Software Development Analytics

C. C. Shelley
OXFORD SOFTWARE ENGINEERING Ltd
9 Spinners Court, 53 West End,
Witney,
Oxfordshire
OX28 1NH
www.osel.co.uk
shelley@osel.netkonect.co.uk
Tel. +44 (0) 1993 700878

"Data, data, data, I'm surrounded by data. Can't anyone give me the information I really need."

Alexander the Great

Objective

- To persuade you that you can get valuable information, *now*, from data *you already have*.

Background

- Why now?
  - Software measurement has traditionally been seen as synonymous with s/w measurement *programmes* putting in place specialist data collection tools and repositories, intended for organizational information needs – estimation, monitoring of productivity, improvement etc., managed by specialists
  - It hasn't worked
  - But recently three essential elements have converged that make genuinely effective (and cost effective) software measurement a real possibility....

The three essential elements

- **Sophisticated development /test/management tools sets, supporting diverse aspects of software and systems development. These collect data by default – lots of data – that is rich in information and largely unregarded (and consequently undistorted)**

- **A good and improving understanding of the nature of software development (due, in part, to the early 'lightweight methodologists'). We can begin to really understand what these data can tell us.**

- **Hidden in plain view we have a powerful and proven set of analytical techniques - well suited to extracting information from our messy, idiosyncratic software development data - that are:**
  - **Widely applicable, robust, progressive, easy to use, teachable**
  - **Tacitly understood by statistics community, and familiar to field scientists and technical workers**
  - **But little known, or valued by the software community**

So what?

applying Benjamin Franklin's 'prudential algebra'...

Pros

Information from analysing our own data:

- will increase understanding
- can improve decision making
- may persuade others
- and increase confidence

Cons

Information from analysing our own data:

- is incomplete
- is a distraction
- tells us nothing new
- can be misused by others
- and incurs costs

A statement of belief

- The ability to measure and analyse software data is useful to *all* software developers, management and other staff as an everyday tactical tool - not just software measurement specialists *
- Therefore measurement and analysis should be pervasive throughout the software community...
- ...in much the same way that it is in other professions - like engineering, social sciences, psychology, medical, etc..
- But we have come to expect and accept dysfunctional 'big metrics'.
- And only a minority of the software community have the technical background, or inclination, to analyse their own software data.

Caveats (there are always caveats)...

- The new software measurement requires:
  - recognition that there is an opportunity to access valuable, extant information
  - A real understanding of the data context
  - discriminating data selection or acquisition
  - careful verification of the data

...other wise GIGA

...Caveats *(there are always caveats)...*

- And, critically:
  - An appreciation of data ownership
    - Whose is it ?
    - Who has access to it – and doesn't or shouldn't ? **
    - How will this be assured ?

...other wise you will not get the chance to do it again

...Caveats *(there are always caveats)*

- There is a real 'information horizon' that attenuates the information content of data as it moves away from its origins and context

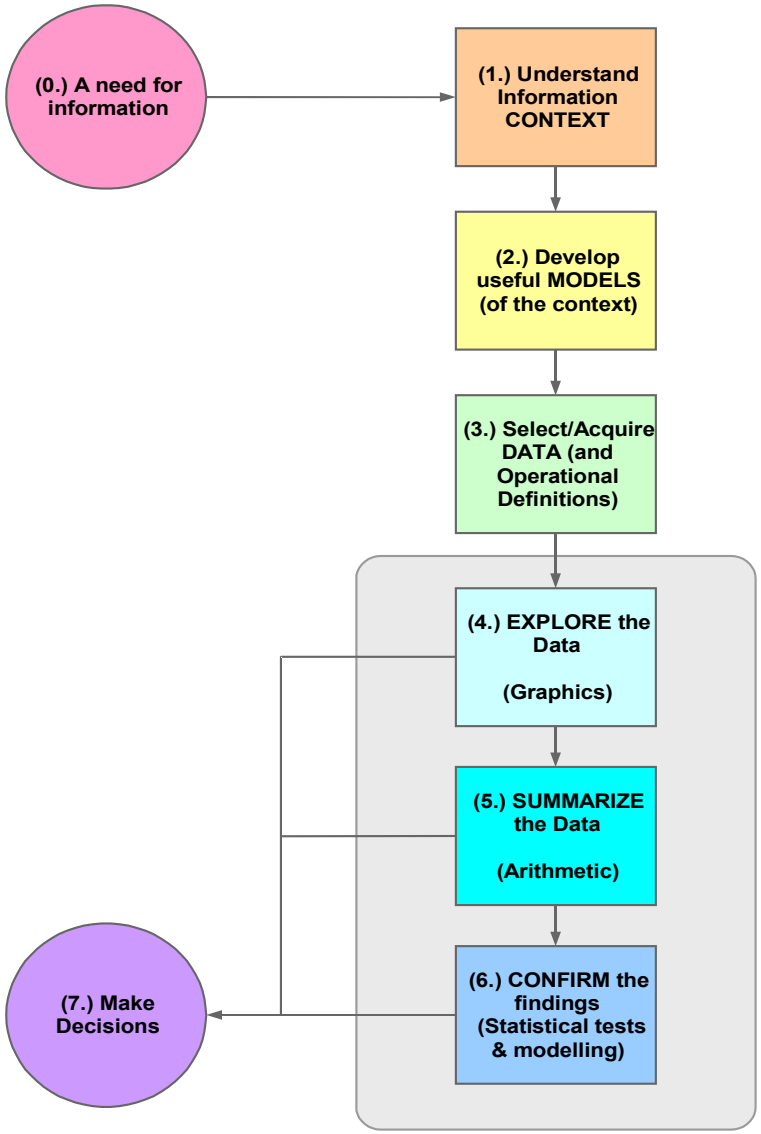
EAF

It has three steps

1) **Explore your data (easy, and most useful, by far)**

2) **Summarize/characterize your data (if needed, and not so easy)**

3) **Confirm your findings (if needed, and can be d****d difficult)**



Exploring the data

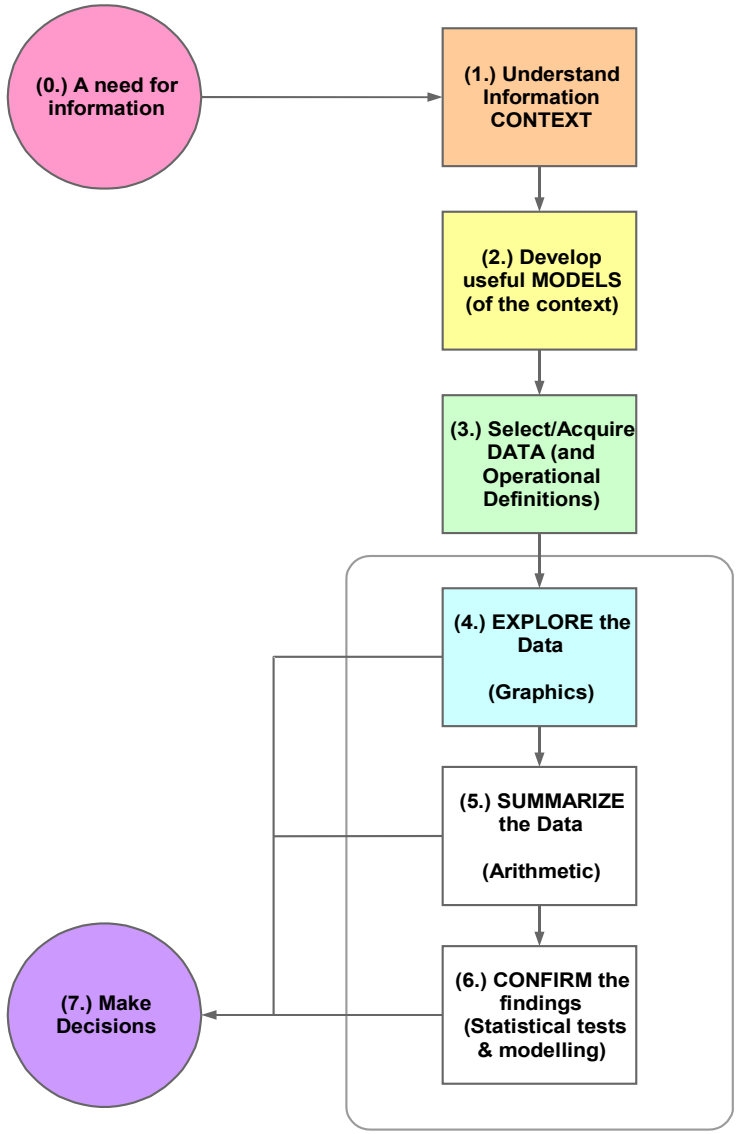
“Oh... that's funny.” *Anon*

...prelude to discovery

Step 4: Explore your data

How?

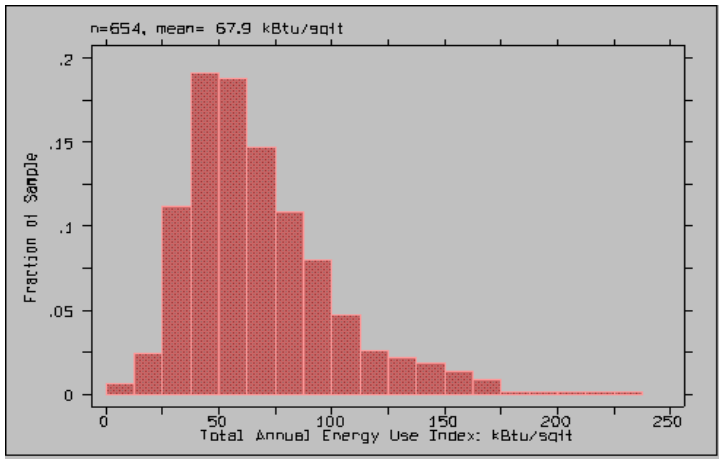
- **Data usually ends up in tables**
  - **rows of instances (say defect reports)**
  - **and columns of variables or attributes: when found, severity, description...**
- ***Explore*** **the individual variables...**
- **....and their relationships with each other ...**
- **....graphically**

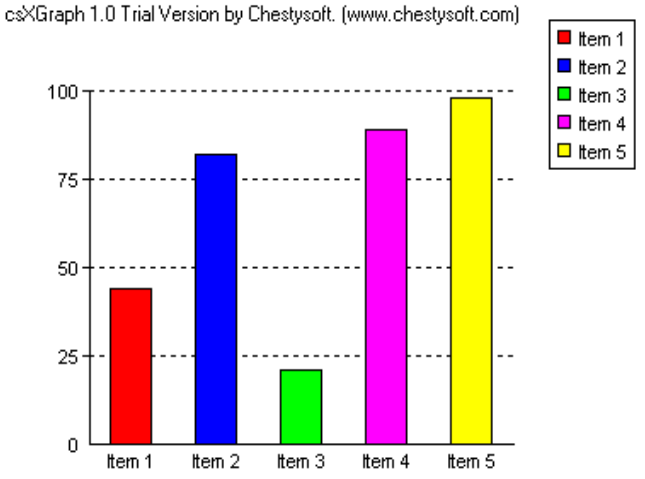


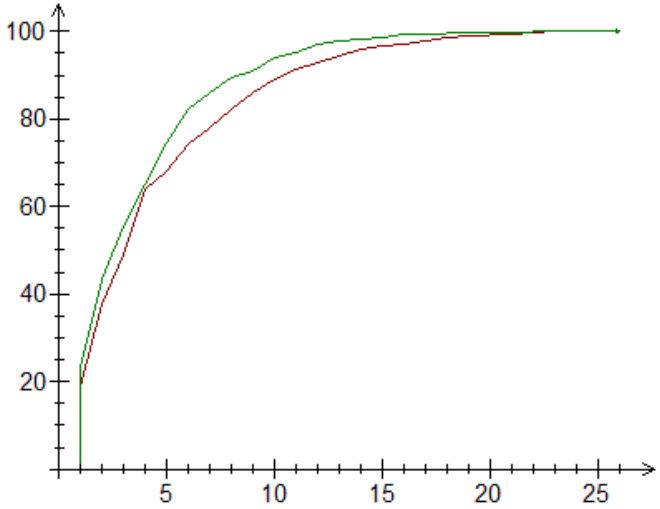
- **"Graphs are friendly"**
- **"Arithmetic often exists to make graphics possible"**
- **"Graphs force us to notice the unexpected; nothing could be more important."**

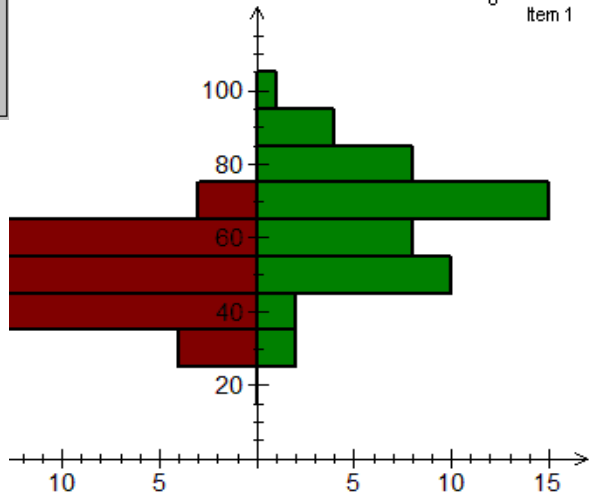
John Tukey, EDA, p157

...Exploring Data...



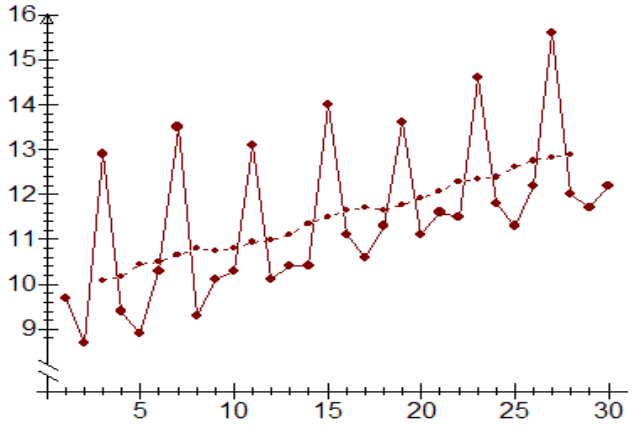


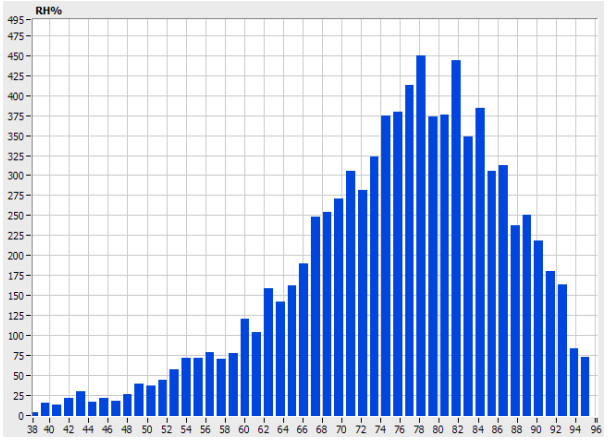


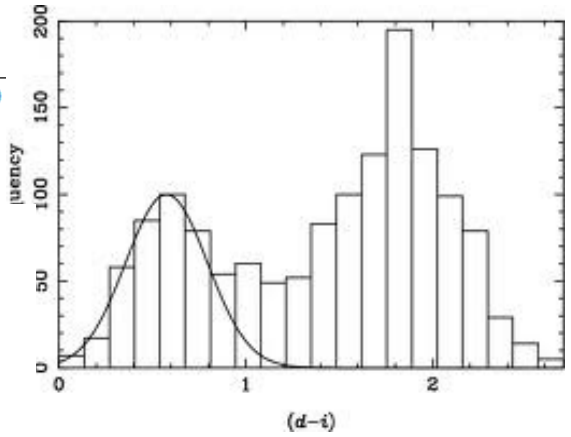


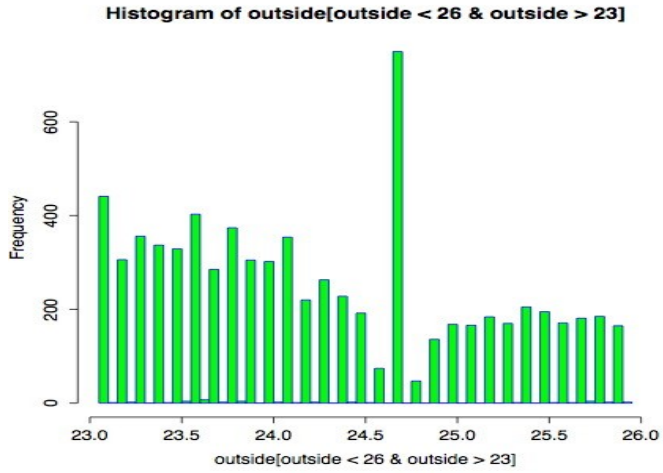
52	0
53	2 4
54	0 2 8
55	0 0 4 4 5 5 5 6 6 7 9
56	0 0 0 0 0 2 4 5 6 8 9 9
57	0 0 0 7 8 9
58	0 0 0 0 4 7
59	0 0 0 0
60	0 0 0 0 3 3 7 8
61	0 5

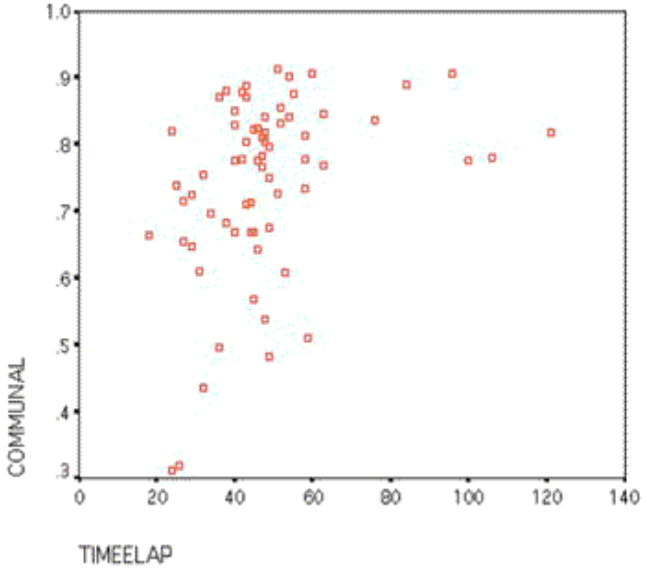
...Exploring Data...





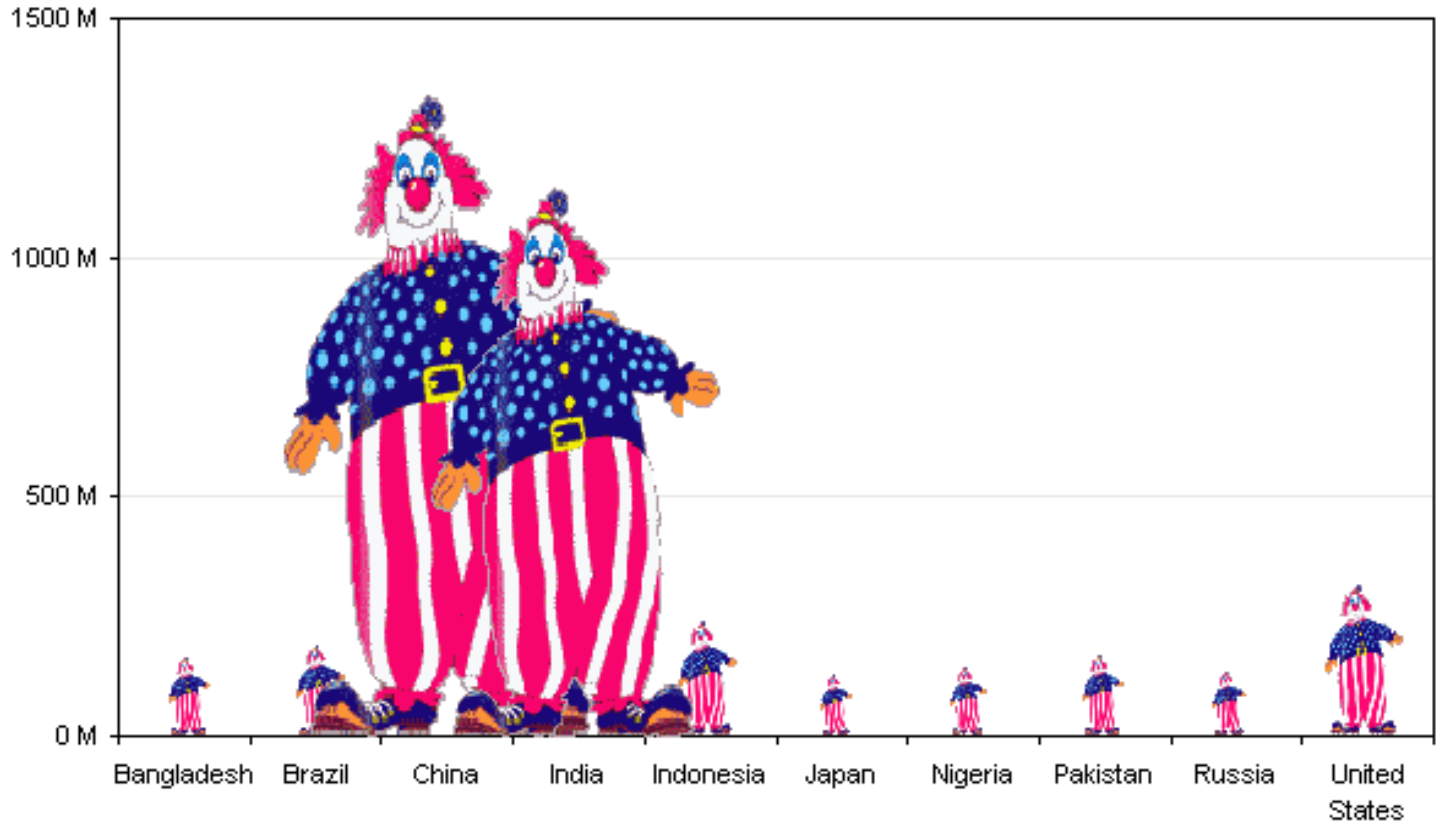




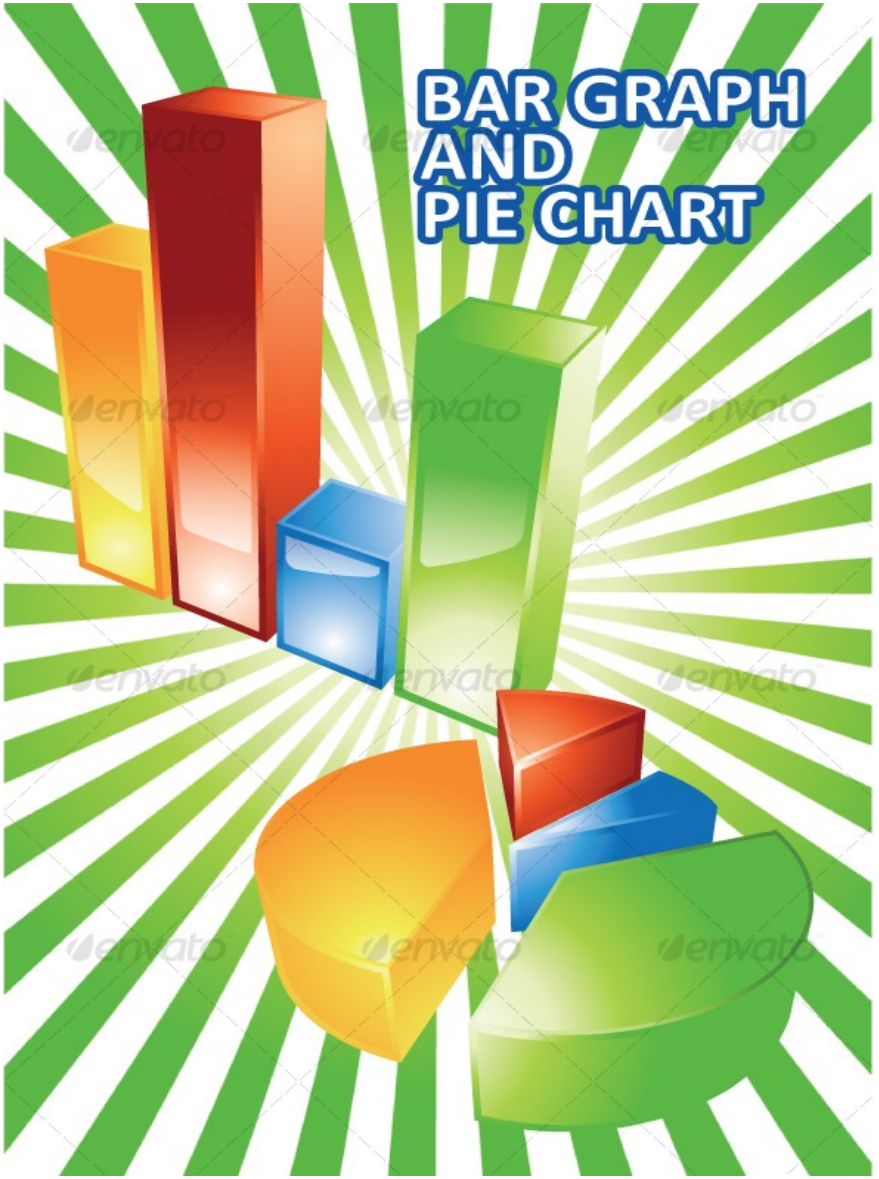


Don't !

Don't !



Don't !



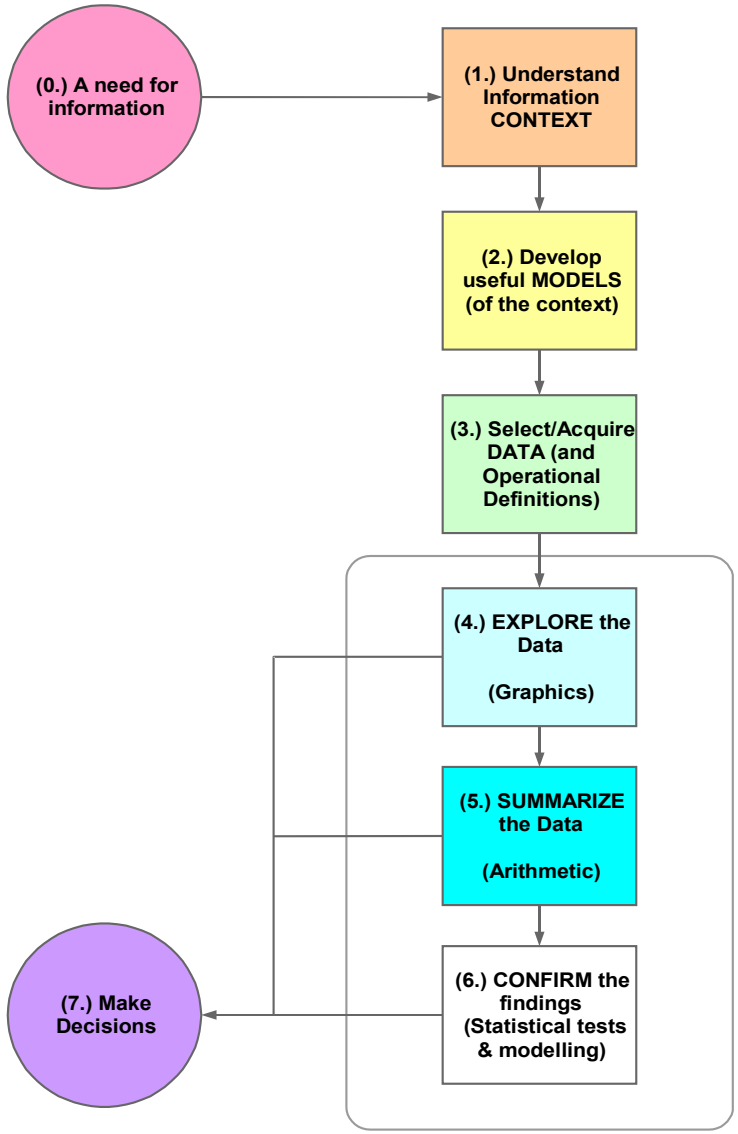
Step 5: Summarize your data

Why?

- **To identify, share and compare the cardinal points**

How?

- **Some arithmetic to find the interesting numbers**
  - **biggest, smallest, middle...**



Summarize...some arithmetic or some counting...

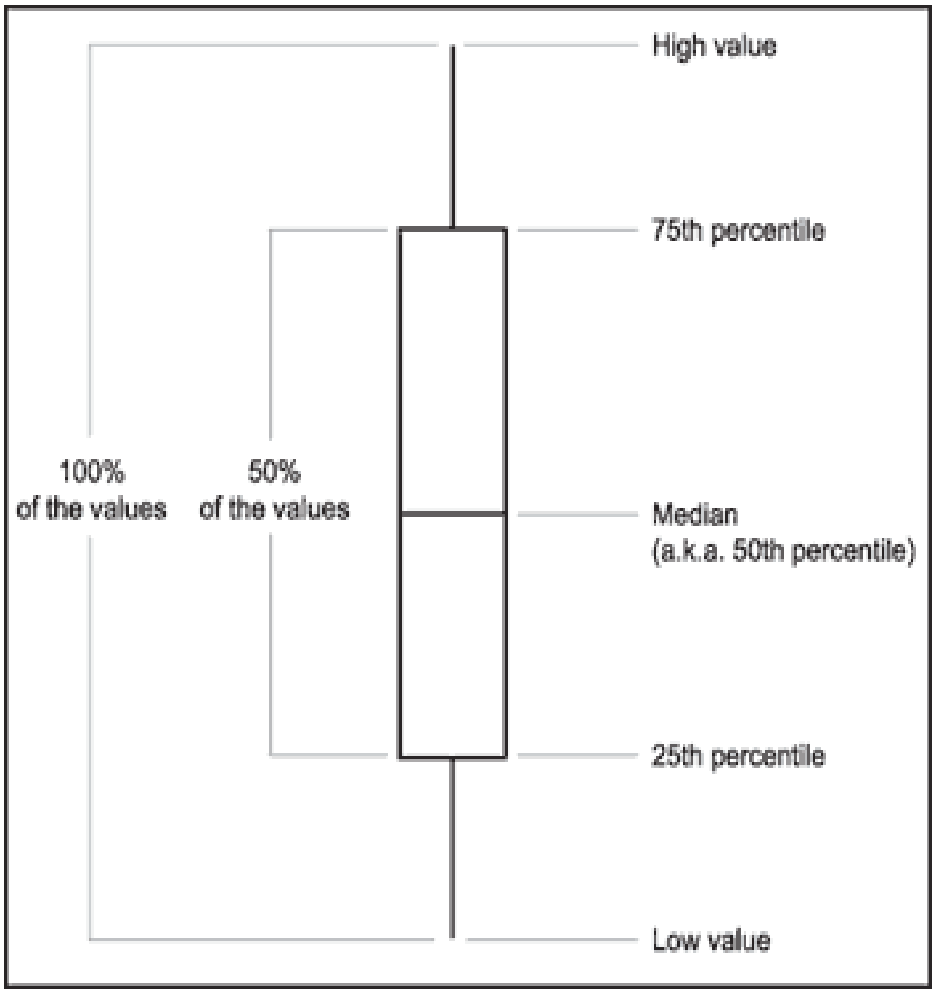
$$y = y_0 e^{\left(\frac{-x^2}{(2\,sd^2)}\right)}$$

$$\bar{x} = \frac{\sum x}{n}$$

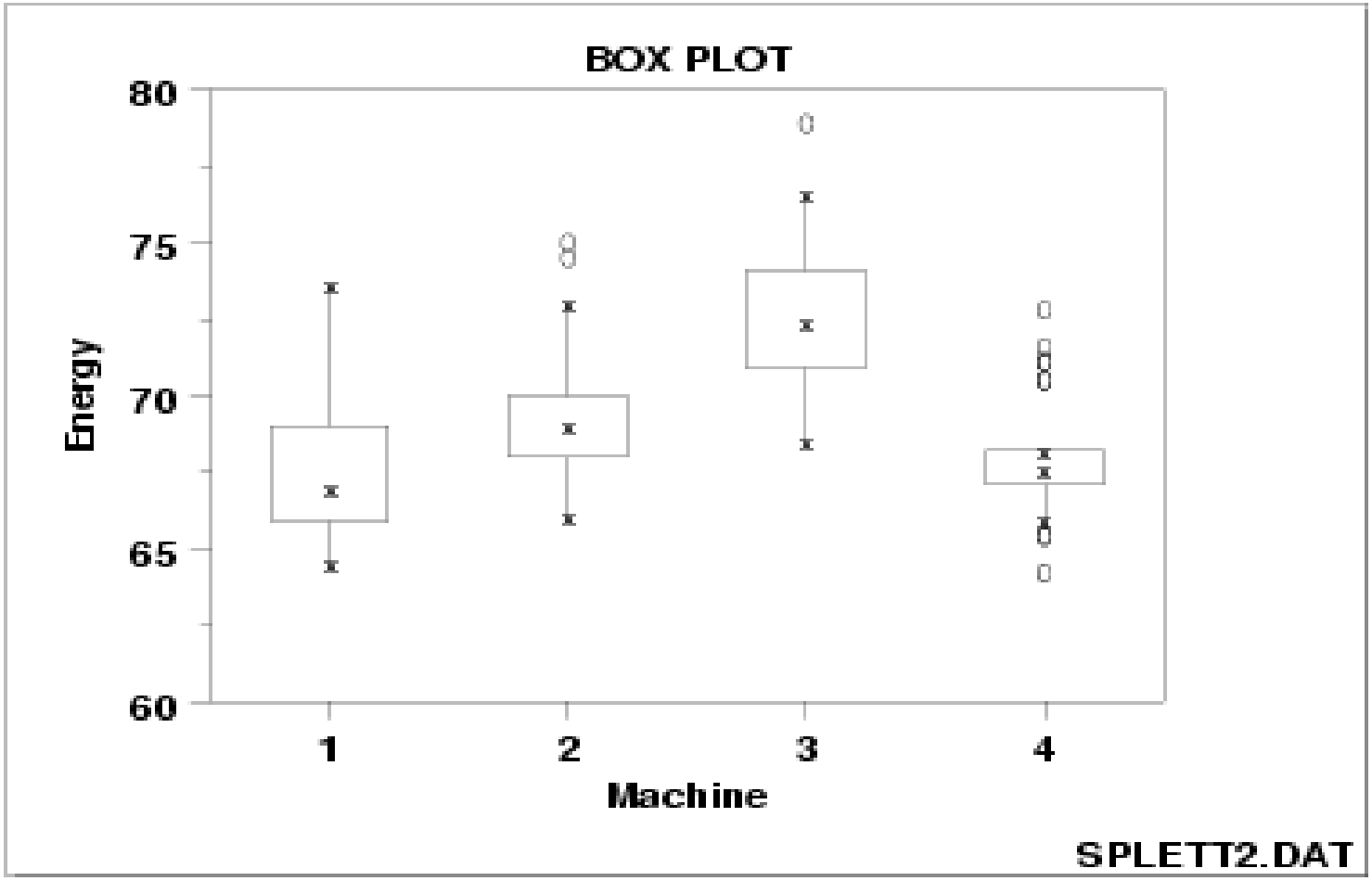
$$sd = \sqrt{\left(\frac{\sum (x_i - \bar{x})^2}{(n-1)}\right)}$$

$$var_i = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$$

$$CI_{95} = \bar{x} \pm t_{0.05} \frac{sd}{\sqrt{n}}$$



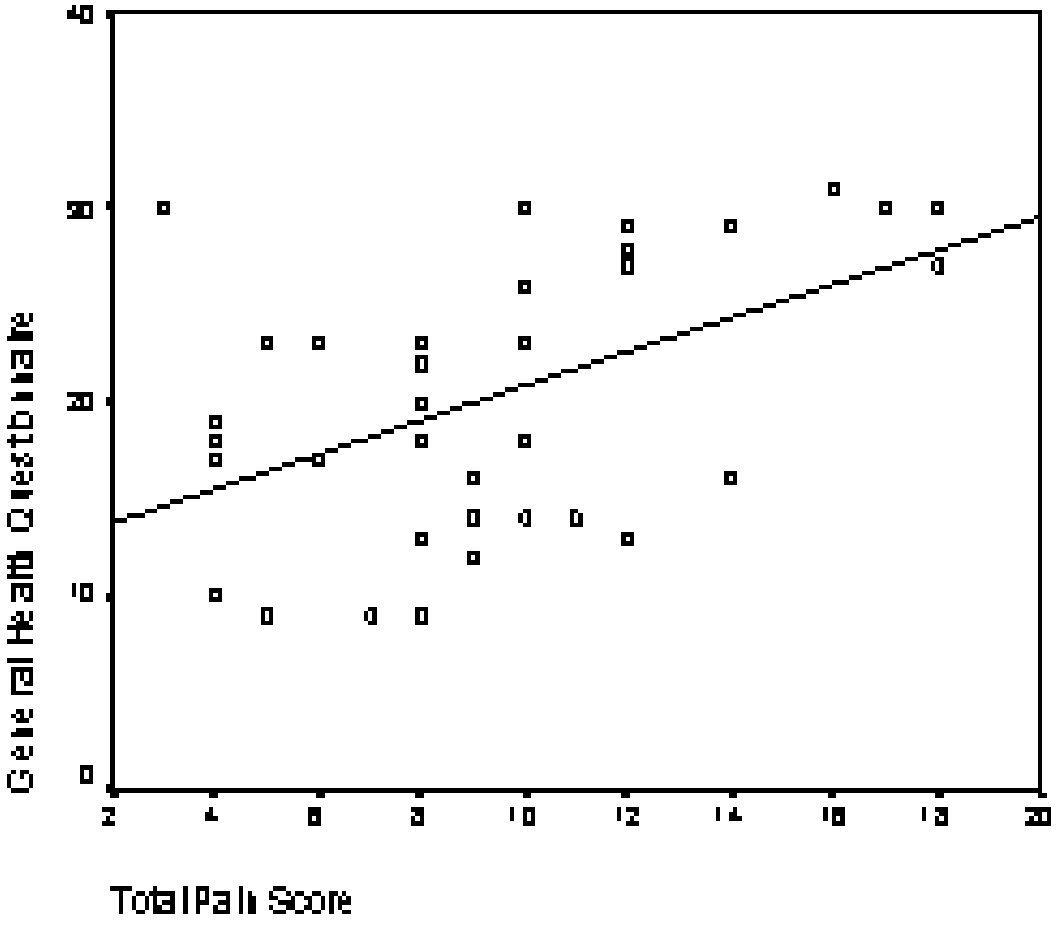
Summarize...some arithmetic or some counting...



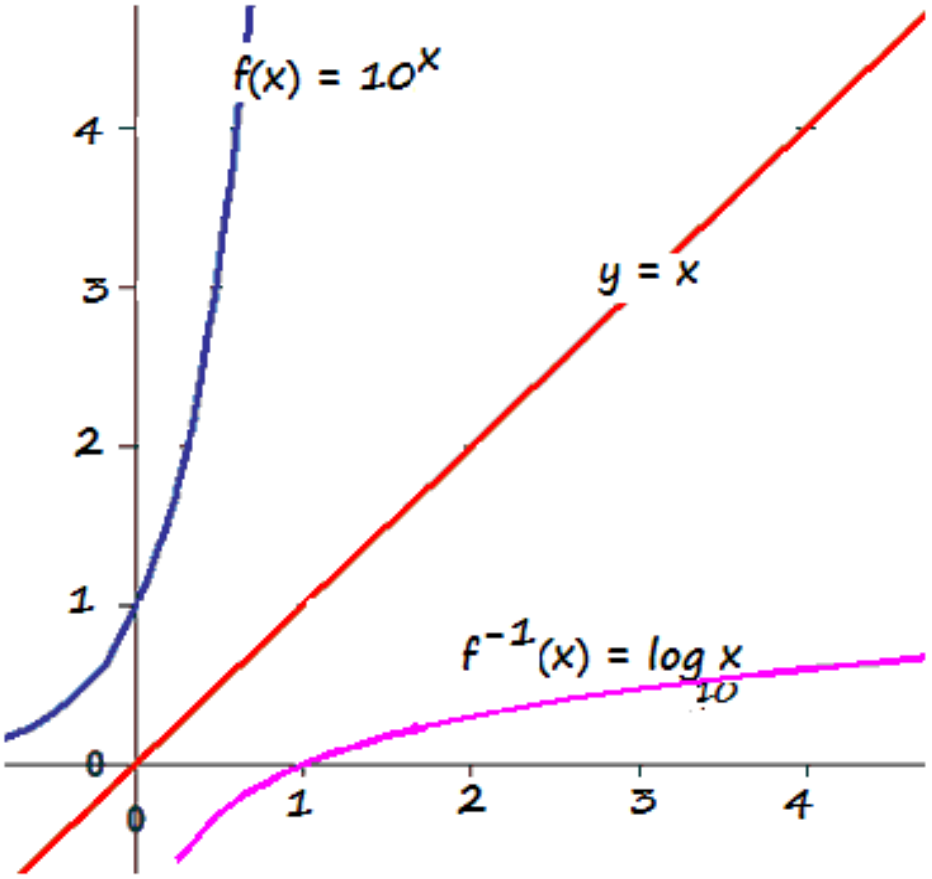
...Summarize…

$$slope = \frac{\sum ((x_i - \bar{x}) \times (y_i - \bar{y}))}{\left(\sum (x_i - \bar{x})\right)^2}$$

$$r = \frac{\sum ((x_i - \bar{x}) \times (y_i - \bar{y}))}{\sqrt{\left(\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2\right)}}$$

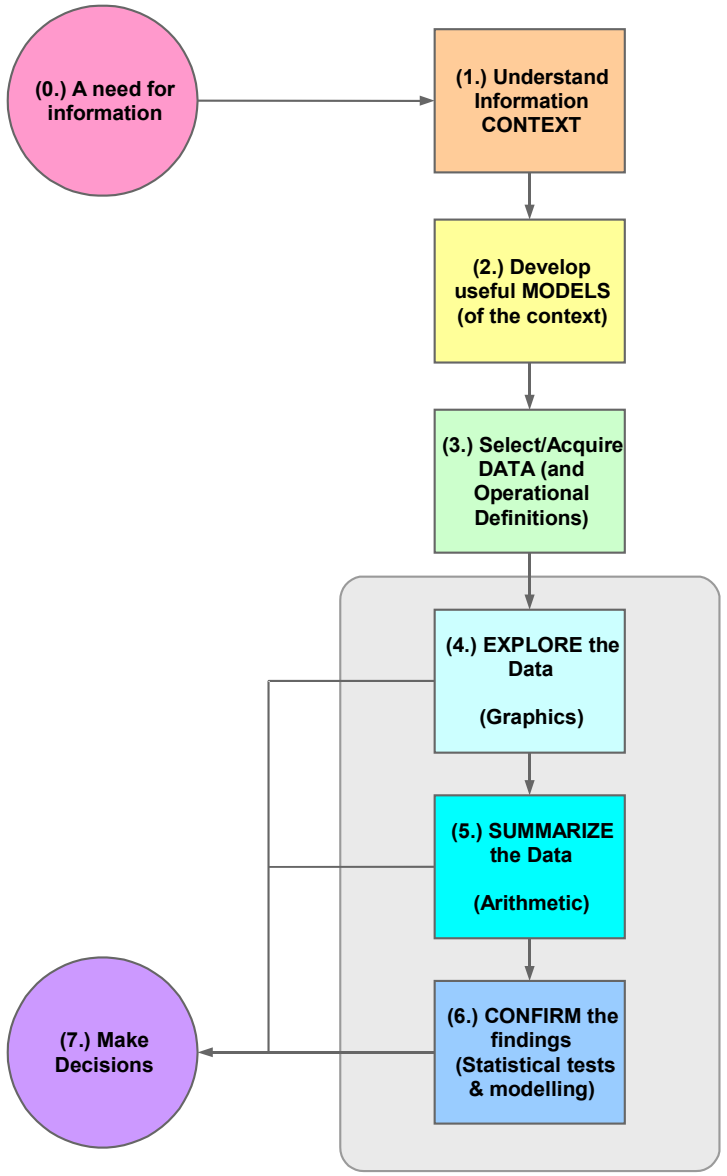


...Summarize…



Step 6: Confirm your findings

- You have explored the data...
- ..found information that may be useful for decision making
- You now wish to confirm what you have found...
- ...to quantify the findings and increase confidence in them by establishing their statistical significance, when you know their contextual significance



The classical statistical testing procedure...

1. State null hypothesis (H0) and its alternative (H1). (Decide what data to collect and under what conditions). Choose a statistical test (with its associated statistical model) for Testing H0.

2. From among the several tests that may be used select the one that best meets the needs and the assumptions on which the test is based

3. Specify a significance level and a sample size.

4. Find the sampling distribution of the statistical test under the assumption that H0 is true

5. On the basis of 1, 2, and 3 above define the region of rejection for the test.

6. Calculate the value of the test statistic. If it is in the region of rejection the decision is to reject H0. If the value is outside the value of rejection the decision is that H0 cannot be rejected at the chosen level of significance.

Adapted from Siegal and Castellan, p7

...The classical statistical testing procedure...

Scale Type	*Measure of Location*	*Measure of Dispersion*	*Measure of Association*	*Tests of Significance*
Nominal	mode	information	contingency correlation	chi square
Ordinal	median	percentiles	rank-order correlation	sign test, run test
Interval	arithmetic mean	standard deviation	product-moment correlation	t test, F test
ratio	geometric mean, harmonic mean	percent variation	correlation ratio	

...The classical statistical testing procedure...

delivers a value of p...

$$p = \langle x \rangle$$

...The classical statistical testing procedure

"...the almost universal reliance on ***merely*** refuting the null hypothesis is a terrible mistake, is basically unsound, poor scientific strategy...."

(Meehl, 1978: 817)

(my italics)

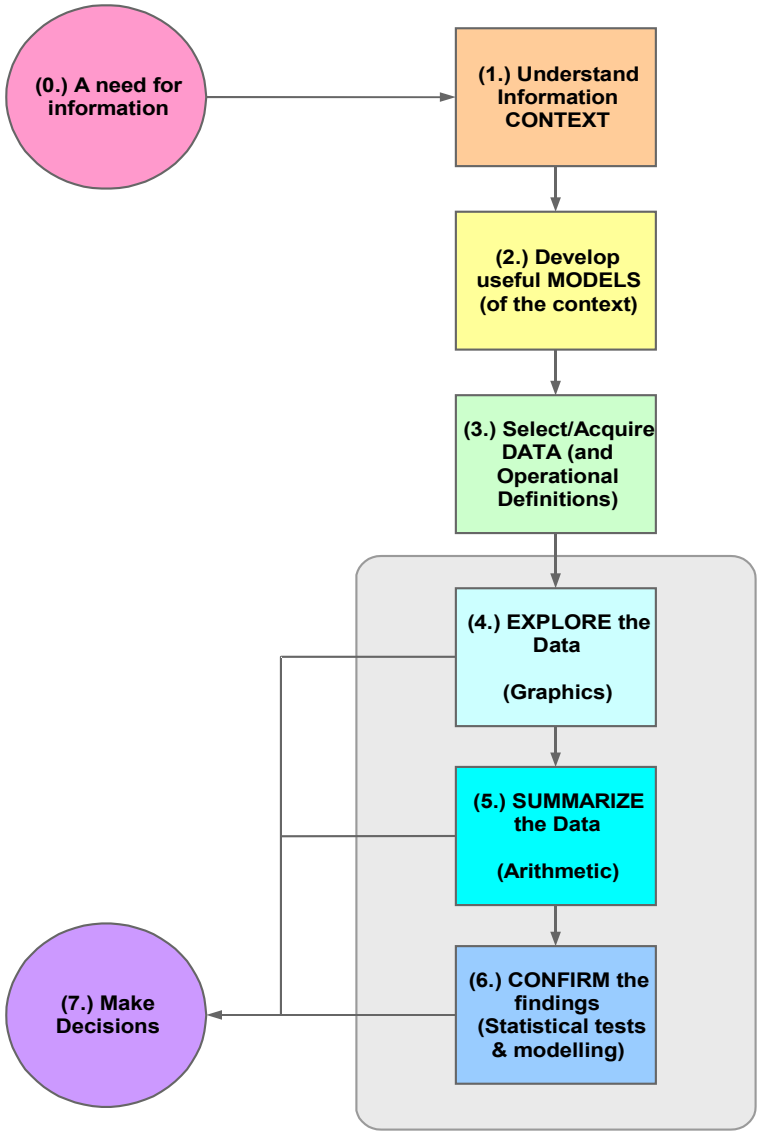
EAF

Recap: the three steps...

1) **Explore your data (easy, and most useful, by far)**

2) **Summarize/characterize your data (if needed, and not so easy)**

3) **Confirm your findings (if needed, and can be d****d difficult)**

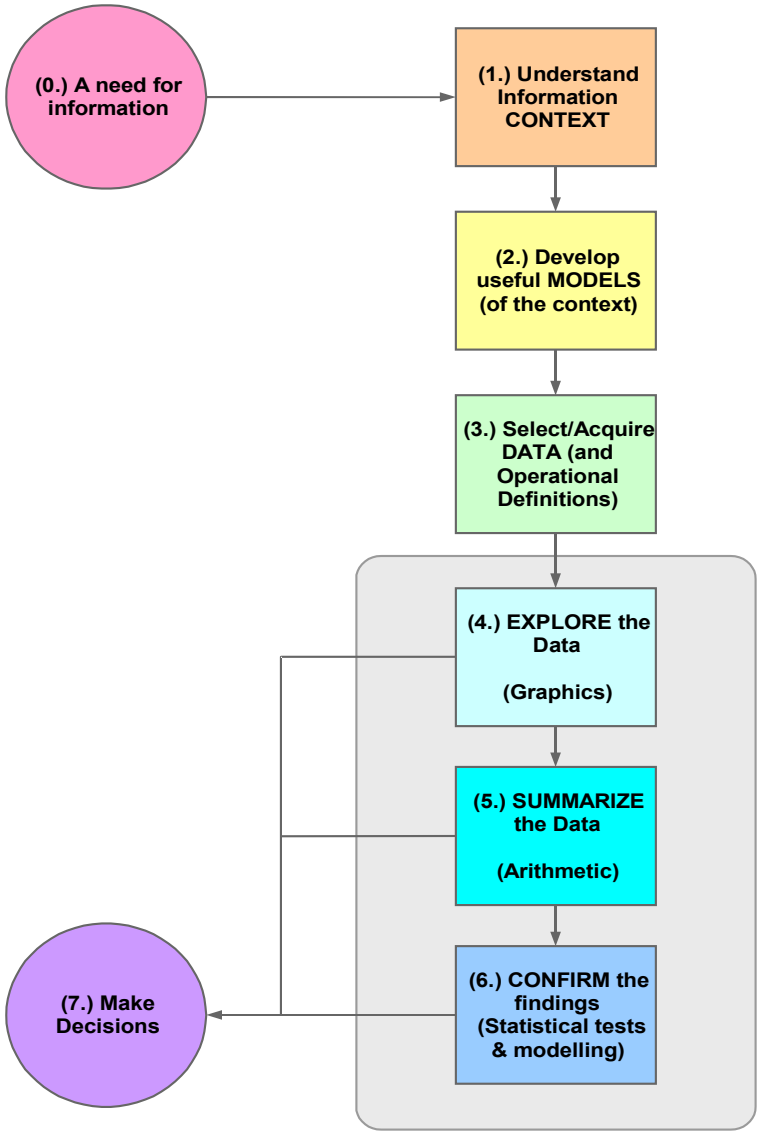


This framework can be extended...

- Refined and adapted graphics, tuned to software (dev) characteristics
- 'Bayesian' statistical models
- 'No data' decision making techniques (minimax, game theory...)
- Modelling and simulation (e.g. Monte Carlo)
- ...

Conclusion

- **We have the data (now)**
- **We have the means to get to the information (now)**
- **Do you want the information ?**



OXFORD
SOFTWARE ENGINEERING
LIMITED

9 Spinners Court, 53 West End,
Witney,
Oxfordshire
OX28 1NH

www.osel.co.uk
shelley@osel.netkonect.co.uk
Tel. +44 (0) 1993 700878

Here are four bivariate (i.e. pairs of variables) data sets....

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

‘Anscombe’s Quartet’ – American Statistician 1973

These four data sets all share
A common set of summary statistics...

N = 11

Mean of X's = 9.0

Mean of Y's = 7.5

Regression line = Y + 0.5X

Standard error of estimate of slope = 0.118

t = 4.24

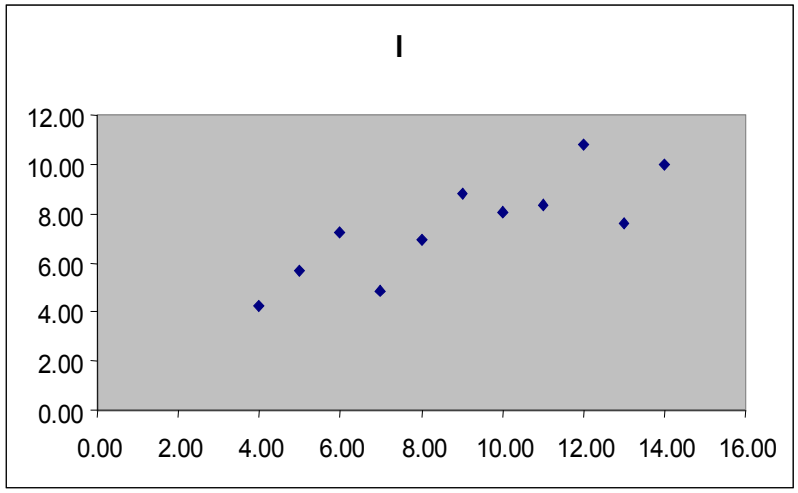
Sum of squares $X - \overline{X} = 110.0$

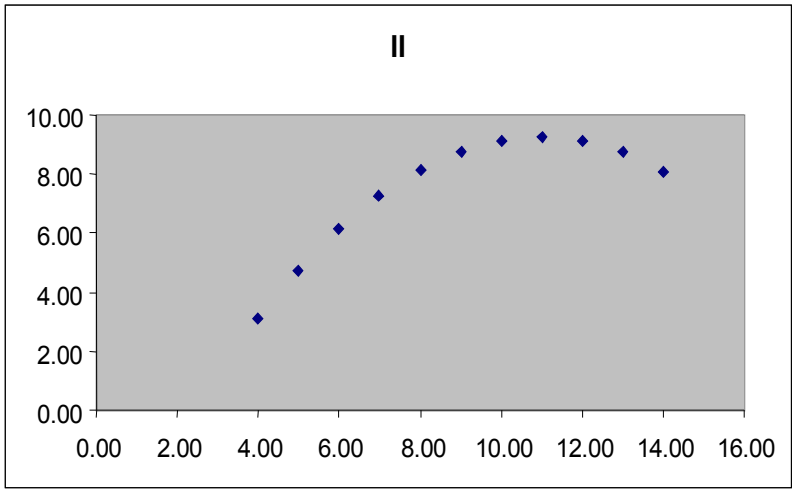
Regression of sum of squares = 27.50

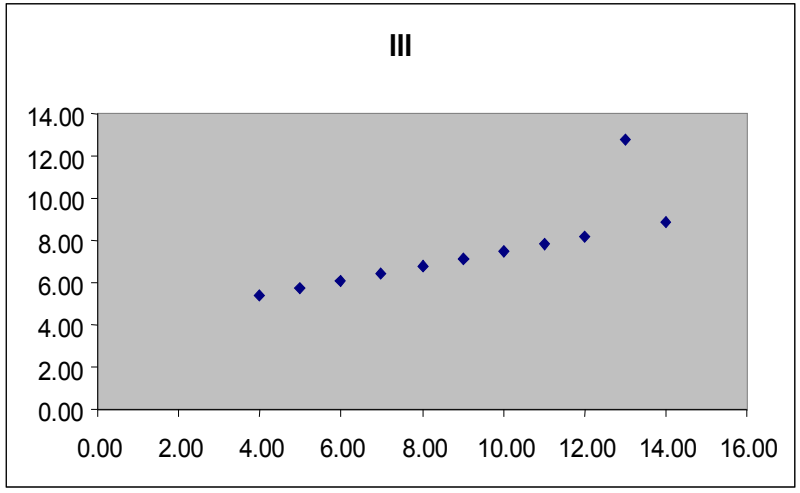
Residual sum of squares of Y = 13.75

Correlation coefficient = 0.82

$R^2 = 0.67$


I
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0.00 2.00 4.00 6.00 8.00 10.00 12.00 14.00 16.00


II
10.00
8.00
6.00
4.00
2.00
0.00
0.00 2.00 4.00 6.00 8.00 10.00 12.00 14.00 16.00


III
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0.00 2.00 4.00 6.00 8.00 10.00 12.00 14.00 16.00

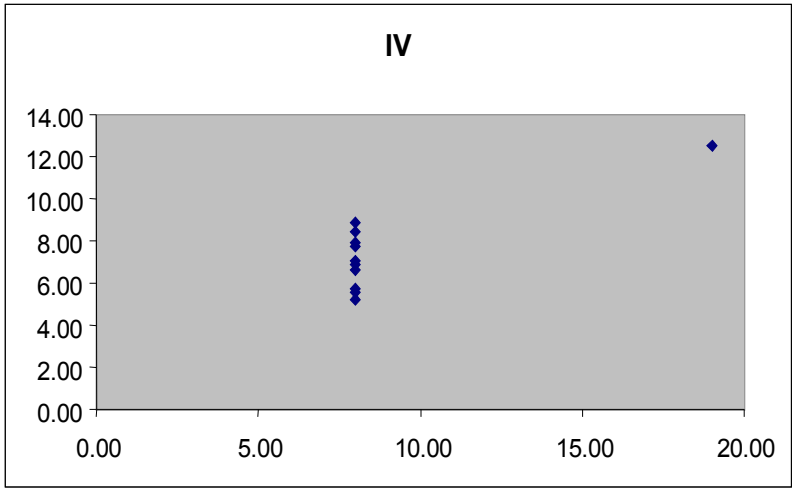
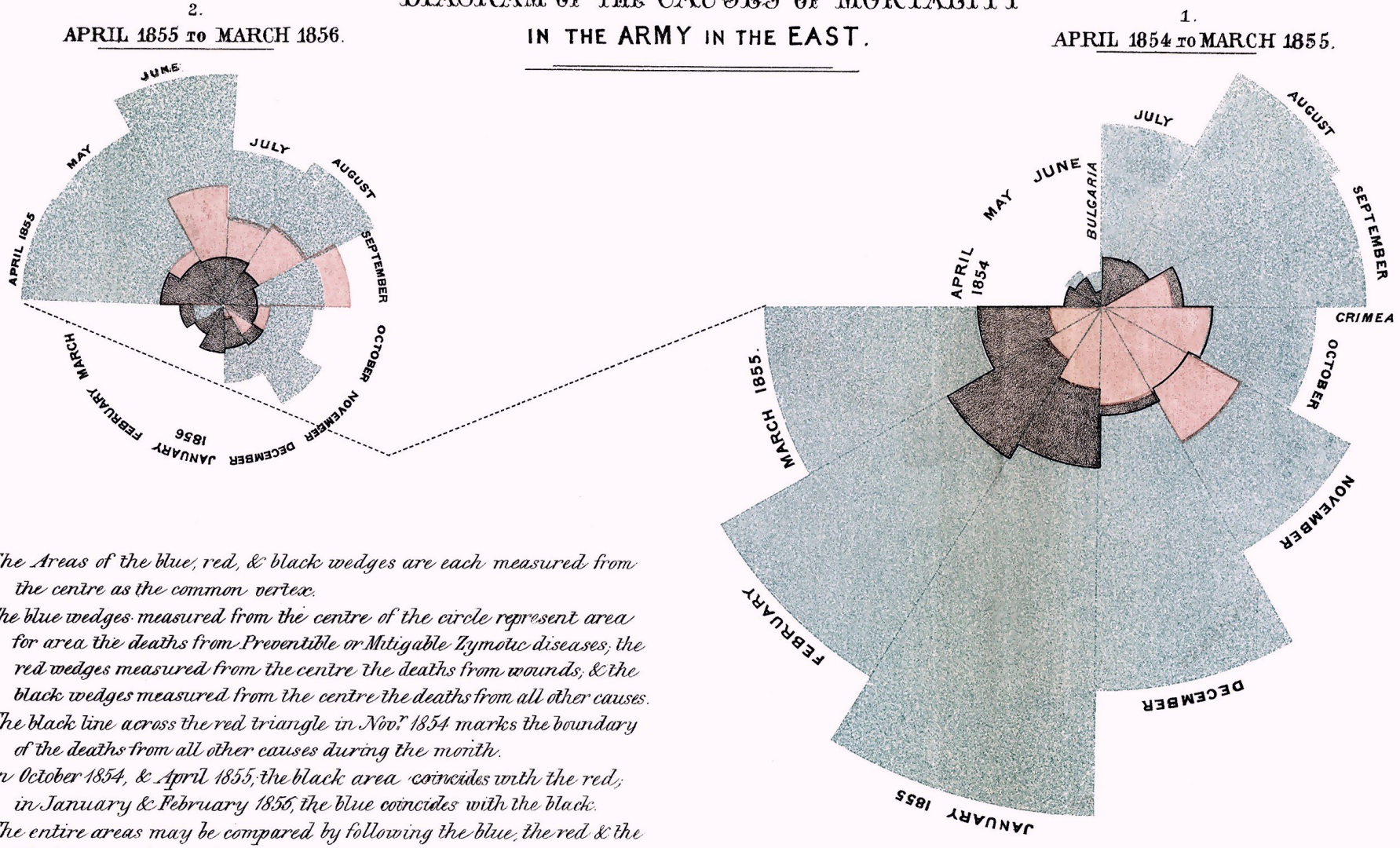

IV
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0.00 5.00 10.00 15.00 20.00

DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

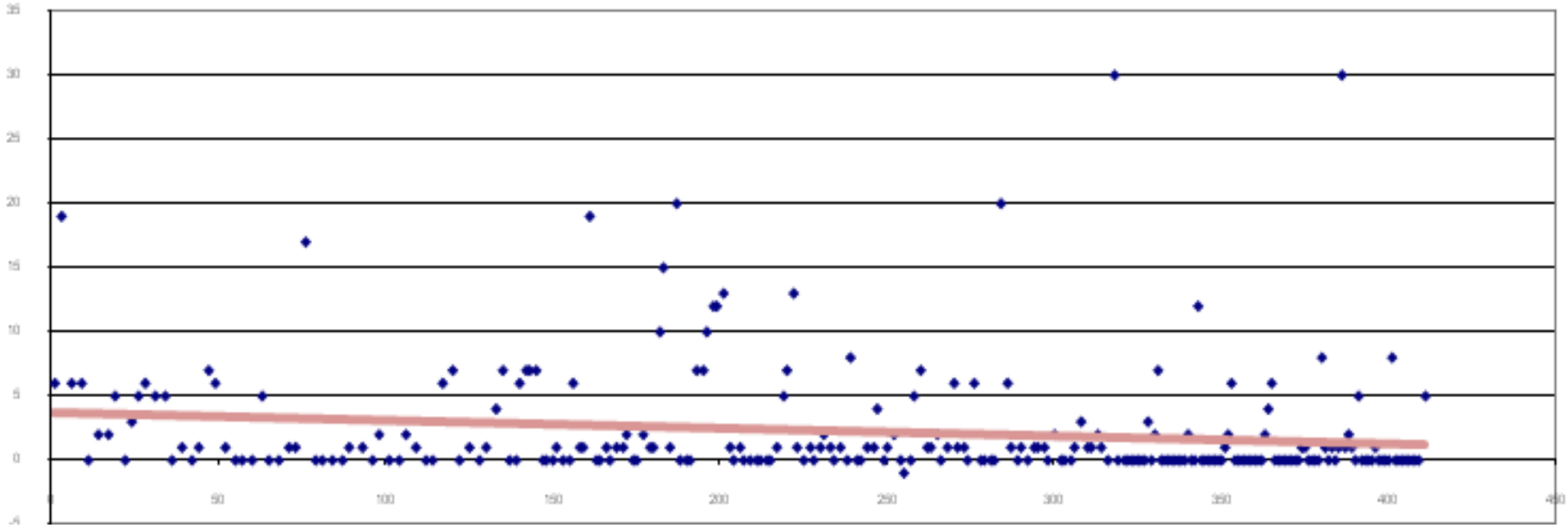


Figure 4.2 – Days delay for deliveries from April 2007 to August 2009

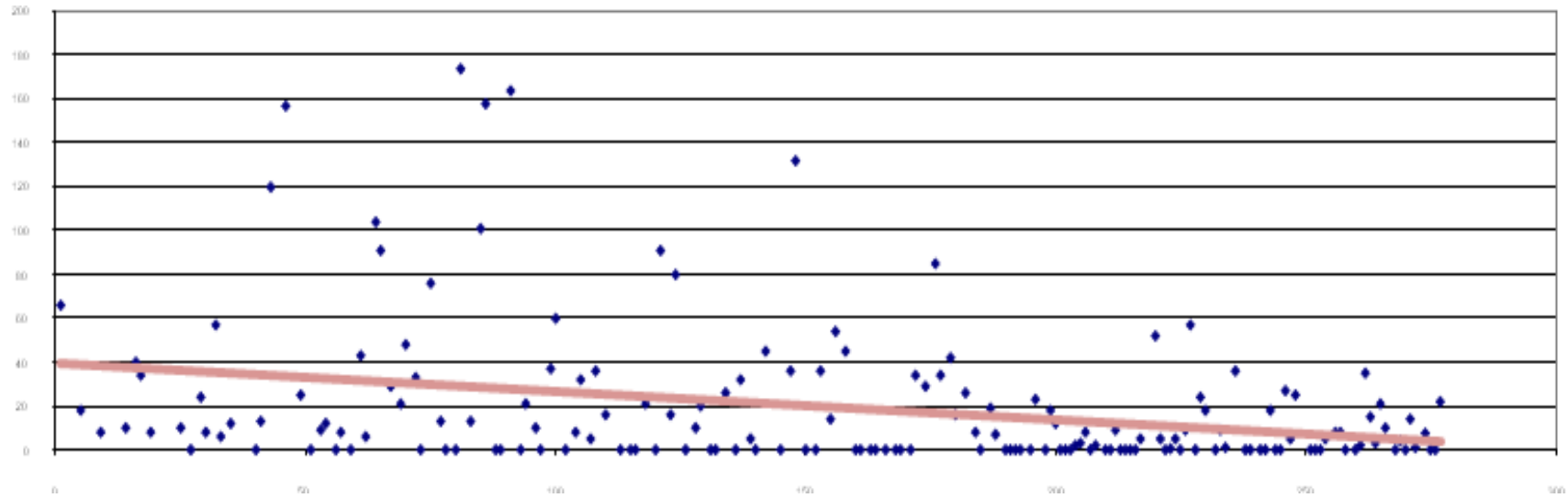


Figure 6.1 – Number of storypoints introduced during sprints from January 2008 through to August 2009.

'You can't control what you can't measure'

Tom DeMarco

'You can't control what you can't measure'

Tom DeMarco

'In my experience absolutely everyone who does it screws it up. So I think the concept's wrong. And I think it's kind of pointless to think, "Well, if they did it right, it would be OK." They don't do it right. The people who want to do it are inclined to do it wrong.'

Tom DeMarco

11a "Eliminate numerical quotas for the work force…… A quota is a fortress against improvement of quality and productivity. I have yet to see a quota that includes any trace of a system by which to help anyone do a better job. A quota is totally incompatible with never ending improvement. There are better ways."

W. Edwards Deming

11b "Eliminate numerical goals for people in management. Internal goals set in management of a company, without a method, are a burlesque… …Focus on outcome is not an effective way to improve process or an activity… …management by numerical goal is an attempt to manage without knowledge of what to do, and is in fact usually management by fear."

W. Edwards Deming

"Ownership of measures should be with the people who do the work."

"Managers tend to think they 'own' the numbers. They should not."

"People do what you count, not necessarily what counts."

"Attention to output can increase costs."

"Using measures for improvement starts with thinking differently."

John Seddon 'I Want You To Cheat'

If Software Metrics leads to something like Taylorian "scientific management" in software development, I for one, will bow my head in shame.

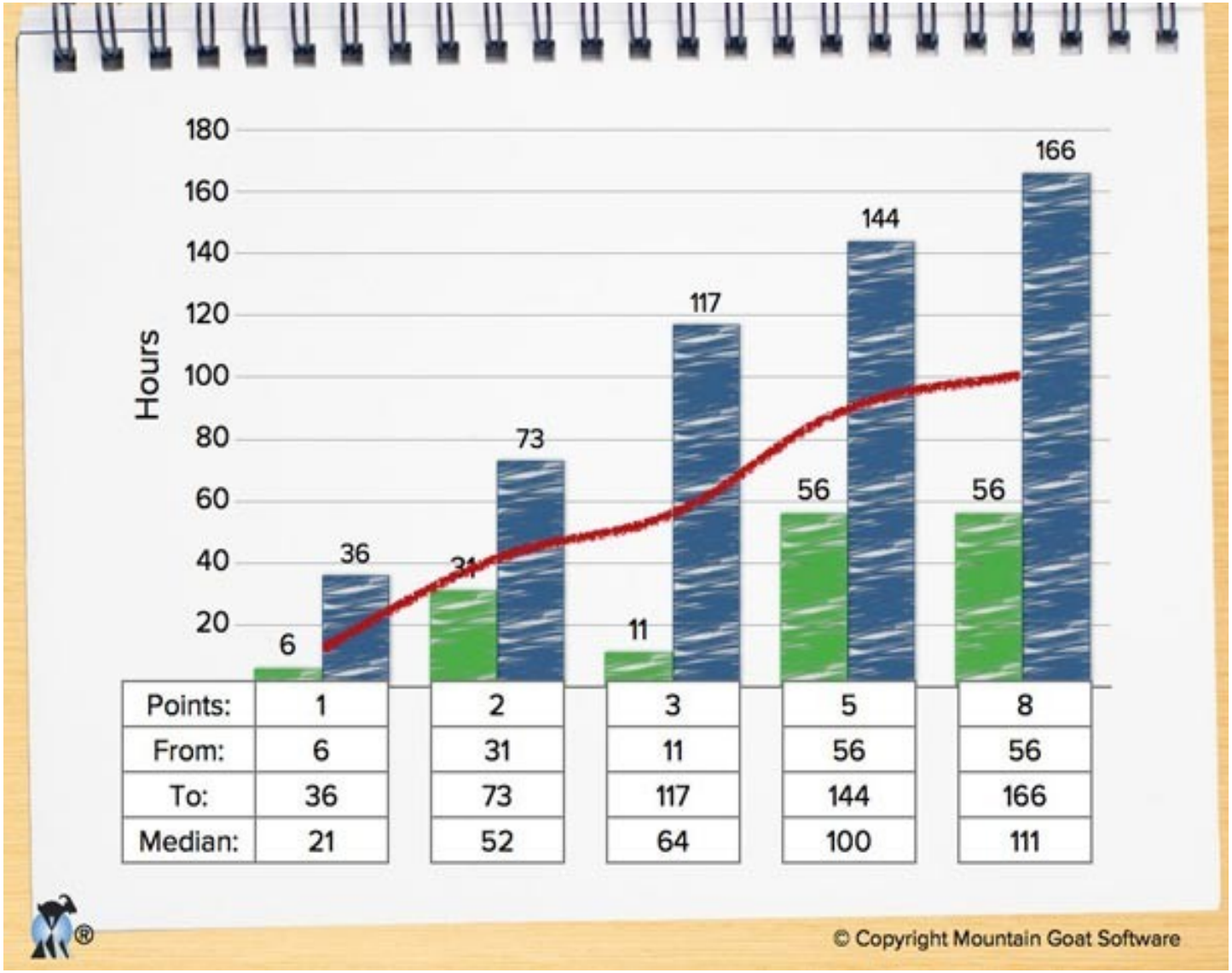
Gerald M. Weinberg

writing in the foreword of Toms Gilb's 'Software Metrics', the first s/w metrics book

'T

'The good news is that you ***can*** succeed in producing a culture conducive to measurement. There are organizations in which people have given themselves completely to the pursuit of organizational goals… …organizations in which members hunger for measurement as a tool that helps get the job done… …To use measurement inappropriately would betray a sacred trust, and no one would consider such a betrayal.'

Robert D. Austin


Hours
180
160
140
120
100
80
60
40
20
6
36
31
73
11
117
56
144
56
166
Points: 1 2 3 5 8
From: 6 31 11 56 56
To: 36 73 117 144 166
Median: 21 52 64 100 111
© Copyright Mountain Goat Software