# *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

## *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

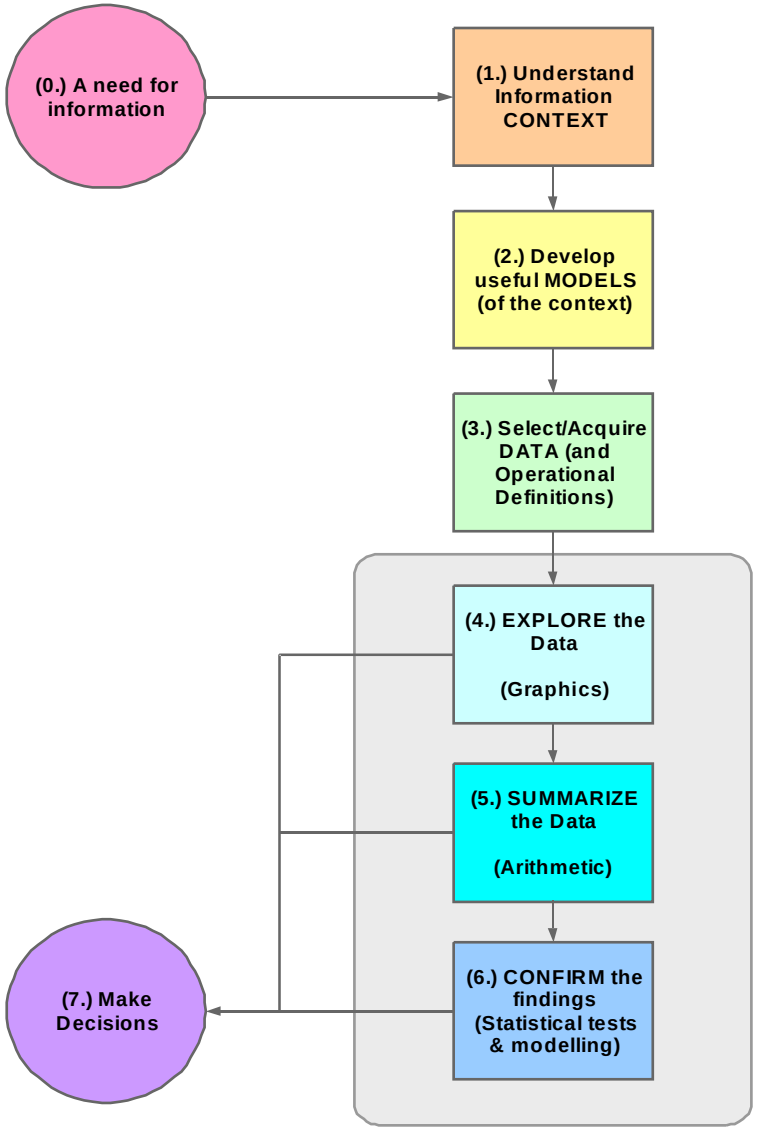
# *EAF*

**It has three steps**

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

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

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



## *Exploring the data*

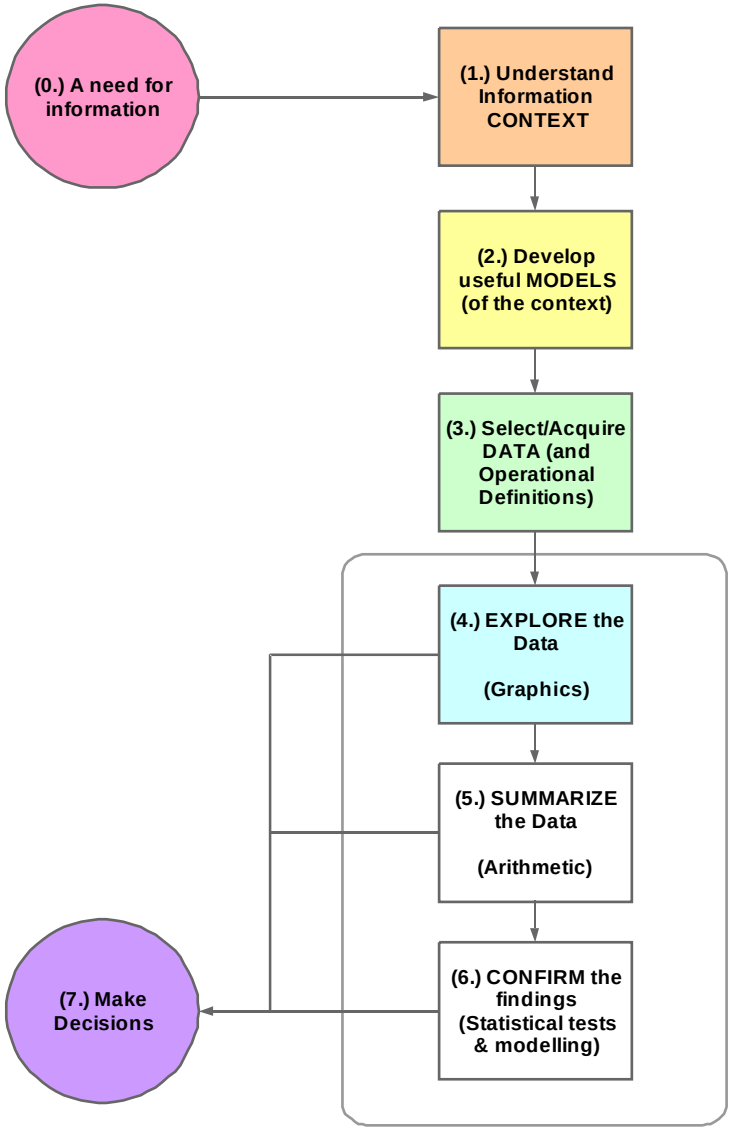
**“Oh... that's funny.”** *Anon*

***...prelude to discovery***

# *Step 4: Explore your data*

**How?**

- **Data is usually ends up in tables (tabulated)**
  - **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**

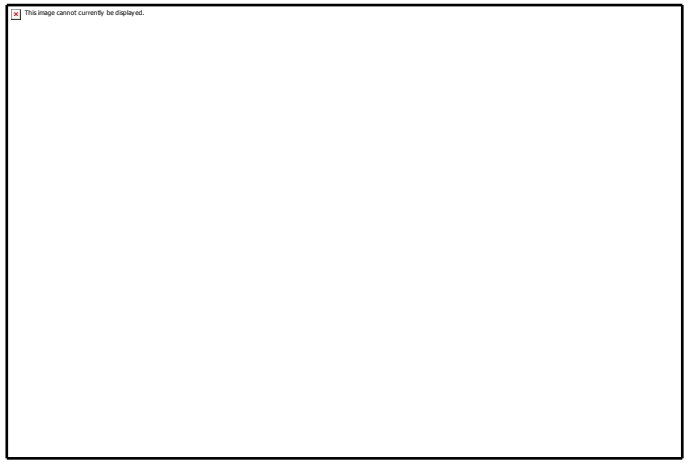


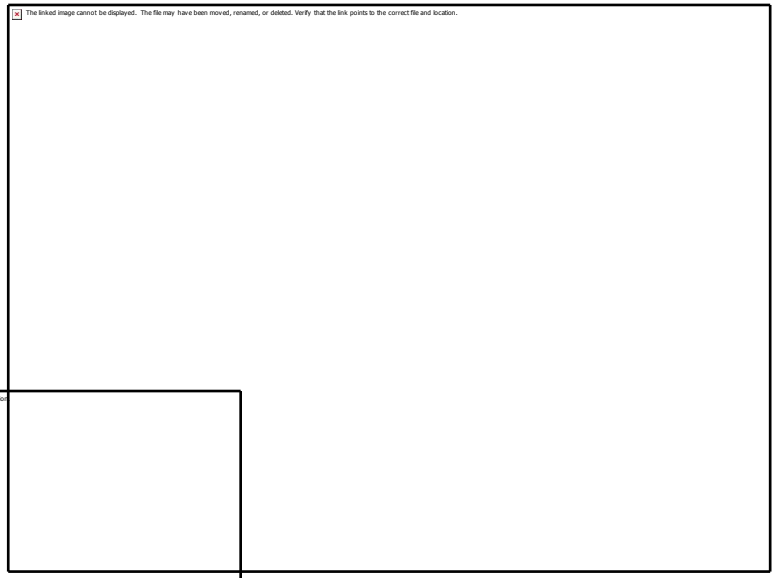
## *Graphical Methods*

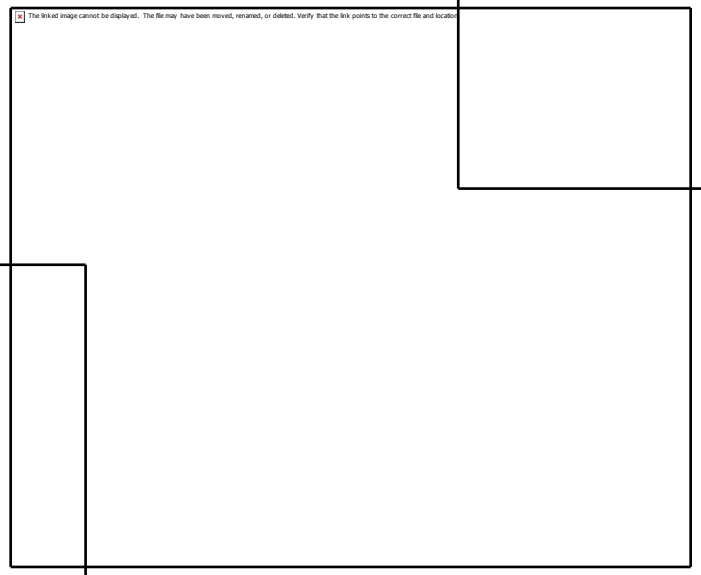
- "Graphs are friendly"
- "Arithmetic often exists to make graphics possible"
- "Graphs force us to notice the unexpected; nothing could be more important."

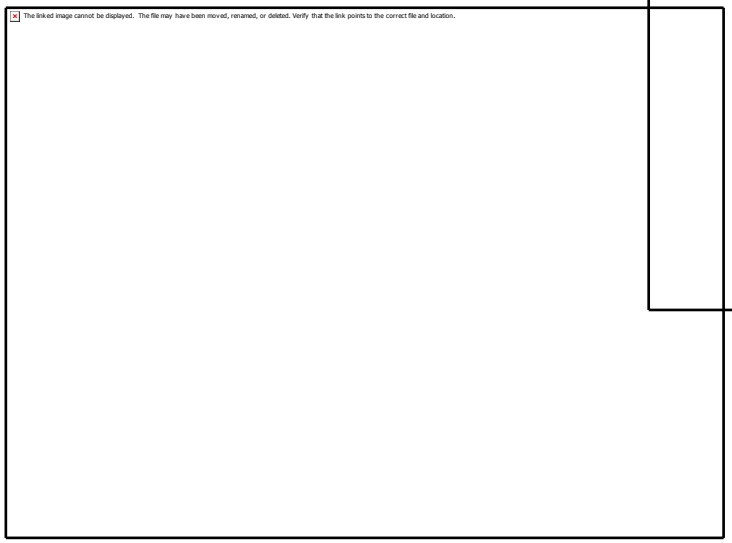
John Tukey, EDA, p157

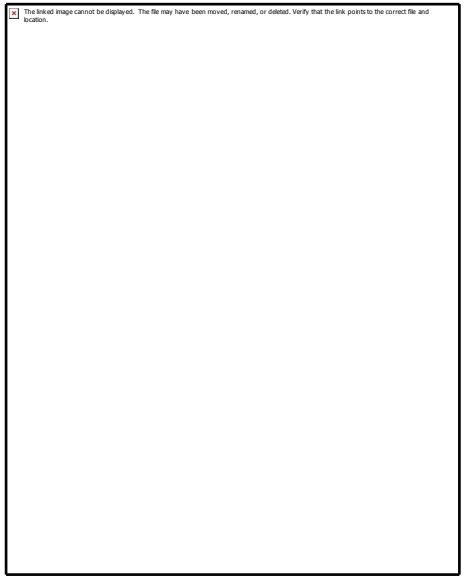
## ...Exploring Data...



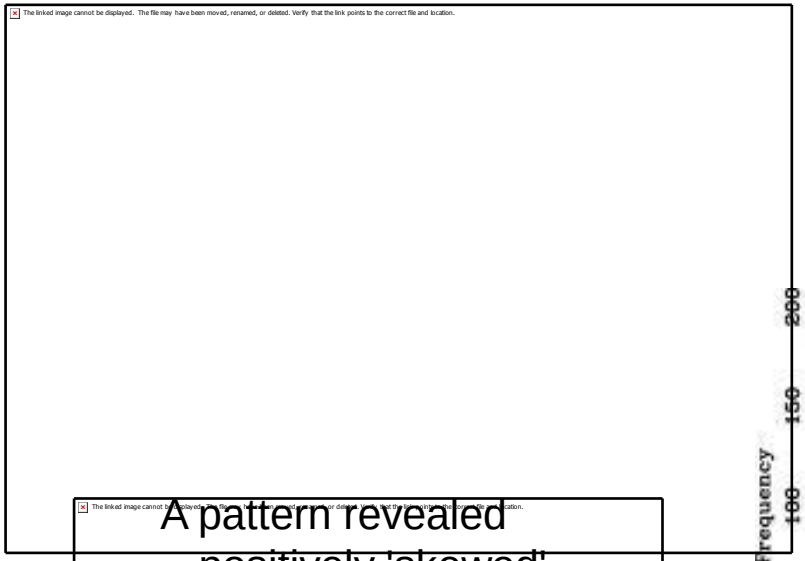


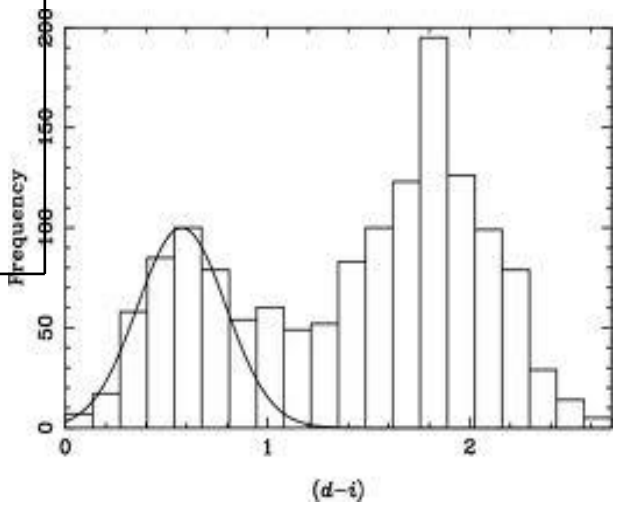


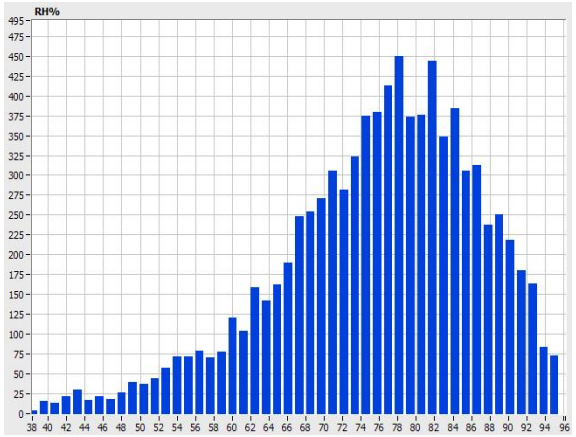


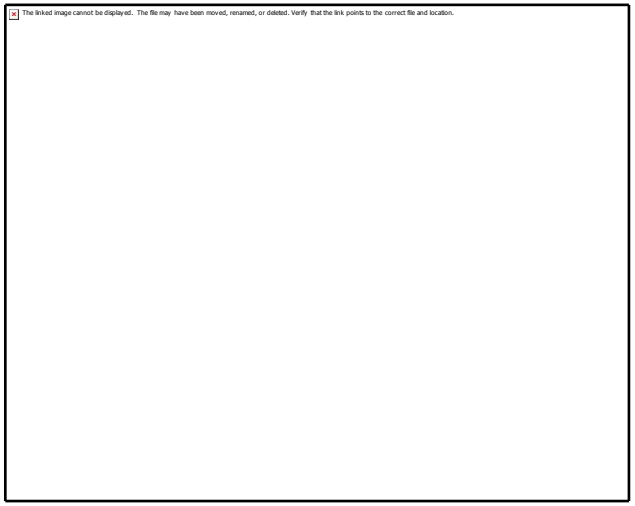


## ...Exploring Data...





A pattern revealed
- positively 'skewed'

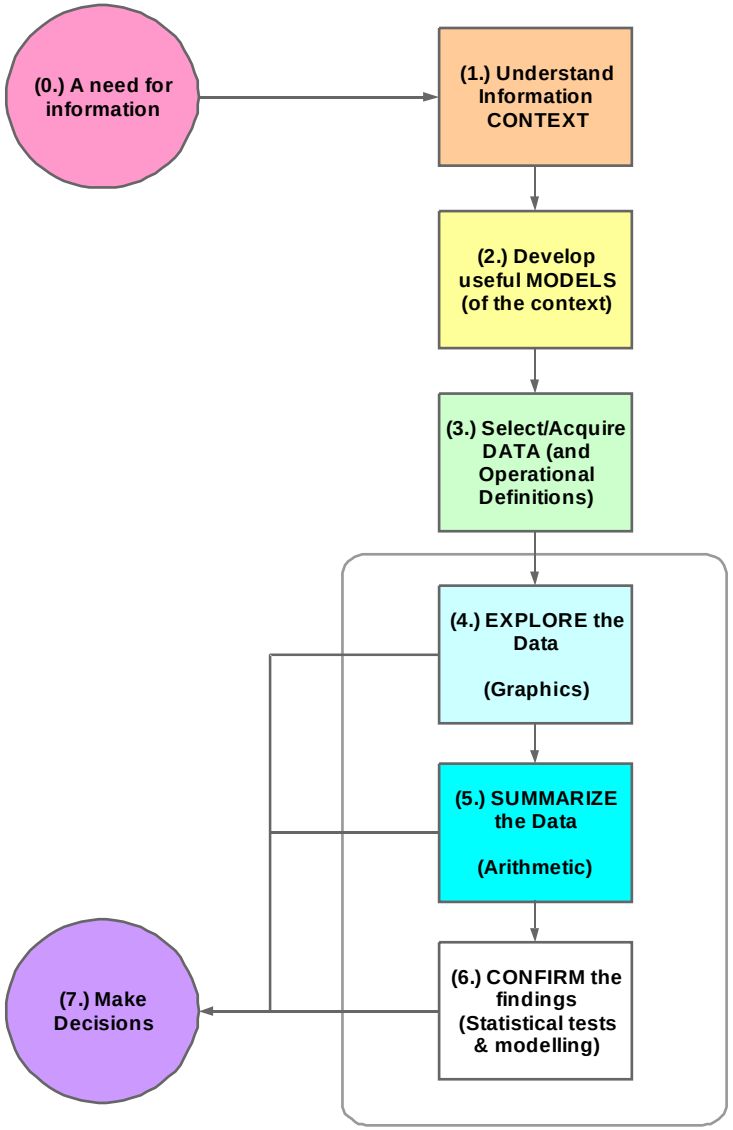
# *Step 5: Summarize your data*

**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...*

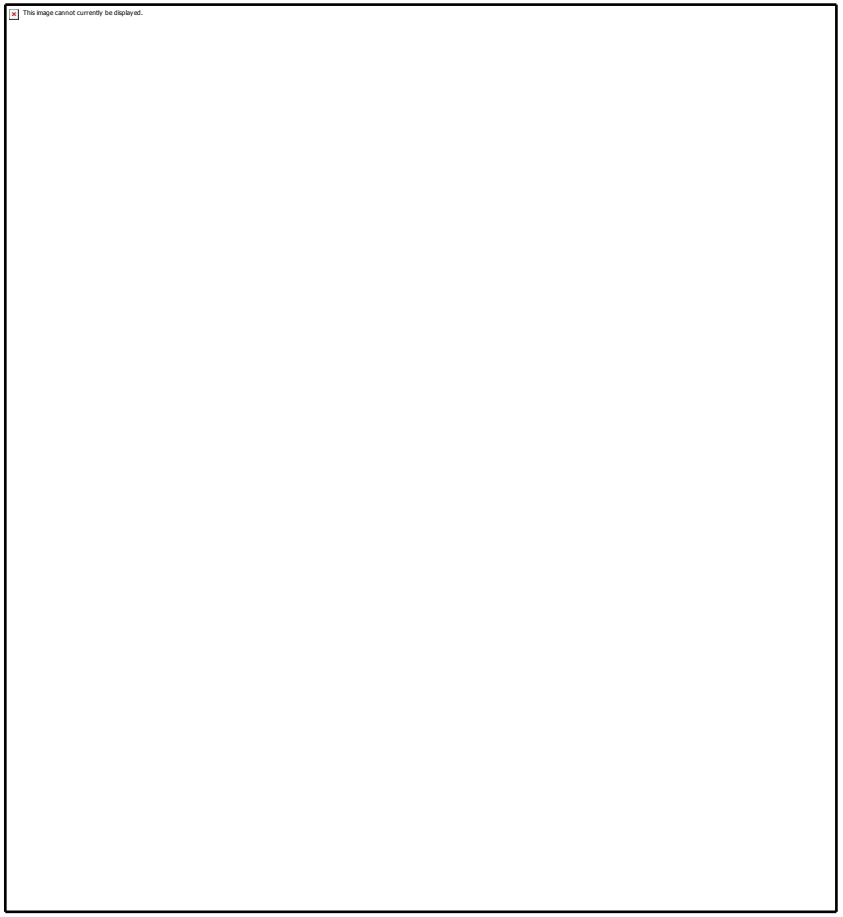
$$y = y_0 e^{\frac{x^2}{2\,sd^2}}$$

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

$$sd = \sqrt{\frac{x_i \quad x^{\,2}}{n \quad 1}}$$

$$var_i = \frac{x_i \quad x^{\,2}}{n \quad 1}$$

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

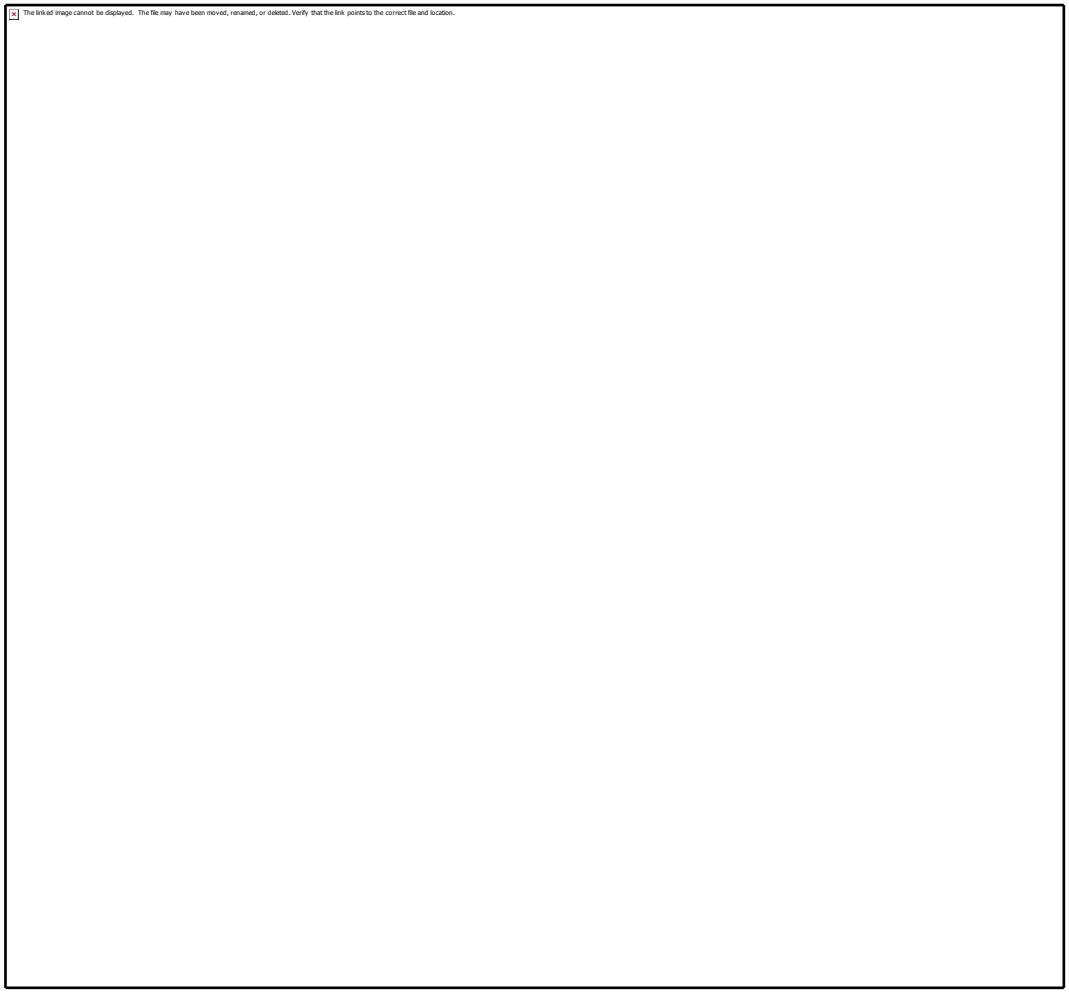


# *Summarize...some arithmetic or some counting...*

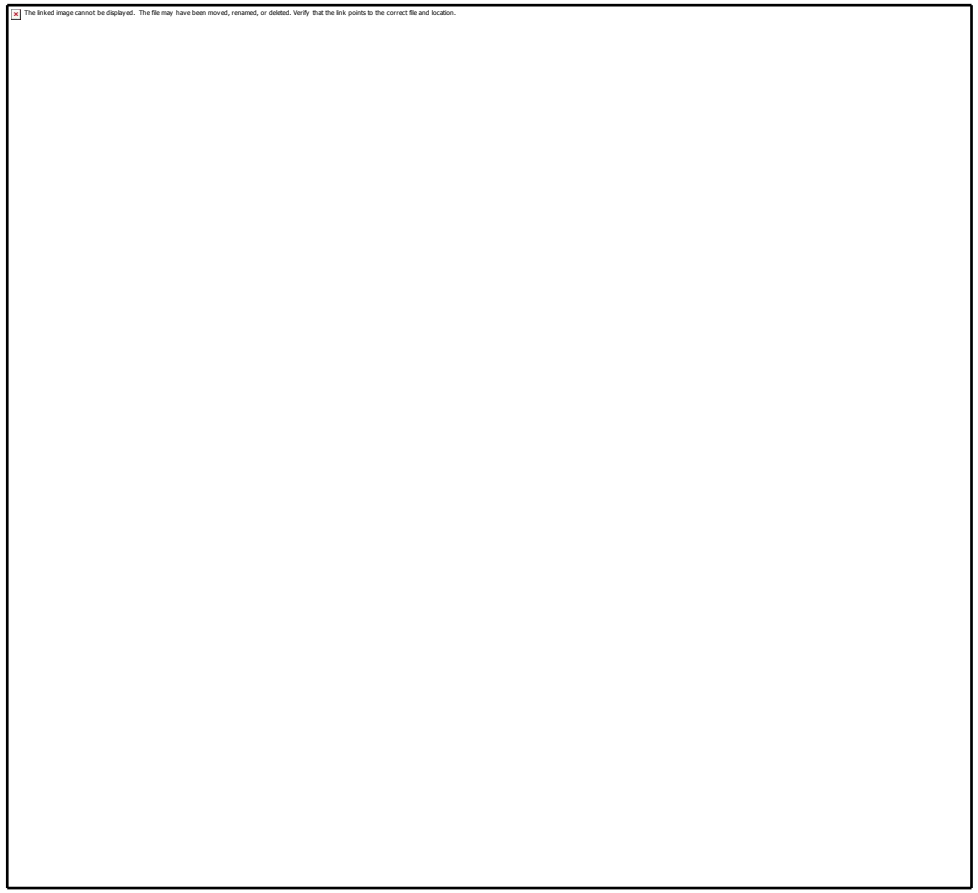
# ...Summarize…

$$slope = \frac{x_i \quad x \times y_i \quad y}{x_i \quad x^{2}}$$

$$r = \frac{x_i \quad x \times y_i \quad y}{x_i \quad x^{2} \times \quad y_i \quad y^{2}}$$

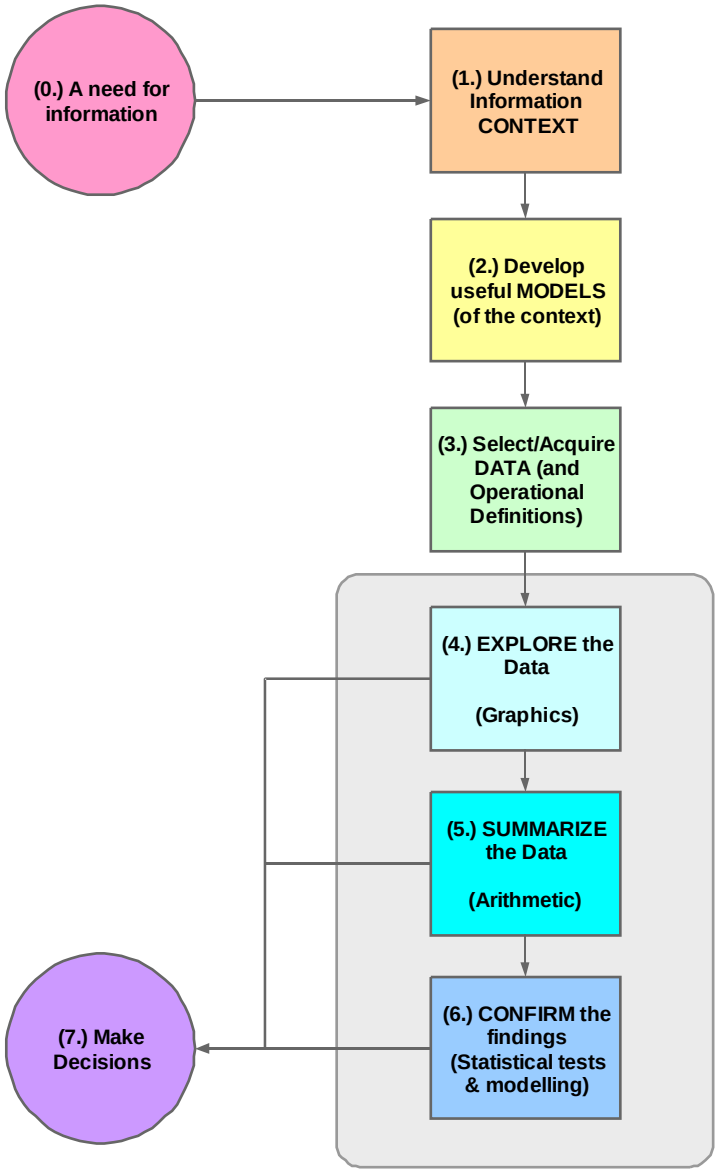


# ...Summarize…



# *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)

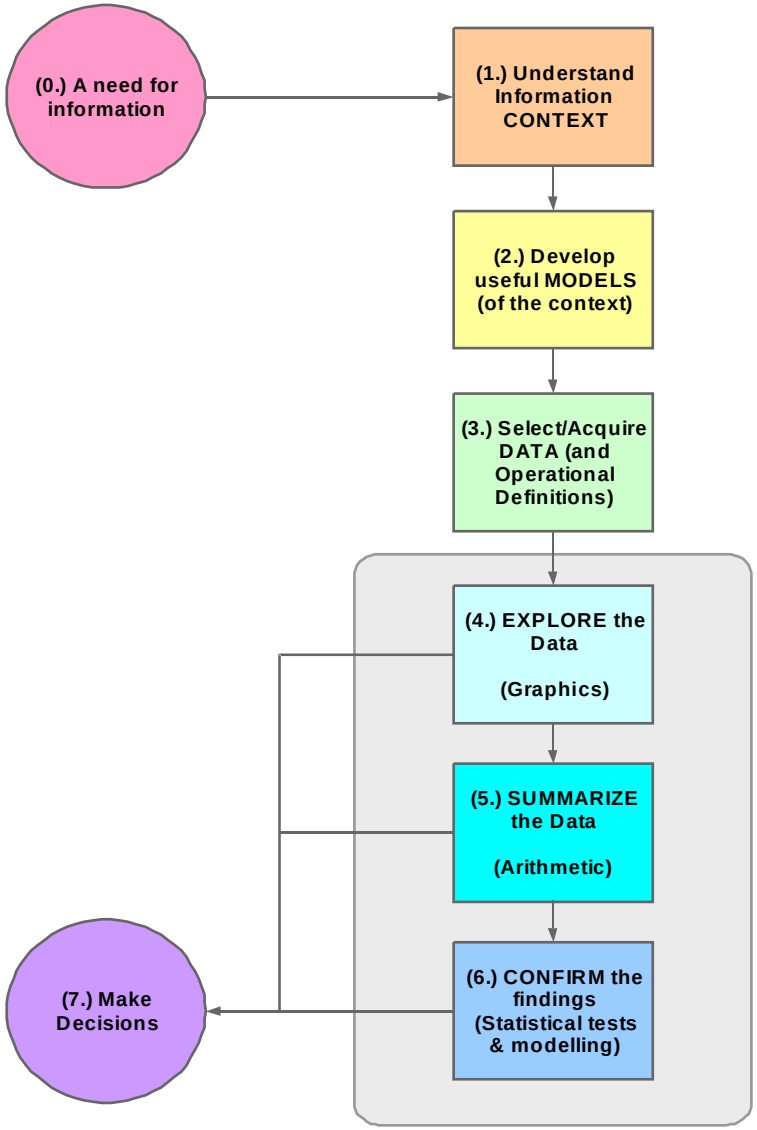
# *EAF*

**Recap: the three steps...**

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

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

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

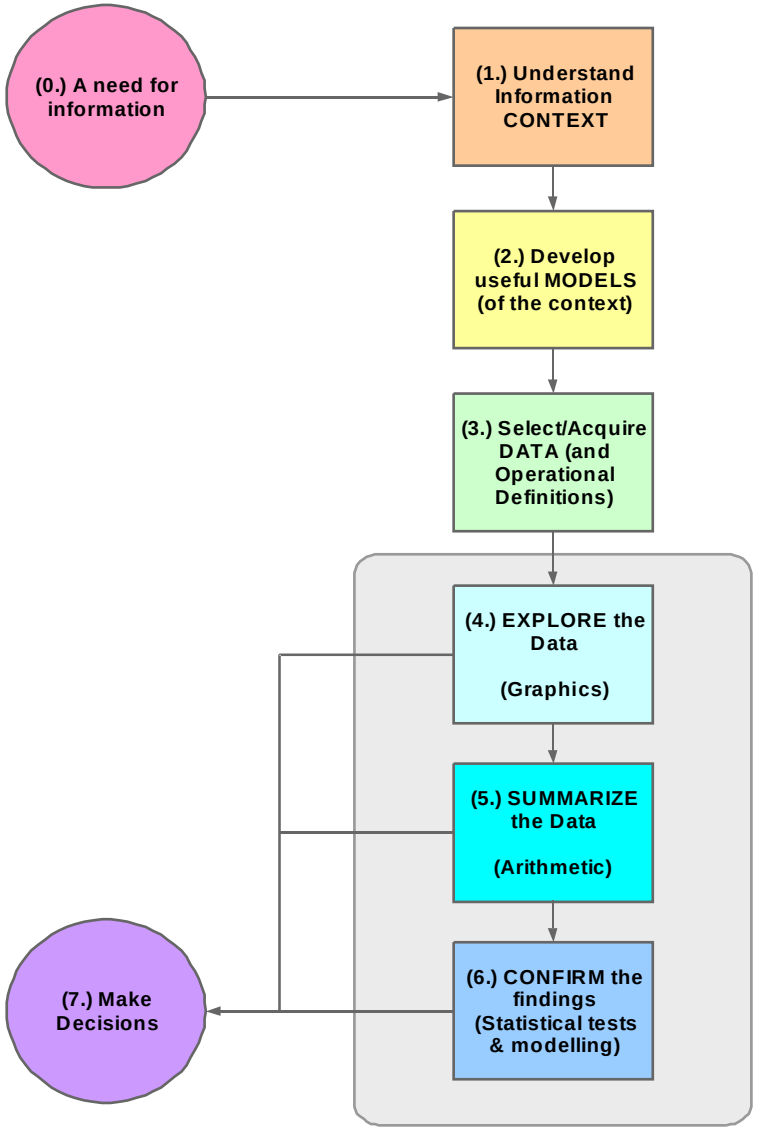


## *This framework can be extended...*

- 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 ?**



O X F O R D
S O F T W A R E  E N G I N E E R I N G
L I M I T E D

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

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