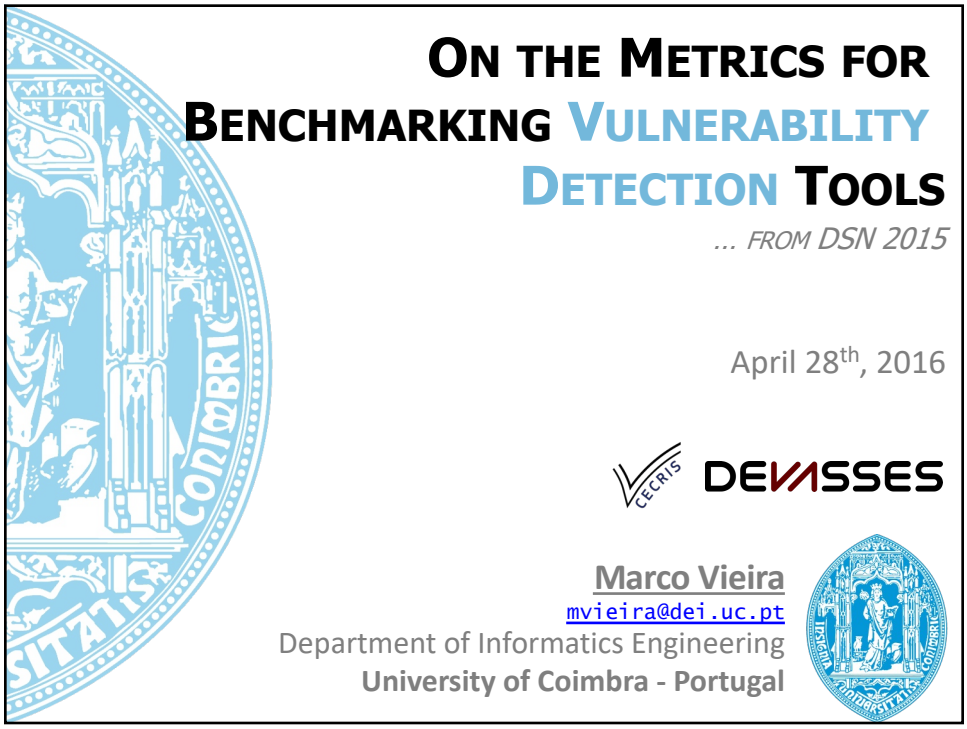


1

Marco's Background (1)

- BSc and MSc in Informatics Engineering
- PhD in Computer Science
- Associate professor at UC
  - Teaching experience of 17 years
    - Hum… I'm getting old ☹
- Adjunct Associate Teaching Professor at CMU
- Large experience in projects with industry
  - Portugal Telecom, Critical Software, ESA, …

2

Marco's Background (2)

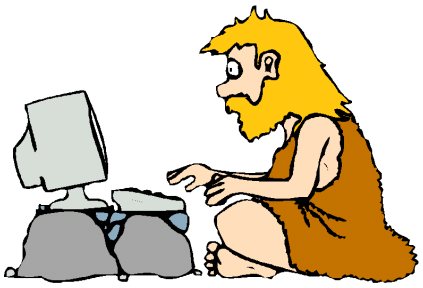
- Research areas:
  - Dependable and secure computing
  - Experimental evaluation
  - Software engineering

- Teaching areas:
  - Software engineering
  - Databases
  - And many other things…
    - Discrete math, telecommunications, data analysis, strategic information systems, programming languages, etc., etc.

3

Marco's Background (3)

- With industry:
  - Software engineering
  - Databases and data warehousing
  - Decision support systems



4

Coimbra

- City in the center of Portugal
  - 200 Km to the north of Lisbon
- ~ 150 000 people
- Most activity around the University
- Many centuries of history

5

6

7

8

9

10

11

University of Coimbra

- University of Coimbra
  - One of the oldest in the world
    - Created in 1290
  - 9 schools (faculties)
    - Sciences and Technology
    - Law
    - Medicine
    - Pharmacy
    - Arts and Humanities
    - Economics
    - Psychology and Education Sciences
    - Sport Sciences and Physical Education
  - About 23000 students
    - 18% of which are foreigners, including > 2000 Brazilians

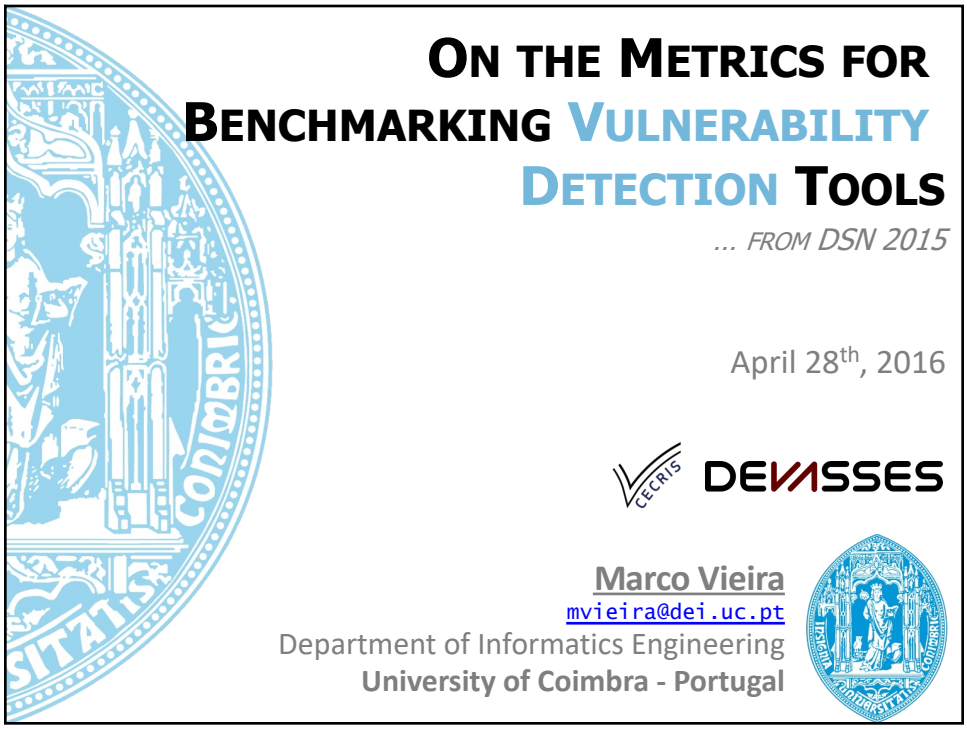
www.uc.pt

12

Software and Systems Engineering (SSE)

- Part of CISUC
  - Coordinated by Marco Vieira
- Key people:
  - Coordinated by Henrique Madeira
  - 15 PhDs (Full Members) + 8 PhDs (Associate Members)
  - 30 PhD students
- Areas of interest:
  - Trustworthy and Resilient Software and Systems
  - Critical Services on the Cloud
  - Efficiency in Software Development
  - Reconfigurable Hardware for Resilient Systems

13



14

OUTLINE

Context and Motivation

- Studying the Selection of Metrics
  - Metrics Gathering and Analysis
  - Empirical Analysis and Metrics Selection
  - Use Expert Knowledge with MCDA for validation
- Demonstration
- Conclusions

MOTIVATION

«If you cannot measure, you cannot improve»

«[W]hen you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind»

Lord Kelvin

«You can't control what you can't measure»

DeMarco

«To measure is to know»

Maxwell

«Measurement motivates»

Galbraith

[Jaquith07] Security metrics: replacing fear, uncertainty, and doubt

VULNERABILITY DETECTION TOOLS

Used by developers to detect software vulnerabilities
- Allow to save time, effort and money

- Often very expensive
- Many tools generate conflicting results
- Due to time constraints or resource limitations…
  … developers have to select a tool from the ones available…
  … and rely on that tool to detect vulnerabilities

- However, practice and experience shows that the effectiveness of many tools is **low**

17

HOW TO SELECT THE TOOL TO USE?

The value of existing evaluations is limited:
- By the short number of tools assessed
- By the representativeness of the experiments

- Developers urge practical ways to compare alternative tools concerning their ability to detect vulnerabilities
  - *We believe that this is important for improving the current state in software vulnerability*

- The solution: **Benchmarking!**

18

BENCHMARKING VD TOOLS

Benchmarks are standard approaches to evaluate and compare different systems
- according to specific characteristics

- Evaluate and compare the existing tools
- Select the most effective one(s)
- Guide the improvement of methodologies
  - In the same way performance benchmarks have contributed to improve performance of systems…

19

EXISTING APPROACHES

«Benchmark for SQL Injection Vuln. Detection Tools»

[Antunes15]

- Based on a well defined set of components:
  - Procedure:
    - Procedures and rules that must be followed during the execution
  - Metrics:
    - *Precision, recall, F-Measure*
    - Characterize the effectiveness of the tools
    - Easy to understand
    - Allow the comparison among different tools
  - Workload:
    - Services to exercise the Vulnerability Detection Tools

20

EXISTING APPROACHES

«Benchmark for SQL Injection Vuln. Detection Tools»

[Antunes15]

Practice show that the metrics generally used are **inadequate** in many scenarios

- Metrics:
  - *Precision, recall, F-Measure*

In fact, we argue that a single set of metrics *cannot be used in every scenario!*

«One size does not fit all!»

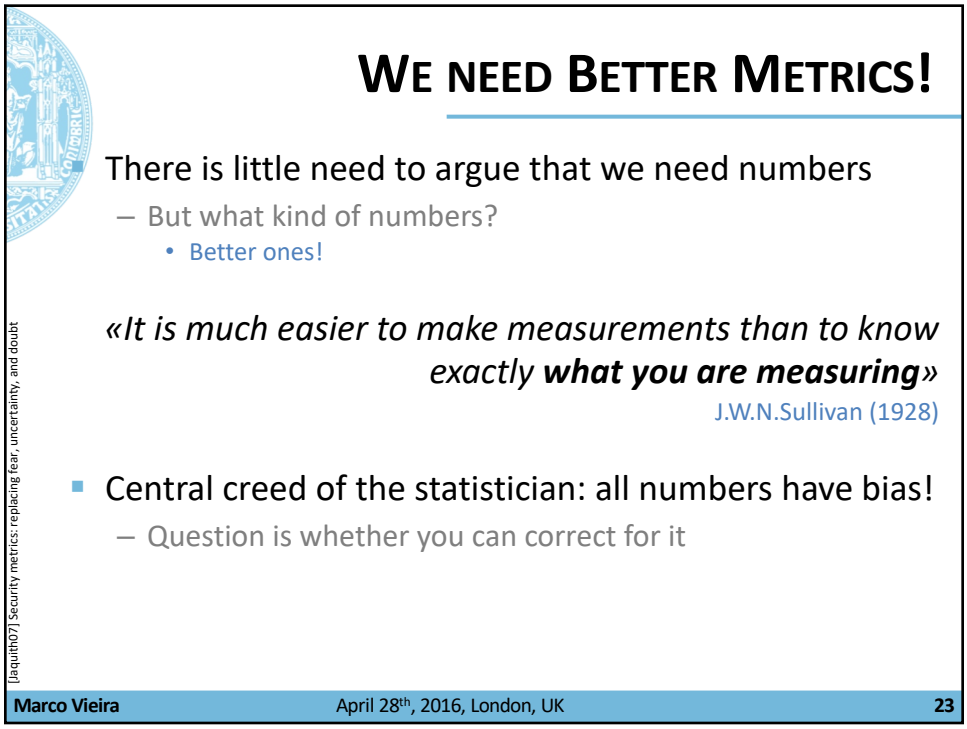
21

AN EXAMPLE (FROM IDS)…

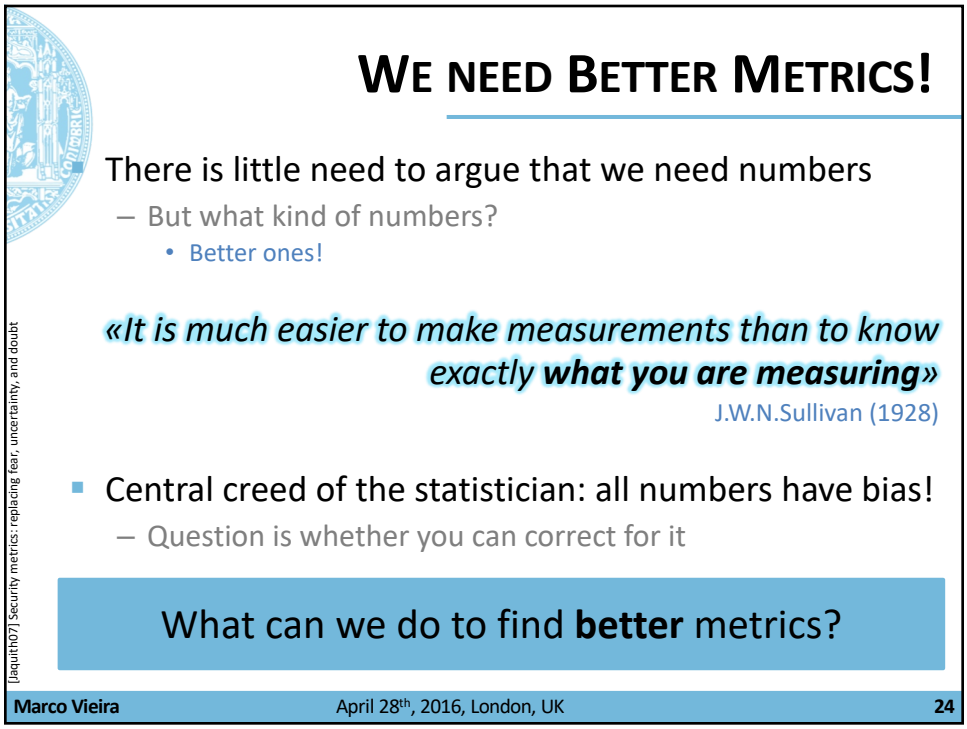
These values are not made up...

Review			Reported				prec	recall	F-M	acc	fpr
P	N	Pop	TP	TN	FN	FP					
120	79	199	118	31	2	48	0,711	0,983	0,825	0,749	0,608
		199	32	58	88	21	0,604	0,267	0,370	0,452	0,266
		199	24	78	96	1	0,960	0,200	0,331	0,513	0,013
		199	24	79	96	0	1,000	0,200	0,333	0,518	0,000
		199	10	79	110	0	1,000	0,083	0,154	0,447	0,000
		199	16	79	104	0	1,000	0,133	0,235	0,477	0,000
##	80	185	30	80	75	0	1,000	0,286	0,444	0,595	0,000
77	122	199	0	122	77	0	#DIV/0!	0,000	#DIV/0!	0,613	0,000
		199	77	122	0	0	1,000	1,000	1,000	1,000	0,000
		199	77	57	0	65	0,542	1,000	0,703	0,673	0,533
62	123	185	0	123	62	0	#DIV/0!	0,000	#DIV/0!	0,665	0,000
		185	0	123	62	0	#DIV/0!	0,000	#DIV/0!	0,665	0,000

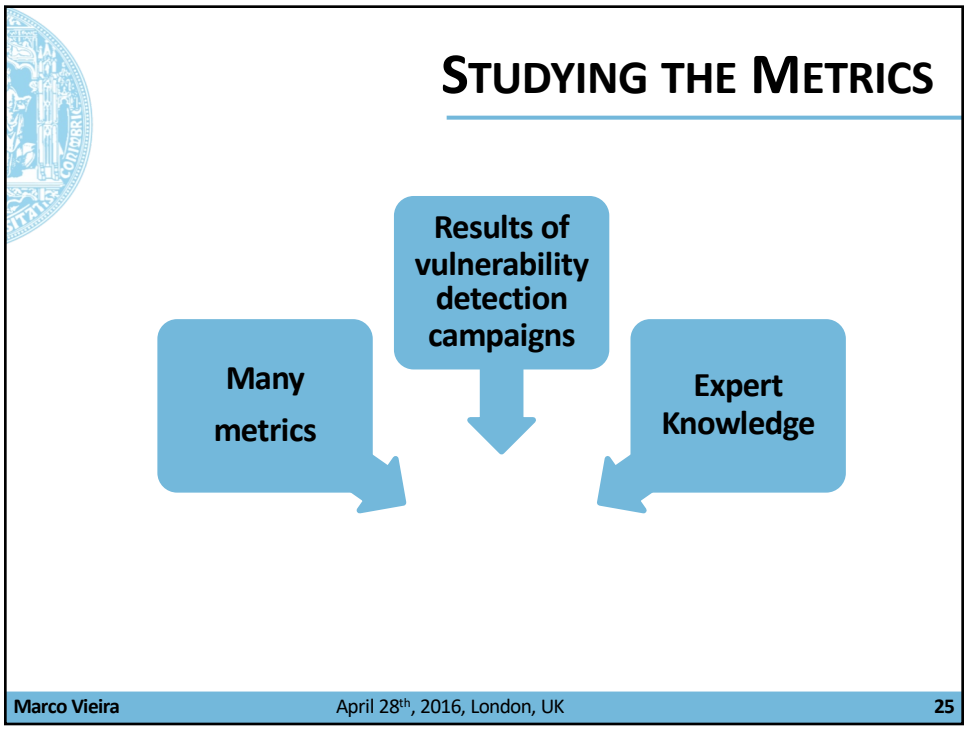
22


WE NEED BETTER METRICS!
There is little need to argue that we need numbers
– But what kind of numbers?
• Better ones!
«It is much easier to make measurements than to know exactly **what you are measuring**»
J.W.N.Sullivan (1928)
Central creed of the statistician: all numbers have bias!
– Question is whether you can correct for it
[Jaquith07] Security metrics: replacing fear, uncertainty, and doubt
Marco Vieira
April 28th, 2016, London, UK
23

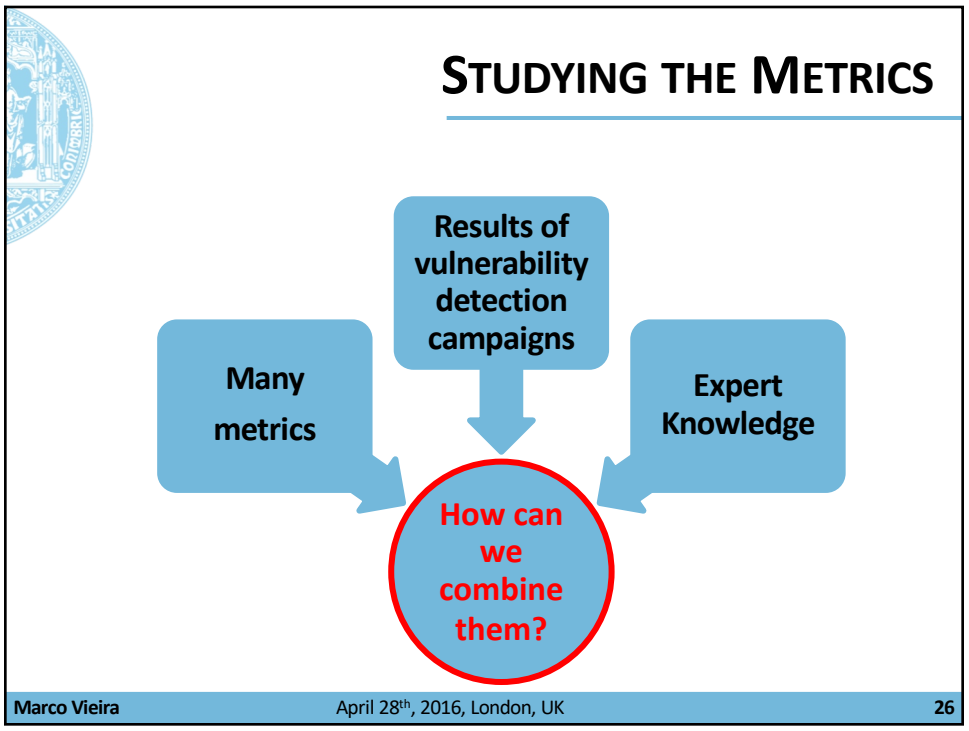
23


WE NEED BETTER METRICS!
There is little need to argue that we need numbers
– But what kind of numbers?
• Better ones!
«It is much easier to make measurements than to know exactly **what you are measuring**»
J.W.N.Sullivan (1928)
Central creed of the statistician: all numbers have bias!
– Question is whether you can correct for it
What can we do to find **better** metrics?
[Jaquith07] Security metrics: replacing fear, uncertainty, and doubt
Marco Vieira
April 28th, 2016, London, UK
24

24


STUDYING THE METRICS
Results of vulnerability detection campaigns
Many metrics
Expert Knowledge
Marco Vieira
April 28th, 2016, London, UK
25

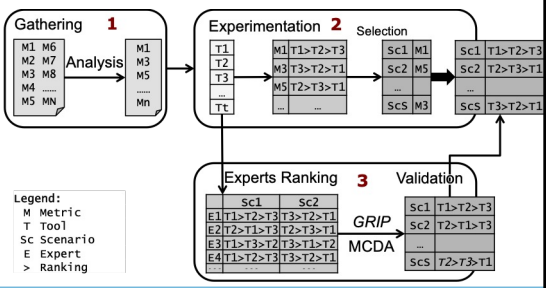
25


STUDYING THE METRICS
Results of vulnerability detection campaigns
Many metrics
Expert Knowledge
How can we combine them?
Marco Vieira
April 28th, 2016, London, UK
26

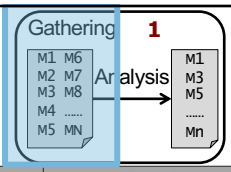
26

APPROACH FOLLOWED

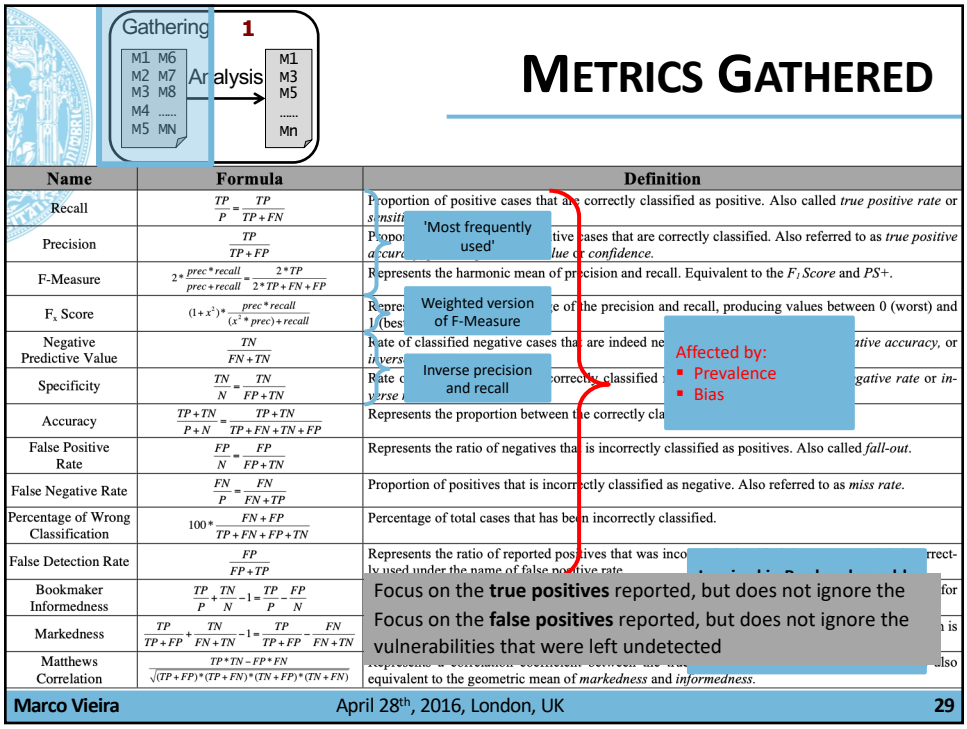
1. Metrics gathering and analysis
   - Gather a large number of metrics from different domains
   - Analyze them according to the characteristics of good metrics
     - In general sense and also in the vulnerability detection context
2. Empirical Analysis and Metrics Selection
   - Use the metrics in a set of specific scenarios using real data
   - Understand their effectiveness and select the best metric for each case
3. Use Expert Knowledge with MCDA to validate the selection
   - Generalized Regression with Intensities of Preference (GRIP)



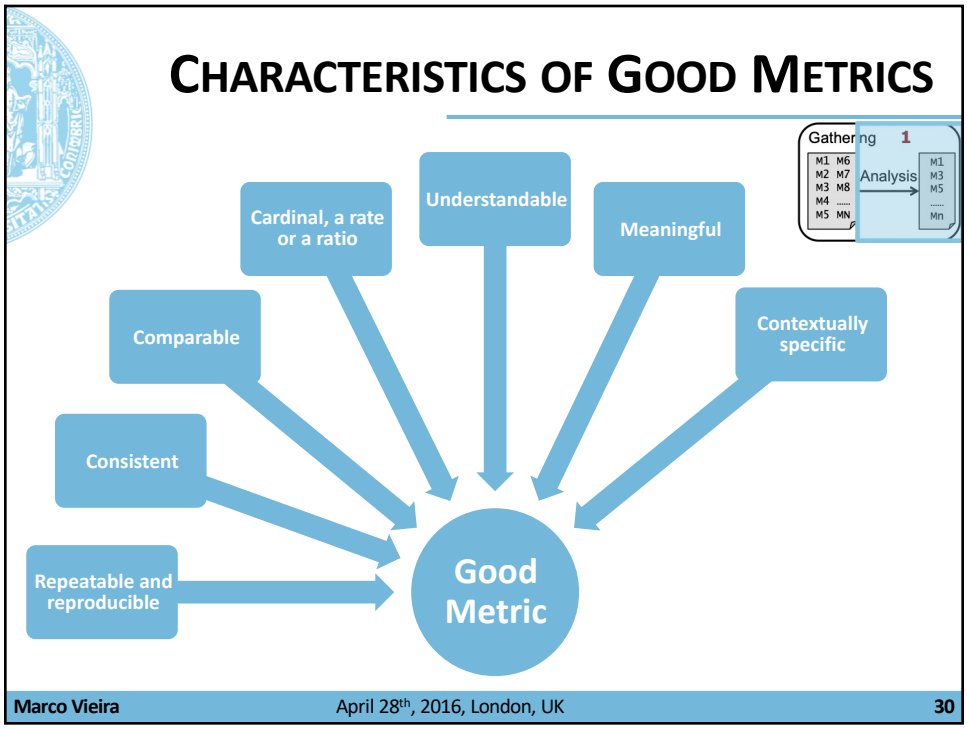
METRICS GATHERED



Name	Formula	Definition
Recall	$\frac{TP}{P}=\frac{TP}{TP+FN}$	Proportion of positive cases that are correctly classified as positive. Also called *true positive rate* or *sensitivity*.
Precision	$\frac{TP}{TP+FP}$	Proportion of the classified positive cases that are correctly classified. Also referred to as *true positive accuracy*, *positive predictive value* or *confidence*.
F-Measure	$2*\frac{prec*recall}{prec+recall}=\frac{2*TP}{2*TP+FN+FP}$	Represents the harmonic mean of precision and recall. Equivalent to the *F_1 Score* and *PS+*.
F_x Score	$(1+x^2)*\frac{prec*recall}{(x^2*prec)+recall}$	Represents the weighted average of the precision and recall, producing values between 0 (worst) and 1 (best).
Negative Predictive Value	$\frac{TN}{FN+TN}$	Rate of classified negative cases that are indeed negatives. Also known as *true negative accuracy,* or *inverse precision*.
Specificity	$\frac{TN}{N}=\frac{TN}{FP+TN}$	Rate of negative cases that is correctly classified negative. Also known as *true negative rate* or *inverse recall*.
Accuracy	$\frac{TP+TN}{P+N}=\frac{TP+TN}{TP+FN+TN+FP}$	Represents the proportion between the correctly classified cases and the total case.
False Positive Rate	$\frac{FP}{N}=\frac{FP}{FP+TN}$	Represents the ratio of negatives that is incorrectly classified as positives. Also called *fall-out*.
False Negative Rate	$\frac{FN}{P}=\frac{FN}{FN+TP}$	Proportion of positives that is incorrectly classified as negative. Also referred to as *miss rate*.
Percentage of Wrong Classification	$100*\frac{FN+FP}{TP+FN+FP+TN}$	Percentage of total cases that has been incorrectly classified.
False Detection Rate	$\frac{FP}{FP+TP}$	Represents the ratio of reported positives that was incorrectly classified. In some cases it is incorrectly used under the name of false positive rate.
Bookmaker Informedness	$\frac{TP}{P}+\frac{TN}{N}-1=\frac{TP}{P}-\frac{FP}{N}$	Quantifies how consistently the predictor predicts the outcome, i.e. how informed a predictor is for the specified condition, versus chance.
Markedness	$\frac{TP}{TP+FP}+\frac{TN}{FN+TN}-1=\frac{TP}{TP+FP}-\frac{FN}{FN+TN}$	Quantifies how consistently the outcome has the predictor as a marker, i.e. how marked a condition is for the specified predictor, versus chance. Equivalent to *DeltaP*.
Matthews Correlation	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$	Represents a correlation coefficient between the true classes and the classified results. It is also equivalent to the geometric mean of *markedness* and *informedness*.

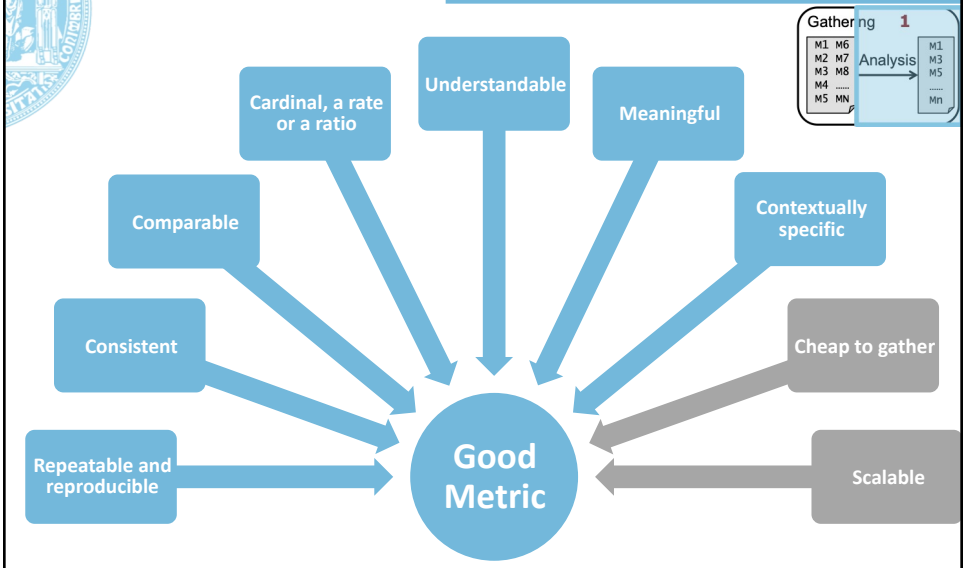

Gathering
1
M1 M6
M2 M7
M3 M8
M4 .....
M5 MN
Analysis
M1
M3
M5
.....
Mn
METRICS GATHERED
Name
Formula
Definition
Recall
Precision
F-Measure
F_x Score
Negative Predictive Value
Specificity
Accuracy
False Positive Rate
False Negative Rate
Percentage of Wrong Classification
False Detection Rate
Bookmaker Informedness
Markedness
Matthews Correlation
'Most frequently used'
Weighted version of F-Measure
Inverse precision and recall
Affected by:
Prevalence
Bias
Represents the harmonic mean of precision and recall. Equivalent to the F1 Score and PS+.
Represents the proportion between the correctly cla
Represents the ratio of negatives that is incorrectly classified as positives. Also called fall-out.
Proportion of positives that is incorrectly classified as negative. Also referred to as miss rate.
Percentage of total cases that has been incorrectly classified.
Focus on the true positives reported, but does not ignore the
Focus on the false positives reported, but does not ignore the
vulnerabilities that were left undetected
equivalent to the geometric mean of markedness and informedness.
Marco Vieira
April 28th, 2016, London, UK
29

29


CHARACTERISTICS OF GOOD METRICS
Gathering
1
Analysis
Cardinal, a rate or a ratio
Understandable
Meaningful
Comparable
Contextually specific
Consistent
Repeatable and reproducible
Good Metric
Marco Vieira
April 28th, 2016, London, UK
30

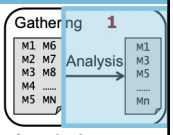
30

Characteristics of Good Metrics



31

Preliminary Analysis of the Metrics



Based on the presented characteristics and the following context-specific characteristics:

Contextually specific

- Ability to detect the maximum real vulnerabilities (TP) while reporting the minimum of false positives (FP)
- Leaving vulnerabilities undetected is in many cases worse than reporting false positives
- Support decision based on a small amount of information
- Comparison of different types of tools

- Filter out metrics that report results of little utility
  - Manual and subjective step
- Reduced the pool of metrics to ~half
  - Allows focusing the following analysis

32

APPLYING THE METRICS TO REAL DATA

Use the metrics with the results of 10 different tools in a benchmark of applications with vulnerabilities

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	recall	F-M	Fx Score 0,5	Fx Score 1,5	acc	fpr	%WC	fdr	Infor	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00	0,79	0,88	0,95	0,85	0,96	0,00	4,02	0,00	0,79	0,95	T01
				T02	48	357	39	4	0,92	0,55	0,69	0,81	0,63	0,90	0,01	9,60	0,08	0,54	0,82	T02
				T03	87	312	0	49	0,64	1,00	0,78	0,69	0,85	0,89	0,14	10,94	0,36	0,86	0,64	T03
				T04	13	334	74	27	0,33	0,15	0,20	0,26	0,18	0,77	0,07	22,54	0,68	0,07	0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00	0,75	0,86	0,94	0,82	0,93	0,00	7,46	0,00	0,75	0,90	T05
				T06	51	304	107	61	0,46	0,32	0,38	0,42	0,35	0,68	0,17	32,12	0,54	0,16	0,20	T06
				T07	38	305	120	60	0,39	0,24	0,30	0,35	0,27	0,66	0,16	34,42	0,61	0,08	0,11	T07
				T08	3	365	155	0	1,00	0,02	0,04	0,09	0,03	0,70	0,00	29,64	0,00	0,02	0,70	T08
				T09	38	336	120	29	0,57	0,24	0,34	0,45	0,29	0,72	0,08	28,49	0,43	0,16	0,30	T09
				T10	117	365	41	0	1,00	0,74	0,85	0,93	0,80	0,92	0,00	7,84	0,00	0,74	0,90	T10

33

APPLYING THE METRICS TO REAL DATA

Use the metrics with the results of 10 different tools in a benchmark of applications with vulnerabilities

- Accuracy and percentage of wrong classification do not seem very useful
  - Too influenced by the TN
  - Are symmetric of each other

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	recall	F-M	Fx Score 0,5	Fx Score 1,5	acc	fpr	%WC	fdr	Infor	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00	0,79	0,88	0,95	0,85	0,96	0,00	4,02	0,00	0,79	0,95	T01
				T02	48	357	39	4	0,92	0,55	0,69	0,81	0,63	0,90	0,01	9,60	0,08	0,54	0,82	T02
				T03	87	312	0	49	0,64	1,00	0,78	0,69	0,85	0,89	0,14	10,94	0,36	0,86	0,64	T03
				T04	13	334	74	27	0,33	0,15	0,20	0,26	0,18	0,77	0,07	22,54	0,68	0,07	0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00	0,75	0,86	0,94	0,82	0,93	0,00	7,46	0,00	0,75	0,90	T05
				T06	51	304	107	61	0,46	0,32	0,38	0,42	0,35	0,68	0,17	32,12	0,54	0,16	0,20	T06
				T07	38	305	120	60	0,39	0,24	0,30	0,35	0,27	0,66	0,16	34,42	0,61	0,08	0,11	T07
				T08	3	365	155	0	1,00	0,02	0,04	0,09	0,03	0,70	0,00	29,64	0,00	0,02	0,70	T08
				T09	38	336	120	29	0,57	0,24	0,34	0,45	0,29	0,72	0,08	28,49	0,43	0,16	0,30	T09
				T10	117	365	41	0	1,00	0,74	0,85	0,93	0,80	0,92	0,00	7,84	0,00	0,74	0,90	T10

34

APPLYING THE METRICS TO REAL DATA

Use the metrics with the results of 10 different tools in a benchmark of applications with vulnerabilities

- *fdr*, *precision* and *markedness* seem effective too characterize the ability to avoid false positives

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	recall	F-M	Fx Score 0,5	Fx Score 1,5	acc	fpr	%WC	fdr	Infor	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00	0,79	0,88	0,95	0,85	0,96	0,00	4,02	0,00	0,79	0,95	T01
				T02	48	357	39	4	0,92	0,55	0,69	0,81	0,63	0,90	0,01	9,60	0,08	0,54	0,82	T02
				T03	87	312	0	49	0,64	1,00	0,78	0,69	0,85	0,89	0,14	10,94	0,36	0,86	0,64	T03
				T04	13	334	74	27	0,33	0,15	0,20	0,26	0,18	0,77	0,07	22,54	0,68	0,07	0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00	0,75	0,86	0,94	0,82	0,93	0,00	7,46	0,00	0,75	0,90	T05
				T06	51	304	107	61	0,46	0,32	0,38	0,42	0,35	0,68	0,17	32,12	0,54	0,16	0,20	T06
				T07	38	305	120	60	0,39	0,24	0,30	0,35	0,27	0,66	0,16	34,42	0,61	0,08	0,11	T07
				T08	3	365	155	0	1,00	0,02	0,04	0,09	0,03	0,70	0,00	29,64	0,00	0,02	0,70	T08
				T09	38	336	120	29	0,57	0,24	0,34	0,45	0,29	0,72	0,08	28,49	0,43	0,16	0,30	T09
				T10	117	365	41	0	1,00	0,74	0,85	0,93	0,80	0,92	0,00	7,84	0,00	0,74	0,90	T10

35

APPLYING THE METRICS TO REAL DATA

Use the metrics with the results of 10 different tools in a benchmark of applications with vulnerabilities

- *recall* and *informedness* seem effective to characterize the ability to detect vulnerabilities

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	recall	F-M	Fx Score 0,5	Fx Score 1,5	acc	fpr	%WC	fdr	Infor	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00	0,79	0,88	0,95	0,85	0,96	0,00	4,02	0,00	0,79	0,95	T01
				T02	48	357	39	4	0,92	0,55	0,69	0,81	0,63	0,90	0,01	9,60	0,08	0,54	0,82	T02
				T03	87	312	0	49	0,64	1,00	0,78	0,69	0,85	0,89	0,14	10,94	0,36	0,86	0,64	T03
				T04	13	334	74	27	0,33	0,15	0,20	0,26	0,18	0,77	0,07	22,54	0,68	0,07	0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00	0,75	0,86	0,94	0,82	0,93	0,00	7,46	0,00	0,75	0,90	T05
				T06	51	304	107	61	0,46	0,32	0,38	0,42	0,35	0,68	0,17	32,12	0,54	0,16	0,20	T06
				T07	38	305	120	60	0,39	0,24	0,30	0,35	0,27	0,66	0,16	34,42	0,61	0,08	0,11	T07
				T08	3	365	155	0	1,00	0,02	0,04	0,09	0,03	0,70	0,00	29,64	0,00	0,02	0,70	T08
				T09	38	336	120	29	0,57	0,24	0,34	0,45	0,29	0,72	0,08	28,49	0,43	0,16	0,30	T09
				T10	117	365	41	0	1,00	0,74	0,85	0,93	0,80	0,92	0,00	7,84	0,00	0,74	0,90	T10

36

APPLYING THE METRICS TO REAL DATA

Use the metrics with the results of 10 different tools in a benchmark of applications with vulnerabilities

- *F-Measure or F_1 Score is balanced*
- $F_{0.5}$ favors *precision*
- $F_{1.5}$ favors *recall*

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	recall	F-M	Fx Score 0,5	Fx Score 1,5	acc	fpr	%WC	fdr	Infor	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00	0,79	0,88	0,95	0,85	0,96	0,00	4,02	0,00	0,79	0,95	T01
				T02	48	357	39	4	0,92	0,55	0,69	0,81	0,63	0,90	0,01	9,60	0,08	0,54	0,82	T02
				T03	87	312	0	49	0,64	1,00	0,78	0,69	0,85	0,89	0,14	10,94	0,36	0,86	0,64	T03
				T04	13	334	74	27	0,33	0,15	0,20	0,26	0,18	0,77	0,07	22,54	0,68	0,07	0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00	0,75	0,86	0,94	0,82	0,93	0,00	7,46	0,00	0,75	0,90	T05
				T06	51	304	107	61	0,46	0,32	0,38	0,42	0,35	0,68	0,17	32,12	0,54	0,16	0,20	T06
				T07	38	305	120	60	0,39	0,24	0,30	0,35	0,27	0,66	0,16	34,42	0,61	0,08	0,11	T07
				T08	3	365	155	0	1,00	0,02	0,04	0,09	0,03	0,70	0,00	29,64	0,00	0,02	0,70	T08
				T09	38	336	120	29	0,57	0,24	0,34	0,45	0,29	0,72	0,08	28,49	0,43	0,16	0,30	T09
				T10	117	365	41	0	1,00	0,74	0,85	0,93	0,80	0,92	0,00	7,84	0,00	0,74	0,90	T10

ONE SIZE DOES NOT FIT ALL!

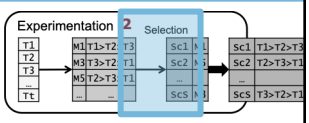
Different ***organizations or teams*** have **different goals**

- That are valid under **different assumptions!**

- A single set of metrics may not be enough
  - But it is unfeasible to repeat this analysis in every different situation or organization

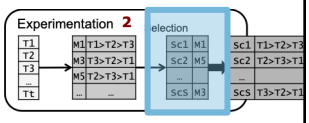
- We need to define ***typical detection*** scenarios
  - Representative of situations where vulnerability detection tools are used *and should be benchmarked* under **different assumptions** and with **different goals**
  - Organizations or teams may later look at the scenario that is closer to their own needs

DEFINED DETECTION SCENARIOS



1. Business-critical applications
   - Detect the highest number of vulnerabilities
2. Heightened-critical applications
   - Detect the highest number of vulnerabilities while avoiding *too many* false positives
3. Best effort
   - Detect a high number of vulnerabilities while reporting a low number of false positives
4. Minimum effort
   - Report the lowest number of false positives

SELECTING METRICS



- Considering the data presented
- We selected the metrics that seem most adequate to each scenario
  - Including the main recommended metric and a tiebreaker

Obvious!

> F_x Score ($x > 1$)

Good balance

> precision

Scenario	Recommended metric	Recommended Tiebreaker
1 Business-critical	recall	precision
2 Non-critical	informedness	recall
3 Best effort	F-Measure	recall
4 Minimum effort	Markedness	precision

CLOSE LOOK: SCENARIO 4

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	fpr	fdr	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00			0,95	T01
				T02	48	357	39	4	0,92			0,82	T02
				T03	87	312	0	49	0,64			0,64	T03
				T04	13	334	74	27	0,33			0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00			0,90	T05
				T06	51	304	107	61	0,46			0,20	T06
				T07	38	305	120	60	0,39			0,11	T07
				T08	3	365	155	0	1,00			0,70	T08
				T09	38	336	120	29	0,57			0,30	T09
				T10	117	365	41	0	1,00			0,90	T10

41

CLOSE LOOK: SCENARIO 4

Type	Review P	Review N	Review Pop	Tool	Reported TP	Reported TN	Reported FN	Reported FP	prec	fpr	fdr	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00			0,95	T01
				T02	48	357	39	4	0,92			0,82	T02
				T03	87	312	0	49	0,64			0,64	T03
				T04	13	334	74	27	0,33			0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00			0,90	T05
				T06	51	304	107	61	0,46			0,20	T06
				T07	38	305	120	60	0,39			0,11	T07
				T08	3	365	155	0	1,00			0,70	T08
				T09	38	336	120	29	0,57			0,30	T09
				T10	117	365	41	0	1,00			0,90	T10

42

CLOSE LOOK: SCENARIO 4

Balanced!

Type	P	N	Pop	Tool	TP	TN	FN	FP	prec	fpr	fdr	Mark	Tool
Lines	87	361	448	T01	69	361	18	0	1,00			0,95	T01
				T02	48	357	39	4	0,92			0,82	T02
				T03	87	312	0	49	0,64			0,64	T03
				T04	13	334	74	27	0,33			0,14	T04
Inputs	158	365	523	T05	119	365	39	0	1,00			0,90	T05
				T06	51	304	107	61	0,46			0,20	T06
				T07	38	305	120	60	0,39			0,11	T07
				T08	3	365	155	0	1,00			0,70	T08
				T09	38	336	120	29	0,57			0,30	T09
				T10	117	365	41	0	1,00			0,90	T10

VALIDATION WITH EXPERT KNOWLEDGE

Multiple-criteria Decision Analysis (MCDA)

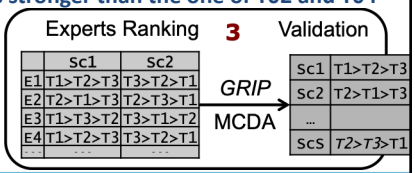
– *Generalized Regression with Intensities of Preference* (GRIP)

- Requires knowledge of multiple experts, translated into a set of preferences between the a
- A decision maker inputs *"intensities of*
  - Define the relation between pairs of prefe *stronger than*, *similar to*, or *stronger than*
    - E.g. (T01, T04) > (T02, T04)
      - **relation between T01 and T04 is stronger than the one of T02 and T04**

Only for a subset of the alternatives, and the algorithm extrapolates for the remaining…

…otherwise we wouldn't need the algorithm ☺

- Used the *D2* tool
  - http://www.decision-deck.org/

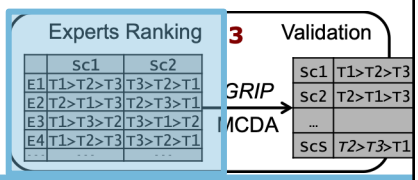


EXPERT QUESTIONNAIRES

6 external researchers

– With no information about the objective of the experiment

- In order not to influence the opinions, each questionnaire contained only:
  - Description of the scenario
  - The raw measurements for each tool
    - *P, N, TP, TN, FP, FN, | tpr* and *fpr* (normalized versions of TP and FP)
- Were asked to rank the tools for each scenario based on their opinion on ***how adequate they thought each tool would be***



45

EXPERT QUESTIONNAIRES

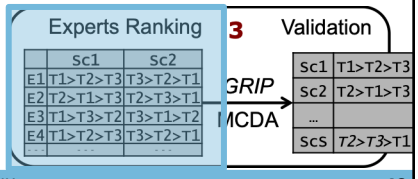
6 external researche

– With no informatio

- In order not to influ
  questionnaire conta
  - Description of the scenario
  - The raw measurements for each tool
    - *P, N, TP, TN, FP, FN, | tpr* and *fpr* (normalized versions of TP and FP)
- Were asked to rank the tools for each scenario based on their opinion on ***how adequate they thought each tool would be***

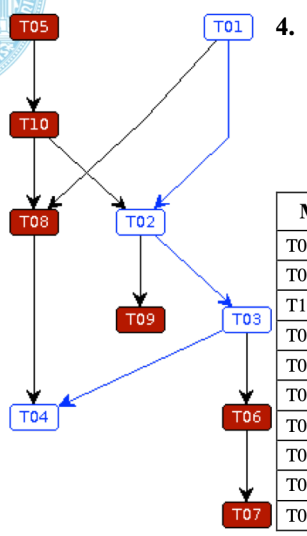
Tool	Review (true class)			Reported				Basic Metrics	
	P	N	Total	TP	TN	FN	FP	TPR	FPR
T01	87	361	448	69	361	18	0	0,793	0,000
T02	87	361	448	48	357	39	4	0,552	0,011
T03	87	361	448	87	312	0	49	1,000	0,136
T04	87	361	448	13	334	74	27	0,149	0,075
T05	158	365	523	119	365	39	0	0,753	0,000
T06	158	365	523	51	304	107	61	0,323	0,167
T07	158	365	523	38	305	120	60	0,241	0,164
T08	158	365	523	3	365	155	0	0,019	0,000
T09	158	365	523	38	336	120	29	0,241	0,079
T10	158	365	523	117	365	41	0	0,741	0,000

No other metric!!!



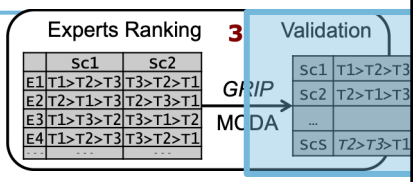
46

GRIP RESULTS: SCENARIO 4



4.

Mark.	
T01	0.95
T05	0.90
T10	0.90
T02	0.82
T08	0.70
T03	0.64
T09	0.30
T06	0.20
T04	0.14
T07	0.11



- We input the preferences for tools T01-T04 (blue boxes)
  - GRIP computed the necessary ranks among the alternatives
    - Arrows mean relationships of sequence
    - Position of each tool in the graphs provides no indication
- Selected for Sc4: *Markedness*

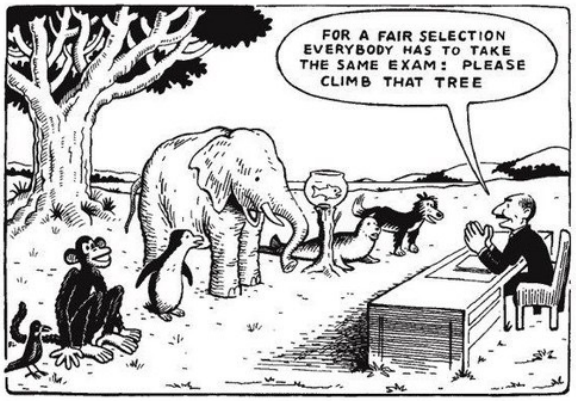
CONCLUSIONS

We presented an approach to select the metrics to use in benchmarks for vulnerability detection tools:

- *Extensive review of the existing metrics*
- An empirical analysis of the results of the metrics
- MCDA algorithms together with expert knowledge

- Nowadays metrics selection process is *inadequate*
  - A single set of metrics does not fit every scenario
    - Often a metric is good for one scenario and **misleading** in another
  - Results revealed that some of the traditional metrics are effective in some vulnerability detection contexts…
    - … in other cases, it is a much better to use alternative metrics: *informedness* and *markedness* may be very effective
      - **Although they are usually not used in benchmarking contexts**

QUESTIONS?



Nuno Antunes, Marco Vieira
Department of Informatics Engineering
University of Coimbra
nmsa@dei.uc.pt | http://eden.dei.uc.pt/~nmsa
mvieira@dei.uc.pt | http://eden.dei.uc.pt/~mvieira

REFERENCES

[Antunes15] N. Antunes and M. Vieira, *"Assessing and Comparing Vulnerability Detection Tools for Web Services: Benchmarking Approach and Examples"*, IEEE Transactions on Services Computing, vol. 8, no. 2, pp. 269–283, 2015.

[Powers11] D. M. Powers, *"Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation,"* Dec. 2011.

[Jaquith07] A. Jaquith, *Security metrics: replacing fear, uncertainty, and doubt*. Upper Saddle River, NJ: Addison-Wesley, 2007.