Upon this gifted age, in its dark hour
Rains from the sky, a meteoric shower
Of facts... they lie, unquestioned, uncombined.
Wisdom enough to leach us of our ill
Is daily spun; but there exists no loom
To weave it into a fabric.

"Huntsman, What Quarry?". 1939, Edna St. Vincent Millay

Original connection to dataviz made by Stephen Few

John Tukey

If we need a short suggestion of what exploratory data analysis is, I would suggest that

1. It is an attitude, AND
2. A flexibility, AND
3. Some graph paper (or transparencies, or both).

No catalog of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is at the heart of exploratory data analysis. The graph paper-and transparencies-are there, not as a technique, but rather as a recognition that the picture-examining eye is the best finder we have of the wholly unanticipated.

John Tukey
1915–2000

Tukey, 1980 . "We need both explanatory and confirmatory" *The American Statistician* 34(1) 23-25

Tukey also invented/discovered: many statistical tests, the word "bit", (maybe) the word "software", the Fast Fourier Transform, etc.

Stem and Leaf Plots

Chapter 6 Test Scores

Class A Stem	Class A Leaves	Class B Stem	Class B Leaves
4	9	4	
5	5, 7	5	2, 7
6	6, 6, 8	6	2, 5, 8, 8
7	2, 8, 8, 8	7	2, 5
8	4, 5, 7, 8, 8	8	1, 4, 5, 7, 7
9	1, 5, 5	9	0, 1, 1, 5, 5, 5
10	0, 0	10	0

Male		Female
5, 2, 0	1	5, 8
5, 1	2	1, 6, 9, 9
5, 5, 5, 3, 1	3	
5, 2	4	1, 2, 6, 8
9, 8, 6, 1, 1	5	5
6, 5, 5, 0	6	0, 1
2, 1, 1, 0, 0	7	2

Stem and leaf plots can be useful for quickly looking at relatively small amounts of data.

Of course, if you turn them sideways, you've got a histogram...

http://www.math6.org/graphing/6.9_quiz.htm
http://www.ck12.org/book/CK-12-Basic-Probability-and-Statistics-Concepts-A-Full-Course/r11/section/8.3/

https://commons.wikimedia.org/wiki/File:Stem-and-leaf_time_tables_in_Japanese_train_stations.jpg#mediaviewer/File:Stem-and-leaf_time_tables_in_Japanese_train_stations.jpg

Boxplots (also invented by Tukey)

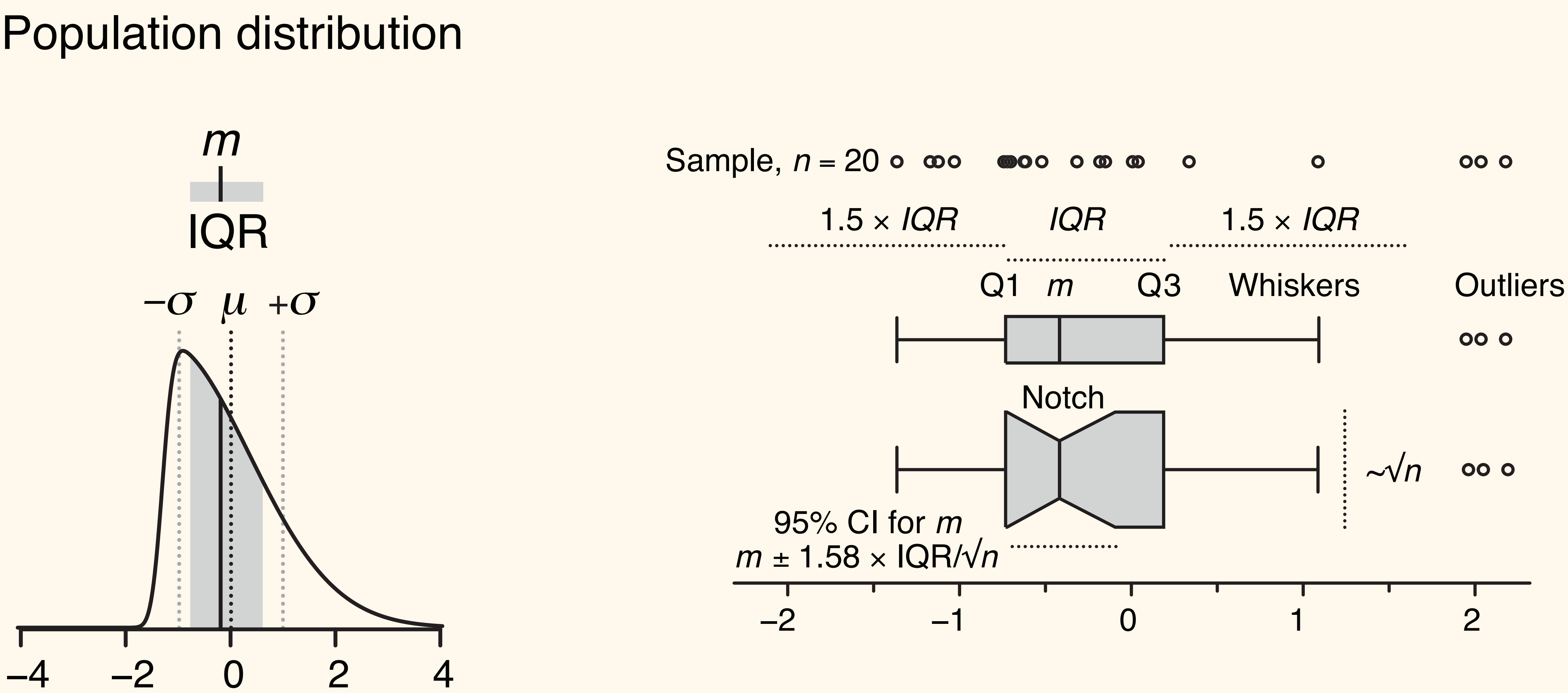
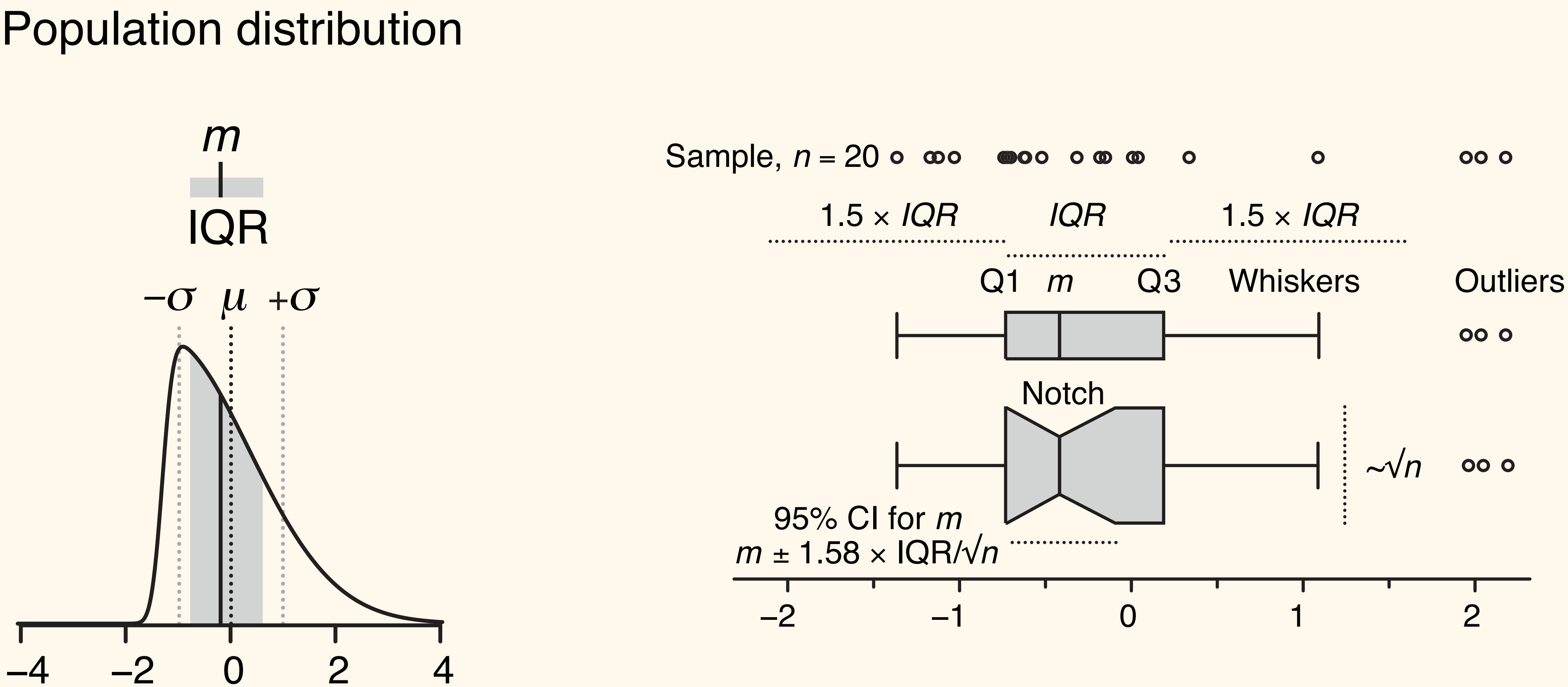


Figure 1 | The construction of a box plot. (**a**) The median ($m = -0.19$, solid vertical line) and interquartile range (IQR = 1.38, gray shading) are ideal for characterizing asymmetric or irregularly shaped distributions. A skewed normal distribution is shown with mean $\mu = 0$ (dark dotted line) and s.d. $\sigma = 1$ (light dotted lines). (**b**) Box plots for an $n = 20$ sample from **a**. The box bounds the IQR divided by the median, and Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. The box width may be scaled by $\sqrt{n}$, and a notch may be added approximating a 95% confidence interval (CI) for the median. Open circles are sample data points. Dotted lines indicate the lengths or widths of annotated features.

Martin Krzywinski, 2014: "Visualizing samples with box plots." *Nature Methods* 11(2) 119–120

Boxplots (also invented by Tukey)

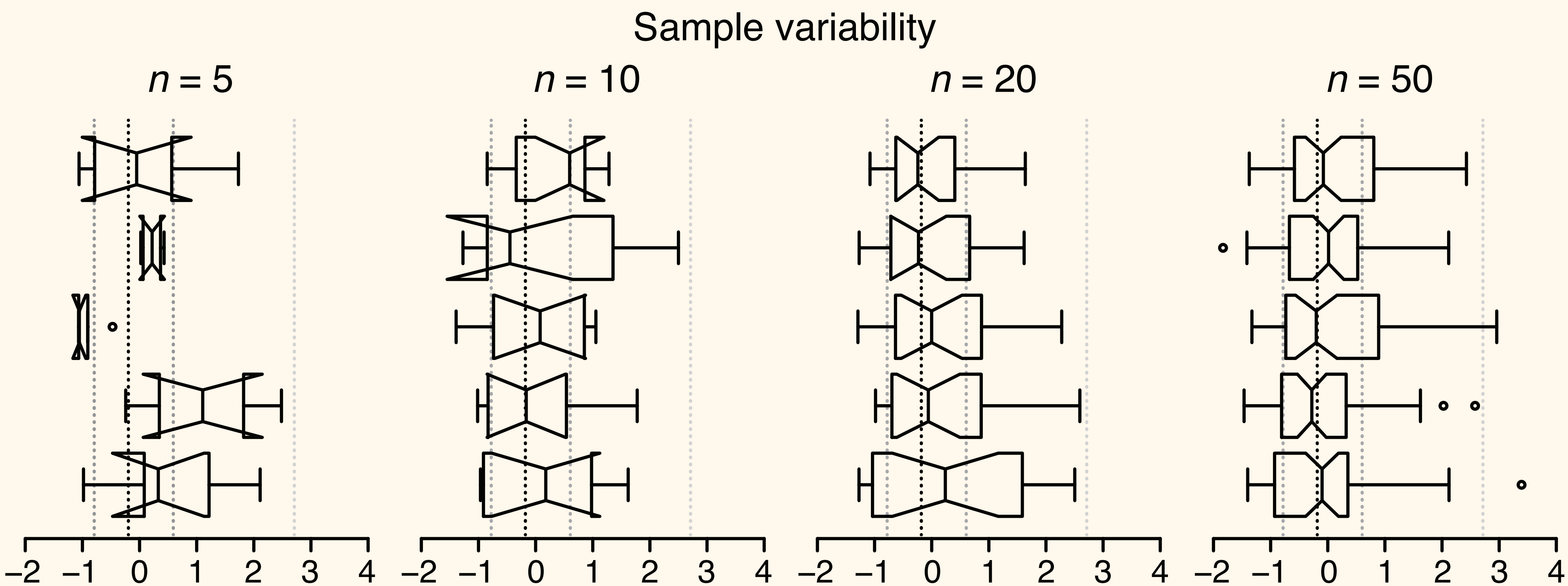


Tukey-style whiskers: the most extreme data point that is no more than 1.5 x IQR from the edge of the box…

Spear-style whiskers: the most extreme values, period (min/max).

Martin Krzywinski, 2014: "Visualizing samples with box plots." *Nature Methods* 11(2) 119–120

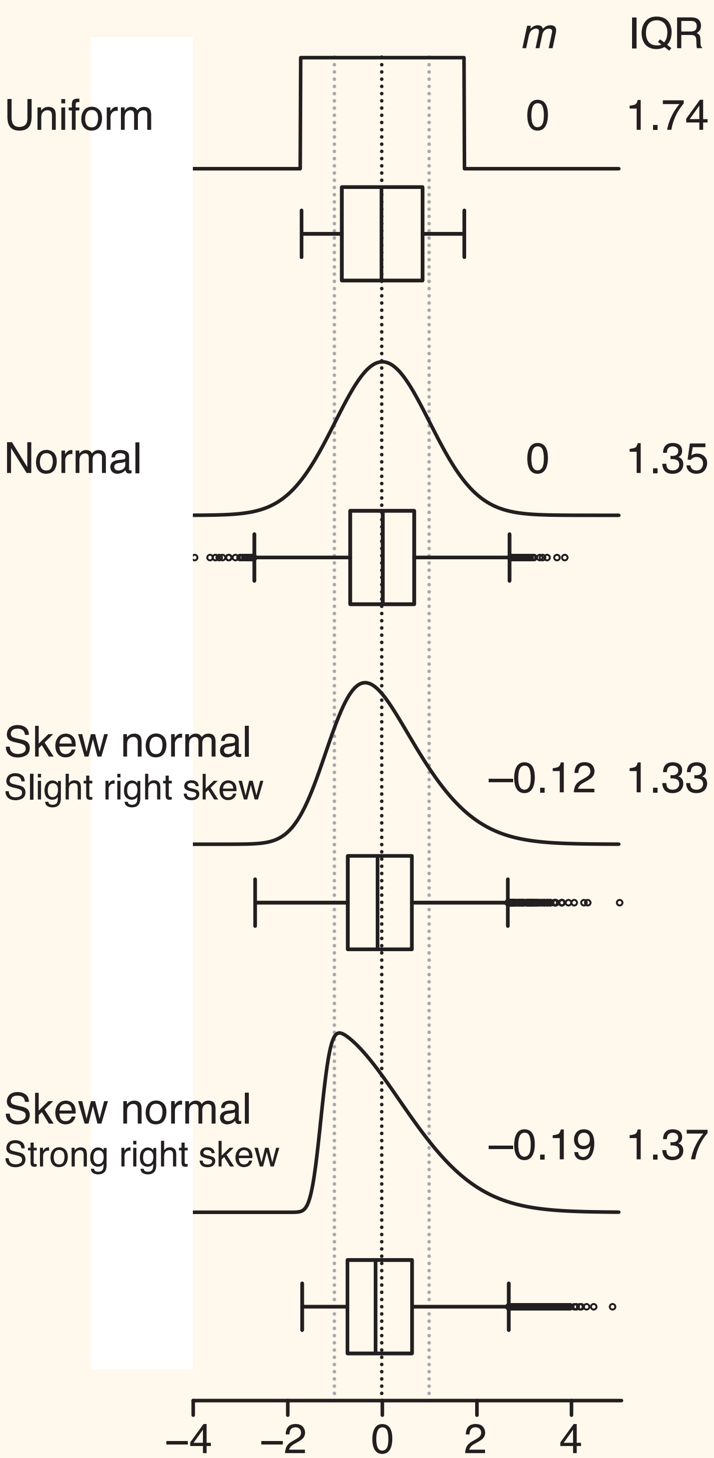
Boxplots (also invented by Tukey)

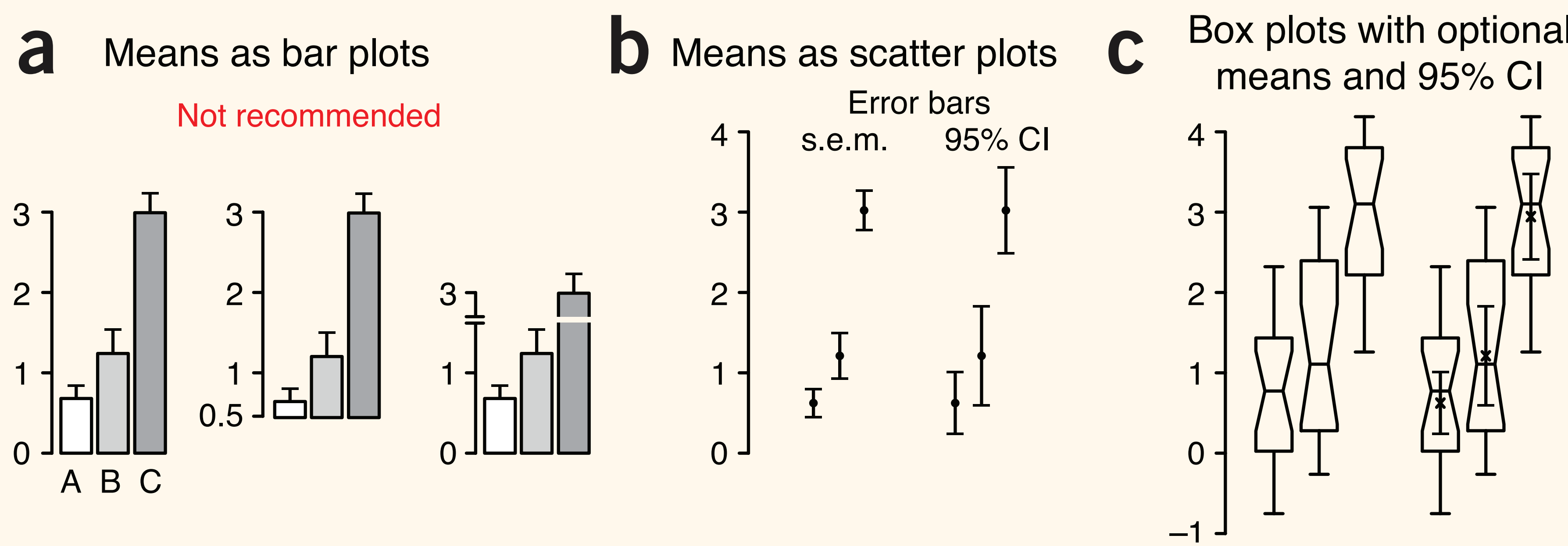


Aspects of the box plot such as width, whisker position, notch size and outlier display are subject to tuning; it is therefore important to clearly label how your box plot was constructed. Fewer than 20% of box plot figures in 2013 *Nature Methods* papers specified both sample size and whisker type in their legends—we encourage authors to be more specific.

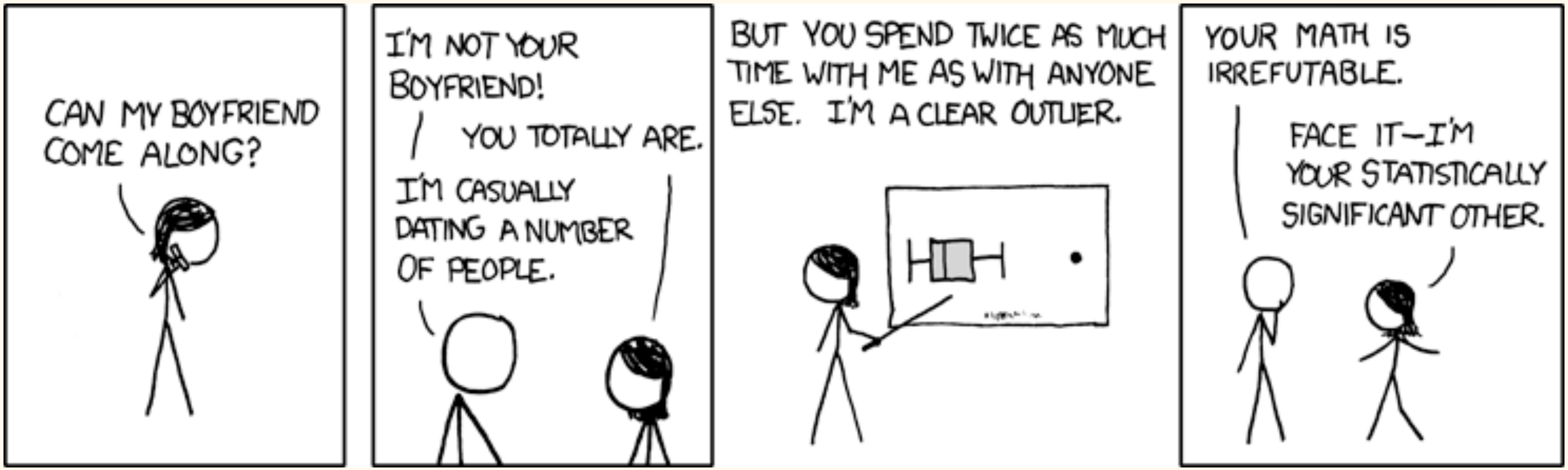
Martin Krzywinski, 2014: “Visualizing samples with box plots.” *Nature Methods* 11(2) 119–120

Boxplots (also invented by Tukey)



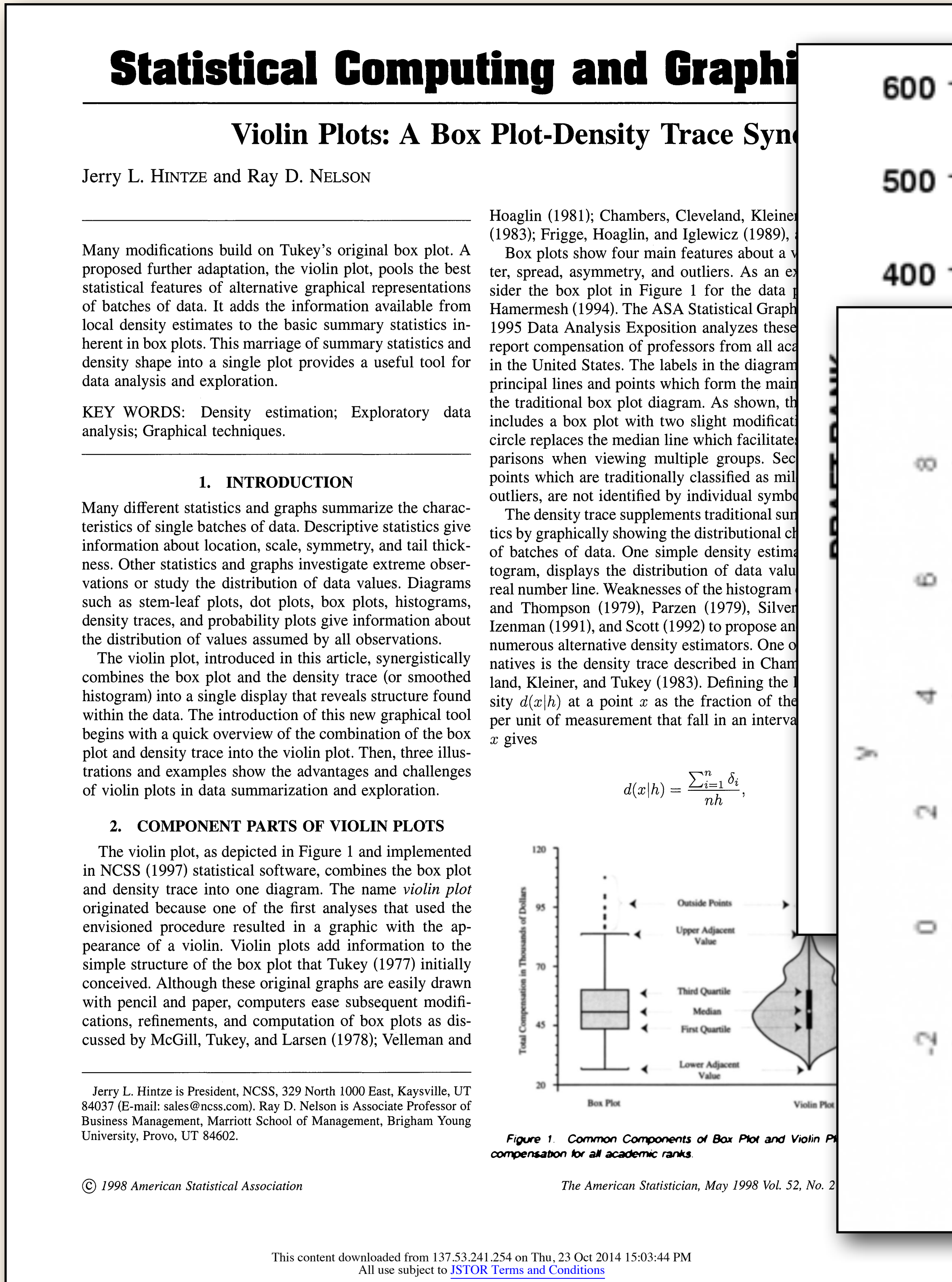


Martin Krzywinski, 2014: "Visualizing samples with box plots." *Nature Methods* 11(2) 119–120



http://xkcd.com/539/

Related (but not invented by Tukey):


Statistical Computing and Graphi

Violin Plots: A Box Plot-Density Trace Syn

Jerry L. Hintze and Ray D. Nelson

Many modifications build on Tukey's original box plot. A proposed further adaptation, the violin plot, pools the best statistical features of alternative graphical representations of batches of data. It adds the information available from local density estimates to the basic summary statistics inherent in box plots. This marriage of summary statistics and density shape into a single plot provides a useful tool for data analysis and exploration.

KEY WORDS: Density estimation; Exploratory data analysis; Graphical techniques.

1. INTRODUCTION

Many different statistics and graphs summarize the characteristics of single batches of data. Descriptive statistics give information about location, scale, symmetry, and tail thickness. Other statistics and graphs investigate extreme observations or study the distribution of data values. Diagrams such as stem-leaf plots, dot plots, box plots, histograms, density traces, and probability plots give information about the distribution of values assumed by all observations.

The violin plot, introduced in this article, synergistically combines the box plot and the density trace (or smoothed histogram) into a single display that reveals structure found within the data. The introduction of this new graphical tool begins with a quick overview of the combination of the box plot and density trace into the violin plot. Then, three illustrations and examples show the advantages and challenges of violin plots in data summarization and exploration.

2. COMPONENT PARTS OF VIOLIN PLOTS

The violin plot, as depicted in Figure 1 and implemented in NCSS (1997) statistical software, combines the box plot and density trace into one diagram. The name *violin plot* originated because one of the first analyses that used the envisioned procedure resulted in a graphic with the appearance of a violin. Violin plots add information to the simple structure of the box plot that Tukey (1977) initially conceived. Although these original graphs are easily drawn with pencil and paper, computers ease subsequent modifications, refinements, and computation of box plots as discussed by McGill, Tukey, and Larsen (1978); Velleman and

Hoaglin (1981); Chambers, Cleveland, Kleine
(1983); Frigge, Hoaglin, and Iglewicz (1989),
Box plots show four main features about a
ter, spread, asymmetry, and outliers. As an e
sider the box plot in Figure 1 for the data
Hamermesh (1994). The ASA Statistical Graph
1995 Data Analysis Exposition analyzes these
report compensation of professors from all ac
in the United States. The labels in the diagram
principal lines and points which form the main
the traditional box plot diagram. As shown, th
includes a box plot with two slight modificat
circle replaces the median line which facilitate
parisons when viewing multiple groups. Sec
points which are traditionally classified as mil
outliers, are not identified by individual symb

The density trace supplements traditional su
tics by graphically showing the distributional c
of batches of data. One simple density estima
togram, displays the distribution of data valu
real number line. Weaknesses of the histogram
and Thompson (1979), Parzen (1979), Silver
Izenman (1991), and Scott (1992) to propose an
numerous alternative density estimators. One o
natives is the density trace described in Cham
land, Kleiner, and Tukey (1983). Defining the
sity $d(x|h)$ at a point x as the fraction of the
per unit of measurement that fall in an interva
x gives

$$d(x|h) = \frac{\sum_{i=1}^{n} \delta_i}{nh},$$

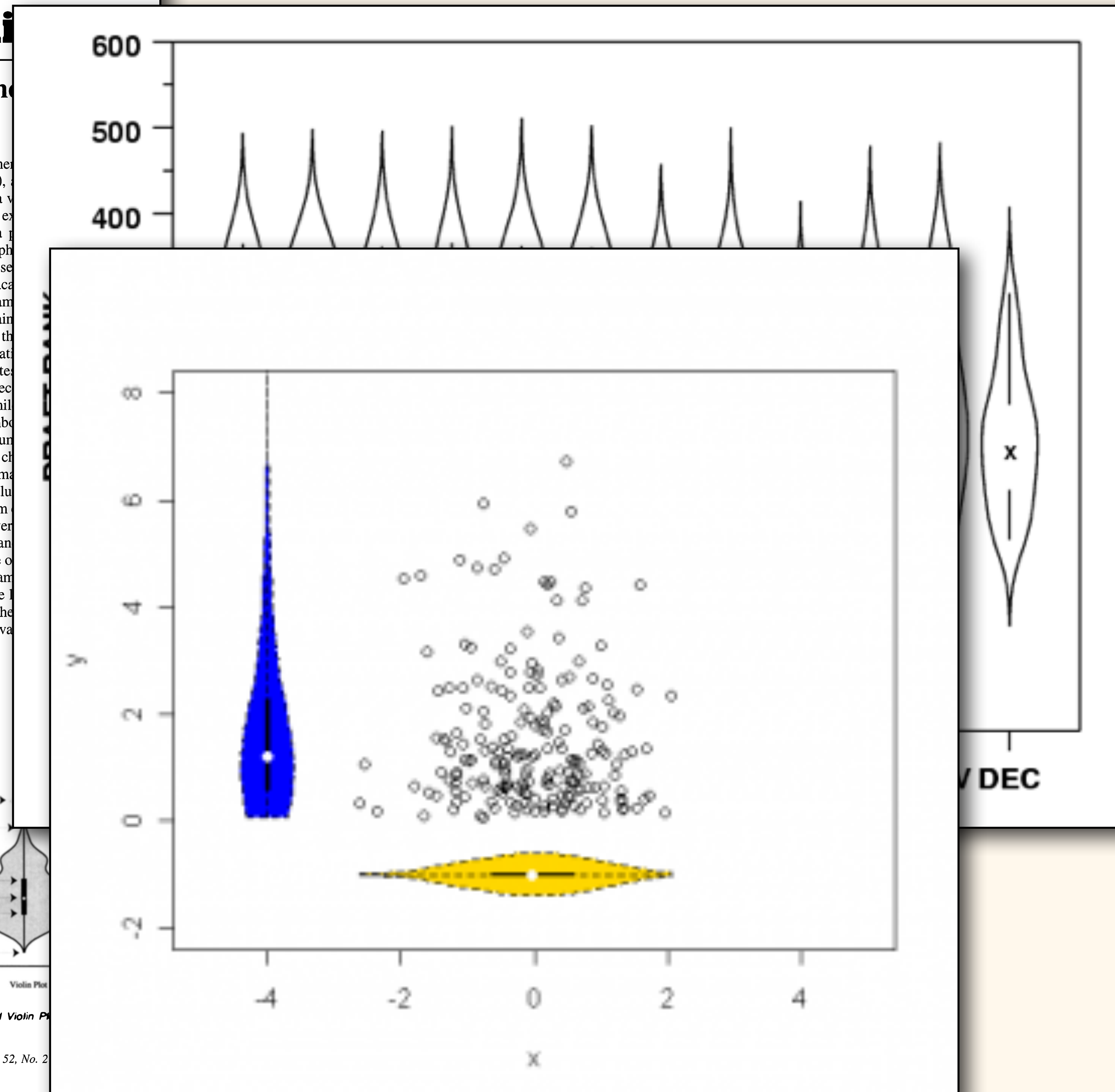
Figure 1. Common Components of Box Plot and Violin Pl
compensation for all academic ranks.

Jerry L. Hintze is President, NCSS, 329 North 1000 East, Kaysville, UT 84037 (E-mail: sales@ncss.com). Ray D. Nelson is Associate Professor of Business Management, Marriott School of Management, Brigham Young University, Provo, UT 84602.

© 1998 American Statistical Association

The American Statistician, May 1998 Vol. 52, No. 2

This content downloaded from 137.53.241.254 on Thu, 23 Oct 2014 15:03:44 PM
All use subject to JSTOR Terms and Conditions

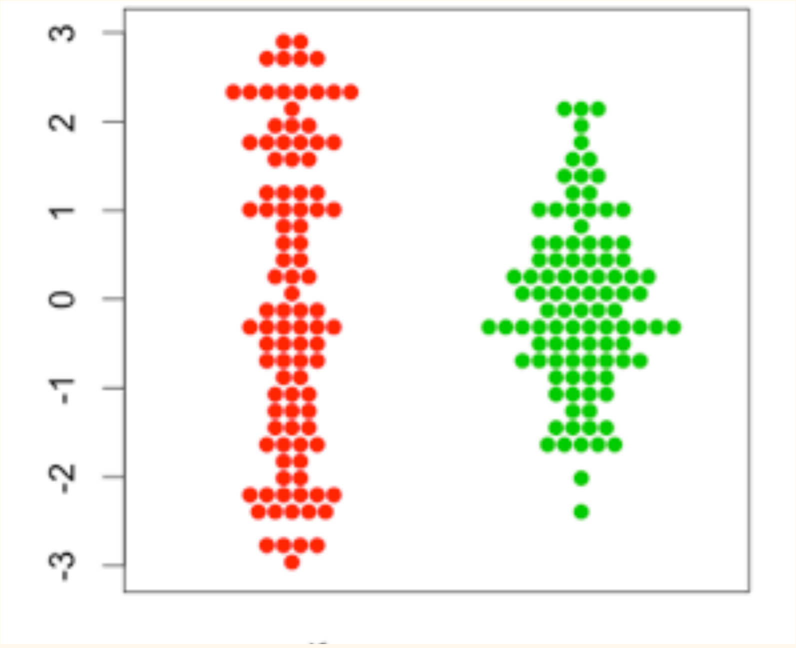


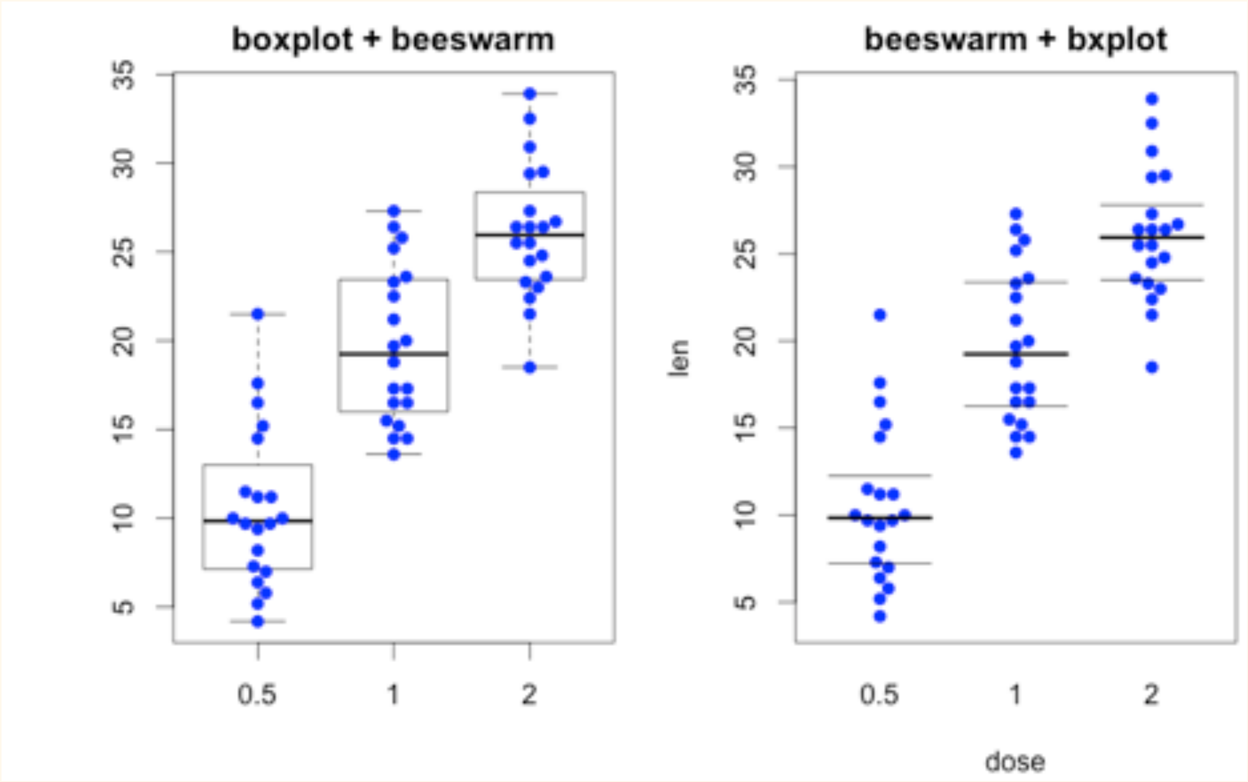
Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. The American Statistician, 52(2), 181-184.

http://commons.wikimedia.org/wiki/File:Violin_plot.gif

http://www.cerebralmastication.com/2009/02/box-plot-vs-violin-plot-in-r/

Beeswarms:


3
2
1
0
-1
-2
-3


boxplot + beeswarm
beeswarm + bxplot
35
30
25
20
15
10
5
0.5
1
2
len
dose

Other examples:

Revenue

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
West	28,384	30,288	34,302	32,039	32,938	34,392	33,923	33,092	34,934	30,384	33,923	37,834	396,433
Central	15,934	16,934	17,173	16,394	17,345	16,384	15,302	14,939	14,039	12,304	11,033	9,283	177,064
East	11,293	12,384	12,938	12,034	11,034	13,983	12,384	12,374	12,384	13,374	14,394	19,283	157,859
Total	55,611	59,606	64,413	60,467	61,317	64,759	61,609	60,405	61,357	56,062	59,350	66,400	731,356

Marketing Expenses

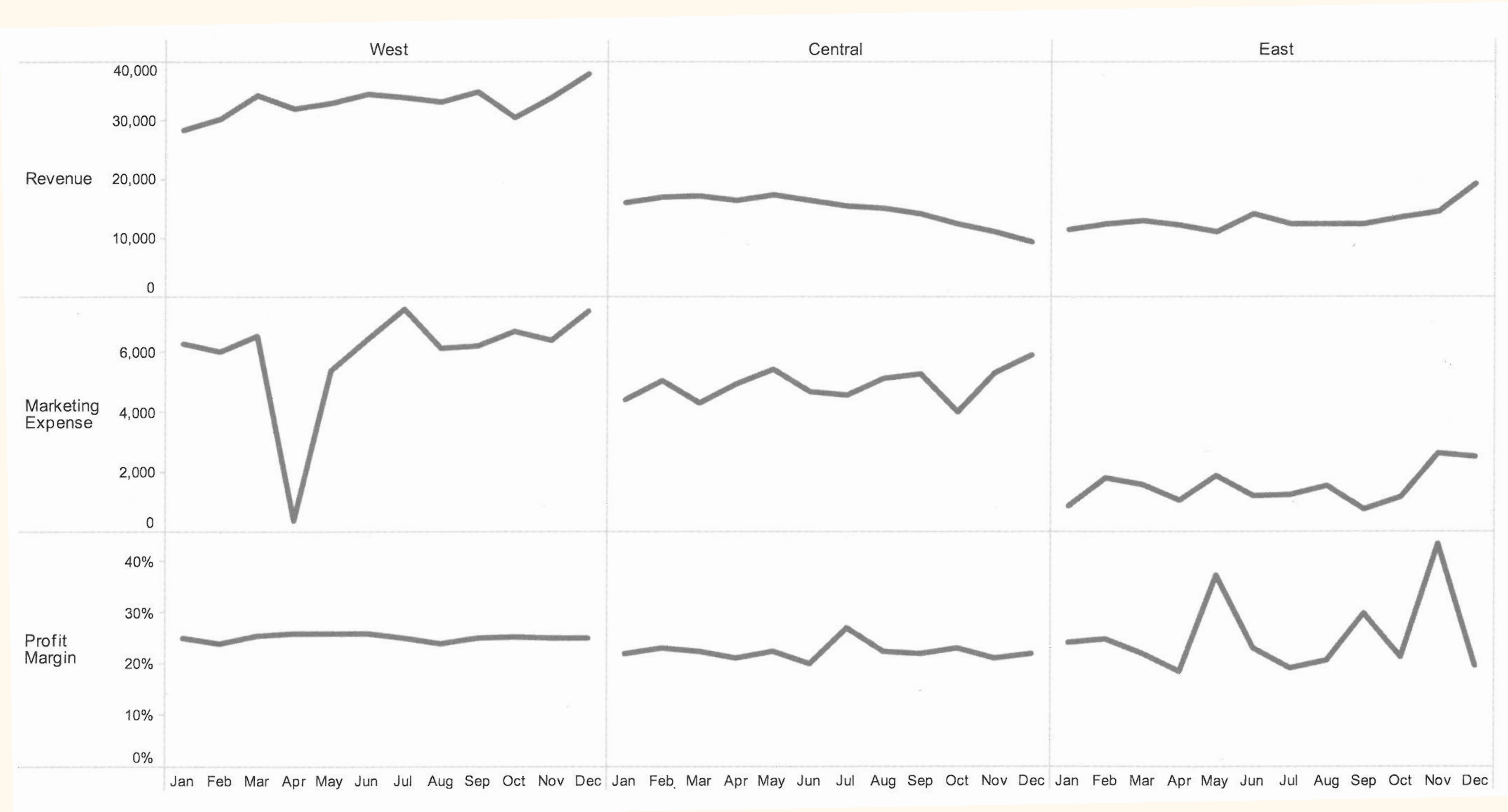
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
West	6,288	6,019	6,555	364	5,407	6,450	7,442	6,150	6,201	6,697	6,408	7,376	71,356
Central	4,429	5,039	4,309	4,951	5,442	4,675	4,558	5,124	5,278	4,016	5,325	5,898	59,044
East	851	1,784	1,542	1,024	1,864	1,173	1,237	1,504	714	1,152	2,620	2,501	17,966
Total	11,568	12,842	12,406	6,339	12,713	12,298	13,237	12,778	12,192	11,865	14,353	15,775	148,367

Profit Margin

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average
West	25.11%	24.07%	25.52%	25.80%	25.93%	26.06%	25.02%	24.41%	25.13%	25.31%	25.12%	25.01%	25.13%
Central	22.13%	23.22%	22.55%	21.08%	22.54%	20.04%	27.08%	22.52%	22.31%	23.32%	21.05%	22.01%	22.38%
East	24.06%	24.80%	21.97%	18.50%	37.16%	23.02%	19.06%	20.60%	29.74%	21.41%	43.29%	19.49%	25.26%
Average	23.69%	23.93%	23.32%	21.77%	28.52%	23.01%	23.69%	22.37%	25.58%	23.24%	29.80%	22.16%	24.26%

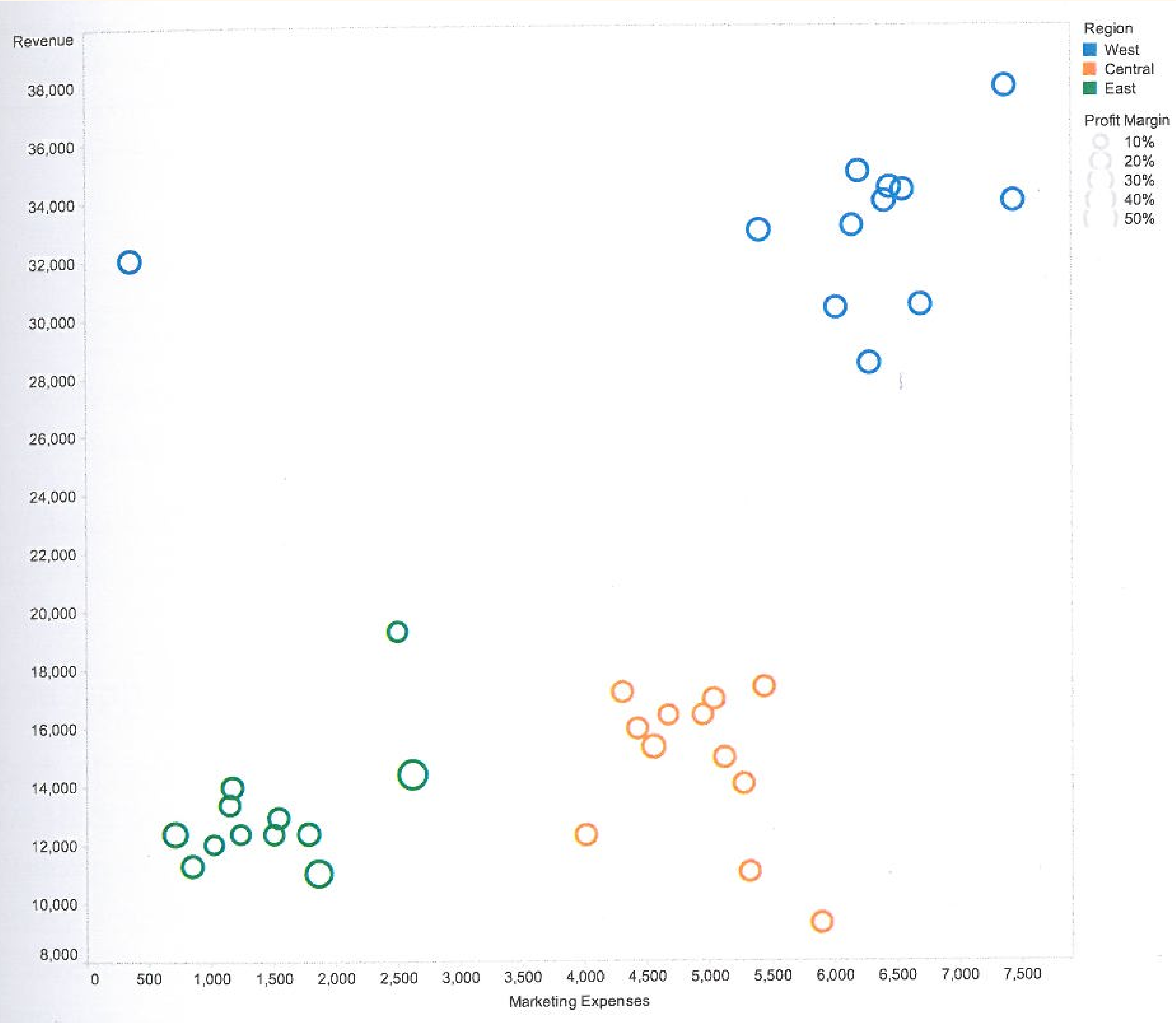
Stephen Few, “Now you see it”

Other examples:



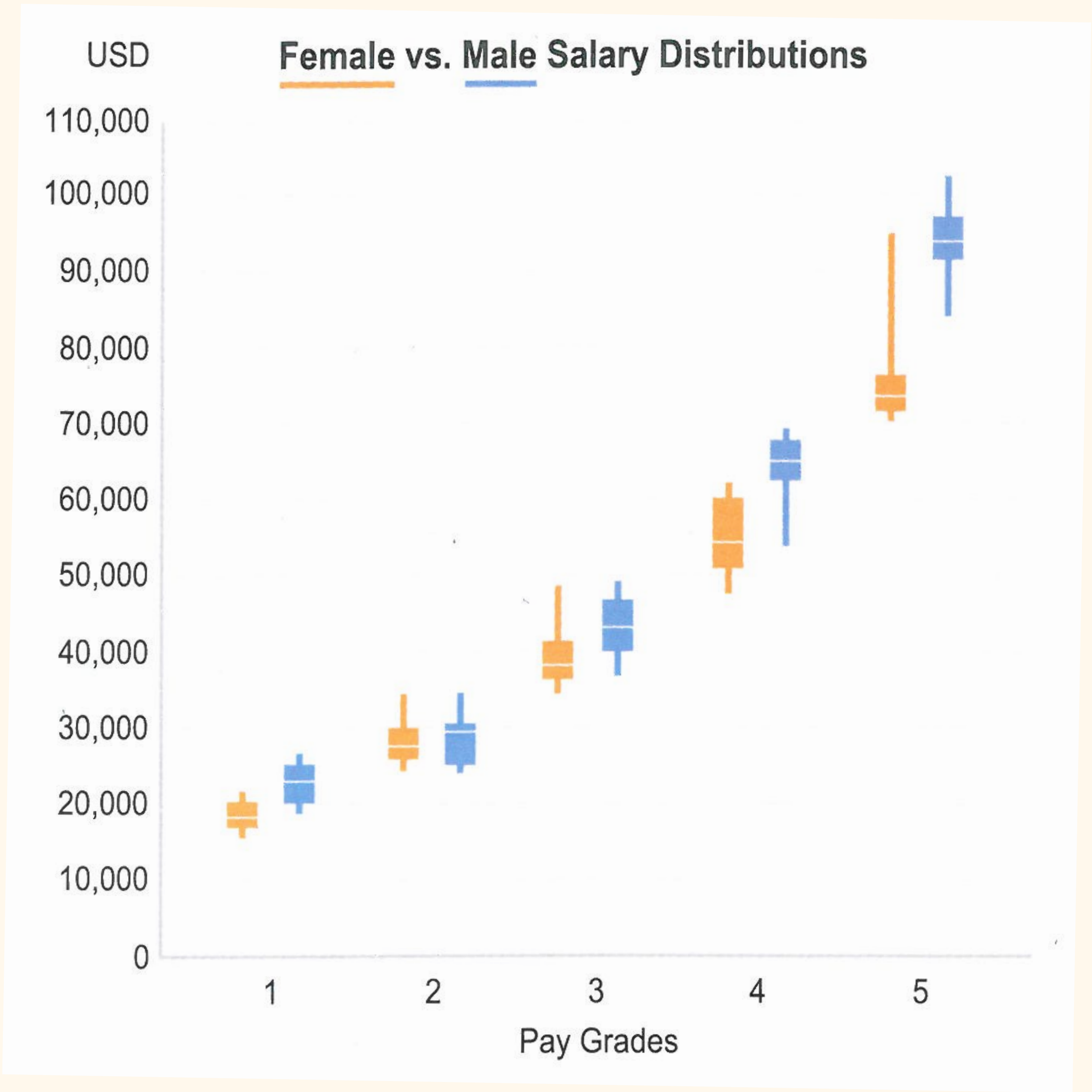
Stephen Few, "Now you see it"

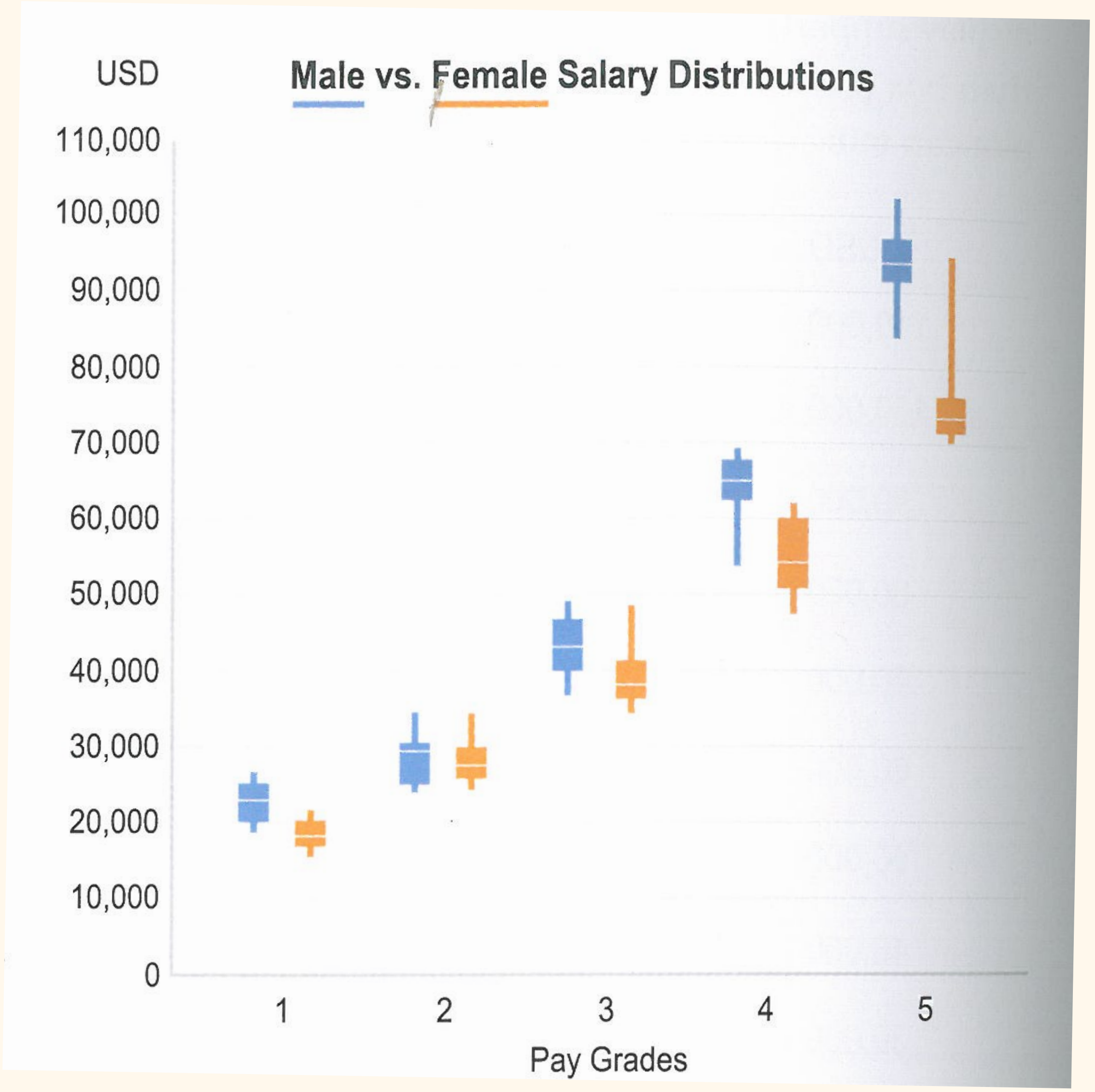
Other examples:



Stephen Few, "Now you see it"

Other examples:





The Gestalt principle of continuation is making the graph on the left look like a much smoother curve than it really is...

Stephen Few, "Now you see it"

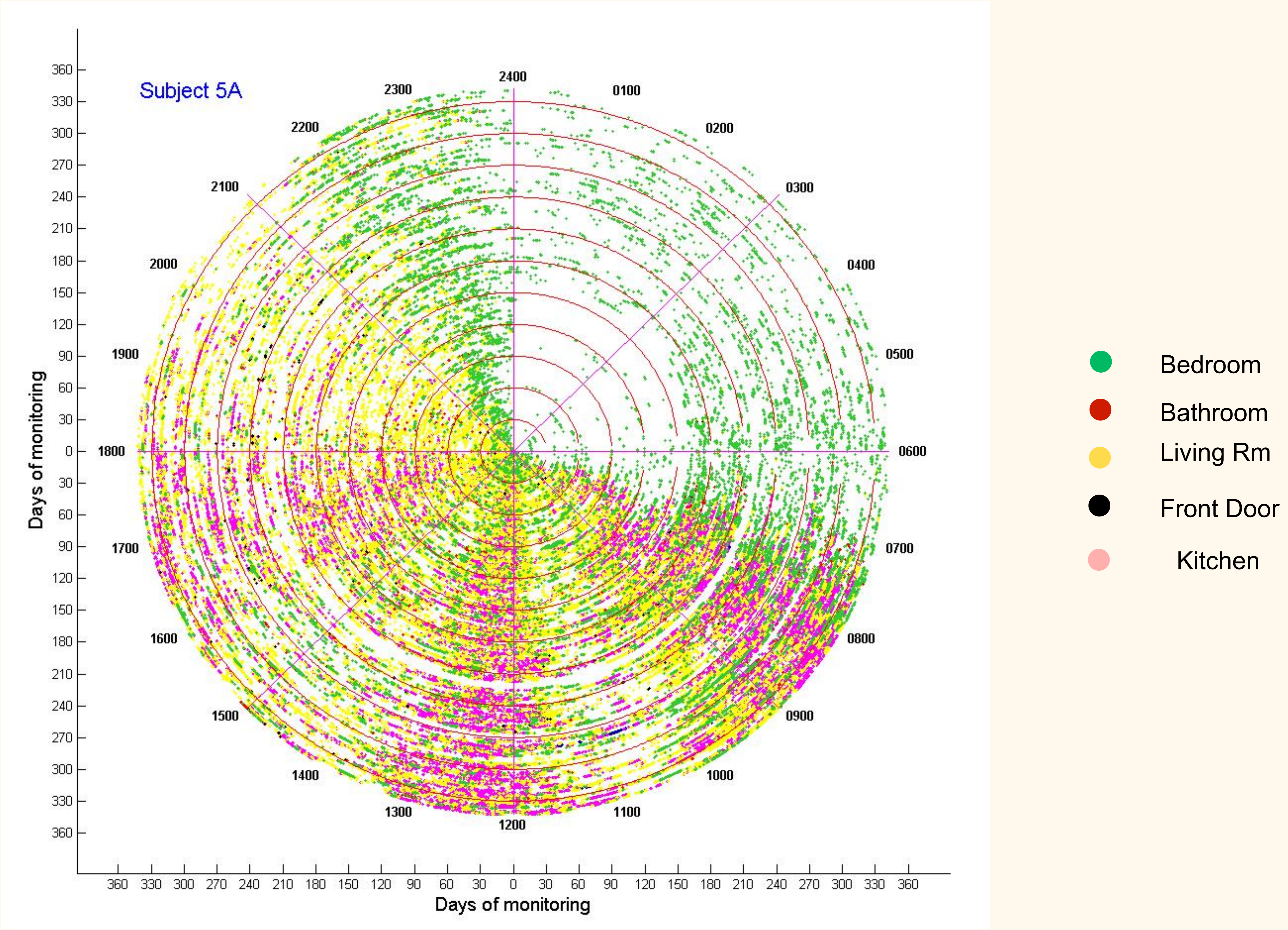


Figure courtesy Holly Jimison

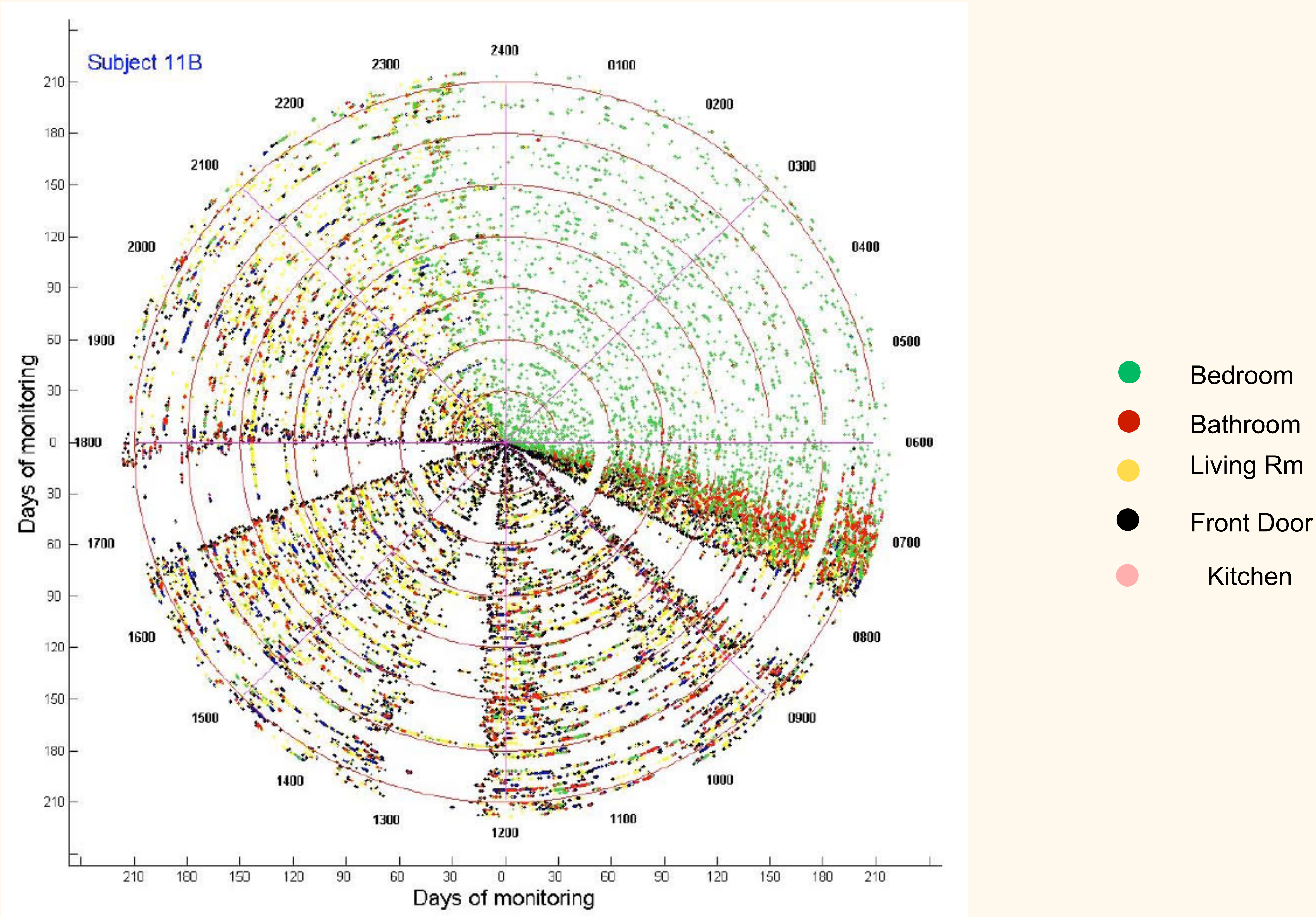


Figure courtesy Holly Jimison