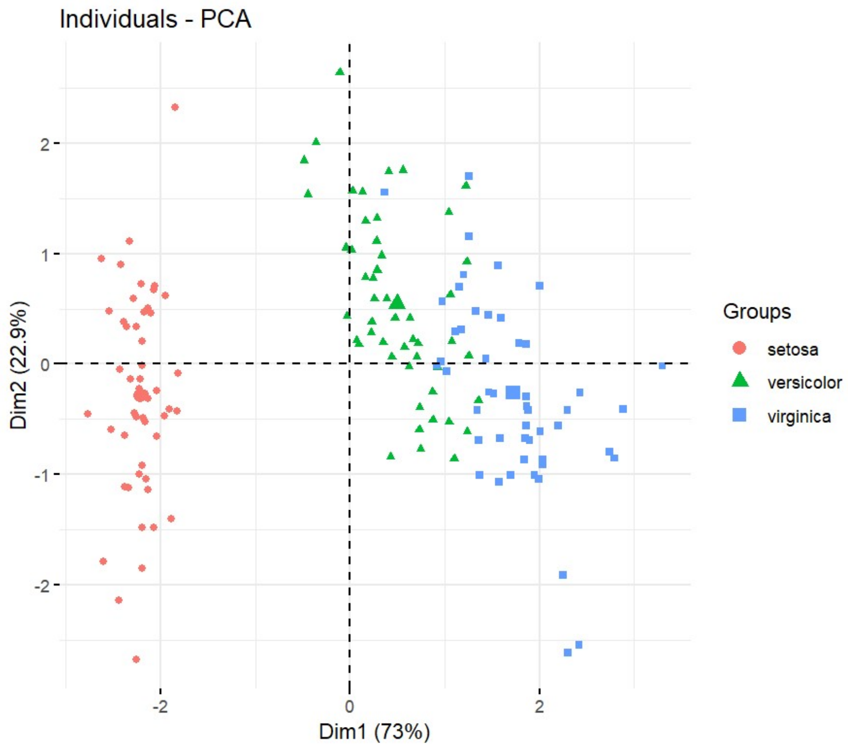
How to do PCA Visualization using R:

```
fviz_pca_ind(iris_pca, geom=“point”)
```

Graph of individuals using only points

```
fviz_pca_ind(iris_pca, label="none",
habillage=iris$Species)
```

To Color individuals by groups

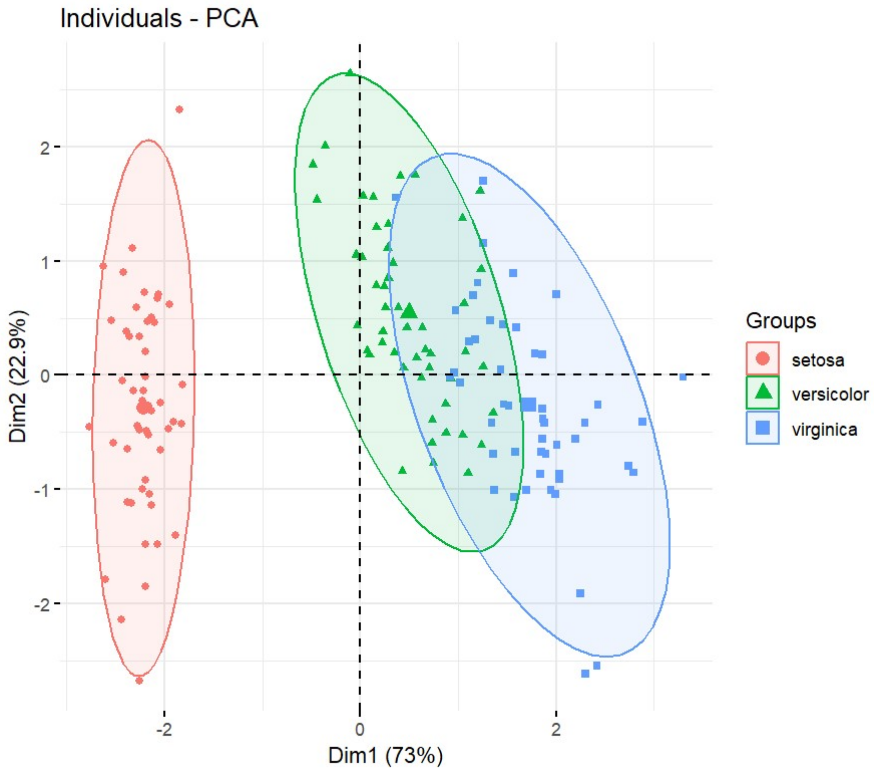


6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```



How to do PCA Visualization using R:

`fviz_pca_ind(iris_pca, geom="point")`

Graph of individuals using only points

`fviz_pca_ind(iris_pca, label="none", habillage=iris$Species)`

To Color individuals by groups

`fviz_pca_ind(iris_pca, label="none", habillage=iris$Species, addEllipses=TRUE, ellipse.level=0.95)`

To add ellipses

6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

More options in: http://www.sthda.com/english/wiki/fviz-pca-quick-principal-component-analysis-data-visualization-r-software-and-data-mining

How to do PCA Visualization using R:

`summary(iris_pca)` # Give us the importance of the components

```
> summary(iris_pca)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
> 
```

6.1.3 Principal Components Analysis (PCA)

Summarizing:

- PCA is a very interpretable method.
- **Each PC** is well-defined as we know that it **is orthogonal to the other dimensions.**
- **We can obtain the variance that is explained by each PC to select an appropriate number of dimensions**

Weakness of PCA:

It tends to be highly affected by outliers in the data

To overcome this issue many robust versions of PCA has been developed: RandomizedPCA, sparsePCA, etc

PCA works best only with continuous data

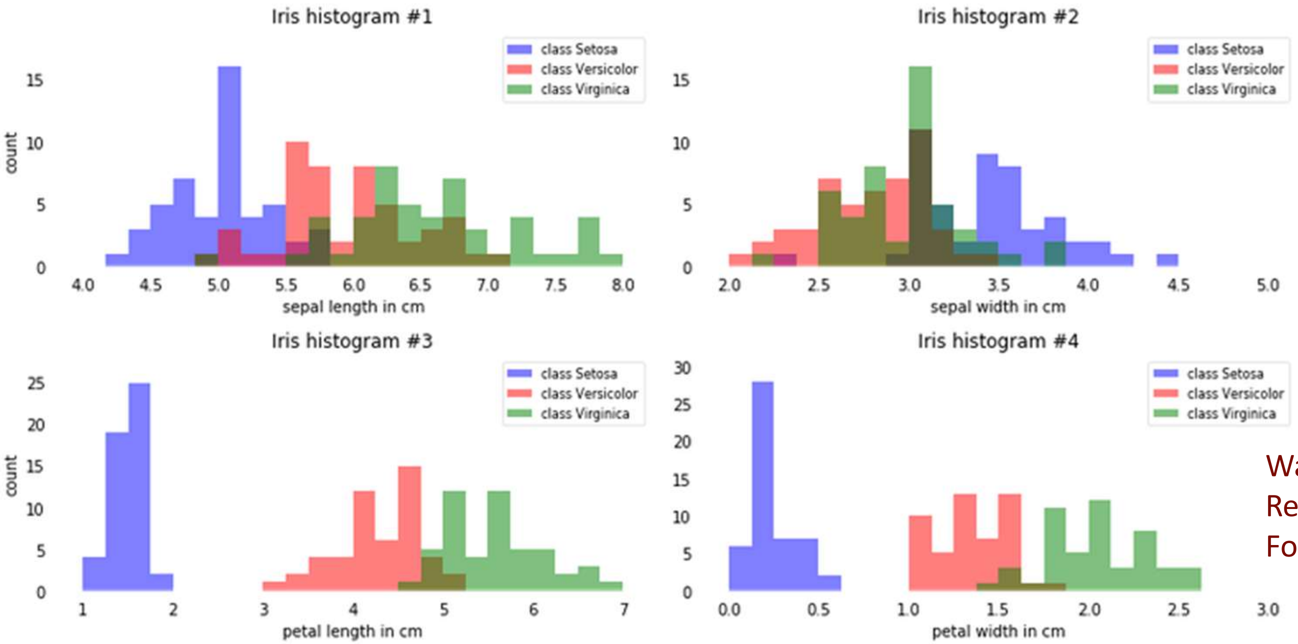
6.1.3 Principal Components Analysis (PCA)

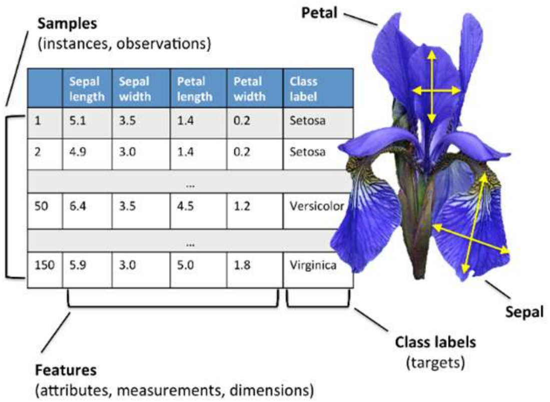
$$\mathbf{X} = \begin{bmatrix} x1_{\text{sepal length}} & x1_{\text{sepal width}} & x1_{\text{petal length}} & x1_{\text{petal width}} \\ \dots & & & \\ x2_{\text{sepal length}} & x2_{\text{sepal width}} & x2_{\text{petal length}} & x2_{\text{petal width}} \end{bmatrix}, y = \begin{bmatrix} \omega_{\text{iris-setosa}} \\ \dots \\ \omega_{\text{iris-virginica}} \end{bmatrix}$$

En numèriques

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width





Wavy AI Research Foundation

! Remark: **For low-dimensional datasets like Iris, those histograms would already be very informative.**

6.1.4 Linear Discriminant Analysis (LDA)

LDA seeks to best **separate** (or discriminate) **the samples** in the training dataset **by their class value.**

The fundamental idea of linear combinations goes back as far as the 1960s

✓ **The idea behind LDA: to find a new feature space to project the data in order *to maximize classes separability***

In 1988, the statiscian Ronald Fisher proposed **:**

- Maximize the function that represents the difference between the means, normalized by a measure of the within-class variability

6.1.4 Linear Discriminant Analysis (LDA)

The Fisher's model seeks to **find a linear combination of input variables** that:

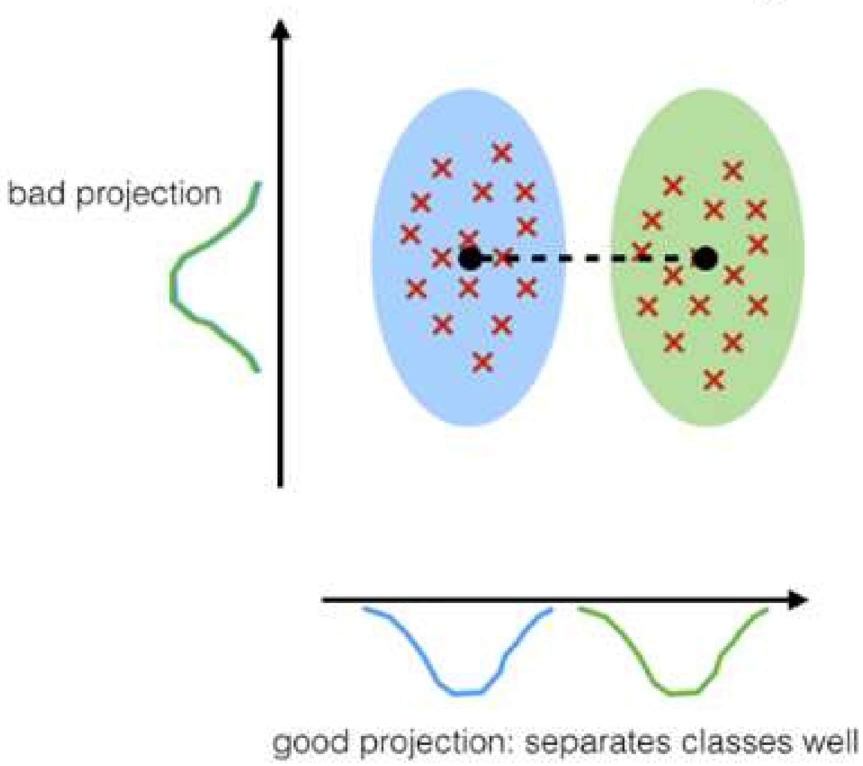
- achieves the **maximum separation** for samples **between classes** (class centroids or means),
- and the **minimum separation** of samples **within each class**.

The LDA takes the mean value for each class and considers variants to make predictions assuming a Gaussian distribution

6.1.4 Linear Discriminant Analysis (LDA)

LDA seeks to best **separate** (or discriminate) **the samples** in the training dataset **by their class value.**

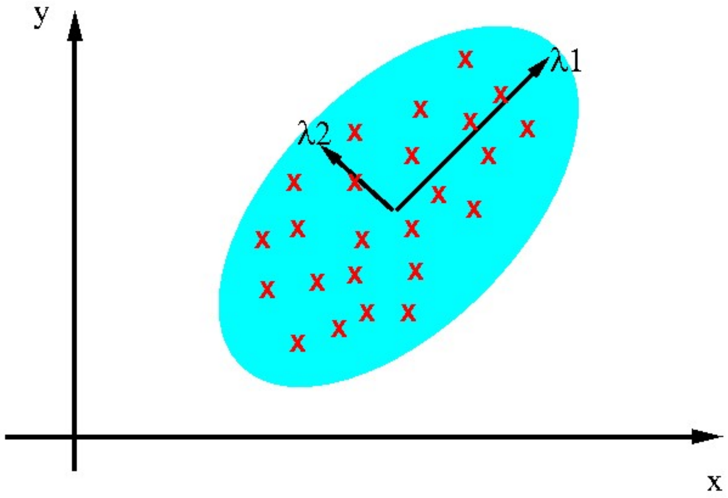
Maximizing the component axes for class-separation:



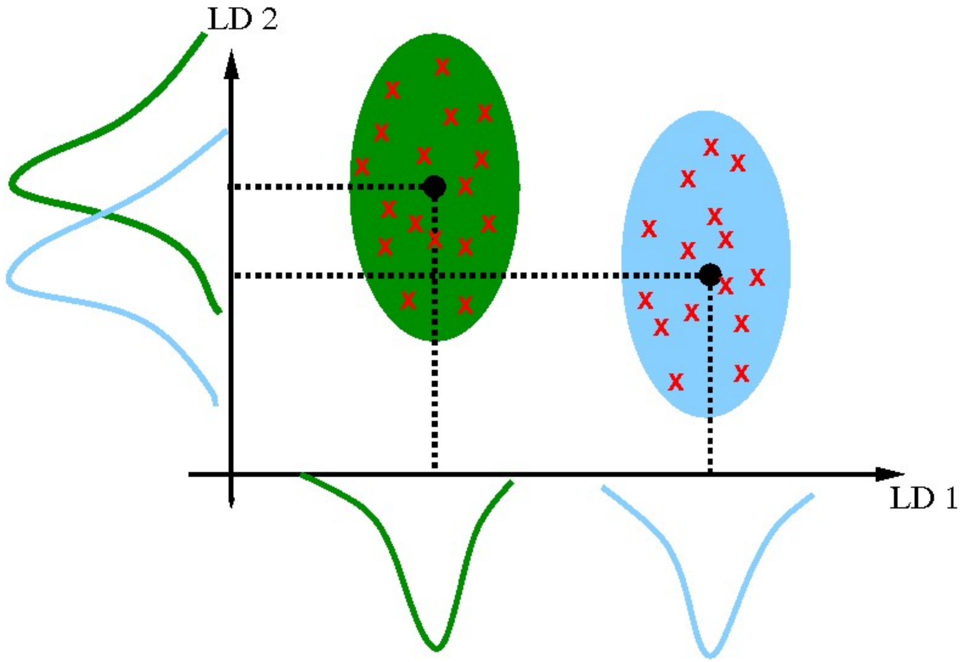
Wavy AI Research Foundation

6.1.4 PCA versus LDA

PCA: component axes that maximize the variance



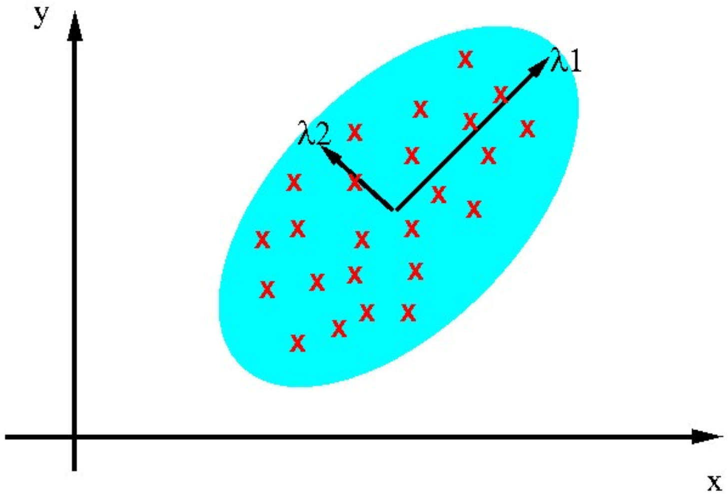
LDA: maximizing the component axes for class−separation



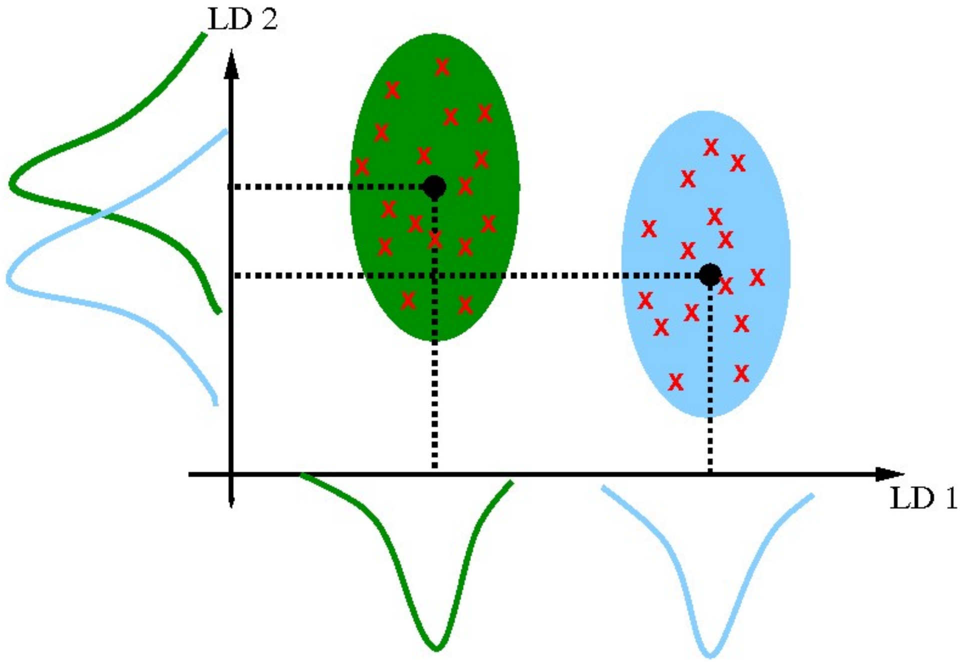
- **Both, LDA and PCA are linear transformation techniques that are commonly used for dimensionality reduction (both are techniques for the data Matrix Factorization)**

6.1.4 PCA versus LDA

PCA: component axes that maximize the variance



LDA: maximizing the component axes for class−separation



- **PCA** is unsupervised algorithm that **attempts to find the orthogonal component axes of maximum variance in a dataset**
- **while the goal of LDA** as supervised algorithm **is to find the feature subspace that optimizes class separability.**

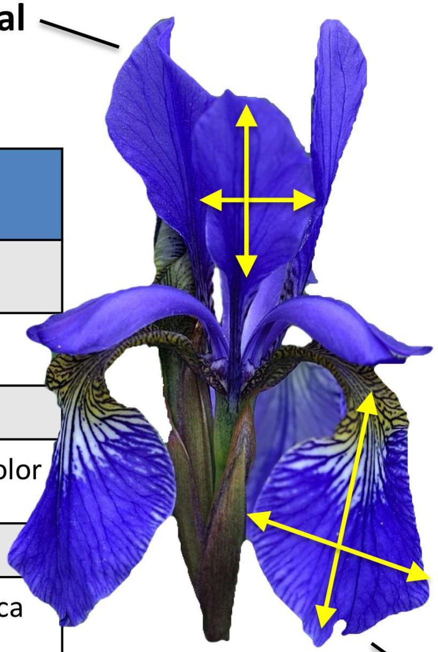
6.1.4. PCA versus LDA

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
…					
50	6.4	3.5	4.5	1.2	Versicolor
…					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)

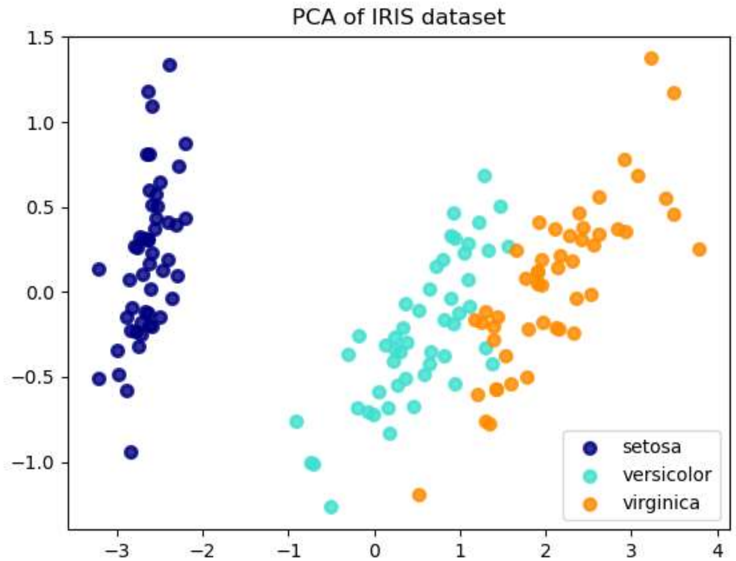


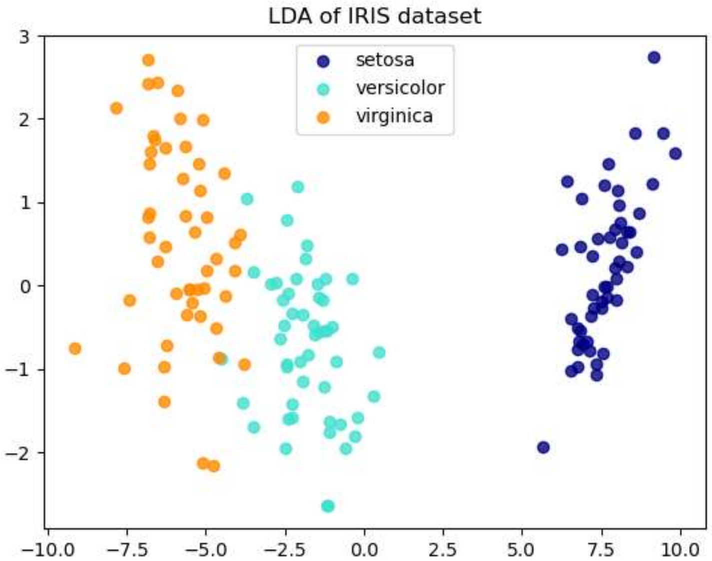

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
```

6.1.4. PCA versus LDA

PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the **most variance in the data.**

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width





LDA: tries to **identify attributes** that account for the most **variance between classes**

6.1.4. Linear Discriminant Analysis (LDA)

Steps: (see this link for an example with iris dataframe)

1. Compute the **d-dimensional mean vector for the different classes** from the dataset. **(in PCA was for each direction)**
2. Compute the Scatter matrix (in between class and within the class scatter matrix)
3. Sort the Eigen Vector by decreasing Eigen Value order and choose *k* eigenvector with the largest eigenvalue to from a *dxk* dimensional matrix W (where every column represent an eigenvector)
4. Used *dxk* eigenvector matrix to transform the sample onto the new subspace. This can be summarised by the matrix multiplication:

$$Y = XW$$

where *X* is a *nxd* dimension matrix representing the *n* samples and you are transformed *nxk* dimensional samples in the new subspace.

6.1.4. Linear Discriminant Analysis (LDA)

LDA **can be useful in areas like image recognition and predictive analysis in marketing**

Weakness of LDA:

- LDA **does not work well if the design is not balanced** (i.e. the number of objects in various classes are (highly) different)
- If the **distribution of your data is significantly non-Gaussian**, the LDA **might not perform very well**.
- It is **sensitive to overfit**
- **LDA is not applicable (inferior) for non-linear problems**

Thanks for your attention!

Judit Chamorro Servent

Departament de Matemàtiques

judit.chamorro@uab.cat

GRAU EN ENGINYERIA DE DADES

104365
TEMA - 7. Sistemes Avançats I

Departament de Matemàtiques

Escola d'Enginyeria UAB

UAB

7. Advanced systems of visualization (I)

7.1 Multiple variables and dimensions

7.2 Networks

7.3 3D Data

7.4 Vector Fields

7.1 Multiple variables and dimensions. Contents

1. Introduction

   Multiple variables and dimensions

2. Visualizing many distributions at once

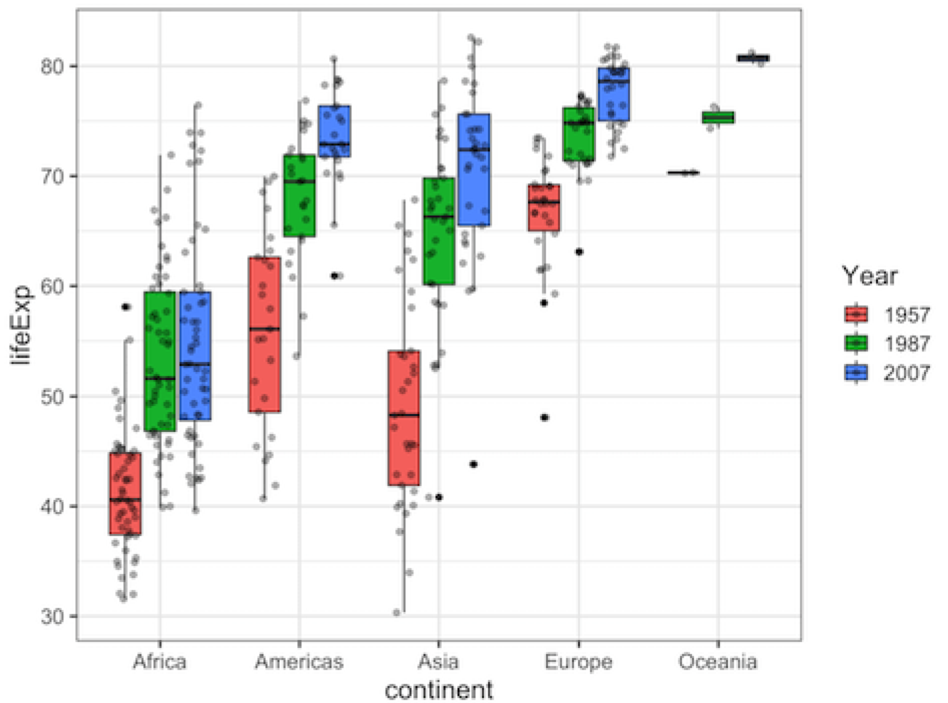
3. Visualizing many proportions at once

4. Visualizing many relations (correlations) at once : Bubble plot and Scatter plot matrices

7.1.1 Multiple variables and dimensions. Introduction

There are **large datasets**, **containing much more information than can be shown in a plot.**

- Some datasets can be shown **in a single figure panel by grouping variables**.



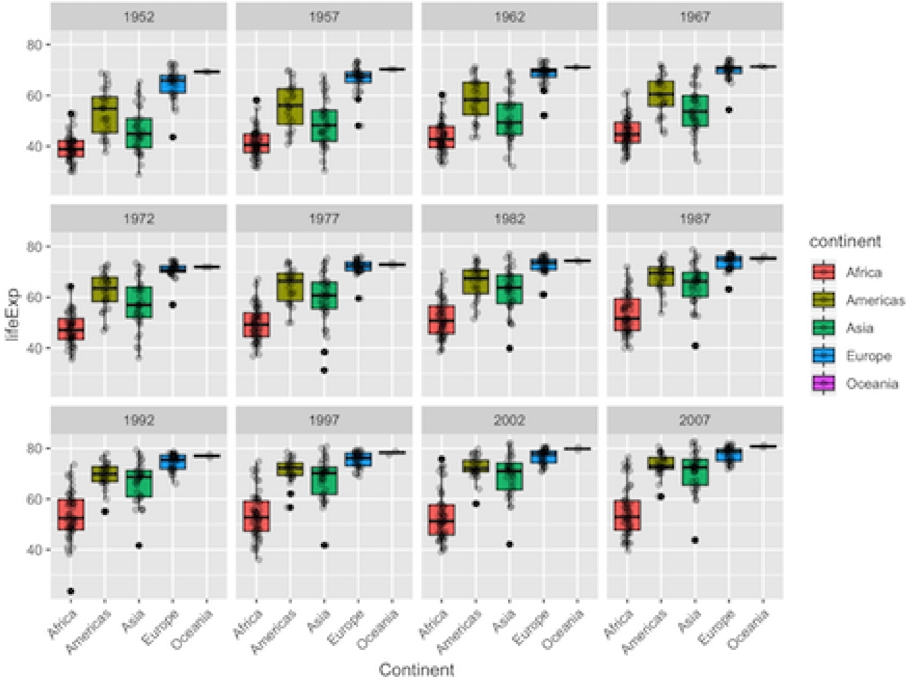
7.1.1 Multiple variables and dimensions. Introduction

There are large datasets, containing much more information than can be shown in a plot.

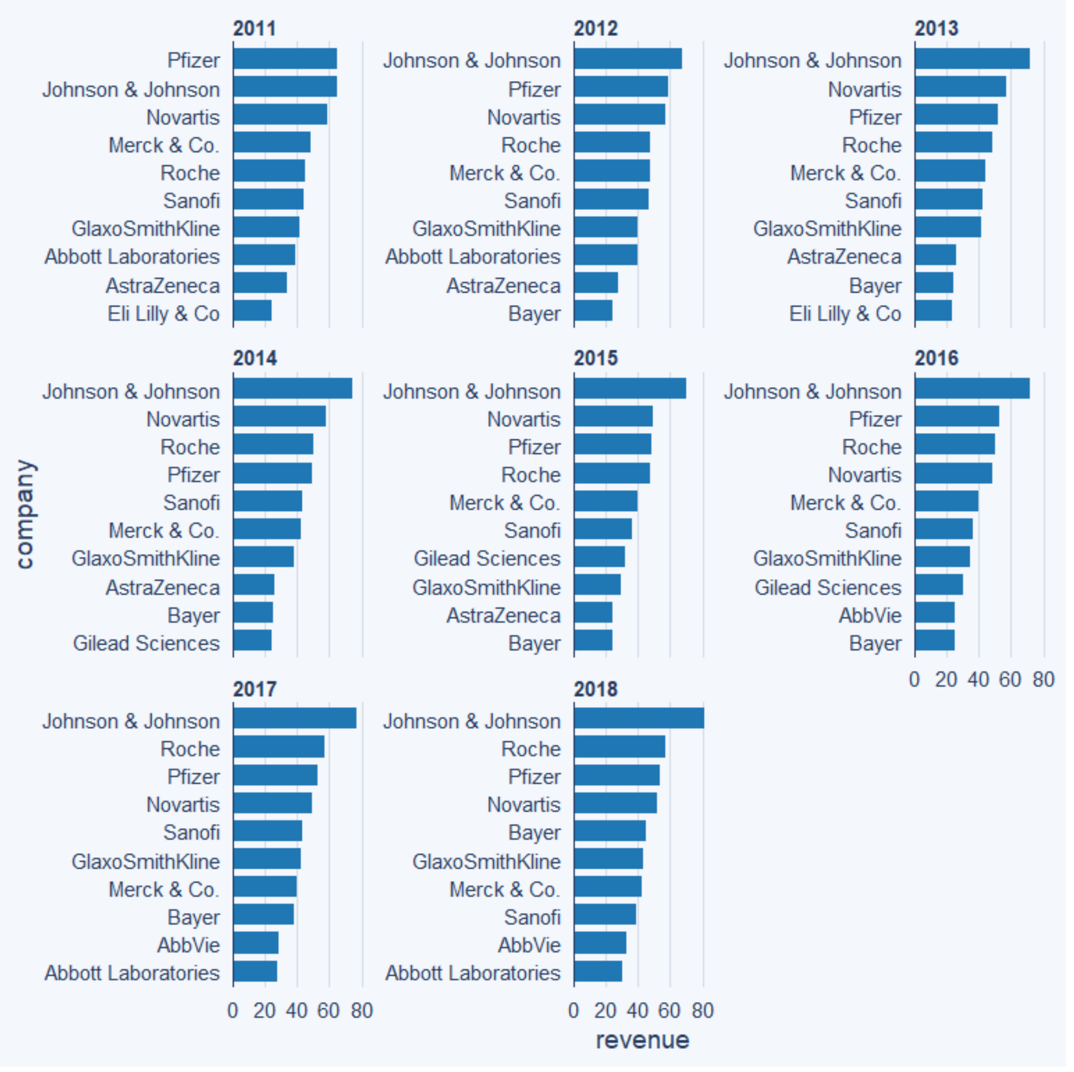
- However, **for more complex datasets, it can be helpful to create multi-panel figures**.

These are figures that consist of *multiple figure panels where* ***each panel shows some subset of the data.***

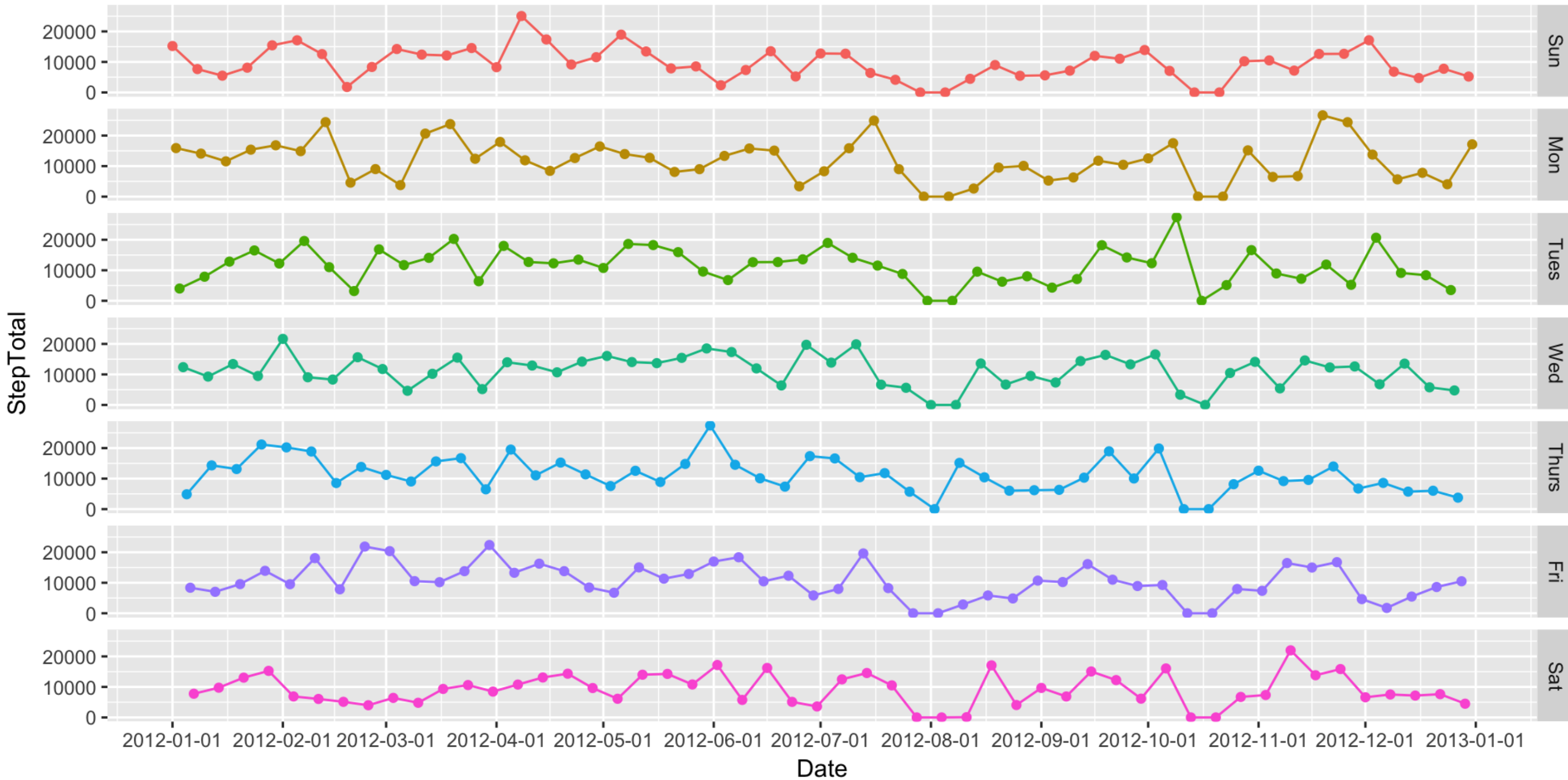
Exemple: Facets en R



7.1.1 Multiple variables and dimensions. Introduction



7.1.1 Multiple variables and dimensions. Introduction



Time Series data (one row for each weekday)

7.1.2 Visualizing many distributions at once

There are many scenarios in which we want **to visualize multiple distributions at the same time.**

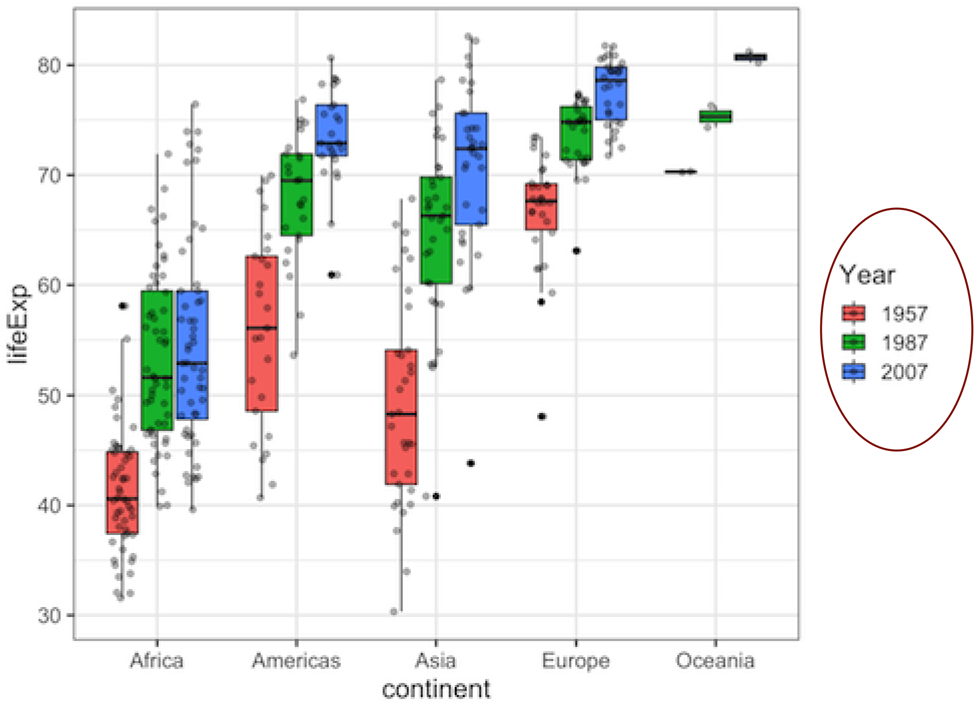
For this, it is helpful to think in terms of the response variable and one or more grouping variables.

- The **response variable** is the variable *whose distributions we want to show.*
- The **grouping variables** define *subsets of the data with distinct distributions of the response variable.*

7.1.2 Visualizing many distributions at once

- The **response variable:** is the variable whose distributions we want to show (***lifeExp***).

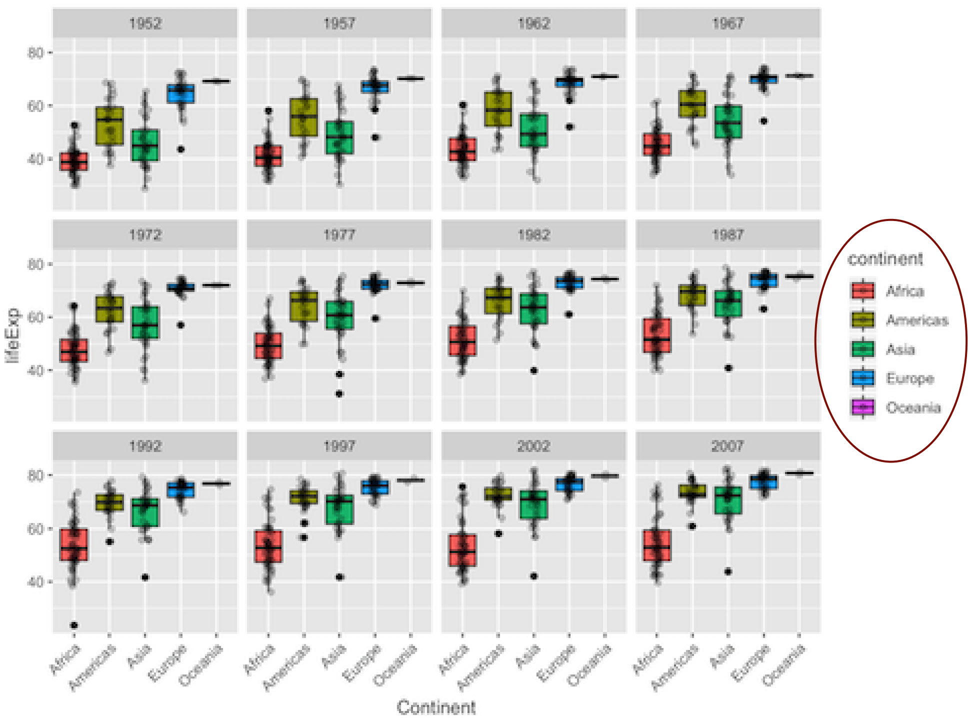
- The **grouping variables:** define subsets of the data with distinct distributions of the response variable ***(continent / year)***



7.1.2 Visualizing many distributions at once

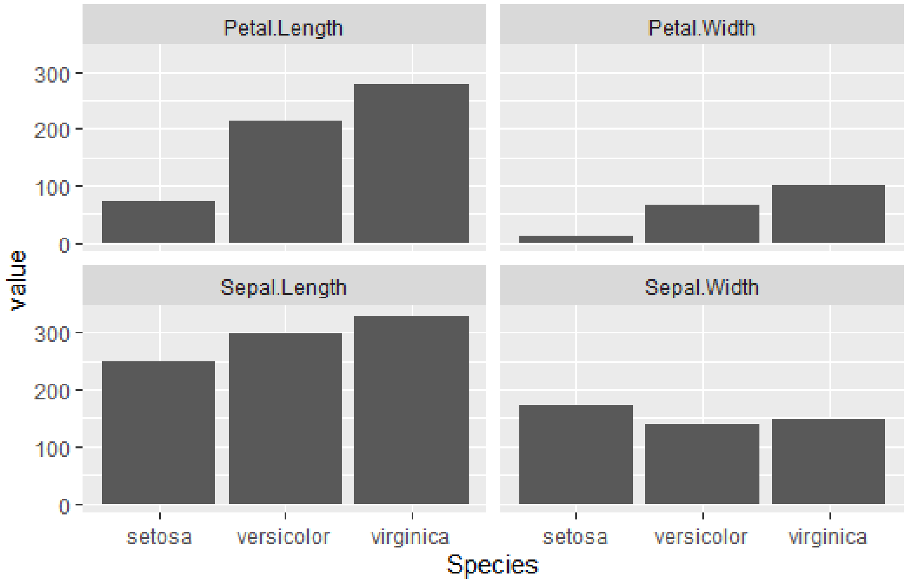
- The **response variable:** is the variable whose distributions we want to show (***lifeExp***).

- The **grouping variables:** define subsets of the data with distinct distributions of the response variable ***(continent / year)***



7.1.2 Visualizing many distributions at once

- The **response variable:** is the variable whose distributions we want to show.
- The **grouping variables:** define subsets of the data with distinct distributions of the response variable.



Seminar 4

```
>iris_long<-gather(iris, metric, value, -Species)
>ggplot(iris_long)+aes(Species, value)+geom_bar(stat='identity')+ facet_wrap(~ metric)
```

7.1.2 Visualizing many distributions at once

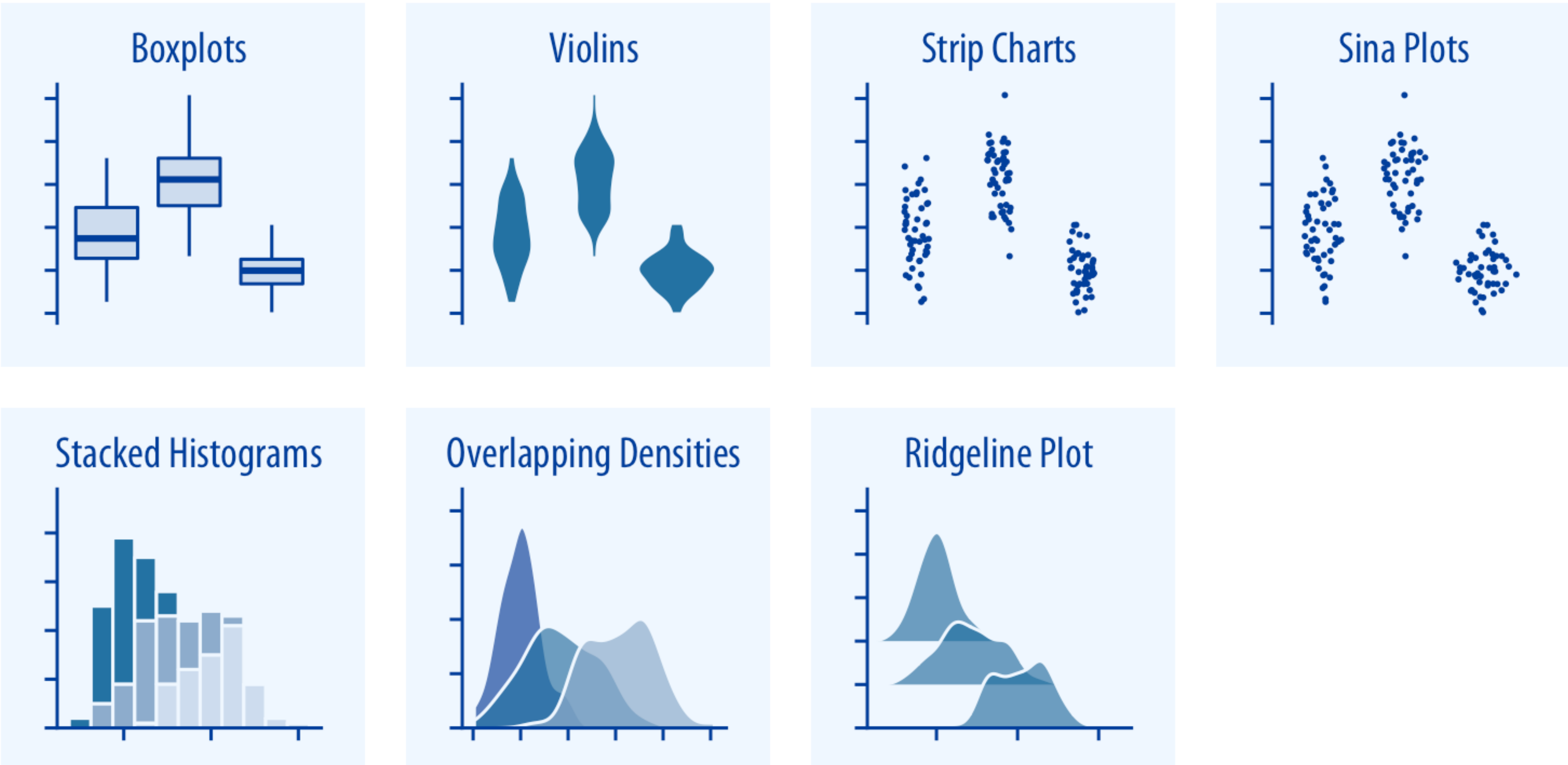
- The **response variable:** is the variable whose distributions we want to show (***value of the metric in cm– length or width***).

- The **grouping variables:** define subsets of the data with distinct distributions of the response variable ***(species / petal/sepal)***

Seminar 4

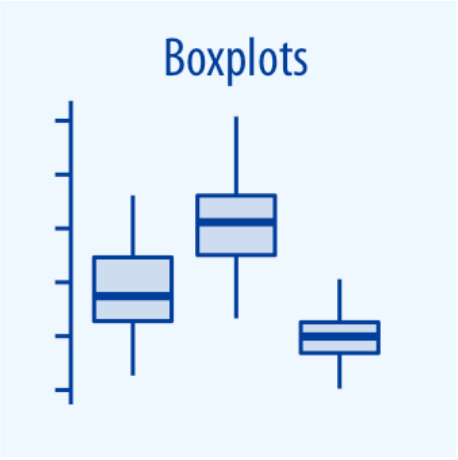
```
>iris_long<-gather(iris, metric, value, -Species)
>ggplot(iris_long)+aes(Species, value)+geom_bar(stat='identity')+
facet_wrap(~ metric)
```

7.1.2 Visualizing many distributions at once



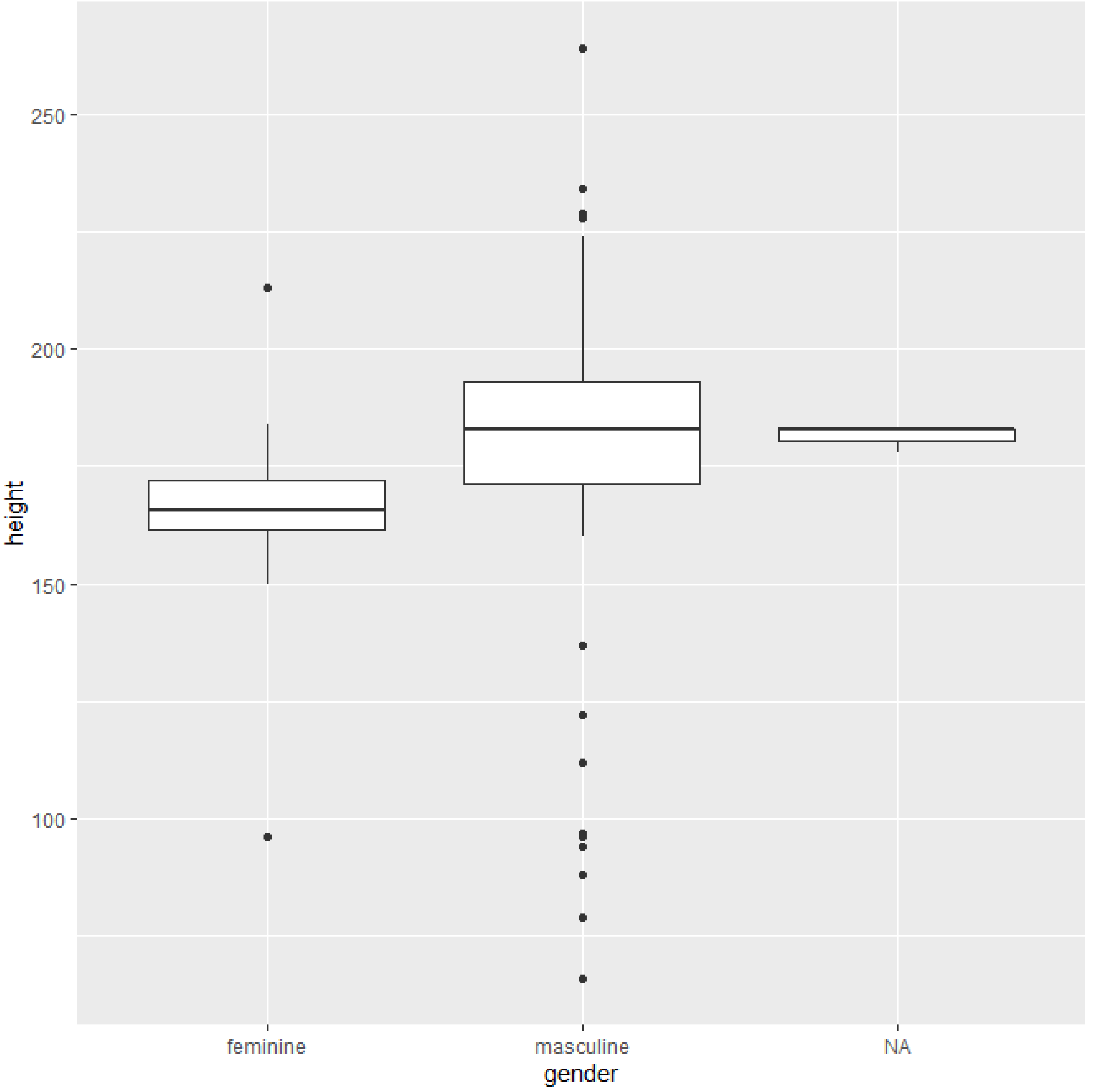
Claus Wilke

7.1.2 Visualizing many distributions at once



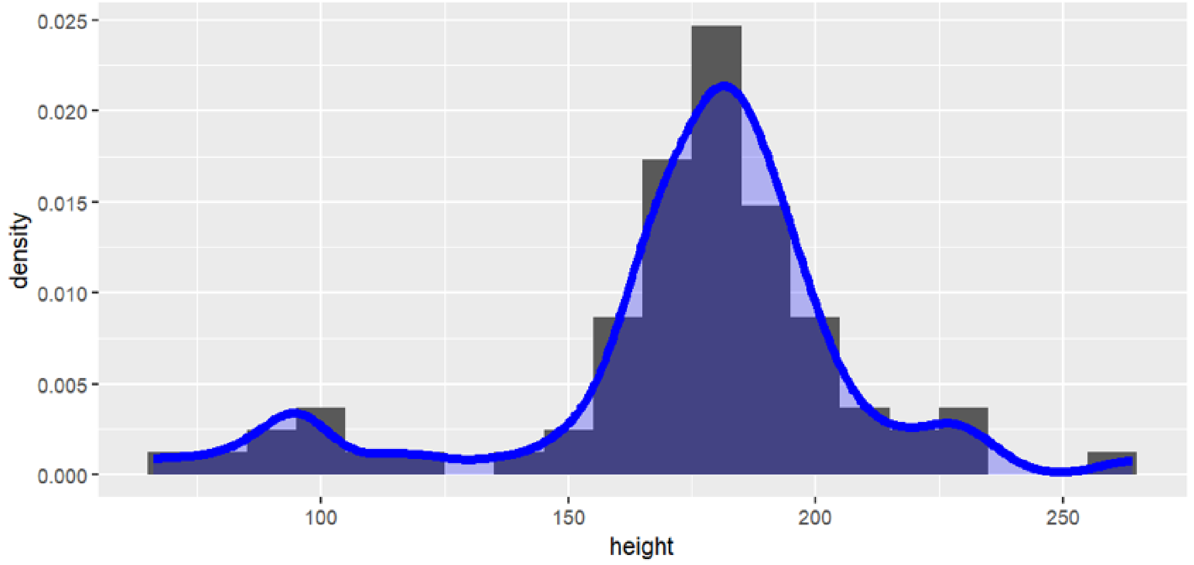
Seminar 3

```
>ggplot(starwars,aes(x=gender,
y=height))+geom_boxplot()
```



7.1.2 Visualizing many distributions at once

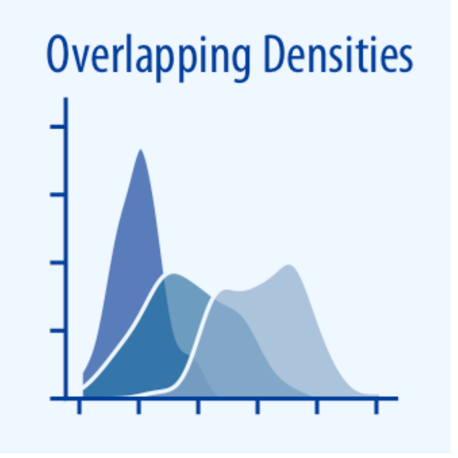
Una altra opció és adjuntar ambdues gràfiques en una fent servir una transparència. Per això podeu posar **`aes(y=..density..)`** en el `geom_histogram`.




```
>ggplot(starwars,aes(x=height))+geom_histogram(binwidth=10, aes(y=..density..))+geom_density(lwd = 2, colour = 'blue', fill = 'blue', alpha = 0.25)
```

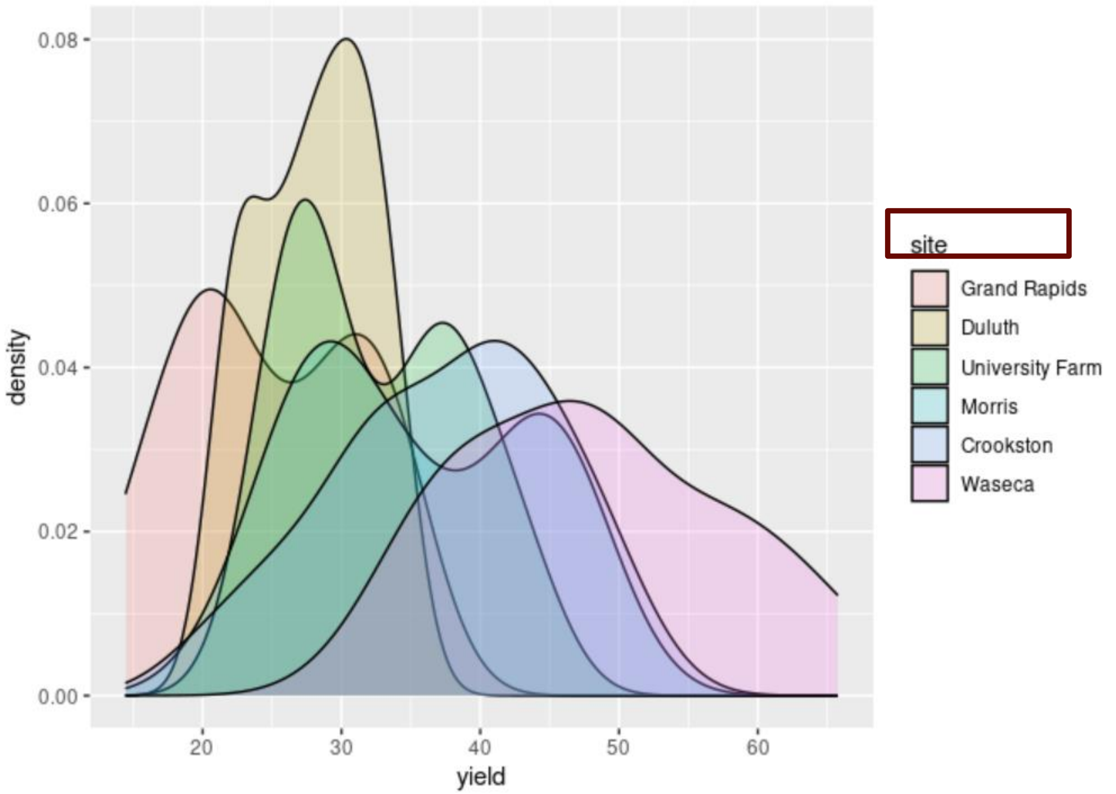
Nota: `lwd` només marca el gruix de la línia de `geom_density`

Seminar 3
Overlapping histogram and density



7.1.2 Visualizing many distributions at once

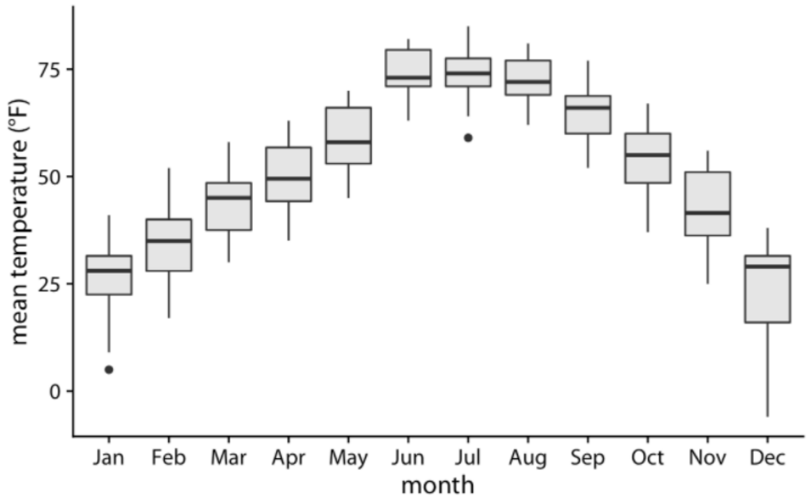
```
ggplot(barley) + geom_density(aes(x = yield, fill = site), alpha = 0.2)
```



https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/histdens.html

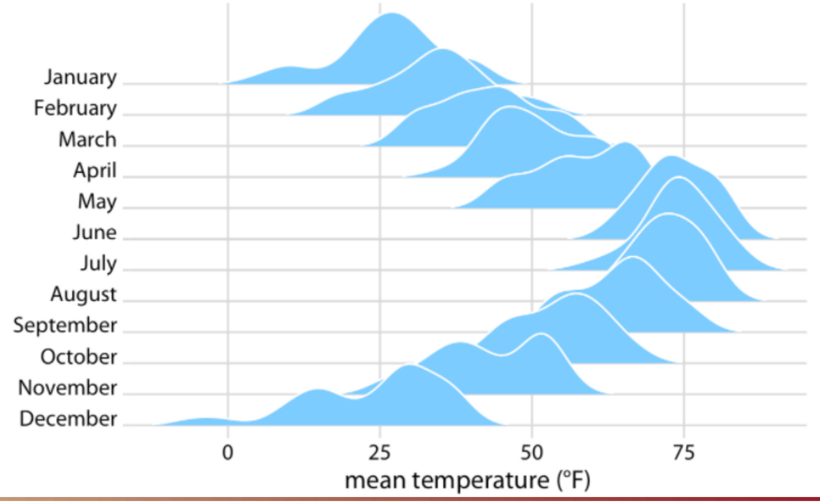
7.1.2 Visualizing many distributions at once

- **Along the vertical axis**

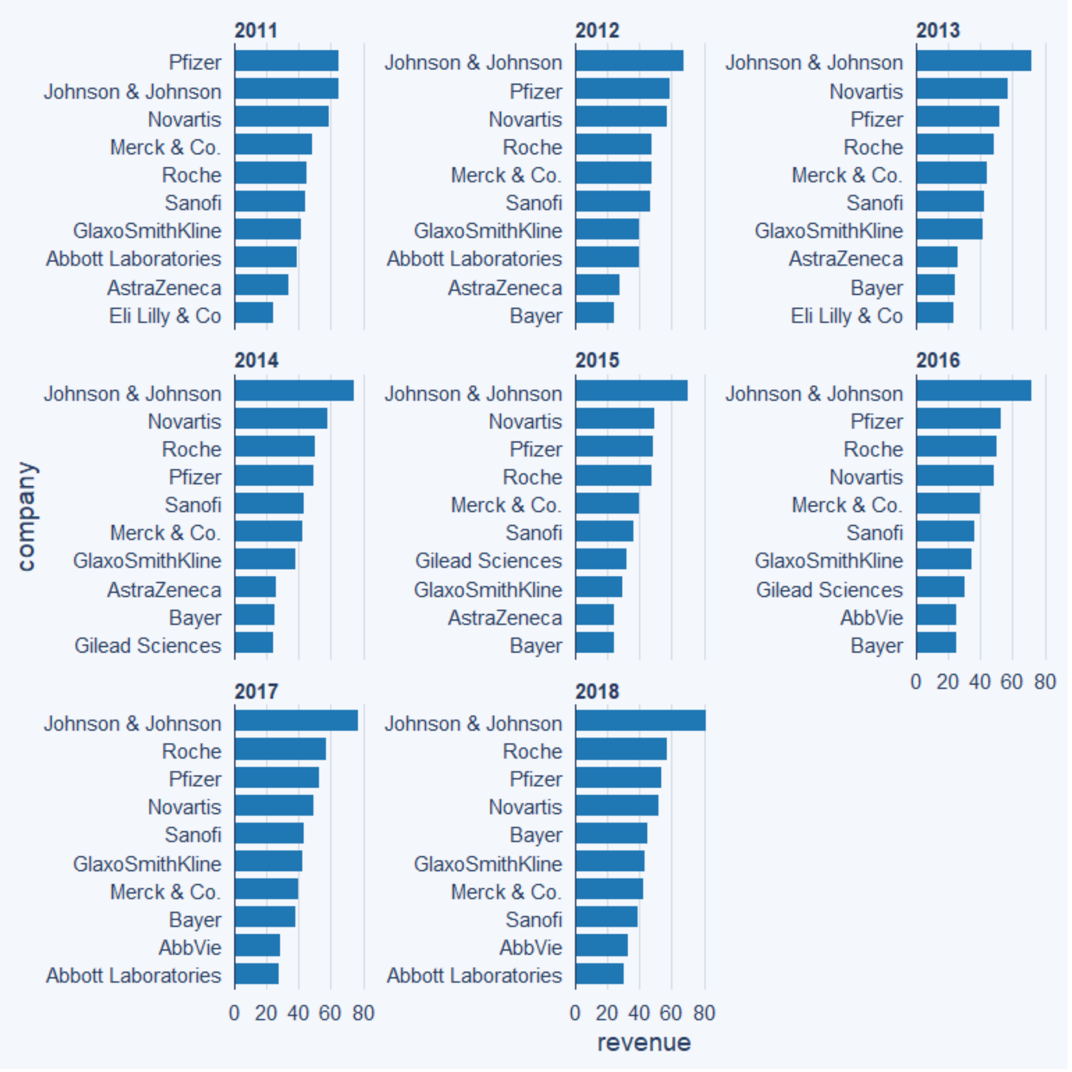


Claus Wilke

- **Along the horizontal axis**

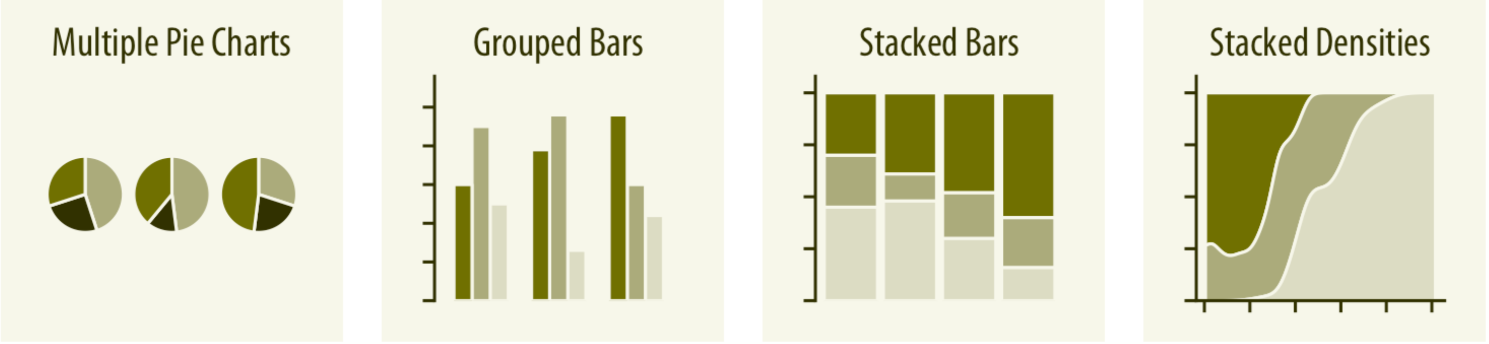


7.1.3 Visualizing many proportions at once



7.1.3 Visualizing many proportions at once

- **Pie charts** tend to be **space-inefficient and often obscure relationships.** ***(Example seminari 3)***
- **Grouped bars** work well as long as **the number of conditions compared is moderate.**
- **Stacked bars** can work **for large numbers of conditions**.
- **Stacked densities** are appropriate when **the proportions change along a continuous variable.**

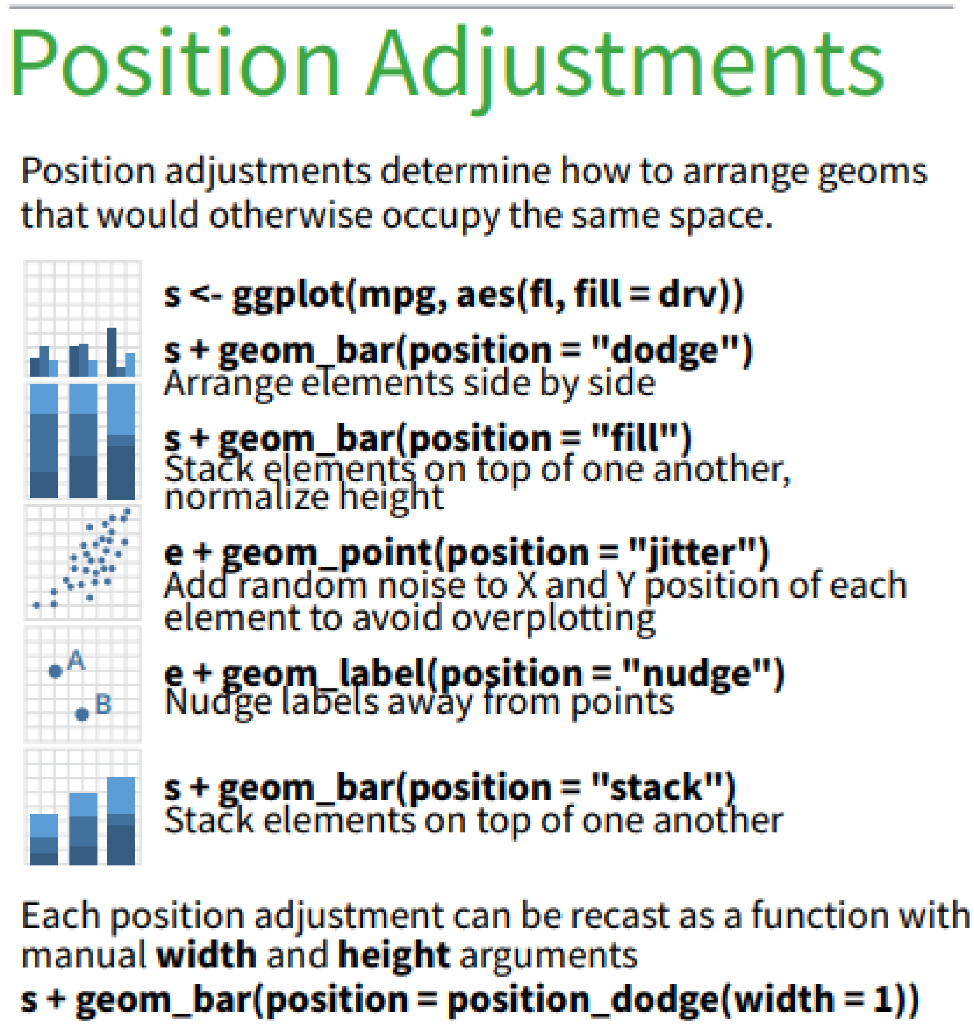


Claus Wilke

7.1.3 Visualizing many proportions at once

- **Stacked bars** can work **for large numbers of conditions**.

Claus Wilke



Position Adjustments

Position adjustments determine how to arrange geoms that would otherwise occupy the same space.

s <- ggplot(mpg, aes(fl, fill = drv))

s + geom_bar(position = "dodge")
Arrange elements side by side

s + geom_bar(position = "fill")
Stack elements on top of one another, normalize height

e + geom_point(position = "jitter")
Add random noise to X and Y position of each element to avoid overplotting

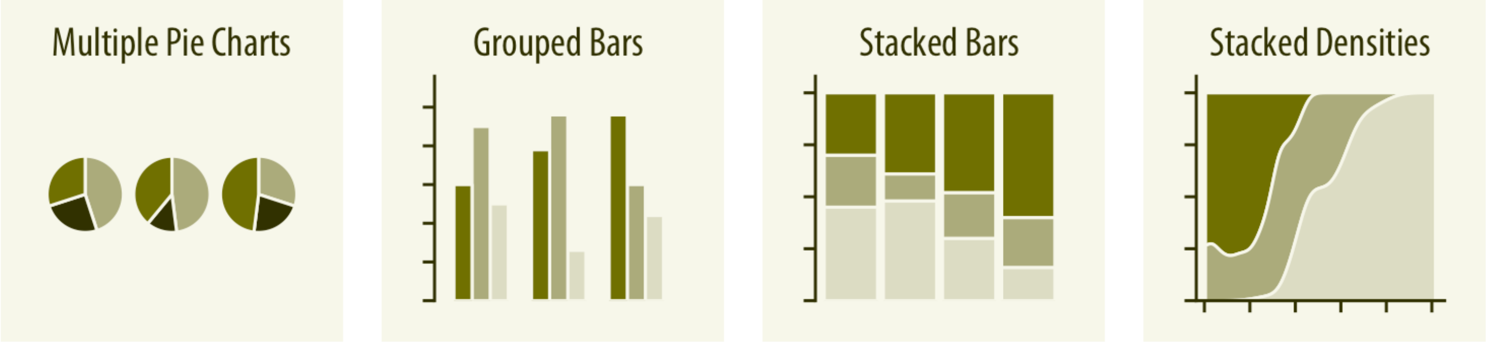
e + geom_label(position = "nudge")
Nudge labels away from points

s + geom_bar(position = "stack")
Stack elements on top of one another

Each position adjustment can be recast as a function with manual **width** and **height** arguments
s + geom_bar(position = position_dodge(width = 1))

7.1.3 Visualizing many proportions at once

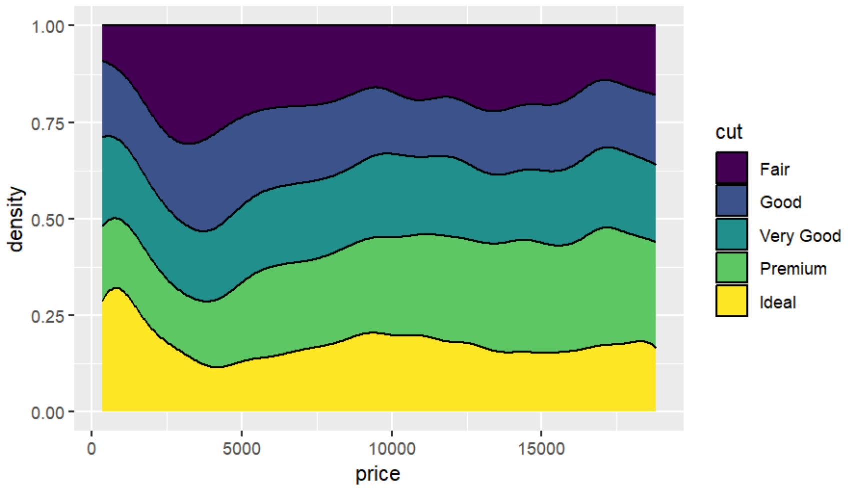
- **Pie charts** tend to be **space-inefficient and often obscure relationships.** ***(Example seminari 3)***
- **Grouped bars** work well as long as **the number of conditions compared is moderate.**
- **Stacked bars** can work **for large numbers of conditions**.
- **Stacked densities** are appropriate when **the proportions change along a continuous variable.**



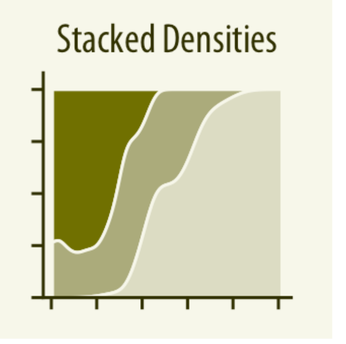
Claus Wilke

7.1.3 Visualizing many proportions at once

- **Stacked densities** are appropriate when **the proportions change along a continuous variable.**
- Stacking is a process where a chart is broken up across more than one categoric variables which make up the whole.



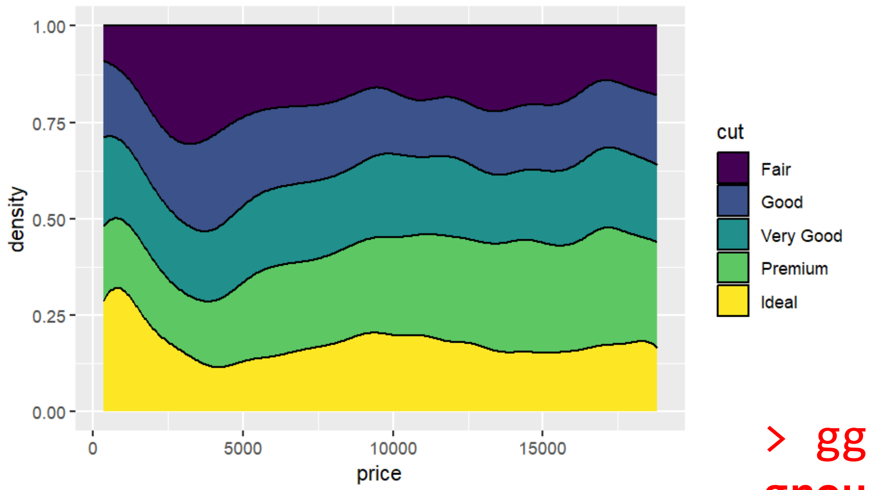
Diamonds contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds



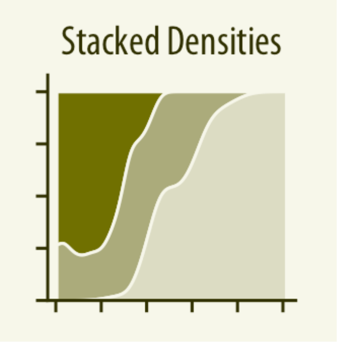
Claus Wilke

7.1.3 Visualizing many proportions at once

- **Stacked densities** are appropriate when **the proportions change along a continuous variable.**
- Stacking is a process where a chart is broken up across more than one categoric variables which make up the whole.



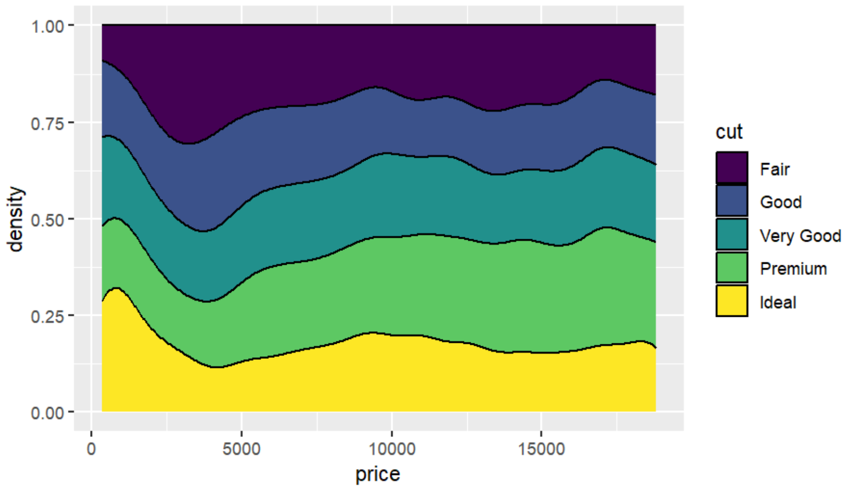
Diamonds contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds



Claus Wilke

```
> ggplot(data=diamonds, aes(x=price,
group=cut, fill=cut)) +
geom_density(adjust=1.5, position="fill")
```

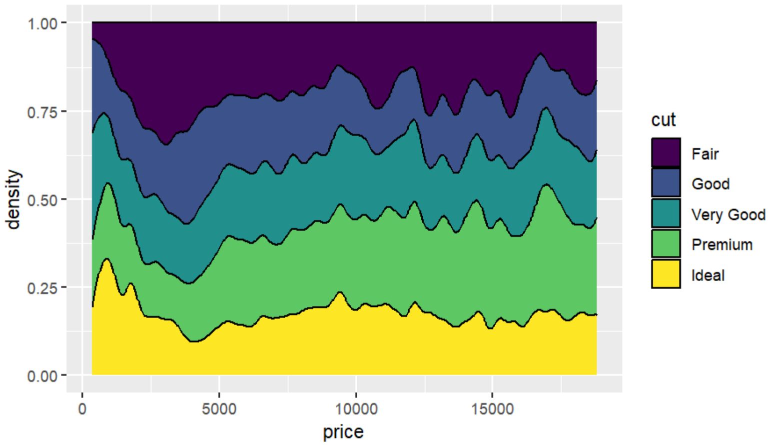
7.1.3 Visualizing many proportions at once



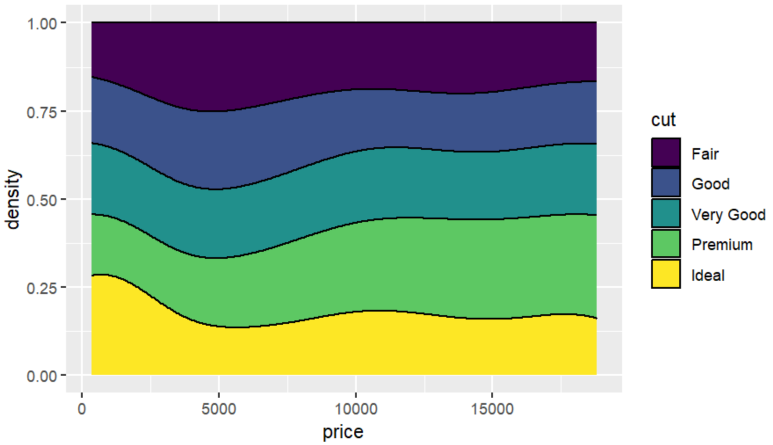
It contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds

```
> ggplot(data=diamonds, aes(x=price,
group=cut, fill=cut)) +
geom_density(adjust=1.5, position="fill")
```

A multiplicative bandwidth adjustment. This makes it possible to adjust the bandwidth while still using a bandwidth estimator



adjust=0.5



adjust=4

7.1.3 Visualizing many proportions at once

When proportions are specified **according to multiple grouping variables:**

Claus Wilke

Mosaic plots, treemaps, or parallel sets are useful visualization approaches

7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

Mosaic Plot: assumes that **each level of one grouping variable** can be **combined with each level of another grouping variable**.

Claus Wilke

7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

- **Mosaic plots** looks similar to the stacked bar plot. However, in a mosaic plot **both the heights and the width of individual shaded areas vary**

Claus Wilke

7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

- **Mosaic plots example:**

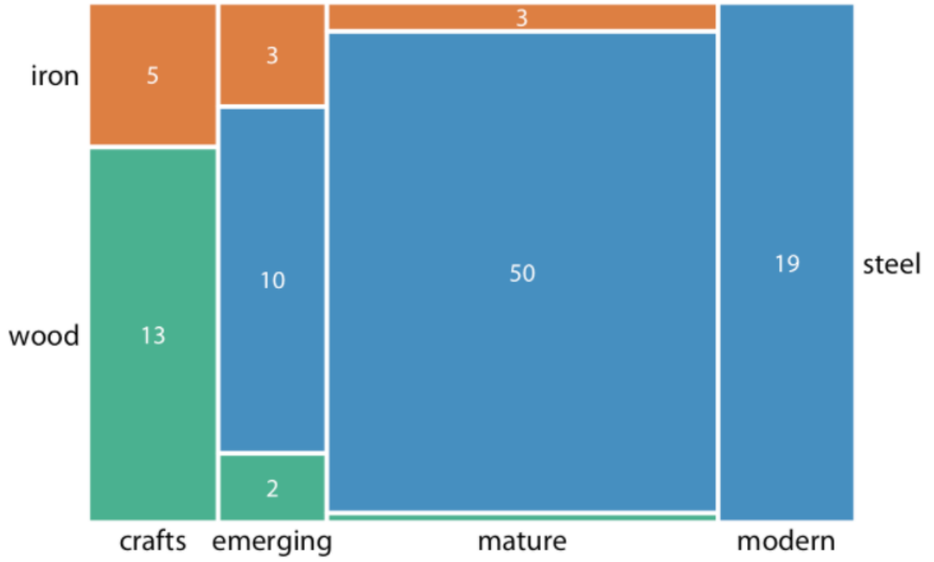


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

Claus Wilke

7.1.3 Visualizing many proportions at once

- **Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable**, whereas treemaps do not make such an assumption.

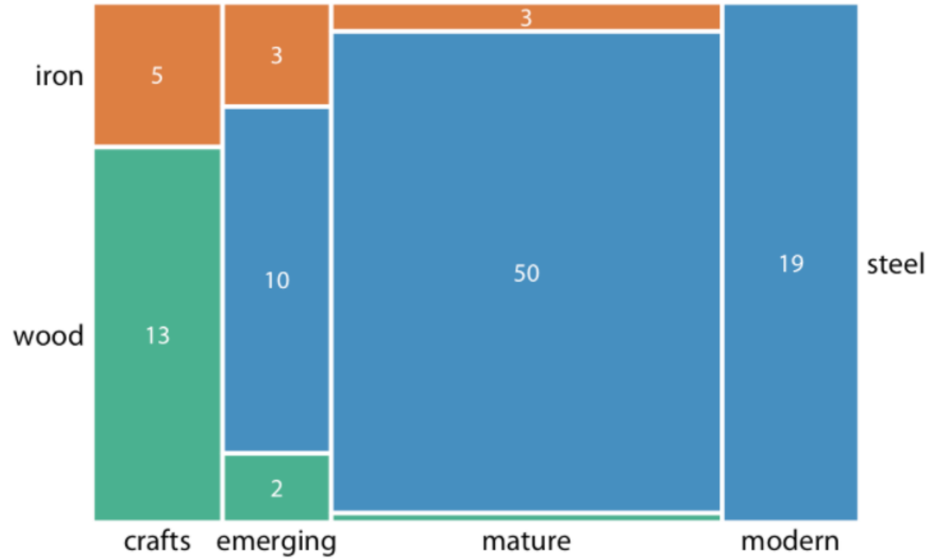


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- **Mosaic plots**

We begin by **placing one categorical variable along the *x* axis** (here, *era* of bridge construction) **and subdivide the *x* axis by the relative proportions that make up the categories.**

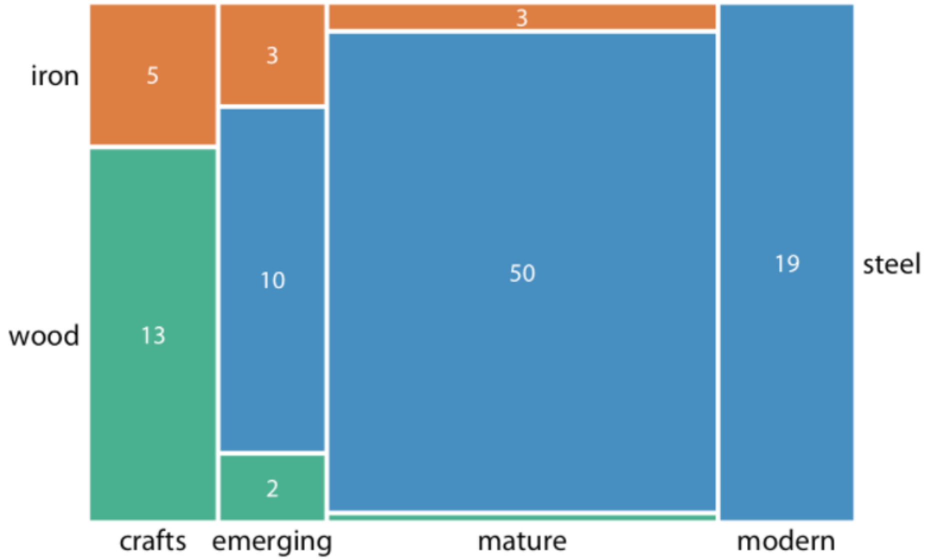


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- **Mosaic plots**

We begin by **placing one categorical variable along the *x* axis** (here, era of bridge construction) **and subdivide the *x* axis by the relative proportions that make up the categories**.

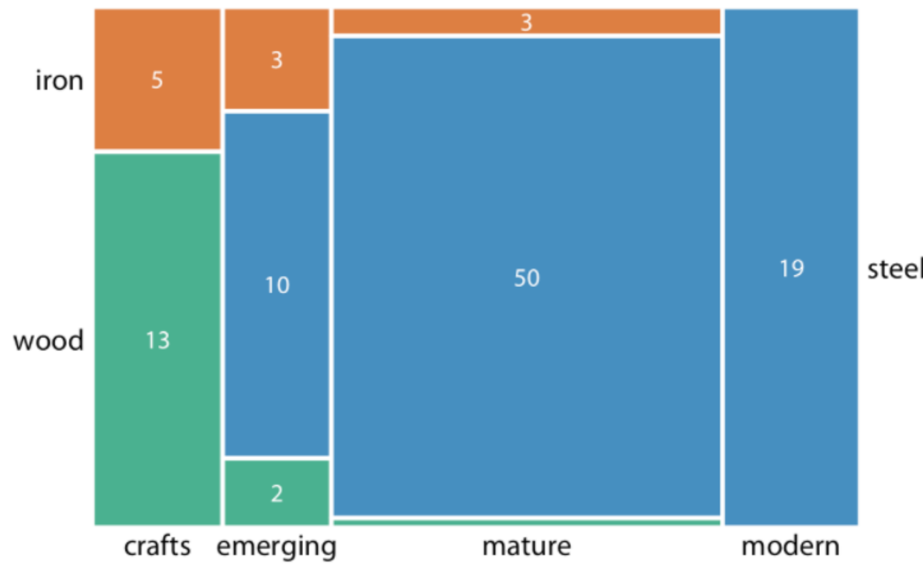


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

We then place the other categorical variable along the *y* axis (here, *building material*) and, **within each category along the *x* axis, subdivide the *y* axis by the relative proportions that make up the categories of the *y* variable.**

7.1.3 Visualizing many proportions at once

- Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable, whereas treemaps do not make such an assumption.

The result is a set of rectangles whose areas are proportional to the number of cases representing each possible combination of the two categorical variables.

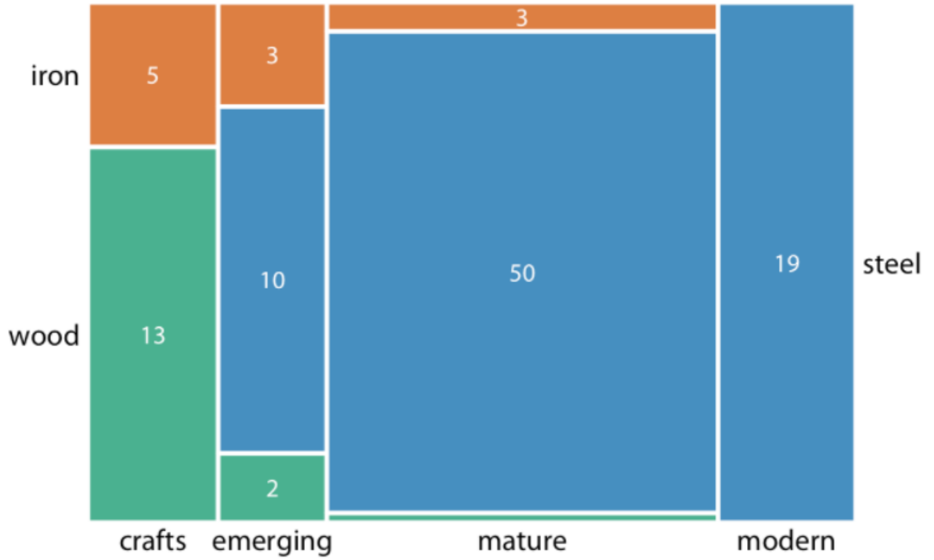


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.2 Visualizing many proportions at once

- **Treemaps** work well **even if the subdivisions of one group are entirely distinct from the subdivisions of another**.

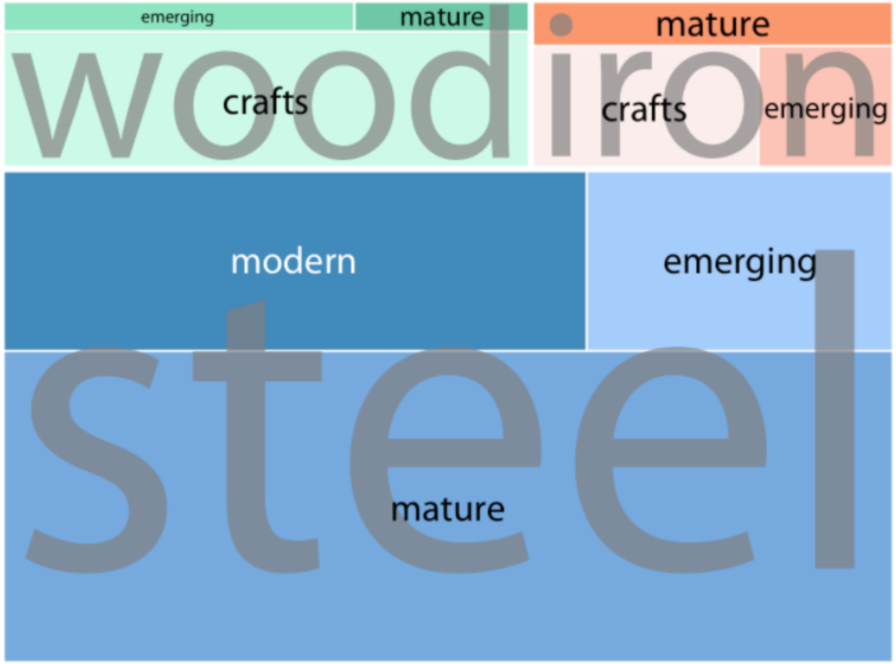


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

In a treemap, **we recursively nest rectangles inside each other.**

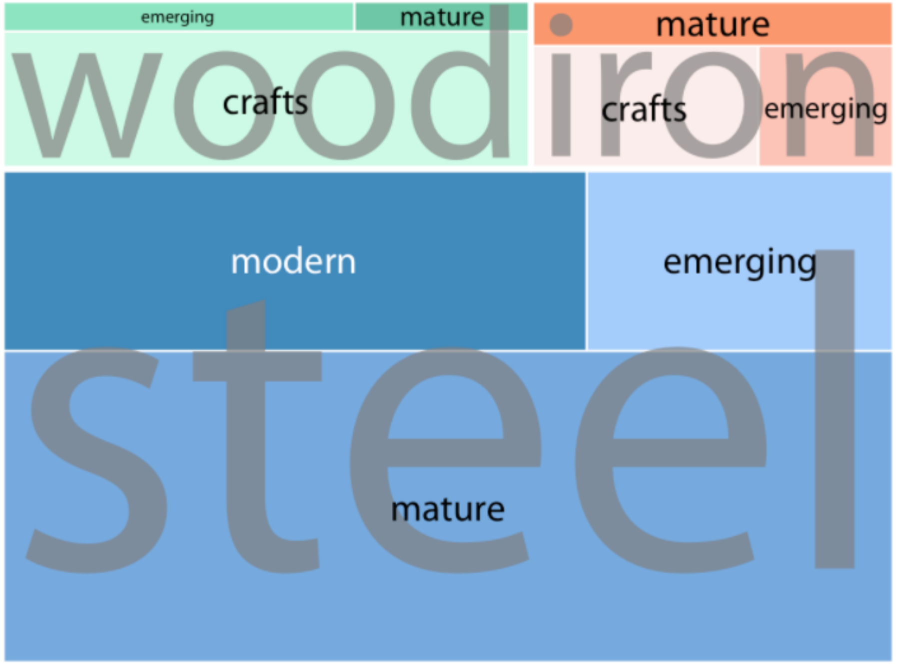


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

For example, in the case of the Pittsburgh bridges, we can **first subdivide the total area into three parts representing the three building materials** *wood, iron, and steel*.

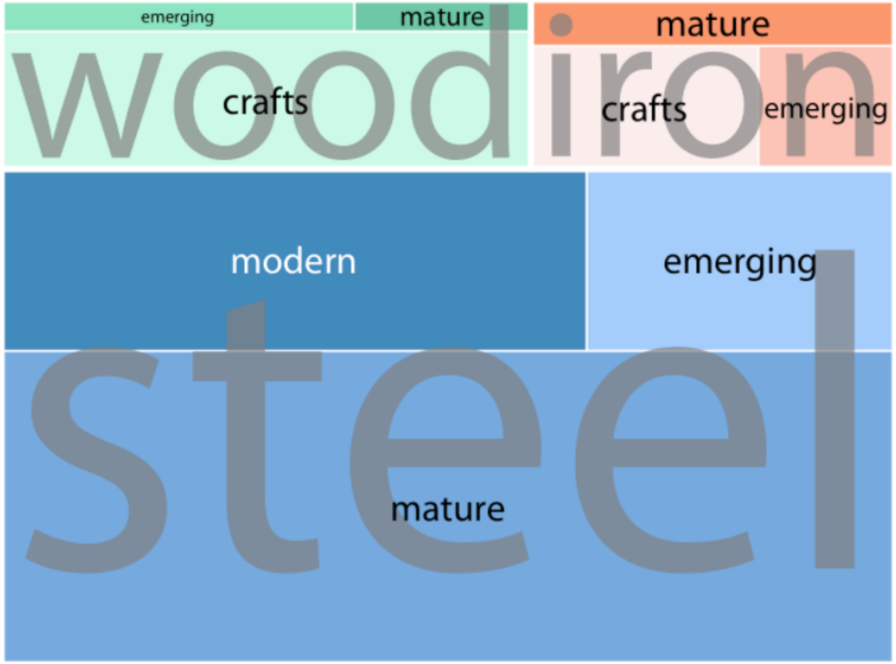
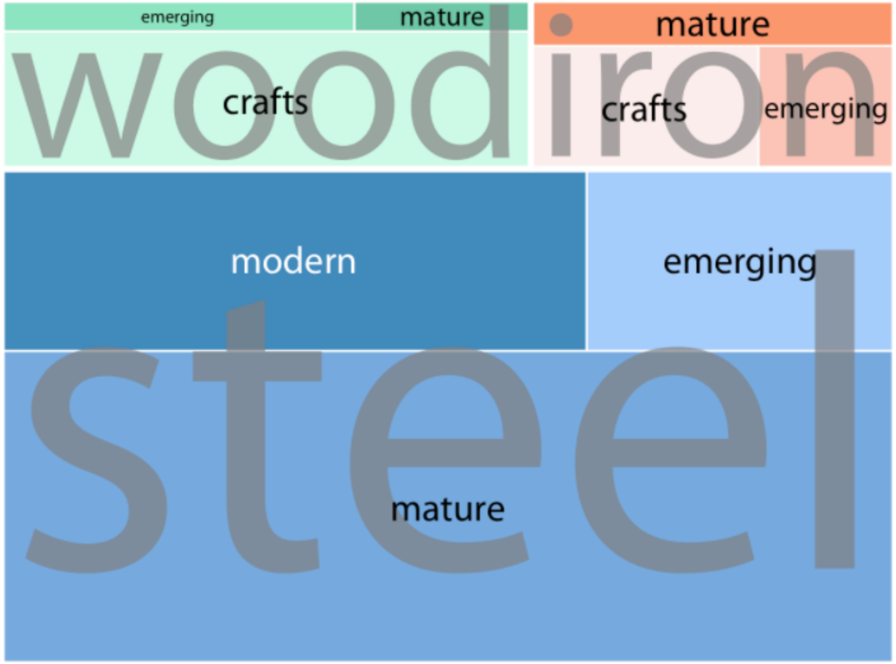


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

For example, in the case of the Pittsburgh bridges, we can first subdivide the total area into three parts representing the three building materials wood, iron, and steel.



Then, we subdivide each of those areas further to represent the construction eras represented for each building material.

Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- **Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable**, whereas treemaps do not make such an assumption. (width/height)

- **Treemaps** work well **even if the subdivisions of one group are entirely distinct from the subdivisions of another**. (area)

- **Parallel sets** work better than either mosaic plots or treemaps **when there are more than two grouping variables**. (next class)

7.1.3 Visualizing many proportions at once

Remember: **A treemap is a rectangular plot divided into tiles, each of which represents a single observation. It is a nice way of displaying hierarchical data** by using nested rectangles.

The relative area of each tile expresses a continuous variable.

- treemapify provides ggplot2 geoms for drawing treemaps
- Install the release version of treemapify from CRAN:

```
>install.packages("treemapify")
> library(treemapify)
```

- geom_treemap () – A 'ggplot2' geom to draw a treemap

Material of interest (extra):

https://cran.r-project.org/web/packages/treemapify/treemapify.pdf
https://cran.r-project.org/web/packages/treemapify/vignettes/introduction-to-treemapify.html

7.1.3 Visualizing many proportions at once

One simple example that you can test (first remember to install the package and load the library):

1. Let’s read the .csv dataframe:
   https://raw.githubusercontent.com/selva86/datasets/master/proglanguages.csv

```
> Proglangs <-
read.csv("https://raw.githubusercontent.com/selva86/datasets/master/proglanguages.csv")
```

This dataframe contains a hierarchical list of programming languages:

```
'data.frame':	40 obs. of  4 variables:
 $ id    : chr  "Java (general)" "PHP (general)" "dotNet (general)" "Python (general)" ...
 $ value : int  423 253 220 219 185 121 91 89 83 79 ...
 $ parent: chr  "Java" "PHP" "dotNet" "Python" ...
 $ rank  : num  40 39 38 37 36 35 34 33 32 31 ...
```

7.1.3 Visualizing many proportions at once

This dataframe contains a hierarchical list of programming languages:

```
'data.frame':	40 obs. of  4 variables:
 $ id    : chr  "Java (general)" "PHP (general)" "dotNet (general)" "Python (general)" ...
 $ value : int  423 253 220 219 185 121 91 89 83 79 ...
 $ parent: chr  "Java" "PHP" "dotNet" "Python" ...
 $ rank  : num  40 39 38 37 36 35 34 33 32 31 ...
```

2. In order to create a treemap, the data must be converted to desired format using treemapify(). The important requirement is, you need to identify in your data:
   - One numerical continuous variable each describes the area of the tiles ('value')
   - One variable for fill color ('parent')
   - One variable that has the tile's label ('id')
   - The parent group ('parent')

7.1.3 Visualizing many proportions at once

2. In order to create a treemap, the data must be converted to desired format using treemapify(). The important requirement is, you need to identify in your data:
   - One variable each describes the area of the tiles ('value')
   - One variable for fill color ('parent')
   - One variable that has the tile's label ('id')
   - The parent group ('parent')

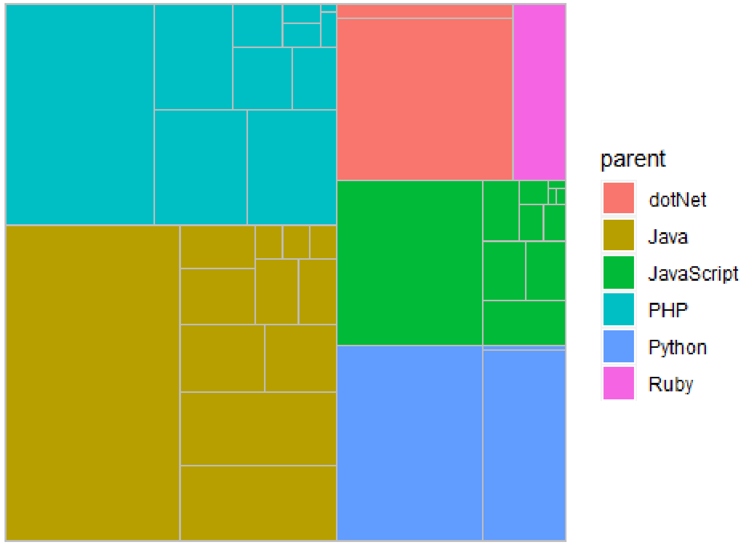
   And map them by using aes:

```
> ggplot(Proglangs, aes(area=value, fill=parent,
label=id, subgroup=parent)) + geom_treemap()
```

3. Use geom_treemap

7.1.3 Visualizing many proportions at once

```
> ggplot(Proglangs, aes(area=value, fill=parent,
label=id, subgroup=parent)) + geom_treemap()
```

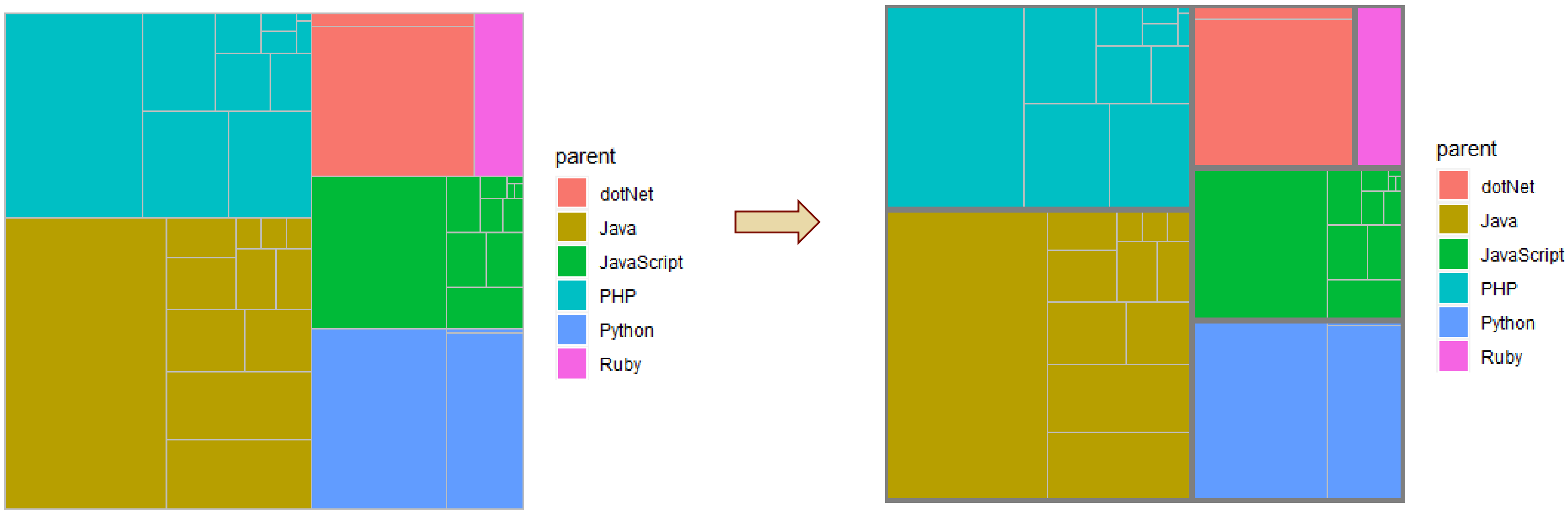


Remember that you can assign this to a variable and add the other optional layers afterwards

4. You can now add the main group bordering

```
> ggplot(Proglangs, aes(area=value, fill=parent,
label=id, subgroup=parent)) + geom_treemap()+
geom_treemap_subgroup_border()
```

7.1.3 Visualizing many proportions at once



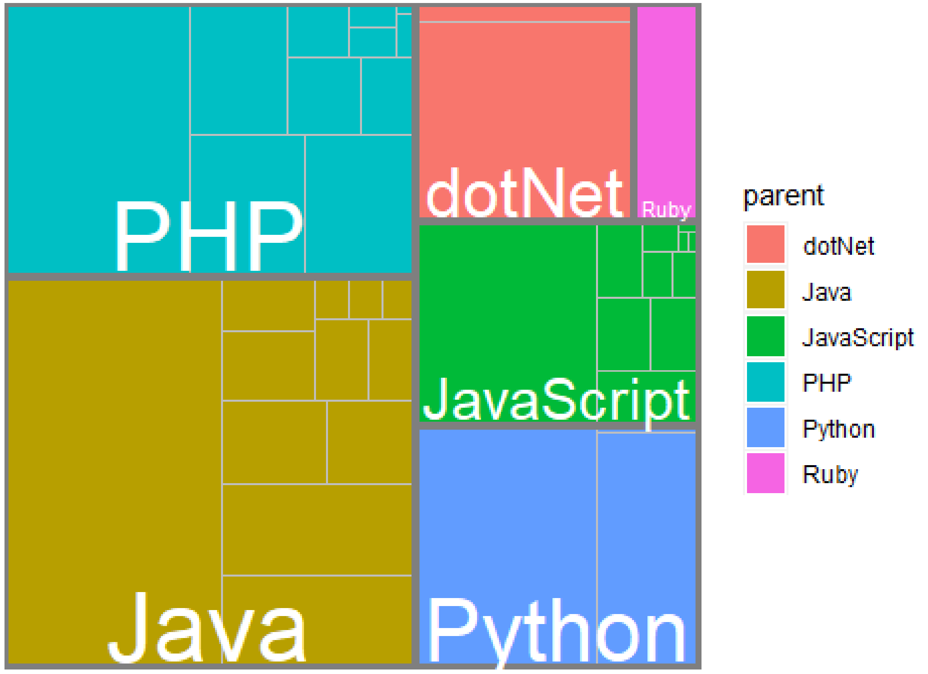
Thanks to:

`geom_treemap_subgroup_border()`

7.1.3 Visualizing many proportions at once

5. We can add subgroup heading in white

```
> ggplot(Proglangs, aes(area=value, fill=parent,
label=id, subgroup=parent)) + geom_treemap()+
geom_treemap_subgroup_border()
+geom_treemap_subgroup_text(color='white')
```

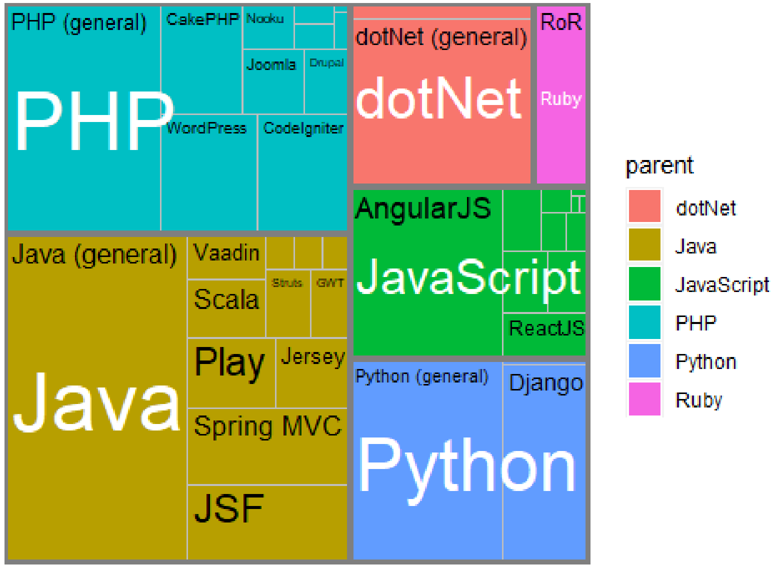


7.1.3 Visualizing many proportions at once

6. Add all other text in black

```
> ggplot(Proglangs, aes(area=value,
fill=parent,label=id,subgroup=parent))+geom_treemap()+
geom_treemap_subgroup_border()+
+geom_treemap_subgroup_text(color="white",place="left")
+geom_treemap_text (aes(label=id))
```

More examples:
https://cran.r-project.org/web/packages/treemapify/vignettes/introduction-to-treemapify.html



7.1.4 Visualizing many relations at once

Many datasets contain **two or more quantitative variables,** and we may be **interested in how these variables relate to each other.**

Scatterplot matrix (SPLOM) uses **multiple scatterplots to determine the correlation** (if any) between a series of variables.

These scatterplots are then ***organized into a matrix*****, making it easy to look at all the potential correlations in one place.**

7.1.4 Correlation between series variables

Bubble plot

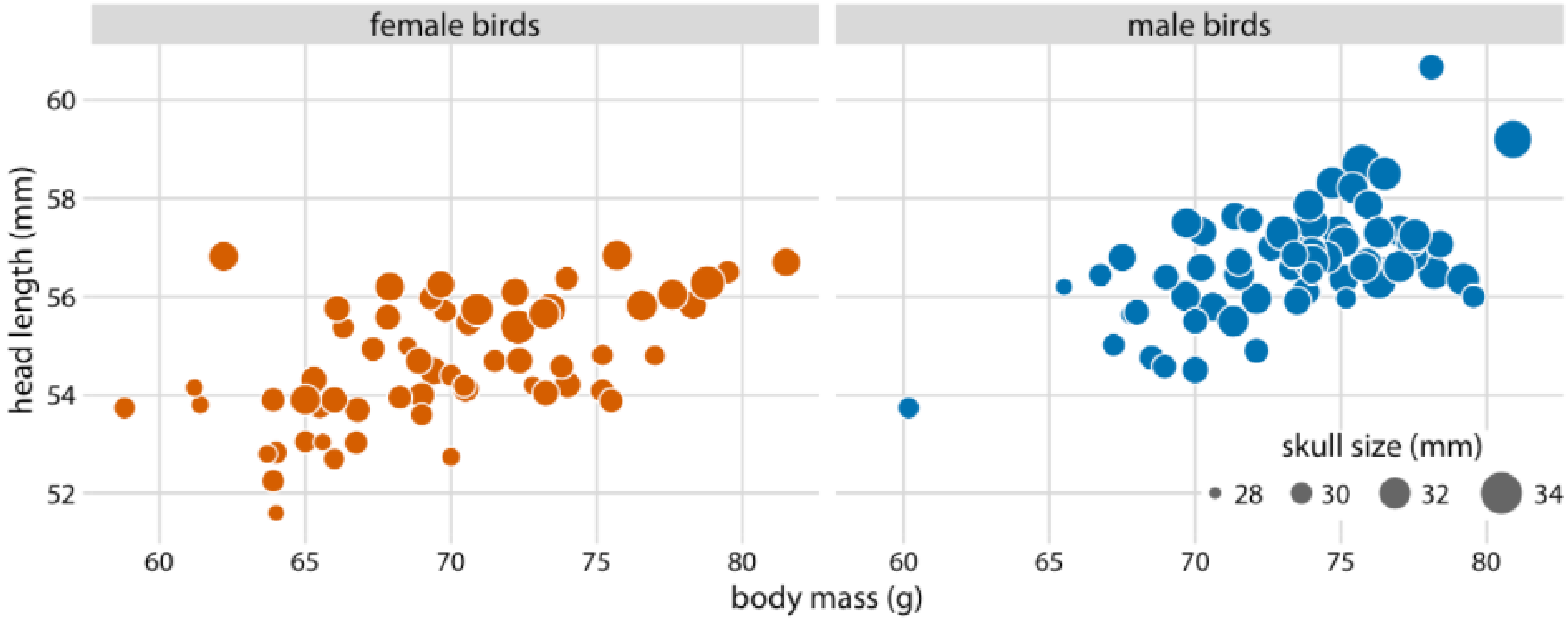


Figure 12.3: Head length versus body mass for 123 blue jays. The birds' sex is indicated by color, and the birds' skull size by symbol size. Head-length measurements include the length of the bill while skull-size measurements do not. Head length and skull size tend to be correlated, but there are some birds with unusually long or short bills given their skull size. Data source: Keith Tarvin, Oberlin College

Claus Wilke

7.1.4 Correlation between series variables

Scatterplot matrix (SPLOM) can provide answers to the following questions:

- Are there **pair wise relationships** between the variables?
- If there are relationships, what is **the nature of these relationships**?
- Are there **outliers** in the data?
- Is there **clustering by groups** in the data?

7.1.4 Scatterplot Matrix

[More SPLOM examples]

Thanks for your attention!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universitat Autònoma de Barcelona

8. Color y Precisión Visual

Sol Bucalo
sol.bucalo@uab.cat

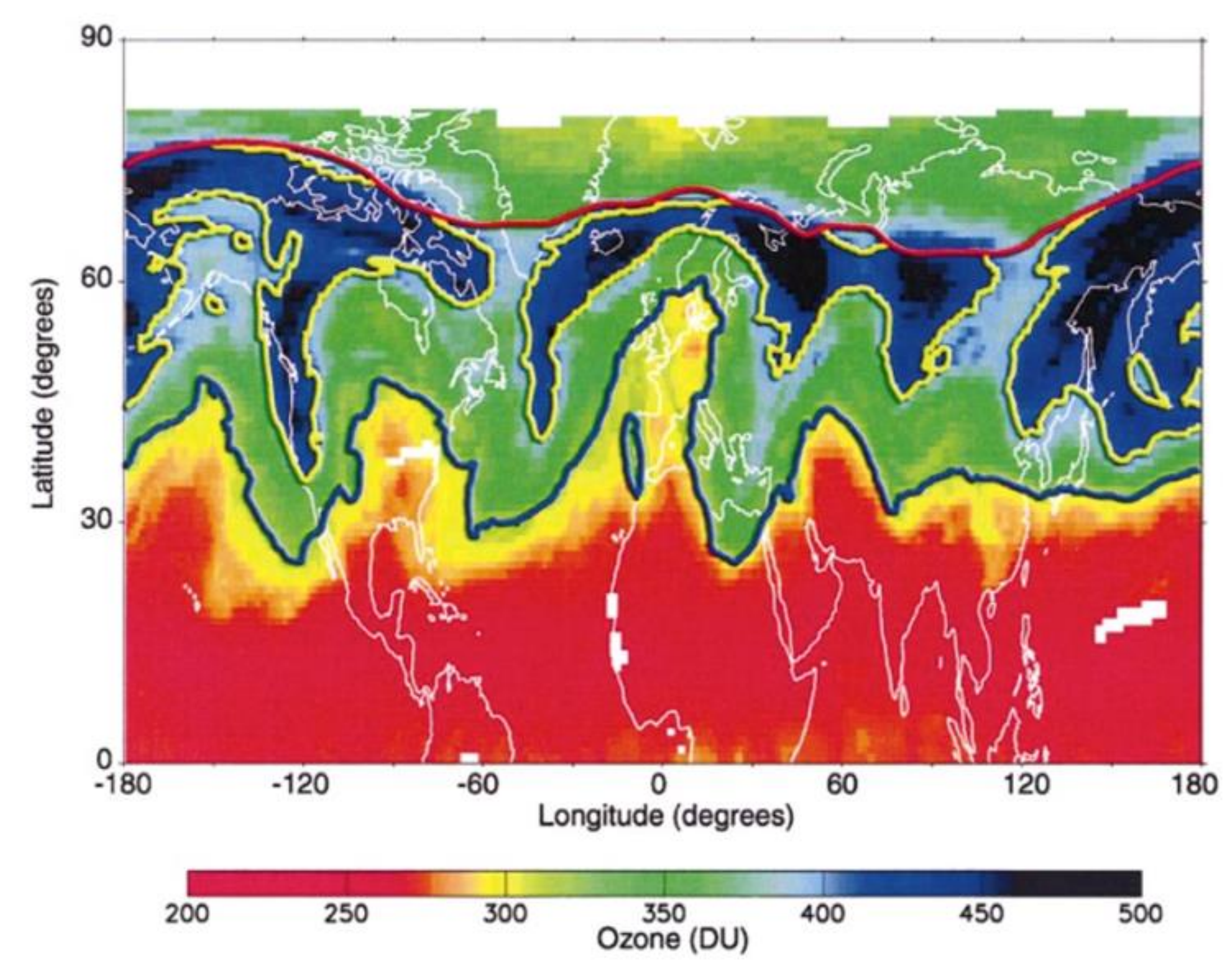
Guillermo Marin
guillermo.marin@uab.cat

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universitat Autònoma de Barcelona

8.1 Color

BEWARE OF THE RAINBOW

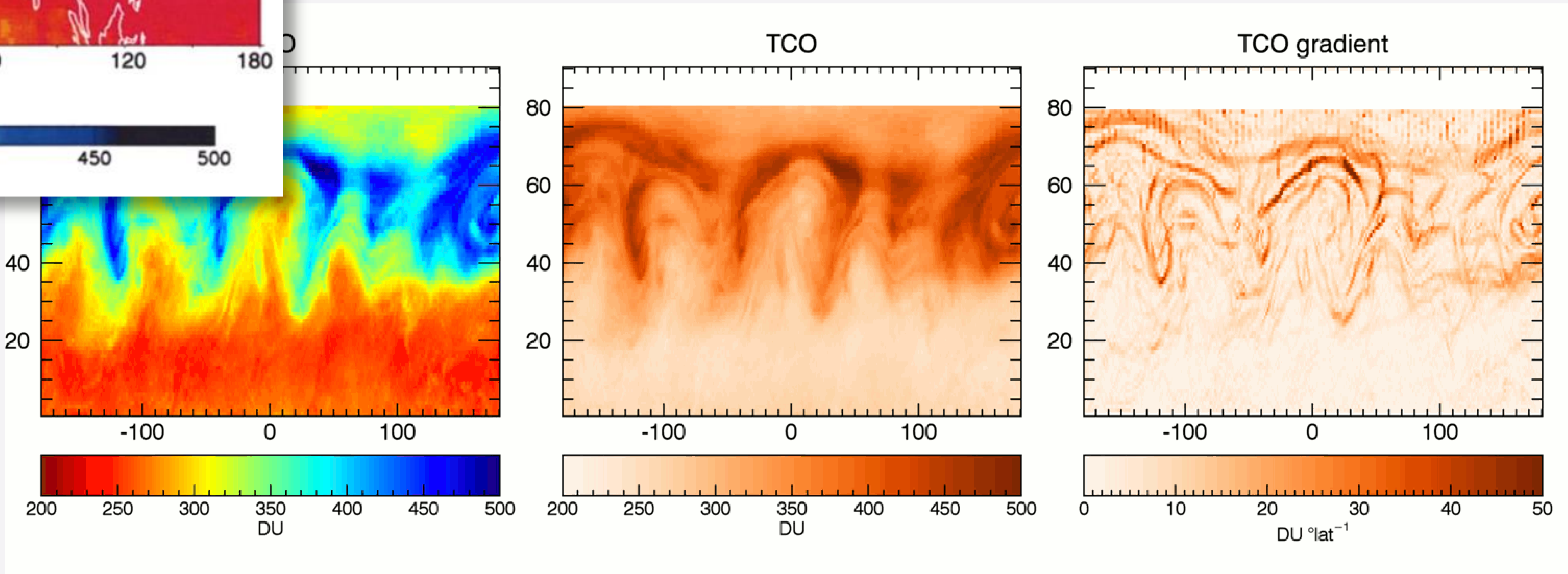


Hudson, R. D., Andrade, M. F., Follette, M. B., and Frolov, A. D.:

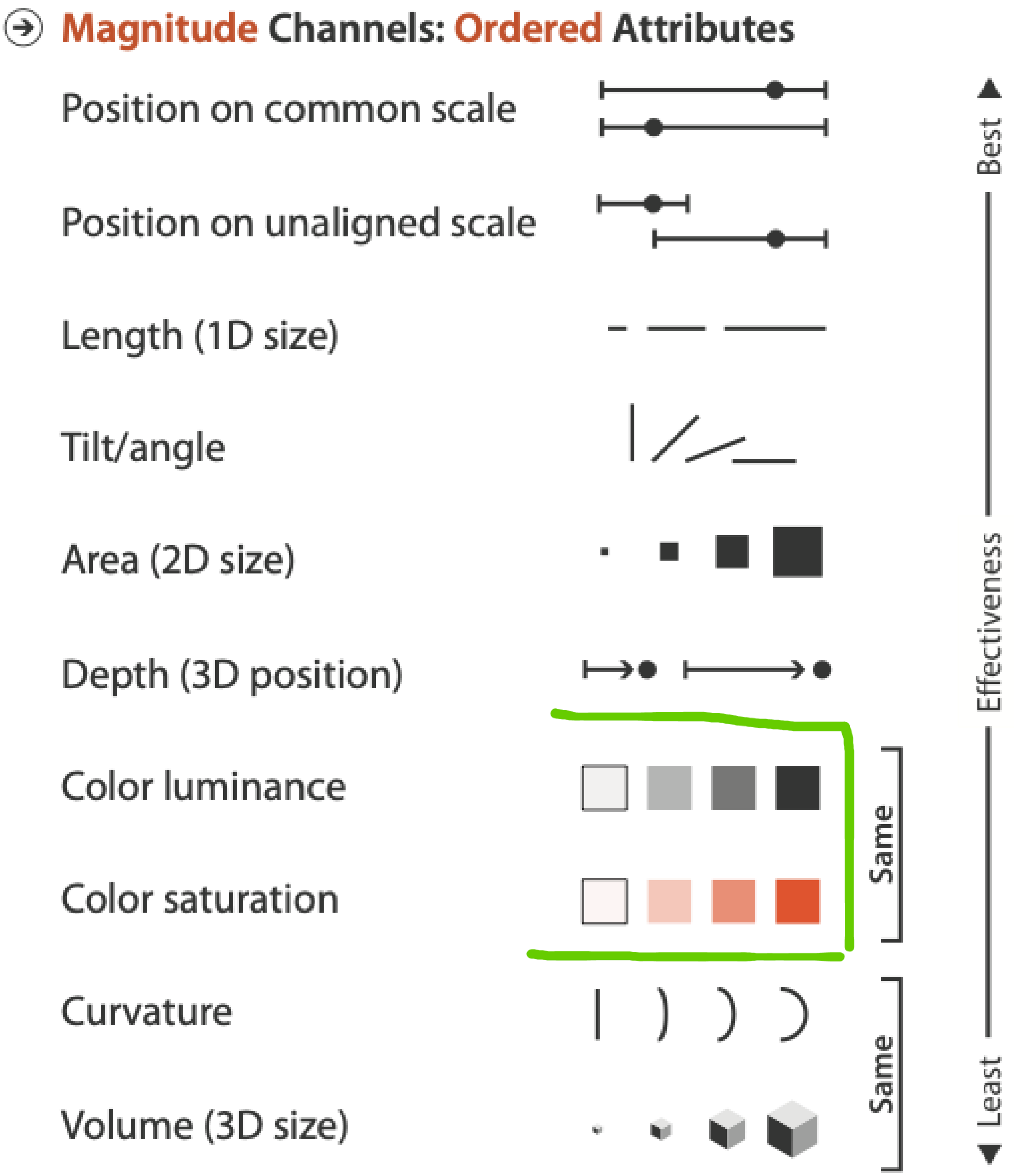
The total ozone field separated into meteorological regimes – Part II: Northern Hemisphere mid-latitude total ozone trends,

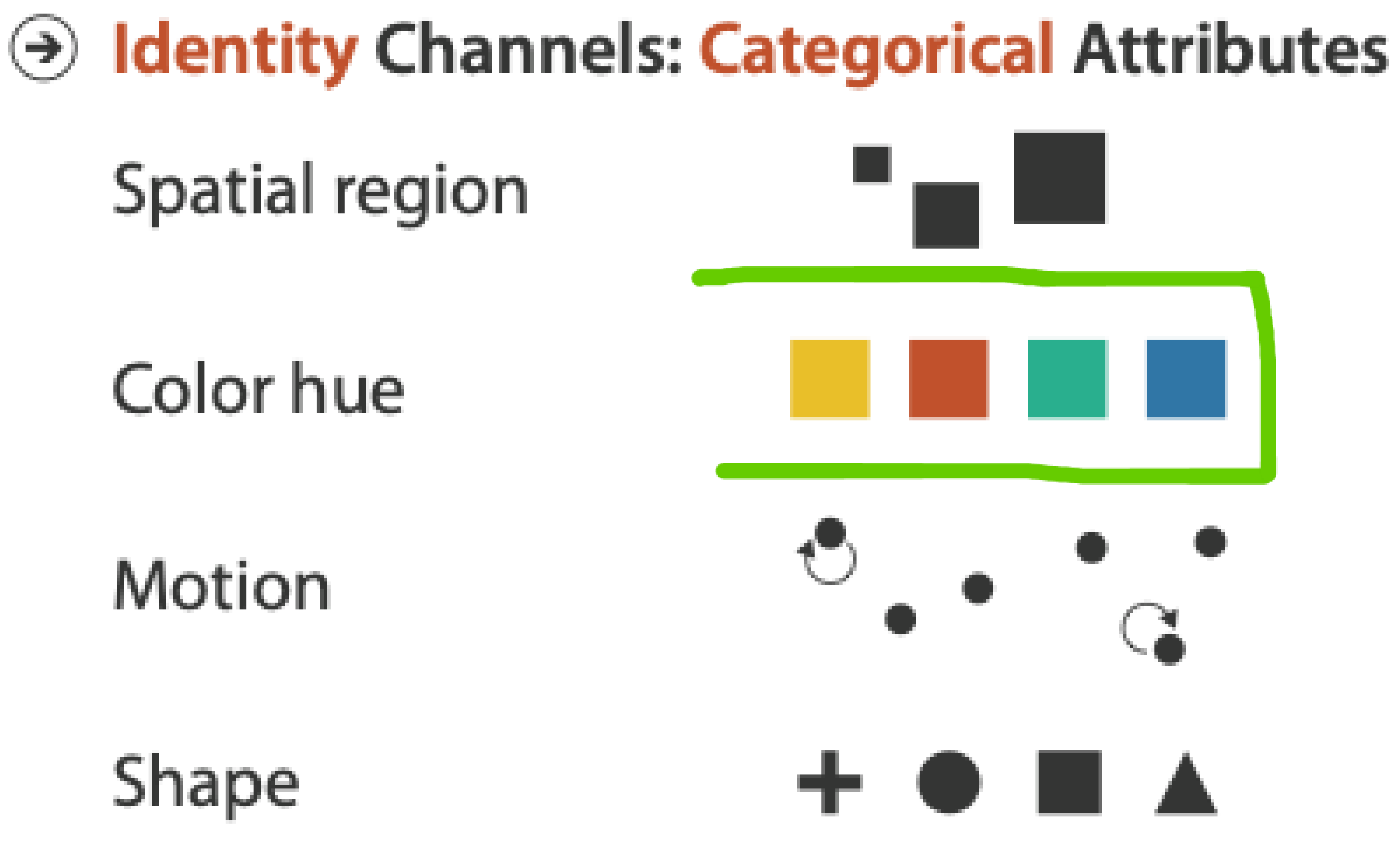
Atmos. Chem. Phys., 6, 5183-5191, https://doi.org/10.5194/acp-6-5183-2006, 2006.

The wrong color palette can show effects in the visualization not present in the data



Sean Davis, NOAA





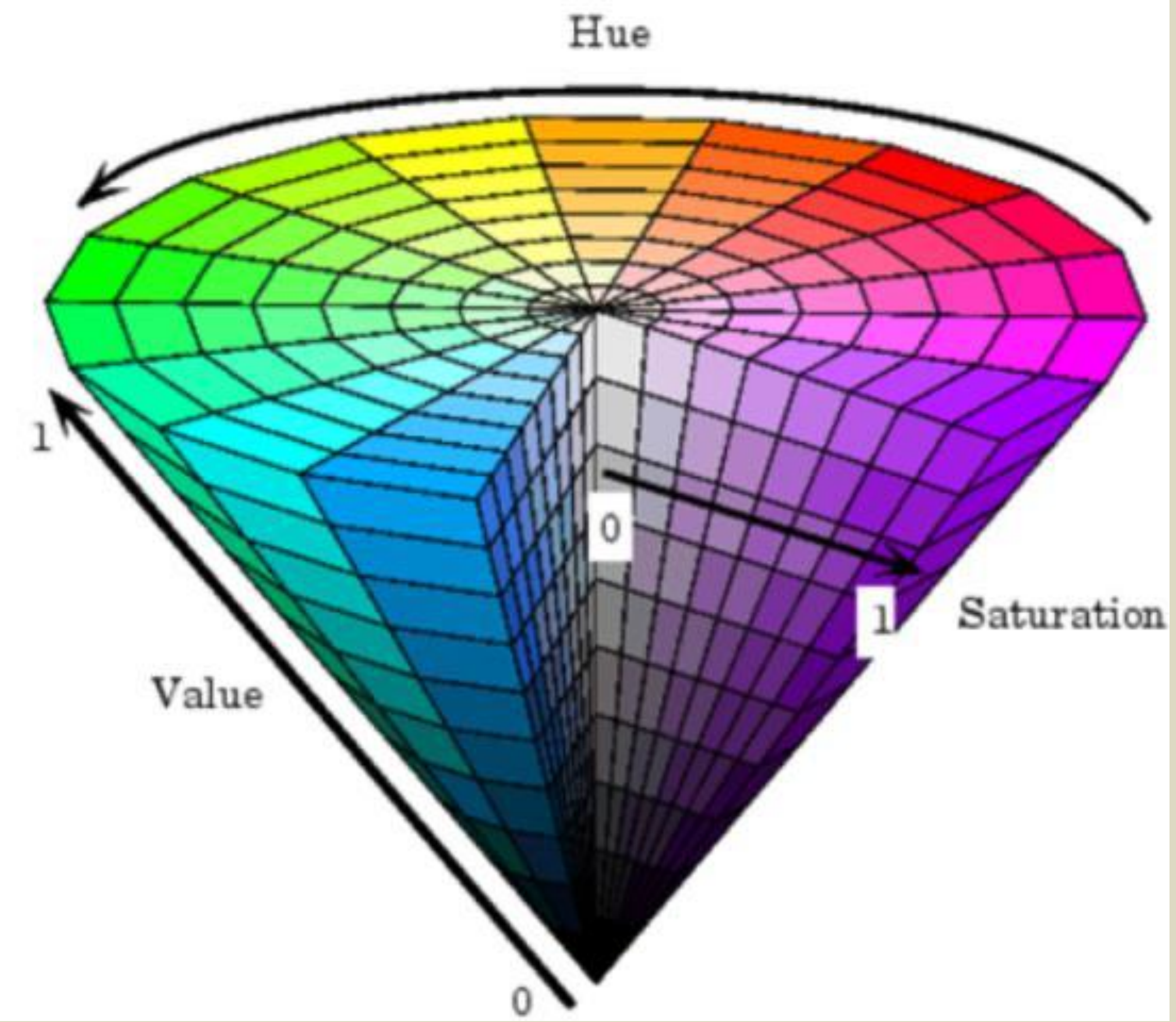
- El color es uno de los canales visuales más potentes para codificar datos
- Dos de sus propiedades son canales de magnitud y una de identidad
- Se pueden usar en combinación con otros: tienen poca o ninguna interferencia con los canales de posición

R cheatsheet

R Color Palettes
This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure….NOW!

A few notes on HSV/HLC
HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model….blah blah blah…

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.



R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors
Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggbb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0
i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

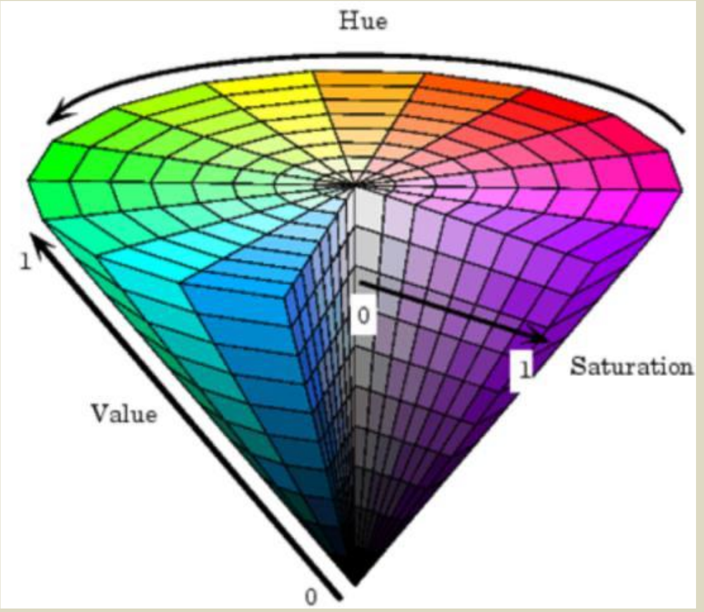
R has 657 built in color names
To see a list of names:
colors()
These colors are displayed on P. 3.

Example:
peachpuff4

R translates various color models to hex, e.g.:
- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.
  rgb(r, g, b, maxColorValue=255, alpha=255)
- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument
  hsv(h, s, v, alpha)
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
  hcl(h, c, l, alpha)

A few notes on HSV/HLC
HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model….blah blah blah…

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.



R can translate colors to rgb (this is handy for matching colors in other programs)
col2rgb(c("#FF0000", "blue"))

R Color Palettes
This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure….NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps
grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

grDevices palettes
cm.colors
topo.colors
terrain.colors
heat.colors
rainbow
see P. 4 for options

heat.colors(4, alpha=1)
> #FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFFF80FF"

For the rainbow palette you can also select start/end color (red = 0, yellow = *1/6*, green = *2/6*, cyan = *3/6*, blue = *4/6* and magenta = *5/6*) and saturation (s) and value (v):
rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)

Package: RcolorBrewer
This function has an argument for the number of colors and the color palette (see P. 4 for options).
brewer.pal(4, "Set3")
> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: display.brewer.all(5)
There is also a very nice interactive viewer:
http://colorbrewer2.org/

My Recommendation

Package: colorspace
These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

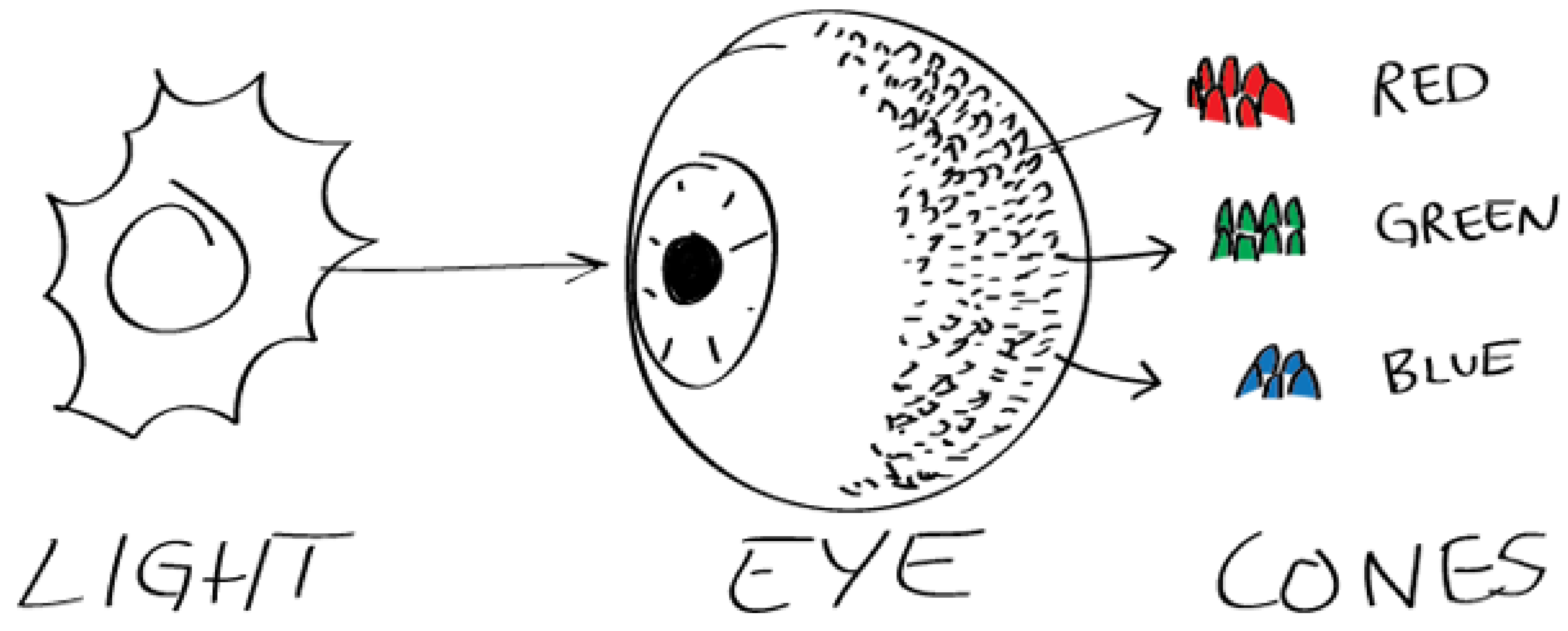
colorspace default palettes
diverge_hcl
diverge_hsl
terrain_hcl
sequential_hcl
rainbow_hcl

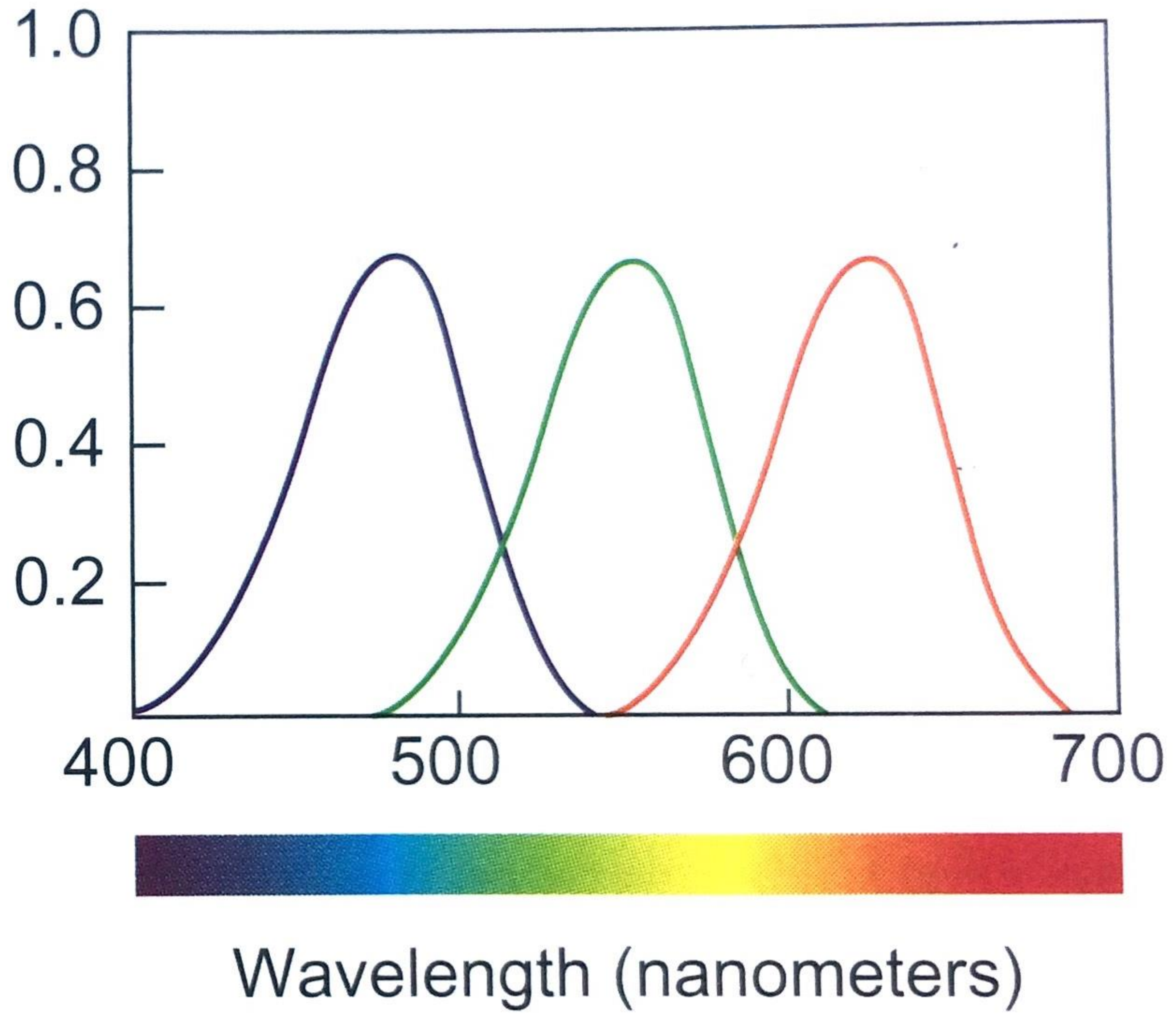
rainbow_hcl(4)
"#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"

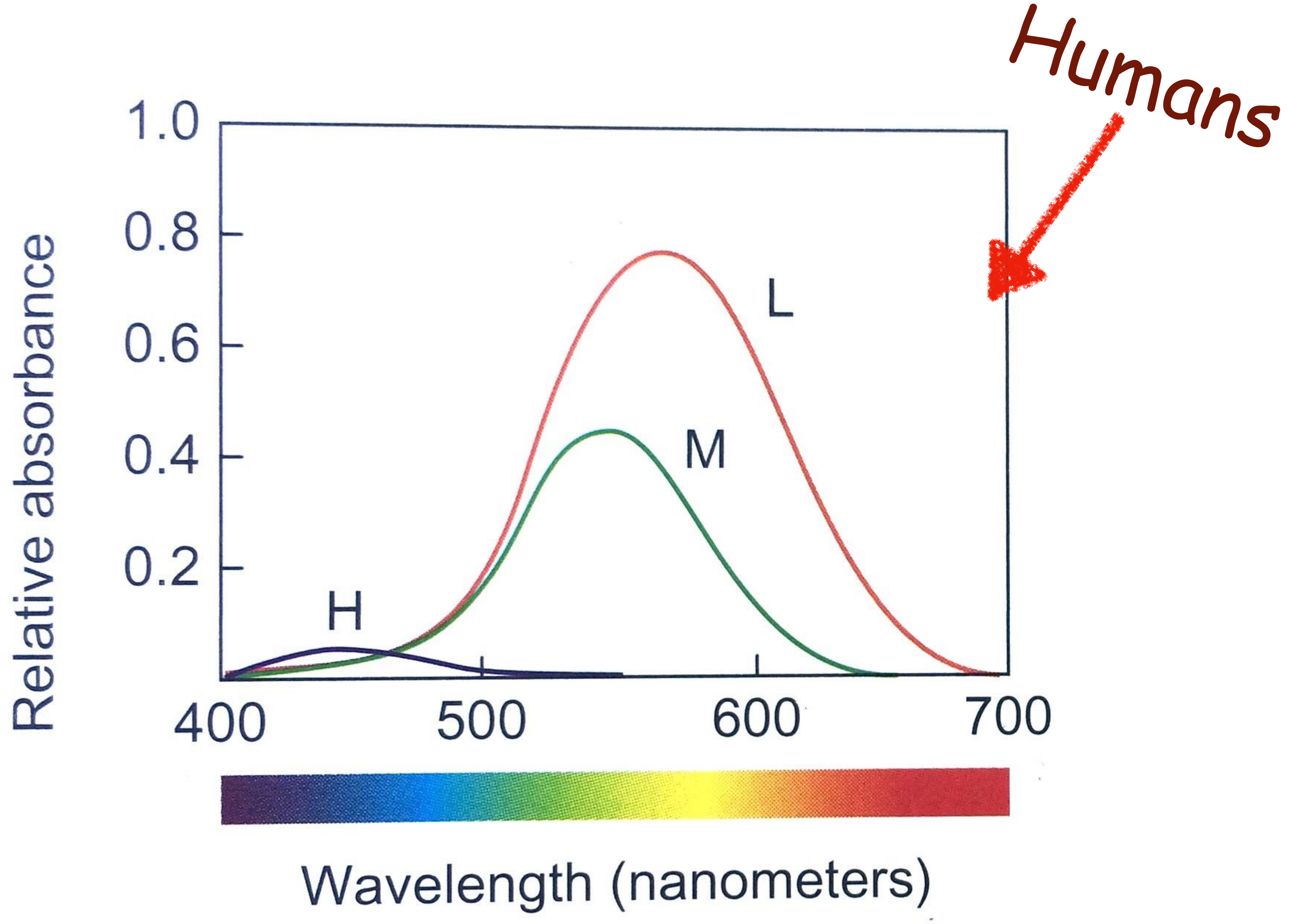
However, all palettes are fully customizable:
diverge_hcl(7, h = c(246, 40), c = 96, l = c(65, 90))
Choosing the values *would* be daunting. But there are some recommended palettes in the colorspace documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
pal <- choose_palette()

Percepción

- Tenemos tres tipos de conos que reaccionan a espectros distintos y de forma desigual
- No son como receptores de una cámara, frecuencias más largas generan estímulos más fuertes





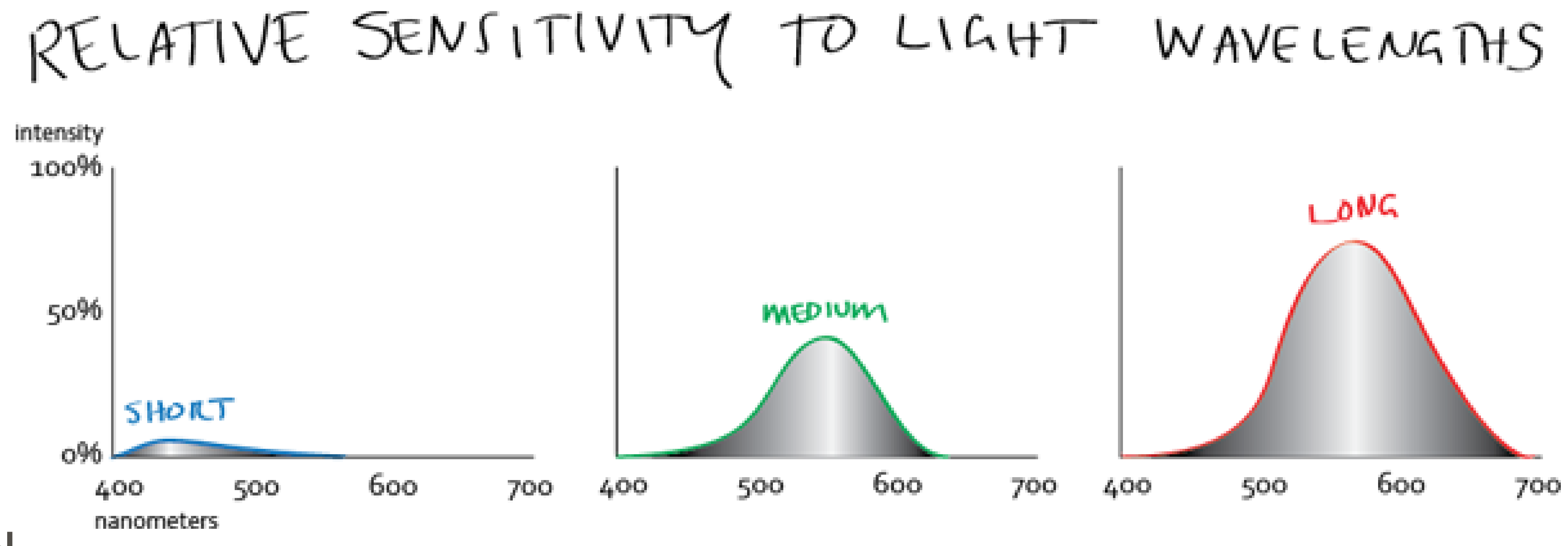


Ware, 2008

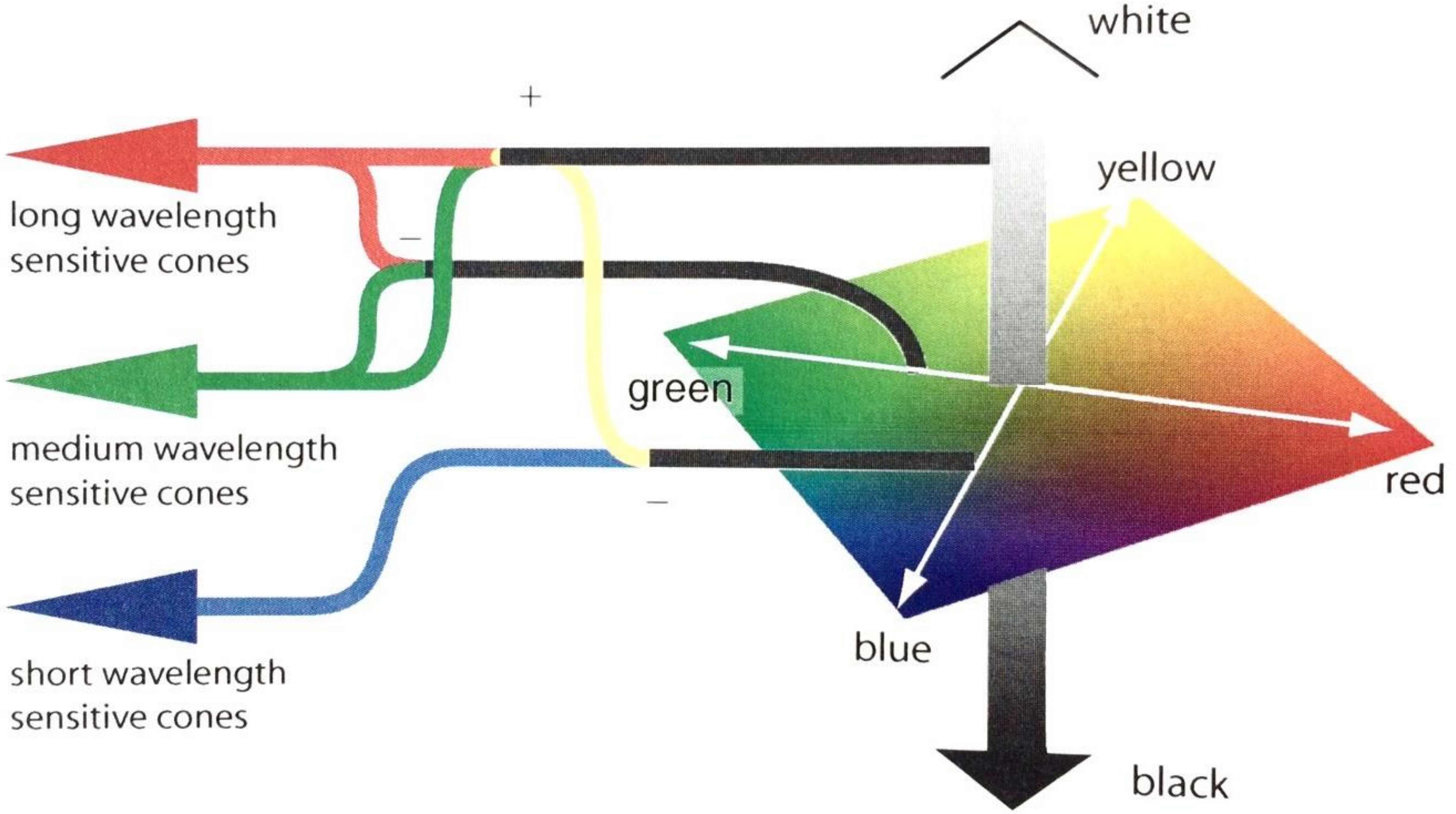
Opponent Process
Theory of Color

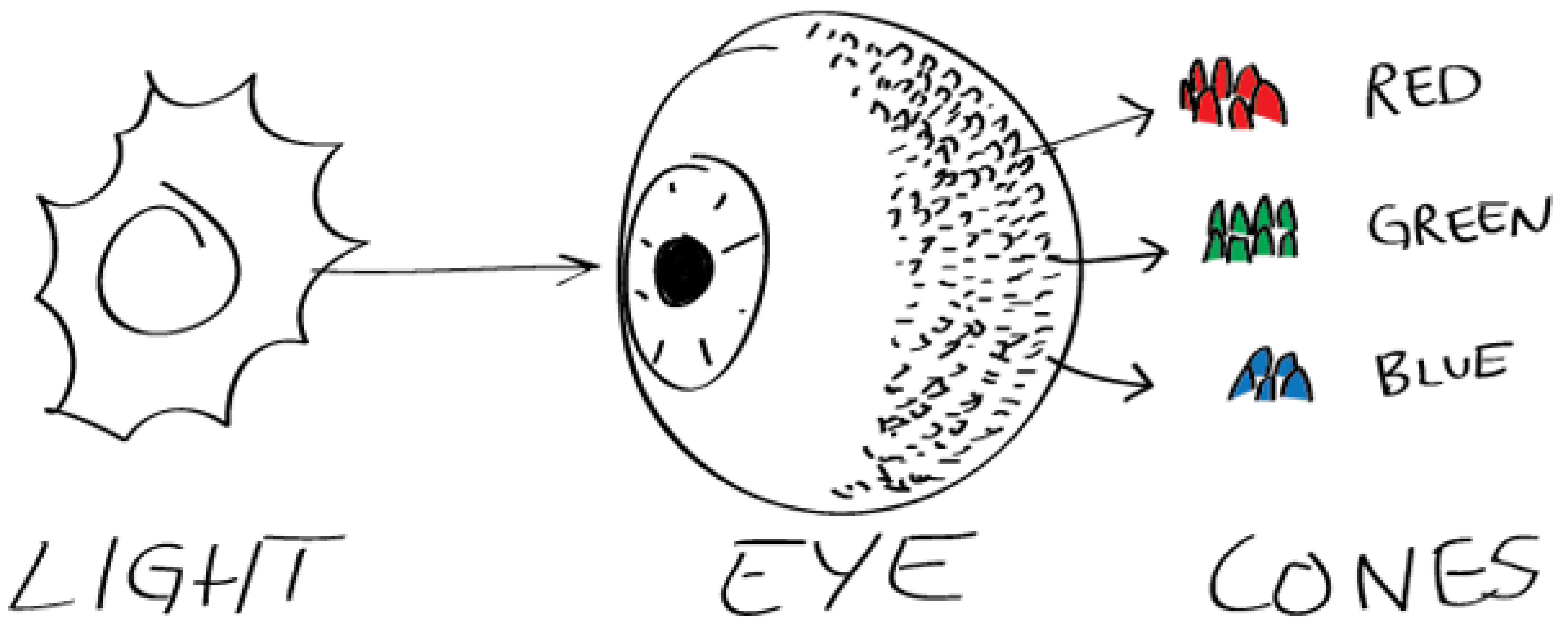
Ewald Hering, 1920

- The brain combines the signals by subtraction: receptors subtract med. and long freq. making a red-green-difference signal channel.
- Other neurones subtract long and short yielding yellow-blue-diff signal channel;
- Third group of neurones ADD long and medium to create luminance (B&W) channel



difference signal channels





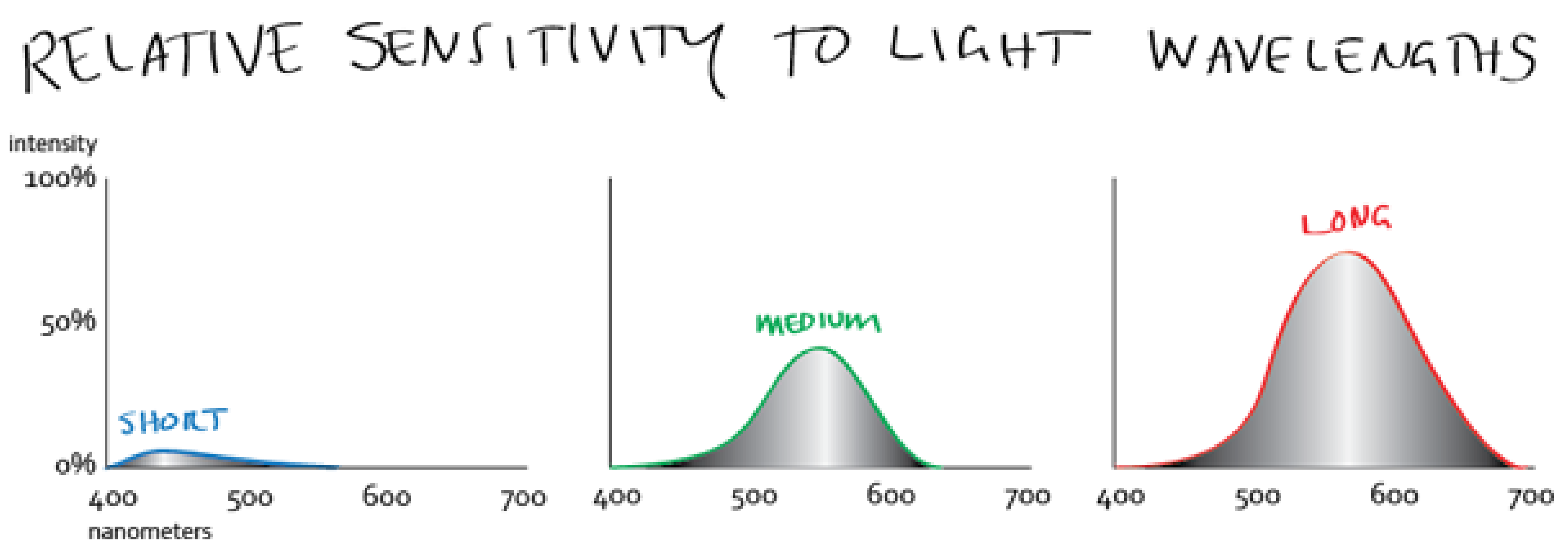
Ware, 2008

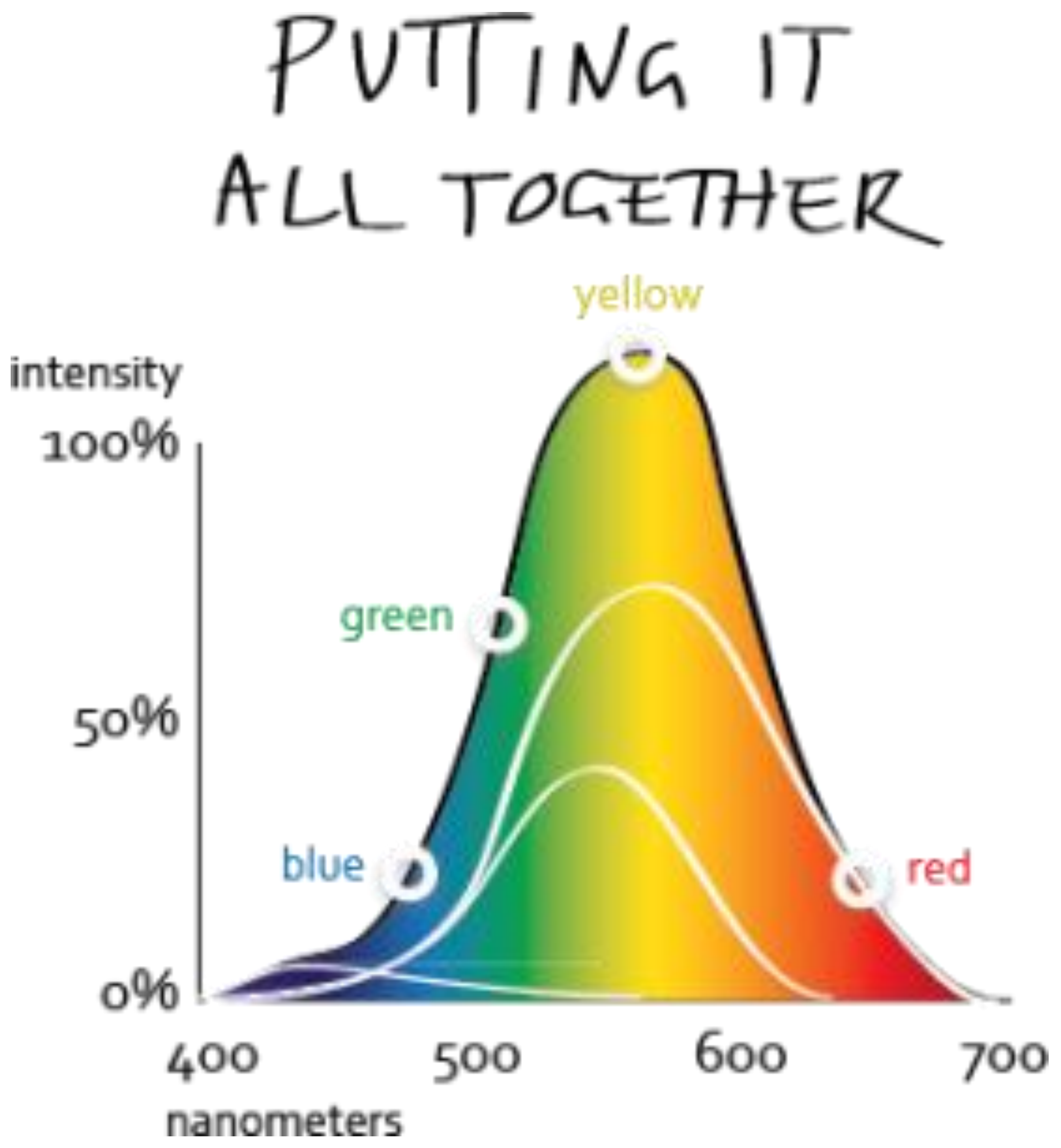
Opponent Process

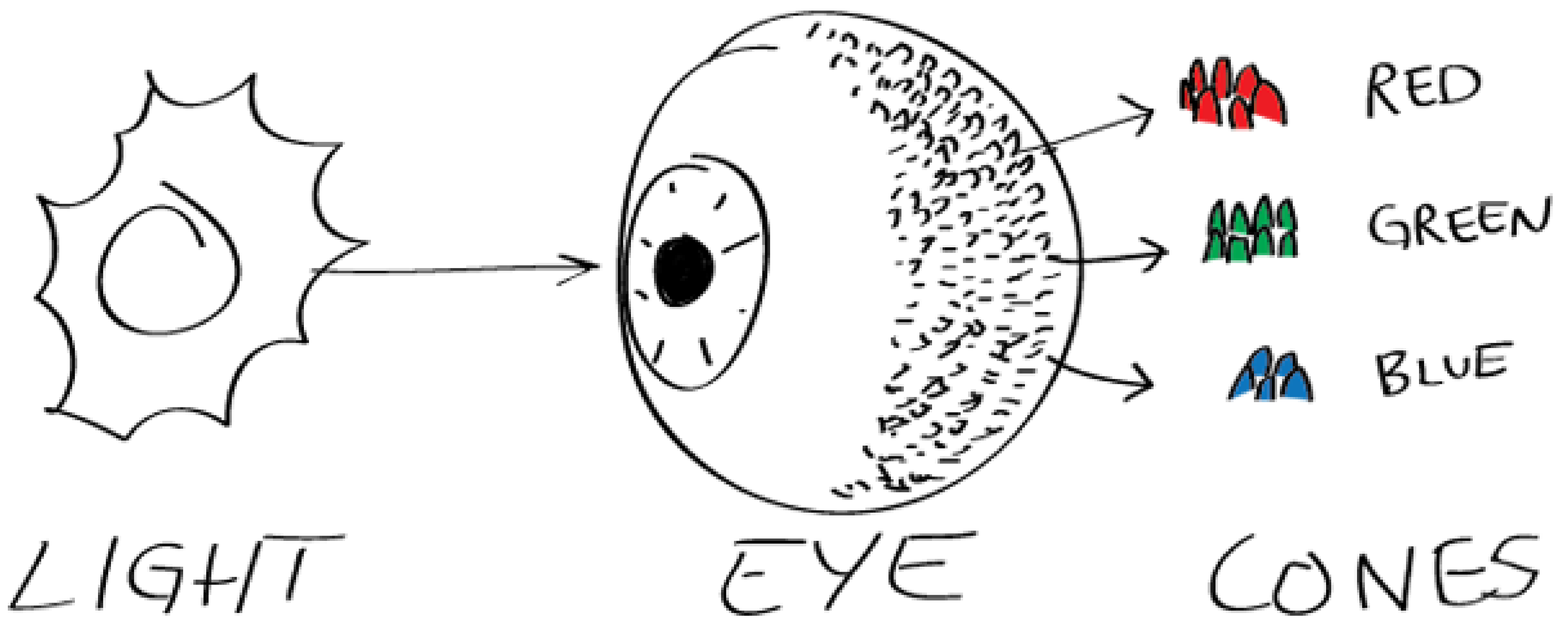
Theory of Color

Ewald Hering, 1920

- Cuando vemos Amarillo hay una sobre excitación de muchos receptores que hacen que lo percibamos más intensamente que otros colores





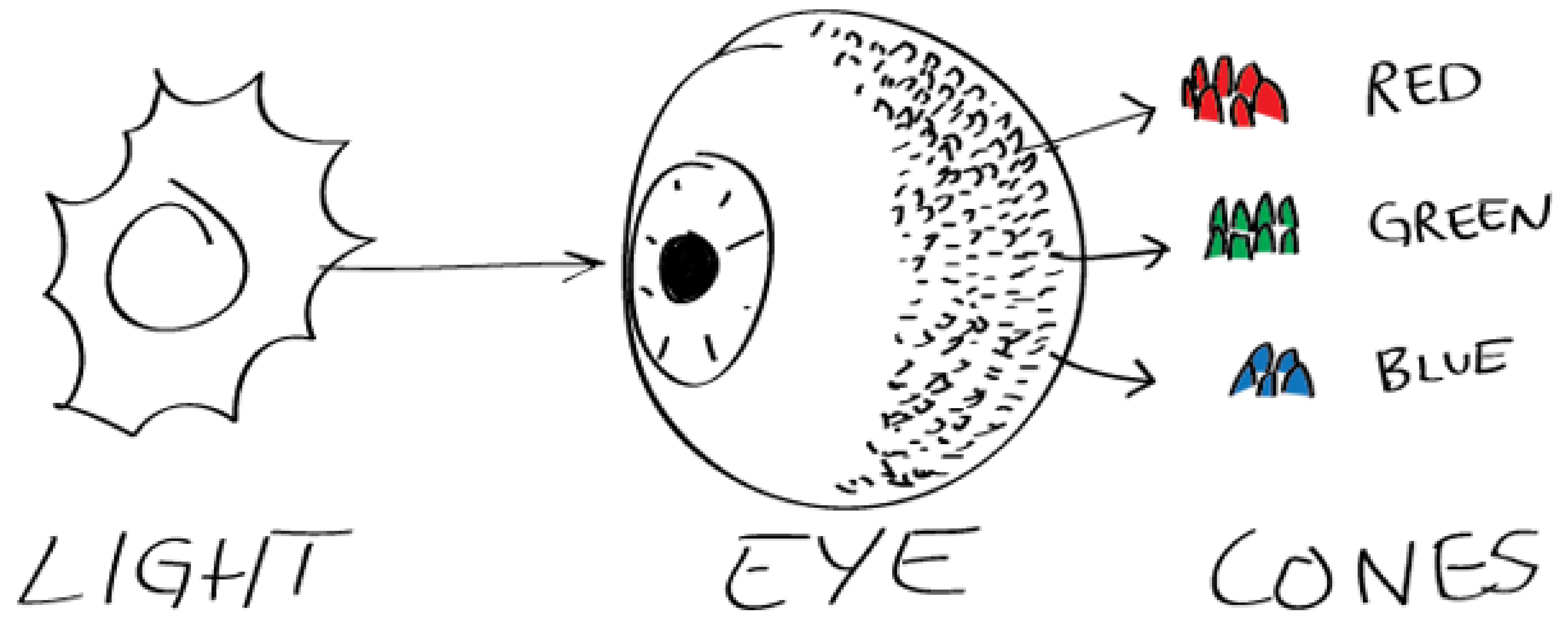


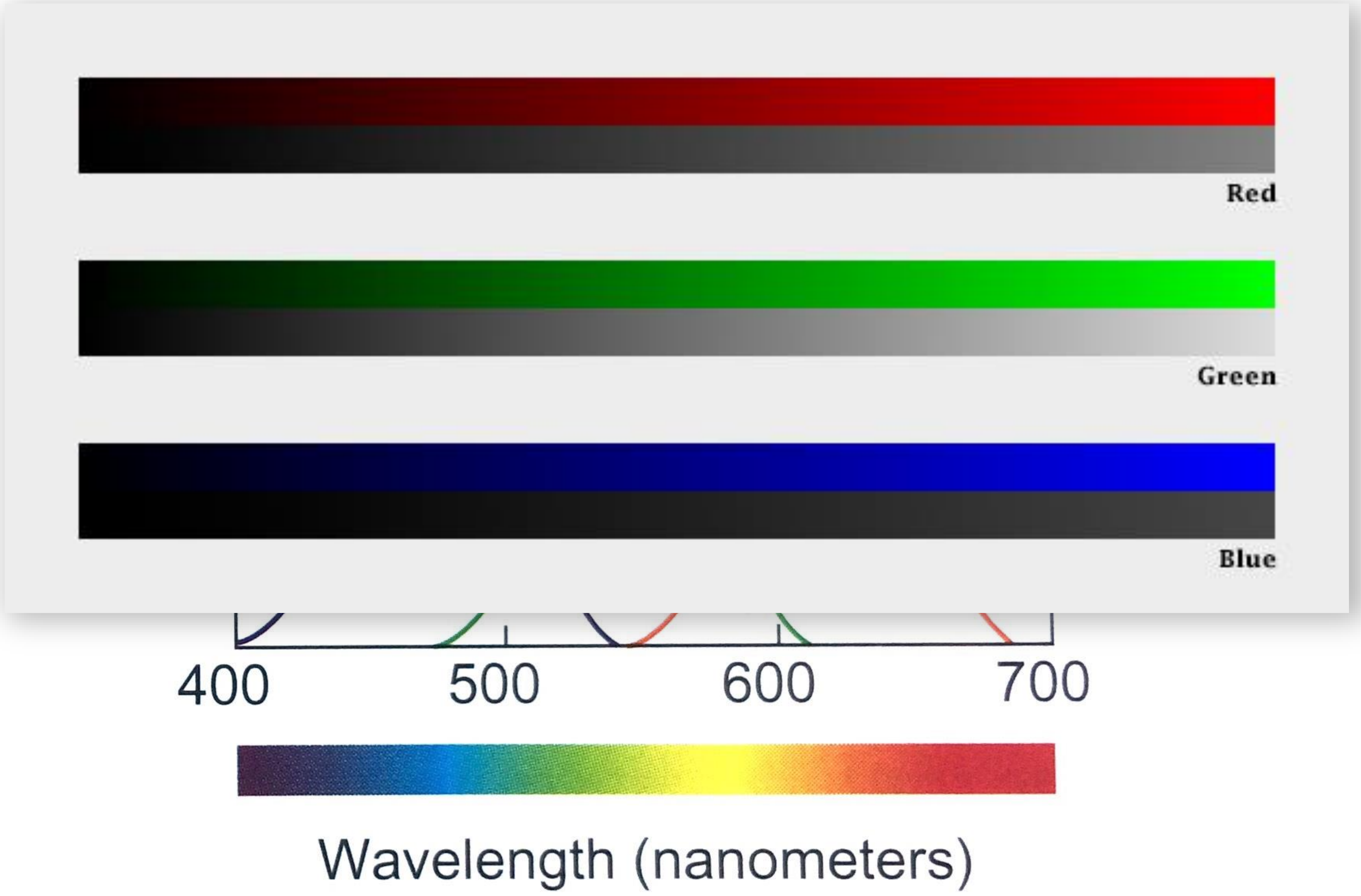
Percepción

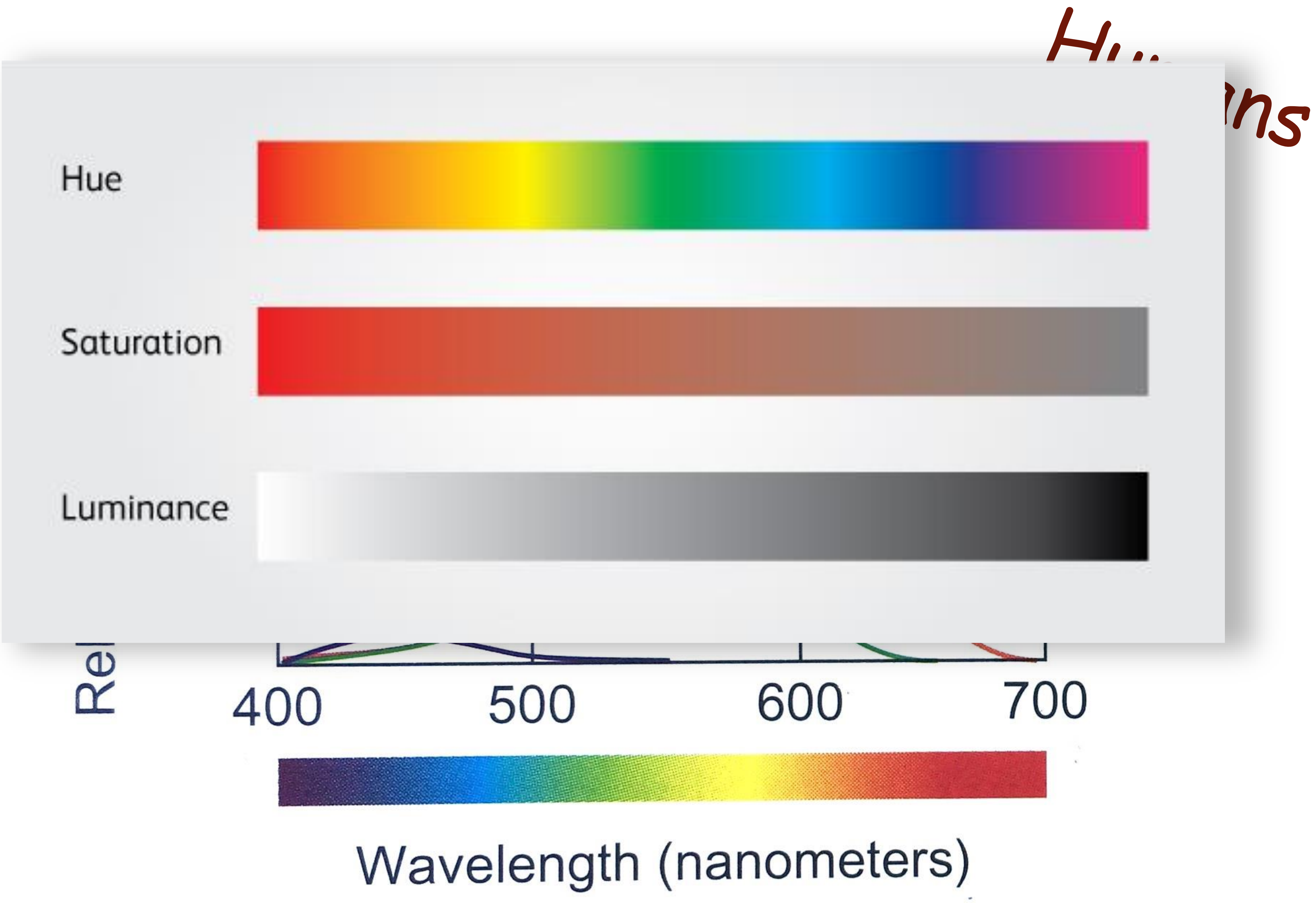
Computers process a combination of very narrow frequency bands of red, green, and blue

Computers calculate light linearly

Our eyes detect red, green, and blue light but we think about colour in terms of lightness, hue, and saturation



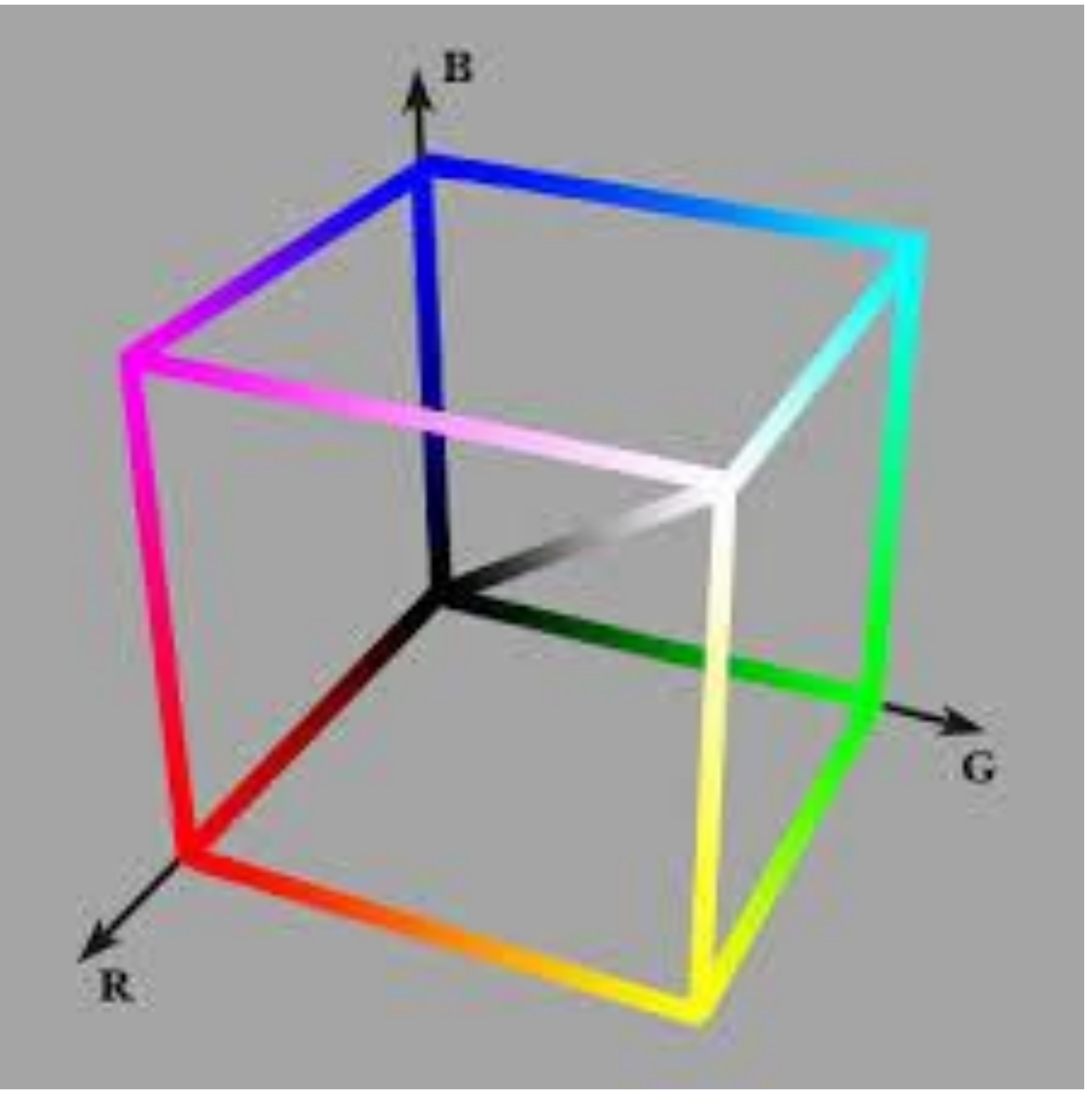


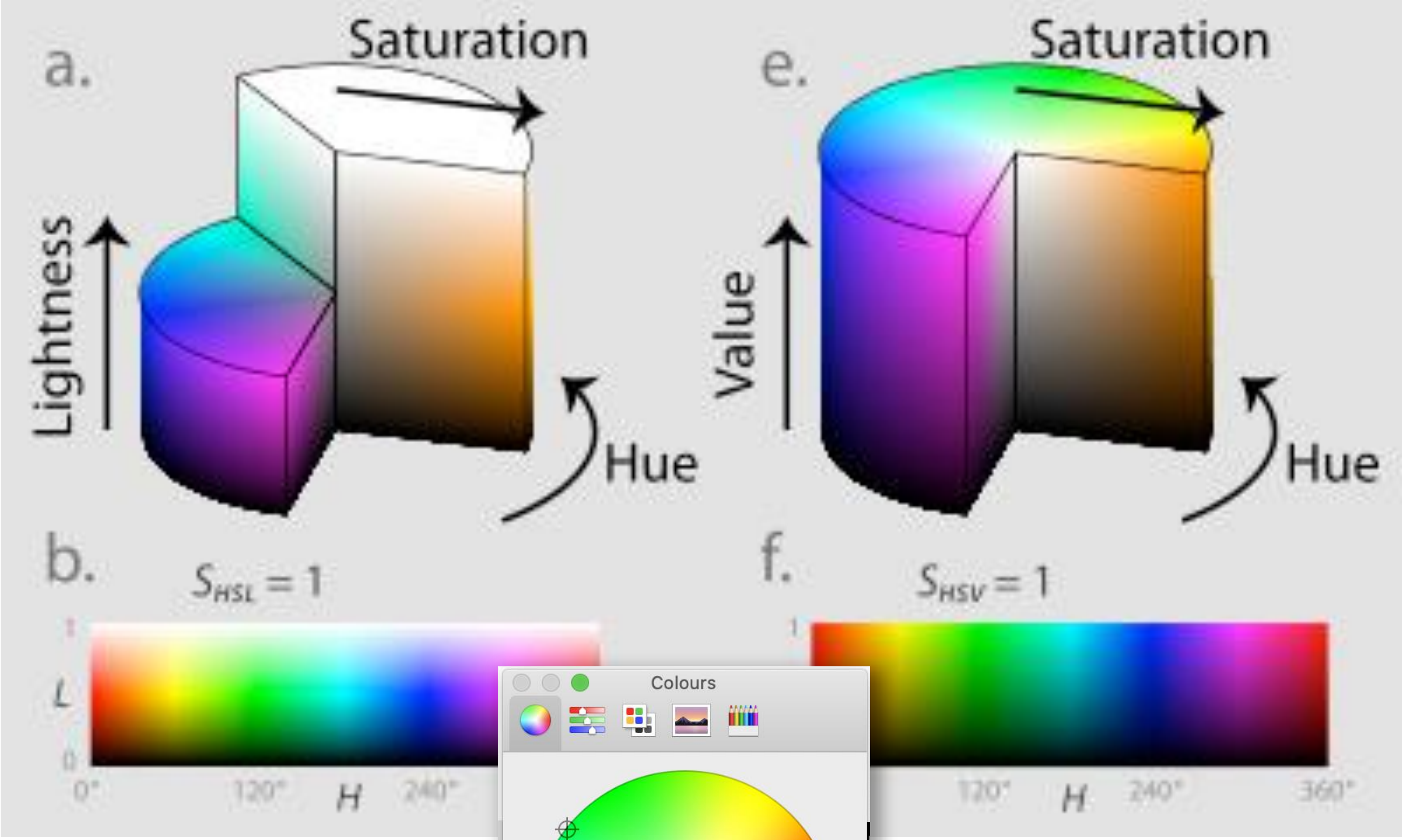


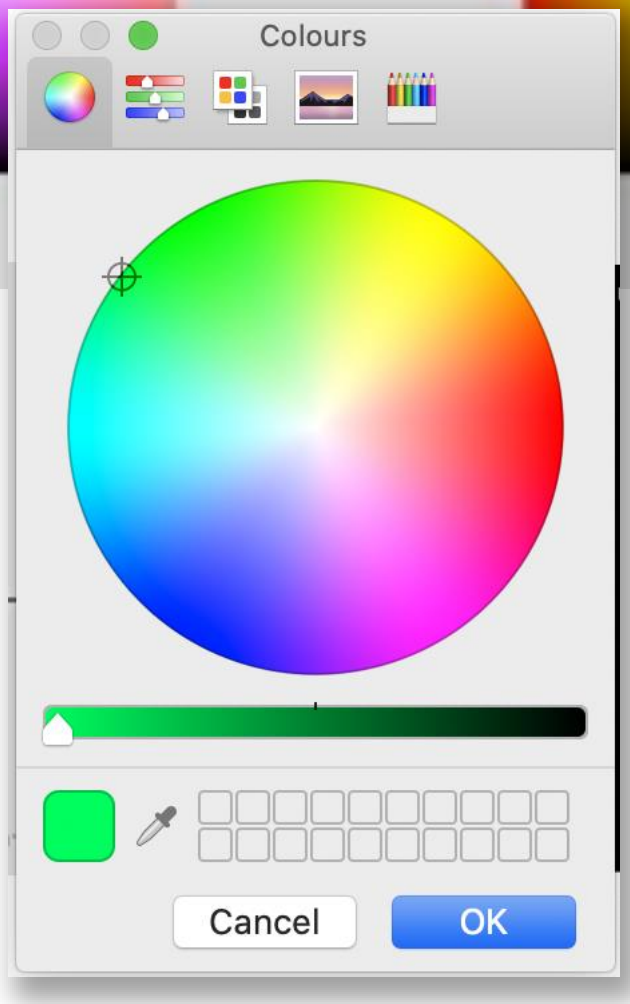
Ware, 2008

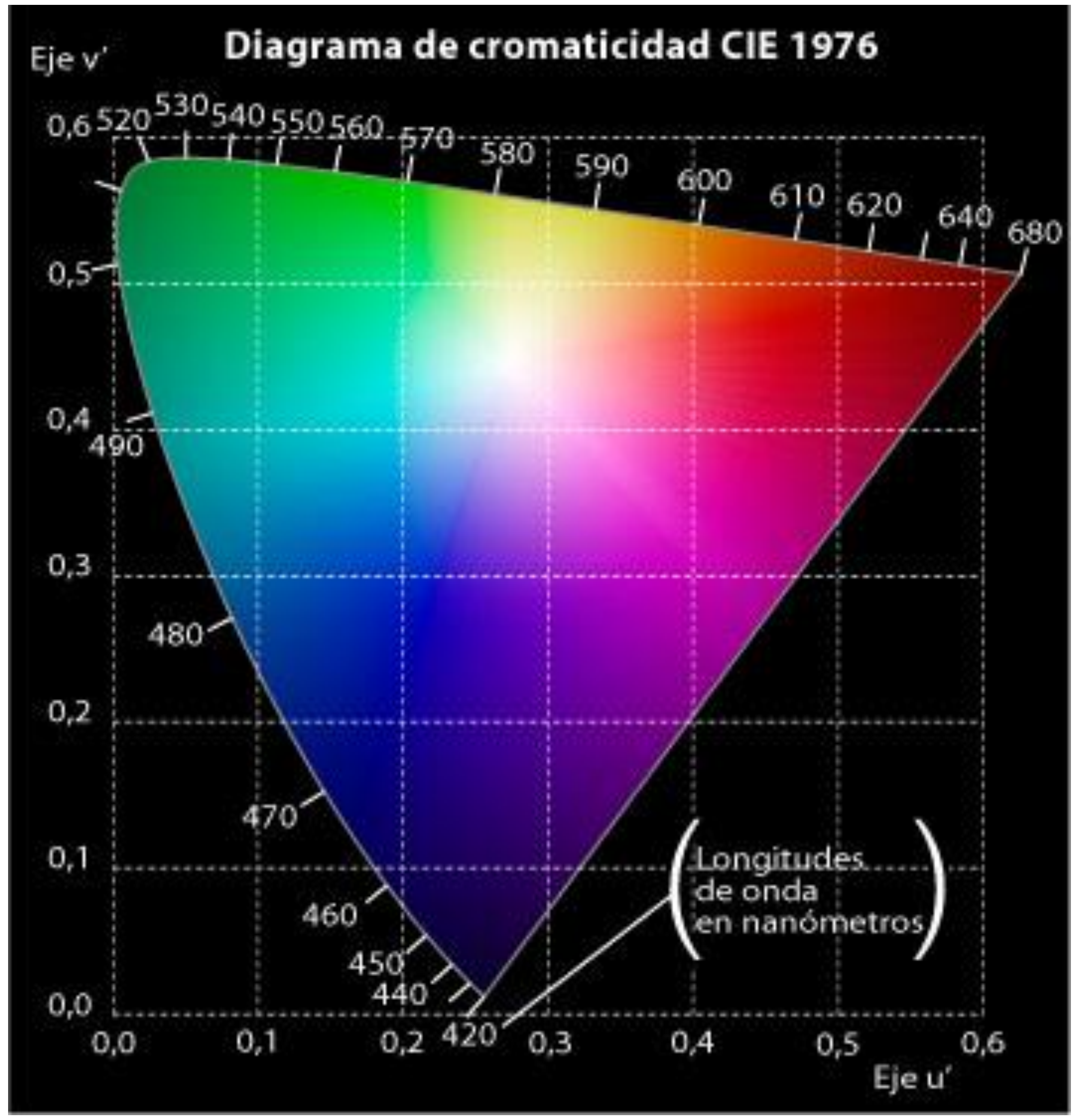
Color spaces

RGB HSL HSV CIEL





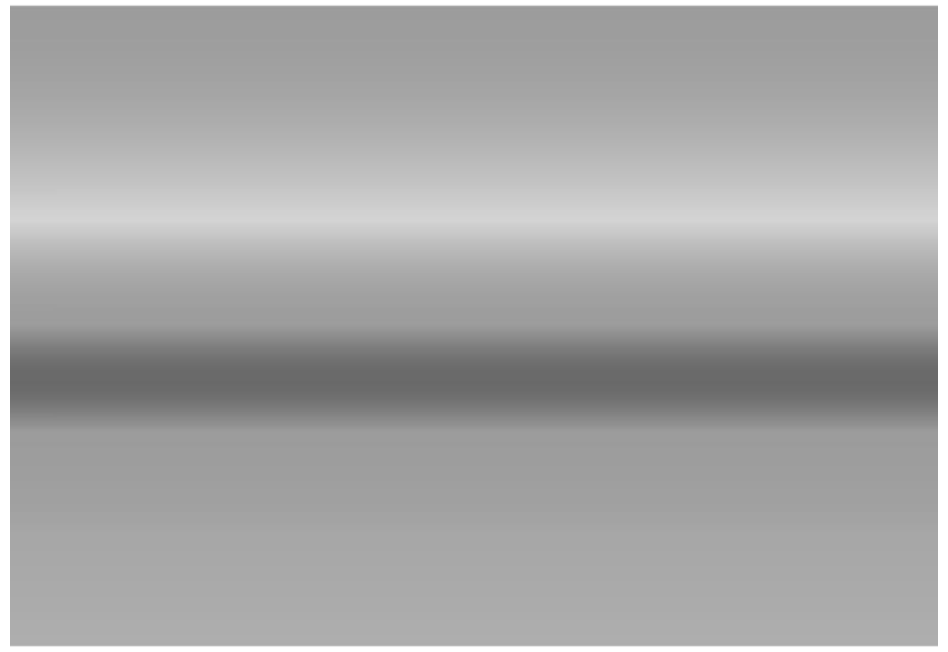


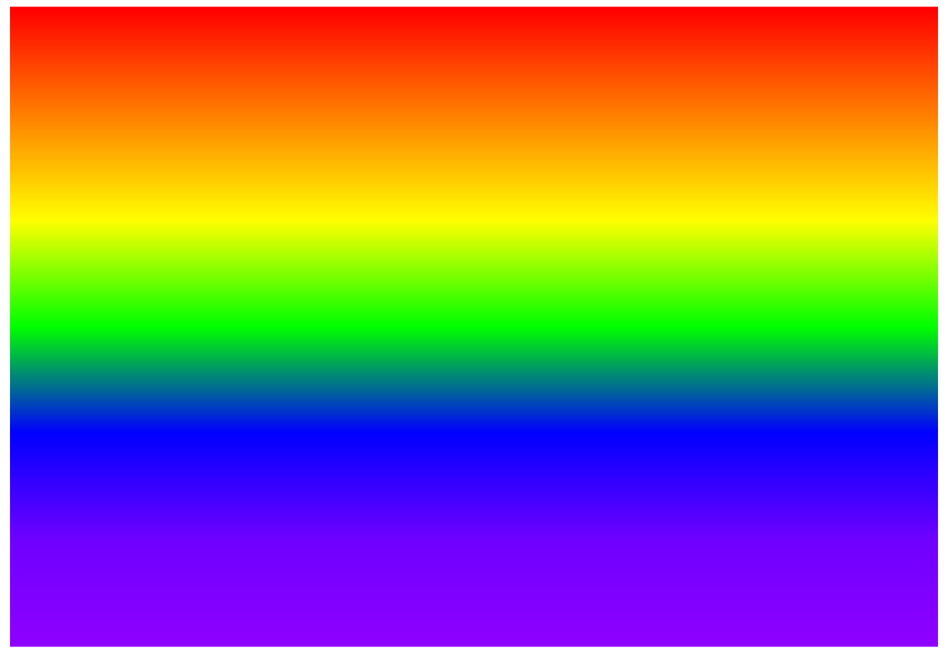


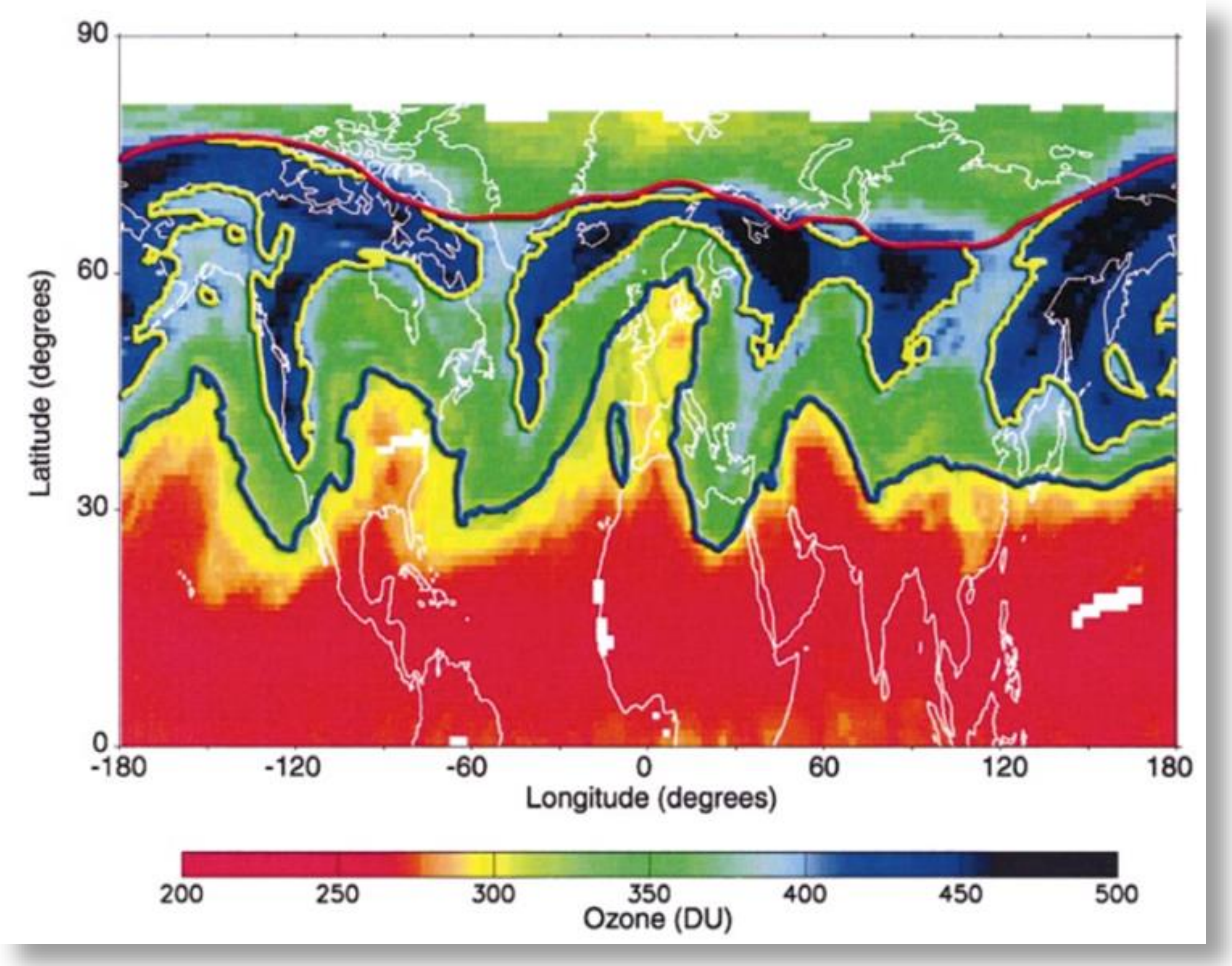
Escala arcoíris

Default en varios softwares de visualización

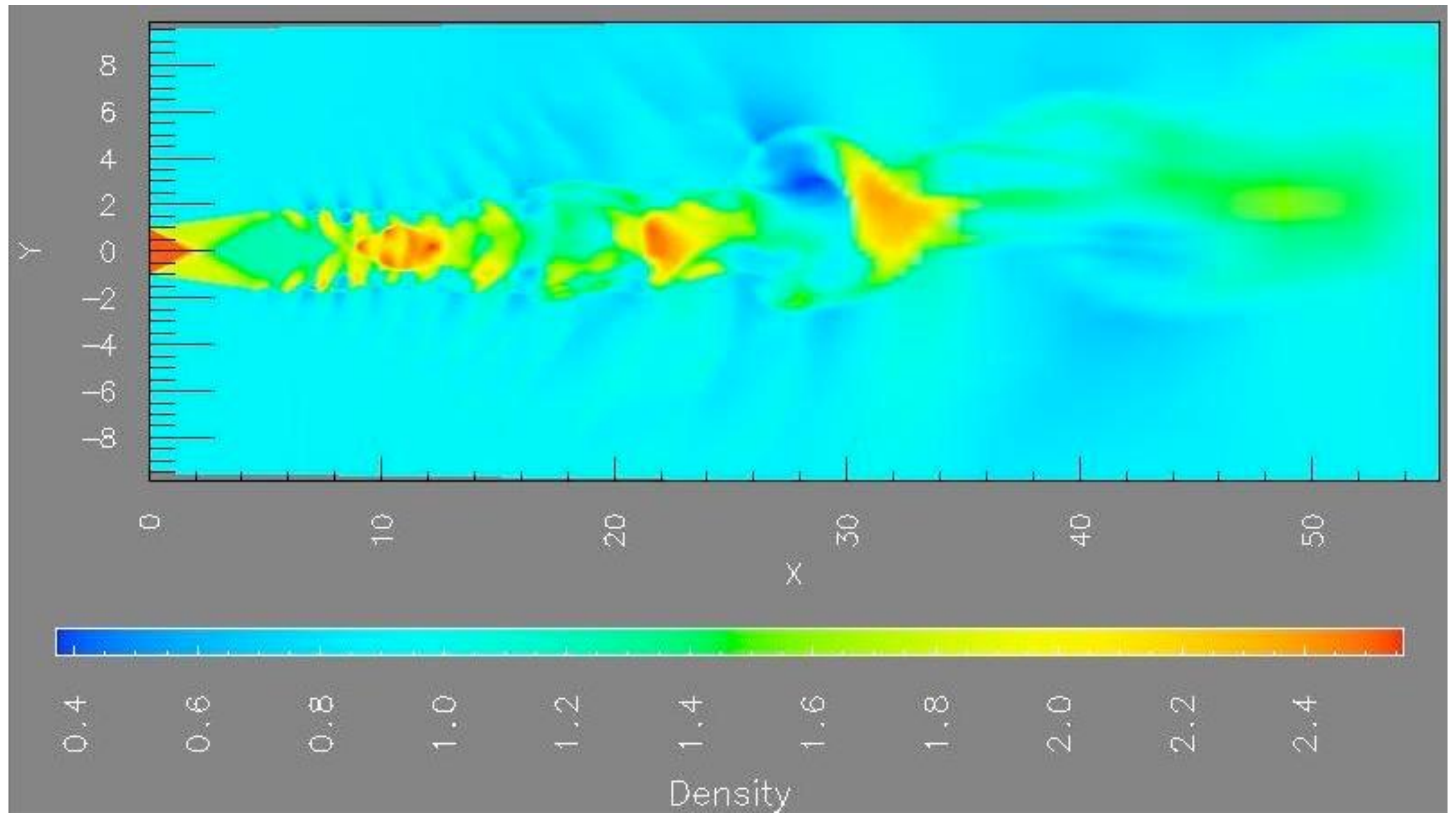
- Los tonos no tienen un orden inherente

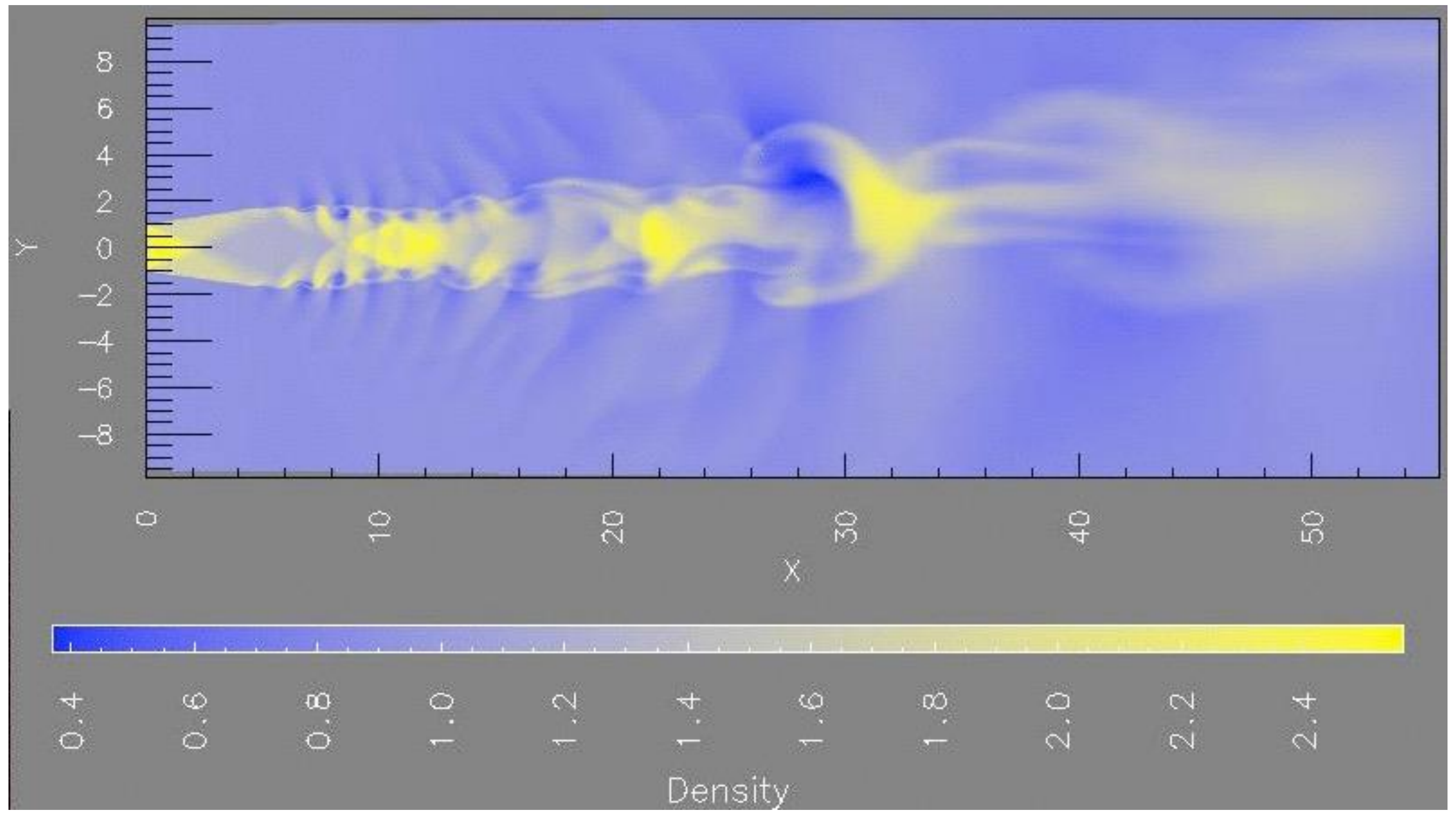






- Es más difícil ver algunos detalles
- Aparecen boundaries que no existen en realidad

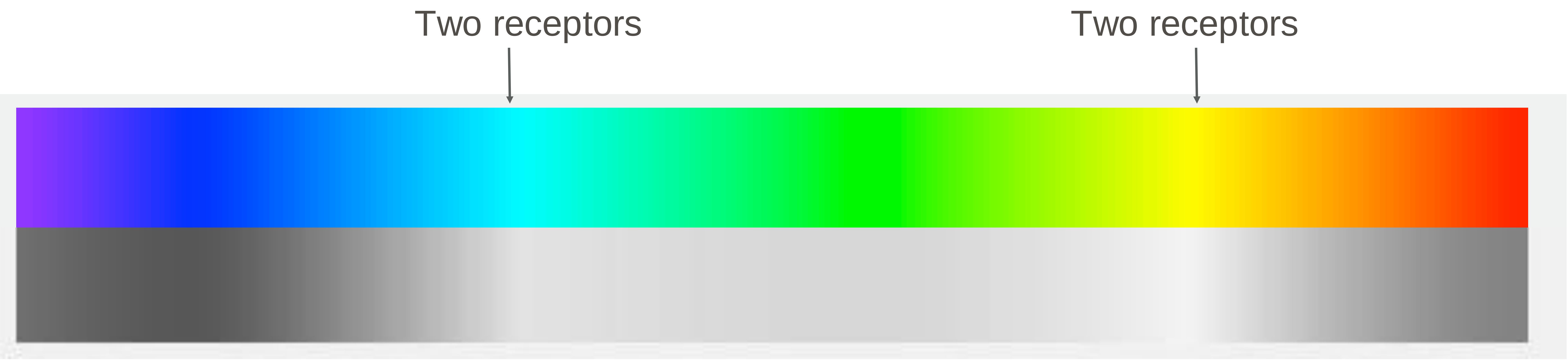


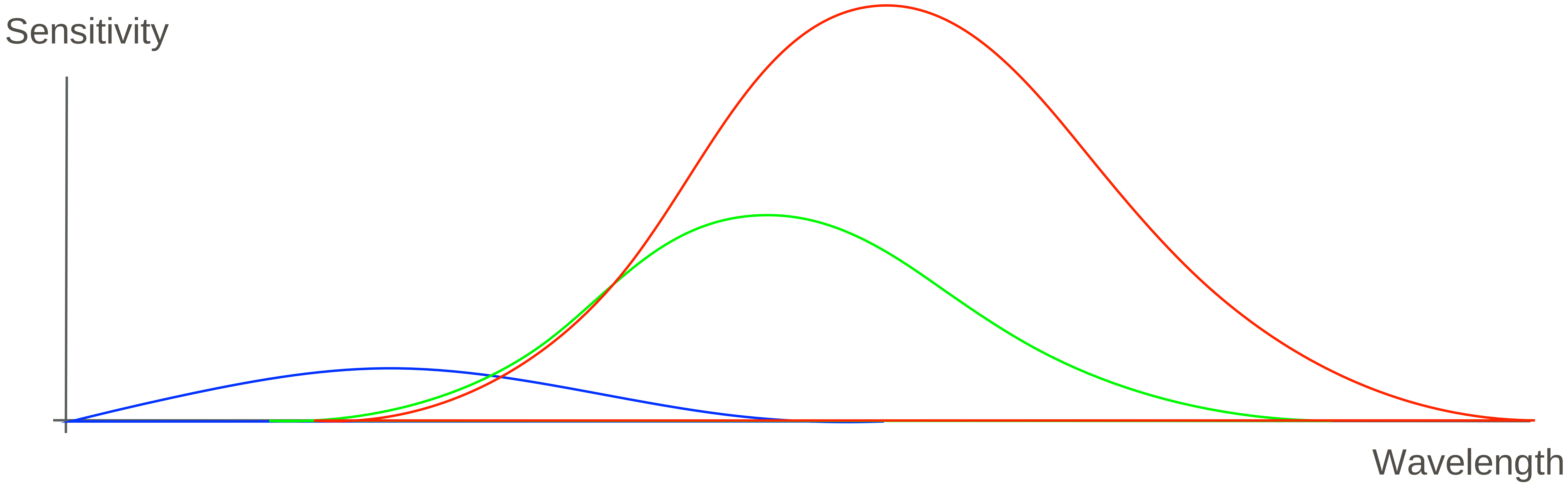


[A Rule-based Tool for Assisting Colormap Selection. Bergman,. Rogowitz, and. Treinish. Proc. IEEE Visualization (Vis), pp. 118–125, 1995.]

Problemas de la escala arcoíris, variación no-lineal y la dificultad para distinguir regiones:

- Sobreexcitación de receptores en algunos puntos, que rompen la linealidad de la escala.

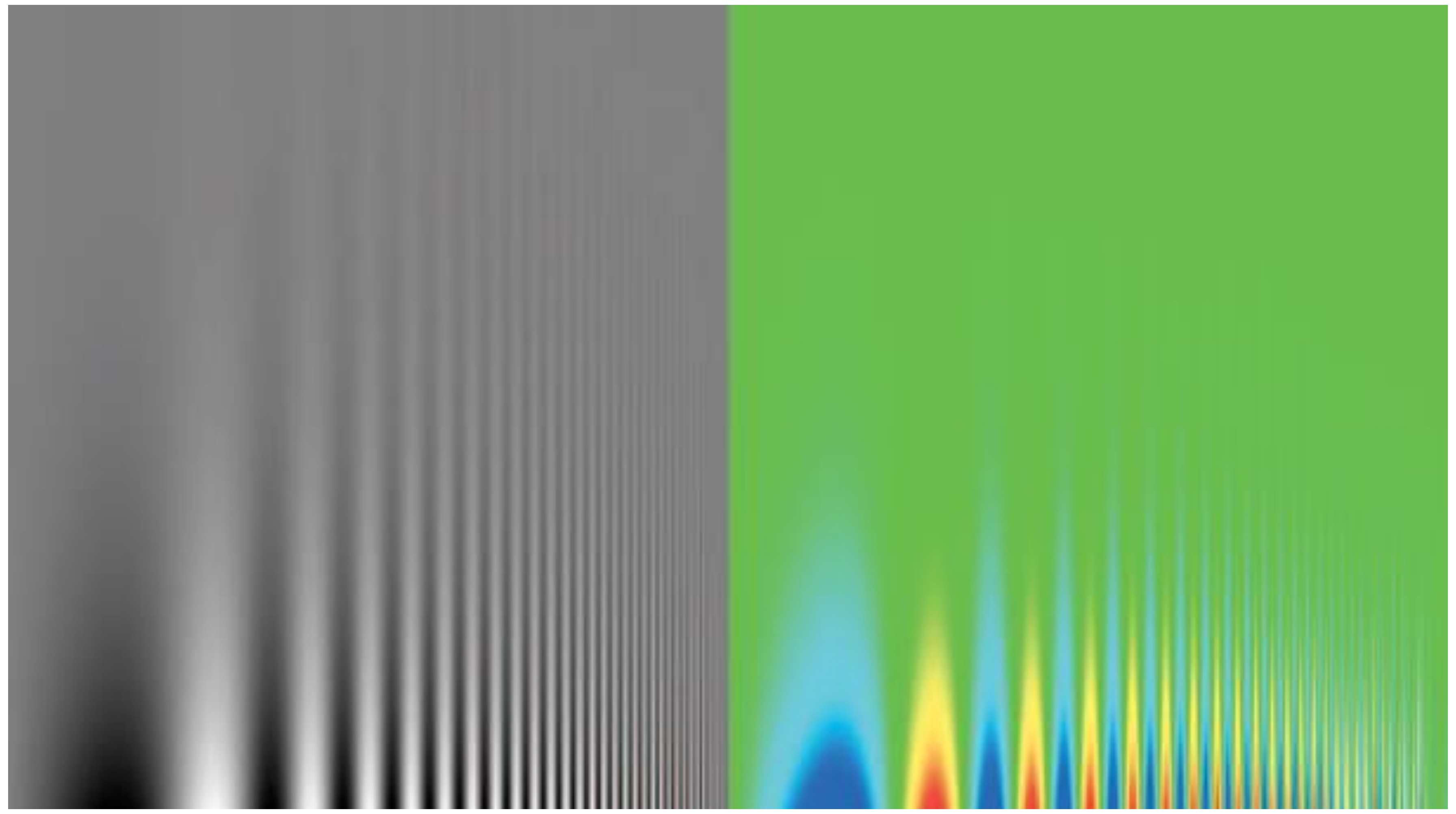


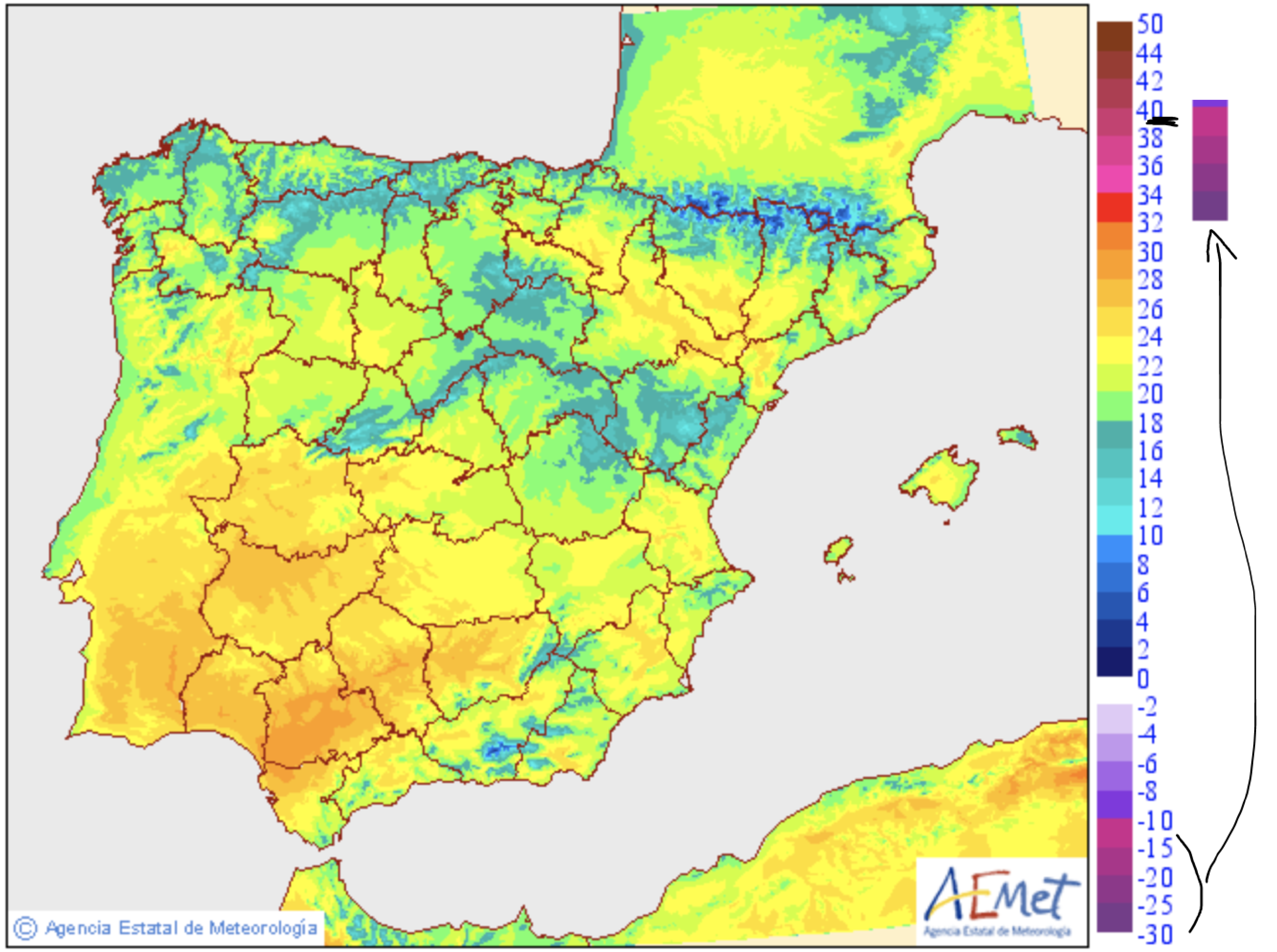


Escala arcoíris

No tiene en cuenta cómo percibimos el color

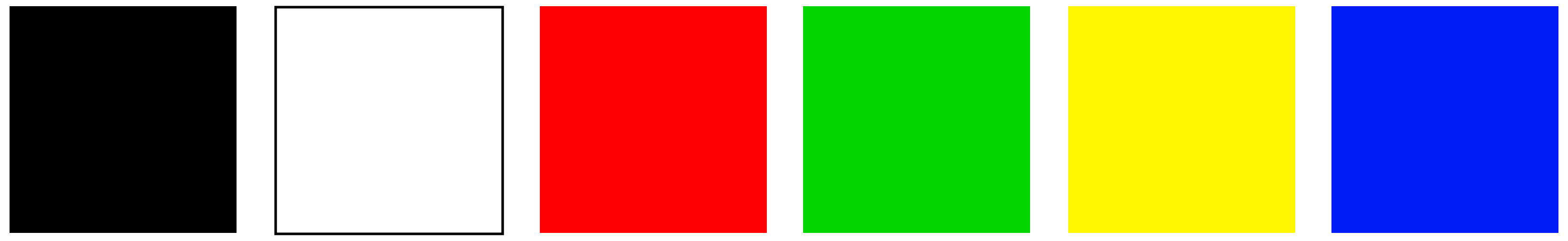
- No tiene un orden inherente
- Hay colores que percibimos más intensamente que otros
- “Bandas” perceptuales
- Pérdida de detalle





Channel Properties

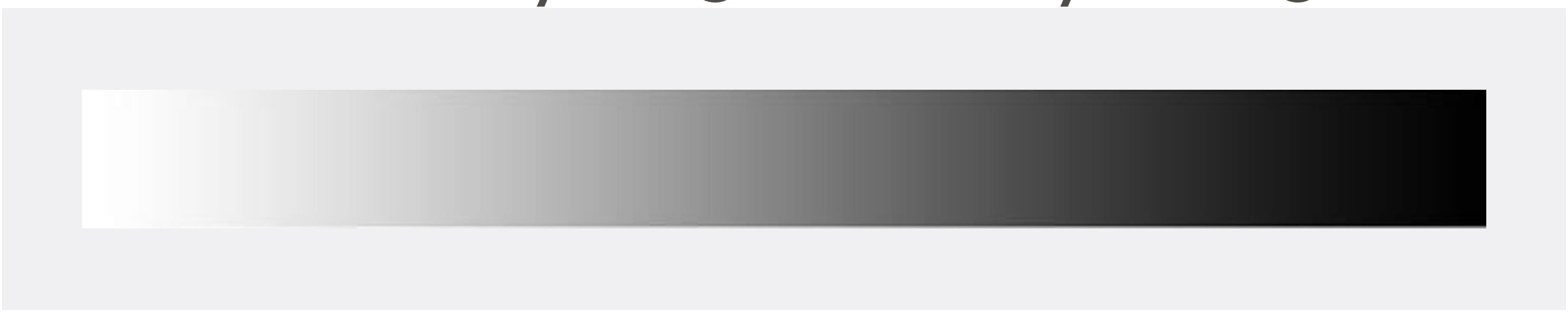
Tono (Hue)



Saturation

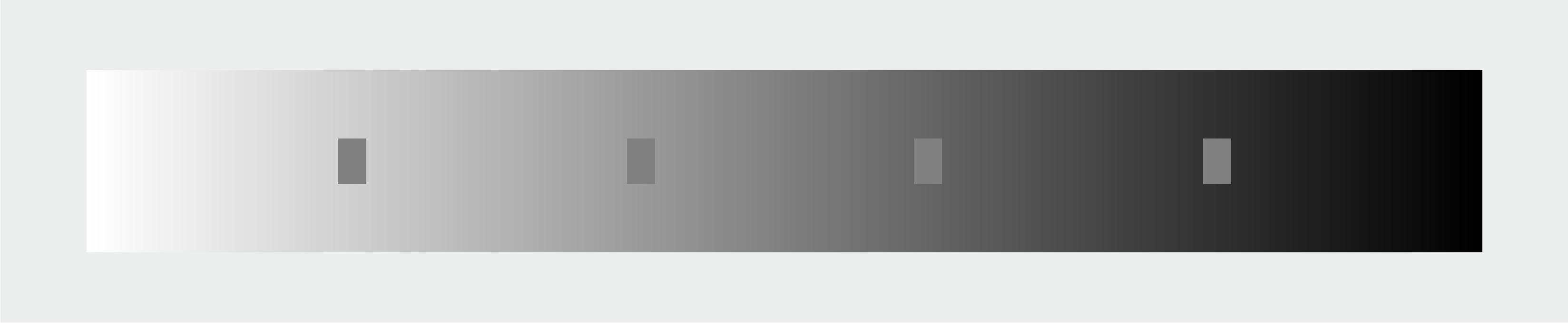
Contrast

Luminance/Lightness/Brightness

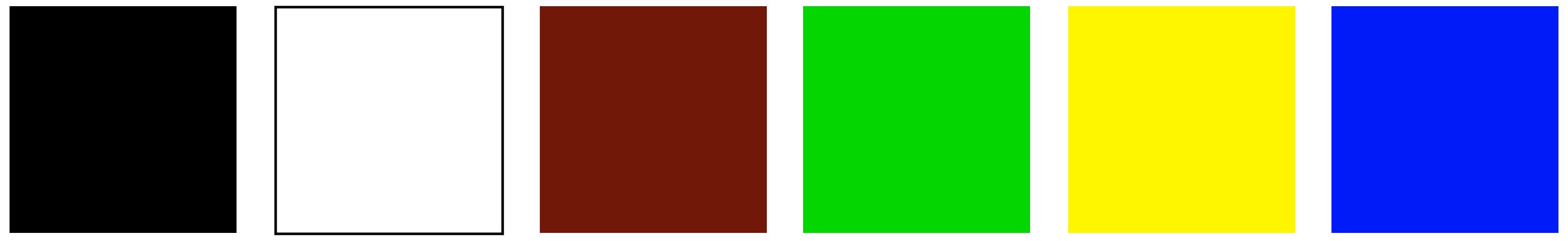


Channel Properties

Contrast

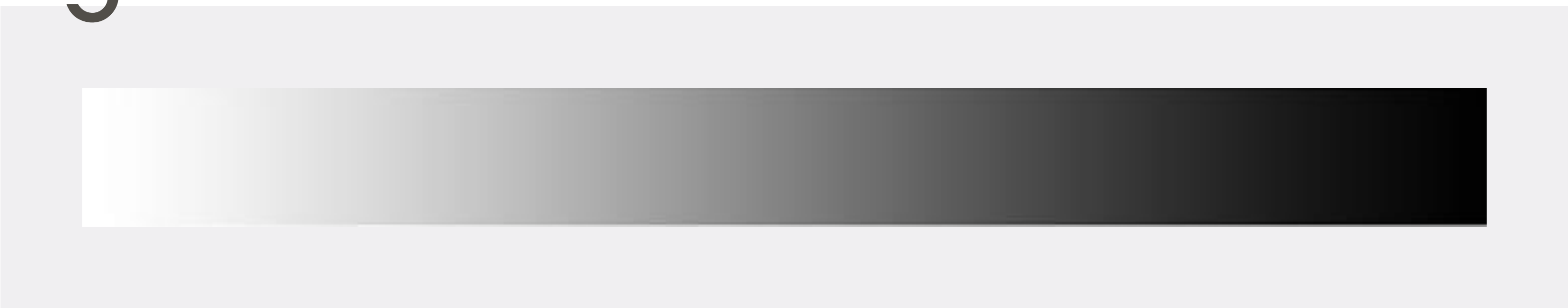


Unique hues

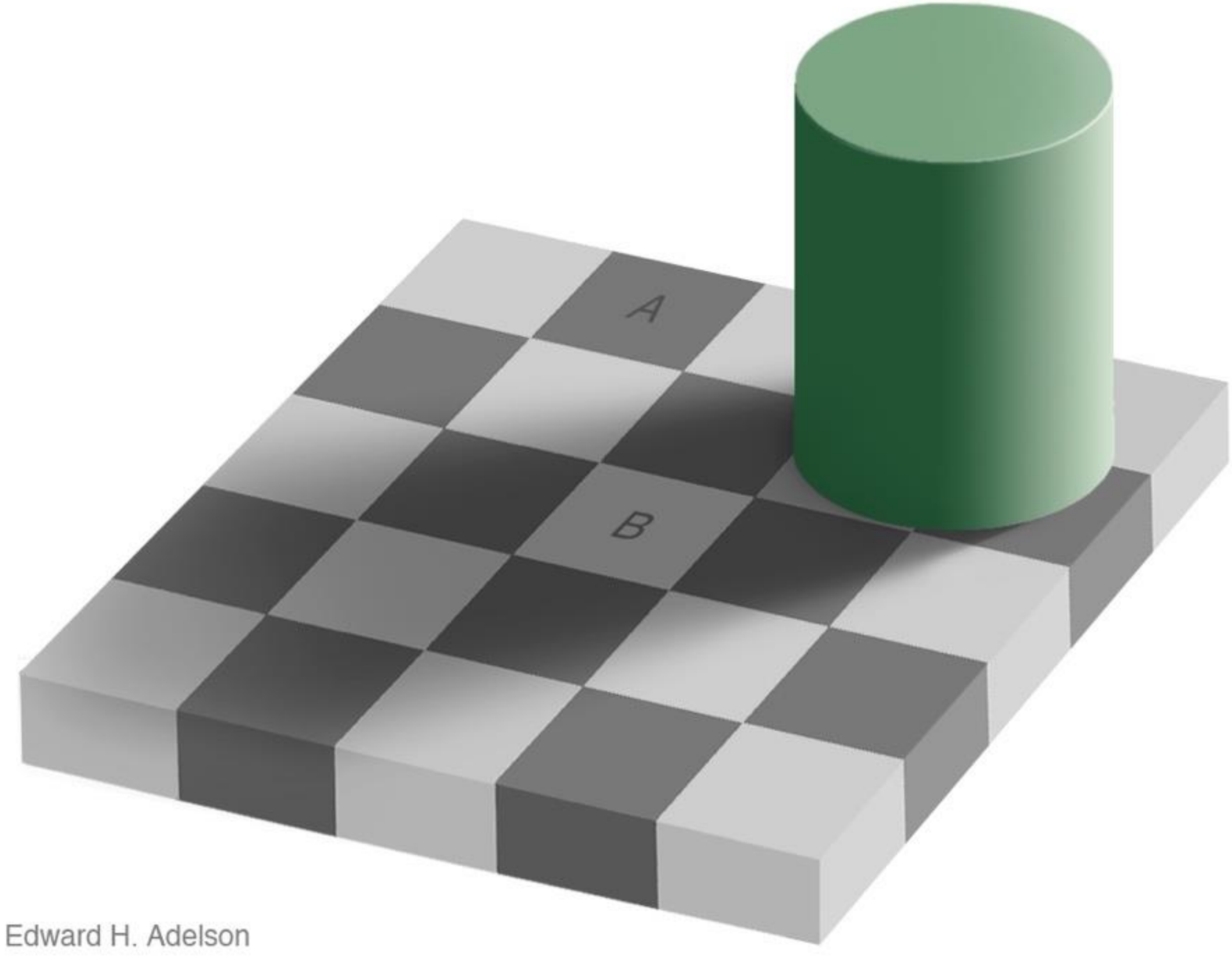


Saturation

Lightness

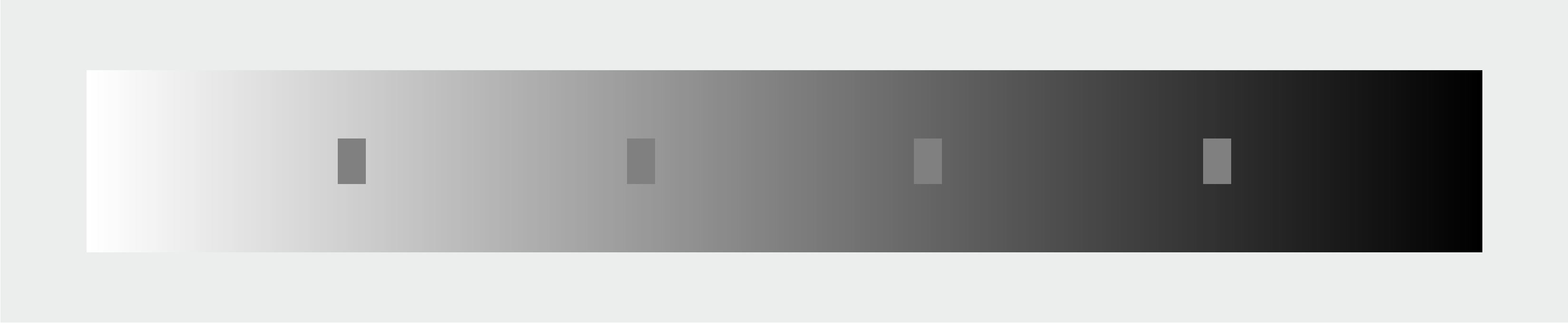


Simultaneous Contrast

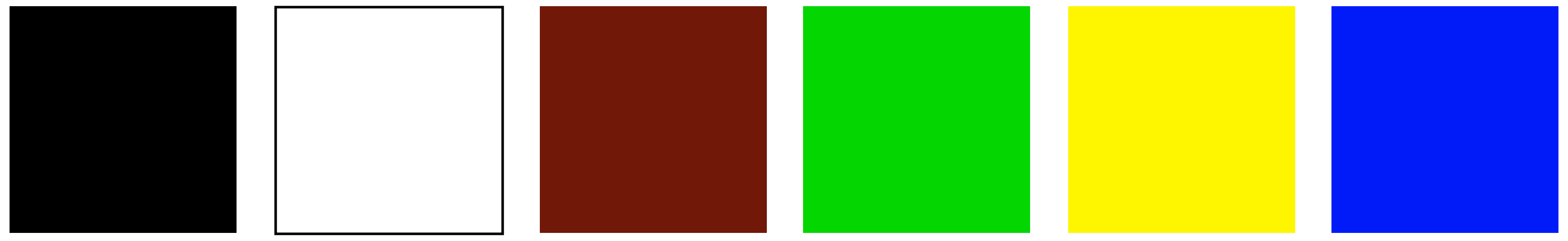


Channel Properties

Contrast

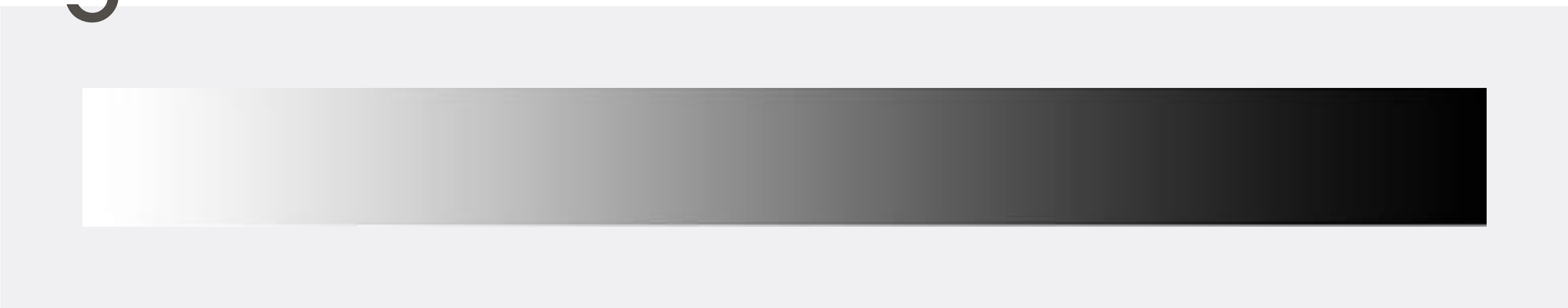


Unique hues

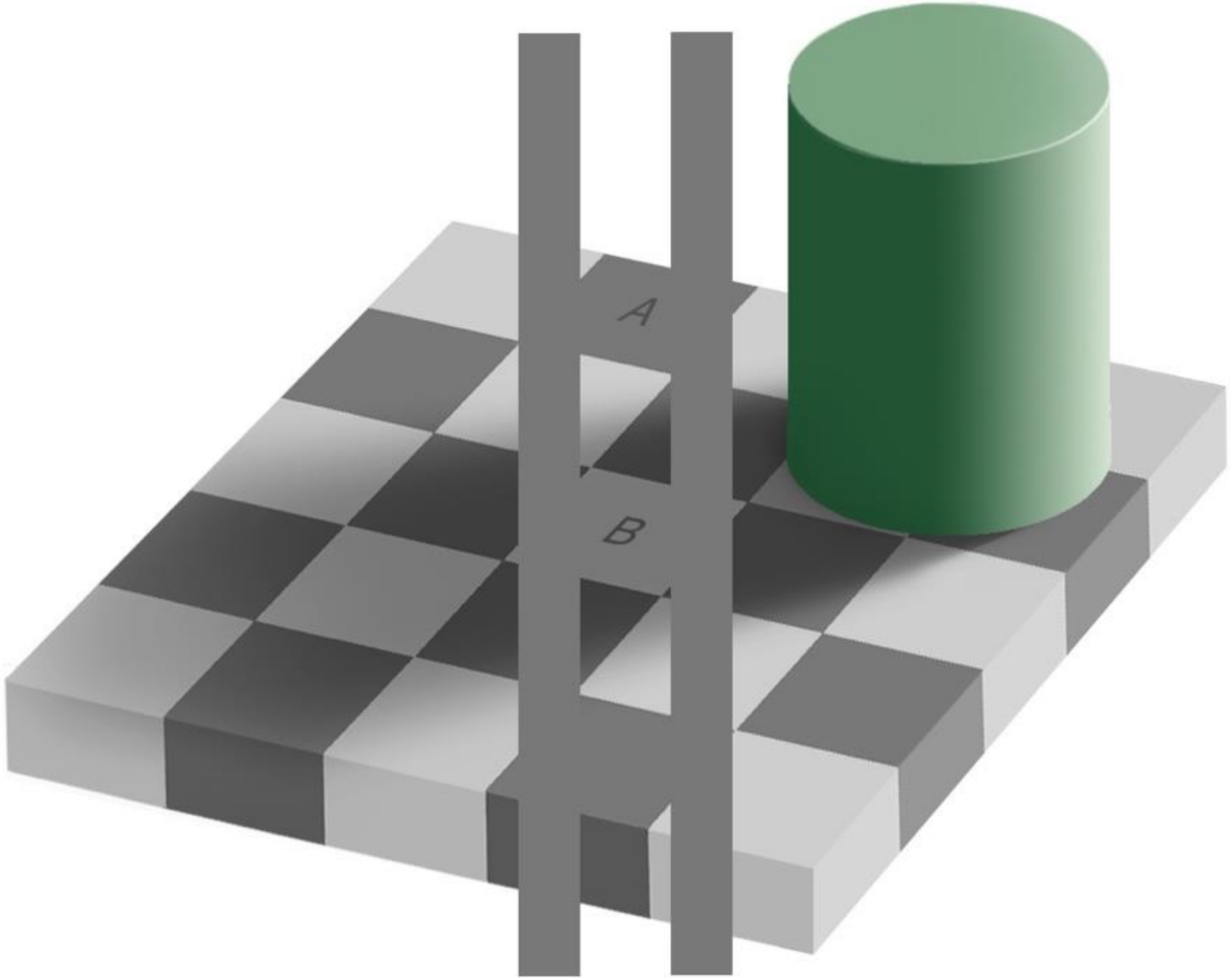


Saturation

Lightness

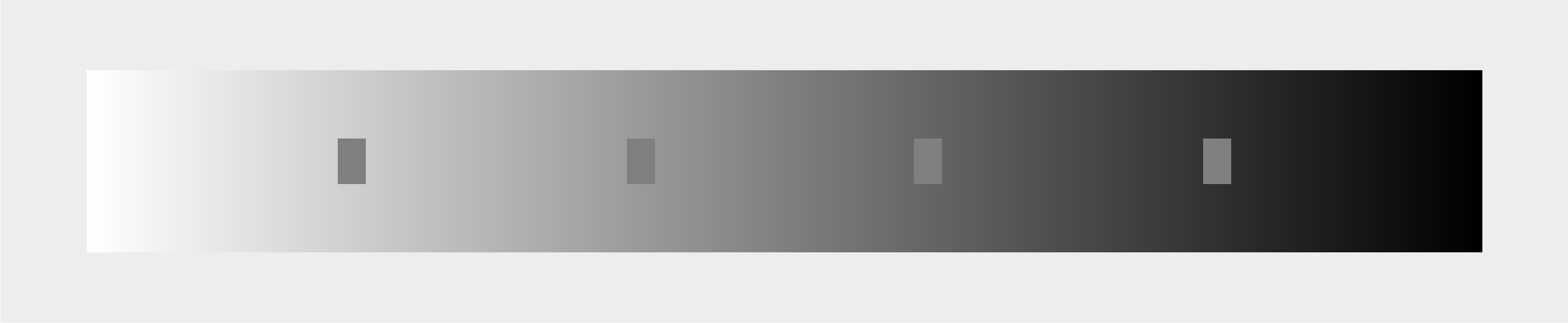


Simultaneous Contrast

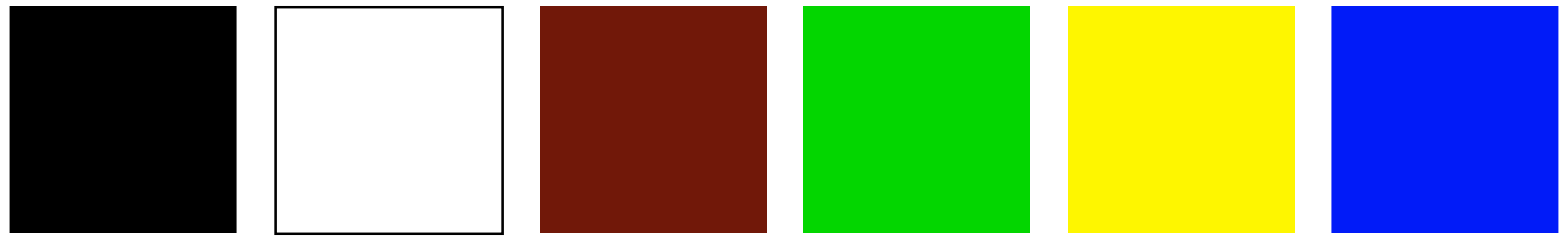


Channel Properties

Contrast

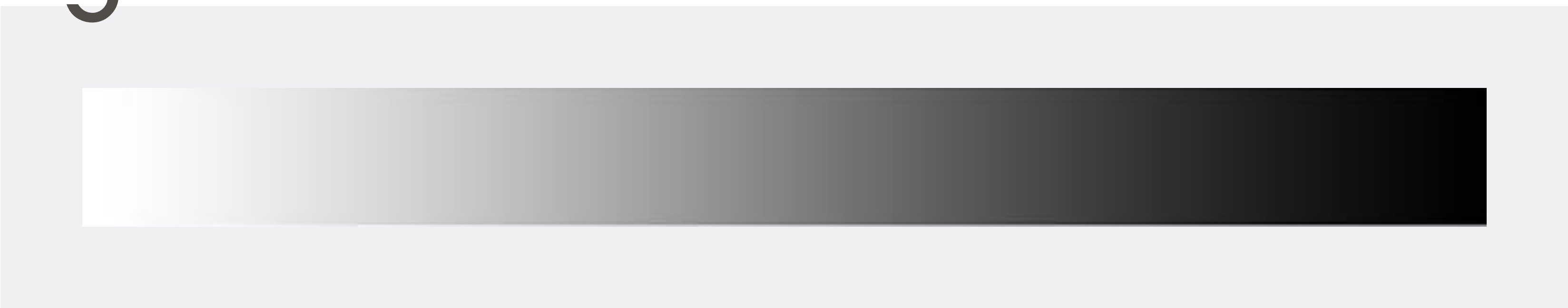


Unique hues

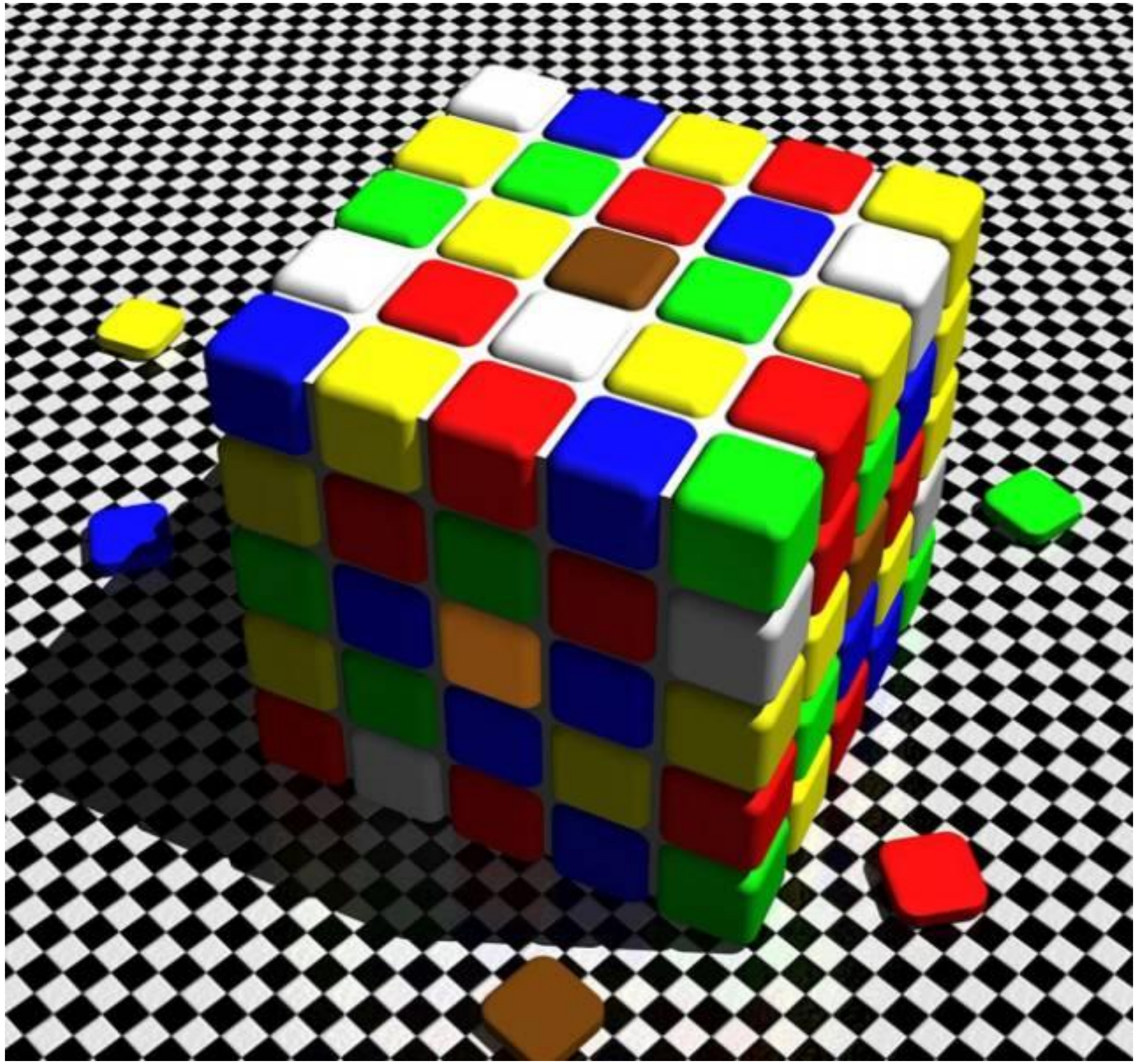


Saturation

Lightness

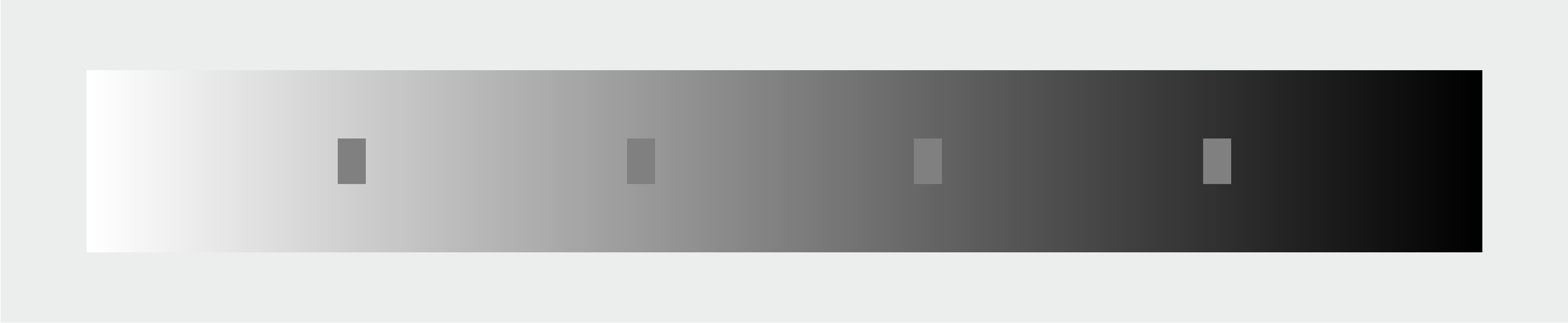


Color appearance

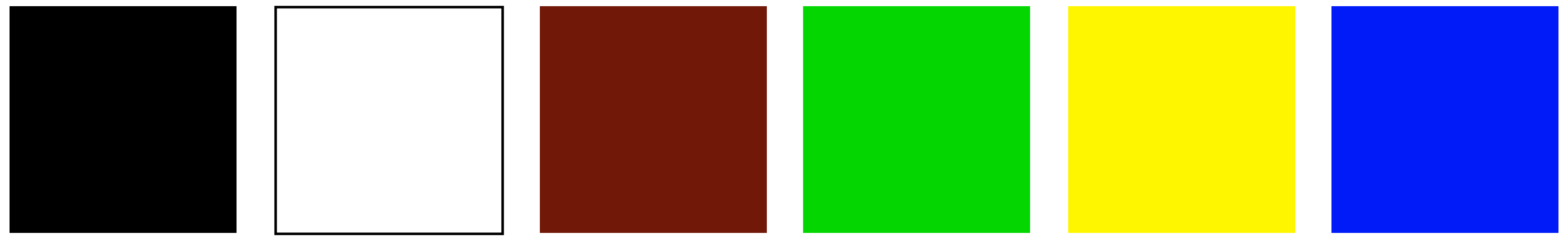


Channel Properties

Contrast

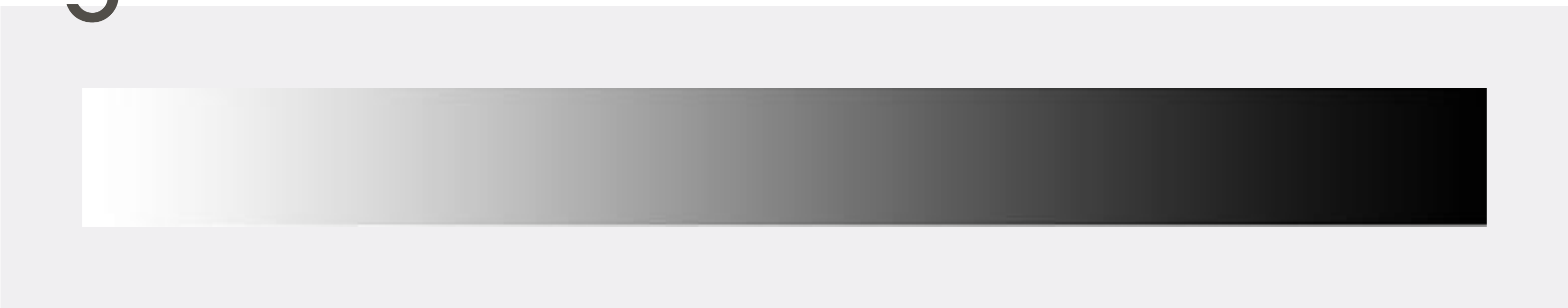


Unique hues

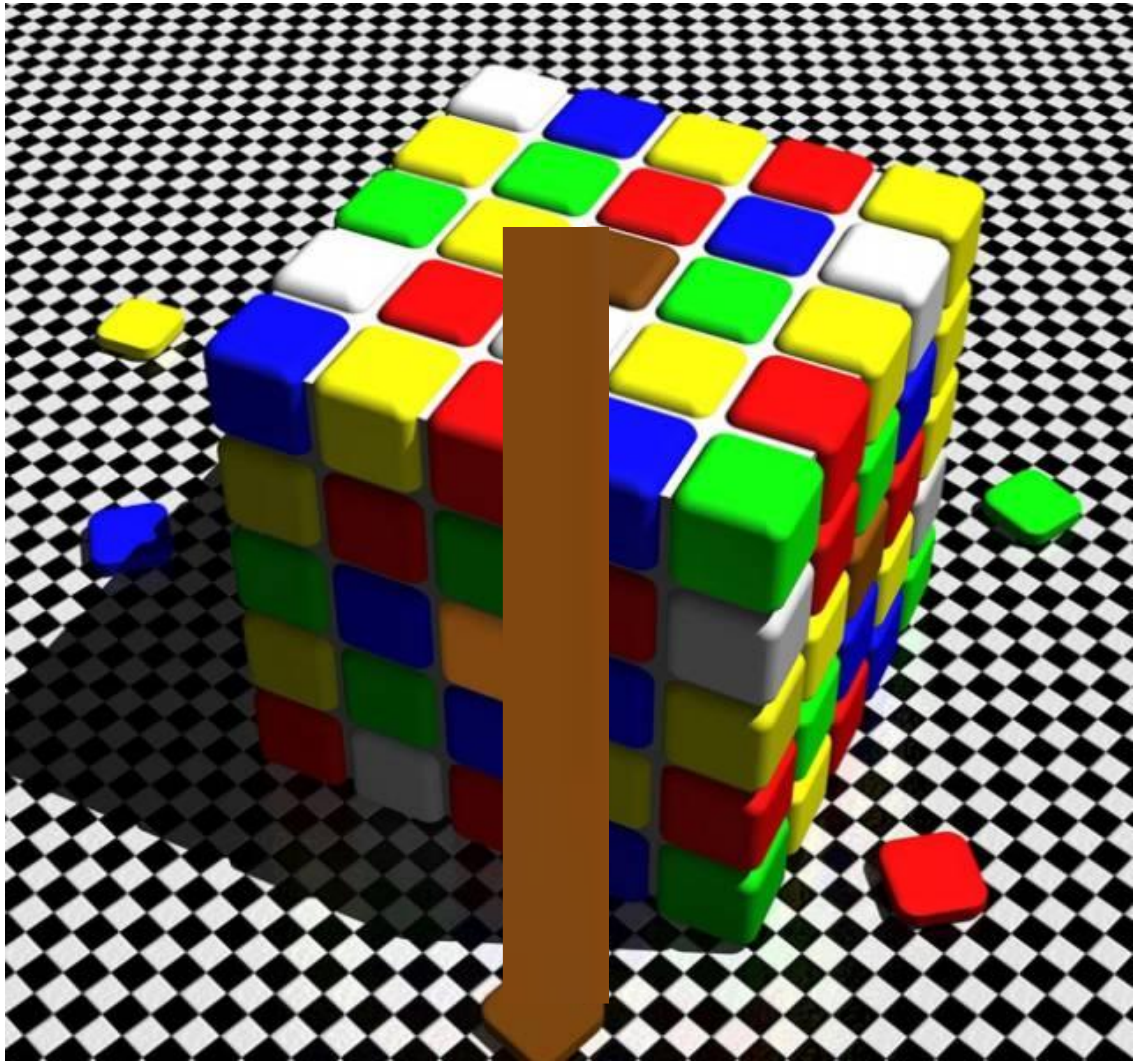


Saturation

Lightness

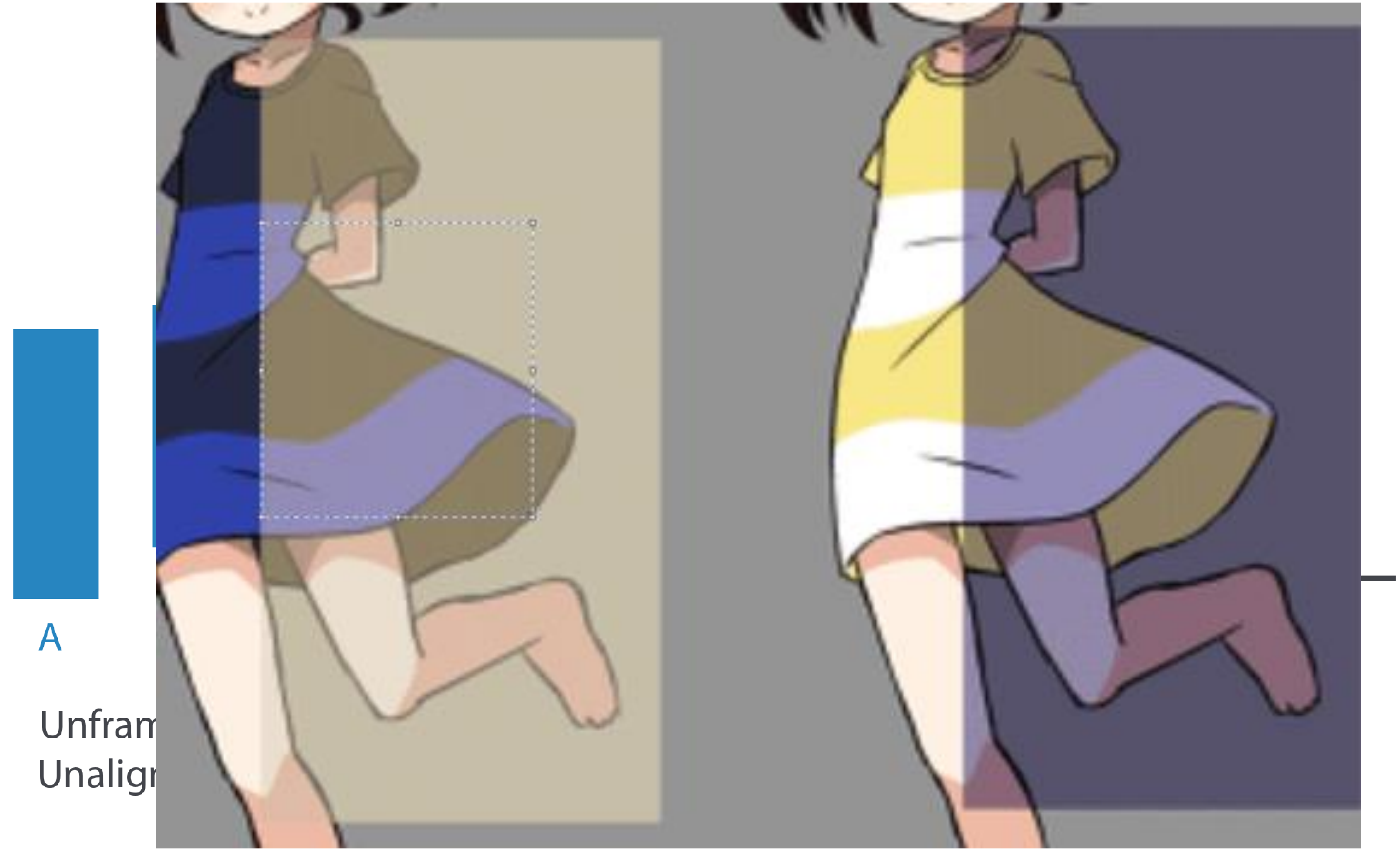


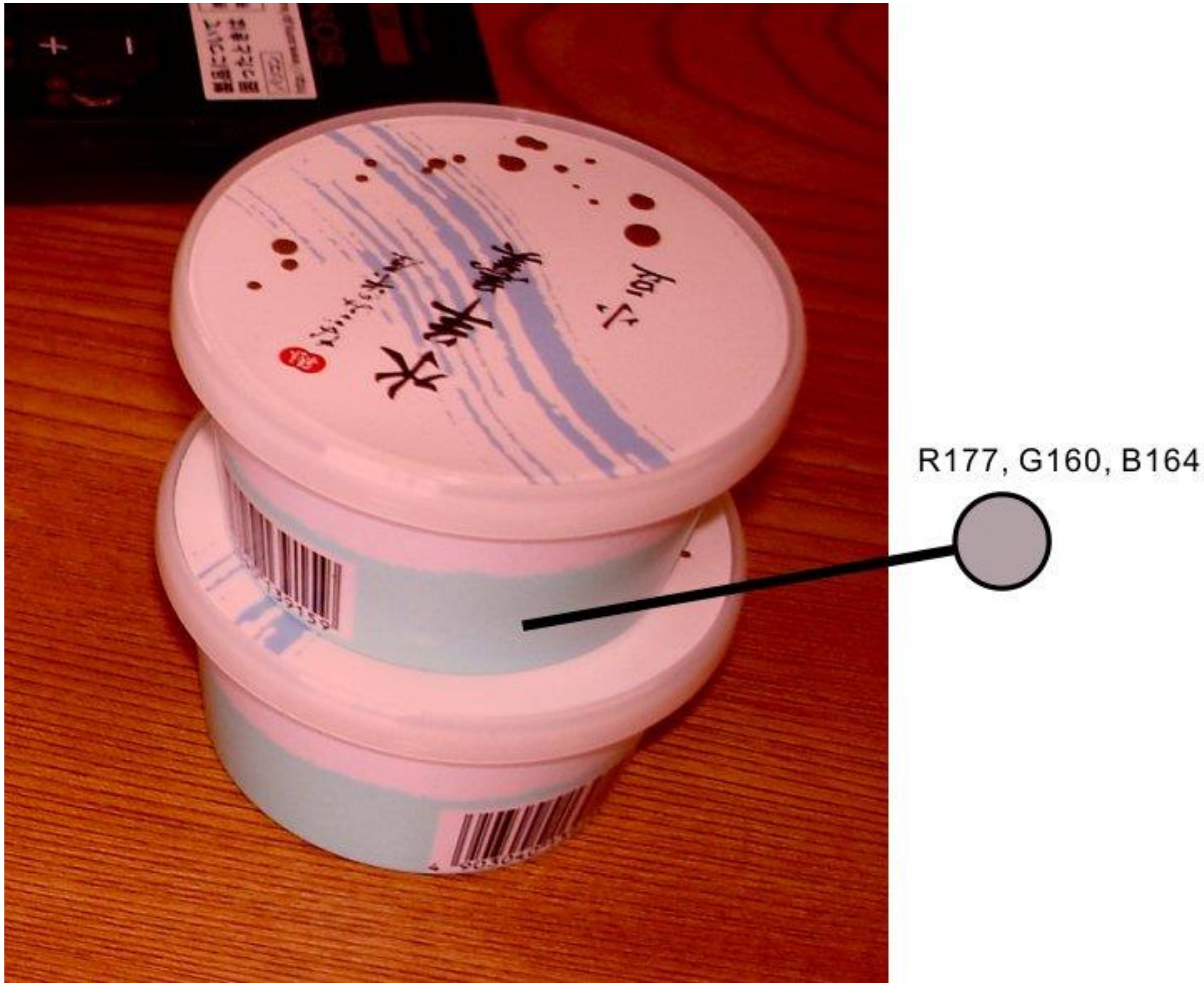
Color appearance



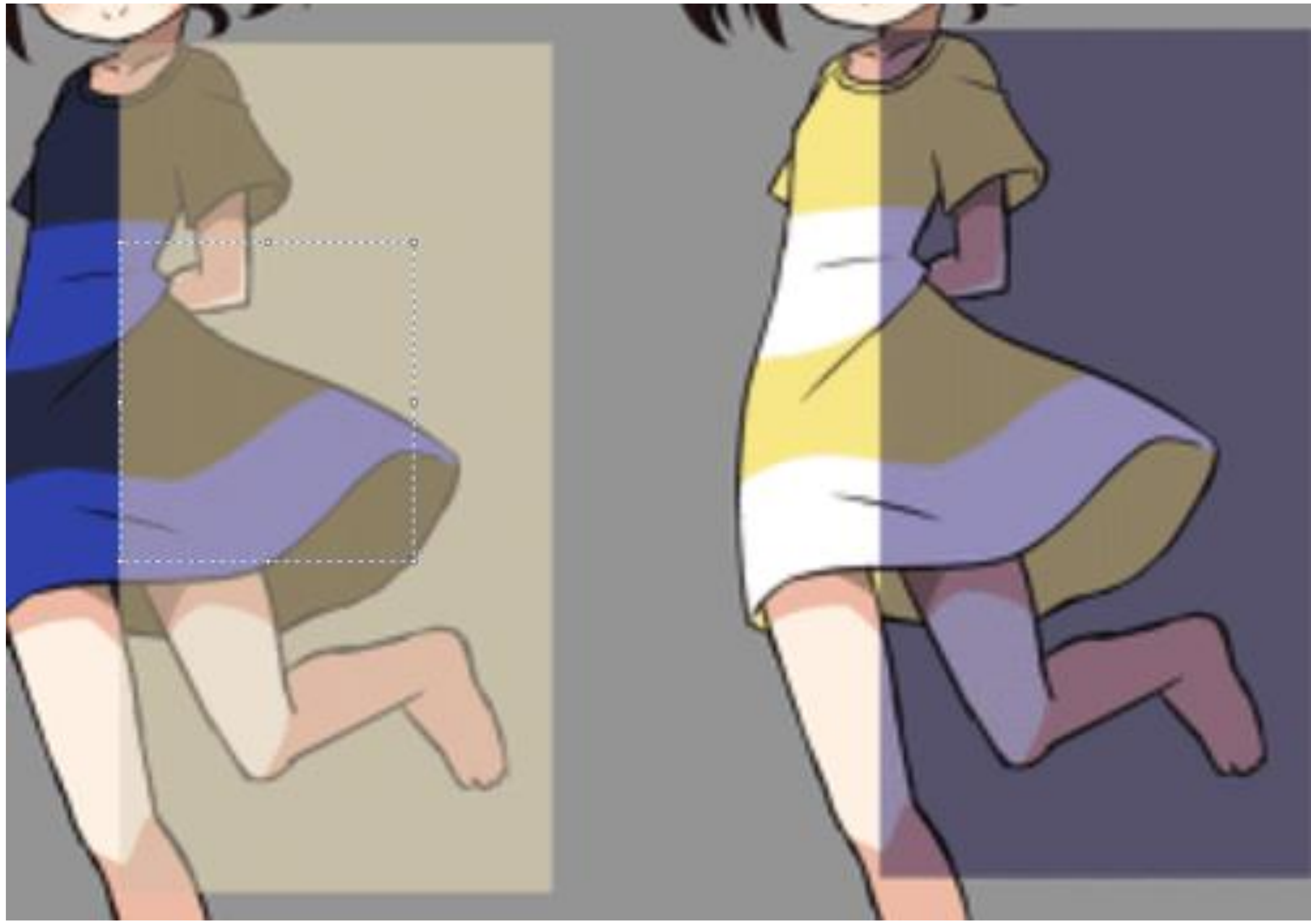
Relatividad

- Nuestro Sistema perceptual se basa fundamentalmente en juicios relativos, no absolutos
- El contexto que rodea a los elementos modula la aplicación de **Discriminabilidad** y **Precisión**



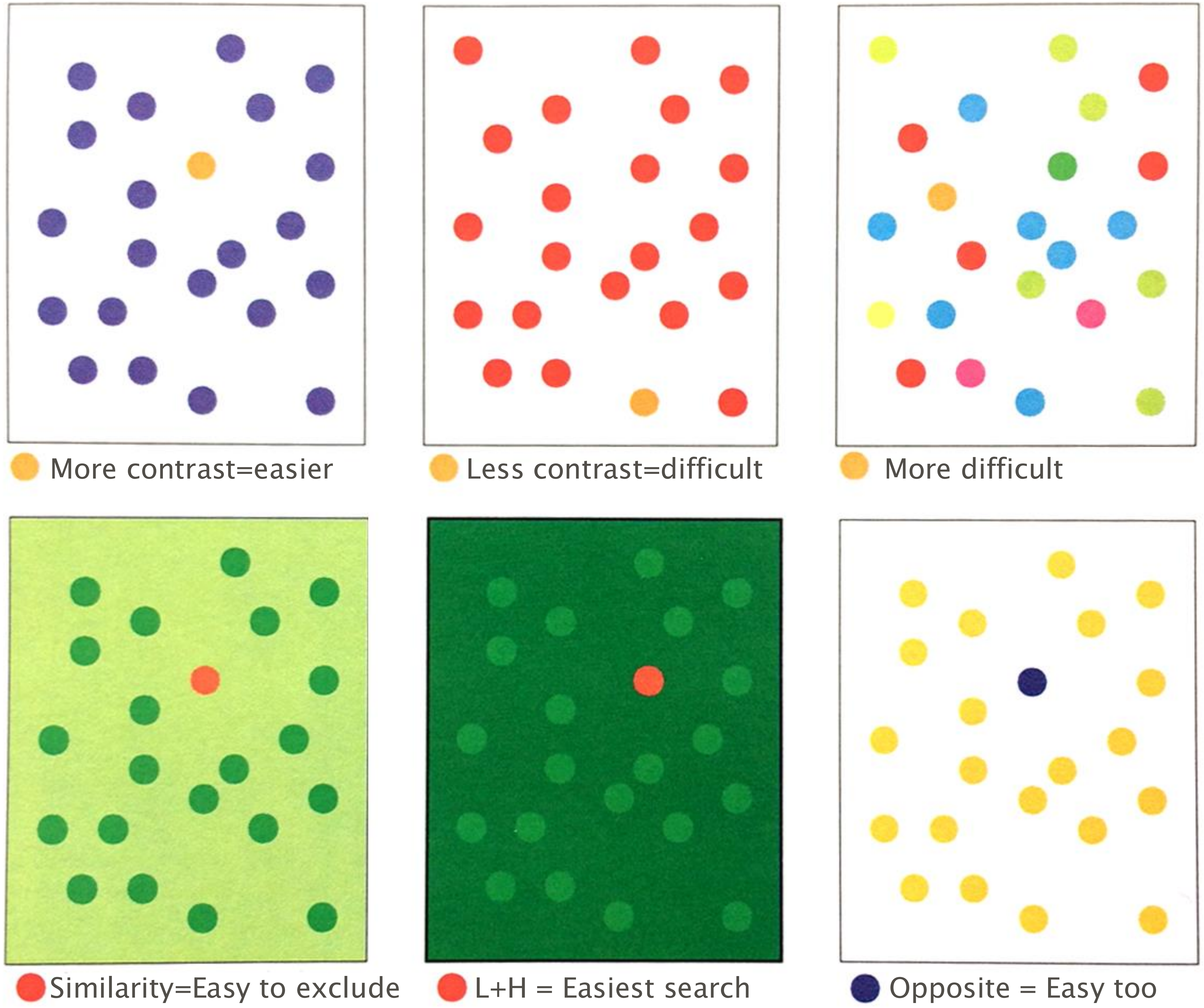


Akiyoshi kitaoka



Akiyoshi kitaoka

Percepción del Color - Contraste



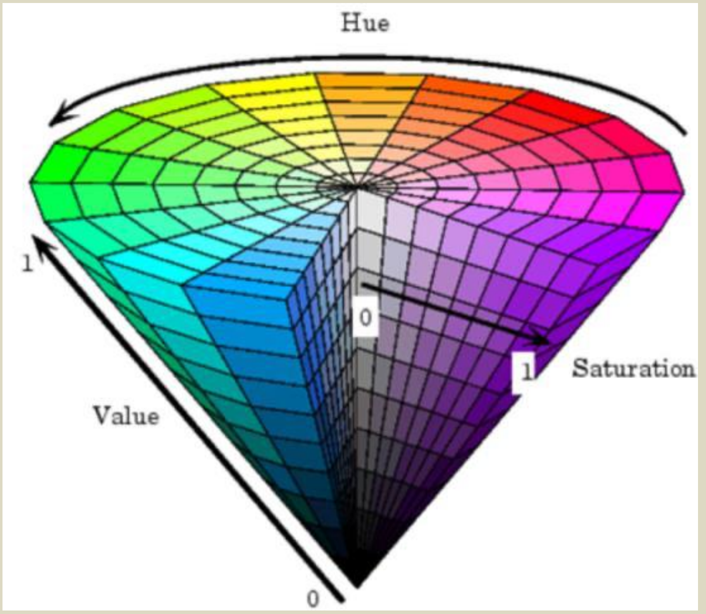
Ware, 2008

- El contraste facilita la detección
- Hue homogéneo
- Variación en Hue y Lightness= Más fácil

R cheatsheet

Todo esto nos sirve para saber que:

- Hay varios modelos de color, no todos son perceptualmente correctos.
- Para escalas de color necesitamos modelos de interpolación no lineal.
- Algunos colores tienen mayor contraste entre si. Podemos usarlo mara dirigir la atención.
- El tamaño modula la percepción de los colores
- Evitar artefactos (contraste simultaneo, apariencia de color)



Page 1, Melanie Frazier

R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggbb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0
i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

R has 657 built in color names Example: peachpuff4

To see a list of names:
colors()
These colors are displayed on P. 3.

R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.
  rgb(r, g, b, maxColorValue=255, alpha=255)
- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument
  hsv(h, s, v, alpha)
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
  hcl(h, c, l, alpha)

A few notes on HSV/HLC

HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model....blah blah blah...

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.

R can translate colors to rgb (this is handy for matching colors in other programs)
col2rgb(c("#FF0000", "blue"))

R Color Palettes

This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure....NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps

grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

grDevices palettes: cm.colors, topo.colors, terrain.colors, heat.colors, rainbow — see P. 4 for options

heat.colors(4, alpha=1)
> #FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFFF80FF"

For the rainbow palette you can also select start/end color (red = 0, yellow = *1/6*, green = *2/6*, cyan = *3/6*, blue = *4/6* and magenta = *5/6*) and saturation (s) and value (v):
rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)

Package: RcolorBrewer

This function has an argument for the number of colors and the color palette (see P. 4 for options).
brewer.pal(4, "Set3")
> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: display.brewer.all(5)
There is also a very nice interactive viewer:
http://colorbrewer2.org/

My Recommendation

Package: colorspace

These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

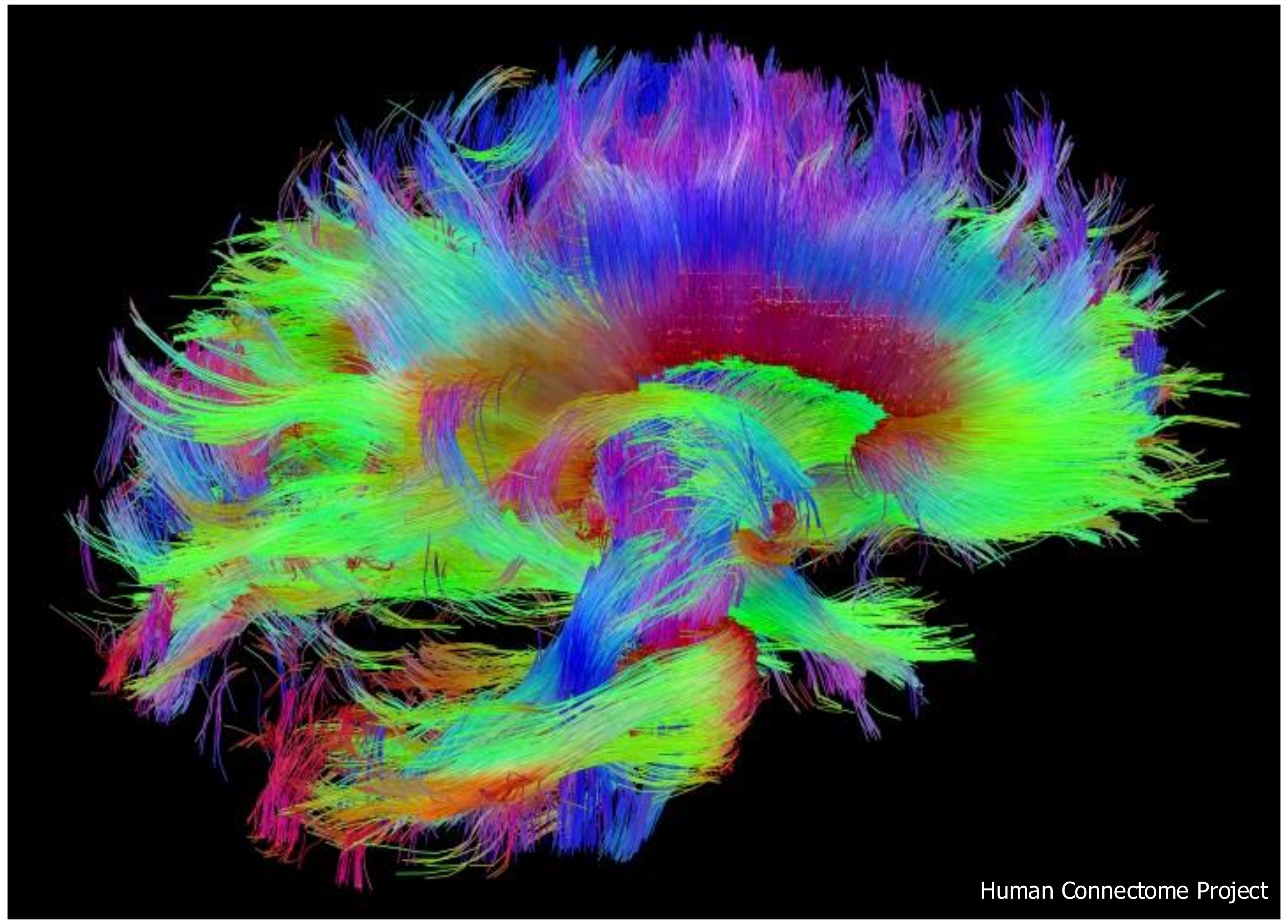
colorspace default palettes: diverge_hcl, diverge_hsl, terrain_hcl, sequential_hcl, rainbow_hcl

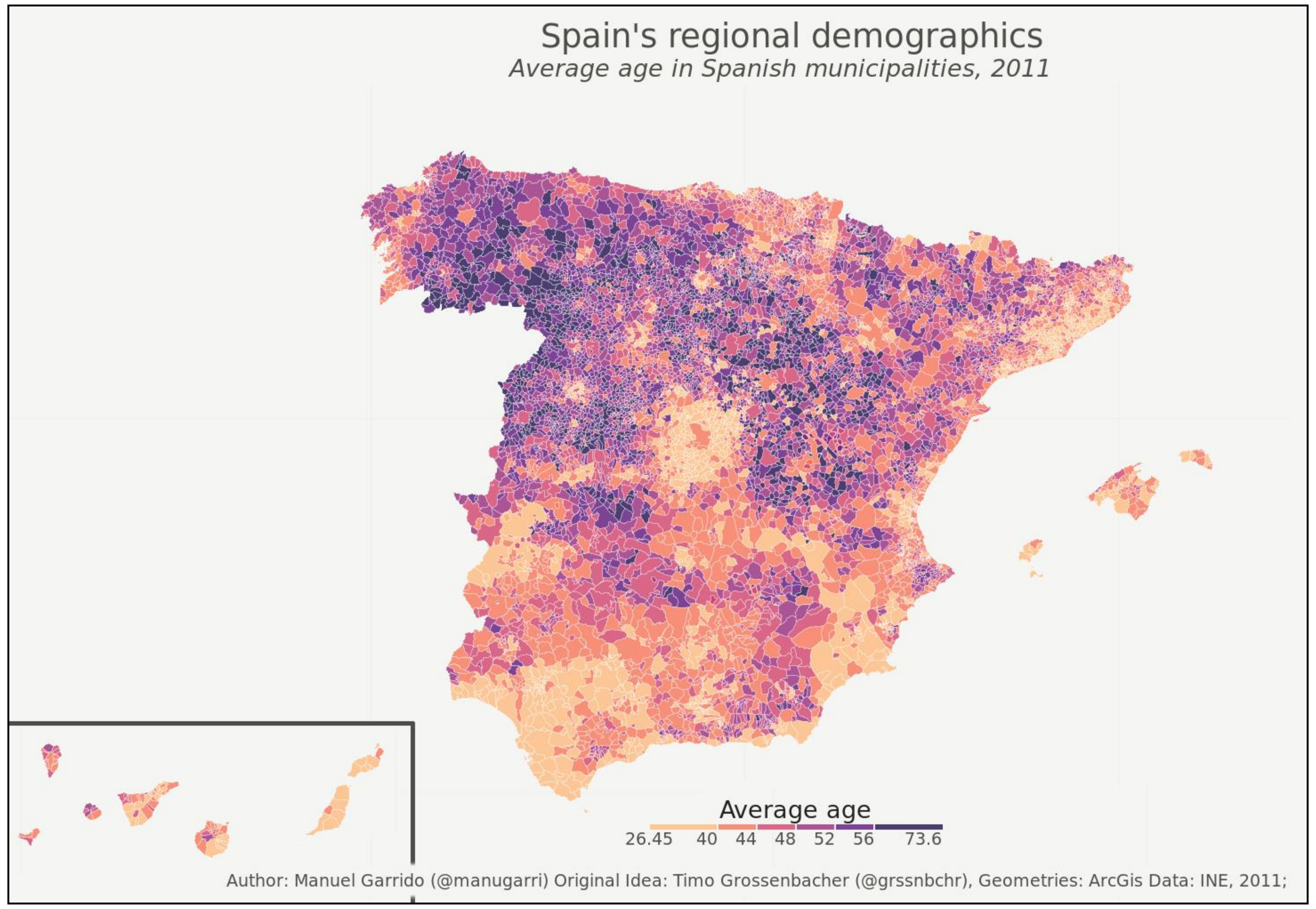
rainbow_hcl(4)
"#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"

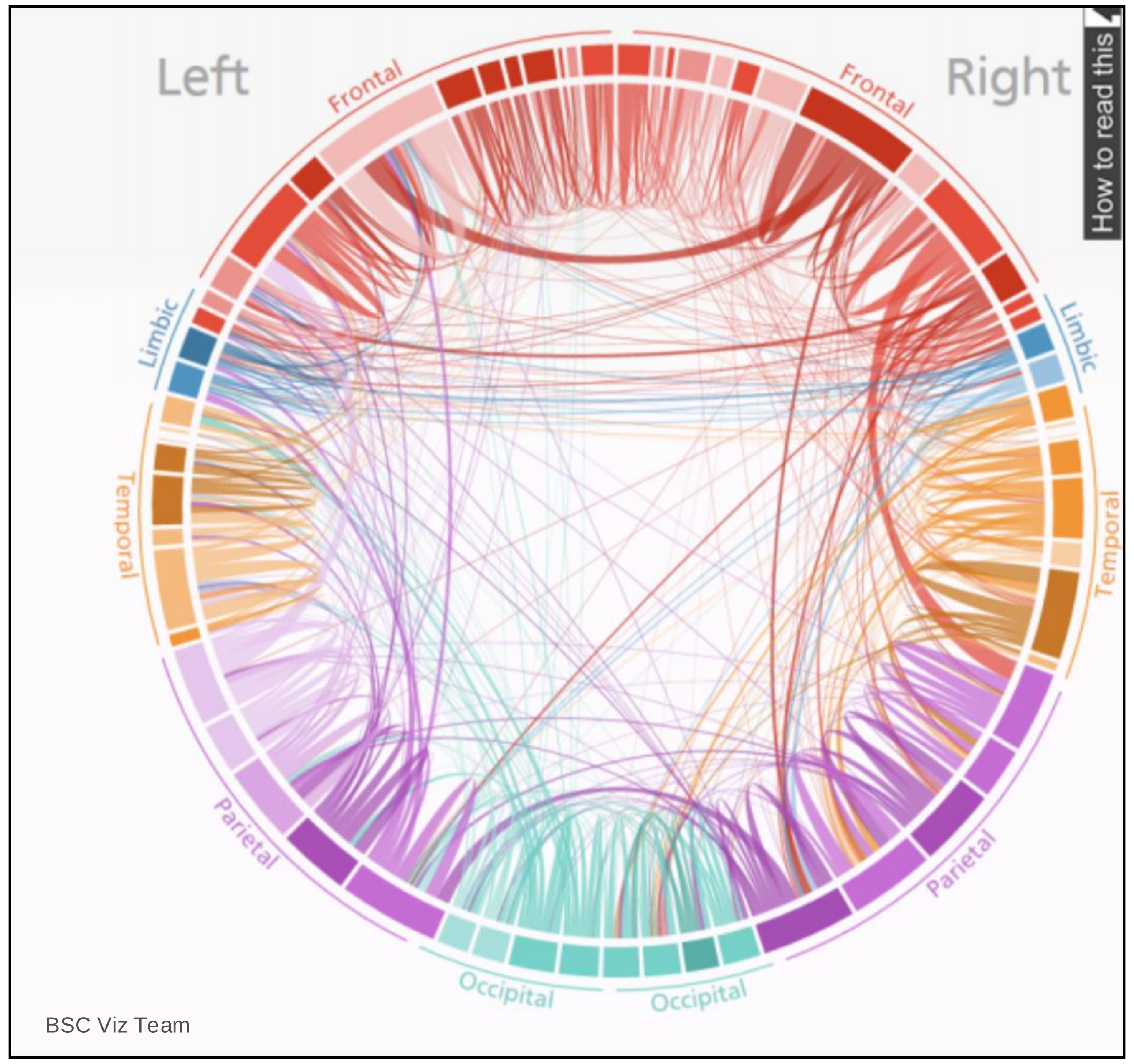
However, all palettes are fully customizable:
diverge_hcl(7, h = c(246, 40), c = 96, l = c(65, 90))
Choosing the values *would* be daunting. But there are some recommended palettes in the colorspace documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
pal <- choose_palette()

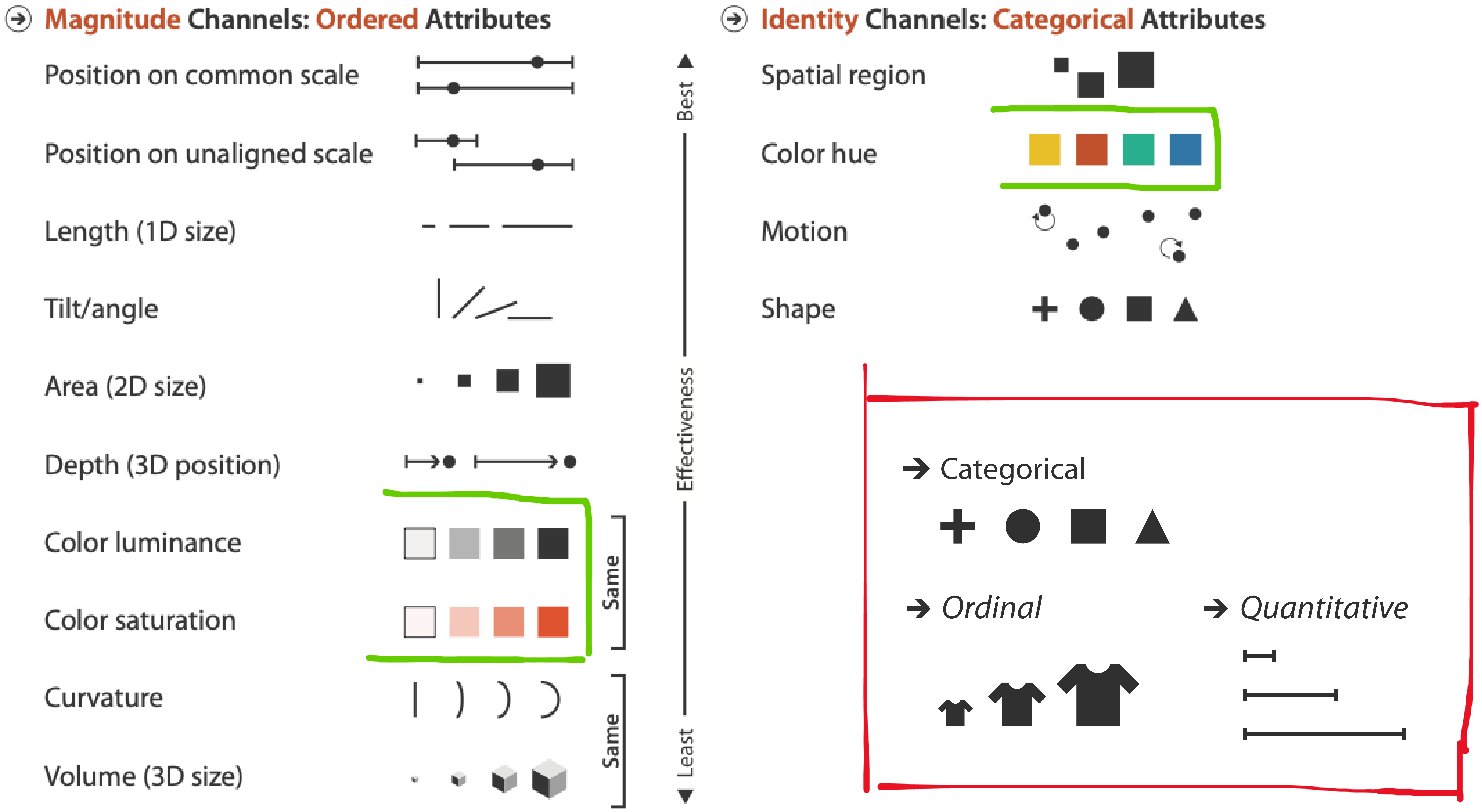
Escalas y paletas de Color

Colorear datos








Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Best
Effectiveness
Least
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape
Categorical
Ordinal
Quantitative

Colorear datos -> Color palette/scheme/scale

Categórica/Cualitativa

Ordering Direction

Sequential

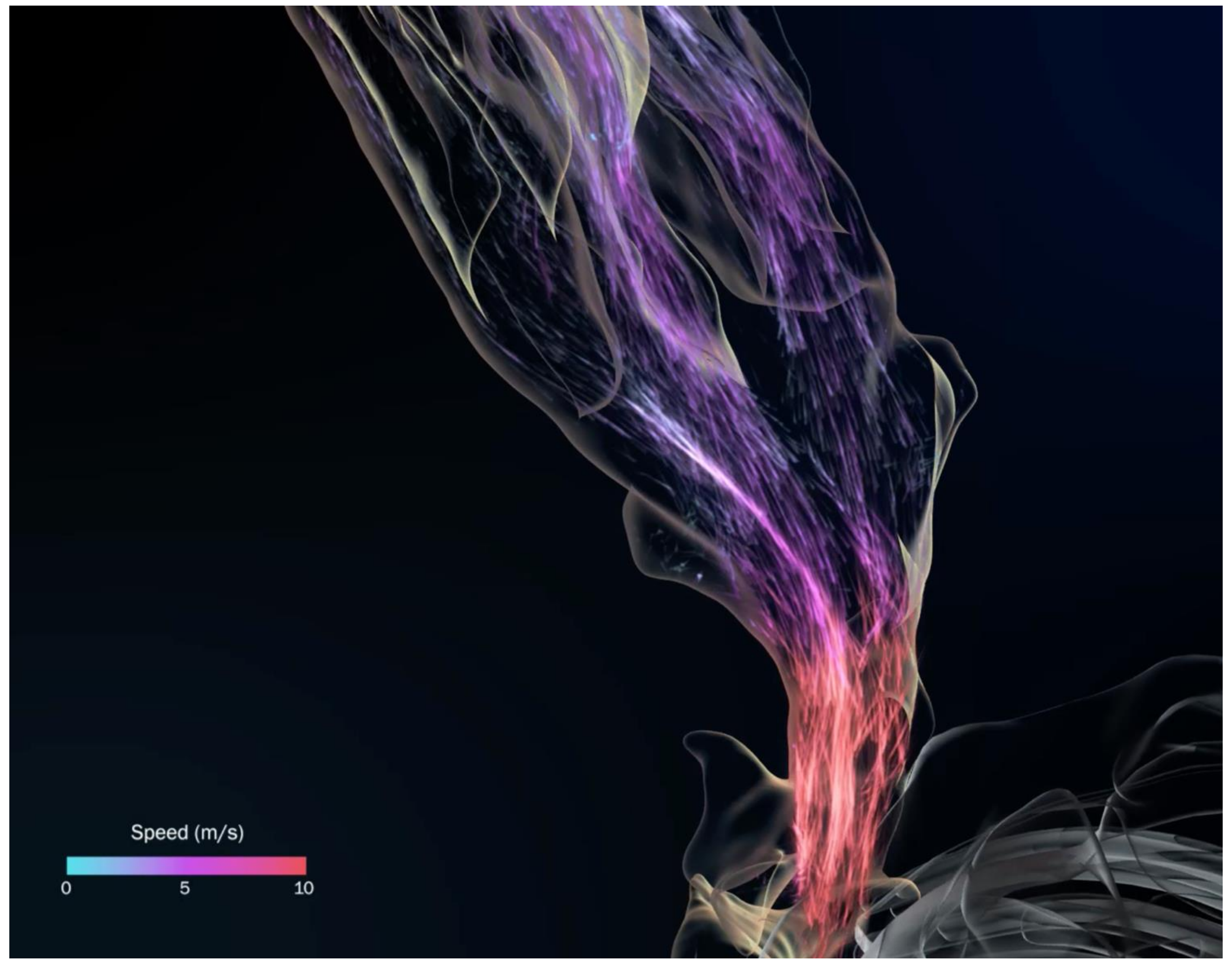
Diverging

Secuencial

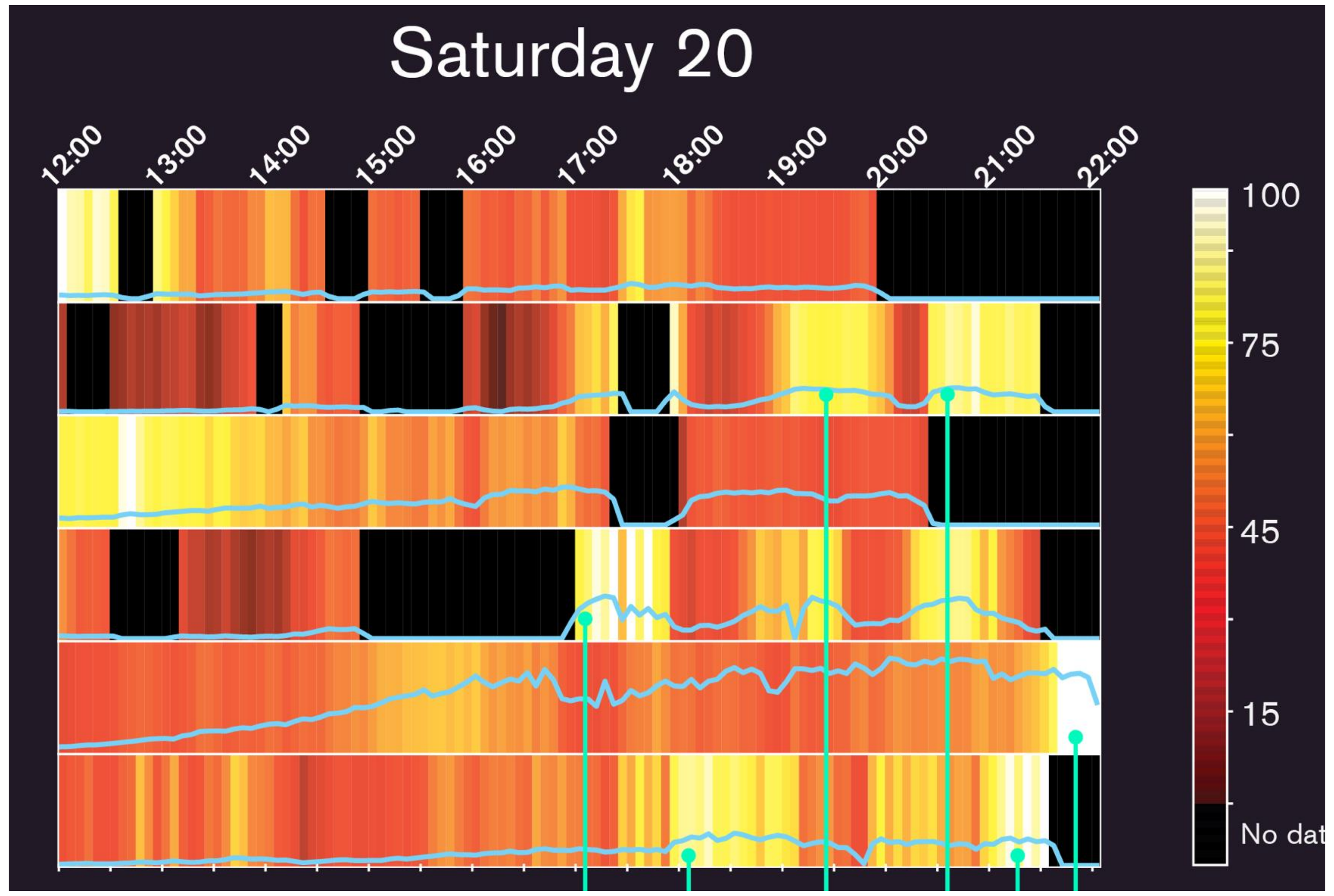
Divergente

Cuantitativo Secuencial

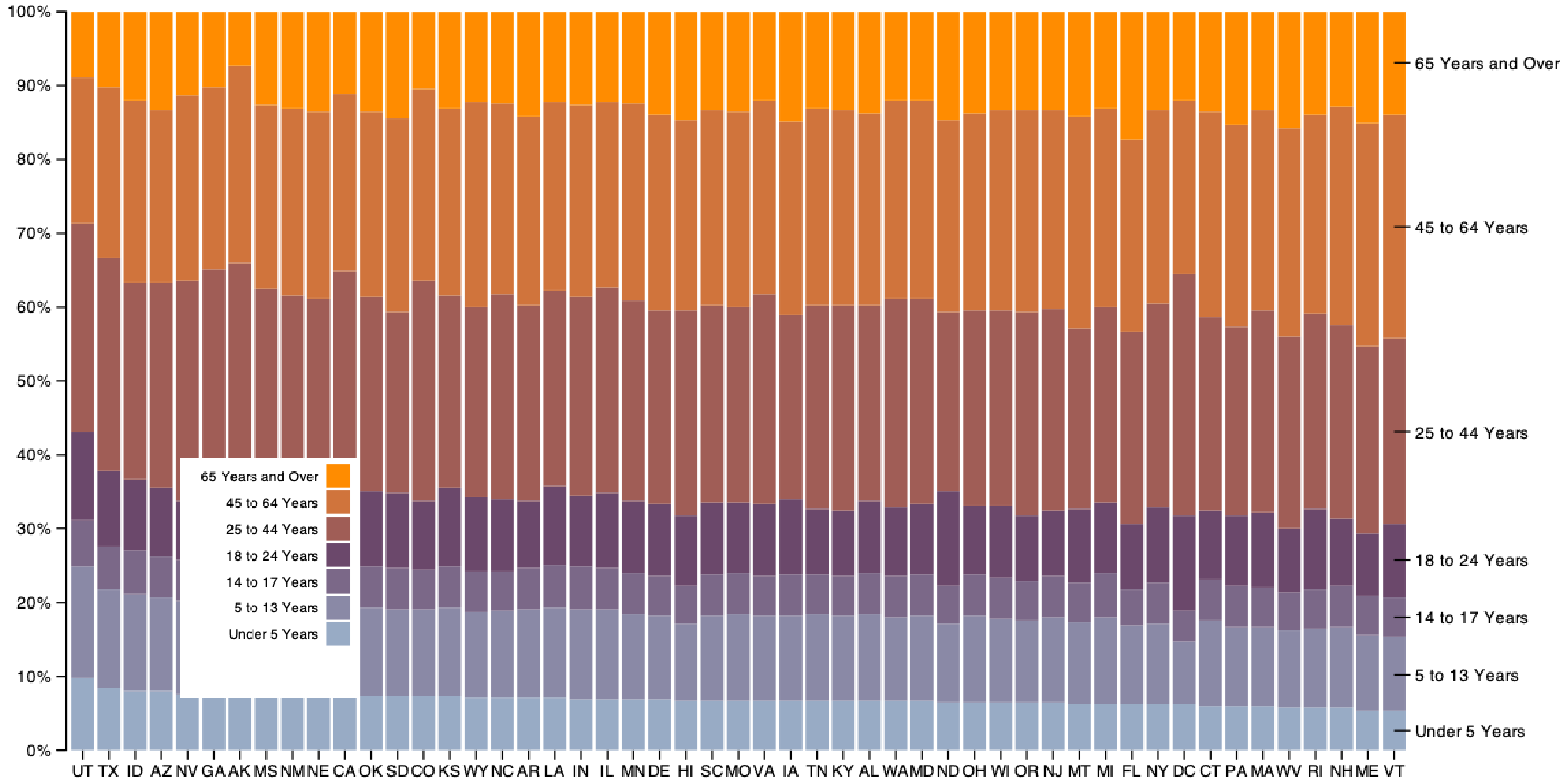
Continua



Discretizada

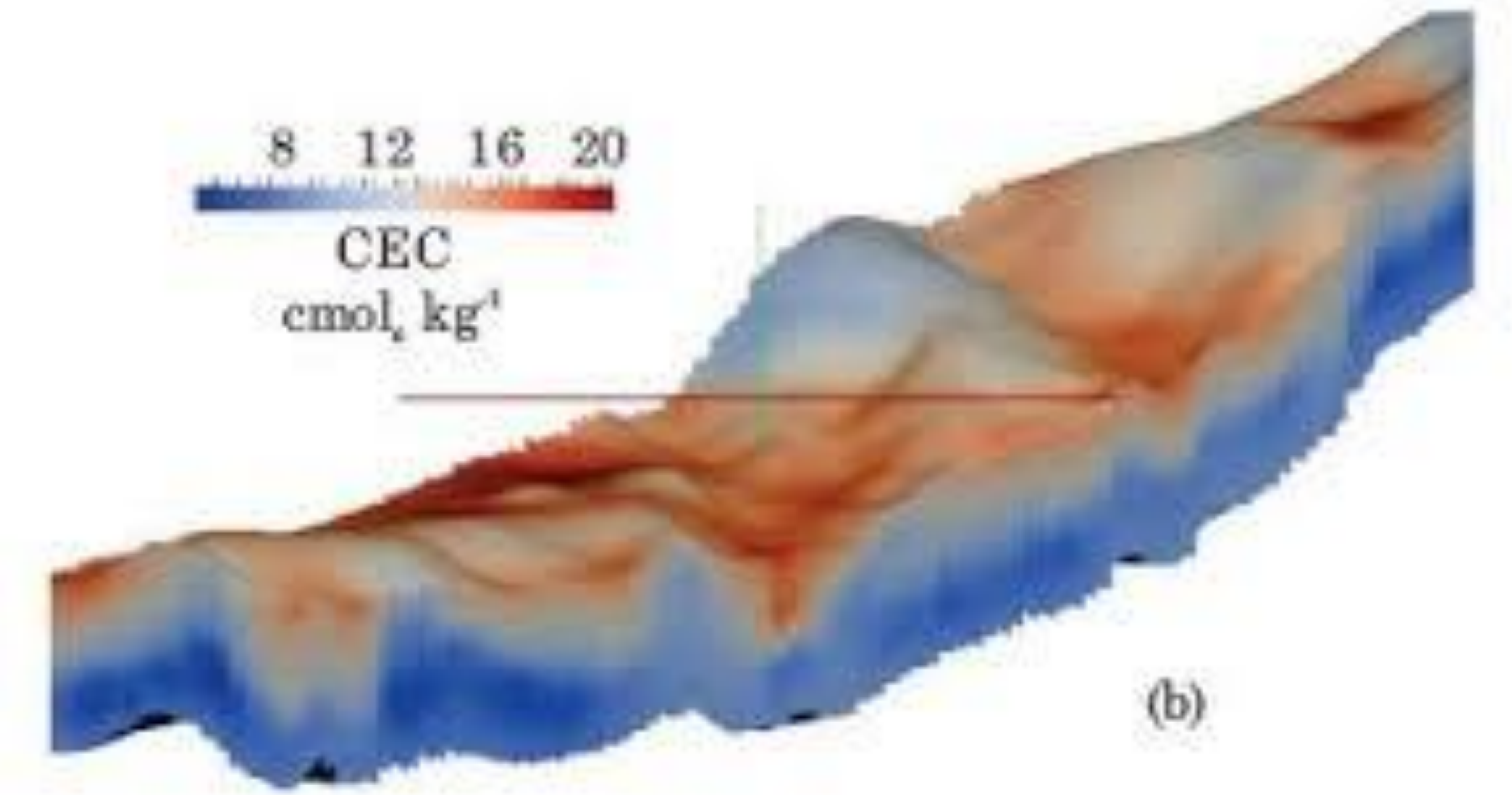


Ordinal = Continua discretizada

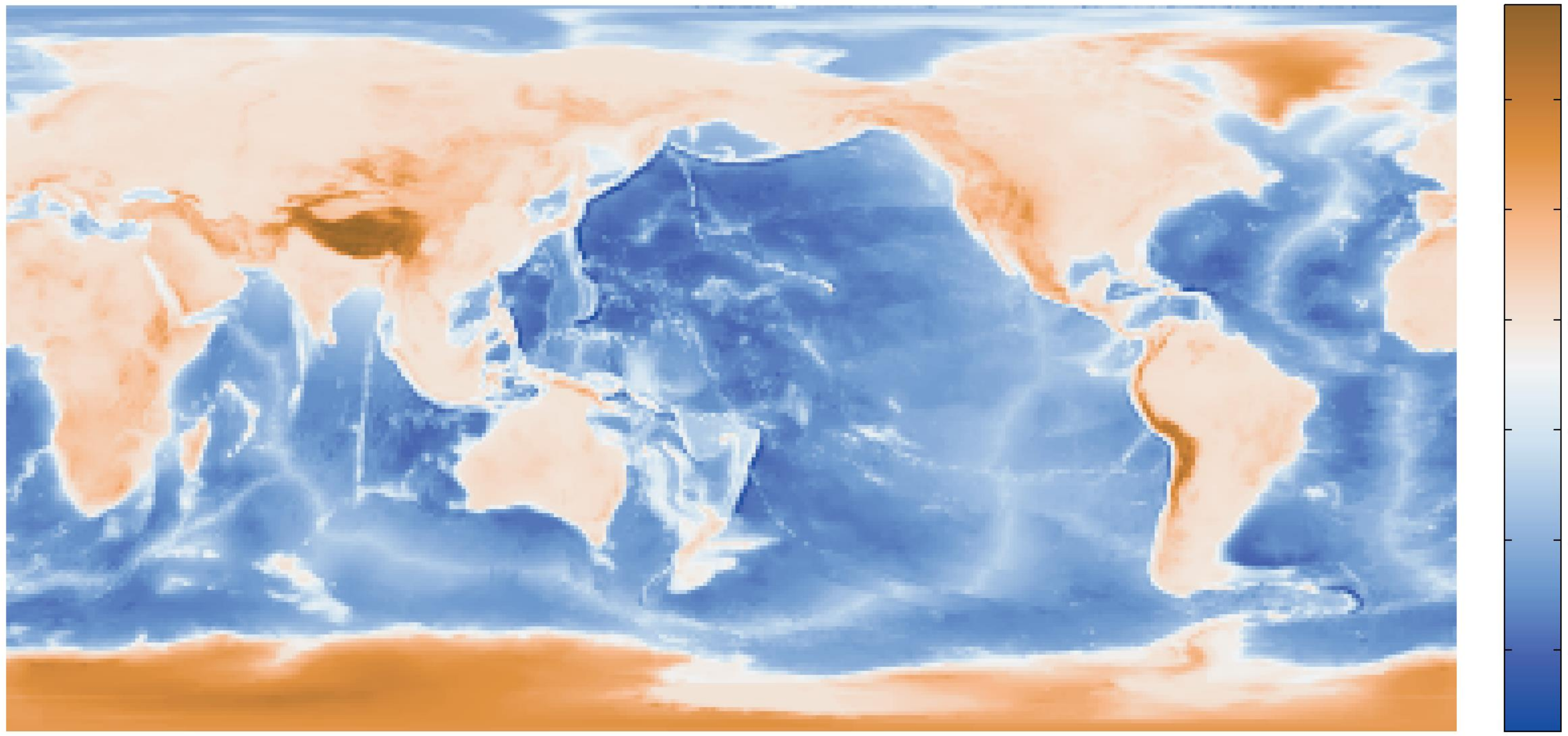


Cuantitativo Divergente

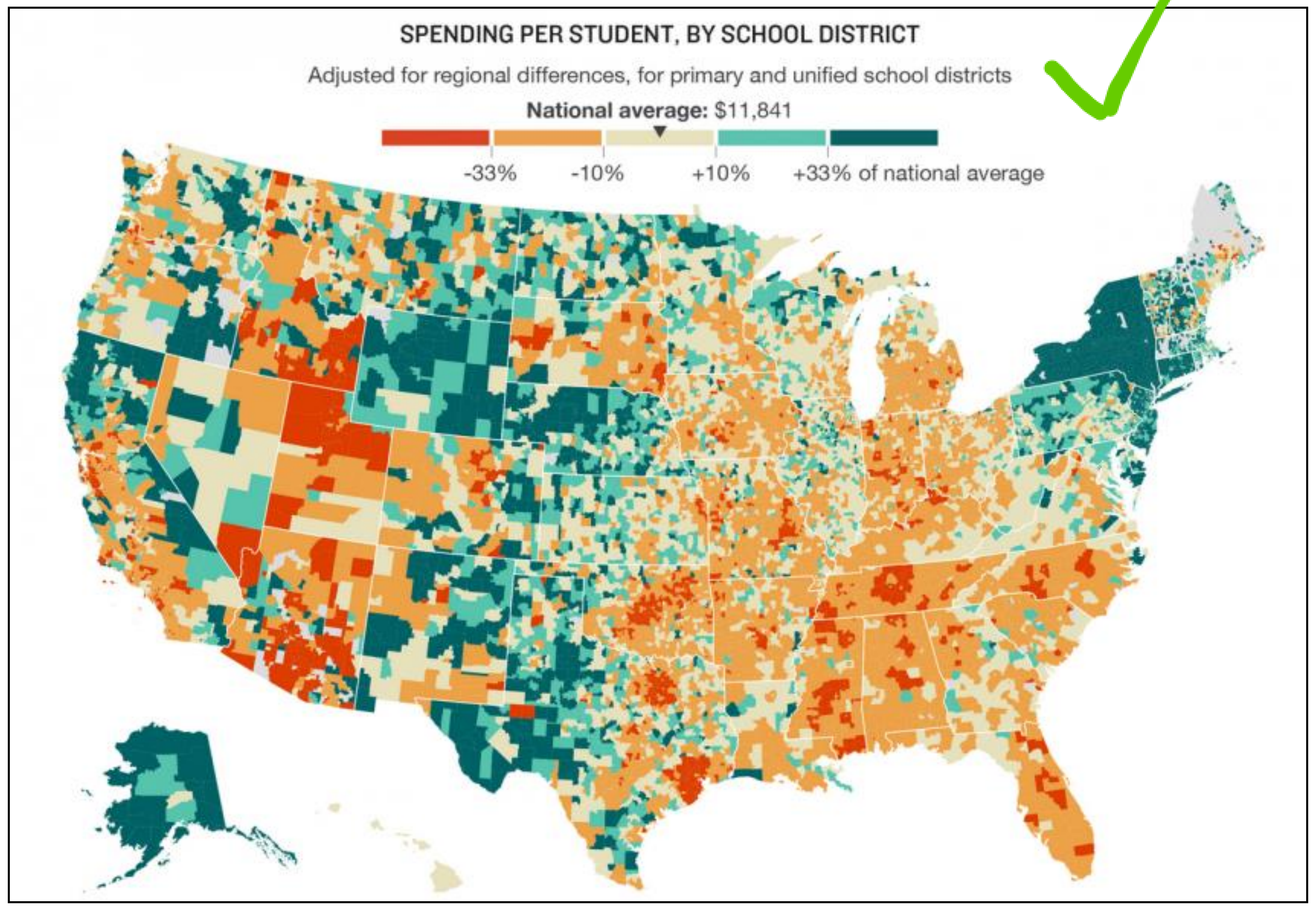
Es importante que el color central (normalmente blanco) represente el valor central en los datos (normalmente 0)



Continua



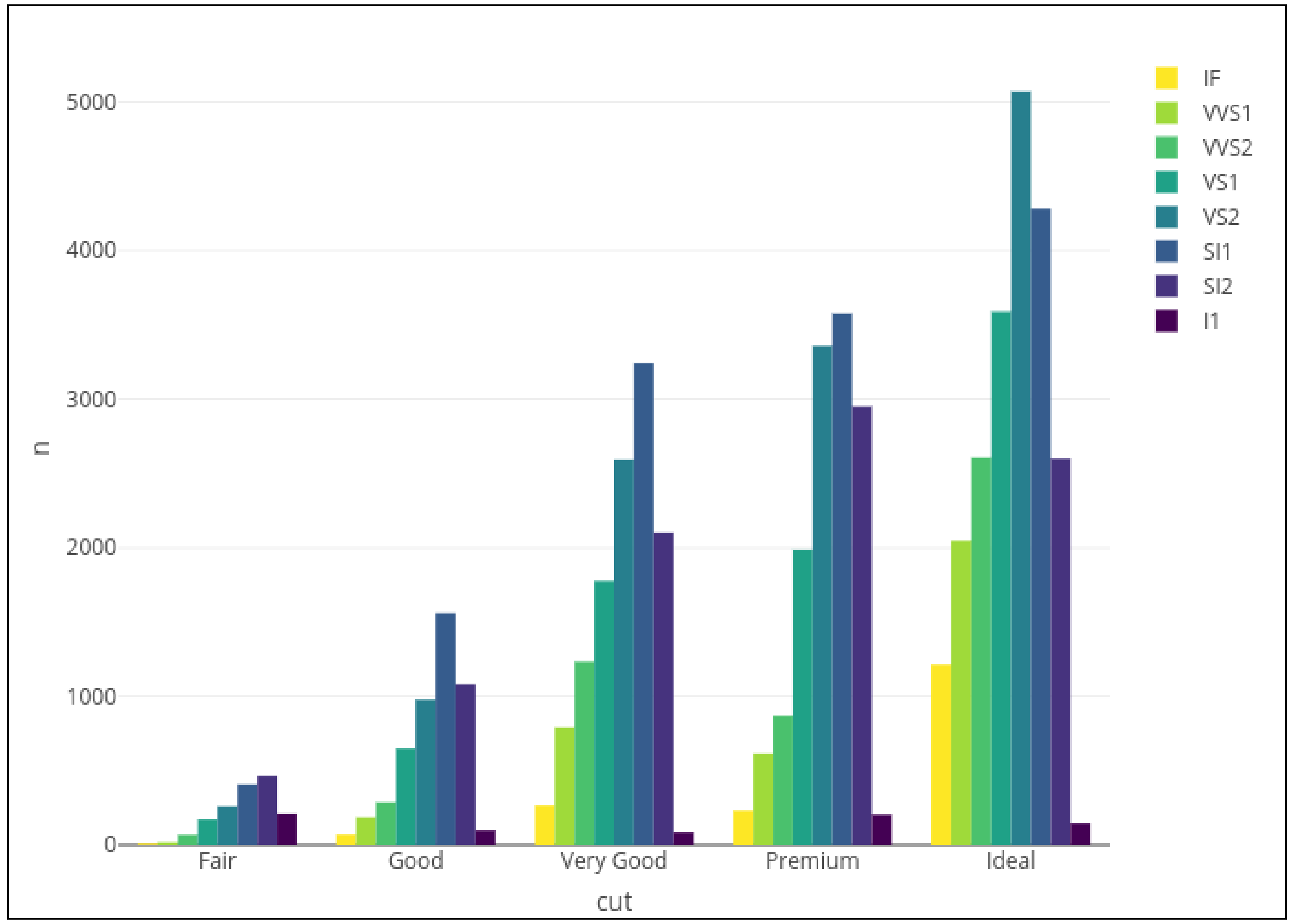
Discretizada

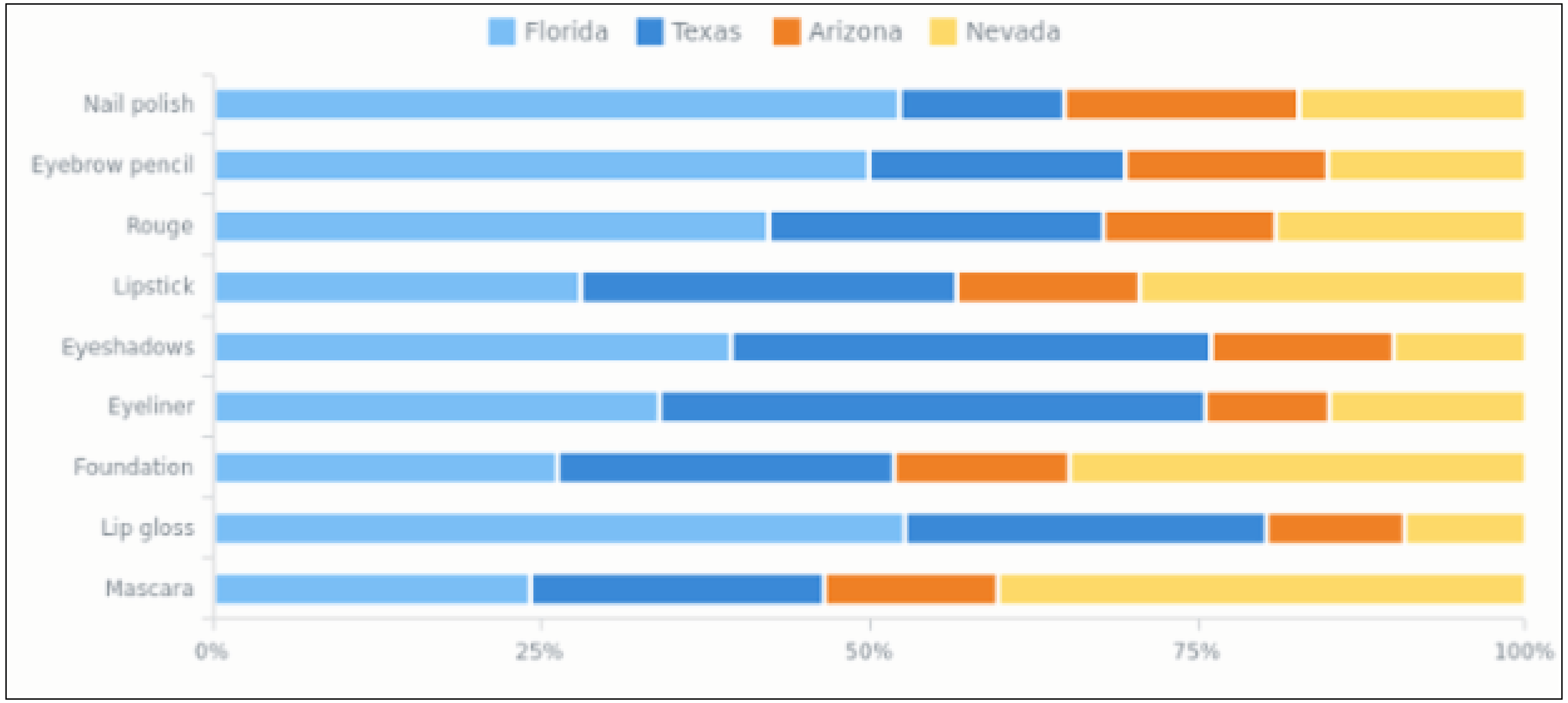


Categórica

Cuidado con colores que formen un gradiente continuo, pueden confundirse con valores ordinales

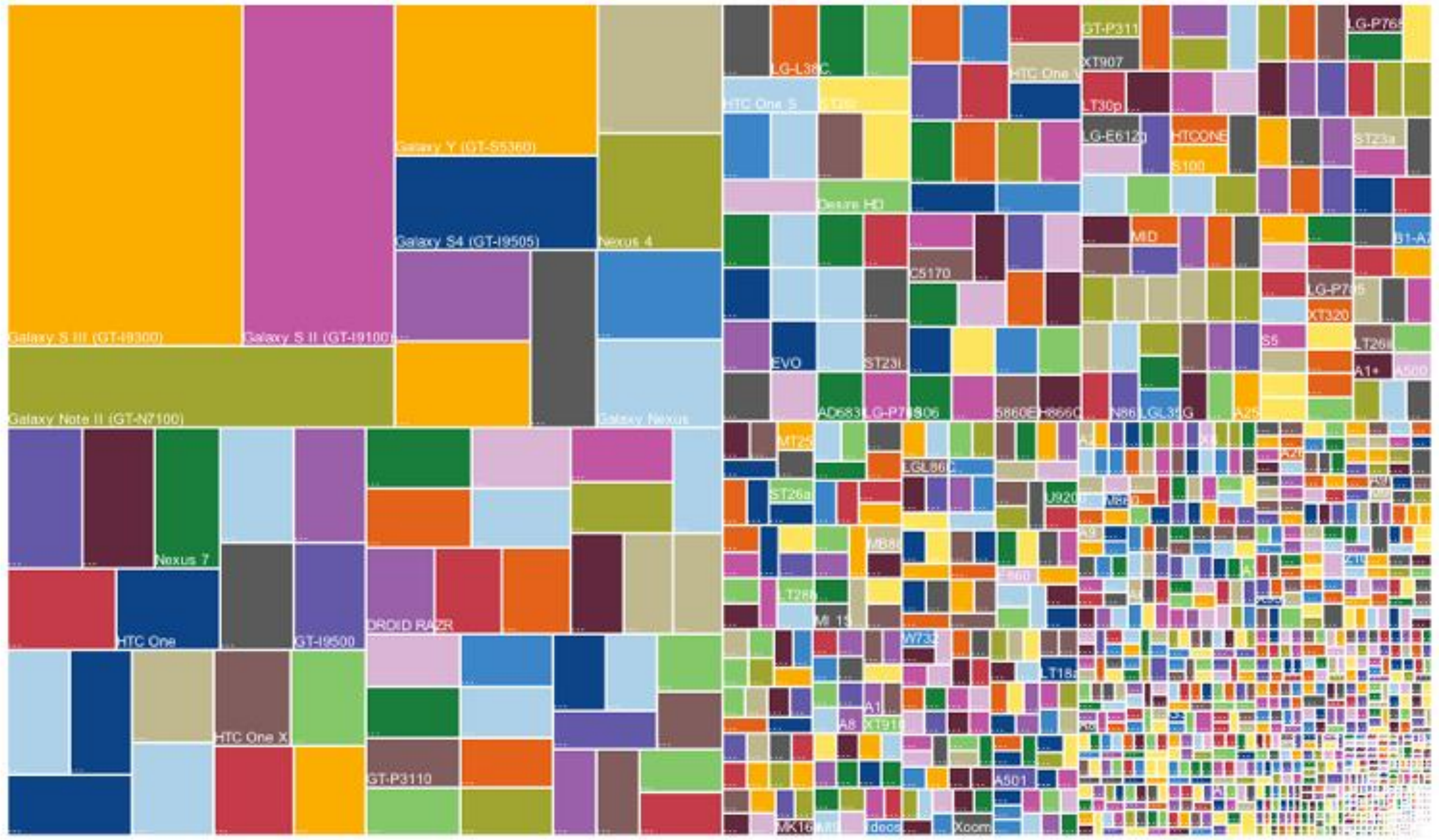
Casi siempre es mejor diferencias grandes en tono, saturación, y/o luminosidad

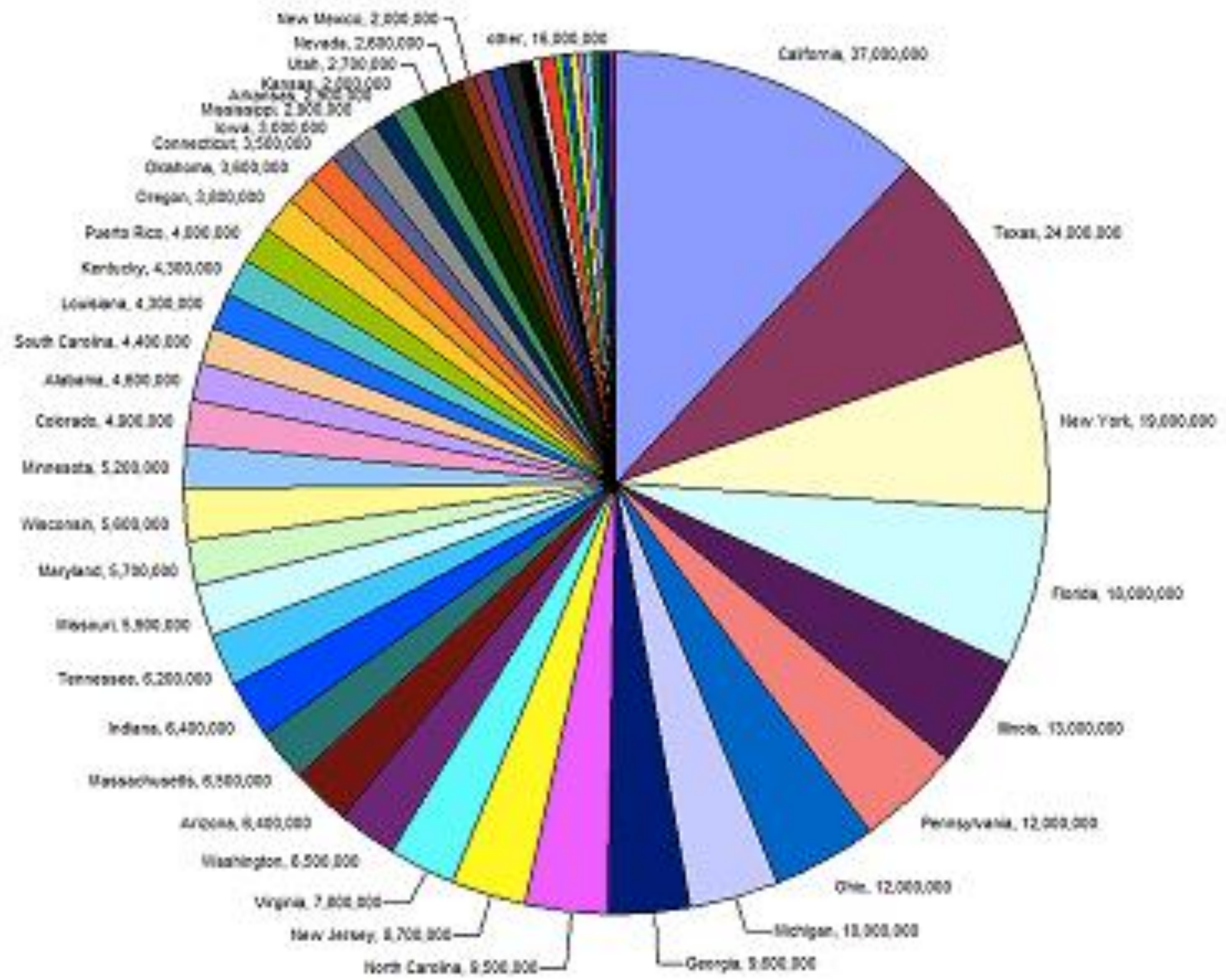


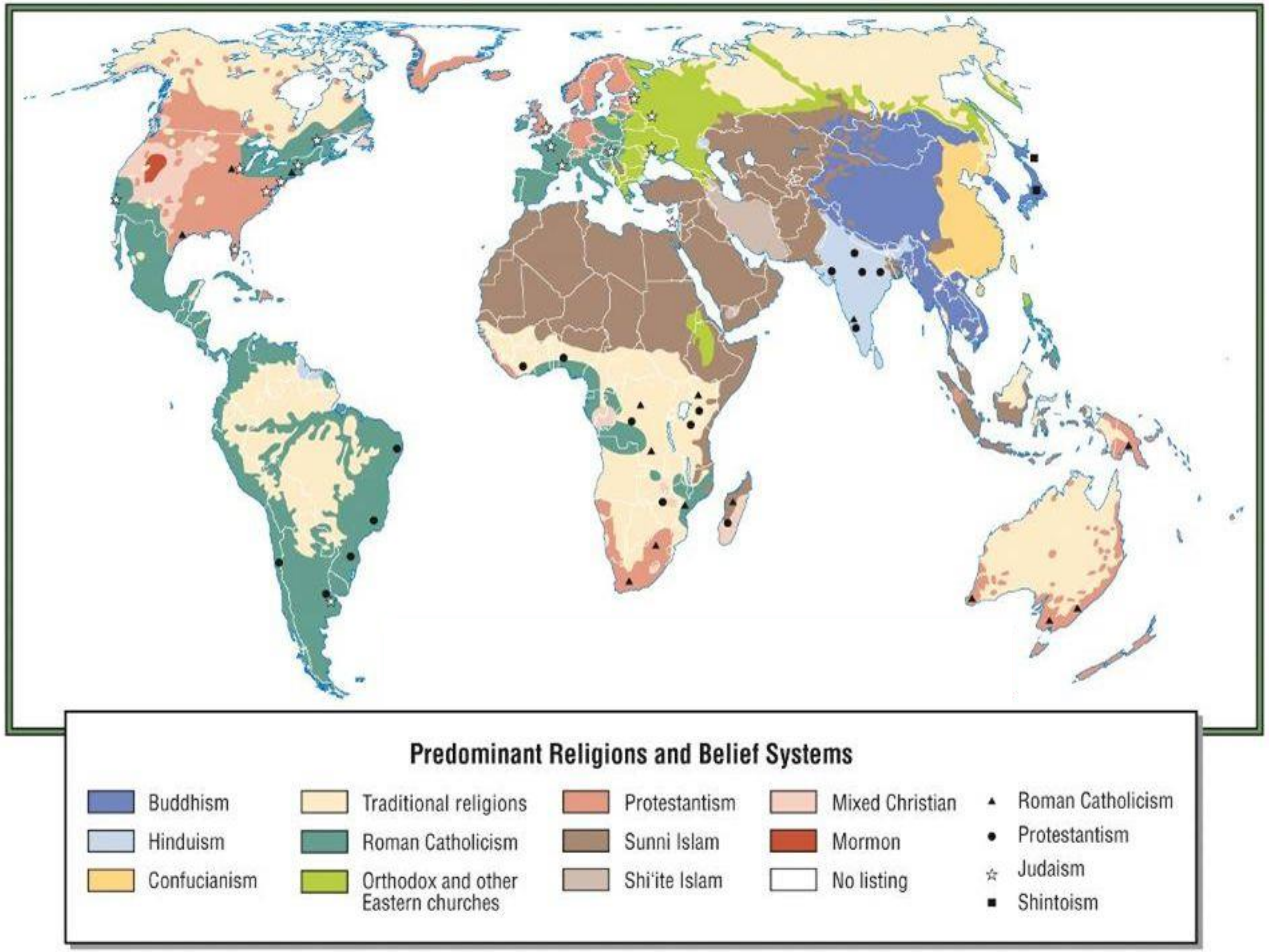


Categórica

No demasiados colores (max. 12)

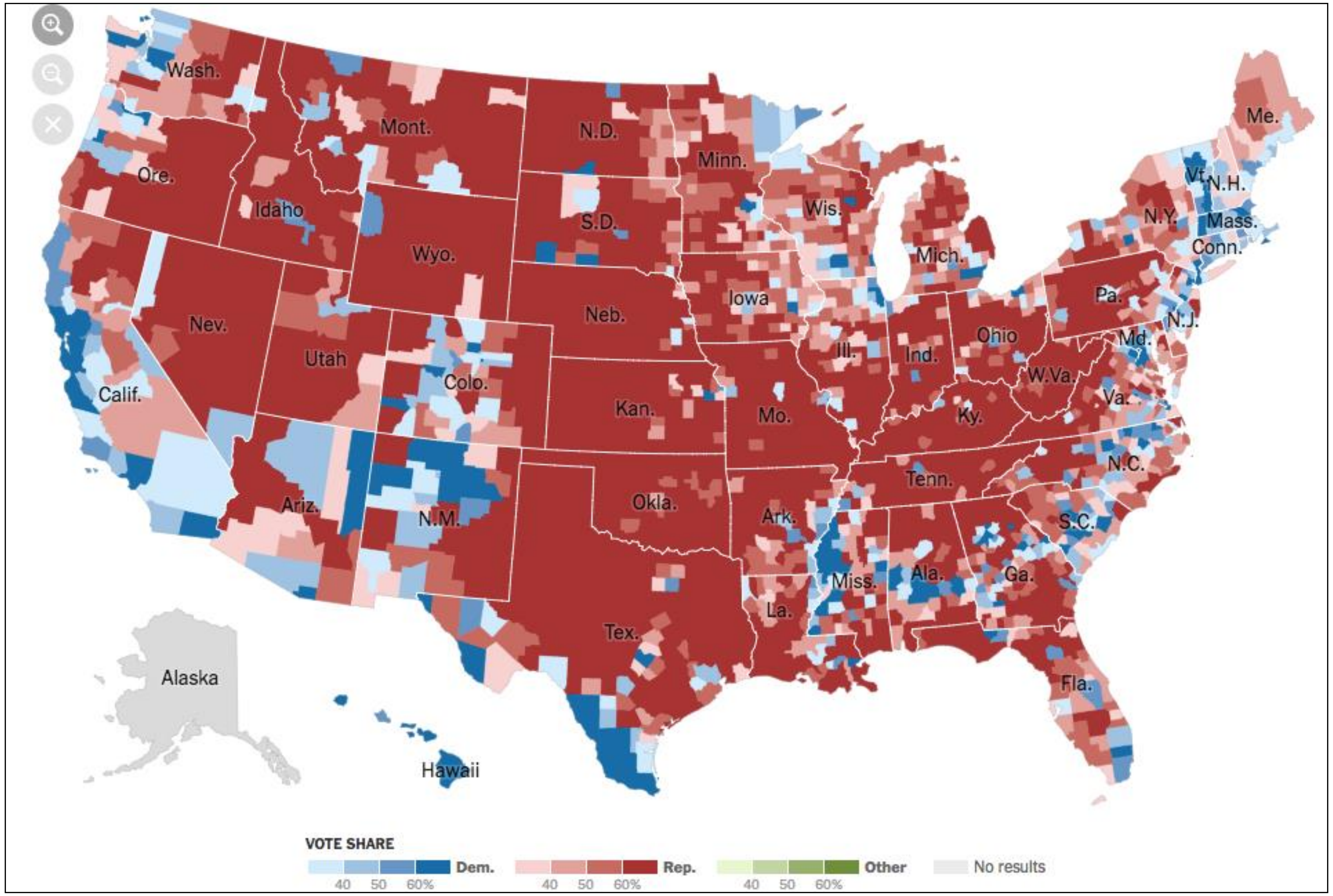






Agrupado

Esquema de color agrupado de 4 categorías (4 tonos), con 4 pasos de saturación y luminosidad cada uno

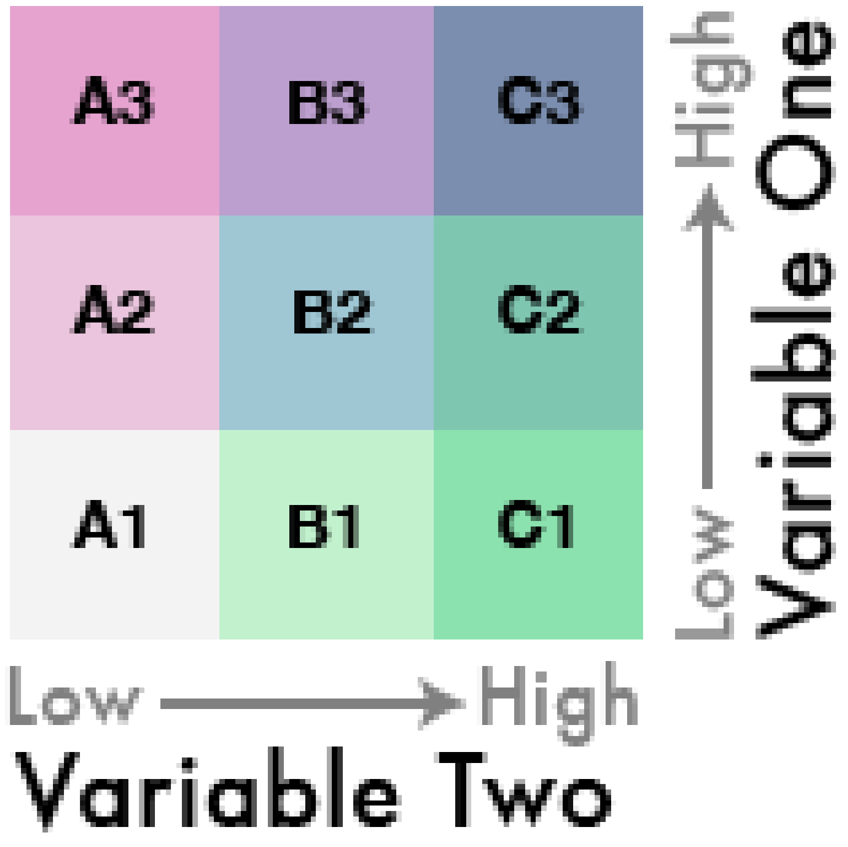


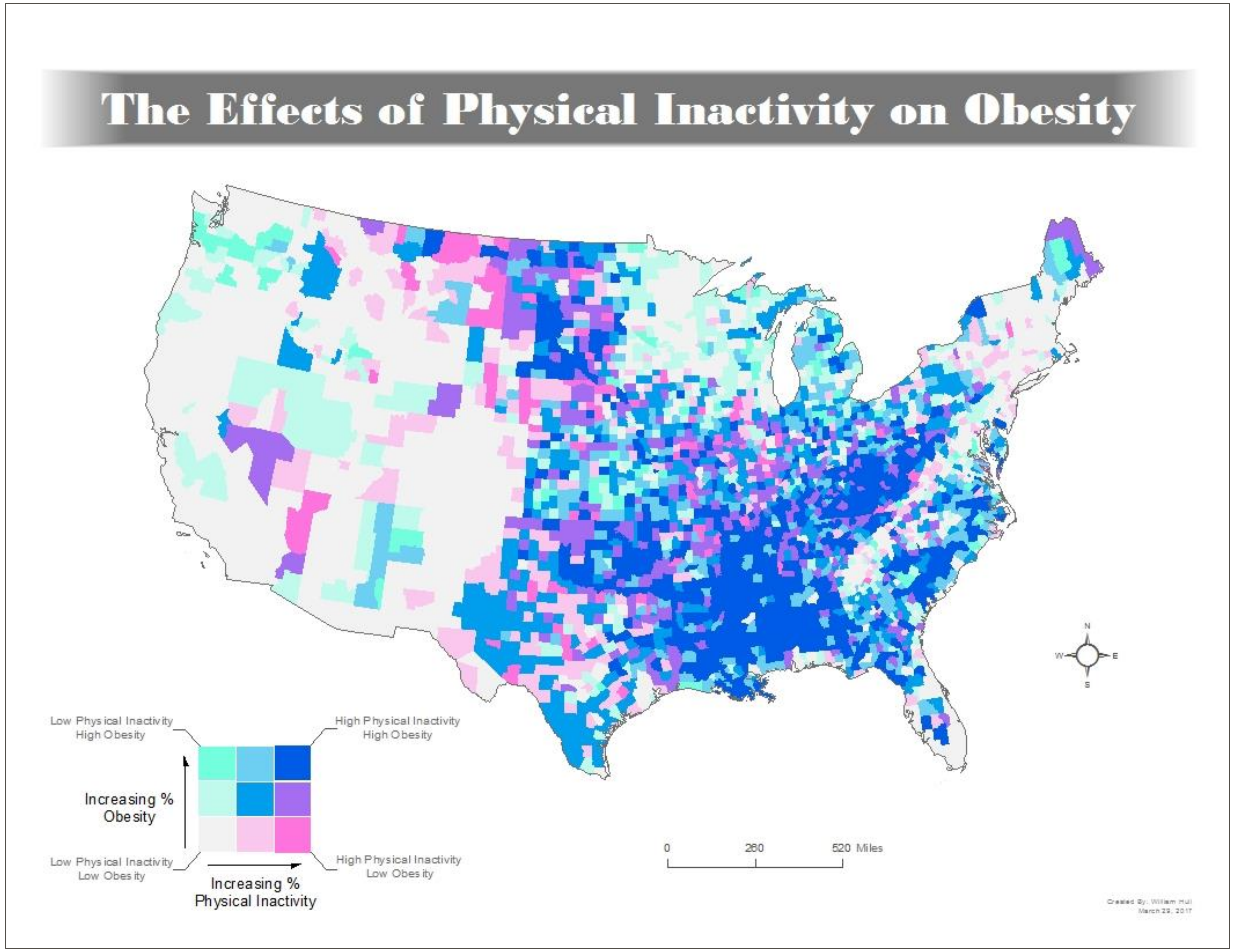
nytimes.com

William Hull

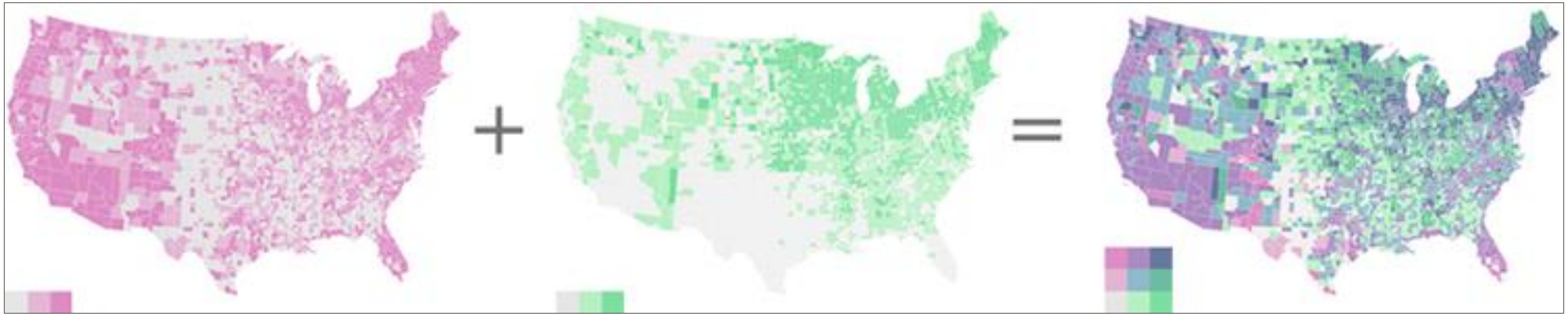
Bi-variate choropleth maps

Esquema de color con dos escalas sobrepuestas en dos direcciones.
Normalmente resulta una matriz de 3x3



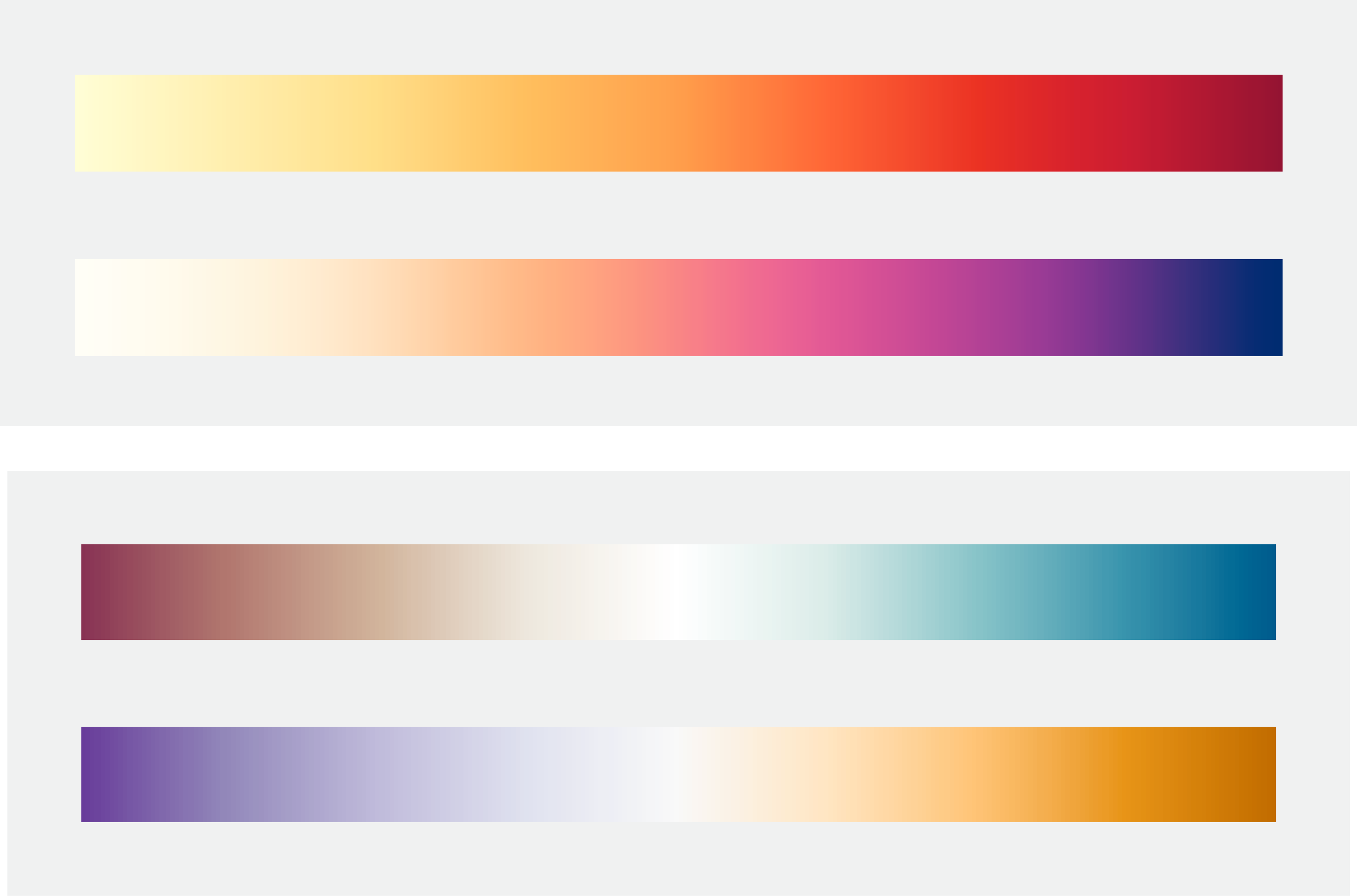


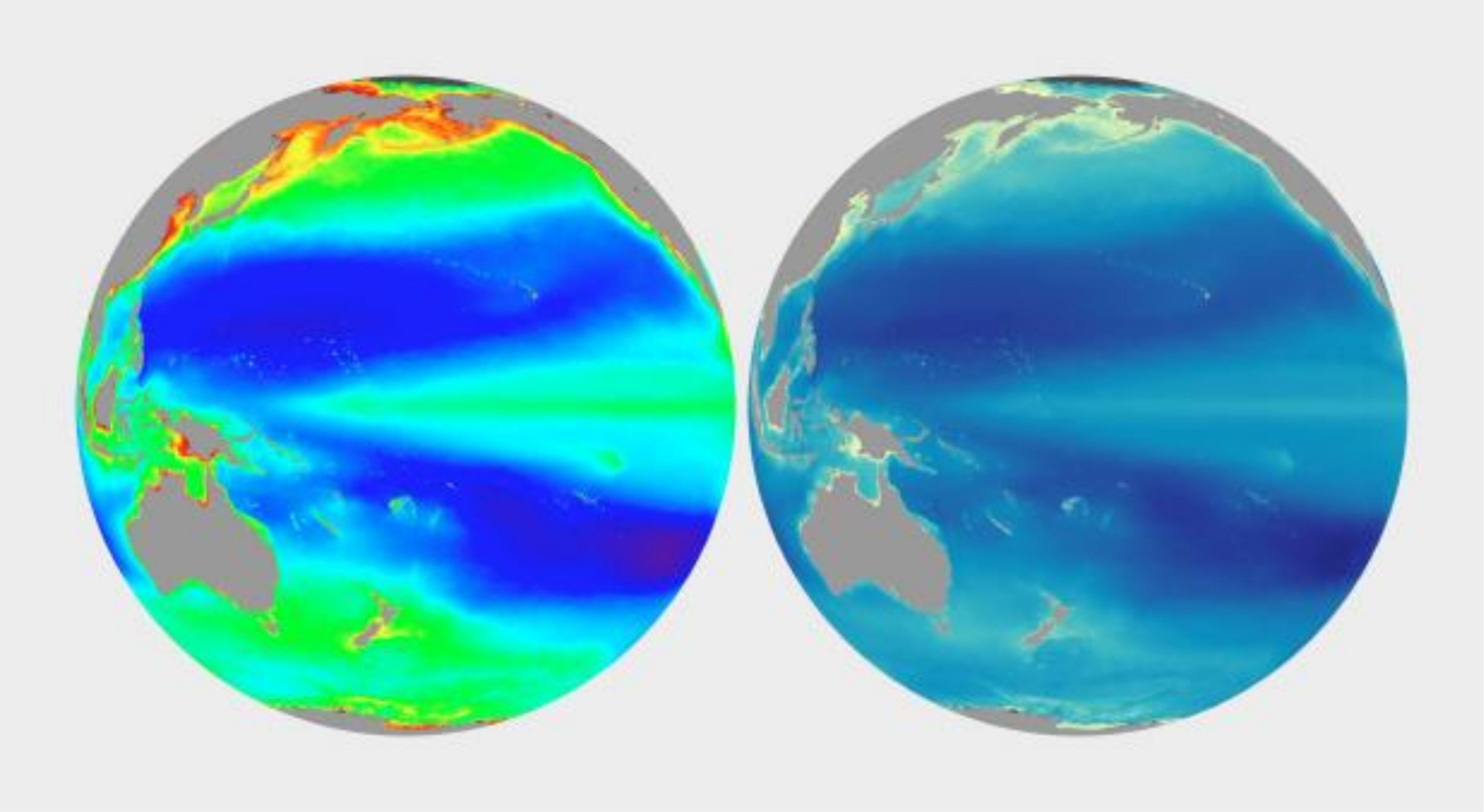
Joshua Stevens, from: www.joshuastevens.net



Perceptually linear scales

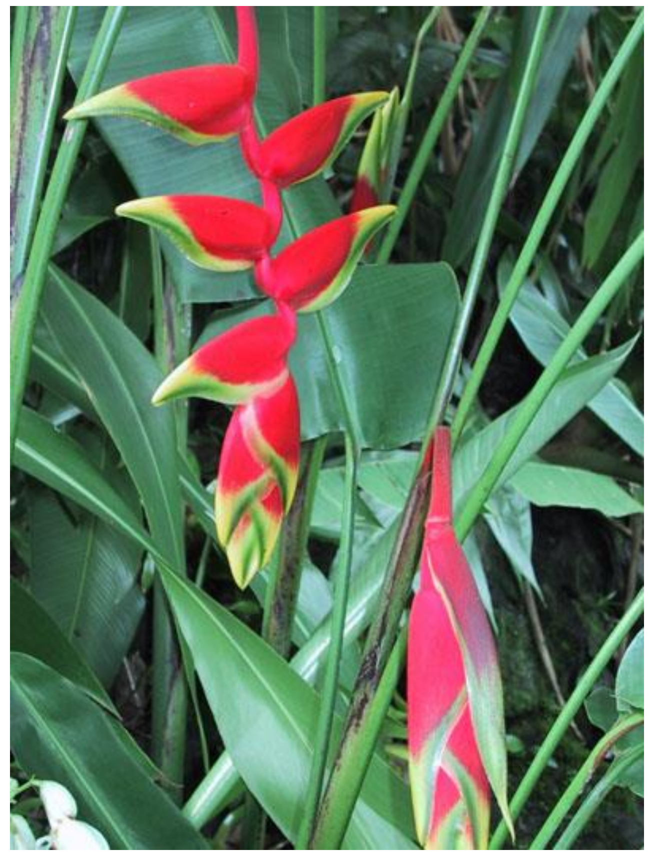
- Cuidado al usar presets
- Una escala de color adecuada debe variar de forma consistente a través del rango de valores
- Usar escalas **perceptualmente correctas**



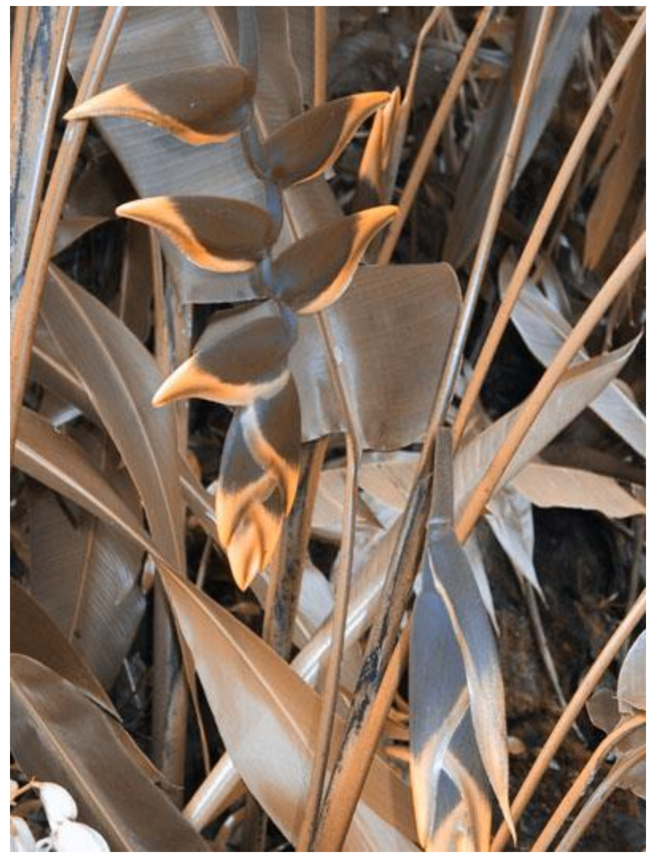


Simmon, 2013

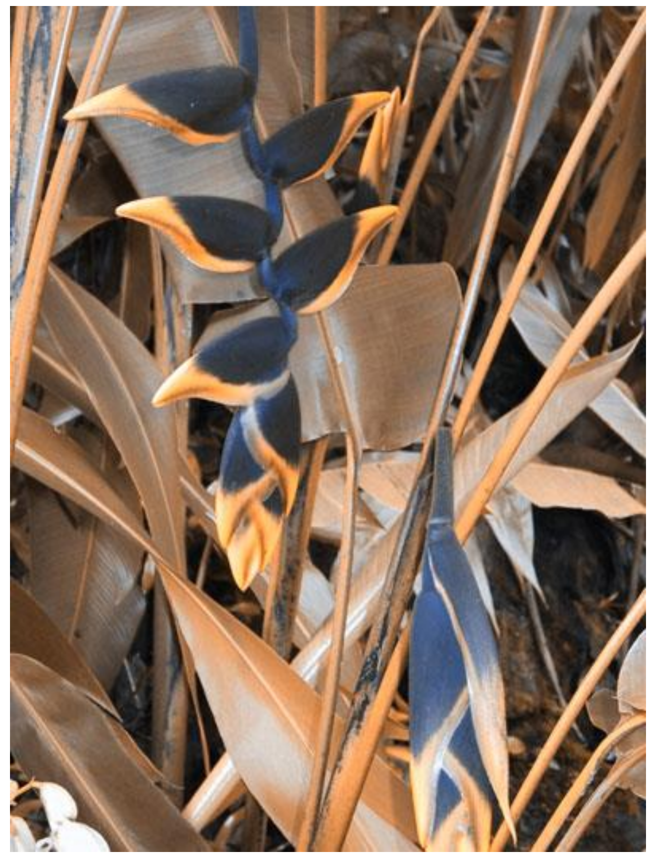
Diseñar para color deficiency: Usar simuladores



Normal vision



Deuteranope

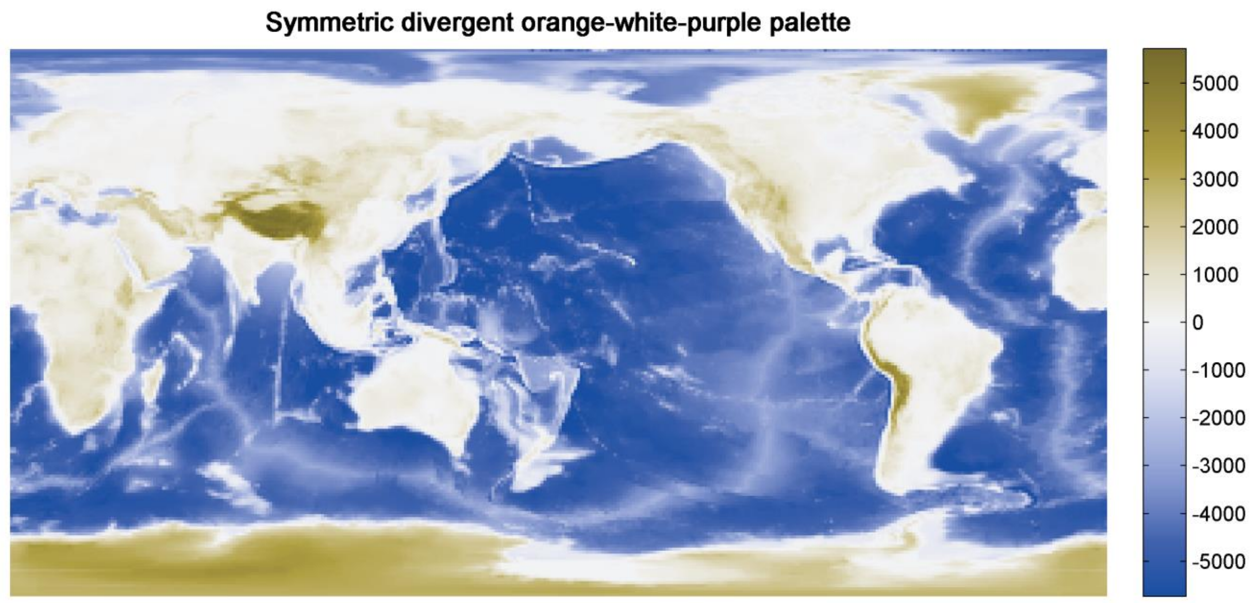


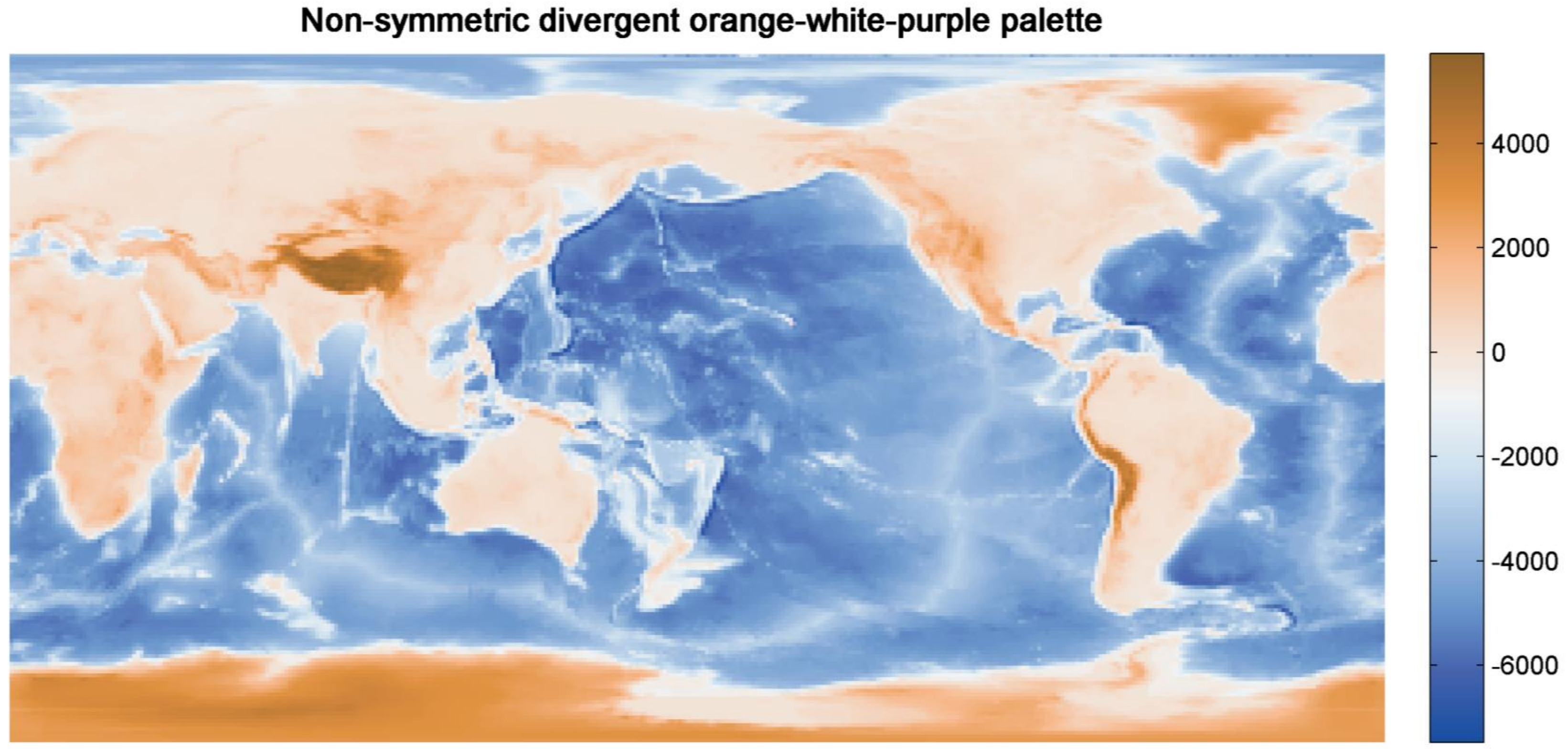
Protanope

Tritanope

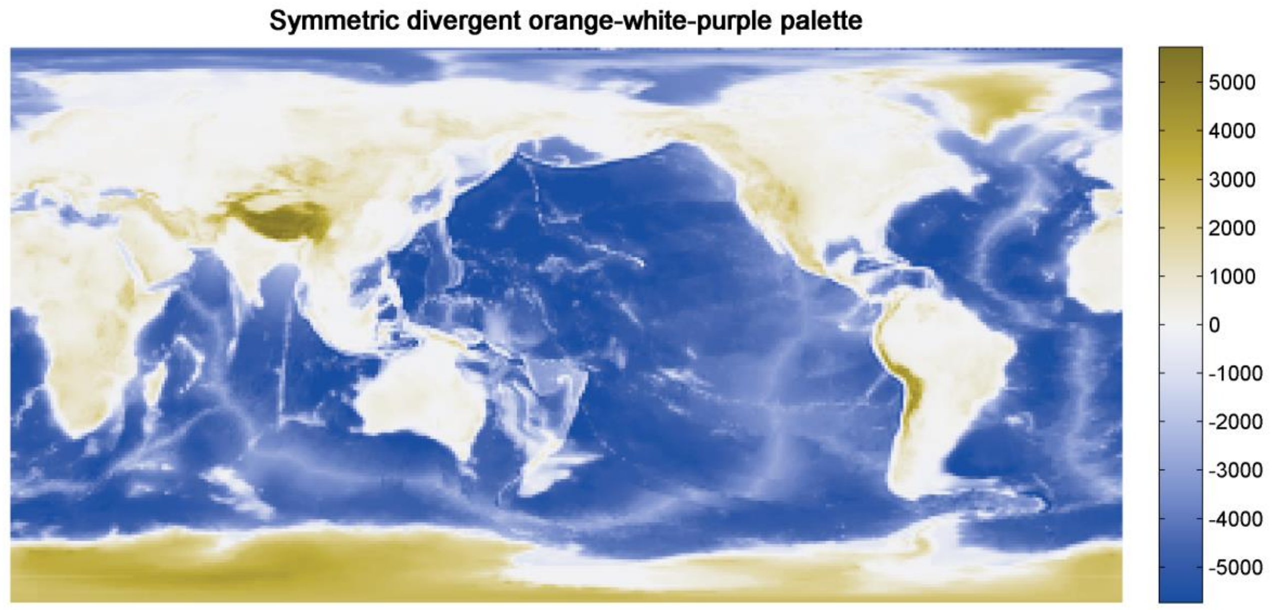
[Seriously Colorful: Advanced Color Principles & Practices. Stone.Tableau Customer Conference 2014.]

Protanope

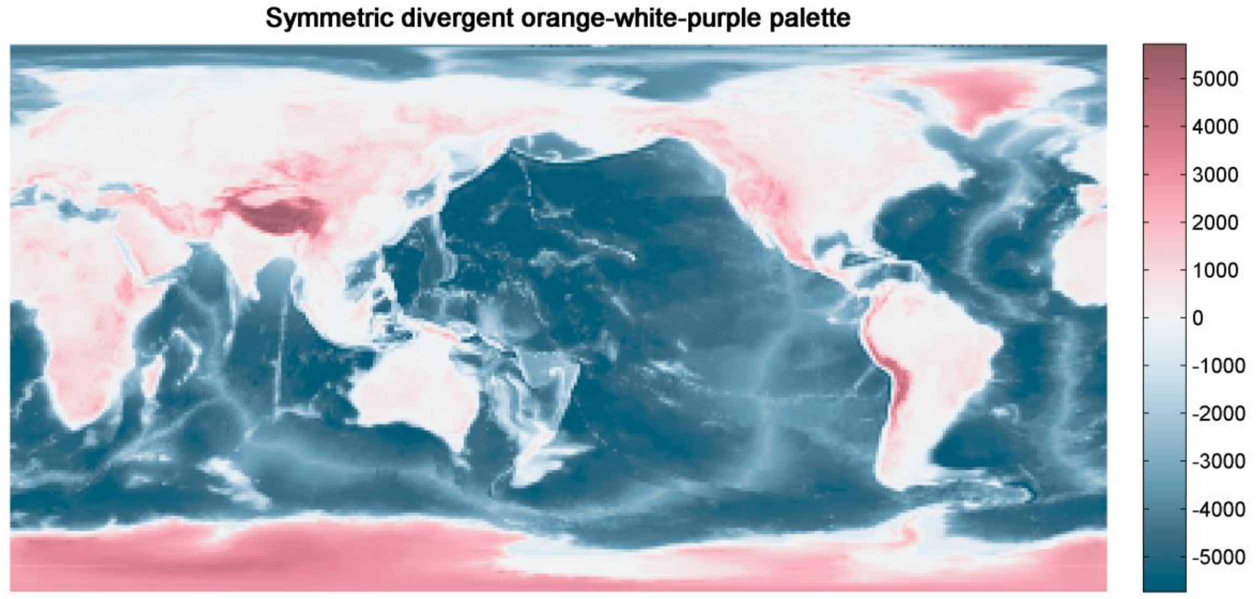




Deuteranope

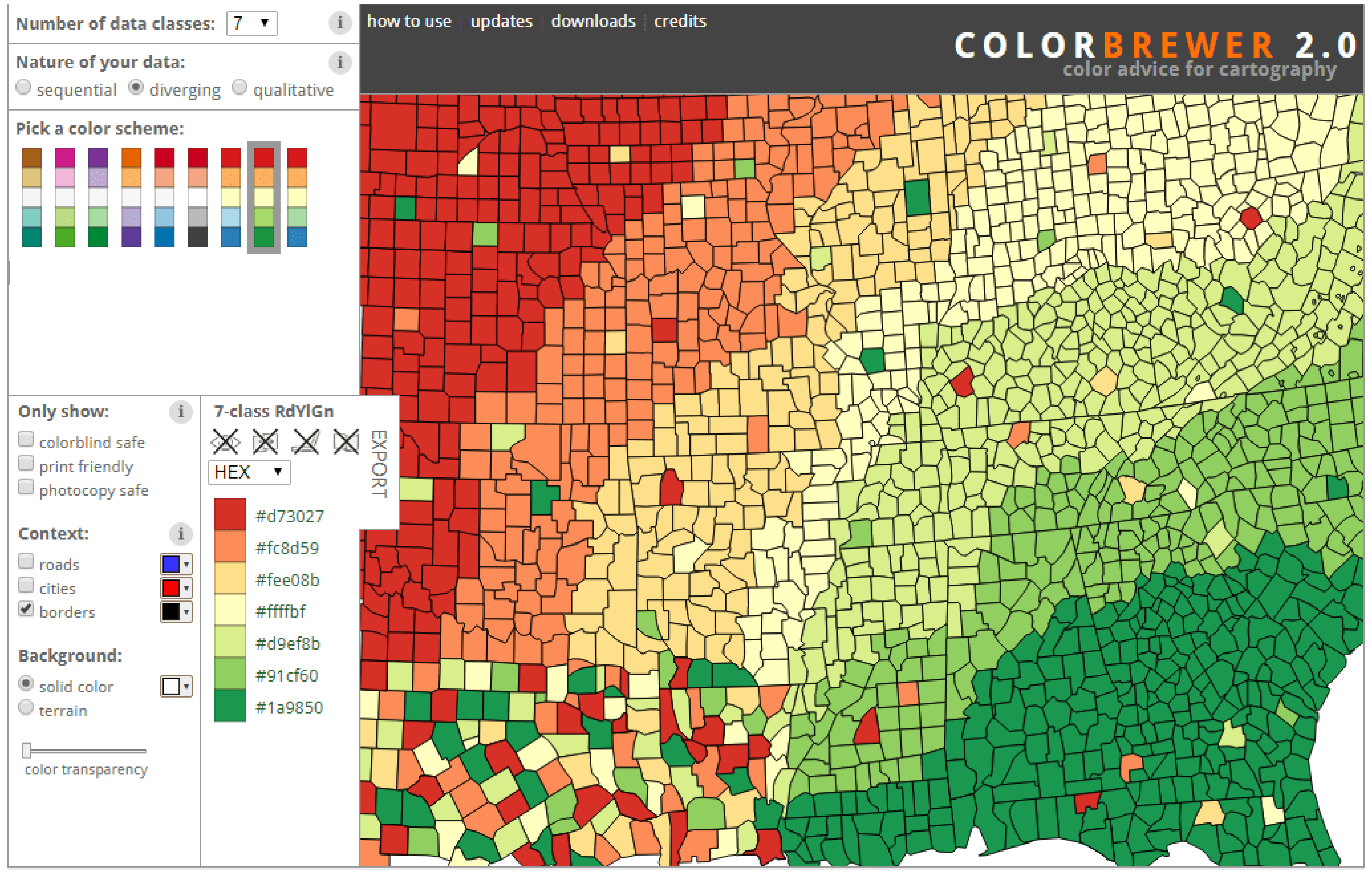


Tritanope



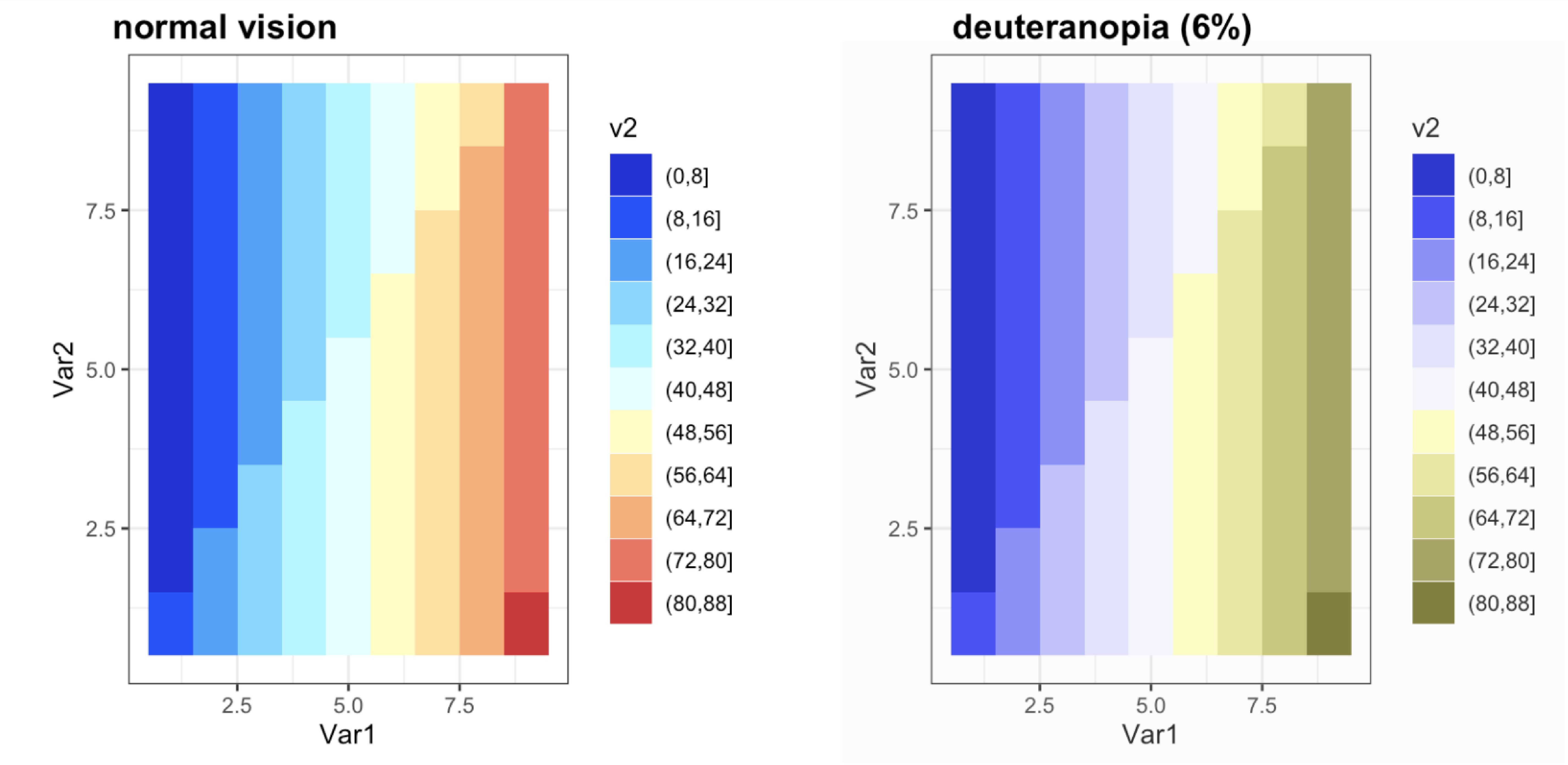
Simuladores de color deficiency

- Colorblindness para R
- Coblis www.color-blindness.com
- Color Brewer 2.0




```r
library(colorBlindness)
mat <- matrix(1:81, nrow = 9, ncol = 9)

library(ggplot2)
library(reshape2)
mat1 <- melt(t(mat[9:1, ]))
len <- length(Blue2DarkRed12Steps)-1
mat1$v2 <- cut(mat1$value,
              breaks = seq(0,ceiling(81/len)*len,
                           length.out = len+1))
ht <- ggplot(mat1) +
  geom_tile(aes(x=Var1, y=Var2, fill=v2)) +
  scale_fill_manual(values=Blue2DarkRed12Steps) +
  theme_bw()
# check the plot by CVD simulator
cvdPlot(ht)
```



Escalas perceptualmente correctas

Viridis. Perceptualmente correctas; corregidas para color deficiency y B/N

https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

Buscar "Perceptually correct"

Viridis color scales (R, plotly, etc.)

The package contains eight color scales: "viridis", the primary choice, and five alternatives with similar properties - "magma", "plasma", "inferno", "civids", "mako", and "rocket" -, and a rainbow color map - "turbo".

The color maps `viridis`, `magma`, `inferno`, and `plasma` were created by Stéfan van der Walt ([@stefanv] (https://github.com/stefanv)) and Nathaniel Smith ([@njsmith](https://github.com/njsmith)). If you want to know more about the science behind the creation of these color maps, you can watch this presentation of `viridis` by their authors at SciPy 2015.

The color map `cividis` is a corrected version of 'viridis', developed by Jamie R. Nuñez, Christopher R. Anderton, and Ryan S. Renslow, and originally ported to R by Marco Sciaini ([@msciain] (https://github.com/marcosci)). More info about `cividis` can be found in this paper.

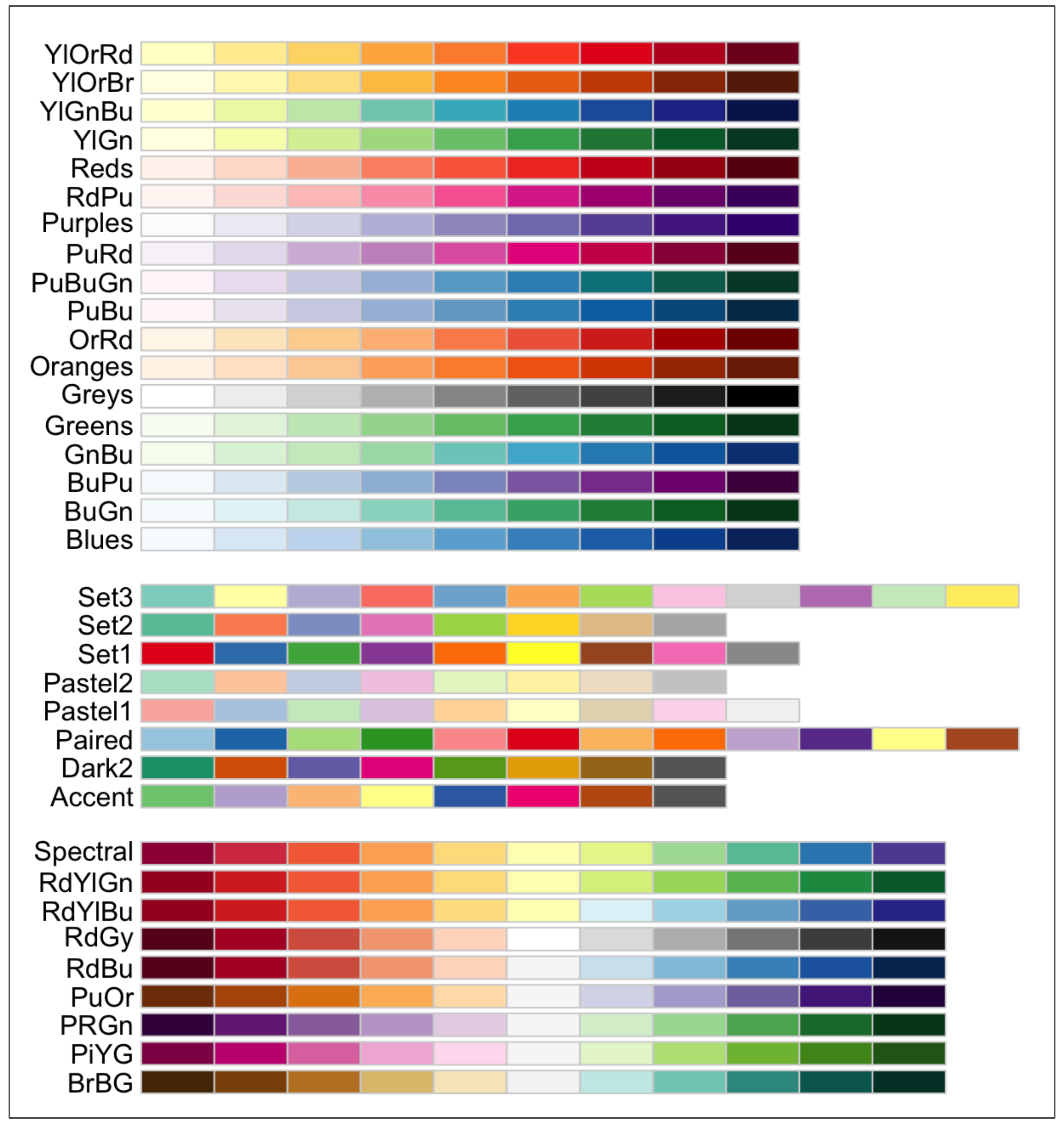
The color maps `mako` and `rocket` were originally created for the `Seaborn` statistical data visualization package for Python. More info about `mako` and `rocket` can be found on the `Seaborn` website.

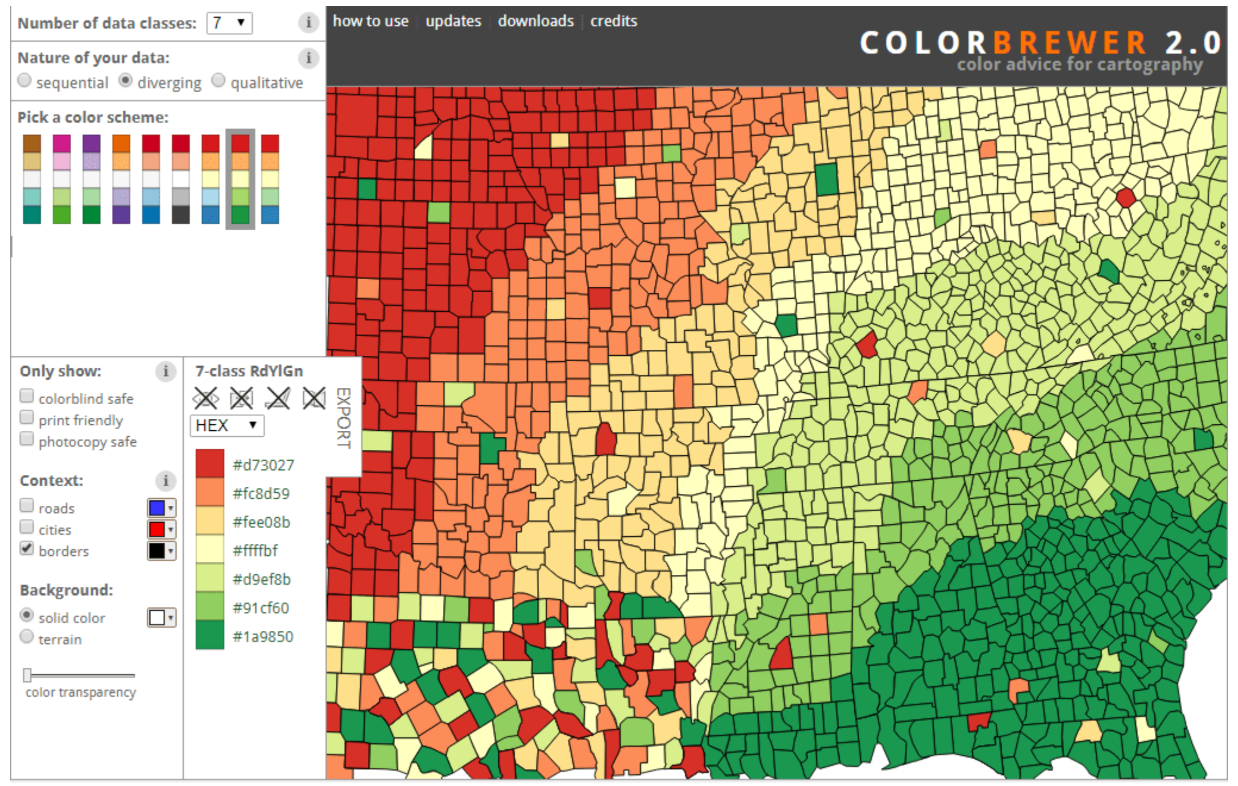
The color map `turbo` was developed by Anton Mikhailov to address the shortcomings of the Jet rainbow color map such as false detail, banding and color blindness ambiguity. More infor about `turbo` can be found here.

R- colorspace

```
library("colorspace")
hcl_palettes(plot = TRUE)
```

Color Brewer





R cheatsheet

- No se trata solo de estética
- Usar esquemas de color que funcionen sobre sistemas basados en percepción:HSV, HSL, CIEL
- Accesibilidad: tener en cuenta color-blindness
- Elegir paletas de color en función del **tipo de datos** (categórico, cuantitativo, ordinal // secuencial, divergente, cíclico)

R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggbb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0
i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

R has 657 built in color names

To see a list of names:
colors()
These colors are displayed on P. 3.

Example:
peachpuff4

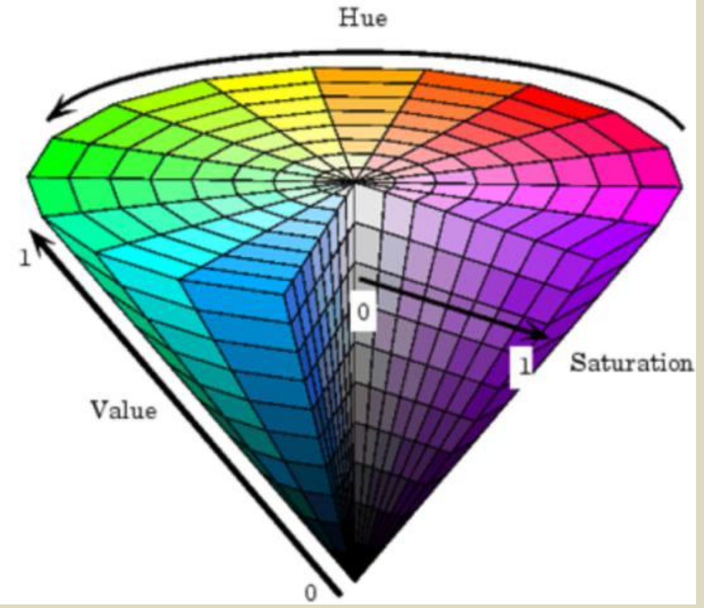
R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.
  rgb(r, g, b, maxColorValue=255, alpha=255)
- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument
  hsv(h, s, v, alpha)
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
  hcl(h, c, l, alpha)

A few notes on HSV/HLC

HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model....blah blah blah...

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.



R can translate colors to rgb (this is handy for matching colors in other programs)
col2rgb(c("#FF0000", "blue"))

R Color Palettes

This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure....NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps

grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

grDevices palettes
cm.colors
topo.colors
terrain.colors
heat.colors
rainbow
see P. 4 for options

heat.colors(4, alpha=1)
> #FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFFF80FF"

For the rainbow palette you can also select start/end color (red = 0, yellow = *1/6*, green = *2/6*, cyan = *3/6*, blue = *4/6* and magenta = *5/6*) and saturation (s) and value (v):
rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)

Package: RcolorBrewer

This function has an argument for the number of colors and the color palette (see P. 4 for options).
brewer.pal(4, "Set3")
> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: display.brewer.all(5)
There is also a very nice interactive viewer:
http://colorbrewer2.org/

My Recommendation

Package: colorspace

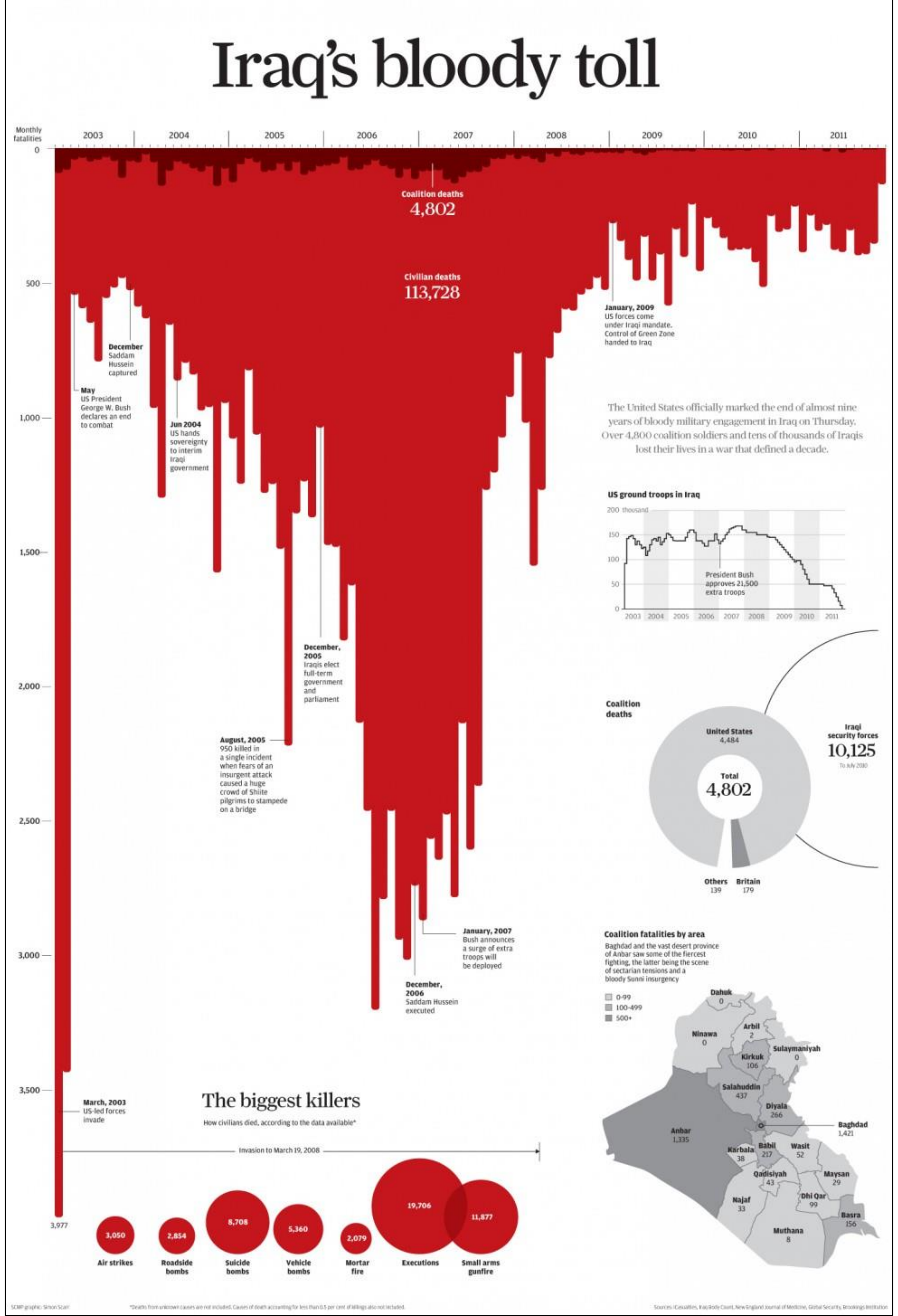
These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

colorspace default palettes
diverge_hcl
diverge_hsl
terrain_hcl
sequential_hcl
rainbow_hcl

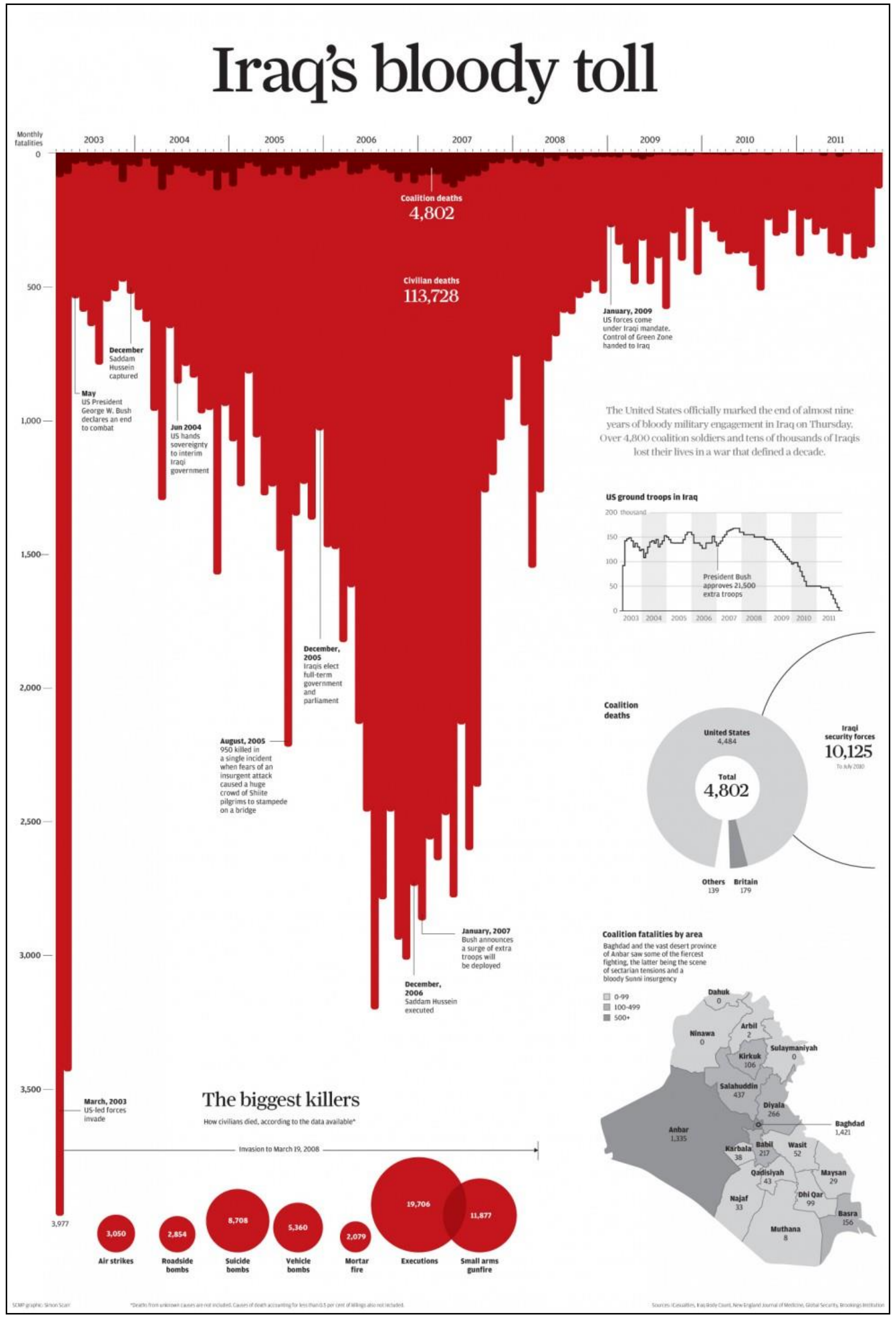
rainbow_hcl(4)
"#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"

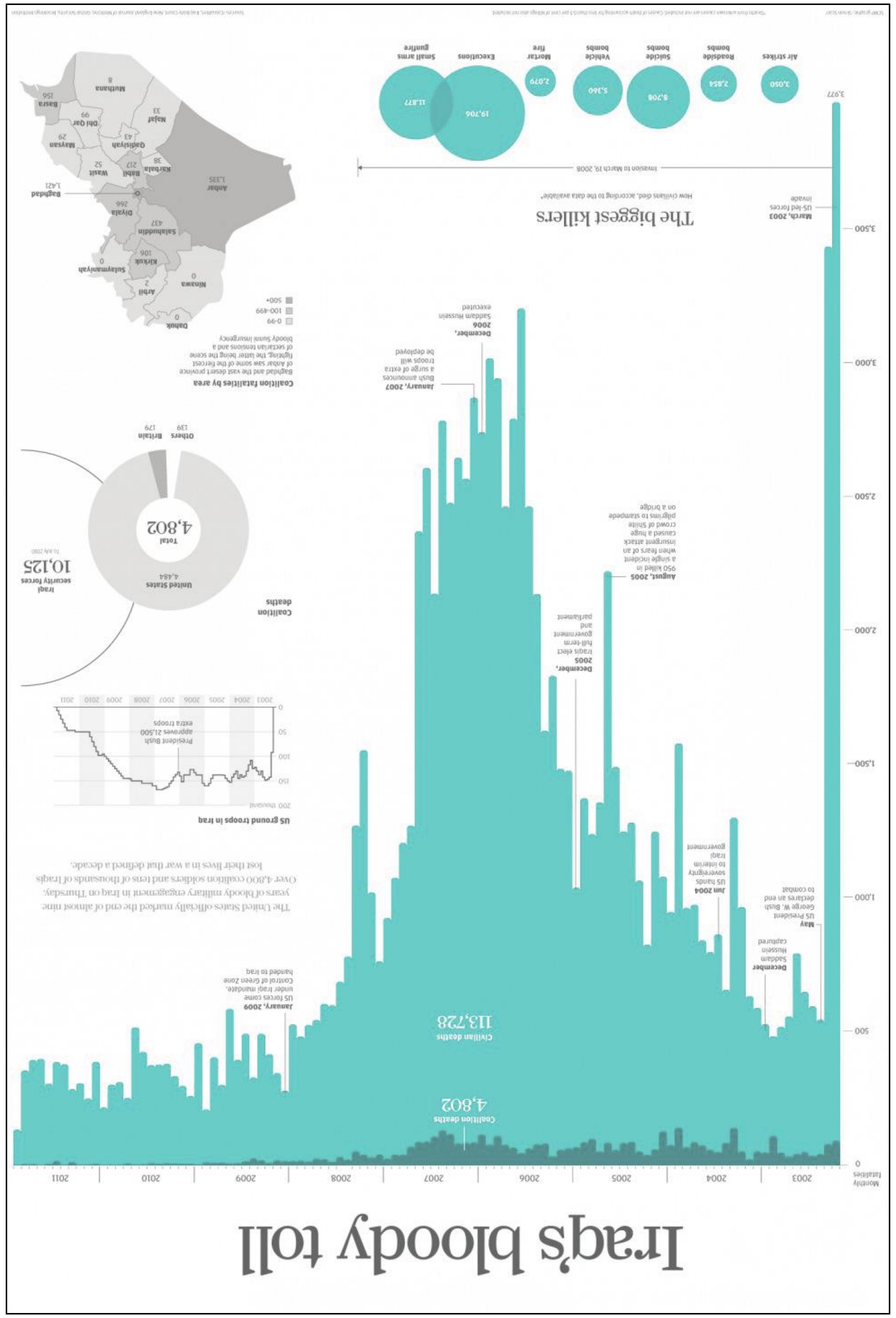
However, all palettes are fully customizable:
diverge_hcl(7, h = c(246, 40), c = 96, l = c(65, 90))
Choosing the values *would* be daunting. But there are some recommended palettes in the colorspace documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
pal <- choose_palette()

Semántica del color

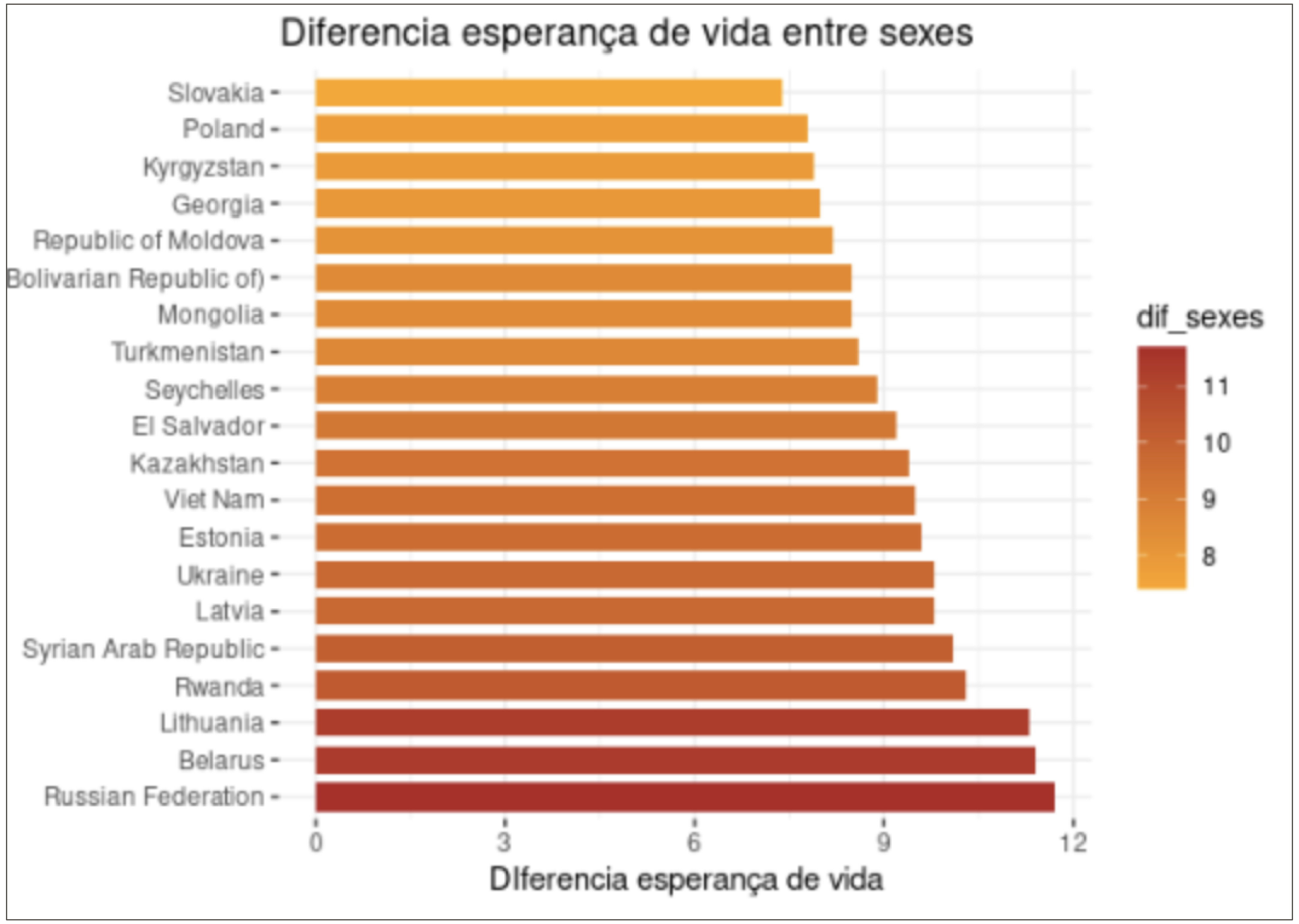


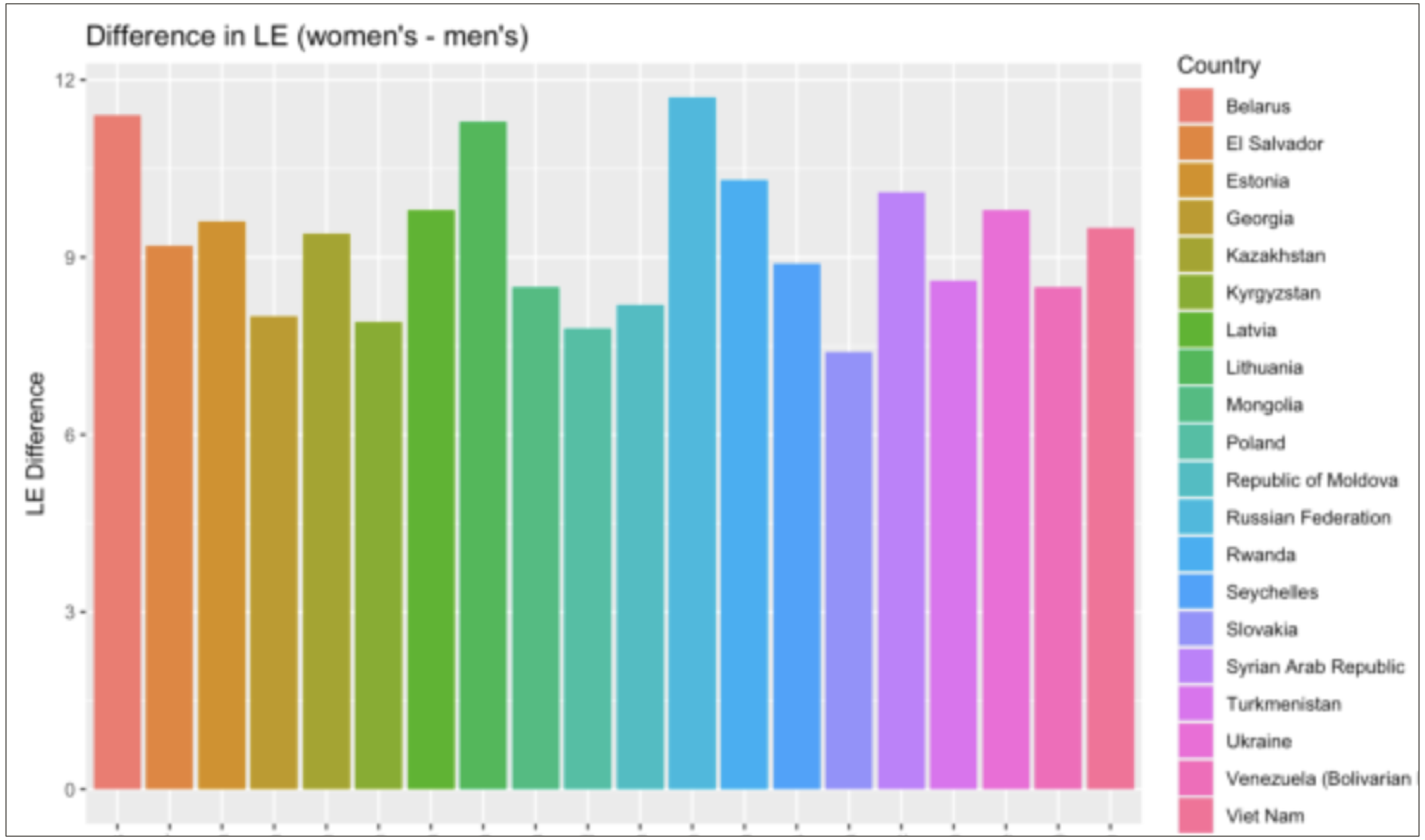
Semántica del color





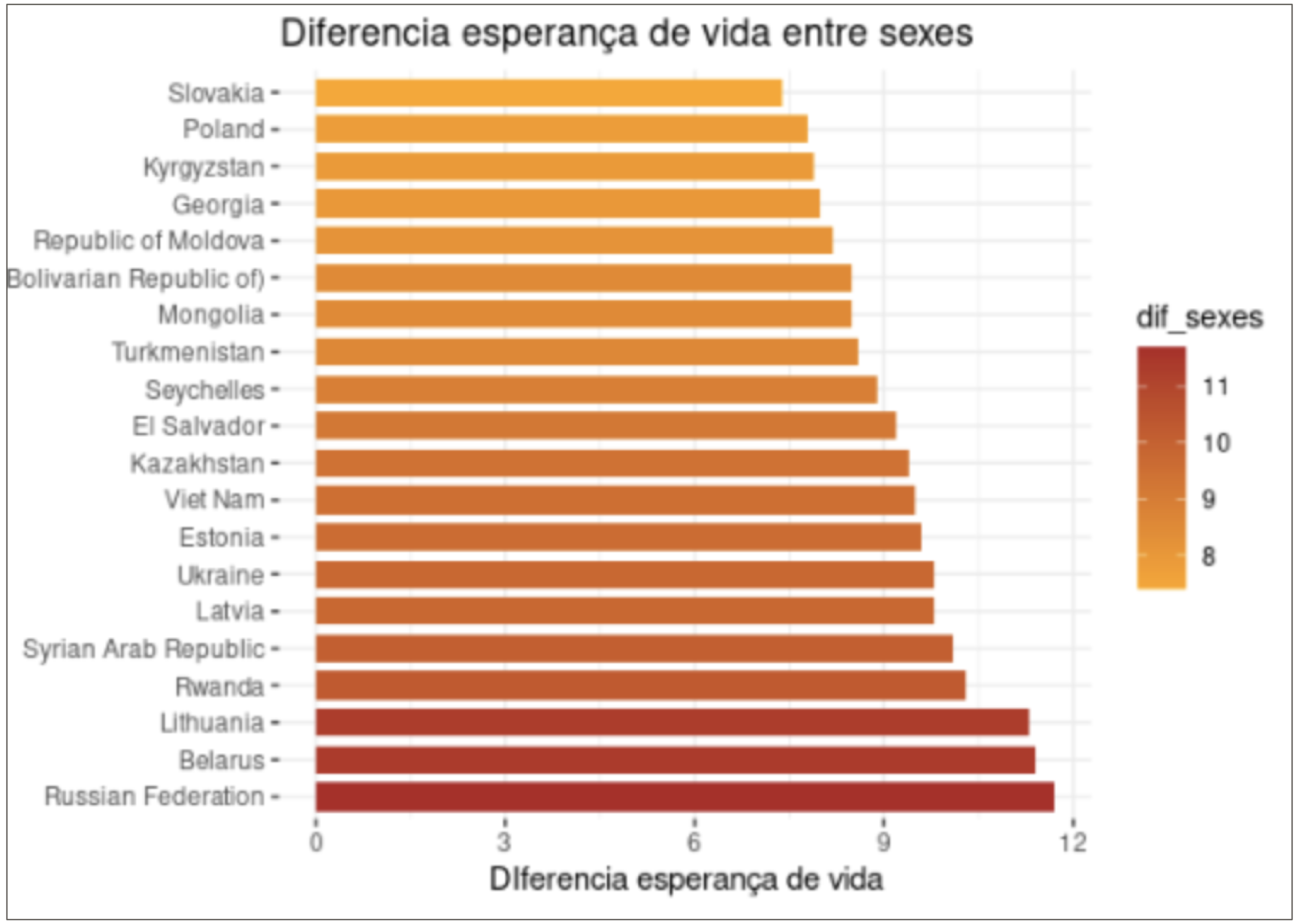
- Gráficas de diferencia en Esperanza de Vida al Nacer – Top 20
- ¿Uso del color correcto?
- ¿Qué fallos tienen?





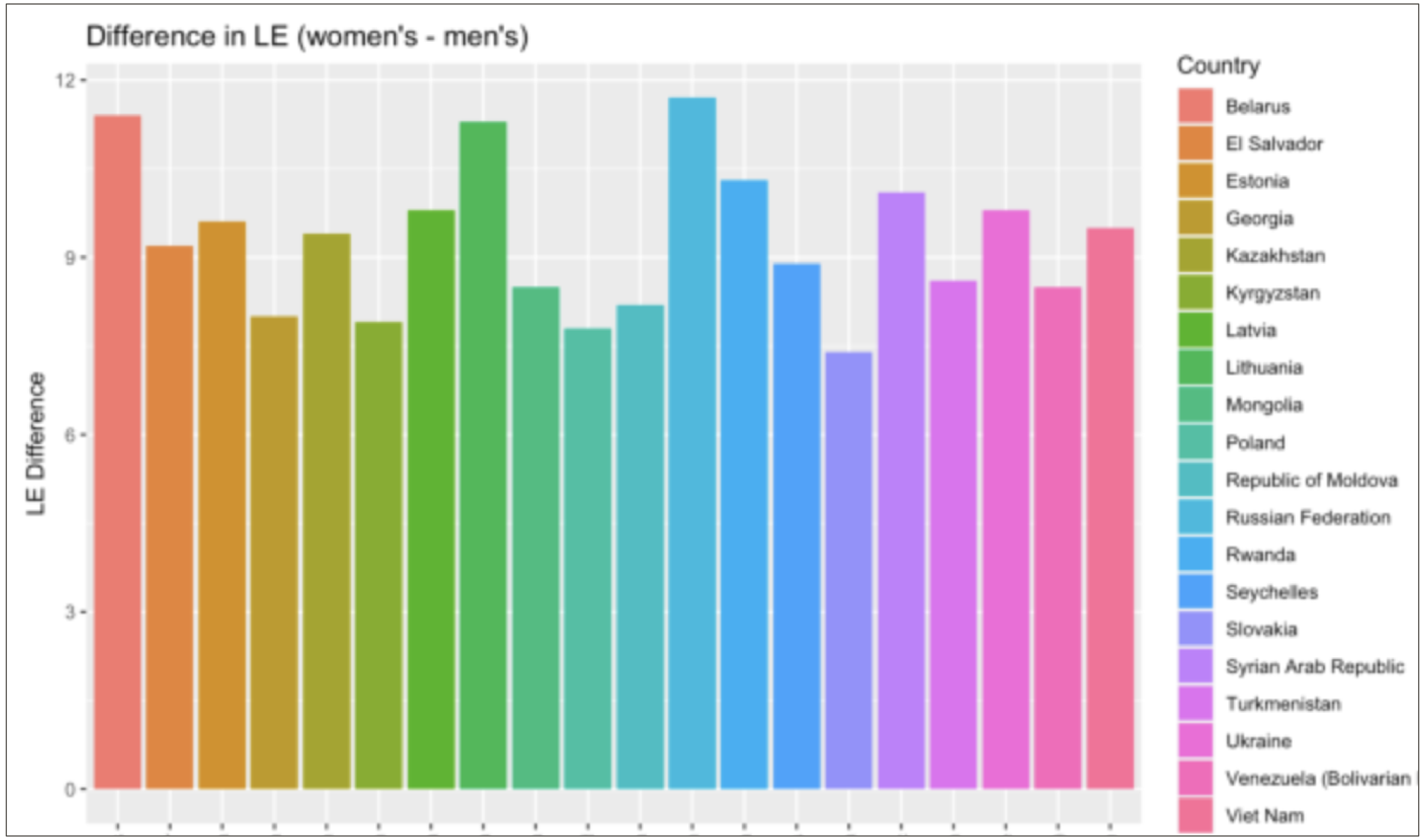
Do's …

- Escala bicromática continua
- Variación en tono y luminosidad
- Escala **continua** usada para **attr. cuantitativo continuo**
- Doble redundancia: Ranking ordenado por tamaño de barras, color



… and dont's

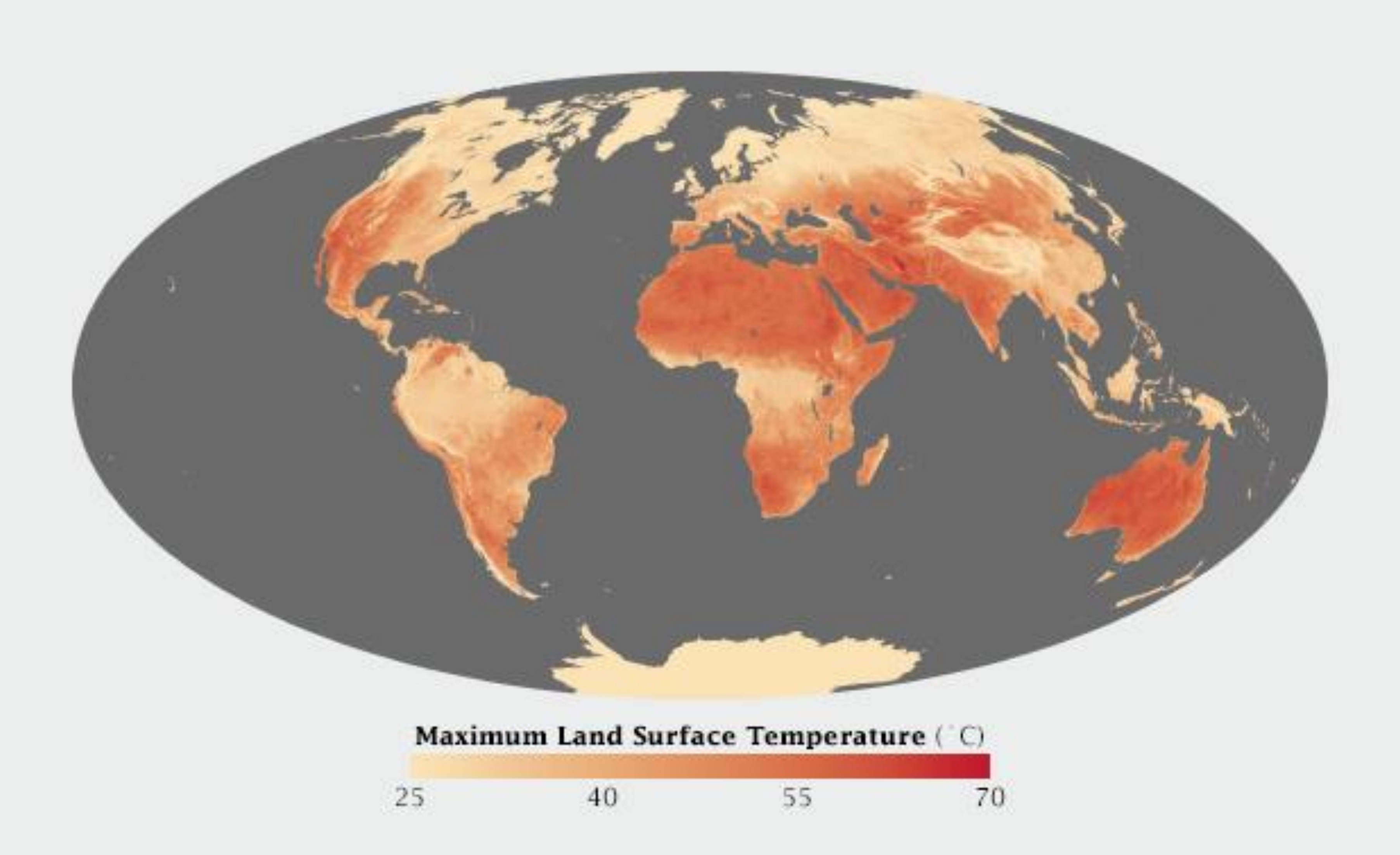
- Escala arcoíris, aunque con luminancia corregida
- Escala **continua** usada como **categórica**
- + de 12 categorías
- Ranking no ordenado
- Etiquetas del barchart fuera de la gráfica

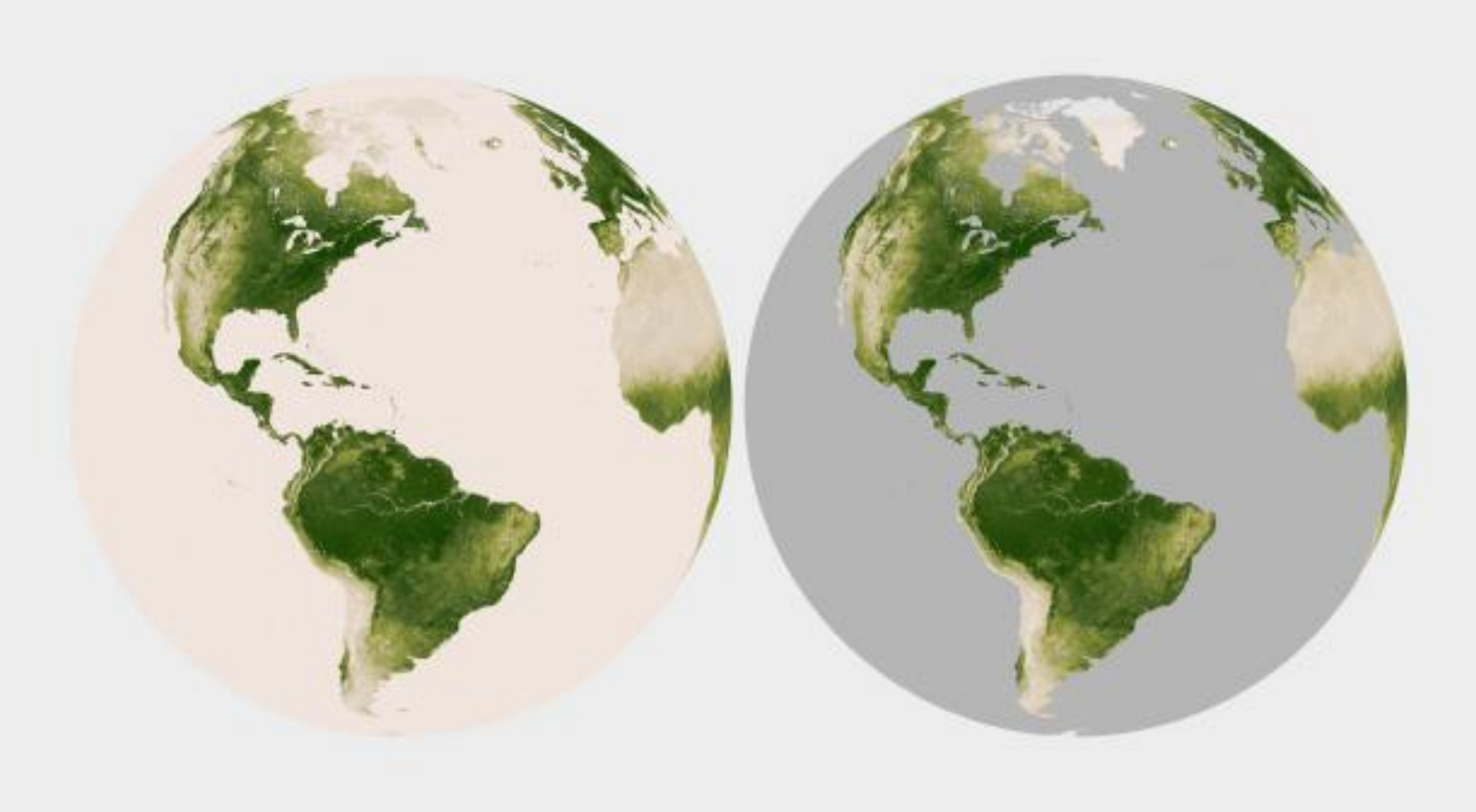


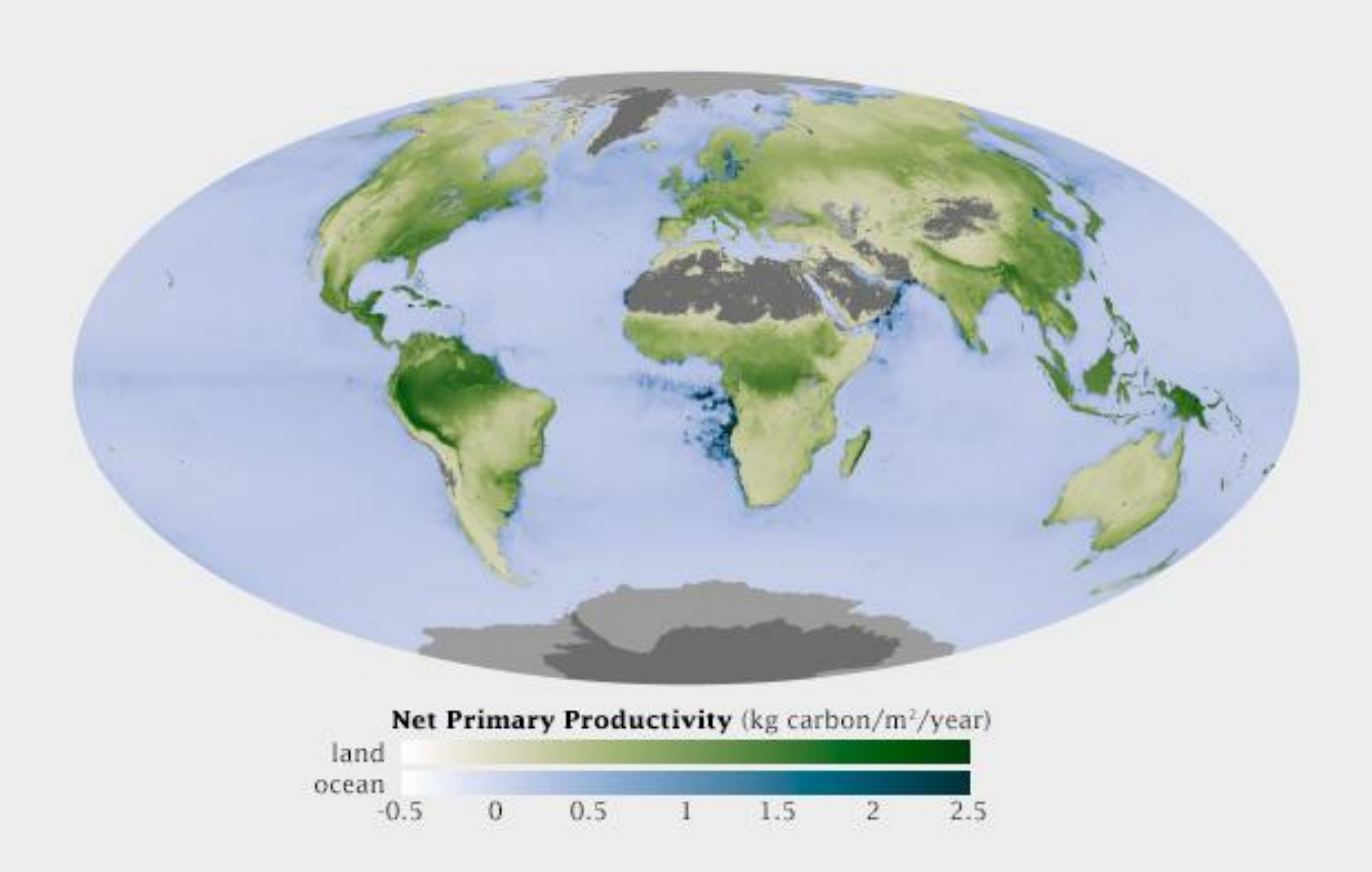
Buenas prácticas

Robert Simmon, 2013

- Asignar color a significado (paletas cálidas para altas temperaturas; divergente azul-rojo para temp negativa a positiva, verde para datos de vegetación, etc)
- Usar la escala correcta para cada tipo de atributo:
  - Si es categórico o cuantitativo/ordinal y
  - si es divergente, secuencial, o cíclico
- Elegir paletas con contraste en luminancia para resaltar detalles.
- Tener en cuenta color-blindness y usos (imprimir, colores web, etc)
- No data, no color







Bibliografia

Color:

- *Subtleties of color:* https://earthobservatory.nasa.gov/blogs/elegantfigures/2013/08/05/subtleties-of-color-part-1-of-6/
- *Tableau color palettes:* https://www.tableau.com/about/blog/2016/7/colors-upgrade-tableau-10-56782
- *Bi-variate choropleth maps:* www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/
- *R colorspace package:* https://cran.r-project.org/web/packages/colorspace/vignettes/colorspace.html

Inspiración

http://www.thefunctionalart.com/

https://eagereyes.org/

http://flowingdata.com/

http://fivethirtyeight.com/

http://truth-and-beauty.net/

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

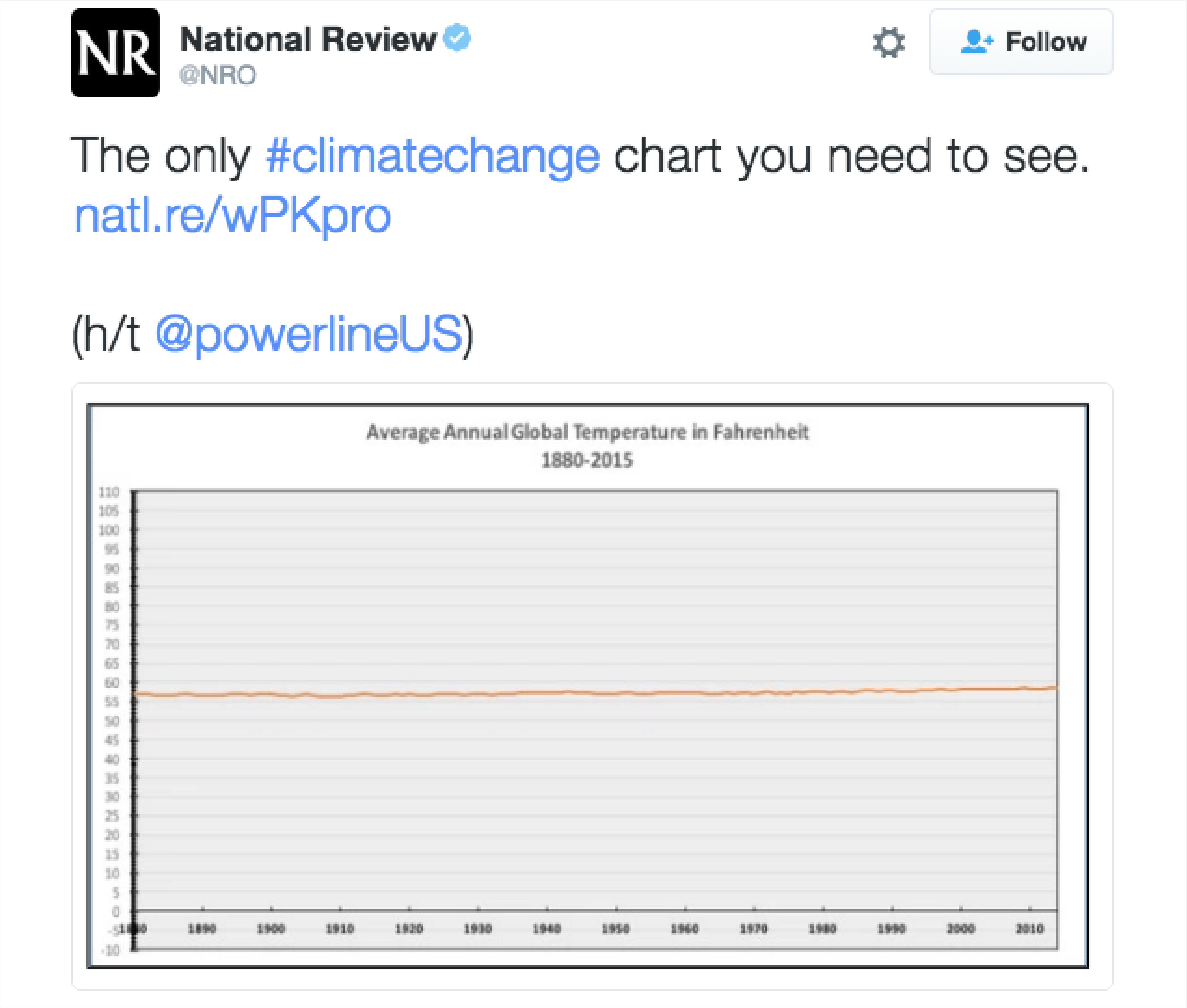
Universitat Autònoma de Barcelona

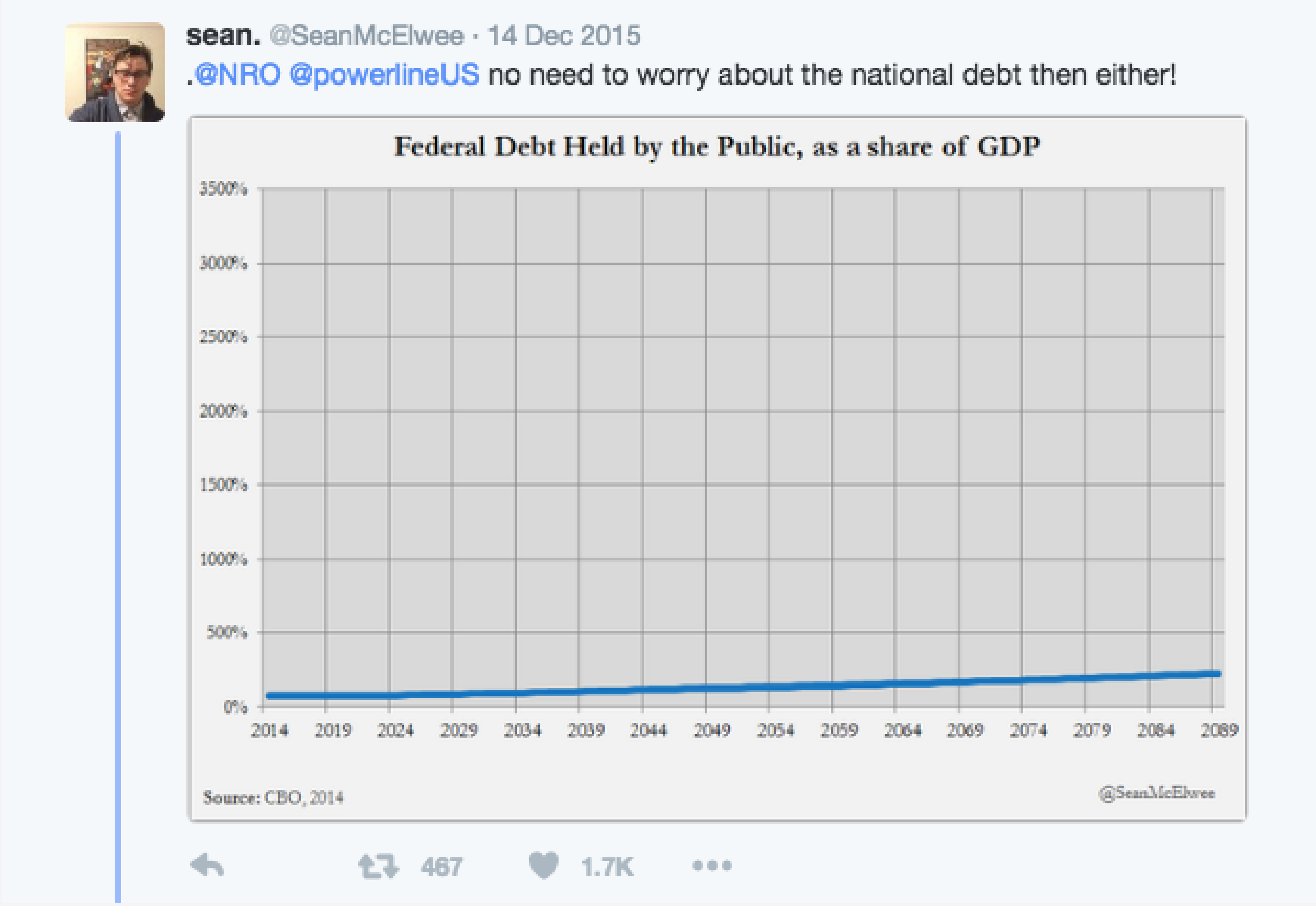
8.2 Honestidad y Precisión visual

Ejes

National Review
@NRO
Follow
The only #climatechange chart you need to see. natl.re/wPKpro
(h/t @powerlineUS)
Average Annual Global Temperature in Fahrenheit
1880-2015
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
-5
-10
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010

Ejes


National Review
@NRO
Follow
The only #climatechange chart you need to see. natl.re/wPKpro
(h/t @powerlineUS)
Average Annual Global Temperature in Fahrenheit
1880-2015
110 105 100 95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0 -5 -10
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010


sean. @SeanMcElwee · 14 Dec 2015
.@NRO @powerlineUS no need to worry about the national debt then either!
Federal Debt Held by the Public, as a share of GDP
3500% 3000% 2500% 2000% 1500% 1000% 500% 0%
2014 2019 2024 2029 2034 2039 2044 2049 2054 2059 2064 2069 2074 2079 2084 2089
Source: CBO, 2014
@SeanMcElwee
467
1.7K

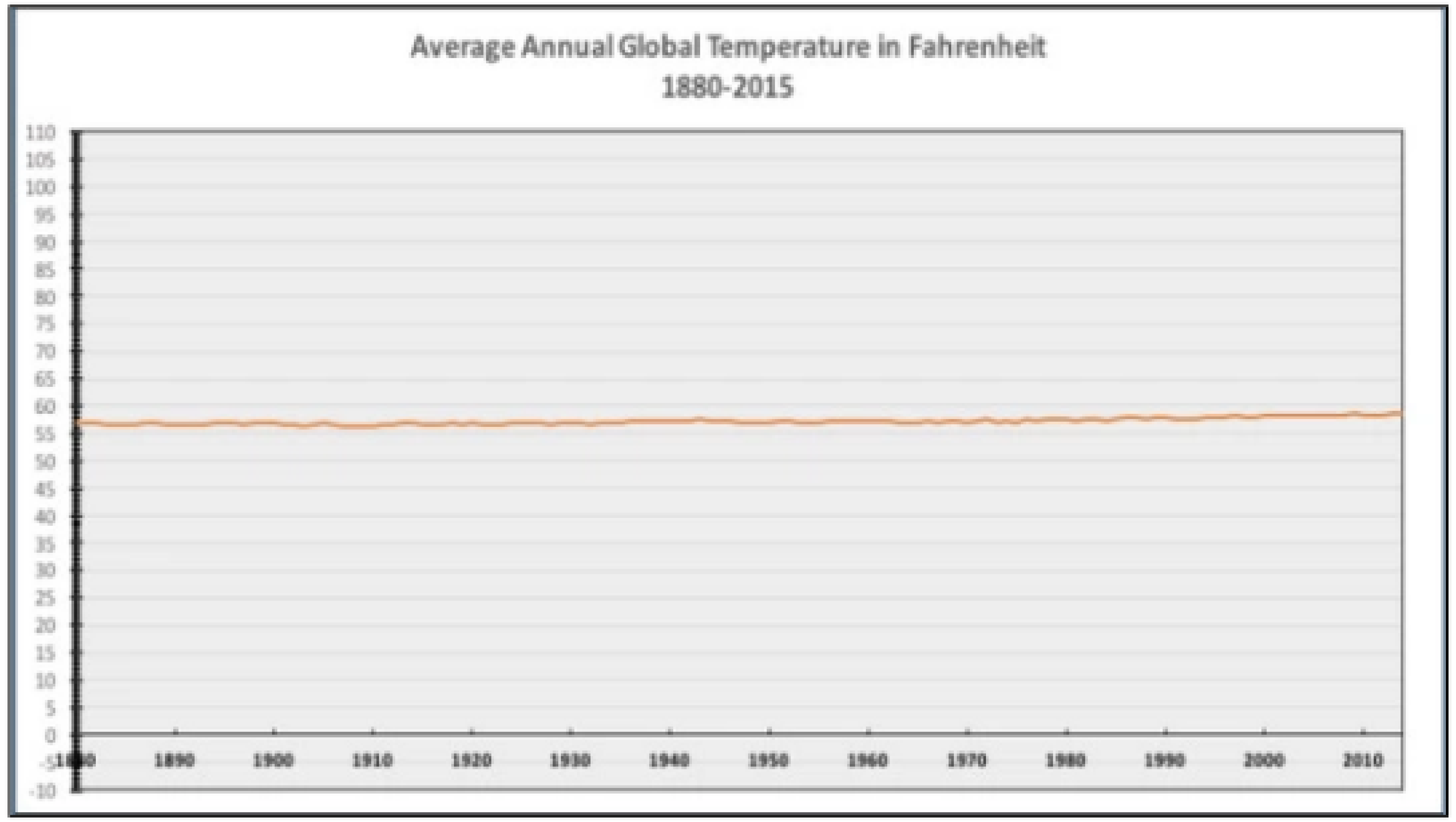
Ejes

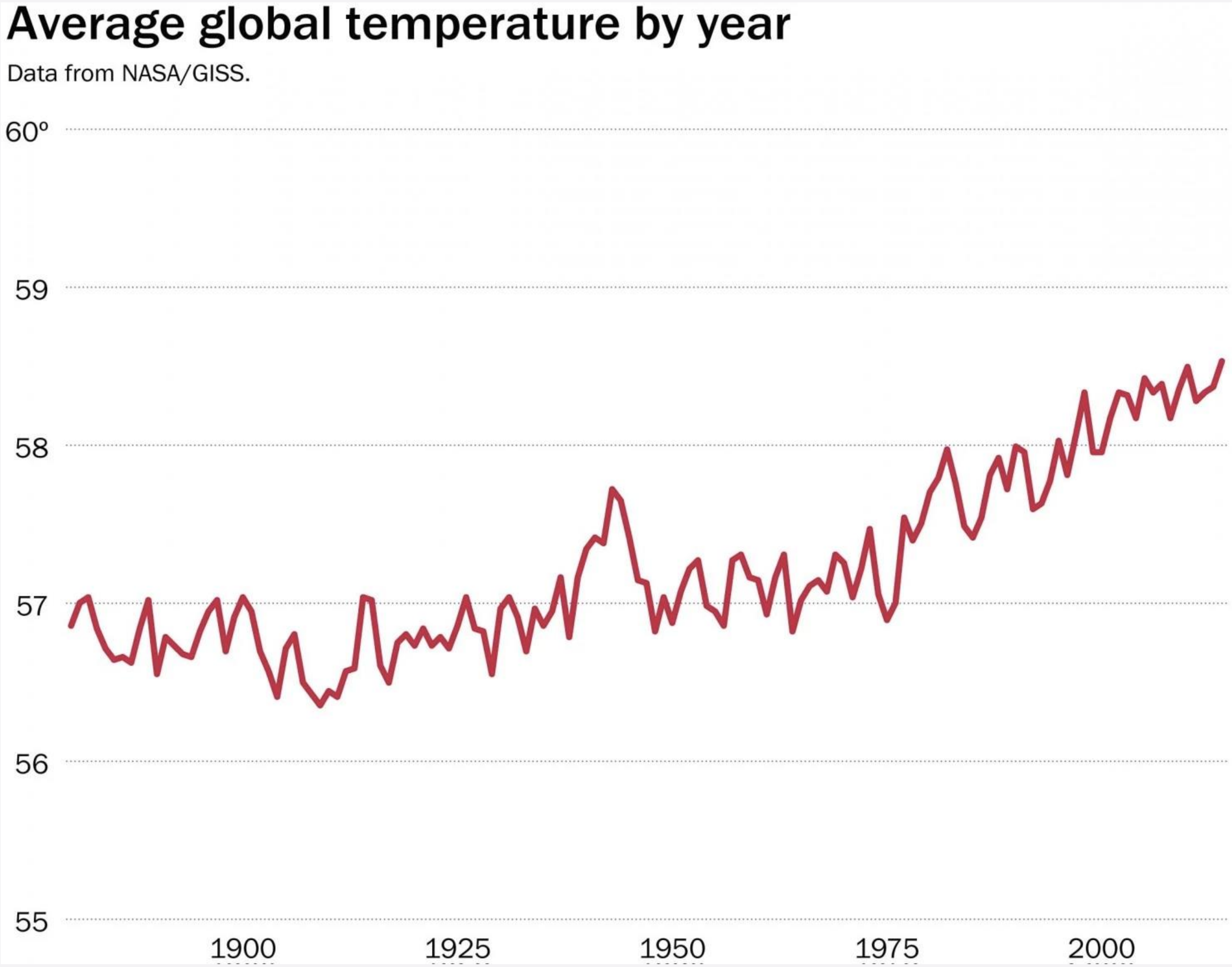
National Review @NRO

Follow

The only #climatechange chart you need to see. natl.re/wPKpro

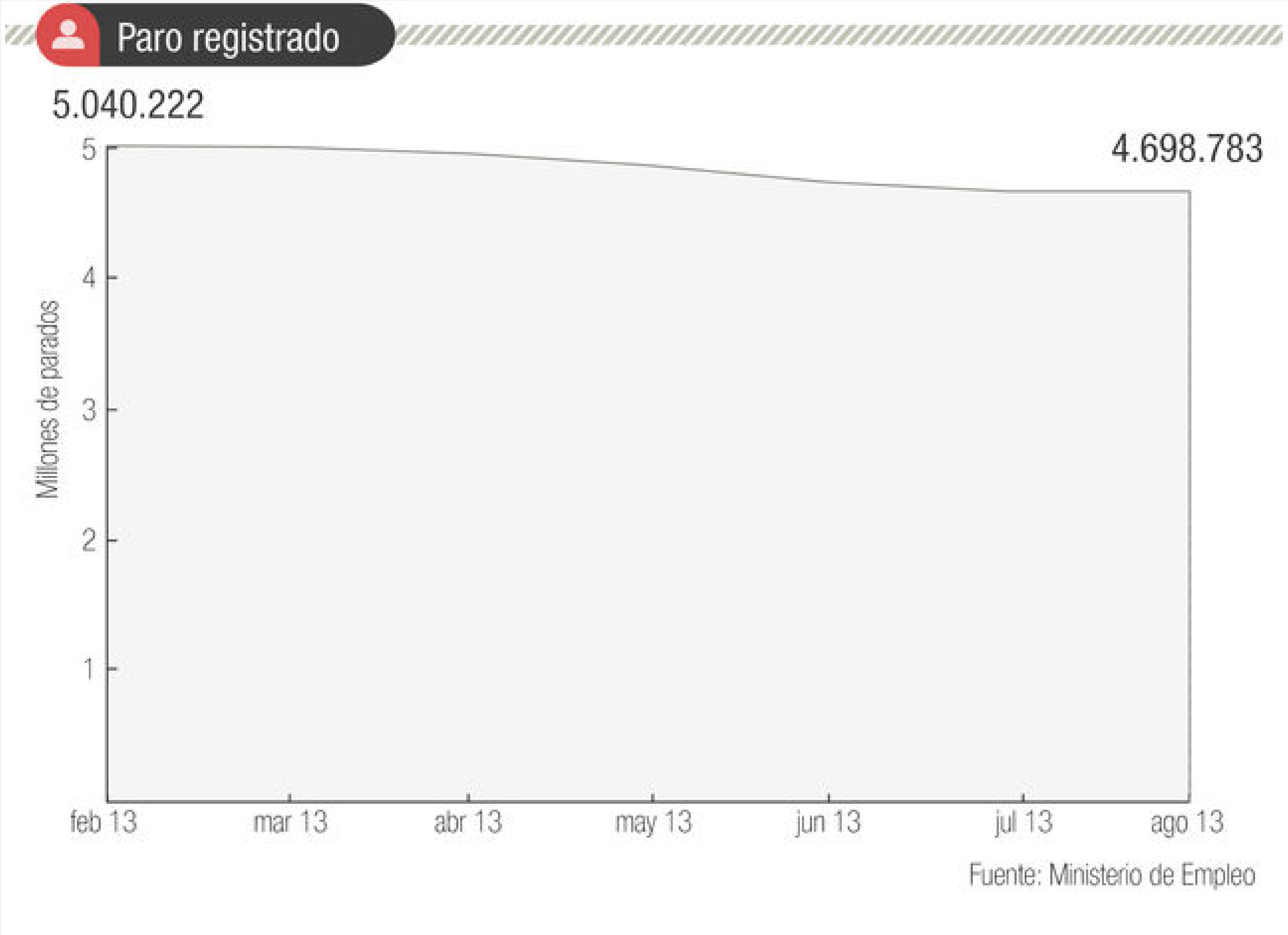
(h/t @powerlineUS)





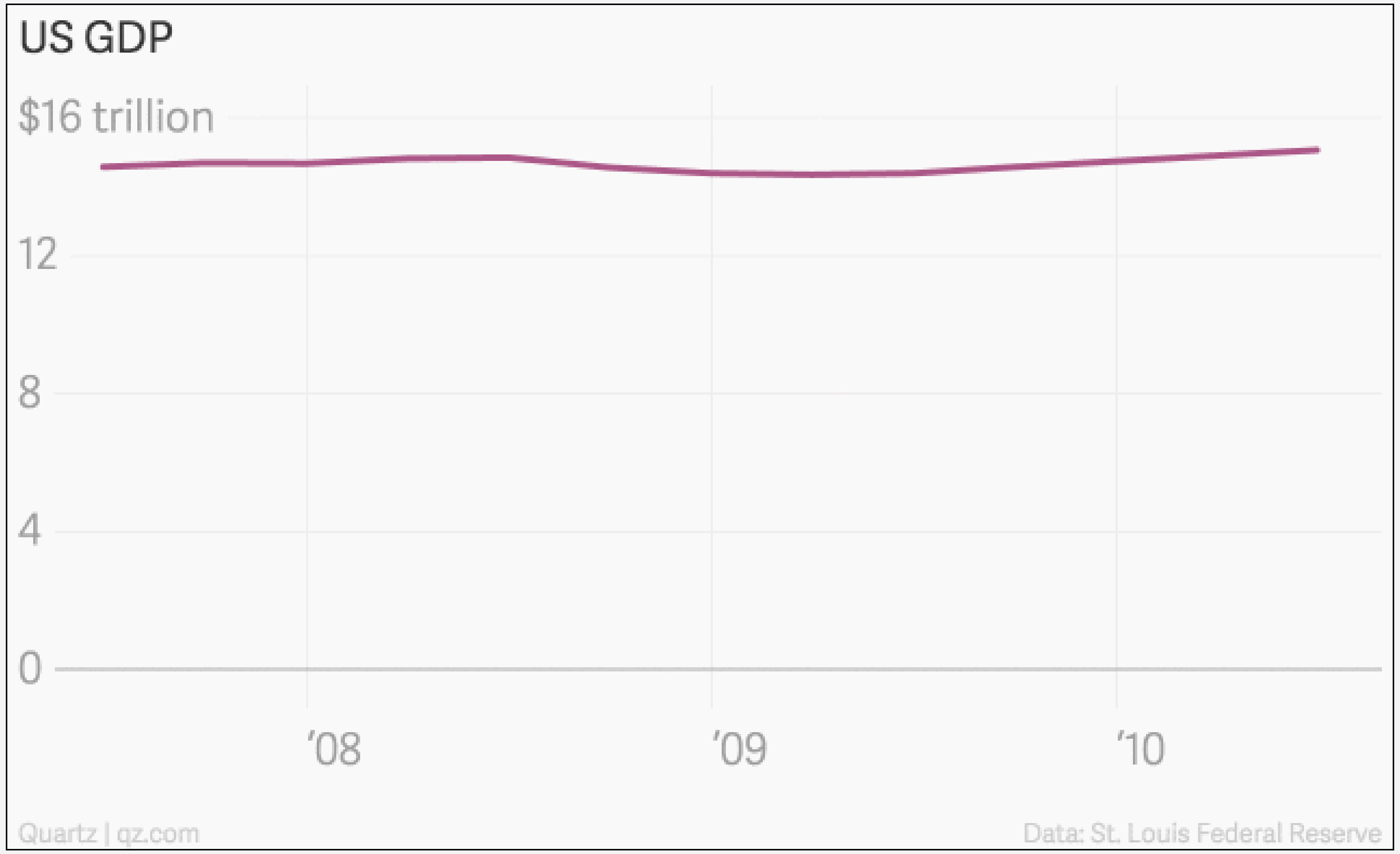
Ejes

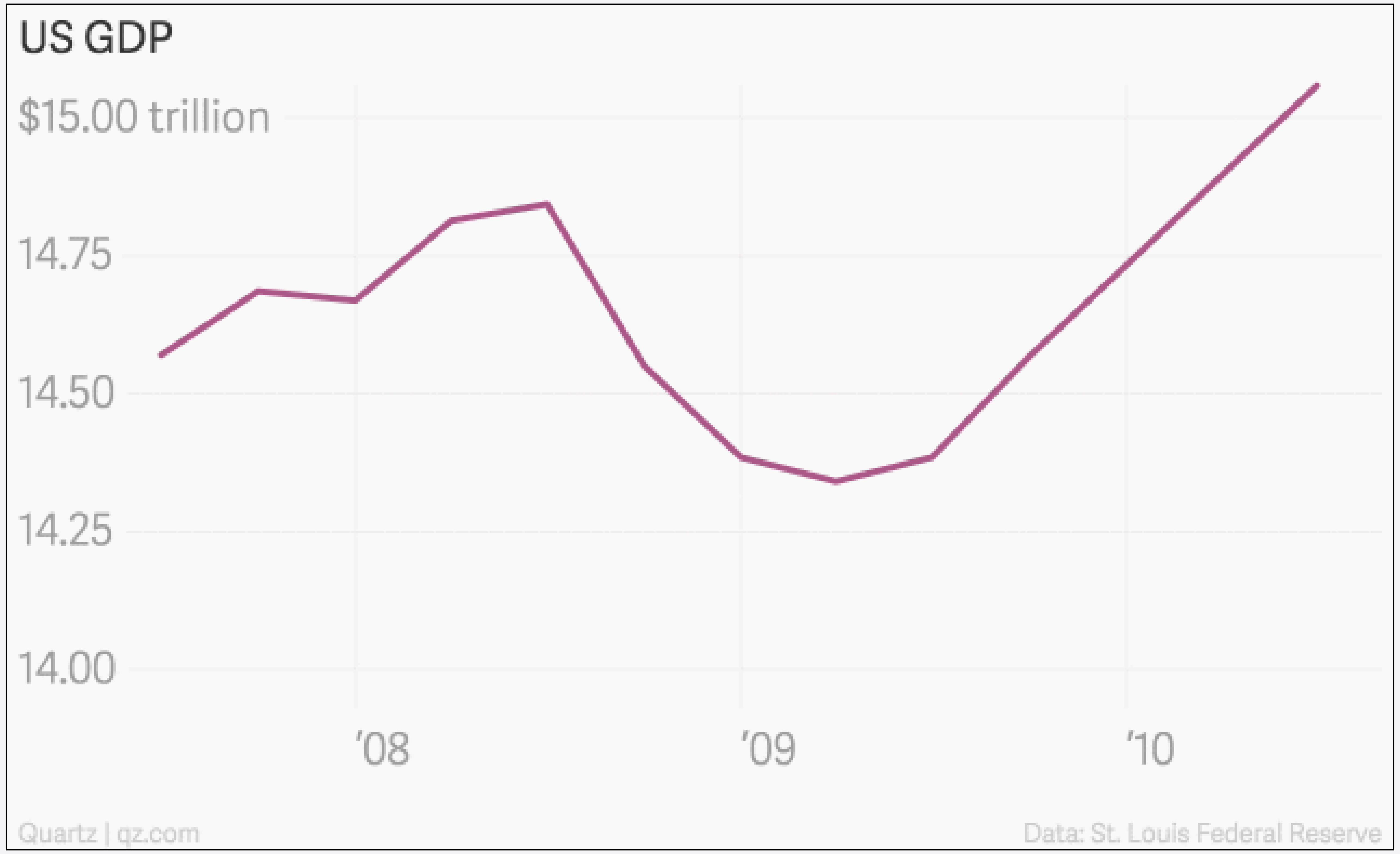
Captura del gráfico con la evolución del paro registrado emitido en el Telediario 1 de TVE



Mostrar eje completo para no magnificar el descenso

Ejes

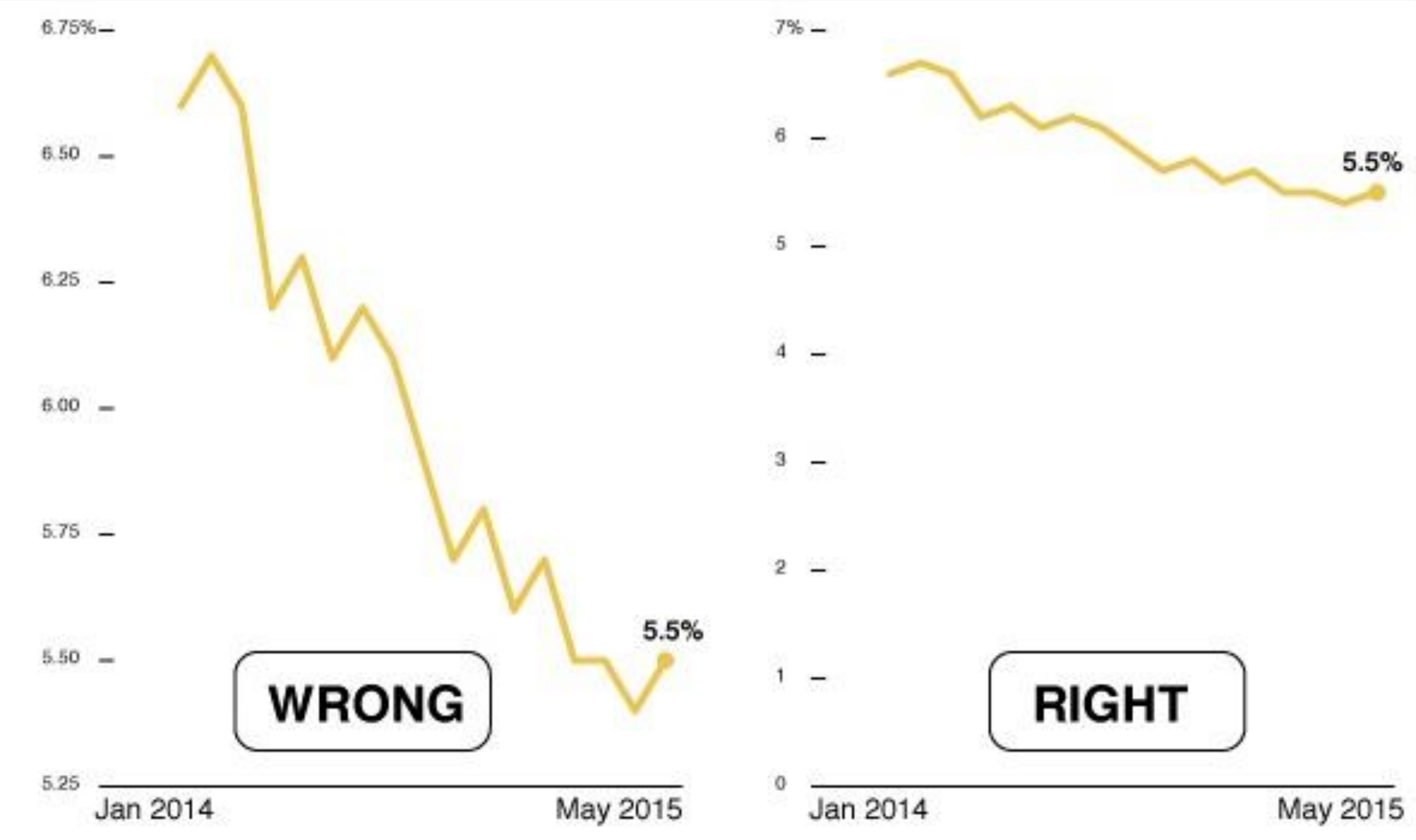




La elección del eje puede enfatizar variación o tendencia

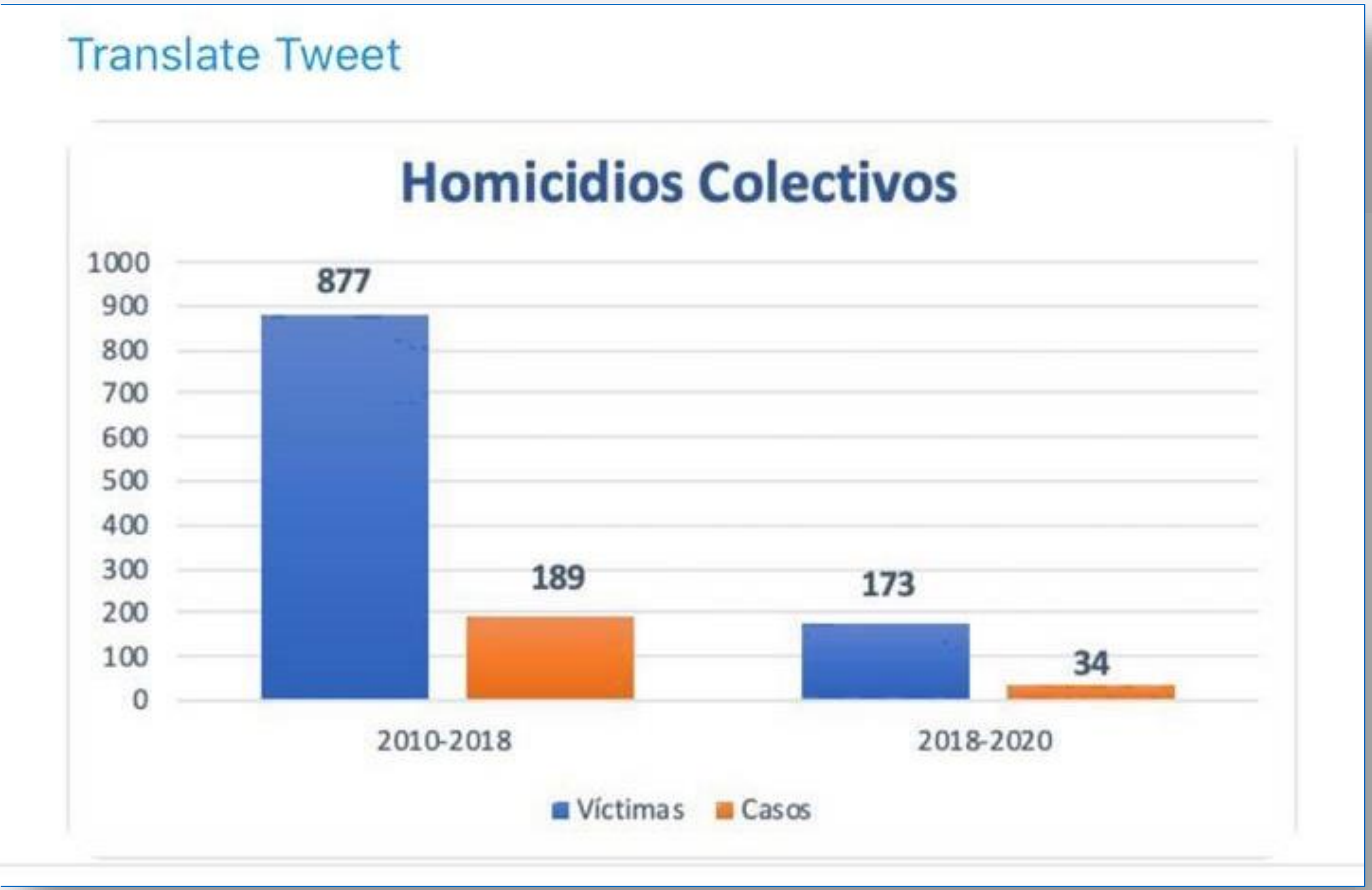
Ejes

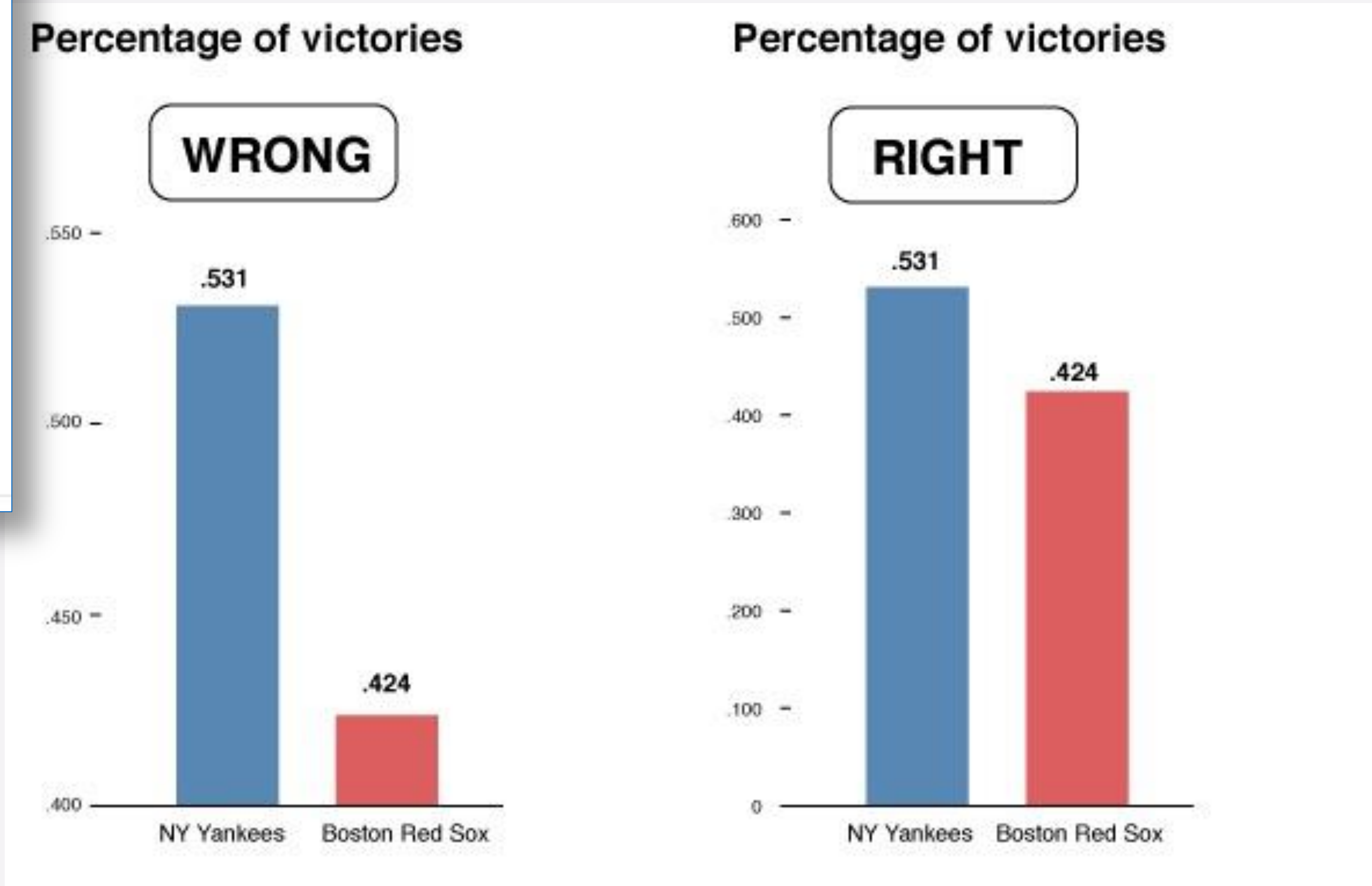
- Linecharts codifican POSICIÓN de puntos
- Barras codifican TAMAÑO
- Las barras necesitan zero baseline, los puntos (linechart) no



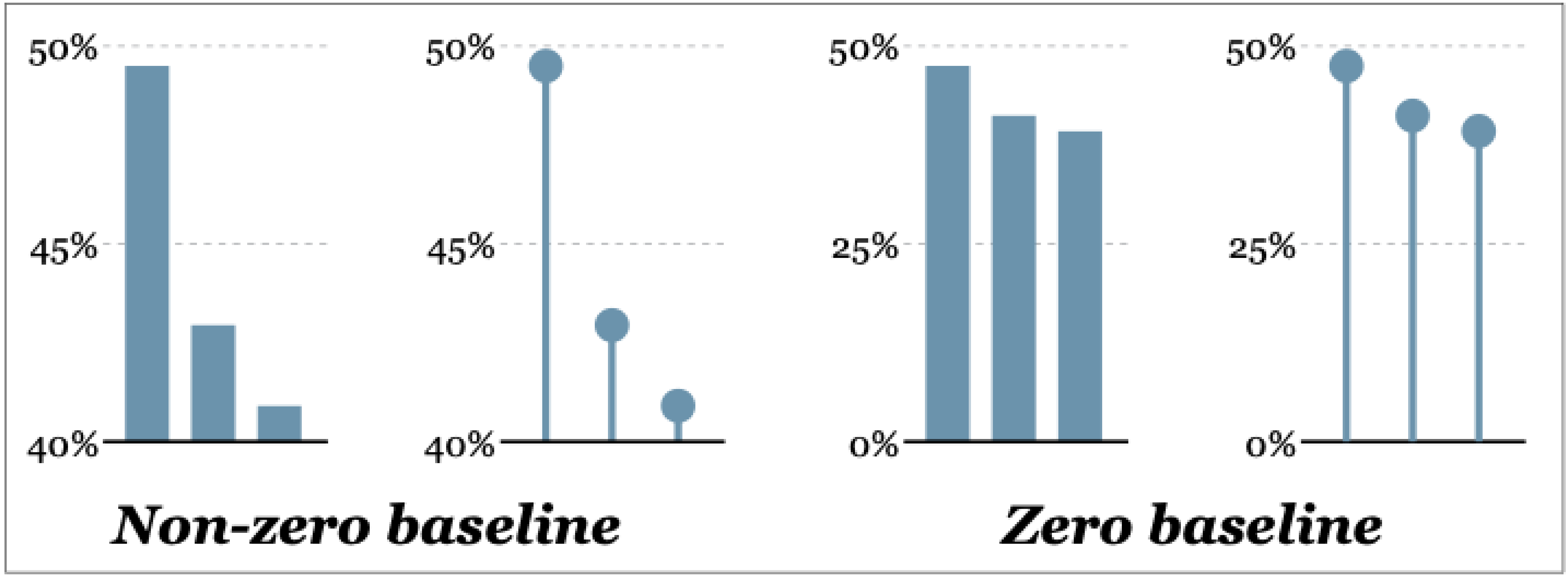
http://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

Ejes



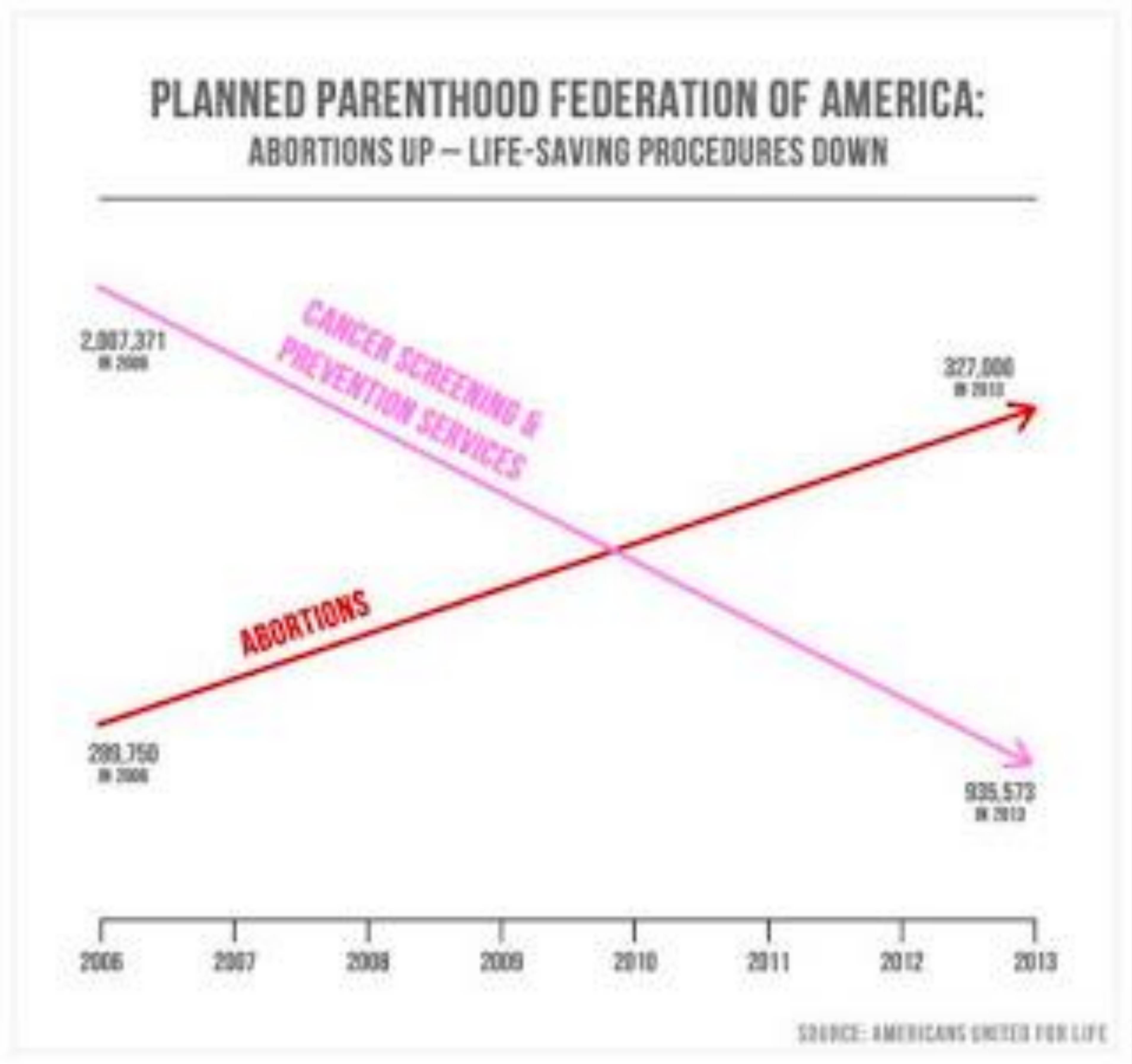


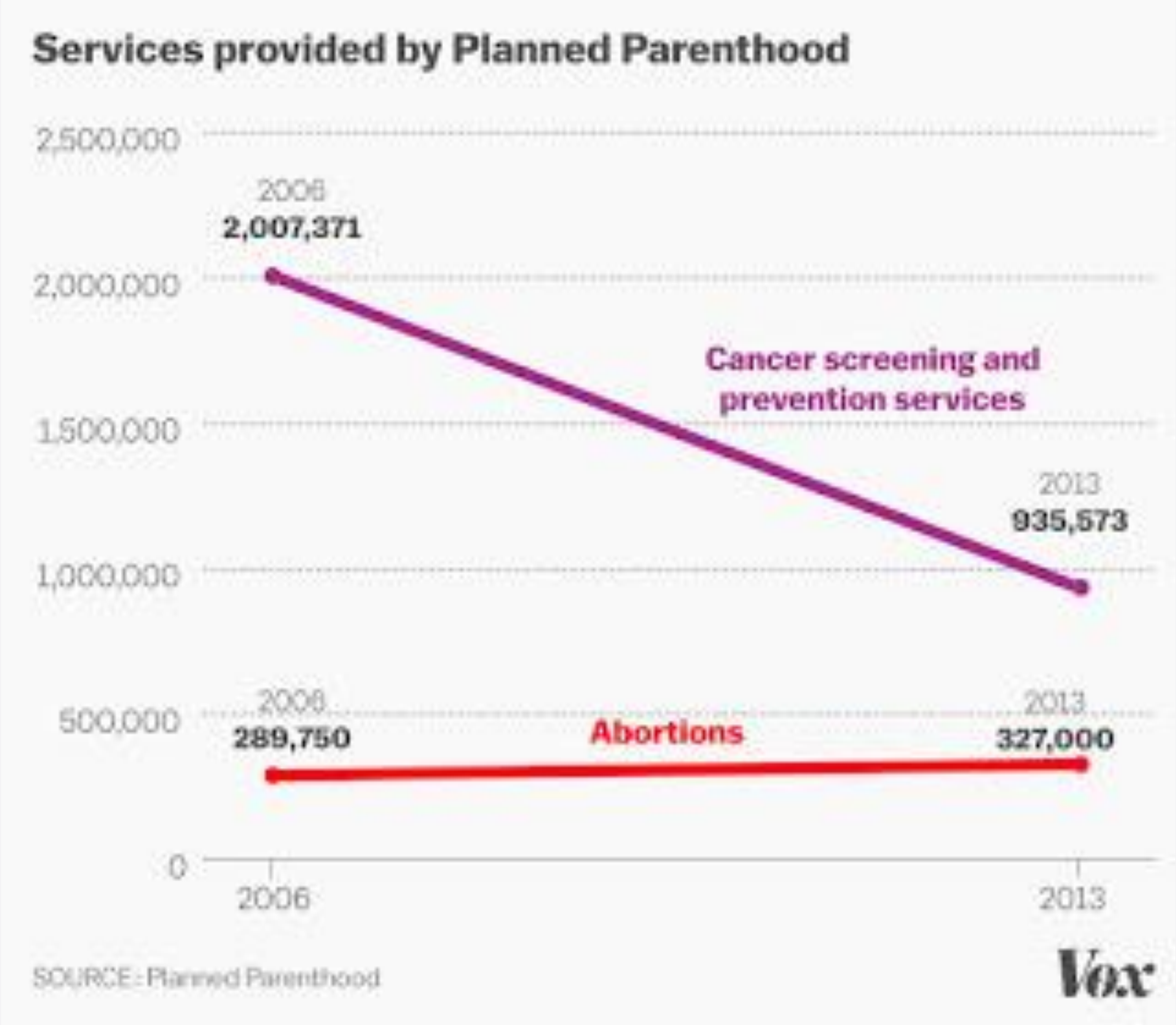
http://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/


50%
45%
40%
50%
45%
40%
50%
25%
0%
50%
25%
0%
Non-zero baseline
Zero baseline

Ejes

- Labels en los ejes son **imprescindibles**



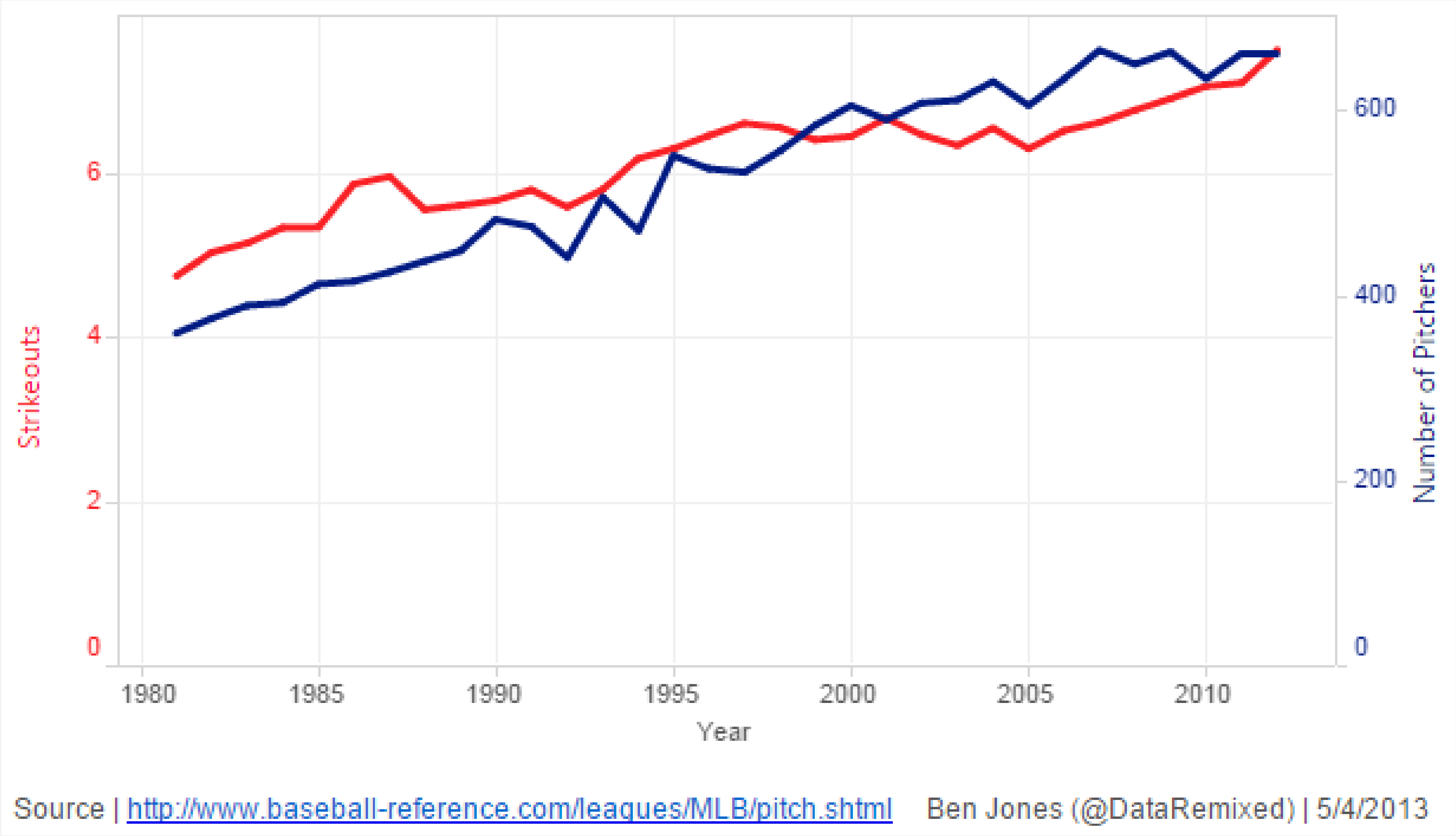


http://www.thefunctionalart.com/2015/10/if-you-see-bullshit-say-bullshit.html

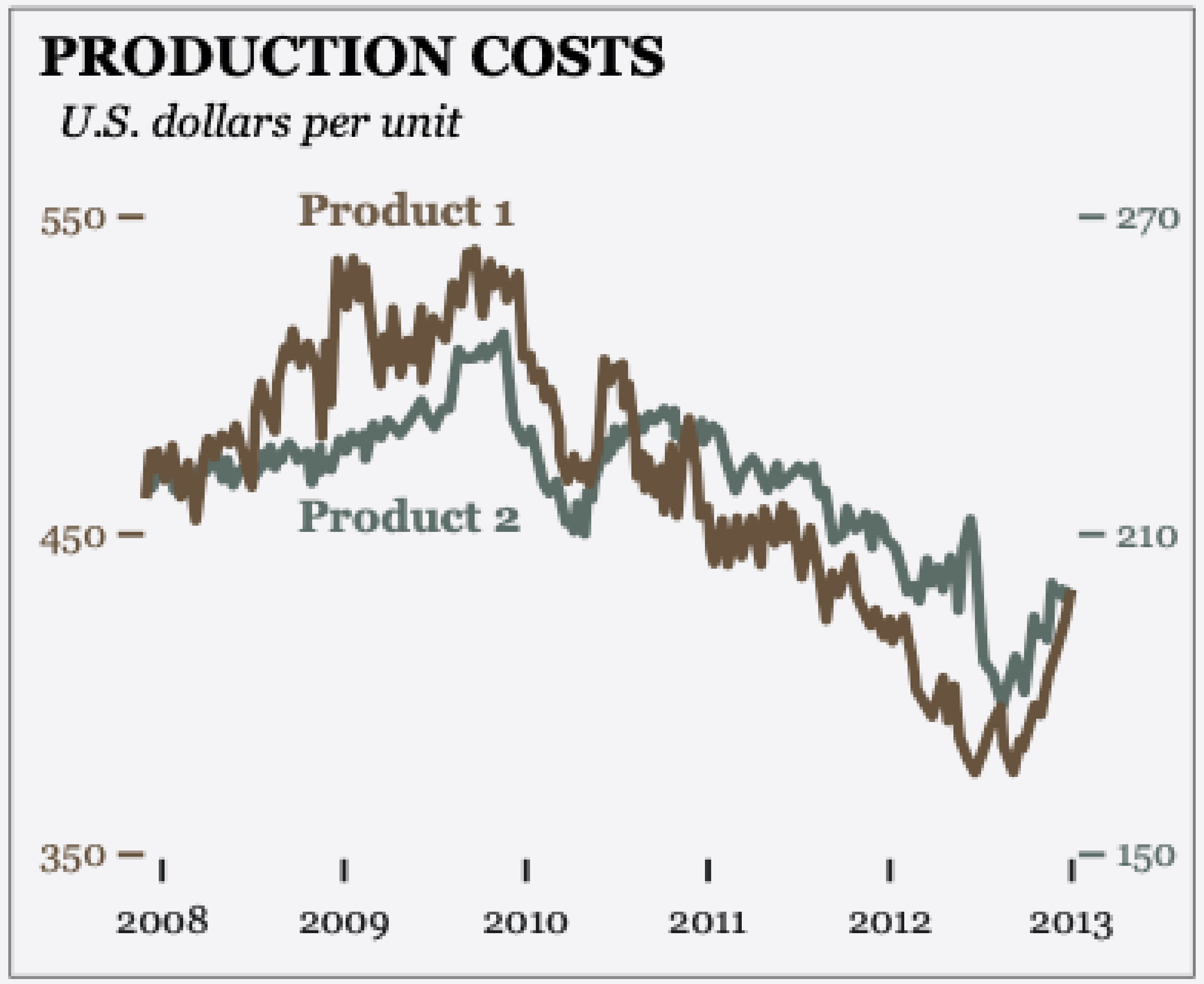
Evitar ejes duales

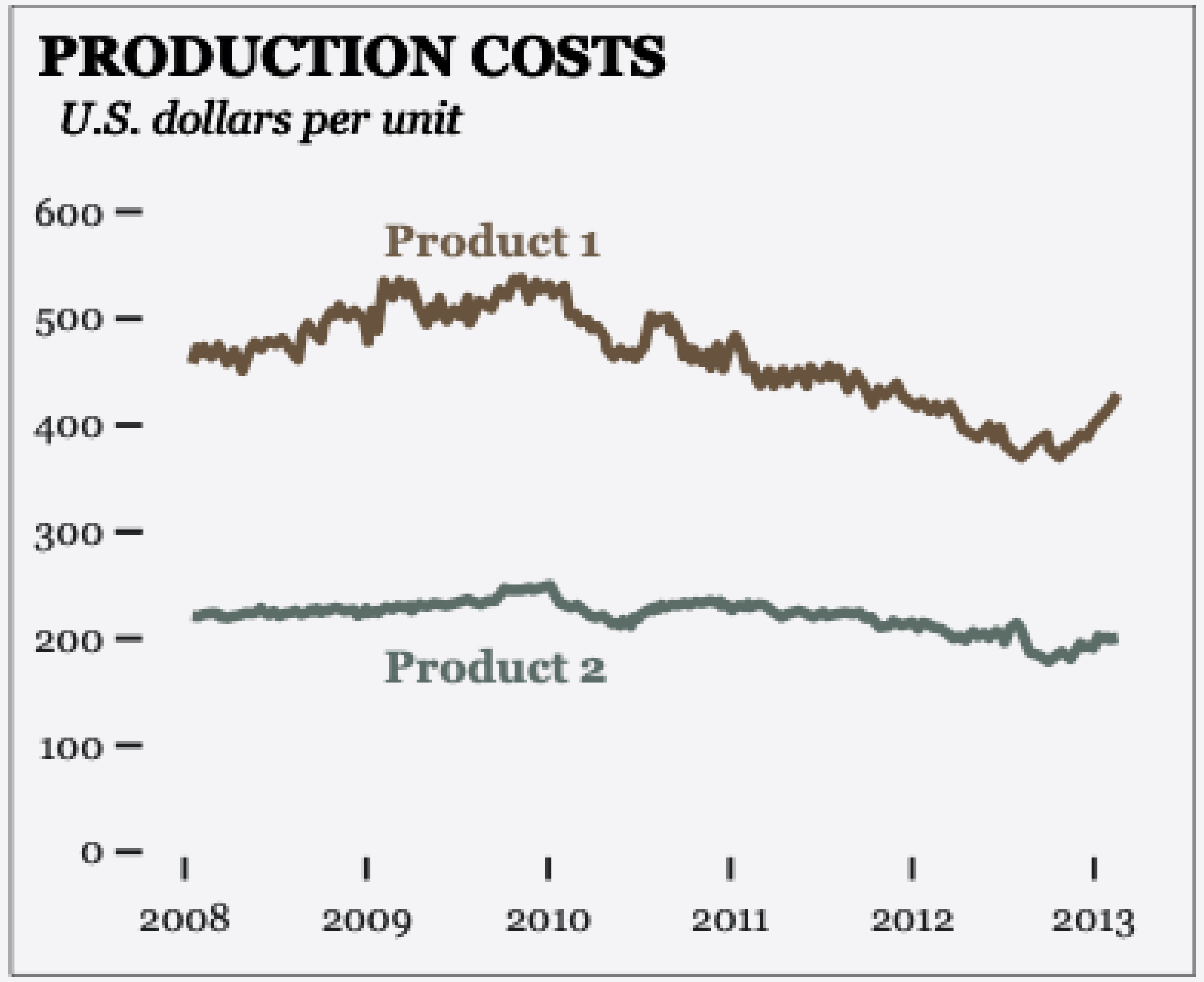
Problemáticos

- Aceptable en casos muy concretos
- Muy fácil llevar a conclusiones equivocadas:
  - Muestra implícita correlación entre líneas

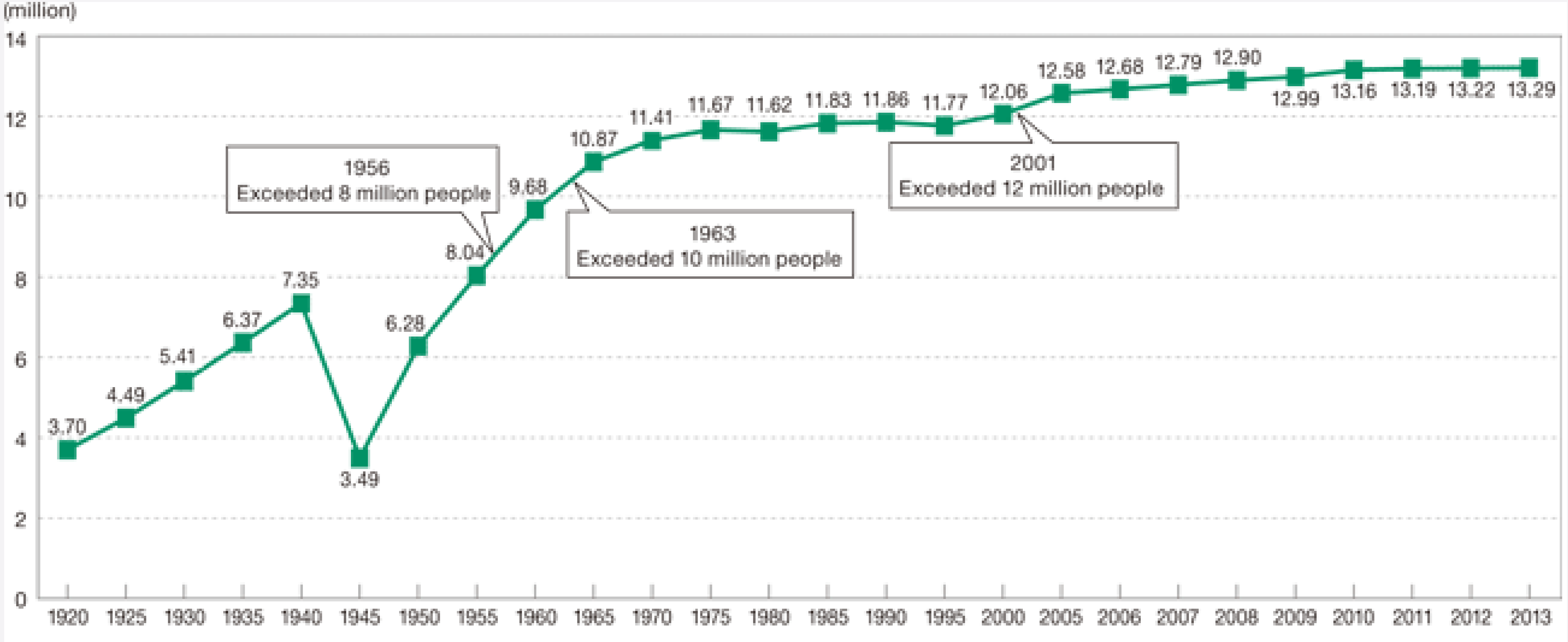


dataremixed.com



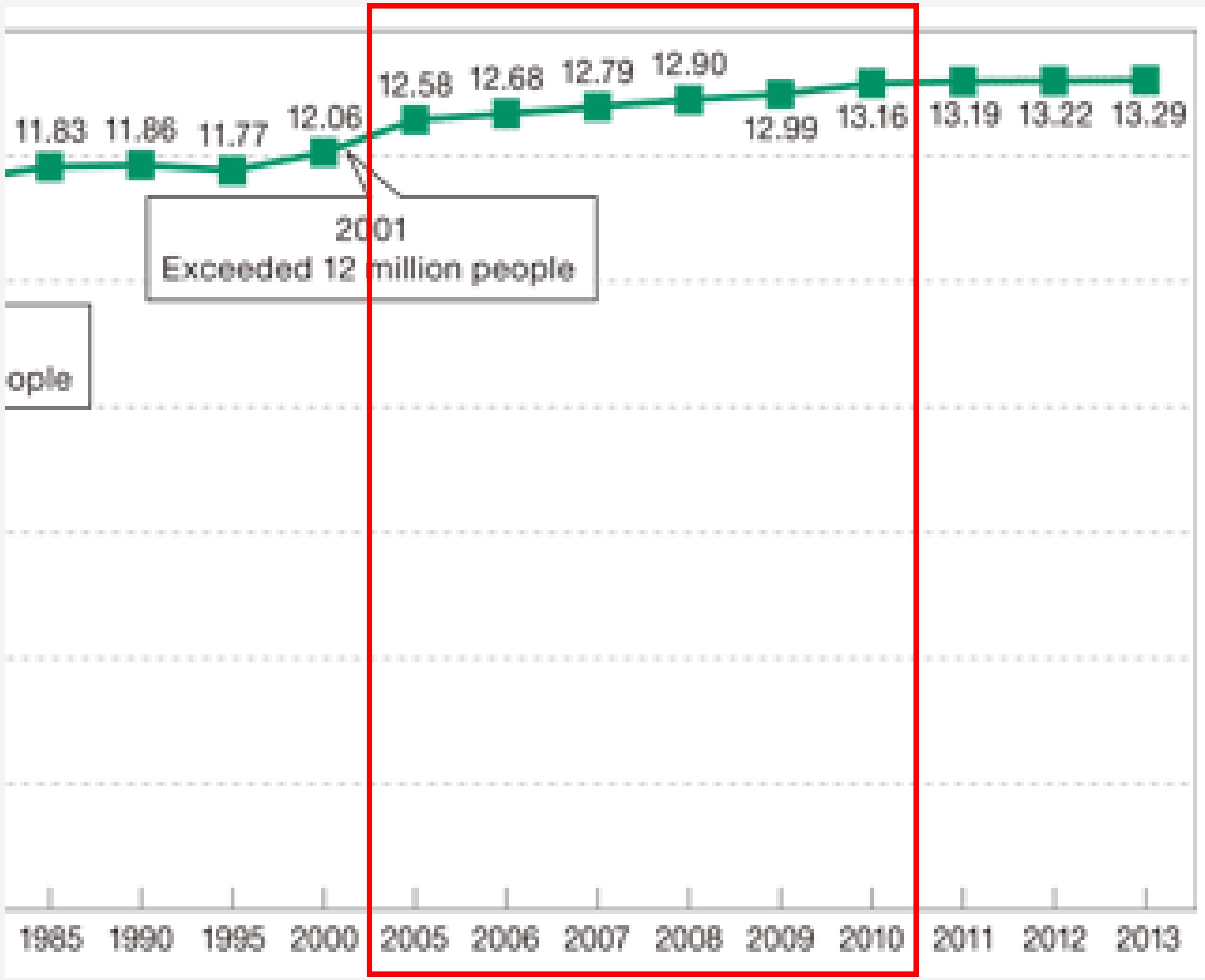


Steps

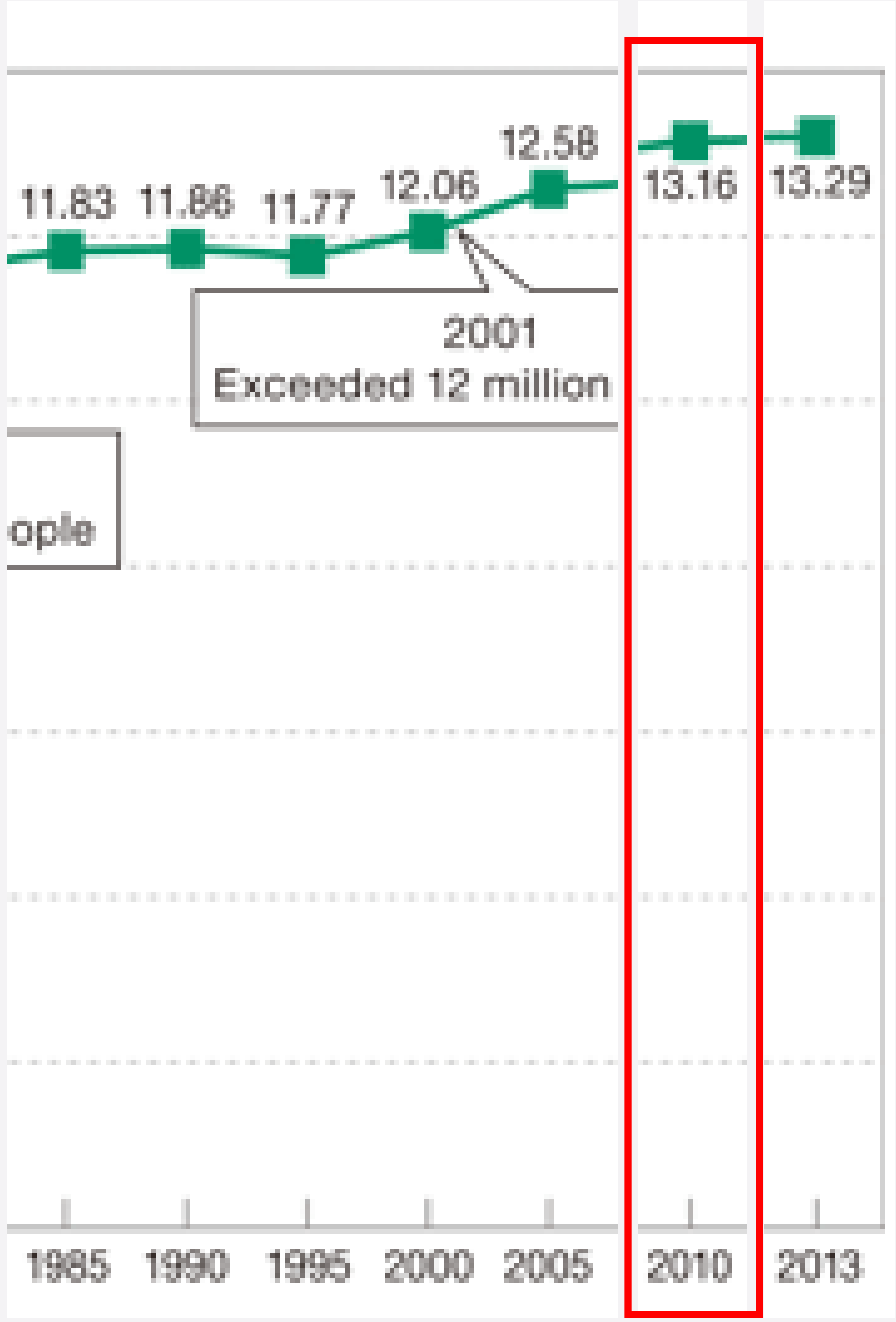


http://www.metro.tokyo.jp

Steps

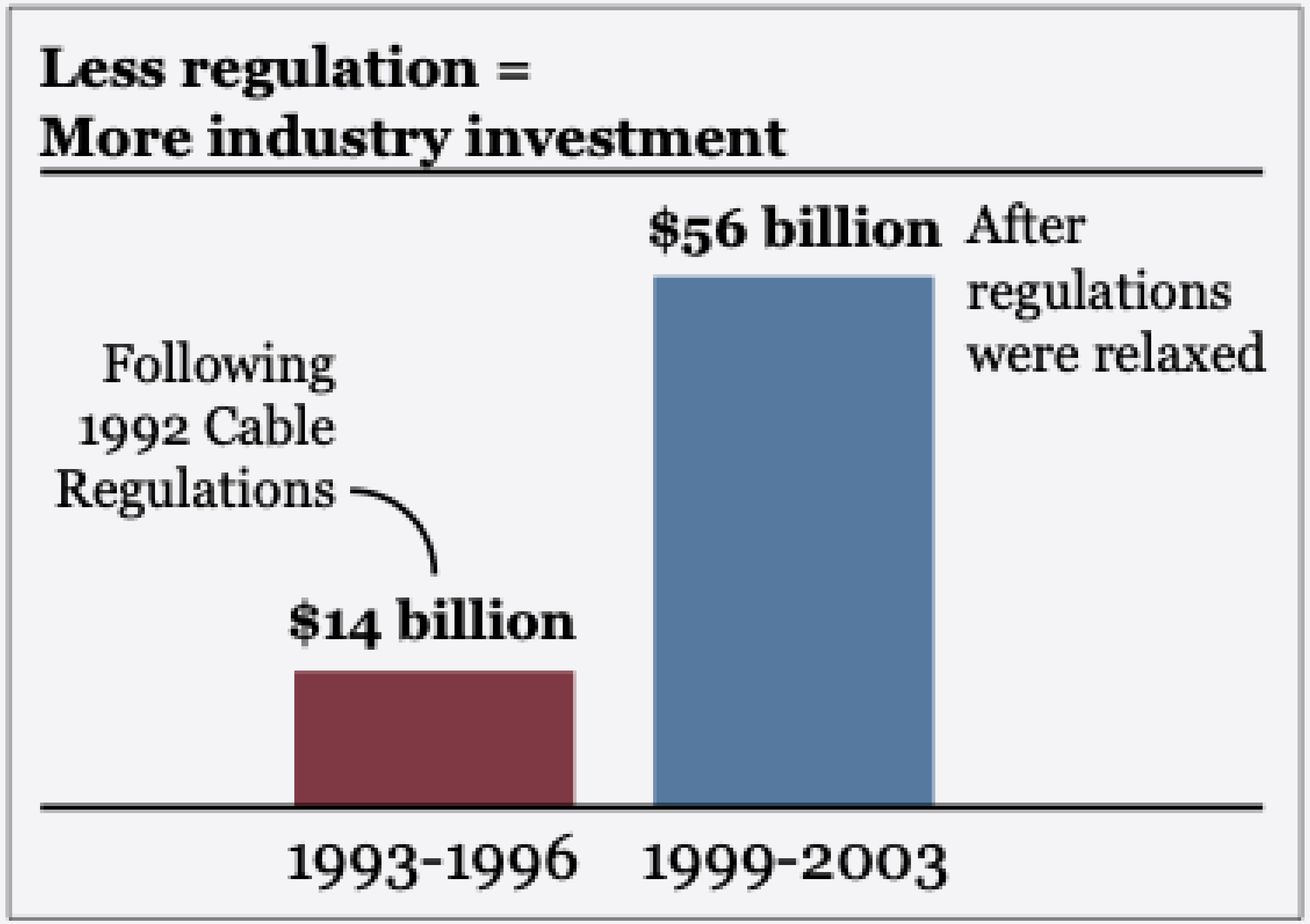


11.83 11.86 11.77 12.06 12.58 12.68 12.79 12.90 12.99 13.16 13.19 13.22 13.29
2001
Exceeded 12 million people
ople
1985 1990 1995 2000 2005 2006 2007 2008 2009 2010 2011 2012 2013

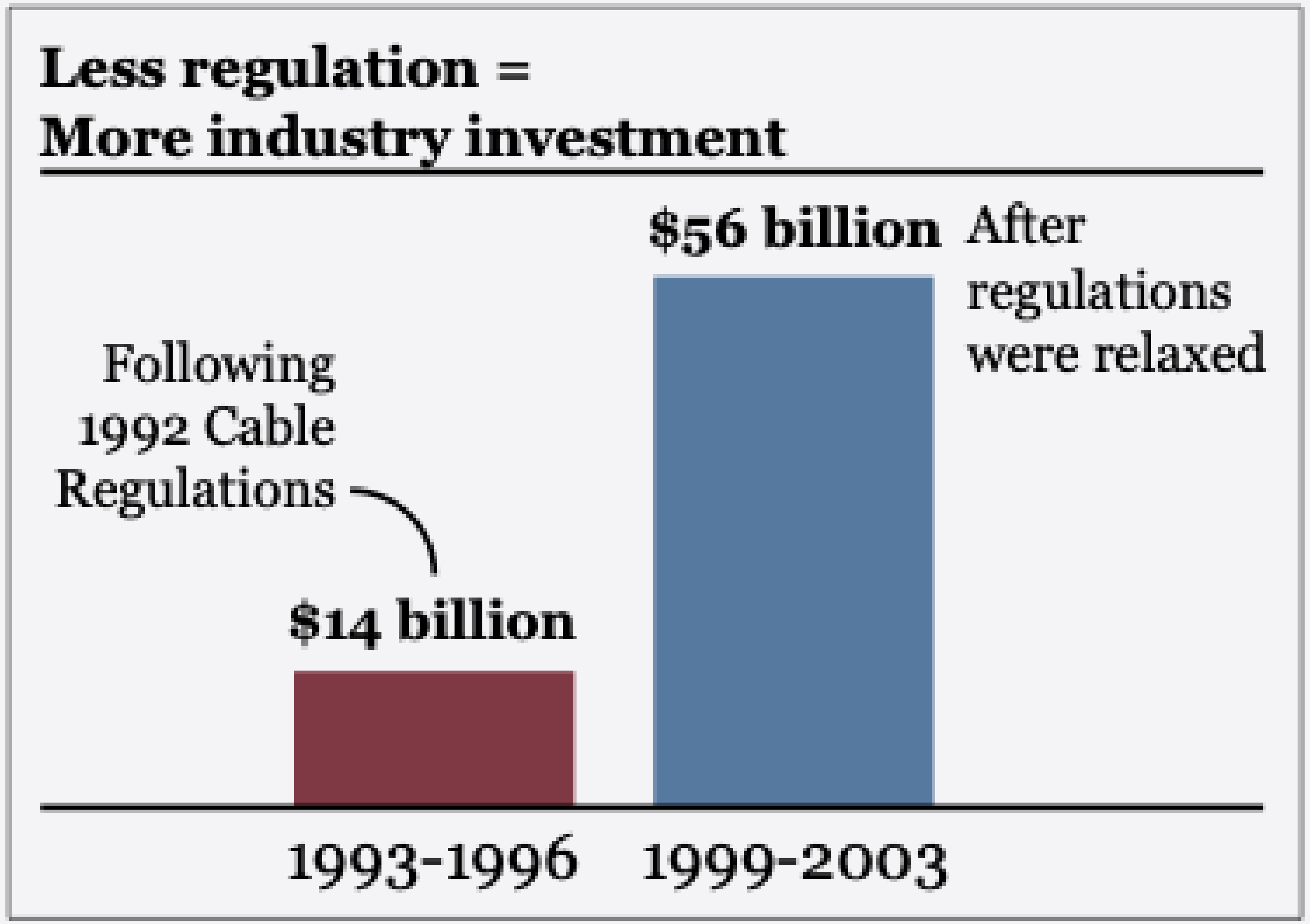


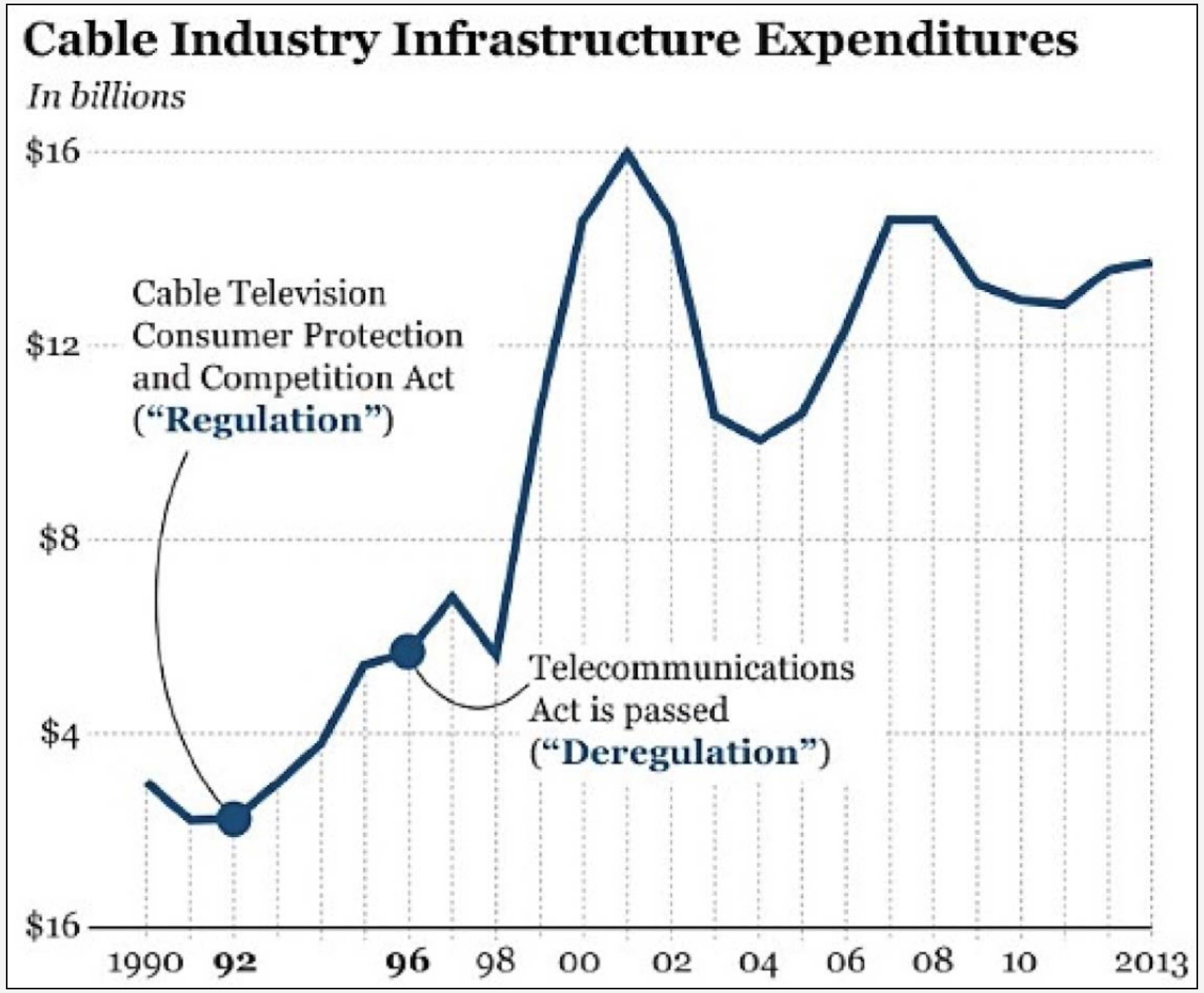
11.83 11.86 11.77 12.06 12.58 13.16 13.29
2001
Exceeded 12 million
ople
1985 1990 1995 2000 2005 2010 2013

Hacer agregaciones equivalentes



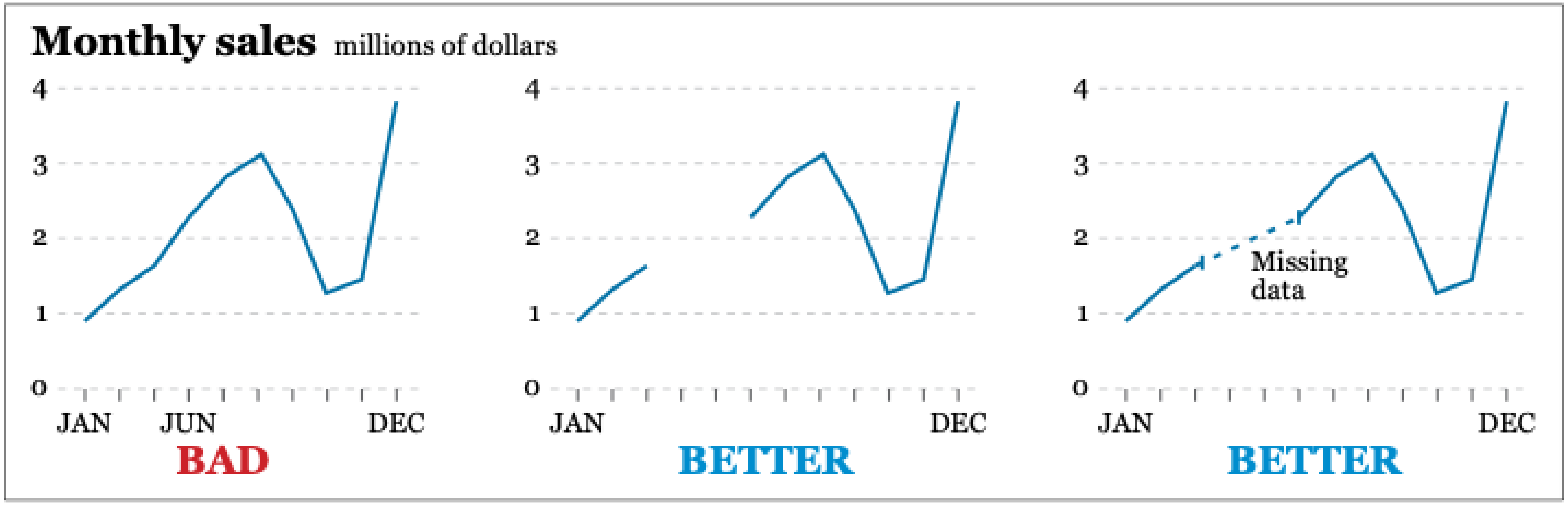
Hacer agregaciones equivalentes



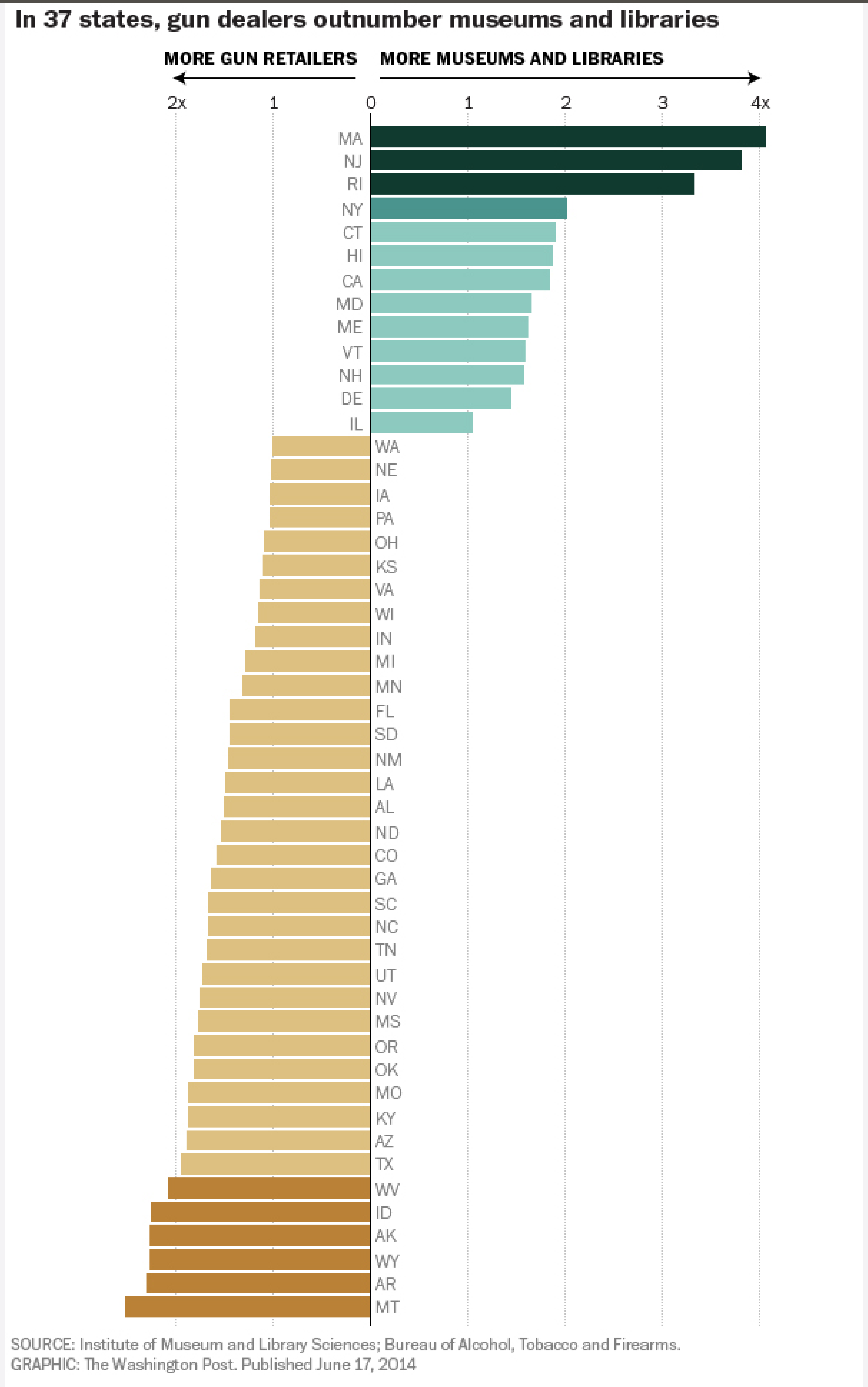


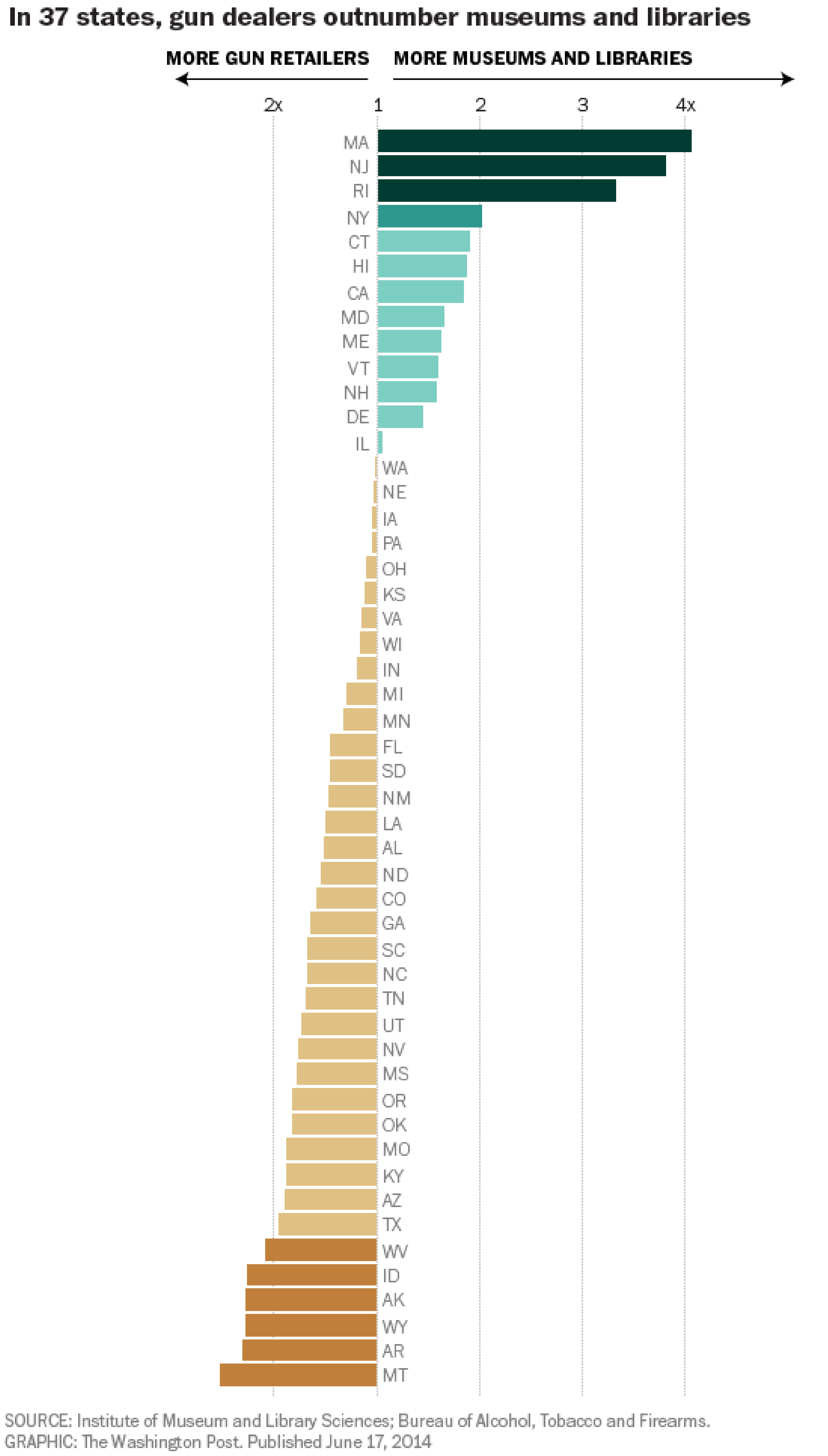
Missing data

- Marcar explícitamente donde no hay datos
- Mantener el espacio de las marcas / usar marcas distintas



Baseline





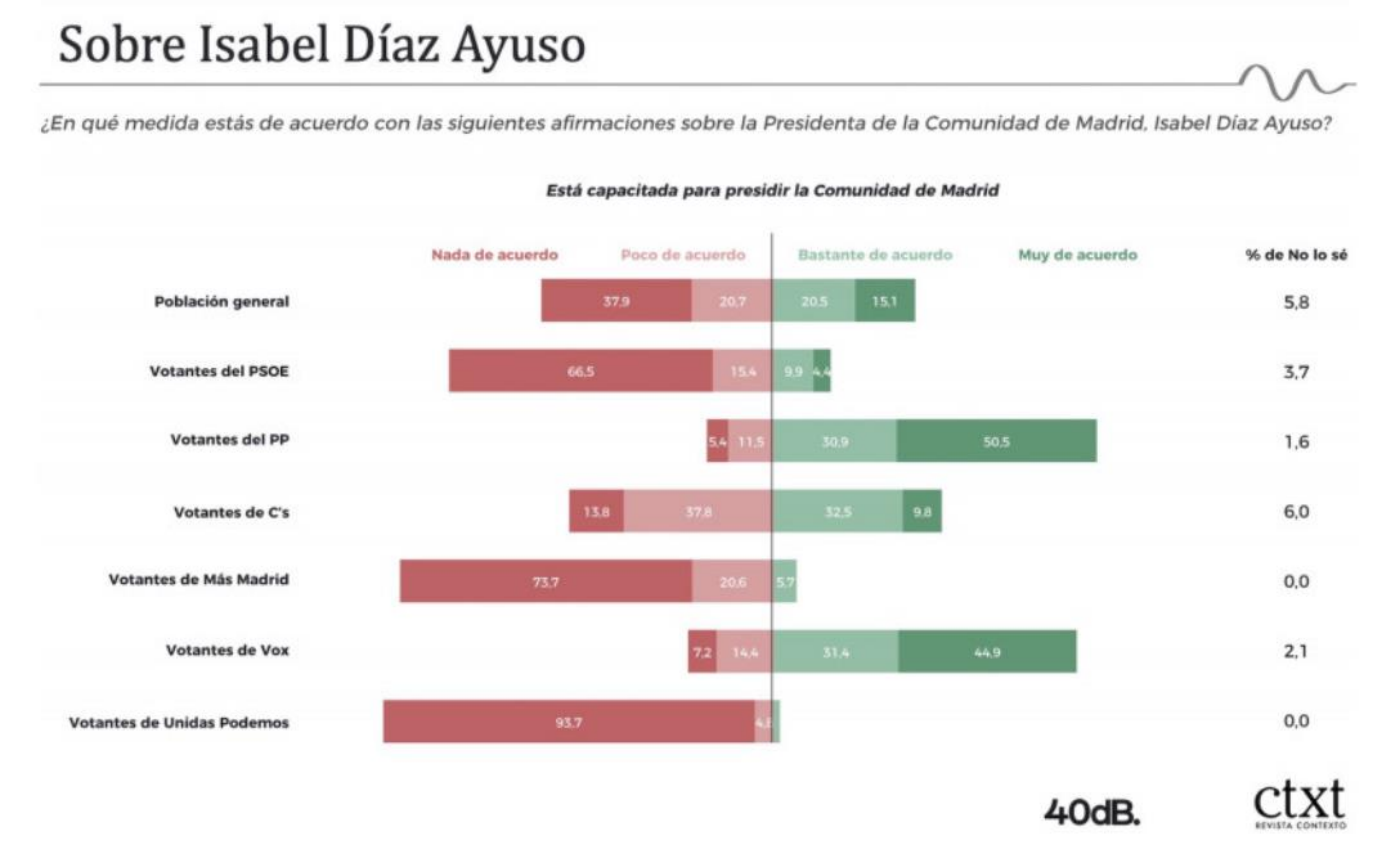
washingtonpost.com

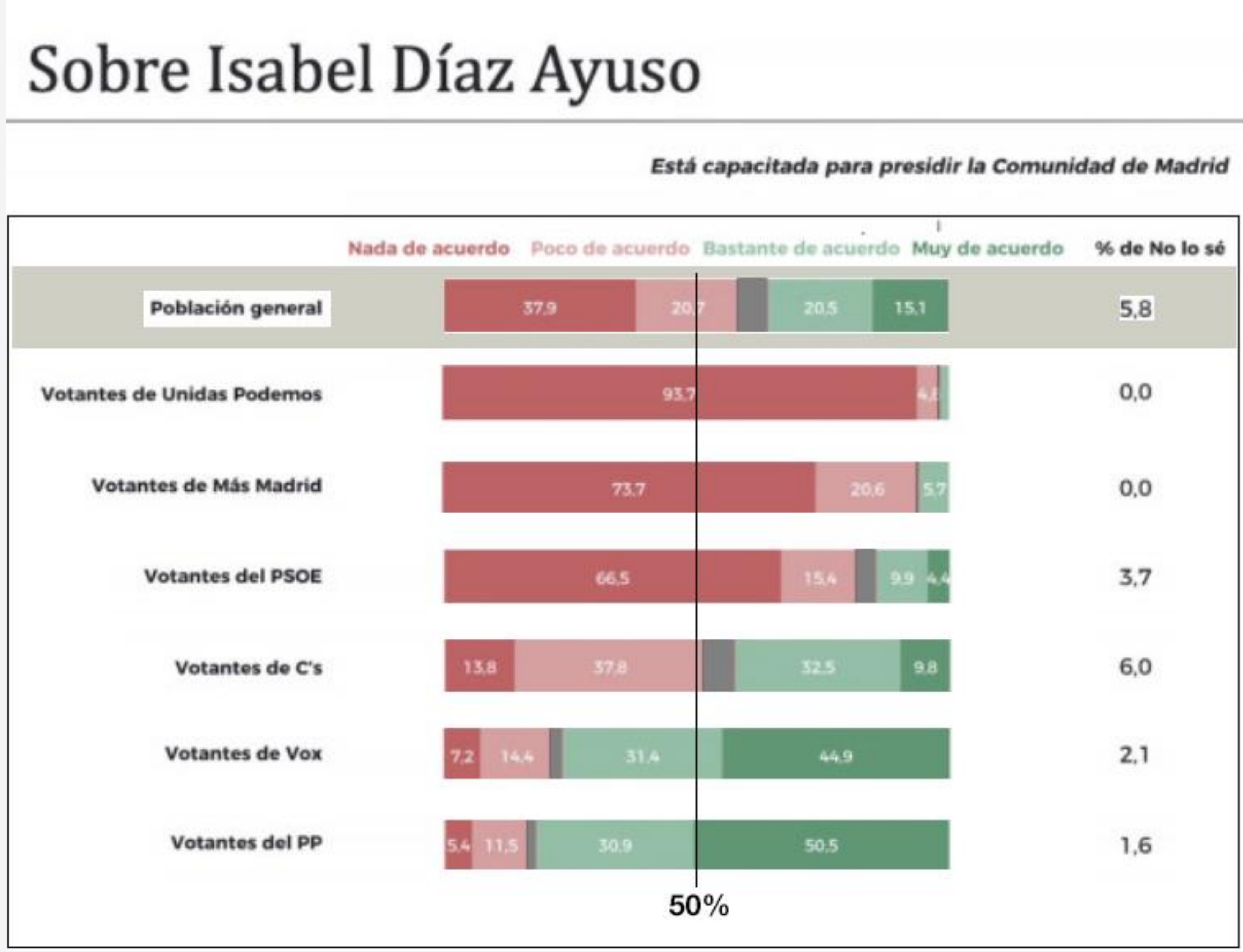
Stacked bars no alineadas

- Problemático. Difícil comparar porcentajes
- Difícil identificar 50%

Más del 50% de los electores de C's suspenden a la presidenta de la Comunidad y al Gobierno regional de su propio partido. El 51,1% de los madrileños, a favor de un cambio. Sánchez sale peor parado que la líder popular

ctxt 6/02/2021

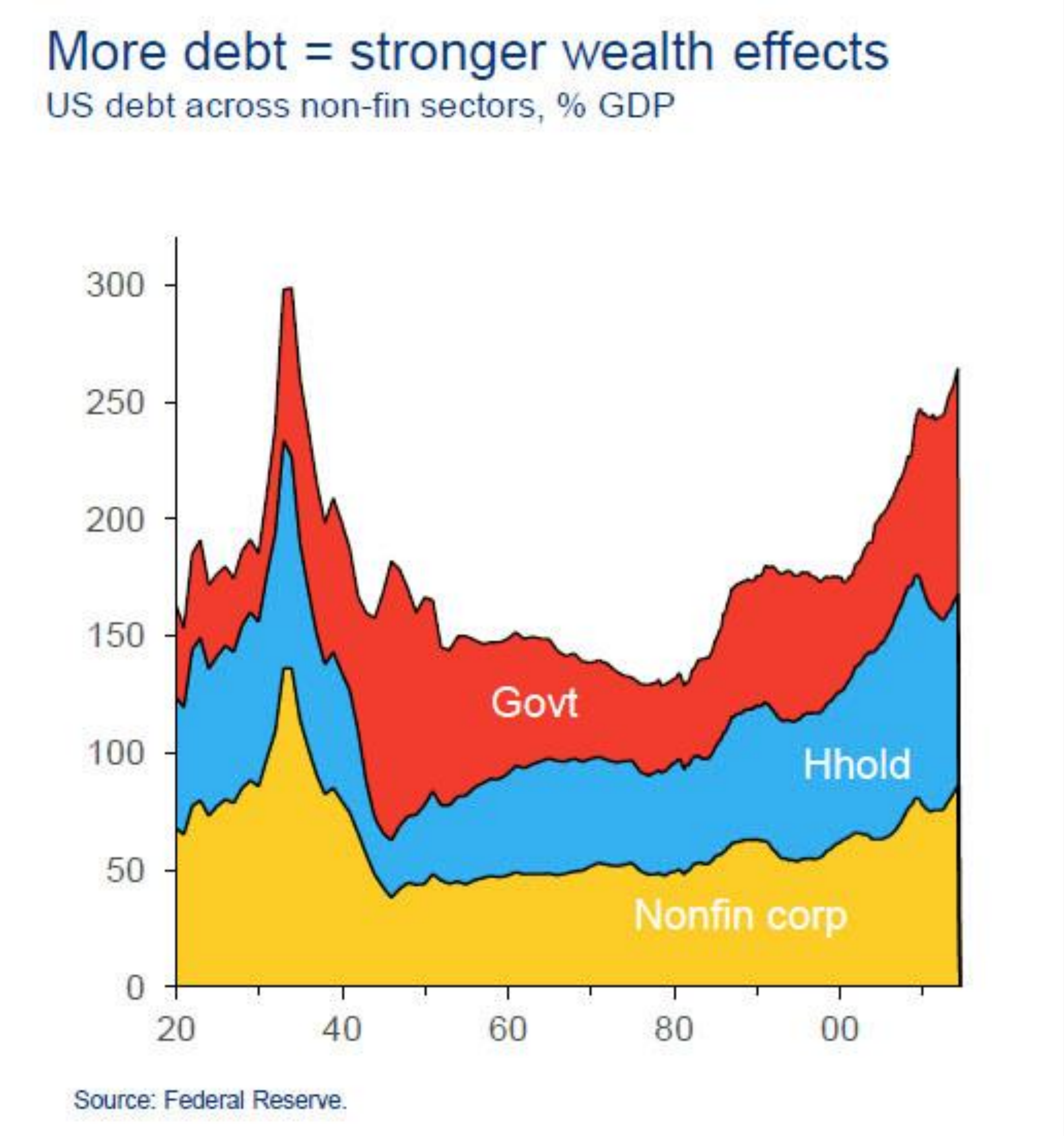




Precisión en area charts

- Mensaje: Niveles de deuda actuales similares a 1930s pero…

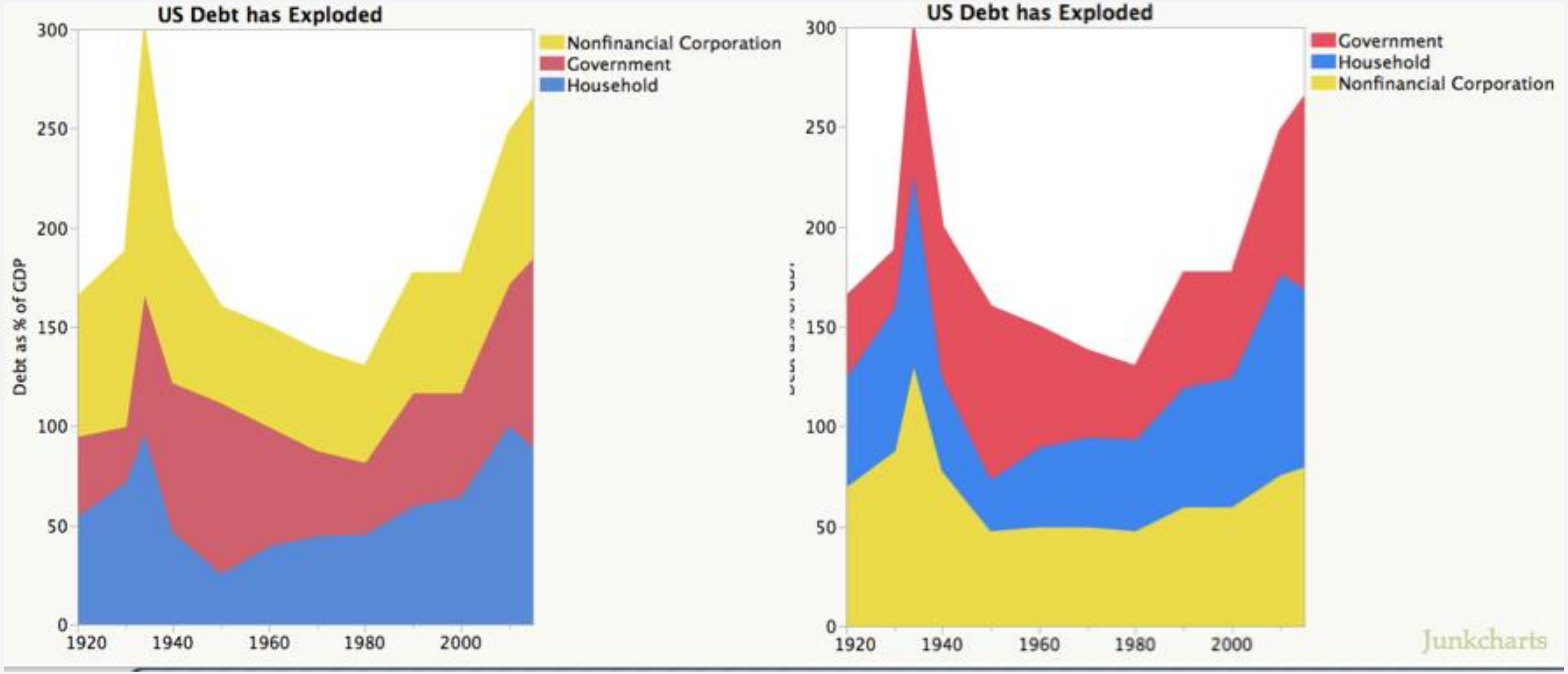
- Muy difícil identificar que categoría es responsable del aumento.
- ¿Otras relaciones entre ellas?



www.junkcharts.typepad.com/

Precisión en area charts

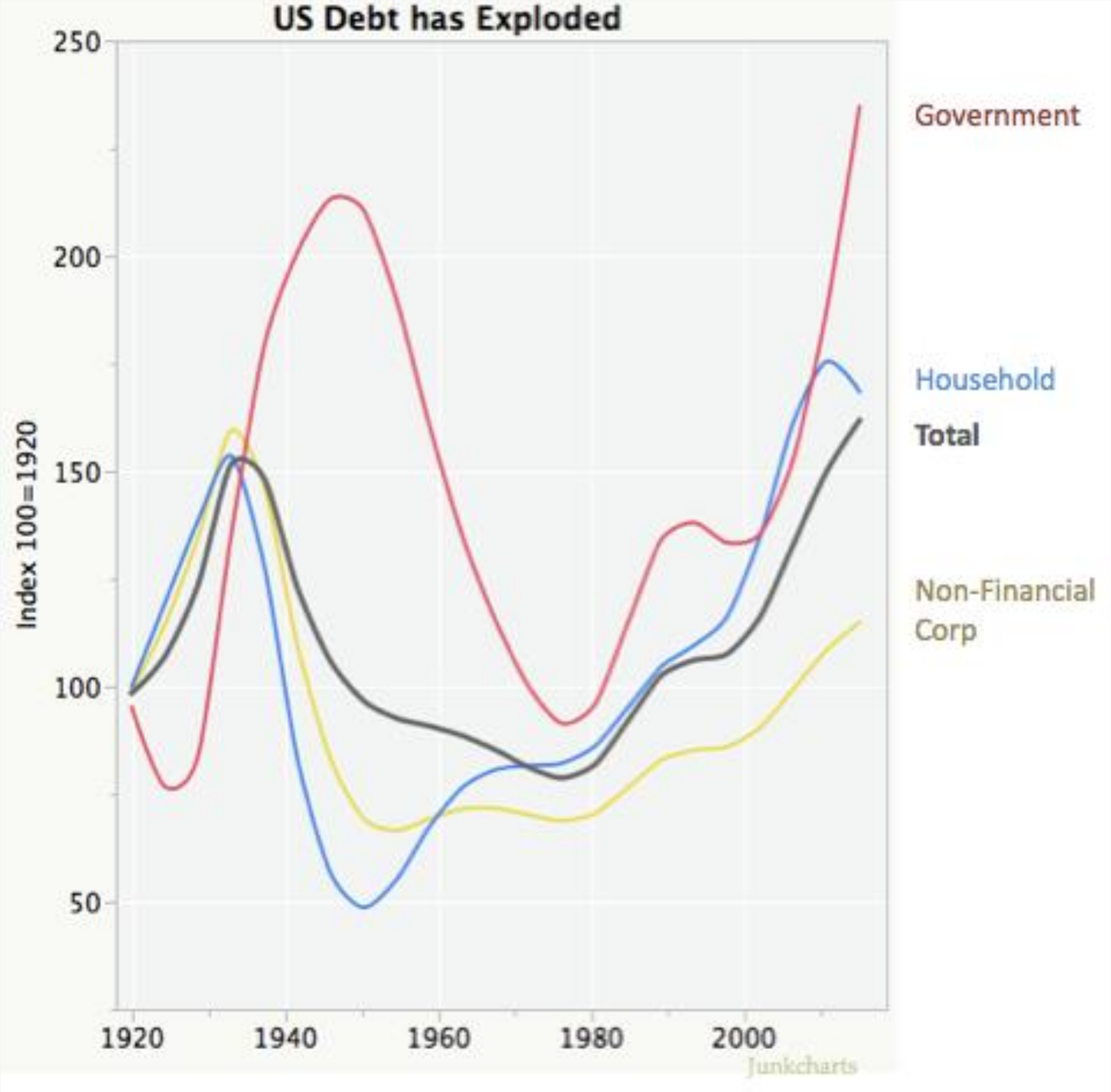
- Dos opciones de 6 posibles
- Picos en capas superiores tienden a magnificarse



www.junkcharts.typepad.com/

Precisión en area charts

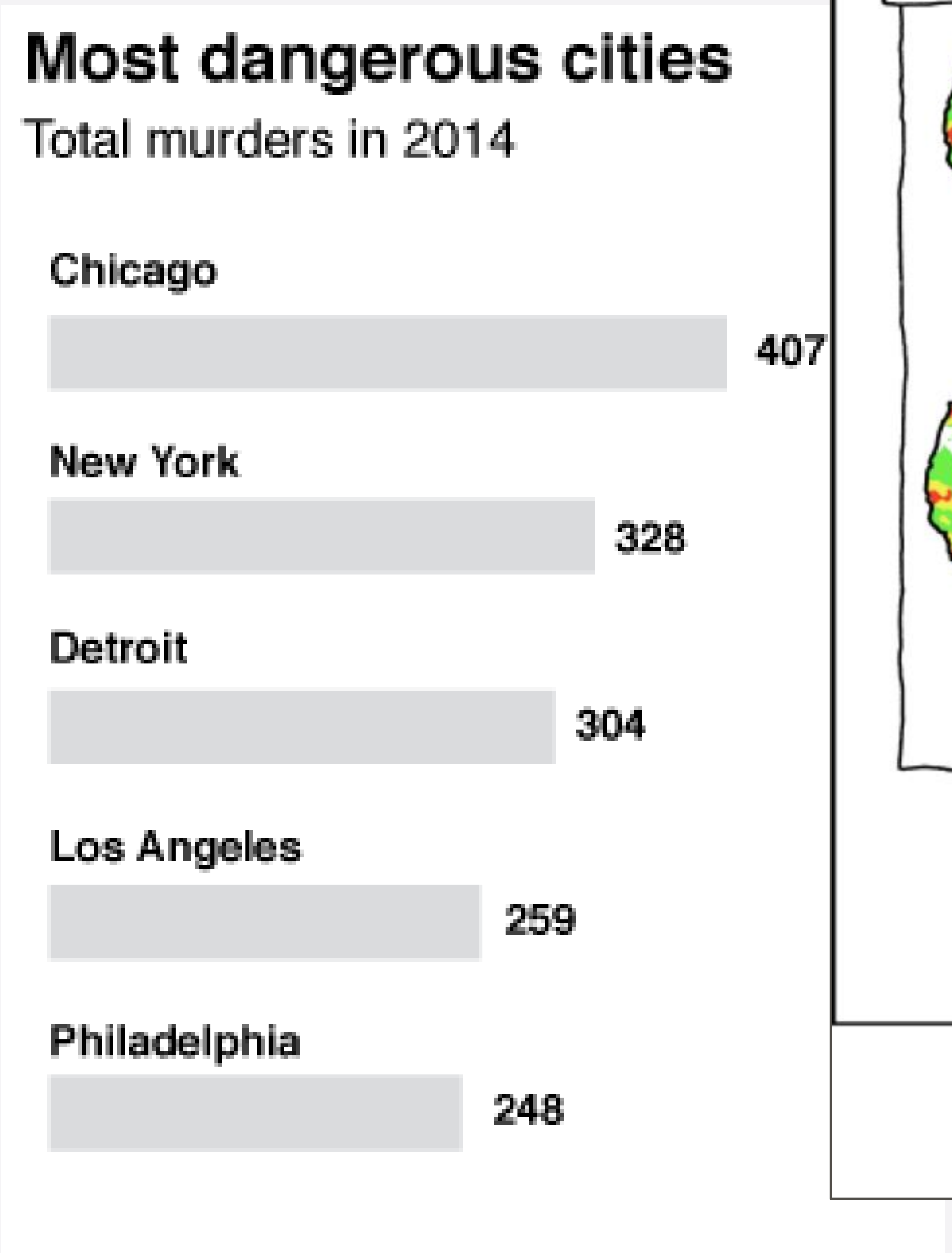
- Solución: no usar área chart
- Calcular índices usando la media de 1920 como referencia

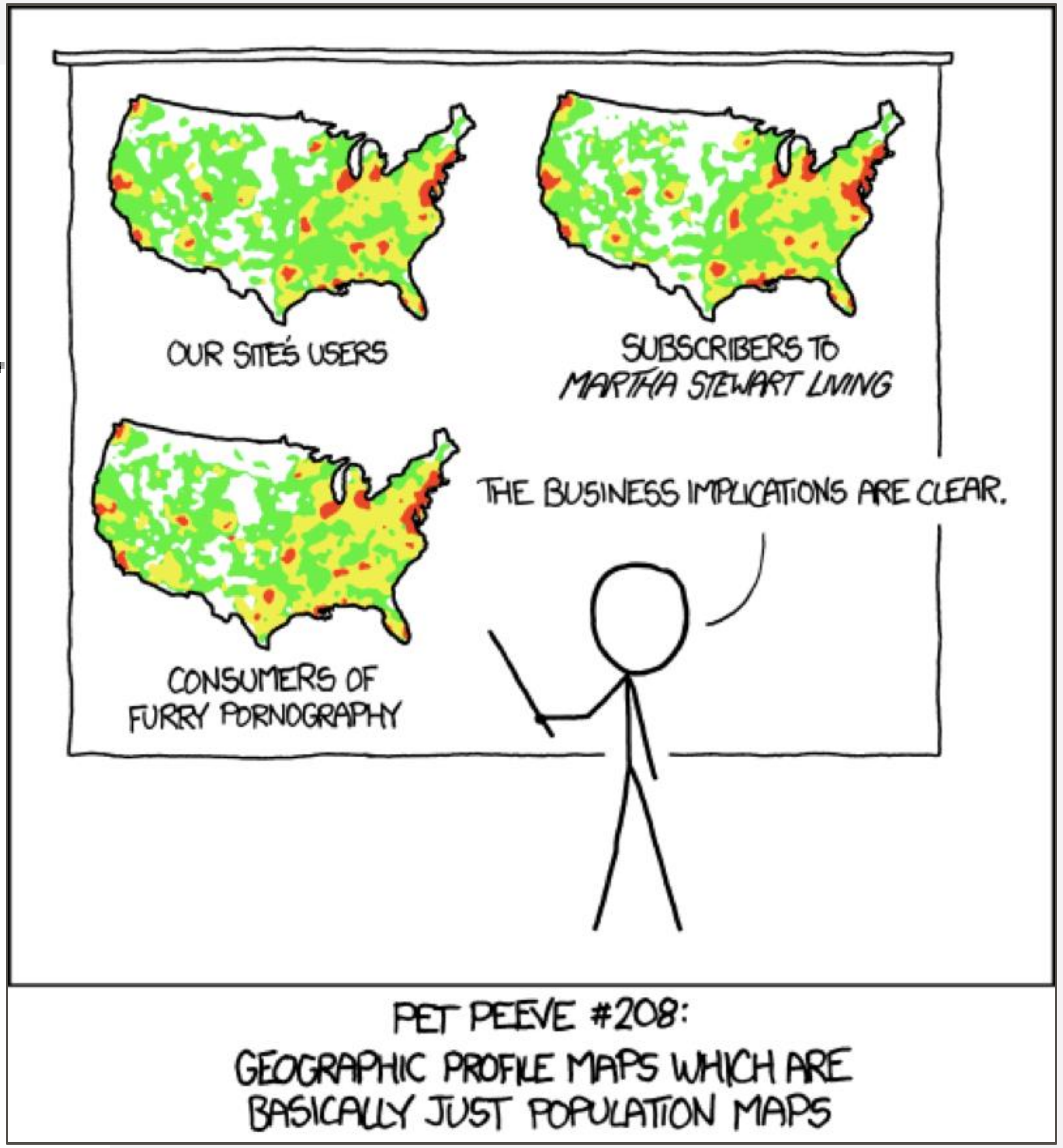


www.junkcharts.typepad.com/

Elegir métricas relevantes

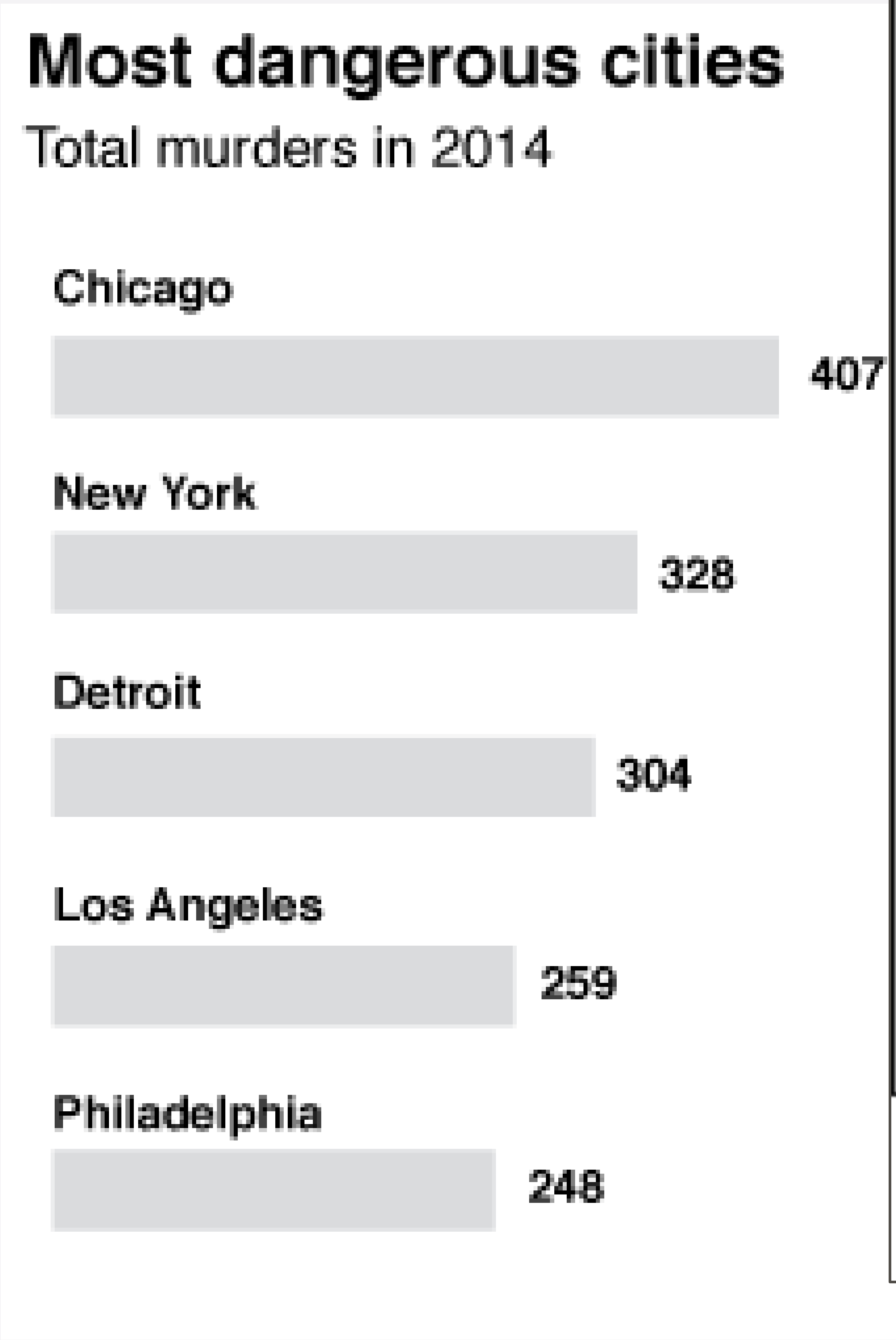
Elegir métricas relevantes

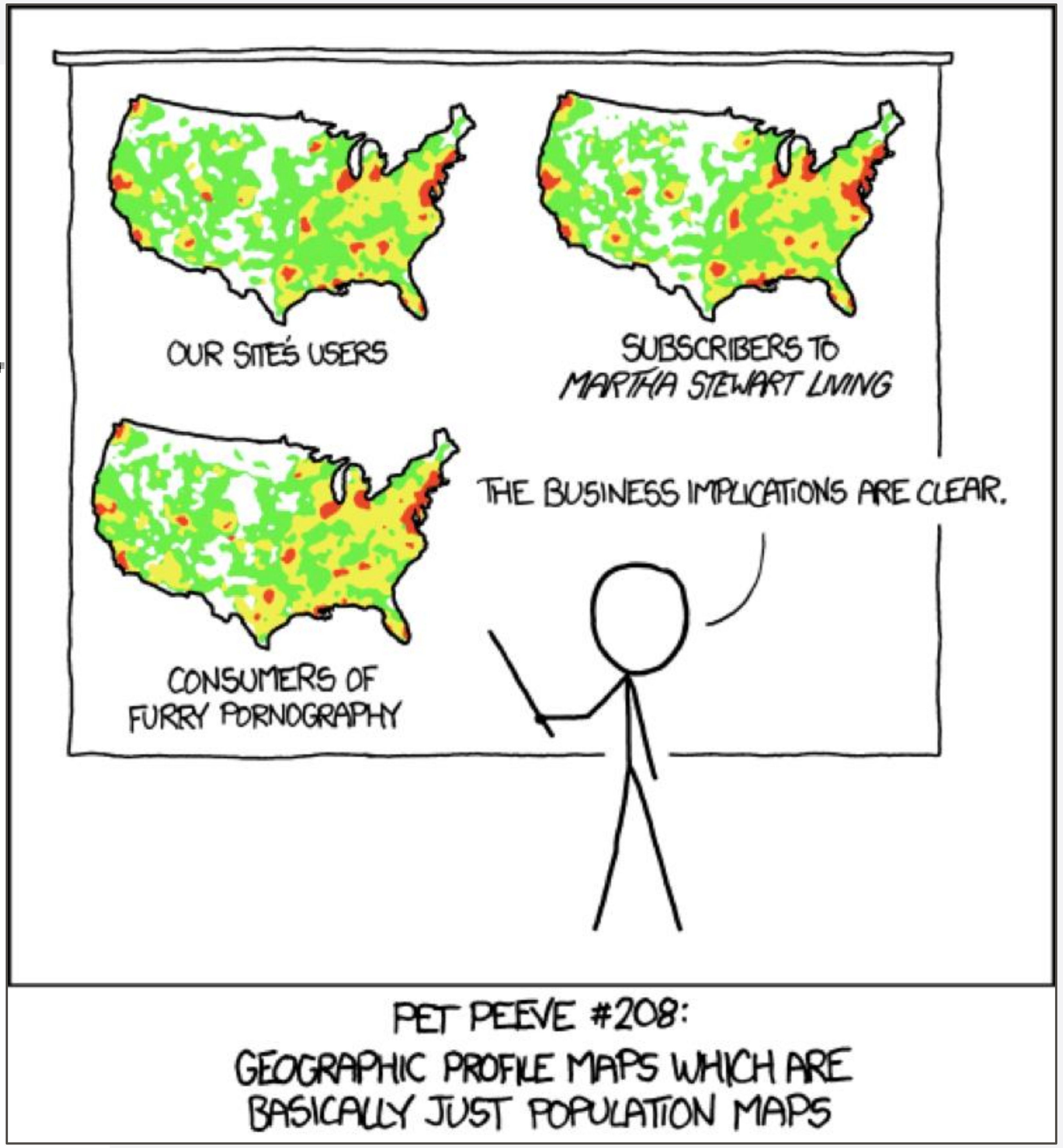


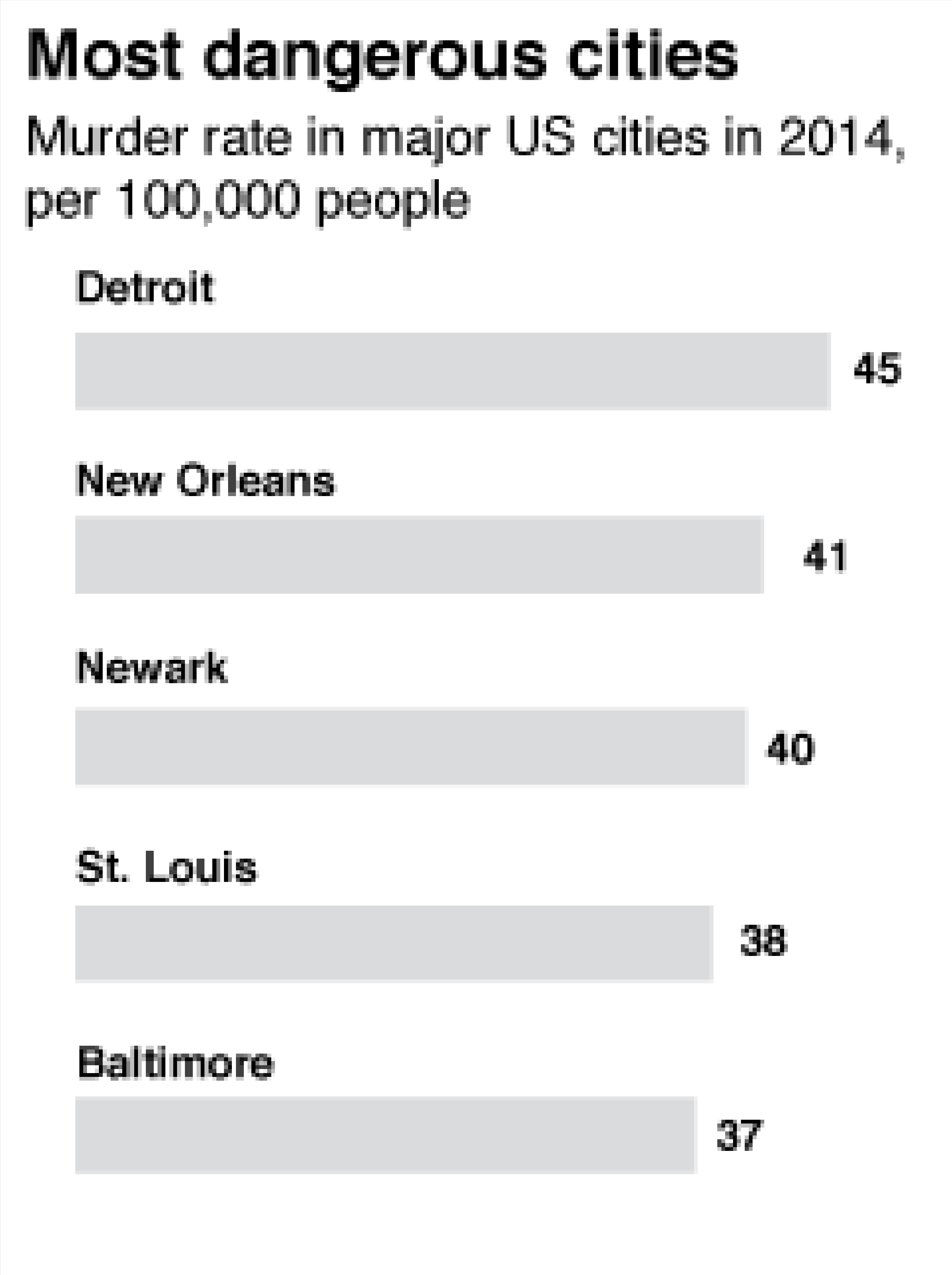


[https://xkcd.com/1138]

Elegir métricas relevantes

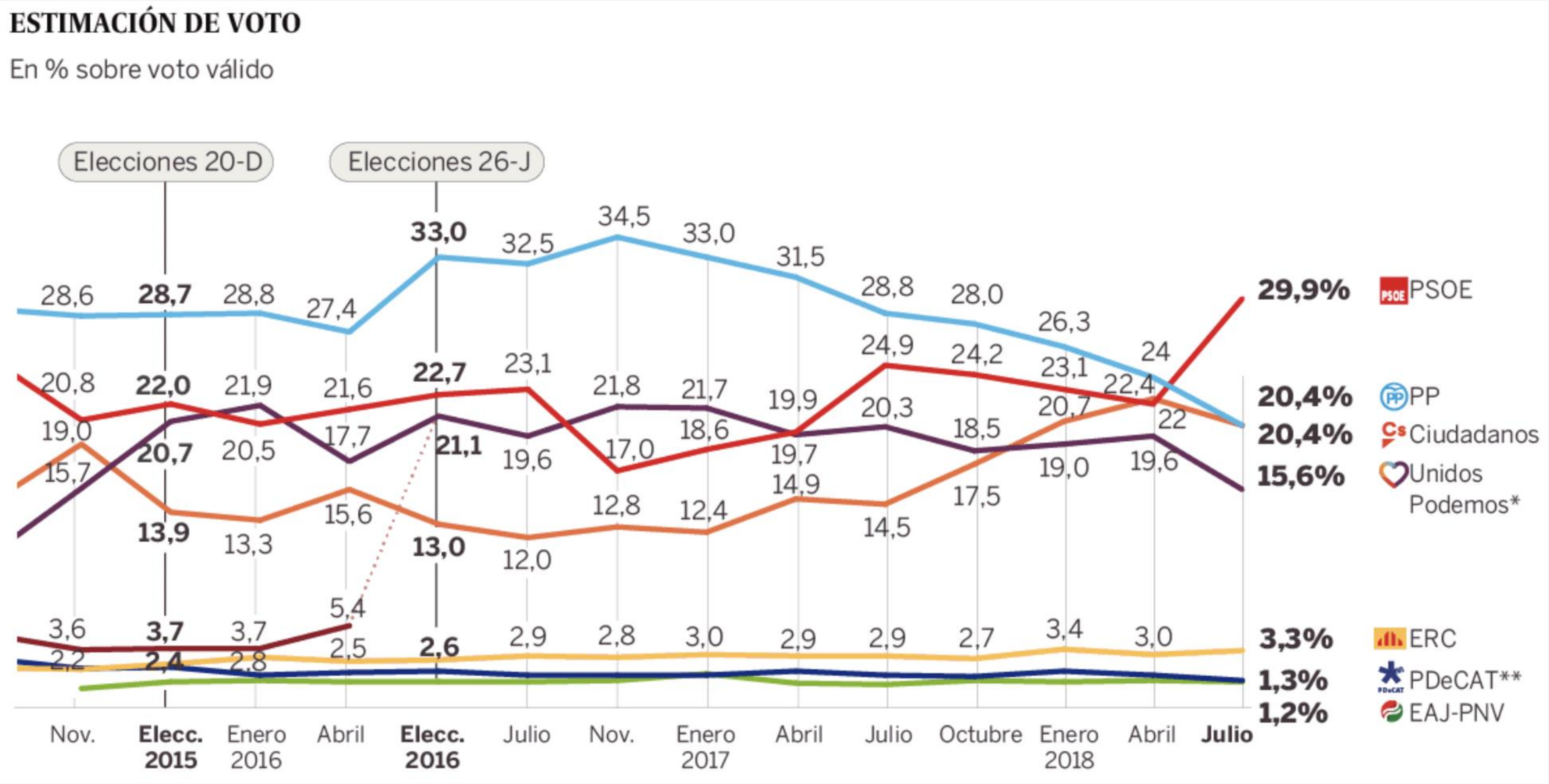






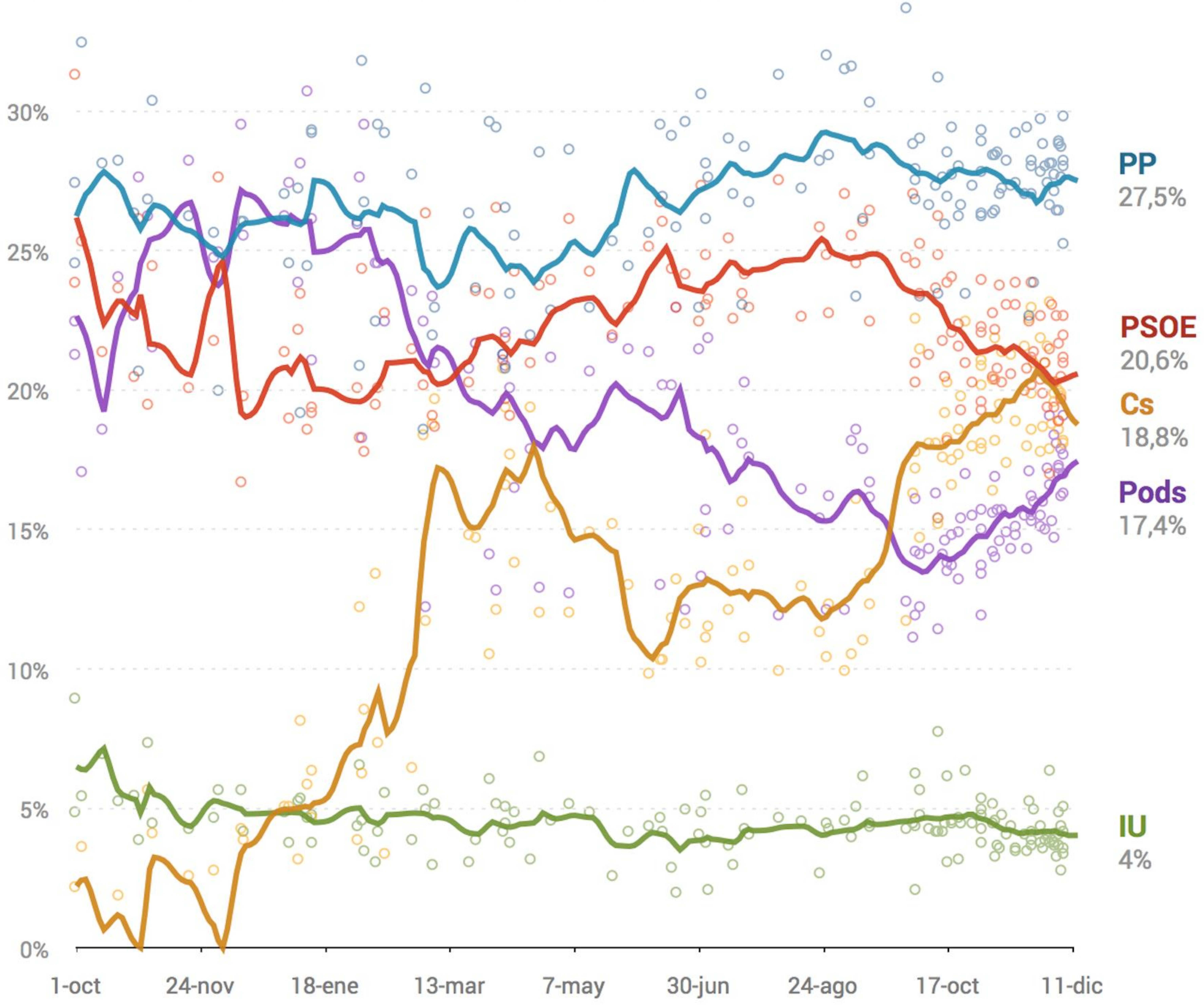
[https://xkcd.com/1138]

FICHA TÉCNICA
Sondeo efectuado mediante entrevista personal a 2.485 personas mayores de 18 años de ambos sexos en 256 municipios de 47 provincias entre el 1 y el 10 de julio. Nivel de confianza: 95,5%. Margen de error: ± 2,0 puntos.



El Pais

Porcentaje de voto según las encuestas. Las líneas representan un promedio ponderado por fecha, tamaño de muestra y empresa encuestadora.

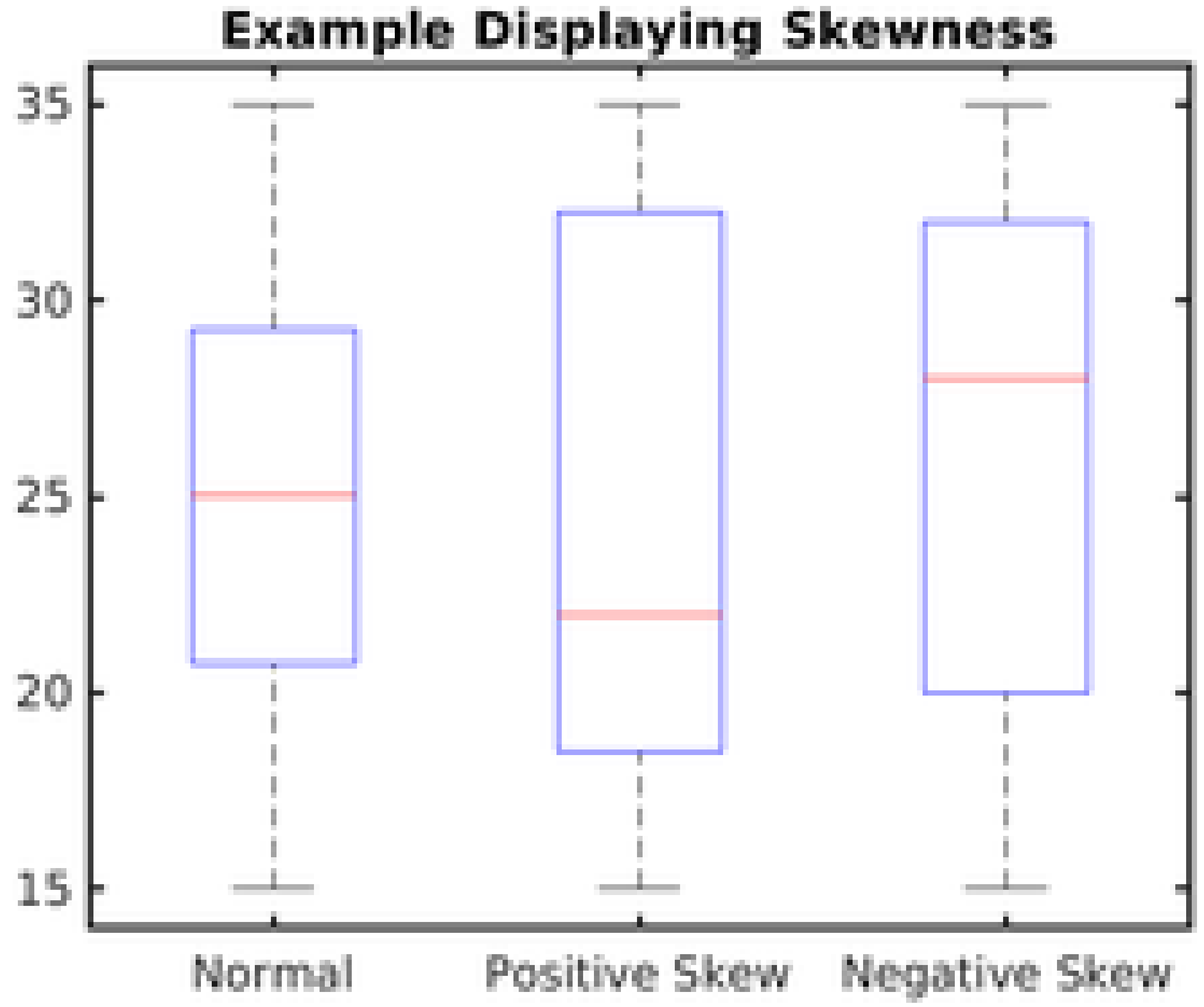


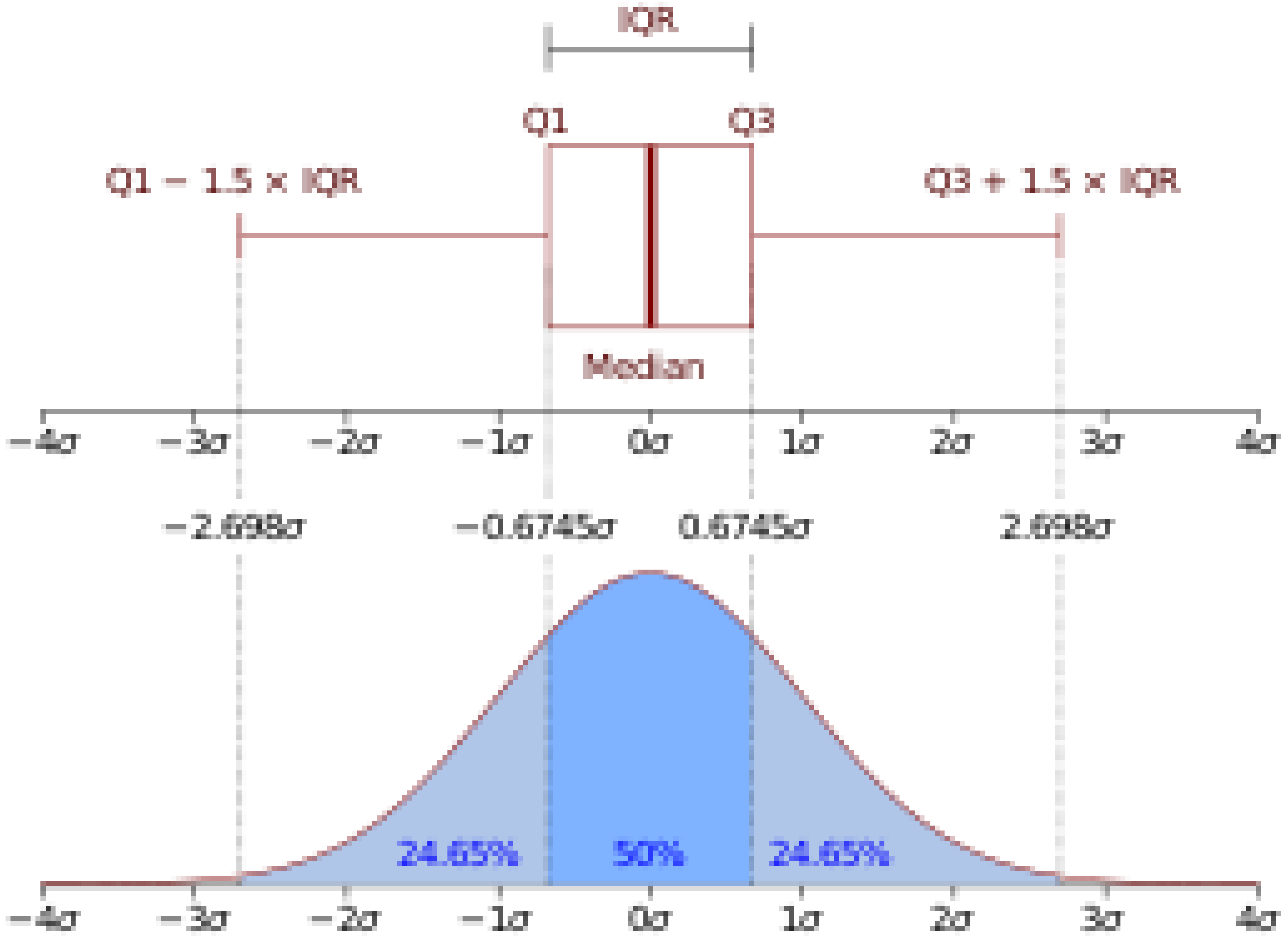
Fuente: Electograph. Fecha actualización: 14 dic. Kiko Llaneras | EL ESPAÑOL

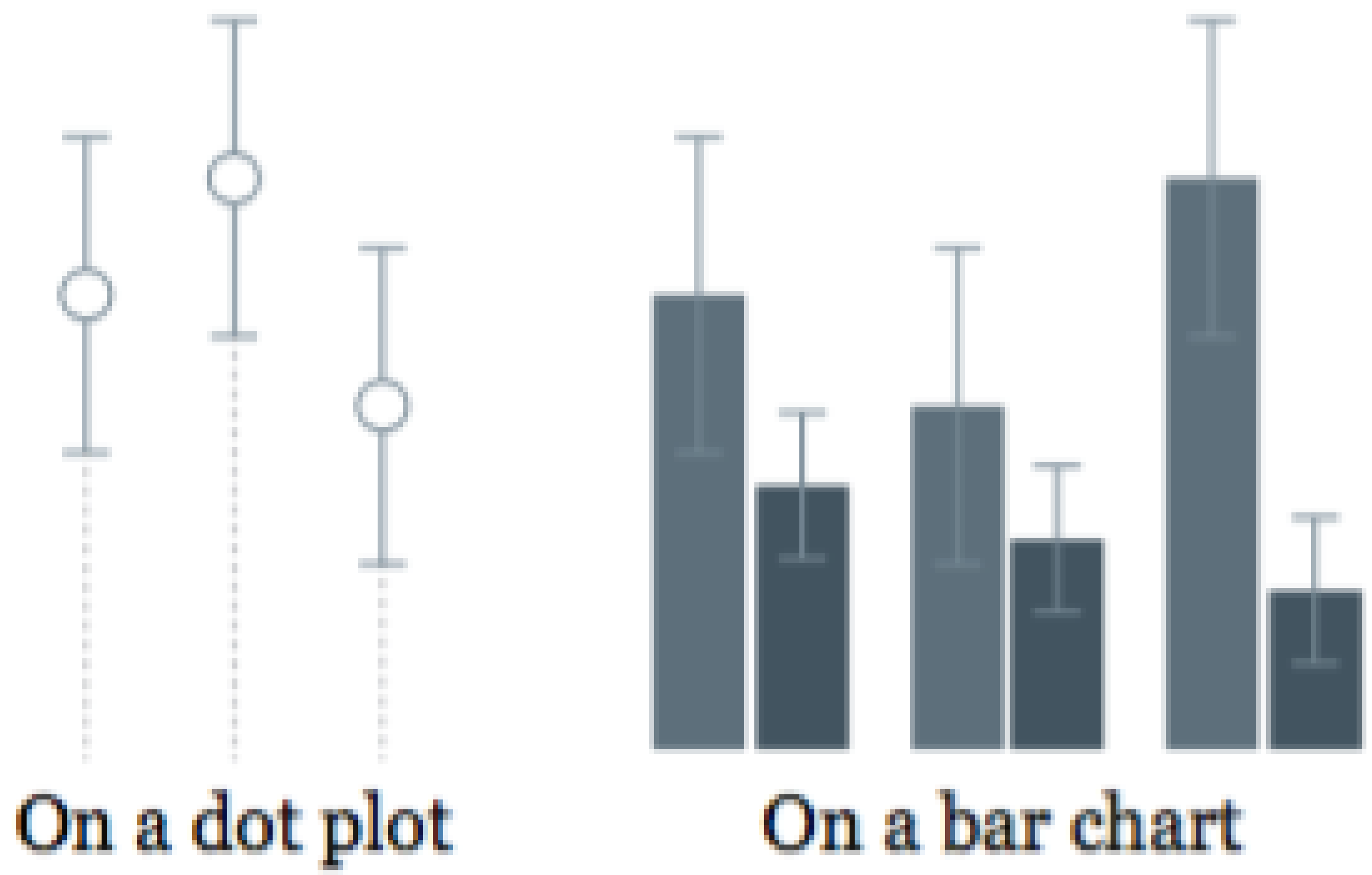
Box and whisker plots

Representación visual de Distribución o:

- error
- intervalos de confianza
- Desviacion standard
- Uncertainty
- …



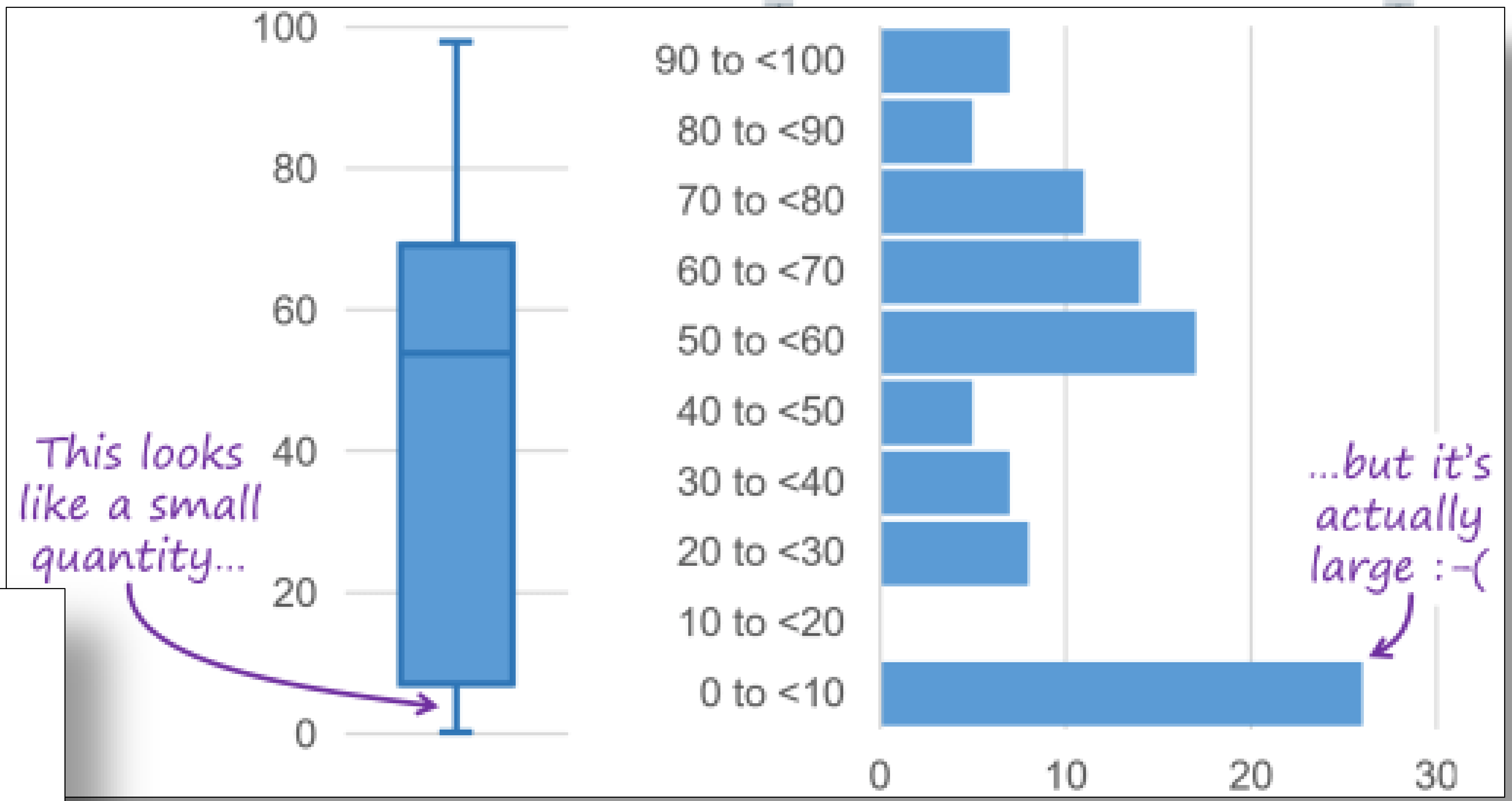


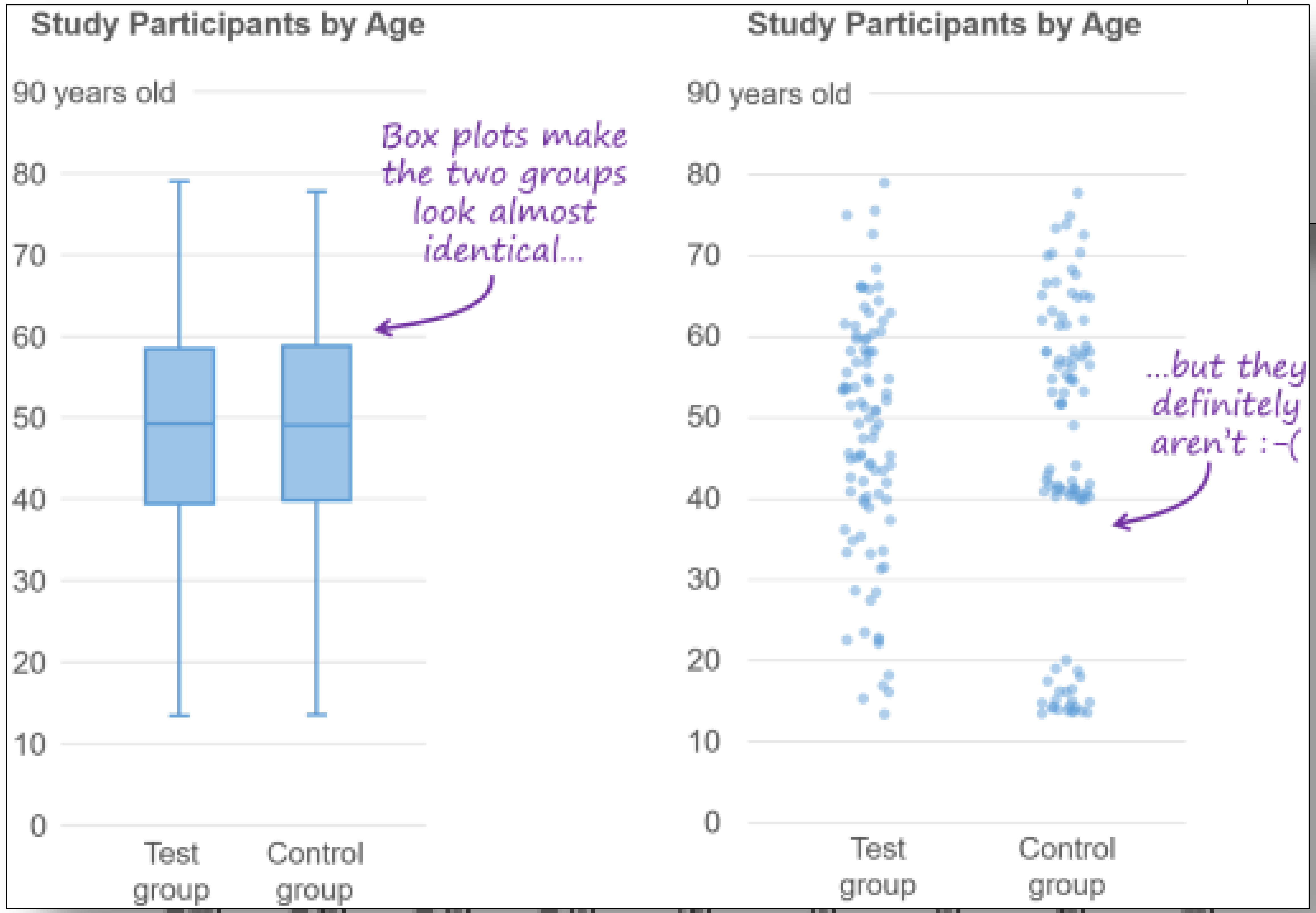


Box and whisker plots

Limitaciones:

- Box se percibe como unidad
- No muestra toda la distribución
- Asociamos tamaño a cantidad

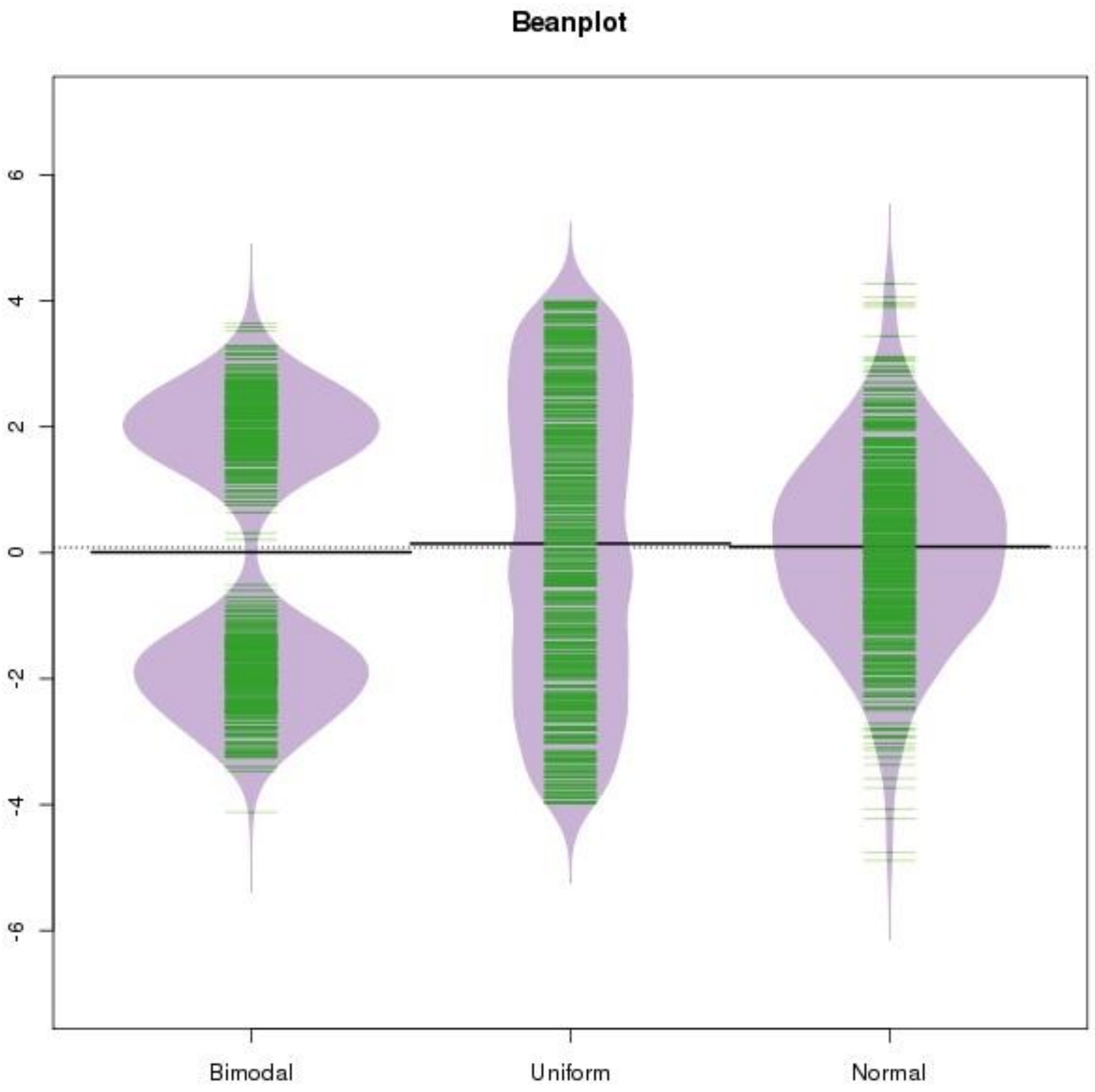




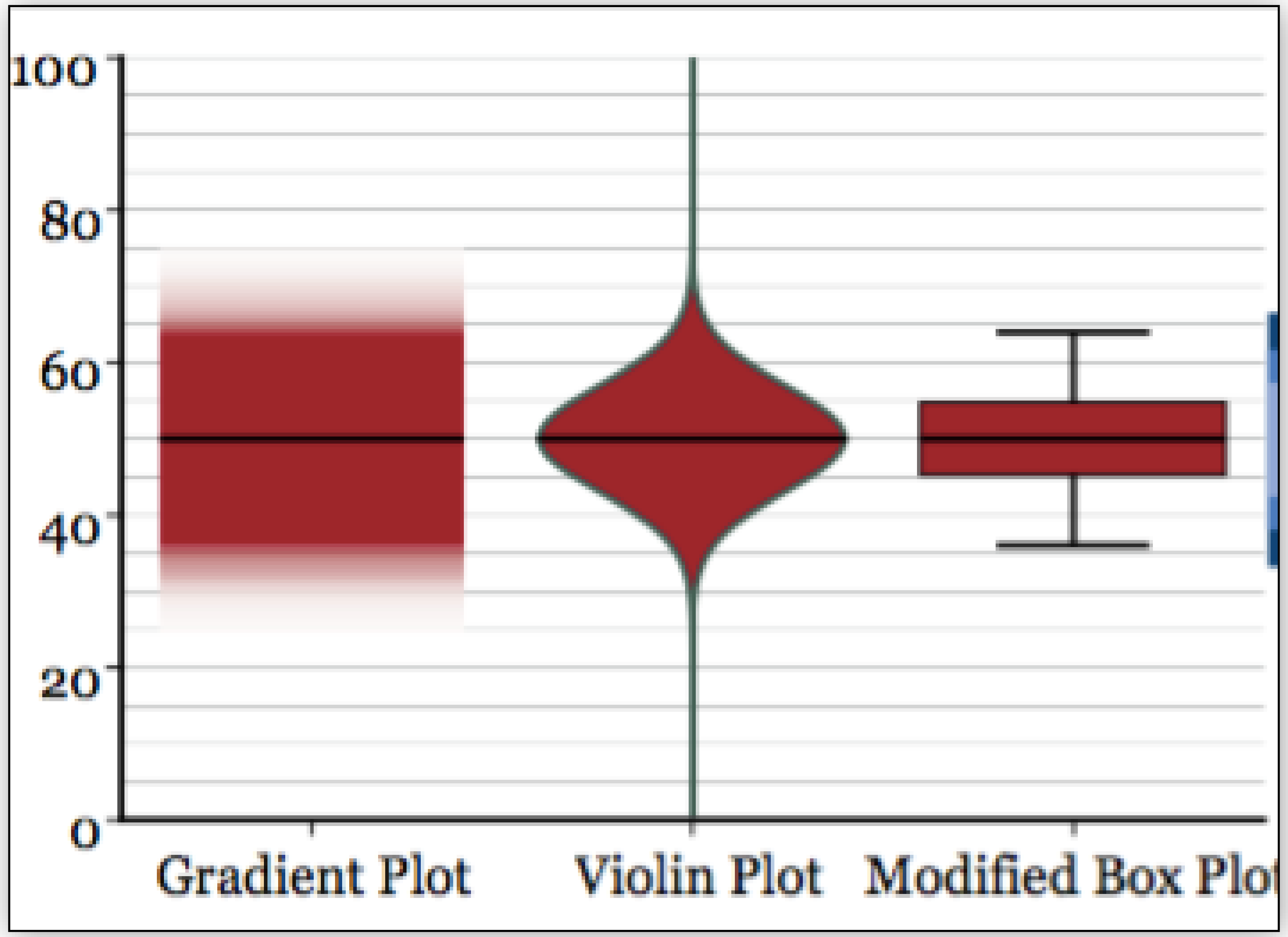
Box and whisker plots

Alternativas:

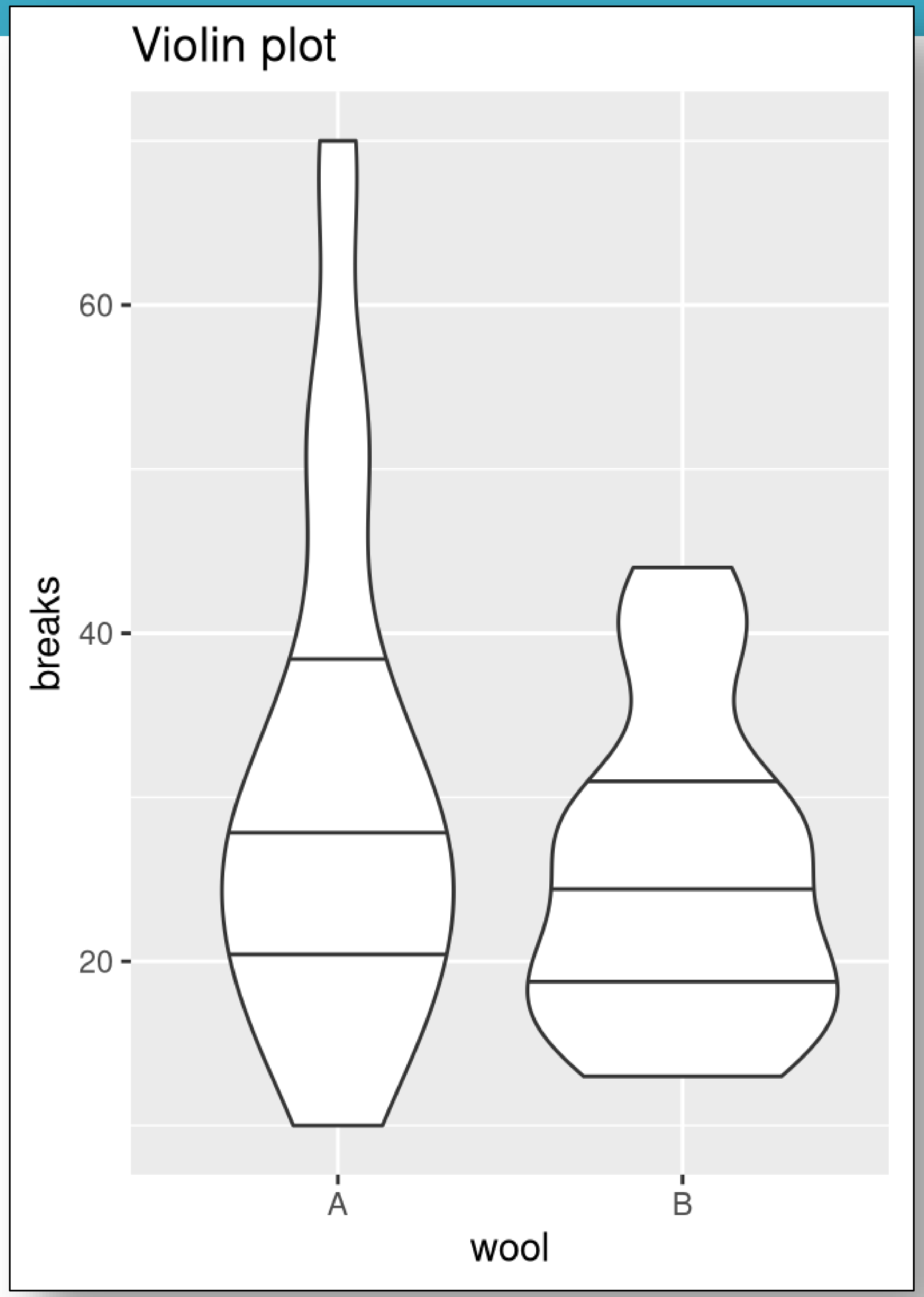
- Violin, gradient, Bean plots, …

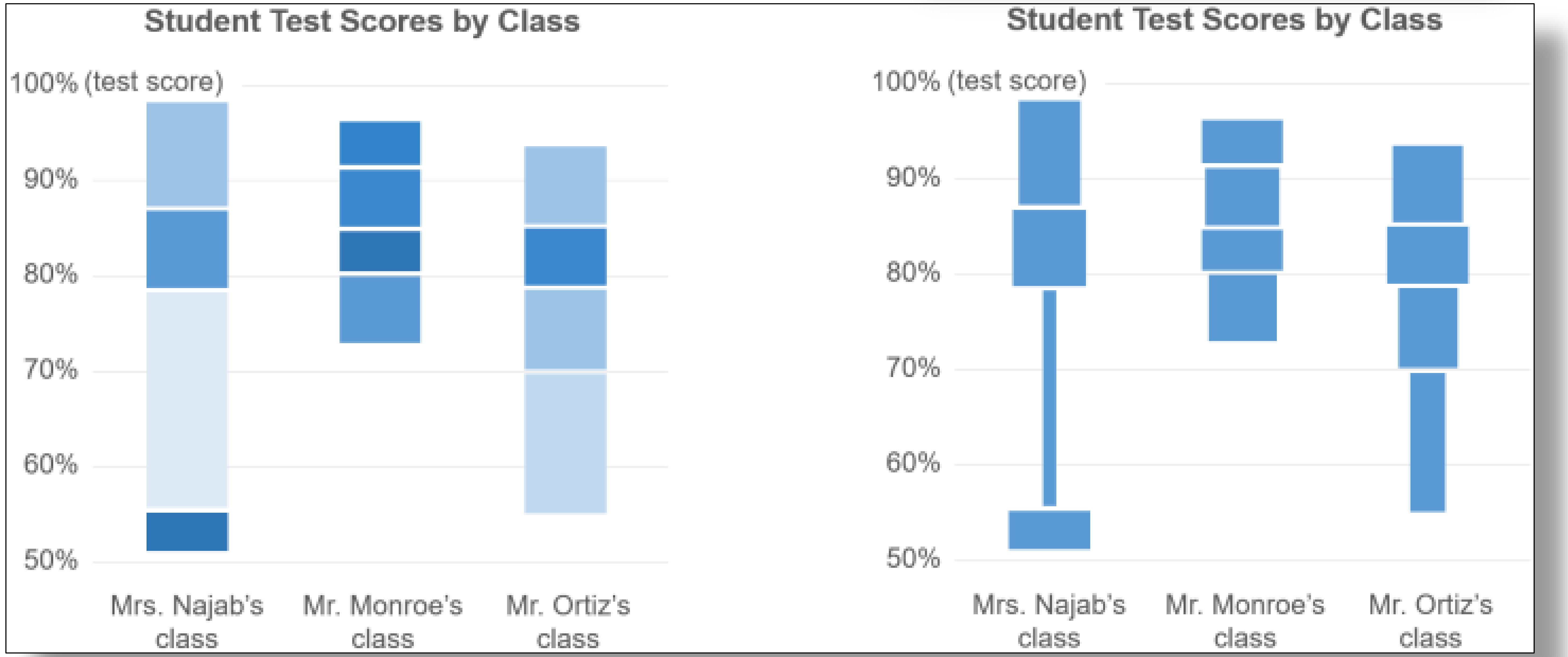


Peter Kampstra, 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, 28



Error bars considered harmful. Correll M., Gleicher M.

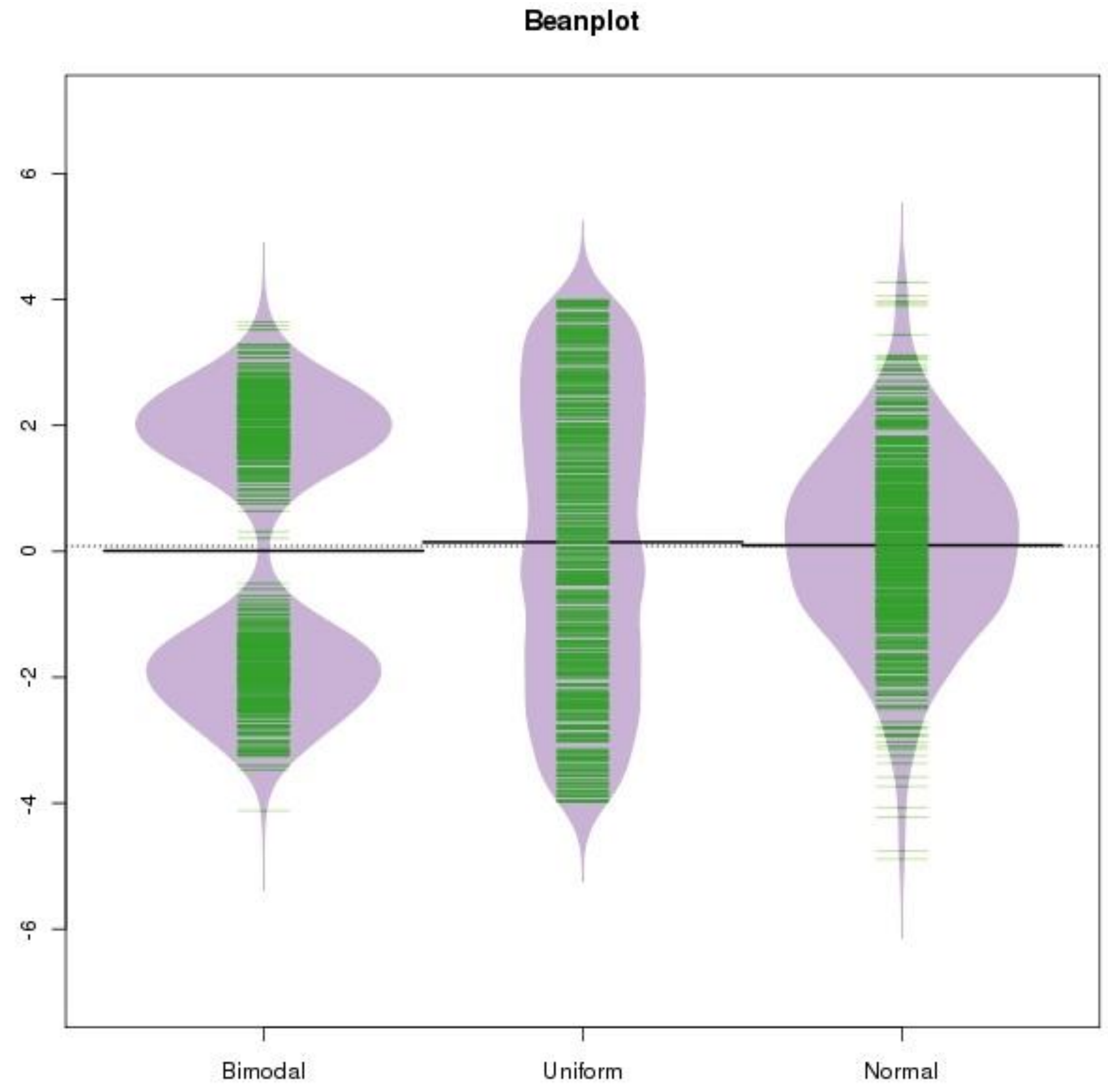




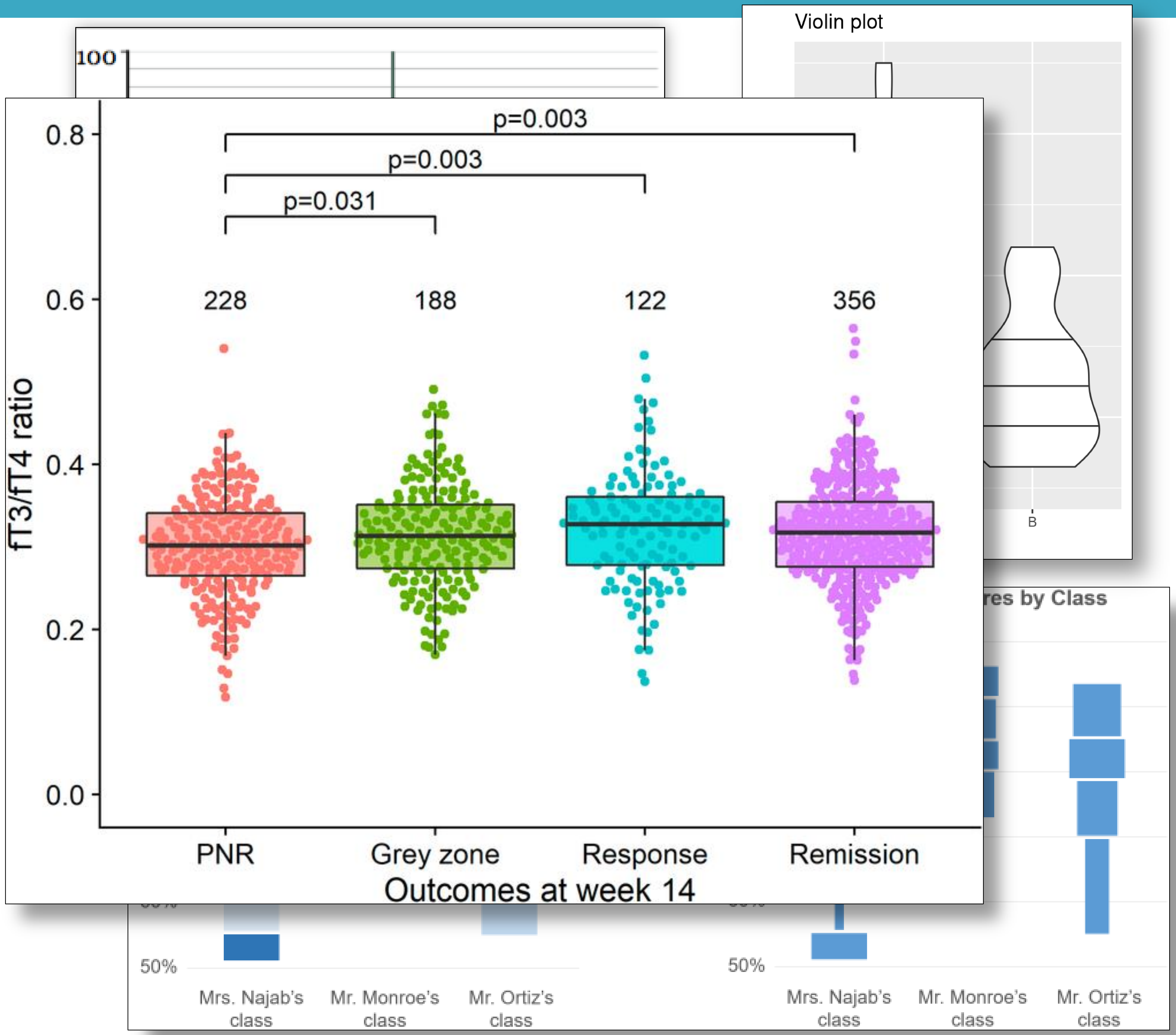
Box and whisker plots

Alternativas:

- Violin, gradient, Bean plots, …
- **Beeswarm**

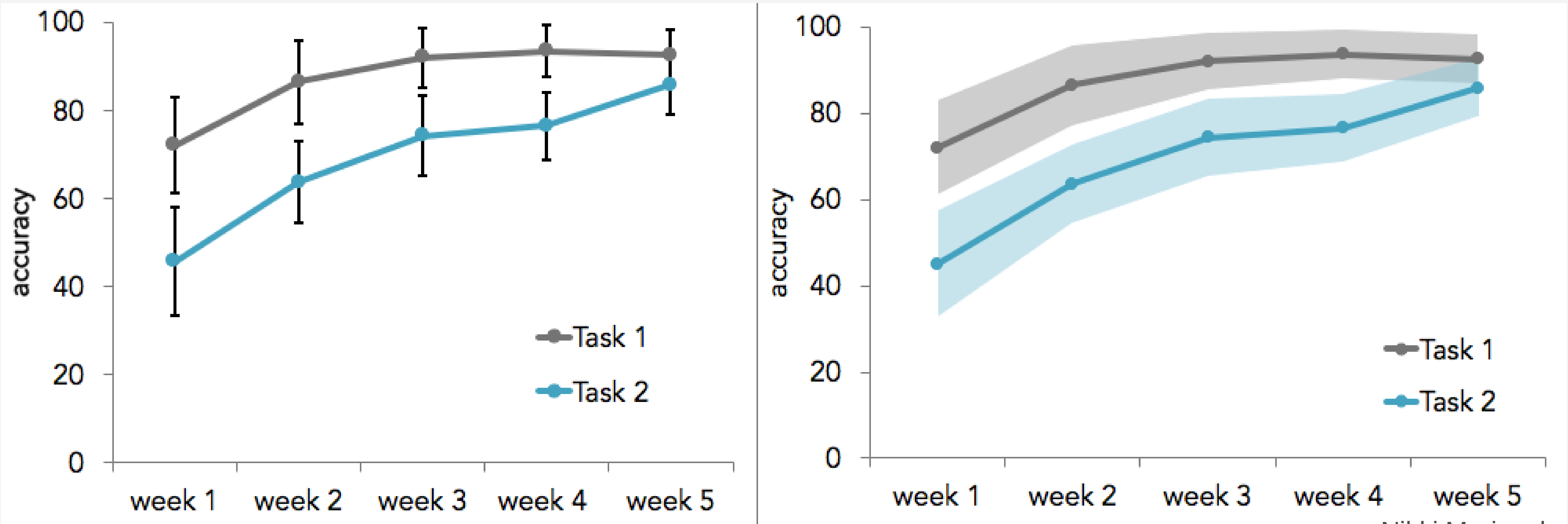


Peter Kampstra, 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, 28

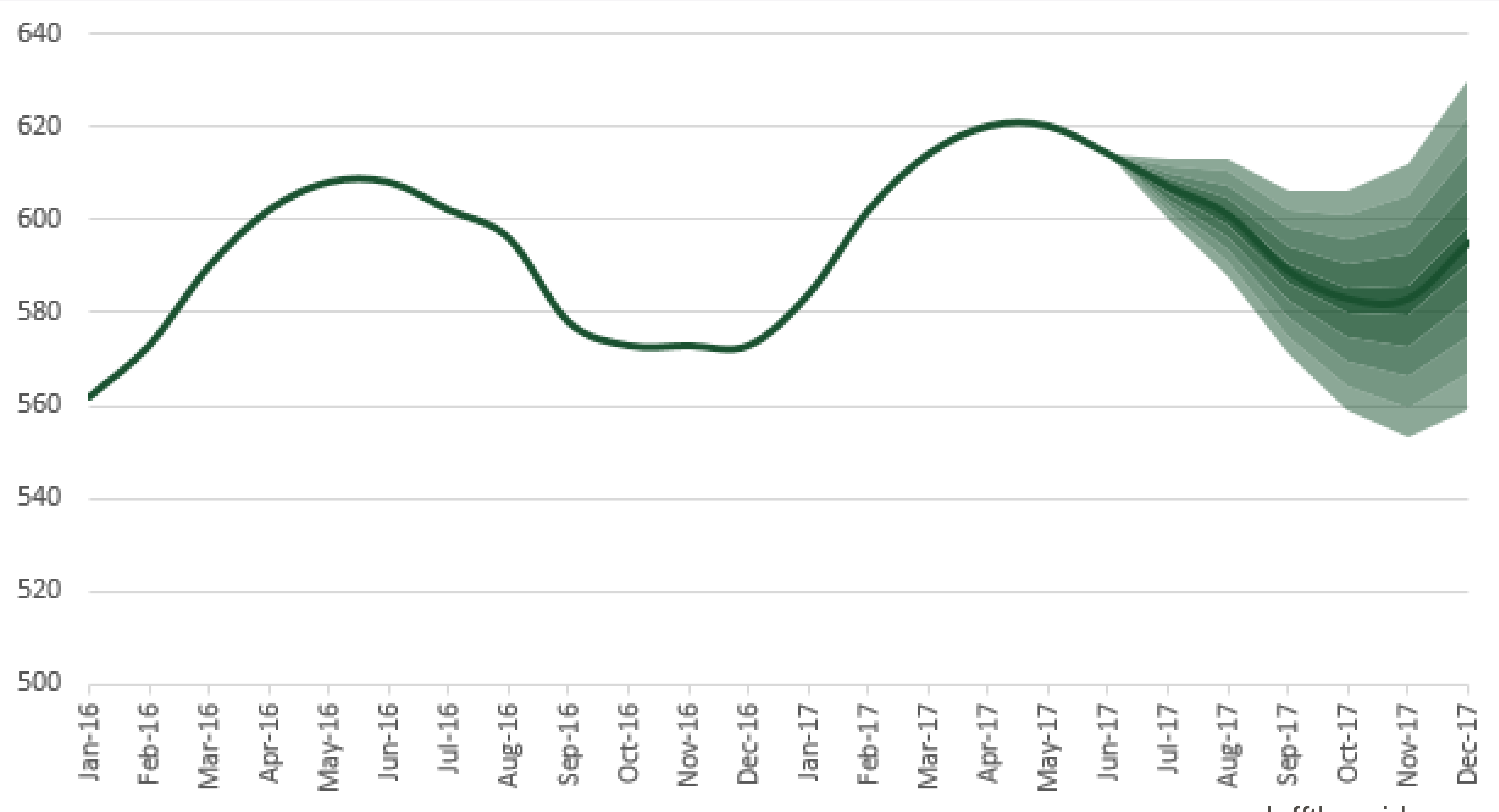


Alternativas a whiskers para error/uncertainty en linecharts:

- Error bands
- Fancharts

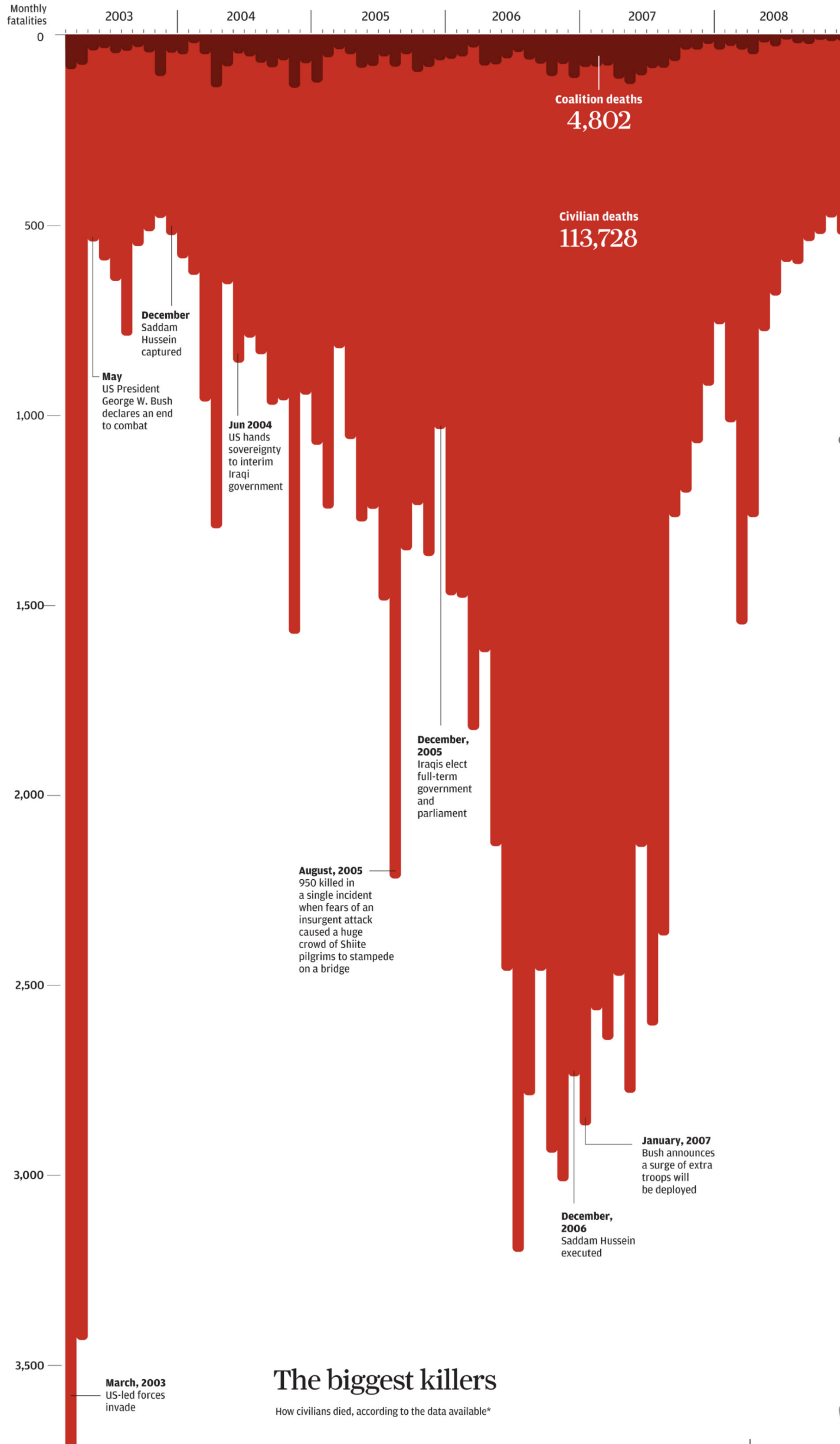


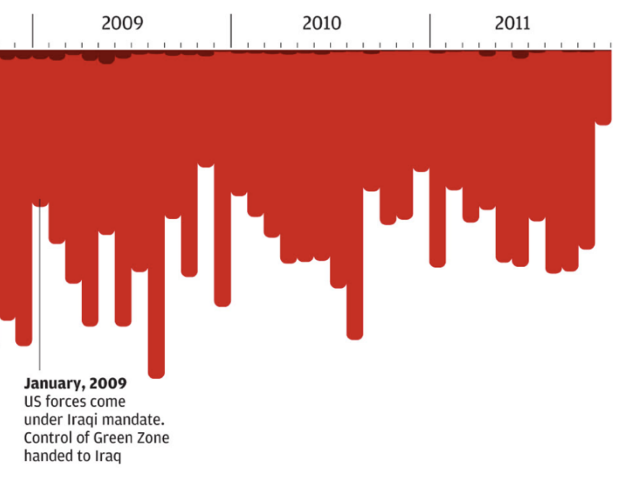
Nikki Marinsek



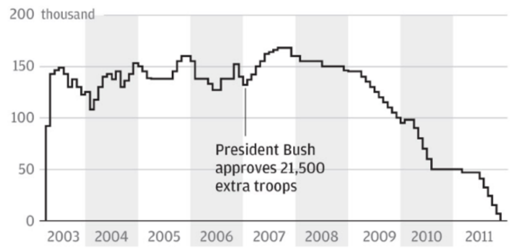
exceloffthegrid.com

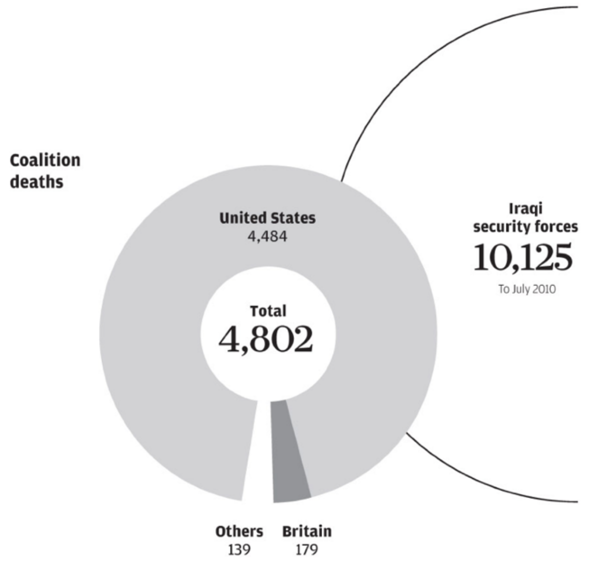
Iraq's bloody toll





The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.



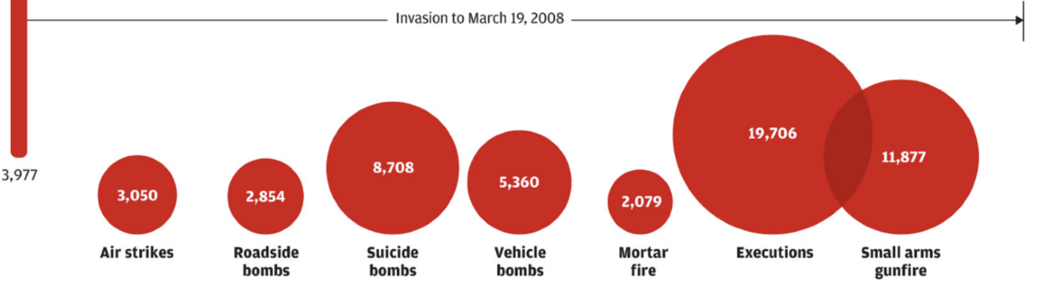


Coalition fatalities by area

Baghdad and the vast desert province of Anbar saw some of the fiercest fighting, the latter being the scene of sectarian tensions and a bloody Sunni insurgency

The biggest killers

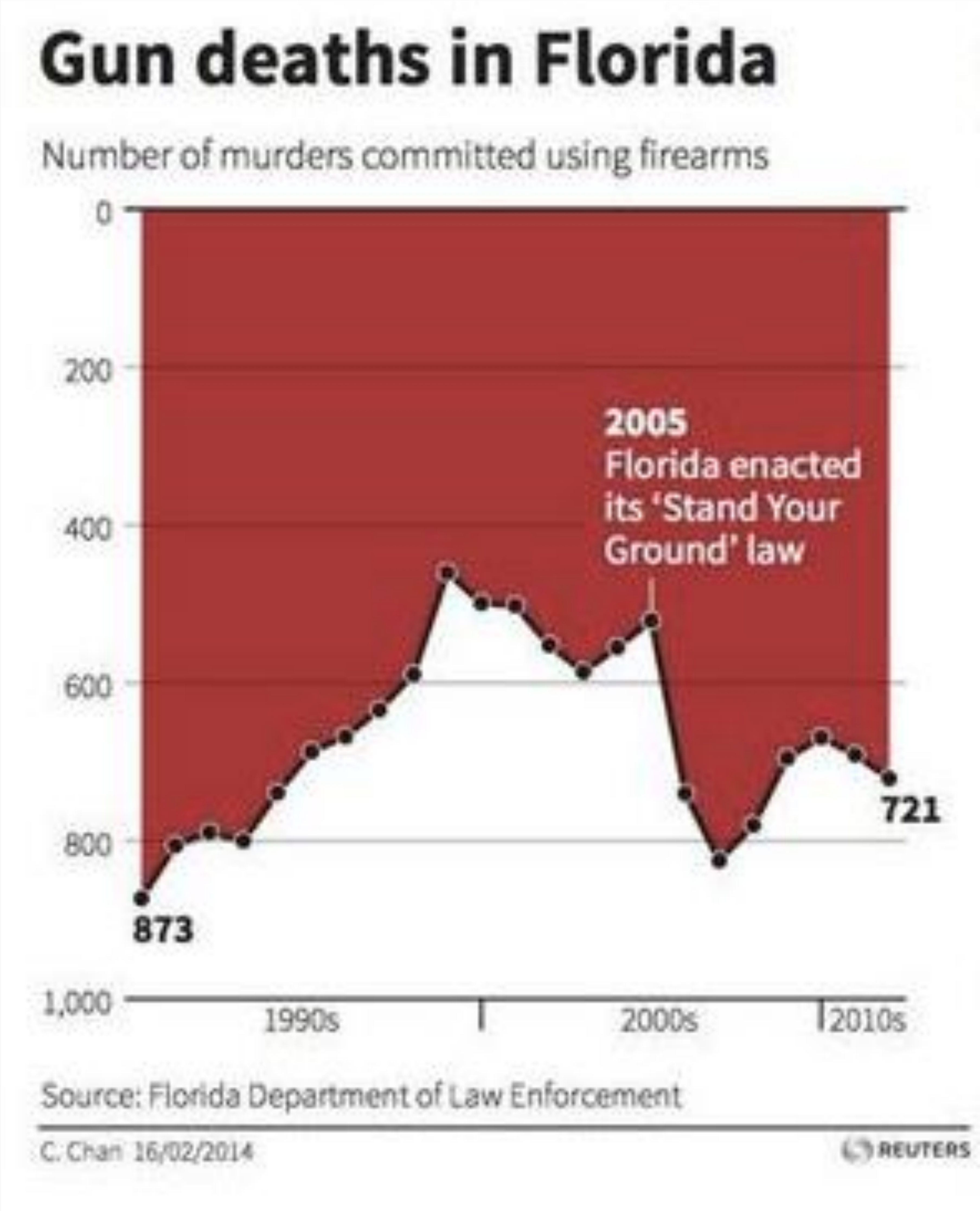
How civilians died, according to the data available*



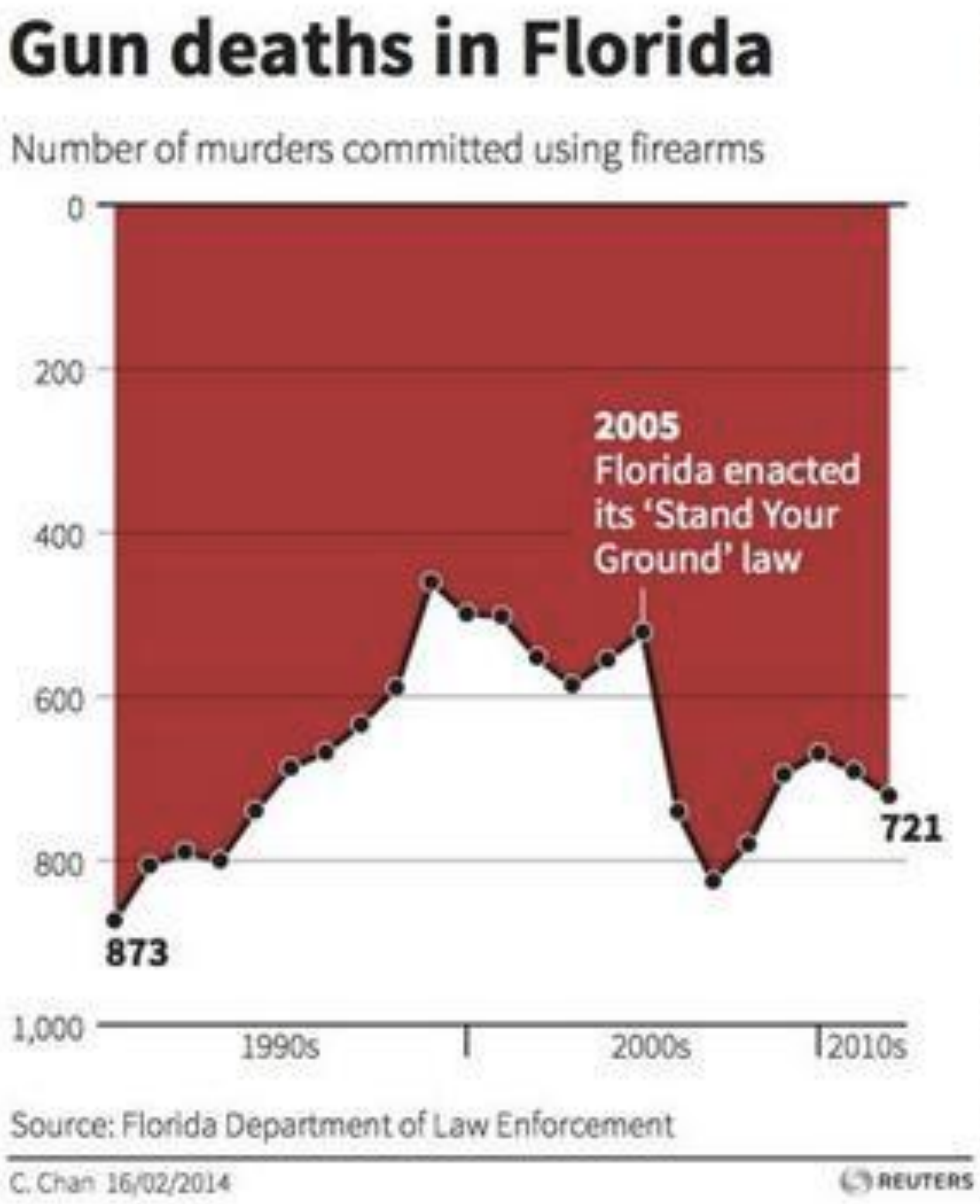
SCMP graphic: Simon Scarr

*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.

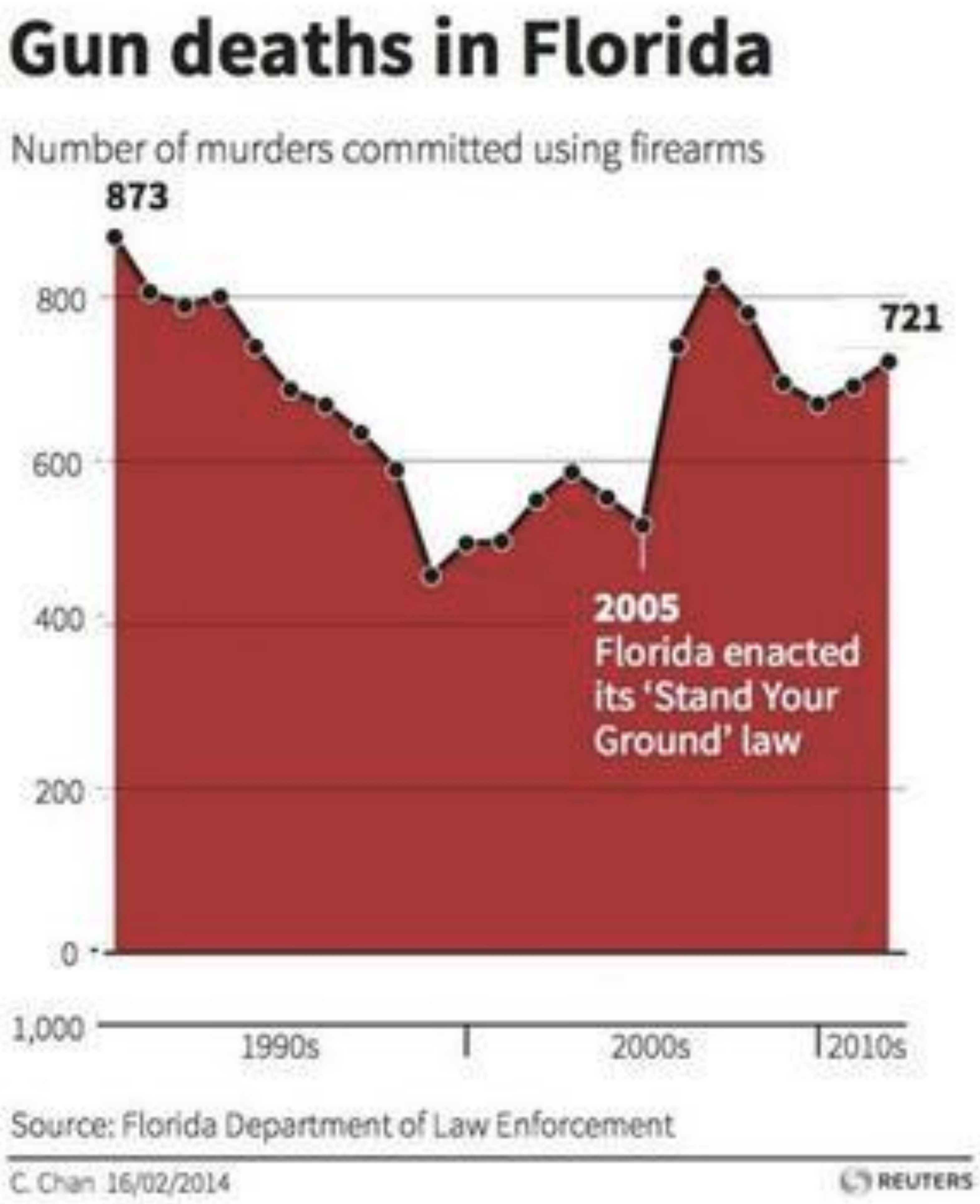
Sources: iCasualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution



Gun deaths in Florida
Number of murders committed using firearms
0
200
400
600
800
1,000
2005
Florida enacted its 'Stand Your Ground' law
873
721
1990s
2000s
2010s
Source: Florida Department of Law Enforcement
C. Chan 16/02/2014
REUTERS



BEFORE



AFTER

Referencias

Cairo, A. (2019). How Charts Lie: Getting Smarter about Visual Information. W.W. Norton & Company.

How Charts Lie by Alberto Cairo (video)

https://youtu.be/oX74Nge8Wkw

GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

Teoria 9. Sistemes Avançats II

Departament de Matemàtiques

judit.chamorro@uab.cat

9. Sistemes Avançats II. Contents:

9.1. Multi-variable (continuation class VII)

9.1.1. Visualizing many proportions at once (continuation of 7.1.3).

9.1.2. Multi-panel figures

9.1.3. Multivariate numerical data

9.2. Dimensionality reduction (continuation class VI)

9.2.1. T-Distributed stochastic neighbour embedding (t-SNE)

9.2.3. Tomography- Slice along a plane, 2D isosurfaces for a 3D field, isocontours

9.3. Maps

9.4. 2D or 3D ?

9.5. Scientific visualization

9.6. Other quality metrics

7.1.3 Visualizing many proportions at once

You saw: When proportions are specified **according to multiple grouping variables:**

Claus Wilke

Mosaic plots, treemaps, or parallel sets are useful visualization approaches

9.1.1 Visualizing many proportions at once

- **Parallel sets** work better than either mosaic plots or treemaps when there are more than two grouping variables **(to visualize proportions described by more than two categorical variables)**

In a parallel sets plot:

- we show **how the total dataset breaks down by each individual categorical variable.**
- **then we draw shaded bands** that show **how the subgroups relate to each other.**

9.1.1 Visualizing many proportions at once

- **Parallel sets** work better than either mosaic plots or treemaps when there are more than two grouping variables

dataset breaks down by each individual categorical variable (*construction material*)

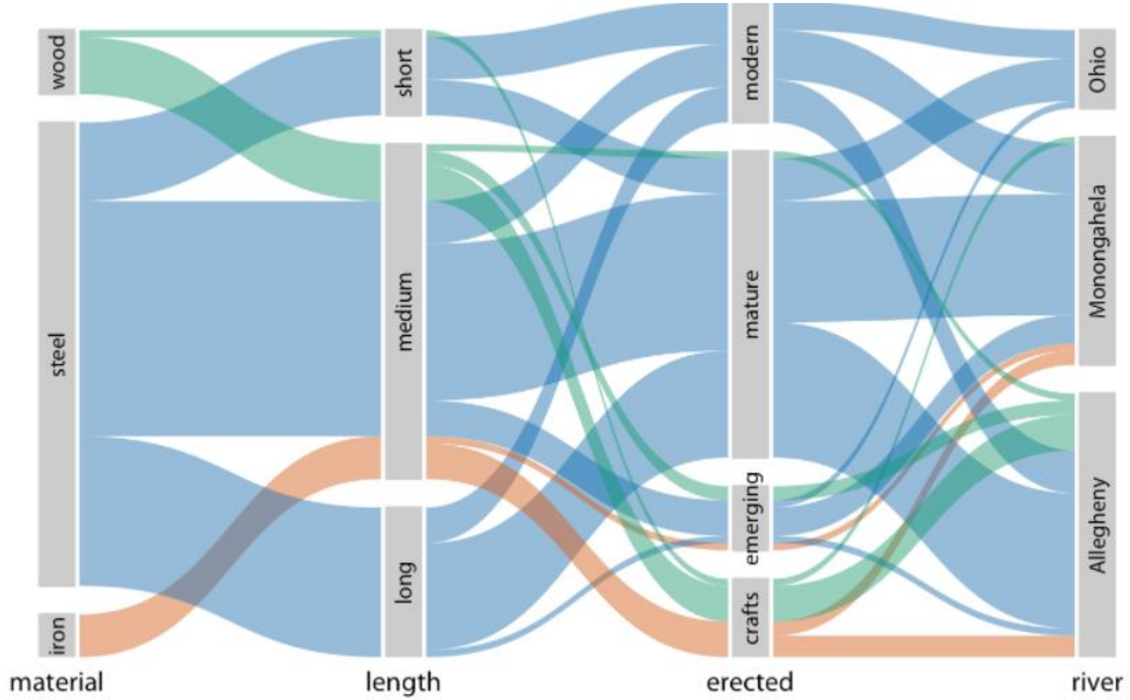
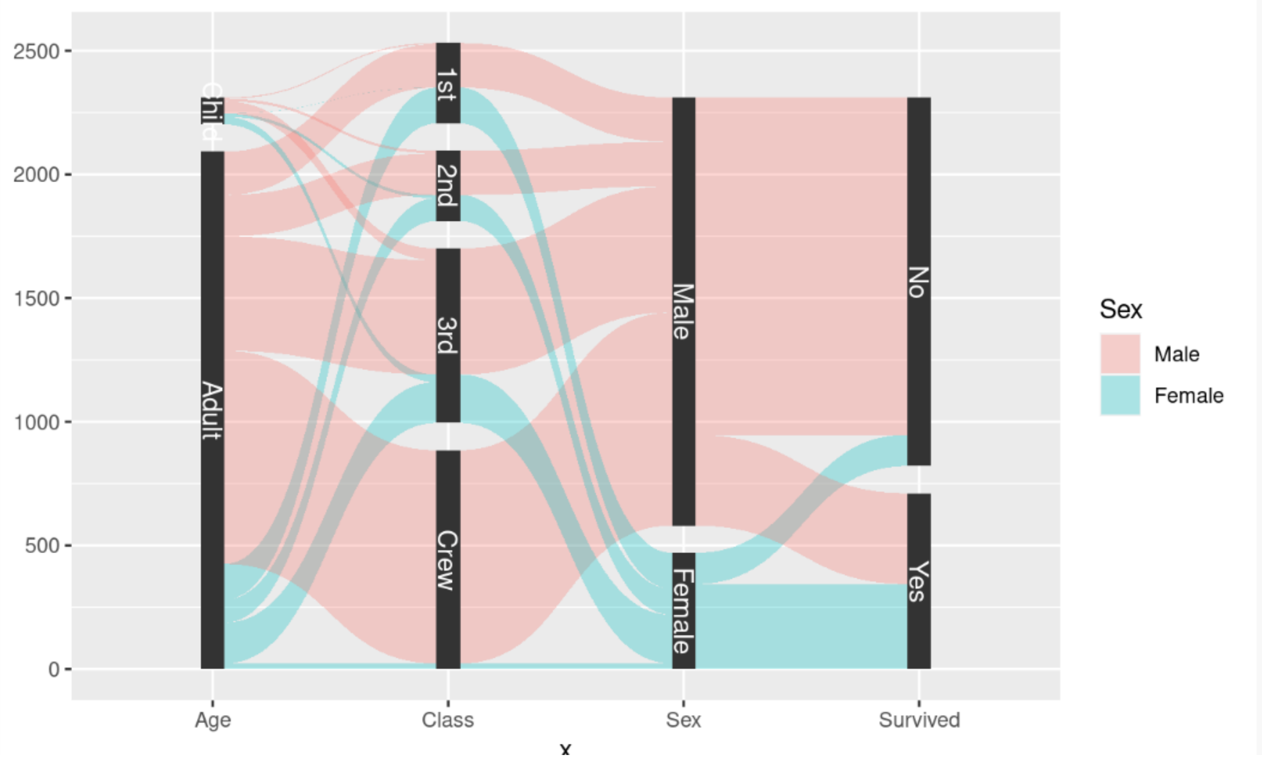


Figure 11.8: Breakdown of bridges in Pittsburgh by construction material, length, era of construction, and the river they span, shown as a parallel sets plot. The coloring of the bands highlights the construction material of the different bridges. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

9.1.1 Visualizing many proportions at once

```
data <- reshape2::melt(Titanic)
data <- gather_set_data(data, 1:4)

ggplot(data, aes(x, id = id, split = y, value = value)) +
  geom_parallel_sets(aes(fill = Sex), alpha = 0.3, axis.width = 0.1) +
  geom_parallel_sets_axes(axis.width = 0.1) +
  geom_parallel_sets_labels(colour = 'white')
```



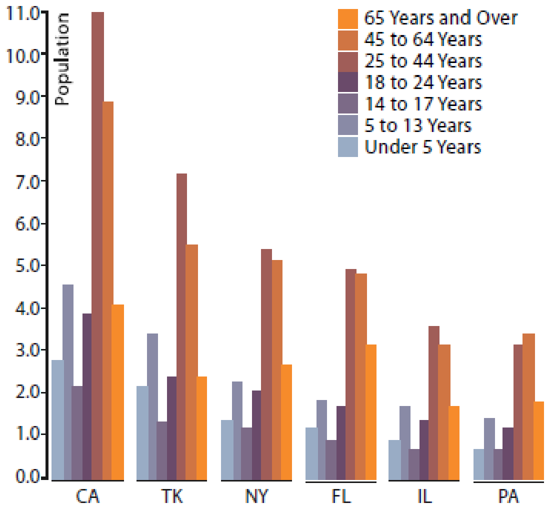
https://ggforce.data-imaginist.com/reference/geom_parallel_sets.html

9.1.2. Multi-panel figures

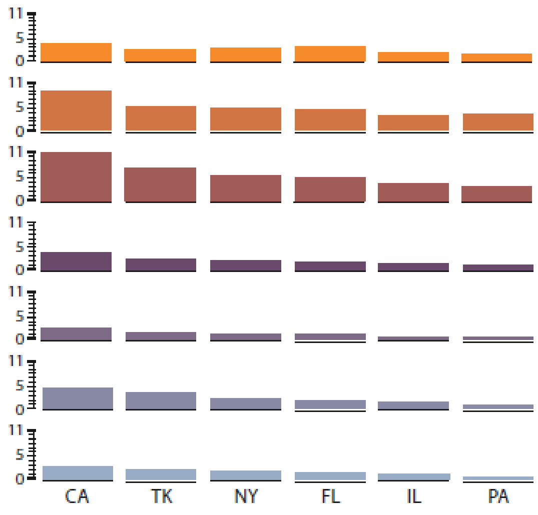
There are two distinct categories of multiple figure panels:

- **Small multiples** - plots consisting of multiple panels arranged in a regular grid. Each panel shows a different subset of the data but *all panels use the same type of visualization.*
- **Compound figures** - separate figure panels assembled in an arbitrary arrangement and *showing entirely different visualizations, or possibly even different datasets.*

9.1.2 Multi-panel figures: Small multiples



(a)



(b)

Grouped bar chart (a) against small multiple (b)

Tamara Munzner

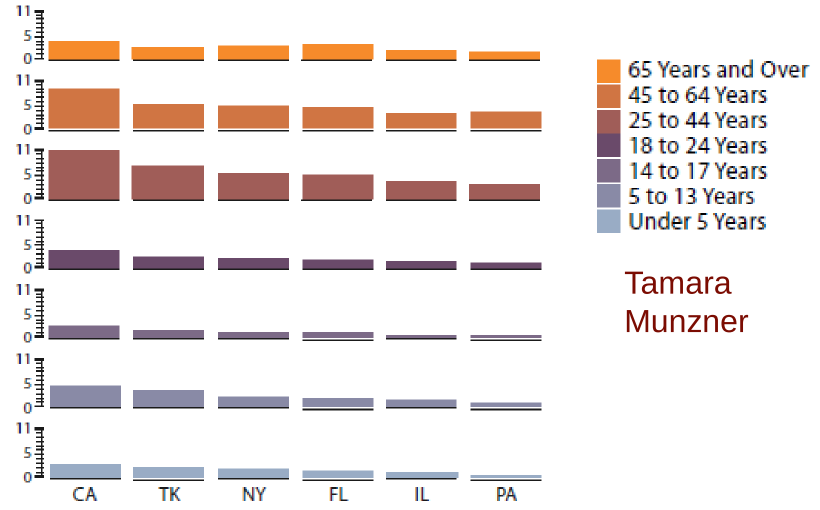
In a **grouped bar chart**, a **"group" of bar charts is drawn within each region**.

In contrast, **the small multiple** design choice simply **shows several standard bar charts, one in each view**.

9.1.2 Multi-panel figures: Small multiples

Small multiples (also known as “trellis plot” or “faceting”) are a means of graphing multivariate data, as an **array of *M X N* panels**.

As a simple example, **two "primary variables" might be chosen to represent a common set of axes (repeated in each panel)**, against which all combinations of two categorical variables (termed "conditioning variables") are plotted.

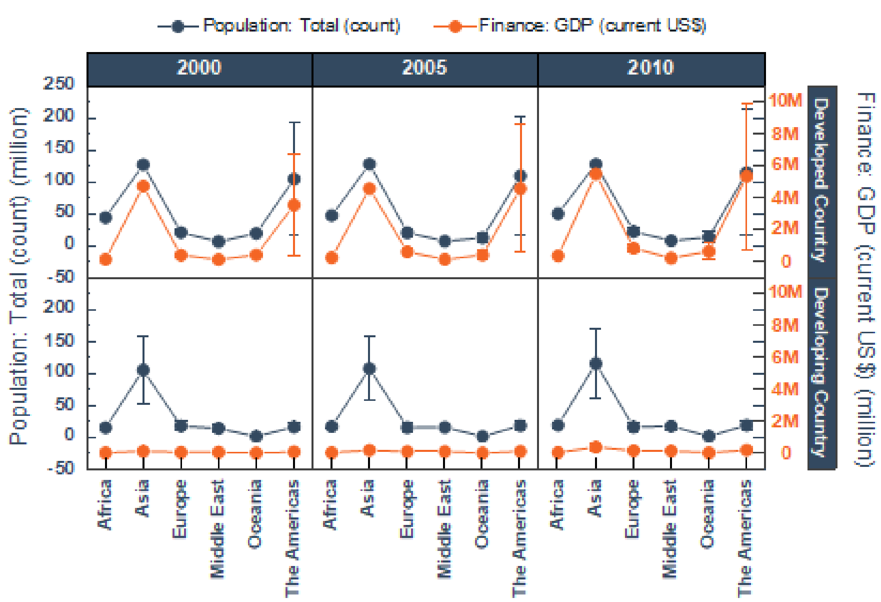


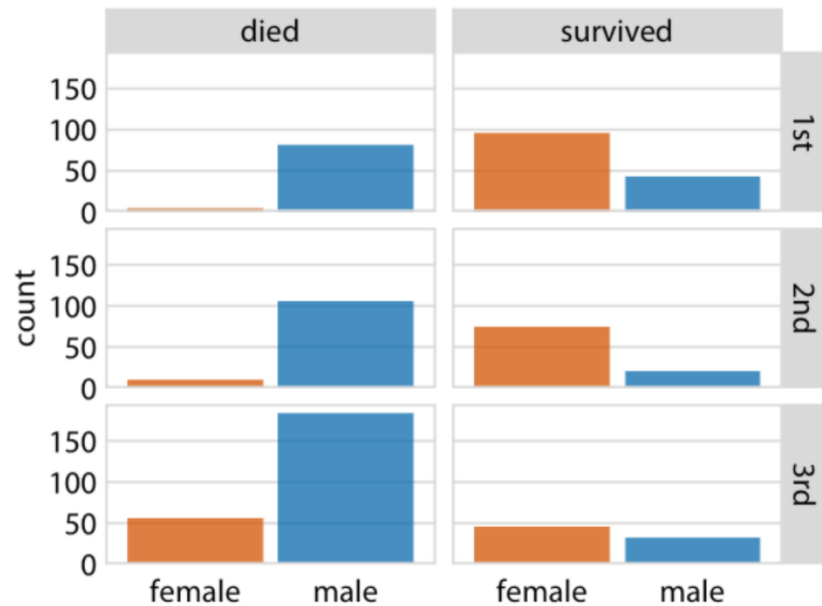
Tamara Munzner

Each combination of categorical values is plotted as a separate panel.

9.1.2 Multi-panel figures: Small multiples

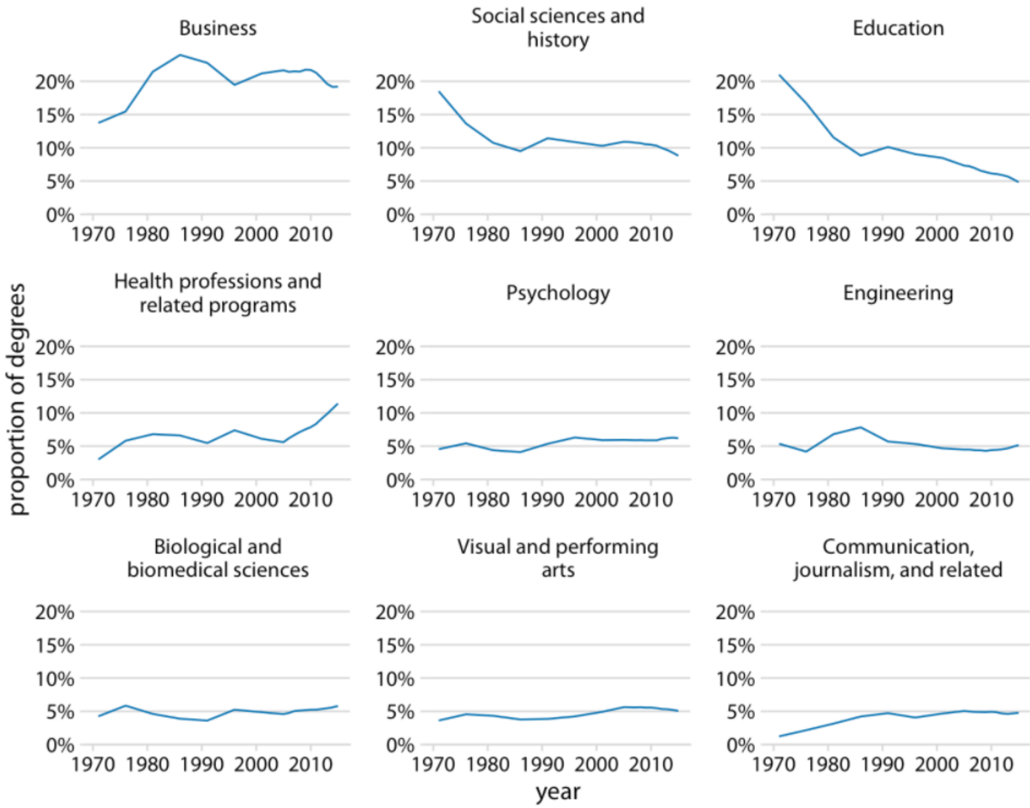
The key idea is to **slice the data into parts according to one or more data dimensions, visualize each data slice separately, and then arrange the individual visualizations into a grid**. **Columns, rows, or individual panels** in the grid are **labelled** by the values of the data dimensions that define the data slices.





9.1.2 Multi-panel figures: Small multiples

Try to avoid to use different axis scaling in separate panels of a small multiples plot. Otherwise, draw the reader's attention to this issue in the figure caption



Always **arrange the panels** in a small multiples plot **in a meaningful and logical order.**

9.1.2 Multi-panel figures: Compound figures

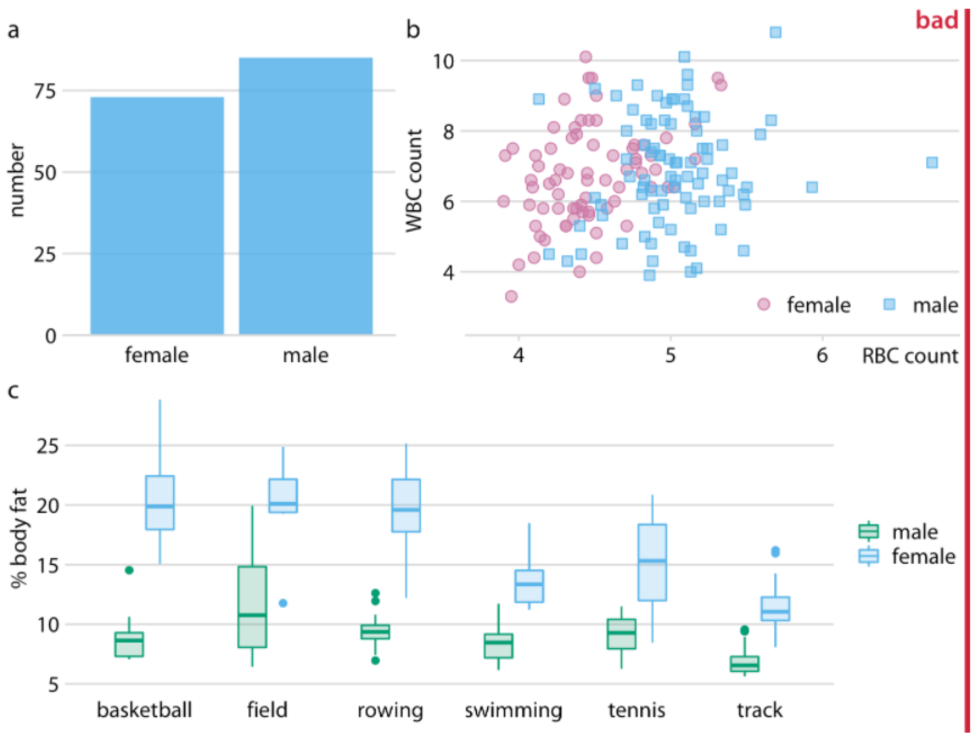
- **Not every figure** with multiple panels **fits** the **pattern** of small multiples.

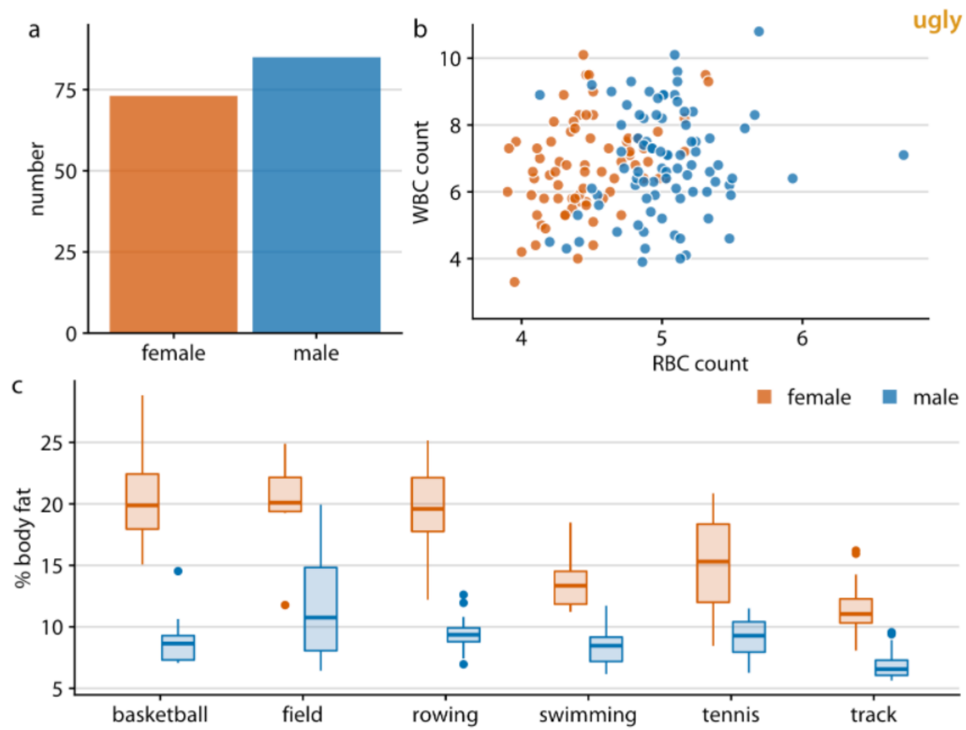
- Sometimes we simply want to **combine several independent panels** into a combined figure that conveys **one overarching point**.

In this case, we can **take the individual plots and arrange them in rows, columns, or other, more complex arrangements, and call the entire arrangement one figure**

9.1.2 Multi-panel figures: Compound figures

We need to **keep constancy between colors, symbols, fonts and so on** that we use to display the data.





Colors do not keep constancy

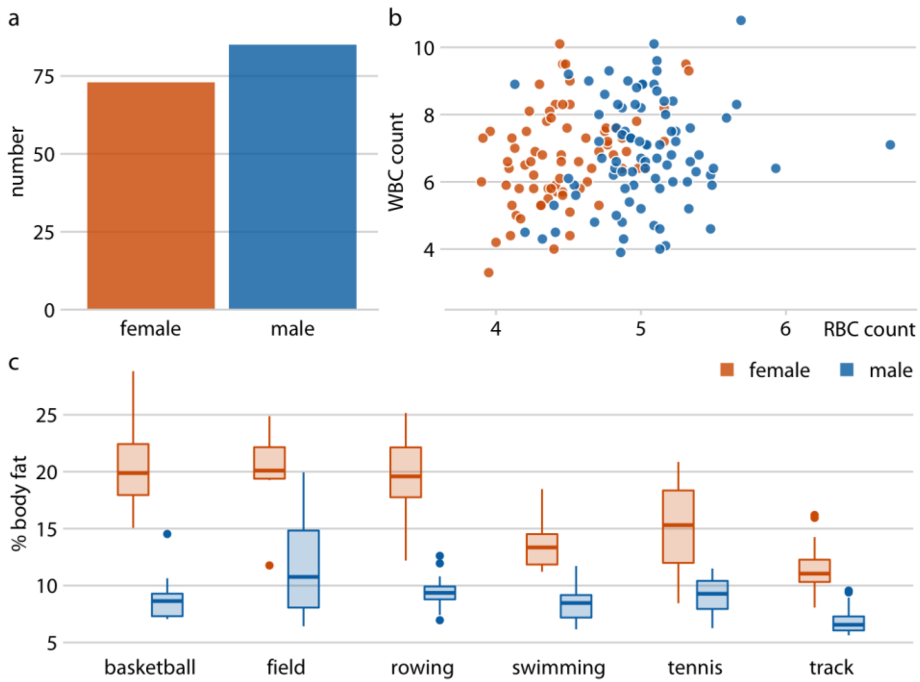
Better?

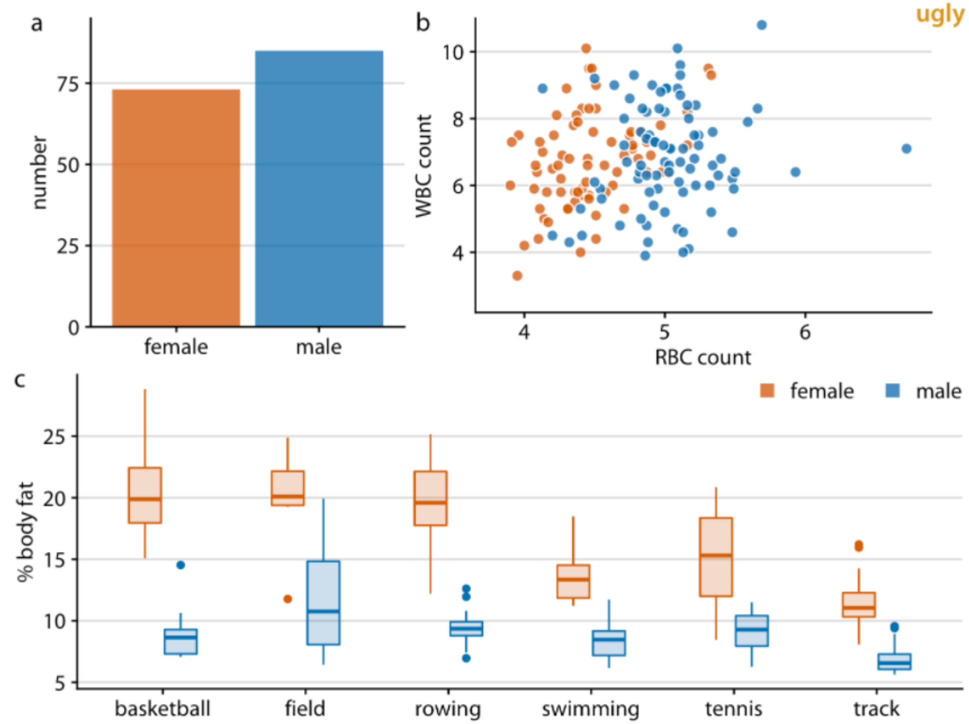
Escola d'Enginyeria UAB

UAB

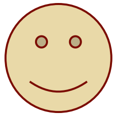
9.1.2 Multi-panel figures: Compound figures

The axis need also to be aligned





ugly



Misalignments are ugly

9.1.3 Multivariate numerical data

A Parallel Coordinate Plot (PCP) is a visualization technique used to analyse **multivariate numerical data**.

- Used **to plot individual data elements across many dimensions**
- It allows **to compare many quantitative variables together looking for patterns and relationships between them**
- The n-dimensional capabilities of the PCP enable **complex relationships** to be plotted with simplicity

7.1.3 Multivariate numerical data: PCP

- This representation does not show the classic Cartesian coordinate plane, but **each numerical variable is given its own axis**.
- **All axes are placed parallel, vertical, and equally spaced.**
- Every data element of the dataset is represented through **connected line segments**, derived from a connected set of points, one on each axis.
- We finally get a **set of lines, each of which is a multi-axis representation of every data record**.

9.1.3 Multivariate numerical data: PCP

- PCP are usually **difficult to understand** for non-technical audiences. There's no way to avoid the typical line jumble.
- **Try not to show many numerical variables at the same time.**
- PCPs are **not suitable for categorical data**.

categorical data?

9.1.3 Multivariate numerical data: PCP

- PCP are usually difficult to understand for non-technical audiences. There's no way to avoid the typical line jumble.
- Try not to show many numerical variables at the same time.
- PCPs are not suitable for categorical data.
- You must use **Parallel Sets Plots (PSPs) with categorical variables.** In the PSPs, **the horizontal bars in the visualization show the absolute frequency of how often each category occurred**

9.1.3 Multivariate numerical data: PCP

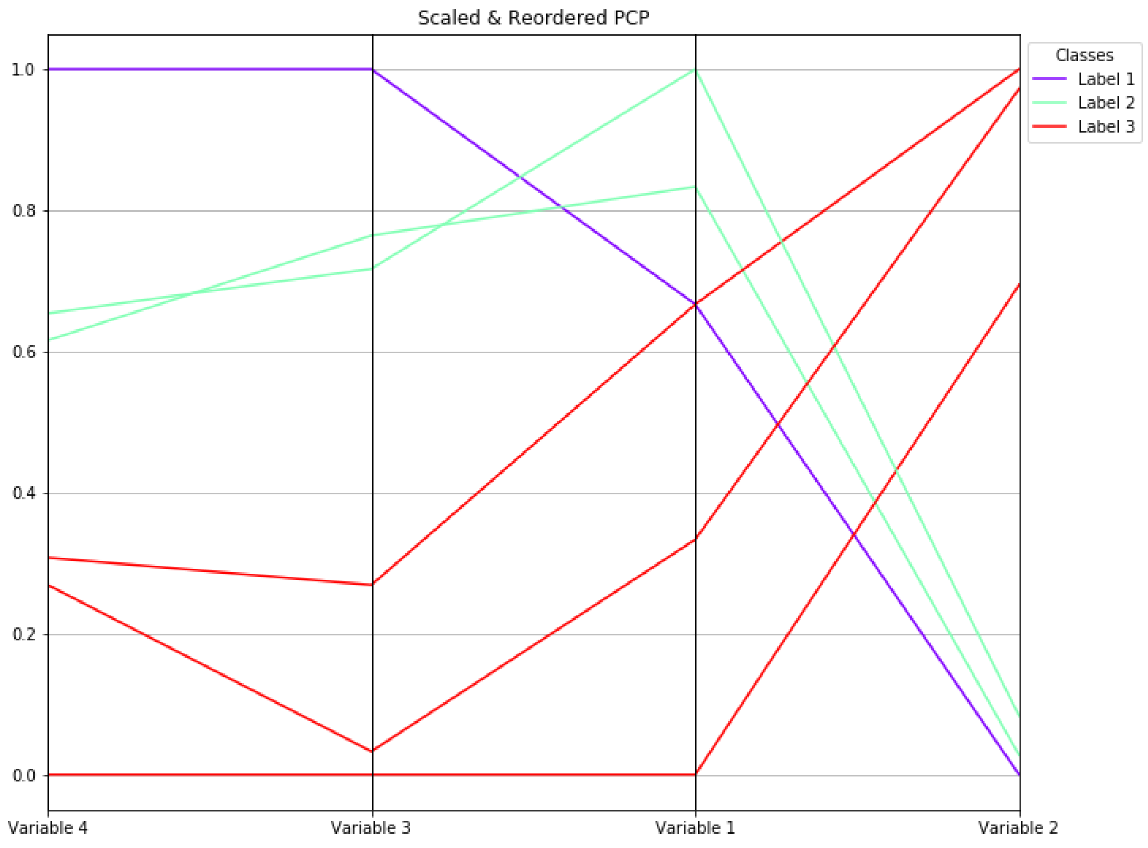
- **PCP's strength** is that **the variables can** even be completely different : have **different ranges and even different units**

- In general:

  - **lots of parallel lines indicate a positive relationship.**

  - lots of crossing lines (X shapes) indicate a negative association.

9.1.3 Multivariate numerical data (MND): PCP

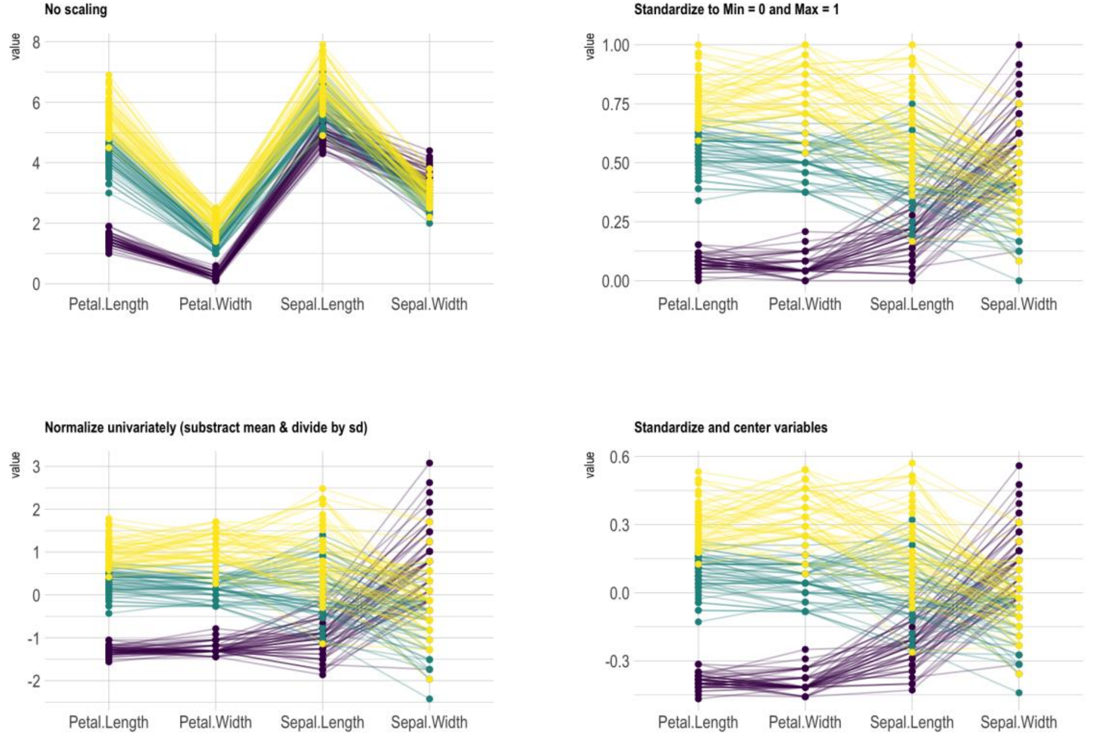


Schematic diagram of a parallel coordinates plot (PCP):

- four numerical variables
- six records related to three different classes

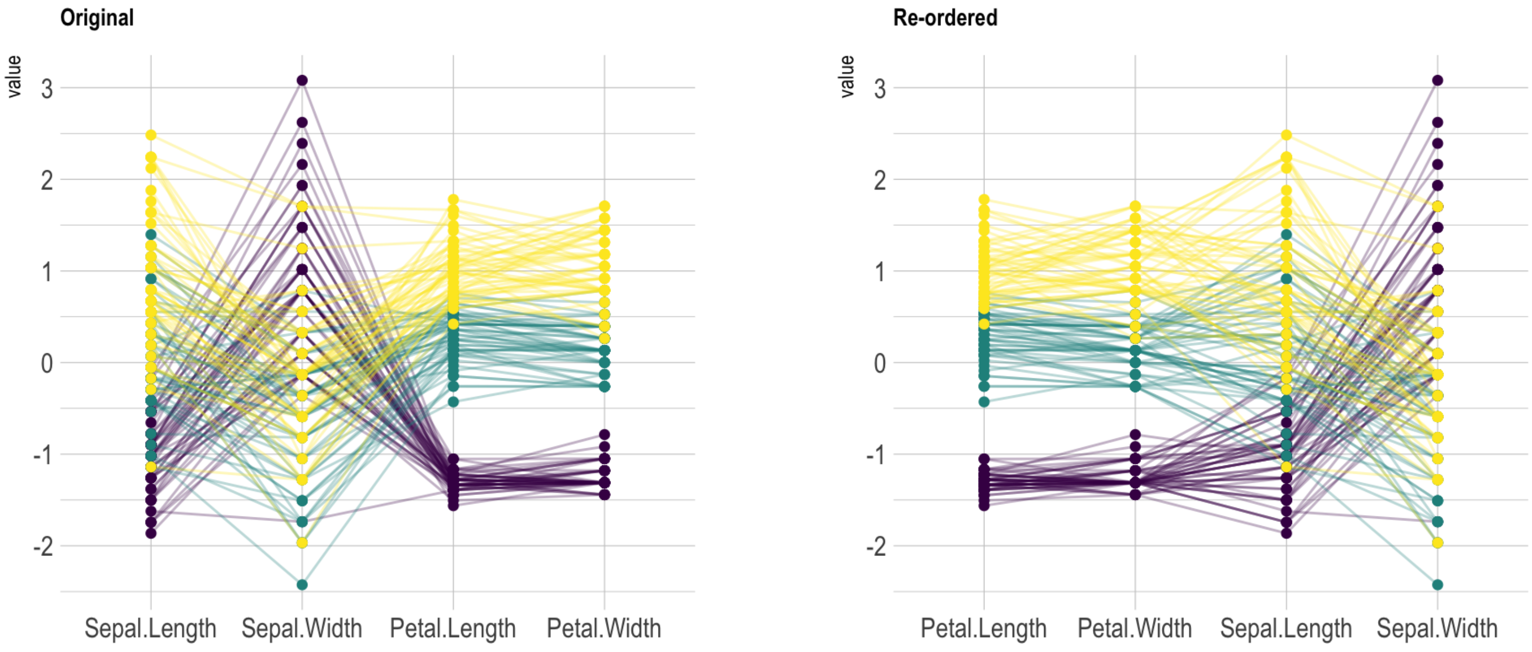
9.1.3 MND: Typical transformations for PCP

- **Scaling** - scaling transforms the raw data to a new scale that is common with other variables. It is a crucial step **to compare variables that do not have the same unit**, but can also help otherwise.



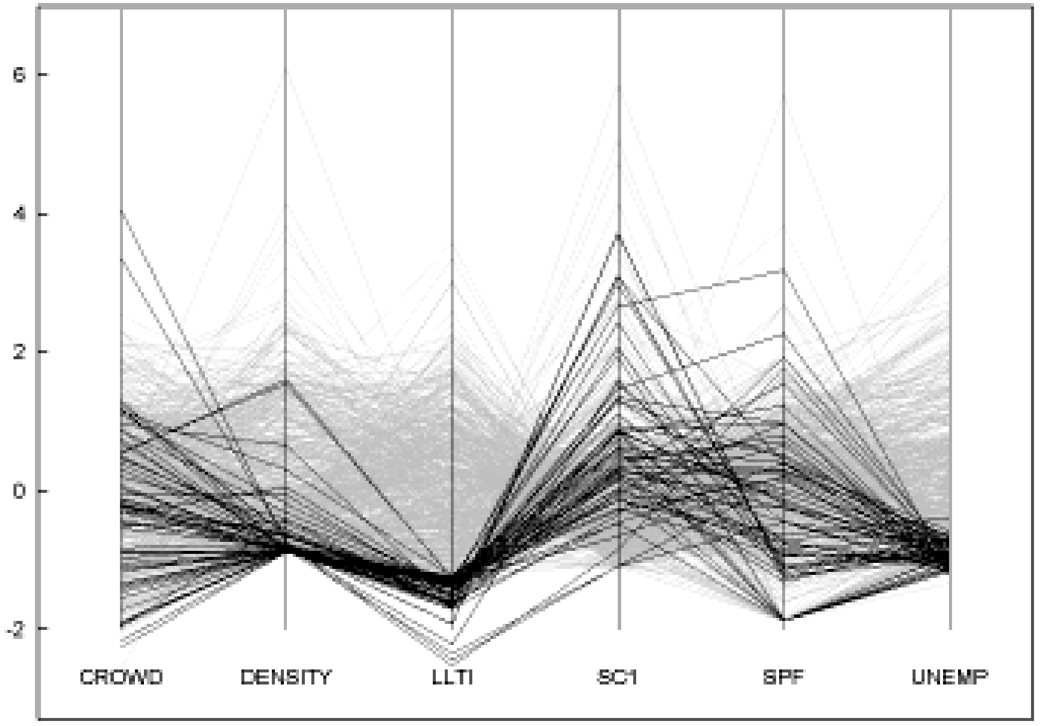
9.1.3 MND: Typical transformations for PCP

- **Axis order – optimizing the order of vertical axis** can decrease the clutter of your parallel plot. The goal is **to minimize the number of cross between series**



9.1.3 MND: Typical transformations for PCP

- **Highlighting**– highlight a specific sample or group of interest can avoid too many lines overlap



More examples of PCPs

9. Sistemes Avançats II. Contents:

9.1. Multi-variable (continuation class VII)

9.1.1. Visualizing many proportions at once (continuation of 7.1.3).

9.1.2. Multi-panel figures

9.1.3. Multivariate numerical data

9.2. Dimensionality reduction (continuation class VI)

9.2.1. T-Distributed stochastic neighbour embedding (t-SNE)

9.2.3. Tomography- Slice along a plane, 2D isosurfaces for a 3D field, isocontours

9.3. Maps

9.4. 2D or 3D ?

9.5. Scientific visualization

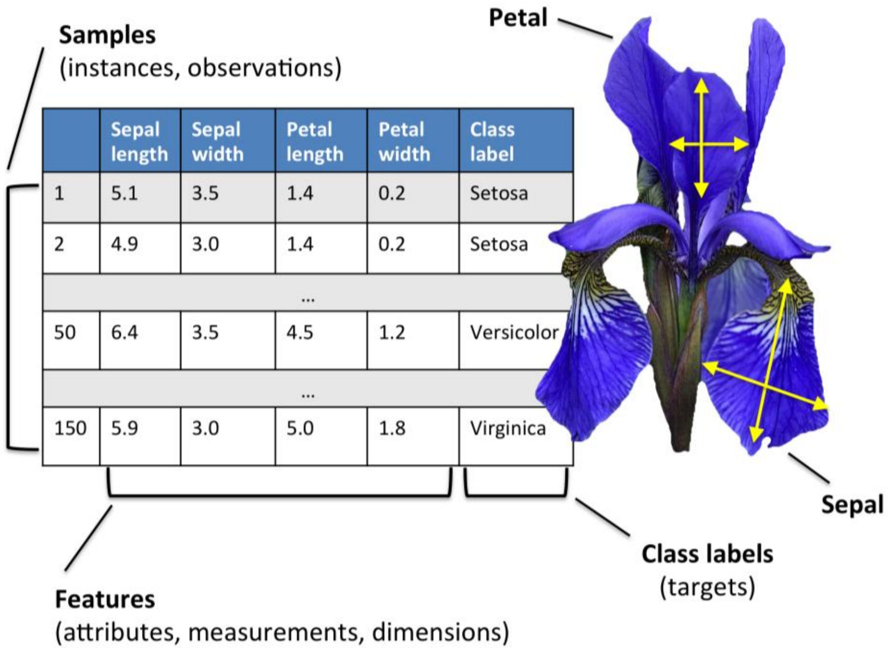
9.6. Other quality metrics

9.2.1 t-Distributed Stochastic Neighbor Embedding

t-SNE is **a non-linear technique for dimensionality reduction** that is most used for visualizing the higher dimensional dataset (More details in this link)

- **T-SNE visualizes high-dimensional data by giving each datapoint a location in 2 or 3D map.**
- **Benefit: It tries to preserve the local structure of the data** so *the distances and grouping of the points we observe in lower dimensions* such as a 2D scatter plot *is as close as possible to the distances we observe in the high-dimensional space*
- **t-SNE optimizes the distances with the local structure in mind.**

9.2 t-SNE iris example



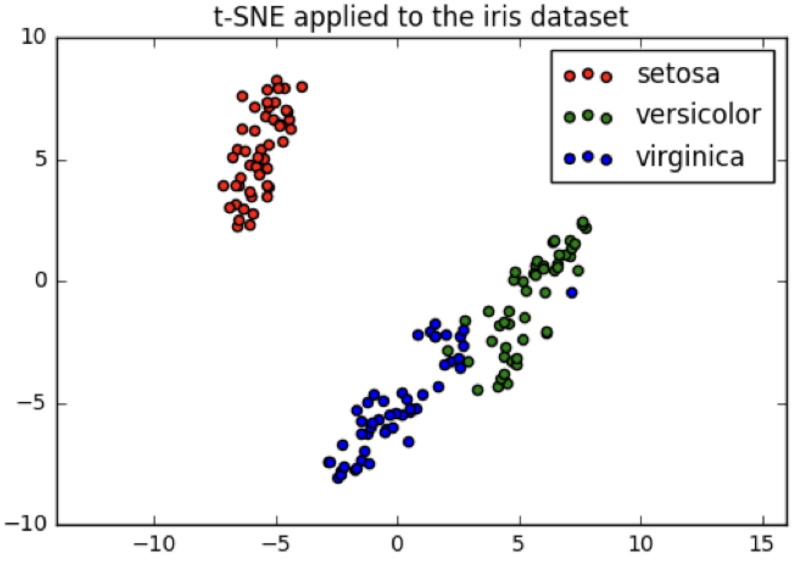

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

- Iris dataset has 4 measurements, so samples are 4-dimensional
- t-SNE maps samples to 2D space
- t-SNE didn't know that there were different species
- Versicolor and virginica are close in space

9.2.1 t-SNE iris example

- Versicolor and virginica are close in space

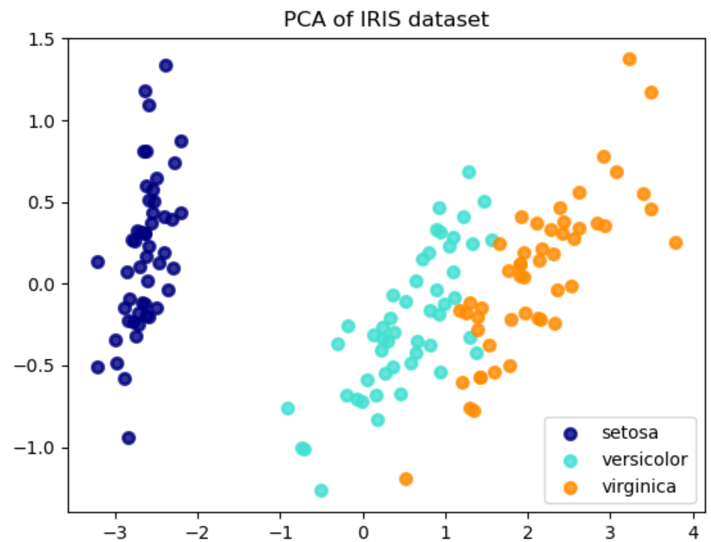
The neighbour embedding property makes **t-SNE effective for identifying local clusters in the data**, **but** the side effect is that **it fails to preserve the global inter-cluster structure** – the embedding distances among clusters are meaningless and the global distribution of clusters is random

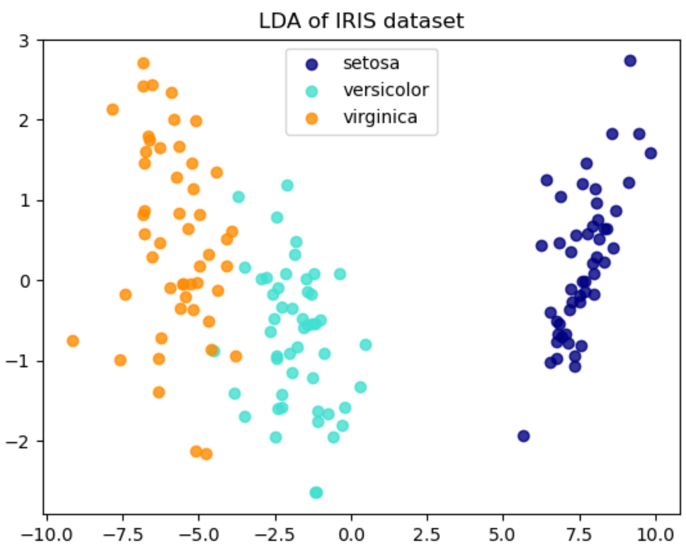


9.2.1 PCA versus LDA (vam veure en 6.3)

PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the **most variance in the data.**

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width





LDA: tries to **identify attributes** that account for the most **variance between classes**

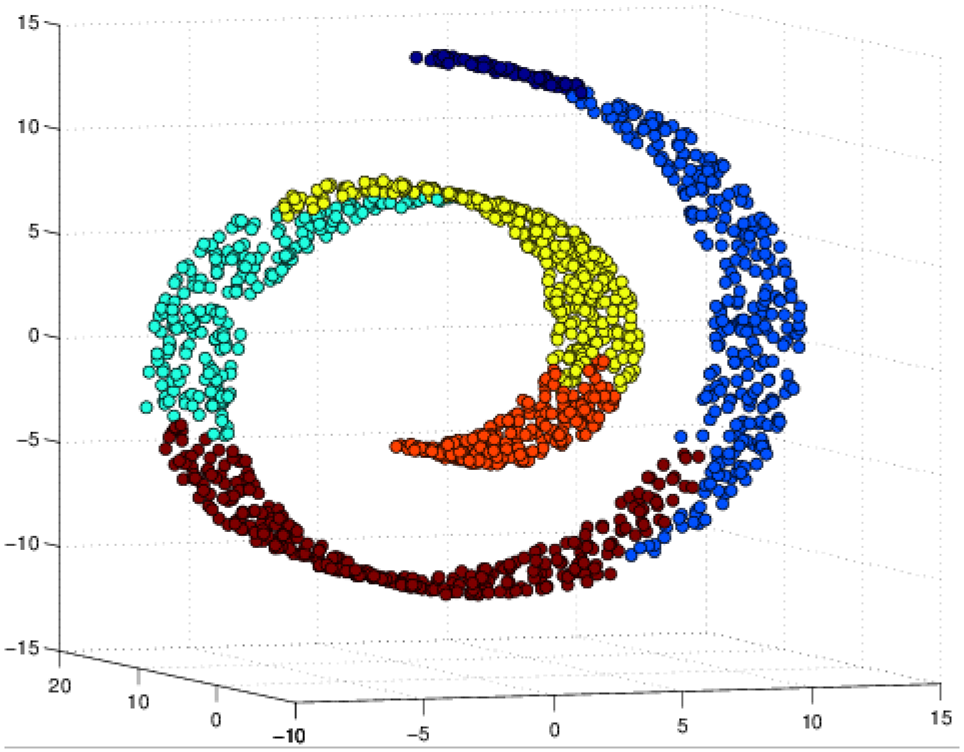
9.2.1 PCA versus t-SNE

The **main difference** between t-SNE (or other manifold learning methods) and PCA is that **t-SNE tries to deconvolute relationships between neighbors in high-dimensional data.**

A classic example is the "swiss roll".

!! **If your dataset is highly nonlinear using PCA will distort the distances** between points and most likely your downstream algorithm will yield misleading results.

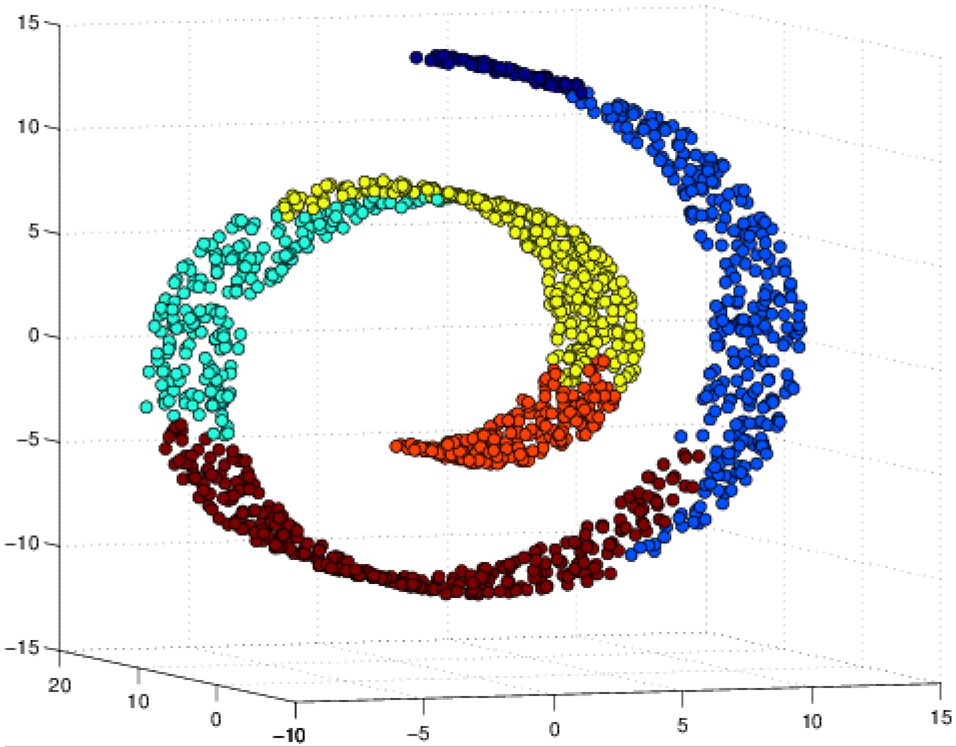
9.2.1 PCA versus t-SNE



"swiss roll"

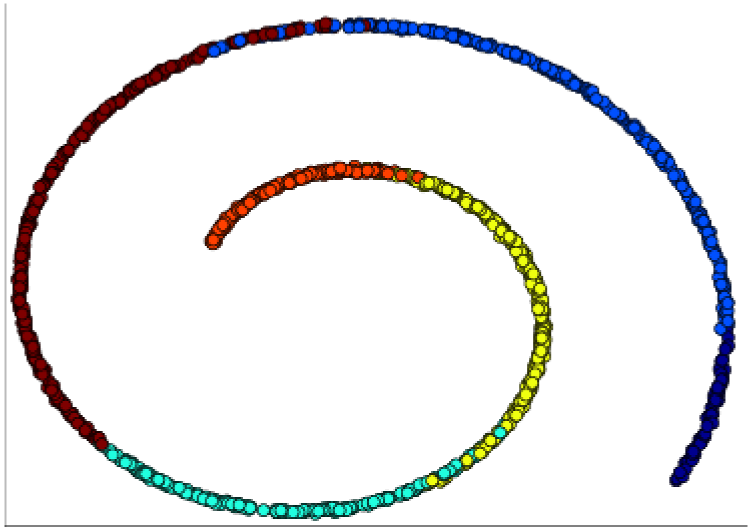
PCA (90% Variance Expaine) **thinks yellow is close to blue when in fact they are far away**

9.2.1 PCA versus t-SNE



"swiss roll"

In contrast, see how t-SNE seems to understand what is going on with this ‘S’

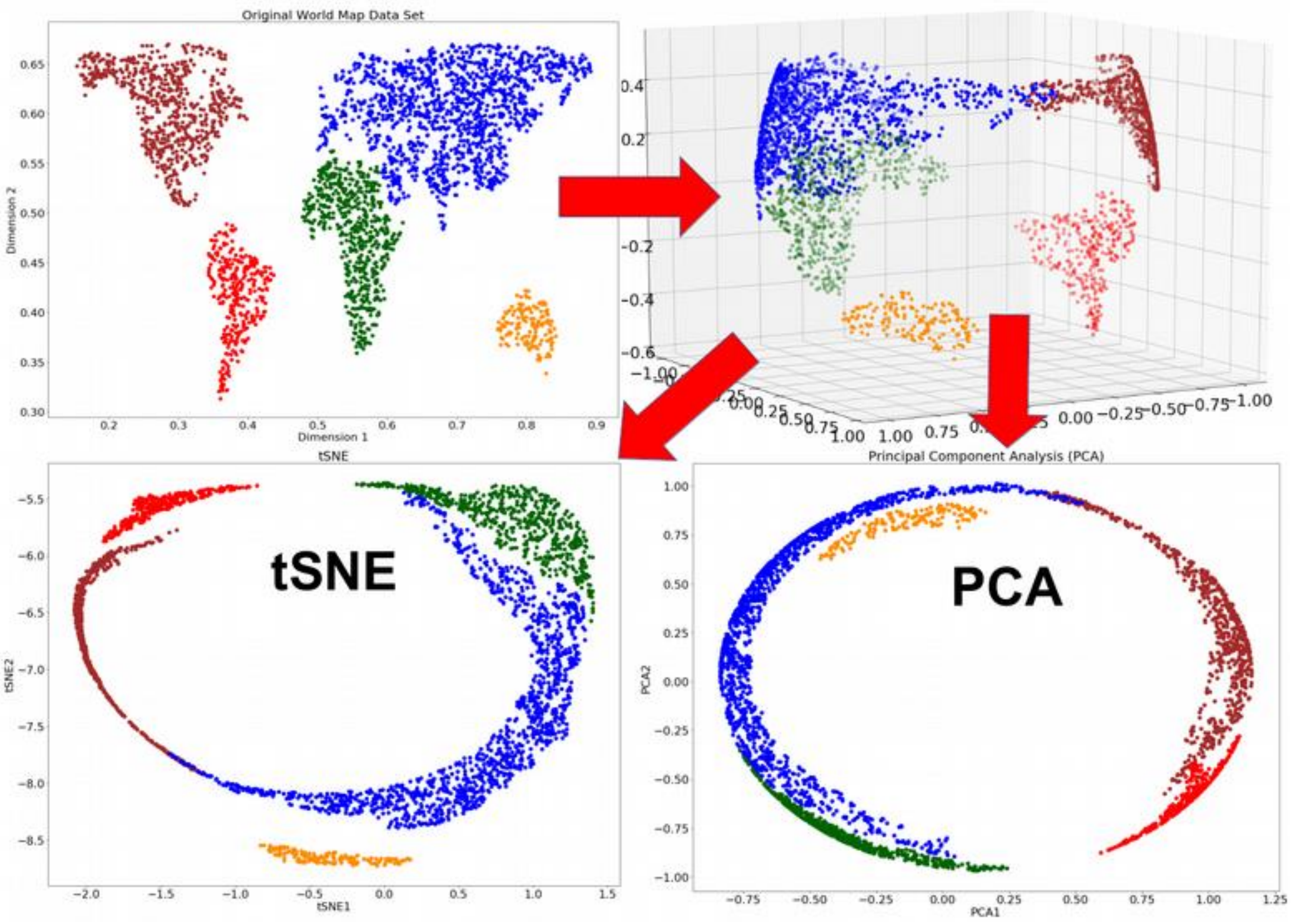


PCA (90% Variance Expaine) **thinks yellow is close to blue when in fact they are far away**

Escola d’Enginyeria UAB

UAB

9.2.1 PCA versus t-SNE



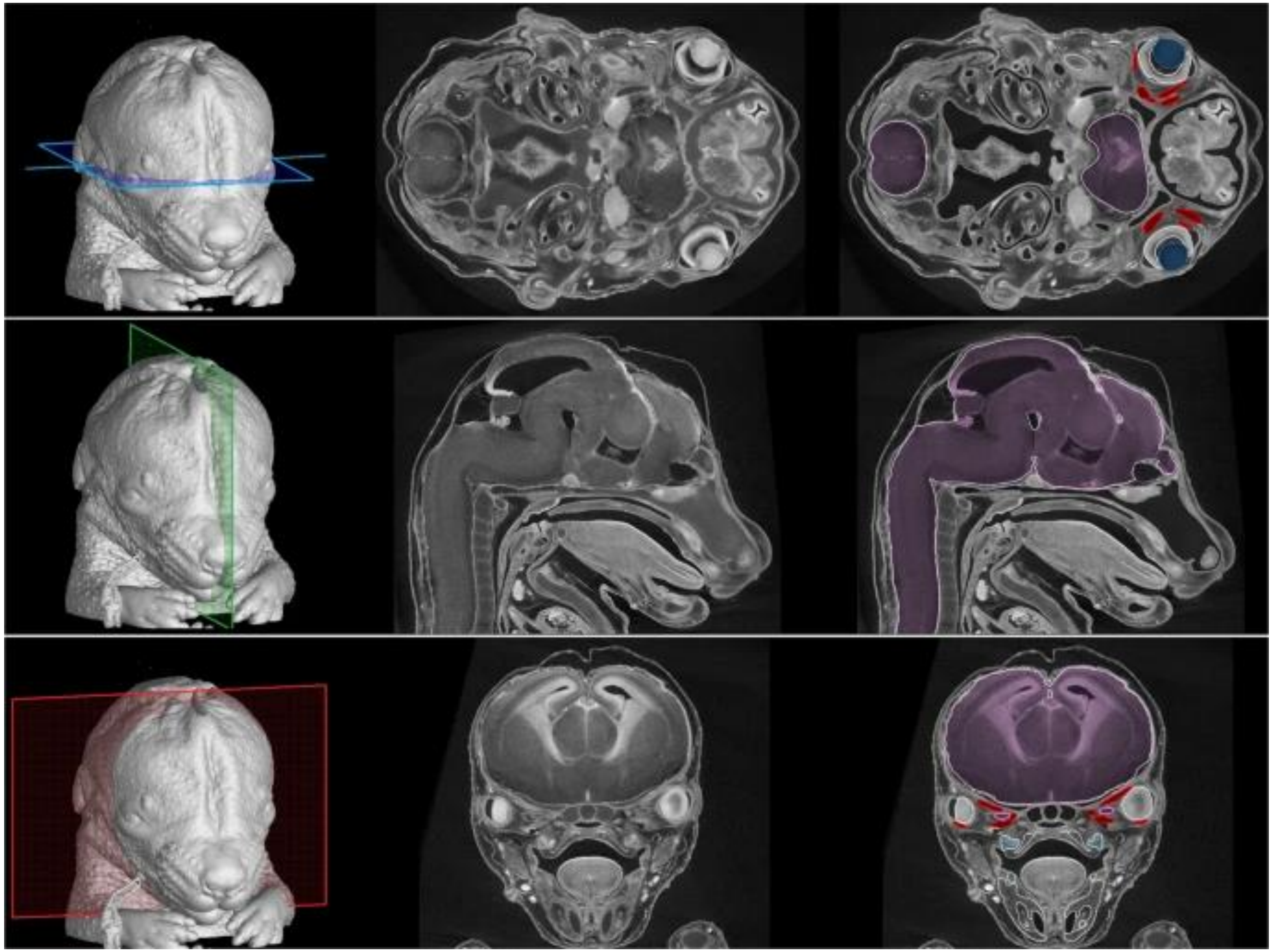
T-SNE degrades to PCA

9.2.1 PCA versus t-SNE

Key differences between PCA and t-SNE can be noted as follows:

- **t-SNE is computationally expensive** and can take several hours on million-sample datasets where PCA will finish in seconds or minutes.
- **PCA is a mathematical technique, t-SNE is a probabilistic one.**
- Linear DR algorithms, like **PCA, concentrate on placing dissimilar data points far apart in a lower dimension representation**. But in order to represent high dimension data on low dimension, non-linear manifold, it is essential that **similar data points must be represented close together, which is something t-SNE does not PCA.**
- **Sometimes in t-SNE different runs with the same hyperparameters may produce different results hence multiple plots must be observed before making any assessment with t-SNE,** while this is not the case with PCA.
- **PCA will not be able to interpret the complex polynomial relationship between features** (due to its linearity) **while t-SNE is made to capture exactly that.**

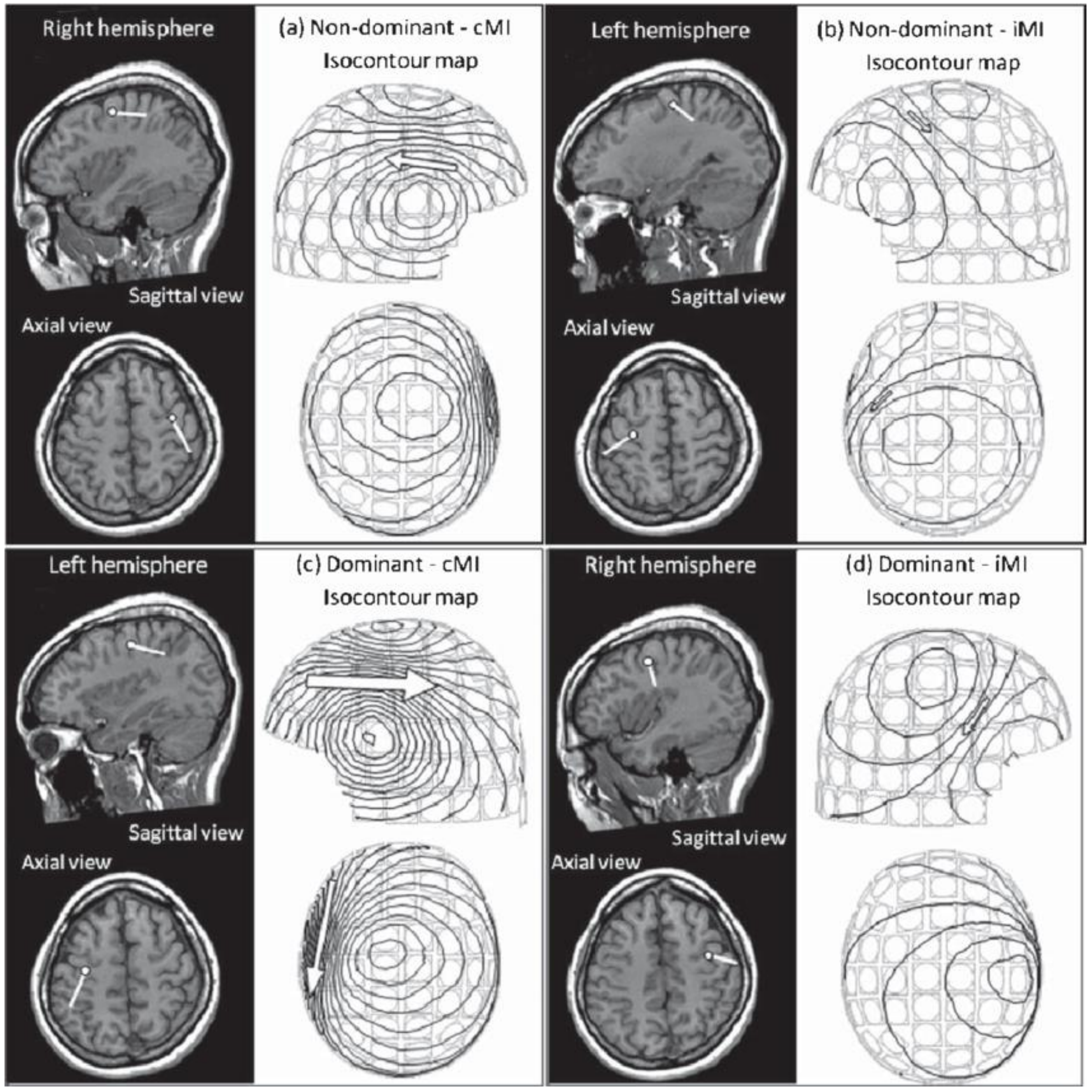
9.2.2 Tomography



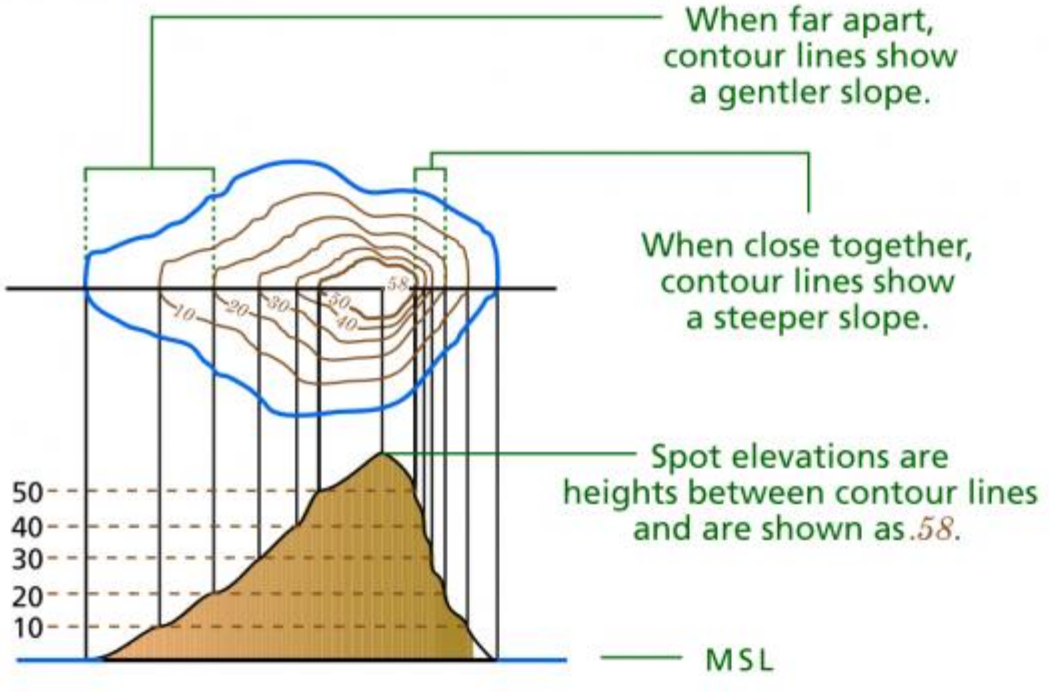
Tomographic measurements and segmentation of craniofacial structures in a mouse embryo 15 days post-fertilisation.

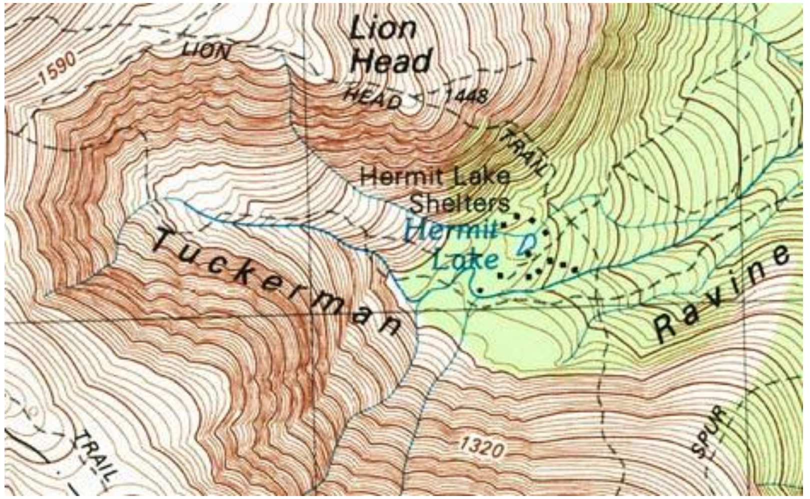
https://www.nature.com/articles/s41598-019-51180-2

9.0.2 Isocontour / Tomographic map contour



9.2.2 Isocontour / Tomographic map contour





9.2.2 Surface plot, slice plane, Isosurface

the dataset boundary
Do not reveal inner part

the same y-coordinate
Only show 2D

the scalar value 65
Ignores all volume points

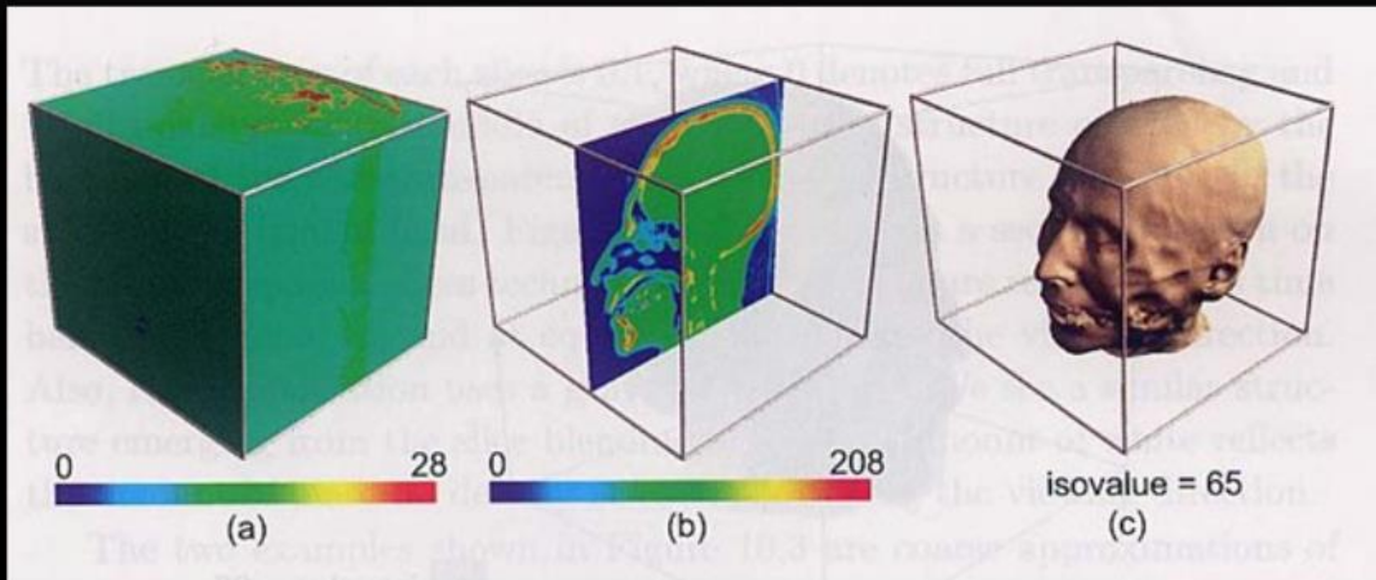
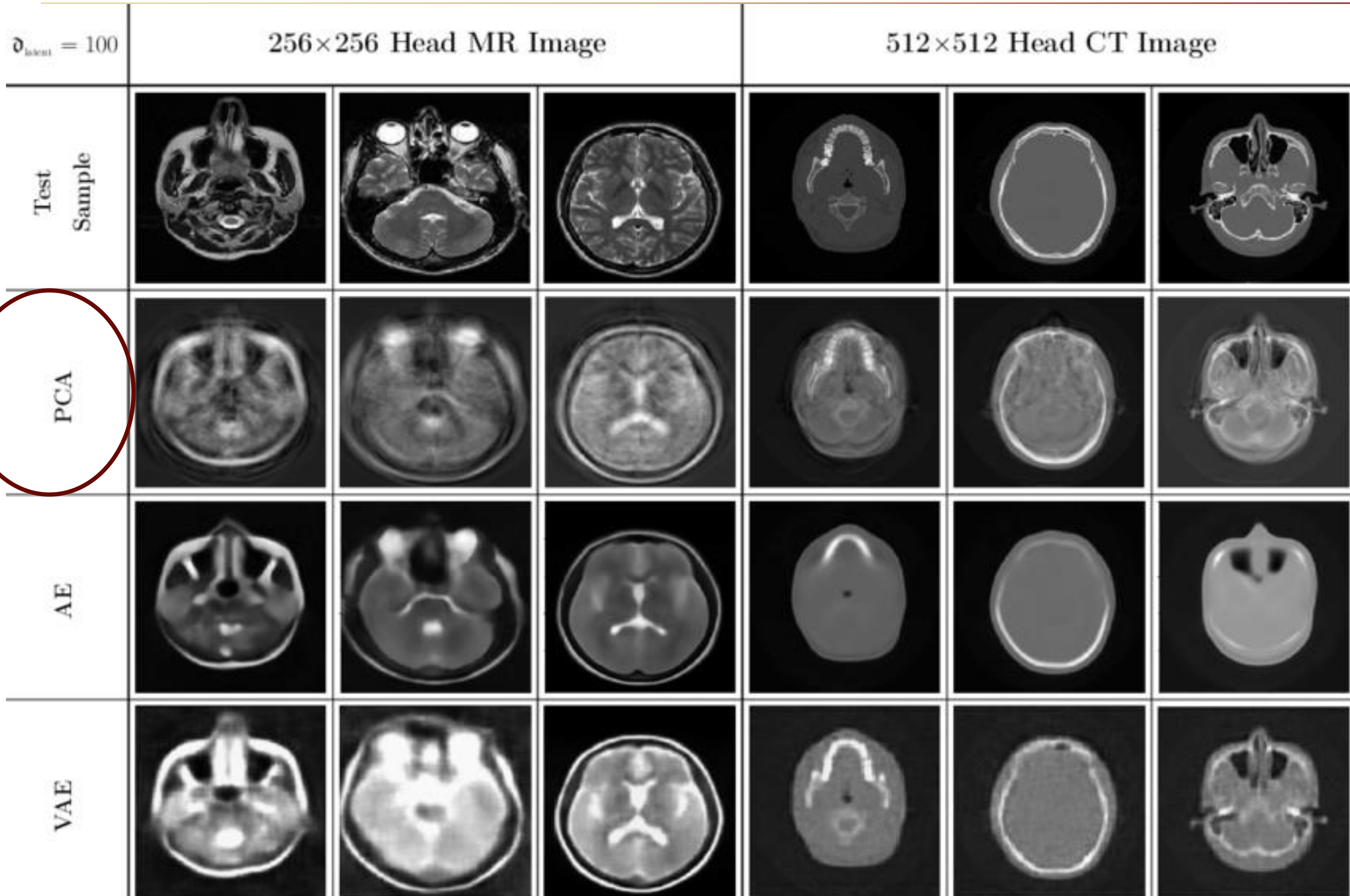


Fig 10.1 Visualizing a 3D scalar dataset (128^3 in 0~255):
(a) Surface plot; **(b)** Slice plane; **(c)** Isosurface (Skin).
The methods reduce data dimensionality from 3D to 2D

More details

9.2.2 Tomography & PCA dimension reduction



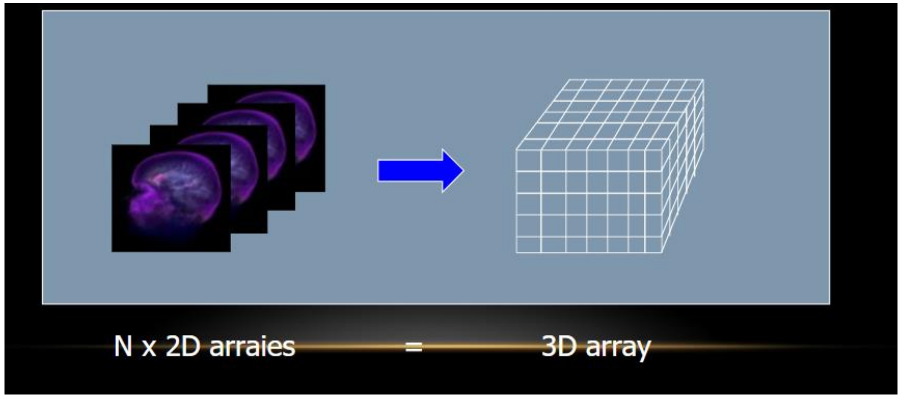
More details

9.2.2 Volume rendering

The basic idea behind volume rendering: Creating a 2D image that reflects at every pixel, the scalar data within a given 3D dataset

Main issue: the choice of the function

- Mapping an entire set of scalar values, for the voxels along the viewing ray, to a single pixel in the resultant 2D image



3D volume data are represented by a finite number of cross-sectional slices

9.2.2 Volume Visualization

The value of the image pixel p

$$I(p) = F(s(t)), \quad t \in [0,1]$$

F : Ray function

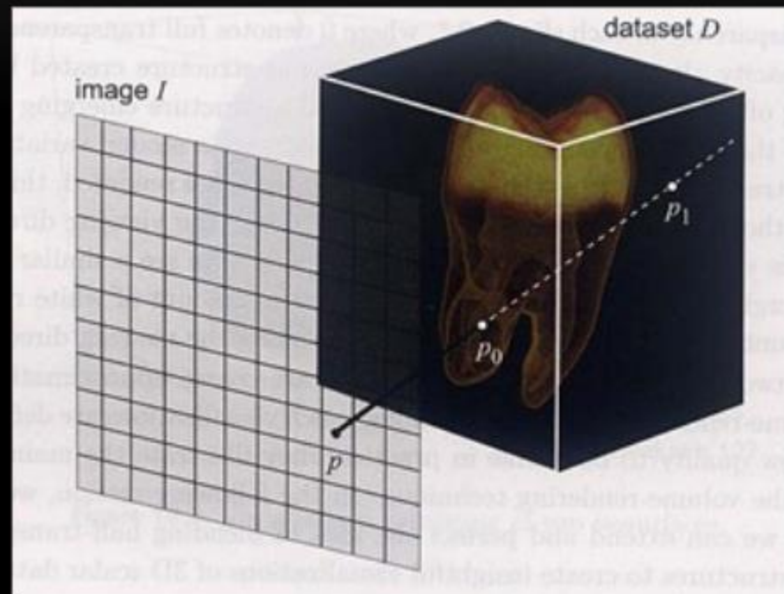
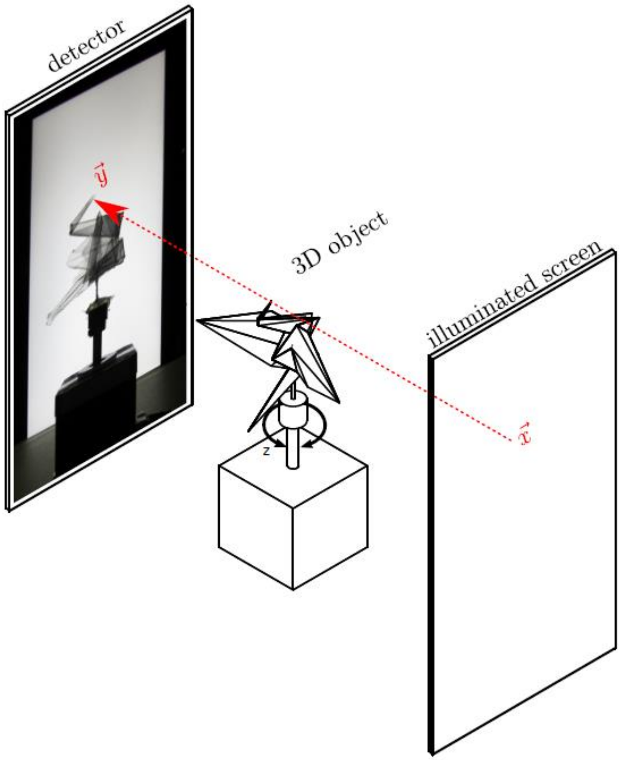


Fig 10.4 Conceptual principle of volume visualization

9.2.2 Tomography



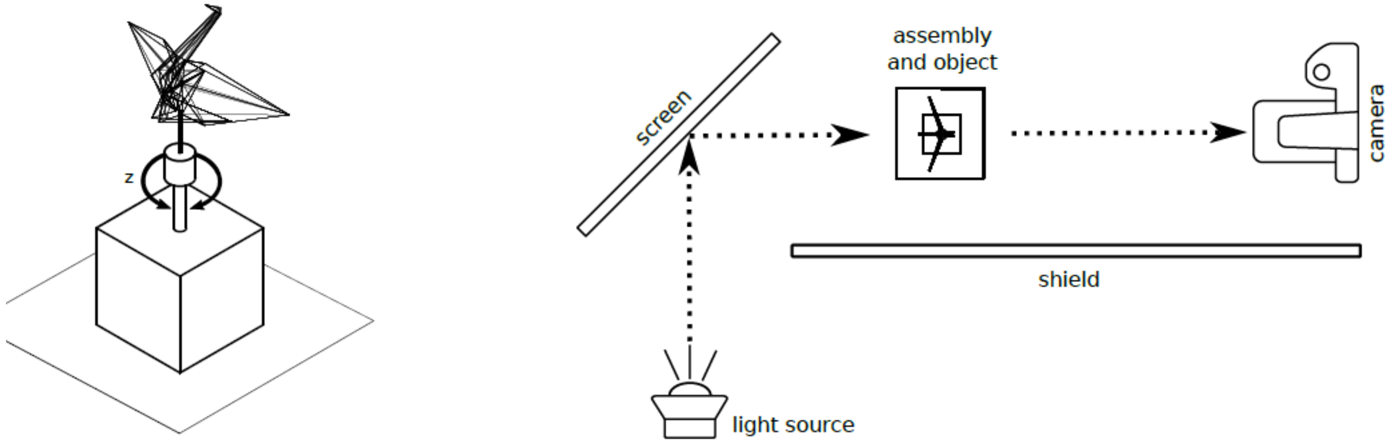
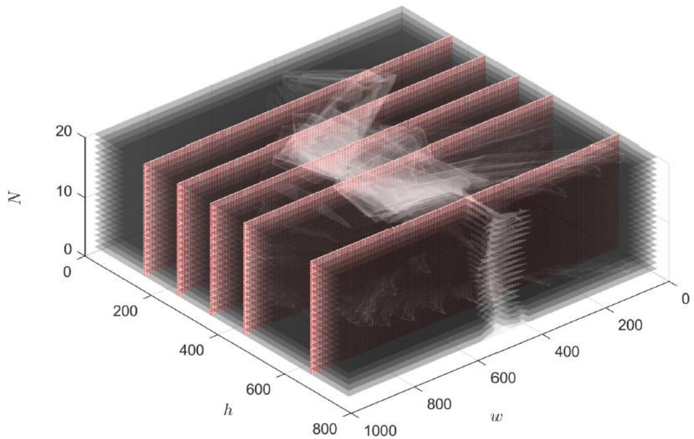
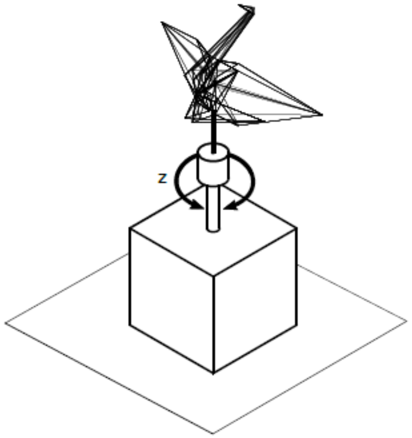


FIGURE 1. *A stepper motor rotates the origami (left) and the proposed imaging system (right).*



Kittenberger, A., Mindrinos, L., & Scherzer, O. (2020). Computed Origami Tomography. *arXiv preprint arXiv:2006.03446*.

9.2.2 Tomography



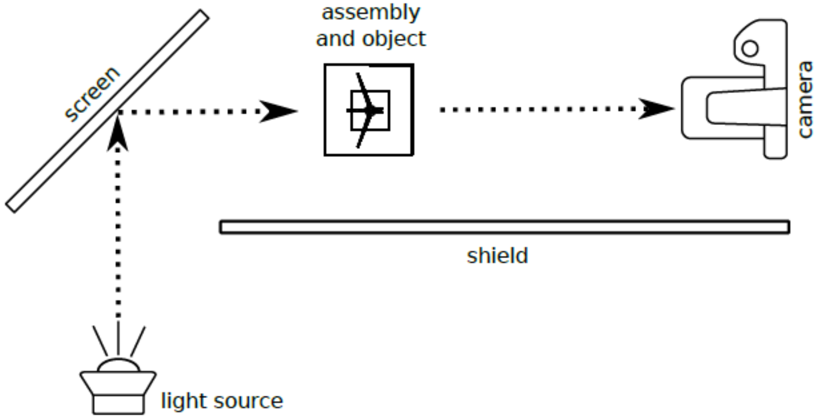


FIGURE 1. *A stepper motor rotates the origami (left) and the proposed imaging system (right).*

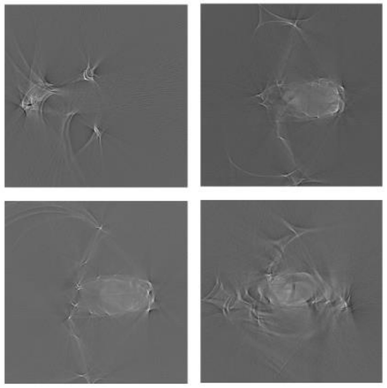
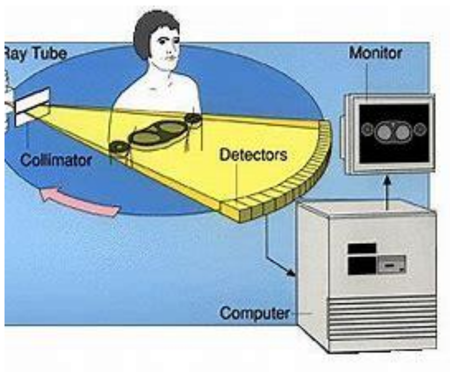


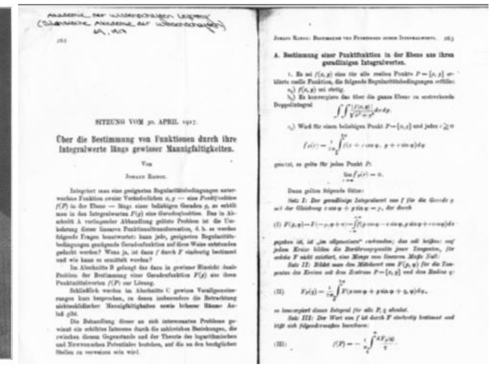
FIGURE 10. *The reconstructed 3D origami crane from the simulated data (400 images). Print-screen of the volume raycaster (left) and its horizontal cross-sections at specific positions (right).*

Kittenberger, A., Mindrinos, L., & Scherzer, O. (2020). Computed Origami Tomography. *arXiv preprint arXiv:2006.03446*.

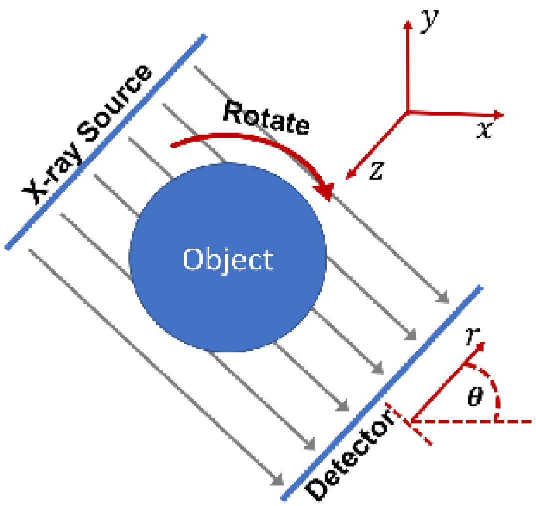
9.2.2 Computed Tomography (CT)

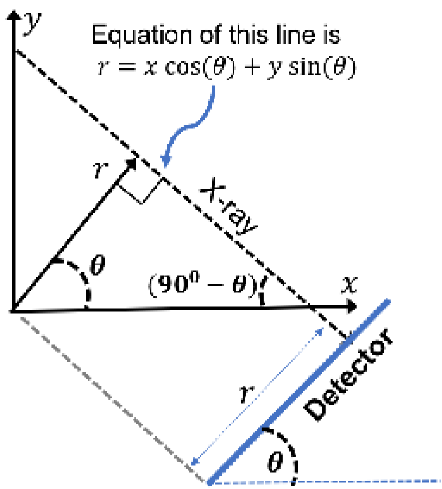


1917
J. Radon

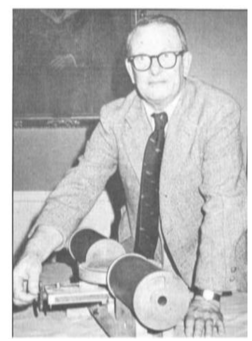


$$\mathcal{R}[f](\theta, s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - s)\,dx\,dy.$$

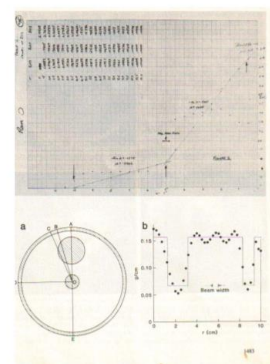




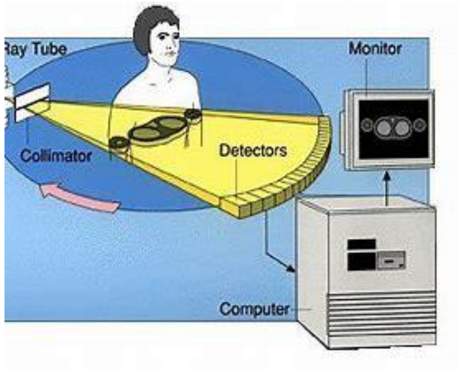
1957…64
A. Cormack

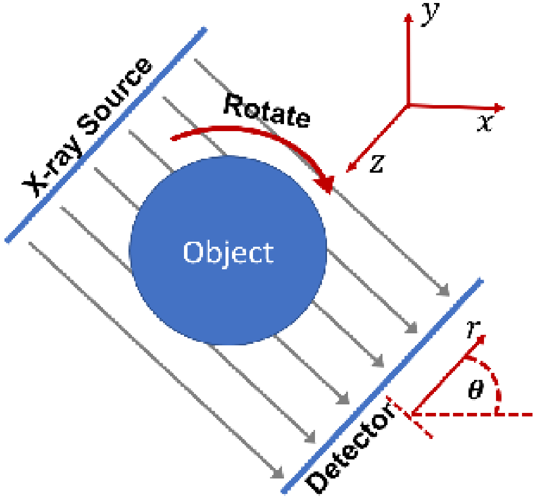


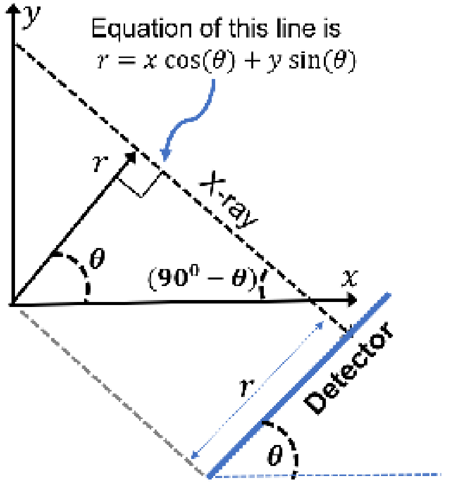
His calculations from the 1960's



9.2.2 Computed Tomography (CT)







The Mathematical Model of X-ray CT

f a function in the plane representing the density of an object
L a line in the plane over which the photons travel.
Parallel Beam Parameterization:
$L = L(\varphi, p) = \{x \in \mathbb{R}^2 \,|\, x \cdot (\cos(\varphi), \sin(\varphi)) = p\}$, $\varphi \in [-\pi, \pi]$, $p \in \mathbb{R}$.
The X-ray (Radon) Transform:

$$\text{Tomographic Data} \sim \mathcal{R}f(\varphi, p) = \int_{x \in L(\varphi,p)} f(x)\,ds$$

–The 'amount' of material on the line the X-rays traverse.

The goal: Recover a picture of the body (values of $f(x)$), from X-ray CT data over a finite number of lines.

Todd Quinto

Tufts

9. Sistemes Avançats II. Contents:

9.3 Maps

- **The maps allow the visualization of geospatial data**
- **By using maps to visualize data, it is hoped that non-technical audiences will be able to comprehend and analyze the involved data more readily.**
- **Yet, there is a science involved in ensuring that the images appropriately reflect the data on which they are based.**

Examples (in Tableau): https://www.tableau.com/learn/articles/interactive-map-and-data-visualization-examples

9.3 Maps

- **Do a good map is not an easy task**
- If you use a map – **people do not remember the data, they only remember how looks like your map**
- **1st question you need to answer – Why is this for?**

(same question – necessary when working on DataViz)

- **2nd question – measure time cost/impact of the outcome**

Good practices: https://youtu.be/nB6dYYGwv2Q

Google Maps offers several APIs for developers, such as Google Earth, Google Maps Images, and Google Places. These tools enable developers to build interactive visual mapping programs for any application or website.

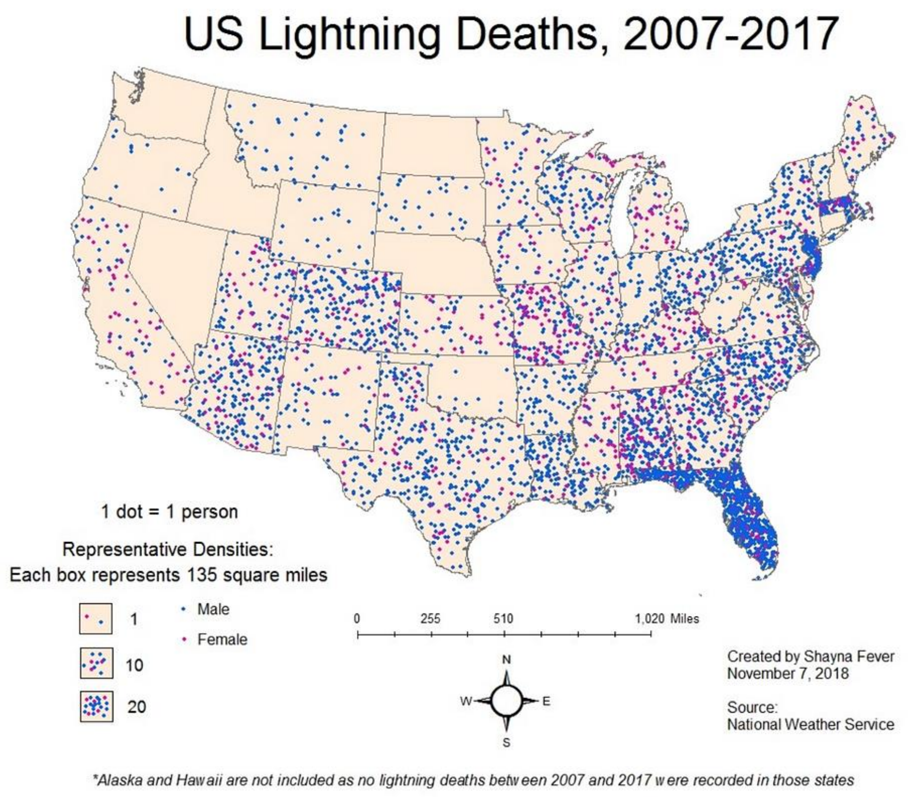
9.3 Maps: Point map

One of the simplest representations of geographical data is a point map. In essence, you set a point at any spot on the map that corresponds to the variable you're attempting to measure (such as a building, e.g.: a hospital, a school, a parking).

Useful for showing distribution and density patterns of things, events.

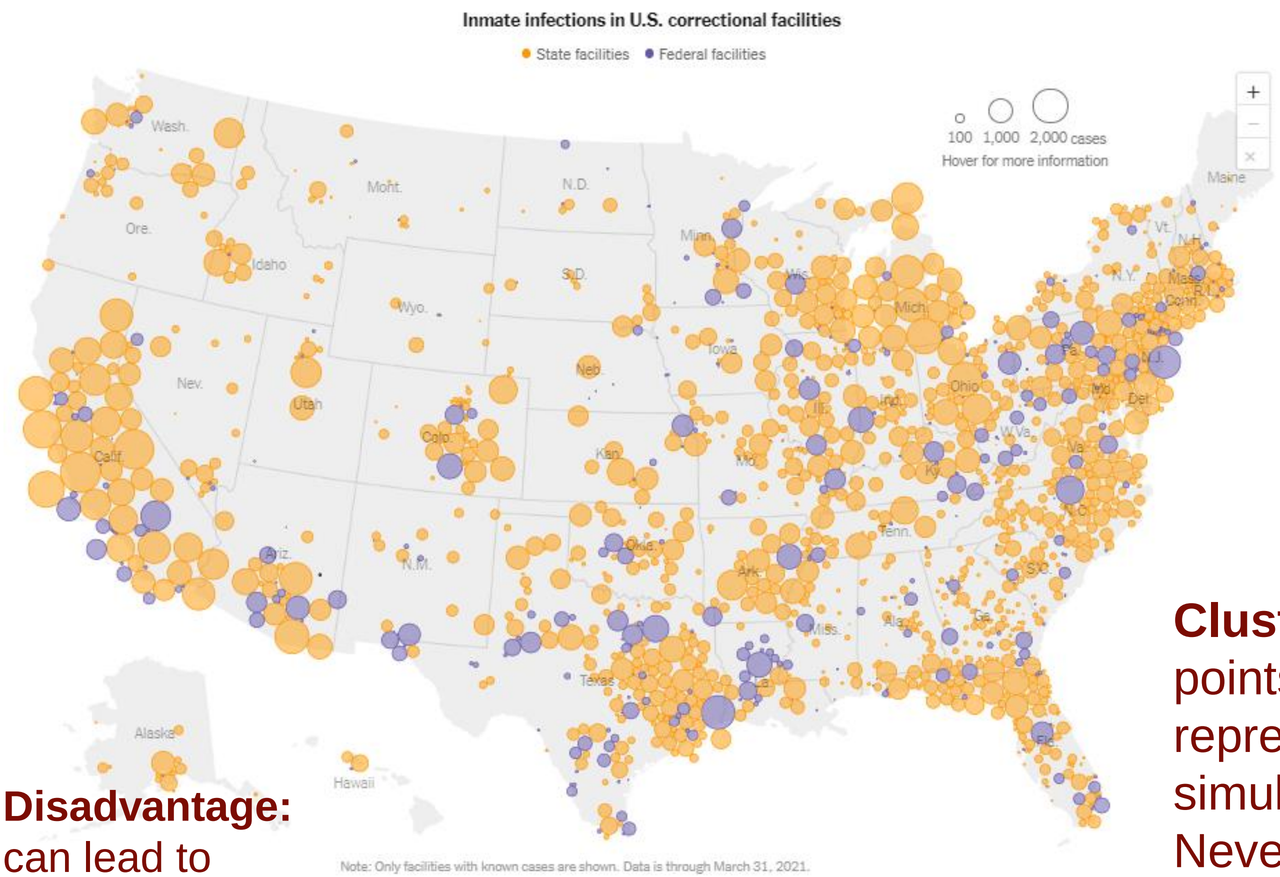
Disadvantages:

- It requires you to collect or geocode location data accurately.
- Depending on zoom levels, points maps can overlap.



https://medium.com/nightingale/data-storytelling-with-maps-5e7845967c9d

9.3 Maps: Proportional symbol Maps



Starting in March of last year, New York Times reporters tracked every known coronavirus case in every correctional setting in the United States, including state and federal prisons, immigrant detention centers, juvenile detention facilities, and county and regional jails.

Coronavirus in American Prisons

https://www.nytimes.com/interactive/2021/04/10/us/covid-prison-outbreak.html

Bubble maps: As bubble charts, it uses circle or other shape to represent data. And based on the point's size and color, it can represent more variables at once.

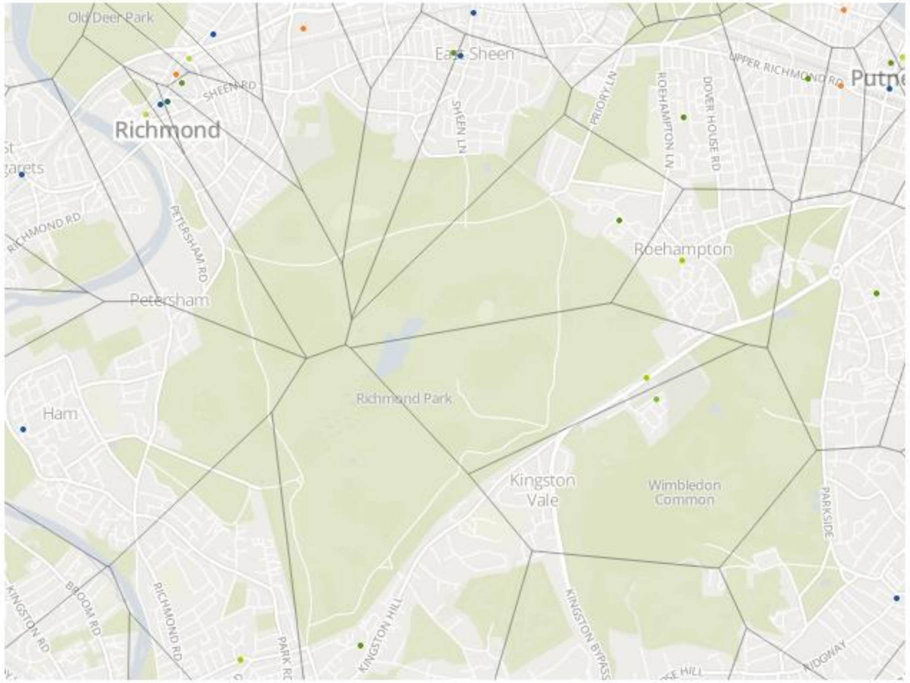
Cluster maps: comparable idea of using points of various sizes and colors to represent numerous types of data simultaneously in one place. Nevertheless, if you zoom in on the map, you may see smaller spots that are represented by these bigger points.

https://www.tableau.com/blog/using-tableaus-clustering-feature-map-layers-get-full-picture

Disadvantage: can lead to overlapping & also need accurate geocodes

9.3 Maps: Voronoi diagrams

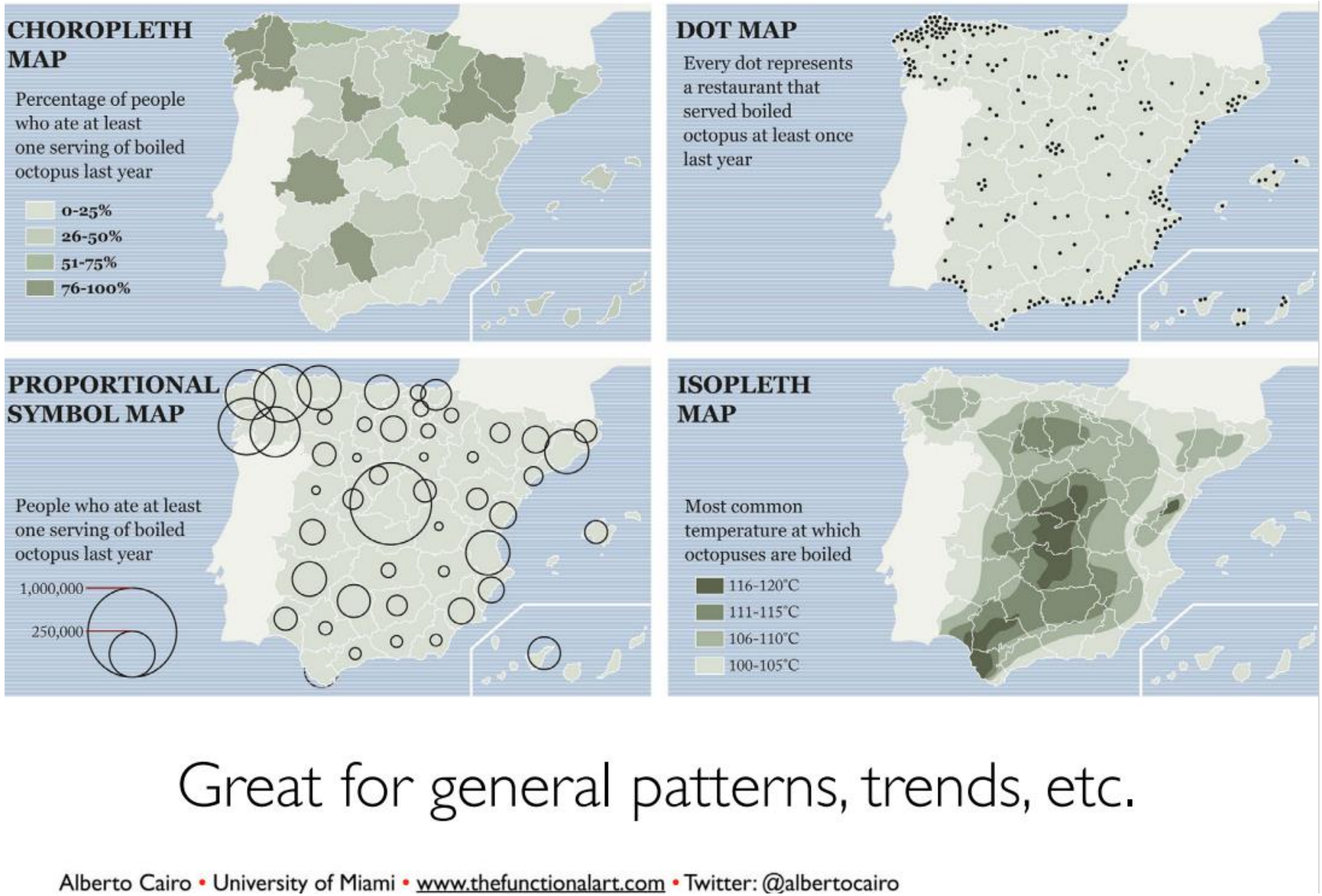
Voronoi diagrams. They are used to find the minimum distance needed to reach a point or a landmark and understand proximity as well as distance features. In the example below, the voronoi map could help choose a new location to build a school that is meant to be far from the existing ones by helping find the largest empty circle amid a collection of points.



Source: Visualising Supermarkets with a Voronoi Diagram

More examples, in "Data Storytelling with Maps": https://medium.com/nightingale/data-storytelling-with-maps-5e7845967c9d

9.3 Maps: Choropleth, dot and isopleth Maps



9.3 Maps: Choropleth, dot and isopleth Maps

It is created by first dividing the area to be mapped into sections, for as by political or geographic boundaries, and then **filling each section with a different hue or shade.**

A different variable, as well as/or a different value or range for a single variable, is represented by each hue or shade. Because of this, **choropleth maps can be used to visualize data clusters over a region while preserving the context of regional boundaries.**

The size of a region may not necessarily have any relationship to the data attributed to it. Therefore, **you need being careful using this style of map with areas where region differ strongly in size**

9.3 Maps: Schematic Maps

Source: CRS route map — Schematic

They are used to simplify complex navigations and system.

Examples: subways, transits, electrical networks

9.3 Much more Maps

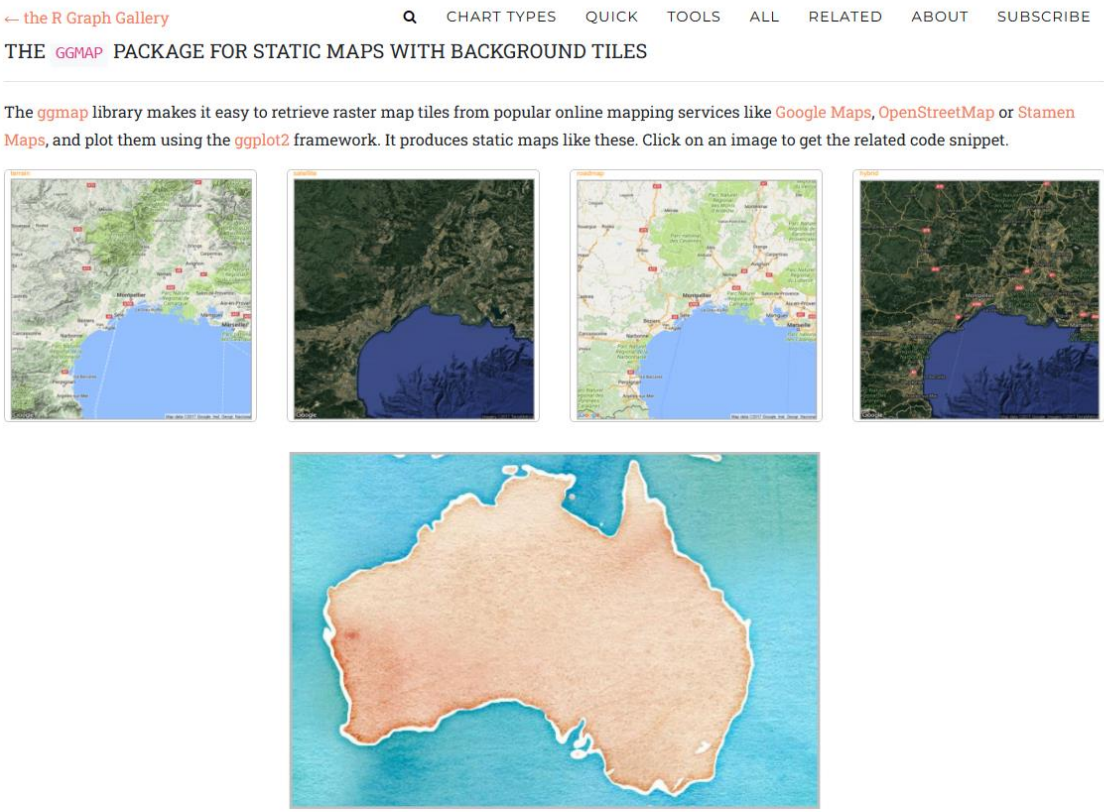
Contents

Some examples in: https://www.finereport.com/en/data-visualization/top-10-map-types-in-data-visualization.html

And more examples in:
https://www.safegraph.com/guides/visualizing-geospatial-data

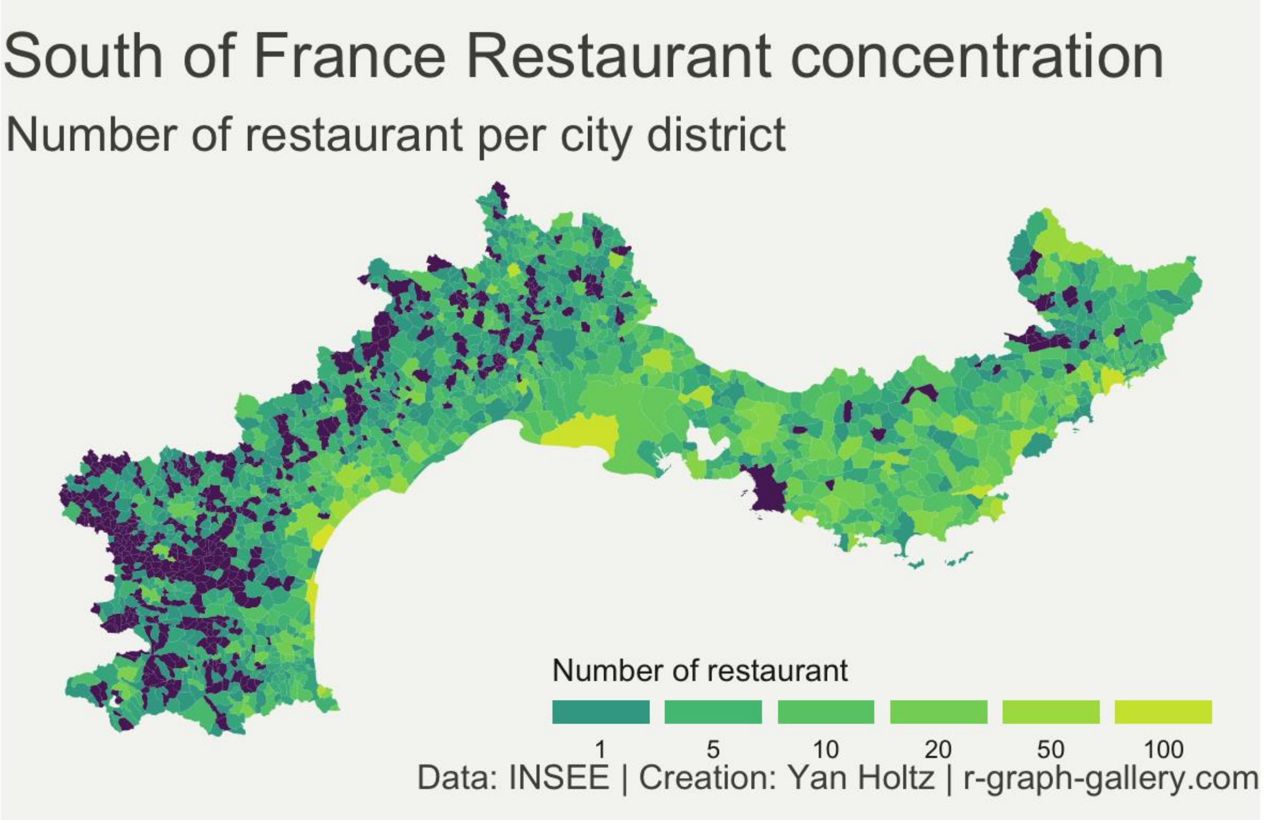
9.3 R graph Gallery

- The background map with R and ggplot2. It explains how to build static maps based on different input map tiles, to use afterwards with ggplot2 to visualize data on it.



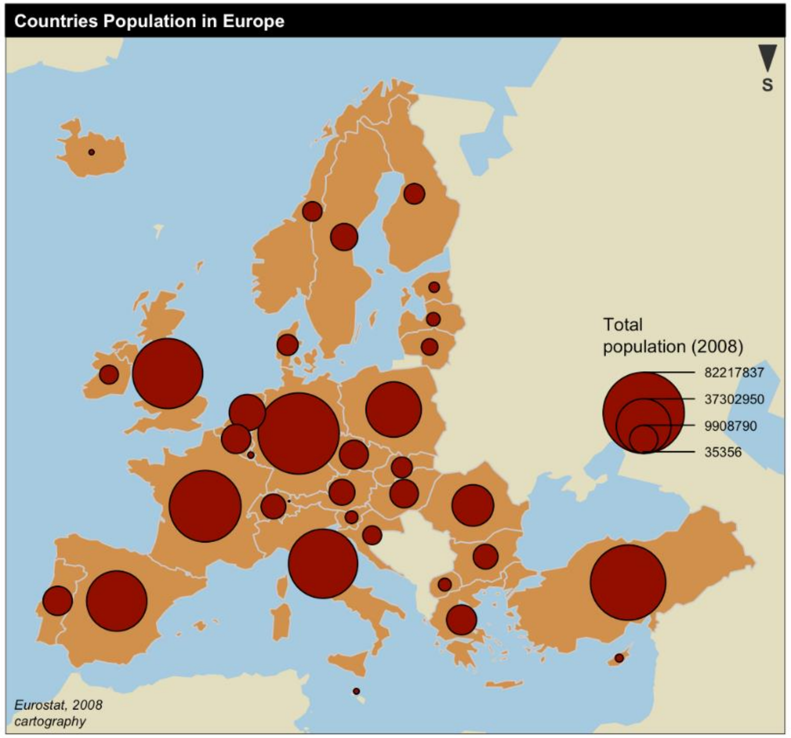
9.3 R graph Gallery

- The cloropleth map with R and ggplot2. It shows how to load geospatial data in R, merge **region** features and build the map.



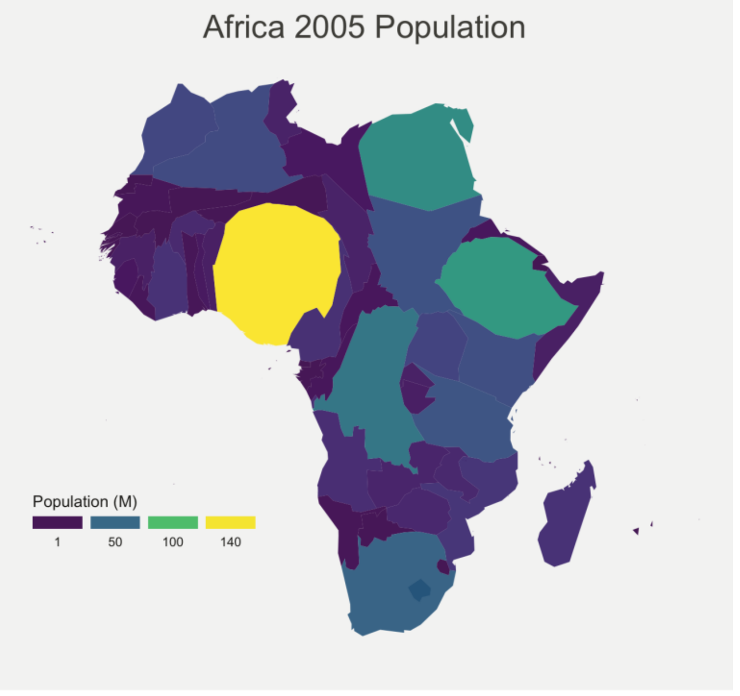
9.3 R graph Gallery

- The bubble map with R and ggplot2. It shows how to turn any of your static bubble map made with ggplot2 interactive (you will see how to do the static ones with me in next seminar and interactivity with Enric). To start to build proportional symbols maps



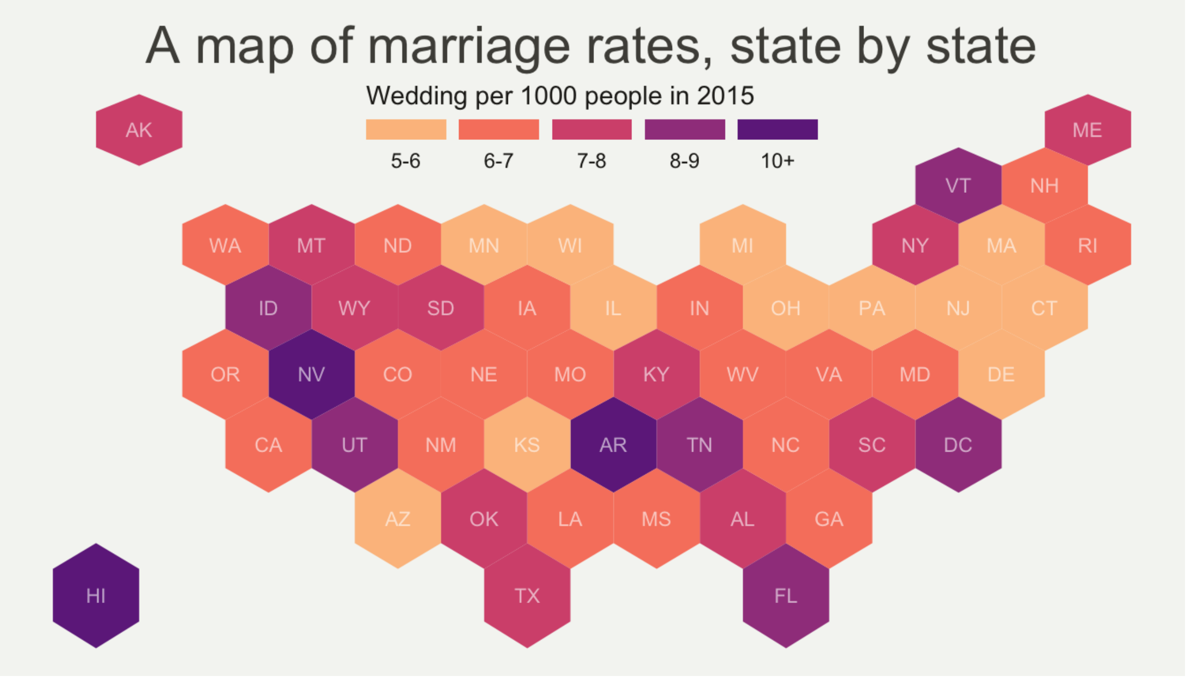
9.3 R graph Gallery

- The cartogram map with R and ggplot2. A cartogram is a map in which **the geometry of regions is distorted in order to convey the information of an alternate variable**.



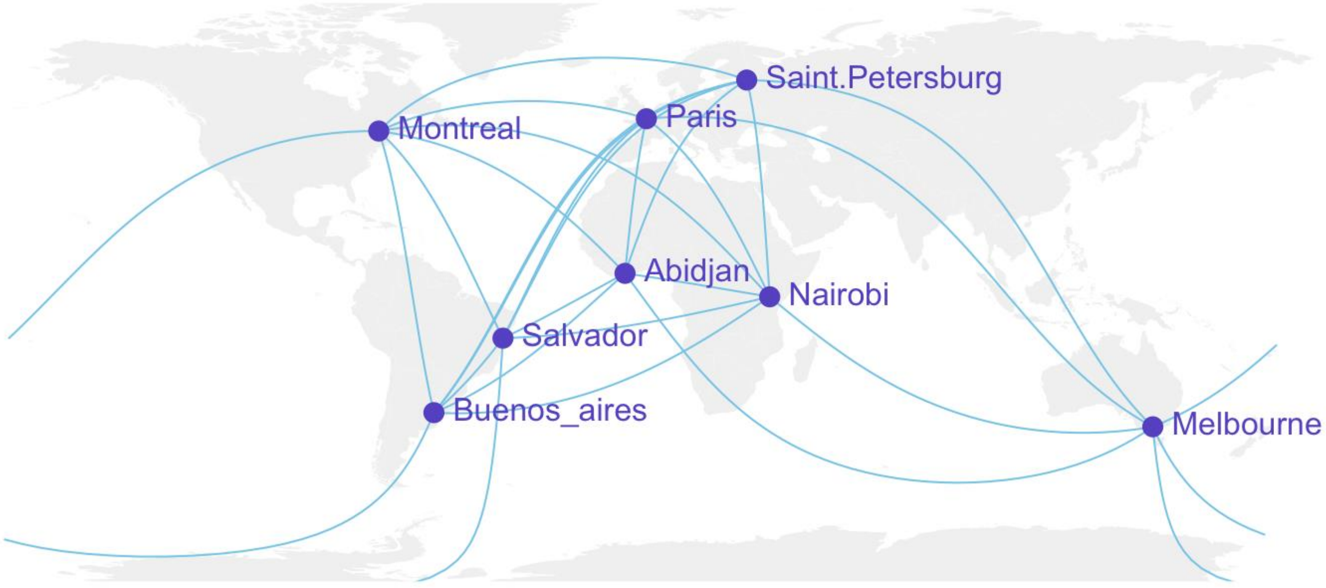
9.3 R graph Gallery

- The hexbin map with R and ggplot2. It refers to two different concepts: i) It can be based on a geospatial object where all regions of the map are represented as hexagons. Or ii) it can refer to a 2d density technique (displaying the relationship between 2 numeric variables)



9.3 R graph Gallery

- The connection map with R and ggplot2. It shows the **connections between several positions on a map.** The link between 2 locations is usually drawn using great circle: the shortest route between them.



9.3 Interactive Maps

- **Example:** The desirability of Boston streets based on the analysis of deviations in pedestrian trajectories obtained from GPS data, comparing the shortest and actual routes between the corresponding points across over 120,000 journeys.
- The most desirable streets in Boston can be easily discovered using a special interactive map recently published by the MIT Senseable City Lab.

See : https://senseable.mit.edu/desirable-streets/

- **Story telling maps:** Pinellas County is the worst place in Florida to be black and go to public school | Failure Factories | Tampa Bay Times

(You will see more about interactive maps with Enric)

9. Sistemes Avançats II. Contents:

9.4 3D or 2D? : Depth

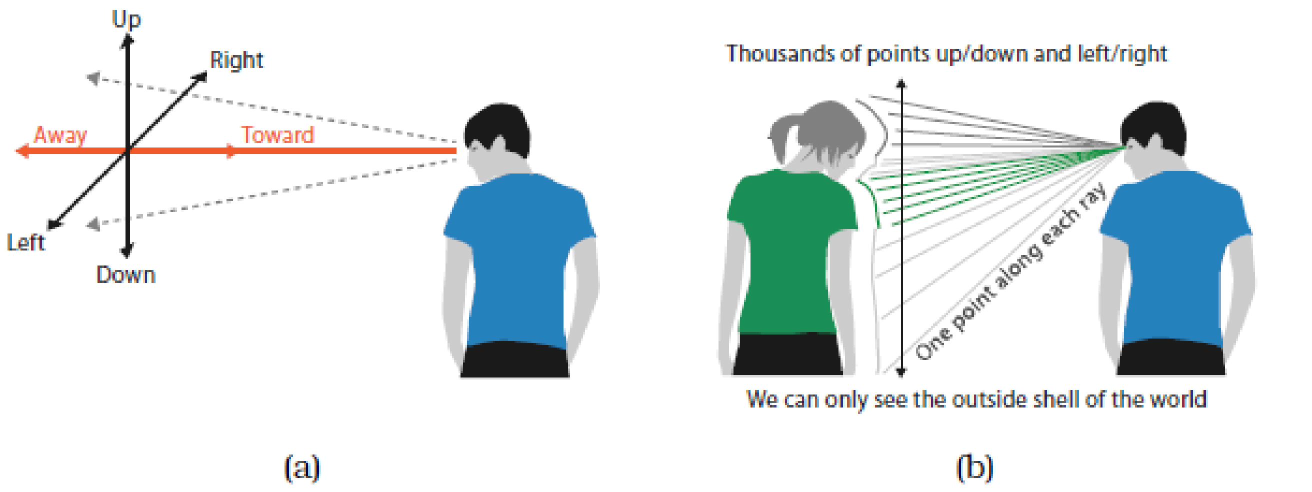


Figure 6.2. Seeing planar position versus depth. (a) The sideways and up–down axes are fundamentally different from the toward–away depth axis. (b) Along the depth axis we can see only one point for each ray, as opposed to millions of rays for the other two axes. After [Ware 08, page 44].

T. Muzner

*Considered perceptually, **the inaccuracy of depth judgements is** also **not surprising**; **the common intuition that we experience the world in 3D is misleading**. We do not really live in 3D, or even 2.5D: to quote Colin Ware, we see in 2.05D [Ware 08].*

9.4 3D or 2D? : Occlusion

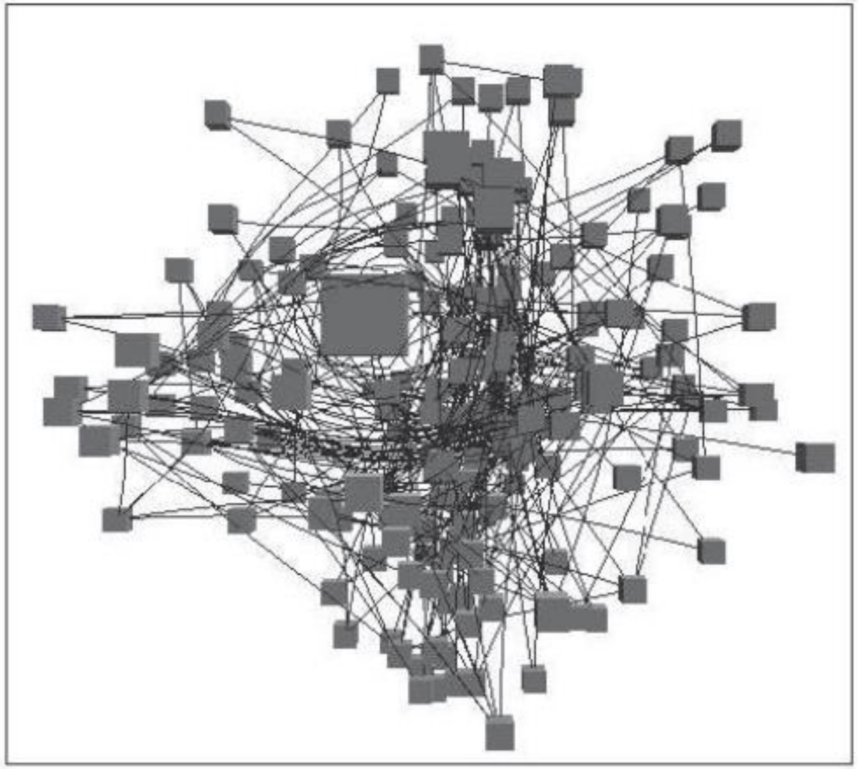


Figure 6.3. Resolving the 3D structure of the occluded parts of the scene is possible with interactive navigation, but that takes time and imposes cognitive load, even when sophisticated interaction idioms are used, as in this example of a node–link graph laid out in 3D space. From [Carpendale et al. 96, Figure 21].

T. Muzner

9.4 3D or 2D? : Perspective Distortion Dangers

The phenomenon of perspective distortion is that **distant objects appear smaller and change their planar position on the image plane.**

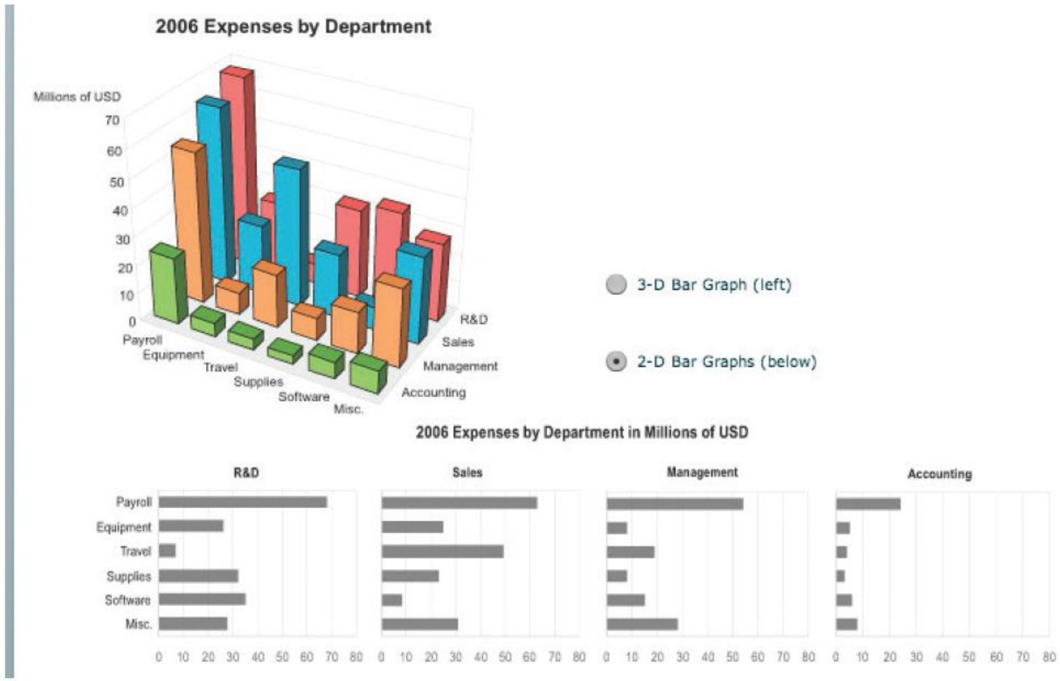
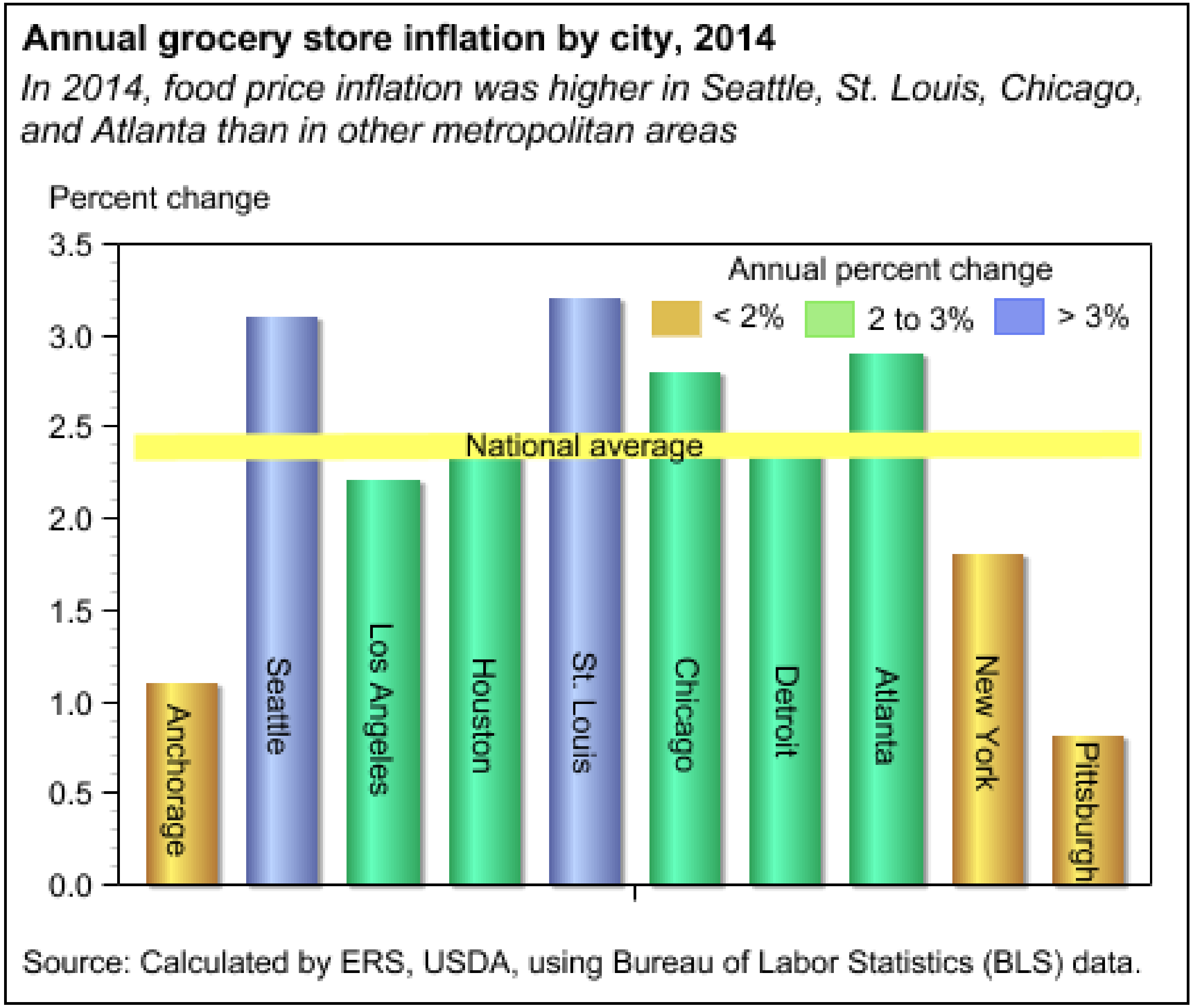


Figure 6.4. 3D bar charts are more difficult than 2D bar charts because of both perspective distortion and occlusion. From [Few 07, Question 7].

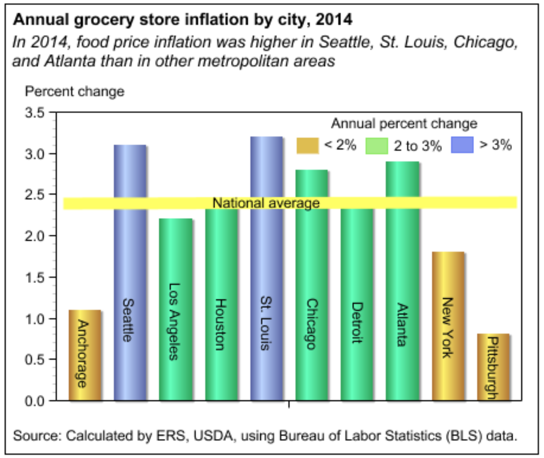
T. Muzner

9.4 3D or 2D? Quality metrics



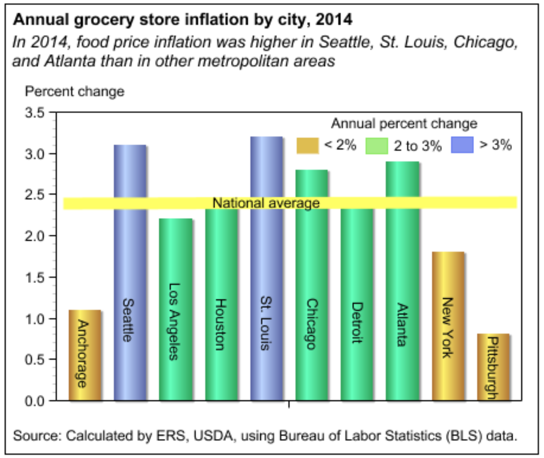
9.4 3D or 2D? Quality metrics

- **Remove 3-D effect**: A 3-D chart showing 2-D data doesn't add value to your charts but does add noise and makes the chart harder to read.
- **Use softer color tones except where you want to draw attention**. Bright colors can be tiring to look at. Use tools like the **Coblis Color Blindness Simulator to determine whether your visual is color-blind friendly.**
- **Pivot the chart:** Use the horizontal bar chart when you have category labels that are many characters long so that users don't have to tilt their heads to read the chart.



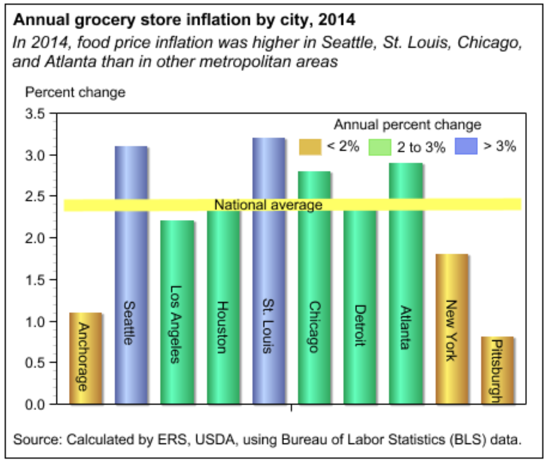
9.4 Other quality metrics

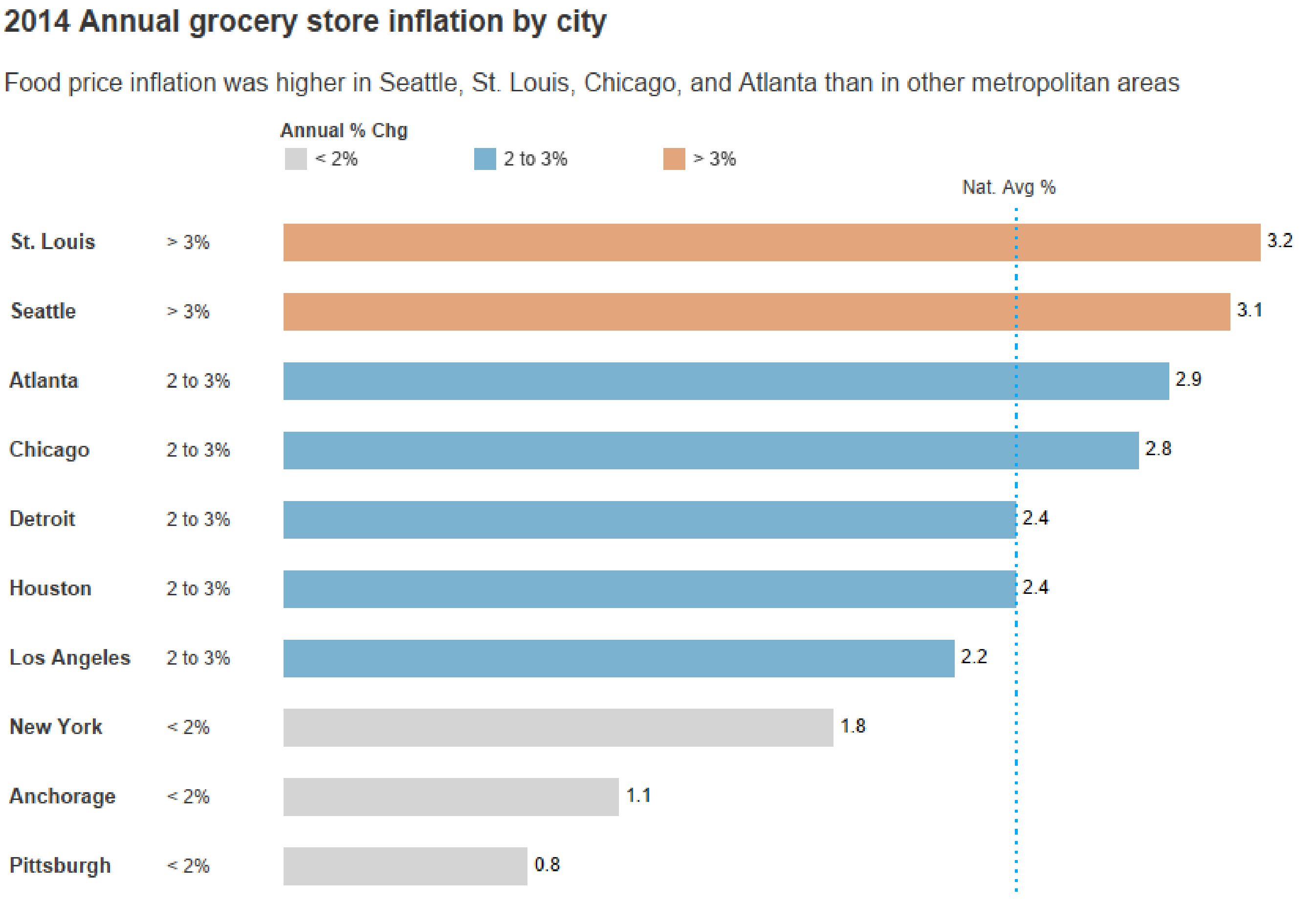
- **Choose an appropriate bar width:** Make bars skinnier to use just enough white space so that the bars are slightly thicker than the white space between them. **When the bars are too close together, your brain will naturally try to evaluate the area vs. length.**



- **Remove the x-axis header and add data labels.** By adding data labels, you can easily see that the national average is 2.4%. Since we added data labels, the x-axis would be redundant and thus may be removed

9.4 3D or 2D? Quality metrics





Melissa Yu 2016

9.4 3D or 2D?

We will use 3d, when the **viewer's task fundamentally requires understanding the three-dimensional geometric structure of objects or scenes**. Otherwise not

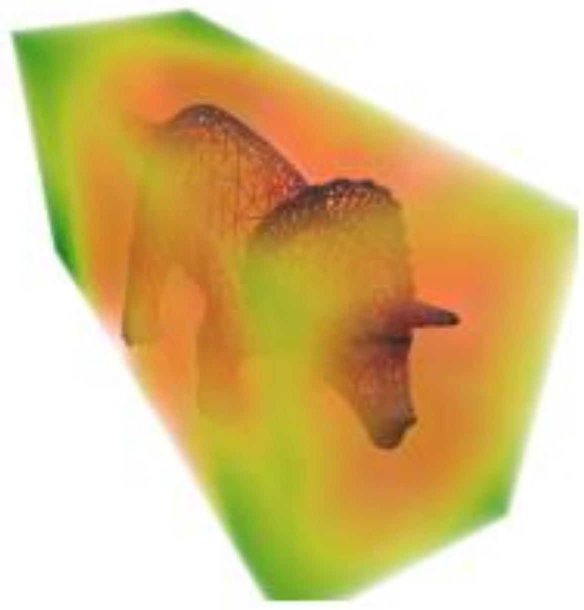


Figure 7: **3D Distance Field**: *Computation of 3D distance field and discrete Voronoi diagram of Triceratops model (5660 polygons). Distance increases from red to green. Grid size* $= 255 \times 111 \times 84$, *Computation time* $= 520ms$.

Sud, A., Govindaraju, N., & Manocha, D. (2005, October). **Interactive computation of discrete generalized voronoi diagrams using range culling.** In *Proc. International Symposium on Voronoi Diagrams in Science and Engineering*.

9.4 3D or 2D?

We will use 3d, when the **viewer's task fundamentally requires understanding the three-dimensional geometric structure of objects or scenes**. Otherwise not

1. Realtime Performance-Based Facial Animation: http://www.youtube.com/watch?v=PlNeS1U0Pwo

3d necessary here. However, the interesting point here is how combining 2D and 3D data adequately can reduce the computational cost. The visualization is in 3D.
Note the use of PCA (seen before).

9.4 3D or 2D?

- The great benefit of using **3D** comes **when the viewer's** task fundamentally **requires understanding the three-dimensional geometric structure of objects or scenes.**

Example: https://vimeo.com/513025319

- In almost all of these cases, **a 3D view with interactive navigation controls** to set the 3D viewpoint will allow users to construct a useful mental model of dataset structure more quickly than simply using several 2D axis-aligned views.
- Measure the cost-effort/time – versus impact of the visualization
- Some 3D Data Visualization Software Tools: http://www.intellspot.com/3d-data-visualization-tools/

9. Sistemes Avançats II. Contents:

9.5 Scientific visualization

Short communication

Ten guidelines for effective data visualization in scientific publications

Christa Kelleher, Thorsten Wagener

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.envsoft.2010.12.006   Get rights and content

Abstract

Our ability to visualize scientific data has evolved significantly over the last 40 years. However, this advancement does not necessarily alleviate many common pitfalls in visualization for scientific journals, which can inhibit the ability of readers to effectively understand the information presented. To address this issue within the context of visualizing environmental data, we list ten guidelines for effective data visualization in scientific publications. These guidelines support the primary objective of data visualization, i.e. to effectively convey information. We believe that this small set of guidelines based on a review of key visualization literature can help researchers improve the communication of their results using effective visualization. Enhancement of environmental data visualization will further improve research presentation and communication within and across disciplines.

https://doi.org/10.1016/j.envsoft.2010.12.006

9.5 Scientific visualization

Short communication

Ten guidelines for effective data visualization in scientific publications

Christa Kelleher, Thorsten Wagener

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.envsoft.2010.12.006   Get rights and content

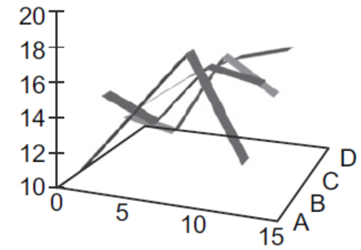
https://doi.org/10.1016/j.envsoft.2010.12.006

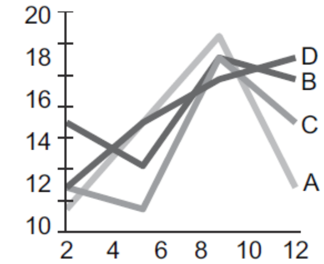
- ***The objective of any graphic in the context of scientific publications and presentations is to effectively convey information.***

- *The ten guidelines proposed here represent an effort to reduce common pitfalls in the pursuit of this objective. Above all, **these guidelines should be taken as general recommendations that can be used to*** **improve visualization design, and not as absolute rules that apply in every case.**

9.5 Scientific visualization

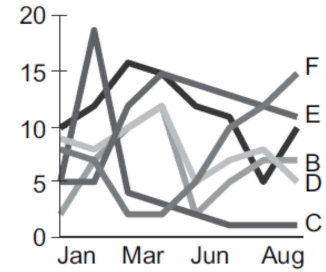
1. Create the **simplest graph** that conveys the information you want to convey (Tufte, 1983 [pp. 91e137])

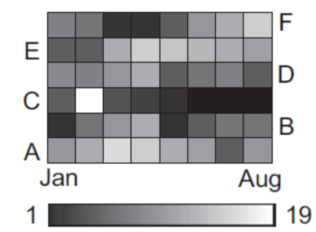




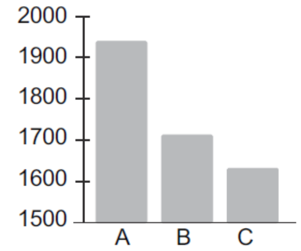
2. Consider the **type of encoding object and attribute** used to create a plot (Chambers et al., 1983 [pp. 137e140]; Cleveland and McGill, 1984)

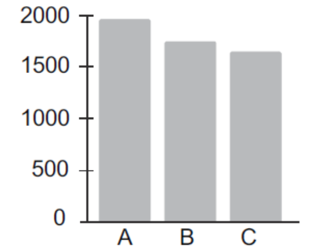
3. **Focus on** visualizing patterns or on visualizing details, **depending on the purpose of the plot** (Few, 2004a; Kosslyn and Chabris, 1992)





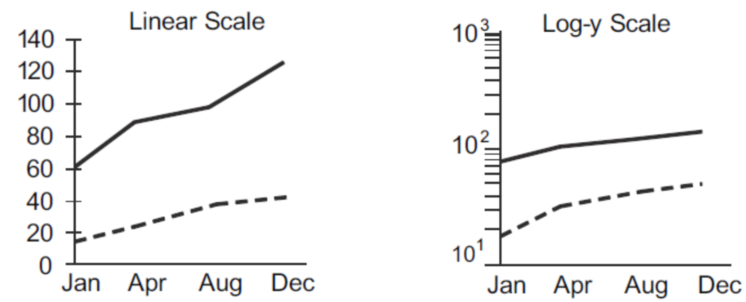
4. Guideline 4: **select meaningful axis** ranges (Robbins, 2005 [pp. 239e241; 285]; Tufte, 2006 [p. 60]; Strange, 2007 [p. 89])





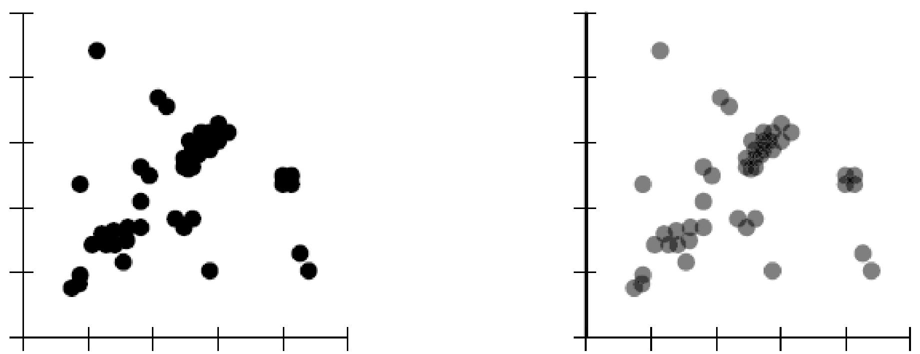
9.5 Scientific visualization

5. Data transformations and carefully chosen **graph aspect ratios** can be used to emphasize rates of change for timeseries data (Cleveland, 1994 [p. 66, 95, 103])



6. Plot **overlapping points** in a way that density differences **become apparent in scatter plots** (Few, 2009 [p. 121]; Cleveland, 1994 [p. 159])

(same -6- would apply to maps)



9.5 Scientific visualization

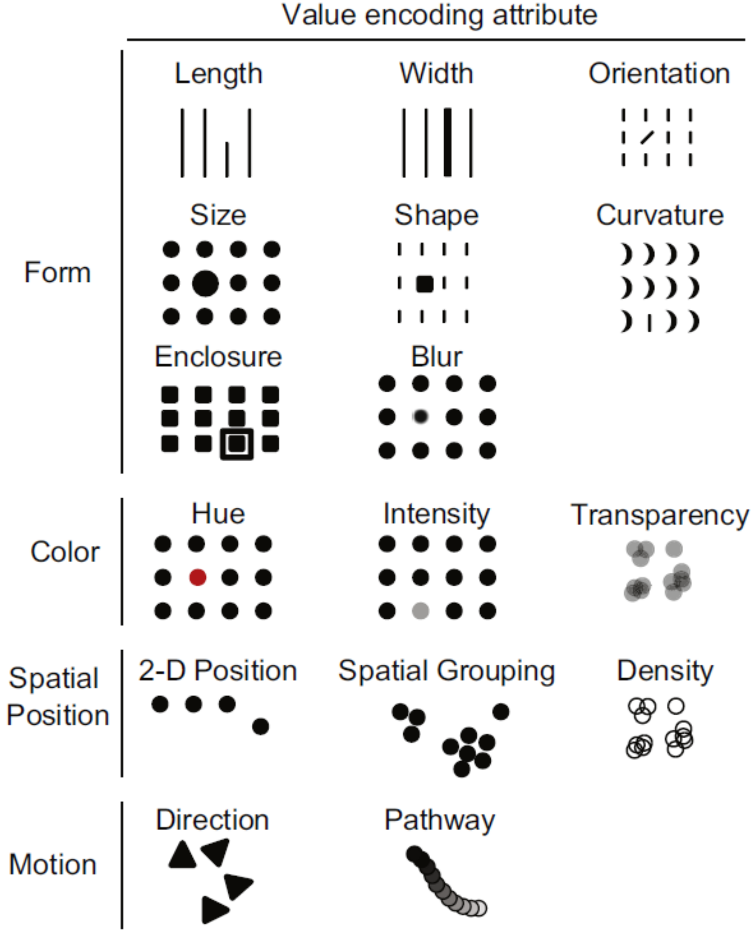
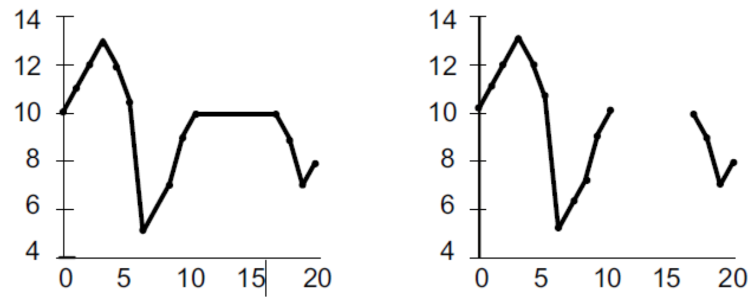


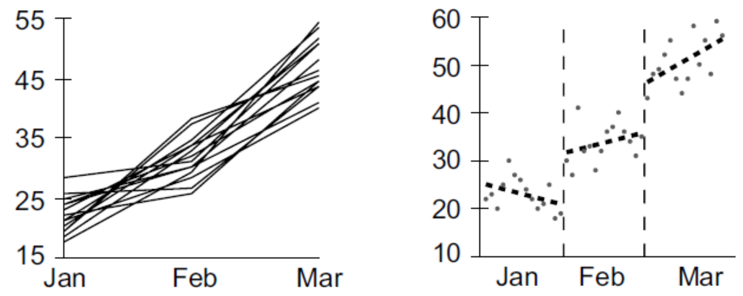
Fig. 2. The following attributes corresponding to different encoding objects can be used in data visualization to highlight contrasts between different parts of a graphic, to represent quantitative differences, or to categorize and group information. Adapted from Few (2009, p. 39).

9.5 Scientific visualization

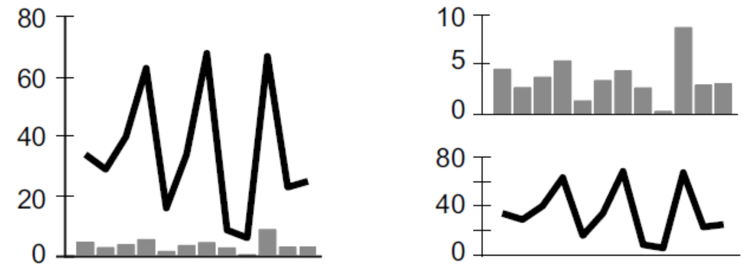
7. Use **lines** when connecting sequential data in **time-series plots** (Strange, 2007 [p. 150])



8. Aggregate larger datasets in meaningful ways (Cleveland and Devlin, 1980; Chambers et al., 1983 [pp. 21e24]; Cleveland, 1994 [p. 187])

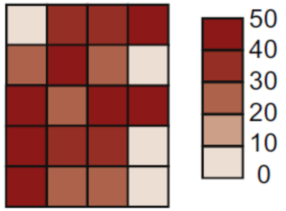


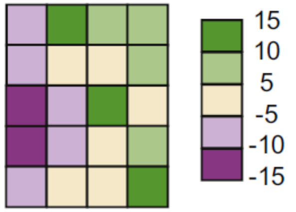
9. Keep **axis ranges as similar as possible** to compare variables (Cleveland, 1994, [pp.86e87]; Few, 2009 [p. 180])



9.5 Scientific visualization

10. Select an **appropriate color scheme** → based on the type of data (Brewer, 1994; Harrower and Brewer, 2003)





(Vist en classe de color amb Guillermo)

*A main practice that has not yet been generally accepted is **the inclusion of uncertainty estimates or error bars in the visualization of both observed and modeled data.***

It has been argued and demonstrated in many places that such estimates improve decision-making and provide a better reflection of our scientific understanding (e.g. Reichert and Borsuk, 2005; Beven,2006).

9. Sistemes Avançats II. Contents:

9.1. Multi-variable (continuation class VII)

9.1.1. Visualizing many proportions at once (continuation of 7.1.3).

9.1.2. Multi-panel figures

9.1.3. Multivariate numerical data

9.2. Dimensionality reduction (continuation class VI)

9.2.1. T-Distributed stochastic neighbour embedding (t-SNE)

9.2.3. Tomography- Slice along a plane, 2D isosurfaces for a 3D field, isocontours

9.3. Maps

9.4. 2D or 3D ?

9.5. Scientific visualization

9.6. Other quality metrics

9.6 Other quality metrics

- **Identifying and removing chart clutter reduces visual "noise", allowing the audience to focus on the information.**
- **Edward Tufte describes such noise as "elements in charts that are not necessary to comprehend the information represented in the graph."**
- Clutter denotes **a disordered collection of graphical entities in information visualization**.
- Clutter **can obscure the structure present in the data**.
- Even in a small dataset, **clutter can make it hard for the viewer to find patterns**, **relationships and structure.**

9.6 Other quality metrics

- **Dimension order is an attribute that can significantly affect a visualization's expressiveness.** By varying the dimension order in a display, it is possible to reduce clutter without reducing information content or modifying the data in any way.
- Common chart clutter items include:
  - 3-dimensional effects
  - Dark gridlines (use soft gray gridlines or eliminate gridlines when possible)
  - Overuse of bright, bold colors
  - Unnecessary use of all uppercase text (uppercase text is only necessary when calling attention to an element)

9.6 Other quality metrics

Conceptualize, create, critique and cut

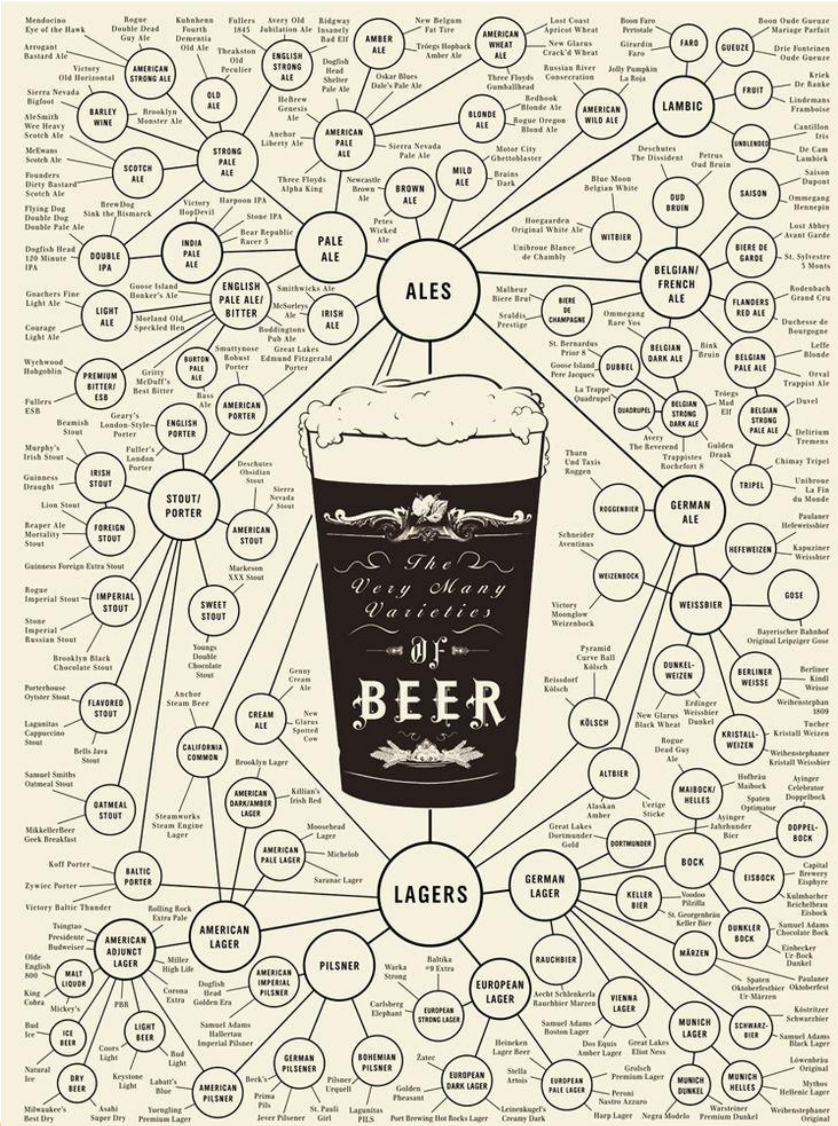
- **Conceptualize —** It's always a good idea to create a **rough plot of your infographic before you begin creating**.
- **Come up with a catchy and descriptive title**.
- **Establish a beginning poi**nt to contextualize the information you are presenting. Similarly, **have an idea of what the ending point** will be and come to some kind of conclusion or call to action. **Figure out what path** you want the reader's eyes to take. (Hypothesis)
- **Create** — Craft your data story. **Have fun with it. Include interesting and surprising information that the reader might not have seen before. Personalize.**

9.6 Other quality metrics

Conceptualize, create, critique and cut

- **Critique —** Ask yourself: **is your data story consistent** (same verb tense, consistent terminology, consistent design)? **Have someone else read your data story before publishing.**
- **Cut — Remove information that is extraneous or unnecessary**. **Keep information** that is **essential and interesting**. Don't overwhelm your reader with too much text or too many images. Ask yourself: can extra information can be removed and set aside for another infographic? Would a certain point be better communicated as an image? Could a part be arranged differently to improve the flow of your data story?

9.6 Other quality metrics



The reader is confronted by a wall of text and shapes.

The title and the image of the beer in the centre of the page looks good and is thematically relevant, but **the overwhelming number of shapes pulls the eyes in too many directions**.

Every space on the page is filled, making it **difficult for the reader to follow any particular narrative path.**

9.6 Other quality metrics

This one is much easier to read.

The subject of the infographic is focused on **one type of beer** and clearly stated in the title.

The text is concise and the whitespace on the page allows the reader to focus on and consider each point.

Colour variation gives the infographic visual interest, while still maintaining a consistent colour scheme.

SARA MCGUIRE, 2015

Thanks for your attention!

Judit Chamorro Servent

Departament de Matemàtiques

judit.chamorro@uab.cat

11. Storytelling

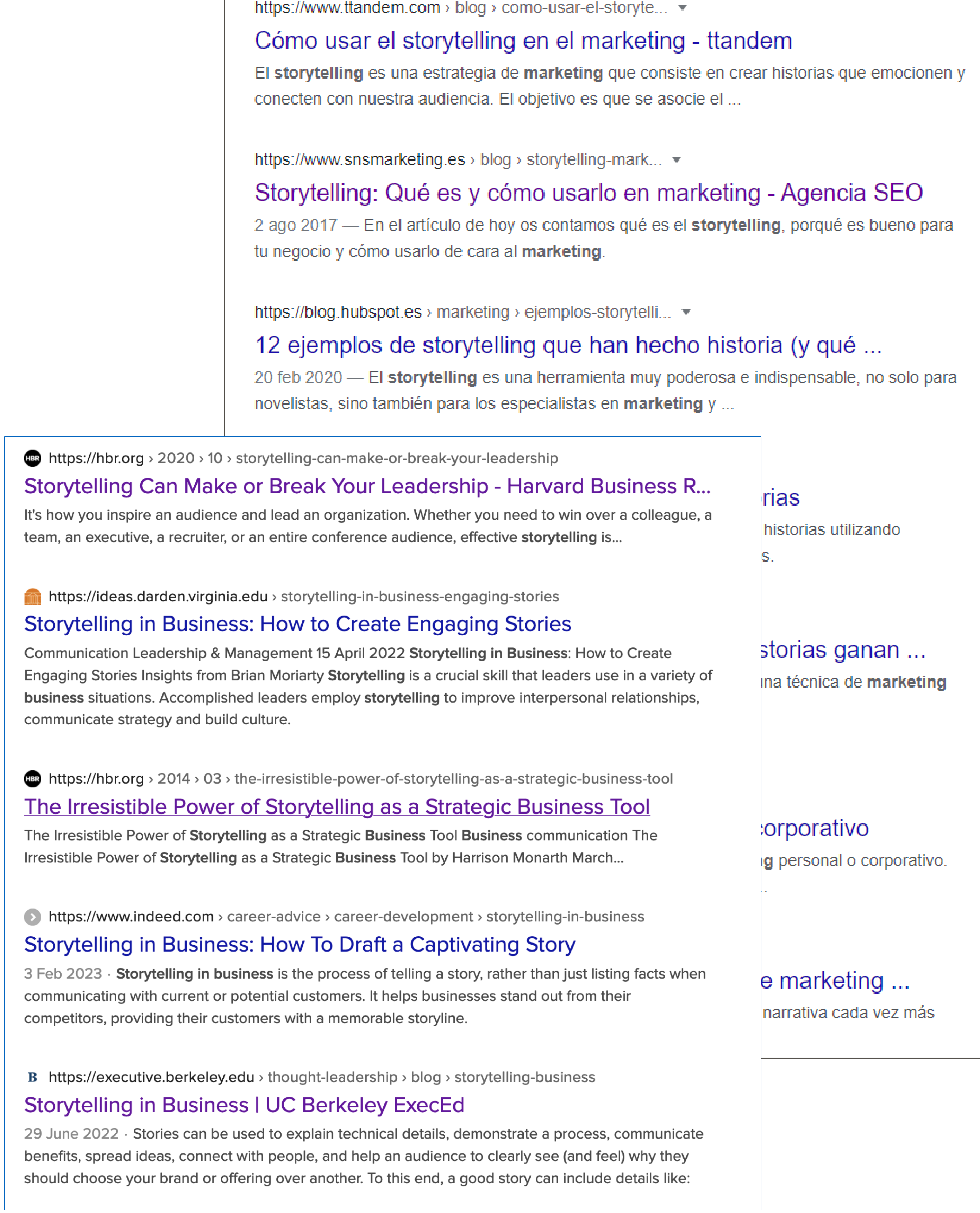
Guillermo Marin
guillermo.marin@uab.cat

Storytelling

- Storytelling = Narrativa
- Técnicas que operan detrás de cualquier tipo de historia
- Narración **estructurada** de una serie de hechos y acciones llevadas a cabo por uno o más personajes
- Técnica de comunicación aplicada en áreas diversas para crear un vínculo emocional con la audiencia
- Aplicación en áreas diversas para sobresalir.

Qué tiene que ver conmigo?

- Se pueden contar historias con datos
- Técnicas útiles para comunicación en general
- Herramientas para dirigir el interés y conectar con la audiencia
- Storytelling + visualizaciones, maximiza el impacto en la comunicación de datos.



Data Analyst
Wallapop
Barcelona, Catalonia, Spain (On-site)
upf. 35 school alumni
3 weeks ago

role preferably for a web/mobile company
- Excellent statistical knowledge, especially concepts related to A/B testing
- Proficiency in SQL
- Experience working with reporting and data visualization tools like Looker, Tableau or Power BI
- Product and business acumen
- Great data visualization with well-structured storytelling

Senior Data Scientist
PepsiCo
Barcelona, Catalonia, Spain (Hybrid)
1 mutual connection with the hiring team
Promoted

development.
- Business storytelling and communicating data insights in business consumable format. Fluent in one Visualization tool.
- Strong communications and organizational skills with the ability to deal with ambiguity while juggling multiple priorities
- Experience with Agile methodology for team work and analytics 'product' creation. Fluent in Jira, Confluence.

Software Engineering Lead (Full Stack) - BCG X
Boston Consulting Group (BCG)
Madrid, Community of Madrid, Spain (Hybrid)
upf. 47 school alumni
Promoted

Job Requirement:
- Bachelor's / Master's degree in computer science engineering/technology or equivalent
- Excellent communication skills and storytelling

Power BI Specialist (m/f/d)
DEHN digital solutions GmbH
Spain (Remote)
Actively recruiting
Promoted · Easy Apply

Key Responsibilities
- Develop and maintain Power BI data models, reports, and dashboards connecting to the cloud data warehouse (Azure Synapse) which is collecting data from various sources such as SAP ERP and Microsoft Dynamics CRM as per the DEHN standards
- Communicate trends, patterns, insights, findings, and recommendations to non-technical stakeholders using data visualizations and storytelling

Data/Python Engineer Intern
TomTom
Madrid, Community of Madrid, Spain (Hybrid)
Your profile matches this job
Promoted · 20 applicants

- You are fluent in English
- You have programming experience with Python and SQL
- You have experience working with the main python libraries for data processing (pyspark, pandas, numpy,...)
- You have experience with data visualization (storytelling) (e.g., dash, streamlit)
- Being familiar with Microsoft cloud computing (Azure & Databricks)

What We Offer

Business Intelligence (Tableau) Developer ESP
Teladoc Health
Barcelona, Catalonia, Spain (Hybrid)
11 mutual connections with the hiring team
Promoted · Easy Apply

-Excellent communication skills, both written and oral - comfortable recommending and presenting solutions to senior managers.
-Proficiency in querying complex data structures using SQL.
-Excels at data storytelling and visualization.
-Experience with multiple BI tools a plus, especially IBM Cognos, PowerBI, Tableau.
-Tableau REST/JS API experience is a plus.

Analytics Engineer
SeQura
Spain (Remote)

- Provide and extract insights and actionable recommendations with the purpose of helping the business units to make the best decisions
- Real time analytics, alert dashboards and storytelling with data
- Contribute to the development of the Data Visualization Platform
- Communicate solutions and present them to both business and technical stakeholders

Staff ML Research Scientist
Crisis Text Line
Greater Barcelona Metropolitan Area (Remote)

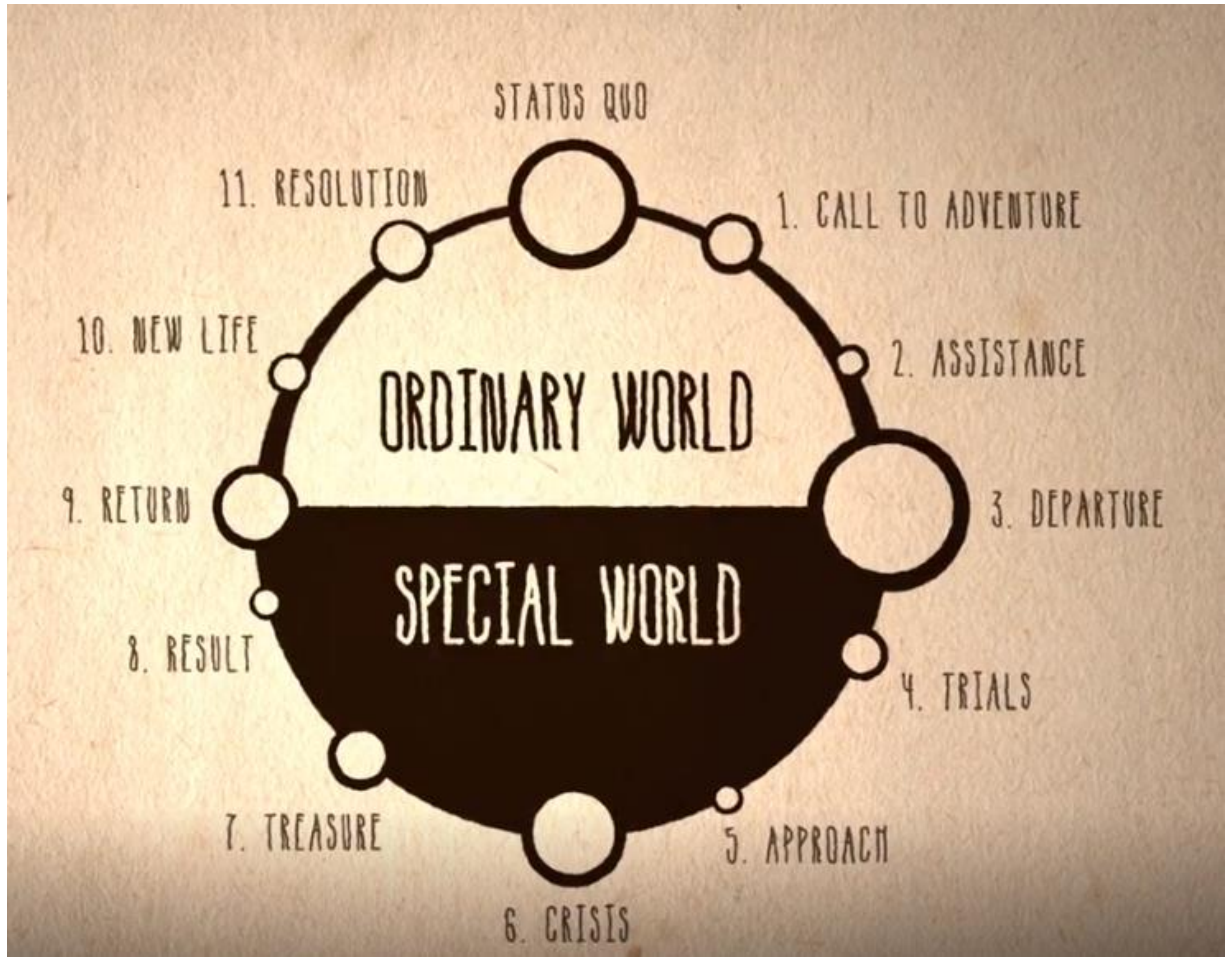
Communicate Results & Support Data Storytelling
- Communicate and share model performance and impact in a digestible way with the rest of the team and organization
- Implement new data visualization techniques
- Support efforts to communicate results for non-technical audiences, including ongoing management of organization-wide internal data insights and research

Data & AI Presales Engineer - Spain
Fujitsu
Greater Madrid Metropolitan Area (Hybrid)
1 mutual connection with the hiring team
Promoted · Easy Apply

Would you like to be a key part of our growth-oriented organization? Do you enjoy designing and building the storytelling of end-to-end AI and Data solutions? Do you have the capacity to handle a lot of information at the same time? Are you customer oriented and have really good communication & presentation skills? **If so, we are looking for you to join our Data Intelligence and AI team in Spain!**

Story

- Story- Narración **estructurada** de una serie de hechos y acciones llevadas a cabo por uno o más personajes
- Good story / well told (Robert McKee). Algo que vale la pena contar / Los medios usados para involucrar profundamente a la audiencia y, en última instancia, recompensarla con una experiencia significativa y conmovedora
    - Comunicar una idea envuelta en emoción=Memorable
- Las historias tienen características universales:
    - **El viaje del héroe.** What makes a hero? https://www.ted.com/talks/matthew_winkler_what_makes_a_hero?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

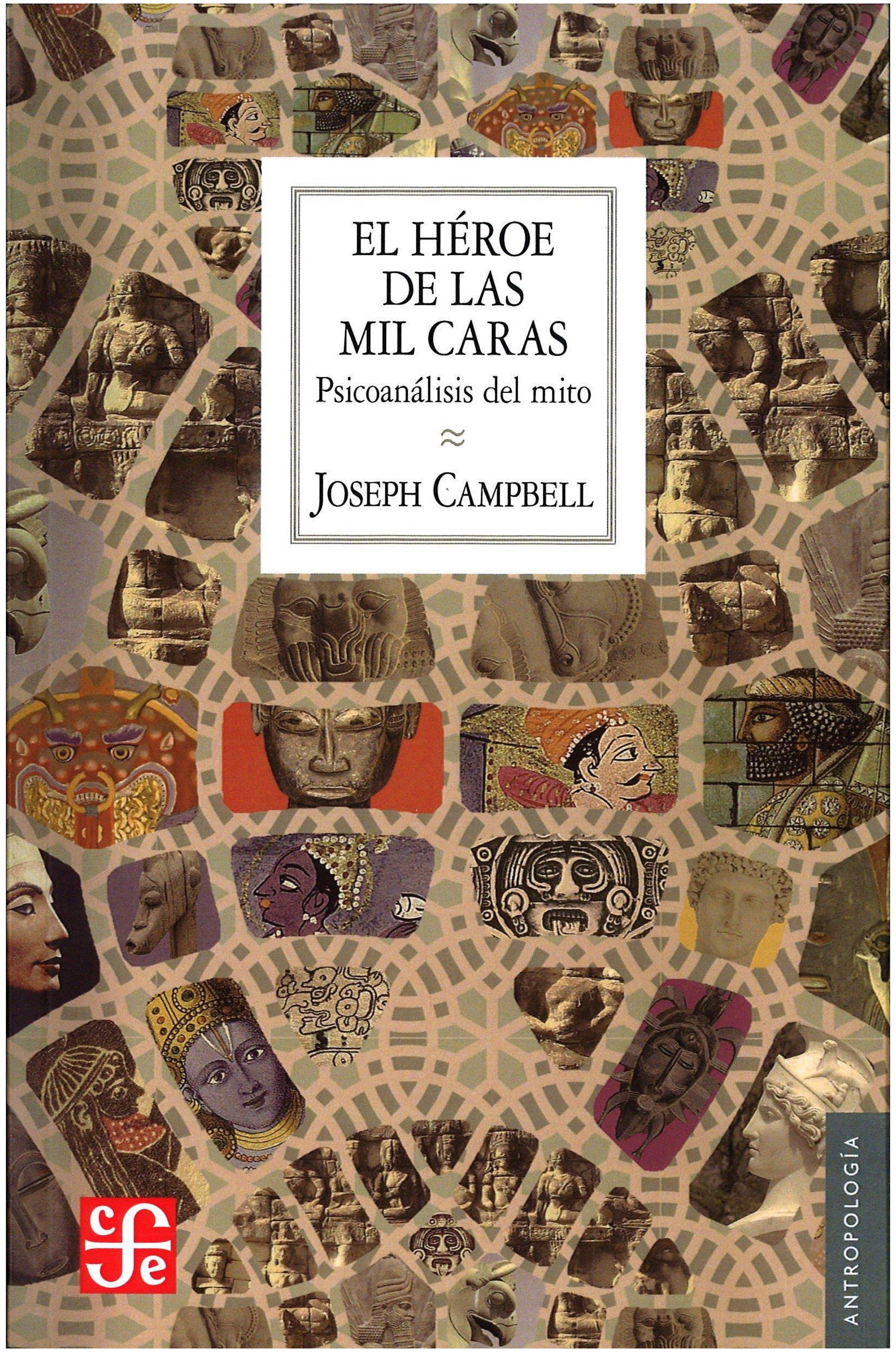


What makes a hero? Kirill Yeretsky

El viaje del héroe

- El héroe de las mil caras, Joseph Campbell
  - Mitos comparten una estructura común fundamental: el monomito
  - “El viaje del héroe”
  - *“El héroe inicia su aventura desde el mundo de todos los días hacia una región de prodigios sobrenaturales, se enfrenta con fuerzas fabulosas y gana una victoria decisiva; el héroe regresa de su misteriosa aventura con la fuerza de otorgar dones a sus hermanos.”*

- Estructura con gran influencia.
- Variaciones y adaptaciones, por ej.:
  - Dan Harmon’s story Circle https://youtu.be/RG4WcRAgm7Y



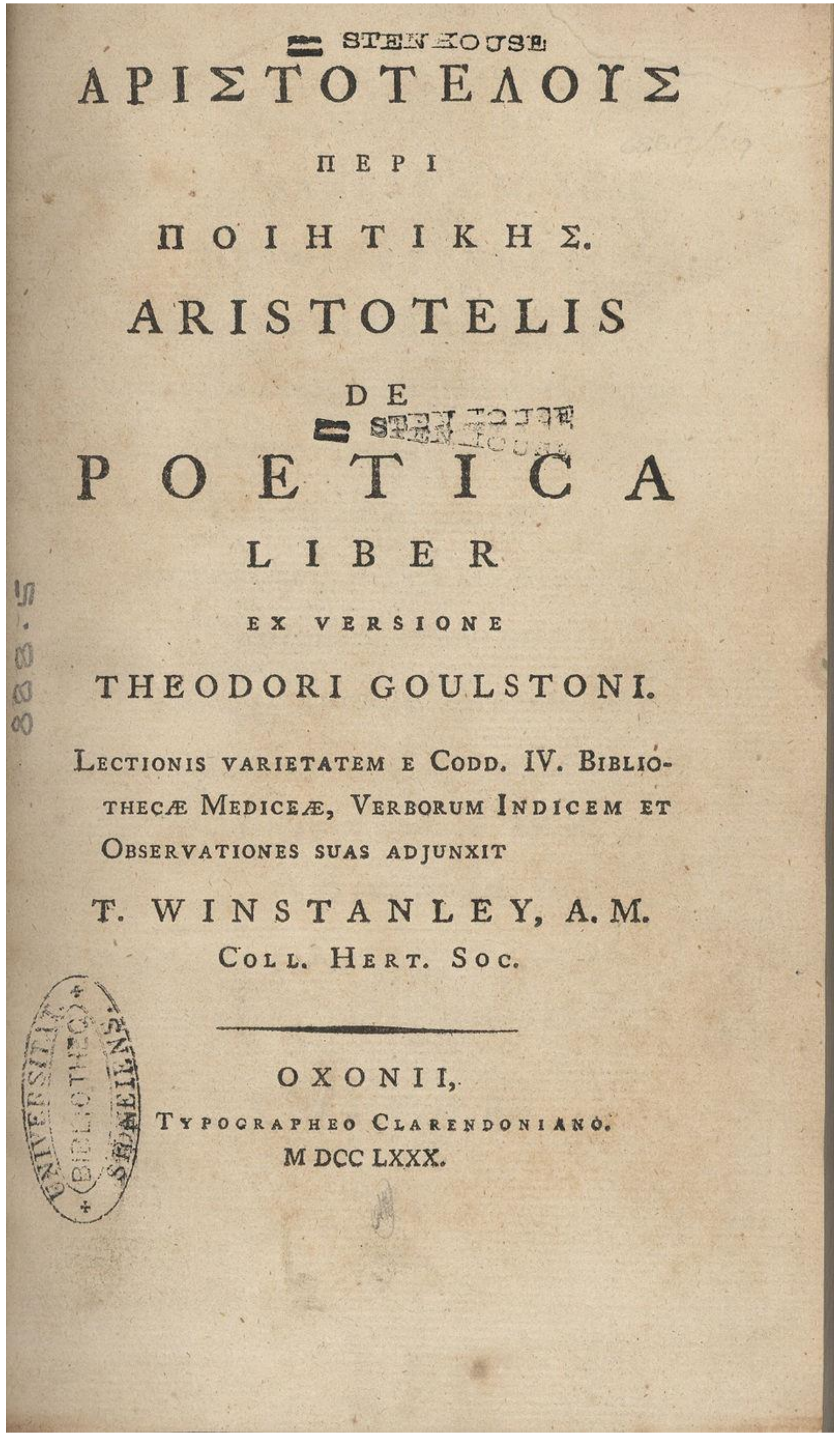
The magical science of storytelling | David JP Phillips

https://youtu.be/Nj-hdQMa3uA

Estructura

- Aristóteles: Estructura clásica para tragedia en 3 actos (introducción-nudo-desenlace)
- Desarrollado como: planteamiento – conflicto - resolución
  - 1. Presentación del personaje y su mundo
    - Inciting incident -> Conflicto
  - 2. Esfuerzos por resolver el conflicto / Cambio
    - Climax
  - 3. Resolución
- Las transiciones están marcadas por situaciones de conflicto que generan tensión
- Otras estructuras (5 actos, story circle, …)

- La estructura en tres actos sirve como modelo para comunicación en general
- El **conflicto** y la **tensión** son una parte integral de las historias y clave para mantener la atención



ΑΡΙΣΤΟΤΕΛΟΥΣ
ΠΕΡΙ
ΠΟΙΗΤΙΚΗΣ.
ARISTOTELIS
DE
POETICA
LIBER
EX VERSIONE
THEODORI GOULSTONI.
Lectionis varietatem e Codd. IV. Bibliothecæ Mediceæ, Verborum Indicem et Observationes suas adjunxit
T. WINSTANLEY, A. M.
Coll. Hert. Soc.

OXONII,
Typographeo Clarendoniano.
M DCC LXXX.

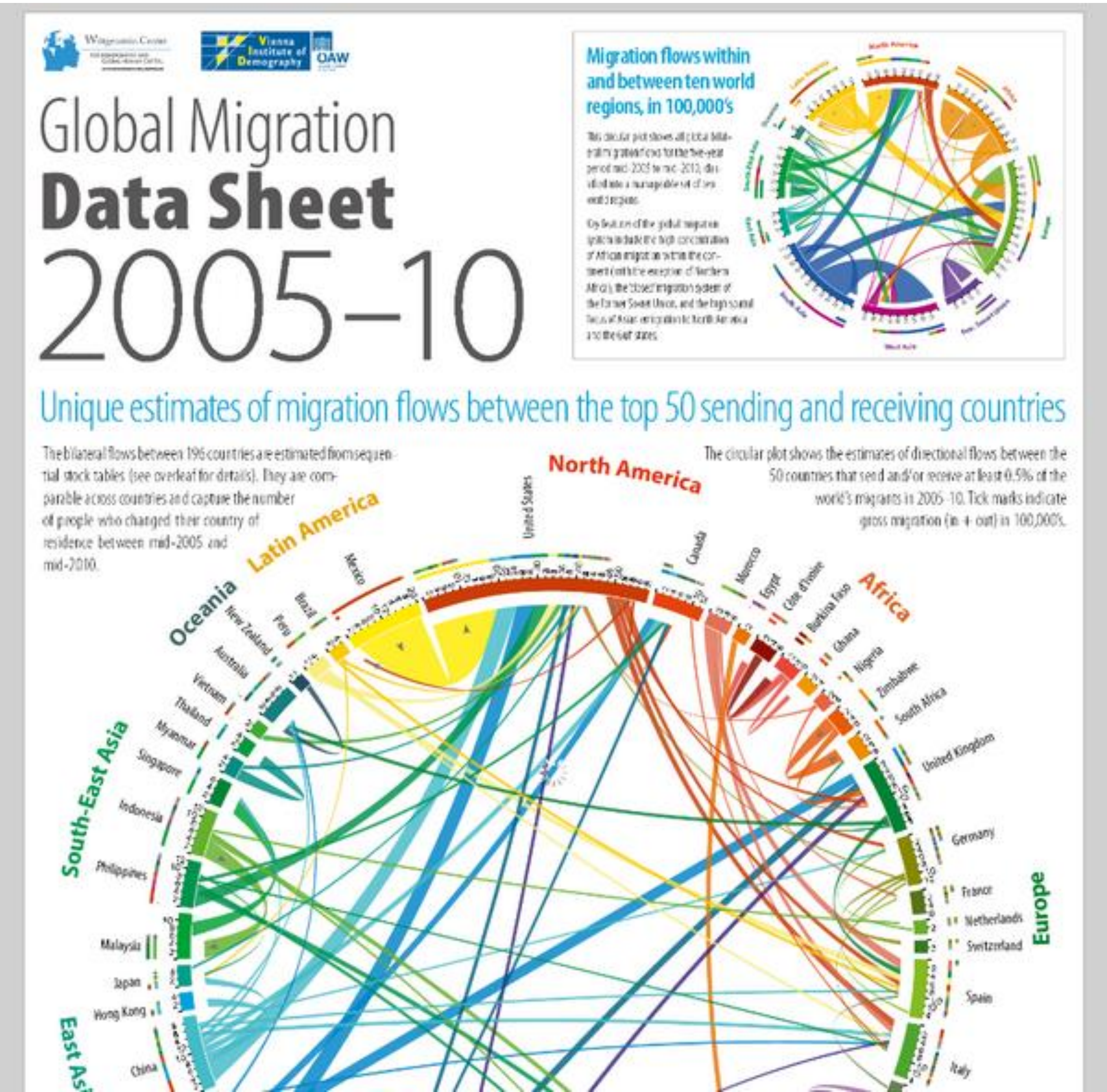
Porqué utilizamos narrativa

- Una historia puede provocar emociones (intriga, empatía…) con las que segregamos hormonas (dopamina, oxitocina, endorfinas,…) que maximizan la atención o ayudan a crear un vínculo con la historia/personaje
  - Para atraer y mantener la atención de una **audiencia**
  - Para transmitir un mensaje de forma memorable a través de la emoción

- Elementos que se pueden aplicar a otras formas de comunicación: por ej., presentaciones en público.
  - Podemos incluir humor, intriga, tristeza, etc. Dramatizar narración.
  - Principios de la narrativa son aplicables a una presentación.

Data viz

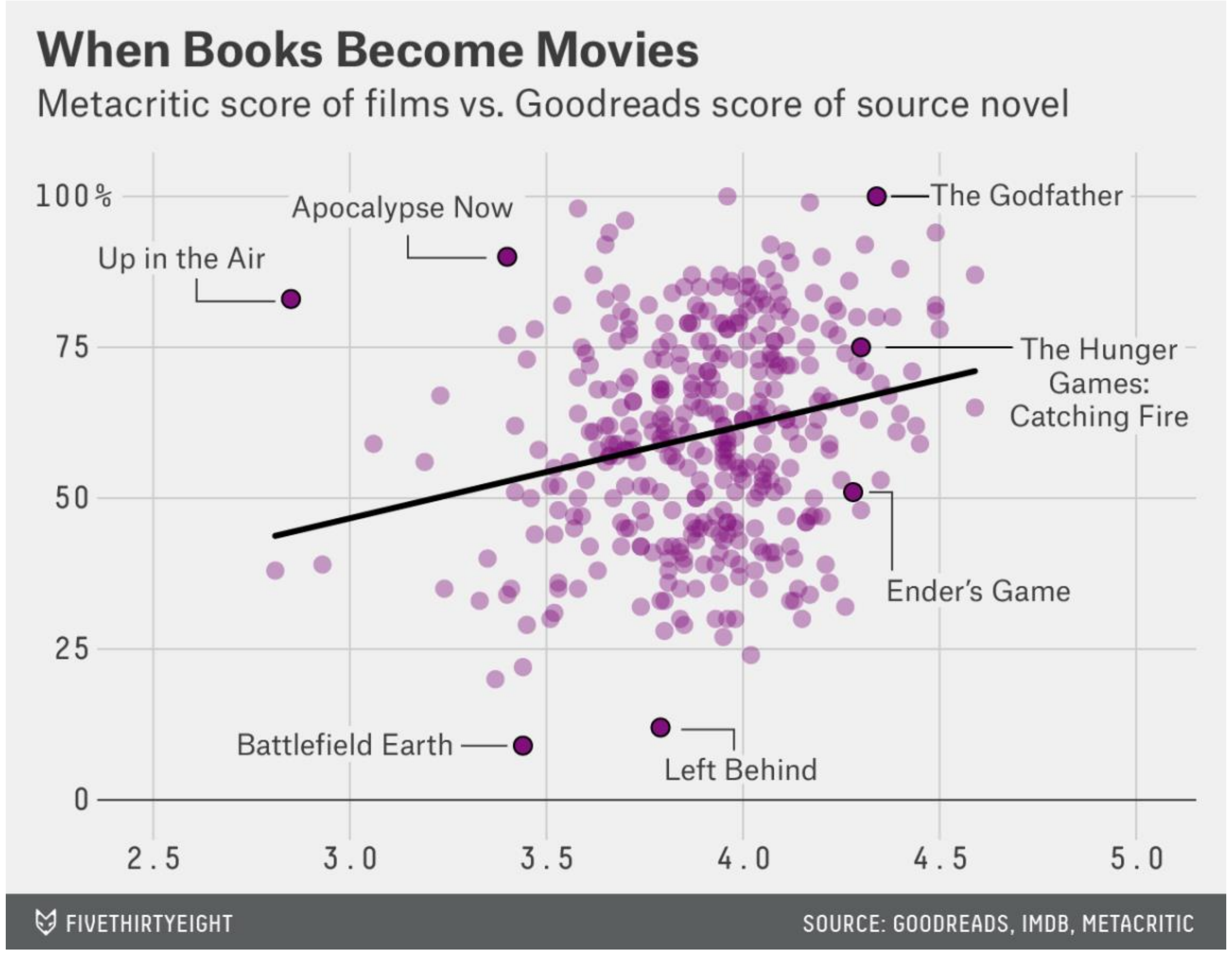
- Storytelling como herramienta para dirigir el interés y conectar con la audiencia
- Modelo para comunicación en general / Estructura y técnicas
- Aplicables a una narración usando gráficas (y otros elementos)



¿Hace falta siempre contar una historia con datos?

¿Hace falta contar siempre una historia?

¿Son/deberían ser todas las visualizaciones una historia?

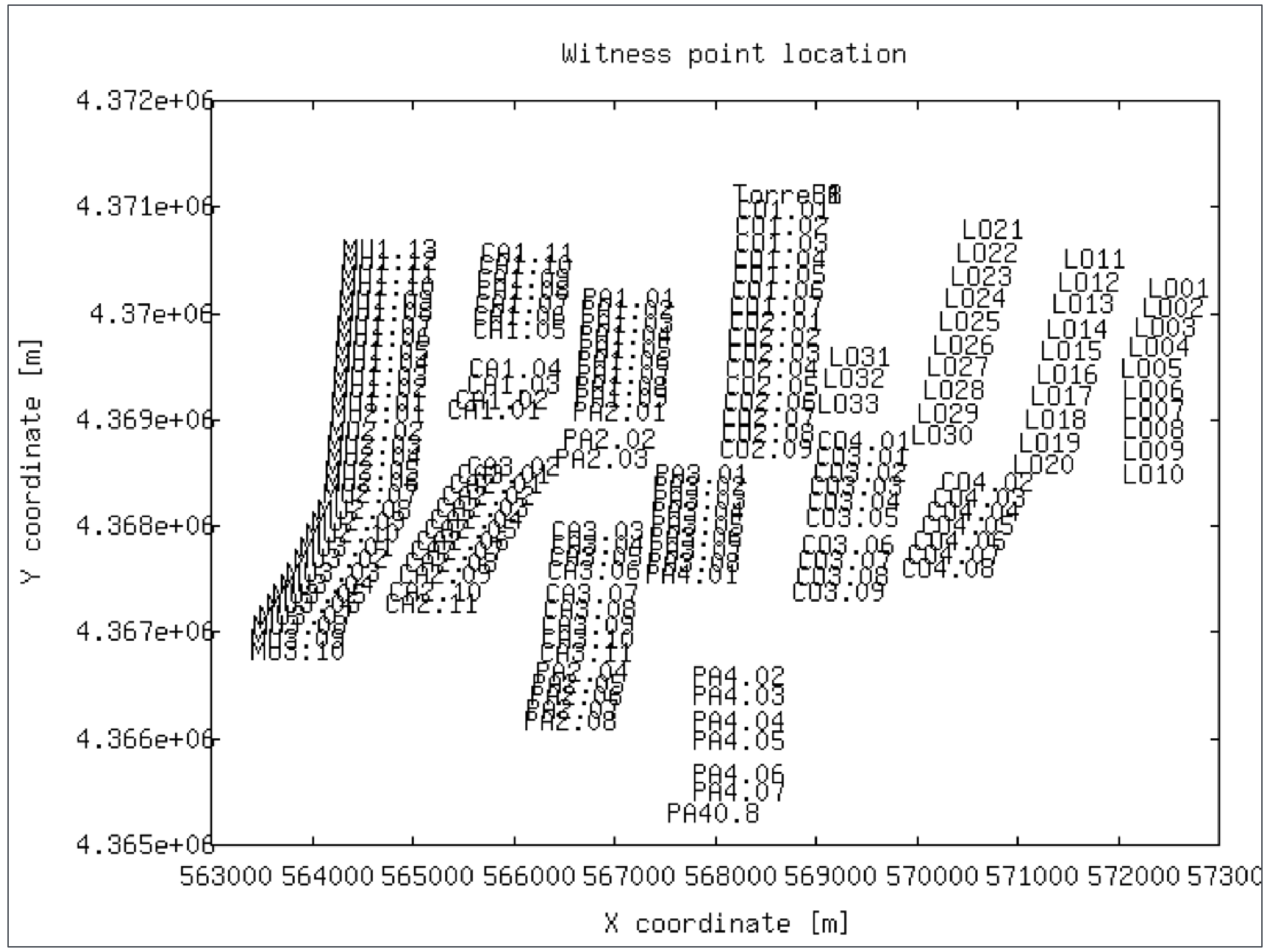


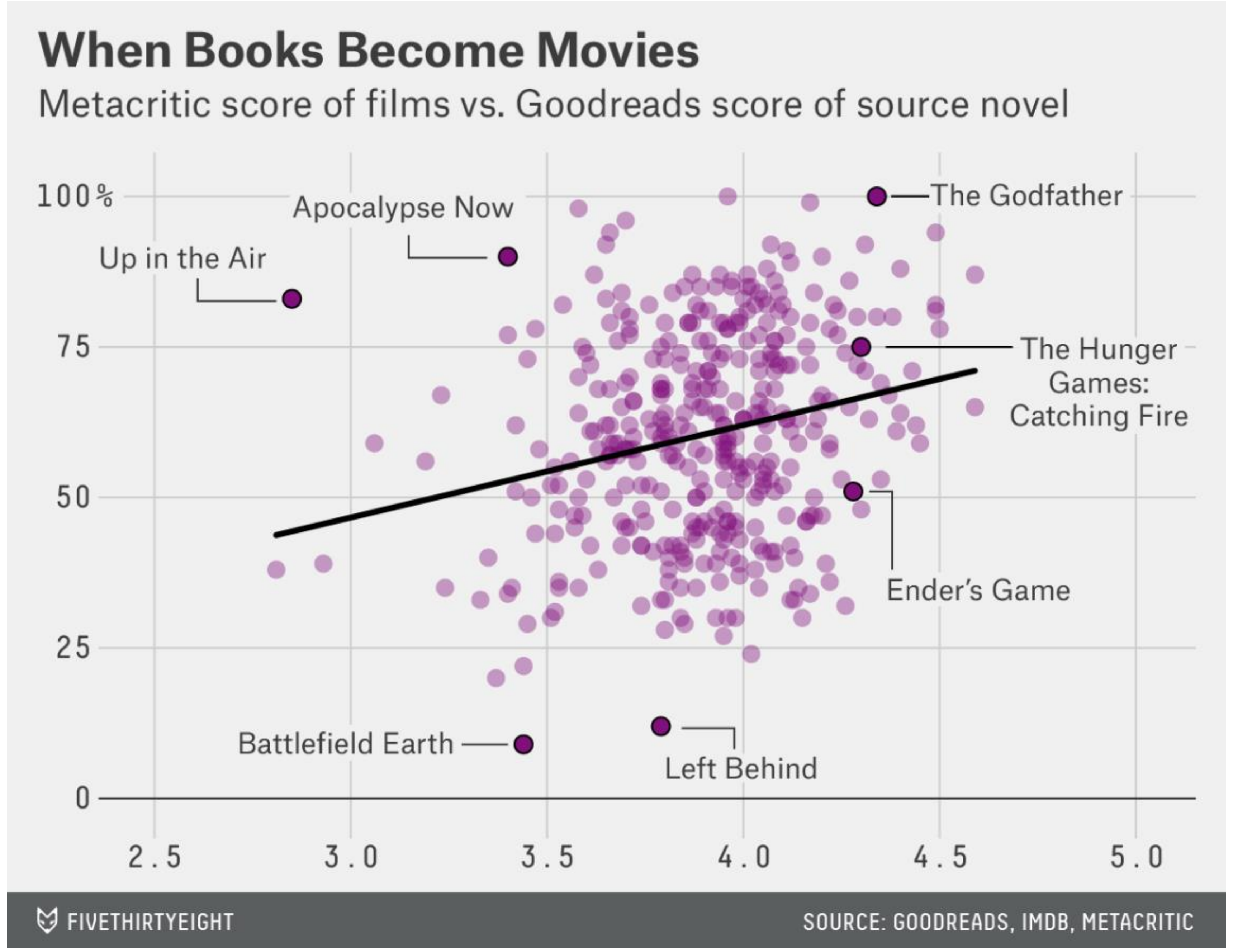
No tiene que ser una historia necesariamente

- Hay varias estrategias para comunicar con datos:
  - **Anotación**. Texto y marcas que ayuden a interpretarla
  - **Narración**. Anotaciones conectadas en una secuencia lógica
  - **Historia**. Una narración que contenga drama/conflicto para obtener una respuesta emocional

¿Cuándo hace falta contar una historia?

- La audiencia no tiene contexto previo sobre los datos/visualización: No puede “leer” la gráfica sin una guía
- Tendemos naturalmente a encontrar patrones, aunque no estén ahí. Una gráfica sin anotaciones, sin guía, puede llevar potencialmente a la audiencia a conclusiones erróneas.
- **Mostrar datos en una estructura narrativa es una forma de guiar al usuario**
- La audiencia puede tener o no un interés previo en el tema. Cuando no lo tiene, podemos crearlo utilizando técnicas de **storytelling**





Tres tipos de estructura narrativa aplicables a data viz

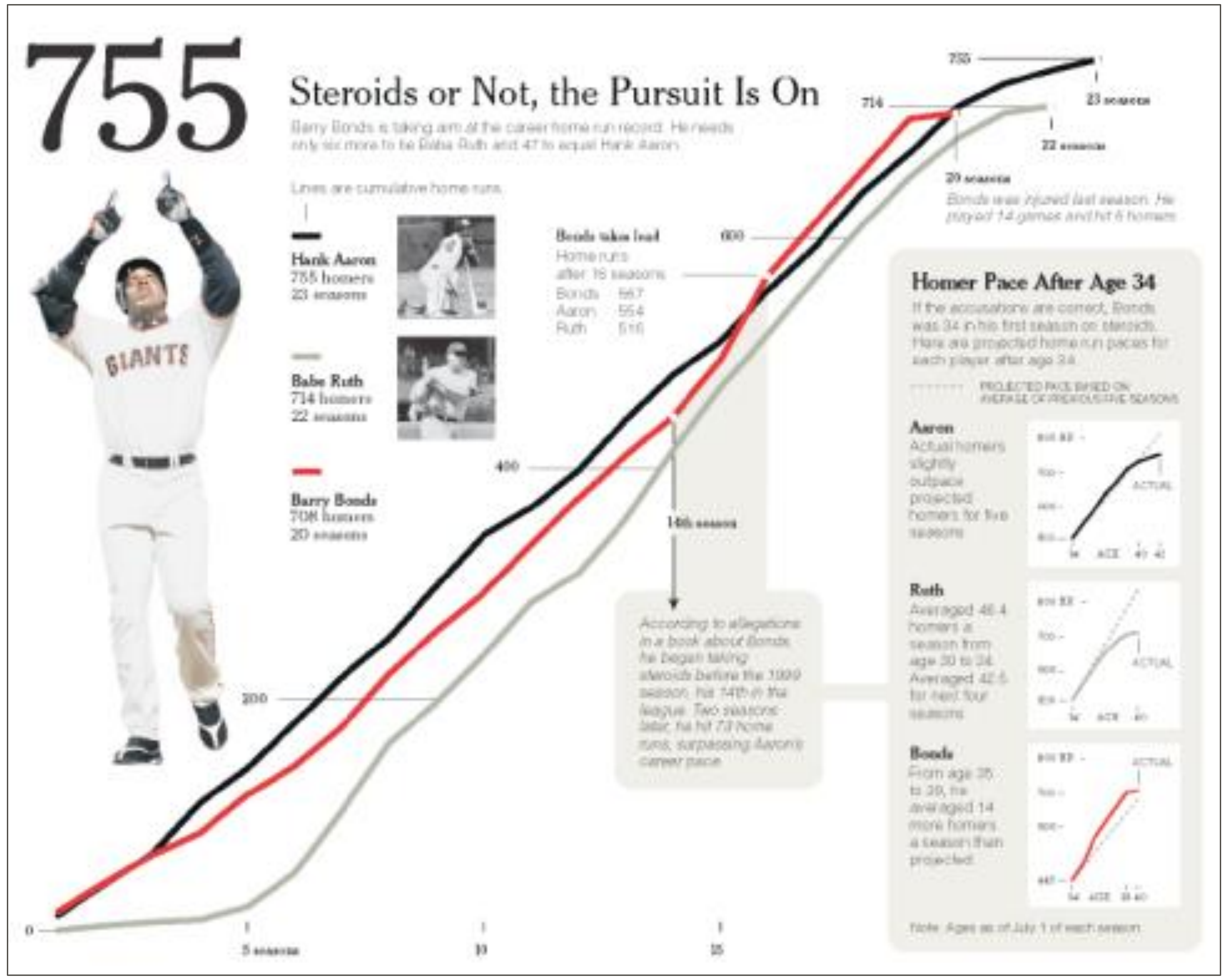
No tiene que ser una historia necesariamente

- Hay varias estrategias para ayudar al usuario a navegar por una visualización:
  - **Anotación**. Texto y marcas que ayuden a interpretarla
  - **Narración**. Anotaciones conectadas en una secuencia lógica
    - **Cerrada.** Estructura secuencial; mensaje definido
    - **Abierta.** Visualización exploratoria. Visualización como herramienta de búsqueda de narrativas
  - **Historia.** Una narración con personajes, factor humano, drama/conflicto pa obtener una respuesta emocional

Tres tipos de estructura narrativa aplicables a data viz

No tiene que ser una historia necesariamente

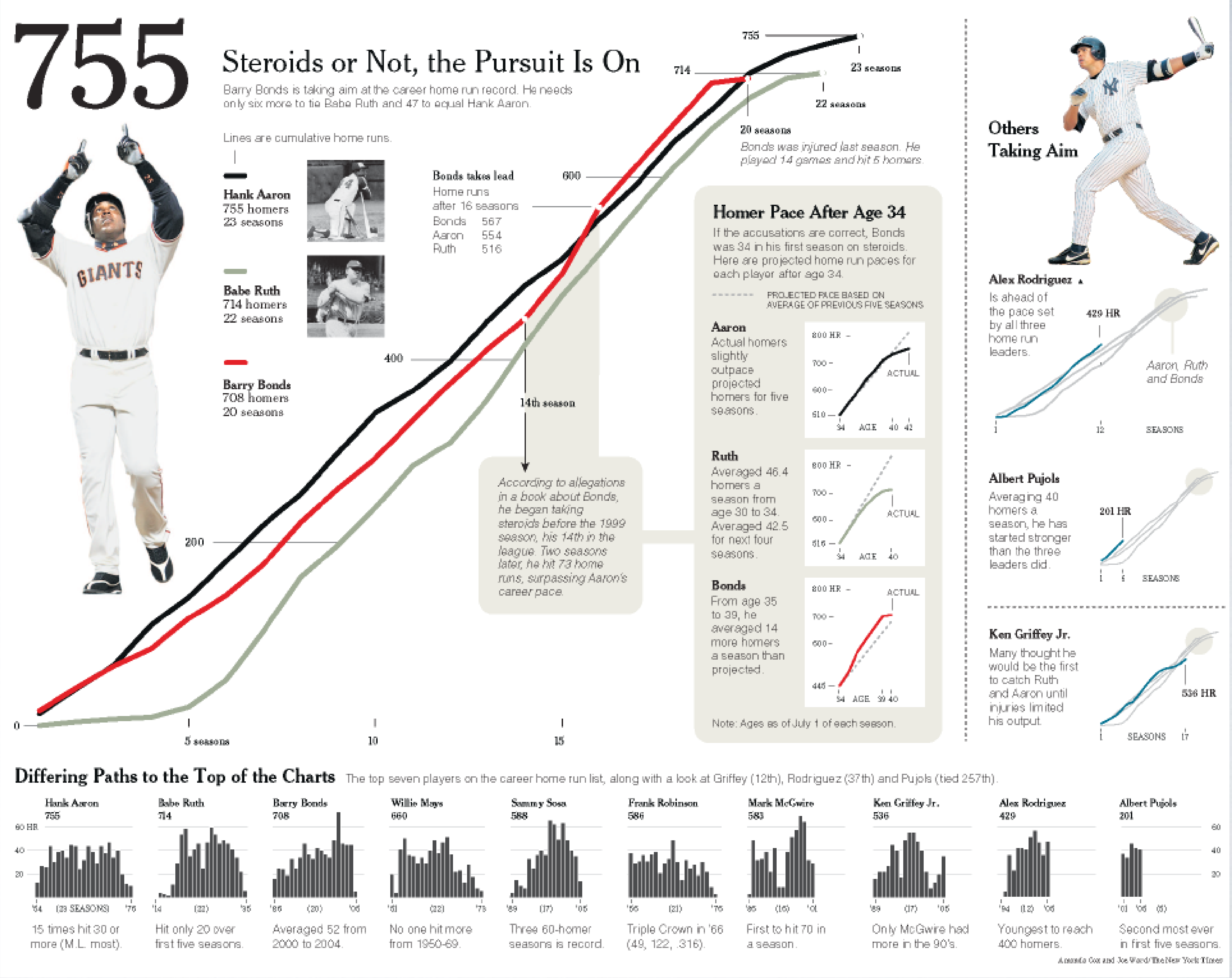
- Hay varias estrategias para ayudar al usuario a navegar por una visualización:
  - **Anotación**. Texto y marcas que ayuden a interpretarla
  - **Narración**. Anotaciones conectadas en una secuencia lógica
    - **Cerrada.** Estructura secuencial; mensaje definido
    - **Abierta.** Visualización exploratoria. Visualización como herramienta de búsqueda de narrativas
  - **Historia**. Una narración con personajes, factor humano, drama/conflicto para obtener una respuesta emocional

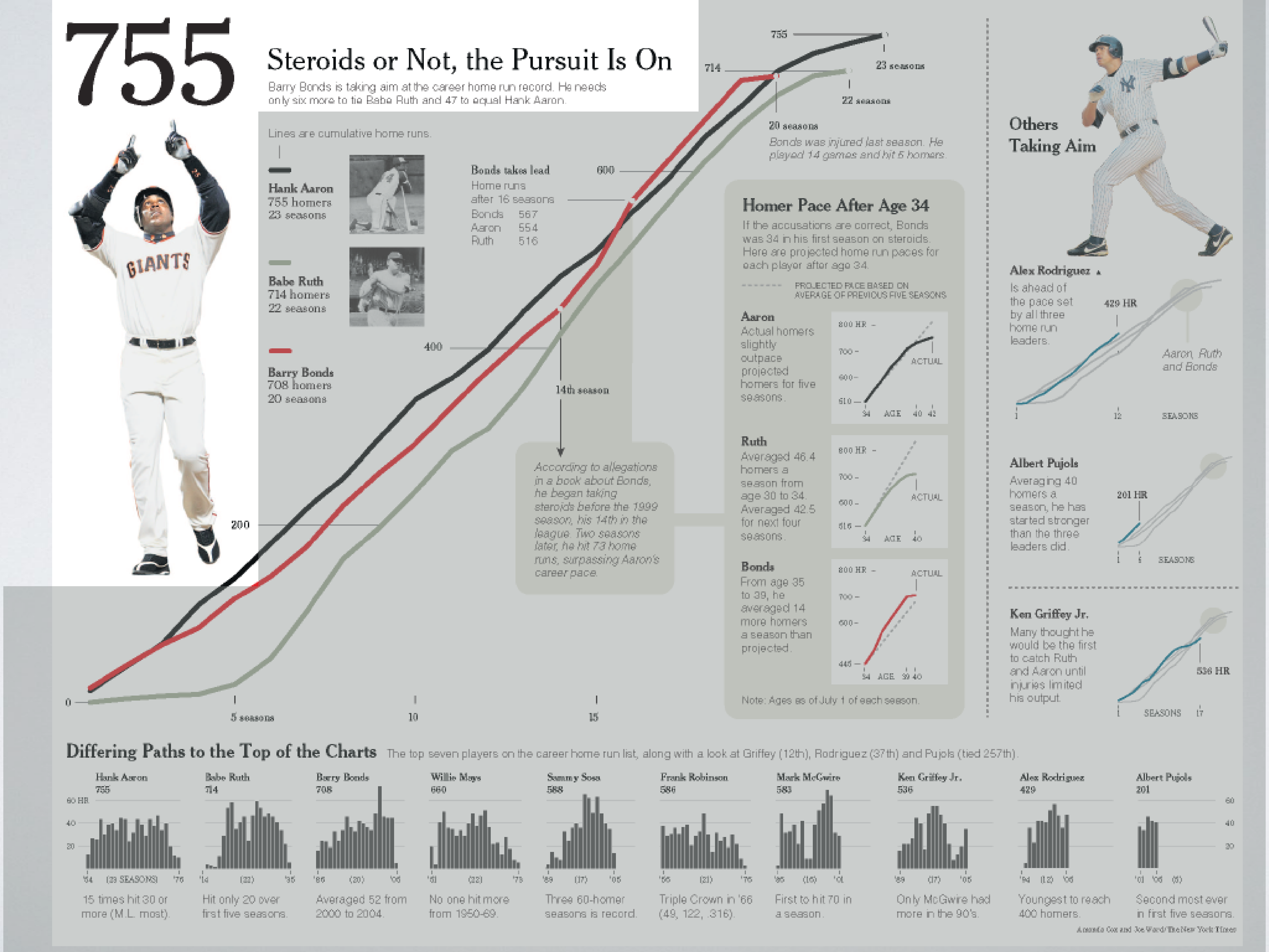


The New York Times

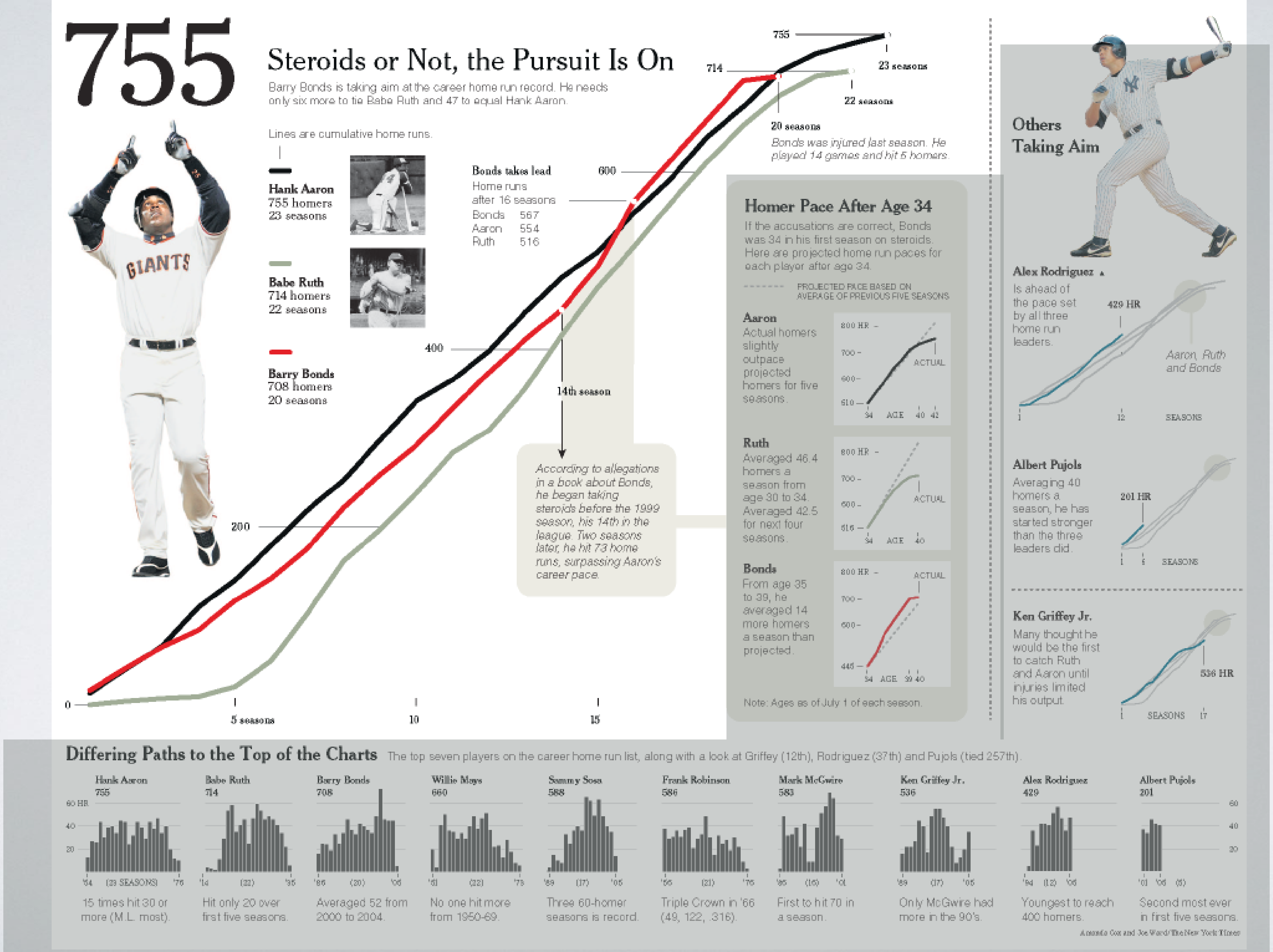
Storytelling

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE transactions on visualization and computer graphics, 16(6), 1139-1148.

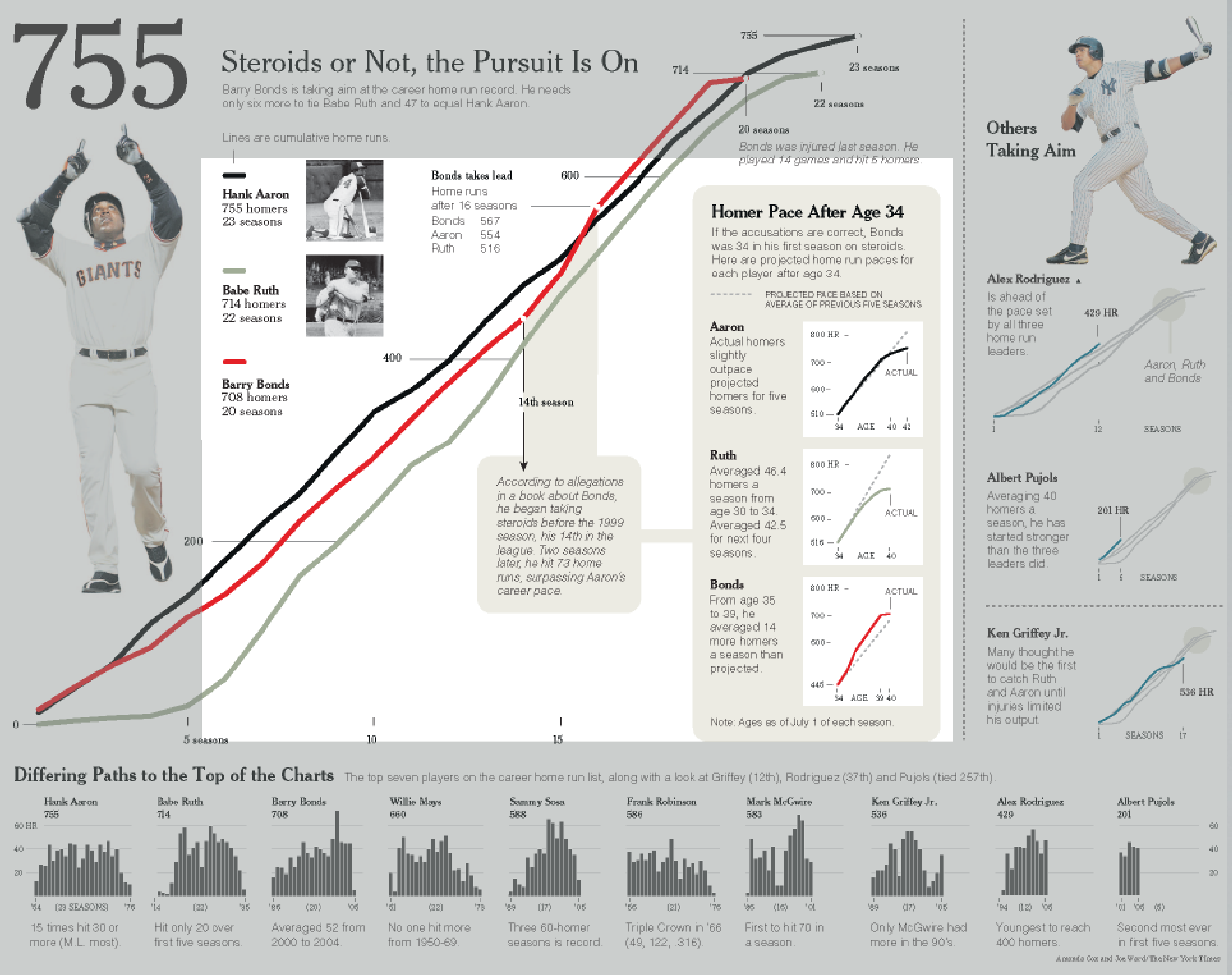




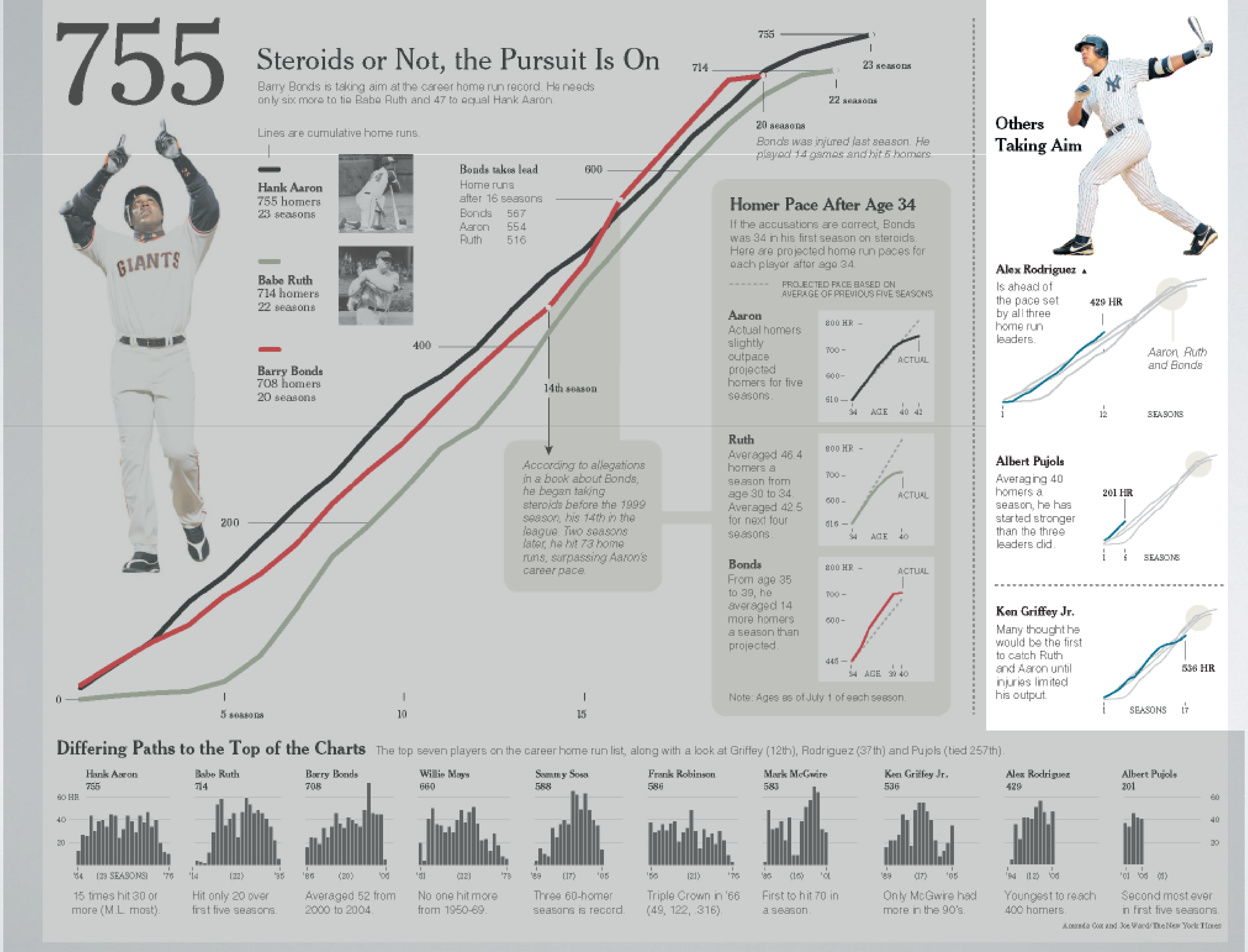
755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
GIANTS
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Bonds takes lead
Home runs after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
5 seasons
10
15
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
800 HR
700
600
510
ACTUAL
34 AGE 40 42
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
800 HR
700
600
516
ACTUAL
34 AGE 40
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
445
ACTUAL
34 AGE 39 40
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
1
12
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
1
6
SEASONS
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
1
SEASONS
17
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
60 HR
40
20
Hank Aaron
755
'54 (23 SEASONS) '76
15 times hit 30 or more (M.L. most).
Babe Ruth
714
'14 (22) '35
Hit only 20 over first five seasons.
Barry Bonds
708
'86 (20) '06
Averaged 52 from 2000 to 2004.
Willie Mays
660
'51 (22) '73
No one hit more from 1950-69.
Sammy Sosa
588
'89 (17) '05
Three 60-homer seasons is record.
Frank Robinson
586
'56 (21) '76
Triple Crown in '66 (49, 122, .316).
Mark McGwire
583
'86 (16) '01
First to hit 70 in a season.
Ken Griffey Jr.
536
'89 (17) '05
Only McGwire had more in the 90's.
Alex Rodriguez
429
'94 (12) '05
Youngest to reach 400 homers.
Albert Pujols
201
'01 '06 (6)
Second most ever in first five seasons.
60
40
20
Amanda Cox and Joe Ward/The New York Times

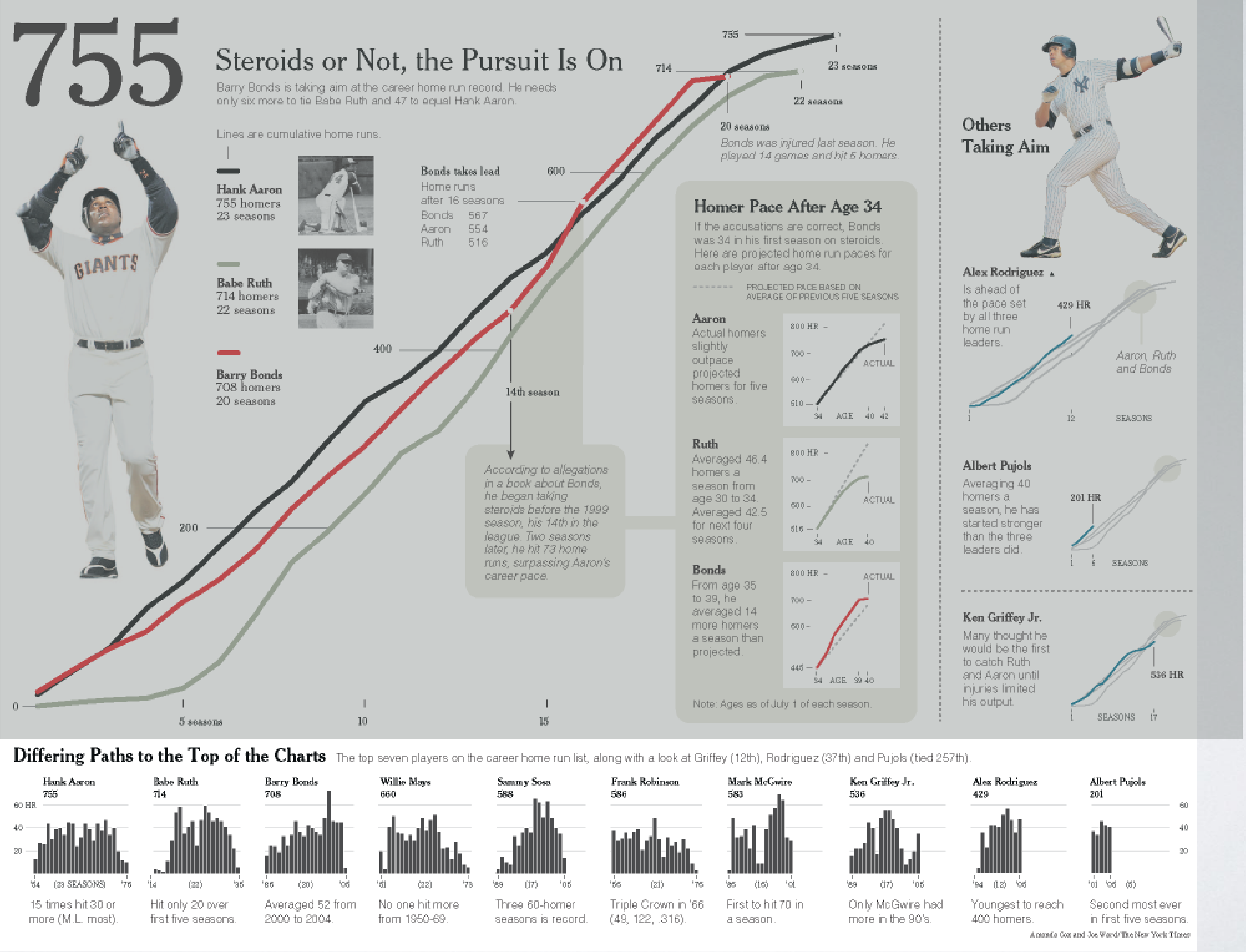


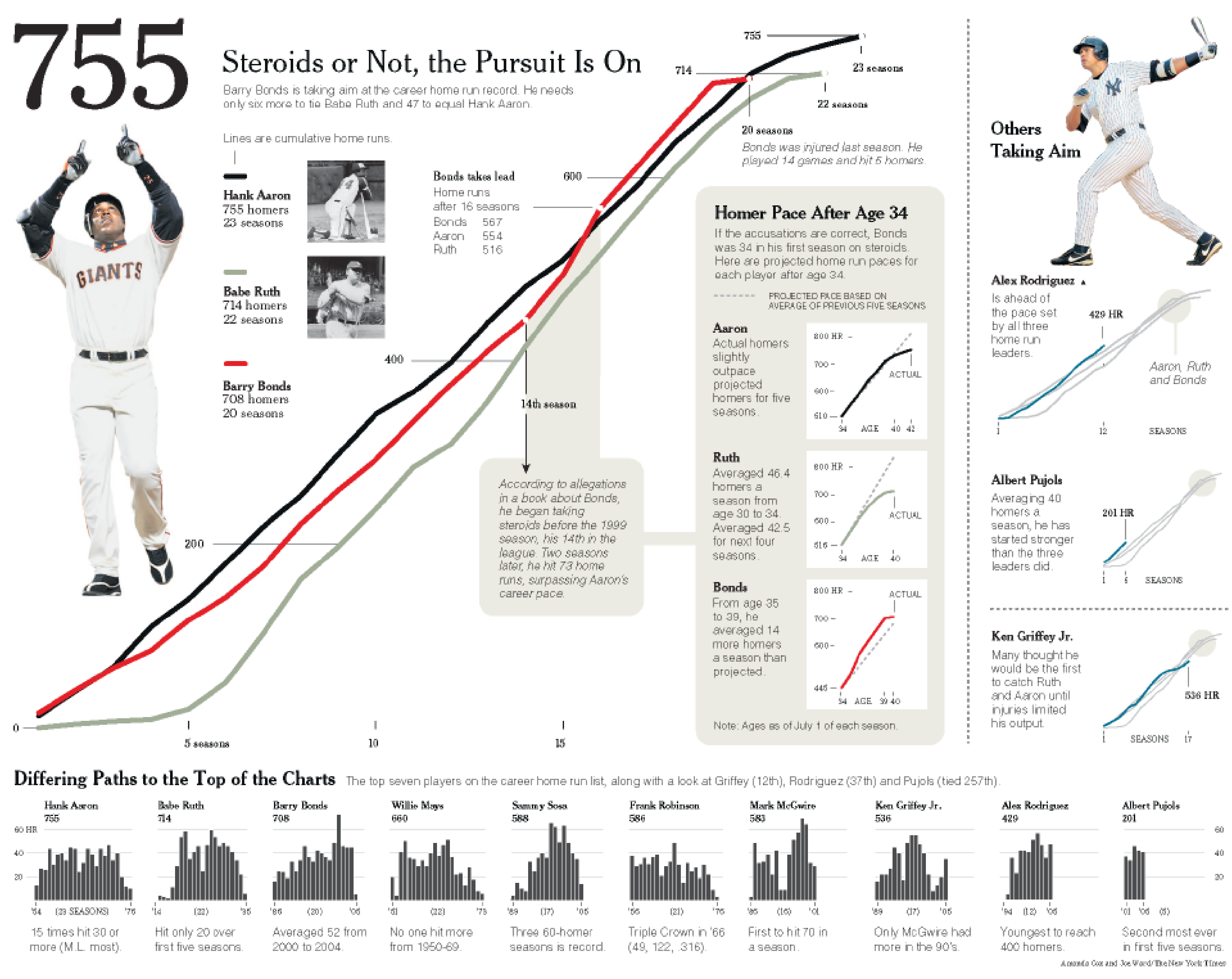
755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Bonds takes lead
Home runs after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
5 seasons
10
15
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
800 HR
700
600
510
ACTUAL
34 AGE 40 42
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
800 HR
700
600
516
ACTUAL
34 AGE 40
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
445
ACTUAL
34 AGE 39 40
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
1
12
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
1
5
SEASONS
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
SEASONS
17
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
60 HR
40
20
Hank Aaron
755
'54 (23 SEASONS) '76
15 times hit 30 or more (M.L. most).
Babe Ruth
714
'14 (22) '35
Hit only 20 over first five seasons.
Barry Bonds
708
'86 (20) '05
Averaged 52 from 2000 to 2004.
Willie Mays
660
'51 (22) '73
No one hit more from 1950-69.
Sammy Sosa
588
'89 (17) '05
Three 60-homer seasons is record.
Frank Robinson
586
'56 (21) '76
Triple Crown in '66 (49, 122, .316).
Mark McGwire
583
'86 (16) '01
First to hit 70 in a season.
Ken Griffey Jr.
536
'89 (17) '05
Only McGwire had more in the 90's.
Alex Rodriguez
429
'94 (12) '05
Youngest to reach 400 homers.
Albert Pujols
201
'01 '05 (5)
Second most ever in first five seasons.
60
40
20
Amanda Cox and Joe Ward/The New York Times



755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Bonds takes lead
Home runs after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
5 seasons
10
15
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
ACTUAL
AGE
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
Hank Aaron 755
Babe Ruth 714
Barry Bonds 708
Willie Mays 660
Sammy Sosa 588
Frank Robinson 586
Mark McGwire 583
Ken Griffey Jr. 536
Alex Rodriguez 429
Albert Pujols 201
15 times hit 30 or more (M.L. most).
Hit only 20 over first five seasons.
Averaged 52 from 2000 to 2004.
No one hit more from 1950-69.
Three 60-homer seasons is record.
Triple Crown in '66 (49, 122, .316).
First to hit 70 in a season.
Only McGwire had more in the 90's.
Youngest to reach 400 homers.
Second most ever in first five seasons.
Amanda Cox and Joe Ward/The New York Times


755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Bonds takes lead
Home runs after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
5 seasons
10
15
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
ACTUAL
AGE
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
Hank Aaron 755
'54 (23 SEASONS) '76
15 times hit 30 or more (M.L. most).
Babe Ruth 714
'14 (22) '35
Hit only 20 over first five seasons.
Barry Bonds 708
'86 (20) '06
Averaged 52 from 2000 to 2004.
Willie Mays 660
'51 (22) '73
No one hit more from 1950-69.
Sammy Sosa 588
'89 (17) '05
Three 60-homer seasons is record.
Frank Robinson 586
'56 (21) '76
Triple Crown in '66 (49, 122, .316).
Mark McGwire 583
'86 (16) '01
First to hit 70 in a season.
Ken Griffey Jr. 536
'89 (17) '05
Only McGwire had more in the 90's.
Alex Rodriguez 429
'94 (12) '05
Youngest to reach 400 homers.
Albert Pujols 201
'01 '05 (5)
Second most ever in first five seasons.
60 HR
40
20
Amanda Cox and Joe Ward/The New York Times


755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
GIANTS
Bonds takes lead
Home runs
after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
5 seasons
10
15
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
800 HR
700
600
510
ACTUAL
34
AGE
40
42
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
800 HR
700
600
516
ACTUAL
34
AGE
40
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
445
ACTUAL
34
AGE
39
40
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
1
12
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
1
5
SEASONS
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
1
SEASONS
17
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
Hank Aaron
755
60 HR
40
20
'54
(23 SEASONS)
'76
15 times hit 30 or more (M.L. most).
Babe Ruth
714
'14
(22)
'35
Hit only 20 over first five seasons.
Barry Bonds
708
'86
(20)
'06
Averaged 52 from 2000 to 2004.
Willie Mays
660
'51
(22)
'73
No one hit more from 1950-69.
Sammy Sosa
588
'89
(17)
'05
Three 60-homer seasons is record.
Frank Robinson
586
'56
(21)
'76
Triple Crown in '66 (49, 122, .316).
Mark McGwire
583
'86
(16)
'01
First to hit 70 in a season.
Ken Griffey Jr.
536
'89
(17)
'05
Only McGwire had more in the 90's.
Alex Rodriguez
429
'94
(12)
'05
Youngest to reach 400 homers.
Albert Pujols
201
'01
'05
(5)
Second most ever in first five seasons.
60
40
20
Amanda Cox and Joe Ward/The New York Times

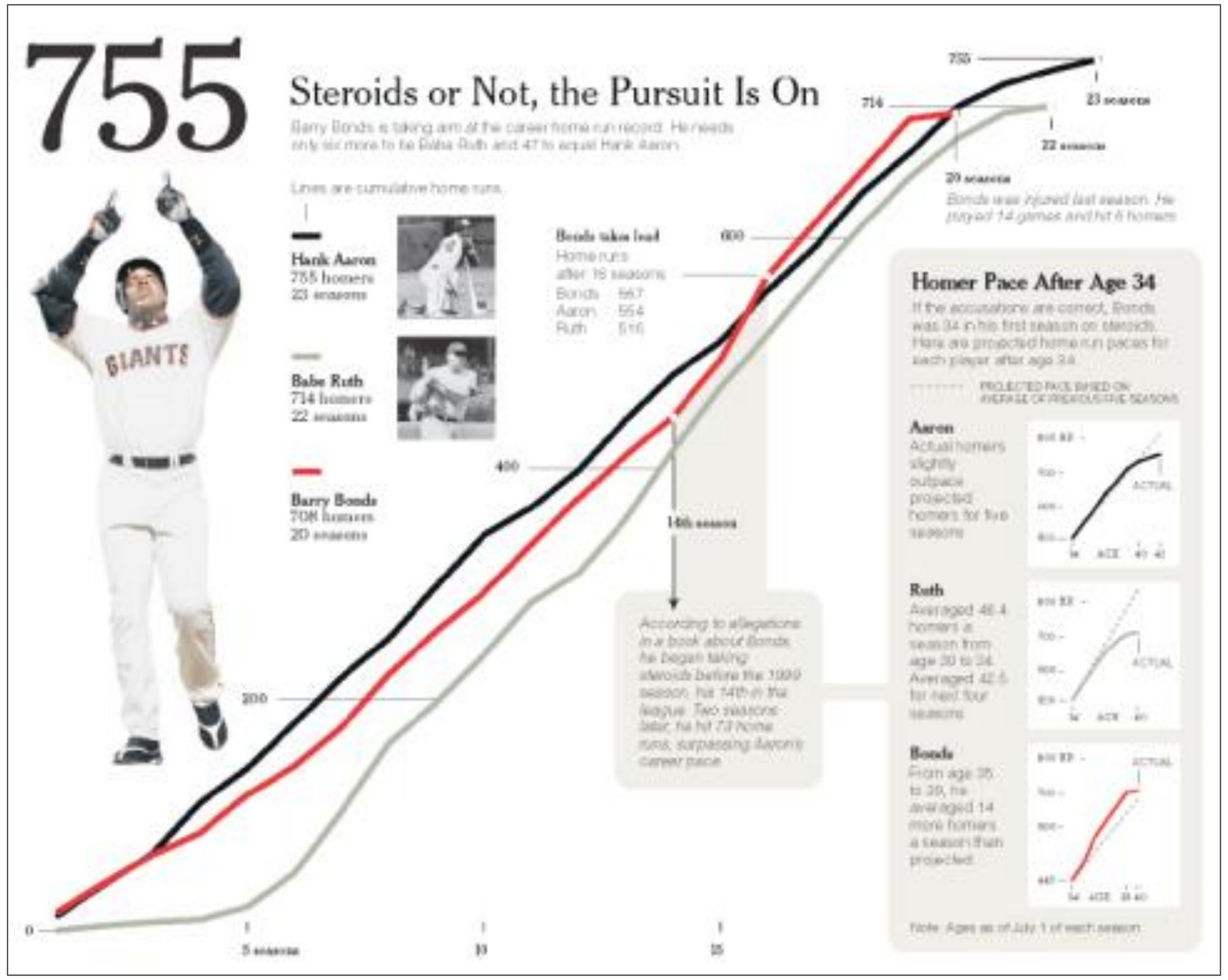


755
Steroids or Not, the Pursuit Is On
Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Lines are cumulative home runs.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Bonds takes lead
Home runs after 16 seasons
Bonds 567
Aaron 554
Ruth 516
755
714
600
400
200
0
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
5 seasons
10
15
Homer Pace After Age 34
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.
PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS
Aaron
Actual homers slightly outpace projected homers for five seasons.
800 HR
700
600
510
ACTUAL
34 AGE 40 42
Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.
800 HR
700
600
516
ACTUAL
34 AGE 40
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.
800 HR
700
600
445
ACTUAL
34 AGE 39 40
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
429 HR
Aaron, Ruth and Bonds
1
12
SEASONS
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
201 HR
1 5 SEASONS
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
1 SEASONS 17
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
Hank Aaron
755
60 HR
40
20
'54 (23 SEASONS) '76
15 times hit 30 or more (M.L. most).
Babe Ruth
714
'14 (22) '35
Hit only 20 over first five seasons.
Barry Bonds
708
'86 (20) '06
Averaged 52 from 2000 to 2004.
Willie Mays
660
'51 (22) '73
No one hit more from 1950-69.
Sammy Sosa
588
'89 (17) '05
Three 60-homer seasons is record.
Frank Robinson
586
'56 (21) '76
Triple Crown in '66 (49, 122, .316).
Mark McGwire
583
'86 (16) '01
First to hit 70 in a season.
Ken Griffey Jr.
536
'89 (17) '05
Only McGwire had more in the 90's.
Alex Rodriguez
429
'94 (12) '05
Youngest to reach 400 homers.
Albert Pujols
201
'01 '05 (5)
Second most ever in first five seasons.
60
40
20
Amanda Cox and Joe Ward/The New York Times

Tres tipos de estructura narrativa aplicables a data viz

No tiene que ser una historia necesariamente

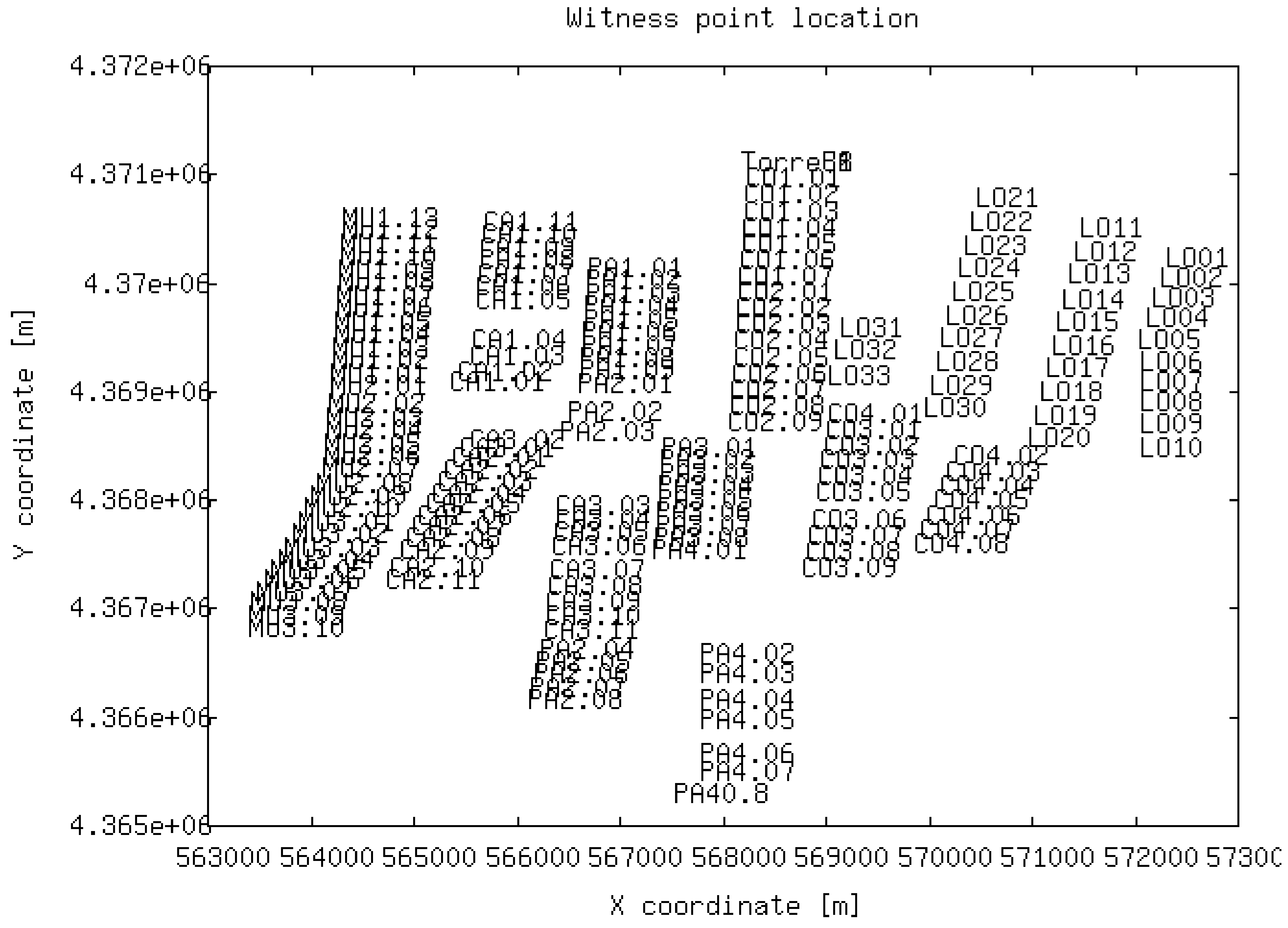
- Hay varias estrategias para ayudar al usuario a navegar por una visualización:
  - **Anotación**. Texto y marcas que ayuden a interpretarla
  - **Narración**. Anotaciones conectadas en una secuencia lógica
    - **Cerrada.** Estructura secuencial; mensaje definido
    - **Abierta.** Visualización exploratoria. Visualización como herramienta de búsqueda de narrativas
  - **Historia**. Una narración con personajes, factor humano, drama/conflicto para obtener una respuesta emocional

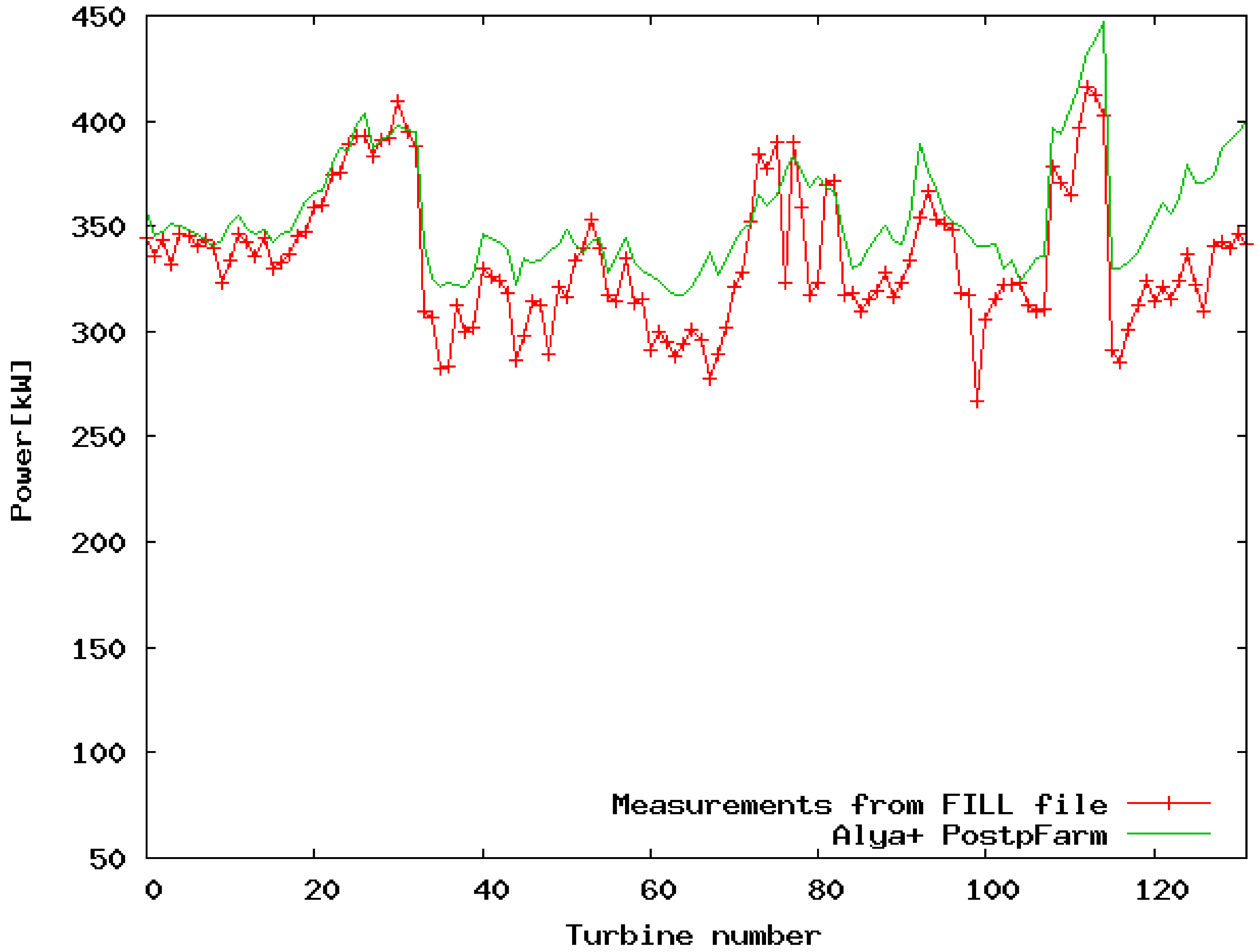


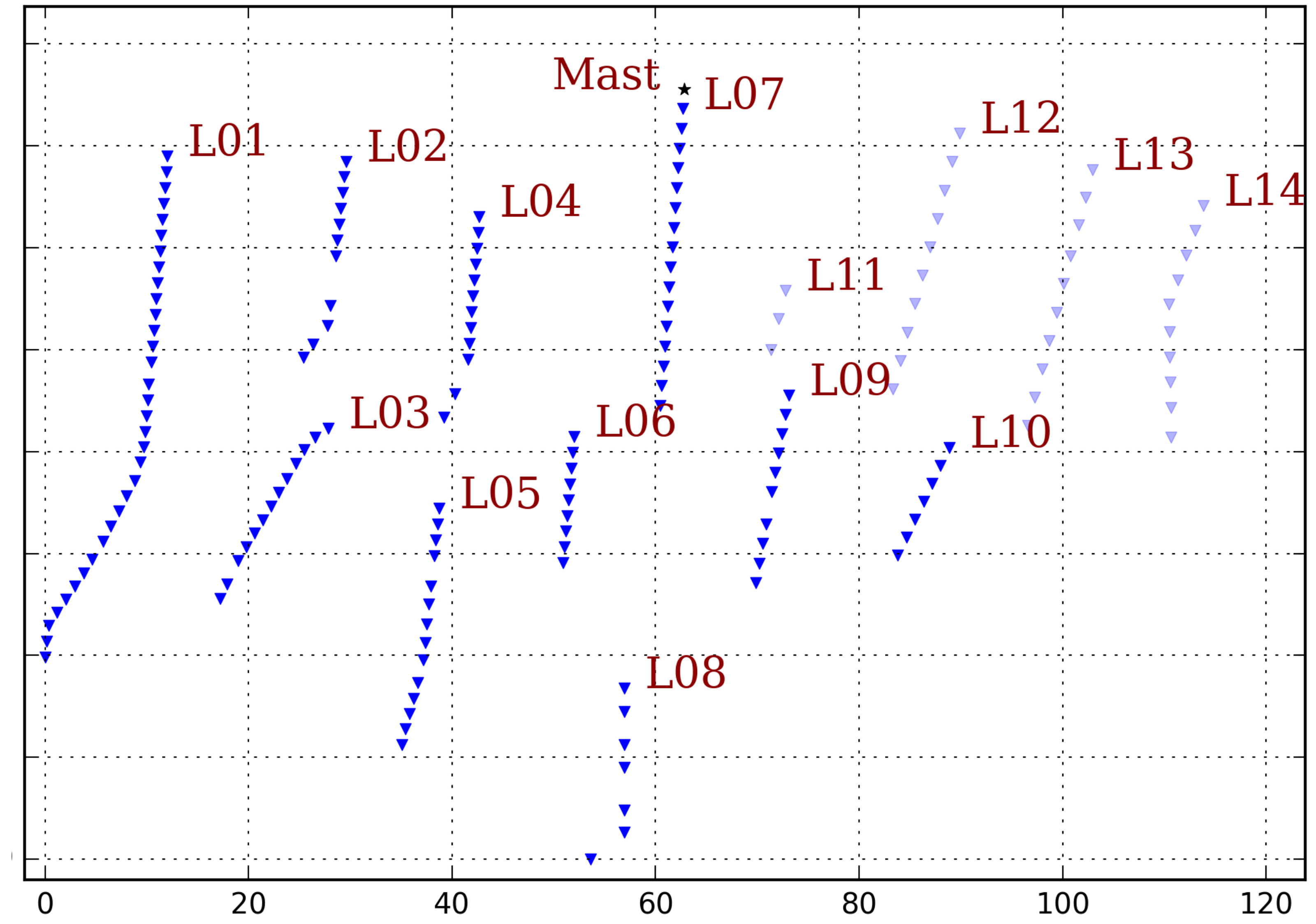
The New York Times

Plots originales

- Simulación del viento para calcular producción de energía en una planta eólica
- Comparación de mediciones reales (rojo) Vs. los resultados de las simulaciones (verde)
- Promedio anual de cada molino (N=132)
- Dataset tiene Potencia (kW), ID, array ID, y localización

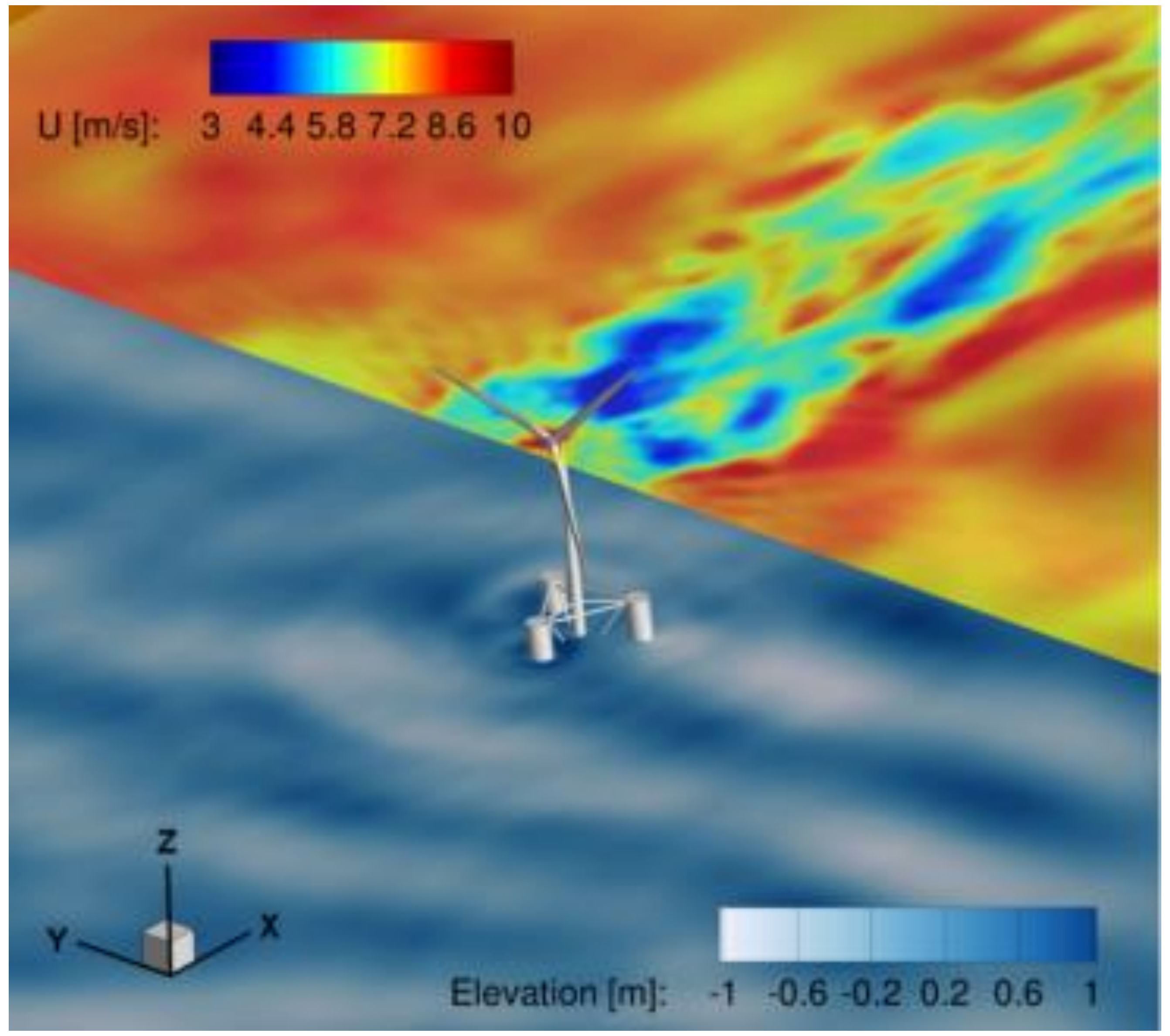




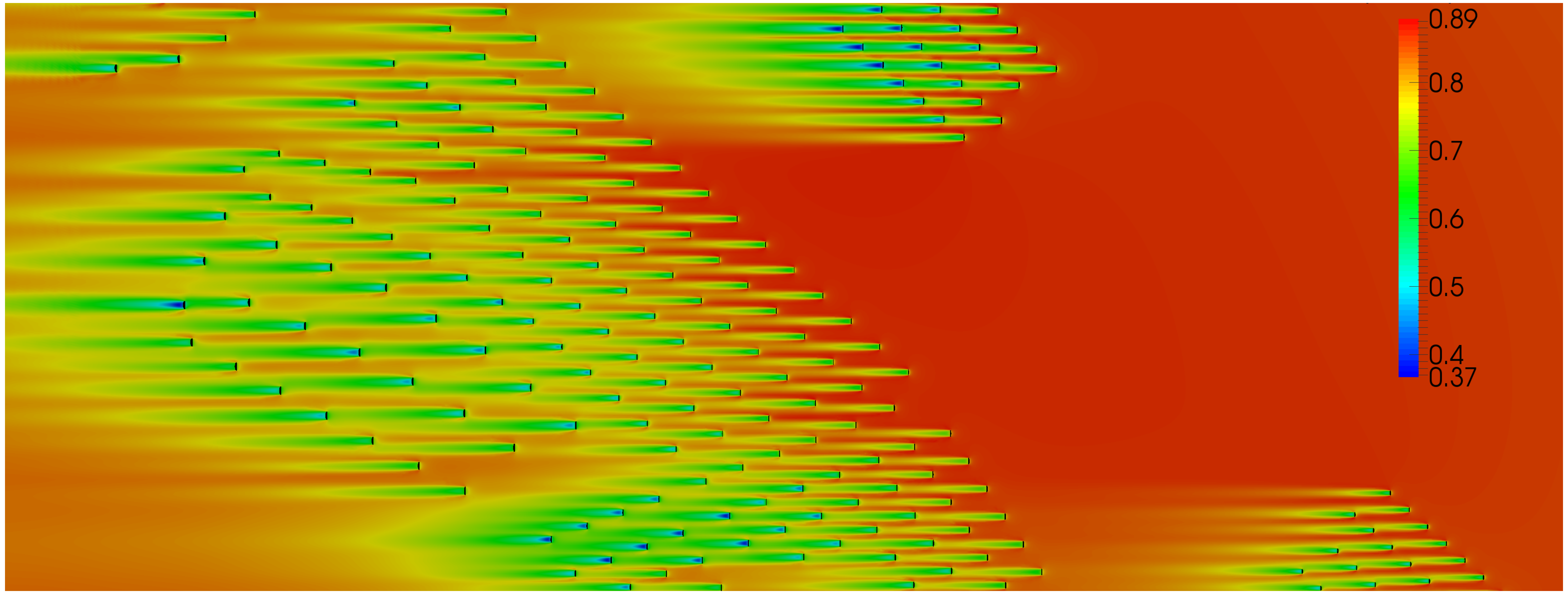


Datos adicionales

- Simulación 3D de la velocidad del viento sobre terreno complejo
- Normalmente graficado como cortes 2D sobre el centro de las aspas

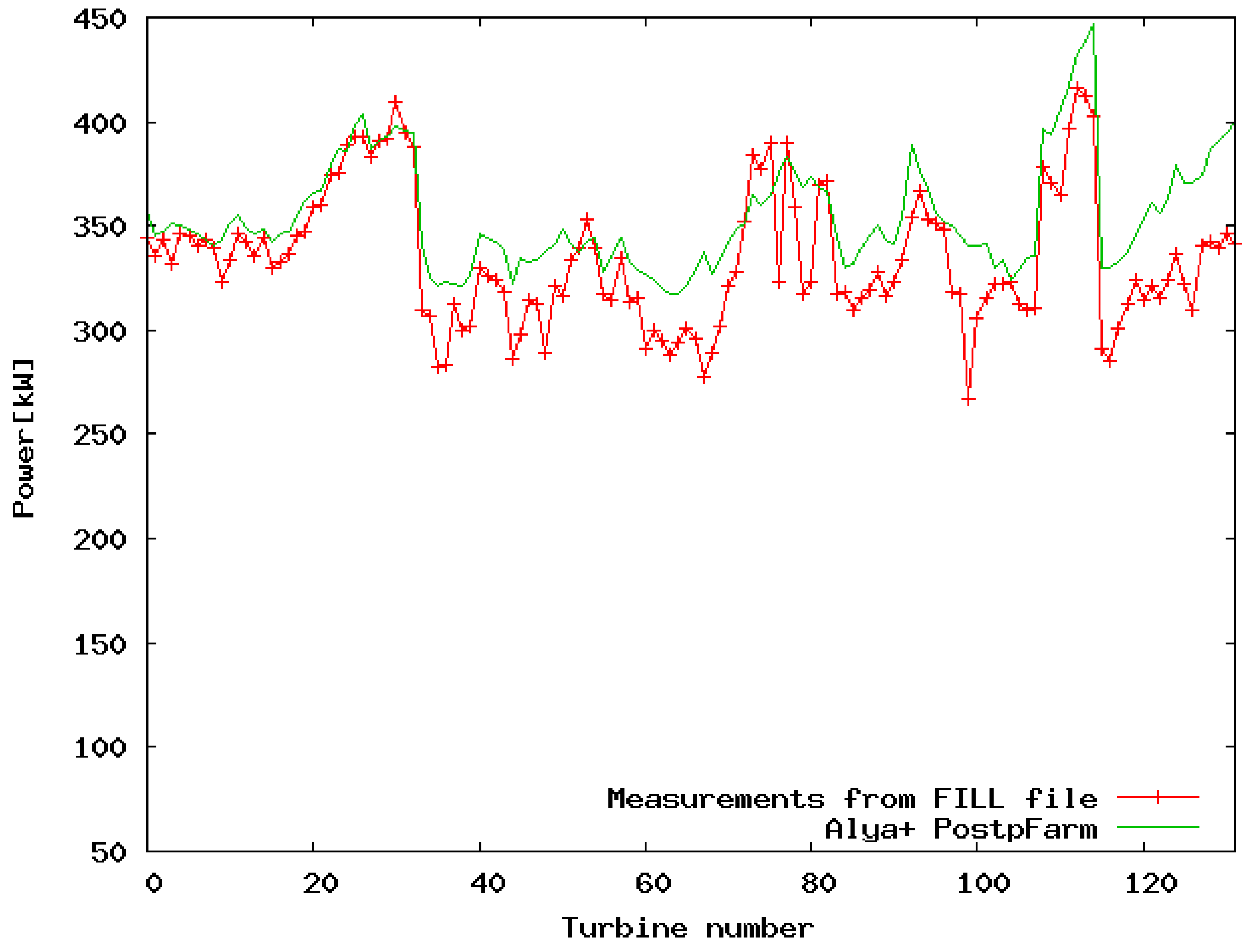


https://www.energy.gov/eere



Barcelona Supercomputing Center

Original plot



Checklist básica

Quién es la audiencia? Project partners en empresa de energía

Cómo se usará? Presentación de resultados/Divulgación

Cual es el objetivo? Comparar resultados vs. mediciones Evaluar la exactitud del modelo

Cual es el desafío? Resaltar las diferencias entre valores y aclarar qué se representa en cada eje

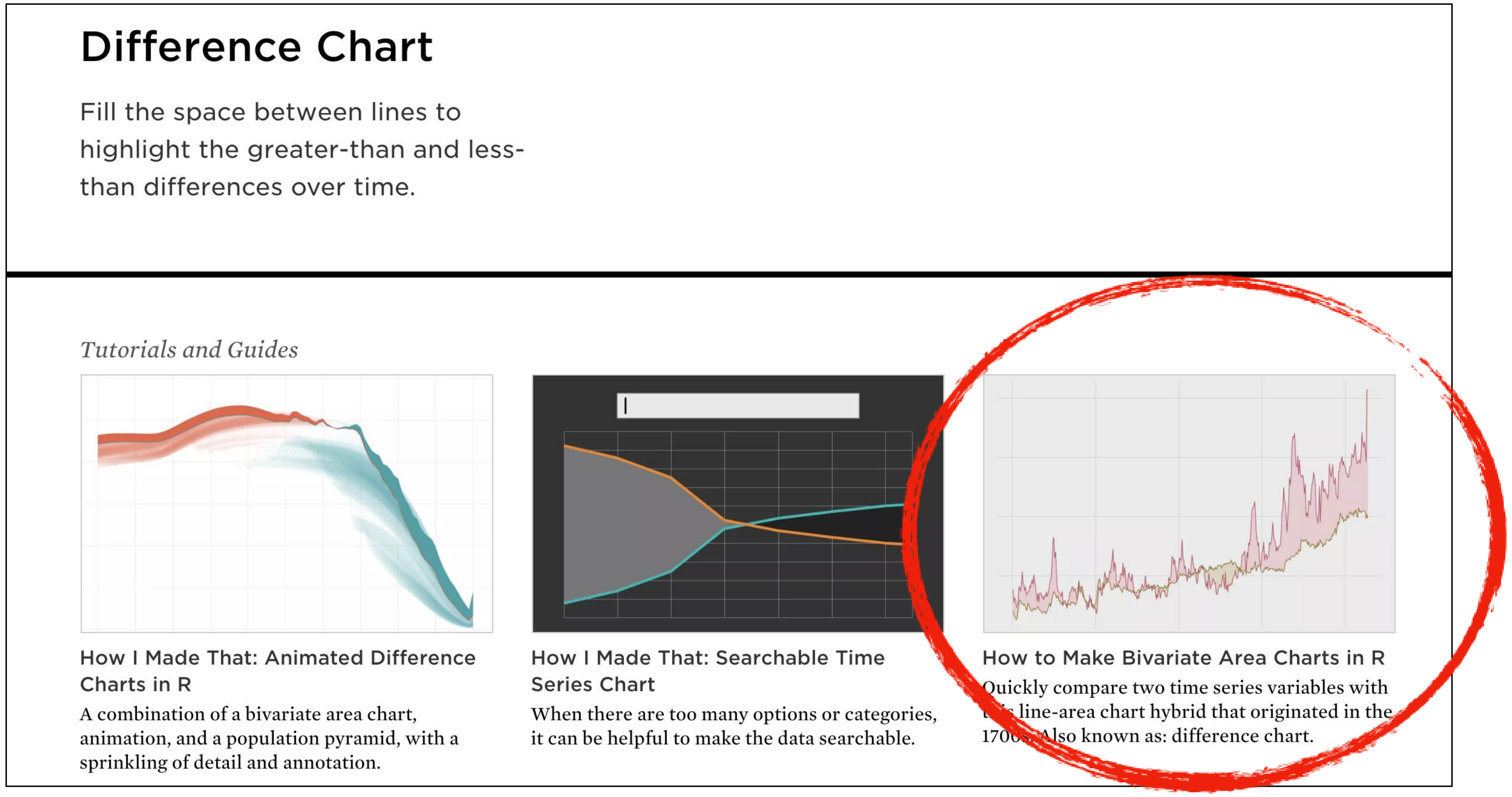
Difference Chart

A.K.A. Bivariate Area Charts

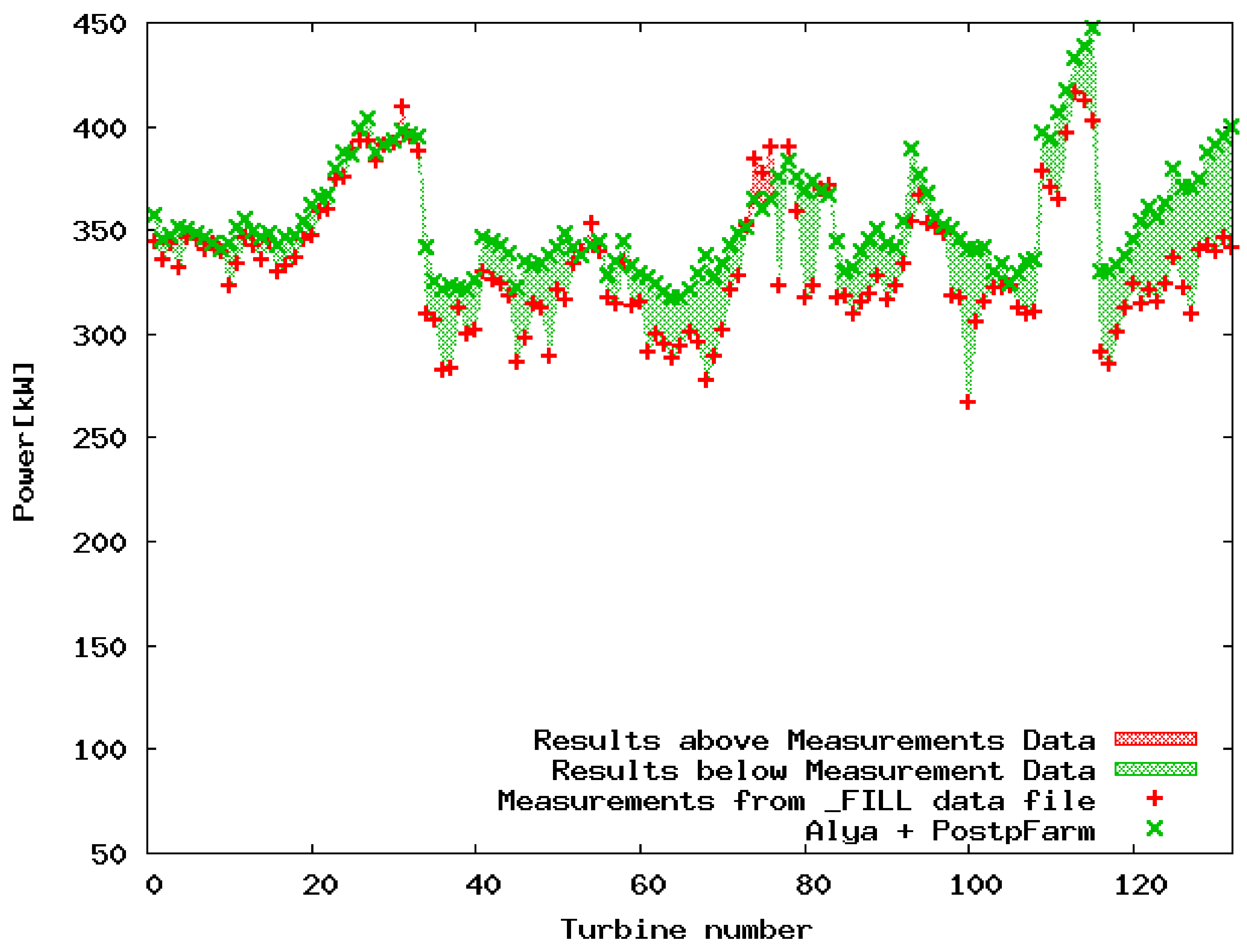
Muestra dos líneas y la diferencia entre ambas

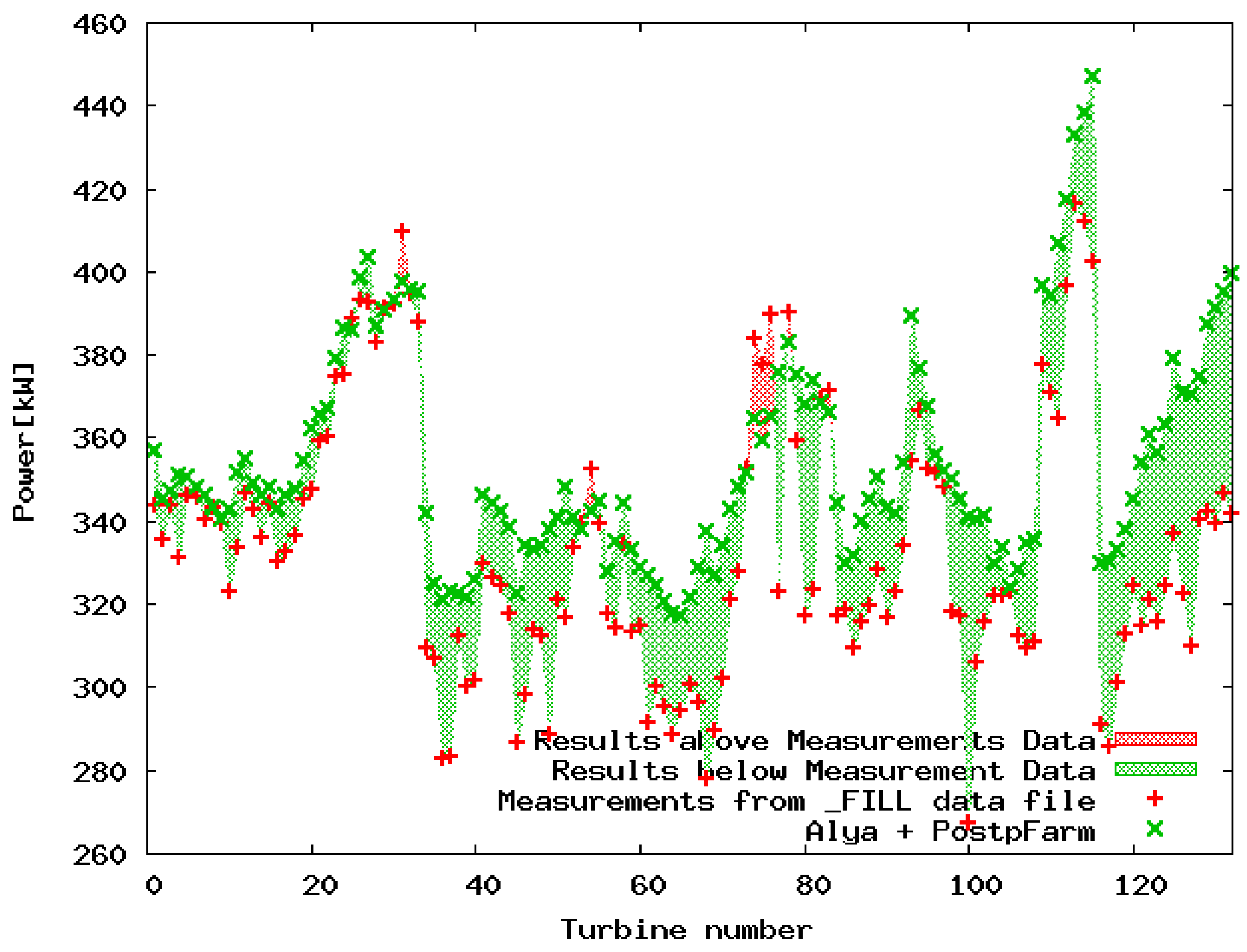
Colores distintos para diferencia positiva y negativa

Generalmente usado para datos continuos, normalmente series temporales



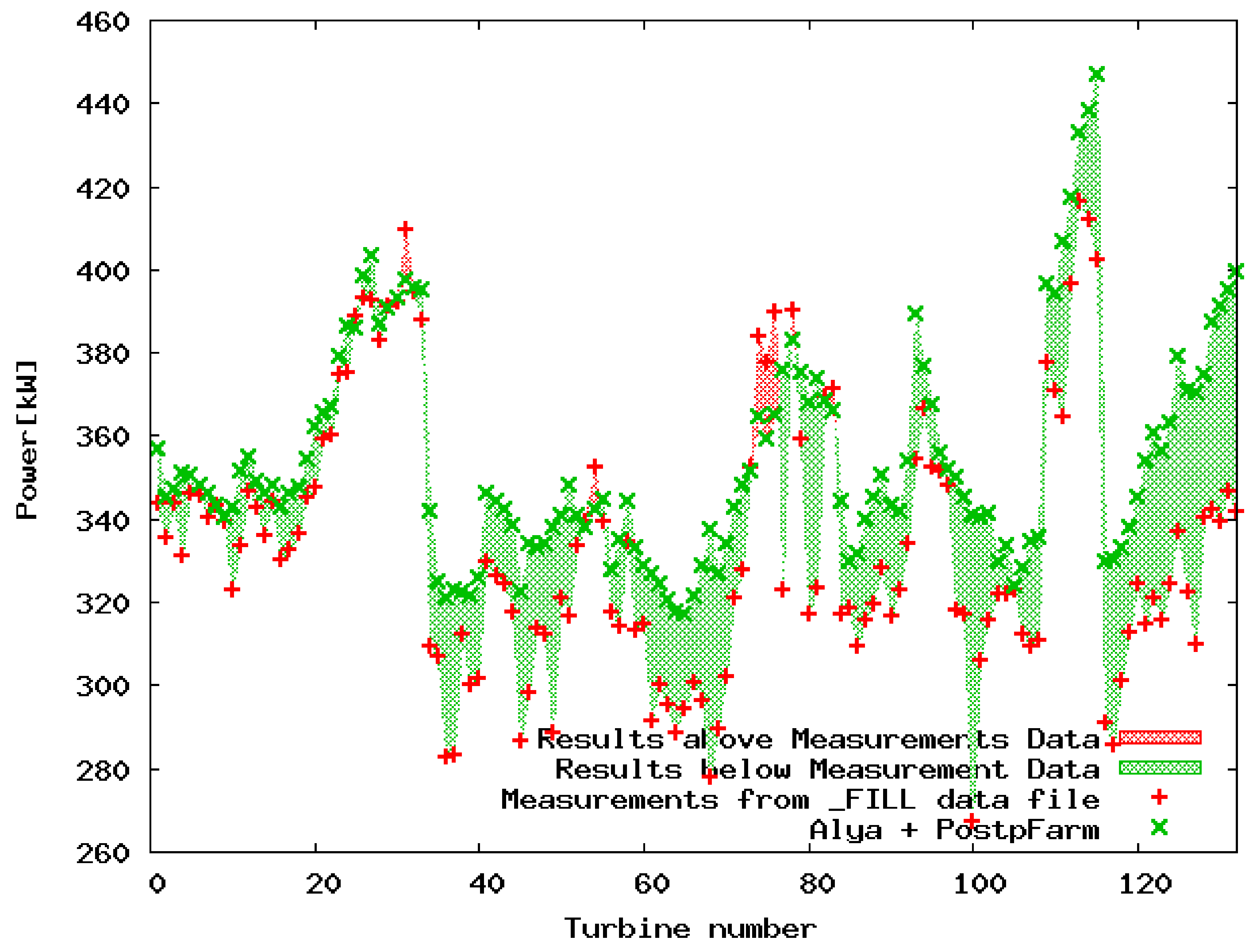
https://flowingdata.com/charttype/difference-chart/

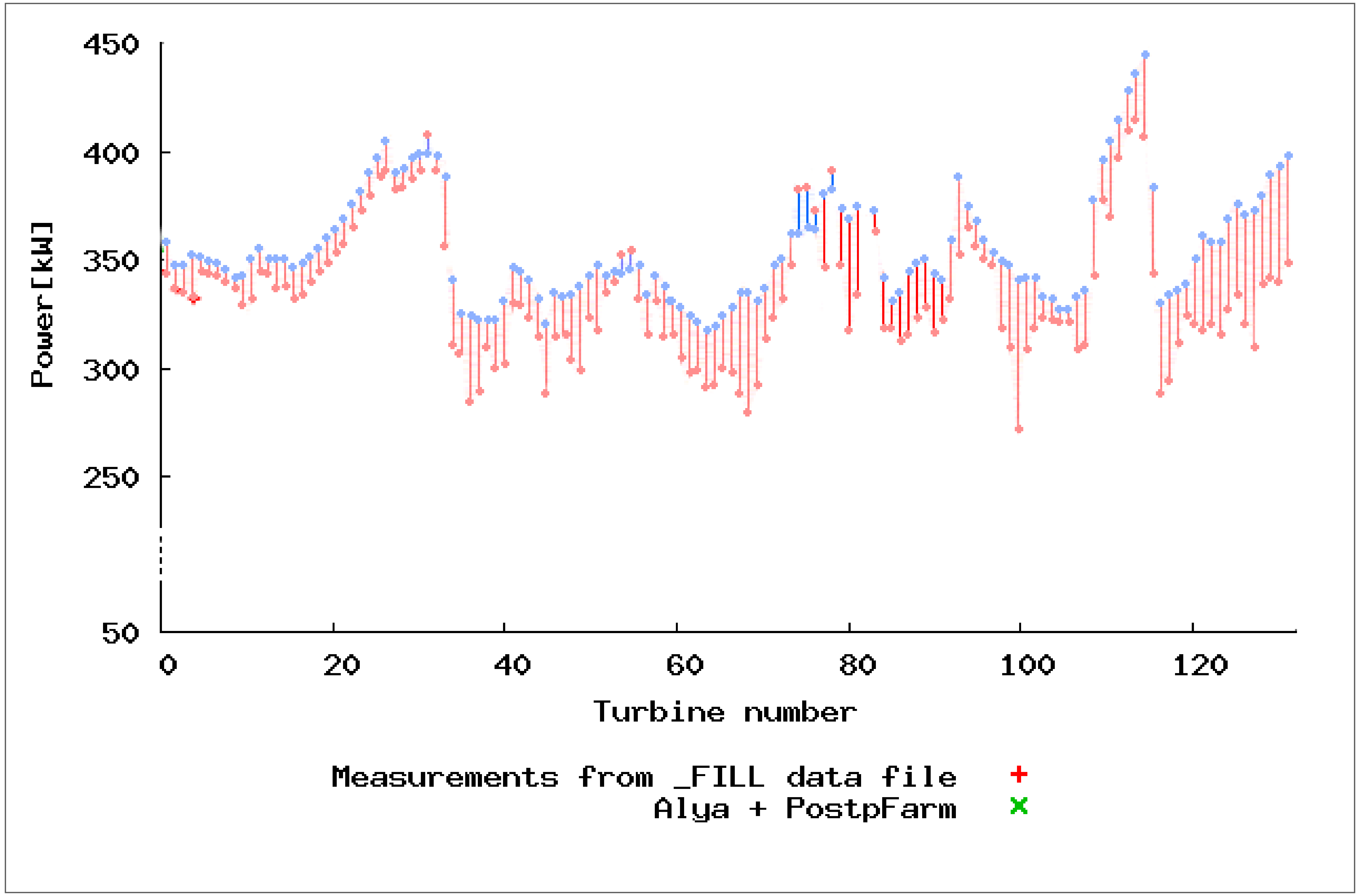




Primer prototipo: Falso difference chart con puntos separados

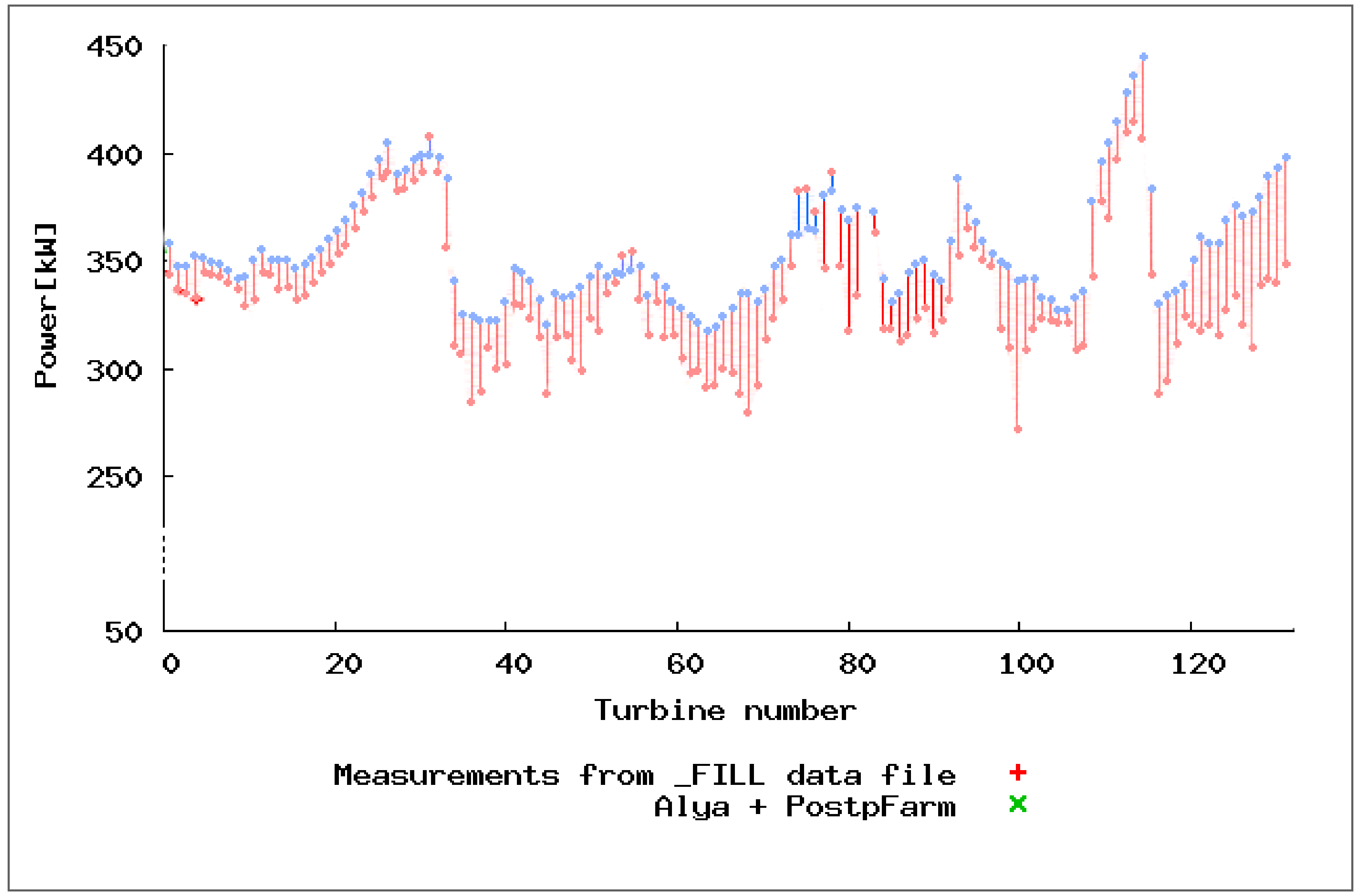
- Hecho a mano en Adobe Illustrator
- Probar esquemas de color diferentes, evitar colores “semáforo”
- Romper la continuidad percibida de un line chart

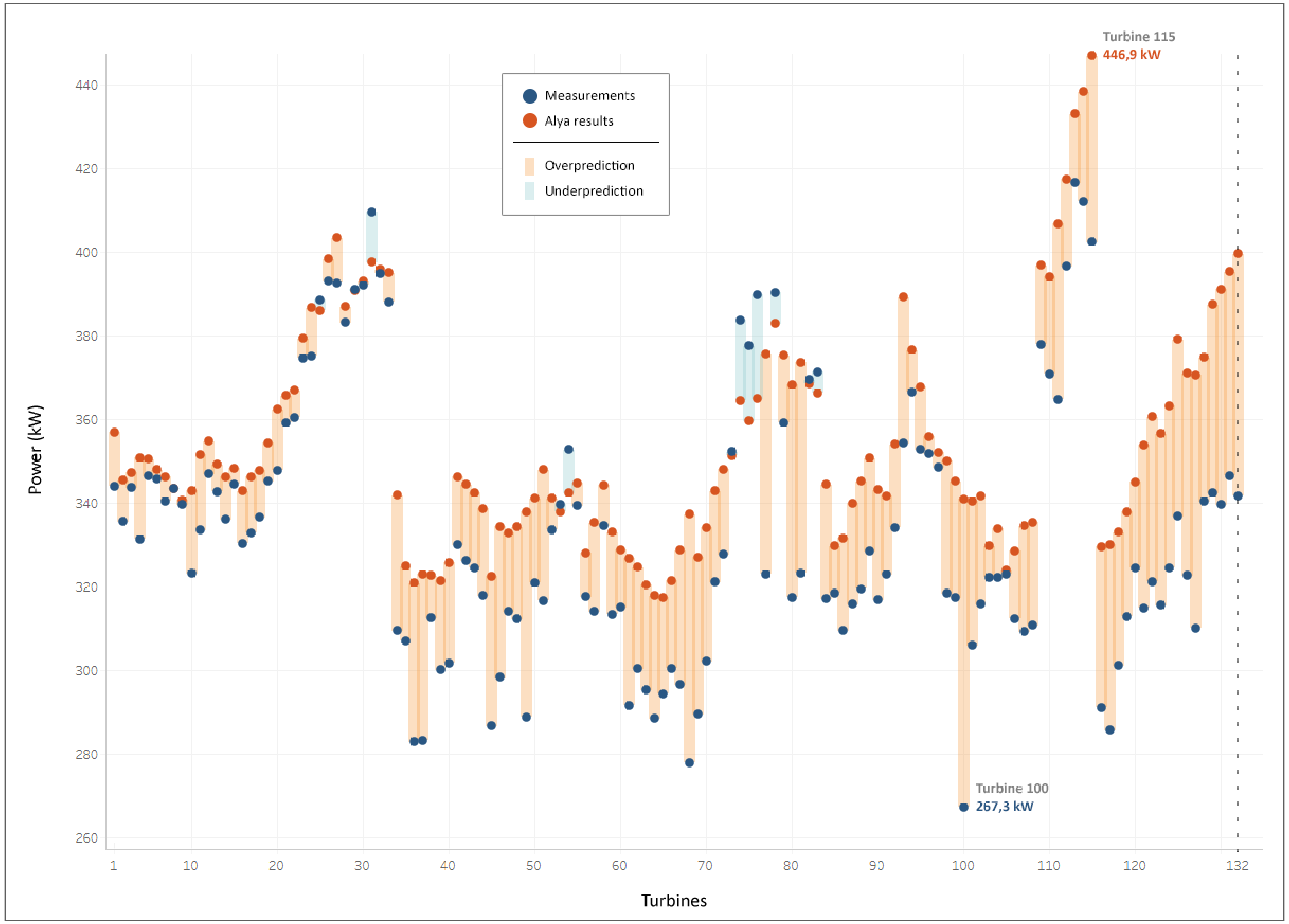




Segundo prototipo: Datos reales en Tableau Public y exportado a Illustrator

- Escoge tu herramienta
- Re-ajustar colores (puntos más oscuros que las líneas)
- **Exportar SVG** a software de diseño
- Añadir anotaciones, crear legend box, ajustar ejes. Cualquier tarea que no dependa del número de elementos



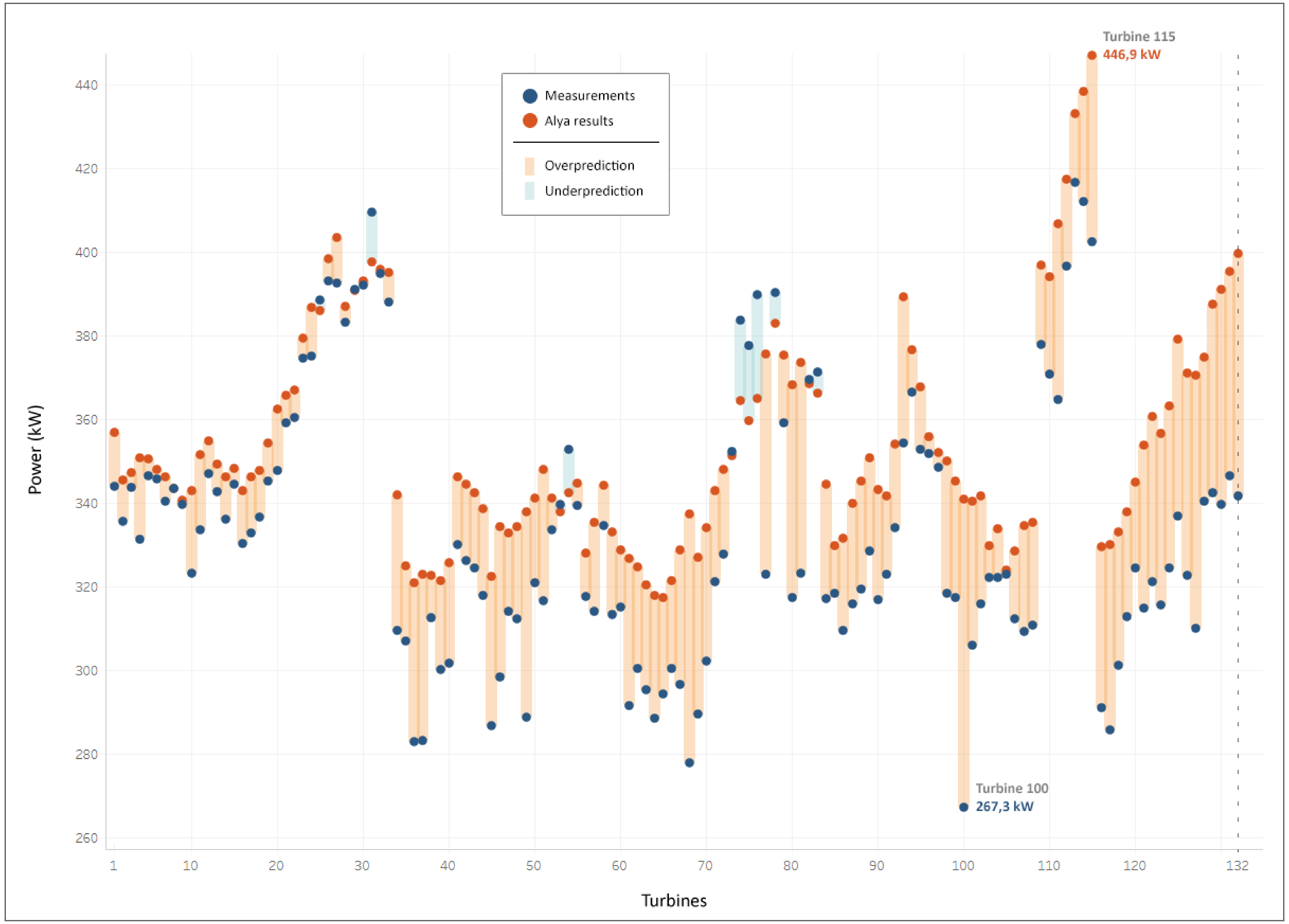


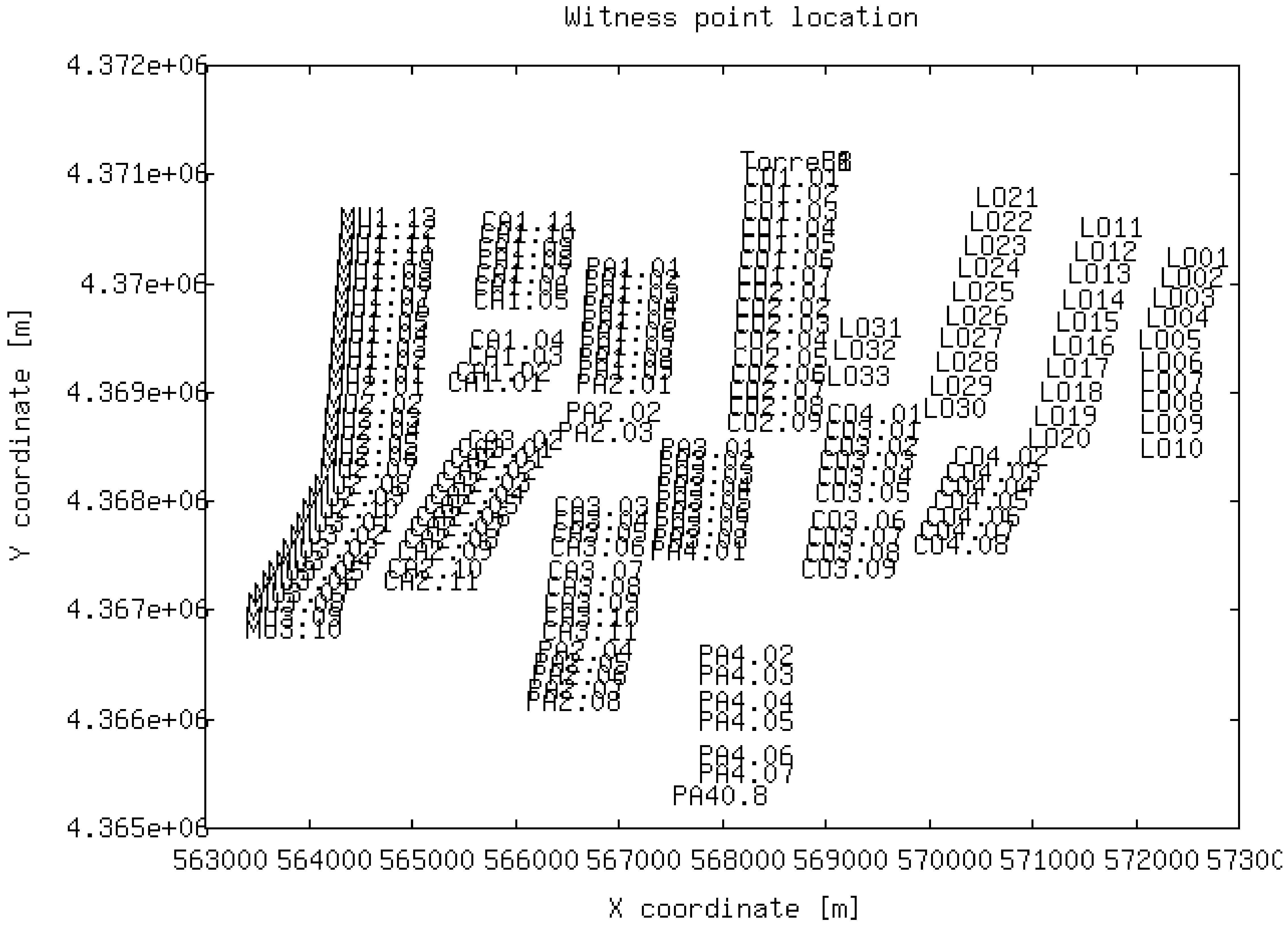
Second prototype: Real data in Tableau Public

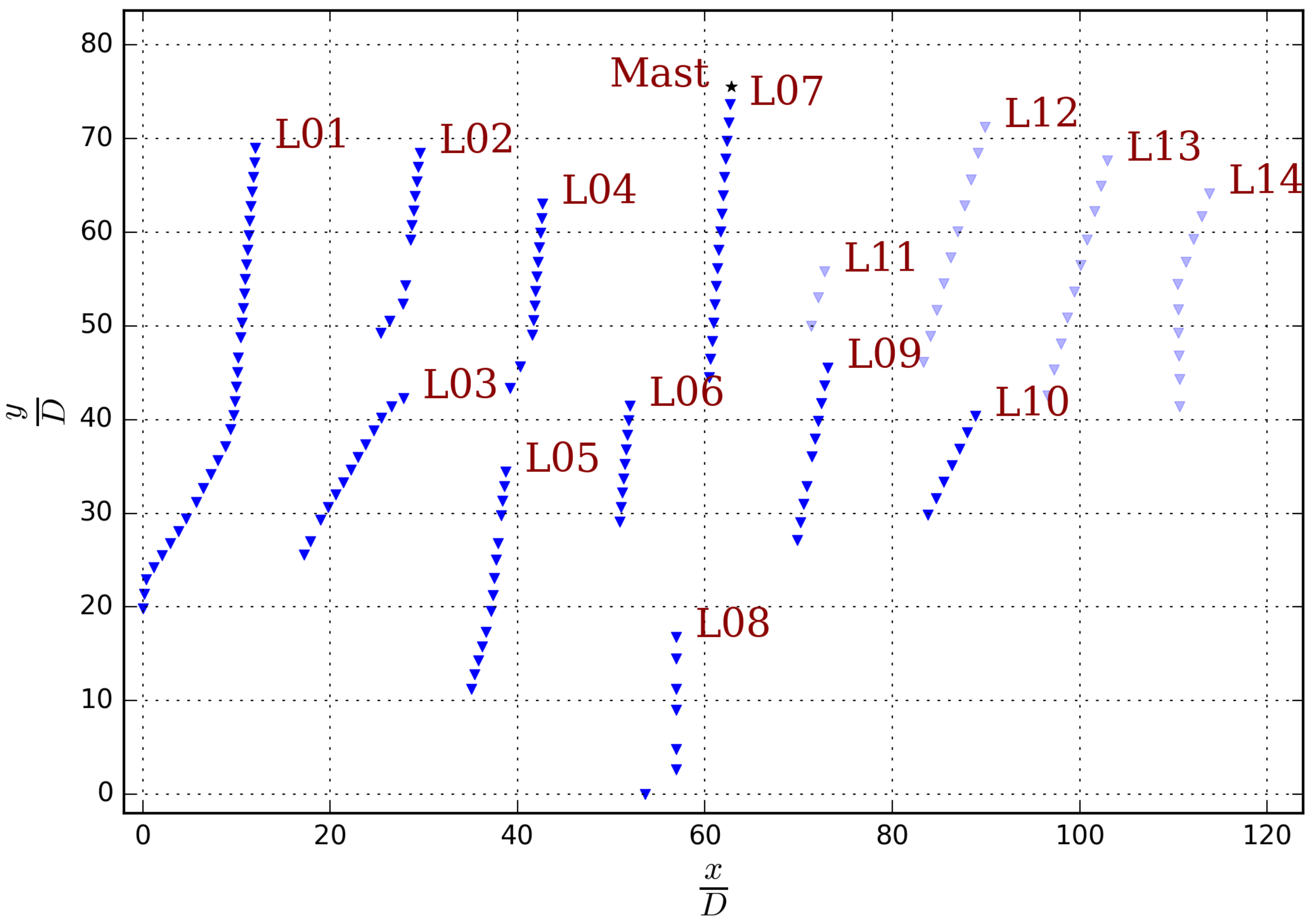
Cual es el desafío?

Resaltar las diferencias entre valores y aclarar qué se representa en cada eje

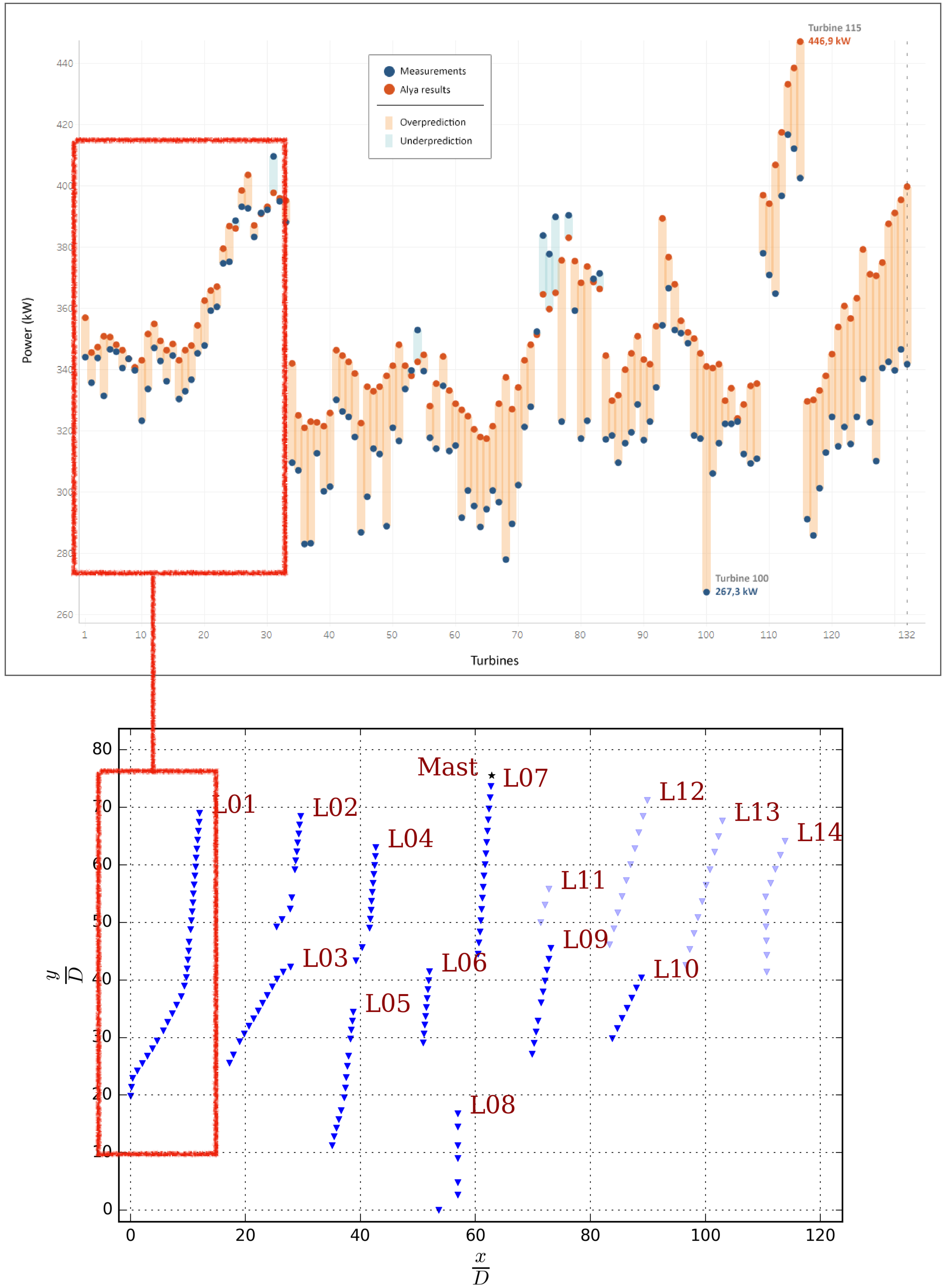
Un eje progresivo no funciona para el número de turbina: se asocia a datos continuos

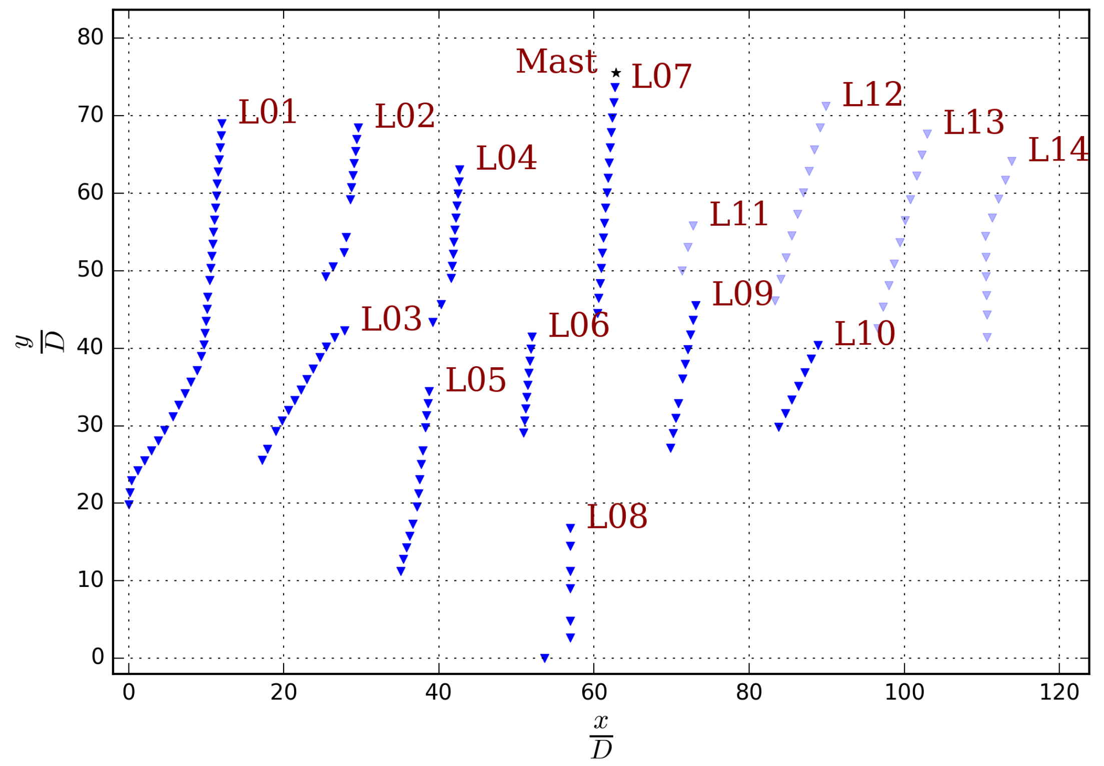



Witness point location
4.372e+06
4.371e+06
4.37e+06
4.369e+06
4.368e+06
4.367e+06
4.366e+06
4.365e+06
Y coordinate [m]
563000 564000 565000 566000 567000 568000 569000 570000 571000 572000 5730C
X coordinate [m]
L001
L002
L003
L004
L005
L006
L007
L008
L009
L010
L011
L012
L013
L014
L015
L016
L017
L018
L019
L020
L021
L022
L023
L024
L025
L026
L027
L028
L029
L030
L031
L032
L033
PA4.02
PA4.03
PA4.04
PA4.05
PA4.06
PA4.07
PA40.8

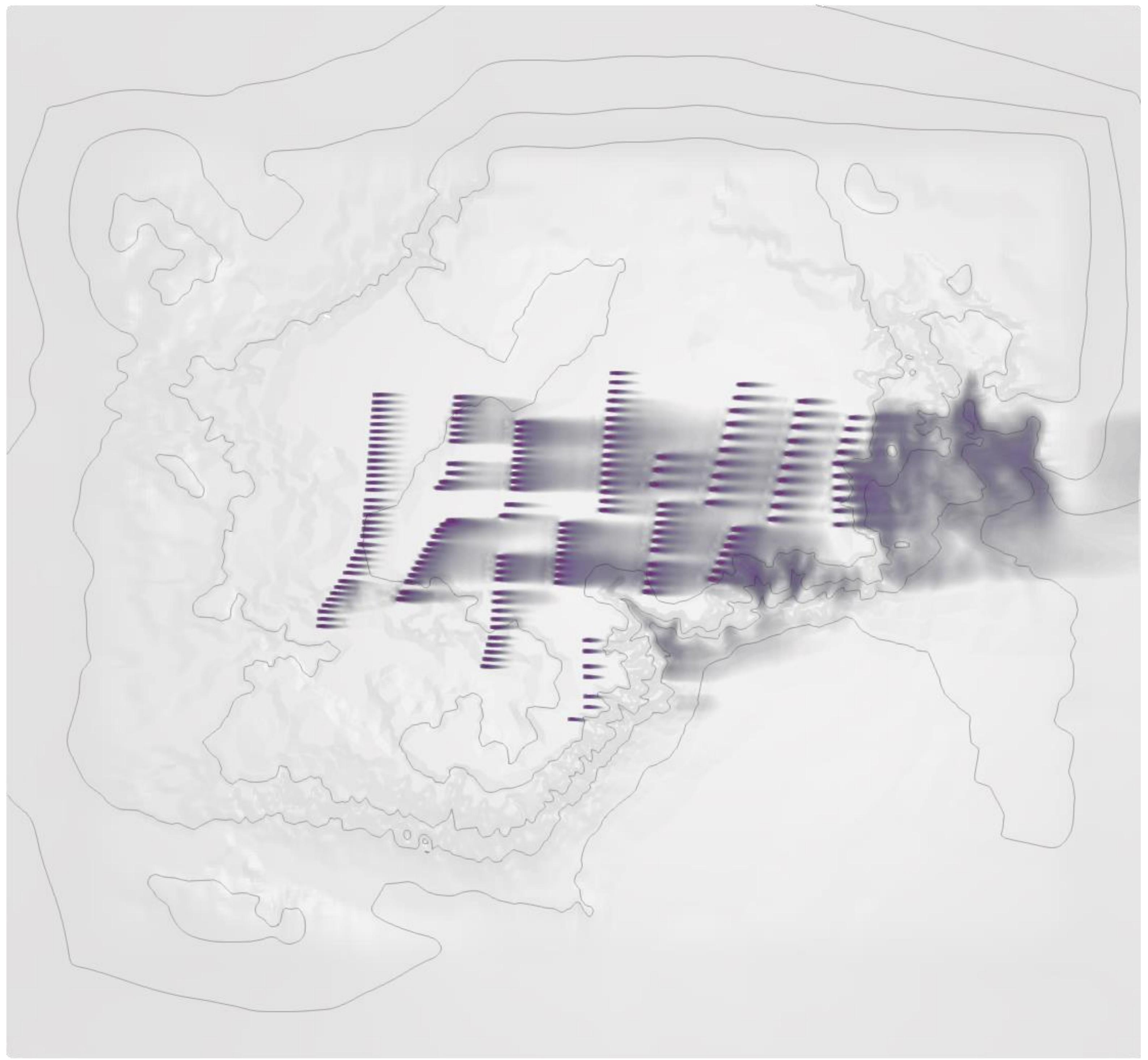

Mast
L01
L02
L03
L04
L05
L06
L07
L08
L09
L10
L11
L12
L13
L14
0
10
20
30
40
50
60
70
80
$\frac{y}{D}$
0
20
40
60
80
100
120
$\frac{x}{D}$

- Unir las gráficas para eliminar el eje X
- Rompe la continuidad y permite comparar diferencias
- Explorar la relación entre potencia generada y localización puede aportar otra perspectiva

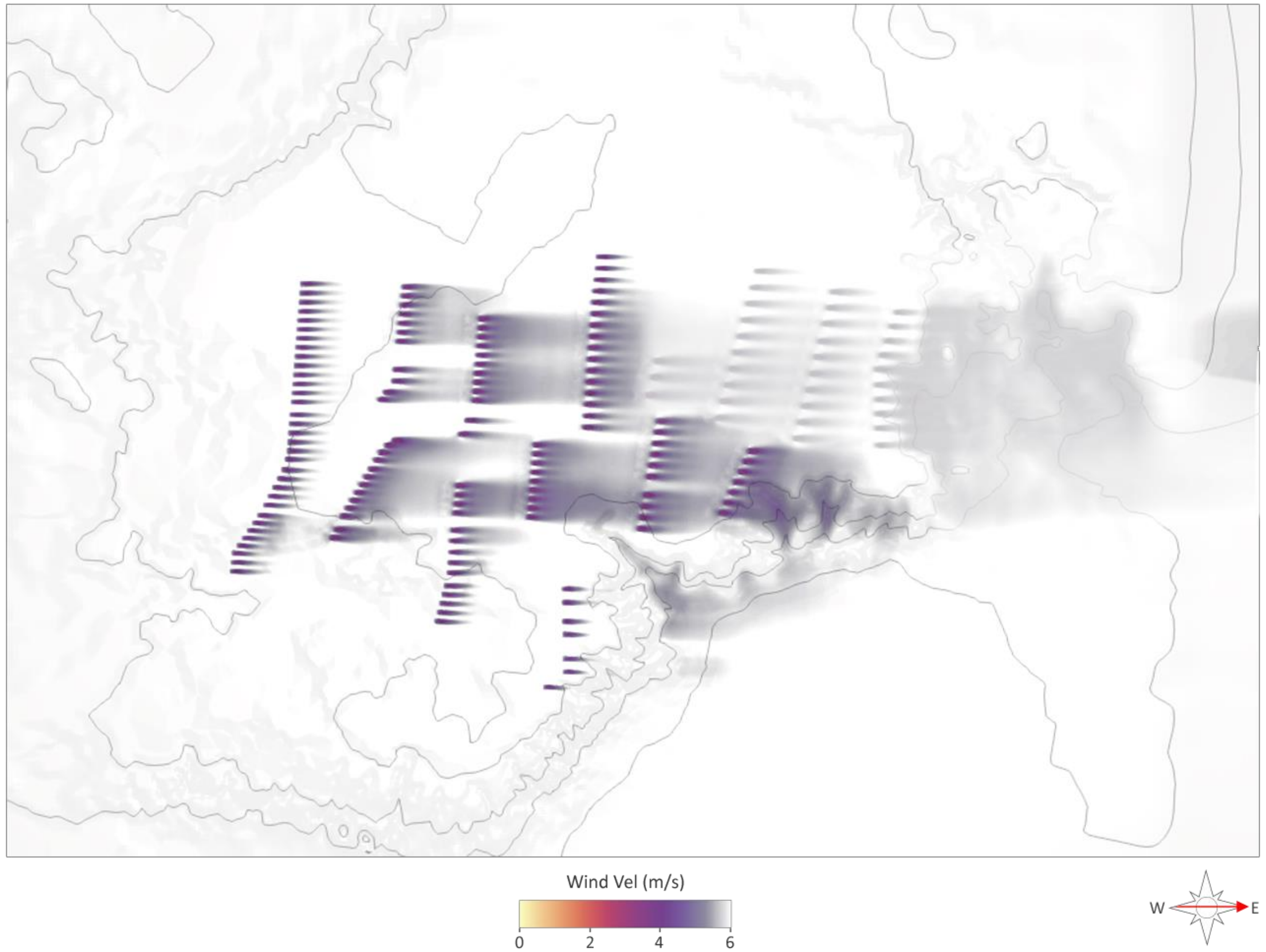




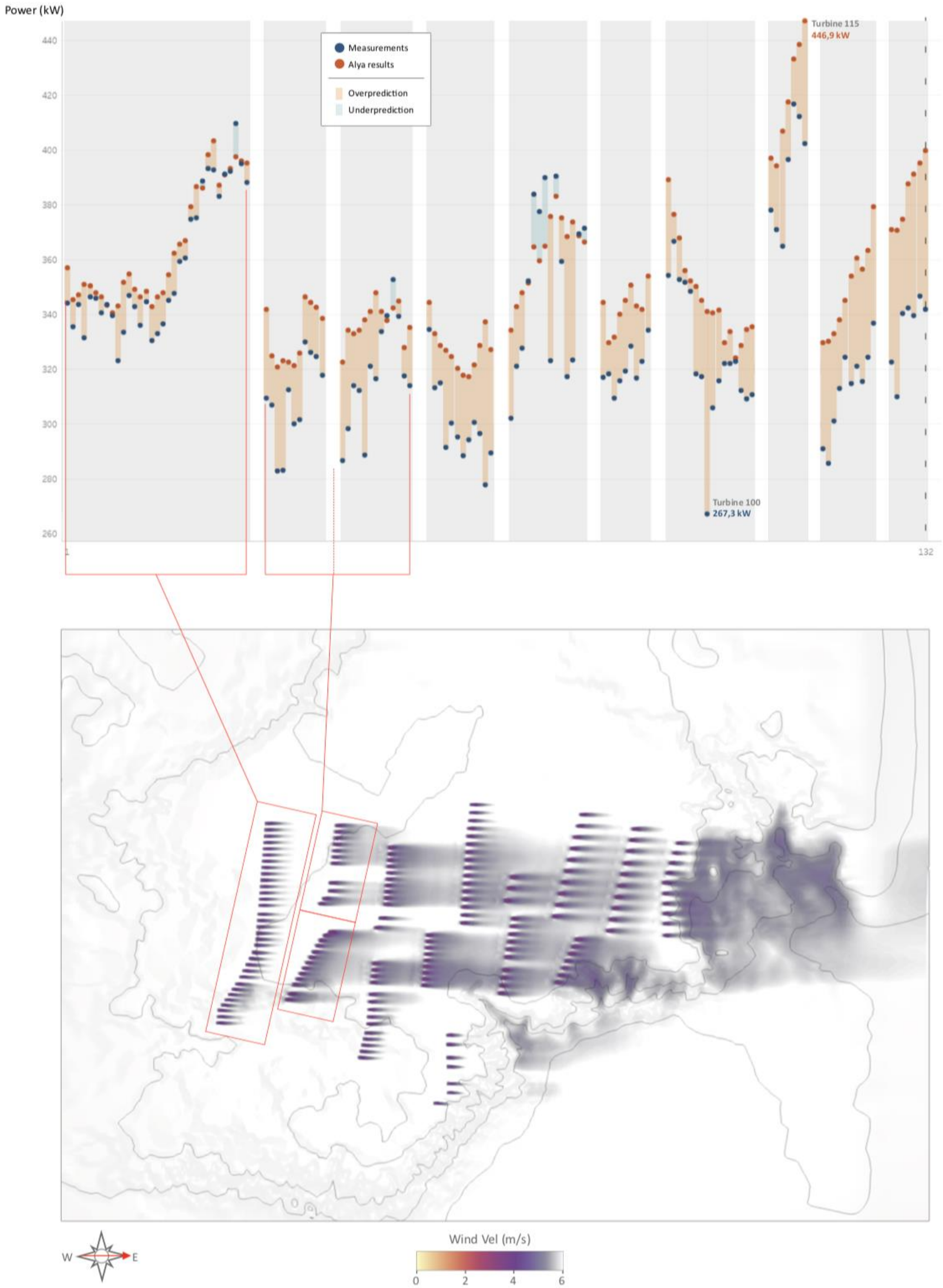
Original de Paraview (Sci-Viz software)



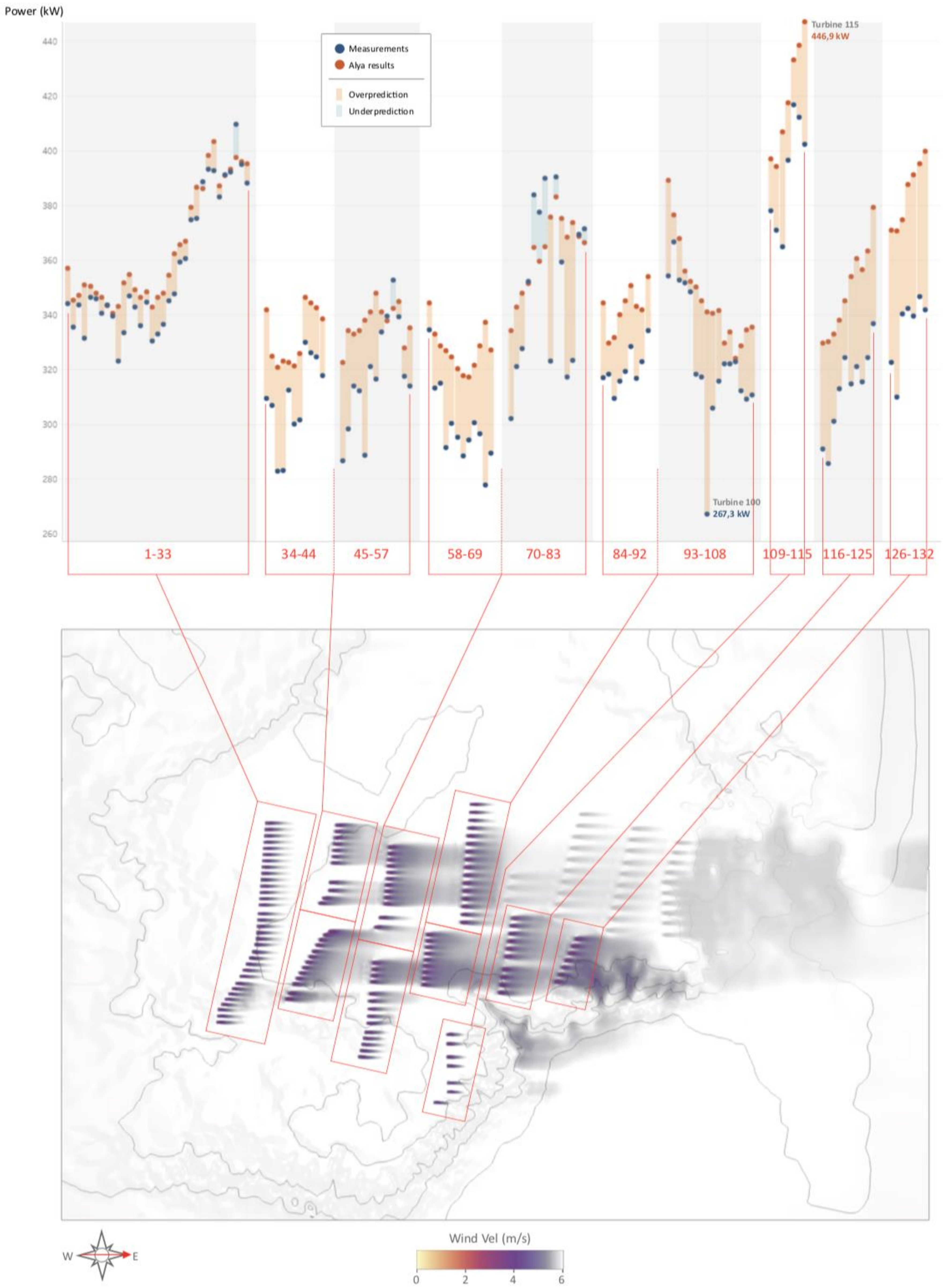
Escala de color, orientación, y ajustes de imagen en Illustrator



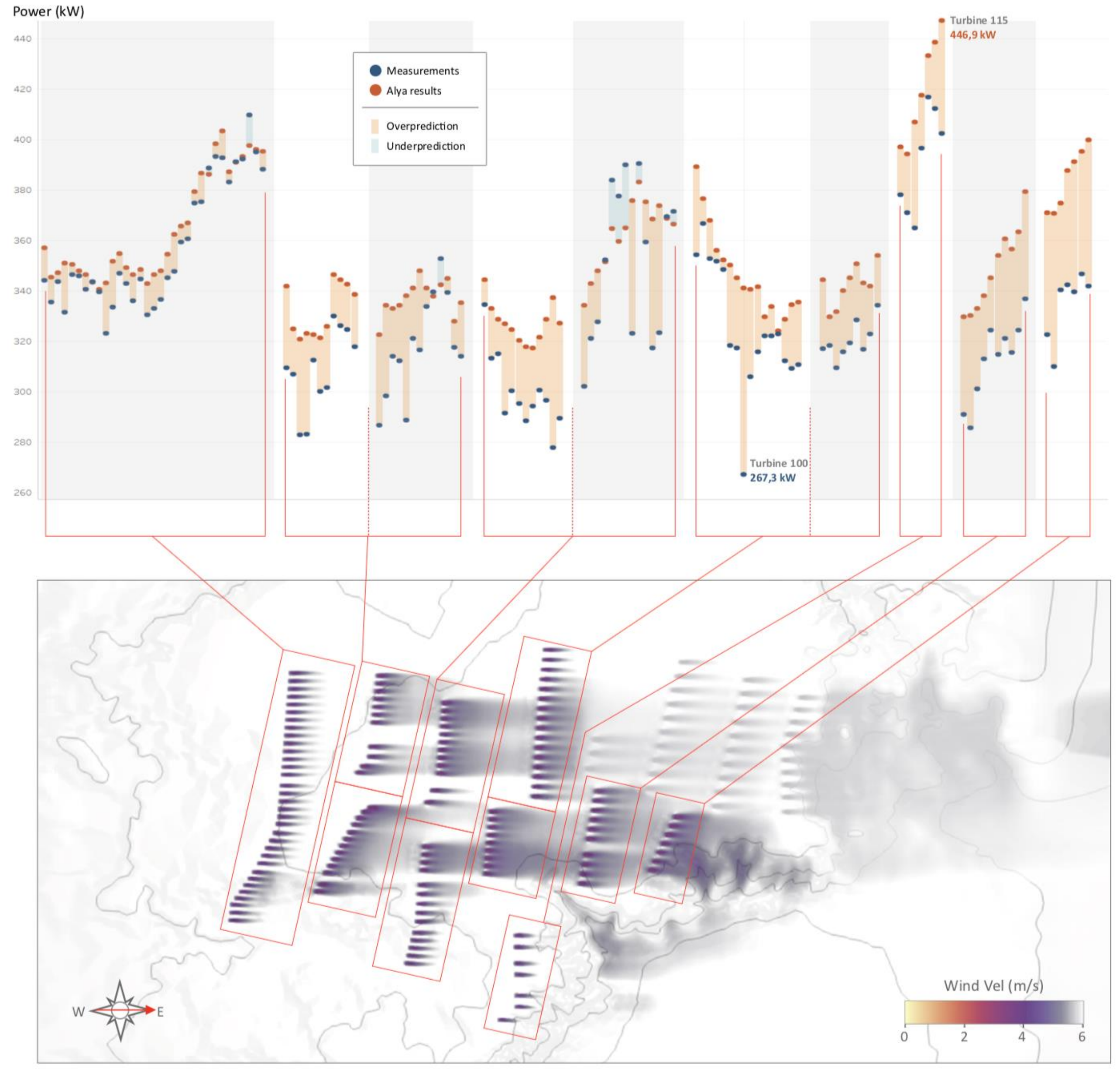
V1


Power (kW)
440
420
400
380
360
340
320
300
280
260
Measurements
Alya results
Overprediction
Underprediction
Turbine 115
446,9 kW
Turbine 100
267,3 kW
1
132
Wind Vel (m/s)
0
2
4
6
W
E

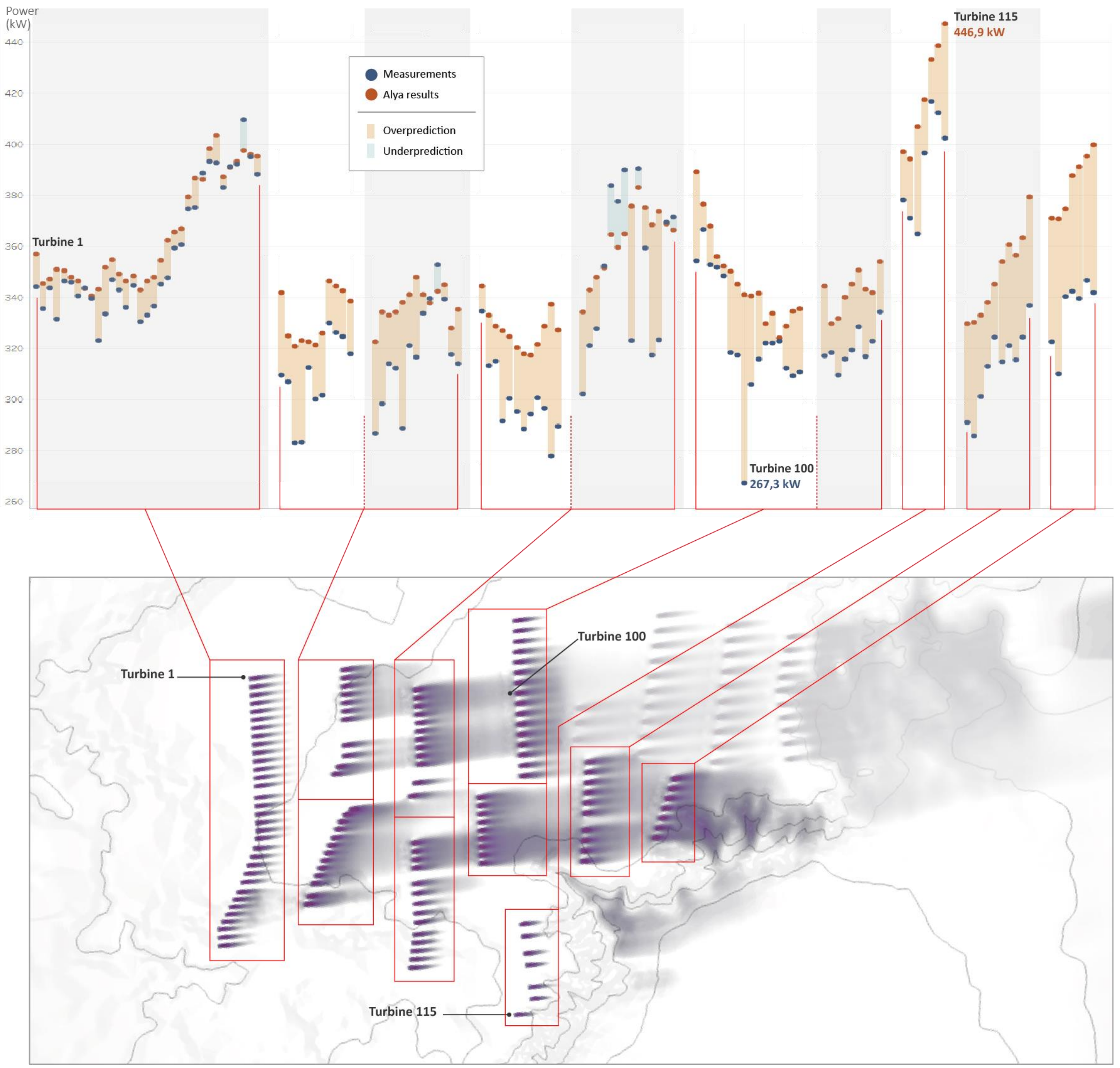
V2


Power (kW)
440
420
400
380
360
340
320
300
280
260
Measurements
Alya results
Overprediction
Underprediction
Turbine 115
446,9 kW
Turbine 100
267,3 kW
1-33
34-44
45-57
58-69
70-83
84-92
93-108
109-115
116-125
126-132
Wind Vel (m/s)
0
2
4
6
W
E

V4



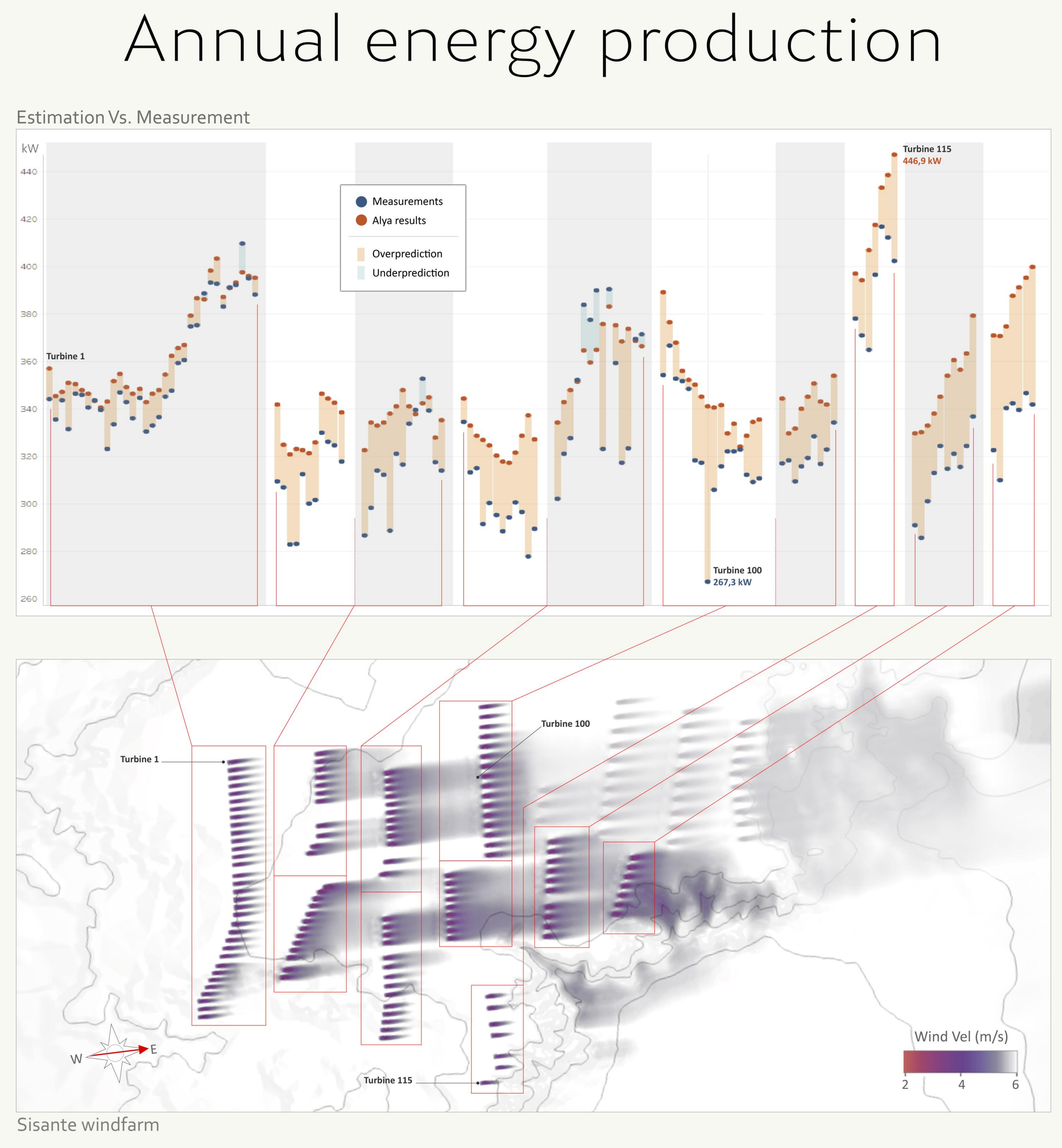
V7



- Rotar mapa para poner los recuadros rectos
- Sustituir eje X por los cuadros rojos
- Marcar el primero, el último y los molinos con valores extremos en ambas gráficas para indicar el orden de lectura

Imagen Final

- Enmarcar y separar las gráficas para facilitar legibilidad: no es un plot único
- Color de fondo claro para resaltar gráficas (poco contraste)
- Título y subtítulo: Jugar con tamaño y cuerpo para establecer jerarquía del póster
- Ventaja: Muestra la relación entre la turbulencia y la exactitud de los resultados. Efecto conocido por los investigadores, pero es un buen modo de explicarlo a stakeholders

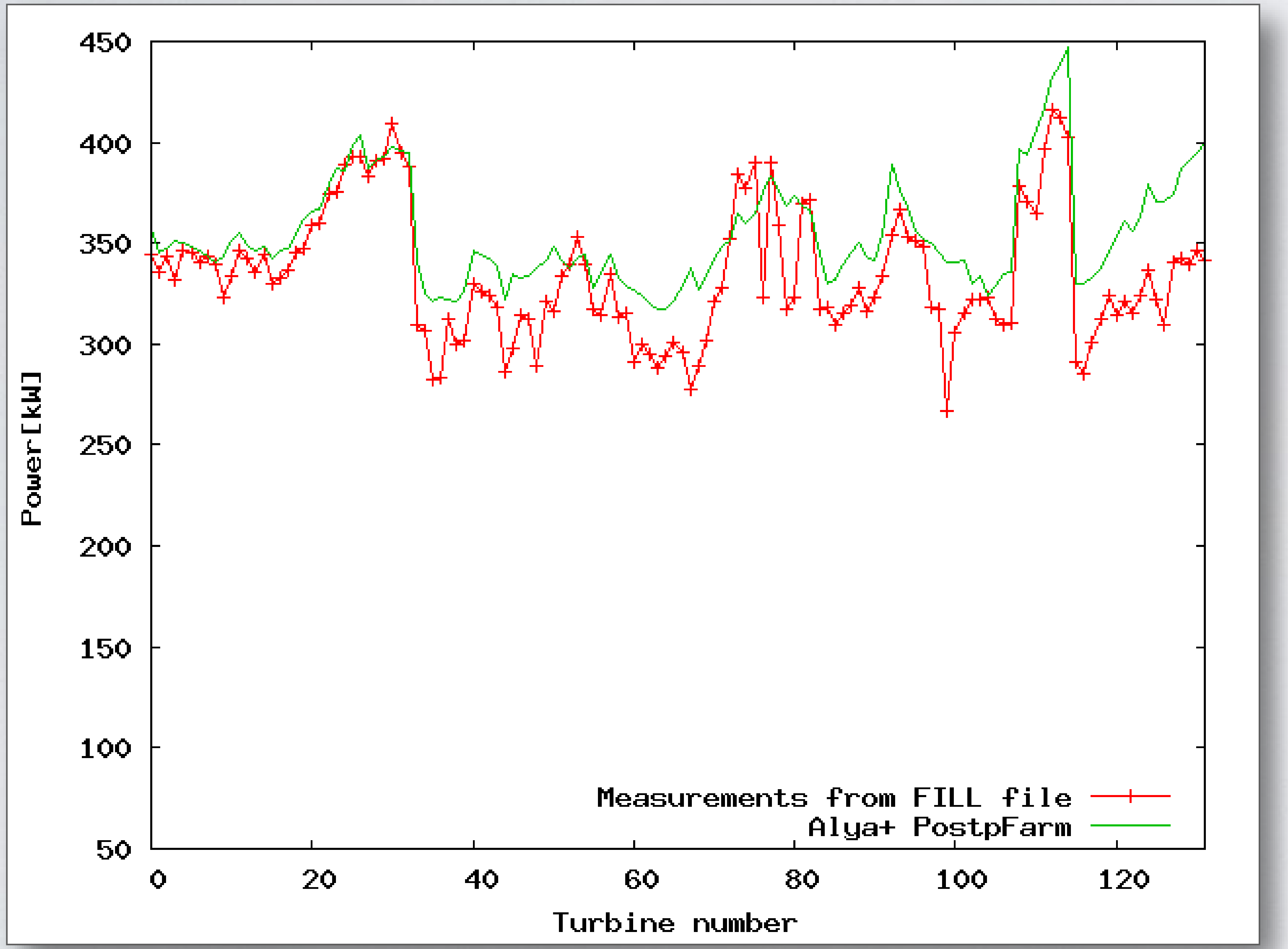


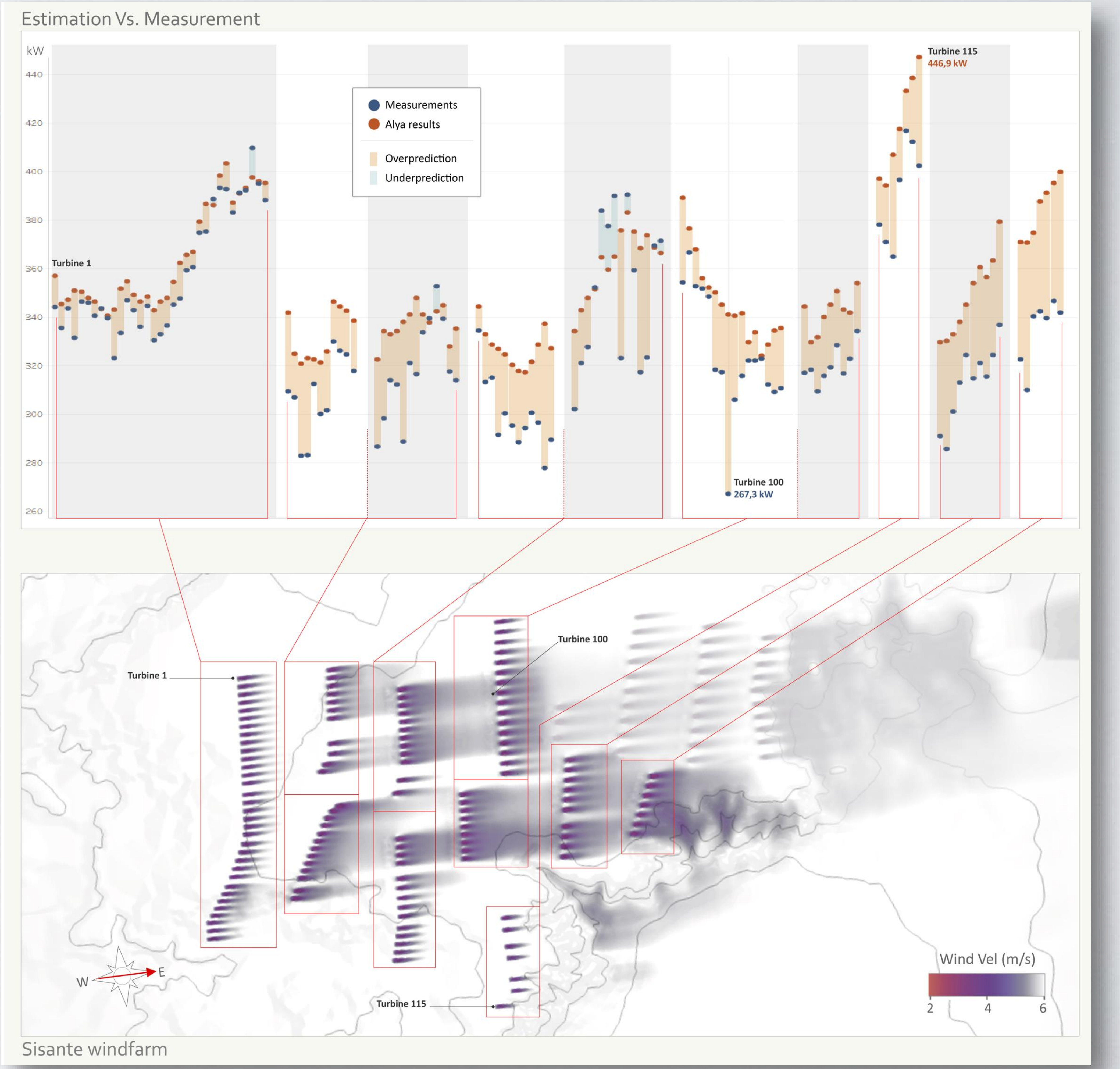
Siempre hay un mensaje

My results somehow correlate to measurements

Vs.

My model is not capturing turbulent wakes accurately

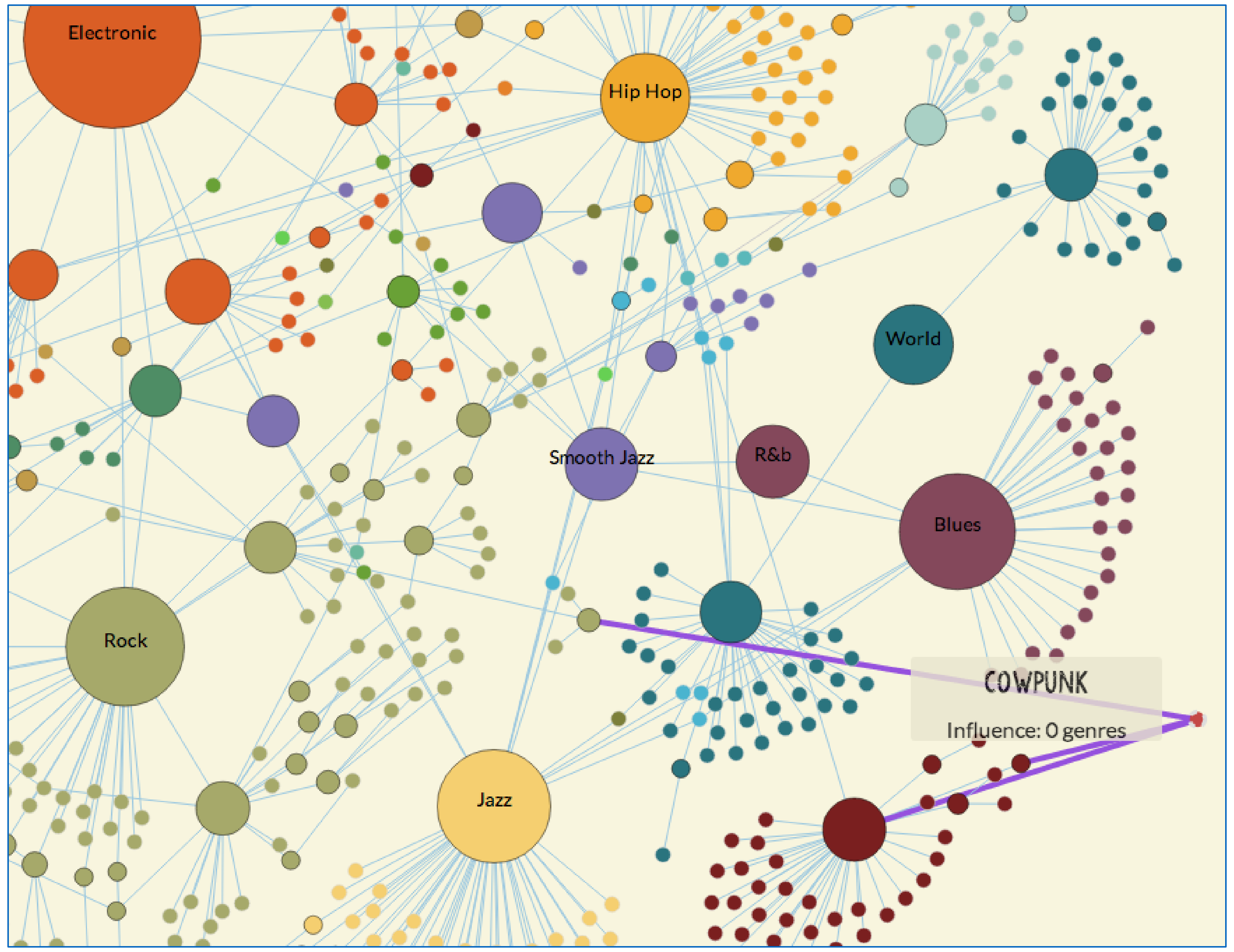




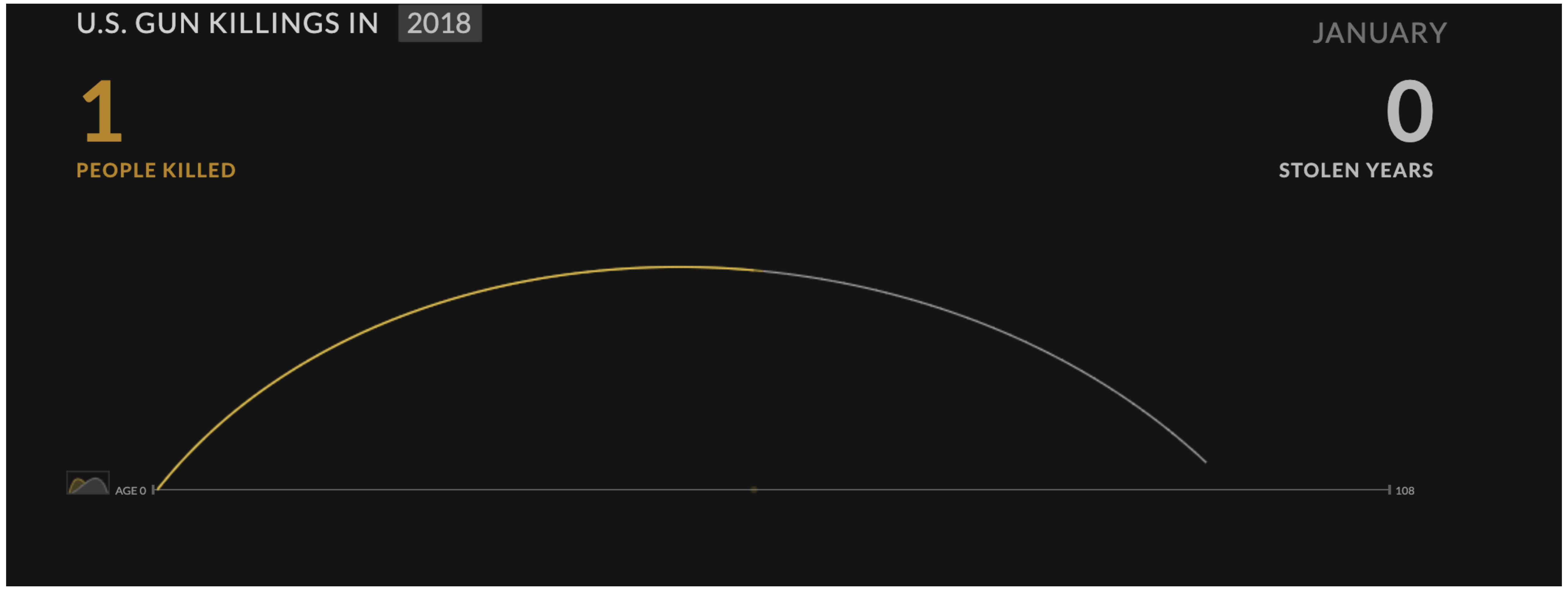
Tres tipos de estructura narrativa aplicables a data viz

No tiene que ser una historia necesariamente

- Hay varias estrategias para ayudar al usuario a navegar por una visualización:
  - **Narración**. Anotaciones conectadas en una secuencia lógica
    - **Cerrada.** Estructura secuencial; mensaje definido
    - **Abierta.** Visualización exploratoria. Visualización como herramienta de búsqueda de narrativas
  - **Historia.** Una narración con personajes, factor humano, drama/conflicto para obtener una respuesta emocional



- Gun killings in 2018. Periscopic https://guns.periscopic.com/



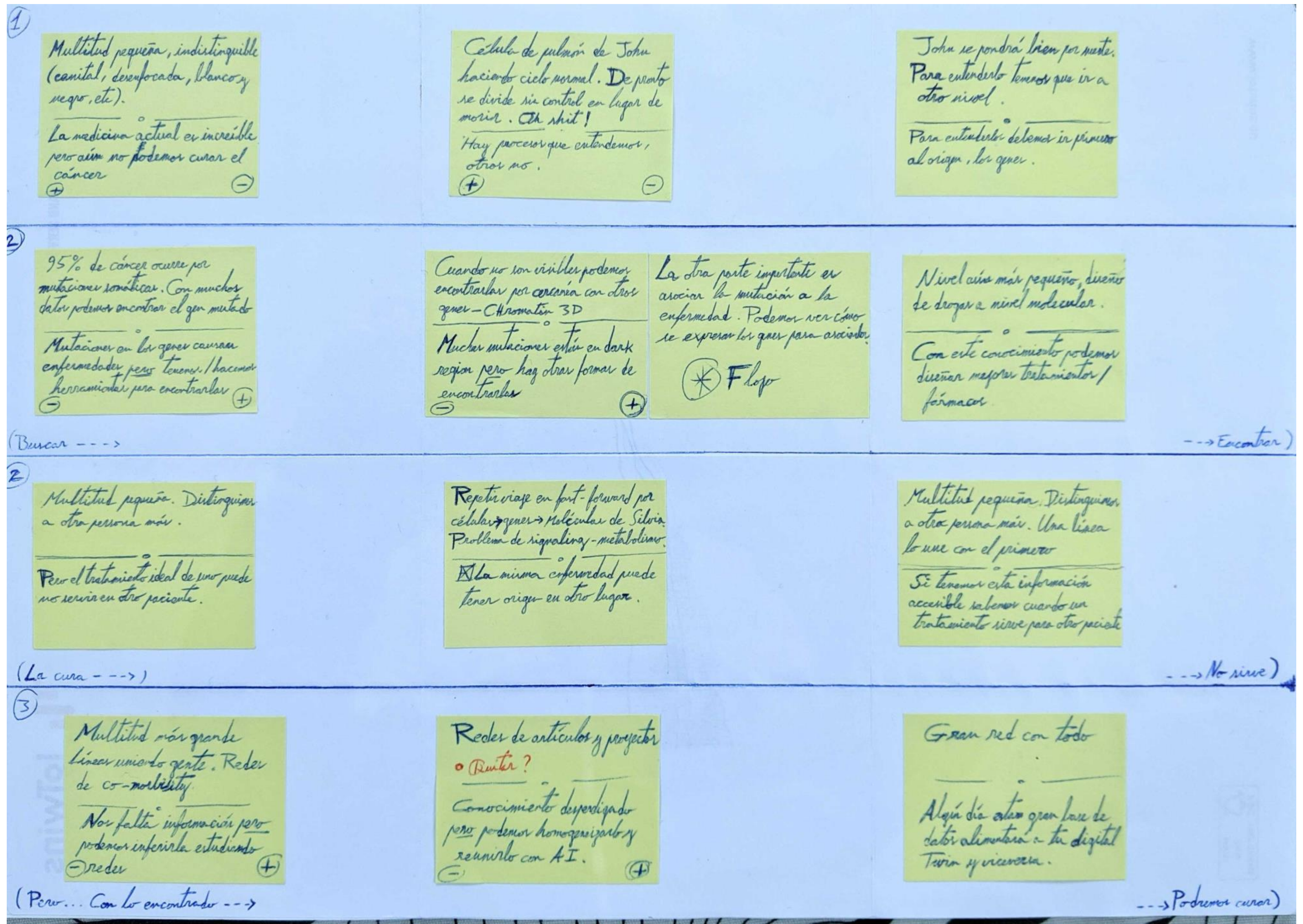
Resumen

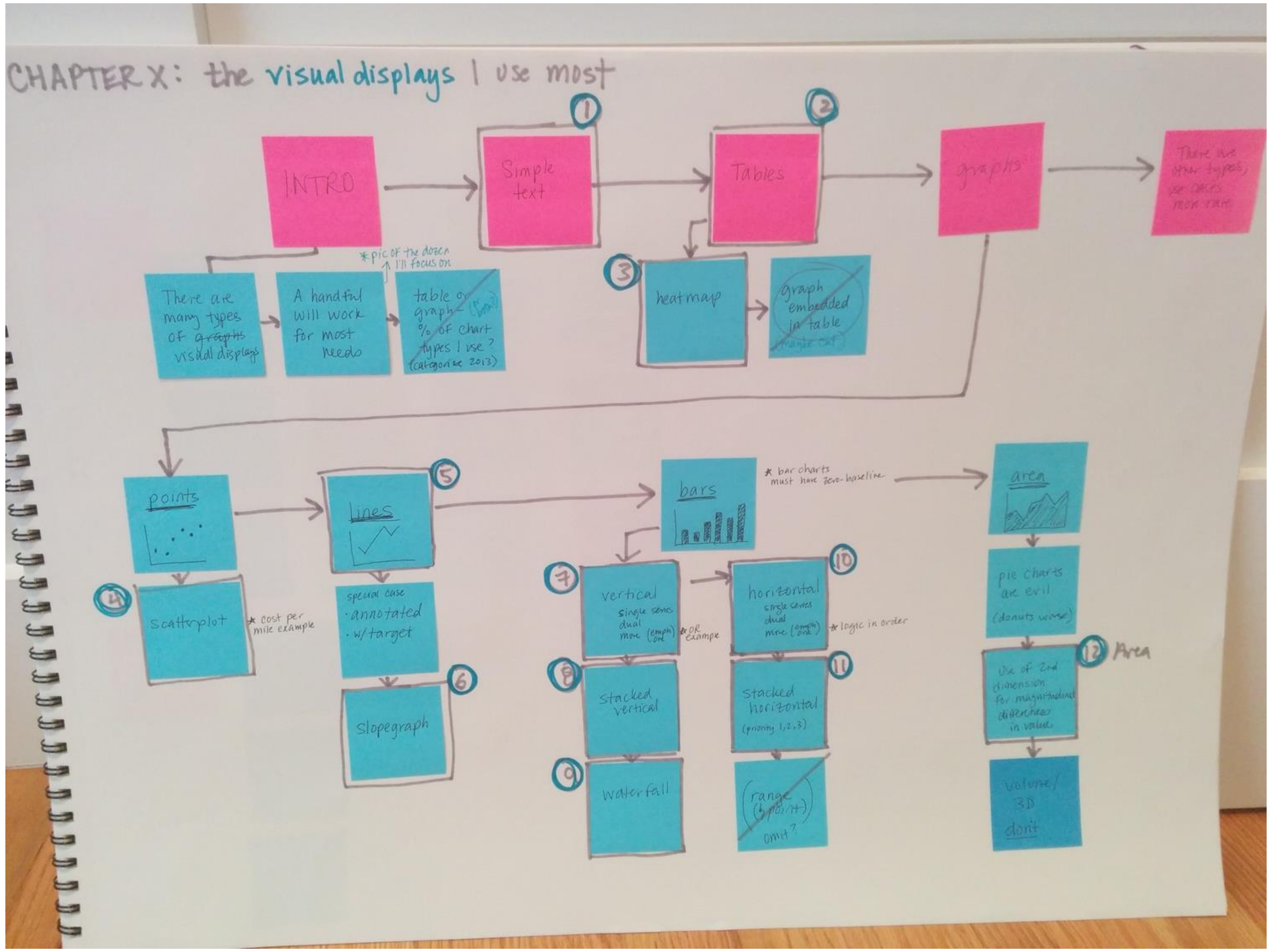
- Es necesario guiar a la audiencia
- La forma de hacerlo depende del contexto previo que tengan
- Hay varias estrategias para ayudar a la audiencia a navegar por una visualización:
  - **Anotación**. Texto y marcas que ayuden a interpretarla
  - **Narración**. Anotaciones conectadas en una secuencia lógica
  - **Historia**. Una narración que contenga drama/conflicto para obtener una respuesta emocional

- La audiencia puede tener o no un interés previo en el tema. Podemos crear interés utilizando técnicas de storytelling: conflicto, tensión, empatía, …

Primer paso de planificación: Storyboard

- Organizar ideas en secciones, subsecciones
- Cada parte está reducida a la mínima expresión (Una acción, una frase, …)
- Storyboard permite ver el flujo de la historia… y cambiarlo
- Multinivel: Esquema de la historia que se puede desarrollar en subsecciones
- Herramientas similares para planificar cualquier tipo de narrativa (guión=board)





www.storytellingwithdata.com

Referencias

- Campbell, J. (2008). *The hero with a thousand faces*
- McKee, R. (2011). Story.
- Snyder, B. (2005). Save the cat.
- Vogler, C. (2007). The Writer's Journey: Mythic Structure for Writers.
- Rachael Stephen. Story circle aplicado https://www.rachaelstephen.com/videolibrary/category/storyfix+episodes

- **David JP Phillips** on science of storytelling para presentaciones https://youtu.be/Nj-hdQMa3uA

- Podcast Data Stories 035: Visual Storytelling. https://datastori.es/podcast/data-stories-35-visual-storytelling-w-alberto-cairo-and-robert-kosara/
- Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE transactions on visualization and computer graphics, 16(6), 1139-1148.
- Stolper, C. D., et al. (2016). Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories.
- Knaflic, C. N. (2015). **Storytelling with data**: A data visualization guide for business professionals. John Wiley & Sons. (capítulos 1 y 7)

- **Tutorial de story en Tableau: https://help.tableau.com/current/pro/desktop/en-us/story_example.htm**

Práctica Tableau

Parte 1.1

Dataset: any_drinking.csv

- Cargar dataset
  - Datasource -> Text File
  - Formateo simple: Ordenar, describir, rename, esconder.
  - ESCONDER COLUMNAS EXCEPTO LAS DE 2012
  - Campos calculados: Crear campo Diff_2012 (Male -Female).

- Hacer una tabla de Both_Sexes_2012 por estado
  - Measures (verde, cuantitativos) y Dimensions (azul, categóricos)
  - Arrastrar State a Rows y Diff_2012 a Marks/Text. Cambiar la forma de agregar a Average (AVG) (por defecto: SUM). Filter out National.
- Barchart simple de Difference por State
  - New sheet. Arrastrar State y Diff 2012 a Rows y Columns.
  - Swap columns/bars, sort, edit axis // Ajustar color, size, tooltip

state_table

State	
Texas	19.085
South Carolina	18.664
West Virginia	18.546
Louisiana	18.326
Arkansas	18.284
Kansas	18.196
Nebraska	18.191
Mississippi	17.873
Alabama	17.532
Oklahoma	17.346
Missouri	17.314
North Dakota	17.272
Hawaii	16.840
Iowa	16.809
North Carolina	16.653
South Dakota	16.600
Indiana	16.408
Virginia	16.245
Kentucky	16.094
Georgia	15.996
Pennsylvania	15.884
Ohio	15.570
New Mexico	15.538
Tennessee	14.986
Nevada	14.178
Illinois	14.075
Minnesota	14.060
Wyoming	14.038
Florida	13.941
District Of Columbia	13.800

Parte 1.2

Dataset: any_drinking.csv

- Boxplot
  - New sheet. Arrastrar State y Diff 2012. // Uncheck Analysis/Aggregate Measures
  - Marks: circles. Ajustar tamaño, color y borde // Arrastrar Location a Tooltip
  - Right-click en eje Y: Add reference Line/ Boxplot

- Histograma
  - Crear **bin** en Data Source -> New Sheet -> Crear gráfica con Diff2012 y el nuevo bin (Probar Bin values)
  - ¿Qué agregación debería tener Diff_2012?

- Barchart Diff_2012 - Location
  - Barchart por defecto agrega Counties con el mismo nombre
  - Filtrar por estado sin crear nuevo campo: En: Work Sheet/Data pane/Create/Combined field, seleccionando Location y State
  - Usar campo combinado en Rows

Parte 1.3

Dataset: any_drinking.csv

- Mapa de Símbolos proporcionales
  - Mapa por estado. Seleccionar State, ir a Show me/map
  - (5') Graficar un atributo Measure (cualquiera) con círculos. Tamaño y color controlados por el attr
  - Doble click en las escalas para modificar rangos
- Mapa por Location
  - Cambiar Geographic Role de Location a "County" antes de hacer el mapa.
  - (5') Graficar dos attr. tipo Measure (uno en color y otro en tamaño) sobre mapa como círculos
- Variación: Coropleto. Crear map, añadir Location y State a Detail. Attr. Cuantitativo en el color.

GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

11. Interactivitat i Animació

Enric Martí Gòdia

Departament de Ciències de la Computació

Continguts

1. Introducció
2. Interactivitat
3. Animació
4. Factors per a implementar la interactivitat

1. Introducció

Tecnologia ha canviat la forma en que les visualitzacions es produeixen, comparteixen, consumeixen.

La capacitat dels dispositius (ordinadors, *smartphones*, etc.) ha facilitat la producció d'experiències interactives i animades.

Interactivitat transforma a l'observador → usuari, donant-li més protagonisme i ser més pro-actiu en la percepció de la informació, menys tutelada per visualitzador.

Animació mostra dades amb dimensió temporal de forma dinàmica. Permet apreciar patrons que no es poden veure en imatges estàtiques.

1. Introducció

Malgrat l'observador sigui pro-actiu, visualitzador no pot renunciar a la seva responsabilitat en presentar la informació per a facilitar a comprensió.

Cal que la incorporació d'interactivitat i/o animació en la visualització de dades sigui justificat i que no distregui o impedeixi la comprensió.

2. Interactivitat

Avantatges d'incorporar interactivitat en projectes de Visualització de Dades:

- Expandeix límits físics del que es pot veure en un espai determinat.
- Cobreix una major varietat d'anàlisis per a atendre diferents necessitats d'informació.
- Facilita manipulacions de dades per a contestar diferents preguntes.
- Amplifica el control i la customització de l'observador.
- Amplia l'interès a diferents grups d'usuaris, especialment els joves, acostumats a ser més actius en el consum d'informació.

2. Interactivitat

Previ a decidir les tècniques interactives, cal estudiar què es necessita, de quins mitjans es disposa i què es pot utilitzar.

Conèixer els mètodes per a

- interrogar
- manipular
- navegar

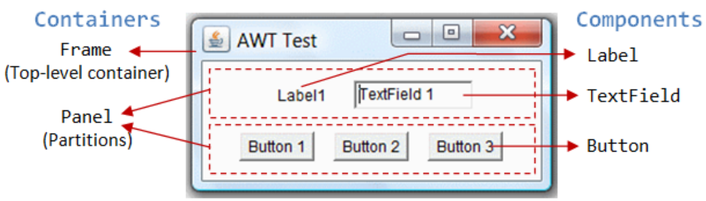
en experiències digitals

Conceptes importants en interactivitat:

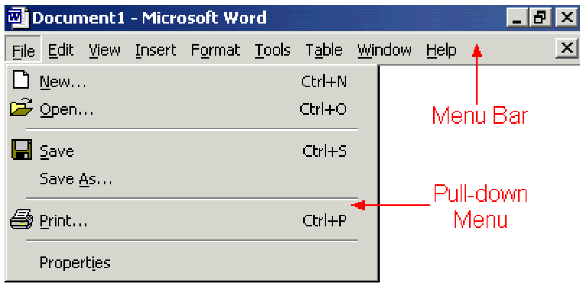
- **Event:** Acció de l’usuari. Prémer tecla, click mouse.
- **Control:** Reacció a l’event, què volem que passi. Desplegar pop up, ressaltar opció, etc.
- **Acció:** Operació a realitzar en l’aplicació. Filtrar o ressaltar informació, etc.).

2. Interactivitat

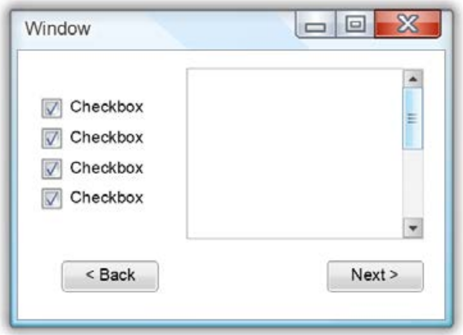
Tipus de controls:



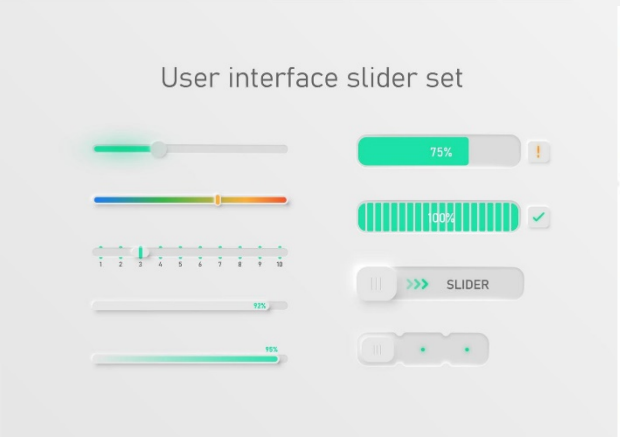
Botó



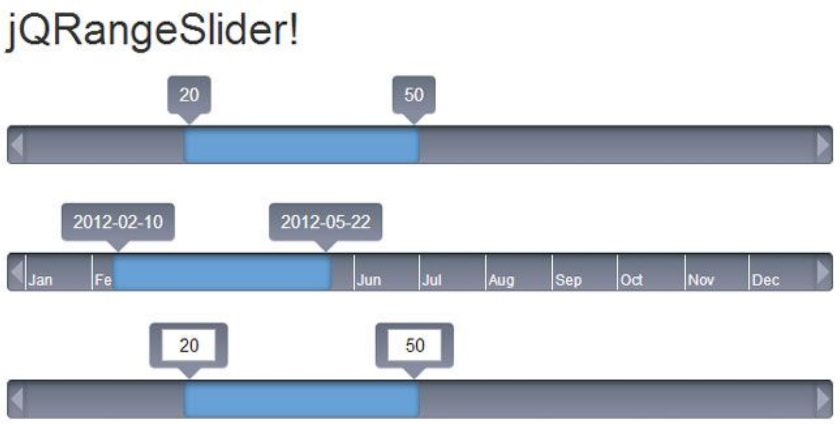
Pop-up – Pull-down menu



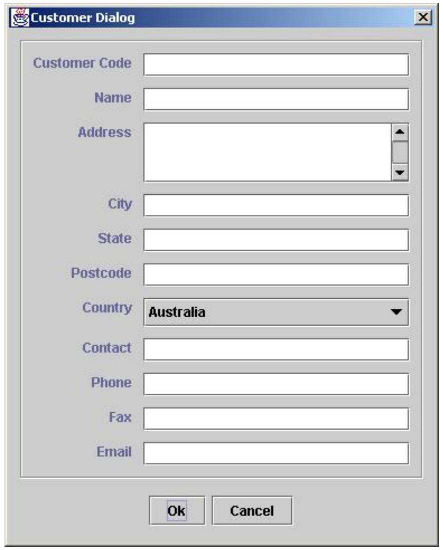
CheckBox



Slider



Rang de valors

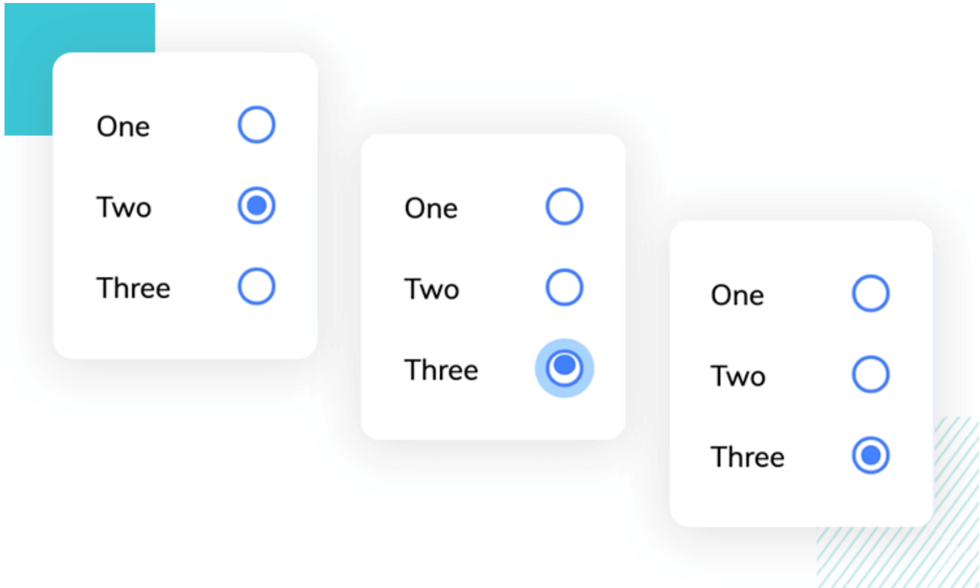


TextBox

2. Interactivitat

Tipus de controls (2):

Toggle



Radio button

2. Interactivitat

Tipus de controls (3):

2. Interactivitat

Factors que influeixen en incorporar tècniques d'interacció en projectes de Visualització de Dades:

- **Restriccions:** Tecnologies disponibles i competències que té l'observador o l'audiència.
- **Objectius:** Quina experiència volem crear i quines opcions interactives tenim?.
- **Representació de Dades:** Quines gràfiques poden necessitar interactivitat per a facilitar la interpretació.
- **Disseny fiable.**
- **Disseny accessible:** Minimitzar tecles o clicks de mouse per a una interactivitat no obstrussiva i útil.
- **Disseny elegant:** Interacció fluida i potenciar la jugabilitat, si possible.

2. Interactivitat

Categories d'interacció per a Visualització de Dades:

- **Filtrar:** Quines dades volem veure i quines no.
- **Seleccionar:** Ressaltar o trackejar dades.
- **Participar / Col·laborar:** Que usuaris puguin aportar dades al data set per a customitzar l'experiència.
- **Resumir:** Ajustar el nivell d'anotacions de les dades.
- **Observar i Explorar:** Vista general i poder ajustar el nivell d'abstracció i detall de les dades.
- **Connectar (*data linking*):** Relacionar dades en diferents gràfiques.
- **Repetir:** Refer o repetir exploració de les dades.
- **Extreure característiques:** Permetre a l'usuari extreure dades.
- **Reconfigurar:** Diferents presentacions de les dades.
- **Codificar:** Diferents representacions de les dades.
- **Jugar (gamificació):** Mostrar dades d'una forma jugable.

2.1. Filtrar

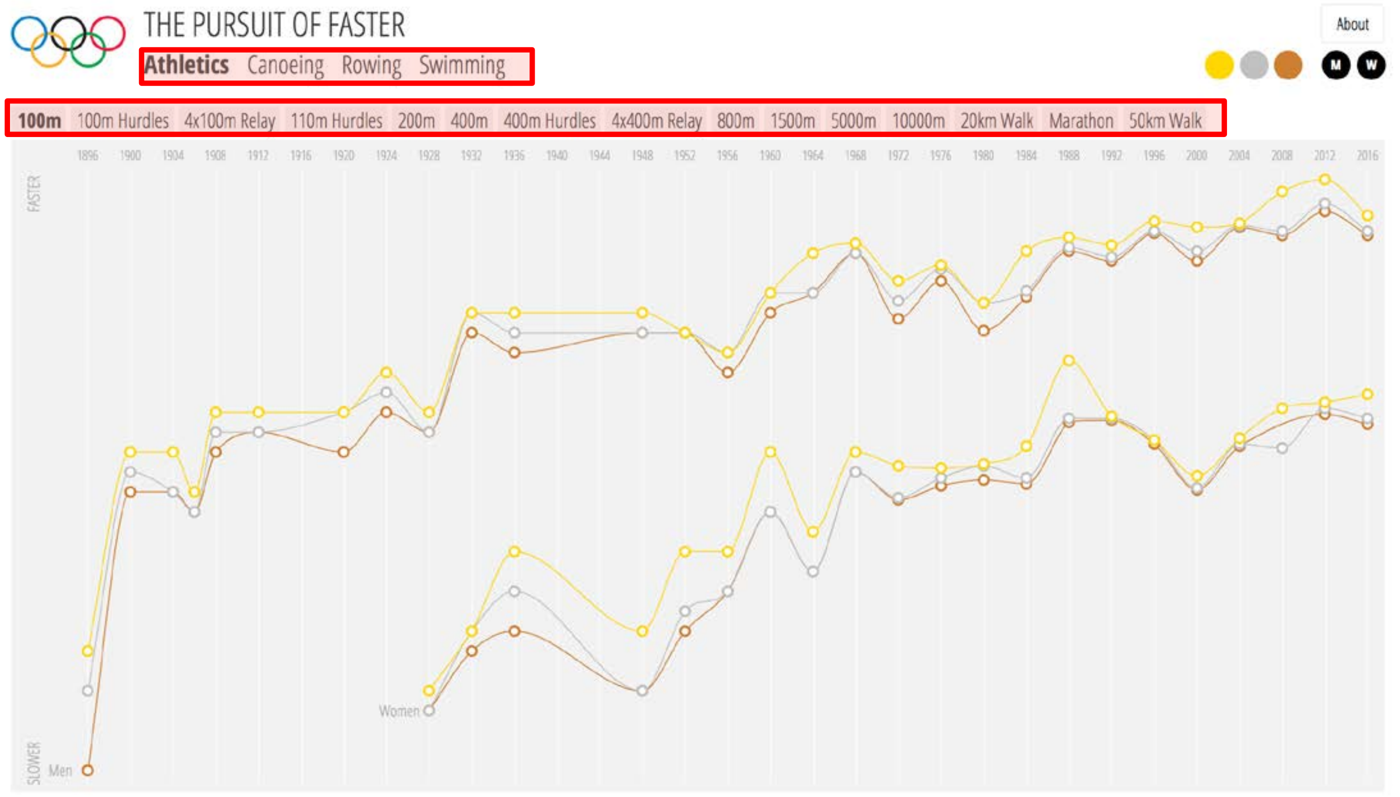
Fer possible que l'usuari pugui especificar quines dades vol veure, i quines no.

Es pot modificar el missatge a comunicar si donem control a l'usuari.

Events i controls	Accions
Seleccionar un botó	Aplicar filtre de dades per categories
Seleccionar item d'un pop up	Aplicar filtre de dades quantitatiu (rang de valors)
Seleccionar ítems d'un checkbox o menu	Reset de tots els valors per a una configuració inicial
Alterar commutador o slider	
Definir rang de valors	
Introduir valor en textbox	

2.1. Filtrar

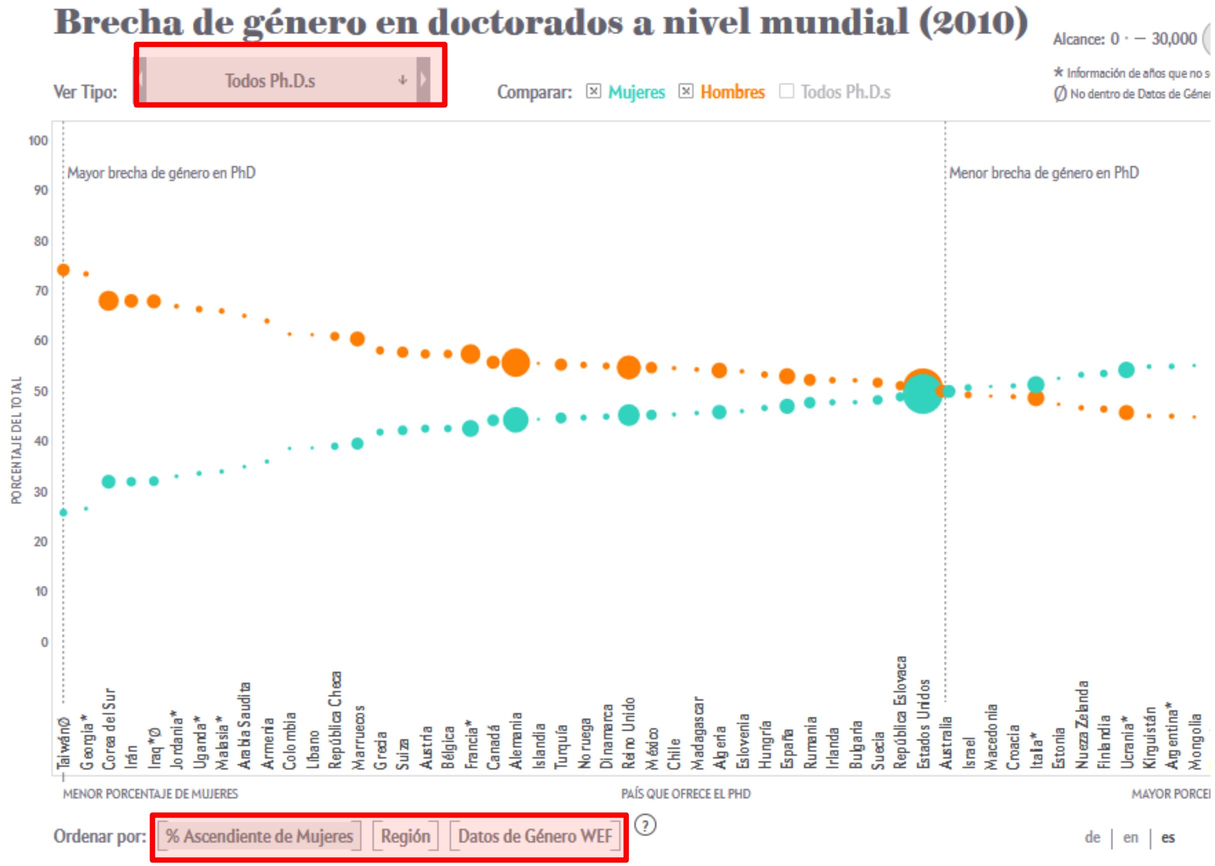
Exemple F1: Resultats en proves d'atletes de Jocs Olímpics esportius.



En línies superiors hi ha pestanyes per a seleccionar l'esport i prova en que es vol mostrar gràfiques (en negreta).

2.1. Filtrar

Exemple F2: Diferències de gènere en doctorats a nivell mundial.

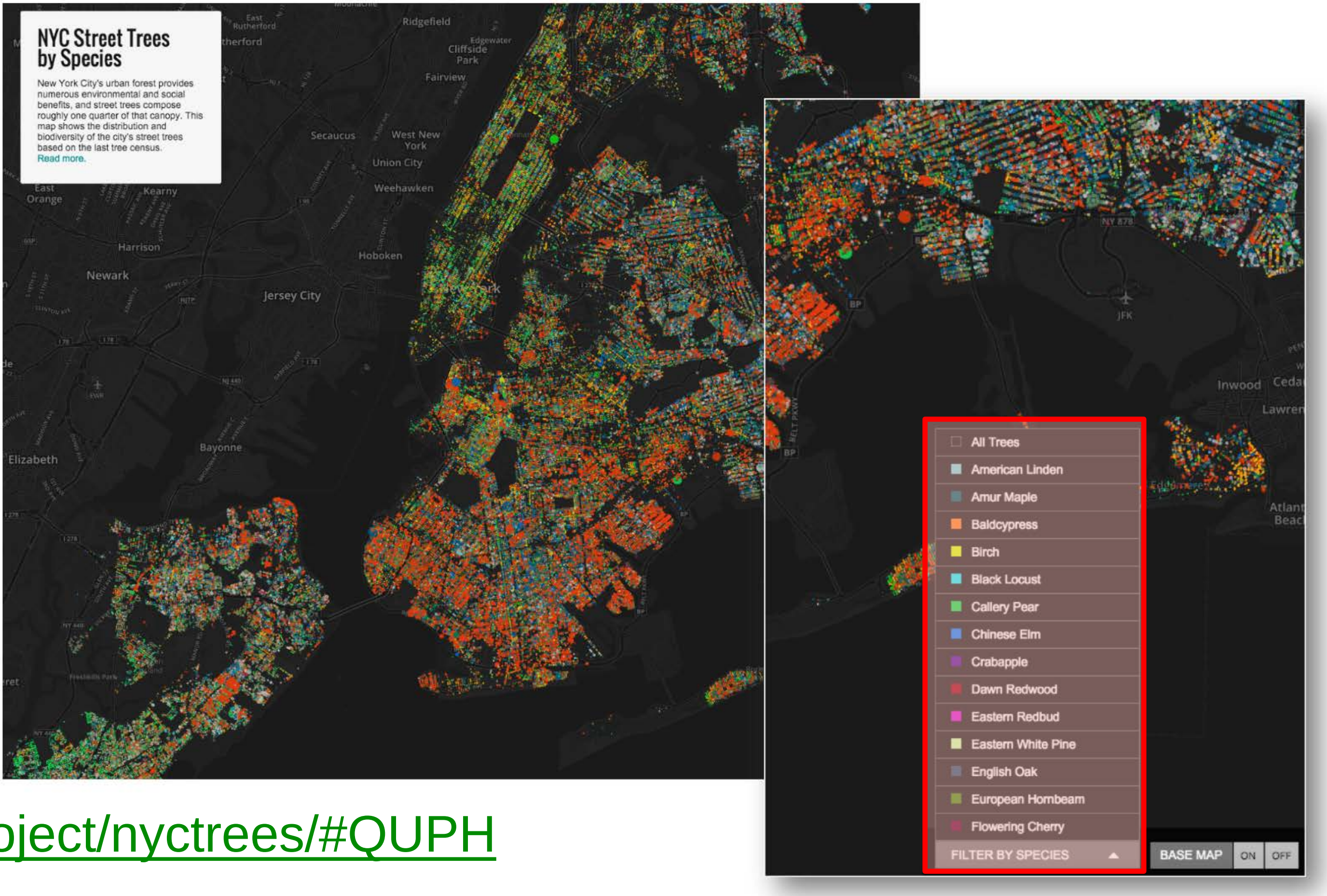


https://www.scientificamerican.com/article/how-nations-fare-in-phds-by-sex-interactive1/

Aplicació de filtres d'ordenació en la part superior i inferior de la pantalla.

2.1. Filtrar

Exemple F3: Espècies d'arbres per zones de color a NY. 52 espècies d'arbres.

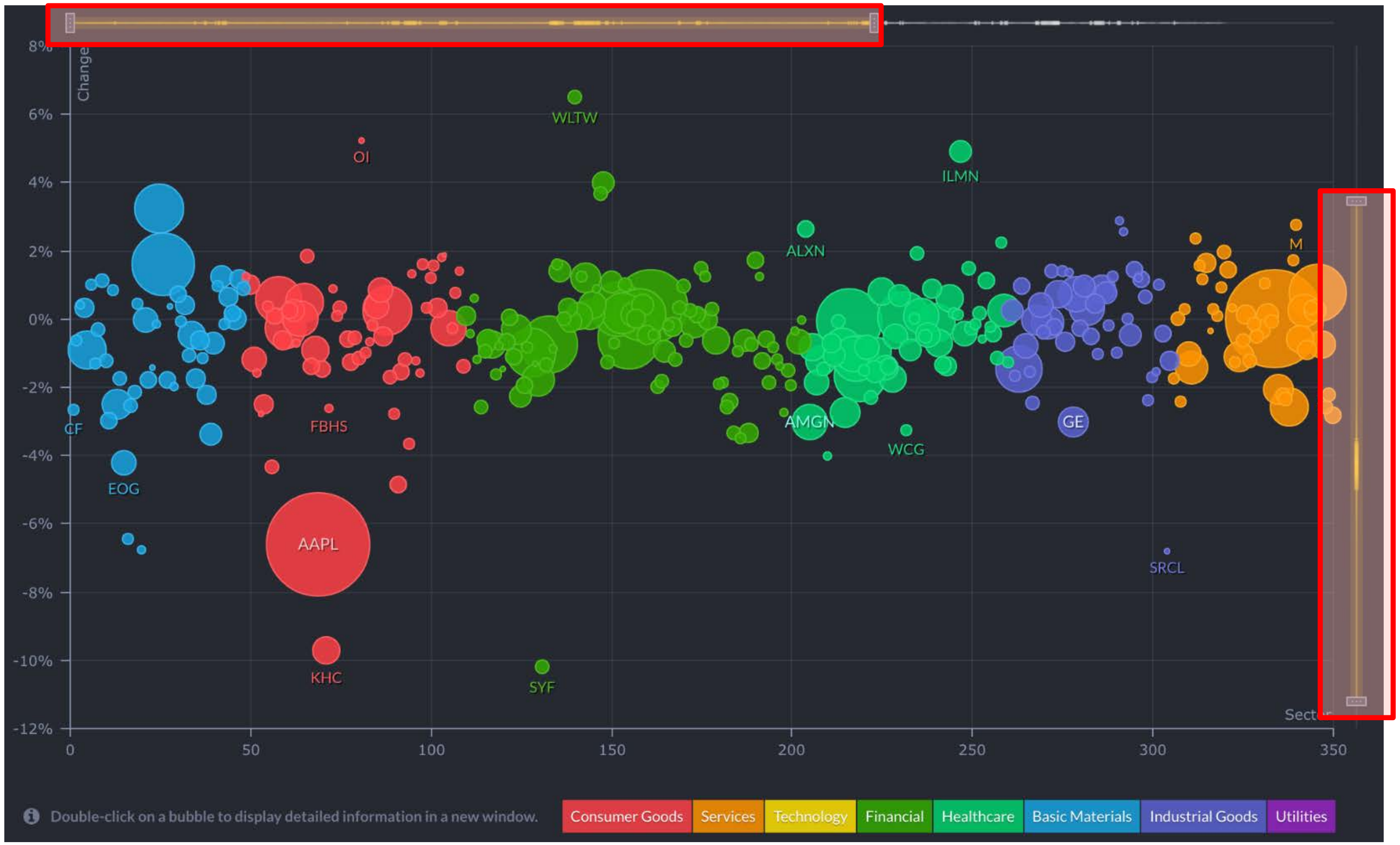


http://jillhubley.com/project/nyctrees/#QUPH

Pop up per a filtrar espècies a visualitzar.

2.1. Filtrar

Exemple F4: Anàlisi de mercats.



https://finviz.com/bubbles.ashx

Sliders per a modificar rangs de valors en eixos X, Y.

2.2. Seleccionar

Ressaltar valors d'interès en la representació, sense eliminar valors.

Es modifica la forma de veure les dades.

Es pot ressaltar:

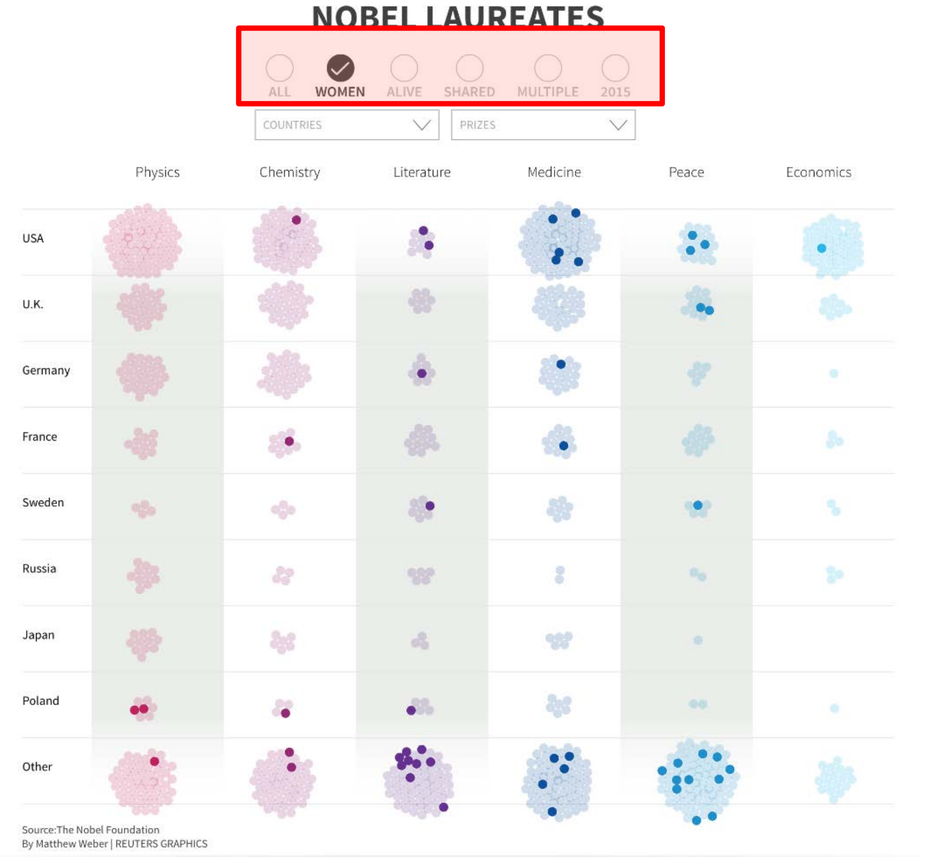
- Modificant atributs de color dels elements
- Reordenant les dades

2.2. Seleccionar

Events i controls	Accions
Seleccionar un botó o link	Ressaltar una selecció
Seleccionar opció de menú	Ressaltar valors associats a una selecció
Seleccionar ítems d’un checkbox o menú	Ressaltar associacions entre valors seleccionats
Seleccionar o alterar toggle o radio button	Reordenar dades
Alterar slider per un valor o rang de valors	Càlculs basats en una selecció
Teclejar valors dins un textbox	

2.2. Seleccionar

Exemple S1: Premis Nobel, per país i premi.



Botons per a seleccionar grups, ressaltats per punts.

2.2. Seleccionar

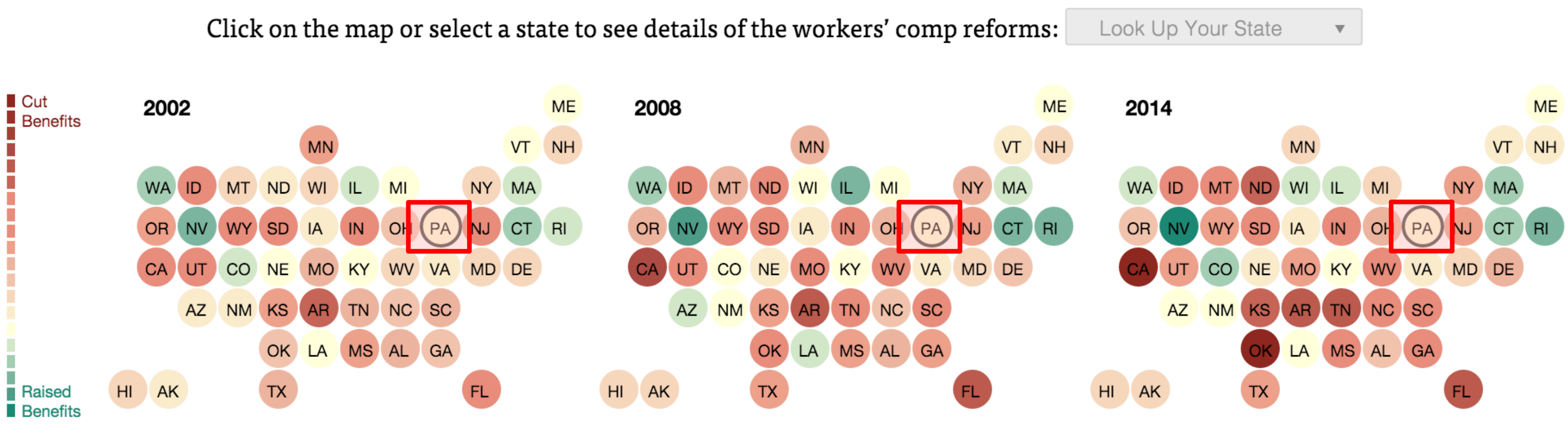
Exemple S2: 100 noms d'infants més populars al llarg dels anys en Anglaterra i Gales.

https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2017

Moltes gràfiques. Ressaltar un màxim de 6.

2.2. Seleccionar

Exemple S3: Compensació als treballadors per estats USA i anys.

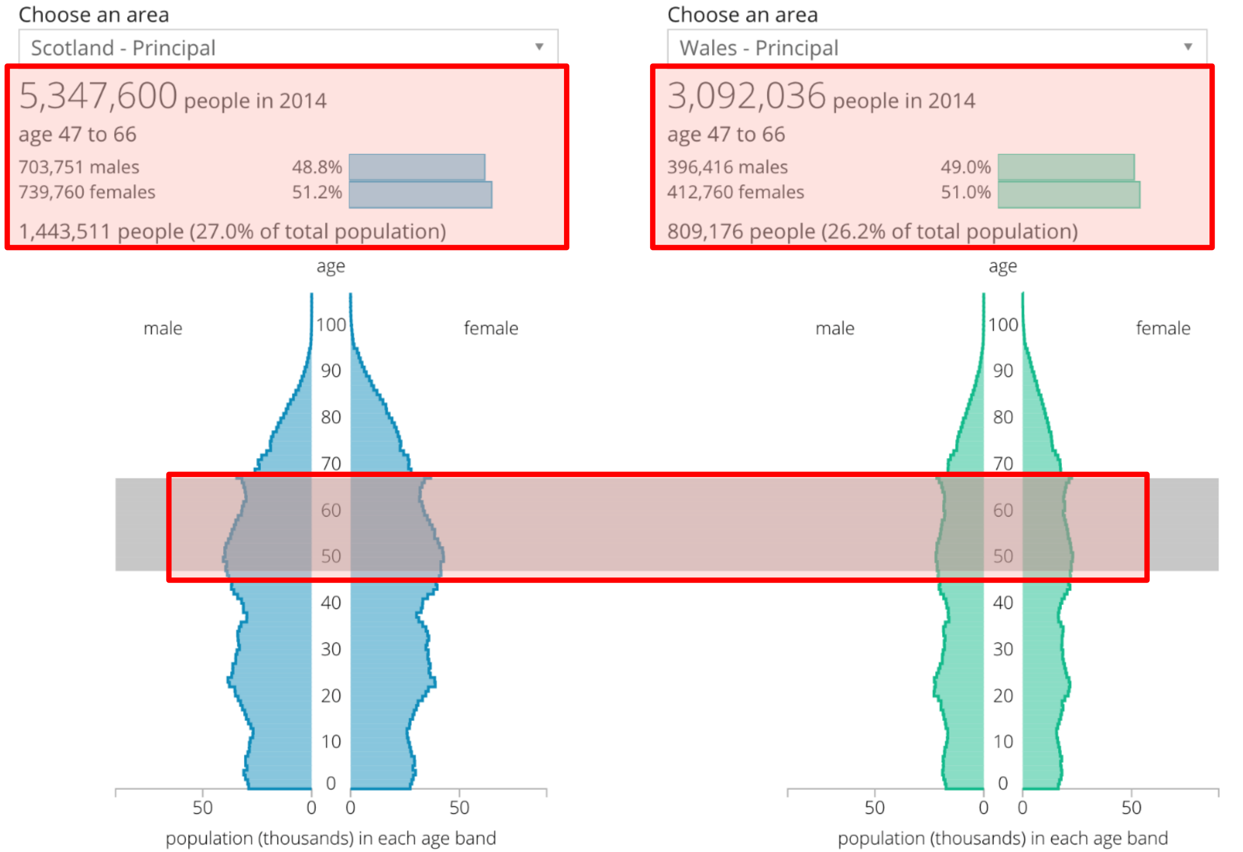


https://projects.propublica.org/graphics/workers-comp-reform-by-state

Data linking: Quan es selecciona una dada en una gràfica, es ressalta mateixa posició (estat) en les mostrades en pantalla.

2.2. Seleccionar

Exemple S4: Histogrames de cens a UK per nacions.

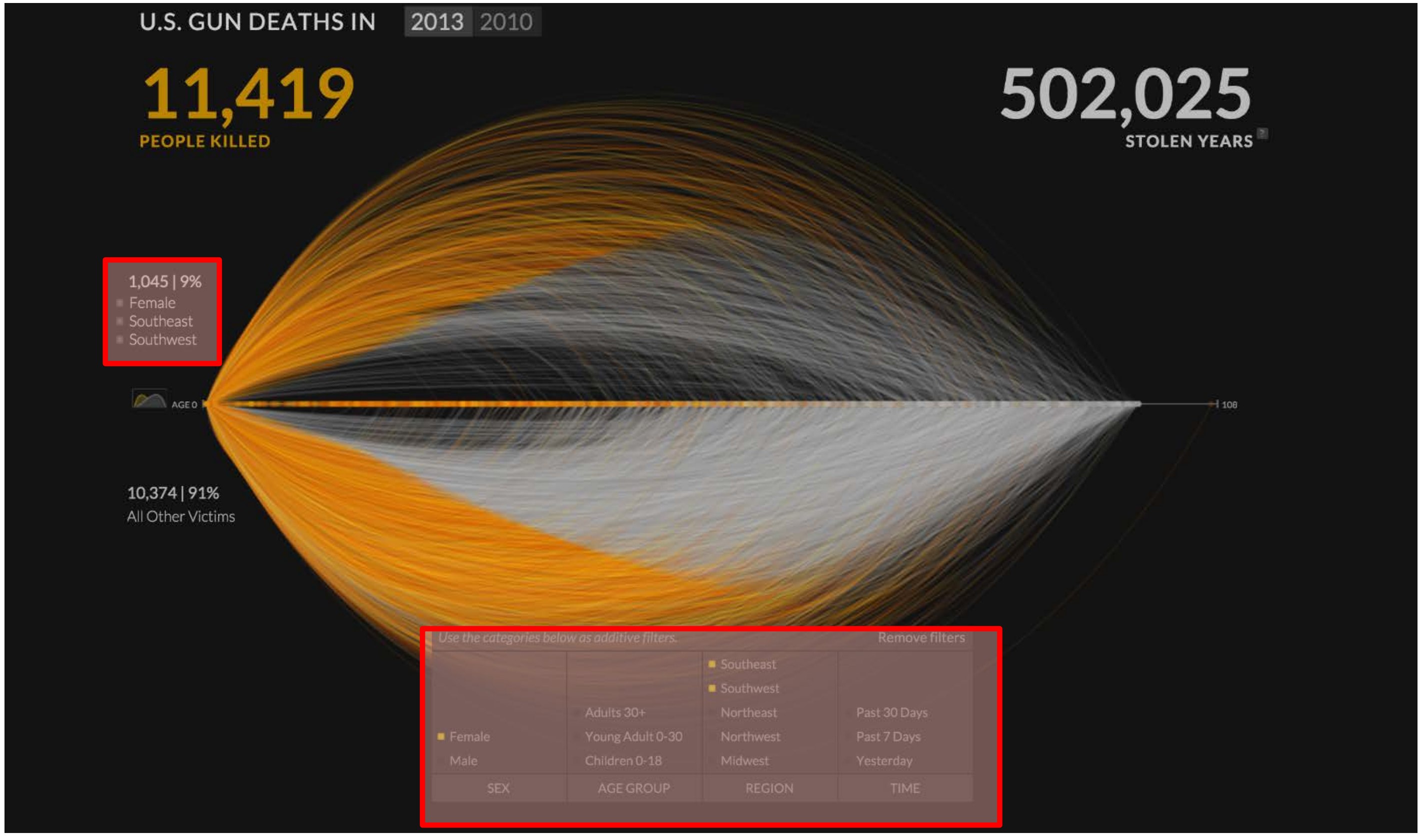


https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/articles/howbigwilltheukpopulationbein25yearstime/2015-10-29

Pintem (*brush*) un rang d'edats que es visualitzen en la part superior.

2.2. Seleccionar

Exemple S5: Morts per armes a USA.



https://guns.periscopic.com/?year=2018

Checkbox per a seleccionar diferents grups a ressaltar, mostrant a banda esquerra dades dels elements ressaltats.

2.2. Seleccionar

Exemple S6: Resultats bancaris.

Test results overview

Click on columns to sort and group overall results. Click on rows to select and compare specific banks.

BANK				ECB ADJUSTMENTS			NEED TO RAISE
Name	Country	Assets end of 2013 (€ bln.)	Ownership	Worst CET1 ratio over stressed scenario (%) Threshold:5.5%	**AQR adjustment (€ mil.)**	Basis points	Capital shortfall post net capital raised (€ mil.)
Monte dei Paschi	Italy	199.1	state (listed)	-0.1%	4,246.0	687	2,110.0
Piraeus	Greece	92.0	state (listed)	4.4%	2,792.0	558	0.0
National Bank of Greece	Greece	109.1	state (listed)	-0.4%	2,257.0	794	930.0
Rabobank	Netherlands	674.1	co-op	8.4%	2,093.0	367	0.0
Banco Popolare	Italy	126.5	state (listed)	4.7%	1,603.0	320	0.0
HSH Nordbank	Germany	109.3	state	6.1%	1,594.0	394	0.0
Commerzbank	Germany	561.4	state (listed)	8.0%	1,522.0	288	0.0
BCPE	France	1,065.4	state (listed)	7.0%	1,517.0	304	0.0

http://graphics.thomsonreuters.com/14/ecbtests/index.html

Fila explicativa de les columnes pot ser seleccionada per a fer ordenació creixent o decreixent de les files.

2.3. Participar

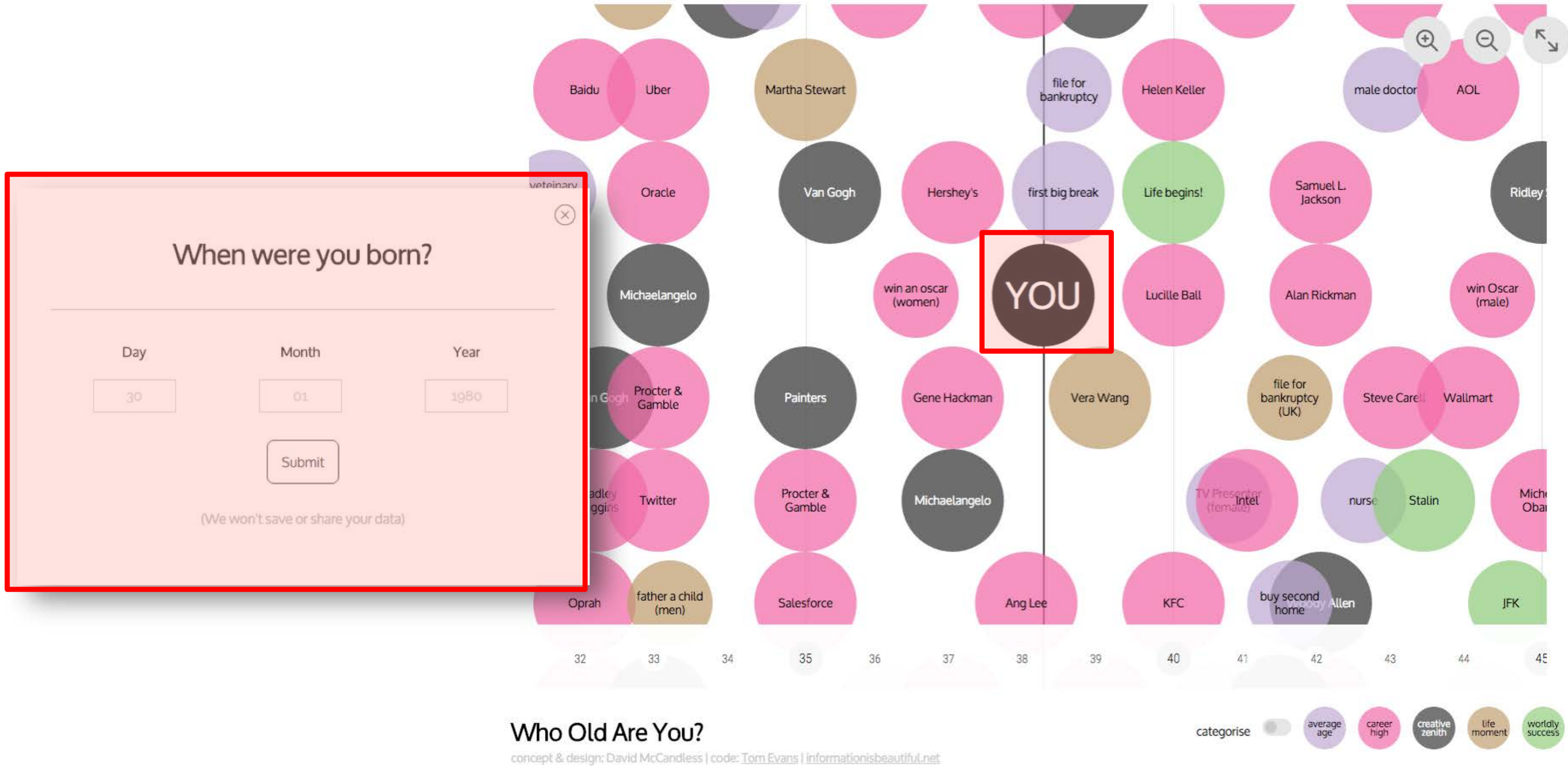
Usuaris prenen un rol més actiu aportant dades pròpies per a customitzar la seva representació visual.

Utilitzar tècniques interactives que facilitin la participació.

Events i controls	Accions
Seleccionar un botó o link	Donar dades per donar feedback (exemple: un quiz)
Seleccionar opció de menú	Donar dades per a customitzar una vista
Seleccionar ítems d'un checkbox o menú	
Seleccionar o alterar toggle o radio button	
Alterar slider per un valor o rang de valors	

2.3. Participar

Exemple P1: “Quina edat tens?”.

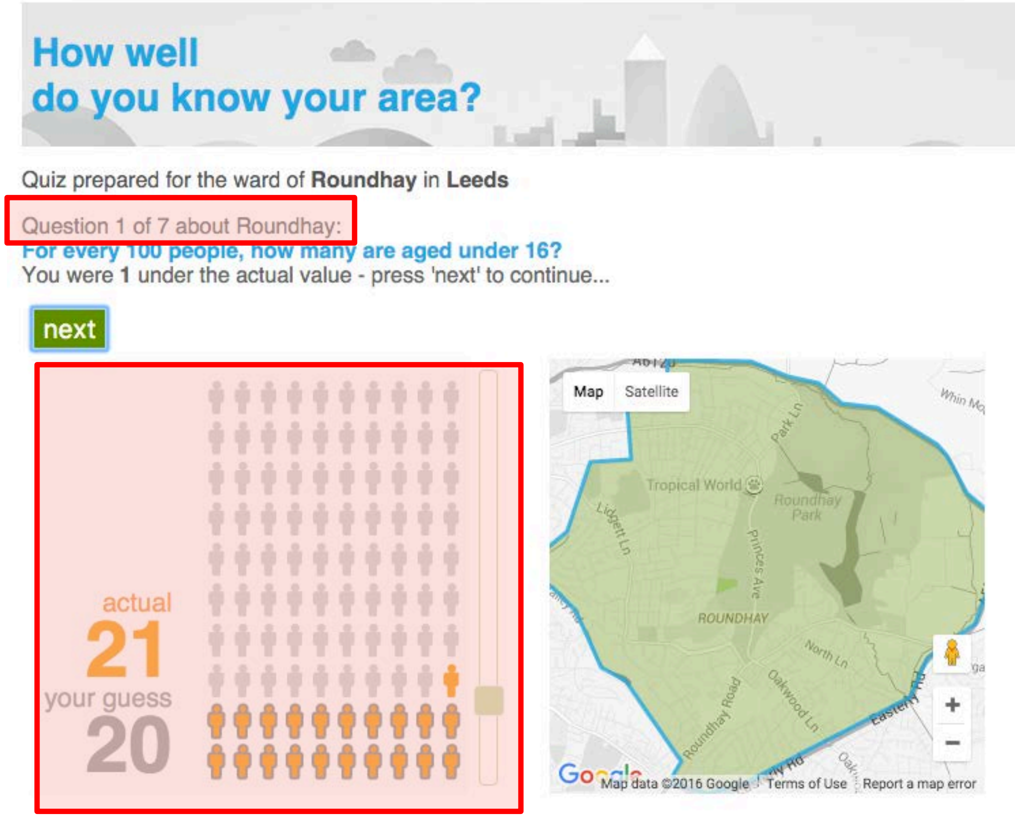


https://www.informationisbeautiful.net/visualizations/who-old-are-you/

Usuaris donen la seva data de naixement i es compara amb celebritats que a la mateixa edat han fet coses importants.

2.3. Participar

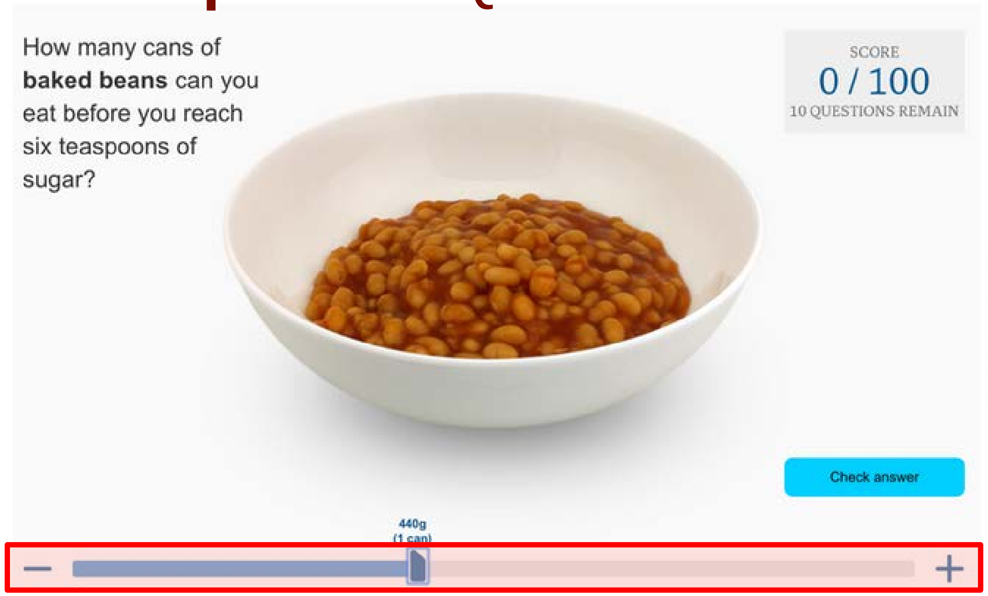
Exemple P2: “Com de bé coneix el seu barri?”.

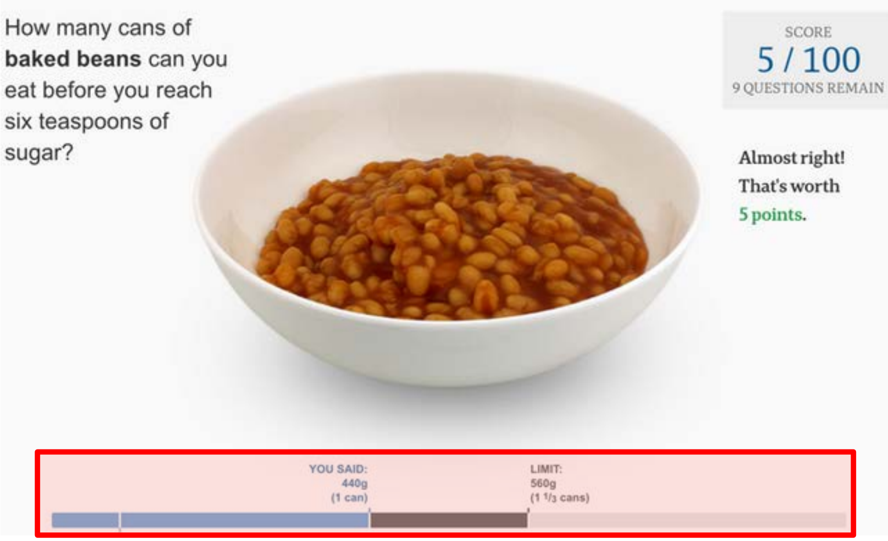


Usuari respon 7 preguntes sobre el seu barri segons codi postal. Segons respostes, es posiciona usuari en un ranking.

2.3. Participar

Exemple P3: “Quant de sucre trobem en el menjar?”.





https://www.abc.net.au/news/2016-03-17/quiz-what-does-six-teaspoons-of-sugar-look-like/7086790?nw=0

10 preguntes sobre sucre en el menjar. Usuari ha de donar les quantitats d’aliment que continguin 6 cullerades de sucre.

2.3. Participar

Exemple P4: “Recordes com estava dividida Alemanya?”.

https://interaktiv.morgenpost.de/innerdeutsche-grenze/

Usuaris han de traçar manualment la línia que separava les dues Alemanyes, que es compara amb la línia real i la feta per altres usuaris.

2.4. Resumir

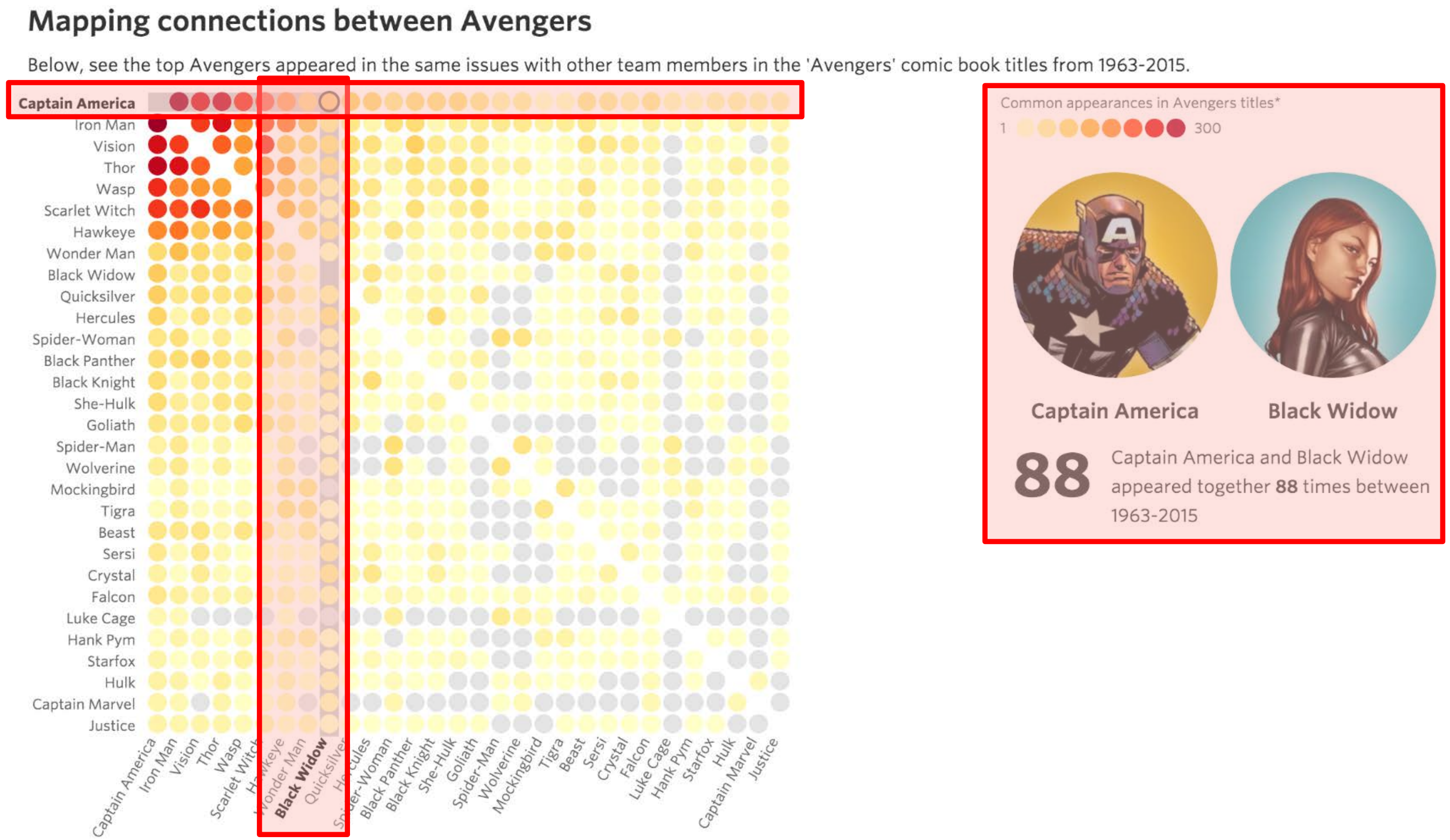
Proporcionar diferents nivells d'anotacions de les dades a demanda de l'usuari, que controla la percepció de les dades.

Obrir pop-ups d'opcions per a no destorbar dades que s'estiguin veient.

Events i controls	Accions
Seleccionar un botó o link	Mostrar anotacions en un pop up
Seleccionar una marca en una gràfica	Mostrar anotacions en un panell o finestra separada
Posar cursor sobre una marca en una gràfica	

2.4. Resumir

Exemple R1: *Heatmap* sobre crossovers de personatges d'Avengers.



Seleccionar una cel.la i apareix informació dels dos així com fila i columna.

2.4. Resumir

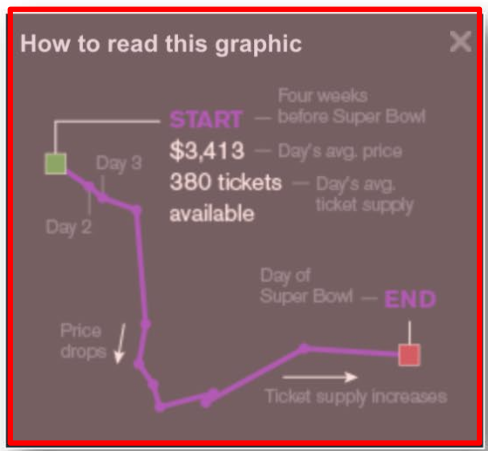
Exemple R2: Retòrica dels presidents USA en discurs estat de la nació.

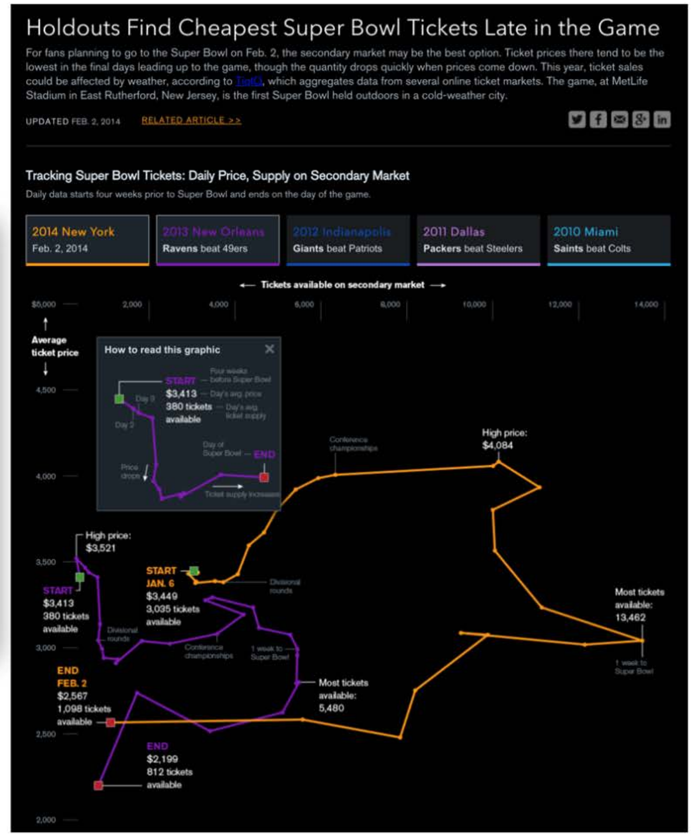
https://www.washingtonpost.com/graphics/politics/2016-sotu/language/

Eix de les X: Presidents USA ordenats per data. Col·locant cursor sobre cercle, ens dona dades quantitatives del president i la paraula dita.

2.4. Resumir

Exemple R3: Visualitzacions Bloomberg proporcionen popup de com llegir el gràfic.





https://www.bloomberg.com/graphics/infographics/tracking-super-bowl-ticket-prices.html

Connected Scatter Plots mostra el significat dels elements gràfics amb el mateix grafisme per a una major comprensió.

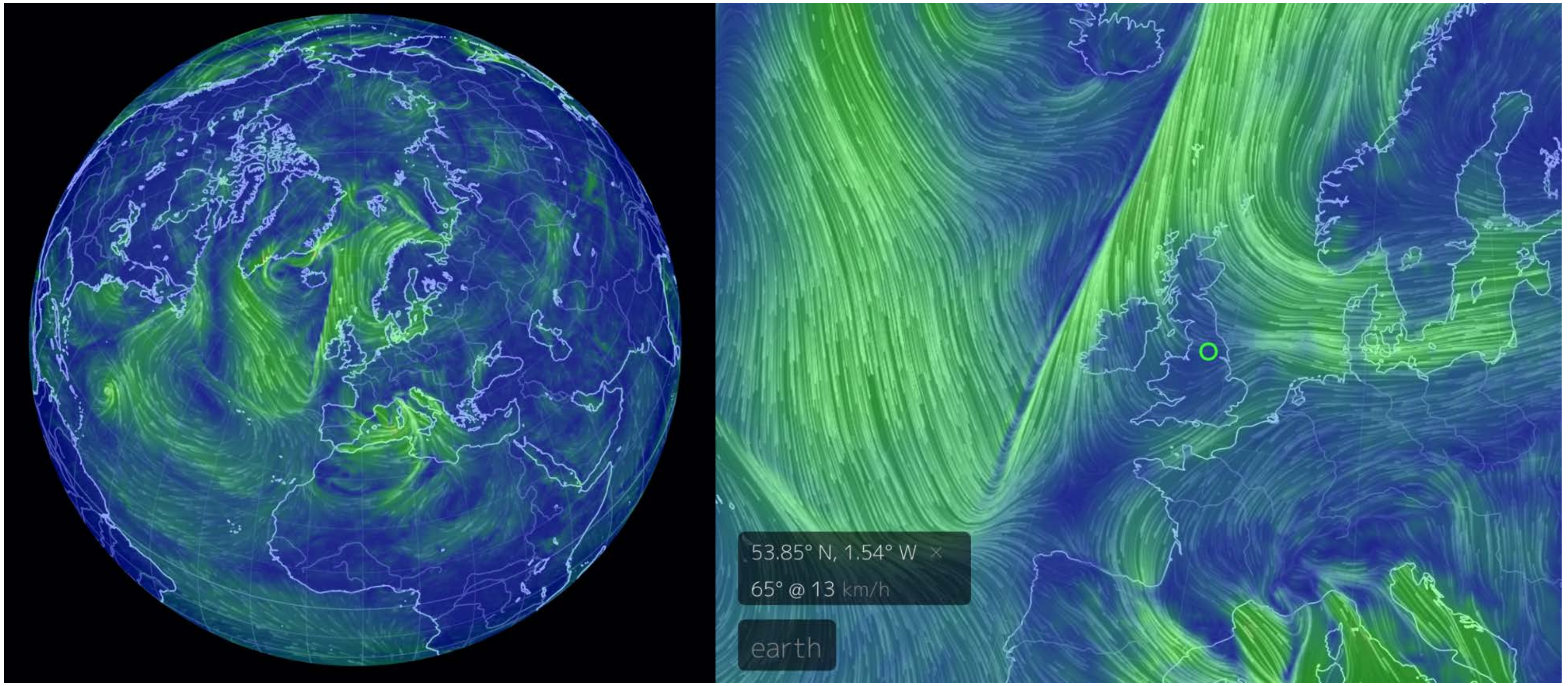
2.5. Observar i Explorar

Permet mostrar múltiples vistes o diferents nivells de detall de forma dinàmica, pel cas de que no es pugui mostrar tot junt en pantalla.

Events i controls	Accions
Seleccionar botó (nivell zoom)	Explorar a través seqüència de pantalles
Seleccionar opcions de tabulador	Explorar a través d'una mostra gradual de visualització
Scroll in o out	
Seleccionar una regió en mapa o menú	
Seleccionar, fixar i pintar una regió d'interès	
Seleccionar, fixar o moure	
Alterar posició d'escala amb slider	
Sideward scroll (només per a trackpads o mouse de Mac)	

2.5. Observar i Explorar

Exemple O1: “*Earth*”. Eina que mostra patrons de vent, clima i oceà en qualsevol lloc del planeta.

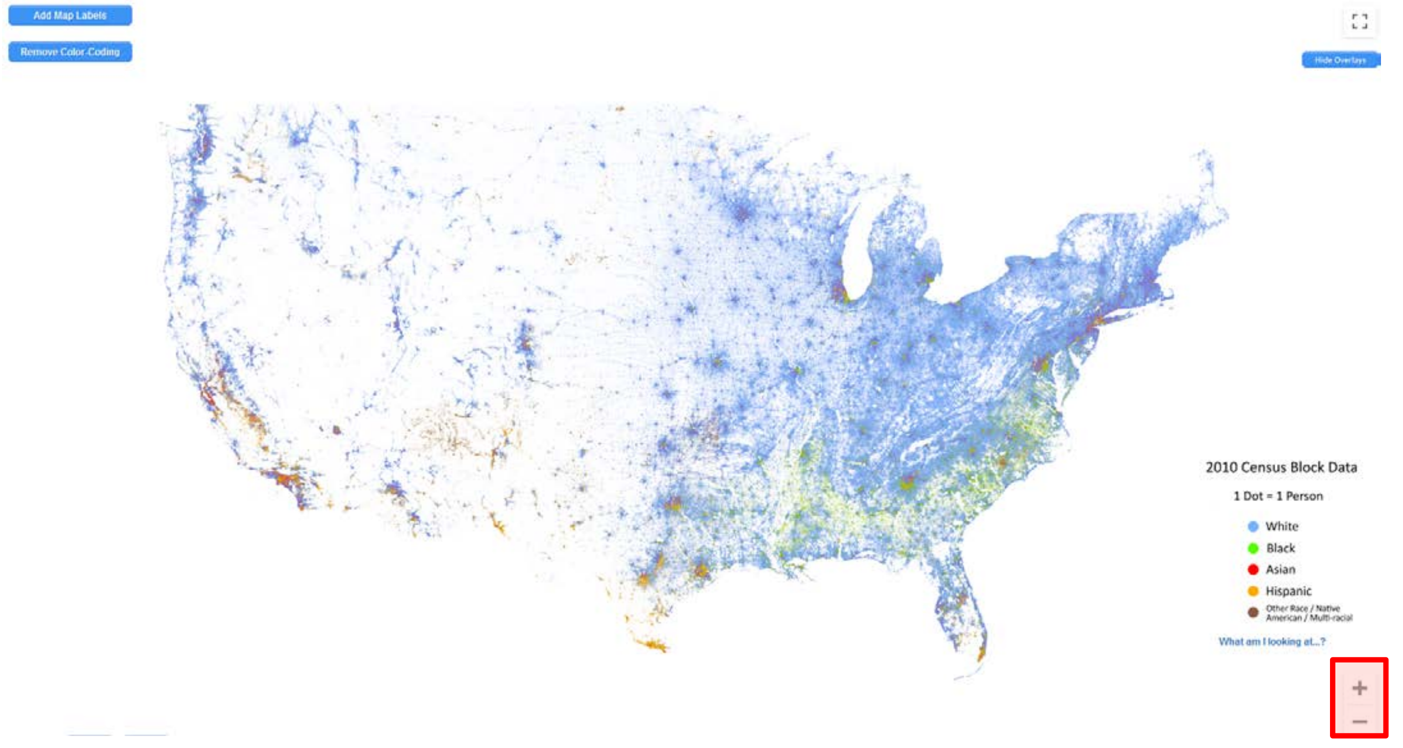


https://earth.nullschool.net/

Permet canviar escala (*zooming*) o desplaçar-se (*panning*).

2.5. Observar i Explorar

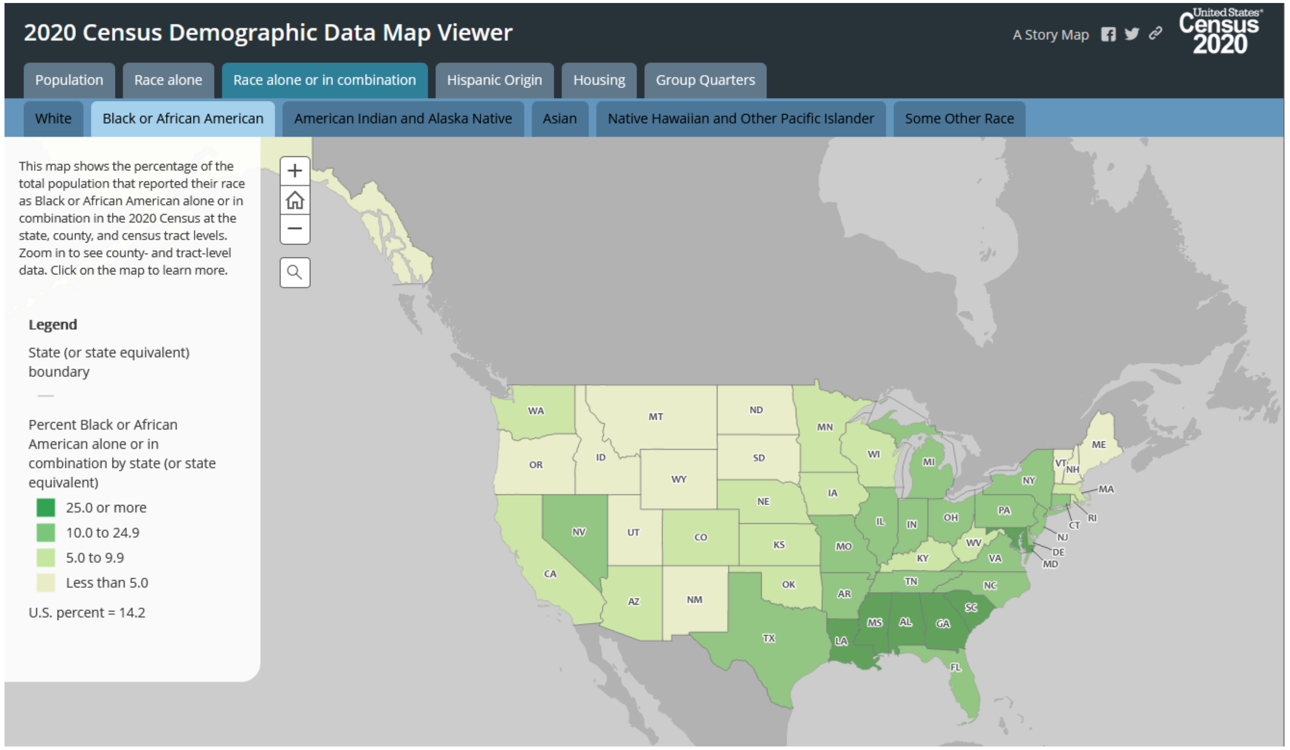
Exemple O2.1: "*Racial Dot Map*". Representació de la densitat de població als USA, amb codificació de color per a races i ètnies.



Permet *geometric zoom*, magnificant el detall de les dades en la zona seleccionada, i panning. **SUPRIMIT A PARTIR DEL GENER 2022.**

2.5. Observar i Explorar

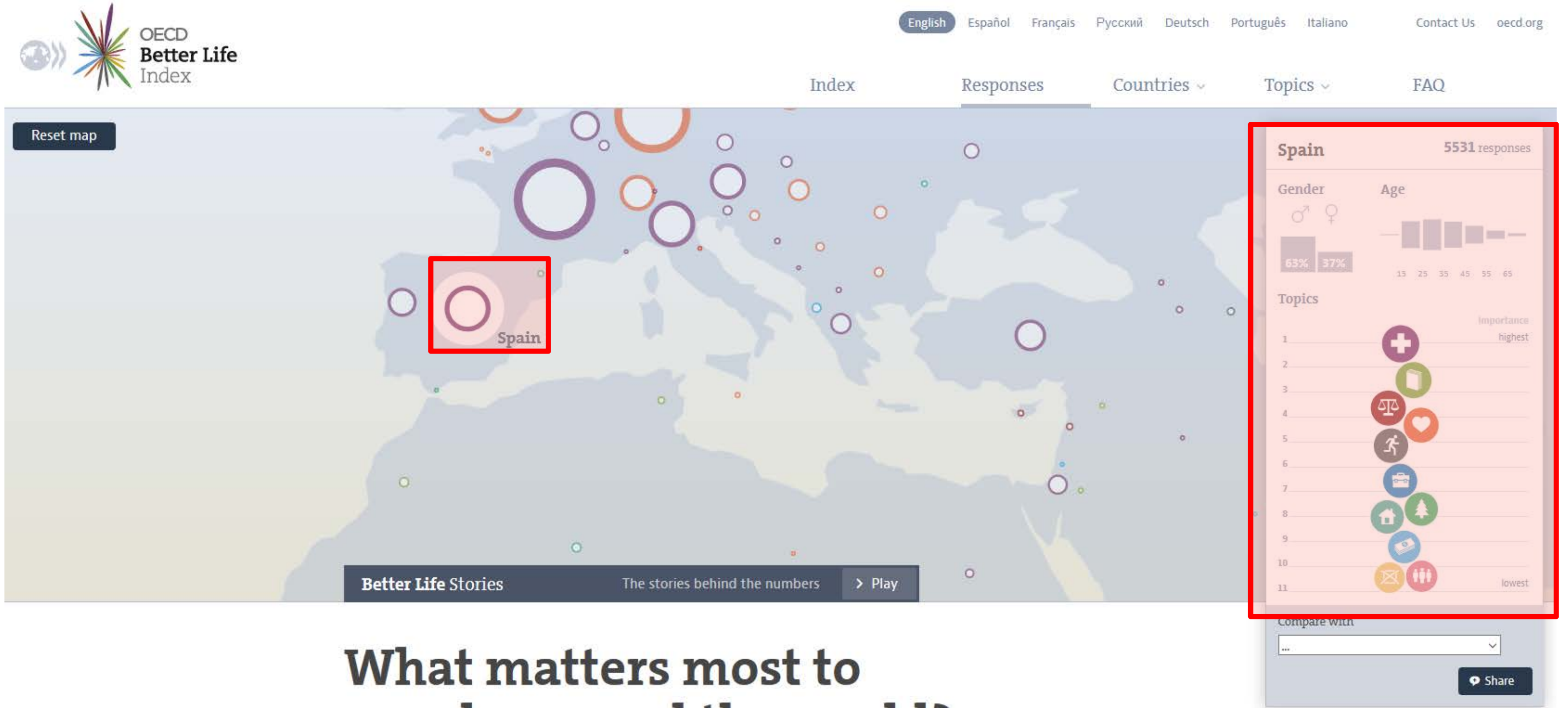
Exemple O2.2: “*2020 Census Demographic data Map Viewer in US*”. Representació de població, races i ètnies per estats i comtats amb codificació de color.



Permet *zoom i panning*, magnificant el detall de les dades en la zona seleccionada.

2.5. Observar i Explorar

Exemple O3: “OECD *Better Life Index*”. Mapa mundi amb rodones seleccionables de països per a mostrar informació.



http://www.oecdbetterlifeindex.org/responses/#

Drill-down: Tècnica que permet crear jerarquies de visualitzacions, de forma que usuari selecciona un element que mostra informació detallada.

2.5. Observar i Explorar

Exemples O1, O2 i O3 són exemples d'exploració per mapes, amb operacions de zoom i pan per a visualitzar a diferents nivells de detall i posició on es vulgui.

Alternativa: Oferir una experiència més lineal, a través d'un seguit de seqüències dins una narrativa.

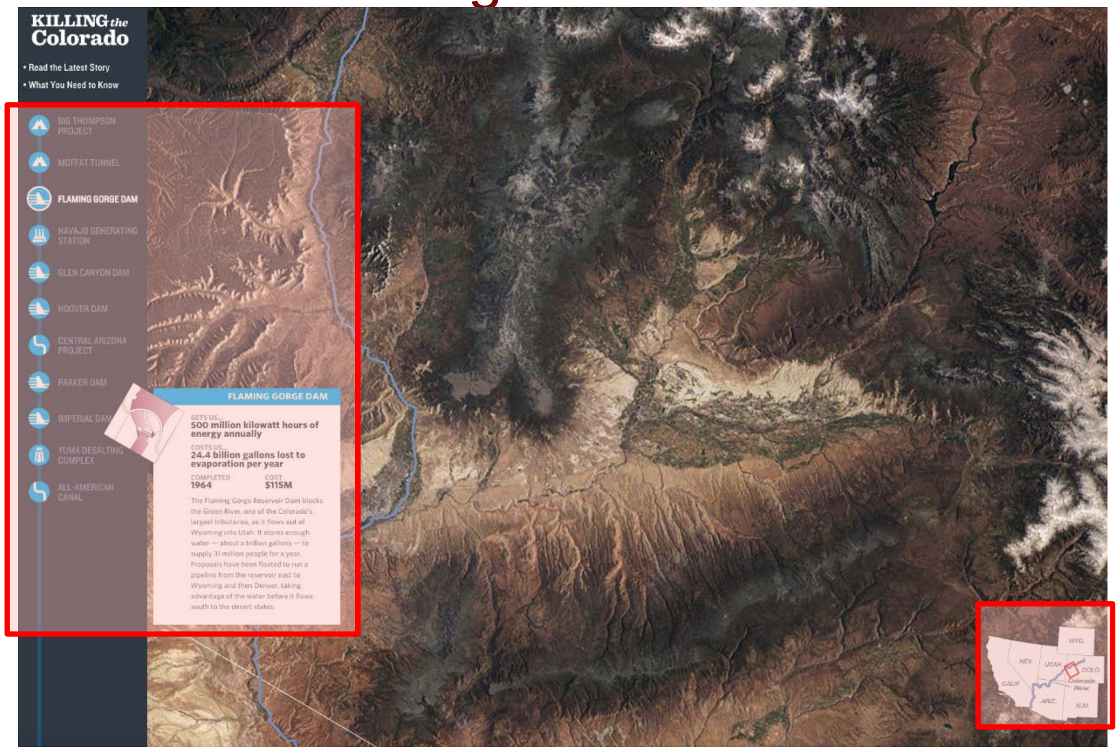
Storytelling: Usuari es desplaça endavant i endarrera dins la narrativa.

Explicar una història facilita comprensió.

Següents exemples són de *storytelling*.

2.5. Observar i Explorar

Exemple O4: "*Killing the Colorado*". Viatge pas a pas pel riu Colorado per a mostrar la degradació del riu.

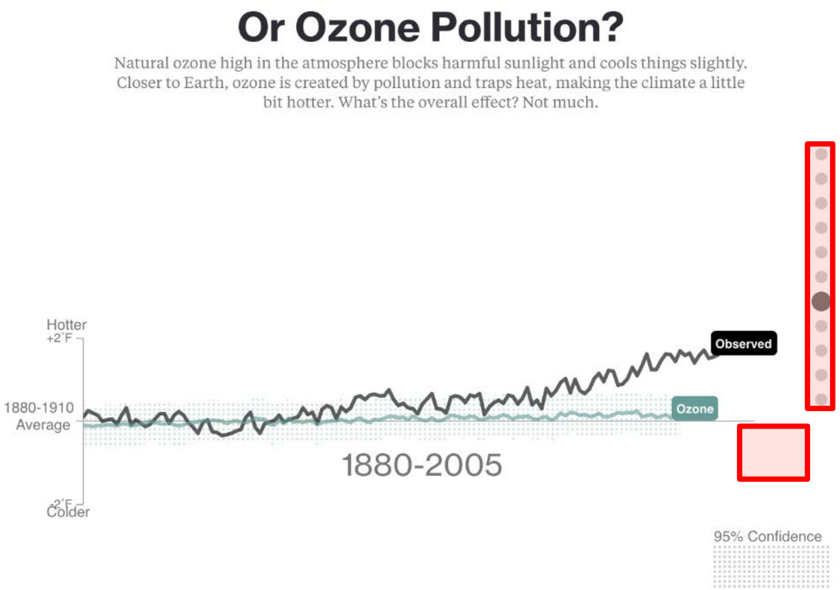


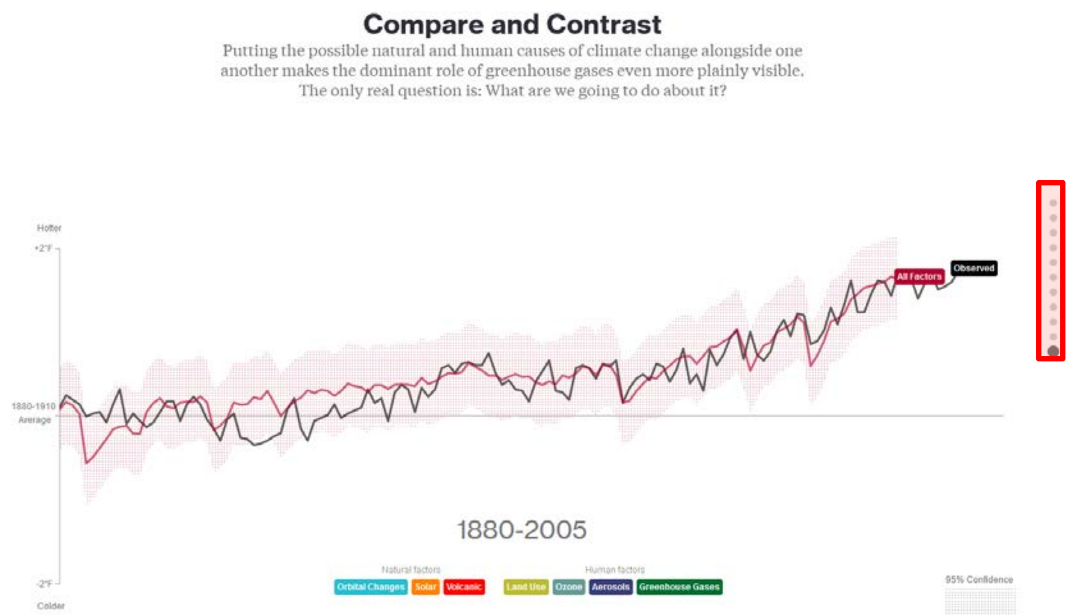
https://projects.propublica.org/killing-the-colorado/explore-the-river#big-thompson-project

Exploració lineal o lliure d'infraestructures fetes vora el riu (columna esquerra) que el degraden. En cada infraestructura hi ha un text explicatiu.

2.5. Observar i Explorar

Exemple O5: “*What’s Really Warming the World?*”. Presenta una seqüència de possibles causes del canvi climàtic.





Scrolling amunt i avall mostra diferents factors naturals o humans que afecten el canvi climàtic mostrant dades i pintant canvis de temperatura.

3. Animació

Seqüència d'imatges amb canvis entre imatges consecutives al llarg de l'espai i temps, creant il·lusió de moviment al cervell.

Genera un impacte visual molt fort en Visualització de Dades (VD), però no a tots els usuaris agrada.

Animació en VD:

- Forma d'incrementar la dimensionalitat de les dades (temps)
- Aporta informacions diferents a gràfiques estàtiques
- Qualitat estètica
- Important la durada entre *frames*: Massa lenta pot ser avorrida, massa ràpida poden perdre's detalls

però...

- pot distreure missatge a transmetre a usuari
- pot portar desafecció a l'usuari

3. Animació

Animació mostra dades amb dimensió temporal de forma dinàmica. Permet apreciar patrons que no es poden veure en imatges estàtiques.

Interacció en animació: Que usuari tingui control per a iniciar i parar l'animació quan vulgui.

Events i controls	Accions
Carregar pàgina web	Iniciar animació automàticament
Seleccionar botons animació (play, pausa, stop, repeat, etc.)	Iniciar animació per botons
Alterar slider	Controlar animació (frames) per slider

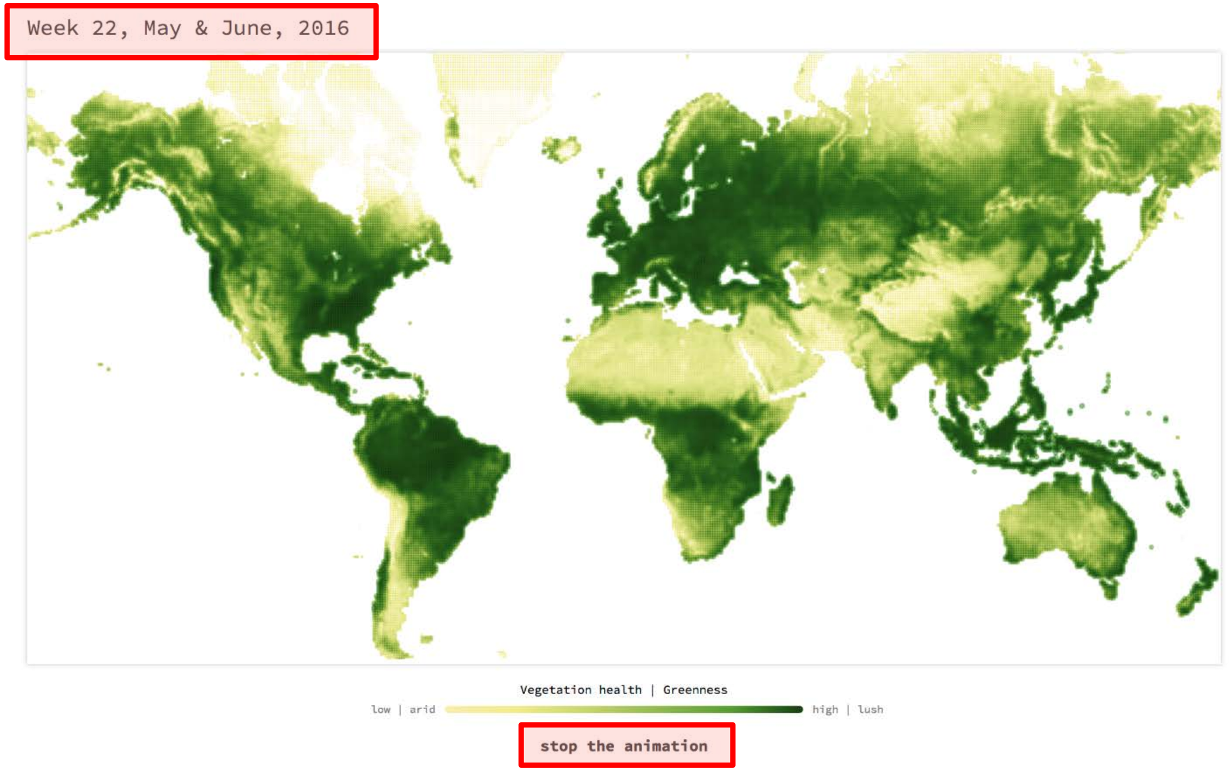
3. Animació

Per a què volem utilitzar l’animació en VD?

- **Coneixement:** Aportar informació contextual fora de l’àrea específica d’atenció
- **Transició:** Guia l’usuari entre diferents gràfiques
- **Descripció funcional:** Relacionat amb el comportament que un objecte animat representa (exemple: *scrolling*)
- **Èmfasi:** Cridar l’atenció sobre un element visible o procés concret.
- **Expressió:** Per a enriquir l’experiència d’usuari amb l’aplicació o tasca a fer
- **Canvi:** Veure com un objecte o procés evoluciona amb el temps
- **Visualització directe:** Correspondències entre atributs de moviment (fase, freqüència) amb variables del data set
- **Associació:** Utilitzar seqüències de moviment per a establir relacions entre grups d’informació

3. Animació

Exemple A1: "*Breathing Earth*". Simulació de les zones de vegetació del planeta al llarg dels anys.

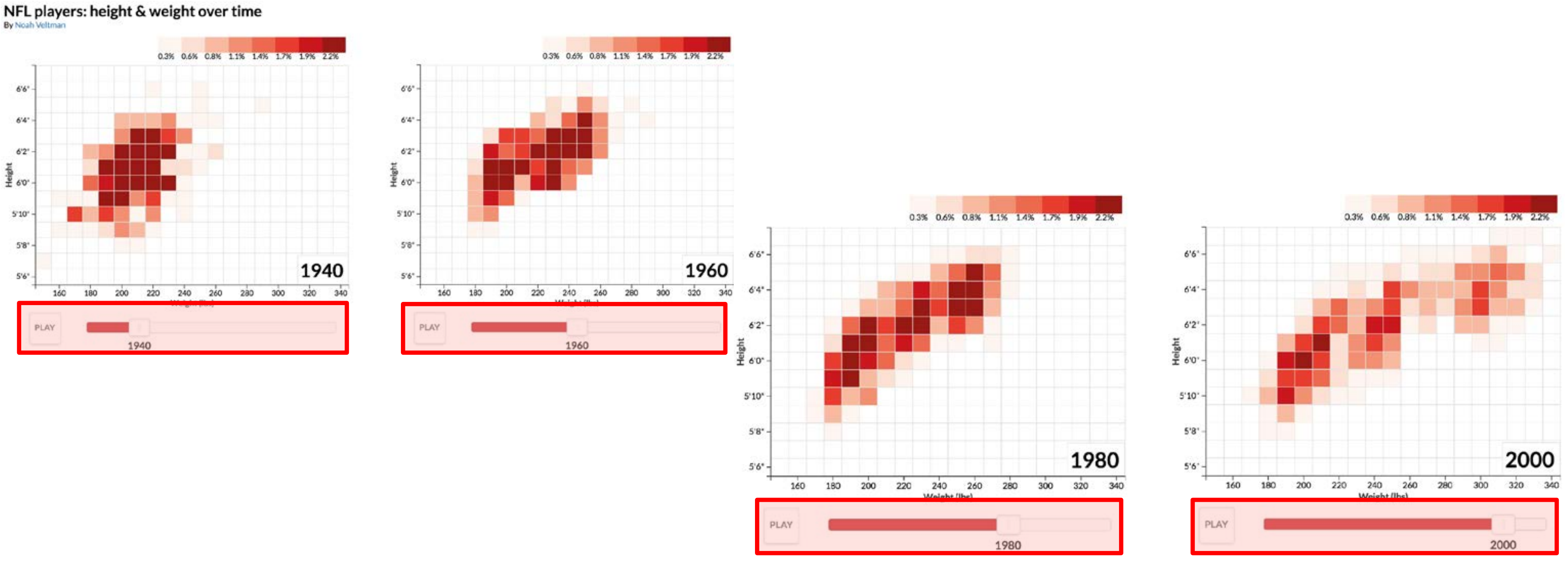


https://breathingearth.visualcinnamon.com/

Animació en loop, en que quan es veu de forma repetida es poden apreciar nous detalls. Usuari pot parar i continuar animació.

3. Animació

Exemple A2: Distribució de l'alçada i pes dels jugadors de la NFL en el temps, mitjançant un *heatmap*.

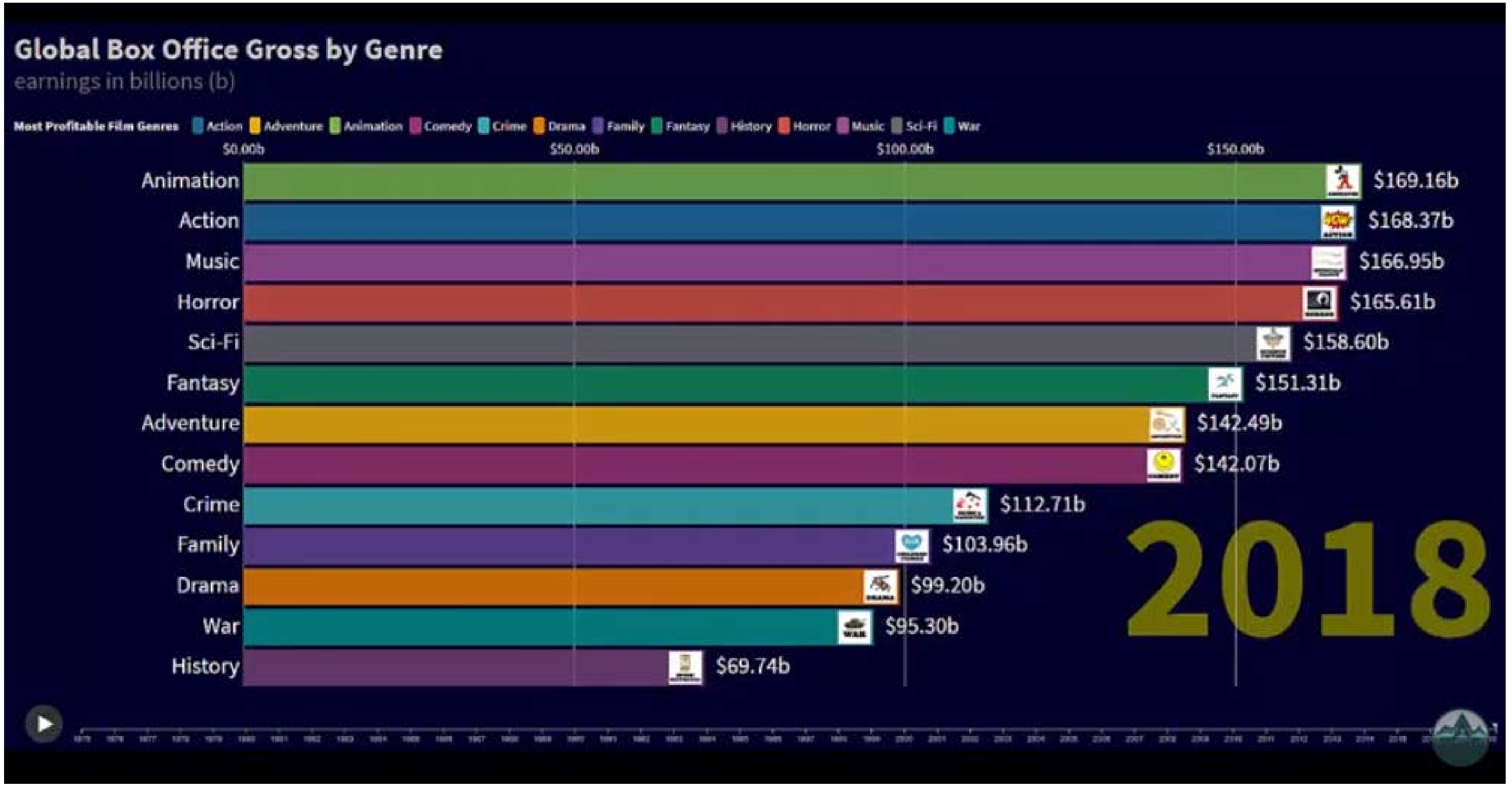


https://noahveltman.com/nflplayers/

Usuari té el control de l'animació (parar o engegar) amb el botó de Play o Stop. Slider mostra els anys.

3. Animació

Exemple A3: Ranking de diagrama de barres animades



https://www.youtube.com/watch?v=GPm9Yzvt25c

Usuari pot tenir control de l'animació (parar o engegar) amb slider.

4. Factors per a implementar la interactivitat

Important tenir en compte factors que poden determinar quines tècniques interactives es poden necessitar i implementar:

1. Restricció
2. Entregables
3. Objectiu
4. Representació de dades
5. Disseny fiable
6. Disseny accessible
7. Disseny elegant

4.1. Restricció

Principal factor per utilitzar interactivitat en els projectes de VD és tenir les competències tècniques i disposar de la tecnologia per a implementar-la.

Si no es tenen, millor no utilitzar la interactivitat per a no comprometre dates d'entrega del projecte.

Ponderar la capacitat que poden aportar les solucions interactives respecte el temps que es disposi per a implementar-les, es tinguin o no competències tecnològiques.

4.2. Entregables

Tenir clar com han de ser els entregables del projecte.

Digital no significa interactiu.

Qüestions principals:

- Audiència té el temps, paciència i capacitat per a utilitzar tècniques interactives?.
- Tipus de dispositius en que ha de funcionar el projecte (*smartphones*, *tablets*, pc's, etc.).
- Nous dispositius defineixen nous paradigmes d'interacció:
  - Mouse, trackball
  - Dispositius tàctils
  - Dispositius naturals no tàctils (RV, Leap Motion, Kinect, etc.)

4.2. Entregables

Comparativa entre tipus d'events en mouse o trackball i dispositius tàctil:

Mouse o trackball	Dispositius tàctils
Botó esquerre o dret	Toc (*tap*) amb un dit o dos dits
Doble click	Doble toc
Click, drag & drop	Toc, arrossegar i aixecar dits
Apuntar	Toc
Scroll de roda	Lliscar, pessigar
Controls de teclat	Rotar o teclat virtual

4.3. Objectiu

No totes les tècniques interactives ofereixen experiències exploratòries.

L'experiència que es vol oferir a l'observador o audiència definirà les tècniques d'interactivitat necessàries.

Mitjançant la interactivitat, es pot proporcionar a l'observador eines d'exploració que el facin interrogar-se sobre les dades a veure per a potenciar certa curiositat pel tema.

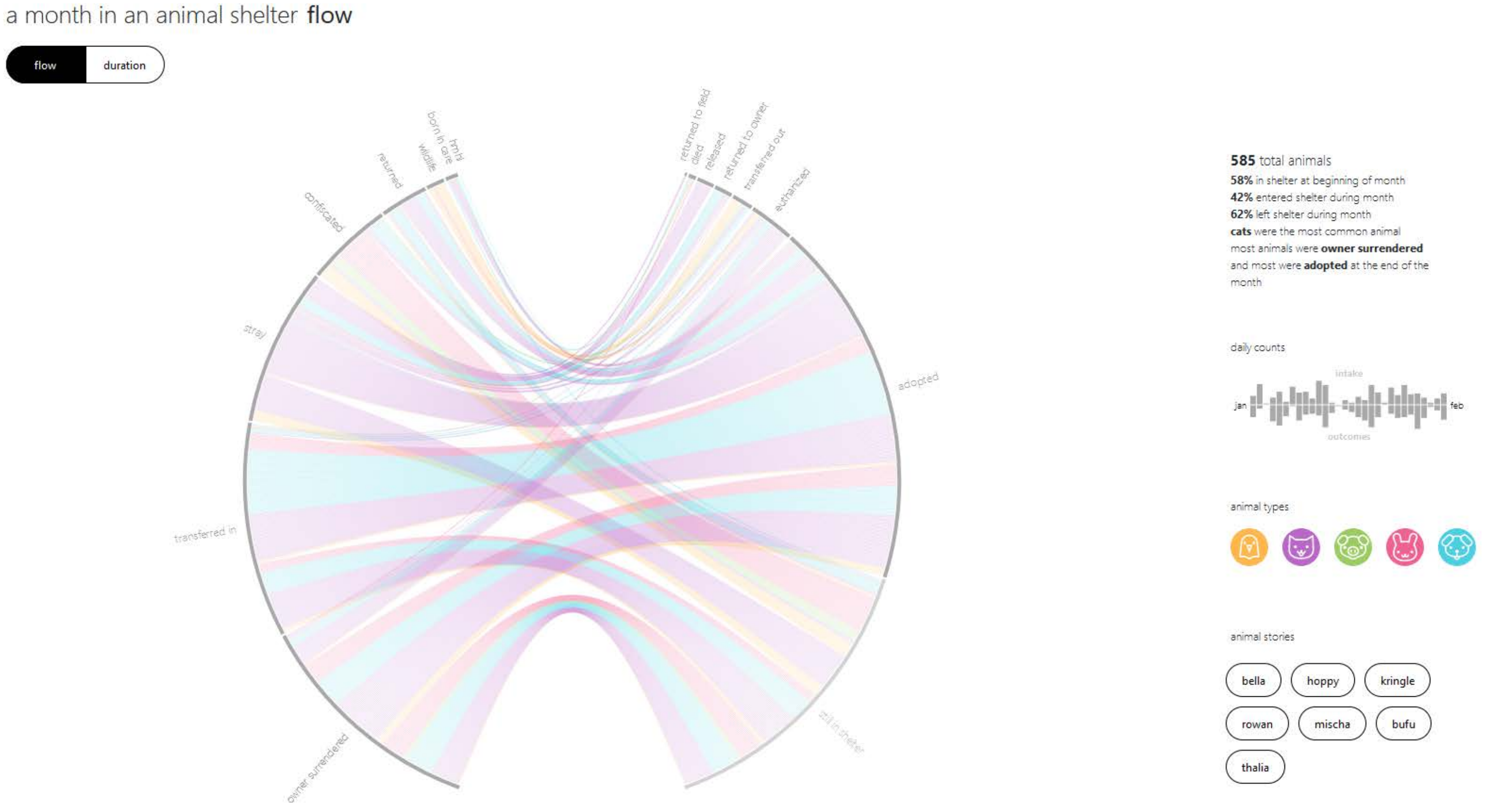
4.4. Representació de Dades

Algunes gràfiques són visualment complexes (molts elements, línies, distribució elements, etc.) i poden dificultar comprensió.

Tècniques interactives poden permetre filtrar o ressaltar certes dades que ajudin a fer més fàcil i interessant la cerca d'informació, la usabilitat de la informació mostrada i la implicació de l'usuari amb el que se li mostra.

4.4. Representació de Dades

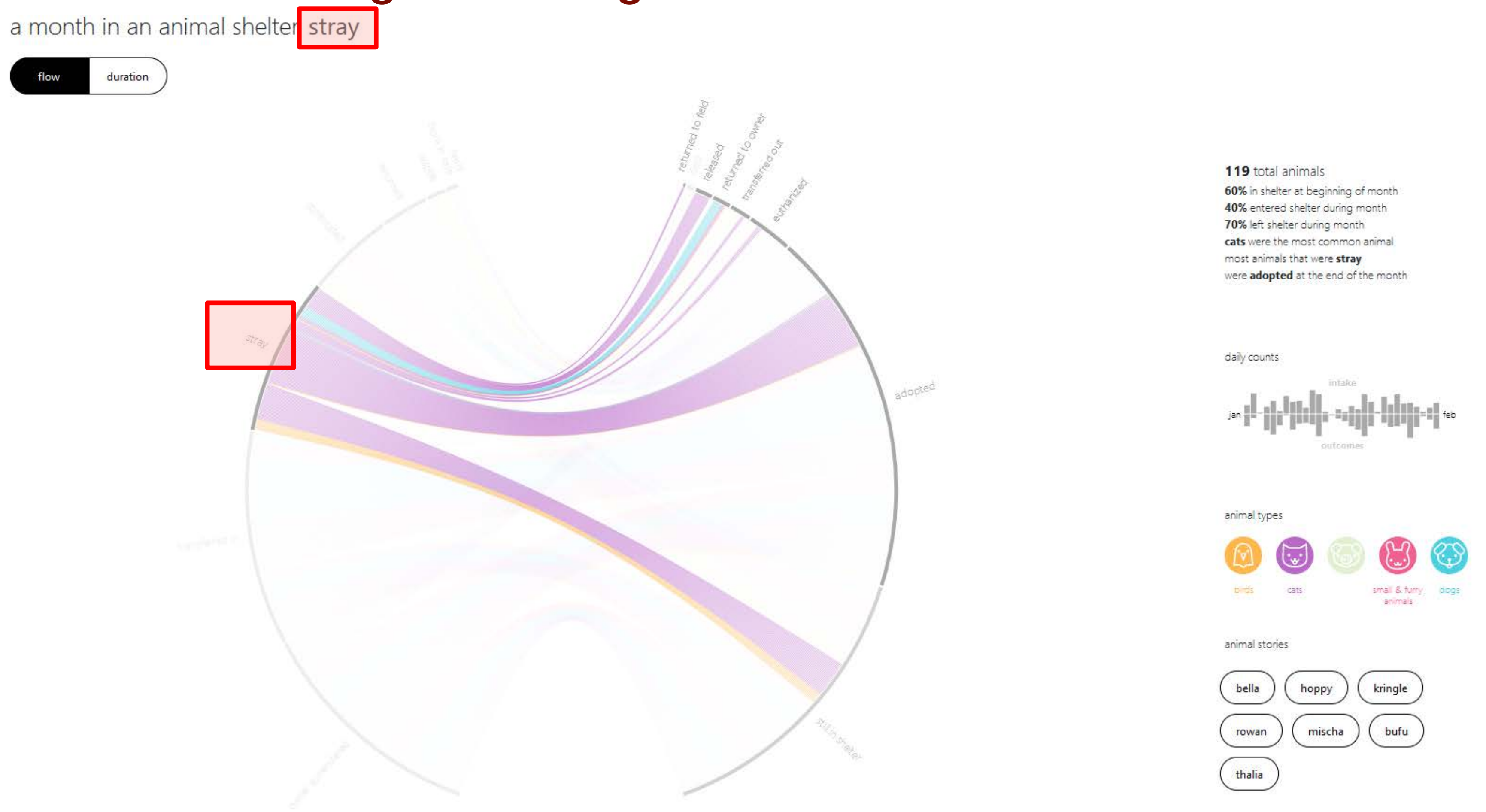
Exemple: "*A month in an animal shelter*". Es mostren les històries dels gossos que entren al refugi en un mes.



https://sarahsoup.github.io/shelter-animals/dist/

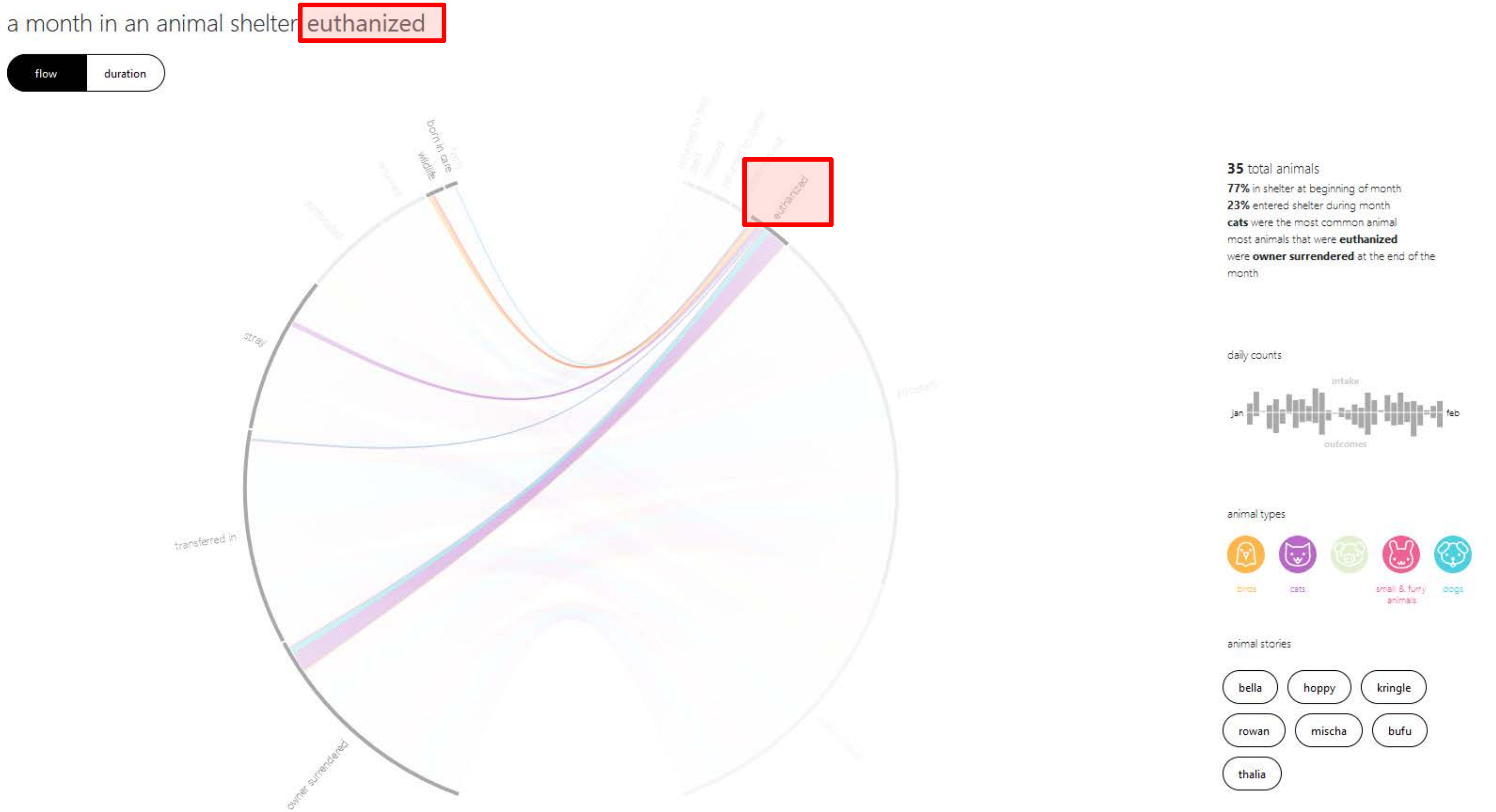
4.4. Representació de Dades

Exemple: “*A month in an animal shelter*”. Ressaltar flux d’informació dels gossos segons la **causa d’entrada**.



4.4. Representació de Dades

Exemple: "*A month in an animal shelter*". Ressaltar flux d'informació dels gossos segons la **causa de sortida**.



4.5. Disseny fiable

Fiabilitat, consistència i rendiment d'una visualització afecta a la confiança del que es percebi.

Preguntes:

- *Funciona tal com s'ha dit?.*
- *Usuari pot confiar en les funcions del sistema?.*

Confidencialitat
Ús
Integritat
Transparència

de les dades afecta a la confiança de l'observador

4.5. Disseny fiable

Usualment les dades proporcionades per l'usuari són utilitzades de forma temporal en la visualització.

Si volem guardar les dades de l'usuari de forma permanent i afegir-les en el data set, cal informar-lo clarament i demanar-li.

Altres preguntes:

- *Disseny és escalable amb noves dades?*
- *Qui mantindrà en el futur el projecte per a que continuï funcionant?*

4.6. Disseny accessible

Cal utilitzar la interactivitat quan haguem esgotat les capacitats que puguin proporcionar les solucions estàtiques.

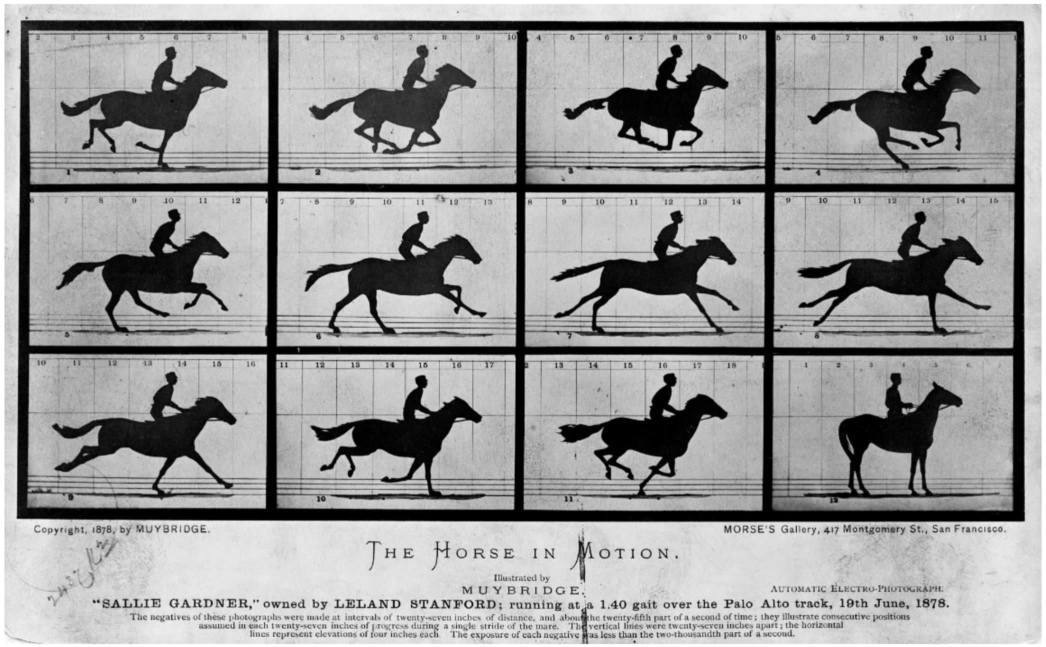
L'animació, no és aconsellable si les dades canvien poc (estàtic) o canvien massa (caos). L'animació es pot utilitzar per a:

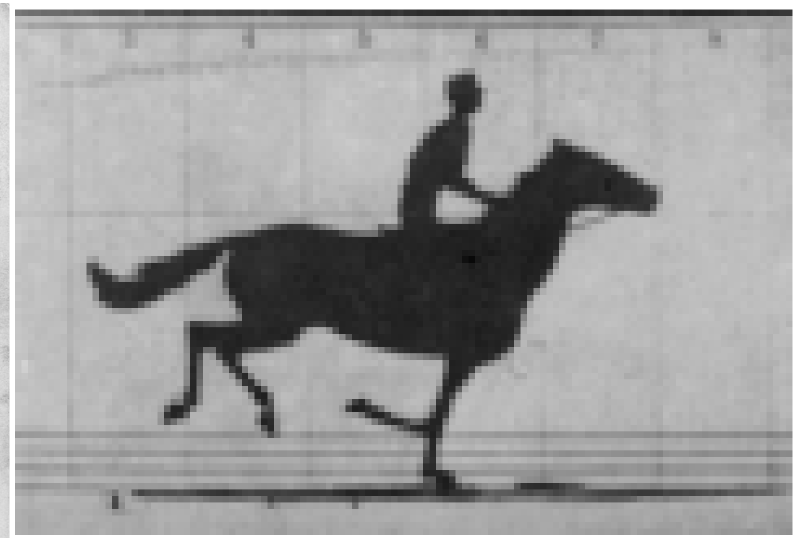
- Mostrar el dinamisme global de les dades
- L'evolució en el temps d'unes dades concretes

Controlar la velocitat de l'animació per a que es percebin els canvis que es volen mostrar a l'observador.

4.6. Disseny accessible

Per a una millor comparativa de les dades i apreciar el dinamisme, en animació cal mostrar els diferents frames sota la mateixa vista (*The Horse in Motion*).





4.7. Disseny elegant

Evitar utilitzar la interactivitat per a lluïment si no aporta res, si no és útil.

Que es pugui fer no significa que calgui.

Elegància en el disseny afecta a l'aparença, i també a com de fàcil s'accedeixi i utilitzin les tècniques interactives.

La implementació de tècniques interactives ve justificat si augmenta l'atenció de l'observador.

Sempre que sigui possible, proporcionar una experiència divertida i jugable, i que no distregui o impedeixi la comprensió a l'observador.

En resum…

La Interactivitat és una aspecte important en la visualització que mostra el dinamisme de les dades i aspectes que no es poden apreciar en les imatges estàtiques.

La interactivitat dona més control a l'observador o usuari en les dades que pot/vol veure. Aquest control ha d'anar orientat a que l'observador experimenti una millor comprensió.

Cal no abusar de la interactivitat i evitar que distregui o aparti a l'observador en l'objectiu principal que és la comprensió.

Les tècniques d'animació pretenen visualitzar dades amb dimensió temporal en forma de seqüència i oferir visions diferents a les gràfiques estàtiques.

En resum…

Consells generals per l’ús de la interactivitat en un projecte de visualització de dades:

- Tenir una bona gestió del temps si implementem una solució interactiva.
- No invertir temps, esforços i recursos en tècniques interactives i/o innovadores si no aporten valor a la comprensió.
- Focalitzar-se en el que és important i rellevant. Qüestionar-se sempre si la interactivitat és necessària per al projecte i li aporta valor.

GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

12. Usabilitat i Experiència d'Usuari (UX)

Enric Martí Gòdia

Departament de Ciències de la Computació

Continguts

1. Introducció

Un cop definit un sistema de Visualització de Dades és important avaluar si val per als usuaris a qui va dirigit, especialment si és un sistema interactiu.

Àrees com **Usabilitat** i **Experiència d'Usuari (UX)** proporcionen eines d'avaluació de sistemes informàtics i interactius.

Ambdós termes (Usabilitat, UX) es barregen sovint, però un és inclòs dins l'altre.

Anem a definir-los i veure quines eines d'avaluació proporcionen.

2. Usabilitat

Usabilitat: Atribut de qualitat que avalua l'ús d'aplicacions, websites, interfícies, sistemes, etc. per part de l'usuari.

Inclou cinc components de qualitat:

1. *Learnability:* Com de fàcil és pels usuaris aprendre les tasques que han de realitzar en el sistema
2. *Efficiency:* Com de ràpid els usuaris poden realitzar les tasques
3. *Memorability:* Com de fàcil els usuaris poden reproduir les tasques en el sistema després d'un temps sense utilitzar-lo
4. *Errors:* Quants i quins errors poden realitzar els usuaris, què de greus poden ser i com el sistema es pot recuperar si es produeixen
5. *Satisfacció:* Com de satisfactori és l'ús del sistema

2. Usabilitat

Conceptes *Utility*, *Usability*, *Useful*:

- *Utility:* Funcionalitat. Sistema fa el que els usuaris esperen?
- *Usability:* Com de fàcil i satisfactori és l'ús del sistema
- *Useful: Utility* + *Usability*

2. Usabilitat

Usabilitat es important, doncs és una condició necessària per a que els usuaris utilitzin el sistema, aplicació, website, etc.

Usuaris abandonen el sistema si:

- És complicat d'utilitzar
- Si el sistema falla sovint
- Si el sistema no permet fer el que els usuaris volen
- Si es perden en el seu ús
- Si la informació o els menús d'ajuda són poc clars i no intuïtius

Usabilitat molt relacionat amb productivitat.

2. Usabilitat

Data Usability: Qualitat de les dades, de la informació, dins el context de Visualització de Dades.

Associat a tres principis:

- ***Data Reliability:*** Qualitat dels processaments de dades, que donin confiança en les dades que es treballen tenint en compte intervals d'errors.
- ***Data Stability:*** Minimitzar l'impacte negatiu de les transformacions o processaments. Per exemple la reducció de dimensionalitat.
- **Suport a la presa de decisions:** Representació ha de ser prou entenedora als usuaris per a que puguin prendre decisions.

2. Usabilitat

Per a avaluar la usabilitat el més important és centrar-se en l'usuari. Vàries formes:

- **Observació:** Veure l'usuari com utilitza el sistema, i com reacciona, comprovant si les seves accions són les esperades o no
- **Qüestionari:** Usuari respon qüestions després d'utilitzar el sistema. La forma més usual

Dos qüestionaris estàndard en usabilitat:

- **SUS (*System Usability Scale*)** – John Brooke, 1986
- **PSSUQ (*Post-Study System Usability Questionnaire*)** – IBM 1988

2.1. SUS (*System Usability Scale*)

Eina ràpida de mesura d'usabilitat de software, website i aplicacions, entre d'altres.

10 preguntes amb 5 escales des de *Strongly Agree* fins a *Strongly Disagree*.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

Avantatges:

- Escala simple i fàcil d'entendre per l'usuari
- Útil per a avaluar sistemes simples i dona resultats fiables
- Distingeix fàcilment entre un sistema útil i no útil

2.1. SUS (*System Usability Scale*)

Consideracions:

- No utilitza les escales de 0-100 com a percentatges
- Per a interpretar resultats, normalitza les valoracions per a produir percentils
- SUS no fa diagnòstic. Només valora la facilitat d'ús d'un sistema, website o aplicació. Avalua:
  - *Effectiveness:* Si l'usuari pot realitzar les accions encomanades
  - *Efficiency:* Esforç de l'usuari per a fer les accions

2.1. SUS (*System Usability Scale*)

Qüestionari (*Effectiveness*, *Efficiency*):

1. *I think that I would like to use this system frequently.*
2. *I found the system unnecessarily complex.*
3. *I thought the system was easy to use.*
4. *I think that I would need the support of a technical person to be able to use this system.*
5. *I found the various functions in this system were well integrated.*
6. *I thought there was too much inconsistency in this system.*
7. *I would imagine that most people would learn to use this system very quickly.*
8. *I found the system very cumbersome to use.*
9. *I felt very confident using the system.*
10. *I needed to learn a lot of things before I could get going with the system.*

2.1. SUS (*System Usability Scale*)

Qüestionari (*Effectiveness*, *Efficiency*):

1. *I think that I would like to use this system frequently.*
2. *I found the system unnecessarily complex.*
3. *I thought the system was easy to use.*
4. *I think that I would need the support of a technical person to be able to use this system.*
5. *I found the various functions in this system were well integrated.*
6. *I thought there was too much inconsistency in this system.*
7. *I would imagine that most people would learn to use this system very quickly.*
8. *I found the system very cumbersome to use.*
9. *I felt very confident using the system.*
10. *I needed to learn a lot of things before I could get going with the system.*

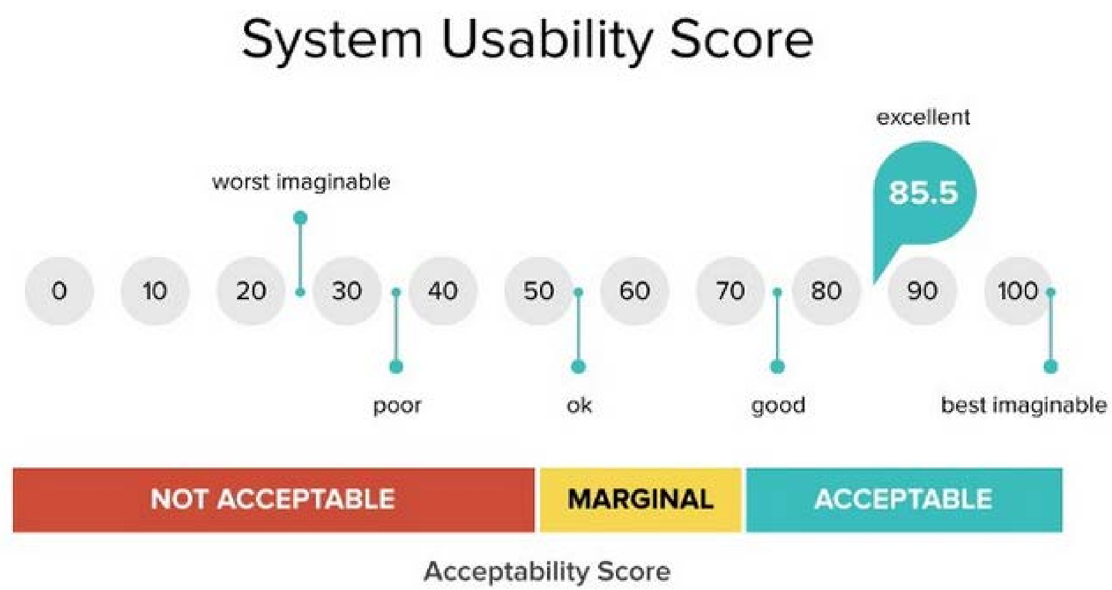
2.1. SUS (*System Usability Scale*)

Valoració SUS:

- Per a cada escala es dona la següent valoració:
  - *Strongly Disagree*: 1 punt
  - *Disagree*: 2 punts
  - *Neutral:* 3 punts
  - *Agree*: 4 punts
  - *Strongly Agree*: 5 punts
- X = (Σ puntuacions preguntes imparells) - 5
- Y = 25 - (Σ puntuacions preguntes parells)
- Valoració SUS = (X + Y) * 2,5

2.1. SUS (*System Usability Scale*)

Valoració SUS (2):



Si la puntuació és menor que 51, cal verificar en el sistema:

- Les estructures de navegació o de jerarquia són prou intuïtives?
- Els missatges són clars i comprensibles?
- Les diferents opcions estan ben categoritzades?
- Les tasques i el flux d'usuari és complicat?
- El seu ús crea frustració o errors de forma contínua?

2.2. PSSUQ (*Post-Study System Usability Questionnaire*)

Molt utilitzat per a mesurar satisfacció d'usuari en una website, sistema o software.

Actualment disponible la PSSUQ v3.0.

16 preguntes amb 7 escales des de *Strongly Agree – Strongly Disagree* més una opció *N.A.*

https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/

2.2. PSSUQ (*Post-Study System Usability Questionnaire*)

	Strongly Agree					Strongly Disagree		
PSSUQ	1	2	3	4	5	6	7	N.A.
1. Overall, I am satisfied with how easy it is to use this system.								
2. It was simple to use this system.								
3. I was able to complete the tasks and scenarios quickly using this system.								
4. I felt comfortable using this system.								
5. It was easy to learn to use this system.								
6. I believe I could become productive quickly using this system.								
7. The system gave error messages that clearly told me how to fix problems.								
8. Whenever I made a mistake using the system, I could recover easily and quickly.								
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.								

2.2. PSSUQ (*Post-Study System Usability Questionnaire*)

	Strongly Agree					Strongly Disagree		
	1	2	3	4	5	6	7	N.A.
10. It was easy to find the information I needed.								
11. The information was effective in helping me complete the tasks and scenarios.								
12. The organization of information on the system screens was clear.								
13. The interface of this system was pleasant.								
14. I liked using the interface of this system.								
15. This system has all the functions and capabilities I expect it to have.								
16. Overall, I am satisfied with this system.								

Questions 1 to 16: Overall
Questions 1 to 6: System Usefulness (SYSUSE)
Questions 7 to 12: Information Quality (INFOQUAL)
Questions 13 to 16: Interface Quality (INTERQUAL)

Source: uiuxtrend.com

2.2. PSSUQ (*Post-Study System Usability Questionnaire*)

Valoració PSSUQ:

- Per a cada escala es dona la següent valoració:
  - *Strongly Disagree*: 1 punt
  - *...*
  - *Strongly Agree*: 7 punts
- Es poden obtenir diferents indicadors:
  - *Overall*: Mitja de valoracions de totes les preguntes 1-16
  - *SYSTEM USEFULLNES (SYSUSE): Mitja* de valoracions de les preguntes 1-6
  - *INFORMATION QUALITY (INFOQUAL): Mitja* de valoracions de les preguntes 7-12. **Millor qüestionari que SUS.**
  - *INTERFACE QUALITY (INTERQUAL): Mitja* de valoracions de les preguntes 13-15

2. Usabilitat

SUS vs PSSUQ: **Quin utilitzar?**

- Mirar les preguntes d'un i altre per a valorar quin s'adapta millor a l'avaluació que es vol fer.
- Quins subapartats es vol valorar?:
  - Si es vol valorar *Learnability* (*Efficiency*), millor SUS
  - Si es vol avaluar qualitat d'informació, millor PSSUQ
- **Fatiga del tester:** PSSUQ té més preguntes i més complex. Si es vol un test amb poques i simples preguntes, millor SUS.

2. Usabilitat

Altres qüestionaris sobre usabilitat referenciats en:

Acronym	Instrument	Institution	Example
QUIS	Questionnaire for User Interface Satisfaction	Maryland	27 questions
PUEU	Perceived Usefulness and Ease of Use	IBM	12 questions
NAU	Nielsen's Attributes of Usability	Bellcore	5 attributes
NHE	Nielsen's Heuristic Evaluation	Bellcore	10 heuristics
CSUQ	Computer System Usability Questionnaire	IBM	19 questions
ASQ	After Scenario Questionnaire	IBM	3 questions
PHUE	Practical Heuristics for Usability Evaluation	OSU	13 heuristics
PUTQ	Purdue Usability Testing Questionnaire	Purdue	100 questions
USE	USE Questionnaire	Sapient	30 questions

També se'n poden trobar en https://www.interaction-design.org/literature/topics/surveys

3. Experiència d’Usuari (UX)

Experiència d’Usuari (UX) tracta, estudia l’experiència d’un usuari en una interacció o sistema.

UX inclou diferents indicadors, siguin objectius, estètics o emocionals:

- *Usabilility*
- *User Interface*
- *Interaction Experience*
- *Interactive Design*
- *Customer Experience*

3. Experiència d'Usuari (UX)

Definicions UX:

- *All the aspects of how people use an interactive product: the way it feels in their hands, how well they understand how it works, how they feel about it while they're using it, how well it serves their purposes, and how well it fits into the entire context in which they are using it (Alben, 1996)*
- *A consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use, etc.)* (Hassenzahl & Tractinsky, 2006)
- *The user's perceptions and responses that result from the use and/or anticipated use of a system, product or service (ISO 9241-11, 2017).*

3. Experiència d'Usuari (UX)

Cinc categories en mètodes avaluació UX:

- **Avaluacions d'Experts:** Avaluacions del sistema fetes per experts. Prèvies a estudis de laboratori o camp.
- **Estudis de Laboratori:** Similar a tests d'usabilitat, útils en fase de prototipatge. Participants proven el sistema i desenvolupador observa el seu comportament.
- **Estudis de Camp:** Participants proven el sistema en l'entorn o context que correspon.
- **Qüestionaris:** Usuaris contesten preguntes sobre el sistema per a formar una opinió.
- **Mixtes:** Utilitzar diferents mètodes per a tenir més indicadors. Exemple: Estudis de Camp + Qüestionaris.

3. Experiència d'Usuari (UX)

Mesures Psico-fisiològiques: Forma objectiva també de mesurar la UX.

- Batecs de cor, transpiració, moviments músculs facials per a mesurar emocions positives o negatives.
- Important prendre mesures amb dispositius no invasius, que no destorbin l'experiència.
- Emocions no són útils en tots els estudis UX.

4. Models d'UX

Models d'avaluació basats en dos aspectes importants:

- Usabilitat
- Experiència, emotivitat

4. Models d'UX

Elements externs són importants per a entendre l'UX com una experiència subjectiva i dinàmica.

3 tipus d'elements externs:

1. Usuari:

- Element important a qui va dirigit el sistema de VD
- Diferents competències: nivells, amb experiència, experts, etc.
- Altres característiques més emocionals: estatus social, personal, objectius personals, necessitats afectives poden influir en l'experiència UX

2. Context:

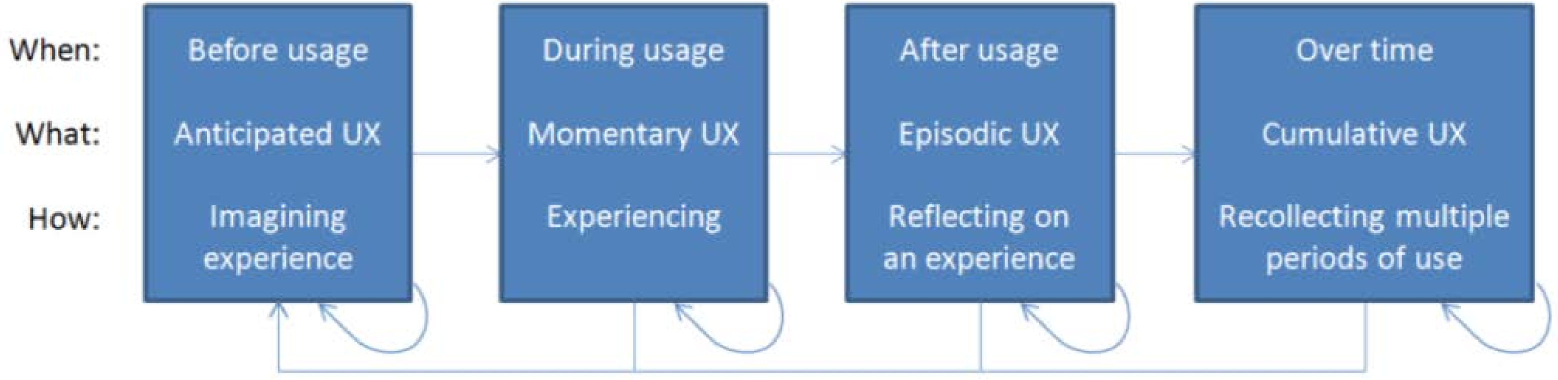
- Entorn físic o localització on s'utilitzarà el sistema (lloc il·luminat, bona connexió Internet, etc.)
- Context sòcio-cultural: Valors, actitud, estil de vida, experiència prèvia en productes semblants, etc.
- Context de mercat: Si producte innovador o existeixen de semblants

4. Models d'UX

3 tipus d'elements externs (2):

3. Aspectes temporals:

- Temps és important en UX
- Qualitat pragmàtica (usabilitat) augmenta amb el temps
- Qualitat hedònica (emotivitat) decreix en el temps, doncs deixa de ser novetat. Una animació pot ser atractiva i novedosa al principi, però al cap del temps deixa de ser novetat
- Avaluació UX d'un sistema es realitza quan està acabat de fer. Pocs estudis d'avaluació UX de sistemes a llarg termini



4. Models d'UX

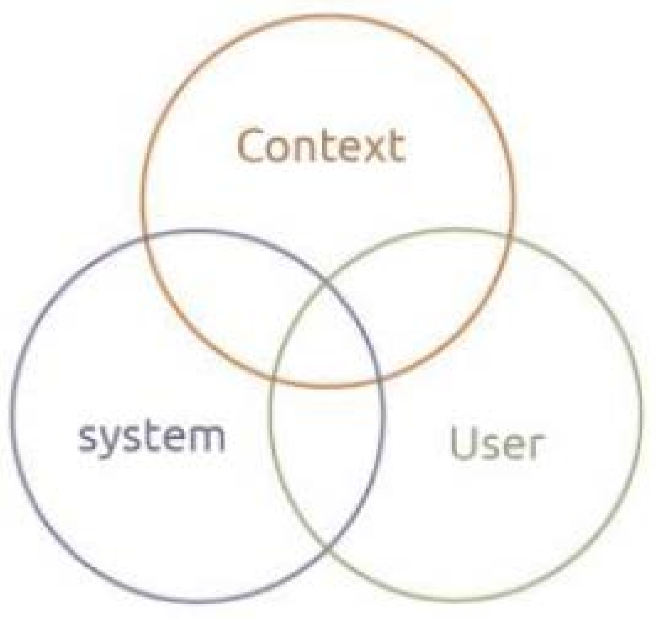
Models més utilitzats **[DLQ-19]**:

1. **Hassenzahl UX** - (Hassenzahl, Burmester, Koller, 2003) **[BHK-11]**
2. ***User Experience Questionnaire* (UEQ)** – (Laugwith, Held, Schrepp, 2005) **[LSH-08]**
3. **meCUE – (**Minge, Thüring, Wagner, Kuhr, 2016) **[MiT-16]**

4.1. Hassenzahl UX

Experiència d'Usuari (UX) basat en 3 factors:

- **Usuari:** Experiència, expectatives, estat d'ànim
- **Sistema:** Propietats, usabilitat, estètica
- **Context:** Entorn, situació social, física, tèrmica, lumínica, etc.)



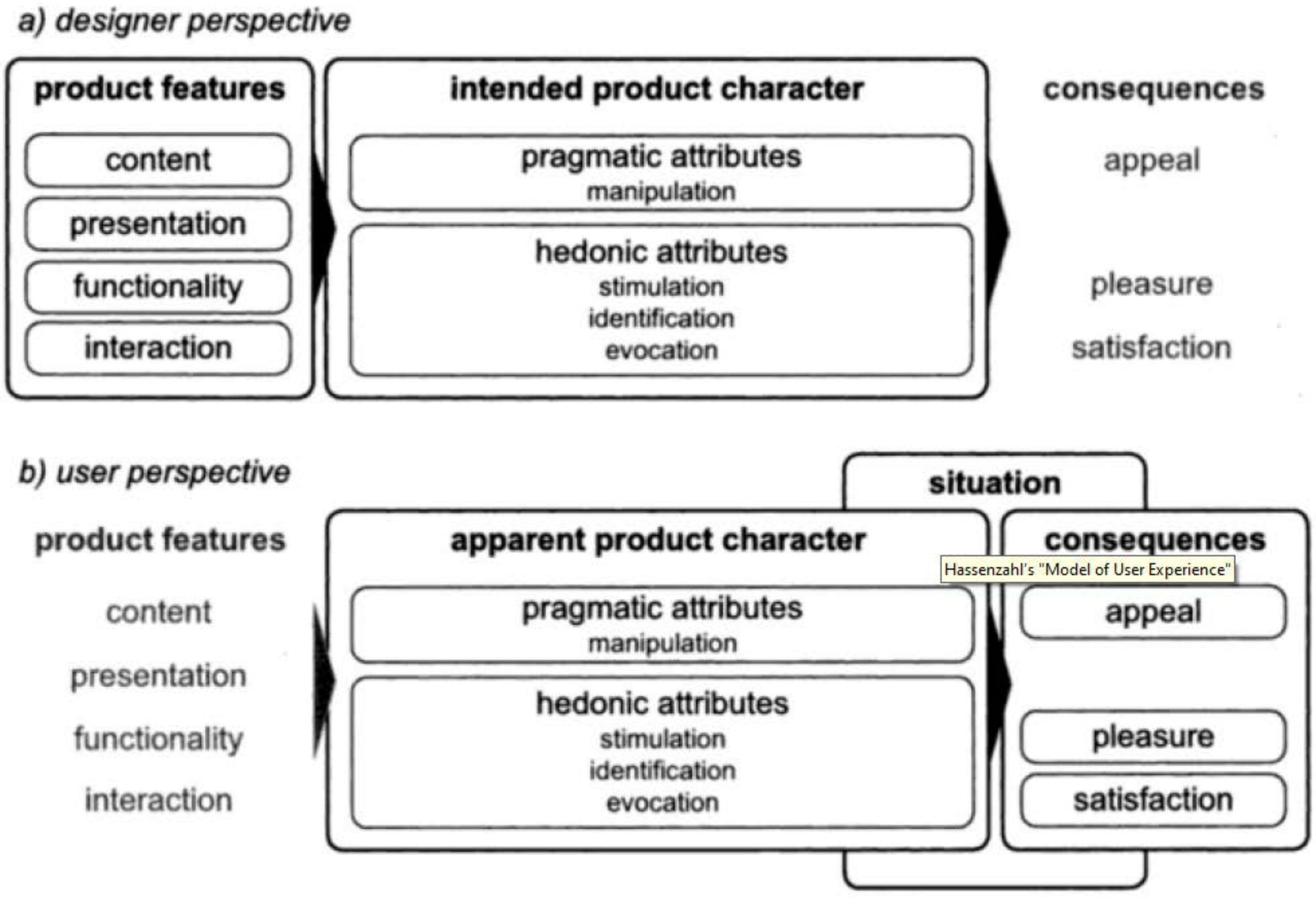
4.1. Hassenzahl UX

Model distingeix entre:

- **Atributs pragmàtics:** Utilitat i la funcionalitat. Mesures d'usabilitat (aprenentatge, eficiència, efectivitat, etc.)
  - **Manipulació** del sistema requereix funcionalitat (utilitat) i formes d'avaluar-la (usabilitat)
- **Atributs hedònics:** Aspectes no instrumentals: Estètics, de comportament o funcionalitat intuïtiva
  - **Estimulació en desenvolupar competències:** Marcar objectius, descobrir sorpreses, etc.
  - **Identificació:** Que el sistema o producte comuniqui identitat, personal (pàgines web personals) o corporativa (webs d'empresa)
  - **Evocació:** Que el sistema evoqui experiències anteriors ja adquirides que facilitin l'ús i la identificació per part de l'usuari

4.1. Hassenzahl UX

Distinció entre els atributs dissenyats i la percepció que en té l’usuari.



4.1. Hassenzahl UX

Aplicació del model: **Questionari Attrakdiff**

Qüestionari de 28 criteris agrupats en quatre tipus de mesures. Cada criteri té 7 escales de pitjor a millor. Exemple de qüestionari:

Your Opinion

In the following, you will find word pairs that are intended to aid you in assessing the product that you have just become acquainted with. The word pairs represent extreme opposites, with seven graduations possible between them.

An example:

bad ☐ ☐ ☐ ☐ ☒ ☐ ☐ good

The above evaluation example suggests that the product is mainly to your liking, but could stand some improvement.

Do not ponder too long over your response. Just offer the first assessment that enters your mind.

Perhaps some of the assessment terms are not quite suitable to the product. Despite this, however, please checkmark at least one box in the scale. Please consider that there are no "correct" or "incorrect" answers – only your own personal opinion counts!

Please provide your impressions of the product you have tested by check marking your impression on the scale between the terms offered in each line.

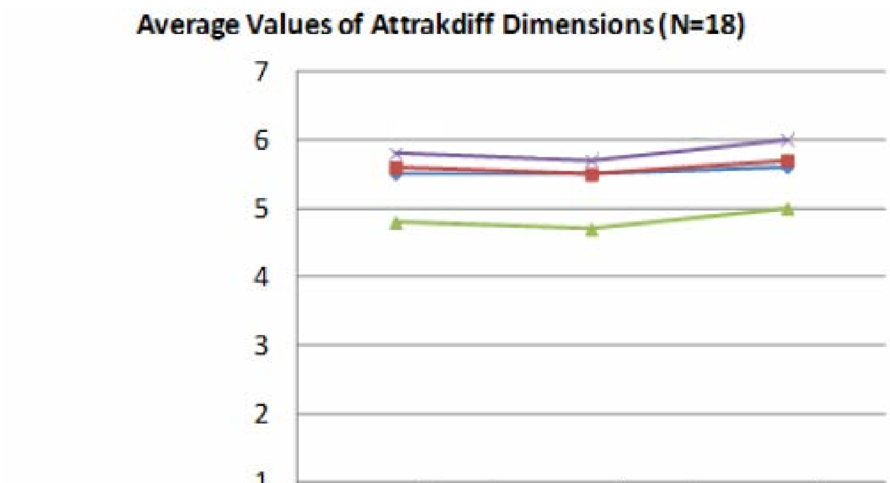
	1	2	3	4	5	6	7	
human	☐	☐	☐	☐	☐	☐	☐	technical
isolating	☐	☐	☐	☐	☐	☐	☐	connective
pleasant	☐	☐	☐	☐	☐	☐	☐	unpleasant
inventive	☐	☐	☐	☐	☐	☐	☐	conventional
simple	☐	☐	☐	☐	☐	☐	☐	complicated
professional	☐	☐	☐	☐	☐	☐	☐	unprofessional
ugly	☐	☐	☐	☐	☐	☐	☐	attractive
practical	☐	☐	☐	☐	☐	☐	☐	impractical
likeable	☐	☐	☐	☐	☐	☐	☐	disagreeable
cumbersome	☐	☐	☐	☐	☐	☐	☐	straightforward
stylish	☐	☐	☐	☐	☐	☐	☐	tacky
predictable	☐	☐	☐	☐	☐	☐	☐	unpredictable
cheap	☐	☐	☐	☐	☐	☐	☐	premium
alienating	☐	☐	☐	☐	☐	☐	☐	integrating
brings me closer to people	☐	☐	☐	☐	☐	☐	☐	separates me from people
unpresentable	☐	☐	☐	☐	☐	☐	☐	presentable
rejecting	☐	☐	☐	☐	☐	☐	☐	inviting
unimaginative	☐	☐	☐	☐	☐	☐	☐	creative
good	☐	☐	☐	☐	☐	☐	☐	bad
confusing	☐	☐	☐	☐	☐	☐	☐	clearly structured
repelling	☐	☐	☐	☐	☐	☐	☐	appealing
bold	☐	☐	☐	☐	☐	☐	☐	cautious
innovative	☐	☐	☐	☐	☐	☐	☐	conservative
dull	☐	☐	☐	☐	☐	☐	☐	captivating
undemanding	☐	☐	☐	☐	☐	☐	☐	challenging
motivating	☐	☐	☐	☐	☐	☐	☐	discouraging
novel	☐	☐	☐	☐	☐	☐	☐	ordinary
unruly	☐	☐	☐	☐	☐	☐	☐	manageable

4.1. Hassenzahl UX

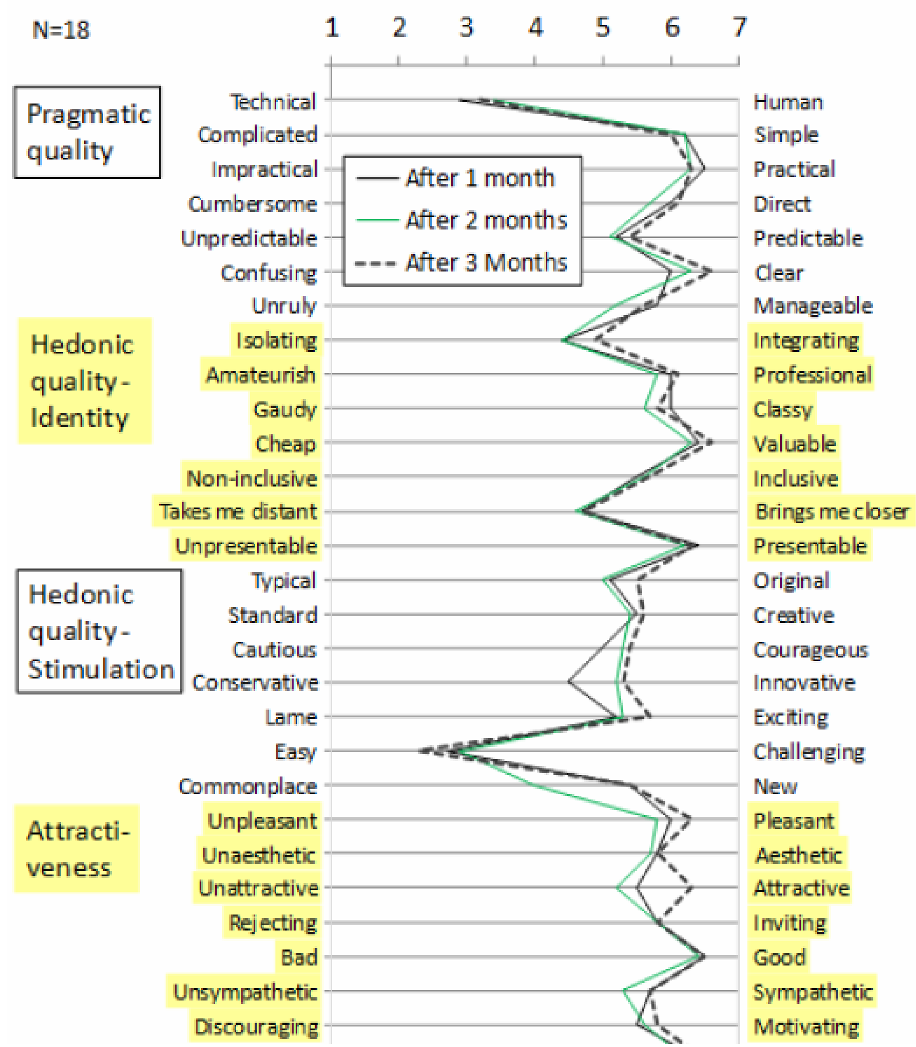
Qüestionari Attrakdiff (2)

Qüestionari de 28 criteris agrupats en quatre tipus de mesures. Cada criteri té 7 escales des de pitjor a millor.

Exemple de Resultats per a N=18:



	1 Month	2 Months	3 Months
Pragmatic quality	5,5	5,5	5,6
Hedonic quality -Identity	5,6	5,5	5,7
Hedonic quality - Stimulation	4,8	4,7	5
Attractiveness	5,8	5,7	6



4.1. Hassenzahl UX

Qüestionari Attrakdiff (3)

Agrupant criteris en dos grans tipus de mesures (atributs pragmàtics i atributs hedònics) valors mínims i màxims de cada tipus es poden representar en un espai 2D per a comparar sistemes:

4.2. User Experience Questionnaire (UEQ)

Avaluació de productes interactius (websites, software, aplicacions mòbils).

Informació qüestionari i eines d'avaluació a https://www.ueq-online.org/ i **[Sch-19]**

Avaluats uns 452 productes per uns 20190 usuaris.

Qüestionari de 26 indicadors, cadascun representat per dos termes oposats (negatiu i positiu) amb 7 escales de valoració, 3 negatives, 3 positives, 1 neutre. Entre 3 i 5 minuts a contestar.

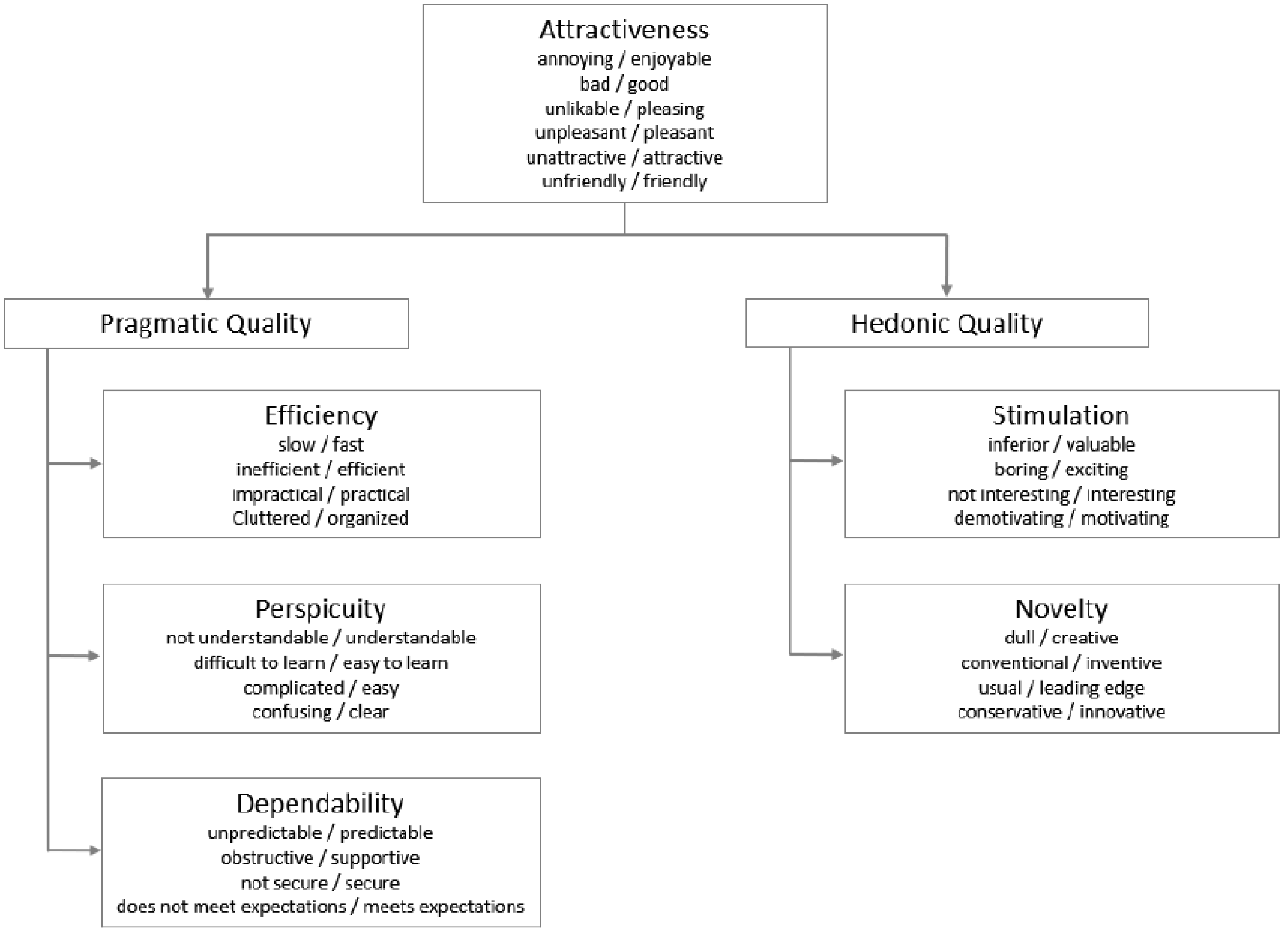
attractive o o o o o o o unattractive

4.2. User Experience Questionnaire (UEQ)

UEQ conté 6 categories de valoració en els 26 indicadors:

1. *Attractiveness:* Impressió general del producte. Agrada o no el producte?
2. *Perspicuity:* És fàcil aprendre com funciona el producte?
3. *Efficiency:* Poden els usuaris resoldre les tasques fàcilment?
4. *Dependability:* L'usuari té la sensació de controlar la interfície?
5. *Stimulaton:* És interesant i motivador l'ús del producte?
6. *Novelty:* És innovador i creatiu el producte?. Capta l'interès dels usuaris?.

4.2. User Experience Questionnaire (UEQ)



4.2. User Experience Questionnaire (UEQ)

Please assess the product now by ticking one circle per line.

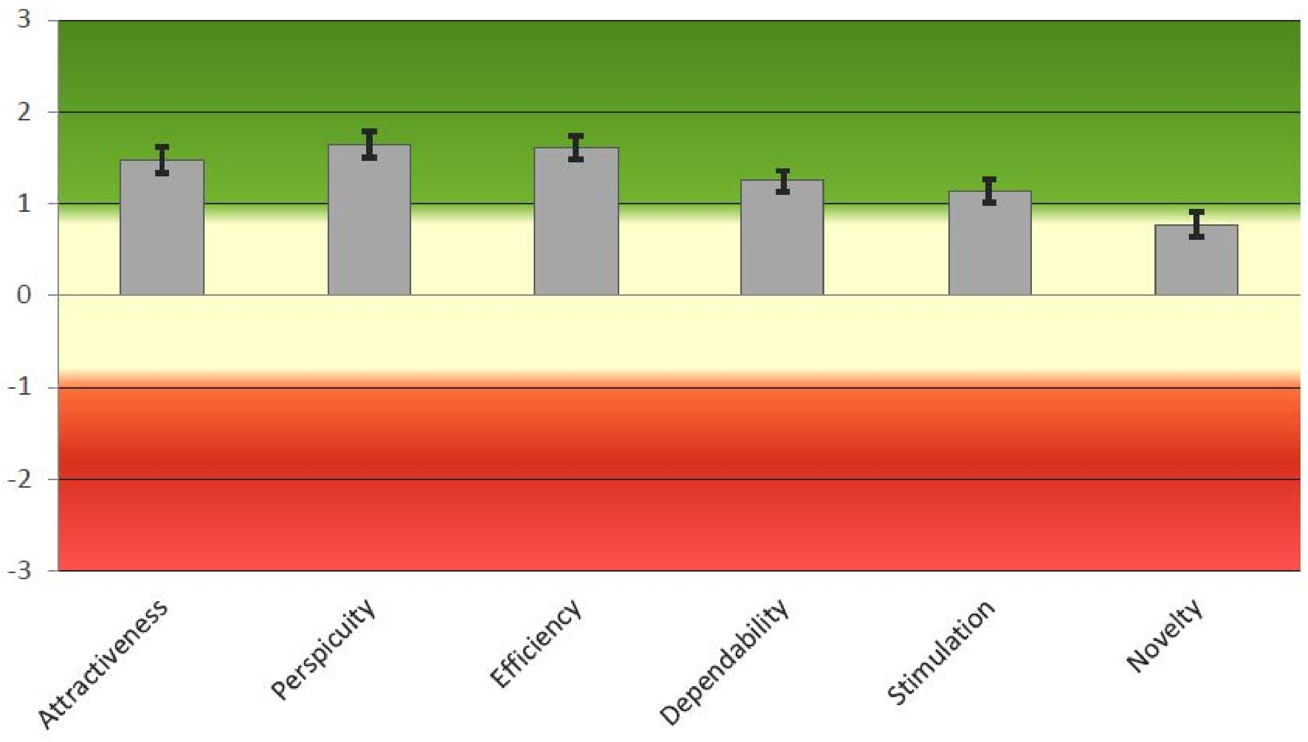
	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

4.2. User Experience Questionnaire (UEQ)

Valoració UEQ sobre un producte

Fer la mitja de les respostes de cada pregunta i calcular la mitja per categories.

Representació gràfica:

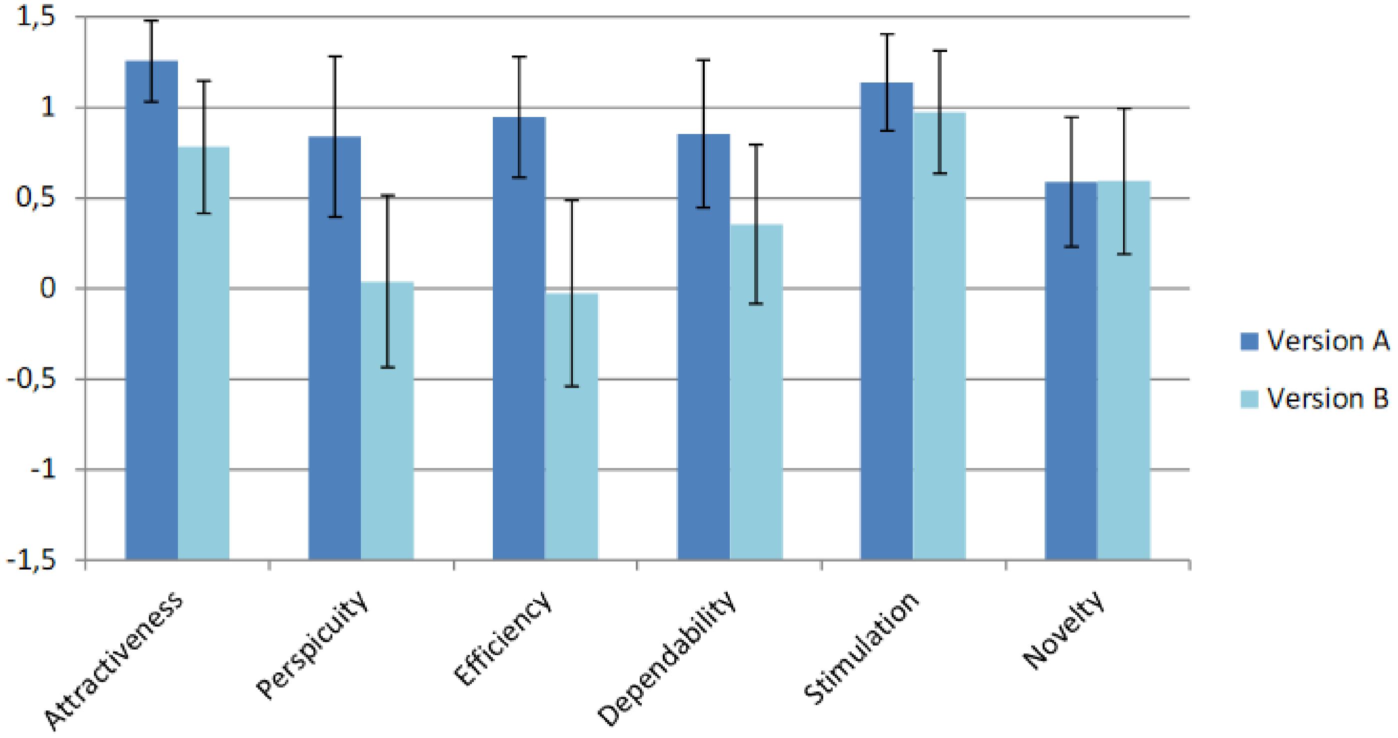


4.2. User Experience Questionnaire (UEQ)

Comparativa entre productes

Fer el mitja de les respostes de cada pregunta i calcular mitja per categories i per a cada producte.

Diagrama de barres per categoria:



4.3. Qüestionari meCUE

modular evaluation of key Components of User Experience (Minge, Thüring, Wagner & Kuhr, 2016)

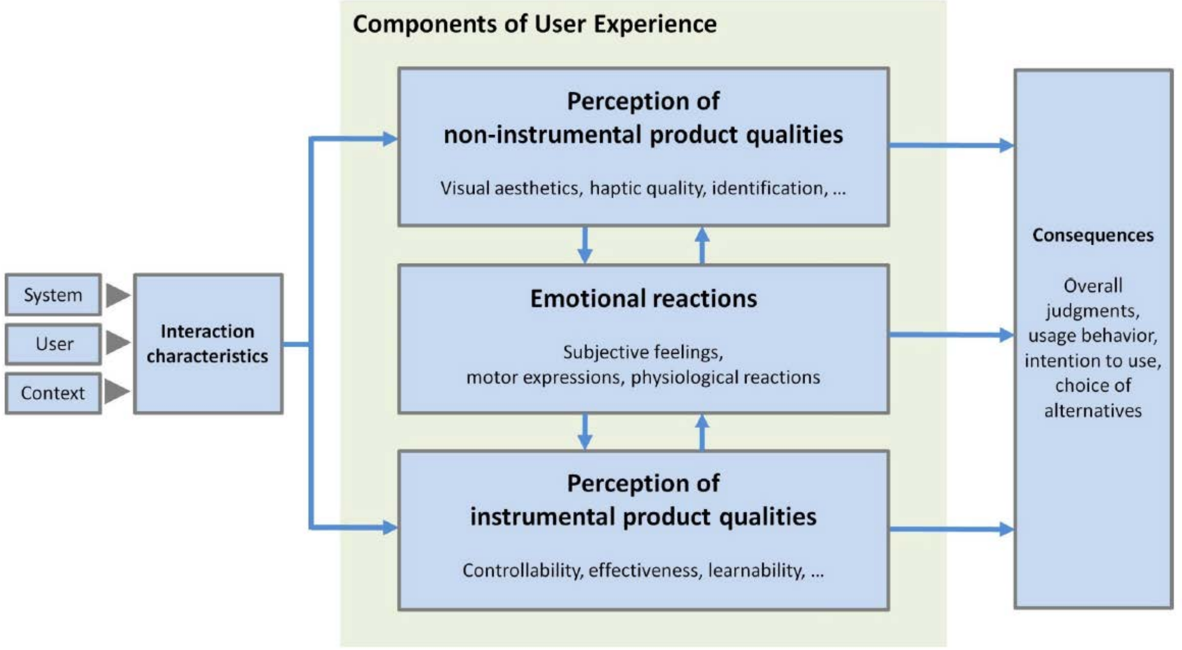
Desenvolupat per a crear una mesura estàndard en UX.

Basat en el model CUE definit per Mahle i Thüring.

4.3. meCUE

CUE (*Components of User Experience*)

Model similar al de Hassenzahl definint característiques instrumentals i no instrumentals similars a les pragmàtiques i hedòniques) que influeixen en les respostes emocionals de l'usuari, tercera característica important del model.



4.3. meCUE

Qüestionari CUE: Aplicació del model per qüestionaris per a cada característica:

- **Instrumentals (usabilitat)**: Qüestionari SUMI (Kirakowsi, Corbett 1993) **[SUMI]**.
- **No instrumentals:** Qüestionari Visual-Estètic (Lavie-Tractinsky 2004). Un dels més vàlids per a mesurar estètica en websites. Distingeix entre estètica clàssica (clar i ordenat) d'un disseny més expressiu (creativitat i innovació) **[LaT-04]**.
- **Reaccions emocionals:** Qüestionari SAM (Bradley-Lang, 1994), basat en pictogrames **[BrL-94]**.

4.3.1. Questionari SUMI

Software Usability Measurement Inventory

SUMI

NB The information you provide is kept completely confidential and no information is stored on computer media that could identify you as a person.

This questionnaire has 50 statements. Please answer them all. After each statement there are three boxes.

- Check the first box if you generally AGREE with the statement.
- Check the middle box if you are UNDECIDED, or if the statement has no relevance to your software or to your situation.
- Check the right box if you generally DISAGREE with the statement.

In checking the left or right box you are not necessarily indicating strong agreement or disagreement but just your general feeling most of the time.

There are also some general questions at the end.

Password:

What, in general, do you use this software for?

Statements 1 - 10 of 50.	Agree	Undecided	Disagree
This software responds too slowly to inputs.	○	○	○
I would recommend this software to my colleagues.	○	○	○
The instructions and prompts are helpful.	○	○	○
This software has at some time stopped unexpectedly.	○	○	○
Learning to operate this software initially is full of problems.	○	○	○
I sometimes don't know what to do next with this software.	○	○	○
I enjoy the time I spend using this software.	○	○	○
I find that the help information given by this software is not very useful.	○	○	○
If this software stops it is not easy to restart it.	○	○	○
It takes too long to learn the software functions.	○	○	○

Statements 11 - 20 of 50.	Agree	Undecided	Disagree
I sometimes wonder if I am using the right function.	○	○	○
Working with this software is satisfying.	○	○	○
The way that system information is presented is clear and understandable.	○	○	○
I feel safer if I use only a few familiar functions.	○	○	○
The software documentation is very informative.	○	○	○
This software seems to disrupt the way I normally like to arrange my work.	○	○	○
Working with this software is mentally stimulating.	○	○	○
There is never enough information on the screen when it's needed.	○	○	○
I feel in command of this software when I am using it.	○	○	○
I prefer to stick to the functions that I know best.	○	○	○

Statements 21 - 30 of 50.	Agree	Undecided	Disagree
I think this software is inconsistent.	○	○	○
I would not like to use this software every day.	○	○	○
I can understand and act on the information provided by this software.	○	○	○
This software is awkward when I want to do something which is not standard.	○	○	○
There is too much to read before you can use the software.	○	○	○
Tasks can be performed in a straight forward manner using this software.	○	○	○
Using this software is frustrating.	○	○	○
The software has helped me overcome any problems I have had in using it.	○	○	○
The speed of this software is fast enough.	○	○	○
I keep having to go back to look at the guides.	○	○	○

Statements 31 - 40 of 50.	Agree	Undecided	Disagree
It is obvious that user needs have been fully taken into consideration.	○	○	○
There have been times in using this software when I have felt quite tense.	○	○	○
The organisation of the menus seems quite logical.	○	○	○
The software allows the user to be economic of keystrokes.	○	○	○
Learning how to use new functions is difficult.	○	○	○
There are too many steps required to get something to work.	○	○	○
I think this software has sometimes given me a headache.	○	○	○
Error messages are not adequate.	○	○	○
It is easy to make the software do exactly what you want.	○	○	○
I will never learn to use all that is offered in this software.	○	○	○

Statements 41 - 50 of 50.	Agree	Undecided	Disagree
The software hasn't always done what I was expecting.	○	○	○
The software presents itself in a very attractive way.	○	○	○
Either the amount or quality of the help information varies across the system.	○	○	○
It is relatively easy to move from one part of a task to another.	○	○	○
It is easy to forget how to do things with this software.	○	○	○
This software occasionally behaves in a way which can't be understood.	○	○	○
This software is really very awkward.	○	○	○
It is easy to see at a glance what the options are at each stage.	○	○	○
Getting data files in and out of the system is not easy.	○	○	○
I have to look for assistance most times when I use this software.	○	○	○

How important for you is the kind of software you have just been rating?

- ○ Extremely important
- ○ Important
- ○ Not very important
- ○ Not important at all

How would you rate your software skills and knowledge?

- ○ Very experienced and technical
- ○ I'm experienced but not technical
- ○ I can cope with most software
- ○ I find most software difficult to use

What do you think is the best aspect of this software, and why?

What do you think needs most improvement, and why?

When you've answered all the questions: please click the 'Send' button.

Send

4.3.2. Qüestionari Visual-Estètic

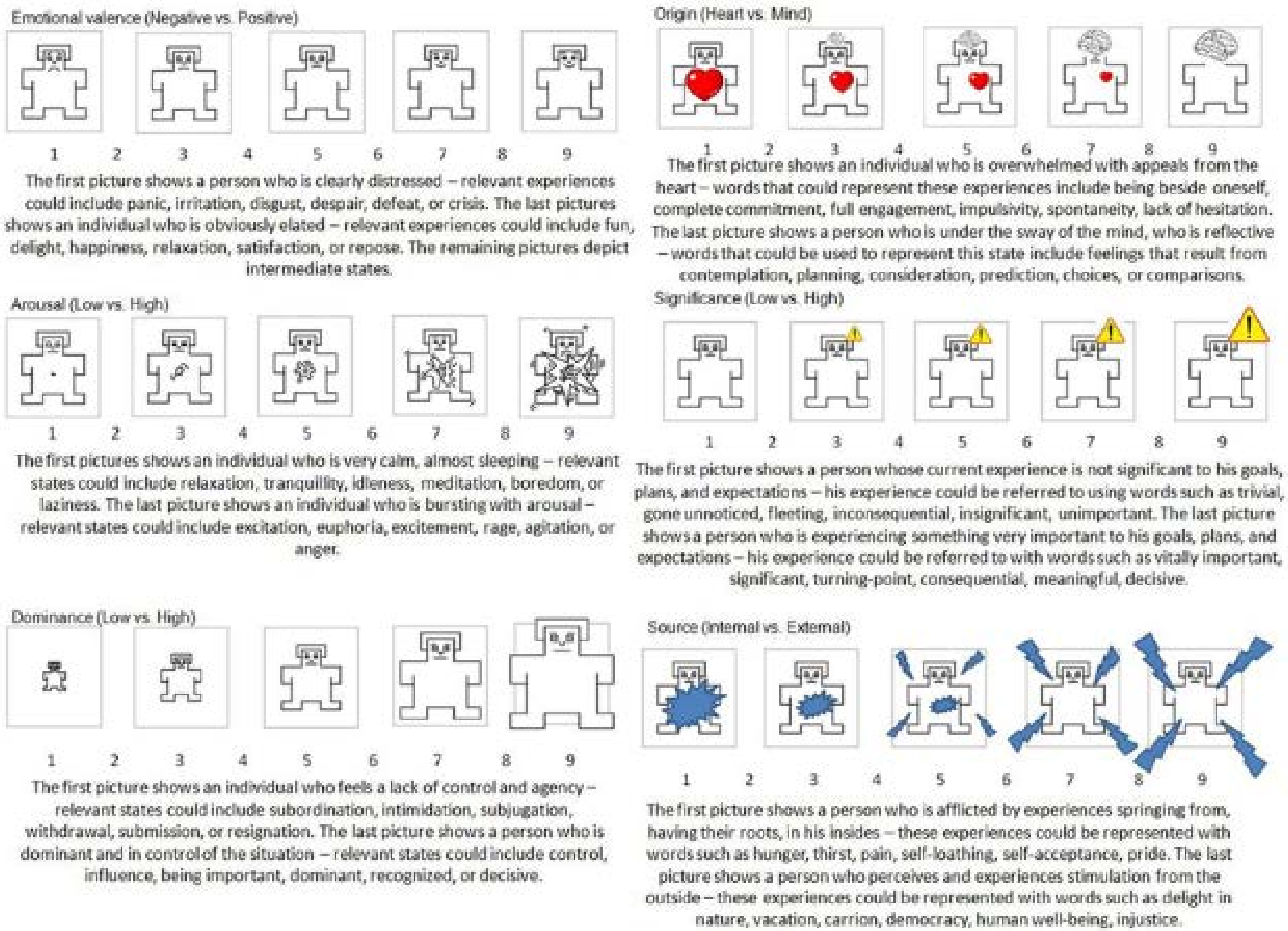
Items d'Estètica Clàssica:

Item	Strongly Disagree 0	1	2	3	4	5	6	Strongly Agree 7
Clean								
Clear								
Pleasant								
Symmetrical								
Aesthetic								

Items d'Estètica Expressiva:

Item	Strongly Disagree 0	1	2	3	4	5	6	Strongly Agree 7
Original								
Sophisticated								
Fascitaning								
Creative								
Uses Special Effects								

4.3.3. Questionari SAM



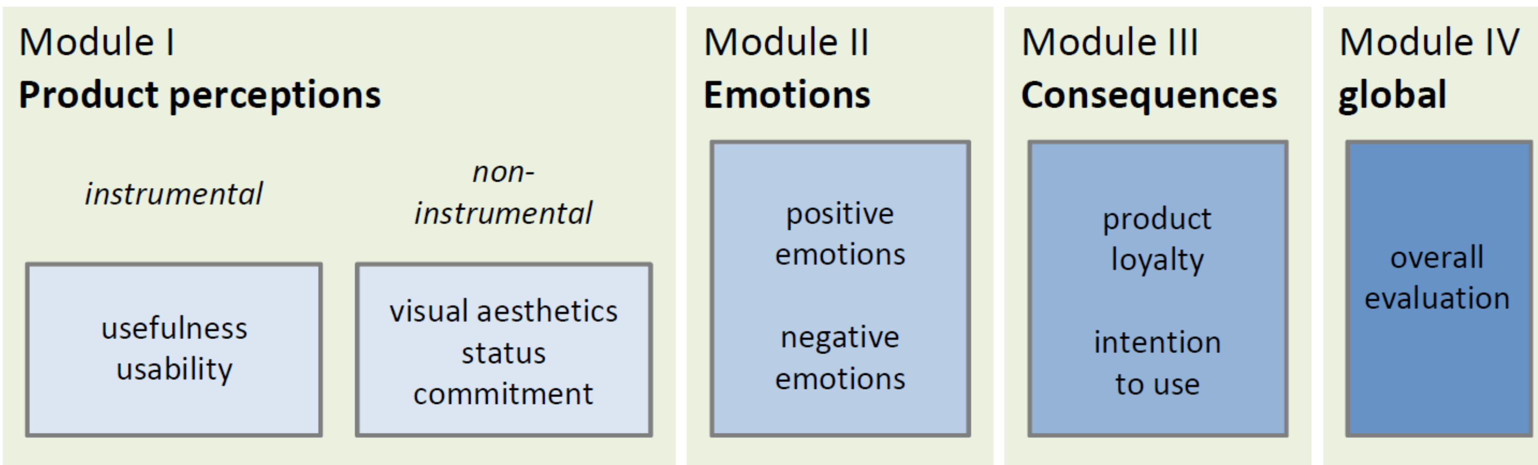
4.3. meCUE

Qüestionari meCUE: Defineix tres mòduls, validats de forma separada:

I. **Percepcions del producte** (instrumentals i no instrumentals)

II. **Emocions**

III. **Conseqüències de la UX**



Flexible: Depenent del que es vulgui mesurar, s'agafen uns mòduls o altres.

Compost de 34 items, contestable entre 2 i 5 minuts. Qüestionari accessible a www.mecue.de/english amb Excel que realitza l'anàlisi.

4.3. meCUE

Qüestionari meCUE:

Instrumental qualities

Item
With the help of this product, I will achieve my goals.
I consider the product extremely useful.
The functions of the product are exactly right for my goals.
The product is easy to use.
The operating procedures are simple to understand.
It is quickly apparent how to use the product.

Non-instrumental qualities

Item
The design looks attractive.
The product is creatively designed.
The product is stylish.
By using the product, I would be perceived differently.
The product would enhance my standing among peers.
I would not mind if my friends envied me for this product.
I could not live without the product.
The product is like a friend to me.
If I ever lost the product, I would be devastated.

4.3. meCUE

Evaluation of interactive products

meCUE 2.0 questionnaire

This questionnaire serves to assess how you experience the product.

On the following pages, there are a number of statements that you can use to judge your experience.

Please, express the degree of your agreement with each statement by ticking the ac-cording circle.

Decide spontaneously and without long contemplation to convey your first impression. Please judge each statement even if you feel that it does not com-pletely match your experience.

There are no right or wrong answers – all that counts is your personal opinion!

id	product	I

	strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
The product is easy to use.	○	○	○	○	○	○	○
The functions of the product are exactly right for my goals.	○	○	○	○	○	○	○
It is quickly apparent how to use the product.	○	○	○	○	○	○	○
I consider the product extremely useful.	○	○	○	○	○	○	○
The operating procedures of the product are simple to understand.	○	○	○	○	○	○	○
With the help of this product I will achieve my goals.	○	○	○	○	○	○	○

id	product	II

	strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
The product is creatively designed.	○	○	○	○	○	○	○
The product would enhance my standing among peers.	○	○	○	○	○	○	○
I could not live without this product.	○	○	○	○	○	○	○
The design looks attractive.	○	○	○	○	○	○	○
By using the product, I would be perceived differently.	○	○	○	○	○	○	○
The product is like a friend to me.	○	○	○	○	○	○	○
The product is stylish.	○	○	○	○	○	○	○
If I ever lost the product, I would be devastated.	○	○	○	○	○	○	○
I would not mind if my friends envied me for this product.	○	○	○	○	○	○	○

Id	product	V

How do you experience the product as a whole?

as bad -5 -4 -3 -2 -1 0 1 2 3 4 5 **as good**

4.3. meCUE

The items were assigned to the following dimensions accordingly:
Moduel I: Usefulness (F), Usability (U)
Module II: Visual aesthetics (A), Status (S), Commitment (C)
Module III: Positive Emotions (PA, PD), Negative Emotions (NA, ND)
Module IV: Intention to use (IN), Product loyalty (L)
Module V: Overall evaluation

id	product	I

	strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree	
The product is easy to use.	O	O	O	O	O	O	O	U.1
The functions of the product are exactly right for my goals.	O	O	O	O	O	O	O	F.1
It is quickly apparent how to use the product.	O	O	O	O	O	O	O	U.2
I consider the product extremely useful.	O	O	O	O	O	O	O	F.2
The operating procedures of the product are simple to understand.	O	O	O	O	O	O	O	U.3
With the help of this product I will achieve my goals.	O	O	O	O	O	O	O	F.3

id	product	II

	strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree	
The product is creatively designed.	O	O	O	O	O	O	O	A.1
The product would enhance my standing among peers.	O	O	O	O	O	O	O	S.1
I could not live without this product.	O	O	O	O	O	O	O	B.1
The design looks attractive.	O	O	O	O	O	O	O	A.2
By using the product, I would be perceived differently.	O	O	O	O	O	O	O	S.2
The product is like a friend to me.	O	O	O	O	O	O	O	B.2
The product is stylish.	O	O	O	O	O	O	O	A.3
If I ever lost the product, I would be devastated.	O	O	O	O	O	O	O	B.3
I would not mind if my friends envied me for this product.	O	O	O	O	O	O	O	S.3

Id	product	V

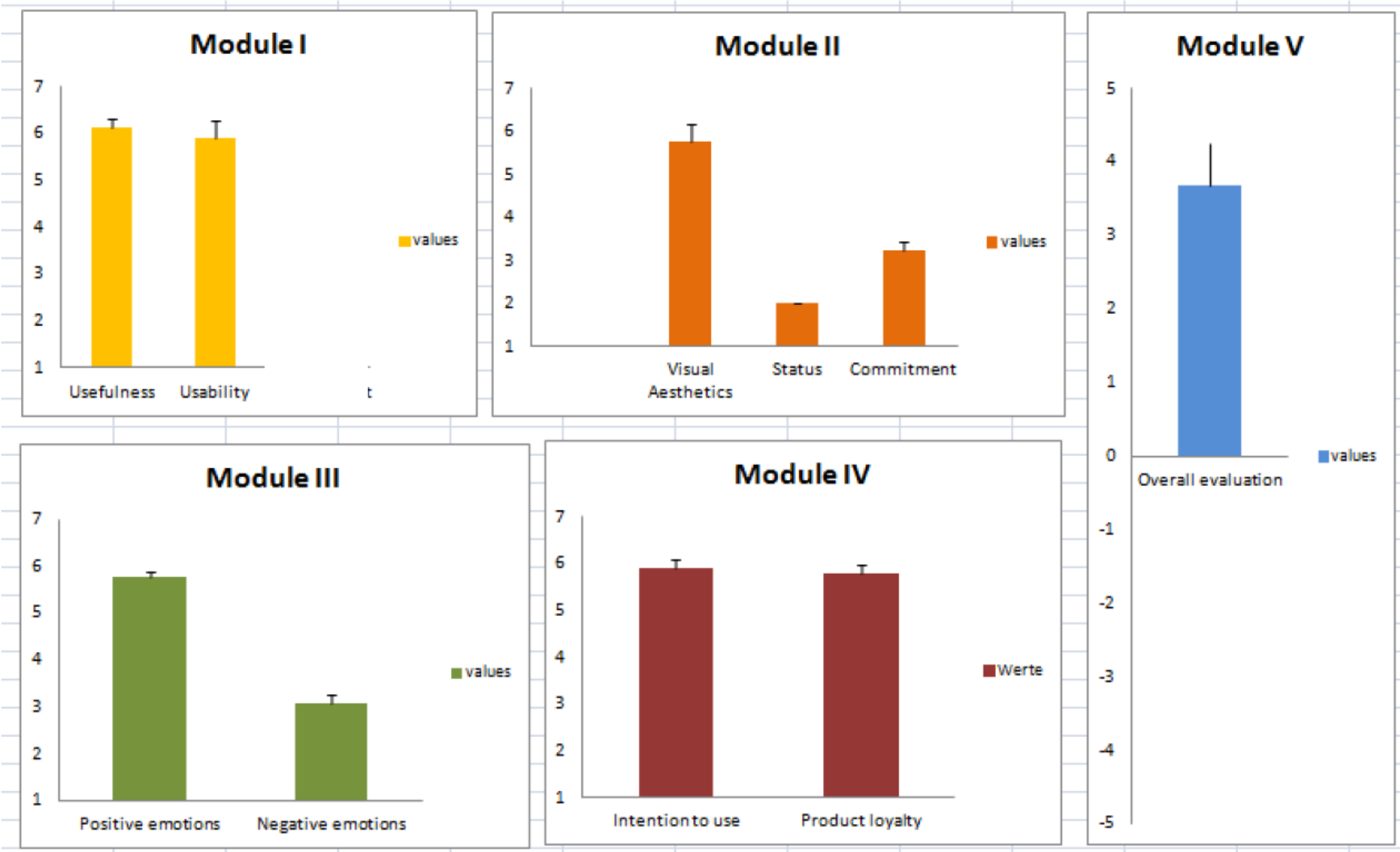
How do you experience the product as a whole?

as bad -5 -4 -3 -2 -1 0 1 2 3 4 5 as good

Overall evaluation

4.3. meCUE

Exemple d'avaluació:



5. Models d'UX per a Visualització de Dades

A l'hora d'avaluar sistemes de Visualització de Dades és important distingir entre característiques instrumentals / pragmàtiques i no instrumentals / hedòniques.

Instrumentals:

- **Utilitat:** Satisfer les necessitats dels usuaris, que les dades siguin útils
- **Completitud:** Que la informació sigui correcte i en quantitat justa
- **Percepció:** Mostrar la informació que sigui comprensible amb el mínim esforç
- **Confiança:** Que la informació sigui vàlida, acurada
- **Intuïció:** Que la informació sigui intuïtiva, fàcil de comprendre

No Instrumentals:

- **Estètica:** Agradable a la vista
- **Atractiva, compromesa:** Que usuari s'interessi en la informació

5. Models d'UX per a Visualització de Dades

El pes d'avaluació entre característiques instrumentals o no instrumentals depèn de l'ús que se'n faci del sistema a avaluar:

- En **websites** predominen **no instrumentals** doncs l'activitat important és la navegació, no realitzar accions específiques.
- En **sistemes basats en accions** a fer (cas dels sistemes de **Visualització de Dades**) les **instrumentals** predominen, doncs cal optimitzar la realització d'aquestes accions.

Per a determinar el model UX millor per a VD, distingir entre websites i sistemes VD, doncs models UX molt basats en avaluar websites:

- **Websites:** Basats en text, es mostren dades qualitatives (vídeos, fotos). El més important és el contingut que no la forma de mostrar-ho.
- **Sistemes VD:** Basats en gràfics, es mostren dades quantitatives, pel que és important com es mostren. Cal que model UX incorpori aspectes visuals. Usuari no s'ha d'identificar amb les dades.

5. Models d’UX per a Visualització de Dades

Per a sistemes de VD, model i qüestionari més idoni és el meCUE:

- Model Hassenzahl dona massa èmfasi a característiques hedòniques, mentre que en VD és molt important usabilitat
- Dóna un pes suficient a característiques no instrumentals i el qüestionari és molt flexible si cal
- Valora aspectes visuals i estètics
- Model i qüestionari funcionen per a diferents tipus d’aplicacions

En resum…

Important avaluar els sistemes de Visualització de Dades definits, especialment si són interactius.

Termes d'usabilitat i experiència d'usuari (UX) s'utilitzen per avaluar sistemes. Ambdós conceptes es barregen sovint.

UX inclou aspectes d'usabilitat però també UX inclou aspectes estètics, visuals i inclús emocionals. Diferents models UX d'avaluació, la majoria associats a qüestionaris.

El model UX i qüestionari meCUE és el més adient per a avaluar sistemes de Visualització de Dades, doncs pondera correctament usabilitat amb estètica visual.

Referències

[BHK-11] M. Burmester, M. Hassenzal, F. Koller, **AttrakDiff: Questionnaire**, 2011, https://www.kompetenzzentrum-usability.digital/kos/WNetz?art=File.download&id=1296&name=AttrakDiff_EN_UID.pdf, 2011 (data darrer accés: Maig 2025).

[BrL-94] M.M. Bradley, P.J. Lang, **Measuring emotion: the self-assessment manikin and the semantic differential**, Journal of behavior therapy and experimental psychiatry, 25(1), 49-59, 1994.

[DLQ-19] I. Diaz-Oreiro, G. López, L. Quesada, L.A. Guerrero, **Standarized Questionnaires for User Experience Evaluation: A systematic Literature Review**, Proceedings MDPI, 2019, DOI: 10.3390/proceedings2019031014 (data darrer accés: Maig 2025).

[LSH-08] B. Laugwitz, M. Schrepp, T. Held, **Construction and evaluation of a user experience questionnaire,** A. Holzinger (ed.), USAB 2008, Lect. Notes in Comp. Sc. 5298, pp. 63-76, 2008.

[LaT-04] B. Lavie, N. Tractinsky, **Assessing dimensions of perceived visual aesthetics of web sites,** Int. Journal Human-Computer Studies 60, 269-298, 2004.

[MiT-16] M. Minge, M. Thüring, I. Wagner, C.V. Kuhr, **The meCUE Questionnaire: A Modular Tool for Measuring User Experience,** Proceedings of the 7th Applied Human factors and Ergonomics Society Conference, pp. 115-128, Switzerland, 2016.

[Sch-19] M. Schrepp, **User Experience Questionnaire Handbook,** https://www.ueq-online.org/Material/Handbook.pdf, 2019 (data darrer accés: Maig 2025).

[SUMI] SUMI Questionnaire, http://sumi.uxp.ie/en/ (data darrer accés: Maig 2025).

[WVK-14] T. Walsh, J. Varsaluoma, S. Kujala, P. Nurkka, H. Petrie, C. Power, **Are UX: Exploring Long-term User Experience with iScale and AttrakDiff**, AcademicMindTrek'14, Finland, 2014, DOI:10.1145/2676467.2676480 (data darrer accés: Maig 2025).