Percentages, Probabilities and Professions of Performance

Jim Alves-Foss

University of Idaho

jimaf@uidaho.edu

How we Report Experimental Cybersecurity Results

- Several papers in cybersecurity compare the results of some tool or algorithm to:
  - Other tools/algorithms
  - Across different datasets

- These papers are written to show successful results, but often with limited information and poor statistics.

Show some percentages…

Table 1: Precision Recall of function start identification (reproduction of Table 2 from Bao et al. [3])

	GCC			ICC		
	Precision	Recall	Time(sec)	Precision	Recall	Time(sec)
Rosenblum et al. [7]	0.4909	0.4312	1172.41	0.6080	0.6749	2178.14
BYTEWEIGHT (3)	0.9103	0.8711	1417.51	0.8948	0.8592	1905.34
BYTEWEIGHT (no-norm)	0.9877	0.9302	19994.18	0.9727	0.9132	20894.45
BYTEWEIGHT	0.9726	0.9599	1468.75	0.9725	0.9800	1927.90

Table 2: Precision Recall of function start identification (reproduction of Table1(a) from Andriesse et al. [1])

	gcc x86	gcc x64	clang x86	clang x64	VS x86	VS x64
IDA Pro 6.7	0.98/0.78	0.97/0.74	0.98/0.78	0.98/0.77	0.84/0.93	1.00/0.94
BAP/ByteWeight 0.9.9	0.68/0.83	0.70/0.66	0.52/0.71	0.73/0.49	0.63/0.74	0.69/0.56
Dyninst 9.1.0	0.93/0.91	0.96/0.74	0.98/0.95	0.88/0.72	–	–
Nucleus	0.98/0.96	0.98/0.96	0.96/0.97	0.96/0.95	0.86/0.96	0.95/0.94
△Nucleus	+0.00/+0.05	+0.01/+0.22	−0.02/ +0.02	−0.02/+0.18	+0.02/+0.03	−0.05/+0.00

What does *mean* mean?

- You use the same dataset for all tools.
- Your numbers are bigger than their numbers.

Win !!

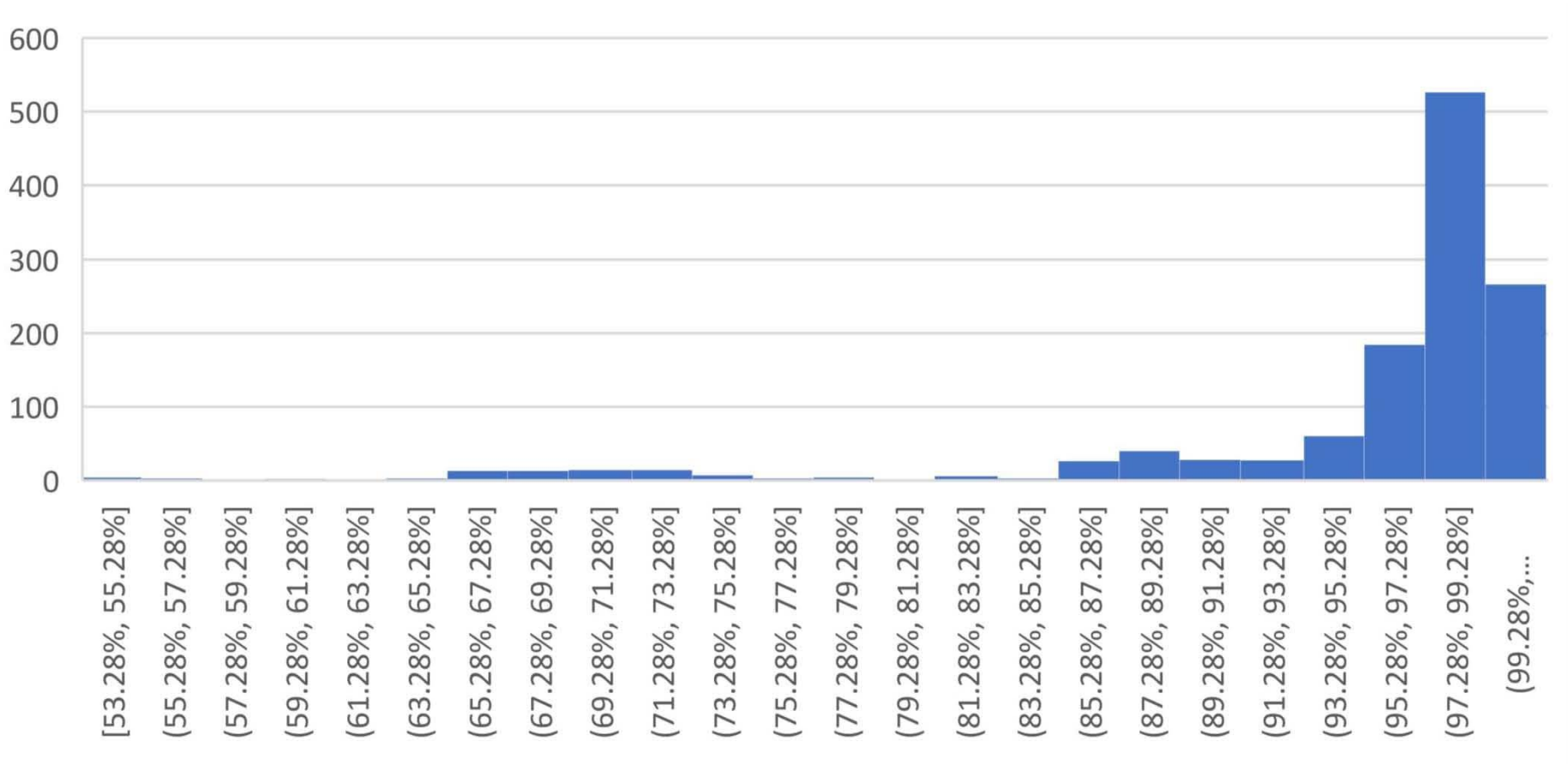
- Are you always better?
  - If not, where and why?
- Are their a few instances where you are a lot better and these skew the means?

What is the nature of your results?

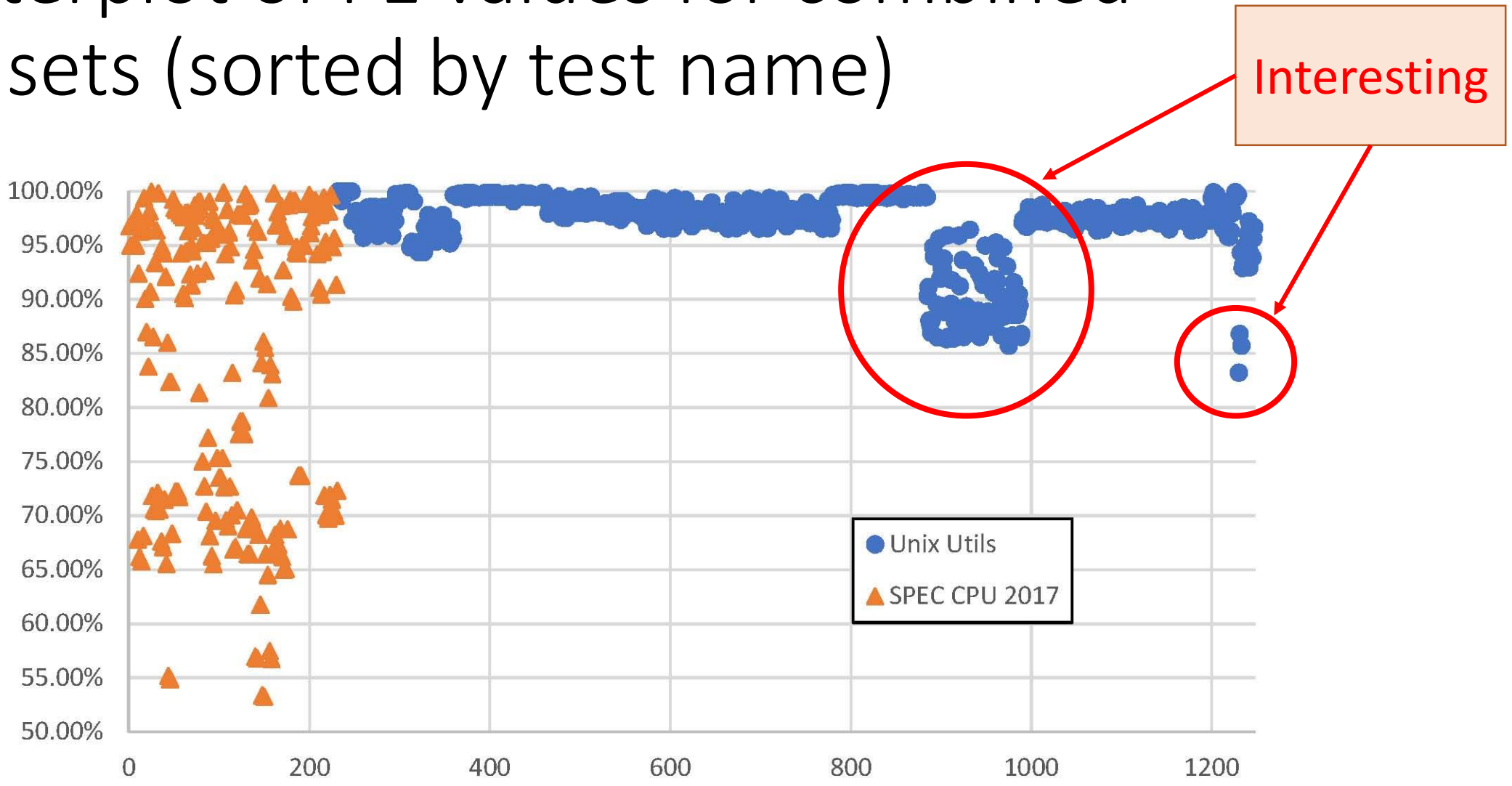
- A mean (an average) is just a single value.
- You can do better than that.

Visualize the Data

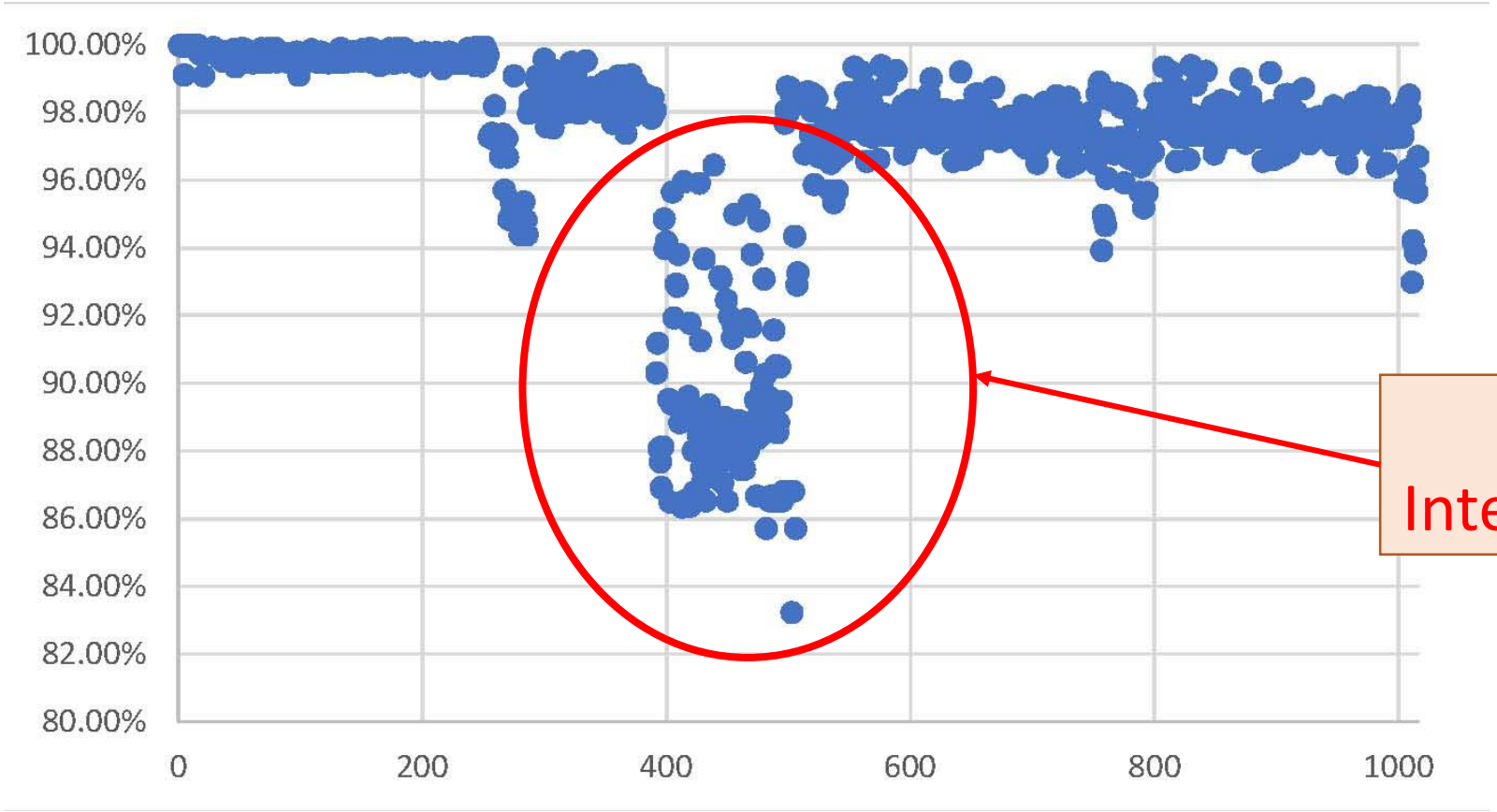
Histogram of F1 values for combined datasets



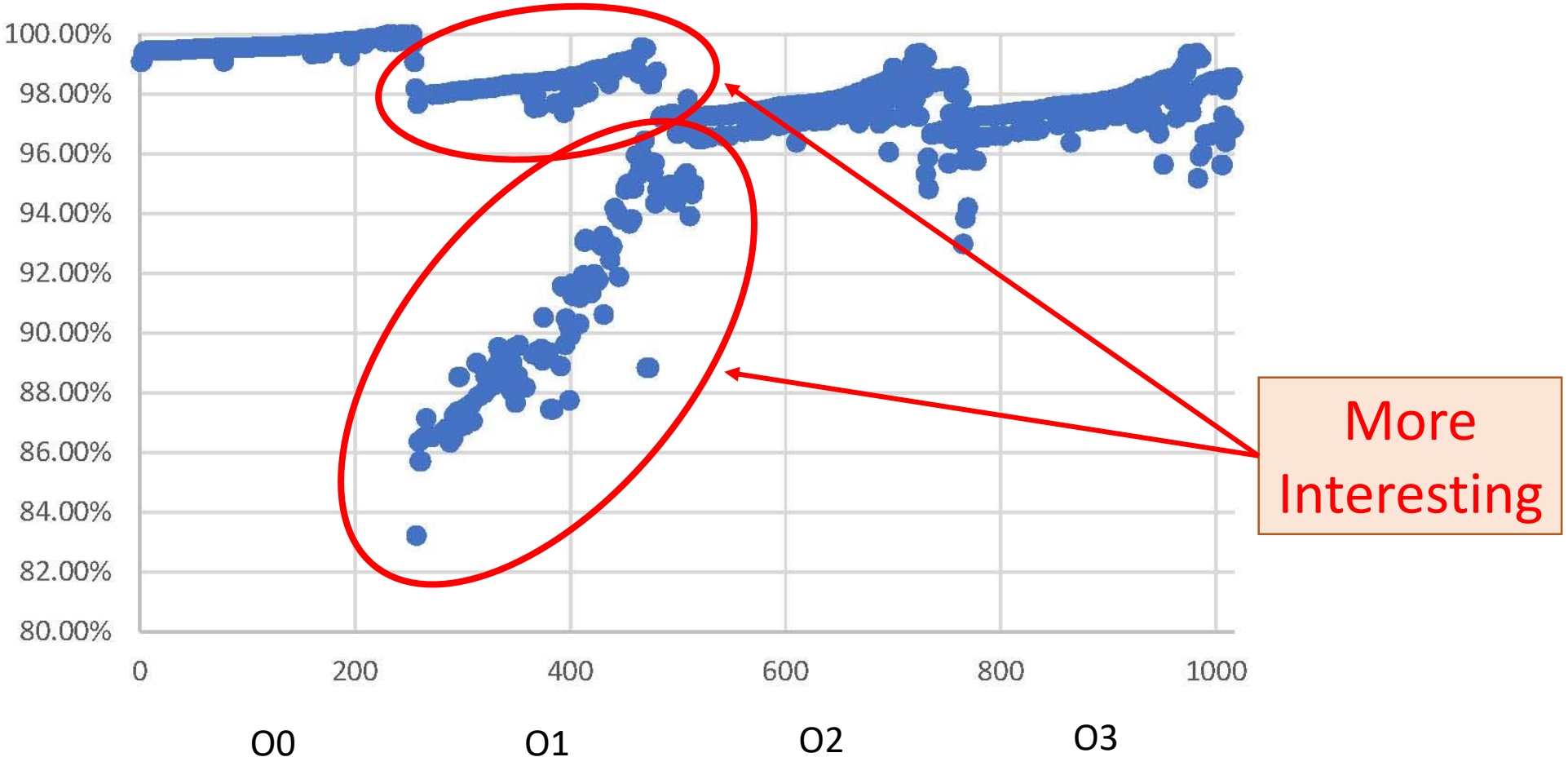
Scatterplot of F1 values for combined datasets (sorted by test name)



Scatterplot for Unix utilities dataset, (grouped by optimization levels and sorted by name)



Scatterplot for Unix utilities dataset, (grouped by optimization levels and sorted by filesize)



Statistical Recommendations

- Distribution Assumptions
  - Make sure you use an appropriate statistical test, many results are not “normal”. A sign test may work best for comparing results of different tools.
- Power
  - Make sure you have enough tests to have valid statistical power
    - Say you want to say Tool A is 3% better than Tool B (what is the confidence interval?):
    - 3% +/- 1% with 95% confidence looks good to me
    - Requires over 9,000 test cases out of a population of 1 Million
- Randomness
  - The above power results assume your samples are randomly selected across the whole population. Not often the case.

How Accurate is Accuracy?

	Condition Positive	Condition Negative
Predicted Positive	True Positive	False Positive (Type 1 error)
Predicted Negative	False Negative (Type 2 error)	True Negative

Confusion Matrix

Some definition from confusion matrix

$$Prevalence = \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total Population}}$$

$$Accuracy = \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total Population}}$$

$$Precision = \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted positive}}$$

$$Recall = \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Consider the Following Ground Truth (rare instance of condition we are looking for)

	Condition Positive	Condition Negative
Predicted Positive	10	
Predicted Negative		990

99% Accurate

	Condition Positive	Condition Negative
Predicted Positive		
Predicted Negative	10	990

You found nothing, and are 99% accurate, but not very useful

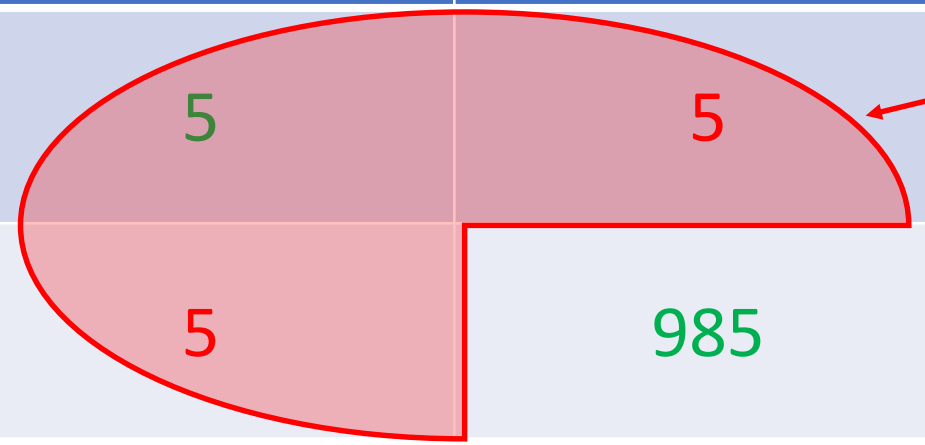
Precision, Recall and F1 value

	Condition Positive	Condition Negative
Predicted Positive		
Predicted Negative	10	990

Precision is 0%, Recall is 0%, F1 is 0%

Precision, Recall and F1 value (New example)

	Condition Positive	Condition Negative
Predicted Positive	5	5
Predicted Negative	5	985



These areas are relevant

Precision is 50%, Recall is 50%, F1 is 50%
(But I am still 99% accurate!)

Conclusions

- If data will have small prevalence (low percentage of true conditions versus false conditions), ignore the true negatives
- Look at the data, visualize it
- Use large random data sets
- Use appropriate statistics (such as paired statistical test, eg. *sign-test)*
- Review the statistics of your own publications
- Consult statistician
- Ask review committees to hold authors accountable on presenting results.

- *“If we shadows have offended,*
Think but this, and all is mended,
That you have but slumbered here
While these visions did appear.
And this weak and idle theme,
No more yielding but a dream,
Gentles, do not reprehend:
If you pardon, we will mend..”

-*Puck in* William Shakespeare,

A Midsummer Night's Dream