How Much Information Do Software Metrics Contain?

Yossi Gil*, Maayan Goldstein, Dany Moshkovich

IBM Research---Haifa

**work done while on sabbatical from the Technion*

Everybody Likes Metrics

- Easy to invent
- Nice to collect
- No validation required
- Provide numbers your can quote

  Quoting numbers makes you sound smart
- You can even draw charts

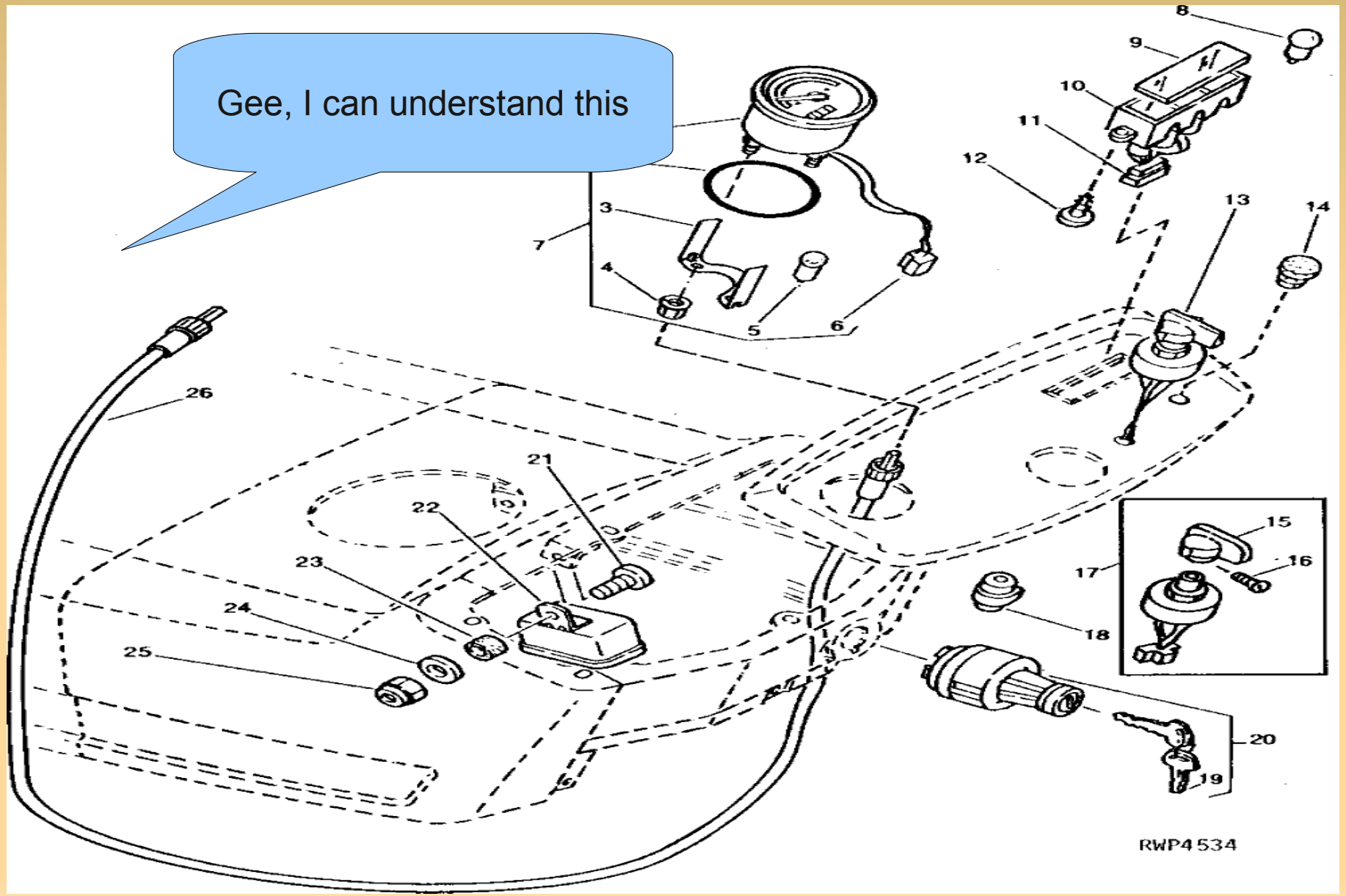
  Presenting charts (especially colorful charts) makes you seem important

Metrics have Many Practical Applications

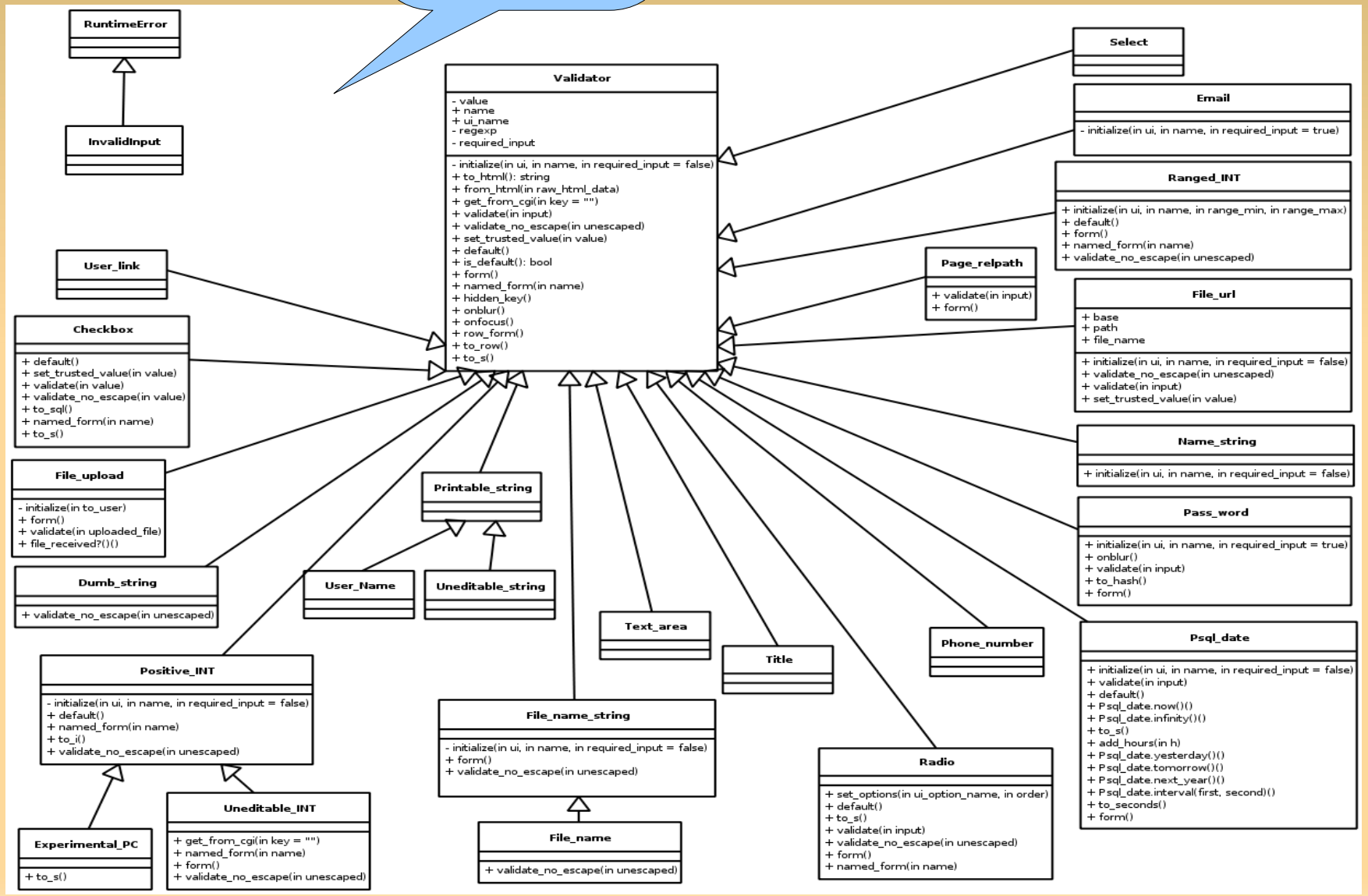
At Least One Useful Application

- Visual Program ”Understanding”
  - Present a program as a diagram
- Many, many tools.
- Problem:
  - All diagrams look the same..
  - Where do I start?

Engineers Want Diagrams



Class Diagrams? Yuk



Systemd Grokking Technology

- An IBM project, whose real name I cannot disclose.
- Visual Modelling Framework for Existing programs
  - C
  - C++
  - Java
  - Cobol
  - Domain Specific Languages
- Targeted at non-software engineers, who do software engineering

SGT Features

- Meta-Modelling (as usual)
- User Extendable (as usual)
- Dedicated Clients (as usual)
- Happy clients (perhaps not so usual)
- Boring to look at:
  - Beautiful diagrams
  - Needs ”livining”
- The big question: where do I start?

Metrics to The Rescue

- Metrics give lots of numbers...
- Are these numbers...
  - Valid? (what do they say about quality?)
  - Reliable? (do they have any ”stability” associated with tme)
- **Main Idea**: Annotate diagrams with metrics information.
  - Which metrics?
  - How to present these?
  - **Are these numbers informative?**

Summary of Work

- Concentrate in Java
- 36 software metrics
- 19 software artifacts
- ~78,000 classes
- Taxonomy of Metrics
- Measurements, giving two clear winners

Metrics Suite

- Chidamber and Kemerer Metrics
- Micro patterns
- Topological Metrics
  - #client classes: direct/indirect
  - #used classes: direct/indirect
  - Strongly Connected Components: Size, depth, height
  - Dominator Tree
- Metrics used in other projects: Belonging, Google page rank, Betweenness,....
- Java keywords: final, abstract, ...

Taxonomy of Metrics

- Semantical vs. Topological
  - Topological metrics can be derived solely from the graph topology
  - Directinal metrics (depend on the direction imposed on the software graph)
- **Locality**: internal, local, global
- **Range**: boolean, discrete, continous

Metrics Actually Used

Metric	Nature	Directed	Scope	Range
final	semantical	undirectional	internal	Boolean
abstract	semantical	undirectional	internal	Boolean
interface	semantical	undirectional	internal	Boolean
sink	topological	directional	local	Boolean
source	topological	directional	local	Boolean
baloon	topological	directional	local	Boolean
wrapper	topological	directional	local	Boolean
pure	semantical	undirectional	internal	Boolean
pool	semantical	undirectional	internal	Boolean
designator	semantical	undirectional	internal	Boolean
function pointer	semantical	undirectional	internal	Boolean
stateless	semantical	undirectional	internal	Boolean
sampler	semantical	undirectional	internal	Boolean
canopy	semantical	undirectional	internal	Boolean
DIT	semantical	undirectional	local	discrete
NOA	semantical	undirectional	local	discrete
NOC	semantical	undirectional	local	discrete
CBO	semantical	undirectional	local	discrete
RFC	semantical	undirectional	local	discrete
WMC	semantical	undirectional	local	discrete
#Incoming	topological	directional	local	discrete
#Clients	topological	directional	global	discrete
#Outgoing	topological	directional	local	discrete
#Descendants	topological	directional	global	discrete
#SCCIncoming	topological	directional	global	discrete
#SCCClients	topological	directional	global	discrete
#SCCOutgoing	topological	directional	global	discrete
#SCCDescendants	topological	directional	global	discrete
SCCSize	topological	undirectional	global	discrete
#DominatedBy	topological	directional	global	discrete
#DominatorHeight	topological	directional	global	discrete
#DominatorWeight	topological	directional	global	discrete
PageRank	topological	directional	global	continuous
Betweeness	topological	directional	global	continuous
Belonging	semantical	undirectional	local	continuous

Shannon's Entropy

- A Measure of Information a Partition Contains
- Given:
  - A set s of n elements.
  - A partition of s into k non-empty substes sized

    $s_1,\ldots, s_k$
- Amount of information (measured in bits):

$$H(n, s_1, \ldots, s_k) = - s_1 \lg (s_1/n) - \ldots - s_k \lg(s_k/n)$$

- **Information density:**
  - How many bits per each set element?
  - Divide by n
- **Normalized Information density α:** divide by $n \lg n$
  - Maximal entropy is when partition is to singletons $0 <= \alpha <= 1$
-

And The Winners Are...

Metric	k	$\widetilde{H}$	α (%)
DIT	5±1	1.6±0.3	19±3
NOA	9±3	2.1±0.3	25±3
NOC	11±4	0.7±0.2	8±3
CBO	38±12	4.1±0.2	48±5
RFC	88±31	5.4±0.3	64±5
WMC	224±111	7.0±0.5	81±5
#Incoming	33±13	3.1±0.2	37±7
#Clients	40±22	3.5±0.6	42±10
#Outgoing	27±9	3.3±0.2	38±4
#Descendants	34±20	3.8±0.7	45±10
#SCCIncoming	19±7	2.6±0.3	30±5
#SCCClients	37±21	3.0±0.6	38±8
#SCCOutgoing	19±7	2.6±0.3	30±5
#SCCDescendants	31±18	3.7±0.6	46±8
SCCSize	4±2	0.9±0.1	10±2
#DominatedBy	4±1	0.9±0.2	11±3
#DominatedBy'	4±1	1.0±0.1	12±2
#DominatorHeight	4±1	0.6±0.1	7±2
#DominatorHeight'	4±1	0.6±0.1	6±1
#DominatorWeight	9±2	0.7±0.1	8±2
#DominatorWeight'	10±2	0.7±0.1	8±2
PageRank	229±129	6.0±1.0	69±11
PageRank'	300±175	7.8±1.0	88±3
Betweeness	92±57	3.0±0.5	33±4
Betweeness'	102±65	3.0±0.5	35±4
Belonging	79±40	4.4±0.5	47±5

Boolean Metrics

Metric	Mean	Median
final	0.47 ± 0.36	0.38 ± 0.33
abstract	0.26 ± 0.10	0.25 ± 0.09
interface	0.45 ± 0.17	0.41 ± 0.17
sink	0.08 ± 0.10	0.06 ± 0.06
source	0.74 ± 0.28	0.87 ± 0.12
balloon	0.44 ± 0.23	0.37 ± 0.17
wrapper	0.80 ± 0.10	0.79 ± 0.05
pure	0.41 ± 0.17	0.41 ± 0.12
pool	0.10 ± 0.07	0.08 ± 0.04
designator	0.03 ± 0.04	0.02 ± 0.02
function pointer	0.01 ± 0.03	0.00 ± 0.00
stateless	0.84 ± 0.12	0.87 ± 0.07
sampler	0.07 ± 0.05	0.07 ± 0.02
canopy	0.62 ± 0.20	0.63 ± 0.19

Further Research

- Metrics Reliability
- Metrics Validity ????
- More Metrics