Data Analysis in Software Engineering

Dr. Javier Dolado
Department of Computer Languages and Systems
University of the Basque Country, Spain

TOK session: ICEBERG Project at Alcalá de Henares

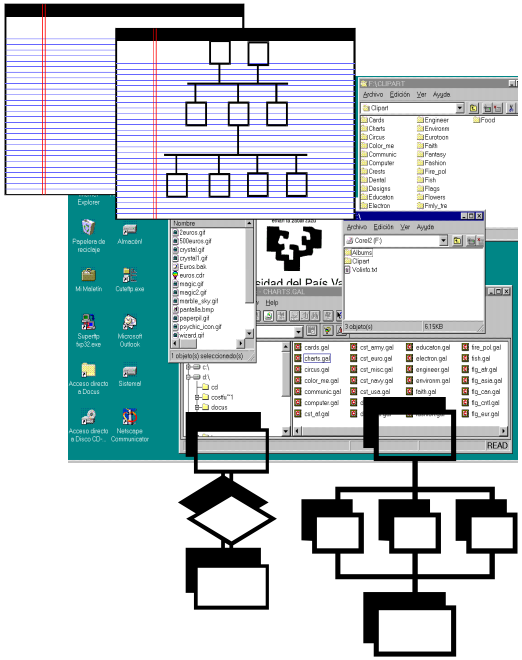
9th February 2016

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°324356

Outline

- **Software Engineering, Data Analysis and Data Mining**
  - Sofware Cost Estimation, Software Size Estimation
  - Process measurement and estimation
- **Methods**
  - **Linear Regression**
  - **Genetic Programming**
  - Curve estimation
  - Clustering, Principal Component Analysis, System Dynamics, etc, etc.
  - Experimentation and Hypothesis Tests
  - Bayesian Networks
- **Tools: R, Weka,** SciPy, etc.
- **Data Sources: Promise database, other public datasets**
- Results and Discussion

Basic Problem: Prediction



Parameters, data collected, previous projects, etc.

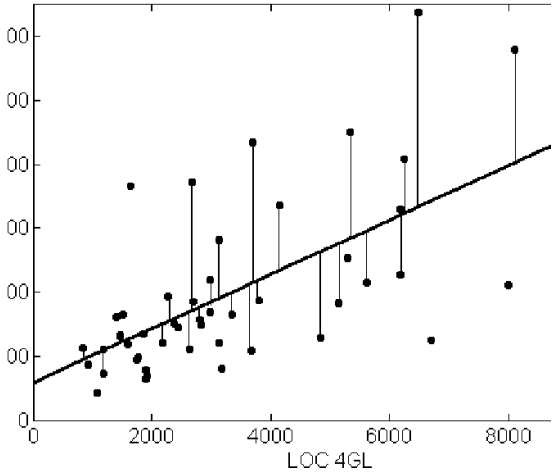
- The estimation of cost, size, defects, quality, etc has always been a problem

Strategy: Build models from data

DATA

Important: data sources must be relevant and reliable

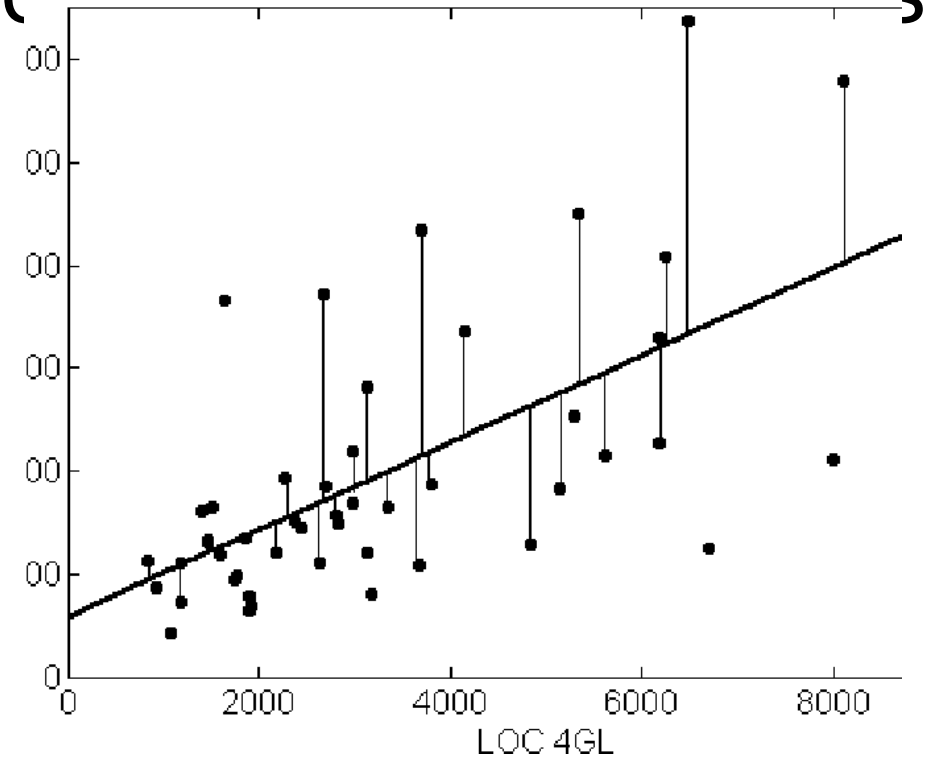
- Different methods are applied depending of the type of problem

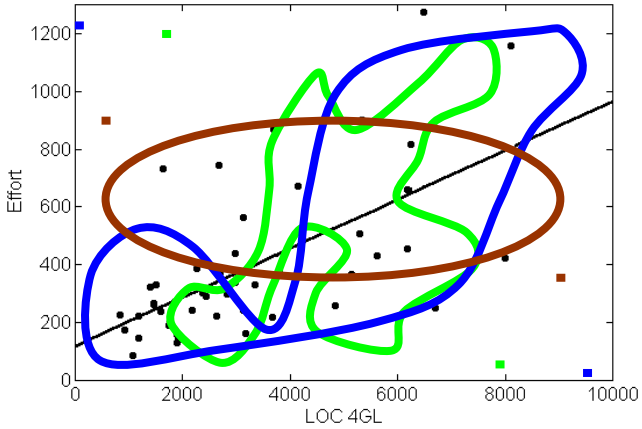


OR
''GUESSTIMATING''

Methods: Linear Regression and Curve Estimation

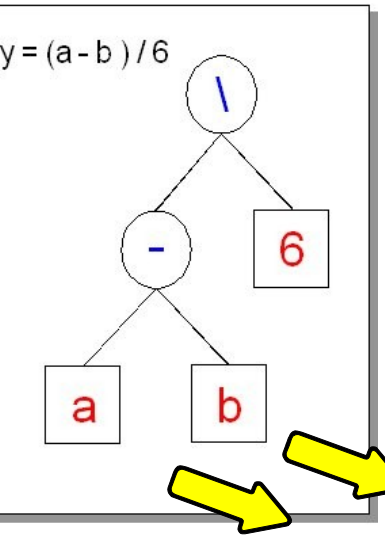
- Probably, the most used method for estimation.
- It is simple and it obtains results as good as other mo[illegible]s

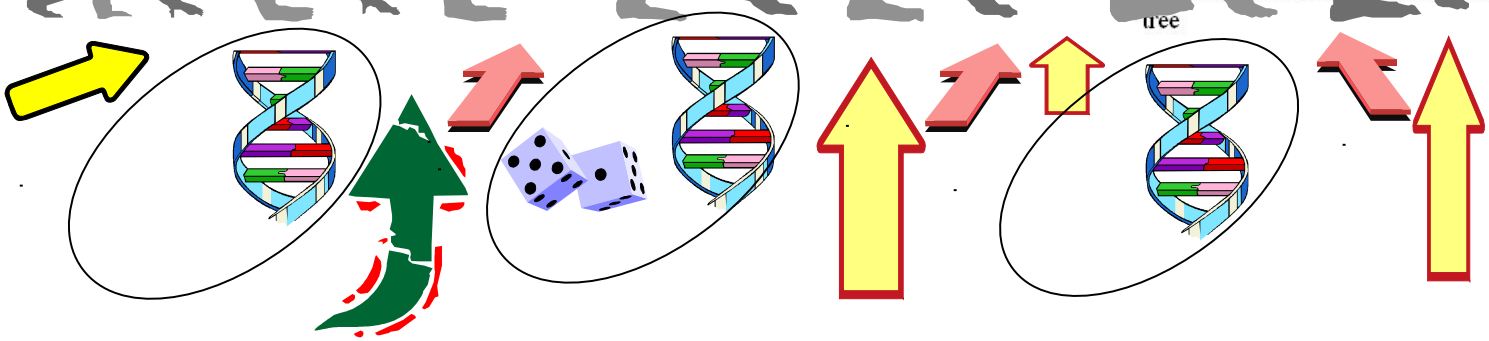


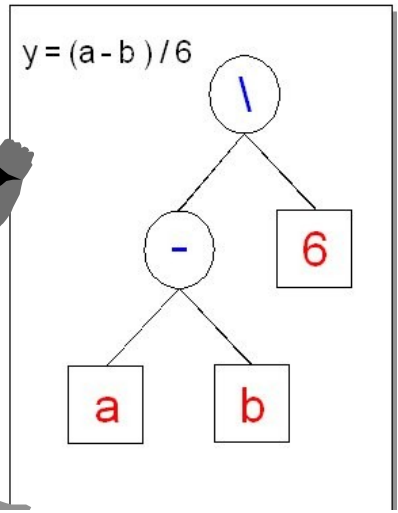


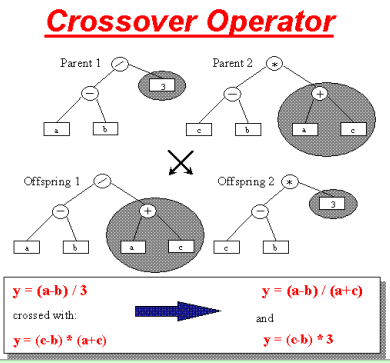
Genetic Programming

- Tries to mimic one of the methods of evolution



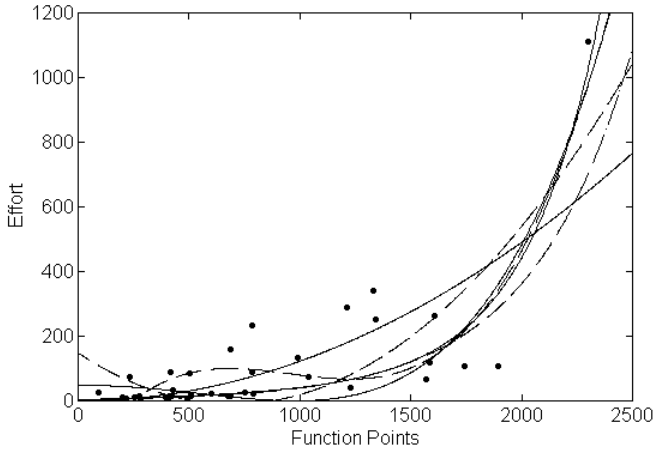


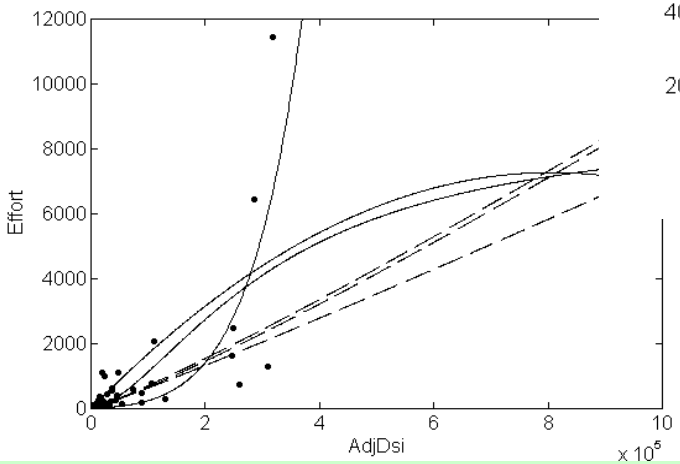


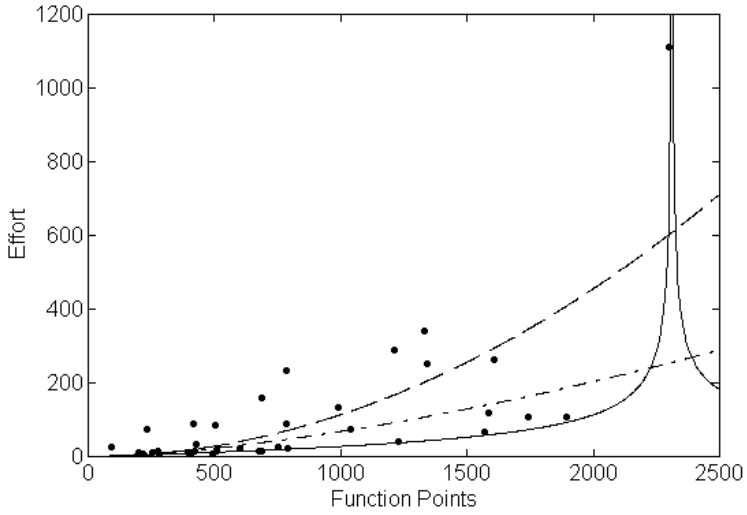


Genetic Programming

- Genetic programming allows us to adjust almost any equation. GP gives always good results, with the proper adjustment of parameters.
- We can always find a “good model“

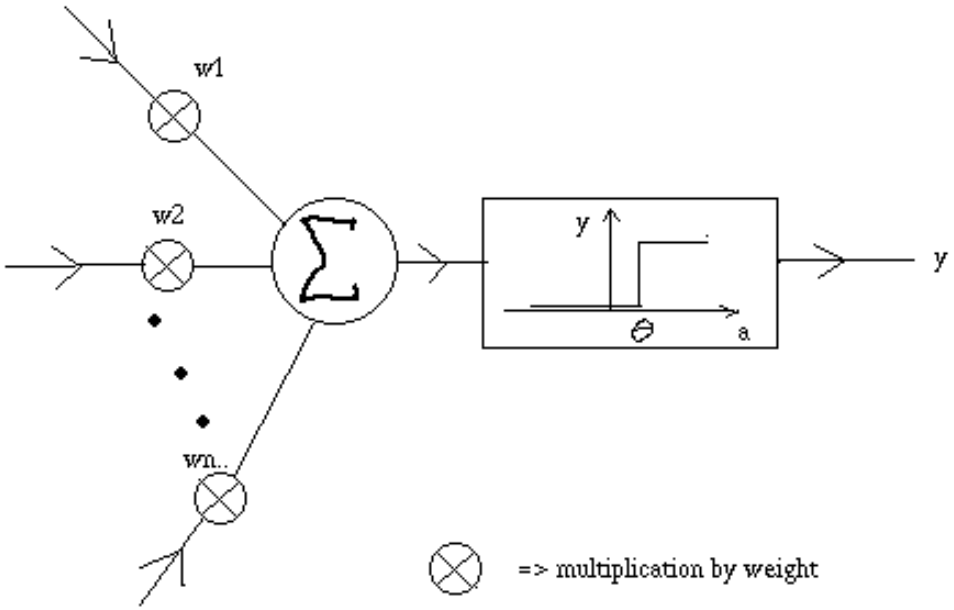


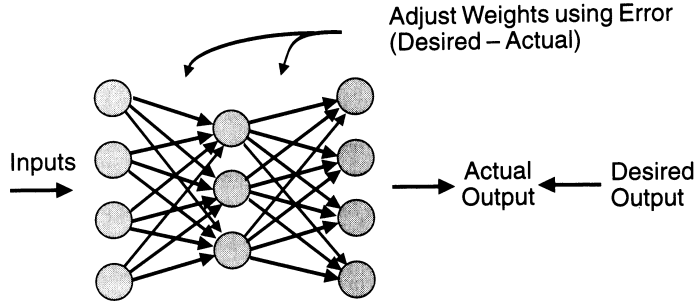




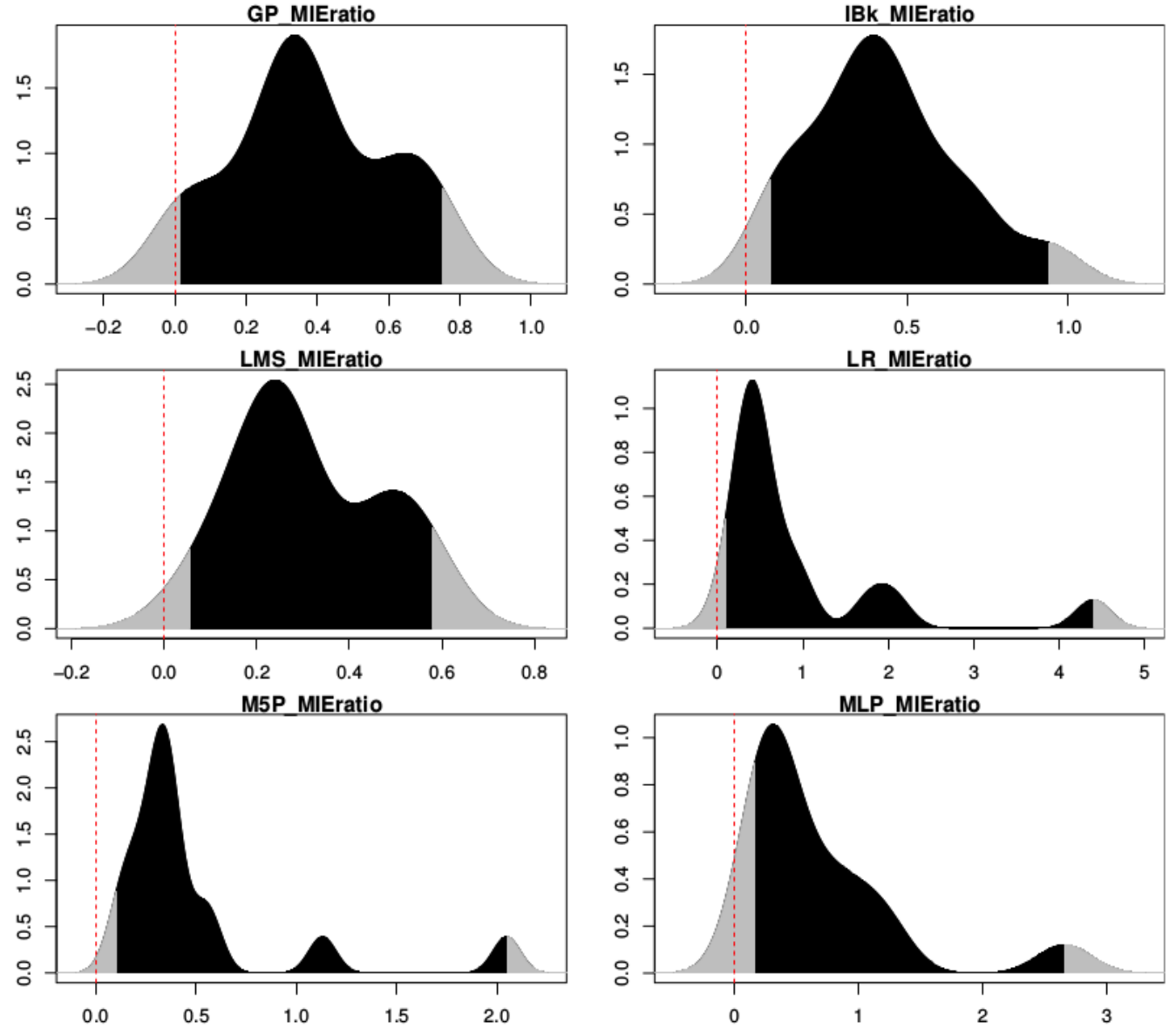
Neural Networks, Clustering, Instance-based methods, etc.

- All methods are based on a specific paradigm and purpose, therefore their application must be carefully examined
- Neural networks provide “moderate good





In software cost estimation there are two methods that perform reasonably well ...



	Qtle. 2.5%-97.5%	HPD low-upper
GP	0.021-0.725	0.015-0.751
IBk	0.096-0.859	0.073-0.943
LMS	0.088-0.566	0.056-0.581
LR	0.162-3.582	0.103-4.397
M5P	0.124-1.727	0.102-2.048
MLP	0.171-2.161	0.168-2.662
RTree	0.169-6.56	0.096-6.841

3: This table shows different probabilistic in
ds ($\alpha = 0.05$) for the data of the MIEratios
ter.

Don't underestimate the value of simple methods...

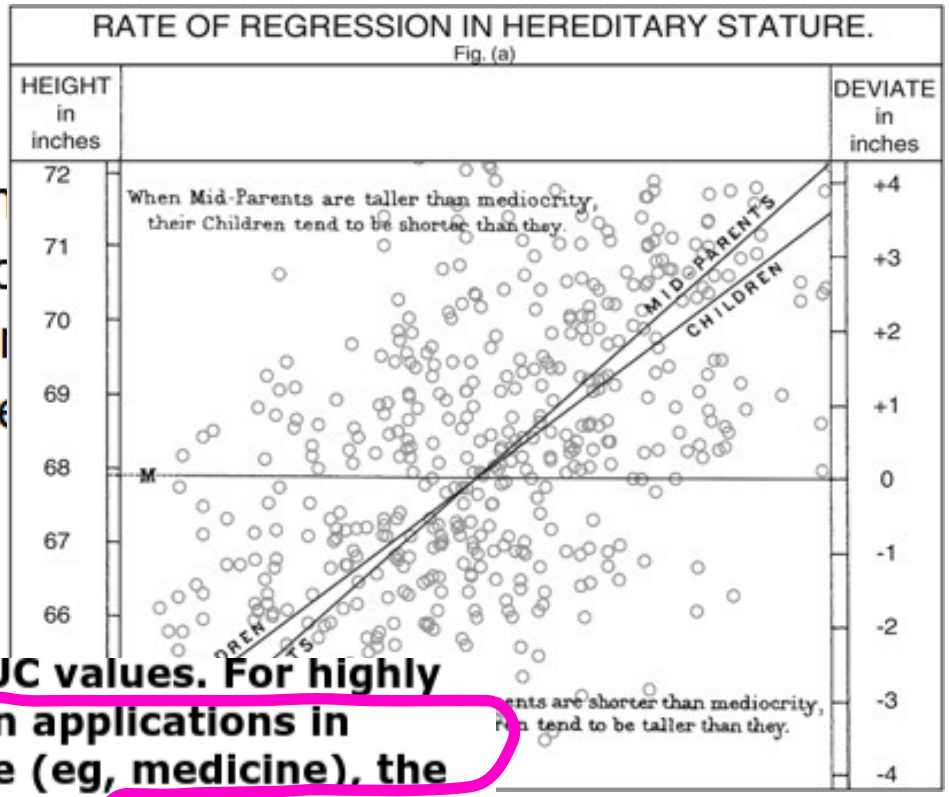
Article

European Journal of Human Genetics (2009) **17,** 1070–1075;
doi:10.1038/ejhg.2009.5; published online 18 February 2009

Predicting human height by Victorian and genomic methods

Yurii S Aulchenko[1,2,7], Maksim V Struchalin
M Belonogova[2,4], Tatiana I Axenovich[2], Mic
Albert Hofman[1], Andre G Uitterlinden[6], Mar
Ben A Oostra[1], Cornelia M van Duijn[1], A Ce
W Janssens[1] and Pavel M Borodin[2,4]

Sir Francis Galton, 1886



genomic profile should explain to reach certain AUC values. For highly heritable traits such as height, we conclude that in applications in which parental phenotypic information is available (eg, medicine), the Victorian Galton's method will long stay unsurpassed, in terms of both discriminative accuracy and costs. For less heritable traits, and in situations in which parental information is not available (eg, forensics), genomic methods may provide an alternative, given that

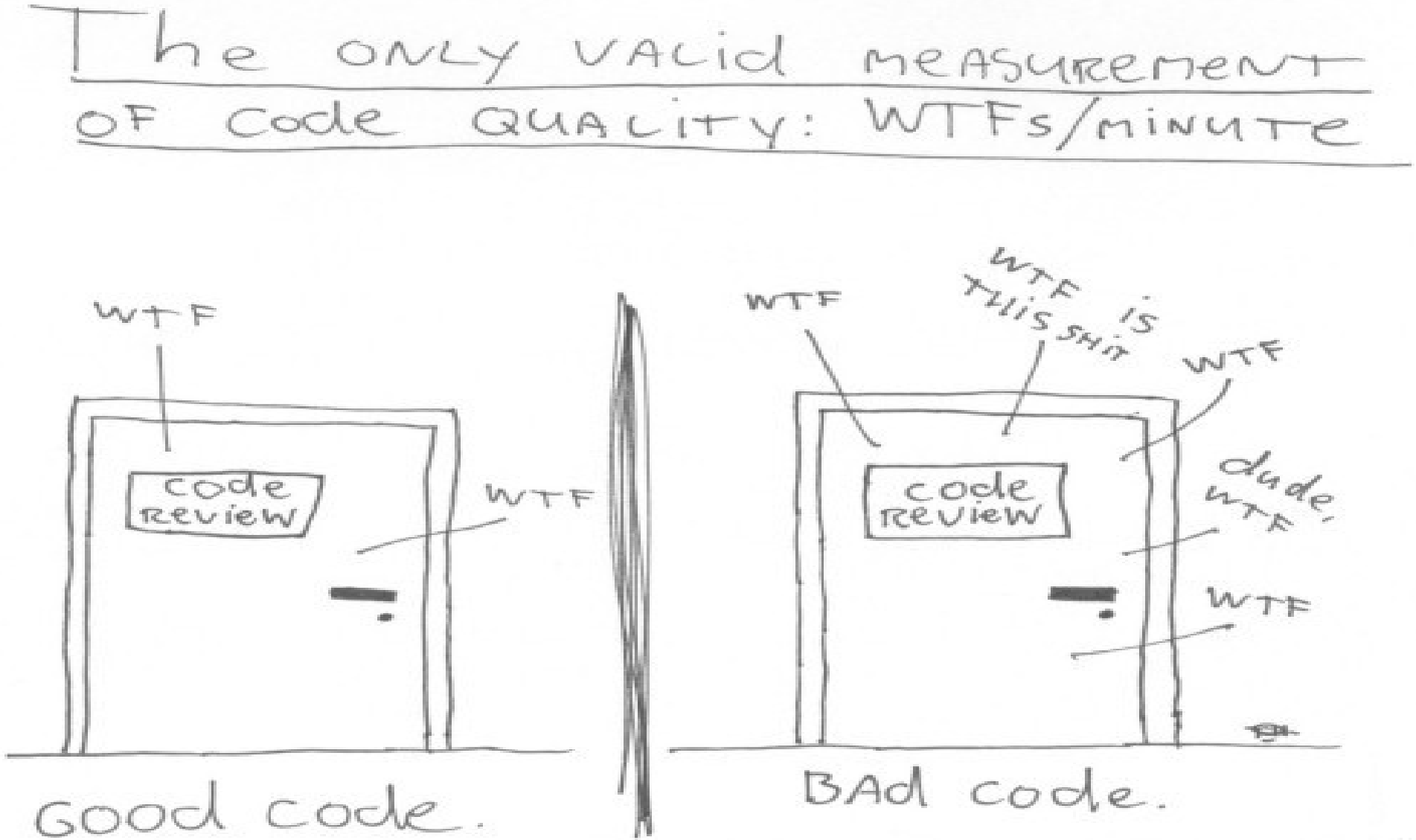
Results

- We apply many statistical methods to different estimation problems in software engineering including, cost, time, defects and others.
- We have applied Equivalence Hypothesis Testing to several software engineering experiments
- Bayesian Networks can be applied in the software testing area
- Main Problem: Data Sources
  - Public data is not always relevant to our specific domain
  - It is much better to collect the data within the organization
- There is no “best method“

Discussion

- Many methods available. The theory may be difficult to understand but they are really easy to apply
- Tools available: R, Weka, ScyPy ...
- Data from public sources cannot be applied to other settings in a straightforward way
- It is unavoidable to use 'within-company' data

- From all the set of methods there may be one that fits to your current problems and endeavours
- Although there are other opinions ...



Acknowledgements

PROJECTS

“Testing of data persistence and user perspective under new paradigms“

“Gamificación y prototipado de procesos para la detección temprana de oportunidades en la producción del software“

PRESI TIN2013-46928-C3-1-R, TIN2013-46928-C3-2-R

Ministerio de Economía y Competitividad