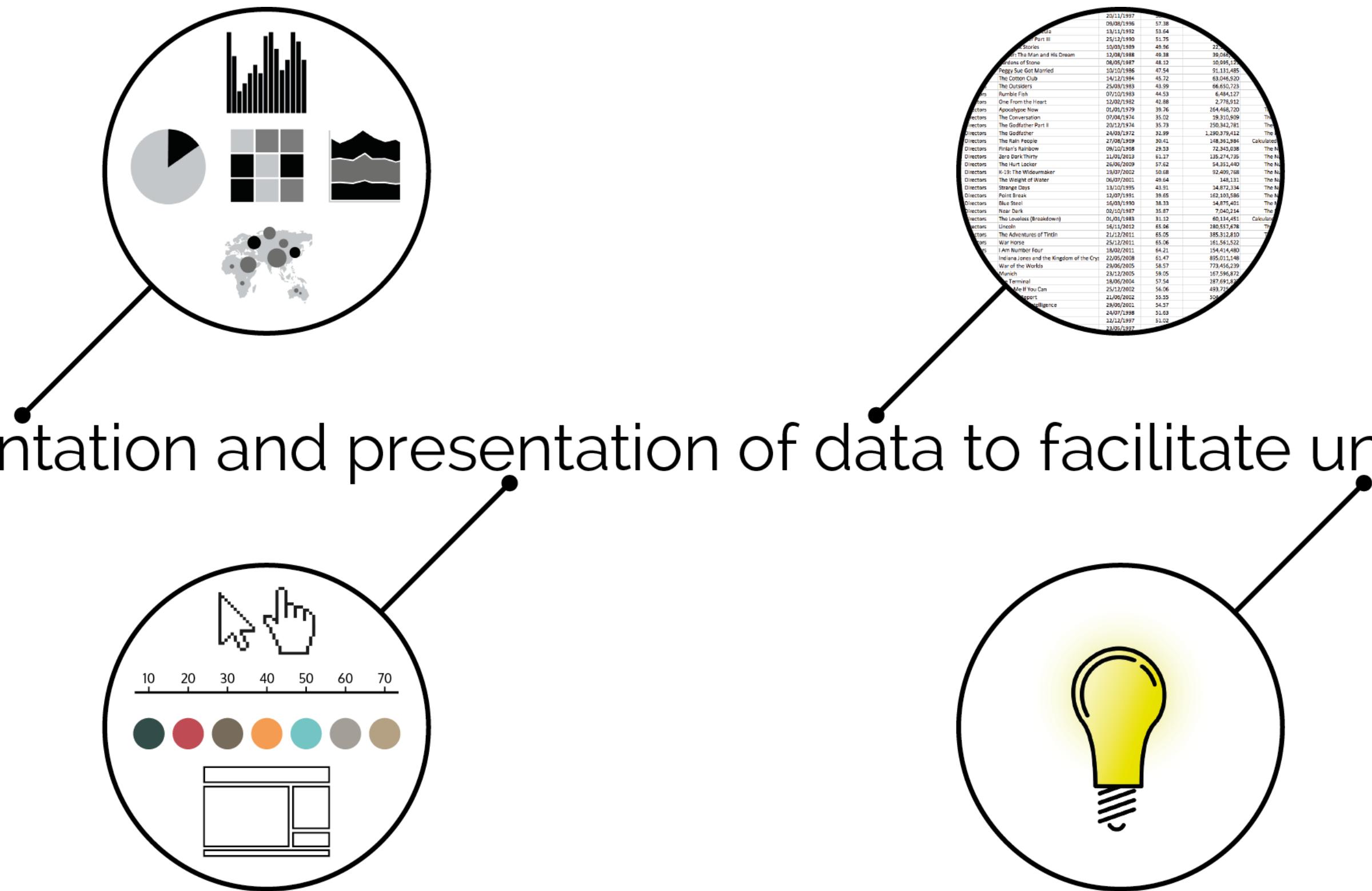


visualisation



The representation and presentation of data to facilitate understanding

ingredients

data is the key actor in the dataviz process

tabular form is what we consider the raw form of data

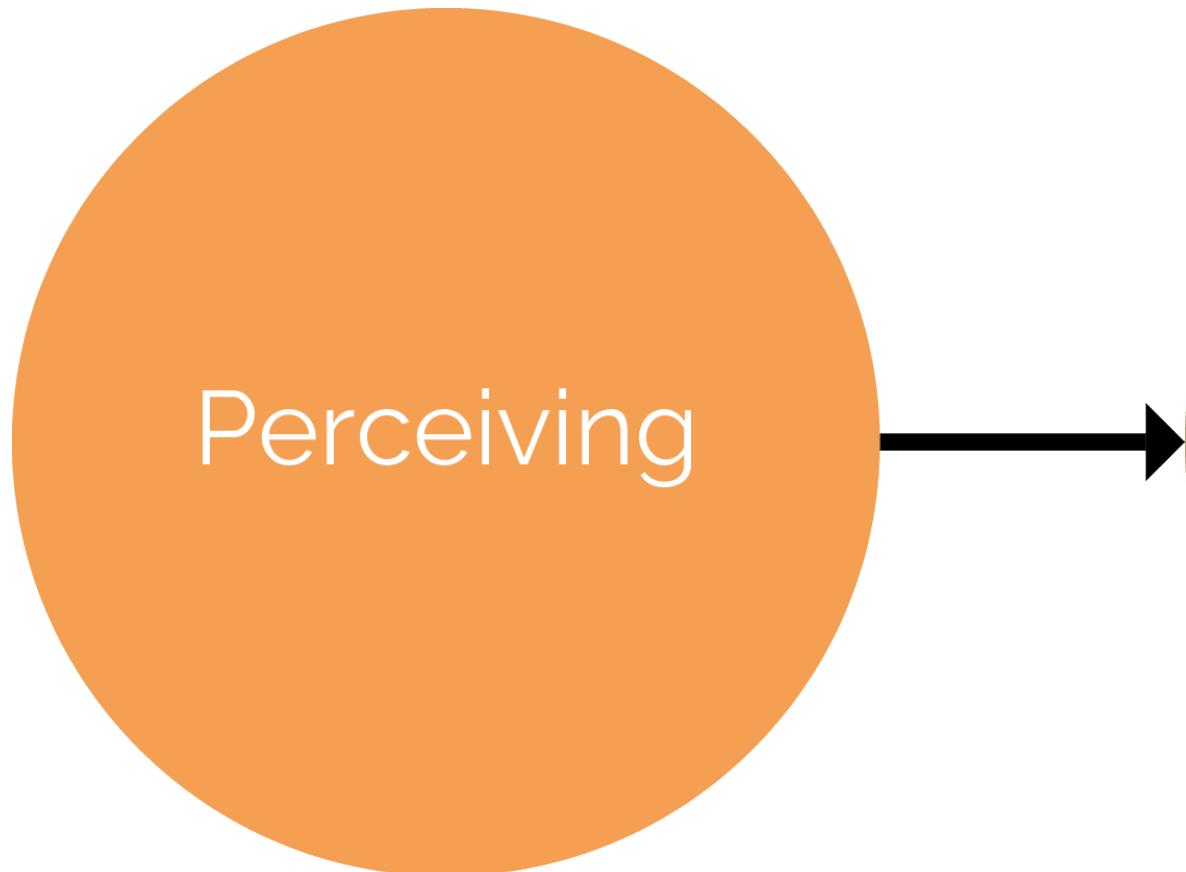
raw data cannot be compared and contrasted effectively

data representation is the visualization of data through the combination of marks and attributes

data presentation concerns all other visible design decision beyond the representation: interactivity, annotations, colours

facilitating **understanding** should be the goal

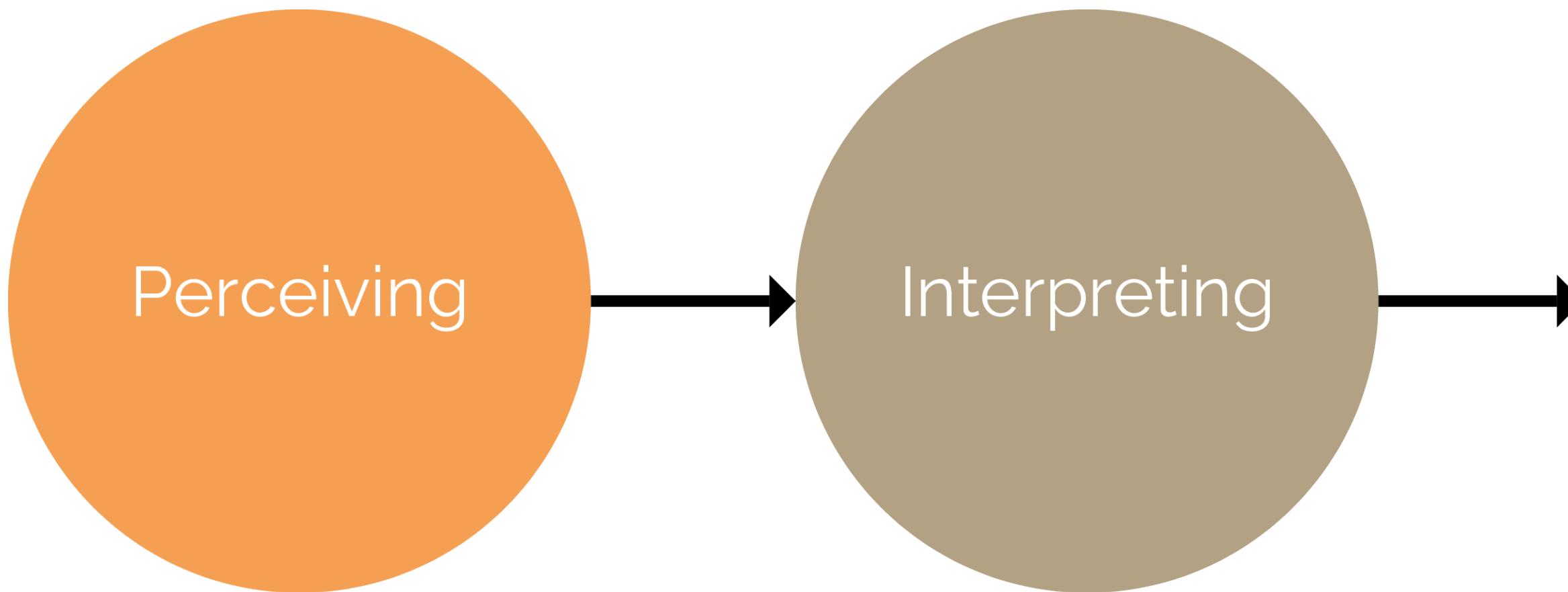
understanding



What does it **show?**

Where is big, medium, small?
How do things compare?
What relationships exist?

understanding



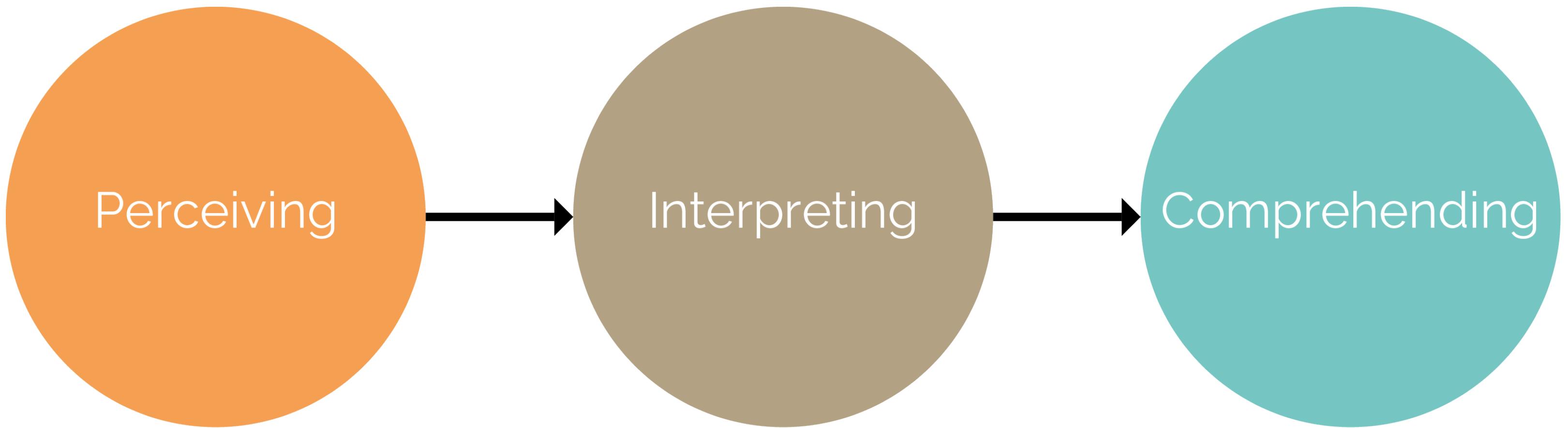
What does it show?

Where is big, medium, small?
How do things compare?
What relationships exist?

What does it mean?

What is good and bad?
Is it meaningful or insignificant?
Unusual or expected?

understanding



What does it show?

Where is big, medium, small?
How do things compare?
What relationships exist?

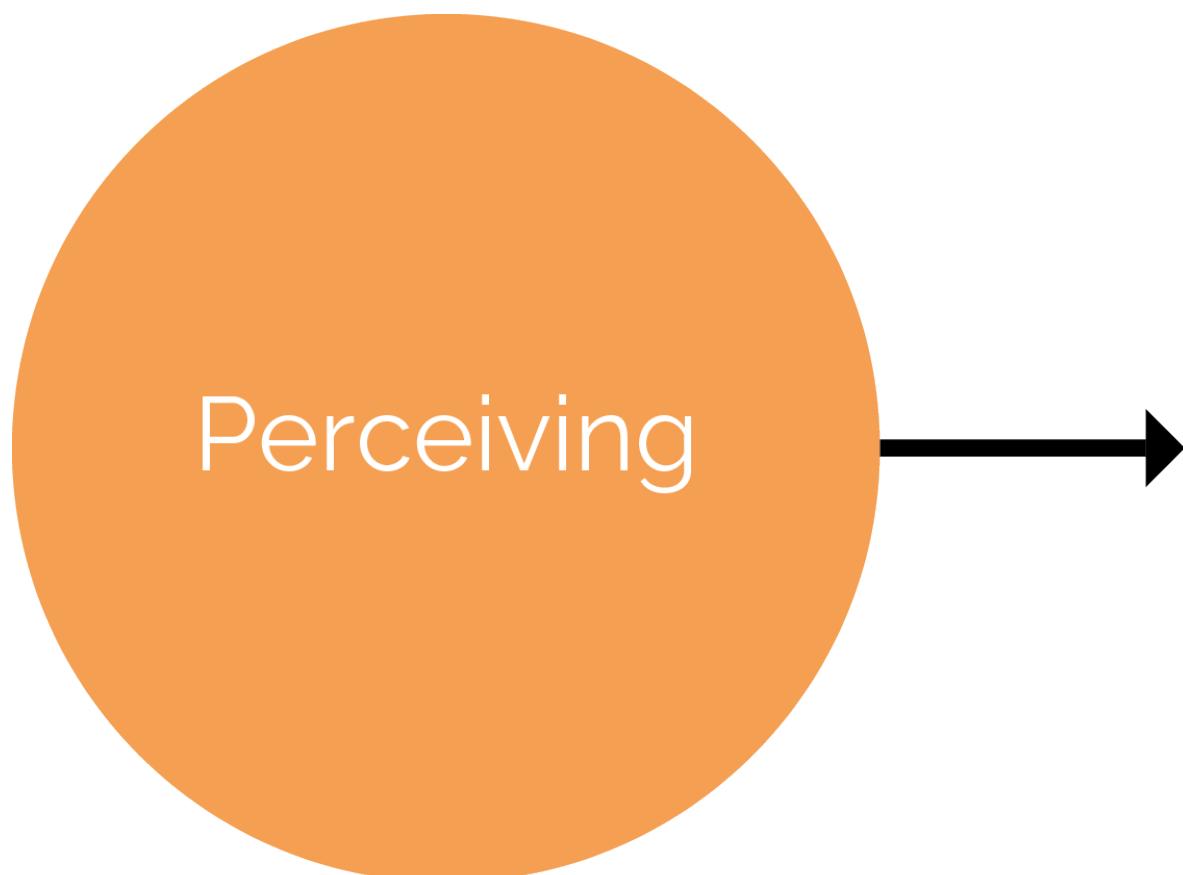
What does it mean?

What is good and bad?
Is it meaningful or insignificant?
Unusual or expected?

What does it mean to me?

What are the main messages?
What have I learnt?
Any actions to take?

perceiving



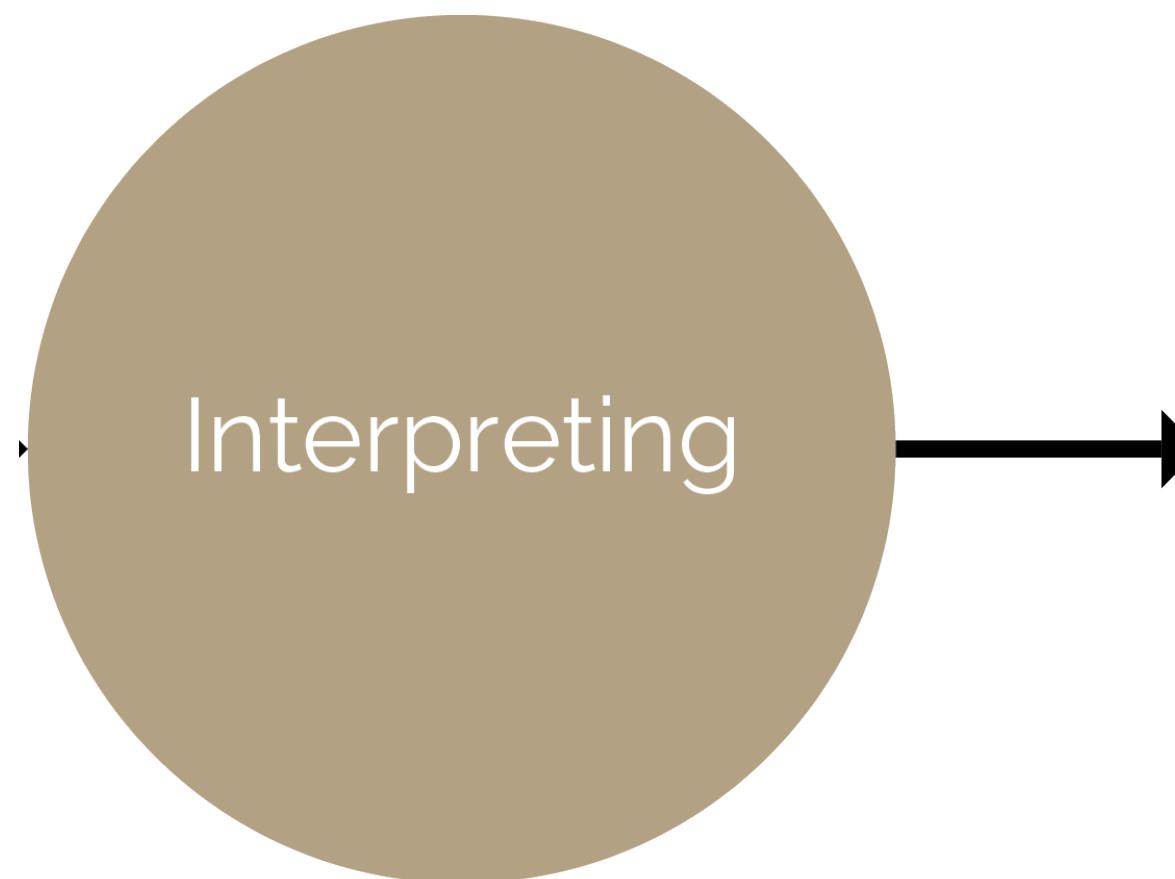
What does it **show**?

Where is big, medium, small?
How do things compare?
What relationships exist?

- where are the largest/middle-sized/smallest values?
- what proportion of the total does that value hold?
- how do these values compare in ranking terms?
- to which other values does this have a connected relationship?

the art of simply being able to read a chart
efficiently decode the representation of the data (shapes/sizes/colors)

interpreting

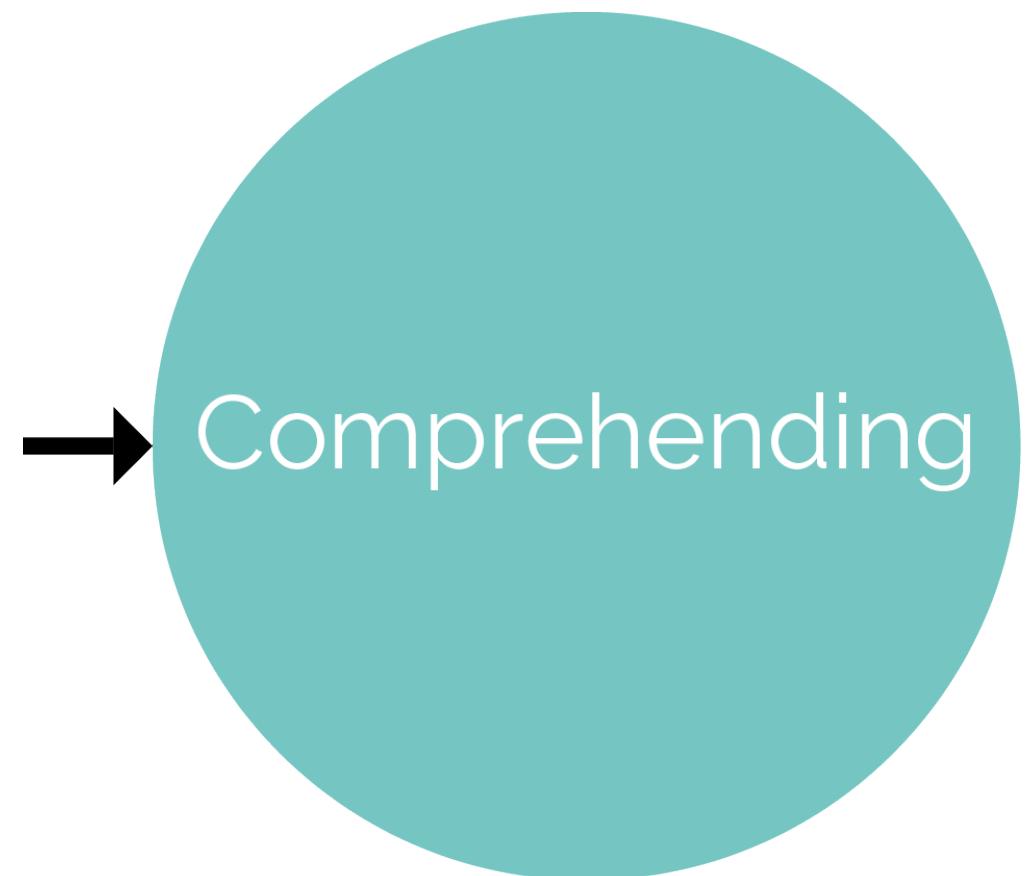


What does it **mean**?
What is good and bad?
Is it meaningful or insignificant?
Unusual or expected?

- is it good to be big or it is better to be small?
- what does it mean to go up or to go down?
- is that relationship meaningful or insignificant?
- is the decline/increase of that category especially surprising?

the art of converting the perception into meaning
use pre-existing knowledge to frame the implication of the viz

comprehending



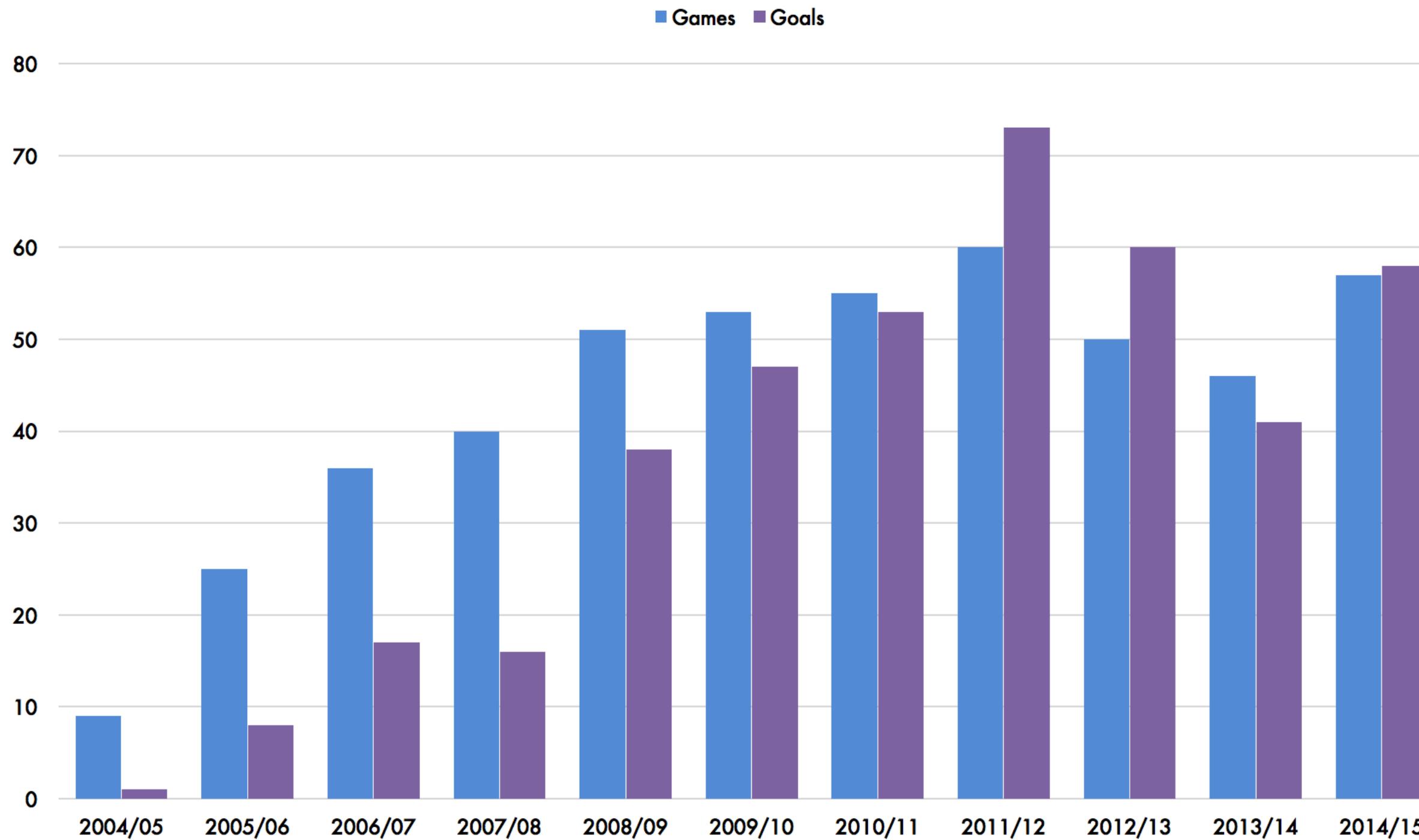
What does it **mean to me?**
What are the main messages?
What have I learnt?
Any actions to take?

- why is this relevant? to whom?
- has it confirmed what I suspected or enlightened me with new knowledge?
- has this impacted me emotionally or left me indifferent?
- does this new understanding force me to take action on the subject?

the act of reasoning about the consequences of interpretation
what is the novelty carried by this dataviz about the subject?

Lionel Messi: Games and Goals for FC Barcelona

example

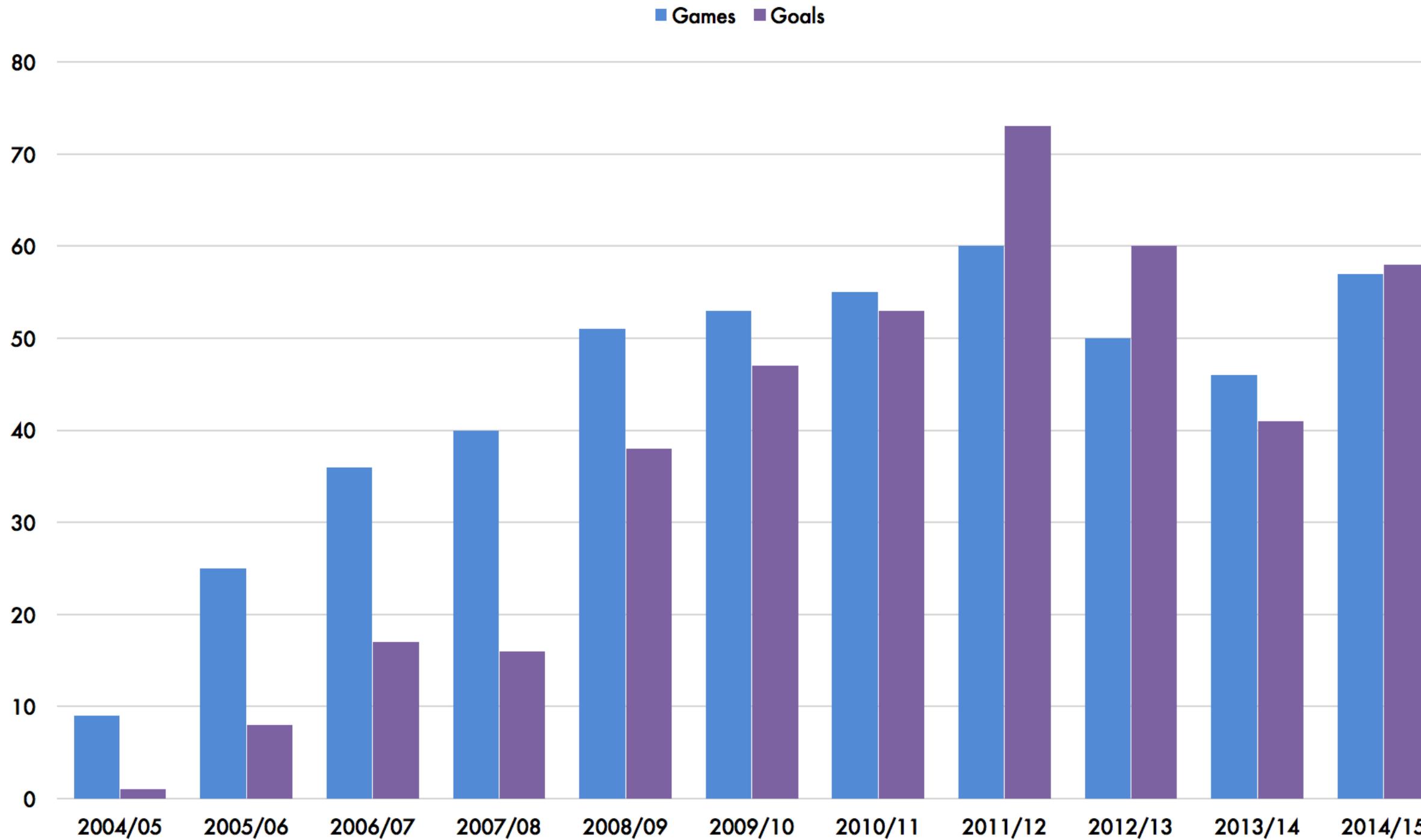


perceiving

- Identifying chart type, axes and labels
- identifying big, small and medium values
- identifying trends for both columns
- making higher/lower comparison

Lionel Messi: Games and Goals for FC Barcelona

example

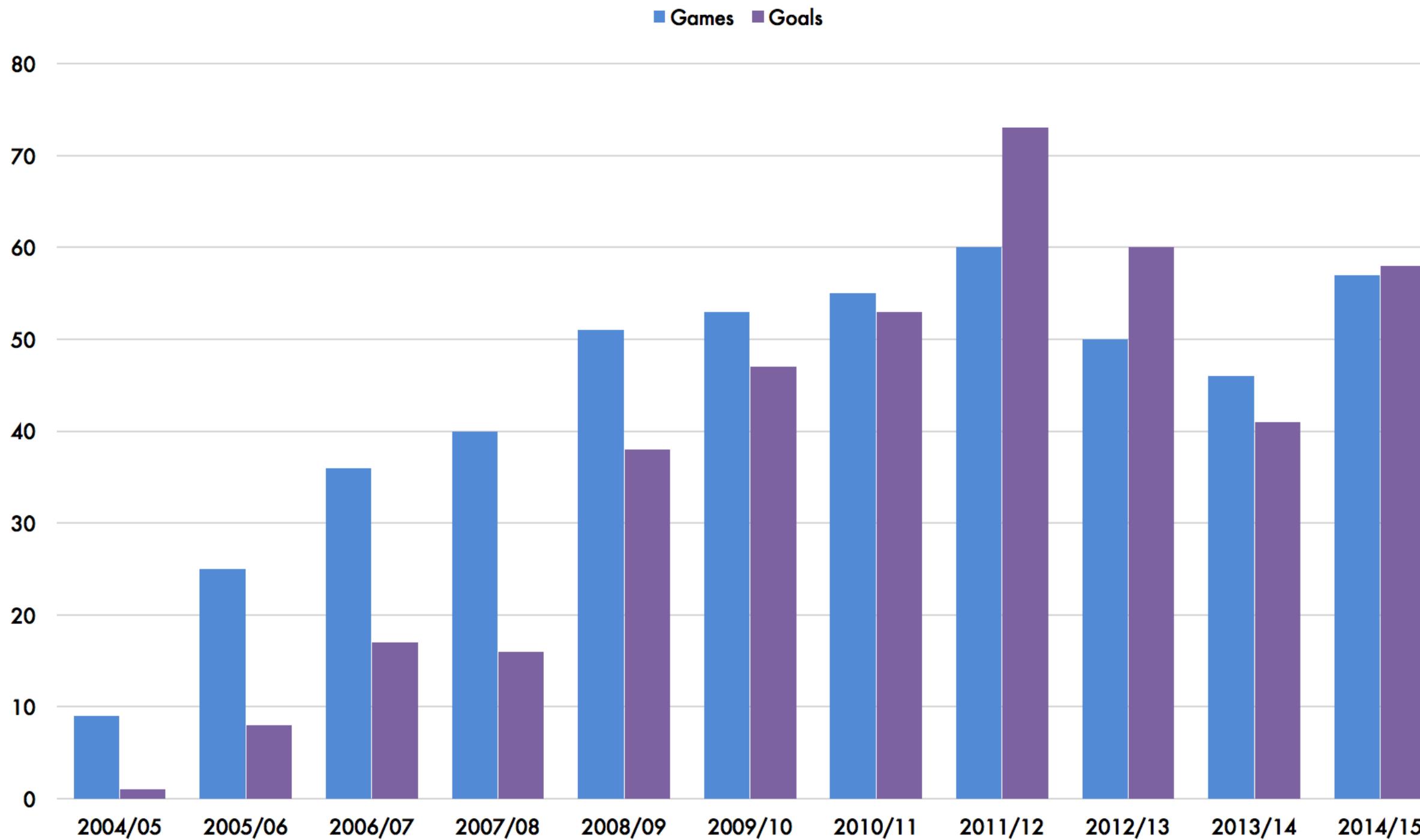


interpreting

- previous knowledge: scoring 25 goals is good, scoring 70 is amazing
- scoring them in la lira / ucl is even more amazing
- a ratio of ~1 goal/game is very rare in football
- even more remarkable if coupled with messi's age

Lionel Messi: Games and Goals for FC Barcelona

example

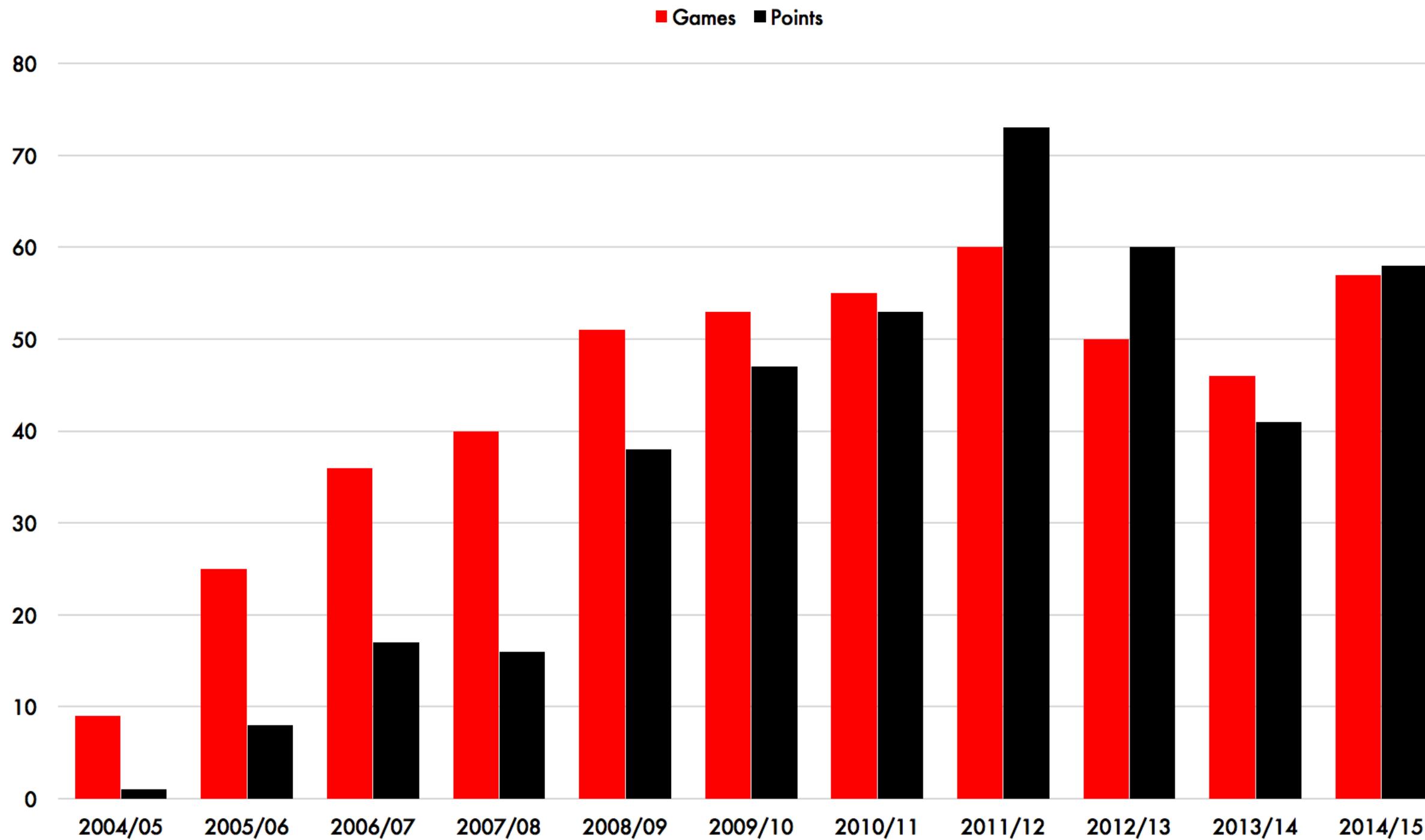


comprehending

- enforce my belief of rating messi one of the top world player
- what is related to the decreasing performances in 2012/2013 & 2013/2014?

Wayne Kane: Games and Points for Toronto Rangers

example



perceiving

interpreting

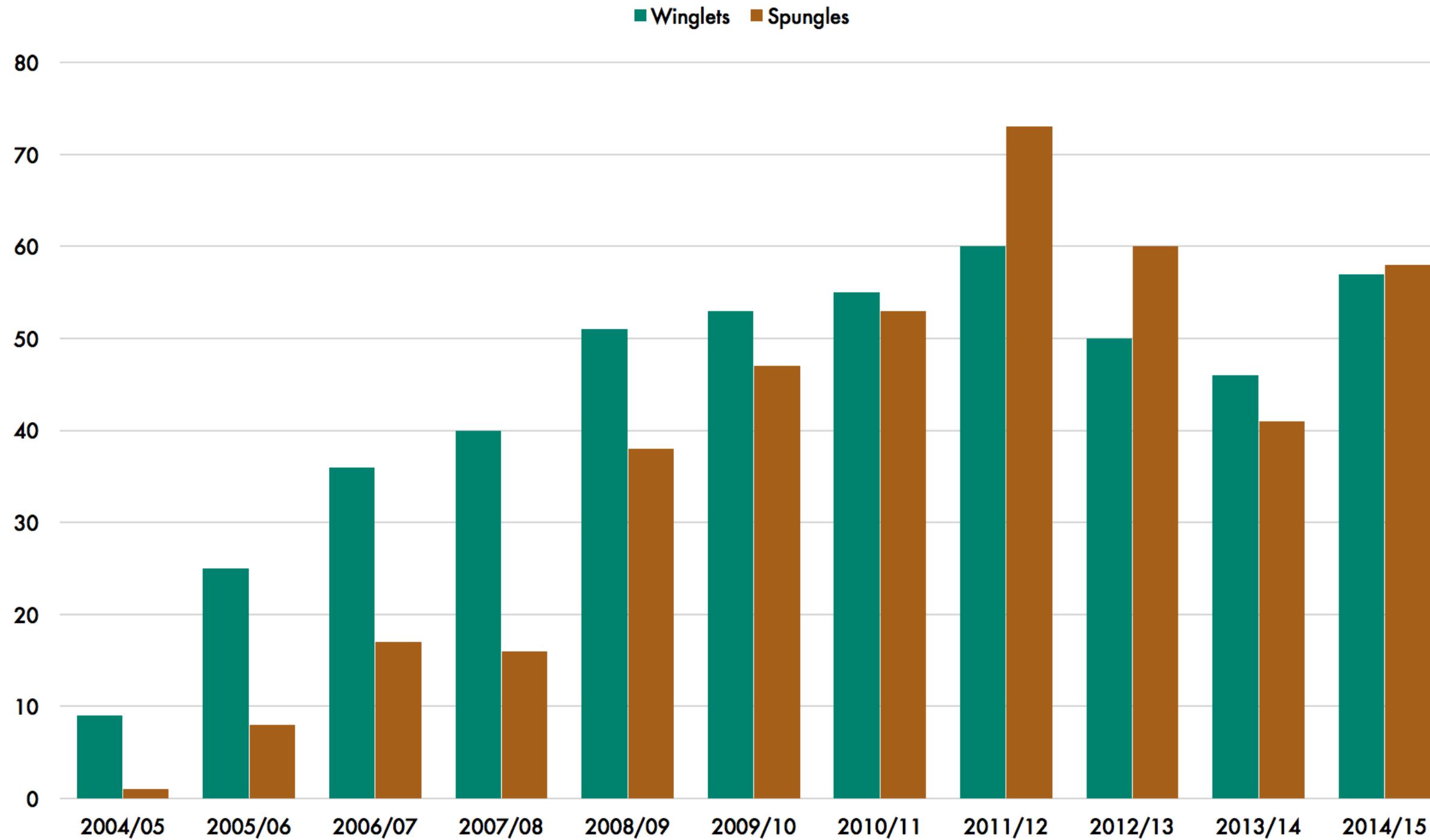
comprehending



if I am no hockey expert

Total Sightings of Winglets and Spungles

example



perceiving

interpreting

comprehending



wwii example



abraham wald



1902



1950

ph.d. maths, uni vienna, 1928

member of columbia us
statistical research group
during wii

how to protect
bombers to get shot by
german planes?

How to optimise position of
additional protecting armor?



wwii example



data...

The hits on the returning planes were distributed as follows:

<u>Part number</u>	<u>Number of hits observed on part</u>	<u>Ratio of number of hits observed on part to total number of observed hits ($\delta(i)$)</u>
1	39	.193
2	78	.386
3	31	.154
4	<u>54</u>	.267
Total number of hits	202	

...model...

Adding $x_{i+1} \left(1 - \frac{1}{p_{i+1}}\right) = \frac{q_{i+1}}{p_{i+1}} x_{i+1}$ to both sides of equation 13, we obtain

$$x_1 + \dots + x_{i+1} + c_{i+1} \cdot \frac{q_{i+1}}{p_{i+1}} x_{i+1} \quad (i = 3, 4, \dots, n). \quad (14)$$

From equations 11 and 14, we obtain

$$x_1 + p_1 \left(c_{i+1} - \frac{q_{i+1}}{p_{i+1}} x_{i+1}\right) + p_1 c_1.$$

Hence,

$$x_1 = p_1 c_1 + c_{i+1} + \frac{p_1 q_{i+1}}{p_{i+1}} x_{i+1} \quad (i = 3, 4, \dots, n).$$

Let

$$d_1 = p_1 c_1 + c_{i+1} = -p_1 x_{i+1} \quad (i = 3, 4, \dots, n)$$

and

$$t_1 = \frac{p_1 q_{i+1}}{p_{i+1}} \quad (i = 3, 4, \dots, n).$$

Then equation 16 can be written as

$$x_1 = d_1 + t_1 x_{i+1} \quad (i = 3, 4, \dots, n). \quad (15)$$

Denote $p_1(1 - a_0)$ by d_1 , $-p_1 x_{i+1}$ by t_1 , and $\frac{p_1 q_{i+1}}{p_{i+1}}$ by t_2 , then we have

$$x_1 = d_1 + x_2 + t_2 x_3 + d_2 + \dots + t_n x_n.$$

for $n = 5$, we obtain

$$\begin{aligned} x_1 &= p_1(1 - a_0) = .030 \\ x_2 &= p_1(1 - a_0 - a_1) = .012 \\ x_3 &= p_1(1 - a_0 - a_1 - a_2) = .004 \\ x_4 &= p_1(1 - a_0 - a_1 - a_2 - a_3) = .002 \\ x_5 &= p_1(1 - a_0 - a_1 - a_2 - a_3 - a_4) = .001 \end{aligned}$$

The value of t_i in the second example is nearly equal to the value in the first example in spite of the fact that the values a_i ($i = 0, 1, \dots, 2$) differ considerably. The difference in the two examples is too example, mainly due to the fact that the probability that a plane will receive a hit is much smaller in the second example than in the first example. Note that the probability that a plane will receive a hit has, of course, no relation to the probability that a plane will be downed if it receives a hit.

PART II
MAXIMUM VALUE OF THE PROBABILITY THAT A PLANE WILL BE DOWNED BY A GIVEN NUMBER OF HITS?

The symbols defined in the previous section in part I will be used here without further explanation. The purpose of this memorandum is to derive the least upper bound of $x_1 + \sum_{j=1}^n x_j$ and that of p_1 ($i = 1, \dots, n$) under the restriction that $q_1 \leq v_2 \leq \dots \leq v_n$.

We note that v_i is a strictly increasing function of p_i for $0 < p_i < 1$. Let us replace p_i by $p_i + \delta$ ($\delta > 0$) and let us study the effect of this change on x_1, \dots, x_n . Denote the changes in x_1, \dots, x_n by $\delta_1, \dots, \delta_n$, respectively. Clearly, $\delta_1 + \dots + \delta_{n-1} = 0$. It follows easily from equation 9 that

$$\delta_2 > 0 \quad \text{and} \quad \delta_{j+1} = -p_{j+1} \delta_j.$$

Hence,

$$\delta_2 + \delta_{j+1} = (1 - p_{j+1}) \delta_j > 0.$$

Similarly, we obtain from equation 9

$$\delta_{j+2} = -p_{j+2}(\delta_j + \delta_{j+1}) = -p_{j+2}(1 - p_{j+1}) \delta_j.$$

Hence,

$$\delta_3 + \delta_{j+1} + \delta_{j+2} = (1 - p_{j+2})(1 - p_{j+1}) \delta_j > 0.$$

In general

$$\delta_j + \delta_{j+1} + \dots + \delta_{j+k} = (1 - p_{j+k})(1 - p_{j+k-1}) \delta_j > 0 \quad (k = 1, \dots, i-1).$$

Hence, we have proved that x_i is a strictly increasing function of p_j ($j = 1, \dots, i$).

This part of "A Method of Estimating Plane Vulnerability Based on Damage of Survivors" was published as SMG memo 67 and AFM memo 74-2.

...solution

ADA091073

LEVEL P2

CRC 422 / July 1980

A REPRINT OF
"A METHOD OF ESTIMATING
PLANE VULNERABILITY BASED
ON DAMAGE OF SURVIVORS"
BY ABRAHAM WALD

Abraham Wald

DTIC ELECTED NOV 3 1980

Approved for Public Release
Distribution Unlimited

MC-FIRE CDR

CCW

CENTER FOR NAVAL ANALYSES ✓
80 10 31 045

wwii example



data...

The hits on the returning planes were distributed as follows:

<u>Part number</u>	<u>Number of hits observed on part</u>	<u>Ratio of number of hits observed on part to total number of observed hits ($\delta(i)$)</u>
1	39	.193
2	78	.386
3	31	.154
4	<u>54</u>	.267
Total number of hits	202	

...model...

X

Adding $x_{i+1} \left(1 - \frac{1}{p_{i+1}}\right) = \frac{\neg q_{i+1}}{p_{i+1}} x_{i+1}$ to both sides of equation 13, we obtain

$$x_1 + \dots + x_{i+1} + c_{i+1} - \frac{q_{i+1}}{p_{i+1}} x_{i+1} \quad (i = 3, 4, \dots, n). \quad (14)$$

From equations 11 and 14, we obtain

$$x_1 + p_1 \left(c_{i+1} - \frac{q_{i+1}}{p_{i+1}} x_{i+1}\right) + p_1 c_1.$$

Hence,

$$x_1 = p_1 (c_1 - c_{i+1}) + \frac{p_1 q_{i+1}}{p_{i+1}} x_{i+1} \quad (i = 3, 4, \dots, n).$$

Let

$$d_1 = p_1 (c_1 - c_{i+1}) + p_2 q_{i+1} \quad (i = 3, 4, \dots, n)$$

and

$$t_1 = \frac{p_1 q_{i+1}}{p_{i+1}} \quad (i = 3, 4, \dots, n).$$

Then equation 16 can be written as

$$x_1 = d_1 + t_1 x_{i+1} \quad (i = 3, 4, \dots, n). \quad (15)$$

Denote $p_1 (1 - a_0)$ by d_1 , $-p_2 q_{i+1}$ by t_2 , and $\frac{p_1 q_{i+1}}{p_{i+1}}$ by t_2 then have

$$x_1 = d_1 + x_2 + t_2 x_3 + d_2 + \dots + t_n x_n.$$

for $n = 5$, we obtain

$$\begin{aligned} x_1 &= p_1 (1 - a_0) = .030 \\ x_2 &= p_2 (1 - a_1 - a_0) = .012 \\ x_3 &= p_3 (1 - a_2 - a_1 - a_0) = .004 \\ x_4 &= p_4 (1 - a_3 - a_2 - a_1 - a_0) = .002 \\ x_5 &= p_5 (1 - a_4 - a_3 - a_2 - a_1 - a_0) = .001 \end{aligned}$$

Example 2n. Let $a_0 = .3$, $a_1 = .2$, $a_2 = .1$, $a_3 = .05$, and $a_4 = .01$. Then the following results are obtained: $q = .17$, $p_1 = .13$, $t_1 = .09$, $x_1 = .05$, $x_2 = .03$, $x_3 = .02$, and $x_4 = .01$.

The value of t_1 in the second example is nearly equal to the value in the first example in spite of the fact that the values a_i ($i = 0, 1, \dots, 3$) differ considerably. The difference in the values of t_1 in the two examples is mainly due to the fact that the probability that a plane will receive a hit is much smaller in the second example than in the first. It is evident from the example that a plane will receive a hit has, of course, no relation to the probability that a plane will be downed if it receives a hit.

PART II
MAXIMUM VALUE OF THE PROBABILITY THAT A PLANE WILL BE DOWNED BY A GIVEN NUMBER OF HITS

The symbols defined in the previous section in part I will be used here without further explanation. The purpose of this memorandum is to derive the least upper bound of $x_i = \sum_{j=1}^i a_j$ and that of p_i ($i = 1, \dots, n$) under the restriction that $q_1 \geq q_2 \geq \dots \geq q_n$.

It is evident that x_i is a strictly increasing function of p_i for $0 < p_i < 1$. Let us replace p_i by $p_i + \delta$ ($\delta > 0$) and let us study the effect of this change on x_1, \dots, x_i . Denote the changes in x_1, \dots, x_i by $\Delta_1, \dots, \Delta_i$, respectively. Clearly, $\Delta_1 + \dots + \Delta_{i-1} = 0$. It follows easily from equation 9 that

$$\Delta_j > 0 \text{ and } \Delta_{j+1} = -p_{j+1} \Delta_j.$$

Hence,

$$\Delta_j + \Delta_{j+1} = (1 - p_{j+1}) \Delta_j > 0.$$

Similarly, we obtain from equation 9

$$\Delta_{j+2} = -p_{j+2} (\Delta_j + \Delta_{j+1}) = -p_{j+2} (1 - p_{j+1}) \Delta_j.$$

Hence,

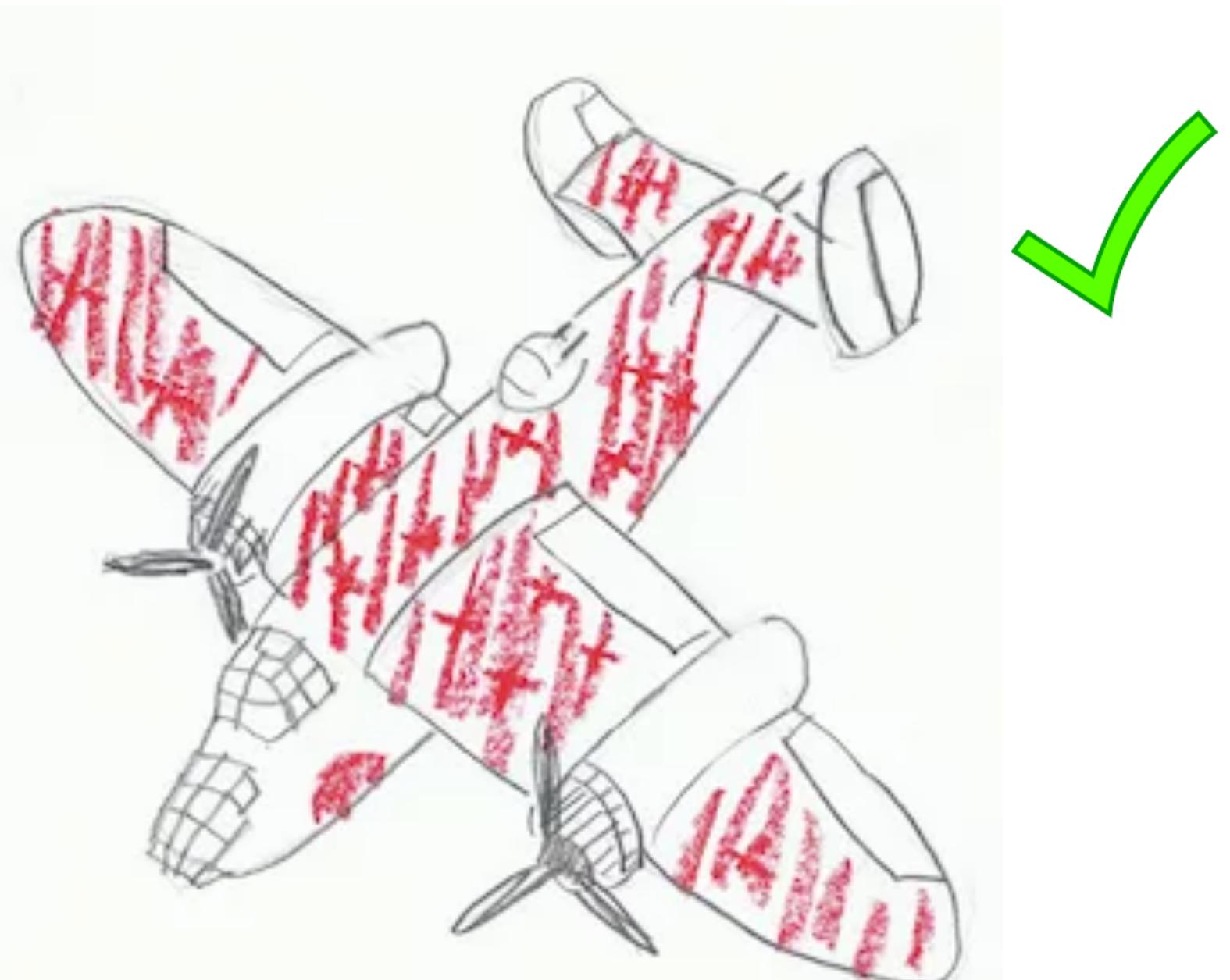
$$\Delta_j + \Delta_{j+1} + \Delta_{j+2} = (1 - p_{j+2}) (1 - p_{j+1}) \Delta_j > 0.$$

In general

$$\Delta_j + \Delta_{j+1} + \dots + \Delta_{j+k} = (1 - p_{j+k}) \dots (1 - p_{j+1}) \Delta_j > 0 \quad (k = 1, \dots, i-1).$$

Hence, we have proved that x_i is a strictly increasing function of p_j ($j = 1, \dots, i$).

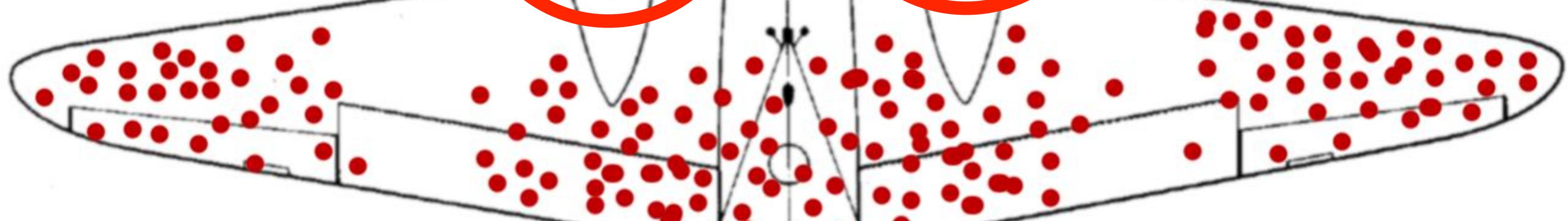
This part of "A Method of Estimating Plane Vulnerability Based on Damage of Survivors" was published as SMG memo 67 and AFM memo 74-2. -11-



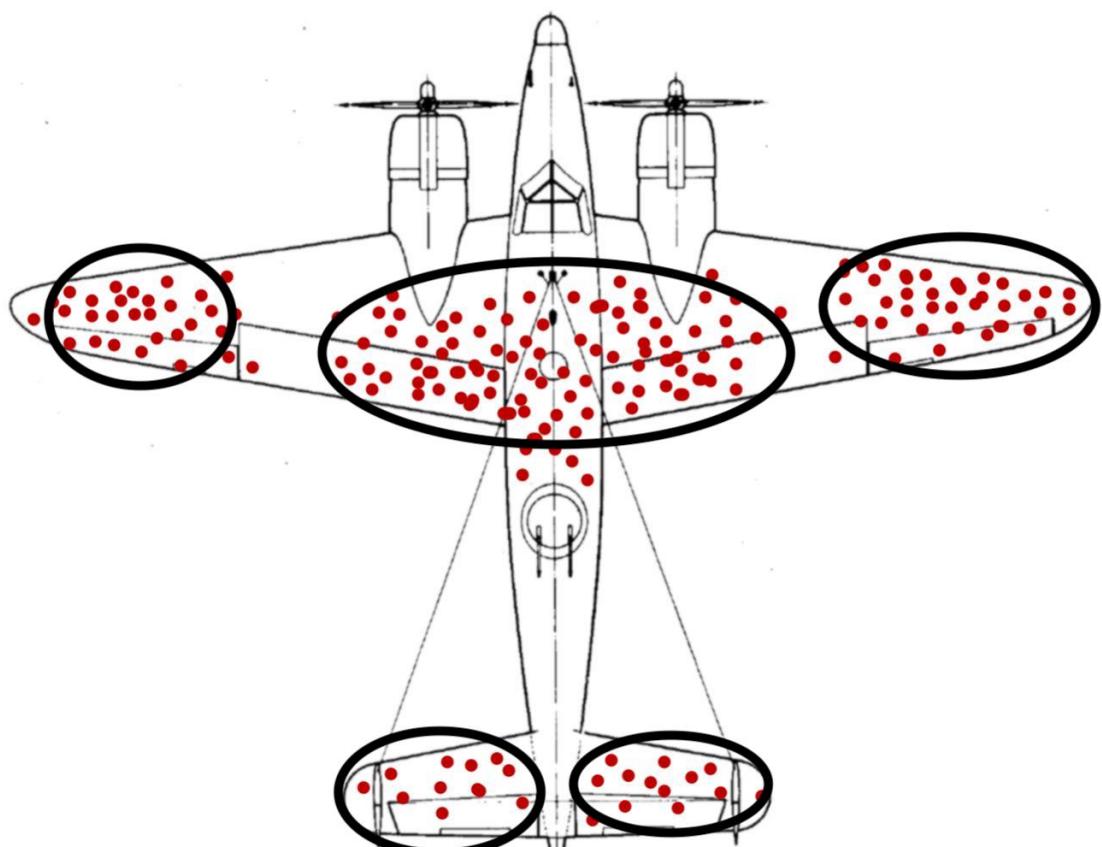


wwii example

where should the new
armors be put?

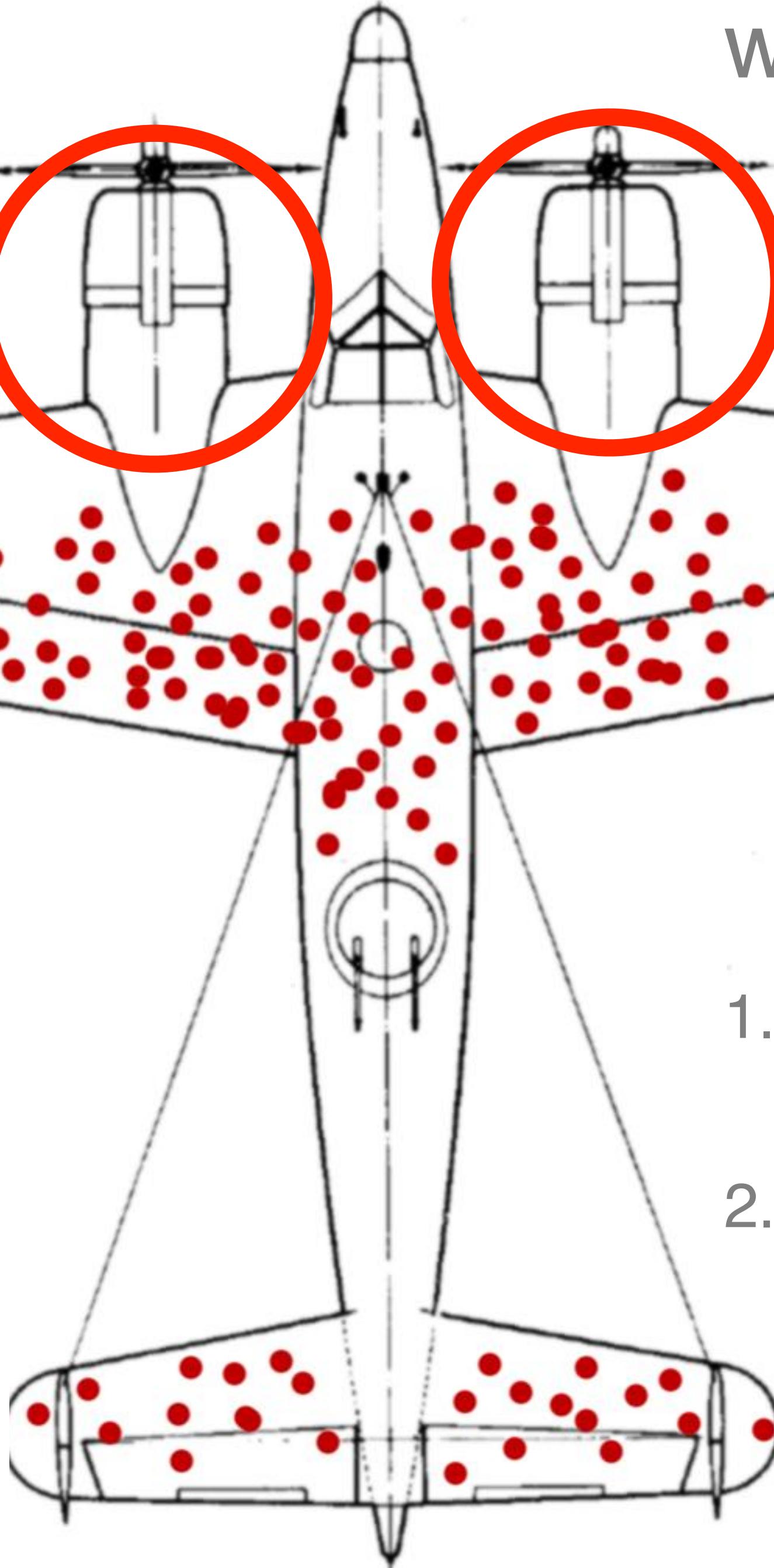


perception



interpretation

1. bullet holes should be equally distributed
2. data only from returning aircrafts

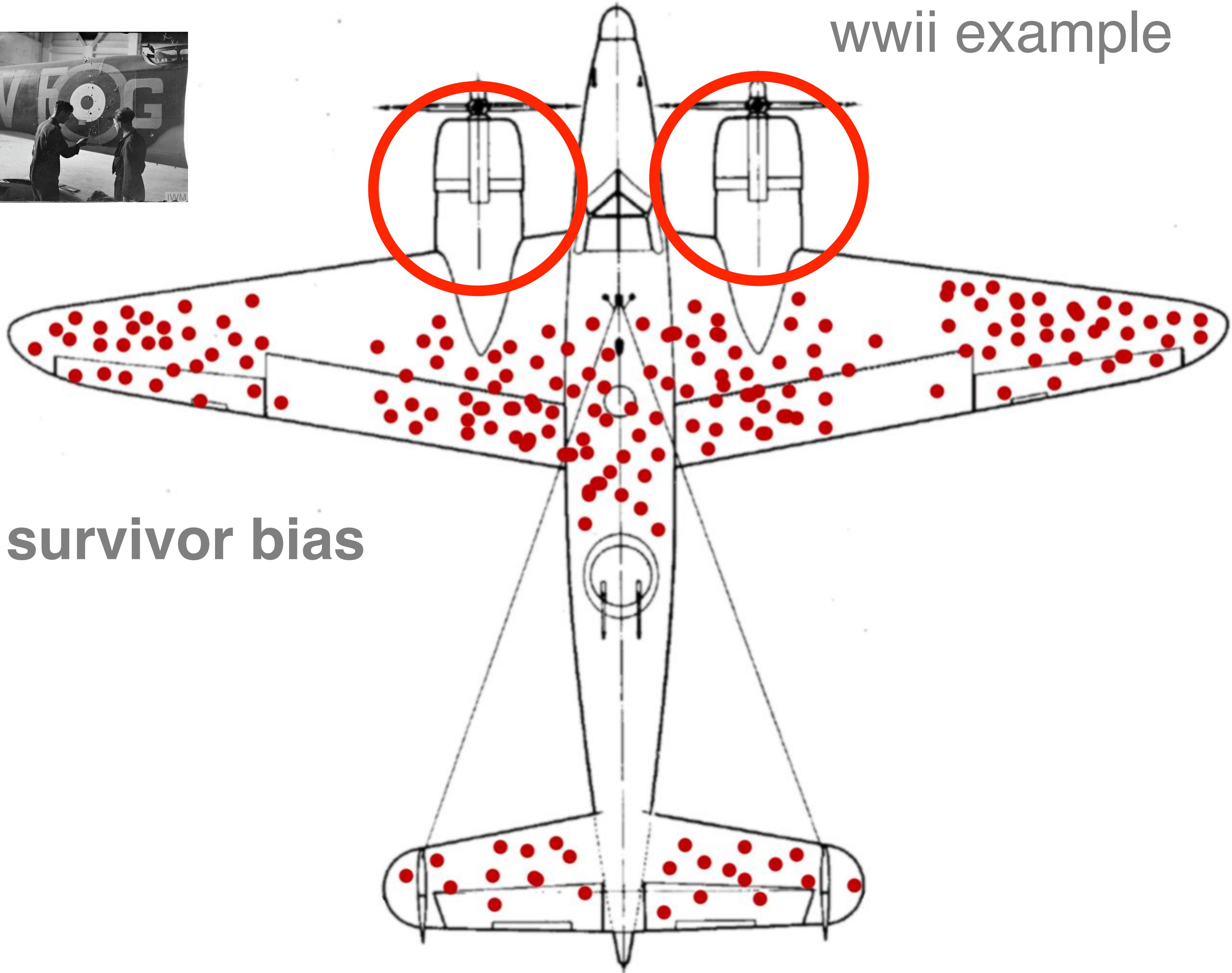


comprehension?



wwii example

survivor bias





principles

inspired by d. rams

principles of good design

good design...

... is innovative

... is honest

... makes a product useful

... is long lasting

... is aesthetic

... is thorough down
to the last detail

... makes a product
understandable

... is environmentally friendly

... is unobtrusive

... is as little design as possible



principles

inspired by d. rams

principles of good dataviz

Principle 1

Good data visualisation
is **TRUSTWORTHY**

Principle 2

Good data visualisation
is **ACCESSIBLE**

Principle 3

Good data visualisation
is **ELEGANT**

Principle 1

Good data visualisation
is **TRUSTWORTHY**

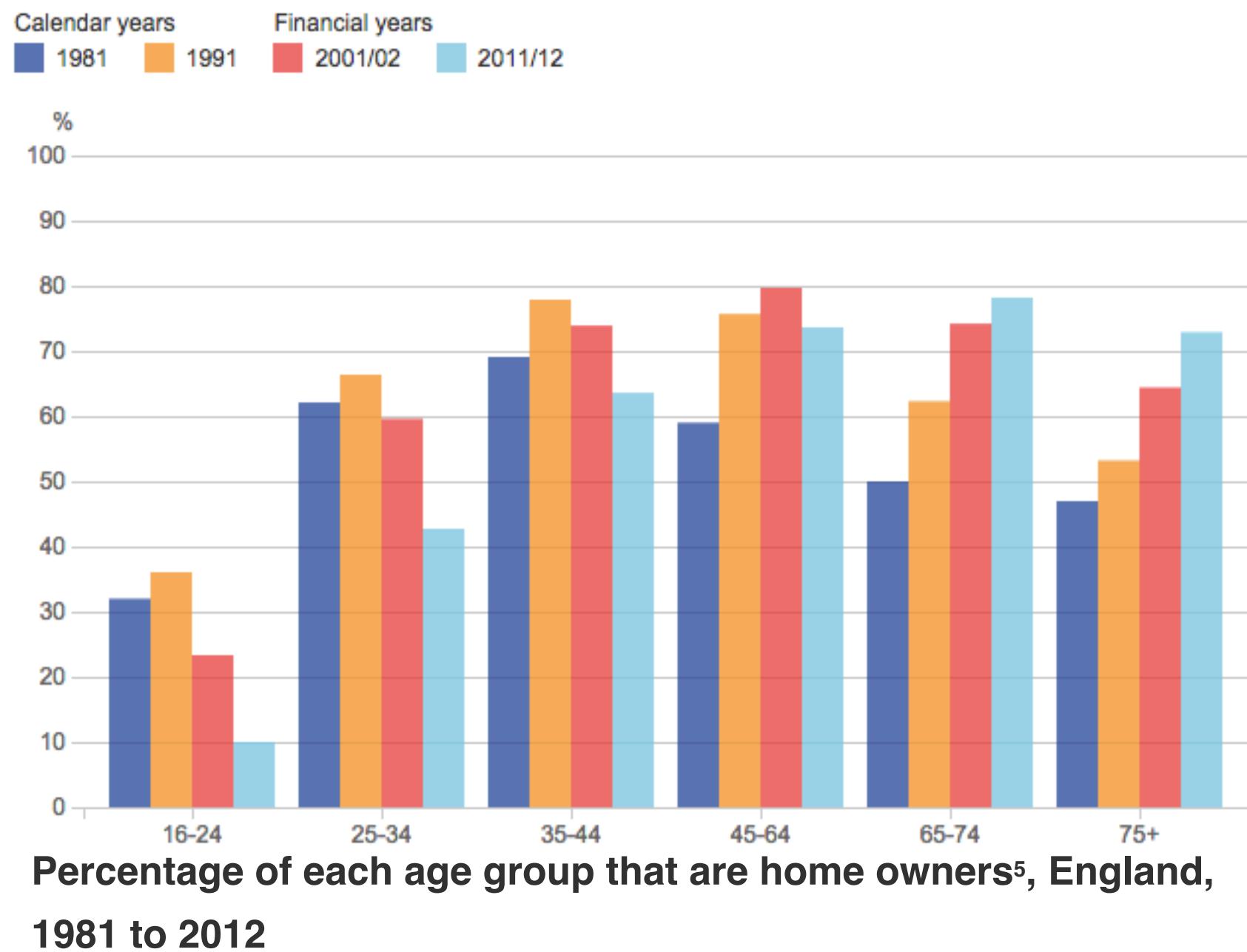
integrity + accuracy + legitimacy

principles

- **truth** is an obligation
- different legitimate versions of truth can exists in dataviz, being the outcome of different pathways
- pure objectivity is rarely possible in dataviz
- there is a need to show that the shown truth is **trustable**
- even a true graph may not be viewed as trustworthy

Principle 1

Good data visualisation
is **TRUSTWORTHY**



Source: English Housing Survey (EHS) 2012 to 2013, Table FC2101, DCLG ; EHS 2001/02, Table S106, DCLG

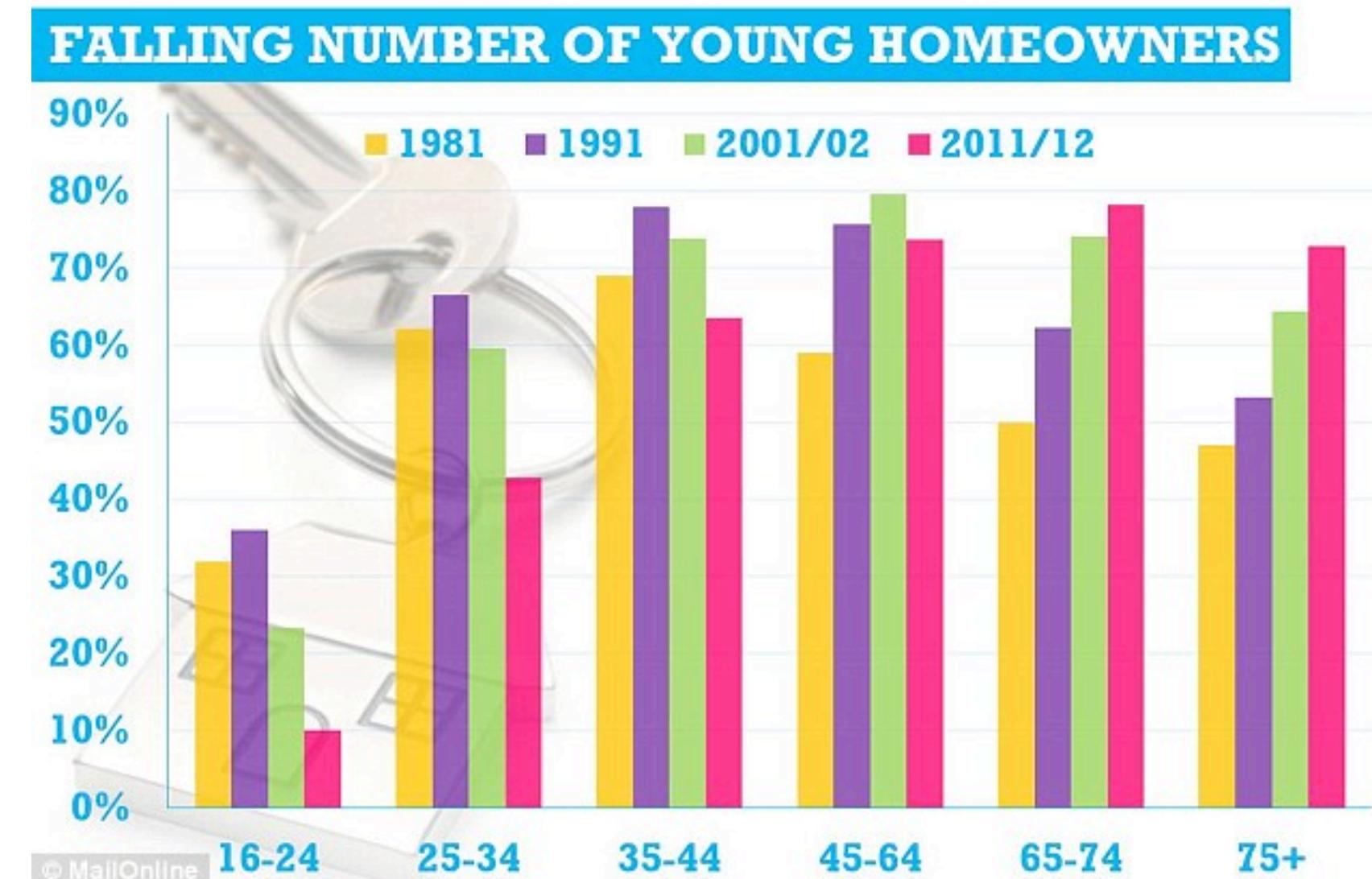
uk office for national statistics

same true data,
different trustworthiness

principles

- colors
- fonts
- background
- data description
- data source

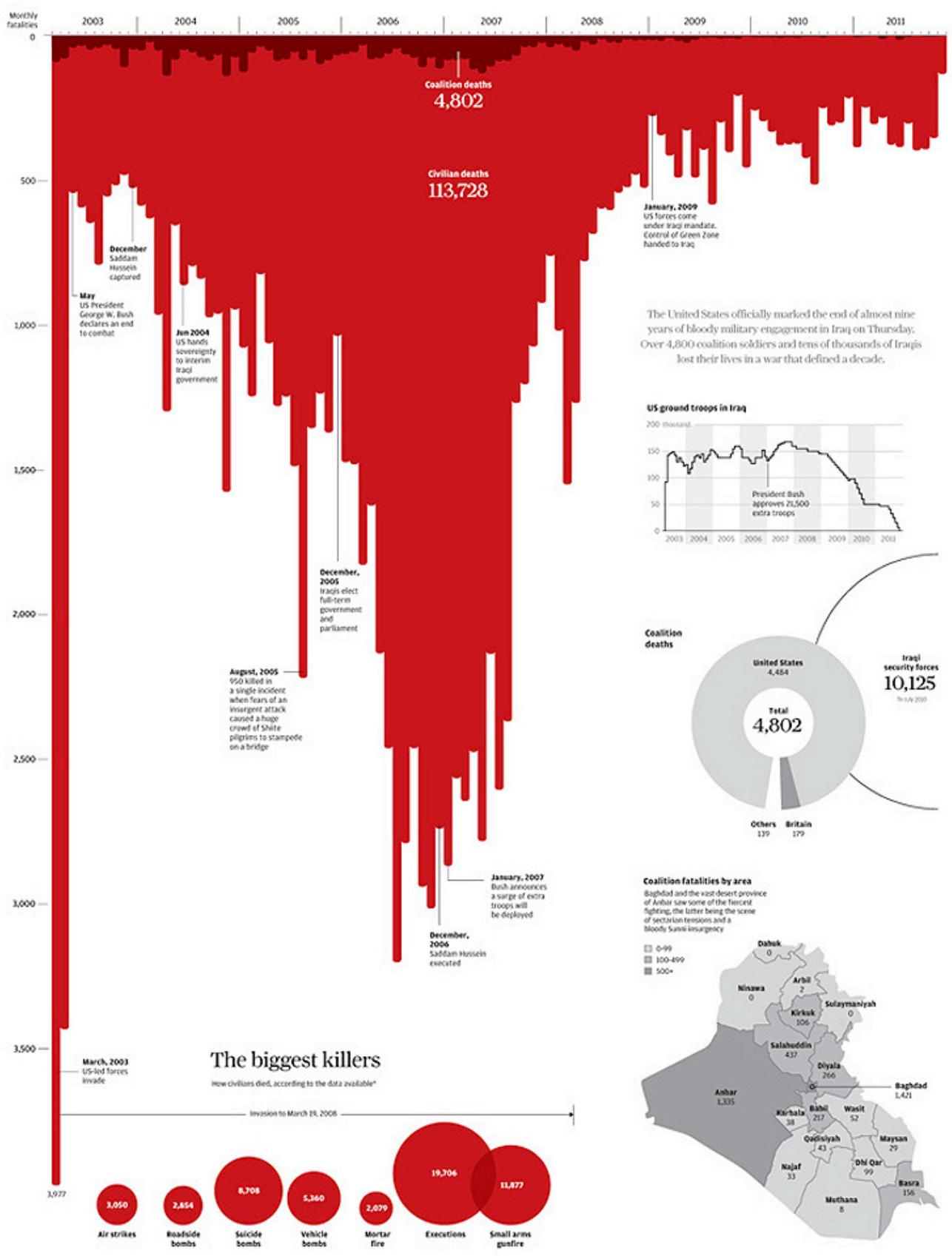
daily mail



Principle 1

Good data visualisation
is **TRUSTWORTHY**

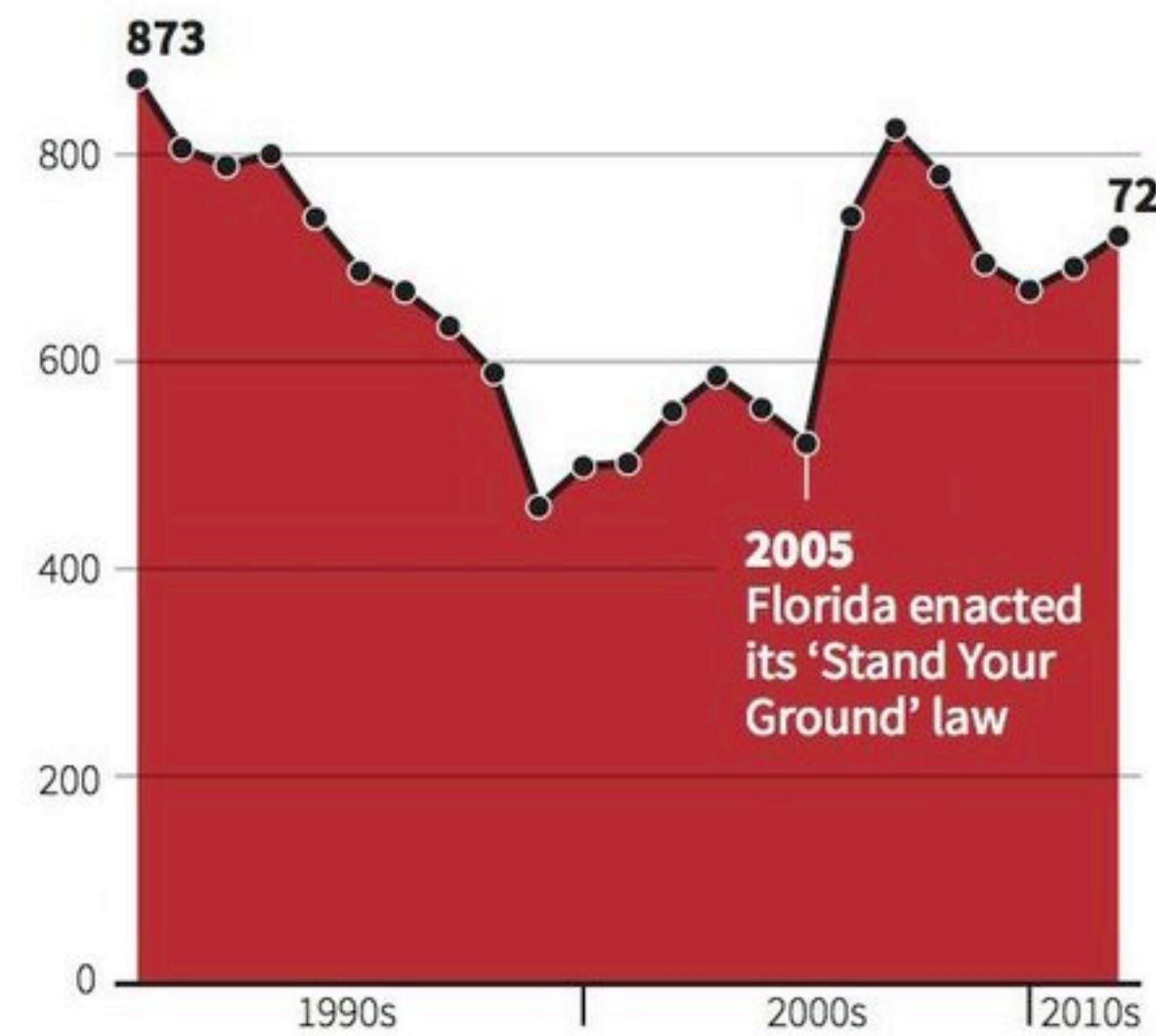
Iraq's bloody toll



principles

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data processing**

- how was data collected: from where & using what criteria?
- what calculation or modification have you applied to it?
- explain the approach in details
- have you made any significant assumption or observed any special counting rules that may not be common?
- have you **removed** or **excluded** any data?
- how representative it is?
- what biases may exist that could distort interpretation?

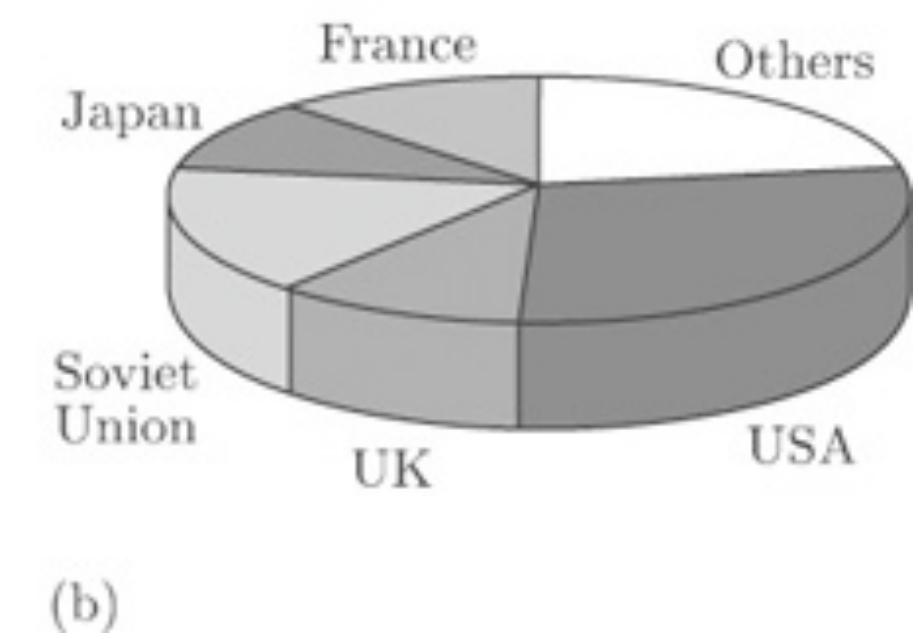
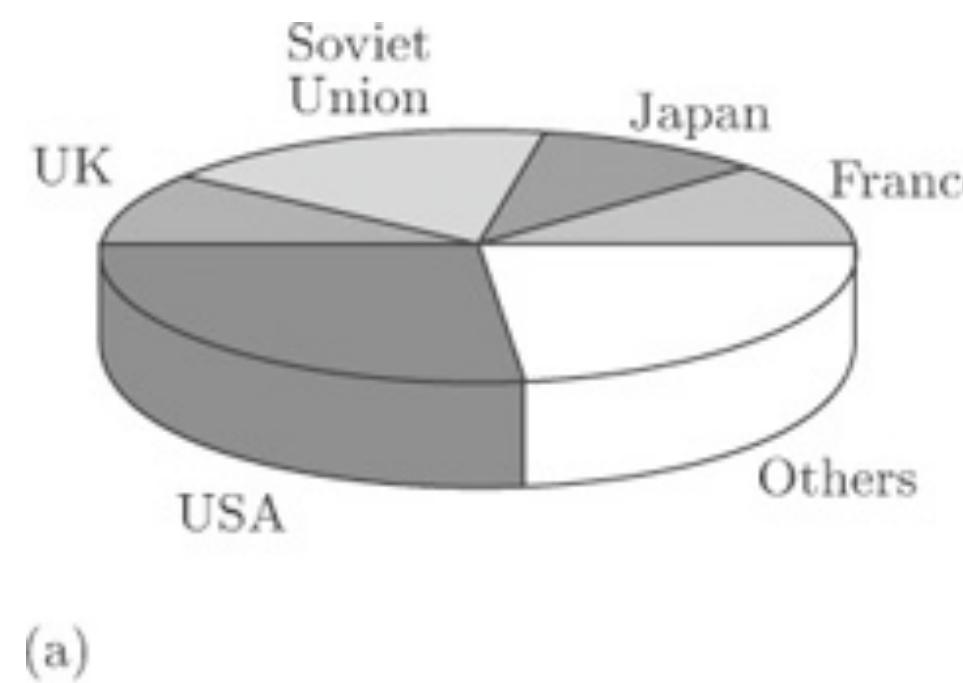
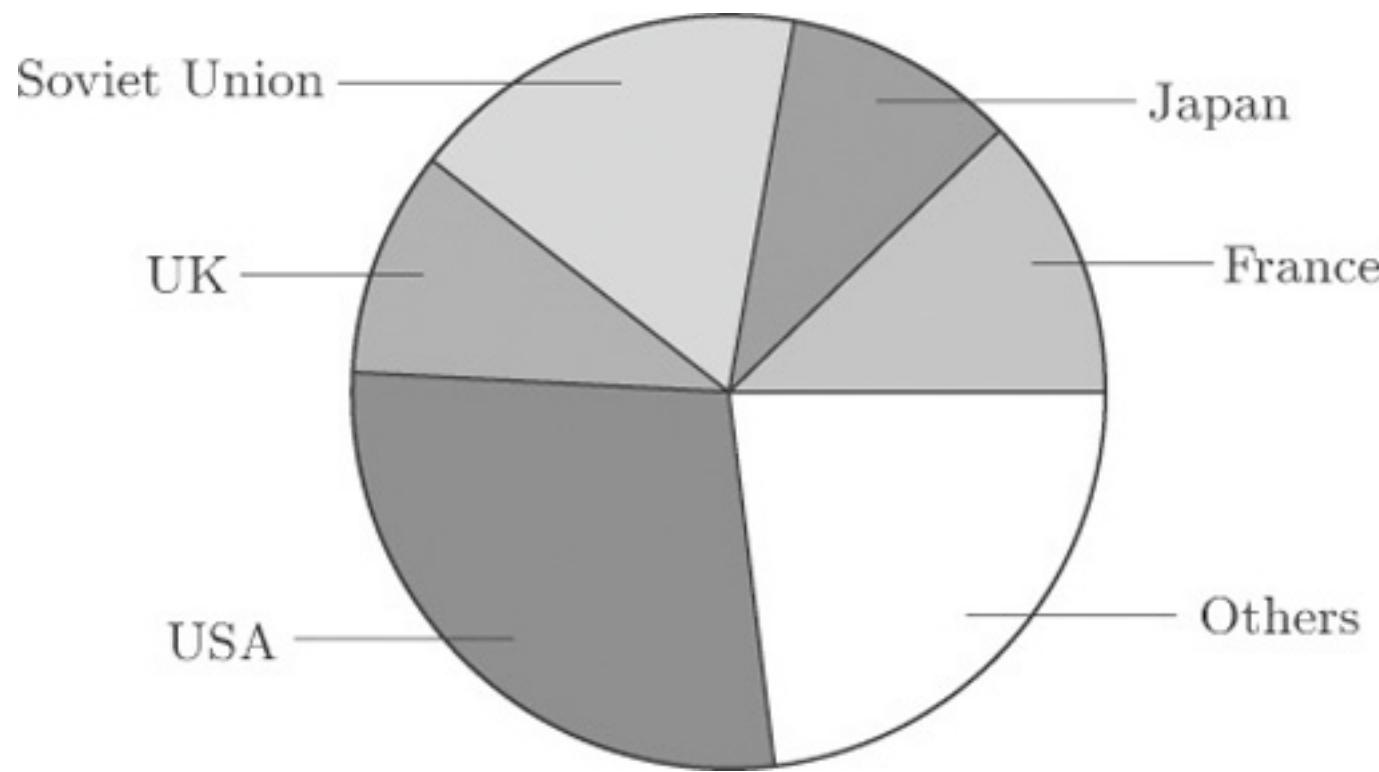
Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data representation**

- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
 - ★ quantitative values represented by areas can be disproportionately perceived

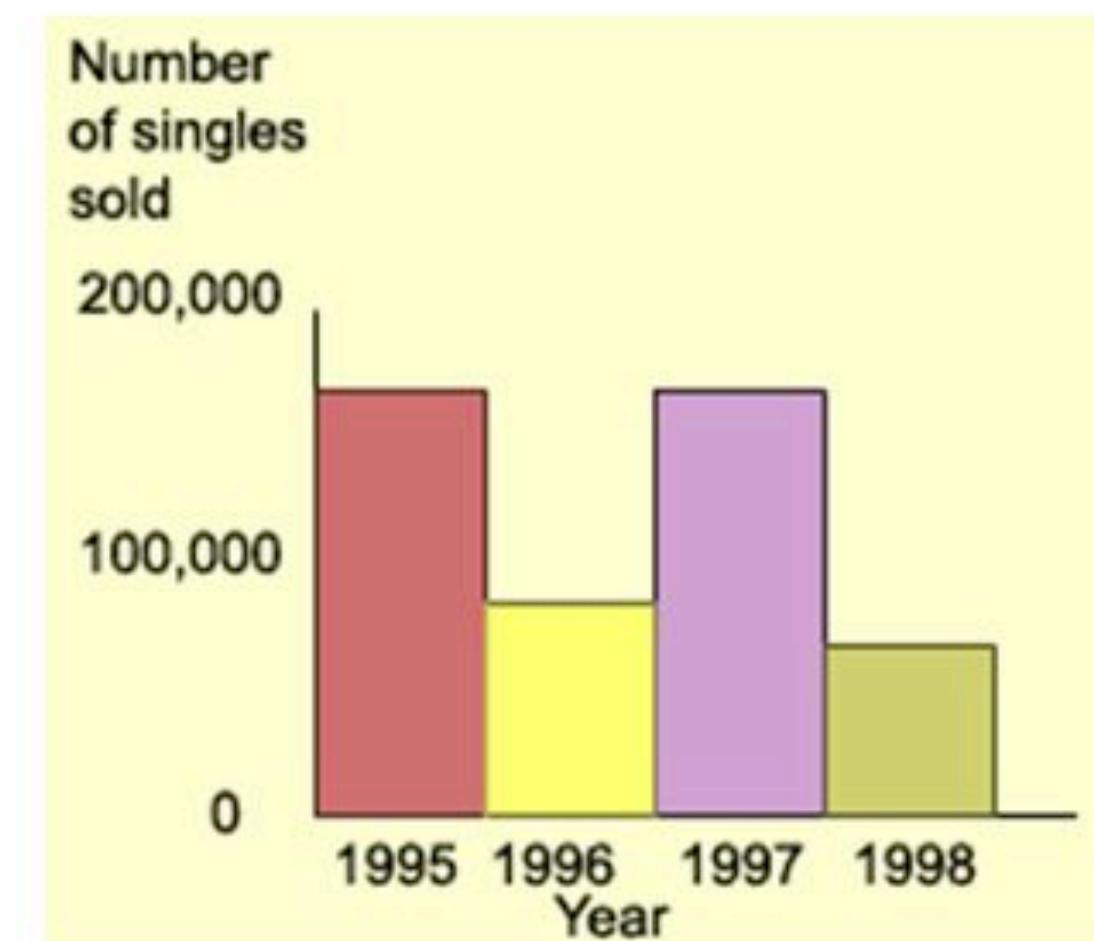
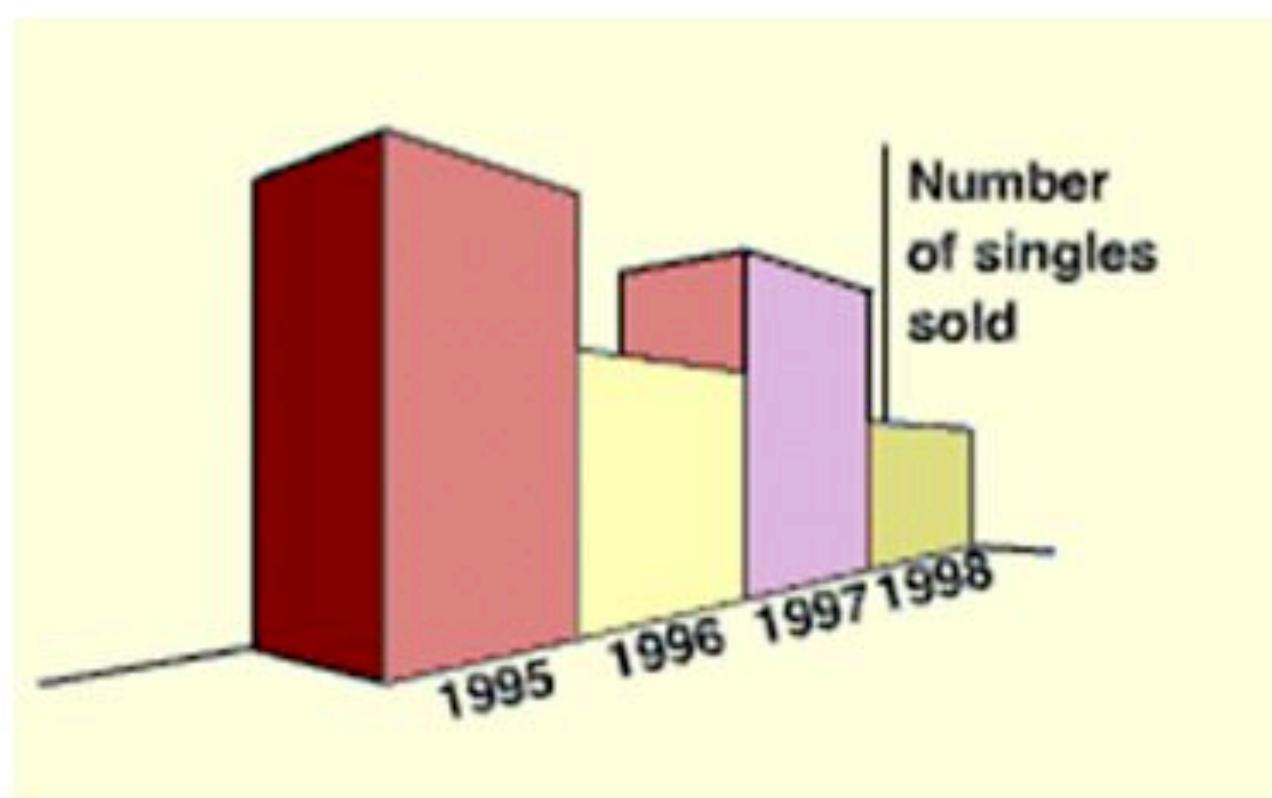


Principle 1

Good data visualisation
is **TRUSTWORTHY**

pursuing trustworthiness in **data representation**

- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
 - ★ 3d representations are often nothing more than distraction — distortion — decoration; use only if there are 3 dimensions in data and the viewer can change point of view to see different 2d perspectives



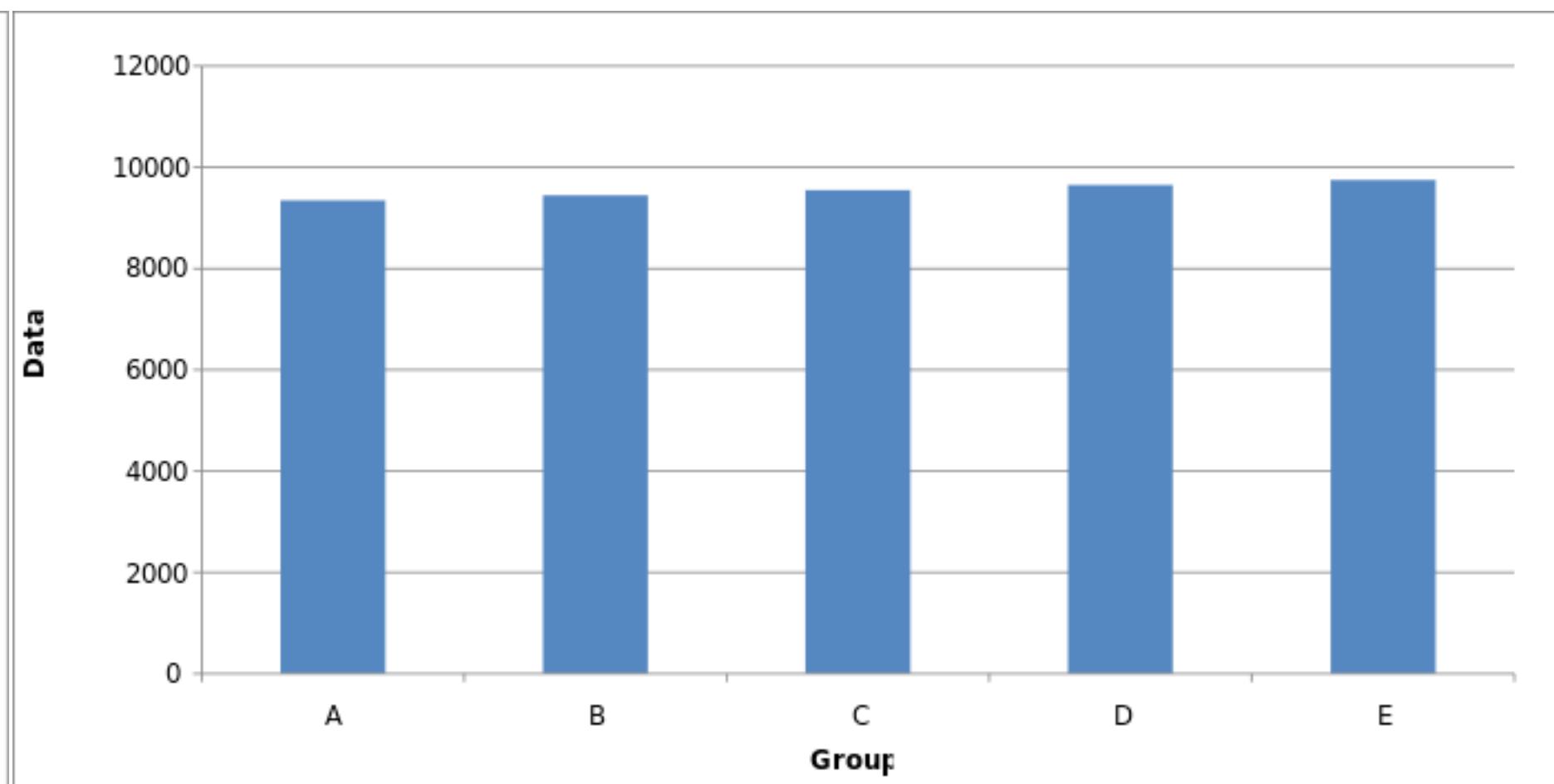
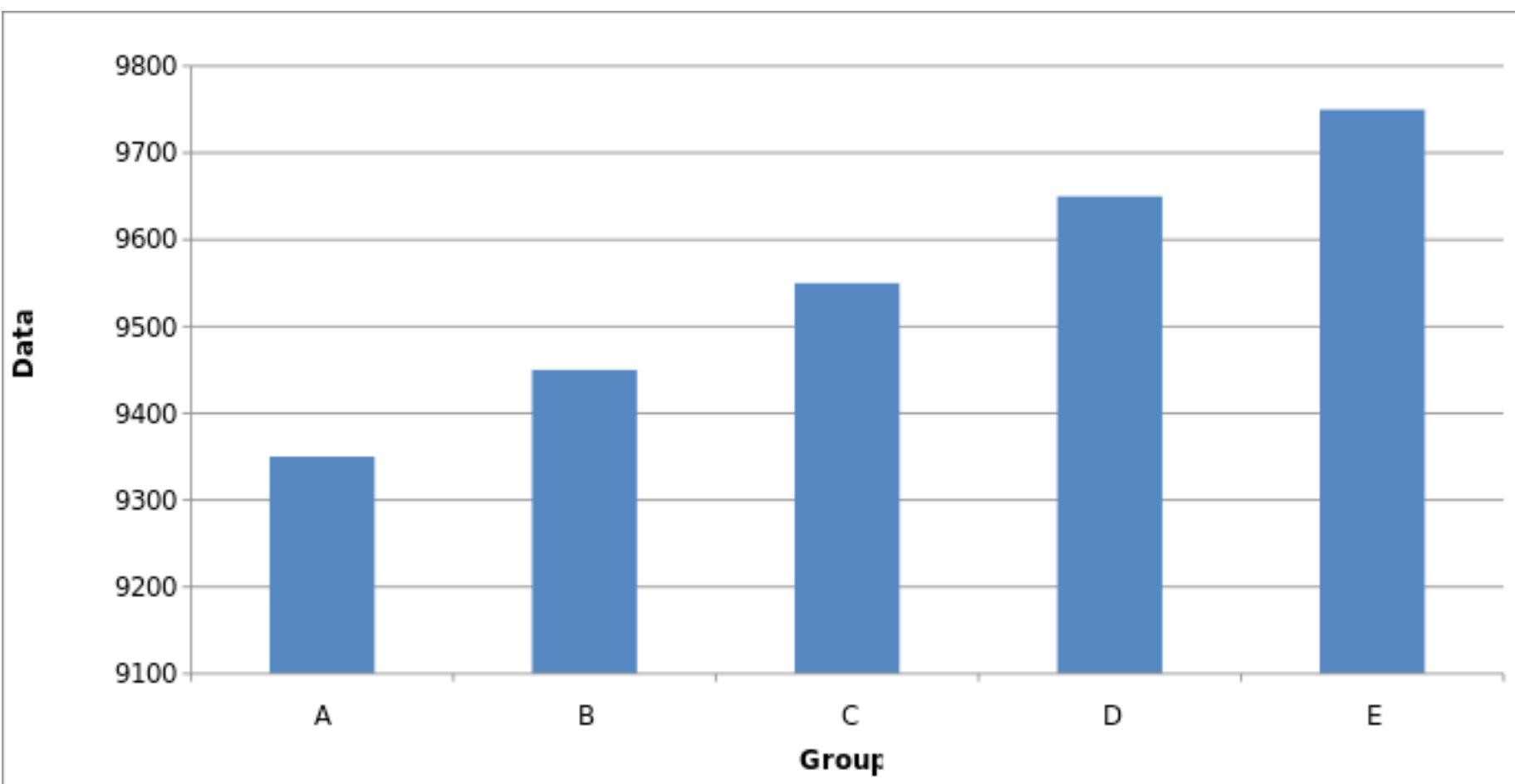
Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data representation**

- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
 - ★ axes (in bar charts) should never be truncated, origin must be 0



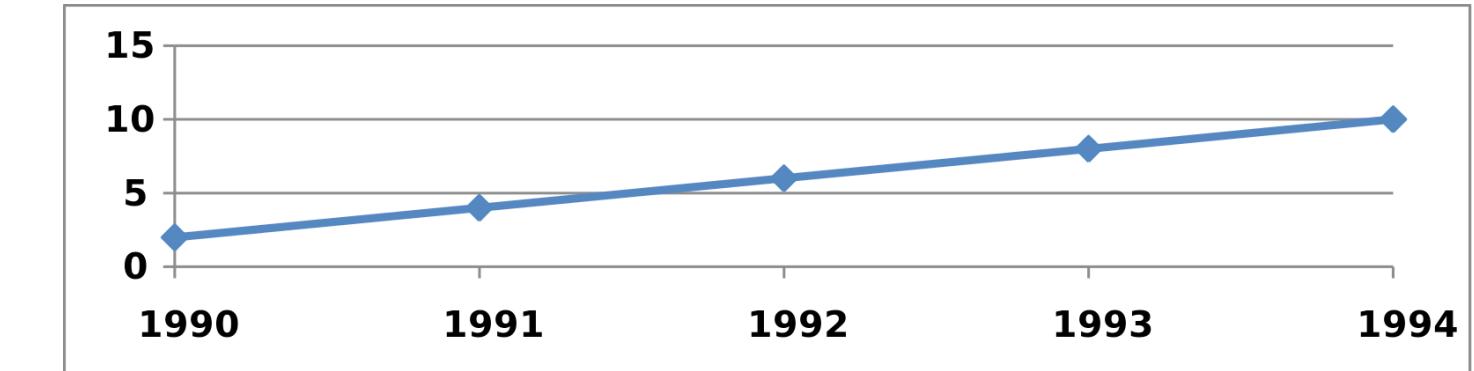
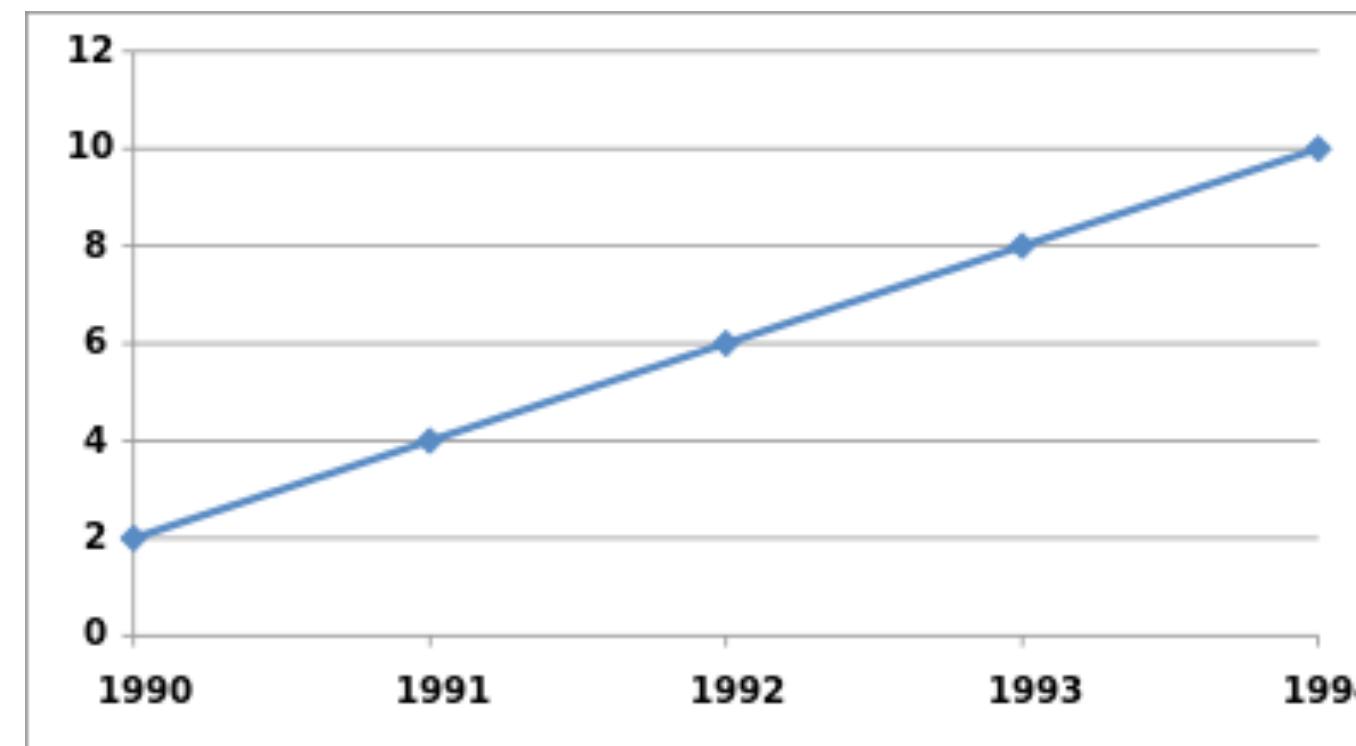
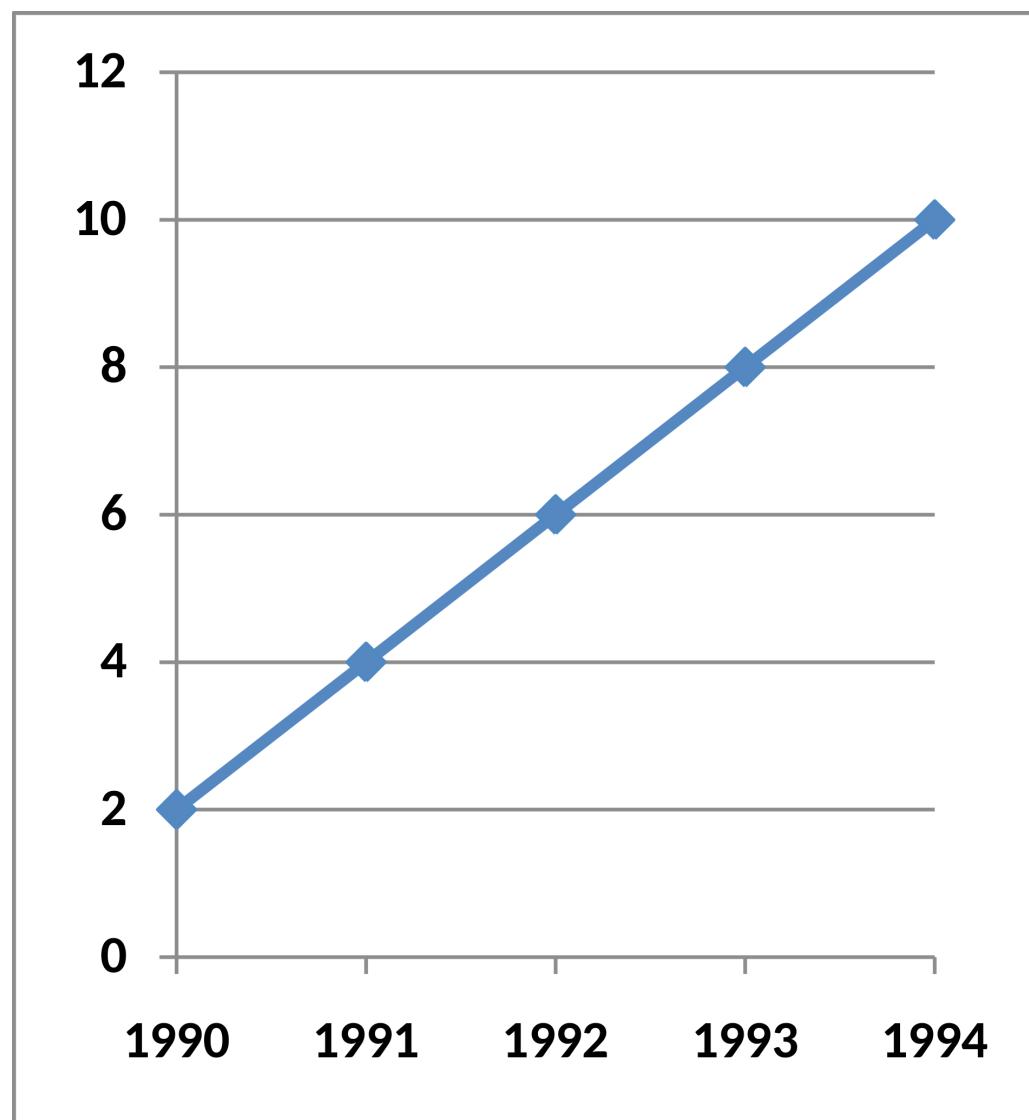
Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data representation**

- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
 - ★ aspect ratio in line chart is influential, since it modifies the perception of the steepness of connecting lines, either embellishing it or dampening it



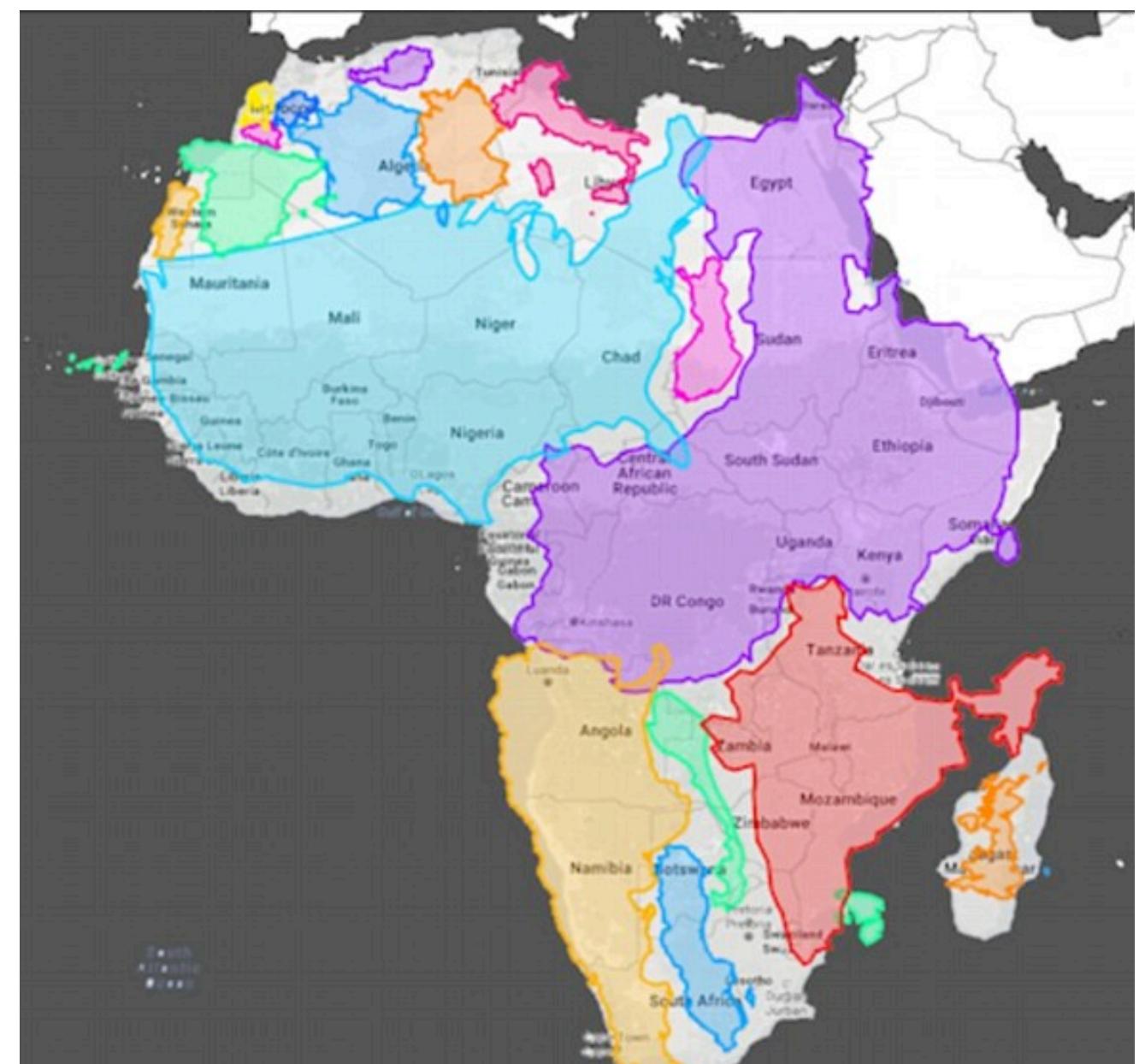
Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data representation**

- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
 - ★ different projections (mercator, etc) may alter the perception of a thematic map, distorting size/shape of regions.



Principle 1

Good data visualisation
is **TRUSTWORTHY**

principles

pursuing trustworthiness in **data presentation**

- if it looks significant, it should be
- absent annotations such as introduction/guides, axis title and labels, footnotes, data sources fail to inform the reader
- inconsistent or inappropriate color schemas
- confusing layouts
- does the chosen solution work, and, specifically, does it work in the way it promises to do?

Principle 2

Good data visualisation
is **ACCESSIBLE**

principles

the viewer should experience minimum friction between the act of understanding (**effort**) and the achieving of understanding (**reward**)

reader's side:

- subject/matter appeal
- dynamic of need
- subject/matter knowledge
- what they need to know
- issue of unfamiliar representation
- time
- format
- personal taste
- attitude & emotion

author's side

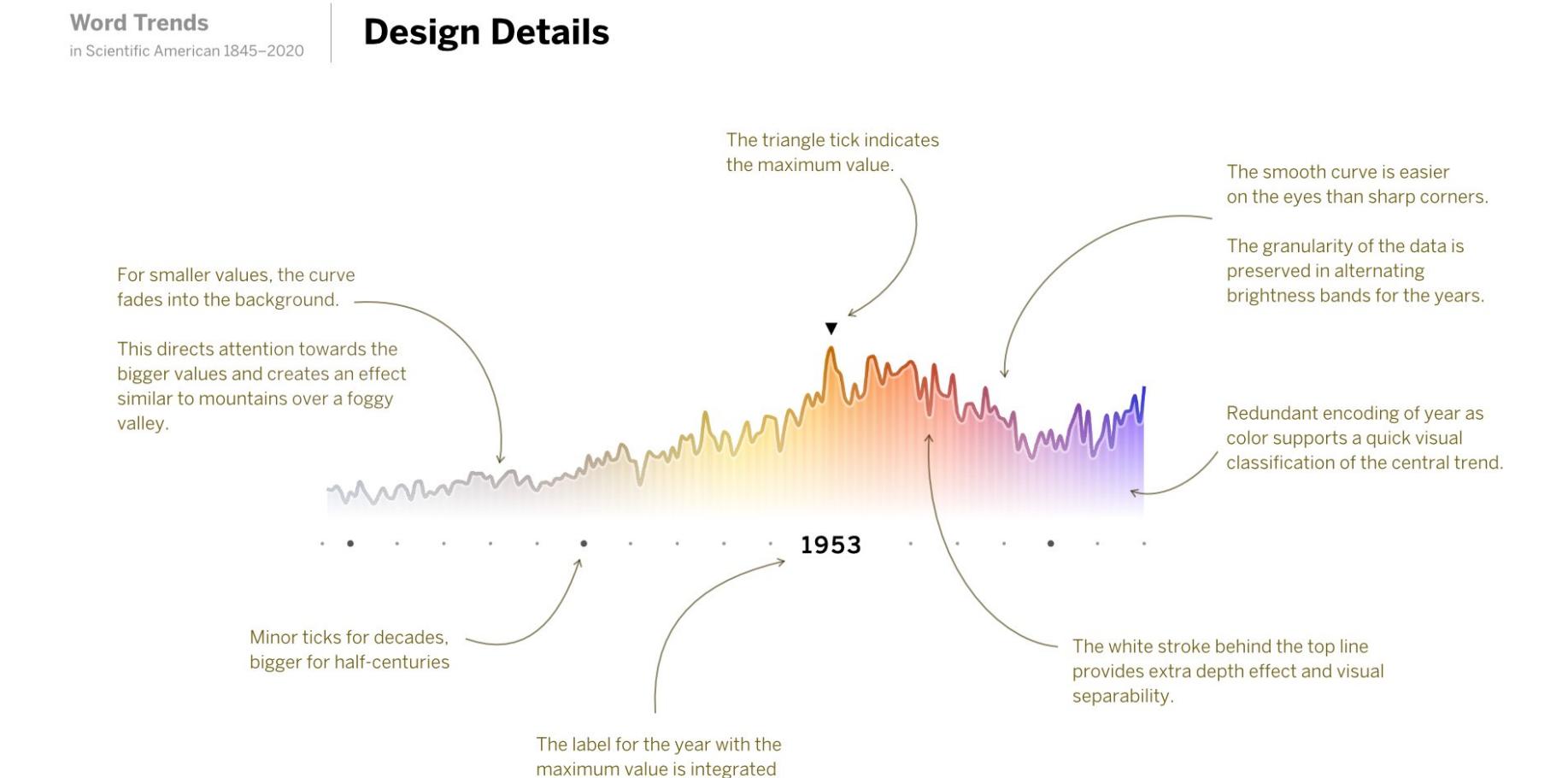
- lack of focus
- not deep enough
- oversimplified representation
- unfit for the setting
- visually inaccessible
- misjudged format, details get lost
- too many interactive options
- too complex representation
- absent annotations

Principle 3

Good data visualisation
is **ELEGANT**

achieving a visual quality to attract your audience, without letting the style overcome the substance

- eliminate the arbitrary: “remove to improve”
- thoroughness
- develop a style
- decoration should be additive, not negative
- offer elegant & appealing presentation congruent with the subject
- do not pursue minimalism at all costs



Source: <https://truth-and-beauty.net/>

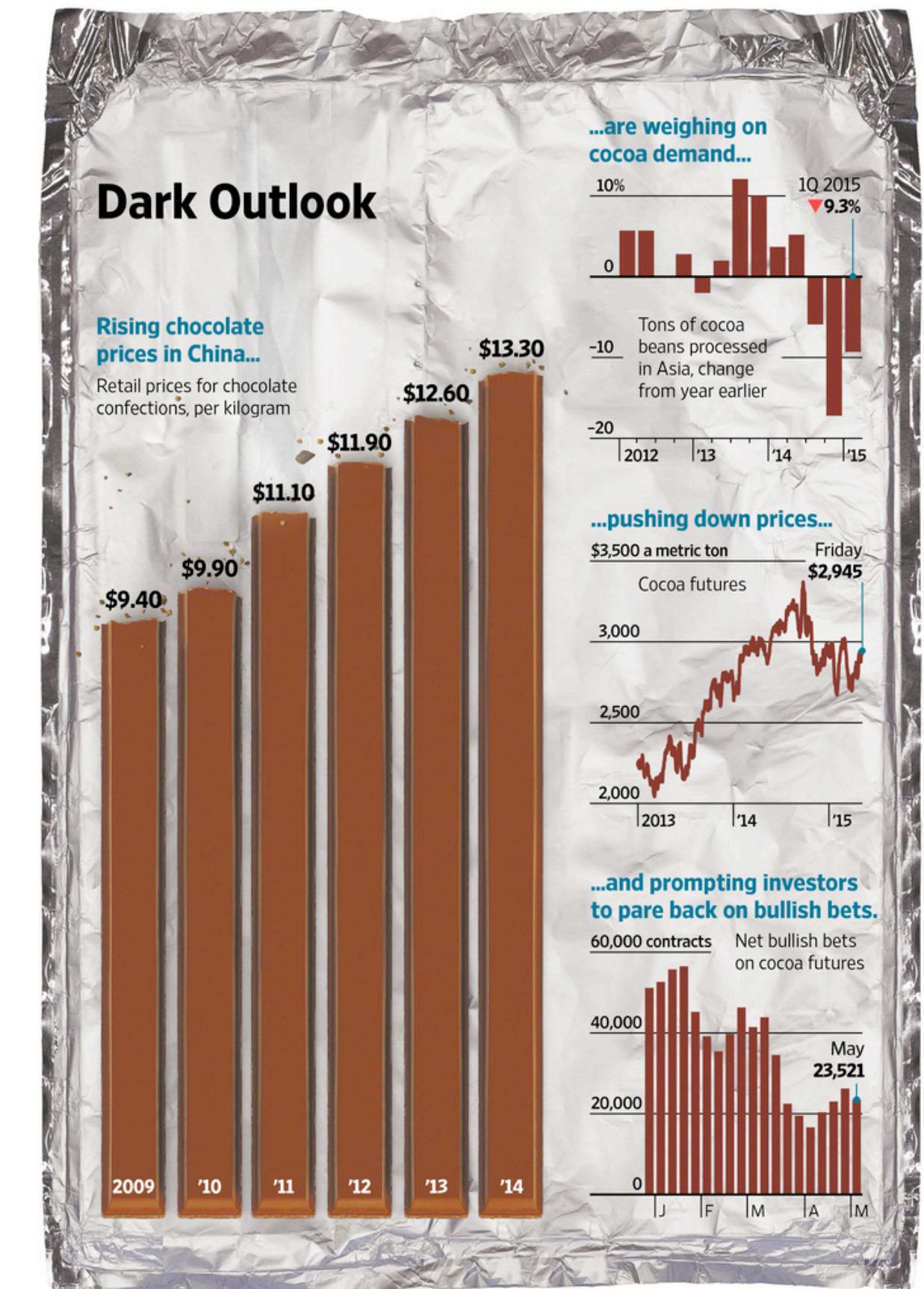
Principle 3

Good data visualisation
is **ELEGANT**

achieving a visual quality to attract your audience, without letting the style overcome the substance

- eliminate the arbitrary: “remove to improve”
- thoroughness
- develop a style
- decoration should be additive, not negative
- offer elegant & appealing presentation congruent with the subject
- do not pursue minimalism at all costs

principles



Sources: Euromonitor (retail prices); Cocoa Association of Asia (beans processed); FactSet (futures); U.S. Commodity Futures Trading Commission (bets)

THE WALL STREET JOURNAL.

Principle 3

Good data visualisation
is **ELEGANT**

achieving a visual quality to attract your audience, without letting the style overcome the substance

- eliminate the arbitrary: “remove to improve”
- thoroughness
- develop a style
- decoration should be additive, not negative
- offer elegant & appealing presentation congruent with the subject
- do not pursue minimalism at all costs



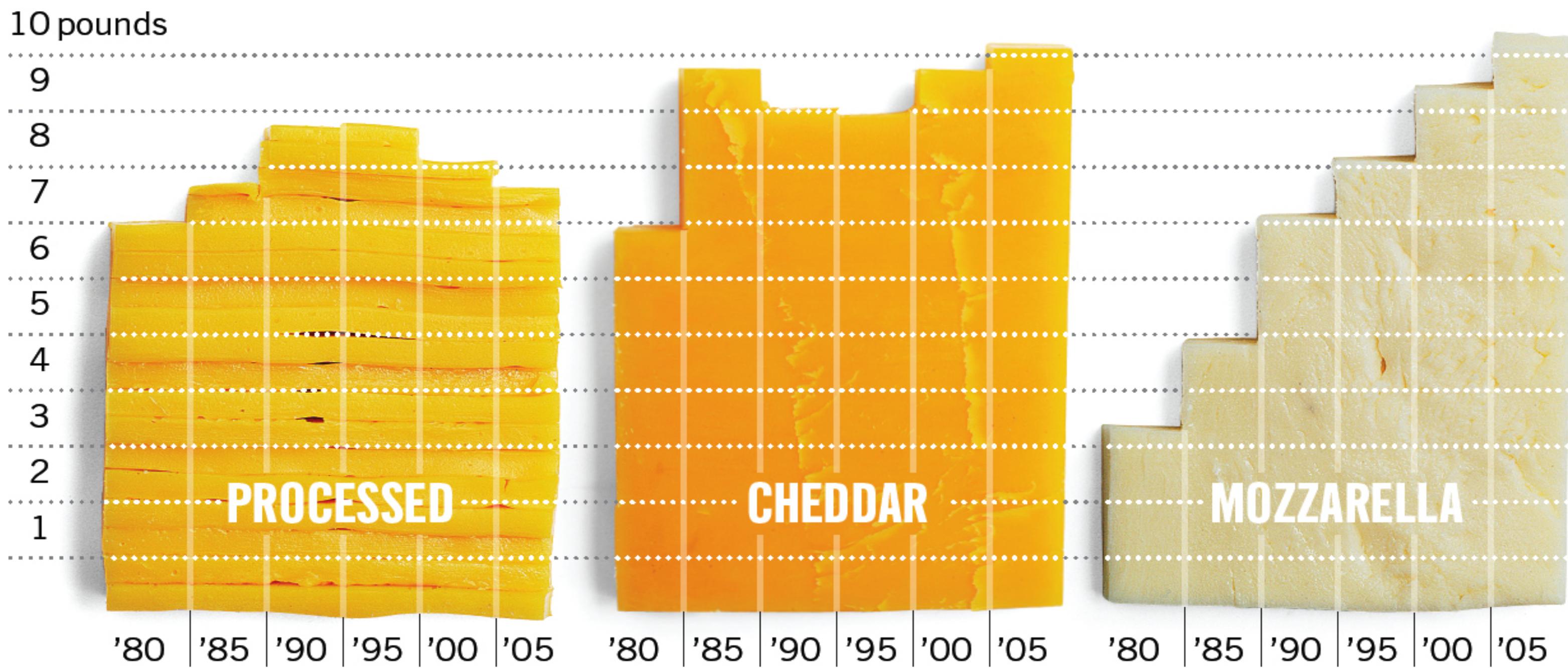
Principle 3

Good data visualisation
is **ELEGANT**

principles

- decoration should be additive, not negative

Per capita cheese consumption in the U.S.



Principle 3

Good data visualisation
is **ELEGANT**

principles

- decoration should be additive, not negative

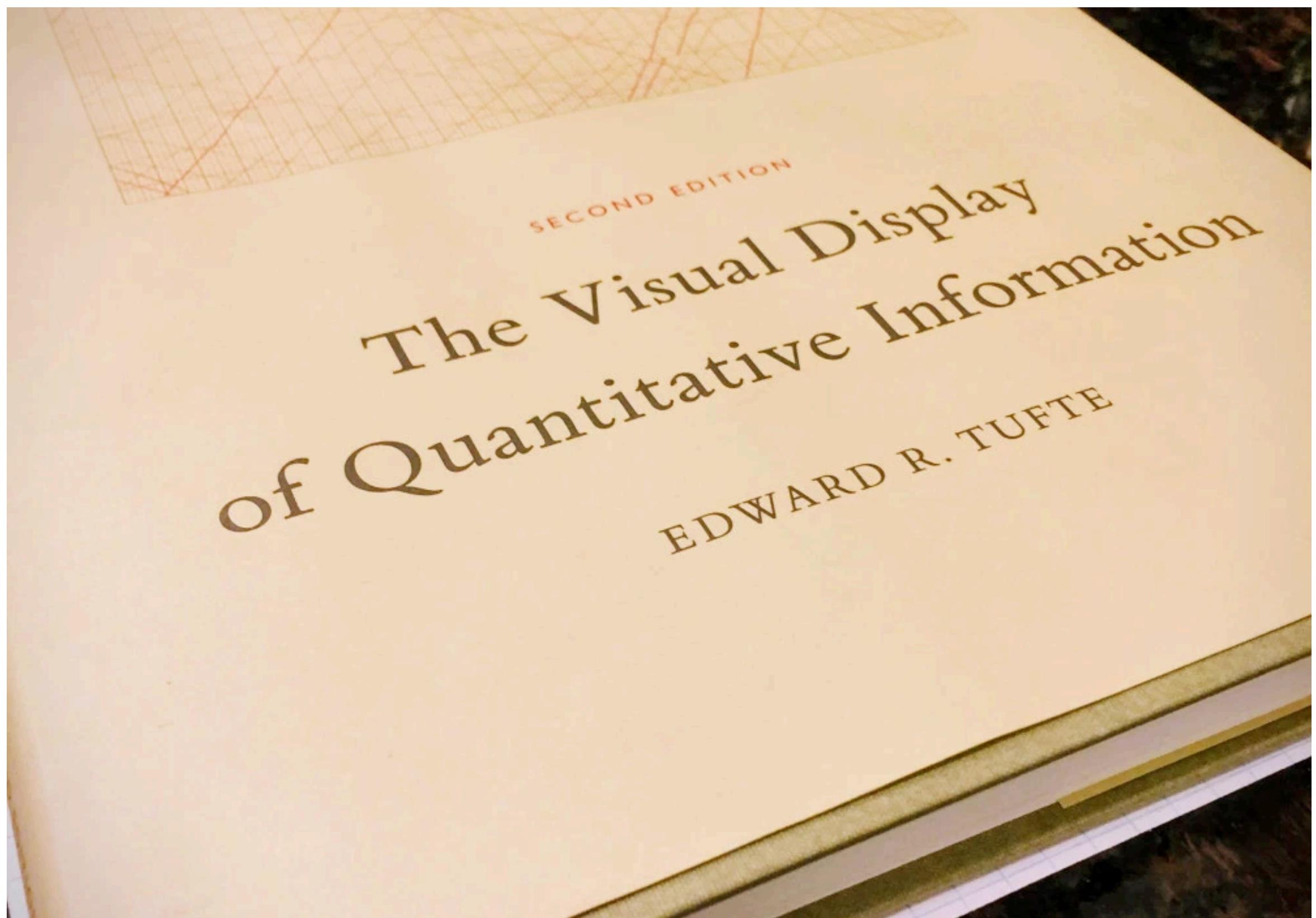


“if vienna would be an apartment”



practice guidelines

by e. r. tufte



practice guidelines

by e. r. tufte

excellence in dataviz consists of complex ideas communicated with clarity, precision and efficiency; graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, technology of production
- avoid distorting what data have to say
- present many numbers in a small space
- make large dataset coherent
- encourage the comparison of different pieces of data
- reveal the data at several levels of detail
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal description of the dataset

ELI Tufte