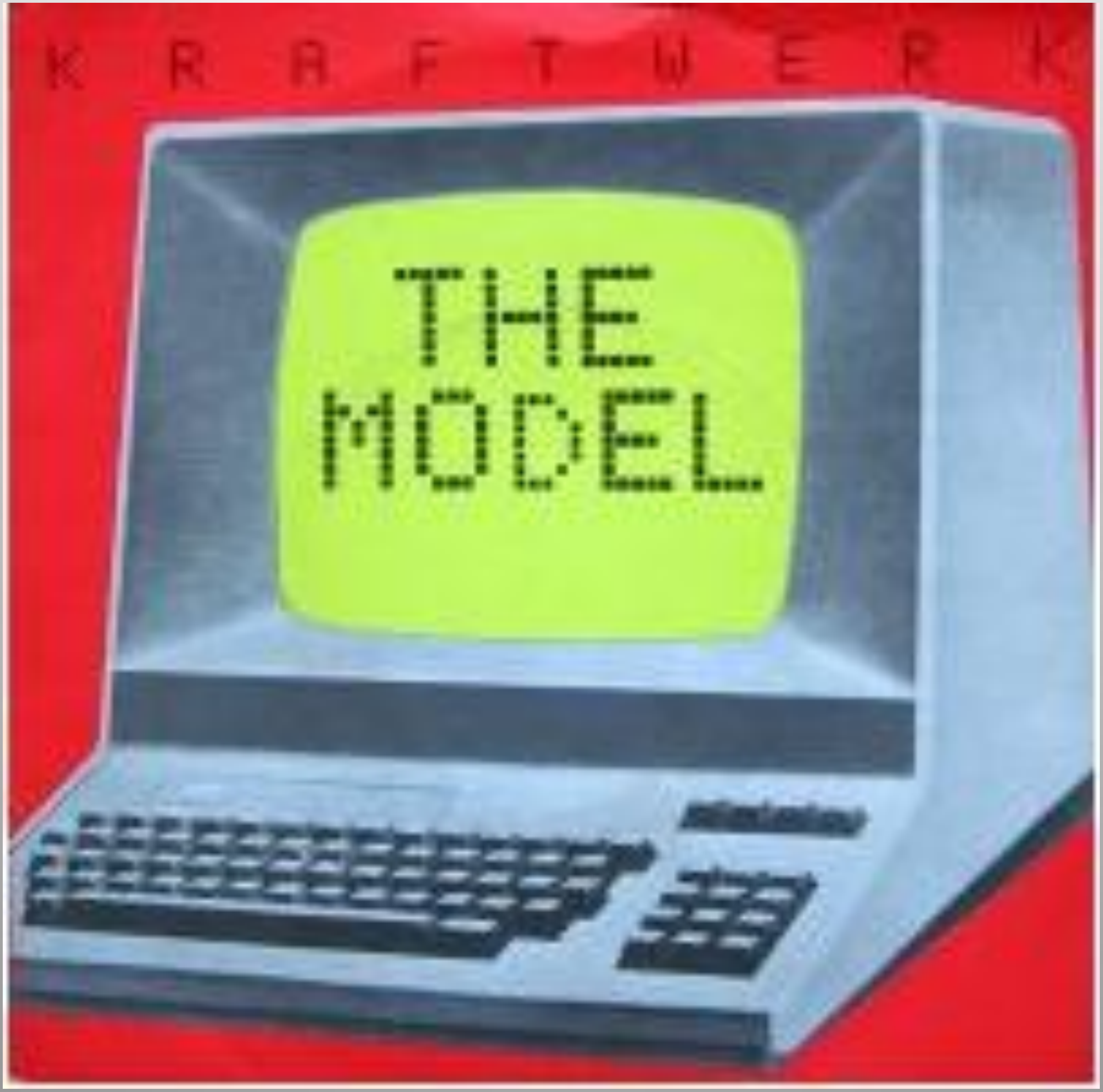
What’s a model

Frances Buontempo

ELECTROLA
KLING KLANG
KRAFTWERK
DAS MODEL
Ralf
Karl
Wolfgang
Florian


KRAFTWERK
THE MODEL

Represents data

- A lookup table
- A clustering algorithm
- An equation
- A decision tree
- An algorithm

$$f(x_1,\ldots,x_n) \Rightarrow y$$

Represents data

- sizeof(Lookup table) == sizeof(training data)
- sizeof(equation, etc) <= sizeof(training data)
  - Shanon’s entropy
  - simple
  - compact
- e.g. F = ma

Represents data

- How well does it fit the data
  - Correlation
  - Cross-validation
  - What if we randomly re-arrange the inputs?
  - What if we randomly generate the inputs?

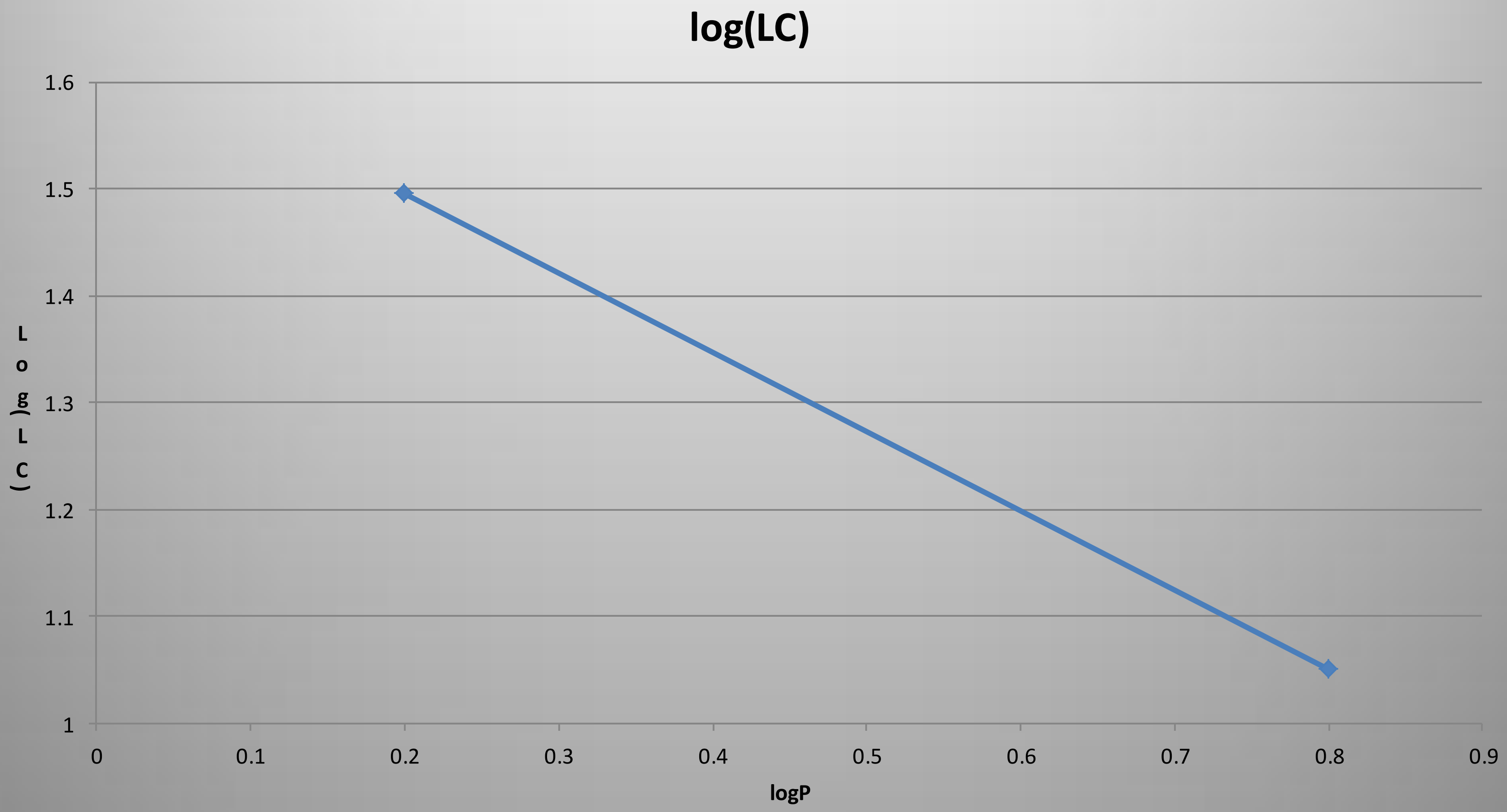
Represents data

- Equation example – fit a line through independent versus dependent data

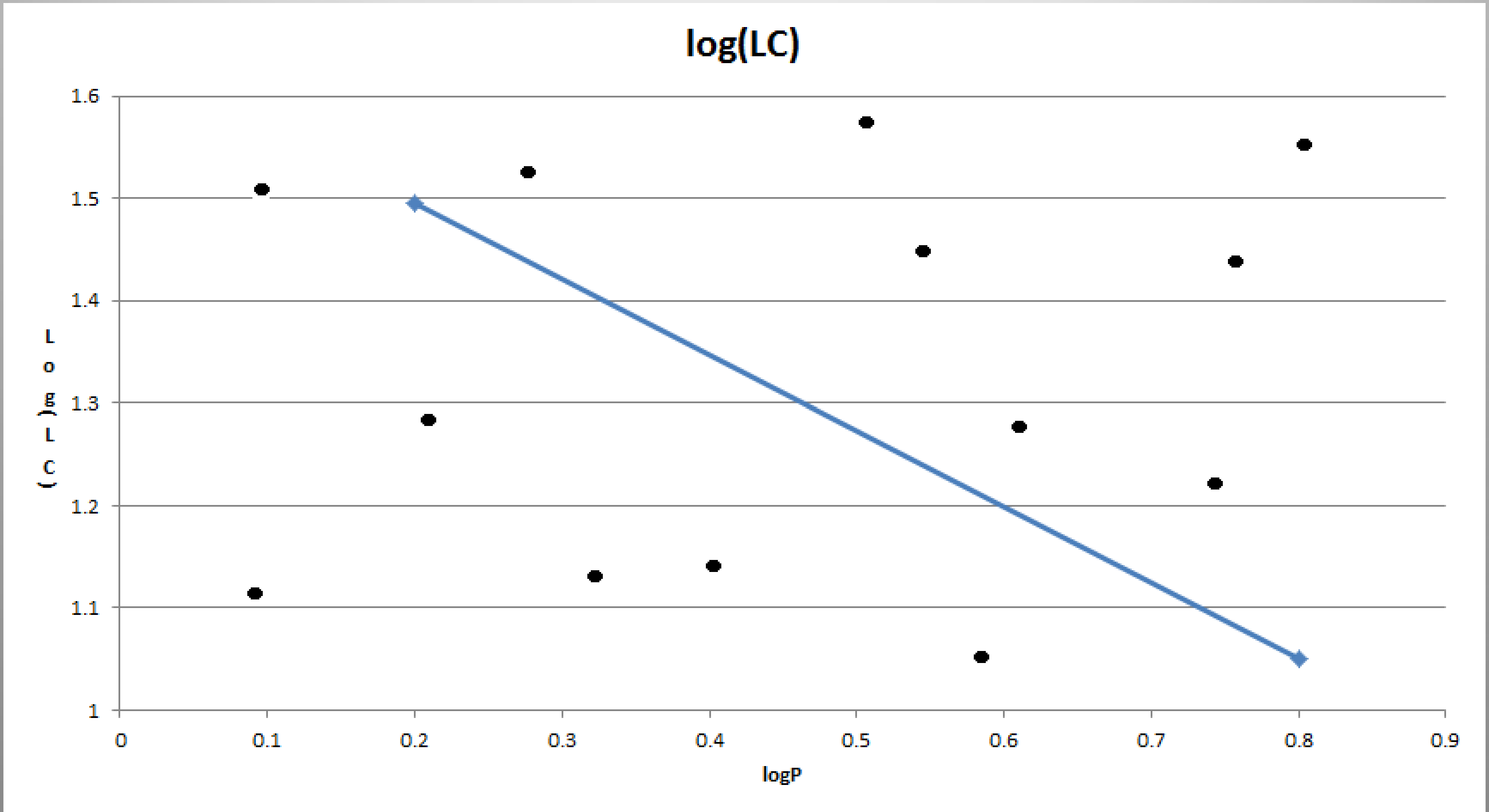
EPA toxicity QSAR “ECOSAR” programme

- http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/information-sources/qsar-document-area/Final_report_BRE_partB.pdf page 12
- **N=2, r^2 = 1.0.**
- Anilines, amino (*meta*- or 1,3-substituted):
  - log(LC) = 0.978 - 0.740×log P 50, N = 2, r^2 = 1.0
- Anilines, amino (*ortho*- or 1,2-substituted):
  - log(LC) = -0.547 - 0.522×log P 50, N = 2, r^2 = 1.0.
- Anilines, amino (*para*- or 1,4-substituted) :
  - log(LC) = -3.337 - 0.123×log P 50, N = 2, r^2 = 1.0.
- Anilines, dinitro: log(LC) = -0.027 - 0.596×log P 50, N = 2, r^2 = 1.0.
- Benzenes, dinitro: log(LC) = -1.867 - 0.333×log P 50, N = 2, r^2 = 1.0.

Represents data



Represents data?



Represents data

Both the training data and any new unseen data

$\Rightarrow$predicting the future

Independent data => dependent data/answer

$f(x_1,\ldots,x_n)$ => y

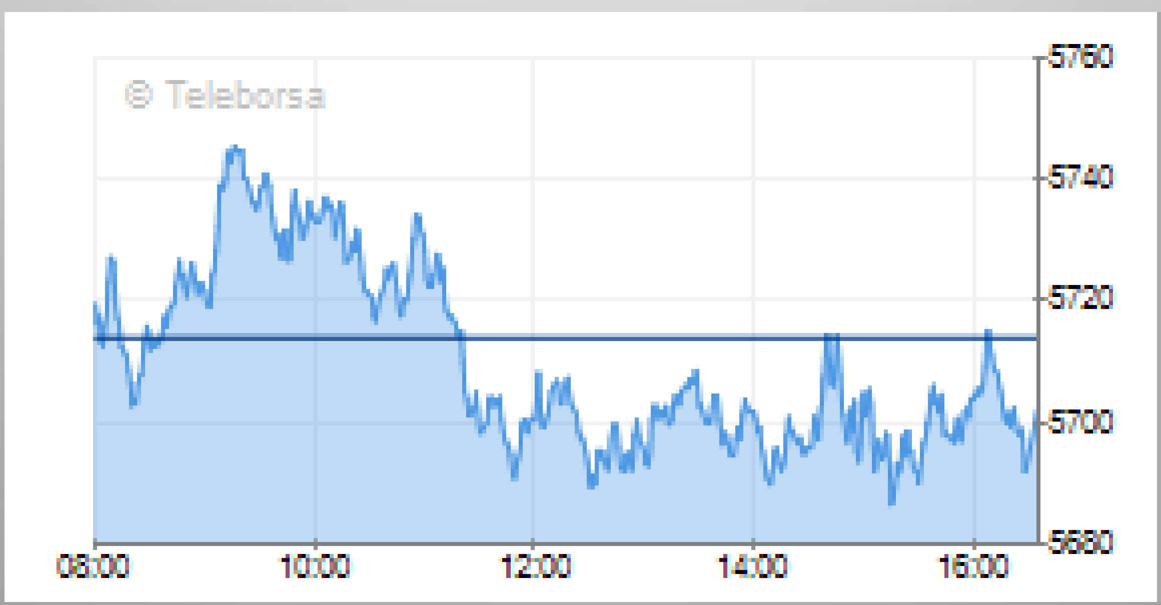
Represents data

- Calibration
- http://www.scientificamerican.com/article.cfm?id=finance-why-economic-models-are-always-wrong

Represents data

- FTSE

(http://www.londonstockexchange.com/home/homepage.htm)



Represents data

$$f(x_1,\ldots,x_n) = \text{sum}(x_1,\ldots,x_n)/\text{fudge\_constant}$$

- Do you reset the “constant” every day or week?
- Is a “model” of this form any good?
- What about F=ma, where *a* is always about 10?

What is a model?

- It represents data
- It (probably) does so compactly and simply
- It gets previously unseen data correct
- It doesn’t need recalibrating