Software Solutions Symposium 2017

March 20–23, 2017

Applied Machine Learning in Software Engineering

Eliezer Kanal

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Copyright 2017 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

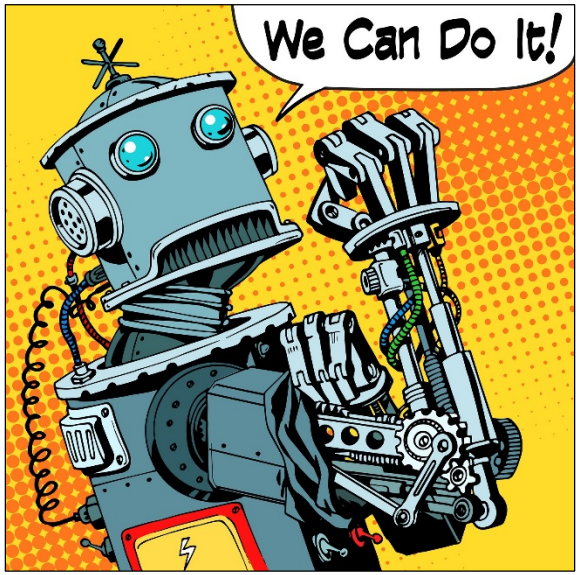
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® and CERT® are registered marks of Carnegie Mellon University.

DM-0004563

What is Machine Learning?



Tom Mitchell, former CMU Machine Learning department chair:

> *The field of Machine Learning asks the question, "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

Machine Learning seeks to automate data analysis and inference.

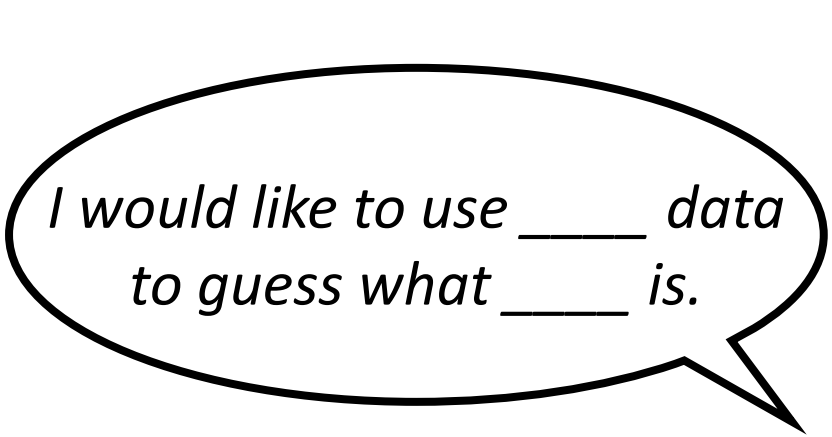
© 2017 Carnegie Mellon University

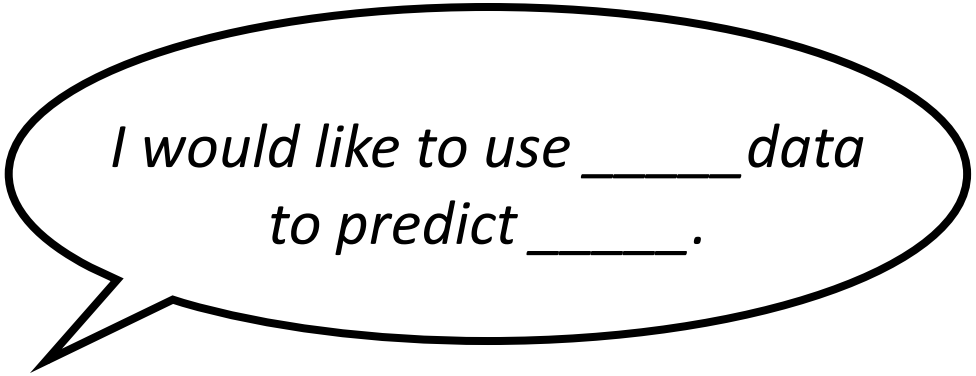
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

What is Machine Learning?

Picture (optional)

If your problem can be stated as either of the following:





…you would likely benefit from machine learning.

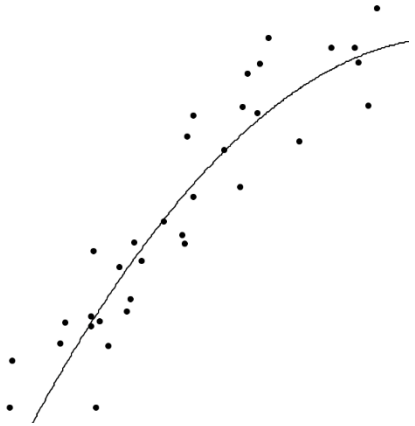
© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

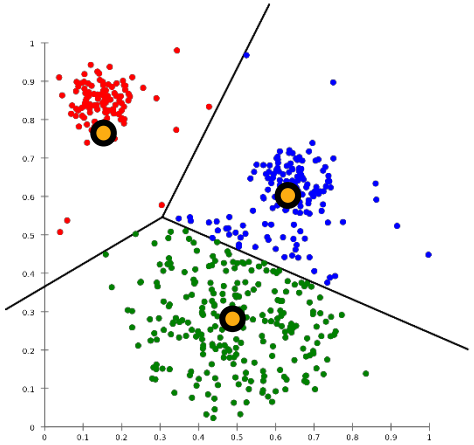
What is Machine Learning?

Sample Techniques:

- Regression



- K-Means Clustering



Clustering image: Weston.pace, https://commons.wikimedia.org/wiki/File:K_Means_Example_Step_4.svg

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

What is Machine Learning?

Feature Engineering:

Using existing data to create **more informative** data

Data Types	
Image	Static Video
Time series	Financial data Event counts
Structured text	Web forms Structured data (JSON, XML) Source code
Free text	News Tweets Email
many more…	

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

What is Machine Learning?

Picture (optional)

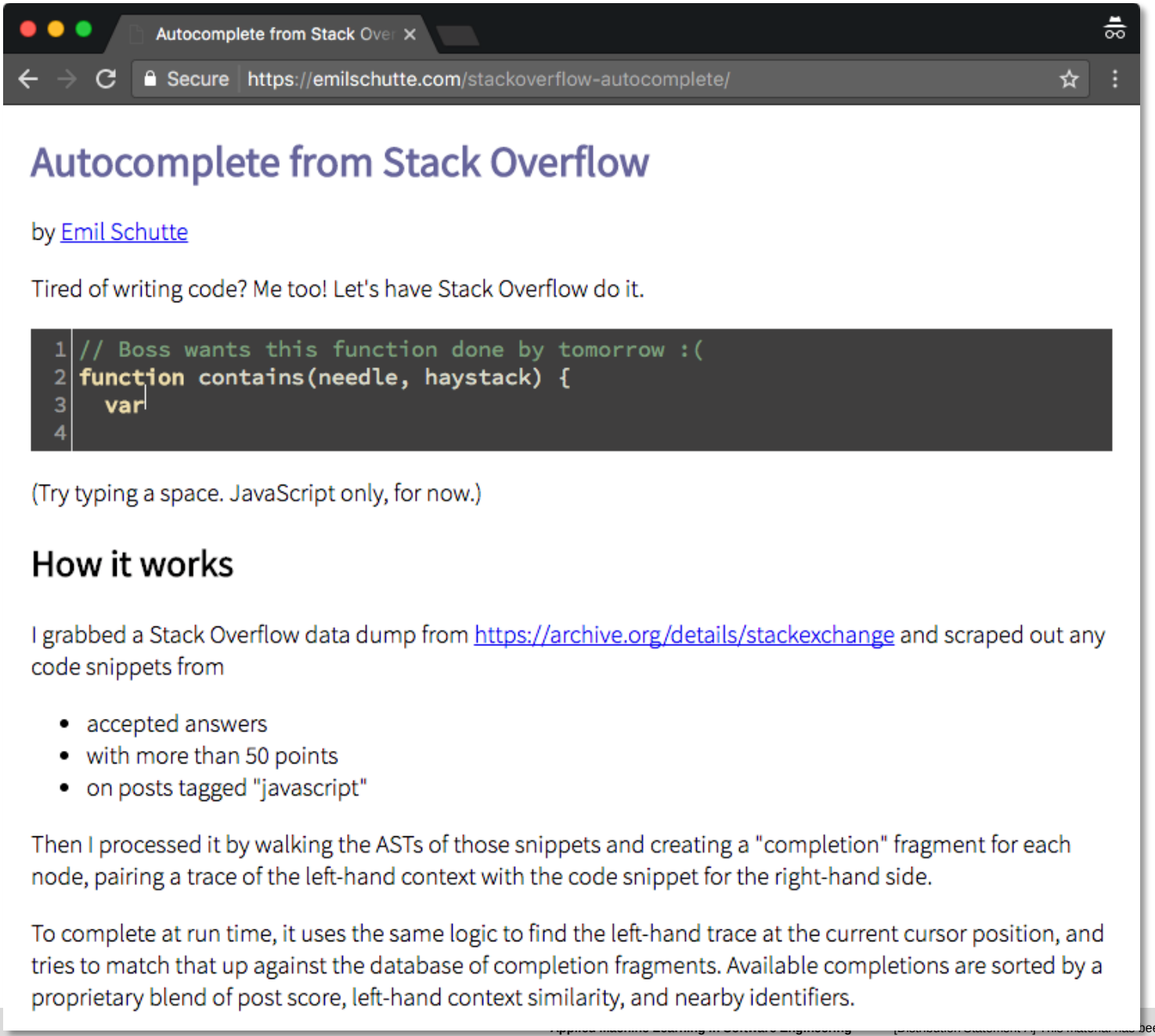
Examples:

- I would like to use **incident ticket** data to predict **customer needs**.
- I would like to use **publicly available code** to predict **what code I will write**.
- I would like to use **bug report** data to guess the location of **undetected bugs in my code**.

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Picture (optional)



© 2017 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

What is Machine Learning?

Picture (optional)

Examples:

- I would like to use **incident ticket** data to predict **customer needs**.
- I would like to use **publicly available code** to predict **what code I will write**.
- I would like to use **bug report** data to guess the location of **undetected bugs in my code**.

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Software Solutions Symposium 2017

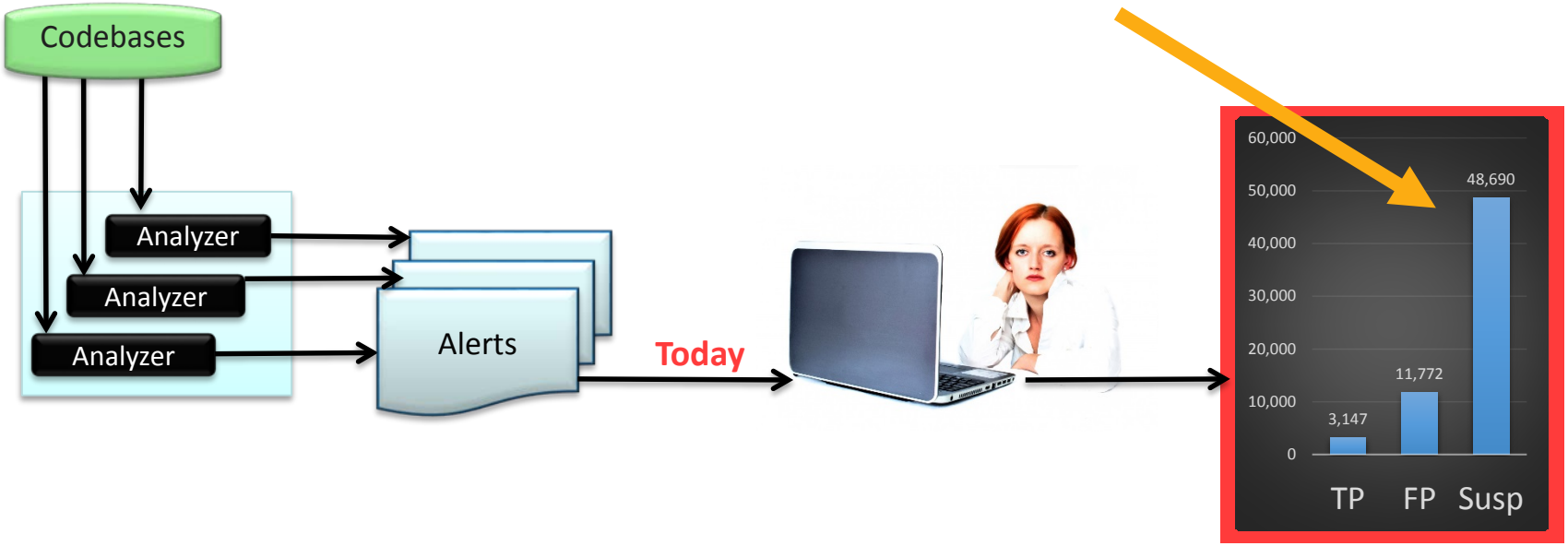
Applied ML: Vulnerability Detection

© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT Please copy and paste the appropriate distribution statement into this space.]

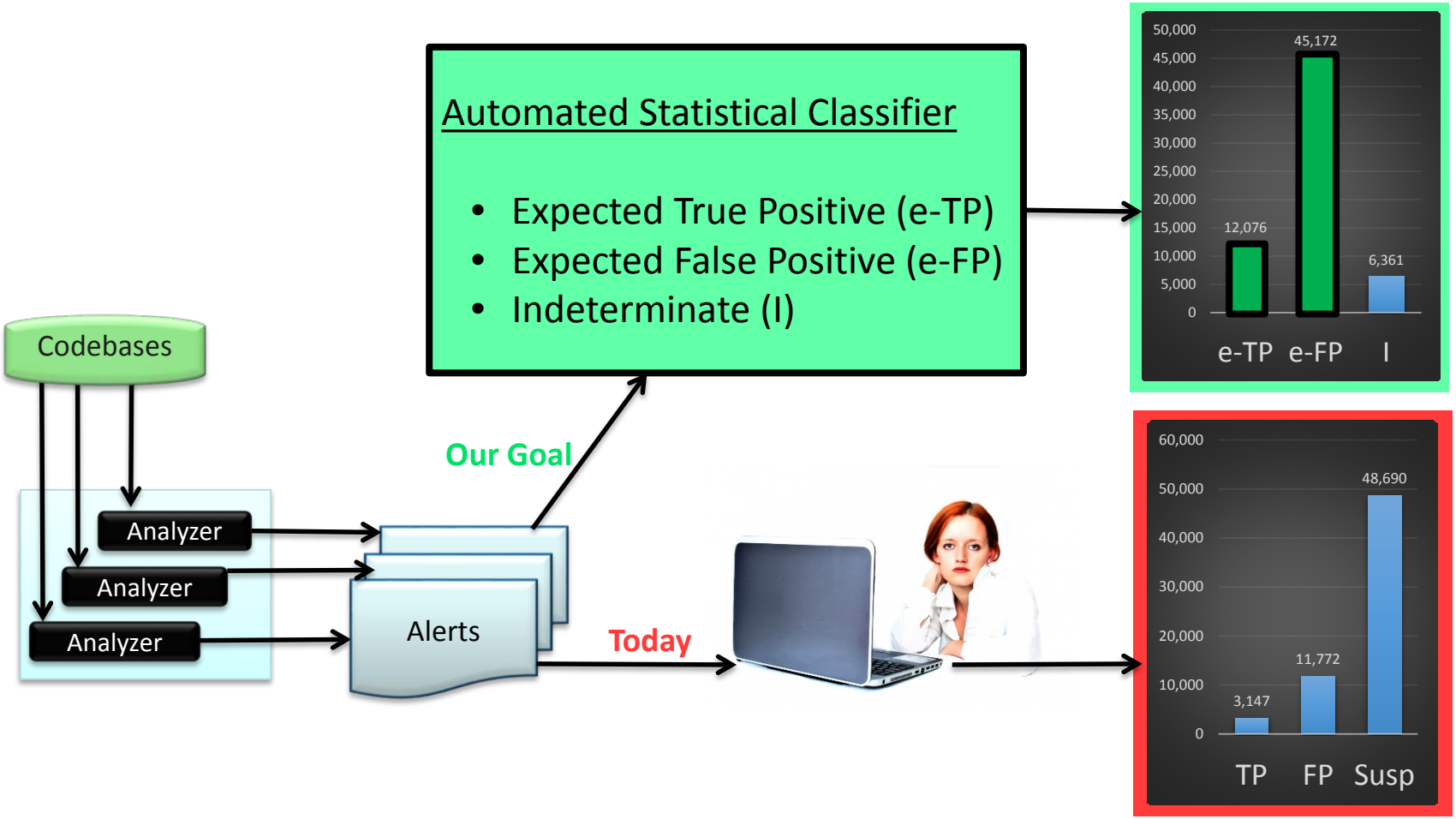
Applied ML: Vulnerability Detection

Many alerts left unaudited!



© 2017 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Vulnerability Detection



© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Vulnerability Detection

Classifiers
Lasso Logistic Regression
CART
Random Forest
Extreme Gradient Boosting (XGBoost)

Some of the features used	
Analysis tools used	Tokens in func/method
Significant LOC	Alerts in func/method
Complexity	Alerts in file
Coupling	Methods in file
Cohesion	SLOC in file
SEI coding rule	Avg Tokens
Function/method length	Avg SLOC
SLOC in func/method	Depth in code repository
# parameters in func/meth.	Cyclomatic complexity (func/meth)

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Vulnerability Detection

Significant improvement!

- 91% Classifier accuracy overall
- Specific rule accuracy at right
- 10x developer time saved!

Rule ID	XGBoost
INT31-C	97%
EXP01-J	74%
OBJ03-J	83%
FIO04-J*	80%
EXP33-C*	83%
EXP34-C*	72%
DCL36-C*	100%
ERR08-J*	100%
IDS00-J*	96%
ERR01-J*	100%
ERR09-J*	88%

* Small quantity of data

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

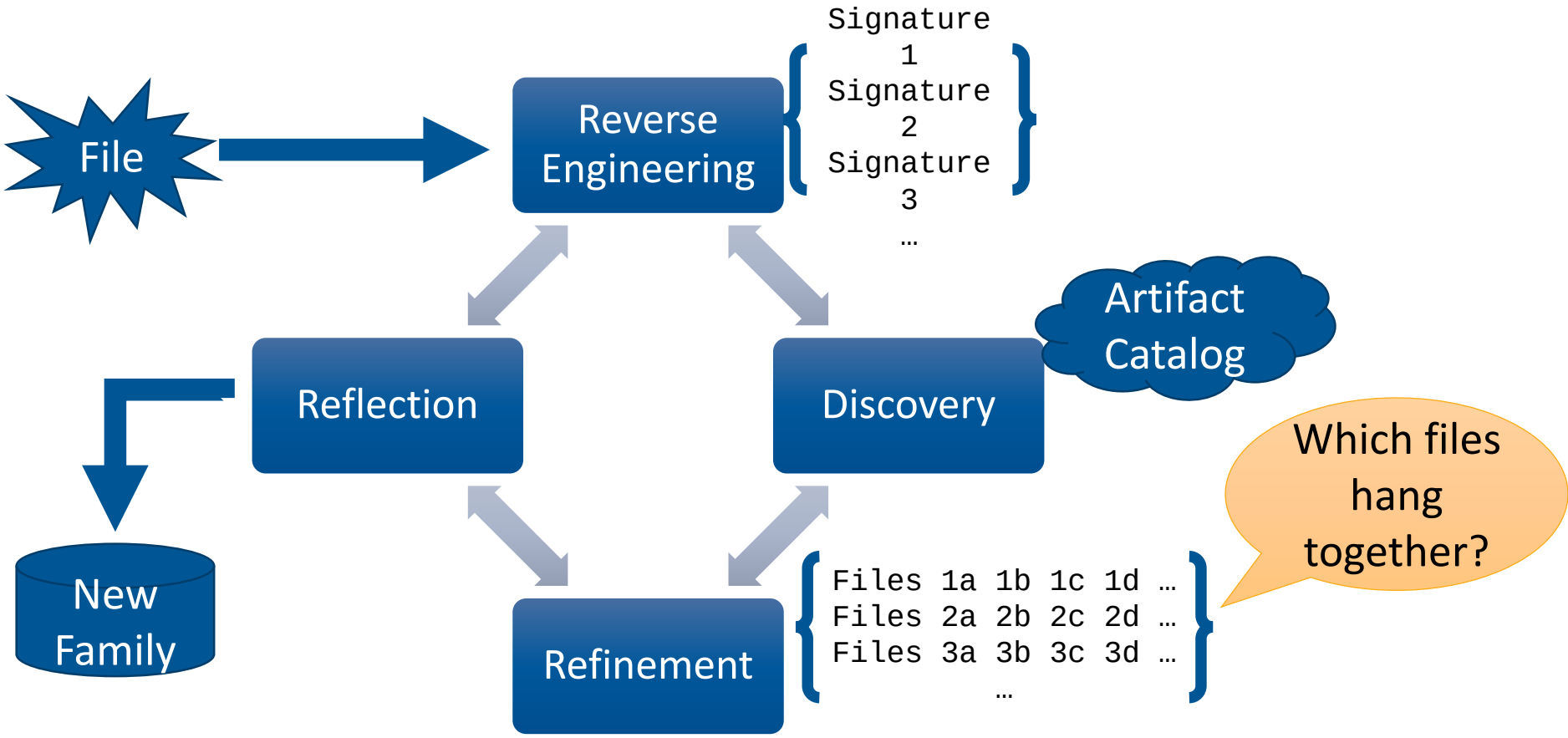
Software Solutions Symposium 2017

Applied ML: Malware family classification

© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT Please copy and paste the appropriate distribution statement into this space.]

Applied ML: Malware family classification

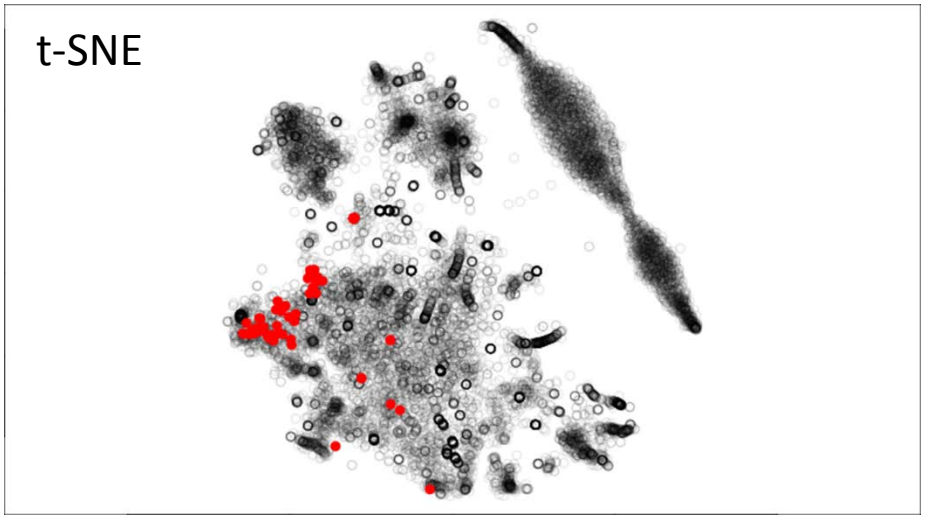


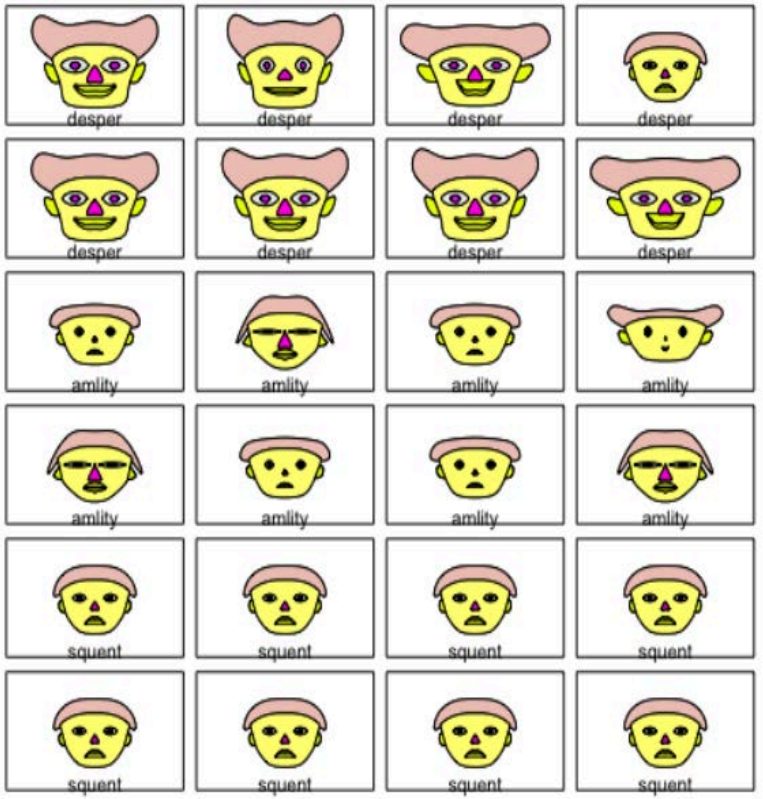
© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Malware family classification

Simplify visualization of extremely complex data through the use of dimensionality reduction and associated visualization techniques



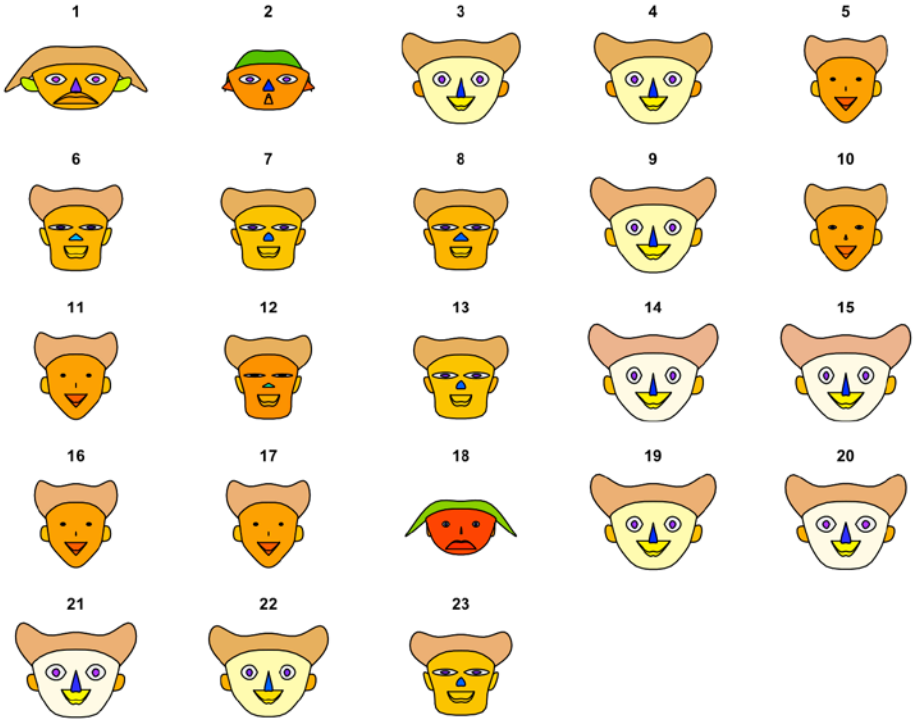


Chernoff face experiment

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Malware family classification



- Discovery found 23 files
- Manual reverse engineering was slow: only two files in 5 days
- Visualizing files with **Chernoff faces** immediately suggests groups of related files.
- Analysis burden shifts from **forming candidate groups** to **verifying groups**
- Faster and cheaper = happy clients

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Software cost estimation

Robert Ferguson, Dennis Goldenson, James McCurley, Robert W. Stoddard, David Zubrow, Debra Anderson. "Quantifying Uncertainty in Early Lifecycle Cost Estimation (QUELCE)". Dec 2011. http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=10039

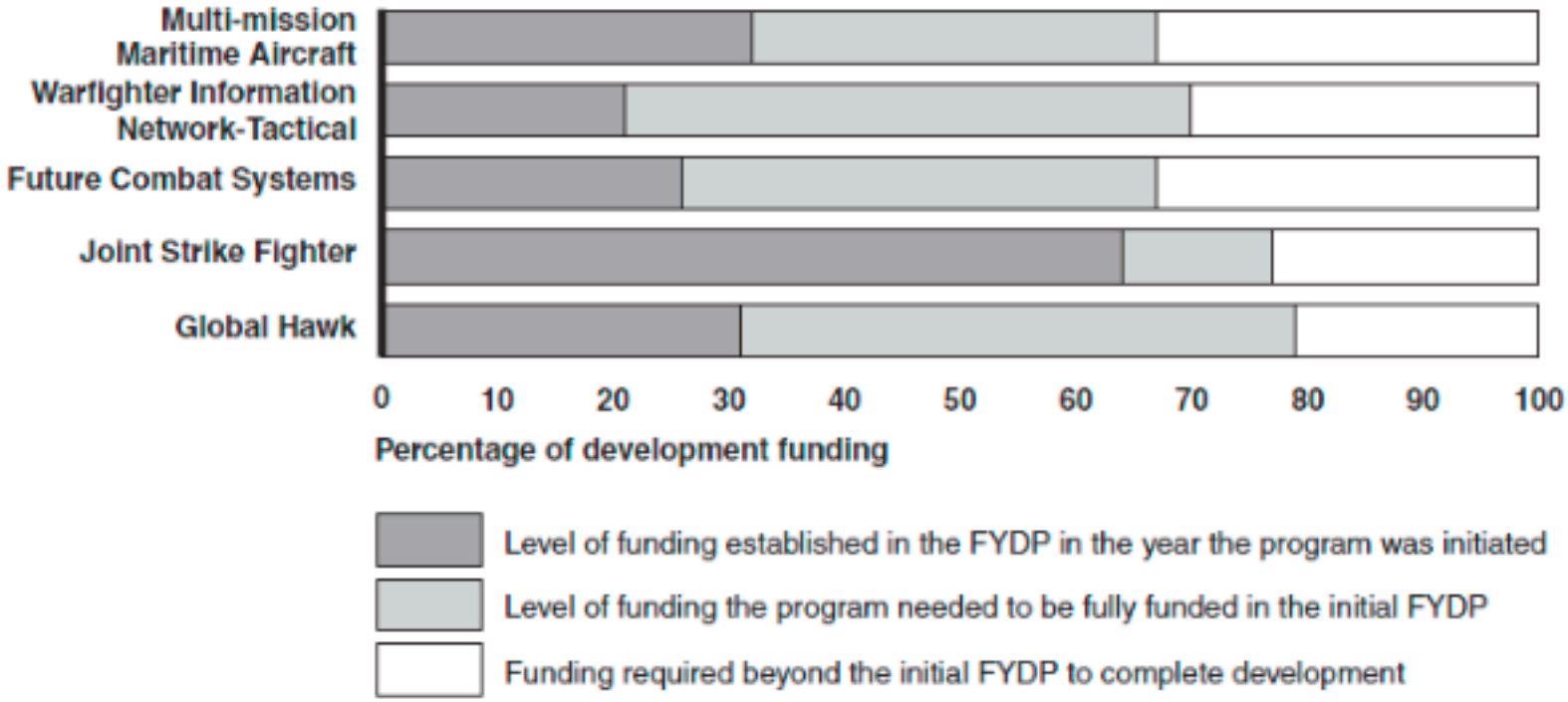
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT Please copy and paste the appropriate distribution statement into this space.]

Applied ML: Software cost estimation

Table 2 Cost Overruns in DoD Acquisitions

Funding Shortfalls at the Start of Development for Five Major Weapon System Programs



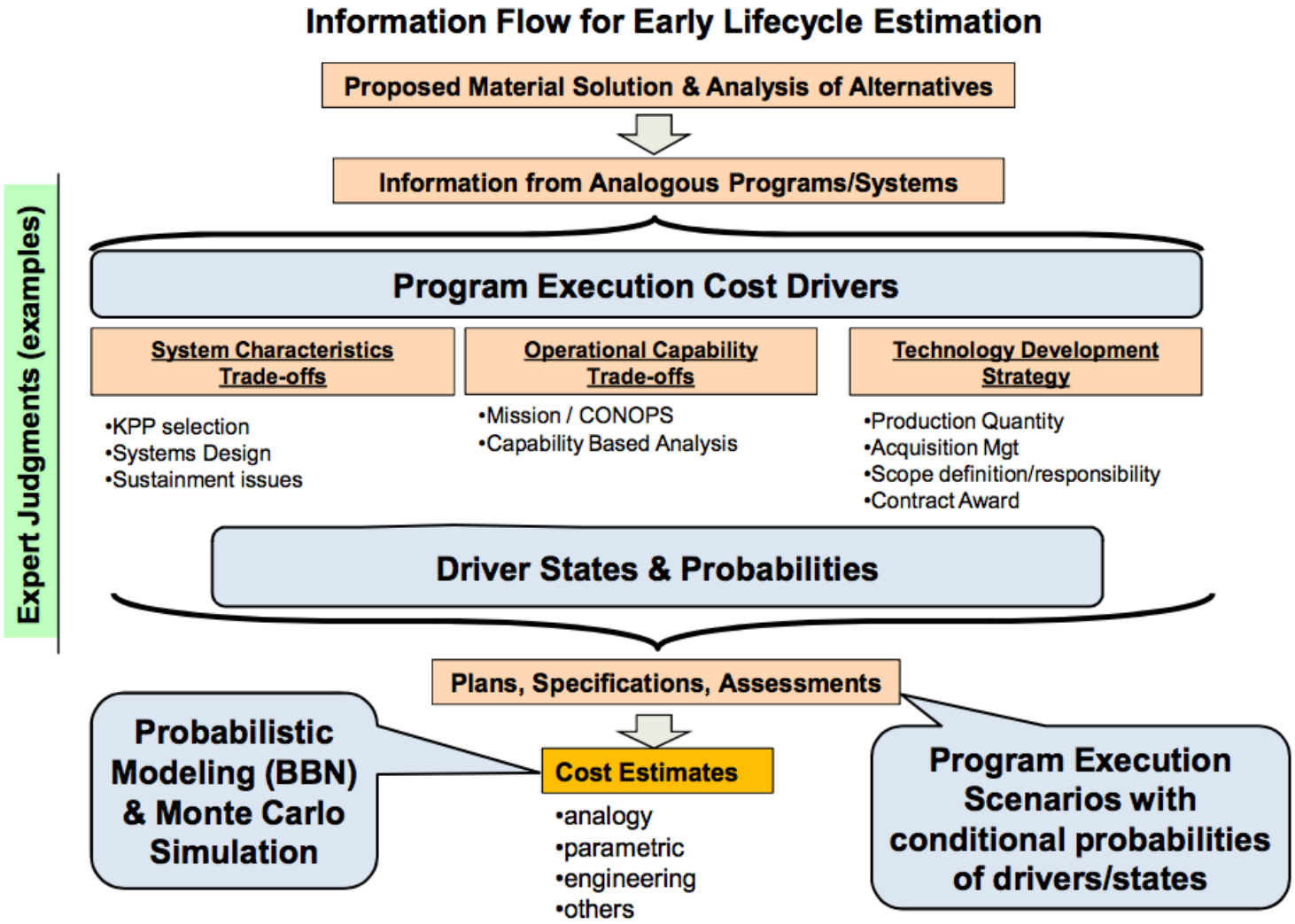
Source: DOD (data); GAO (analysis and presentation).

General Accounting Office. *Defense Acquisitions: A Knowledge-Based Funding Approach Could Improve Major Weapon System Program Outcomes*. Report to the Committee on Armed Services, U.S. Senate, July 2008, GAO-08-619.

© 2017 Carnegie Mellon University

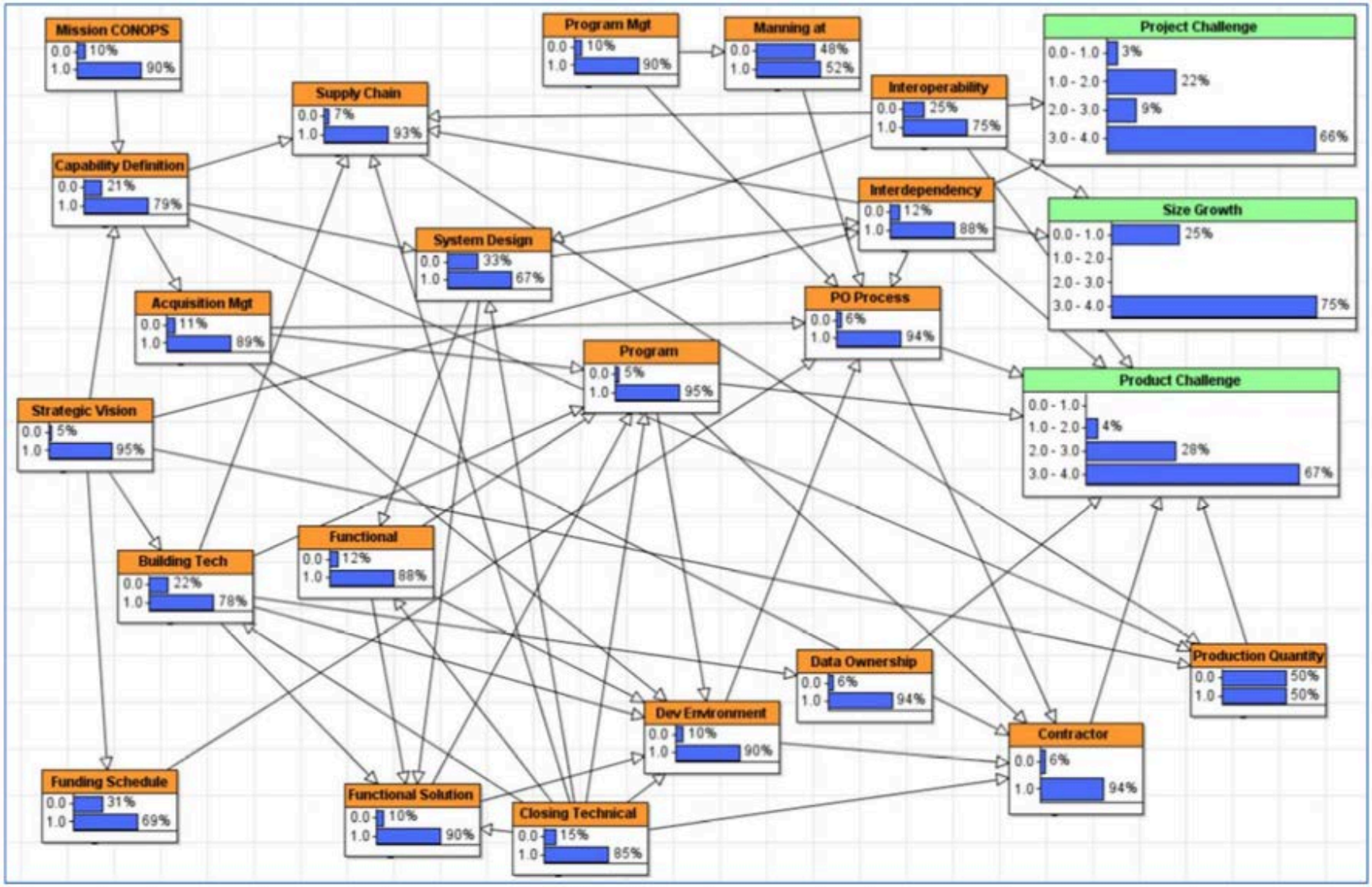
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Software cost estimation



© 2017 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Software cost estimation



© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Applied ML: Software cost estimation

Causes \ Effects	Scope Responsibility	Scope Definition	Mission / CONOPS	Capability Definition	Funding Schedule	Prog Mgt Structure	Manning at program office	Systems Design	Standards/Certifications	Acquisition Management	Program Mgt - Contractor Relations	Project Social / Dev Env	Supply Chain Vulnerabilities	Information sharing	PO Process Performance	Sustainment Issues	Contract Award	Contractor Performance	Production Quantity	Data Ownership	Change in Strategic Vision	Advocacy Change	Industry Company Assessment	Cost Estimate	Test & Evaluation	Closing Technical Gaps (CBA)	Building Technical Capability & Capacity	Functional Measures	Functional Solution Criteria (measure)	Interdependency	Interoperability	Size	Project Challenge	Product Challenge	Total
Scope Responsibility		2				1	1			1	1	1			1																				8
Scope Definition	3					1	1																												5
Mission / CONOPS	0	3		3																															6
Capability Definition					0	0	0	3	2	2	1	1	2	2	0	1	0	0	2	0															16
Funding Schedule							1			1					2			1																	5
Prog Mgt Structure							2					1		1	2																				6
Manning at program office						2								1	2																				5
Systems Design	1												1	1		1				1			2	2	3	2	1	2	2	2	2				23
Standards/Certifications													1			1	1	1					1		3		1		1						10
Acquisition Management											2	3	1	1	2		2	2		1			1	1										1	20
Program Mgt - Contractor Relations												2		1	1	1		2							1		1	1	1	1				2	15
Project Social / Dev Env										1	1			1	2		2	2		1					1		1	1	1	1	1	1	1	1	19
Supply Chain Vulnerabilities					1			1	1	1								1	2																7
Information sharing								1							1	1		1		1						1	1								7
PO Process Performance																		2																2	4
Sustainment Issues																																			0
Contract Award																																			0
Contractor Performance																																		2	2
Production Quantity																																		2	2
Data Ownership																																		2	2
Change in Strategic Vision				3	2									2		2	2		3			3	2	3	2		3			2					29
Advocacy Change	1	2				1	1			1																									6
Industry Company Assessment																																			0
Cost Estimate																																			0
Test & Evaluation																																			0
Closing Technical Gaps (CBA)	1							3	1	1	2	2	2	1	0	2	2	2	1	1	1	2	2	2	1		2	2	2	1	1				37
Building Technical Capability & Capacity (CBA)		1				1				1	2	2	2	3		2	2	1	1	2	0	1	2	1	1	0		1	2	1					29
Functional Measures									1		2	2	1	1	1		1	1		1		1	1		2			2							17
Functional Solution Criteria (measure)									1		2	2						1					1		2					1					10
Interdependency	1	1			1	1	1	2	1		1	1	2	1	2	2		1		1	1	1	1	1	1				1		2	2	2	3	34
Interoperability	1	1						2	1	1	1	1	2	1	1	2		1		1				1	3		1		1	1		2	2	2	29
Size																																			0
Project Challenge																																			0
Product Challenge																																			0
Totals	8	10	0	6	4	7	7	12	8	10	15	18	14	17	17	15	12	19	9	10	2	8	13	11	20	3	11	8	14	11	7	5	5	17	

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Contact Information

Eliezer Kanal

Technical Manager

Telephone: +1 412.268.5204

Email: ekanal@sei.cmu.edu

© 2017 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.