W1

Test Analytics, AI/ ML
Wednesday, October 3rd, 2018 11:30 AM

Fighting Test Flakiness: A Disease that Artificial Intelligence Will Cure

Presented by:

Tariq King

Ultimate Software

Brought to you by:

350 Corporate Way, Suite 400, Orange Park, FL 32073
888-268-8770 · 904-278-0524 - info@techwell.com - http://www.starwest.techwell.com/

Tariq King

Tariq King is the senior director and engineering fellow for quality and performance at Ultimate Software. With more than fifteen years' experience in software testing research and practice, Tariq heads Ultimate Software's quality program by providing technical and people leadership, strategic direction, staff training, and research and development in software quality and testing practices. Tariq is a frequent presenter at conferences and workshops, has published more than thirty research articles in IEEE- and ACM-sponsored journals, and has developed and taught software testing courses in both industry and academia. His primary research interest is engineering autonomous self-testing systems. He is cofounder with Jason Arbon of the Artificial Intelligence for Software Testing Association. Contact Tariq via LinkedIn.

STAR WEST
A TECHWELL EVENT

FIGHTING TEST FLAKINESS

A DISEASE THAT ARTIFICIAL INTELLIGENCE WILL CURE

TARIQ M. KING

Ultimate SOFTWARE

October 3, 2018
Disneyland Hotel, Anaheim, California

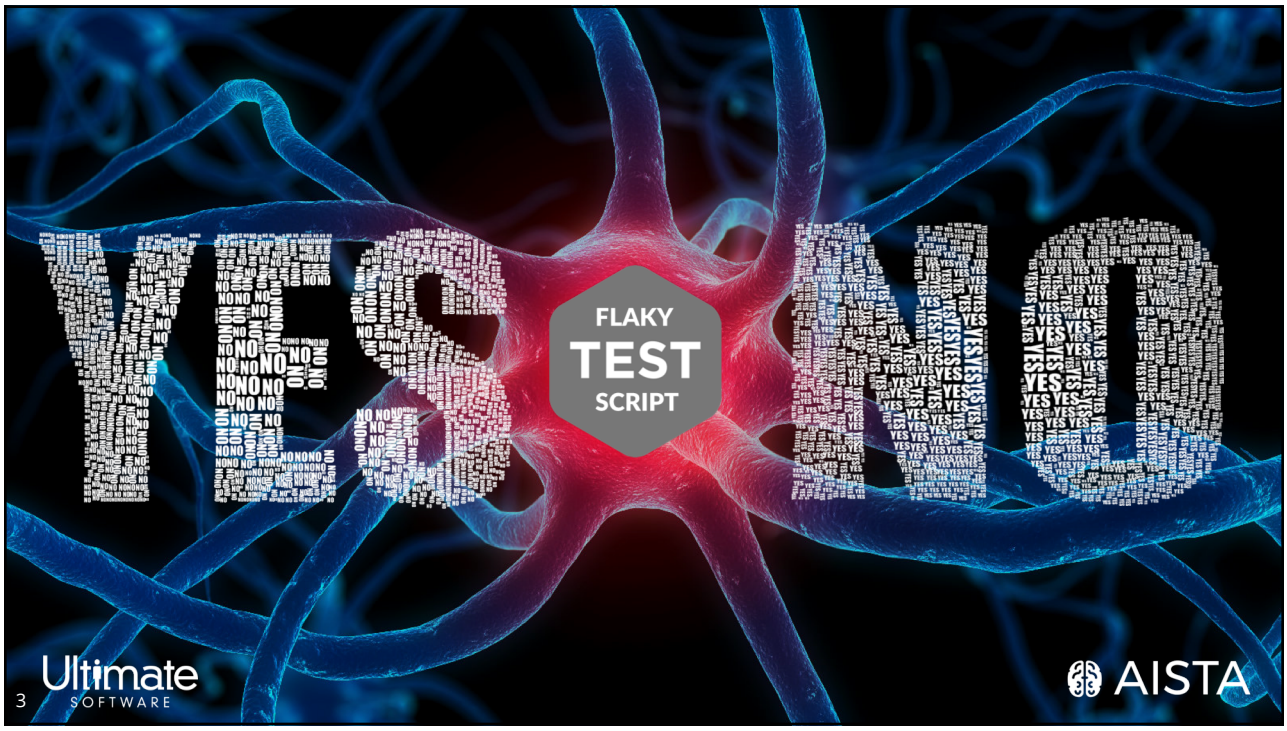
AISTA

AGENDA

- MOTIVATION
  - TEST FLAKINESS PROBLEM
  - AI FOR SOFTWARE TESTING
- FIGHTING TEST FLAKINESS WITH AI/ML
  - TEST FLAKINESS PREDICTION MODEL
  - AI-DRIVEN TEST AUTOMATION
- WRAP UP

Ultimate SOFTWARE

AISTA



3

FLAKY TESTS

EXHIBIT BOTH PASSING AND FAILING RESULTS WITH SAME CODE

HIGH TEST COMPLEXITY

Large test scripts with many steps and assertions tend to be unstable.

IMPLEMENTATION COUPLING

Directly referencing classes and/or web element selectors instead of interfaces.

FLAKY TEST SCRIPT

NON-DETERMINISTIC TEST CODE

Conditional logic such as if-then-else, explicit delays, random number generation.

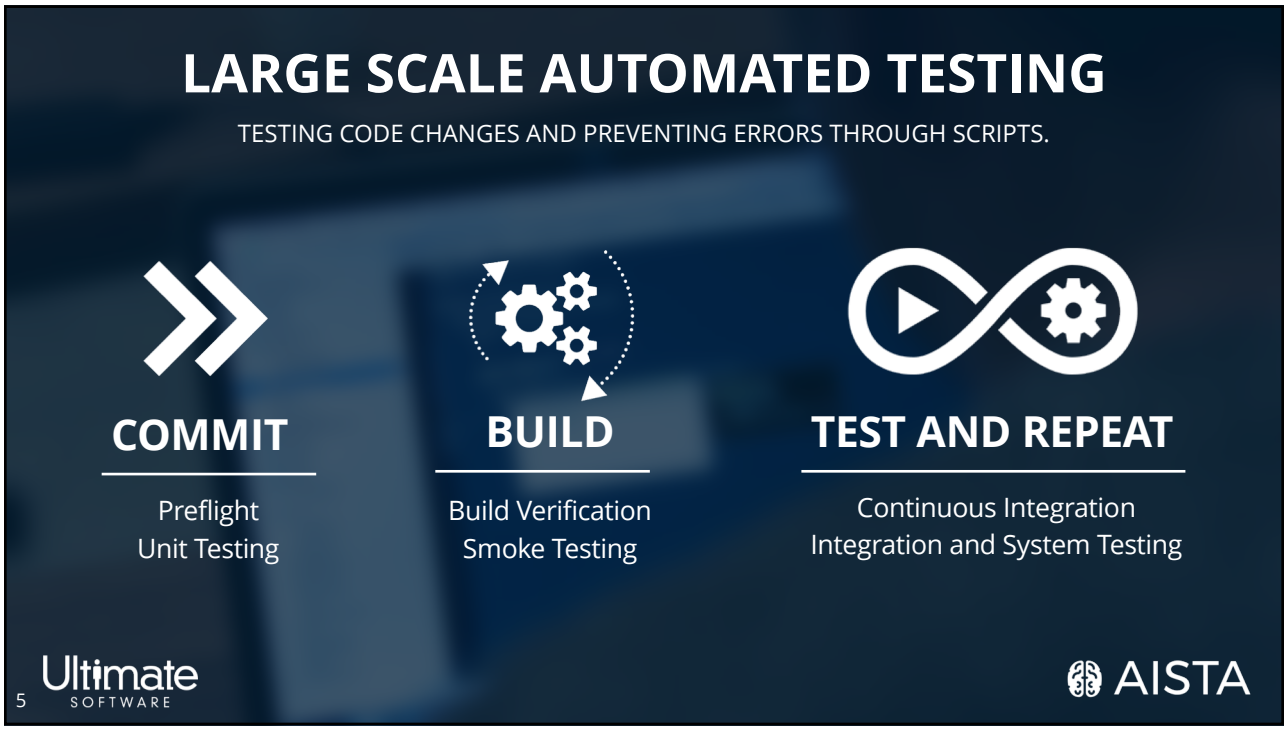
IMPROPER SETUP OR TEARDOWN

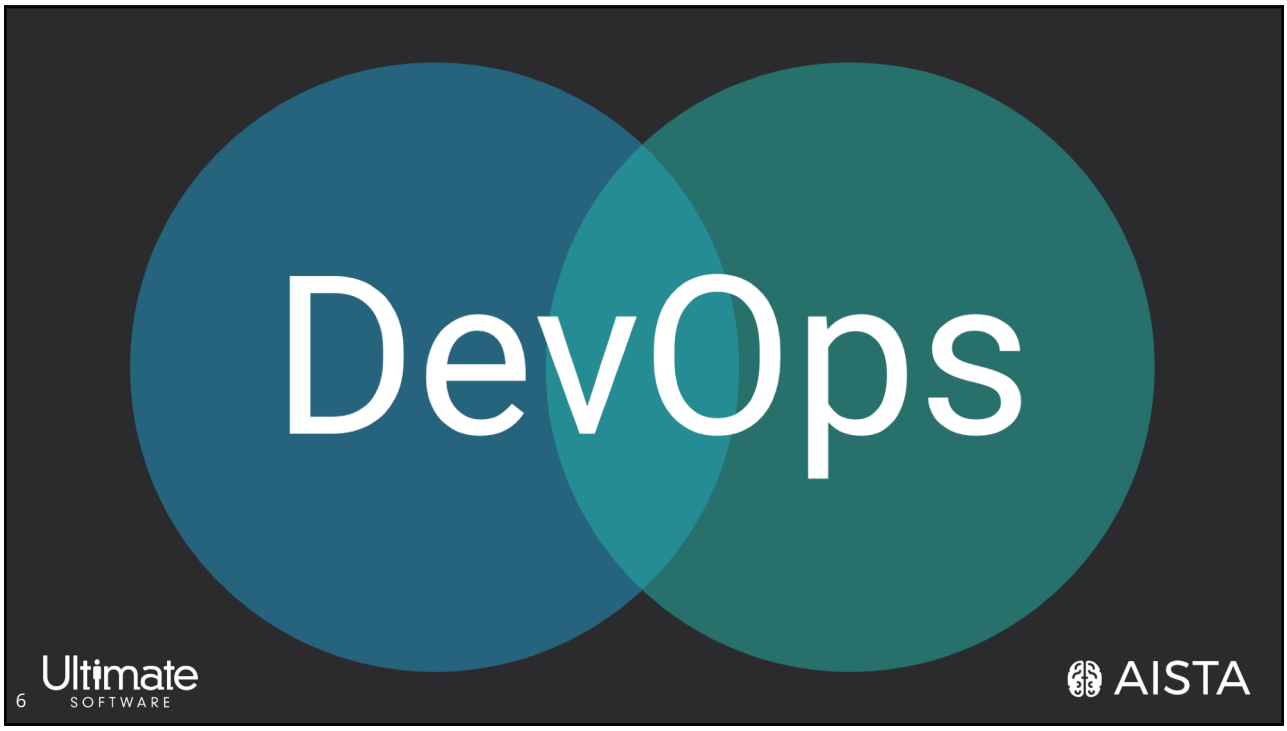
Failure to establish or re-establish prerequisites such as data or environment conditions.

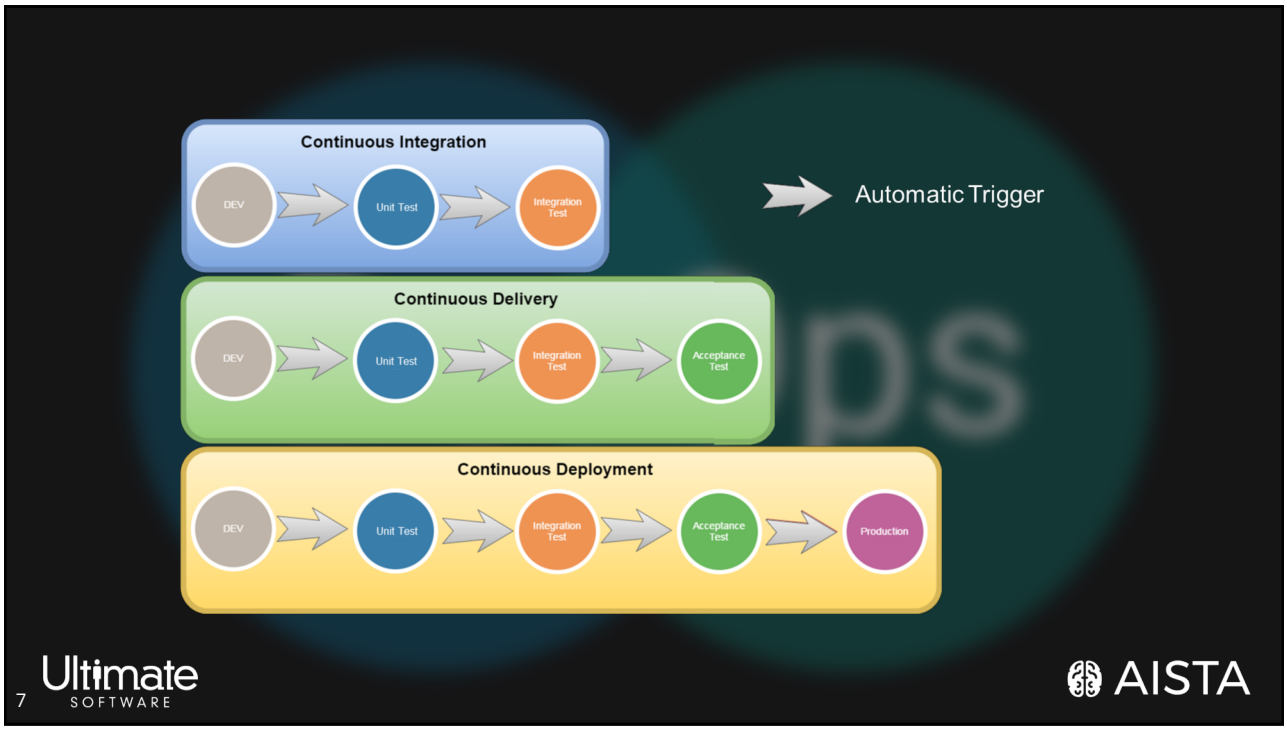
Ultimate SOFTWARE

AISTA

4


LARGE SCALE AUTOMATED TESTING
TESTING CODE CHANGES AND PREVENTING ERRORS THROUGH SCRIPTS.
COMMIT
Preflight
Unit Testing
BUILD
Build Verification
Smoke Testing
TEST AND REPEAT
Continuous Integration
Integration and System Testing
5
Ultimate SOFTWARE
AISTA


DevOps
6
Ultimate SOFTWARE
AISTA


Continuous Integration
DEV
Unit Test
Integration Test
Automatic Trigger
Continuous Delivery
DEV
Unit Test
Integration Test
Acceptance Test
Continuous Deployment
DEV
Unit Test
Integration Test
Acceptance Test
Production
7
Ultimate SOFTWARE
AISTA

Google
8
Ultimate SOFTWARE
AISTA

PRODUCTIVITY FOR DEVELOPERS…

RELIES ON THE ABILITY OF [AUTOMATED] TESTS TO FIND REAL PROBLEMS WITH THE CODE BEING CHANGED OR DEVELOPED IN A TIMELY AND RELIABLE FASHION.

– JOHN MICCO
SENIOR MANAGER, GOOGLE RESEARCH (2016)

WHAT WE FIND IN PRACTICE IS THAT…

84% OF THE TRANSITIONS WE OBSERVE FROM PASS TO FAIL INVOLVE A **FLAKY TEST!**

– JOHN MICCO
SENIOR MANAGER, GOOGLE RESEARCH (2016)

MY SELENIUM TESTS AREN'T STABLE!

THERE ARE A NUMBER OF COMMON CAUSES FOR FLAKY TESTS THAT I'VE OBSERVED…

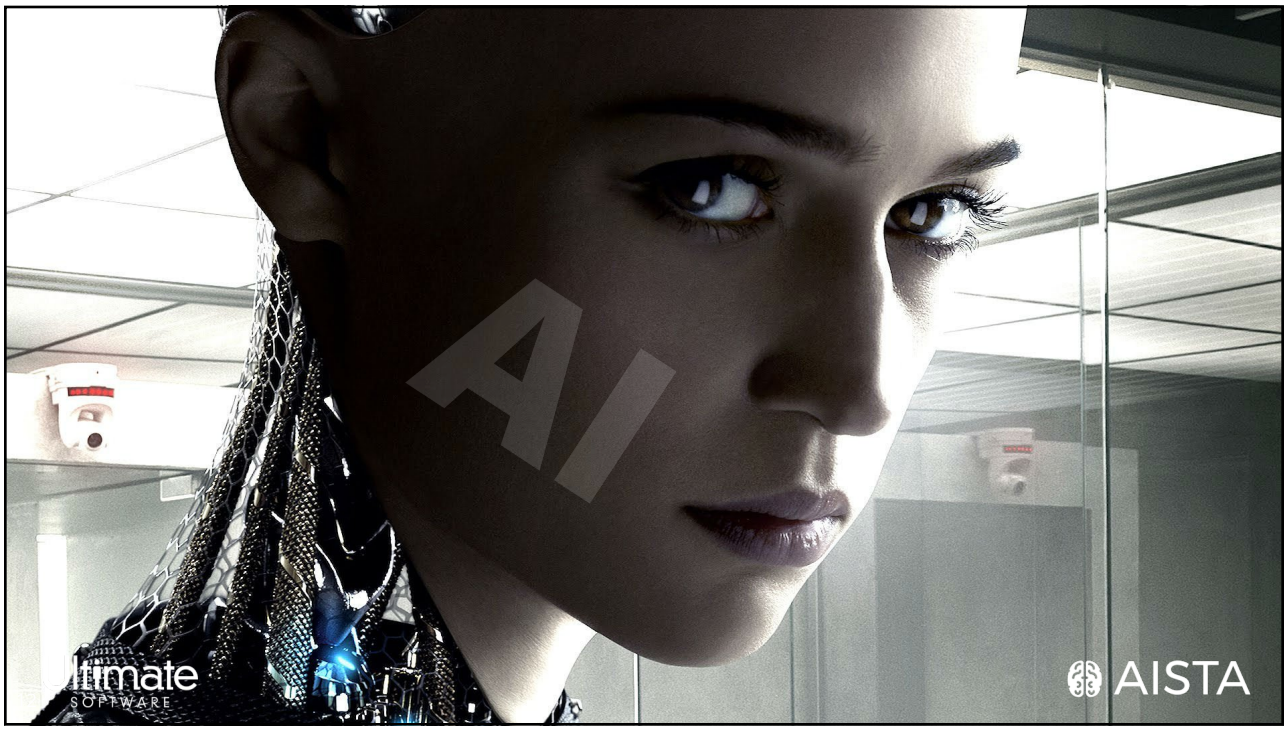
1) POOR TEST ISOLATION
2) FLAKY EXTERNAL SERVICES
3) TIMEOUTS NOT LONG ENOUGH
4) TIMEOUTS TOO LONG

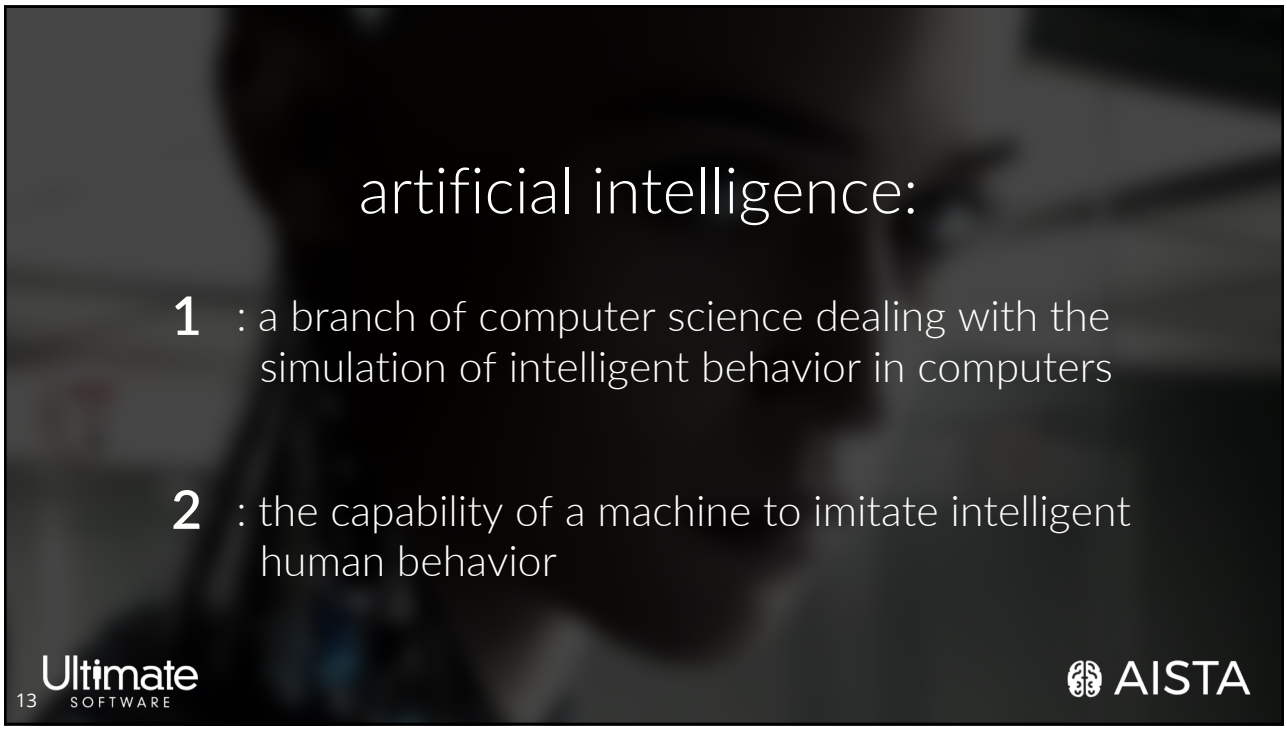
– SIMON STEWART
STAFF SOFTWARE ENGINEER IN TEST, GOOGLE (2009)

Ultimate SOFTWARE

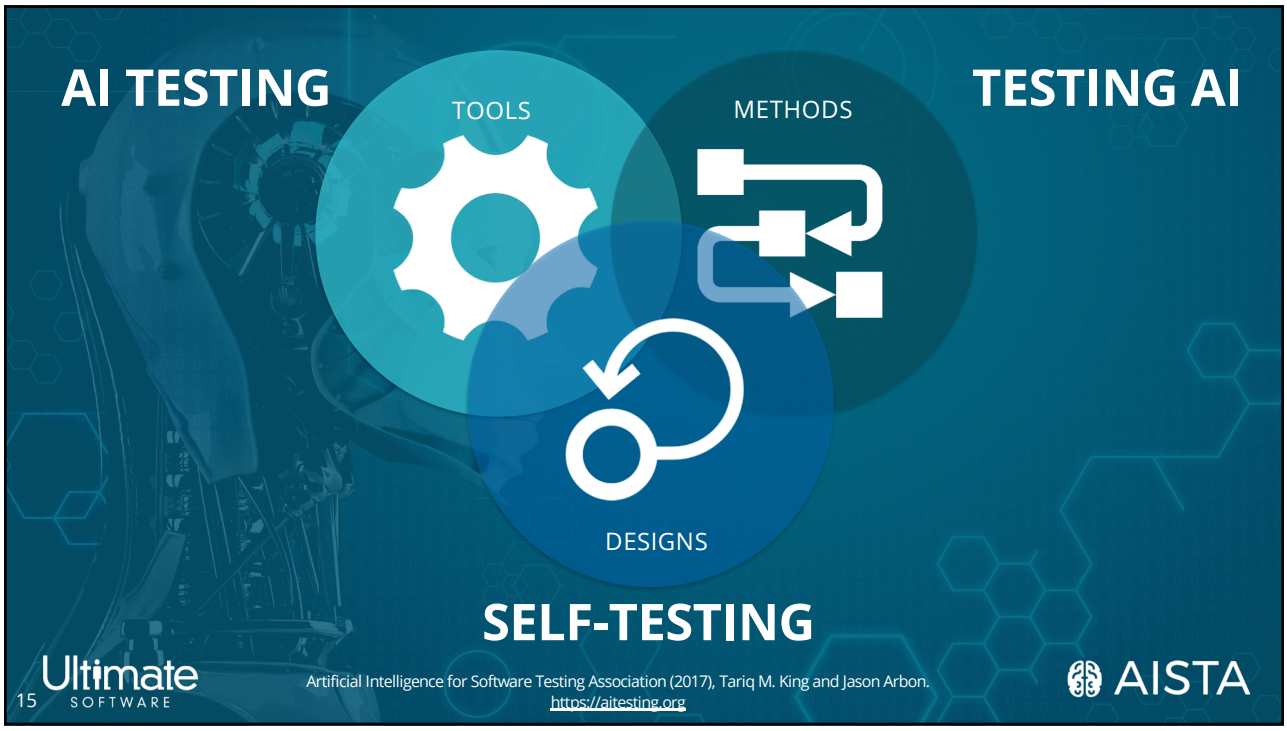
11

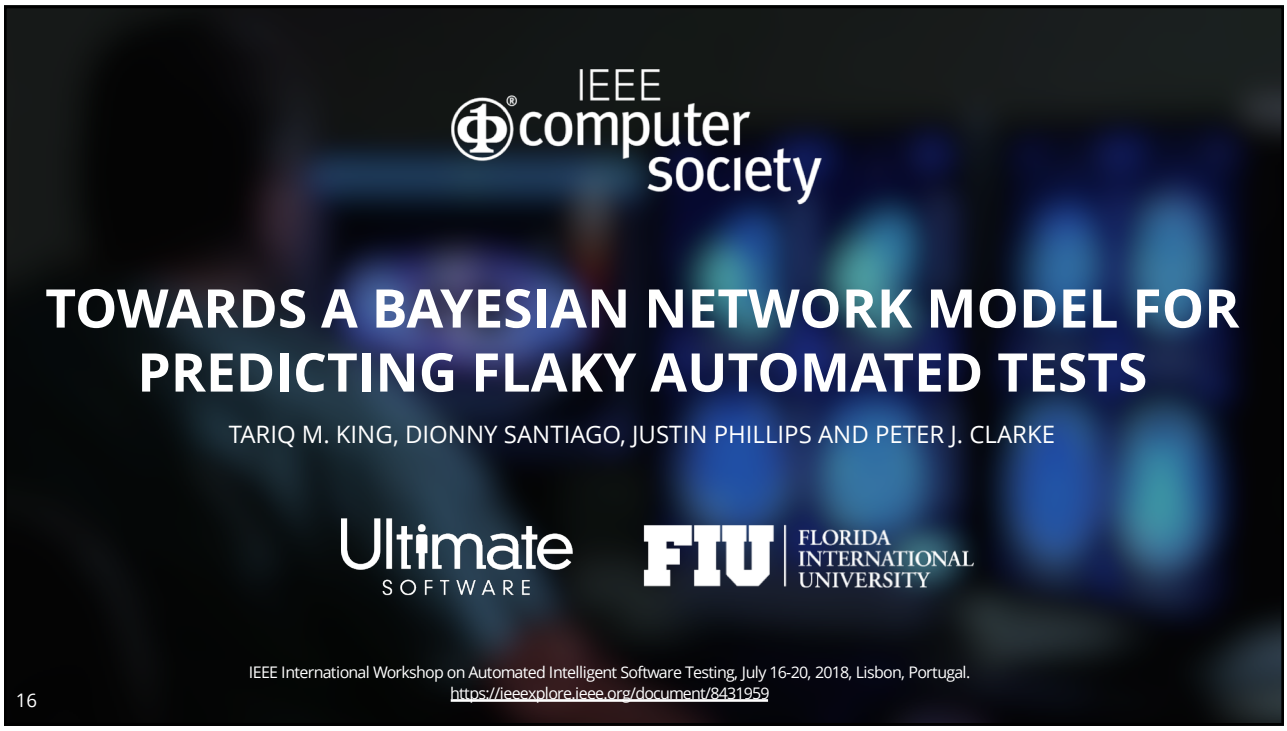
AISTA




artificial intelligence:
1 : a branch of computer science dealing with the simulation of intelligent behavior in computers
2 : the capability of a machine to imitate intelligent human behavior
Ultimate SOFTWARE
13
AISTA

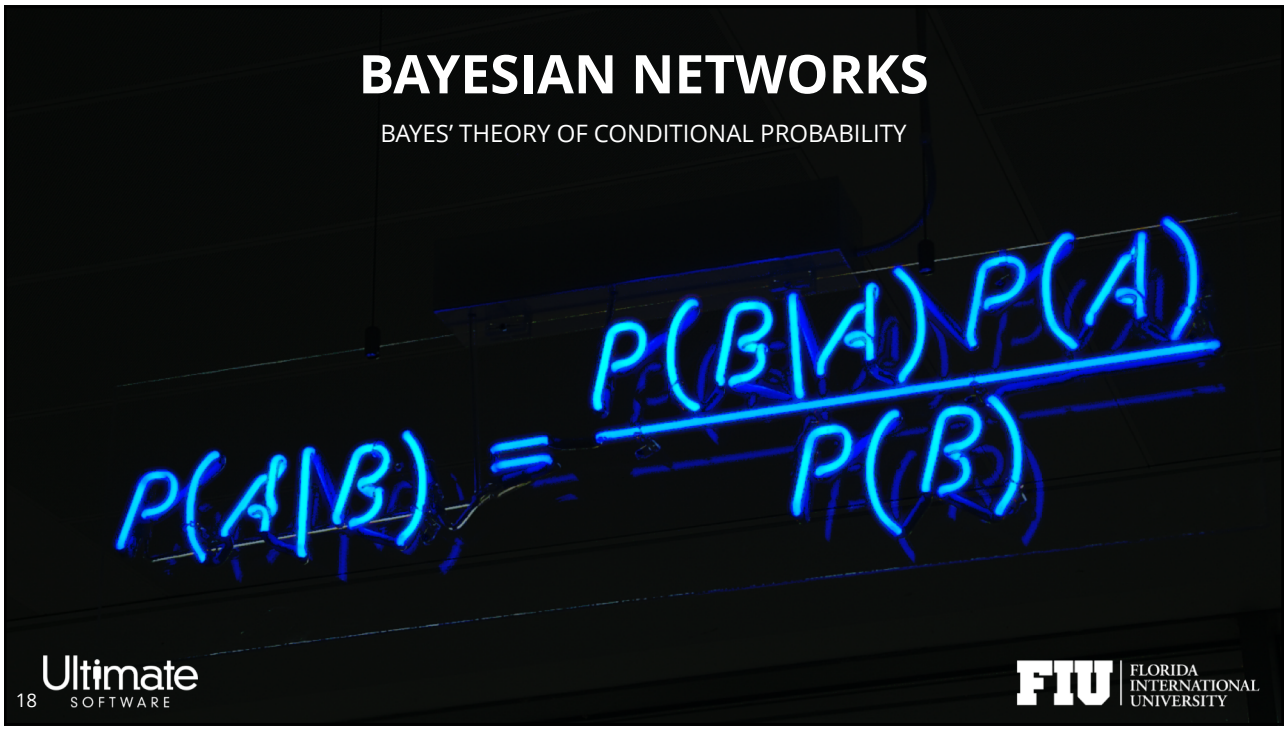
AIST
ARTIFICIAL INTELLIGENCE FOR SOFTWARE TESTING
Ultimate SOFTWARE
14
Artificial Intelligence for Software Testing Association (2017), Tariq M. King and Jason Arbon.
https://aitesting.org
AISTA


AI TESTING
TESTING AI
TOOLS
METHODS
DESIGNS
SELF-TESTING
Ultimate SOFTWARE
Artificial Intelligence for Software Testing Association (2017), Tariq M. King and Jason Arbon.
https://aitesting.org
AISTA
15


IEEE computer society
TOWARDS A BAYESIAN NETWORK MODEL FOR PREDICTING FLAKY AUTOMATED TESTS
TARIQ M. KING, DIONNY SANTIAGO, JUSTIN PHILLIPS AND PETER J. CLARKE
Ultimate SOFTWARE
FIU FLORIDA INTERNATIONAL UNIVERSITY
IEEE International Workshop on Automated Intelligent Software Testing, July 16-20, 2018, Lisbon, Portugal.
https://ieeexplore.ieee.org/document/8431959
16

Ultimate
SOFTWARE
FIU
FLORIDA
INTERNATIONAL
UNIVERSITY

17


BAYESIAN NETWORKS
BAYES' THEORY OF CONDITIONAL PROBABILITY
$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$
Ultimate
SOFTWARE
FIU
FLORIDA
INTERNATIONAL
UNIVERSITY

18

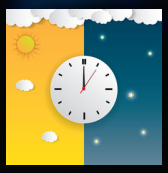
BAYESIAN NETWORKS

BAYES' THEORY OF CONDITIONAL PROBABILITY

Conditional probability is the probability of an event happening, given that it has some relationship to one or more other events.

FINDING AN OPEN PARKING SPACE ?

LOCATION



TIME OF DAY

EVENTS

APPLYING BAYES' THEOREM

EXAMPLE: MEDICAL DIAGNOSIS

Patient's probability of liver disease if they are an alcoholic?

Past data tells you that:

- 10% of patients entering the clinic have liver disease.
- 5% of the patients are alcoholics.
- Among patients diagnosed with liver disease, 7% are alcoholics.

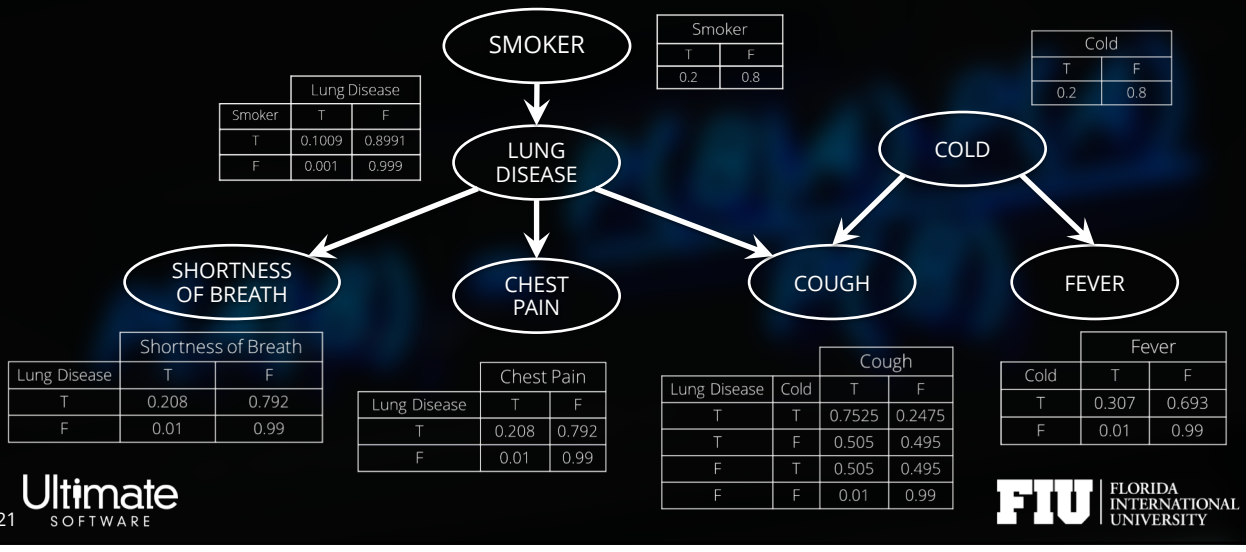
A – Patient has liver disease.	P(A) = 0.1
B – Patient is an alcoholic.	P(B) = 0.05
B|A – Liver disease given alcoholism.	P(B|A) = 0.07

P(A|B) = (0.07 x 0.1) / 0.05 = 0.14

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

BAYESIAN NETWORK MODELING

A DAG ANNOTATED WITH CONDITIONAL PROBABILITIES



Smoker	T	F
	0.2	0.8

Cold	T	F
	0.2	0.8

Smoker	Lung Disease T	Lung Disease F
T	0.1009	0.8991
F	0.001	0.999

Lung Disease	Shortness of Breath T	Shortness of Breath F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Chest Pain T	Chest Pain F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Cold	Cough T	Cough F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

Cold	Fever T	Fever F
T	0.307	0.693
F	0.01	0.99

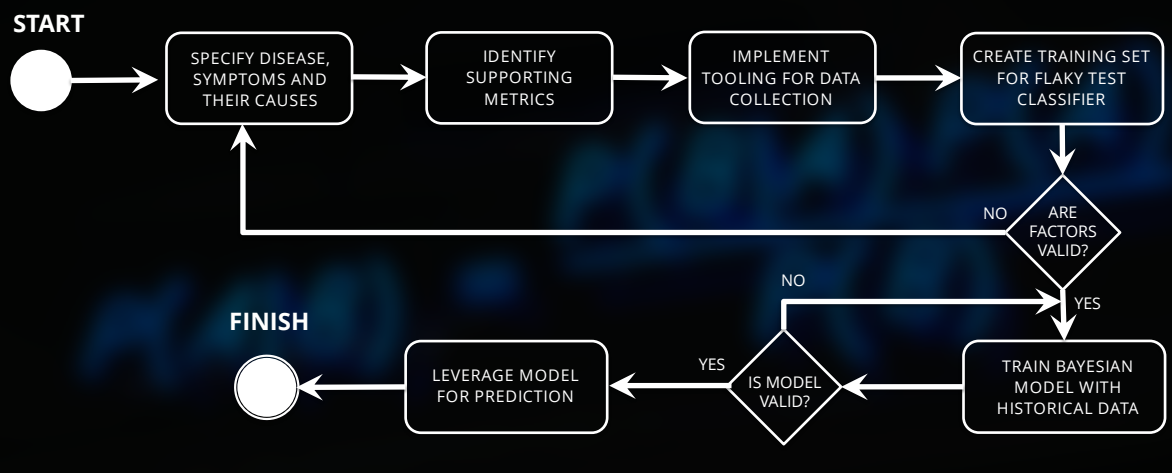
Ultimate SOFTWARE

21

FIU FLORIDA INTERNATIONAL UNIVERSITY

OUR APPROACH

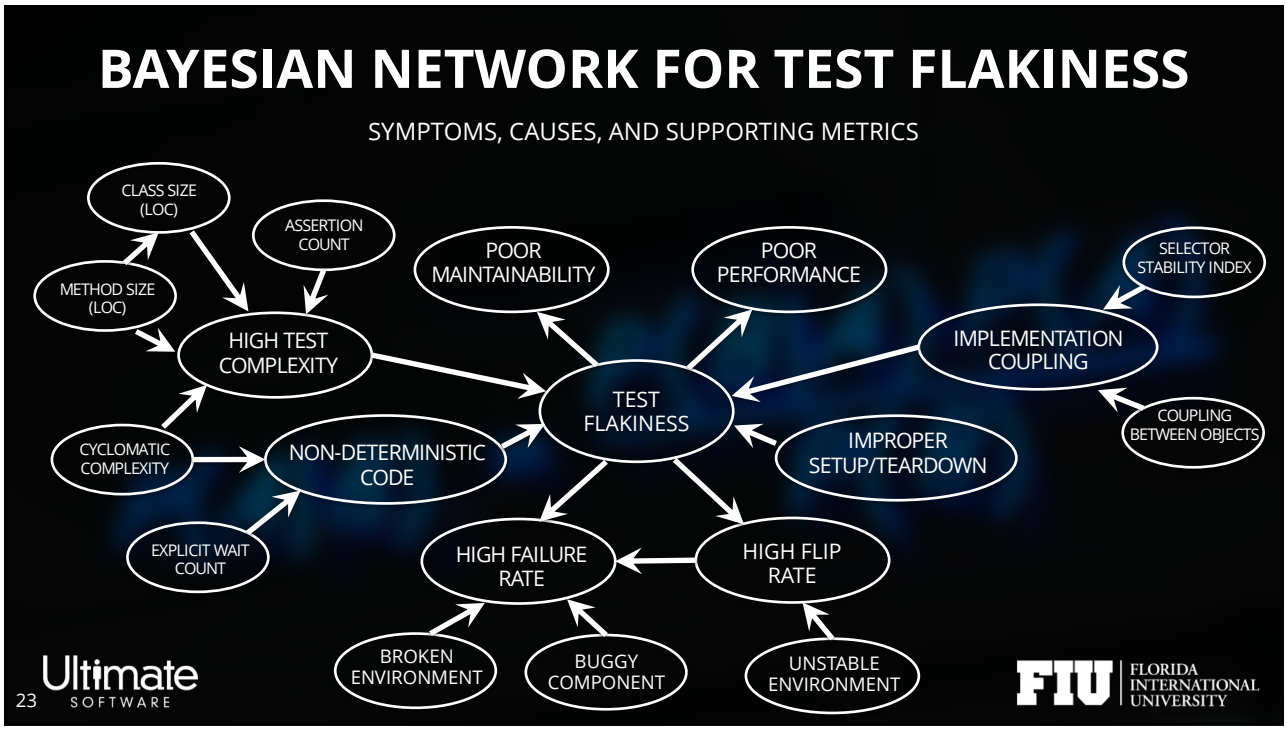
TEST FLAKINESS: AN AUTOMATION DISEASE



Ultimate SOFTWARE

22

FIU FLORIDA INTERNATIONAL UNIVERSITY


BAYESIAN NETWORK FOR TEST FLAKINESS
SYMPTOMS, CAUSES, AND SUPPORTING METRICS
CLASS SIZE (LOC)
ASSERTION COUNT
METHOD SIZE (LOC)
HIGH TEST COMPLEXITY
POOR MAINTAINABILITY
POOR PERFORMANCE
SELECTOR STABILITY INDEX
IMPLEMENTATION COUPLING
COUPLING BETWEEN OBJECTS
TEST FLAKINESS
IMPROPER SETUP/TEARDOWN
CYCLOMATIC COMPLEXITY
NON-DETERMINISTIC CODE
EXPLICIT WAIT COUNT
HIGH FAILURE RATE
HIGH FLIP RATE
BROKEN ENVIRONMENT
BUGGY COMPONENT
UNSTABLE ENVIRONMENT
Ultimate SOFTWARE
23
FIU FLORIDA INTERNATIONAL UNIVERSITY

Ultimate
SOFTWARE
People first.
Ultimate SOFTWARE
24
FIU FLORIDA INTERNATIONAL UNIVERSITY

APPLYING OUR APPROACH IN PRACTICE

CONTEXT AND ENVIRONMENT SETUP

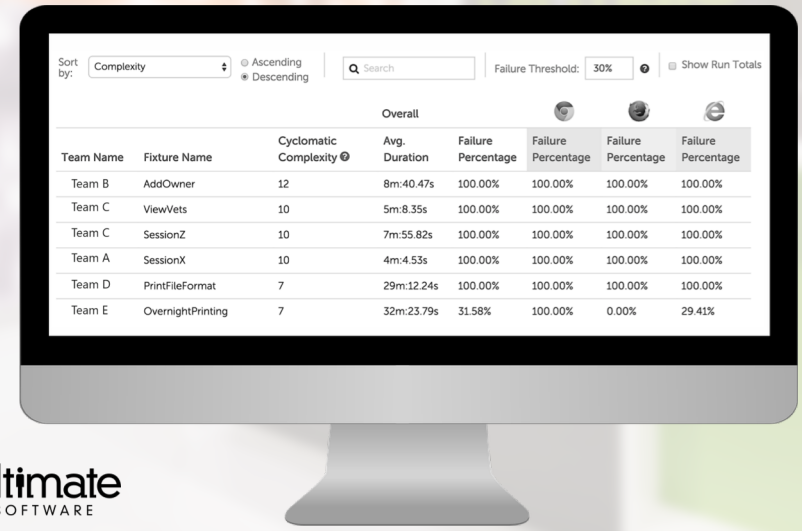
Ultimate SOFTWARE

25

FIU FLORIDA INTERNATIONAL UNIVERSITY

APPLYING OUR APPROACH IN PRACTICE

OBJECTIVES AND DATA COLLECTION TOOLING



Team Name	Fixture Name	Cyclomatic Complexity	Avg. Duration	Failure Percentage	Failure Percentage	Failure Percentage	Failure Percentage
Team B	AddOwner	12	8m:40.47s	100.00%	100.00%	100.00%	100.00%
Team C	ViewVets	10	5m:8.35s	100.00%	100.00%	100.00%	100.00%
Team C	SessionZ	10	7m:55.82s	100.00%	100.00%	100.00%	100.00%
Team A	SessionX	10	4m:4.53s	100.00%	100.00%	100.00%	100.00%
Team D	PrintFileFormat	7	29m:12.24s	100.00%	100.00%	100.00%	100.00%
Team E	OvernightPrinting	7	32m:23.79s	31.58%	100.00%	0.00%	29.41%

BUSINESS GOALS

IDENTIFY FLAKY TESTS
QUARANTINE
FIX OR THROW OUT

RESEARCH GOALS

VALIDATE MODEL
GENERATE TRAINING SET
EVALUATE APPROACH

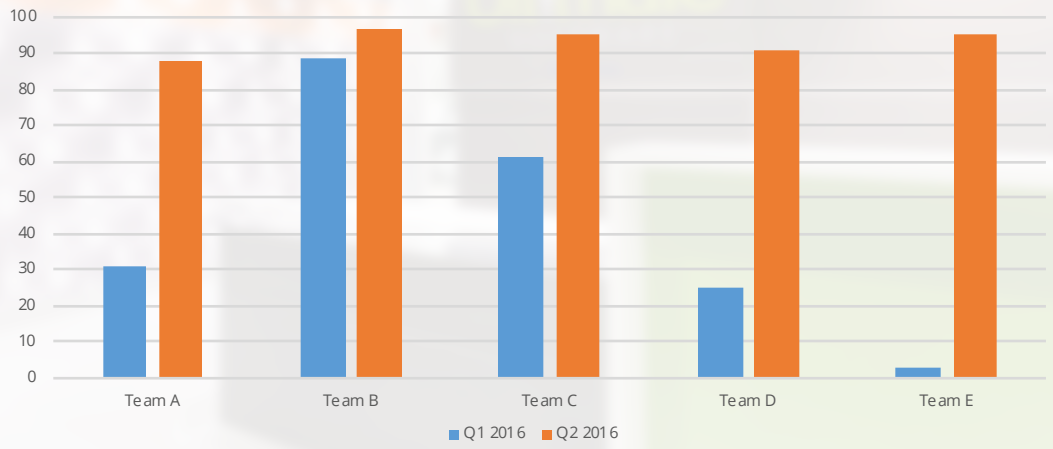
Ultimate SOFTWARE

26

FIU FLORIDA INTERNATIONAL UNIVERSITY

APPLYING OUR APPROACH IN PRACTICE

CI STABILIZATION: PRE- AND POST- TEST CLEANUP RESULTS



APPLYING OUR APPROACH IN PRACTICE

TEST FLAKINESS CLASSIFIER: TRAINING RESULTS

1 TEAM MONTH

714 TEST SUITES

26,298 OBSERVED TEST RUNS

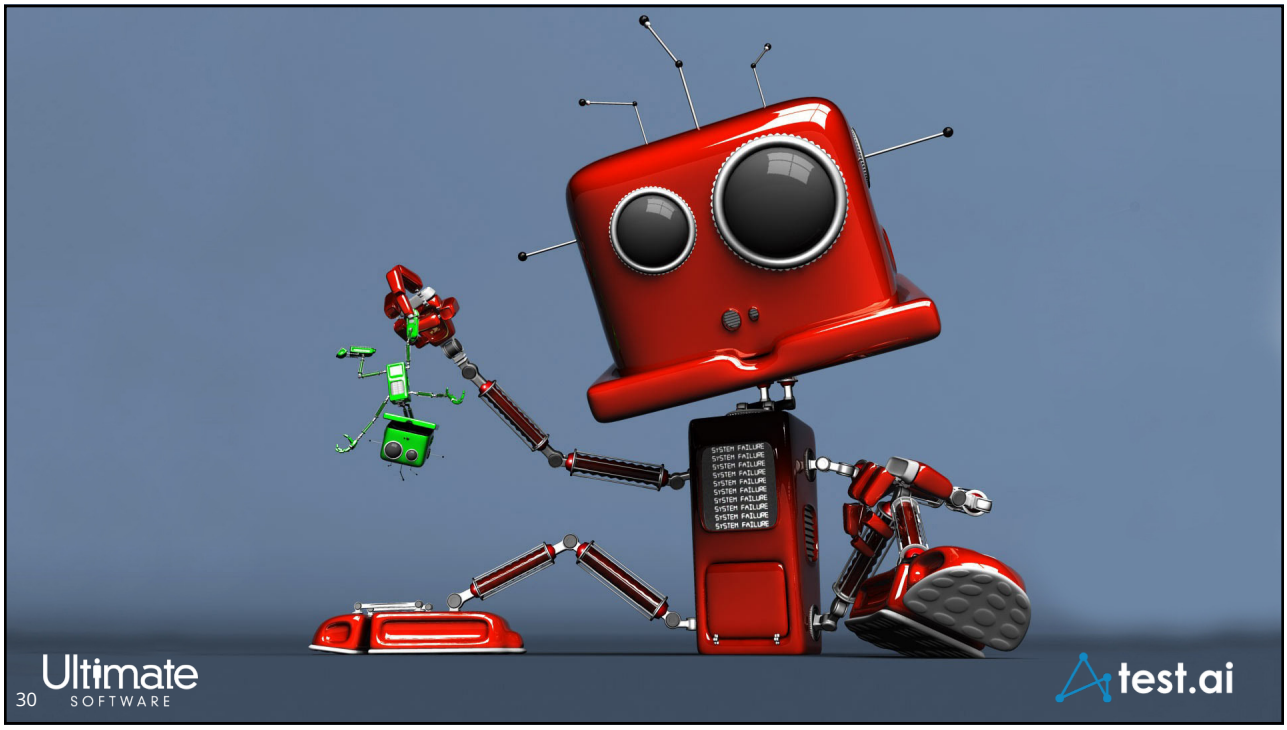
CHARACTERISTIC	ACCURACY	TRUE	FALSE
High Test Complexity	0.889	511/673	2275/2462
Large Test Fixture Size	0.822	404/486	2173/2649
Poor Performance	0.458	919/919	516/2216
Test Flakiness	0.657	188/583	1880/2552

GENIE MODELER

NEXT STEPS

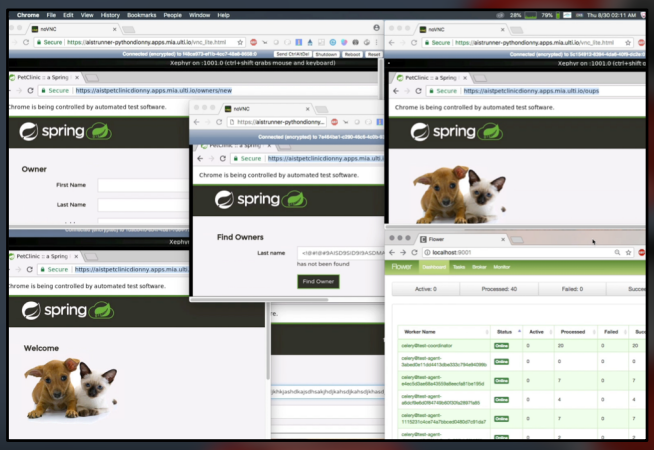
ONGOING AND FUTURE WORK

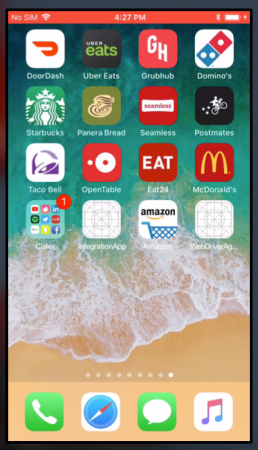
- Our test flakiness prediction tool is still in the alpha stage.
- Currently extending the tool for an beta release with development teams at Ultimate Software.
- In the future would like to conduct a comparative case study in collaboration with other organizations.
- Aspiring to build a self-healing CI/CD pipeline that learns from fixes to autocorrect flaky tests.



AI-DRIVEN TEST AUTOMATION

THE TESTING BOTS ARE ALREADY HERE!





Ultimate SOFTWARE

31

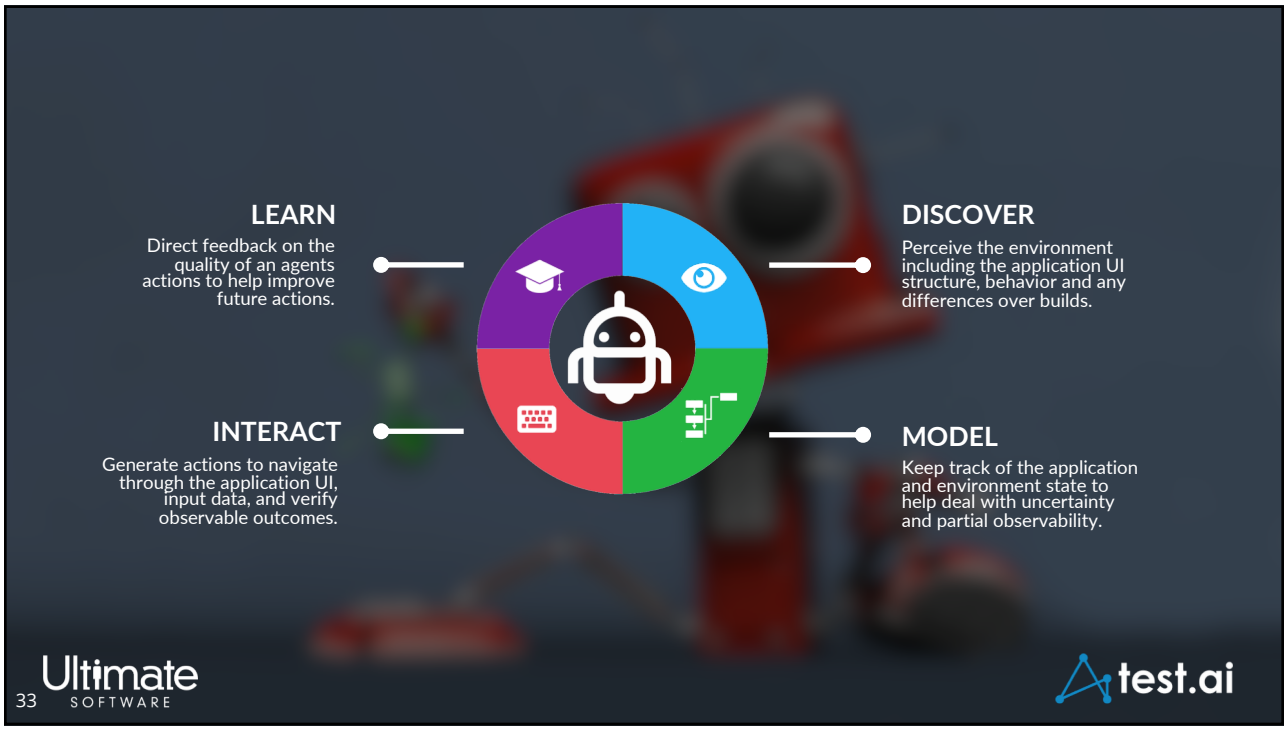
test.ai

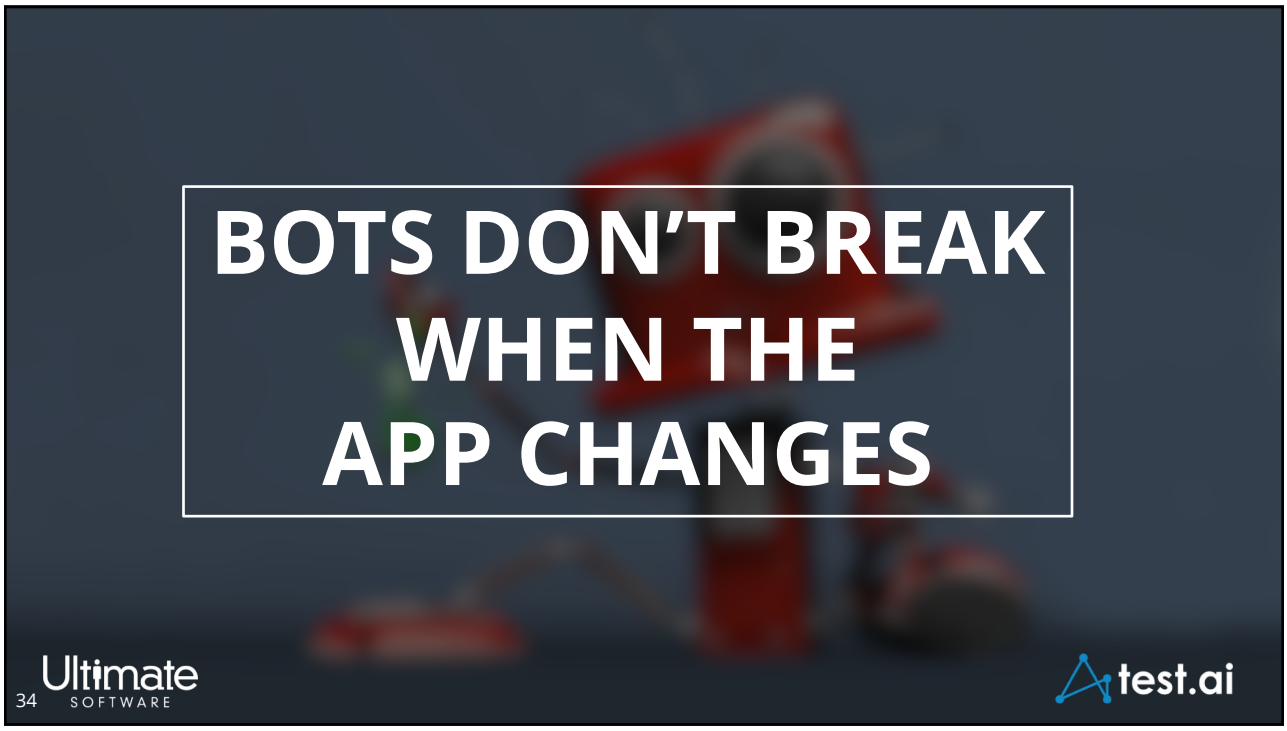
BOTS EXPLORE AND TEST LIKE HUMANS

Ultimate SOFTWARE

32

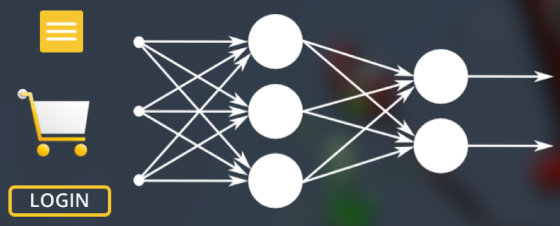
test.ai


LEARN
Direct feedback on the quality of an agents actions to help improve future actions.
DISCOVER
Perceive the environment including the application UI structure, behavior and any differences over builds.
INTERACT
Generate actions to navigate through the application UI, input data, and verify observable outcomes.
MODEL
Keep track of the application and environment state to help deal with uncertainty and partial observability.
Ultimate SOFTWARE
33
test.ai

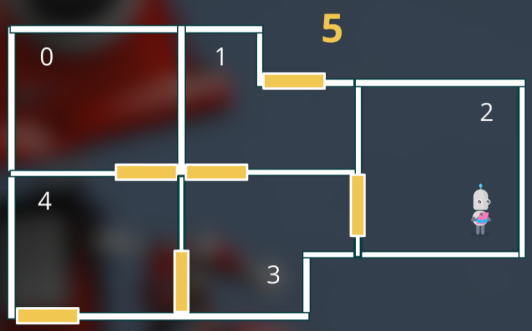

BOTS DON'T BREAK WHEN THE APP CHANGES
Ultimate SOFTWARE
34
test.ai

WHY AUTOMATION WITH BOTS IS STABLE

BOTS CAN ADAPT TO STRUCTURAL AND BEHAVIORAL CHANGES



Find Elements via Image Recognition



Find New Paths via Learning

test.ai

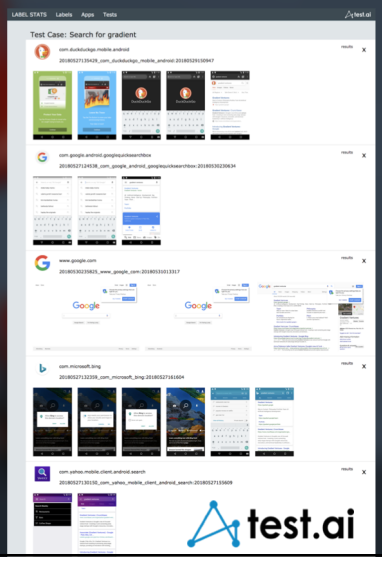
WHY AUTOMATION WITH BOTS IS STABLE

BOT TEST SPECIFICATIONS ARE ABSTRACT AND REUSABLE

```
Test Name: Search for Gradient
Tags: "Single Word", "Search"
Description: Perform a single-word search
             for the term "Gradient"

Steps: Get to Search
StepType: Navigation
Labels: "Search Button"

Step: Enter Search Text
StepType: TextInput
Text: "Gradient"
Labels: "Search Box"
```



test.ai

AVAILABLE SOONER THAN YOU THINK ?

test.ai

```
driver.find_element(By.LABEL, “search_button”)
```


test.ai

39

THANK YOU!

RESEARCH TEAM @ULTIMATE/FIU

DIONNY SANTIAGO
JUSTIN PHILLIPS
PETER CLARKE
DAVID ADAMO
KEITH BRIGGS
PATRICK ALT

TEST LEADS @ULTIMATE

PHILIP DAYE
DAVID SKIRVIN
ANDREW POMERANS
JAIRO PAVA
YUNIOR BETANCOURT
EDUARDO PENA

THE TEAM @TEST.AI

JASON ARBON
JUSTIN LIU
GREGG TABOT
JOANNE SENG
AND MORE…

Ultimate SOFTWARE
AISTA

40