Applied
Software Engineering
Research Group

STATService and EXEMPLAR: SBSE research supporting tools

A not so brief introduction

José Antonio Parejo

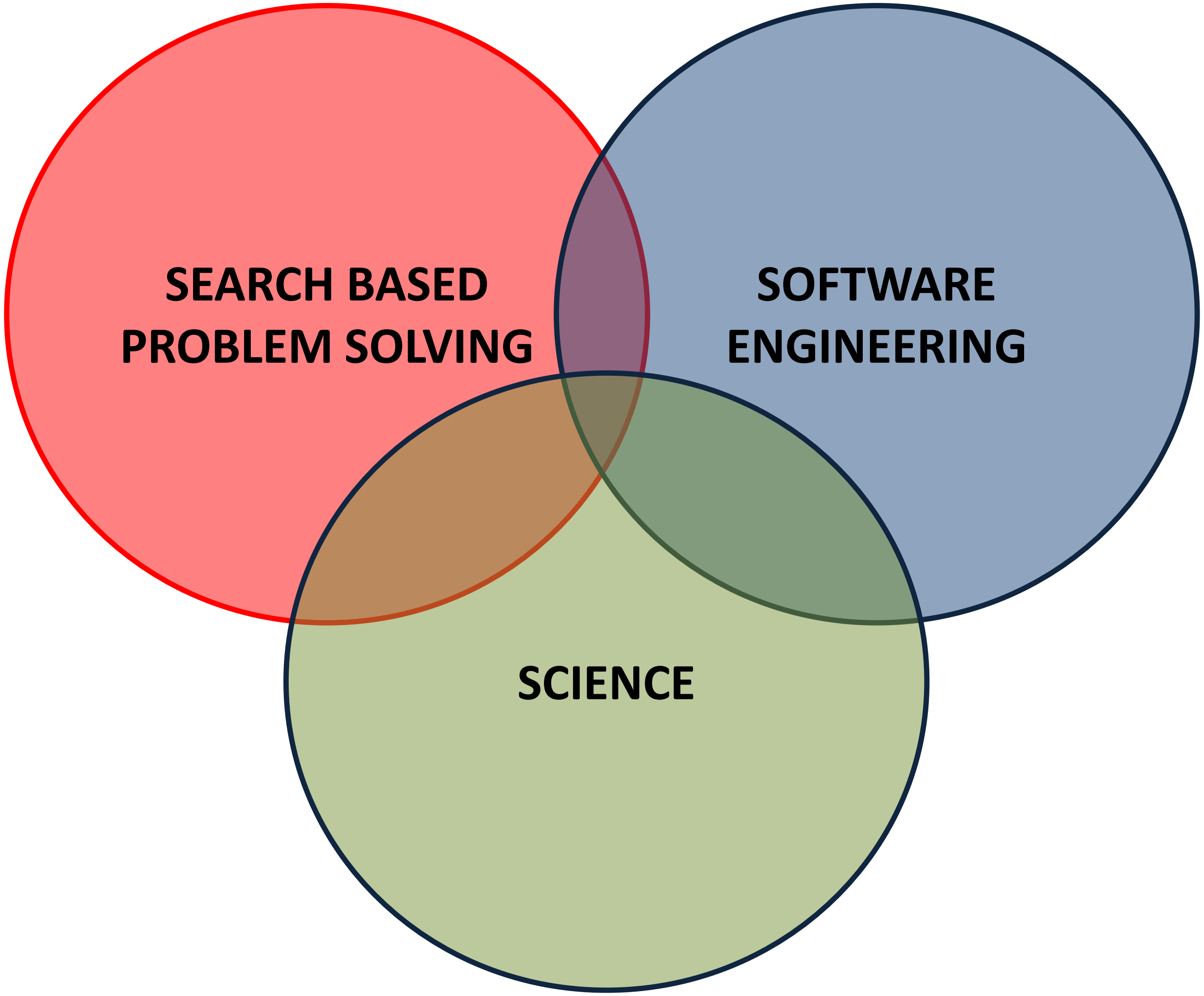
UNIVERSIDAD DE SEVILLA

1st SBSE Summer School 2016

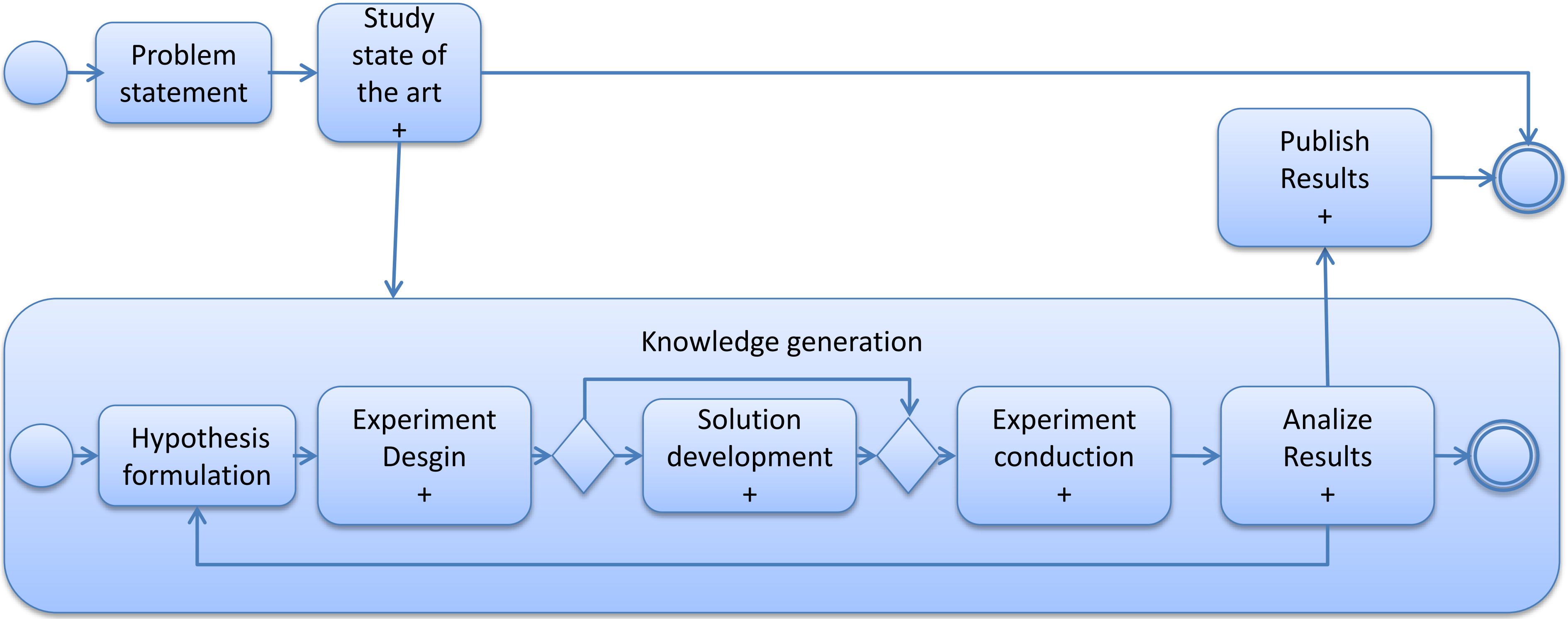
- Introduction/motivation (with survey)
- Background on STH and experimental design
- STATService
- EXEMPLAR
- Conclusions

- **Introduction/motivation (with survey)**
- Background on experimental design and STH
- Currently available tools
- STATService
- EXEMPLAR
- Conclusions

Our field


SEARCH BASED PROBLEM SOLVING
SOFTWARE ENGINEERING
SCIENCE

Our “business” as SBSE researchers



Our target: **The perfect SBSE Researcher**

“Don't only practise your art, but force your way into its secrets; art deserves that, for it and knowledge can raise man to the Divine. “

Ludwig van Beethoven Letter to Emilie, July 17, 1812

Survey

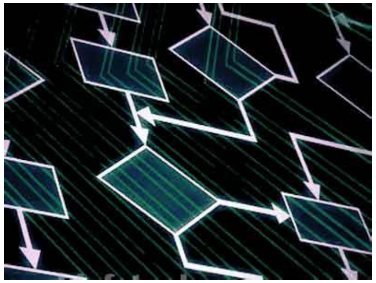
http://goo.gl/forms/YDMANy51IagtHkcp2

Survey Results

https://goo.gl/JWI5Bn

Skills (in Soft. Eng.)

- Understand the methodologies, phases and techniques.
- Evaluate the applicability and the impact of potential improvement in the industry
- Interpret the solutions provided by search methods
- Be good developer and software engineers!!



Skills (in search based problem solving)

- Proper formalization of software engineering challenges as search problems
- Master the search techniques, variants, and extension points, in order to choose those that provide a better fit for your problem
- Develop adaptions for those techniques

Skills (in SCIENCE/RESEARCH)

Furthermore the SBSE researcher should be able to:

- Design experiments in such a way that hypothesis can be refuted of confirmed
- Conduct experiments with minimal threats to the validity of the results.
- Analyze the results of the experiments (using statistical techniques)
- Draw conclusions from the results of such analyses
- Critical thinking even about your own results
- Make your results replicable, communicate and disseminate them

Our experience: Motivation (I)

“Good Ideas, Bad methodology” → 🚫

“Authors should use statistical analysis to support the conclusions drawn” → 🚫

“no statistical tests were performed to validate this claim. Therefore, I don´t endorse this paper” → ⚠

Motivation (II)

Statistical packages (ej: SPSS,R):

- Missign features (for instance non-parametric tests and post-hoc procedures in SPSS)
- Lack of Usability (for non-programmers)
- Lack of interpretation aid

Statistical analysis libraries:

- Lack of usability (for non-programmers)
- Technological constraints
- Data format and structure constraints

The problem behind the problems

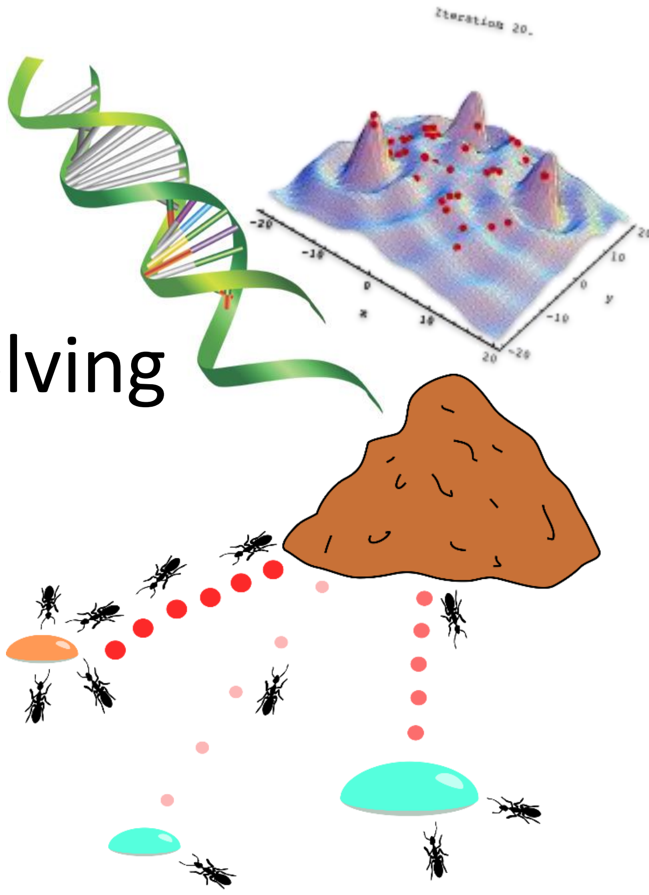
Our target

Michelangelo Buonarotti (Caprese, 1475 - Rome, 1564)

My personal perspective on this issue

Not so bad in:

- Software Engineering
- Search Based Problem Solving



Weak in:

- Empirical Methodology
- Design of Experiments

-Statistics

Motivation for creating tools!

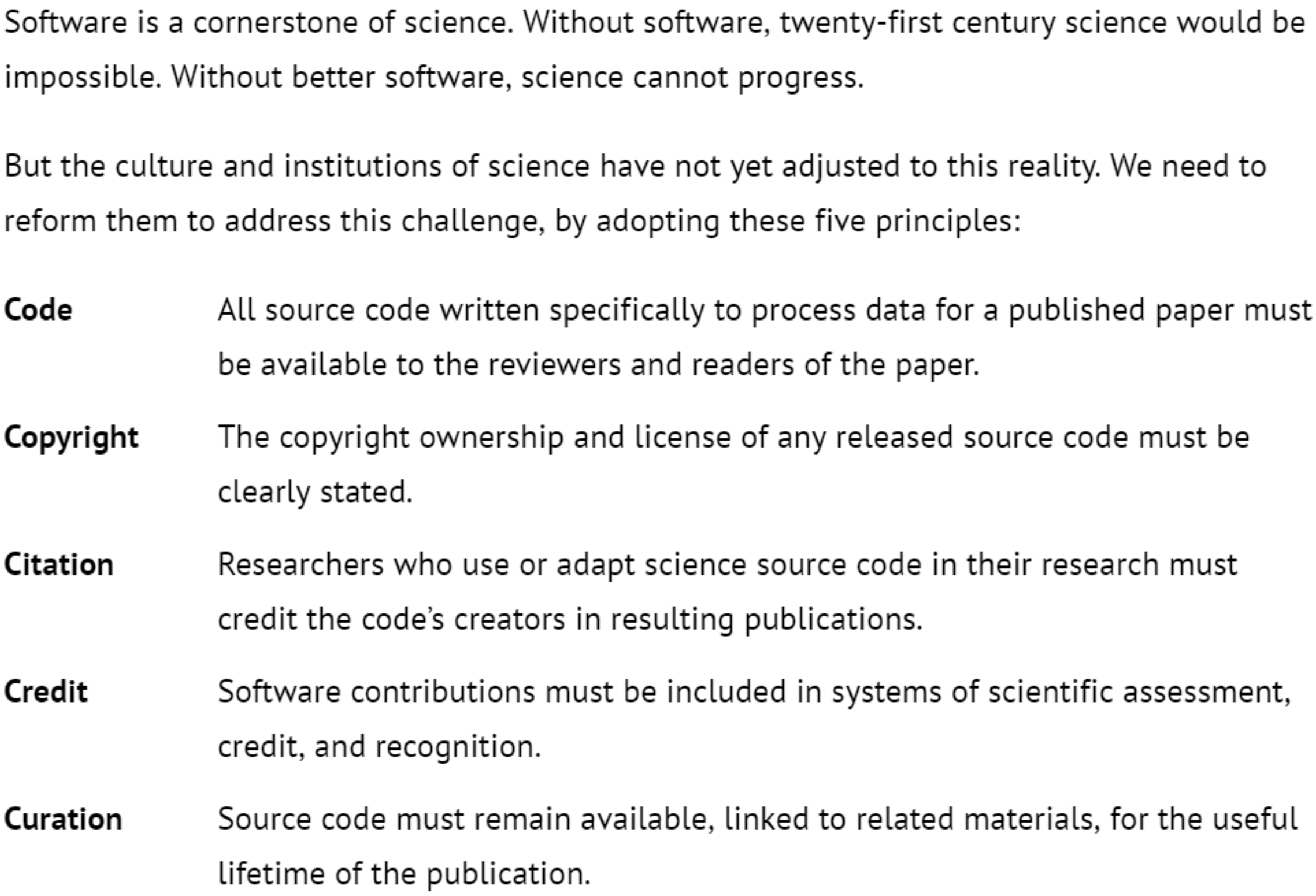
Our "products" as SBSE Researchers…

- Our products are:
  - Papers?
  - Efficient/Performant problem solving algorithms?
  - Algorithm implementations?
  - Verified knowledge?

- What does mean "quality" for such products?

The manifestos (I)

The science code manifesto



Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

Code All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.

Copyright The copyright ownership and license of any released source code must be clearly stated.

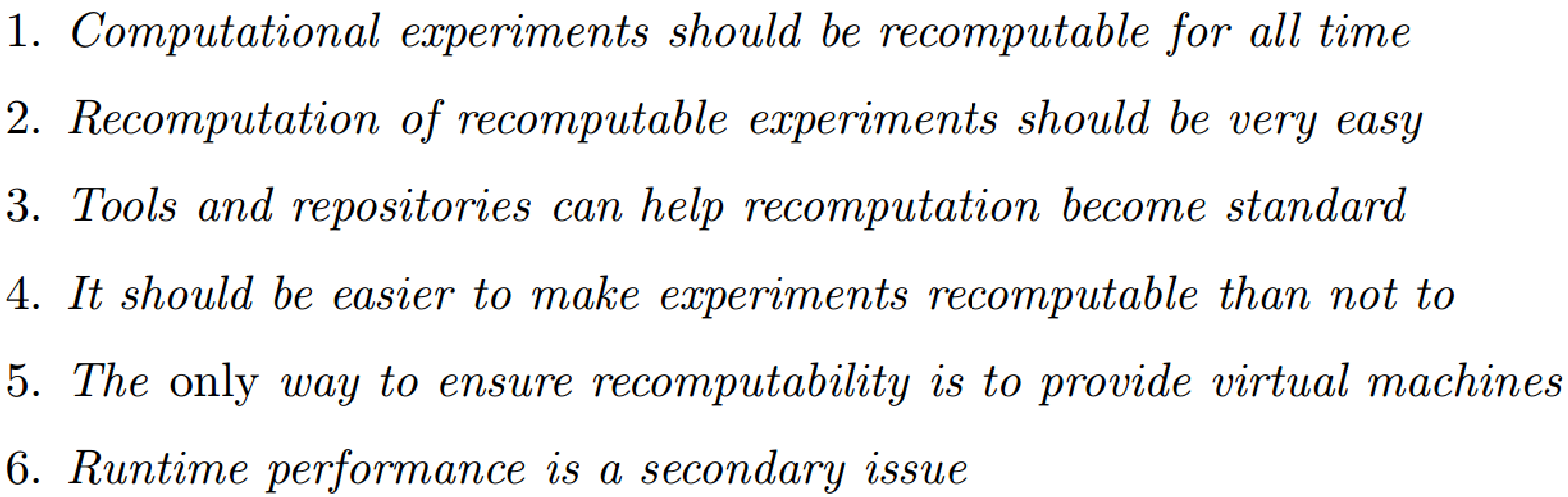
Citation Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.

Credit Software contributions must be included in systems of scientific assessment, credit, and recognition.

Curation Source code must remain available, linked to related materials, for the useful lifetime of the publication.

The manifestos (II)

The recomputation manifesto



1. *Computational experiments should be recomputable for all time*
2. *Recomputation of recomputable experiments should be very easy*
3. *Tools and repositories can help recomputation become standard*
4. *It should be easier to make experiments recomputable than not to*
5. *The* only *way to ensure recomputability is to provide virtual machines*
6. *Runtime performance is a secondary issue*

Questions, questions, questions,…

- Do we endorse the manifestos?
- Can we make our experiments REPRODUCIBLE/RECOMPUTABLE?
- Should we publish the source code of our papers?
  - The data analysis source code?
  - The contribution source code (algorithm, platform, etc.)?

Motivation

"The use of precise, repeatable experiments is the hallmark of a mature scientific or engineering discipline"

Lewis, J.A., Henry, S.M., Kafura, D.G., Schulman, R.S.: On the relationship between the object-oriented paradigm and software reuse: An empirical investigation. Technical report, Blacksburg, VA, USA (1992)

Motivation

- "*Verifying results found in the literature is in practice almost impossible“*
- *“Running a reportedly good algorithm on your own data is an extremely difficult task*"
- “the details presented in a typical paper are insufficient to ensure that one would implement the same algorithm“
  Eiben, A., Jelasity, M.: A critical note on experimental research methodology in EC. Computational Intelligence, Proceedings of the World on Congress on 1 (2002) 582–587
- “most SE experiments results have not been reproduced”
  Natalia Juristo, Omar S. Gómez: Replication of Software Engineering Experiments, chapter of Empirical Software Engineering and Verification. Lecture Notes in Computer Science Volume 7007, 2012, pp 60-88
- “Not only are experiments rarely replicated, they are rarely even replicable in a meaningful way.” Ian P. Gent: The recomputation manifesto. Available online at http://www.recomputation.org/papers/Manifesto1_9479.pdf

Introduction/Motivation

"The use of precise, repeatable experiments is the hallmark of a mature scientific or engineering discipline"

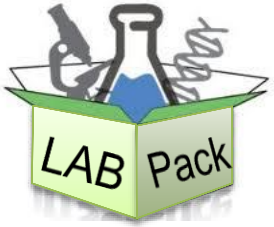
Precission → **~~detailed~~ and ~~unambiguous~~ description** of the experiment

.

→ **Currently? PAPERS**

Repeatability → providing all the **materials used** and an appropiate description of the experimental context.

→ **Currently?**



Summarizing: Two main problems

- Statistical data analysis & Empirical methodology → STAT SERVICE 2.0

- Replicability of results / experiments → EXEMPLAR

- Introduction/motivation (with survey)
- **Background on STH and experimental design**
- STATService
- EXEMPLAR
- Conclusions

Experiment

“a process of systematic inquiry and data collection with the aim to confirm or disprove a hypothesis”

Gliner et al 2012

Scientific Hypothesis

- A “testable” statement that can be falsified through experience and observation
- Scientific hypotheses are defined using variables

Types of Scientific Hypotheses

- **Descriptive hypotheses**

  “The average height of Spanish males is over 1.75m”

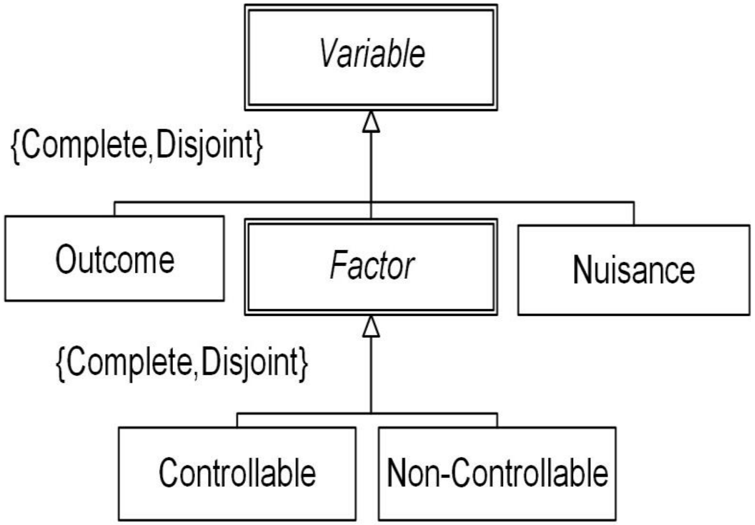
- **Differential hypotheses**

  “The volume of milk that you drink during childhood has an impact on your height”

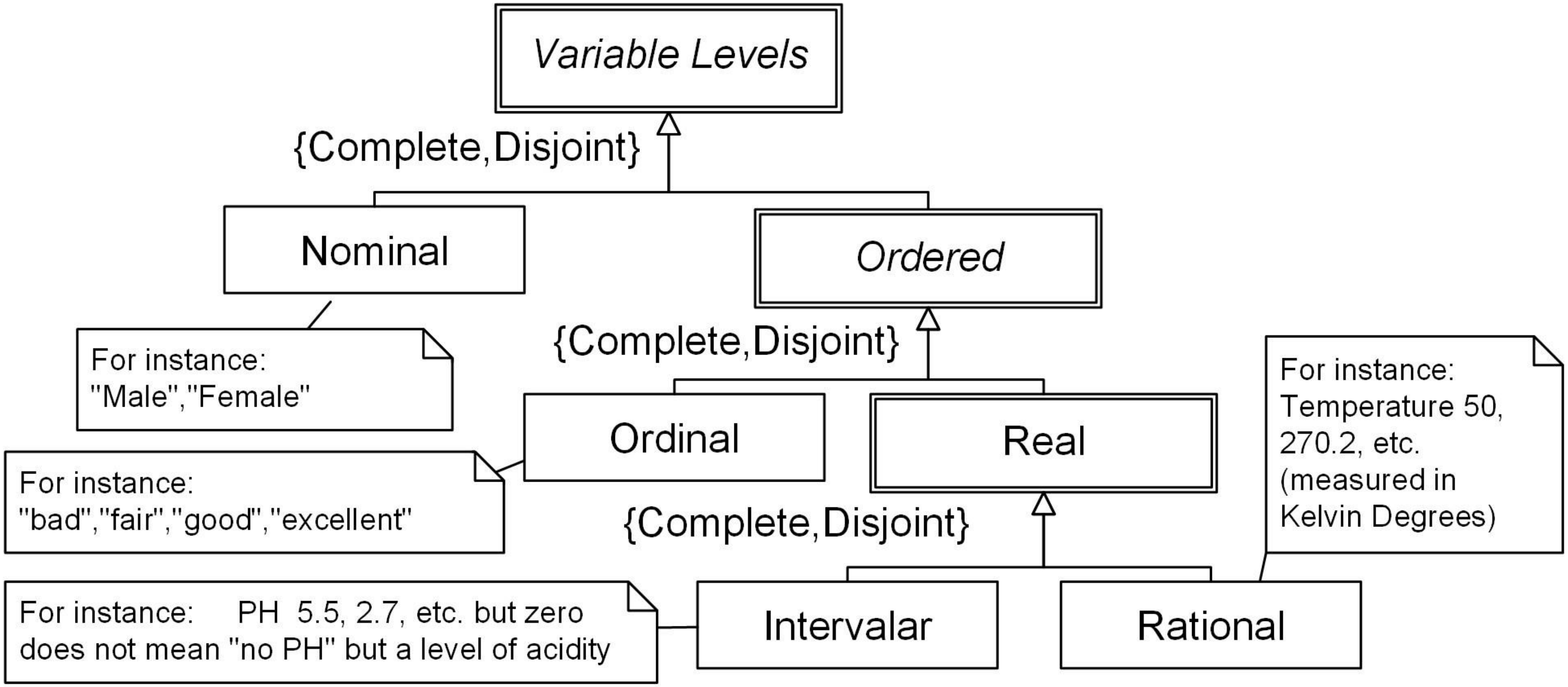
- **Associative hypotheses**

  “The weight of Spanish males is strongly, positively, and linearly correlated with their height”

Role of a variable in the experiment



Variable domains, levels and tpes



For instance: "Male","Female"

For instance: "bad","fair","good","excellent"

For instance: PH 5.5, 2.7, etc. but zero does not mean "no PH" but a level of acidity

For instance: Temperature 50, 270.2, etc. (measured in Kelvin Degrees)

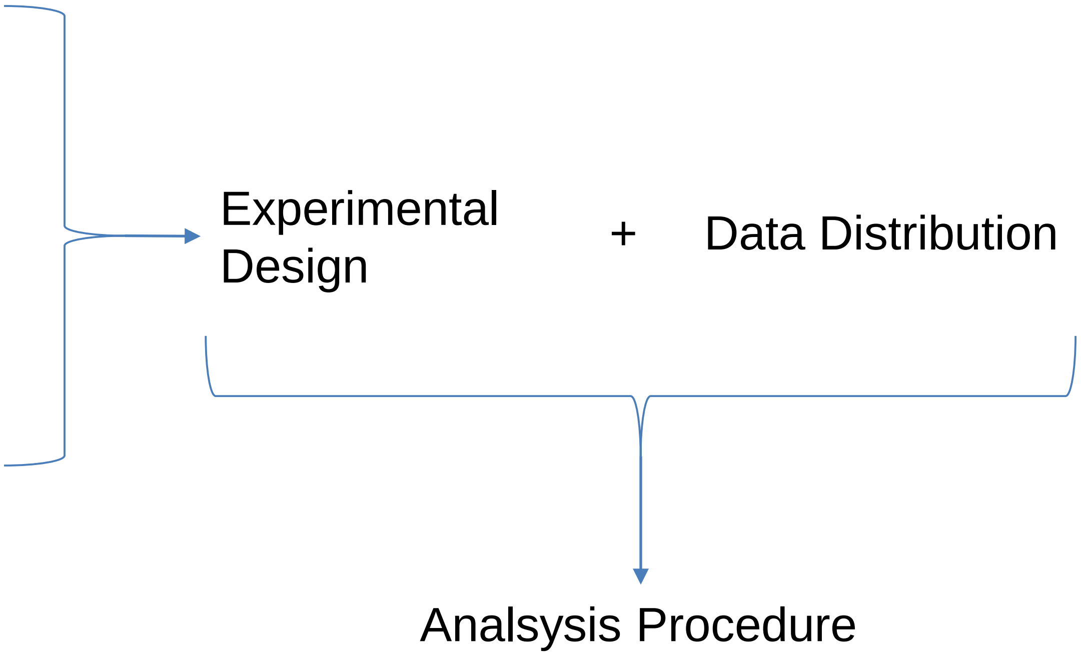
Experimental design

- An experimental design is the specification of the sequence and distribution of modifications of the factors and measurements of the outcomes, such that it allows us to test the hypothesis using a statistical analysis

Principles of Experimental Design

- ***Repetition*.** To reduce the bias introduced by the specific characteristics of every single experimental objects in the observations of the outcome variable.
- ***Randomization*.** To reduce the bias introduced when all the repetitions of a factor level are performed on individuals with similar characteristics
- ***Local Control* or *Blocking***. When a factor makes the outcomes of the experiment non comparable, the selected sample should be partitioned into blocks as homogeneous as possible regarding that factor (or the value of such factor should be randomized)

- Hypothesis type
- Variables
  - Domain
  - Type



Experimental Design + Data Distribution

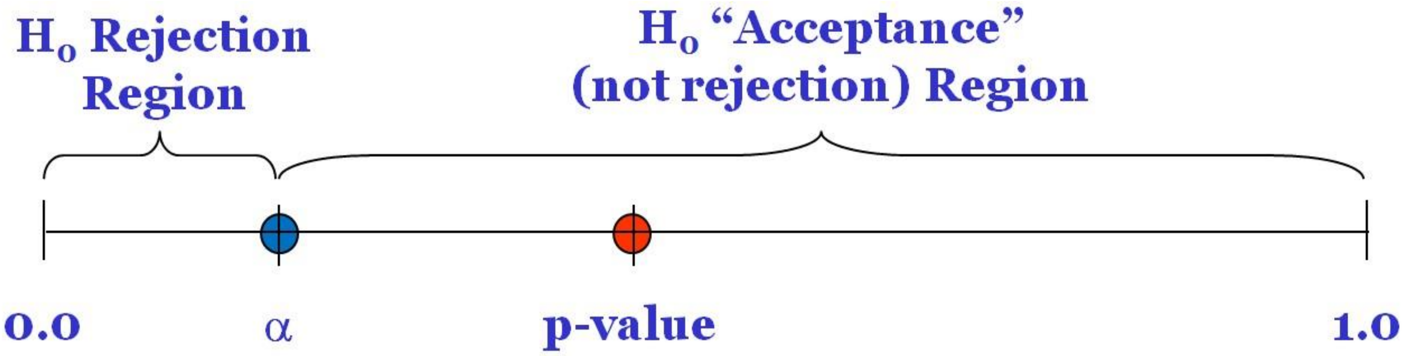
Analsysis Procedure

		Type of Hypothesis		
		Differential	Associational	Descriptive
Number of factors	Zero	-	-	Exploratory analysis and basic STH
	One	Basic STH	Correlation coefficients / regression models	
	More	Complex STH	Complex correlation / regression models	

STH: Stastistical Testing of Hypothesis

- STH works by defining two hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1.
- Both hypothesis are mutually exclusive; i.e., if H_0 holds then H_1 does not hold, and vice-versa
- The null hypothesis is a statement of no effect or no difference, whereas the alternative hypothesis represents the presence of an effect or a difference
- Statistical tests generate a p-value that allows us to discard (or not) H_0 in favour of $H_{1.}$

Interpretation of p-values



WHAT IS THE ACTUAL MEANING OF A P-VALUE?

A p-value is the probability of the observations provided as result of the experiment assuming that H_0 is true

Which STH should I use?

- One factor:

		two-levels factor		three-or-more-levels factor	
		No blocking	Blocking	No blocking	Blocking[4]
Type and distribution of the outcome	Real Normal	Independent Samples t-Test	Paired samples t-Test	Oneway ANOVA	Repeated Measures ANOVA
	Real not-Normal	Mann-withney	Wilcoxon or Sign Test	Kruskal-Wallis	Friedman
	Ordinal				
	Nominal	ChiSquare or Fisher exact Test	McNemar	Chi Square	Cochran Q

- Multiple factors:

	Experimental Design	two-levels factor		three-or-more-levels factor	
		Not blocking	Blocking	Not Blocking	Blocking
Type and distribution of the outcome	Real Normal	Factorial ANOVA	Factorial ANOVA (rep. measures)	Factorial ANOVA	Factorial ANOVA (rep. measures)
	Real not-normal	-	Friedman	-	Friedman
	Ordinal	-	Friedman	-	Friedman

Multiple comparisons and STH

- What is the alternative hypothesis in multiple comparison?

“there are at least one distribution that is different from the rest” ➔ we ignore among which specific pairs of distributions (algorithms)

We need an additional type of statistical technique named post-hoc procedure

Is it enough with the p-values?

- Post-hoc procedures find relationships among a couple of distributions from the associated multiple comparison test.

- They control the accumulation of potential errors that derives for linking a sequence of statistical tests

- They provide a global significance level for all the comparisons performed.

Additional requiremens for differential hypothesis testing

- If you collect enough data, you can prove differential hypothesis between data distributions whose mean is very close

- Statistically significant does not mean relevant in practice

- You must provide an effect size estimator. For instance, for not normal data, you can use A12

- Introduction/motivation (with survey)
- Background on NHST and experimental design
- **STATService**
- EXEMPLAR
- Conclusions

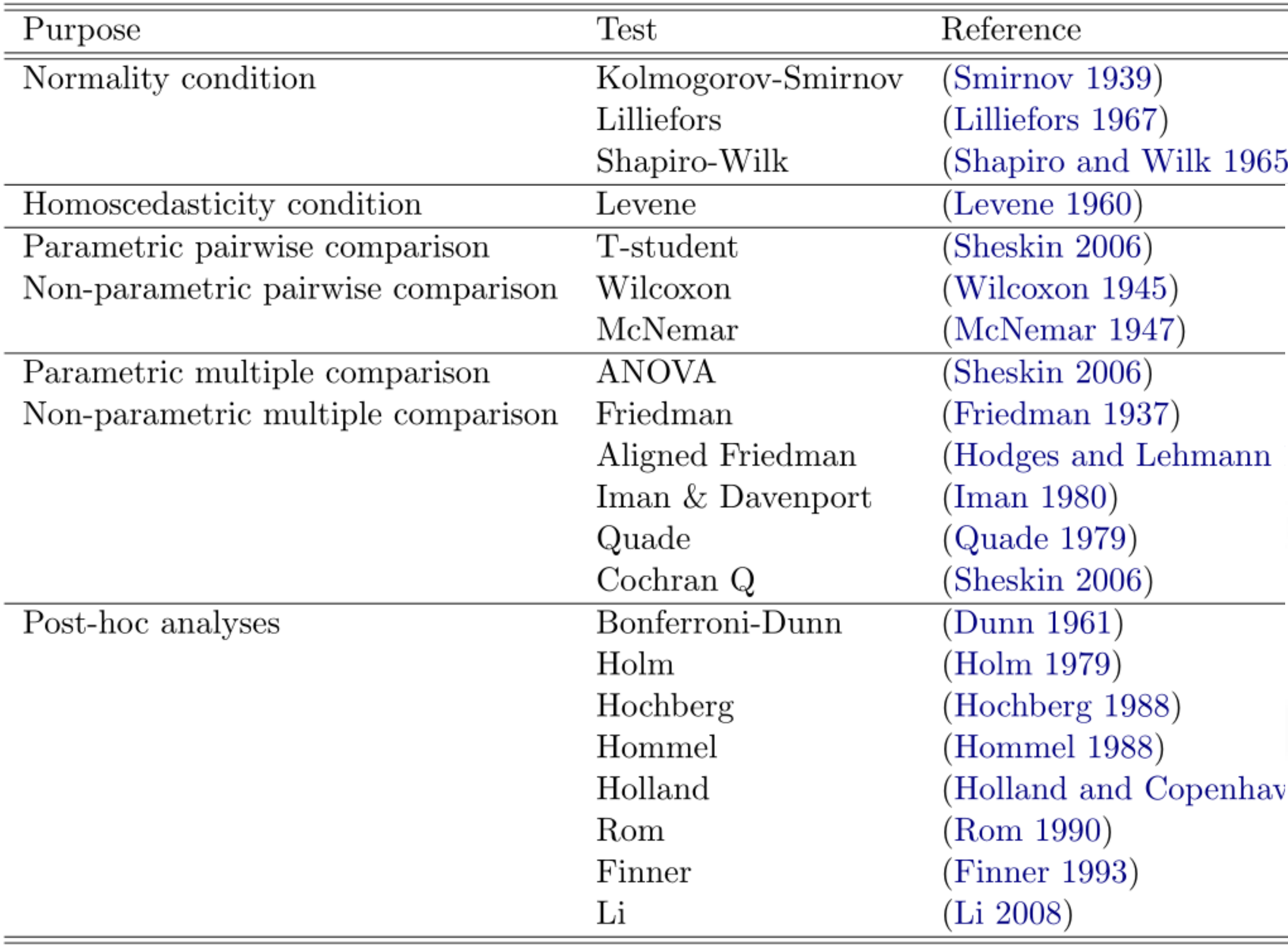
STATService

- A suite of statistical analysis tools that comprises of:
  - A web portal (that support online analysis of datasets).

  - A set XML/SOAP and REST Web Services.
  - A plugin for MS Excel

STATService features

- Supported Test:



Purpose	Test	Reference
Normality condition	Kolmogorov-Smirnov	(Smirnov 1939)
	Lilliefors	(Lilliefors 1967)
	Shapiro-Wilk	(Shapiro and Wilk 1965)
Homoscedasticity condition	Levene	(Levene 1960)
Parametric pairwise comparison	T-student	(Sheskin 2006)
Non-parametric pairwise comparison	Wilcoxon	(Wilcoxon 1945)
	McNemar	(McNemar 1947)
Parametric multiple comparison	ANOVA	(Sheskin 2006)
Non-parametric multiple comparison	Friedman	(Friedman 1937)
	Aligned Friedman	(Hodges and Lehmann
	Iman & Davenport	(Iman 1980)
	Quade	(Quade 1979)
	Cochran Q	(Sheskin 2006)
Post-hoc analyses	Bonferroni-Dunn	(Dunn 1961)
	Holm	(Holm 1979)
	Hochberg	(Hochberg 1988)
	Hommel	(Hommel 1988)
	Holland	(Holland and Copenhav
	Rom	(Rom 1990)
	Finner	(Finner 1993)
	Li	(Li 2008)

STATService features (II) (Web Portal)

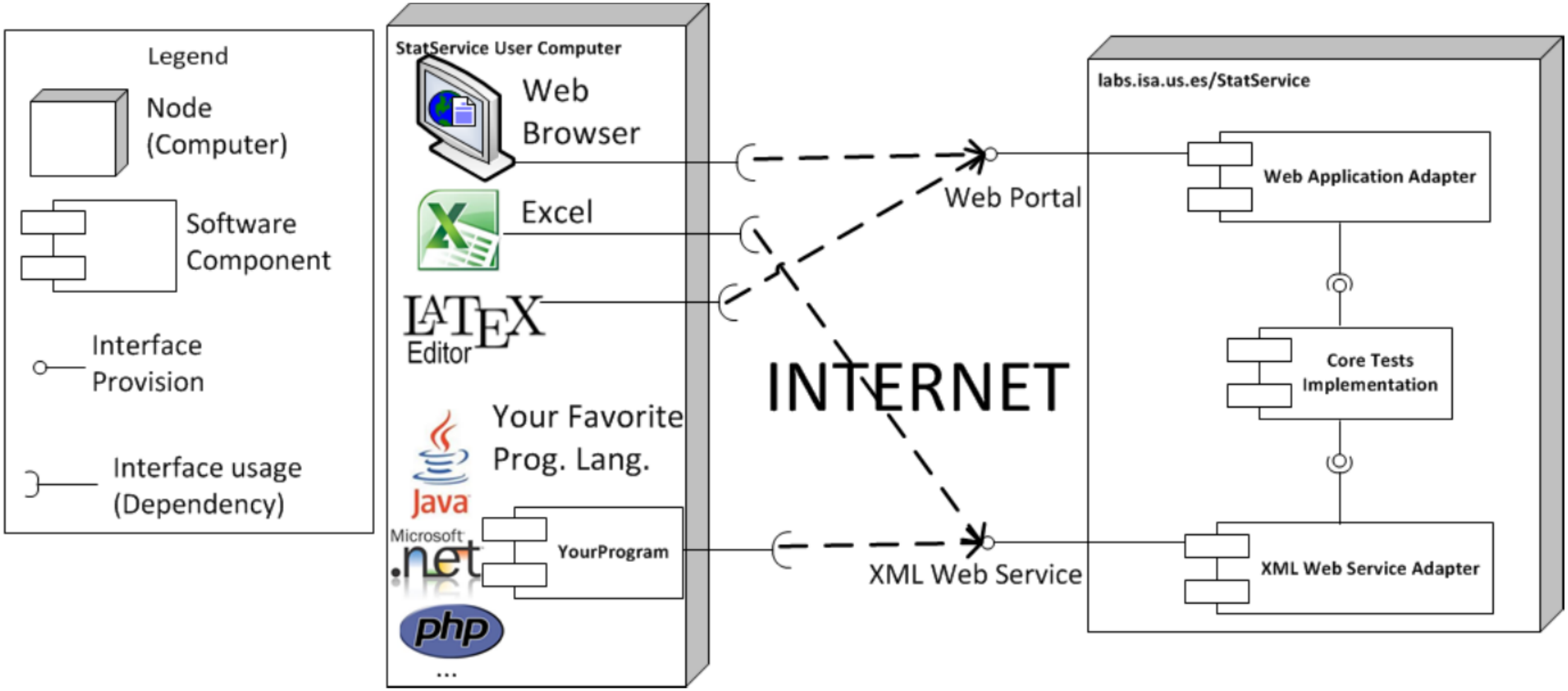
- Versatility:
  - Input Formats (excel, csv, arbitrary text with ad hoc separators).
  - Data transformation.
  - Output formats (XML, HTML, Latex).
- Computer aided test selection (SMARTest) for choosing the appropriate test to be applied. (With some limitations)
- Detailed reporting on decision making and tests results

DEMO

DEMO

http://labs.isa.us.es/apps/statservice

STATService architecture



Apart from SBSE STATService is used for …

Pontes et al. Algorithms for M
http://www.almob.org/conte

RESEARCH

Configura
biclusterin

Beatriz Pontes[1*], Raúl G

Abstract
Background: Bicluster
different subsets of exp
datasets, heuristic searc
techniques is still a chal
conditions, which make
to specify any preferenc
Results: Here, we prese
features in terms of diffe
incorporating new objec
expression patterns, bein
Evolutionary computatio
(Evolutionary Biclusterin
Conclusions: We have d
abilities to obtain meanin
performance with other
also confirm the proper p
Ontology.
Keywords: Gene expres

ELSEVIER

A comparison of ma
for LiDAR-derived e

J. García-Gutiérrez[a,*], F. M

[a] Department of Computer Science, Univers
[b] Department of Computer Science, Pablo d

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form
2 August 2014
Accepted 17 September 2014
Available online 14 May 2015

Keywords:
LiDAR
Machine learning
Regression
Remote sensing

Neural Comput & Applic
DOI 10.1007/s00521-015-2121-7

CrossMark

ORIGINAL ARTICLE

Medium–large earthquake magnitude prediction in Tokyo with artificial neural networks

G. Asencio-Cortés[1] · F. Martínez-Álvarez[1] · A. Troncoso[1] · A. Morales-Esteban[2]

Received: 15 July 2015 / Accepted: 4 November 2015
© The Natural Computing Applications Forum 2015

Abstract This work evaluates artificial neural networks' accuracy when used to predict earthquakes magnitude in Tokyo. Several seismicity indicators have been retrieved from the literature and used as input for the networks. Some of them have been improved and parameterized in order to extract more valuable knowledge from datasets. The experimental set-up includes predictions for five consecutive datasets referring to year 2015, earthquakes with magnitude larger than 5.0 and for a temporal horizon of seven days. Results have been compared to four well-known machine learning algorithms, reporting very promising results in terms of all quality parameters evaluated. The statistical tests applied conclude that differences between the proposed artificial neural network and the other methods are significant.

Keywords Earthquake prediction · Artificial neural networks · Time series

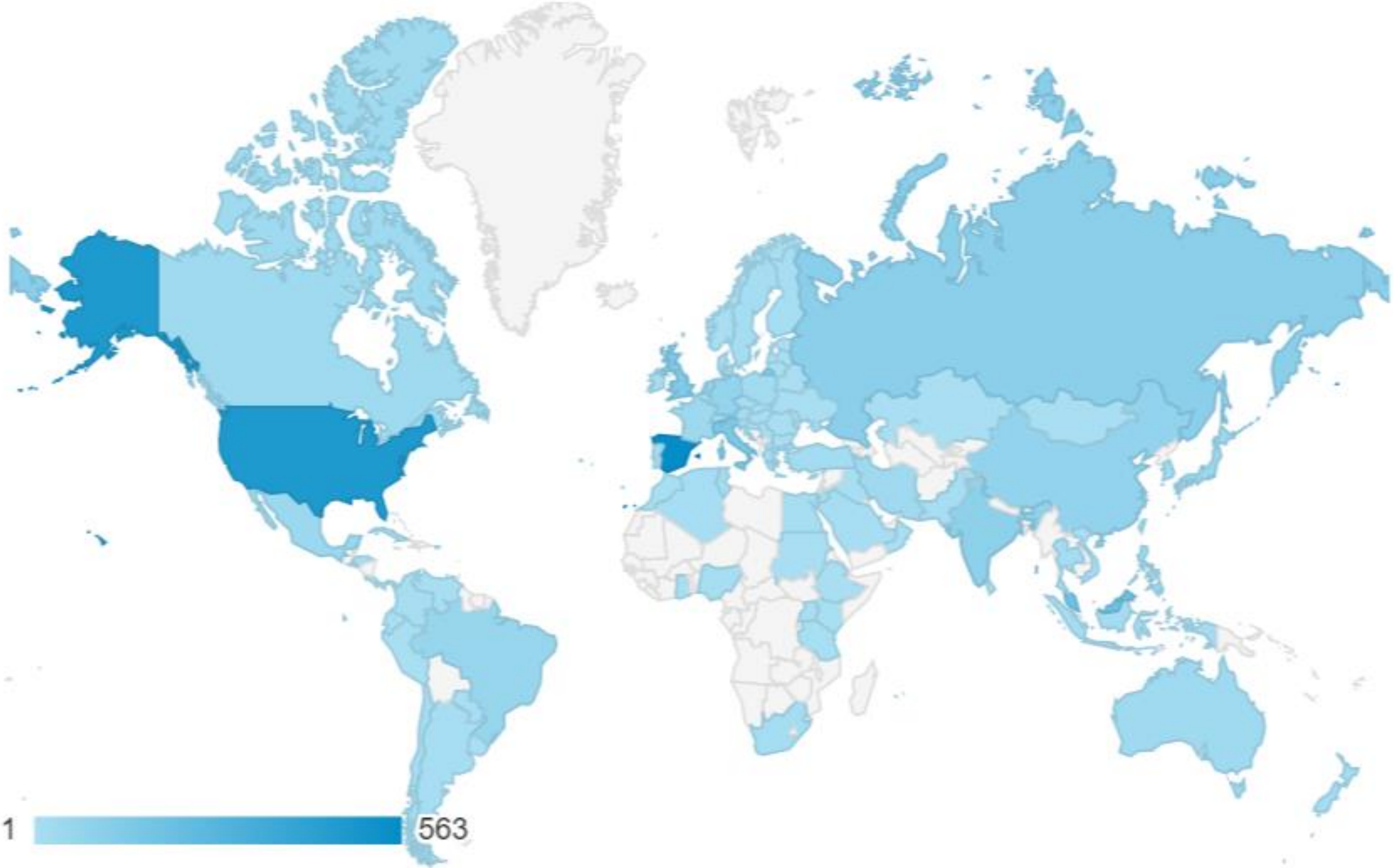
1 Introduction

Earthquakes occur, apparently, after no patterns and can cause huge human and material losses. For this reason, the issue of earthquake prediction has been widely addressed by means of many different strategies. Nevertheless, results are not very convincing to date because no developed method can be used worldwide.

Japan emerges as one of the countries with larger seismic activity, with more than 5000 quakes per year, being 1000 of them felt by the population. Destructive earthquakes, often resulting in tsunamis, occur several times a century. The most recent major quakes include the 2011 Tohoku earthquake and tsunami, the 2004 Chuetsu earthquake and the Great Hanshin Earthquake of 1995.

The important seismic and volcanic activity of Japan is due to the tectonic plate movement. It is also responsible for the shape and contents of the Japan archipelago. Fifteen million years ago Japan formed part of Eurasia. The sub-

Where is used STATService?



Alternatives

- Statistical analysis systems:
  - SPSS,SAS, Minitab
  - R
  - Mathlab, Mathematica, etc.
- Libraries (for Java):
  - JavaNPST
  - Support libraries (Garcia et al. 2009 y 2010).
  - Apache Math Commons

- Introduction/motivation (with survey)
- Background on NHST and experimental design
- Currently available tools
- STATService
- **EXEMPLAR**
- Conclusions

Our Approach

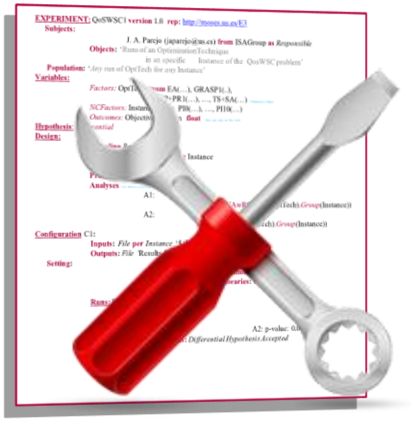
EXpEriments Management PLAtfoRm

Online Repository

Exp. descriptions & lab-packs authoring

Automated Analysis Tools

+


EXPERIMENT: QoSWSC1 version 1.0 rep: http://moses.us.es/E3
Subjects:
J. A. Parejo (japarejo@us.es) from ISAGroup as Responsible
Objects: 'Runs of an OptimizationTechnique in an specific Instance of the QosWSC problem'
Population: 'Any run of OptTech for any Instance'
Variables:
Factors: OptTe… enum EA(…), GRASP1(..), …+PR1(…), …, TS+SA(…)
NCFactors: Instan… PI0(…), …, PI10(…)
Outcomes: Objective… float
Hypothesis: …ential
Design:
Instance
Analyses
A1:
A2:
…Tech).Group(Instance))
…ech).Group(Instance))
Configuration C1:
Inputs: File per Instance
Outputs: File Results
Setting:
…braries:
Runs:
A2: p-value: 0.0…
…s: Differential Hypothesis Accepted

+

Related work

- Exp. Information repositories:

my experiment

research object.org
Supporting Reproducible Research

- Experimental Workflow platforms:

Taverna

Kepler
Your Science. Enabled.

  – R. Salado-Cid, J.R. Romero, S. Ventura. "**Metaherramienta para la generación de aplicaciones científicas basadas en workflows**". Actas de *X Jornadas de Ciencia e Ingeniería de Servicios* (JCIS 2014). pp. 96-105. Cádiz (España). ISBN:

    978-84-697-1153-8

EXEMPLAR in github

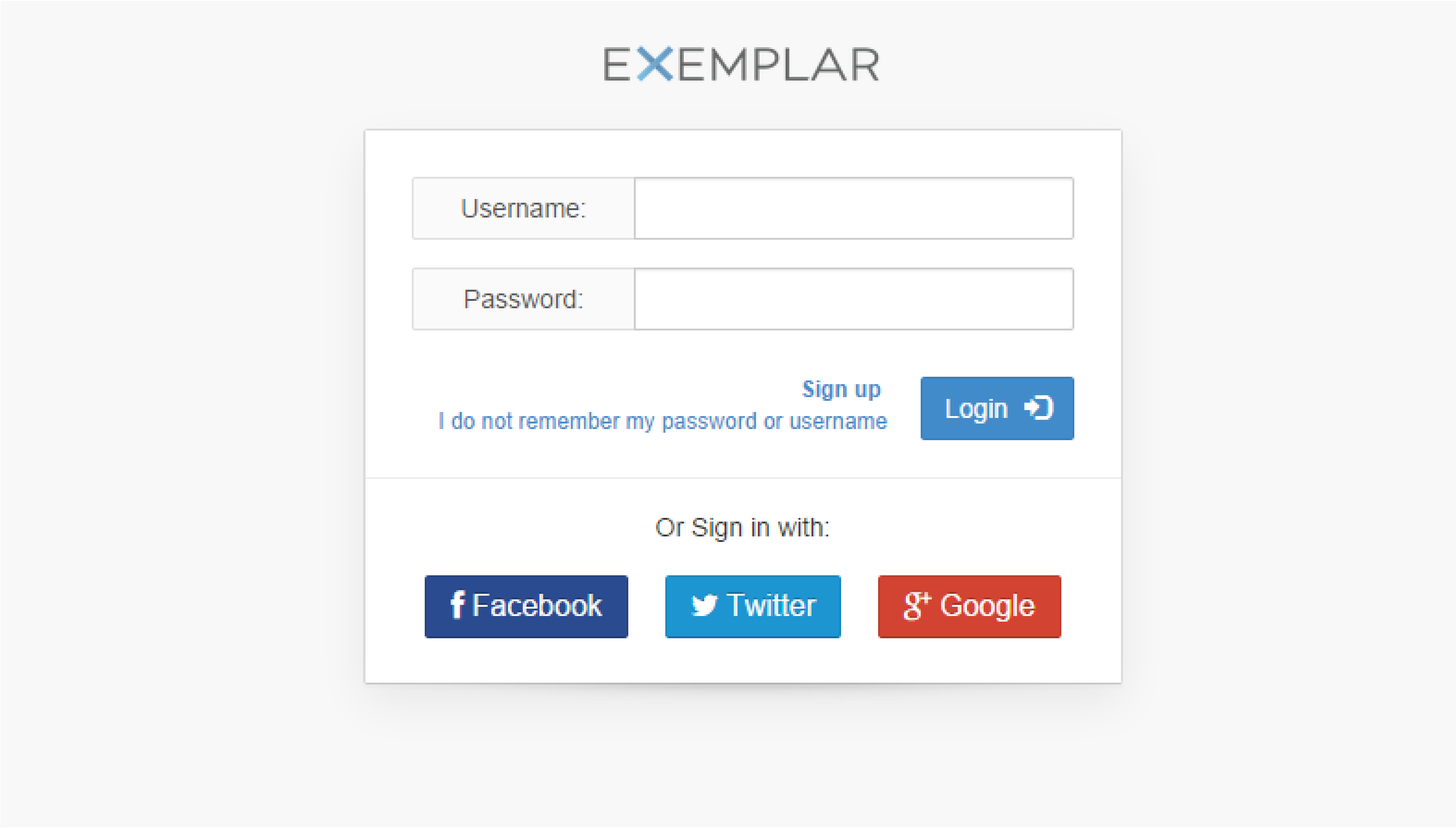
- IDEAS Studio (online editor & repository) https://github.com/isa-group/ideas-studio
- SEDL Module (Experiments description language): https://github.com/isa-group/ideas-sedl-module

  https://github.com/isa-group/sedl

  https://github.com/isa-group/sedl-analyzer
- R Module: https://github.com/isa-group/ideas-r-module

Grupo de investigación en Ingeniería del Software Aplicada

Exp. Inf. Rep. – Social Login



EXEMPLAR

Username:

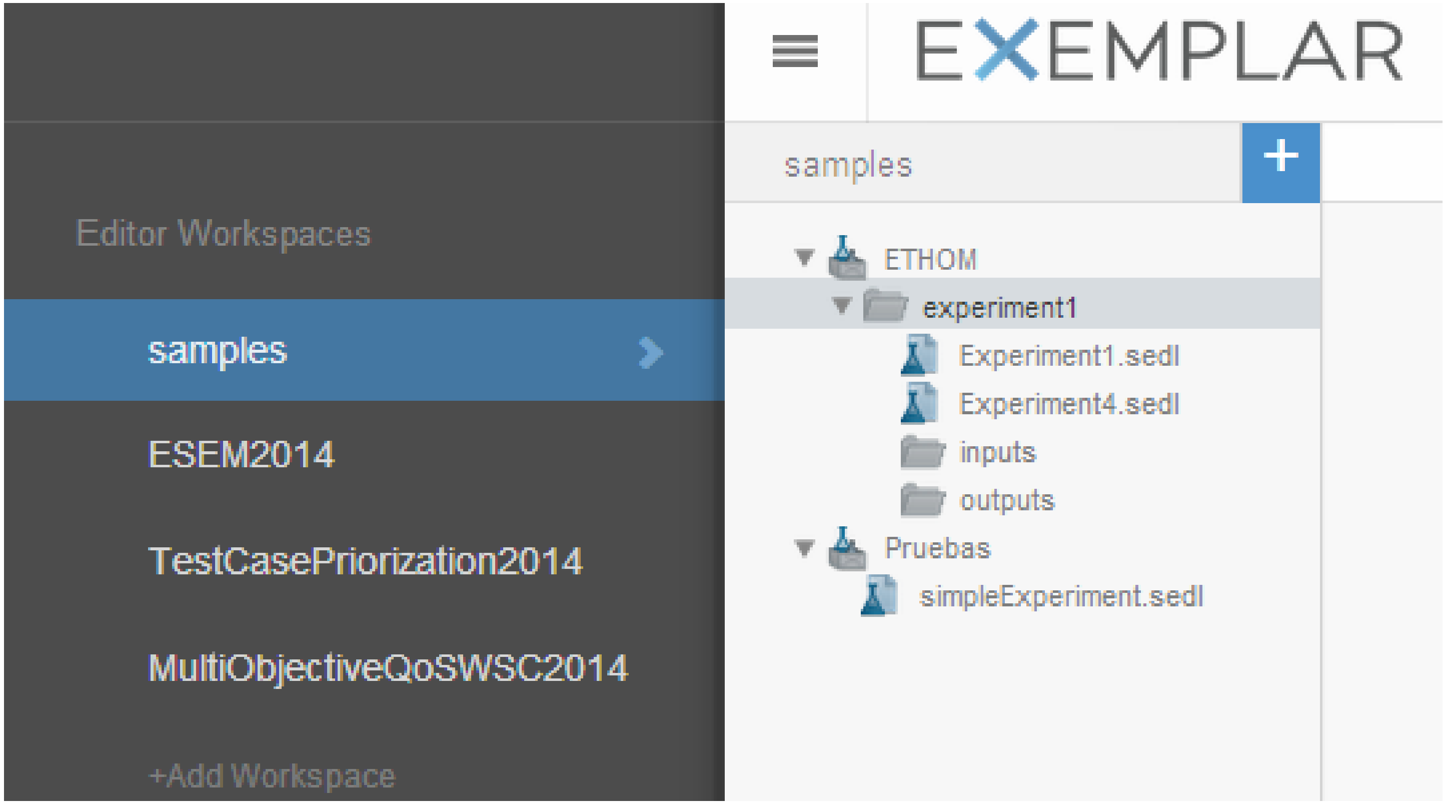
Password:

Sign up

I do not remember my password or username

Or Sign in with:

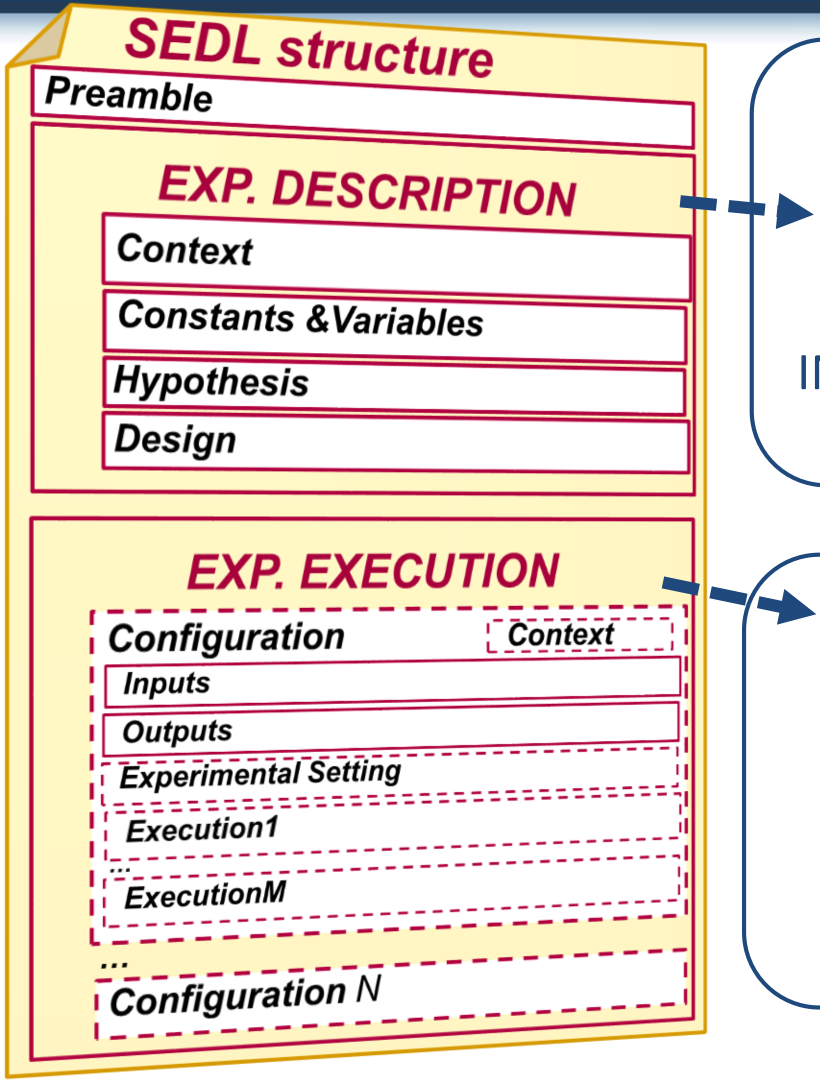
Exp. Inf. Rep. – Workspaces & Projects



Experimental information repository

DEMO

SEDL in a nuthsell



WHO?

WHAT?

TO WHOM?

IN WHICH ORDER?

Human readable & editable

HOW?

INPUT DATA?

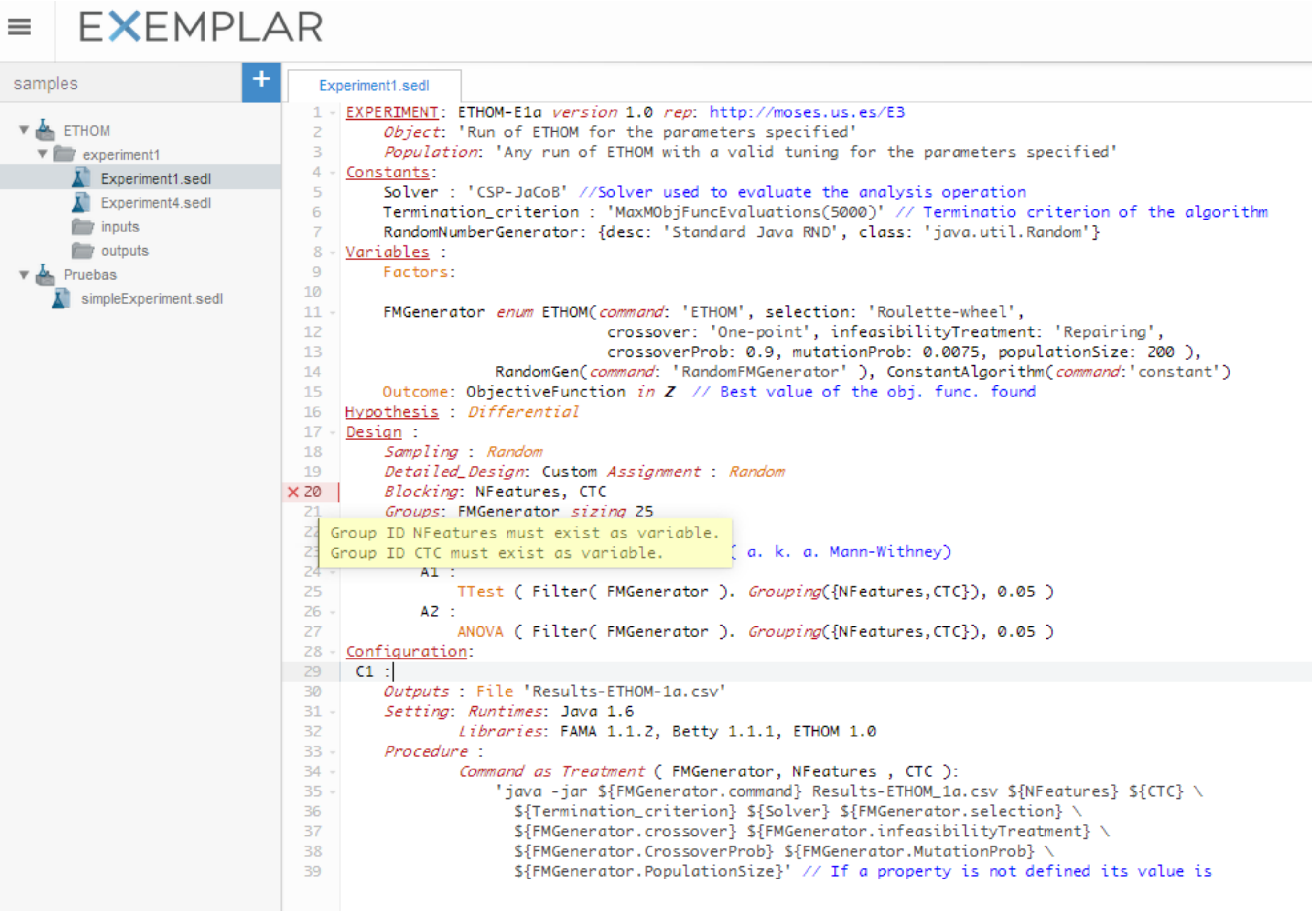
WHEN?

WHERE ARE THE

RESULTS

Human readable, but usually generated automatically

SEDL Editor



```
EXPERIMENT: ETHOM-E1a version 1.0 rep: http://moses.us.es/E3
    Object: 'Run of ETHOM for the parameters specified'
    Population: 'Any run of ETHOM with a valid tuning for the parameters specified'
Constants:
    Solver : 'CSP-JaCoB' //Solver used to evaluate the analysis operation
    Termination_criterion : 'MaxMObjFuncEvaluations(5000)' // Terminatio criterion of the algorithm
    RandomNumberGenerator: {desc: 'Standard Java RND', class: 'java.util.Random'}
Variables :
    Factors:

    FMGenerator enum ETHOM(command: 'ETHOM', selection: 'Roulette-wheel',
                                crossover: 'One-point', infeasibilityTreatment: 'Repairing',
                                crossoverProb: 0.9, mutationProb: 0.0075, populationSize: 200 ),
                RandomGen(command: 'RandomFMGenerator' ), ConstantAlgorithm(command:'constant')
    Outcome: ObjectiveFunction in Z  // Best value of the obj. func. found
Hypothesis : Differential
Design :
    Sampling : Random
    Detailed_Design: Custom Assignment : Random
    Blocking: NFeatures, CTC
    Groups: FMGenerator sizing 25

                                      ( a. k. a. Mann-Withney)
        A1 :
            TTest ( Filter( FMGenerator ). Grouping({NFeatures,CTC}), 0.05 )
        A2 :
            ANOVA ( Filter( FMGenerator ). Grouping({NFeatures,CTC}), 0.05 )
Configuration:
 C1 :
    Outputs : File 'Results-ETHOM-1a.csv'
    Setting: Runtimes: Java 1.6
            Libraries: FAMA 1.1.2, Betty 1.1.1, ETHOM 1.0
    Procedure :
            Command as Treatment ( FMGenerator, NFeatures , CTC ):
                'java -jar ${FMGenerator.command} Results-ETHOM_1a.csv ${NFeatures} ${CTC} \
                  ${Termination_criterion} ${Solver} ${FMGenerator.selection} \
                  ${FMGenerator.crossover} ${FMGenerator.infeasibilityTreatment} \
                  ${FMGenerator.CrossoverProb} ${FMGenerator.MutationProb} \
                  ${FMGenerator.PopulationSize}' // If a property is not defined its value is
```

```
Group ID NFeatures must exist as variable.
Group ID CTC must exist as variable.
```

Why automated analysis?

- Are we using the appropriate statistical test for our design, variables and hypothesis?
- Do we have enough students / individuals / algorithm runs (given the analysis that we plan to perform)?
- …

Experimental descriptions authoring

DEMO

Experimental analysis replication

- R module for EXEMPLAR:
  - R Script editor with syntax coloring an linter.
  - R Script execution.
  - Plots generation.
  - One-click, online replicability of your analyses without installation.

- Introduction/motivation (with survey)
- Background on NHST and experimental design
- STATService
- EXEMPLAR
- **Conclusions**

Conclusions on the skills of an SBSE researcher

- We are not geniuses of the Renaissance so…
- Team work and collaboration is essential
  ➔ SEBASE Net is a good idea!!
- Newcomers need to acquire a wide set of skills and practice
  ➔Masters/PhD courses are good ways to acquire those skills but a summer school on SBSE can be even better!!

Personal Conclusions on Tool Creation

- Tools (if successful) are worthy in terms of:
  - Citations & Visibility
  - Pride & non-academic curriculum
- Tools are not worthy in terms of:
  - Academic curriculum, i.e. Number of publications / effort required (in development and maintenance)
- Eat your own dog food and be happy with it

Final conclusions

- STATService can ease the task of test selection and application
- STATService does not provide effec size
- EXEMPLAR & SEDL + R can improve the replicability of your experiments
- We are introducing some complexity and overhead ☹

Thank you!!!

Questions?

José Antonio Parejo

japarejo@us.es

Departamento de Lenguajes y Sistemas Informáticos
E.T.S. Ingeniería Informática, Universidad de Sevilla, España