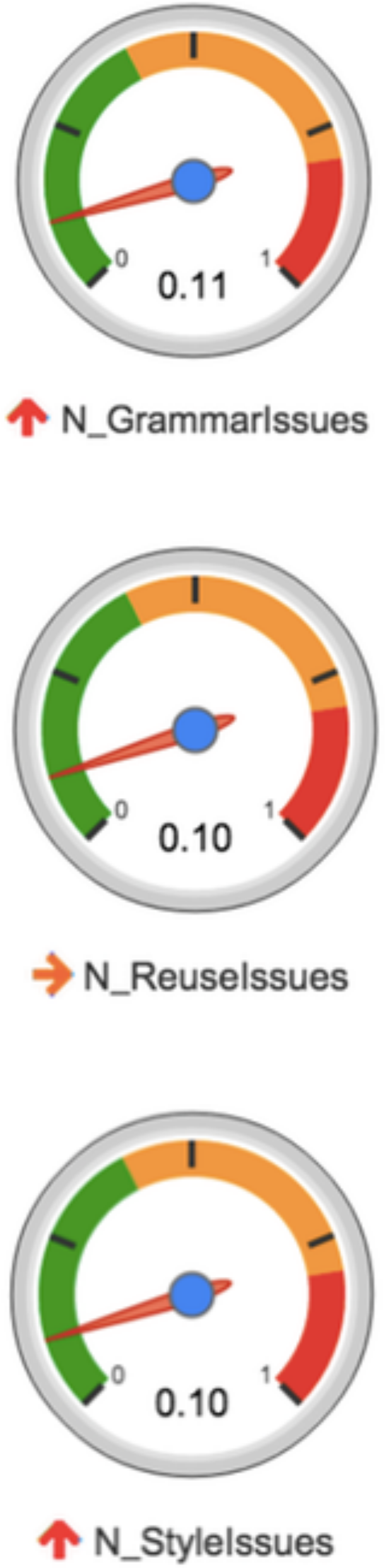
Probabilistic Quality Models to Improve Communication and Actionability

Morgan Ericsson[1], Welf Löwe, Anna Wingkvist
Linnaeus University, Sweden

[1] email: *morgan.ericsson@lnu.se*, twitter: *@morganericsson*

Our Data/Analytics Bucket

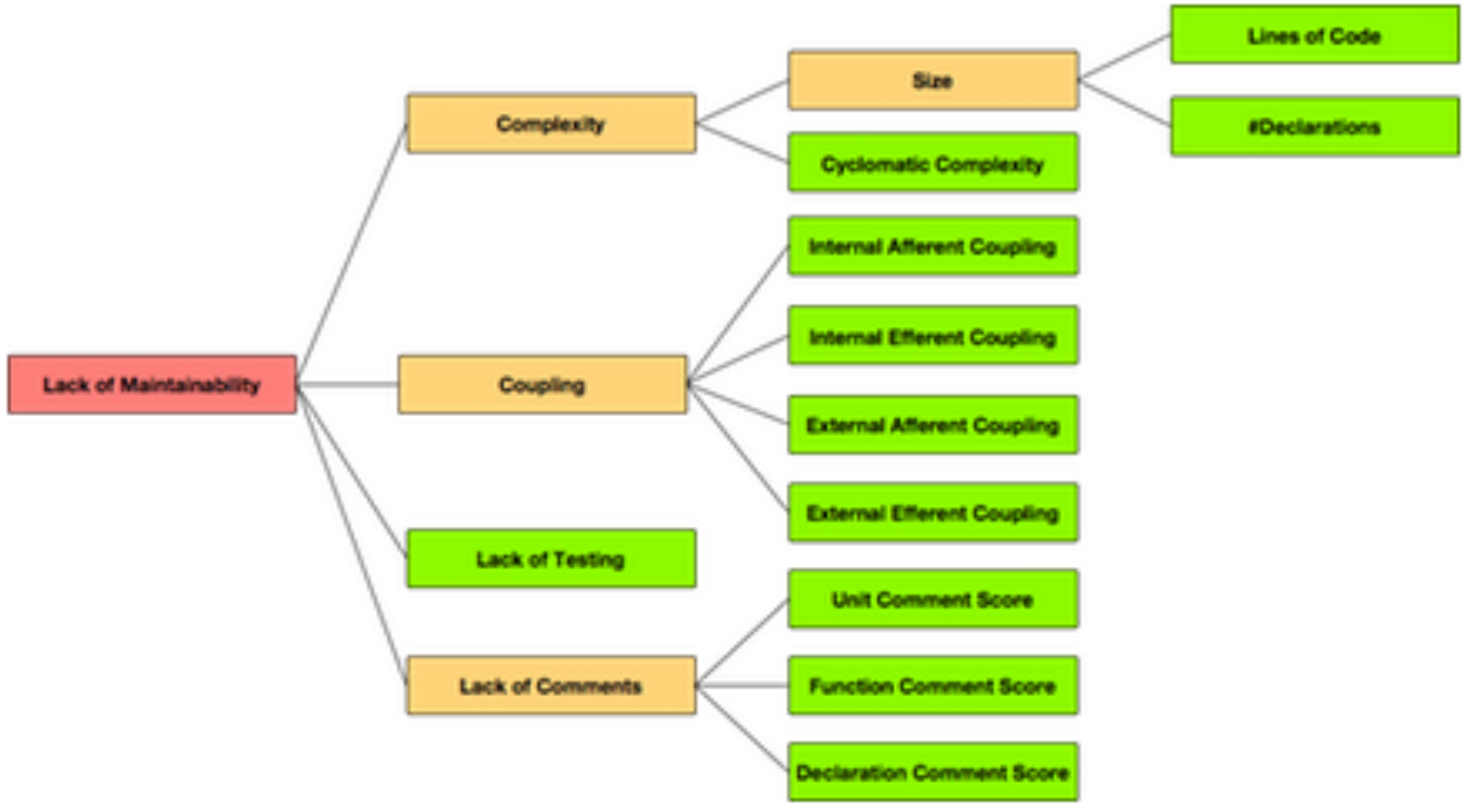
- Source code metrics
- Documentation metrics
- Fail logs and debug data
- Documentation usage information
- Developer and project metrics
- …


0
1
0.11
N_GrammarIssues
0
1
0.10
N_ReuseIssues
0
1
0.10
N_StyleIssues


0
0.11
1


Cyclomatic Complexity


0
0.11
1


Lack of Maintainability
Complexity
Coupling
Lack of Testing
Lack of Comments
Size
Cyclomatic Complexity
Internal Afferent Coupling
Internal Efferent Coupling
External Afferent Coupling
External Efferent Coupling
Unit Comment Score
Function Comment Score
Declaration Comment Score
Lines of Code
#Declarations

Purpose is to **describe**, not predict!

Models are **handcrafted** by domain experts.

Problems

- Data from different sources, different scales, different (coverage) of the population.
- Data is collected in different ways, e.g., collected from users/execution, collected during project/development, measured on artifacts.

Difficult to aggregate (correctly)

Our previous efforts include

- Assume everything is normally distributed
- Use non-parametric methods
- Estimate distributions for all data
- …

These do **not match users expectations**, do **not provide enough information**, or are **difficult to communicate**. (Any maybe **false**…)

Probabilities

(i.e., *empirical distributions*)

Gauges become:

Given all that we know

how likely is X?

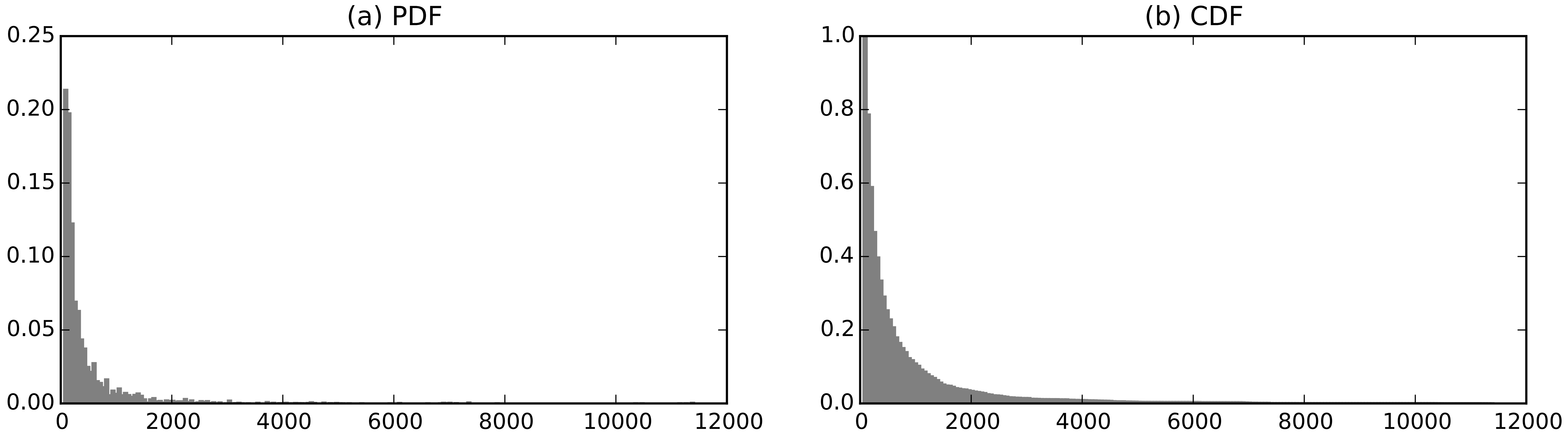
(e.g., X = a file of 1k LoC)

Data aggregation

We compute the empirical distribution PDF and CDF for each *metric* from a data set

- If values are discrete (much of our data is), we can either use a histogram and approxiate from buckets or continuity correct.

In our data sets, we reshape data so larger values are worse. To reflect this, we use $1 - CDF$ rather than the CDF.


(a) PDF
0.25
0.20
0.15
0.10
0.05
0.00
0
2000
4000
6000
8000
10000
12000
(b) CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
12000

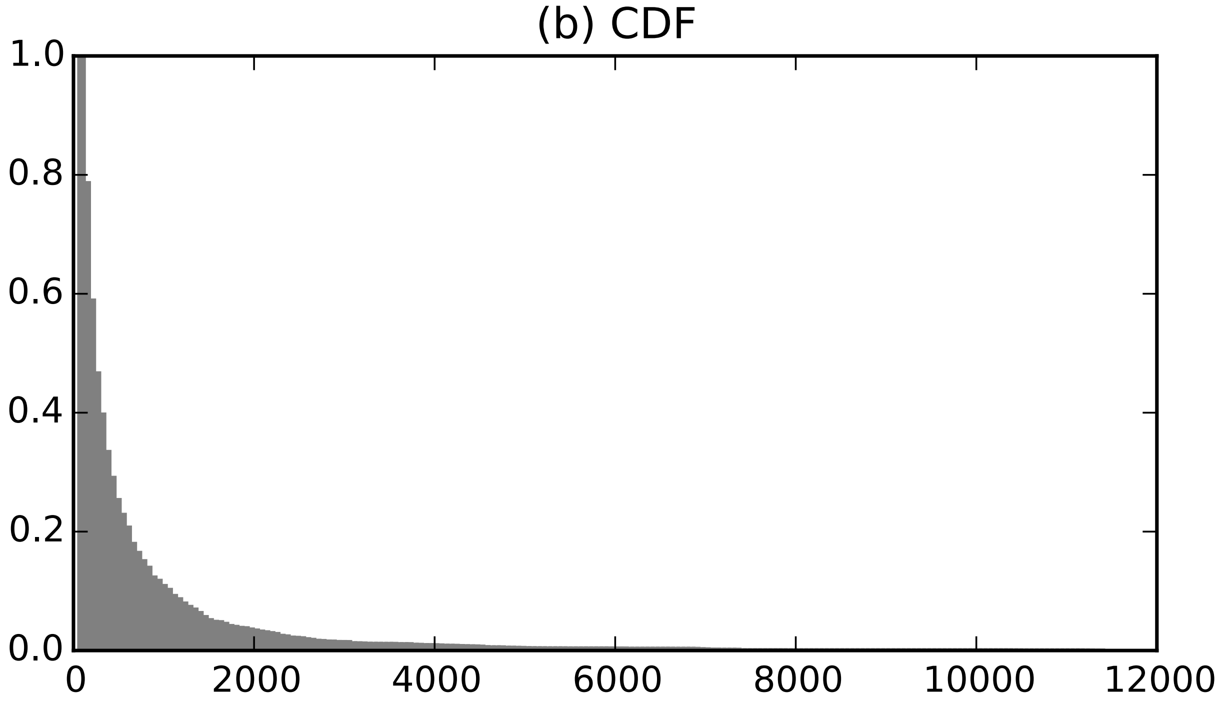
Data aggregation

We can now determine the likelihood of getting this exact value, or this or a larger value. We use the latter to define the probability, P of a value, e.g., $P(LoC)$.

We define indirect values in our model as the joint probability distribution for the direct and indirect metrics it aggregates, e.g.,

$$P(Size) = P(LoC)P(NID)$$

LOC	NID	P(LOC)	P(NID)	P(Size)
23	0	1.000000	1.000000	1.000000
67	4	0.786563	0.197914	0.155672
739	0	0.150740	1.000000	0.150740
303	12	0.334465	0.052874	0.017685
7286	284	0.001819	0.000728	0.000001

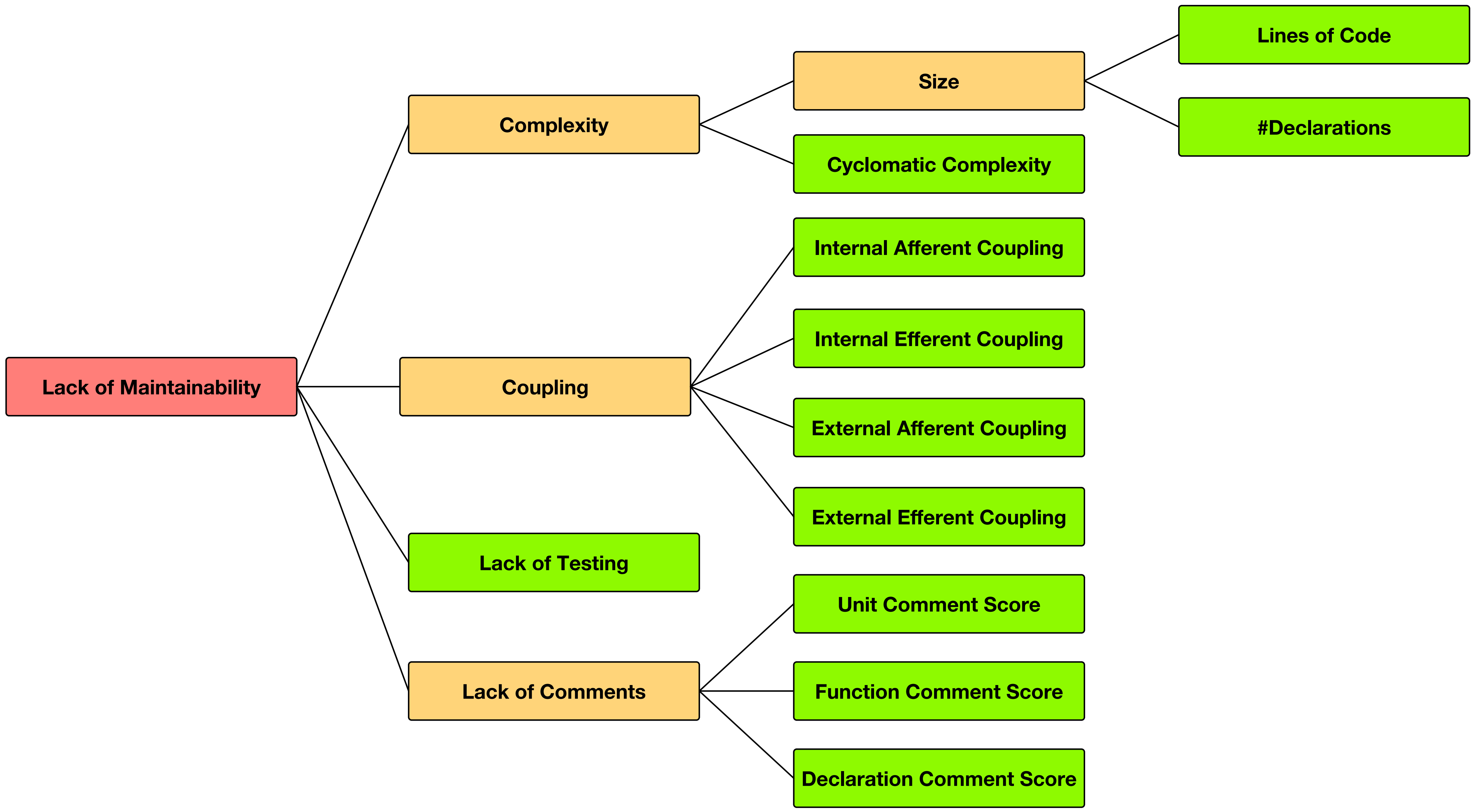


Example

Data and (part of the) model from a large mobile hydraulics companys **integrated development tool**. Their tool is used to develop **hydraulics control software** in a **domain-specific language**.

Data is from **1 year** and contains 8 minor versions.

The tool is implemented in **Delphi** and consists of 110 Delphi Units. A single Unit contains 1 to 266 files. The number of files vary between the 8 minor versions, but each has about 1,000 files.


Lack of Maintainability
Complexity
Coupling
Lack of Testing
Lack of Comments
Size
Cyclomatic Complexity
Internal Afferent Coupling
Internal Efferent Coupling
External Afferent Coupling
External Efferent Coupling
Unit Comment Score
Function Comment Score
Declaration Comment Score
Lines of Code
#Declarations

Position	Name	$P(LackOfMaintainability)$	Pos. Normalize	Pos. BestFit
1	P1DiagDiagnosticNavigatorUnit.pas	2.376713e-15	688	52
2	CgMainUnit.pas	2.966665e-15	686	204
3	CgProcUnit.pas	3.639642e-14	685	37
4	DesignController.pas	7.411084e-12	6	4
5	CgFileUnit.pas	1.255789e-11	693	77

Method/Rank	LOC	NID	MCC	AIC	AEC	EIC	EEC	UCS	FCS	DCS	LOT
Prob/1	7060	38	23.694019	23	3	0	30	0	20	1	0
Prob/2	693	57	4.844187	22	16	0	9	3	37	1	0
Norm/1	4710	1	437.782330	6	0	39	0	0	0	0	1
Norm/2	3642	12	13.072488	4	0	3	3	4	0	0	1
Best/1	4710	1	437.782330	6	0	39	0	0	0	0	1
Best/2	2660	1	209.846604	2	3	21	0	0	0	0	1
Max	11350	284	437.782330	249	51	340	91	4	71	6	1

Findings from the example

A method that seems sound and can be explained to stakeholders.

The empirical distributions can be calculated from any data set, so it is easy to check against various systems or answer *what if*-questions.

If we use many variables in our quality model, the overall quality will likely be low (e.g., $0.9^{11} = 0.31$).

Current and Future Work

We are currently converting systems to use the new model. In the process, we are gathering data to compute empirical distributions for the various measurements.

We need to better understand what is going on in the various models, and why we get certain results.