Some Methods and Tools for Software Data Analysis

Javier Dolado
U. País Vasco/Euskal Herriko Unibertsitatea

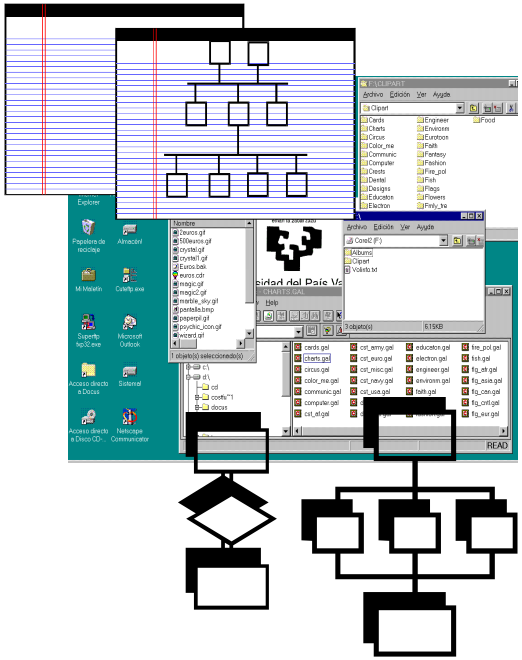
Daniel Rodríguez
Universidad de Alcalá

Javier Tuya
Universidad de Oviedo

Outline

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

Basic Problem: Prediction



Parameters, data collected, previous projects, etc.

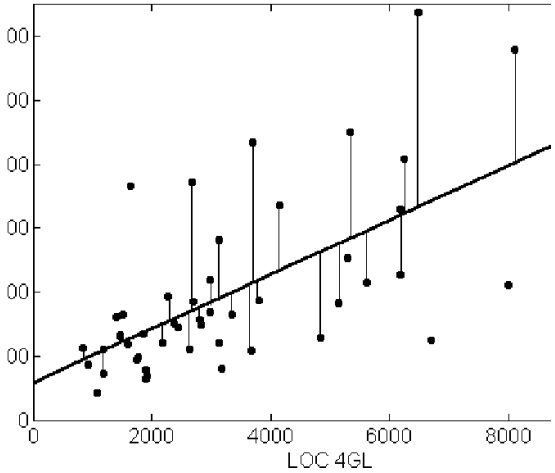
- The estimation of cost, size, defects, quality, etc has always been a problem

Strategy: Build models from data

DATA

Important: data sources must be relevant and reliable

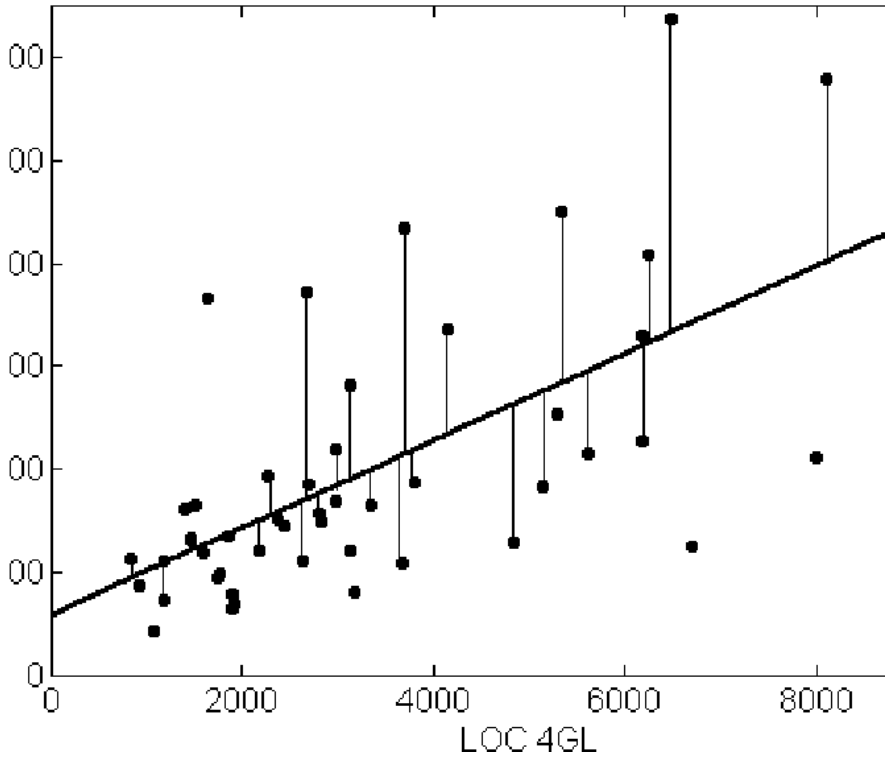
- Different methods are applied depending of the type of problem

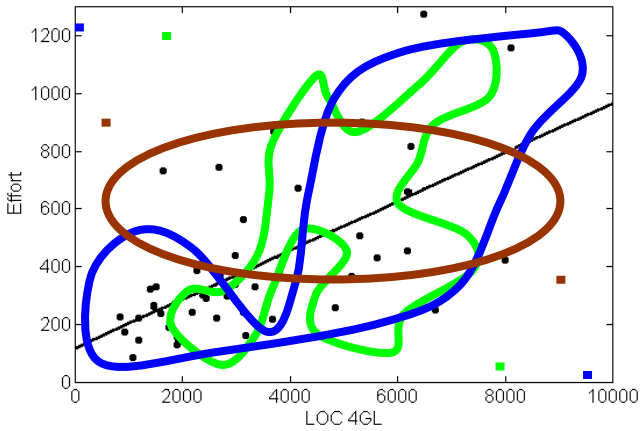


OR
"GUESSTIMATING"

Methods: Linear Regression and Curve Estimation

- Probably, the most used method for estimation.
- It is simple and it obtains results as good as other more complex methods





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

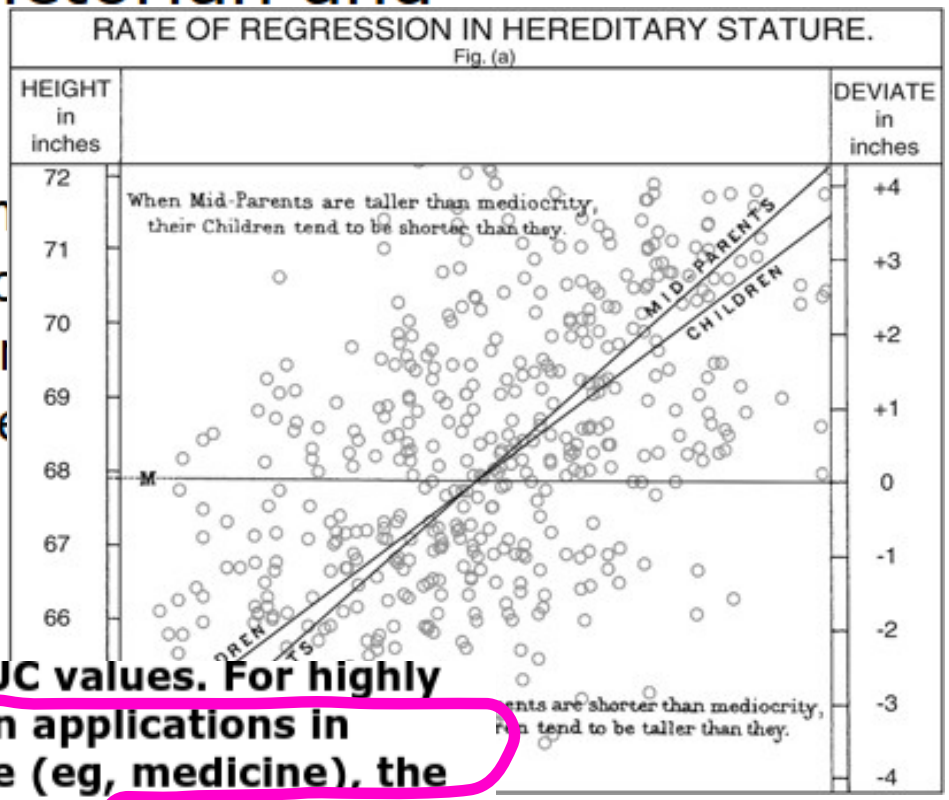
Article

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

Sir Francis Galton, 1886

Predicting human height by Victorian and genomic methods

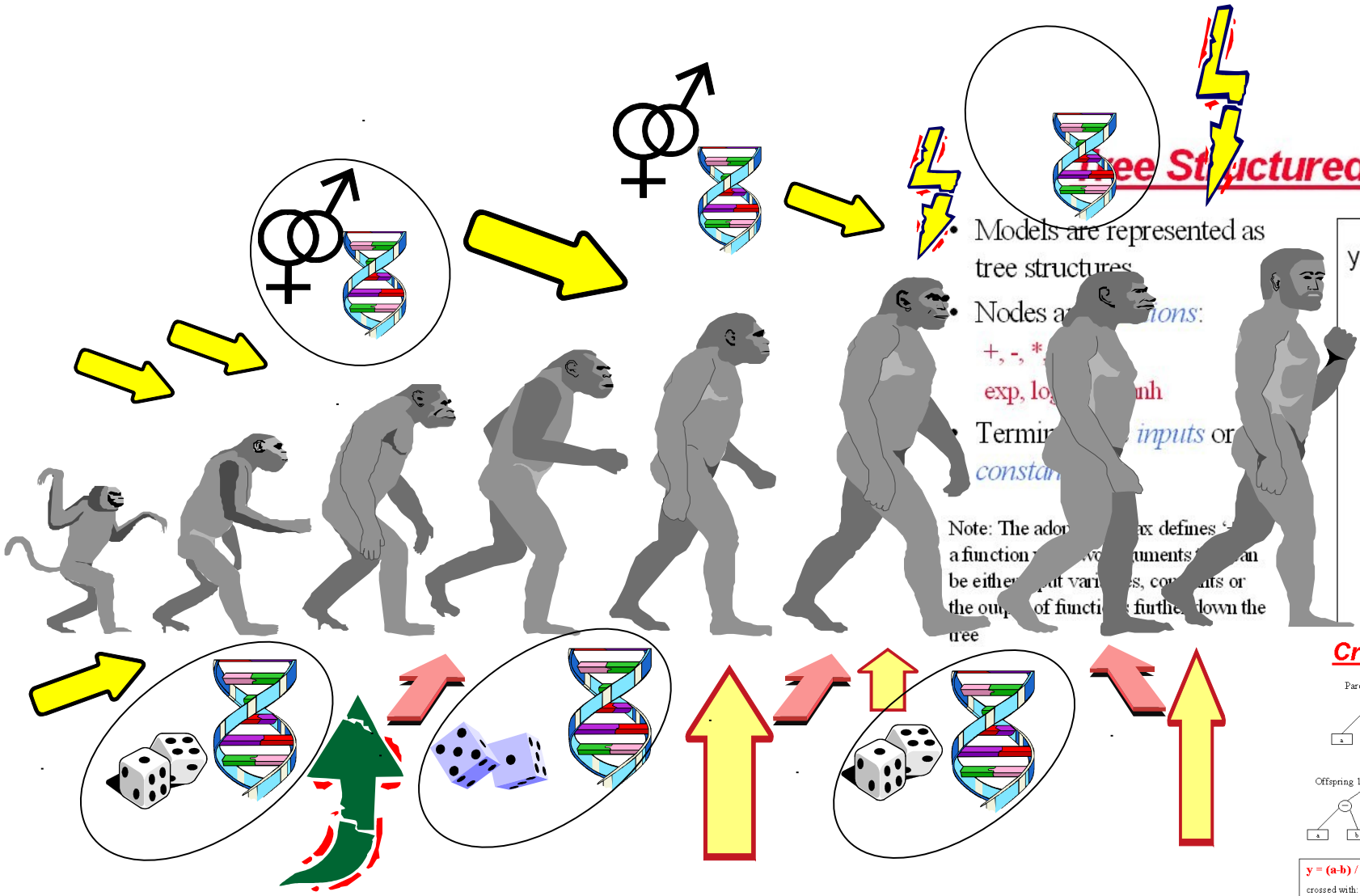
Yurii S Aulchenko[1,2,7], Maksim V Struchalin
M Belonogova[2,4], Tatiana I Axenovich[2], Mic
Albert Hofman[1], Andre G Uitterlinden[6], Mar
Ben A Oostra[1], Cornelia M van Duijn[1], A Ce
W Janssens[1] and Pavel M Borodin[2,4]



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

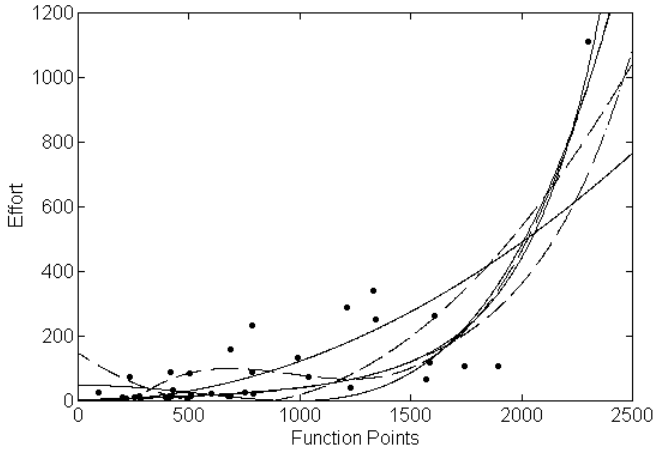
Genetic Programming

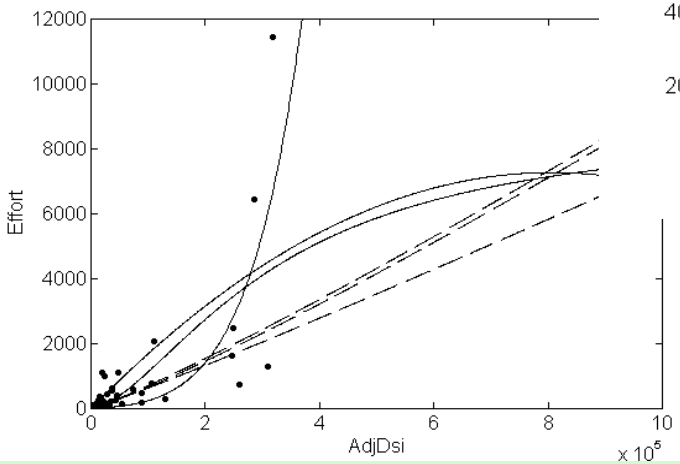
- Tries to mimic one of the methods of evolution

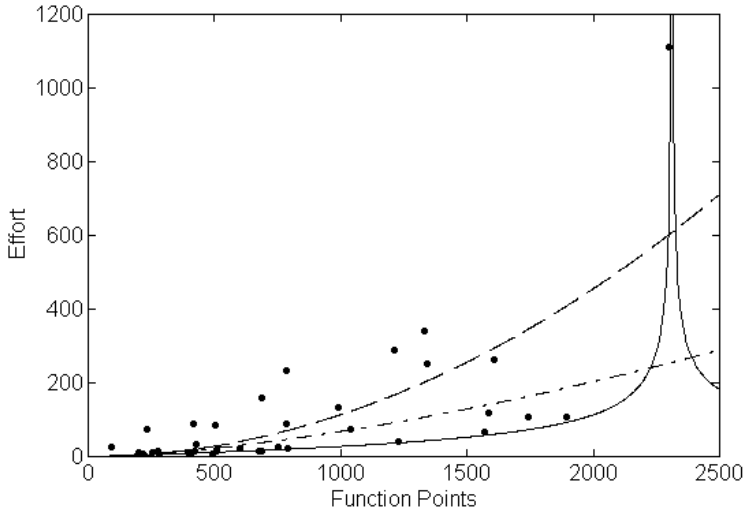


Genetic Programming

- 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“

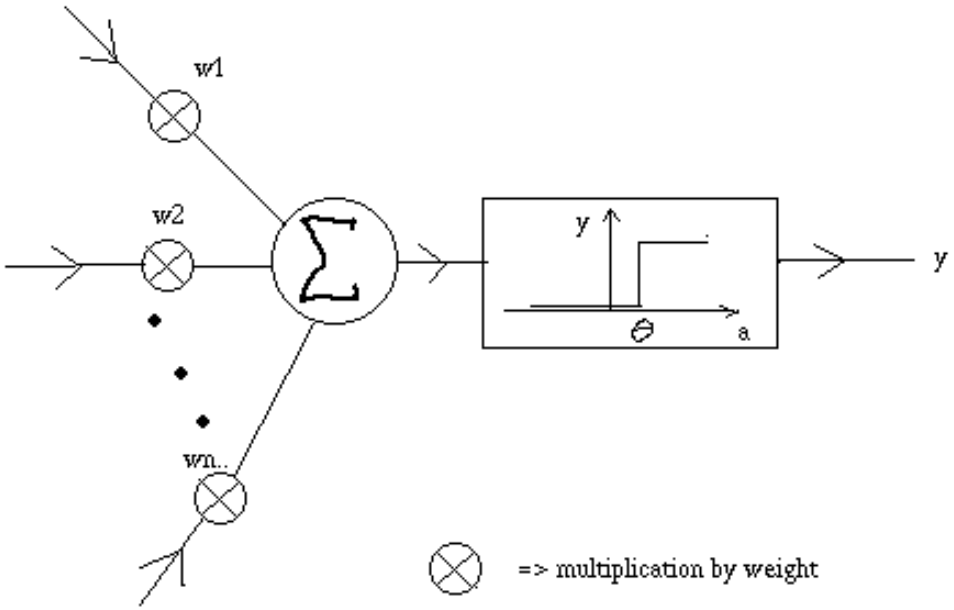


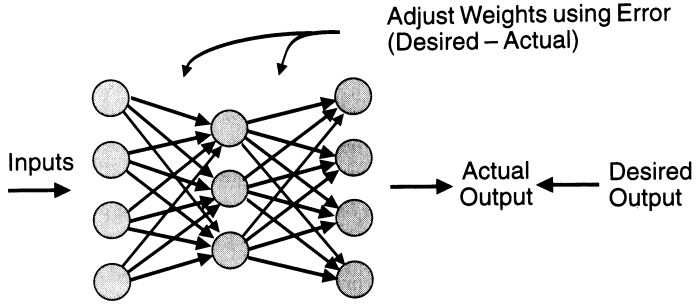




Neural Networks, Clustering, etc.

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





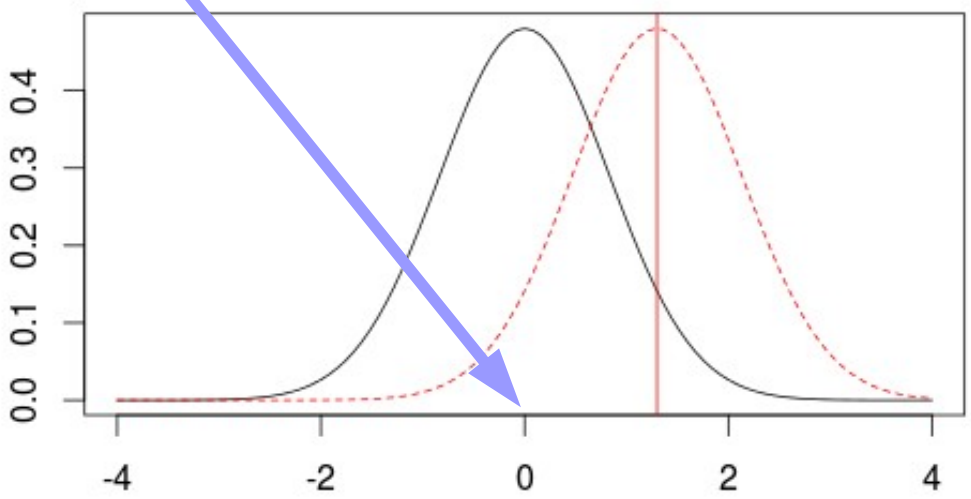
Experimentation and Hypothesis Testing

Equivalence Hypothesis Testing

- With this method we try to prove that two things are "equivalent"
- It is a method in the drug and food industry to show that two things can replace each other
- Equivalence is usually known as bioequivalence in pharmacological studies
- Examples:
  - Substitution of brand-name drugs by their generic counterparts
  - Identify food products with similar properties (sensorial or others)

NHST versus Equivalence Hypothesis

NHST: Null hypothesis significance test

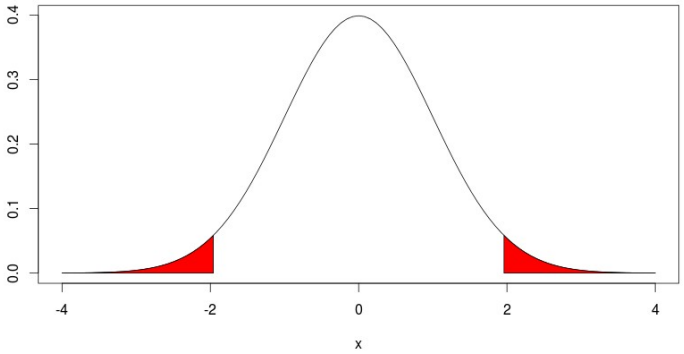


Compare two things

115.8	111.9
96	55.2
79	50.9
90.8	75.4
39.6	75.4
98.4	89.5
18.9	14.9
10.3	14.3
28.5	32.8
7	5.5
9	4.5
7.3	9.7
5	2.1
8.4	5.2
98.7	85.4
15.6	10.2
23.9	14.8
138.3	110.3

Null hypothesis that is a Nil hypothesis: the means are equal

$$H_0 : \mu_T - \mu_R = 0$$
$$H_a : \mu_T - \mu_R \neq 0 \quad \text{for the two-tailed test}$$



Equivalence Hypothesis Testing

$$H_0 : |\mu_T - \mu_R| \geq \Delta$$
$$H_a : |\mu_T - \mu_R| < \Delta$$

This is the range of values that consider "acceptable"

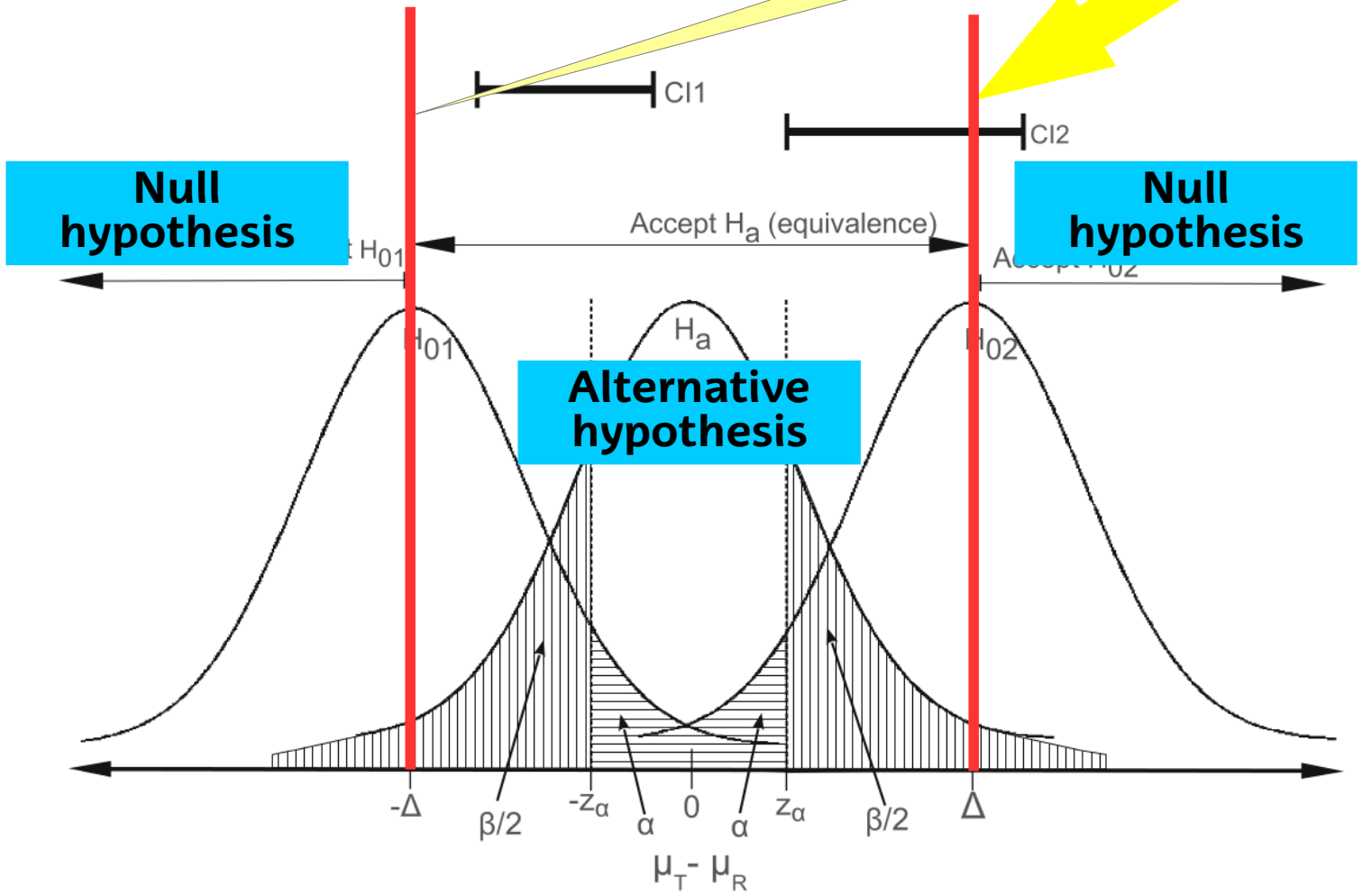
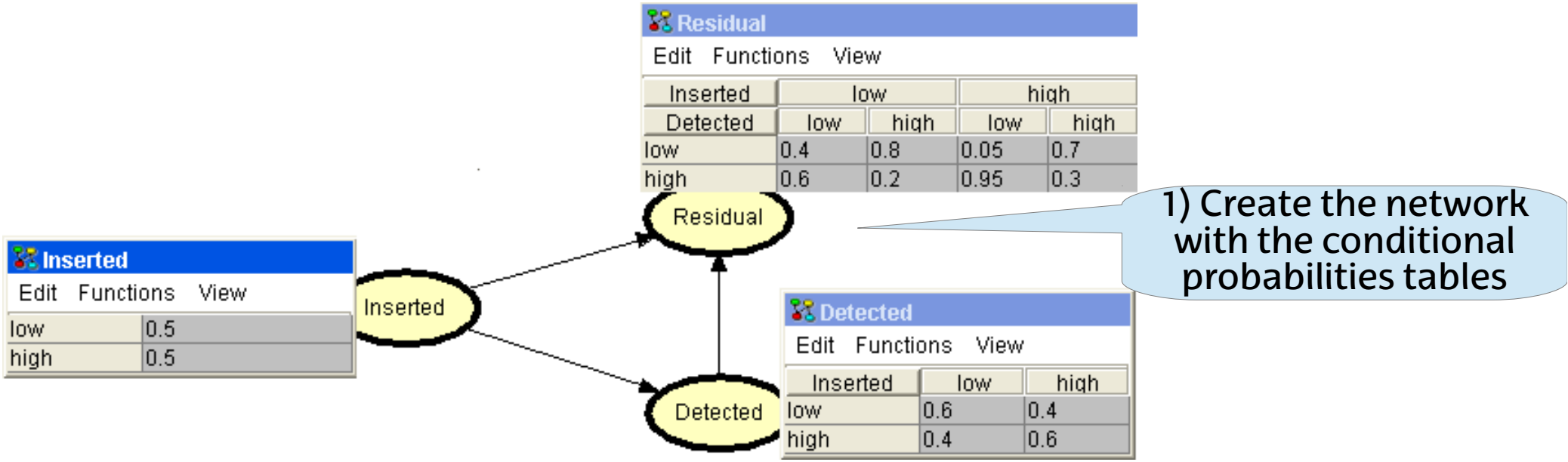


Fig. 2 Plot of the TOST and the confidence intervals CI1 and CI2 for the difference of means

Bayesian Networks

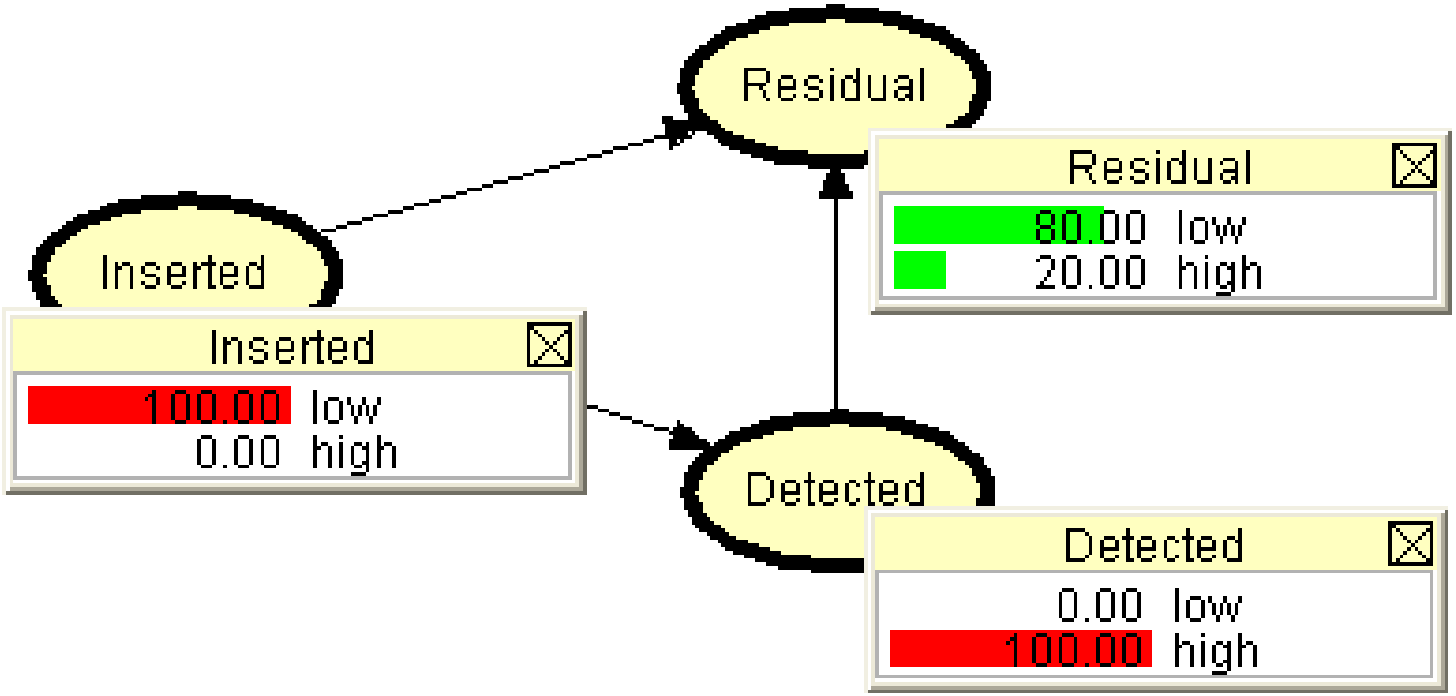
- A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.
- It allows to make inferences from causes to symptoms and from symptoms to causes



Basic bayesian network for defect estimation

Bayesian Networks

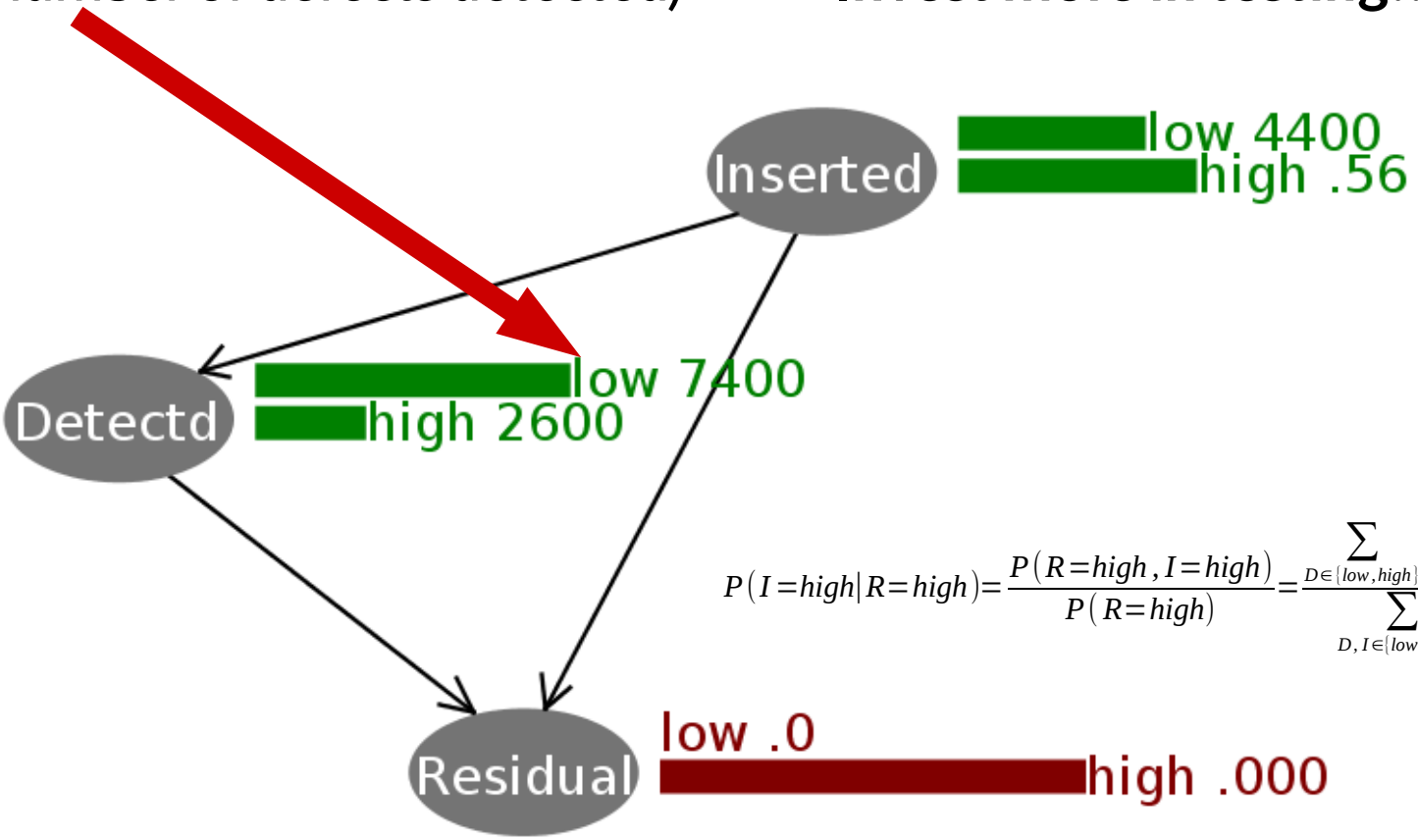
- If we introduce "the evidences" we get the output probabilities (estimates)



Evidences (in red) are introduced in the network

Bayesian Networks

- From symptoms we can get to the causes: "What is the probability that the number of inserted defects was high, given that the number of residual defects is high?" → 56%
  - Compare that value with the probabilities of Defects Detected (74% probability of low number of defects detected) !!!! → **Invest more in testing!!!!!!!**



$$P(I=high|R=high)=\frac{P(R=high,I=high)}{P(R=high)}=\frac{\sum_{D\in\{low,high\}}P(R=high,D,I=high)}{\sum_{D,I\in\{low,high\}}P(R=high,D,I)}$$

If we fix the evidence that we have had a high number of residual defects, we can deduce, by means of the conditional probability formula, the probability (high or low) of the defects inserted.

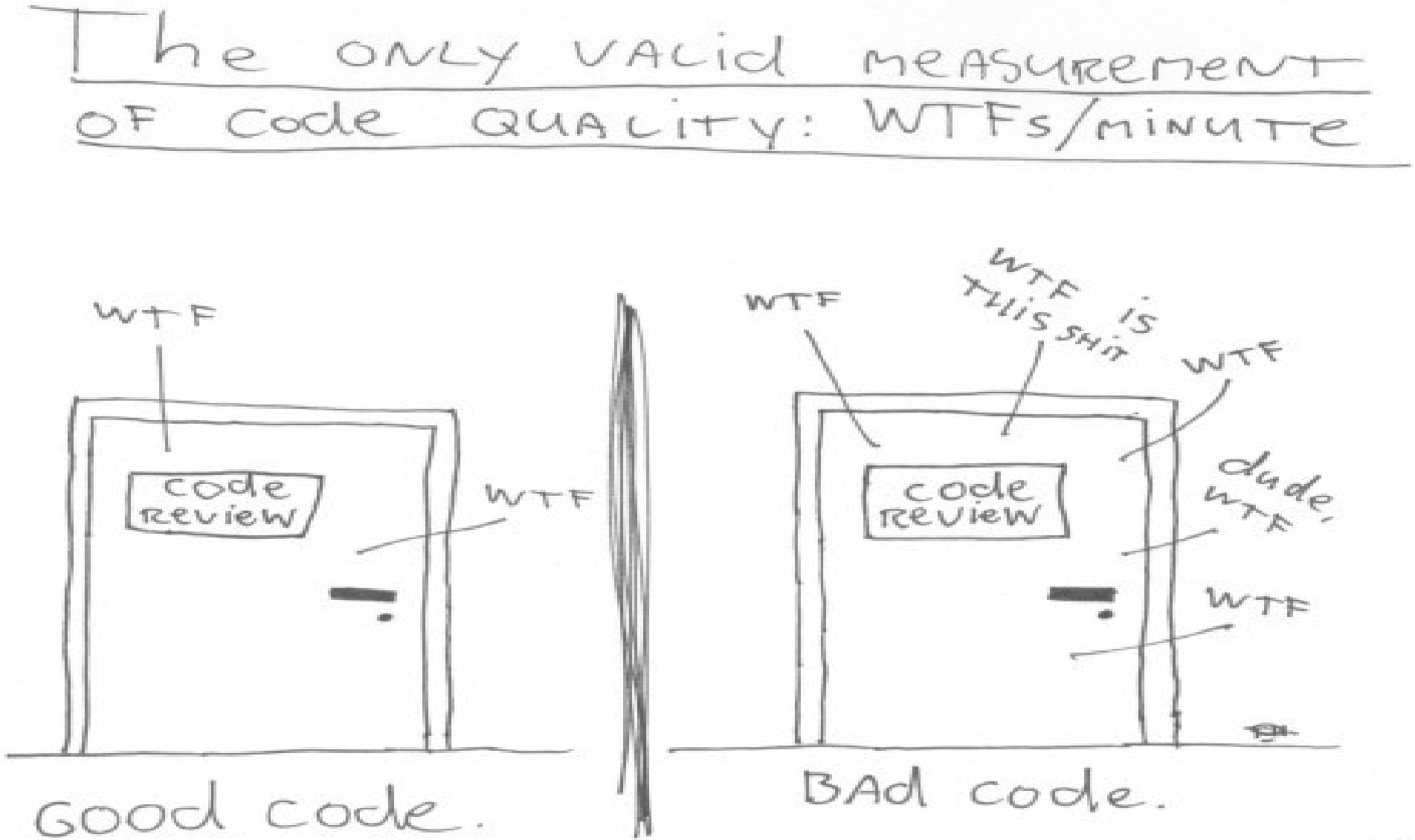
Results

- We have applied many statistical methods to different problems
- 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 it within the organization
- There is no “best method“

Discussion

- Many methods available. The theory may be difficult but they are really easy to apply
- 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