# CUTTING INTEL PRODUCTS TIME TO MARKET USING BIG DATA & MACHINE LEARNING

Nufar Gaspar

Advanced Analytics, Intel

# ABOUT US: ADVANCED ANALYTICS INTEL

**Machine Learning & Big Data Experts**

**Global Organization**

**100 Employees (all in Israel)**

**30 Data Scientist (PHD & MSc)**

**Solve strategic Business Problems**

**Leverage Analytics to Grow Revenue**

**Machine Learning Products Planning**

# ABOUT US: ADVANCED ANALYTICS INTEL

Machine Learning &
Big Data Experts

Global
Organization

100 Employees
(all in Israel)
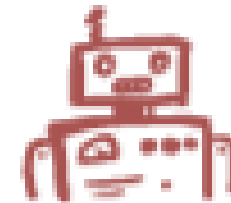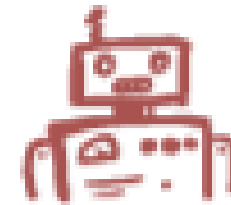
30 Data Scientist
(PHD & MSc)

Solve strategic
Business Problems

Leverage Analytics
to Grow Revenue

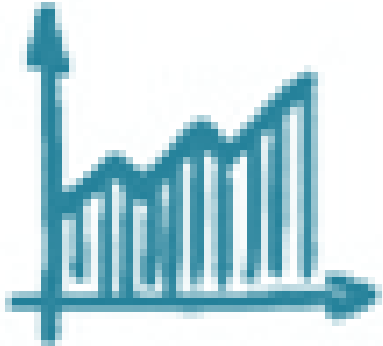Machine Learning
Products Planning

# ABOUT ME



Lead a Team of Data Scientist and Big Data Developers to Create Solutions for Intel R&D Organizations

Ben-Gurion Industrial Engineering MSc. Grad.

# ADDITIONAL CAPABILITIES EXAMPLES



**Active Learning & Advanced Surface Fitting to Project Design Parameters**
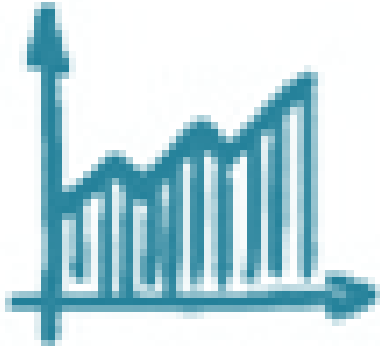
**Embedded Online Machine Learning Algorithms for Chip Performance Improvements**

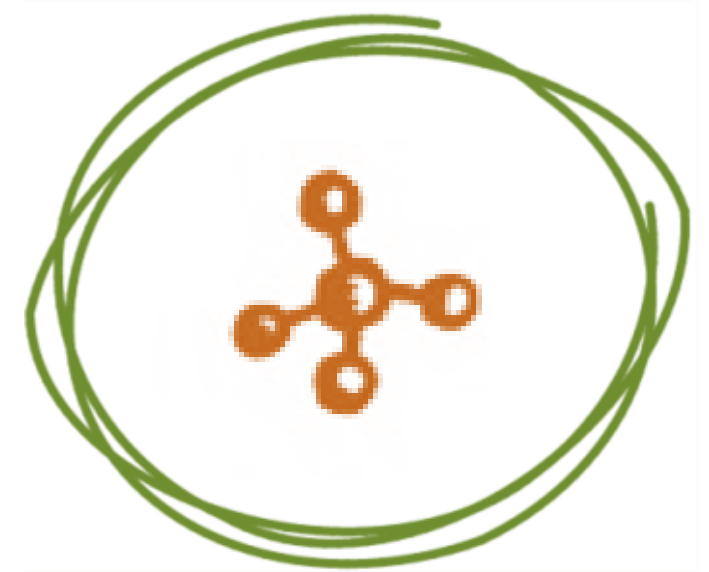**Pattern Recognition for Improving Products Quality**

# ADDITIONAL CAPABILITIES EXAMPLES

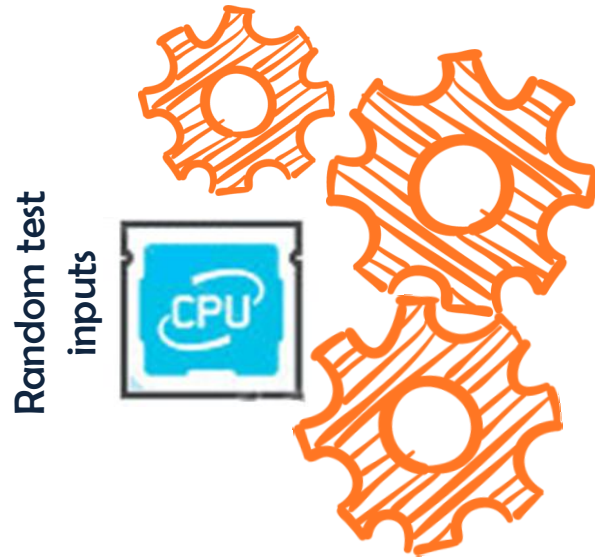**Active Learning & Advanced Surface Fitting to Project Design Parameters**

**Embedded Online Machine Learning Algorithms for Chip Performance Improvements**

**Pattern Recognition for Improving Products Quality**

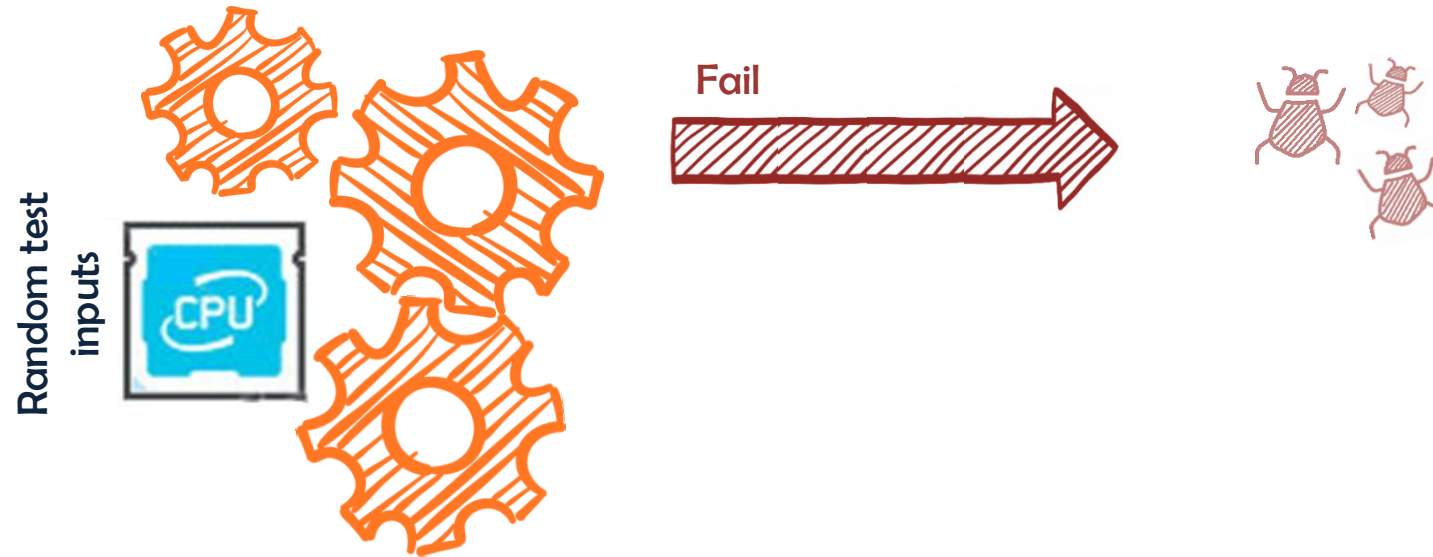# PROBLEM STATEMENT: HIDDEN BUGS FOR LACK OF VALIDATION

**CPU Validation Process:**

Random test inputs

CPU

**Many thousands of random tests**

# PROBLEM STATEMENT: HIDDEN BUGS FOR LACK OF VALIDATION

**CPU Validation Process:**



Random test inputs

Fail

Many thousands of random tests

# PROBLEM STATEMENT: HIDDEN BUGS FOR LACK OF VALIDATION

**CPU Validation Process:**
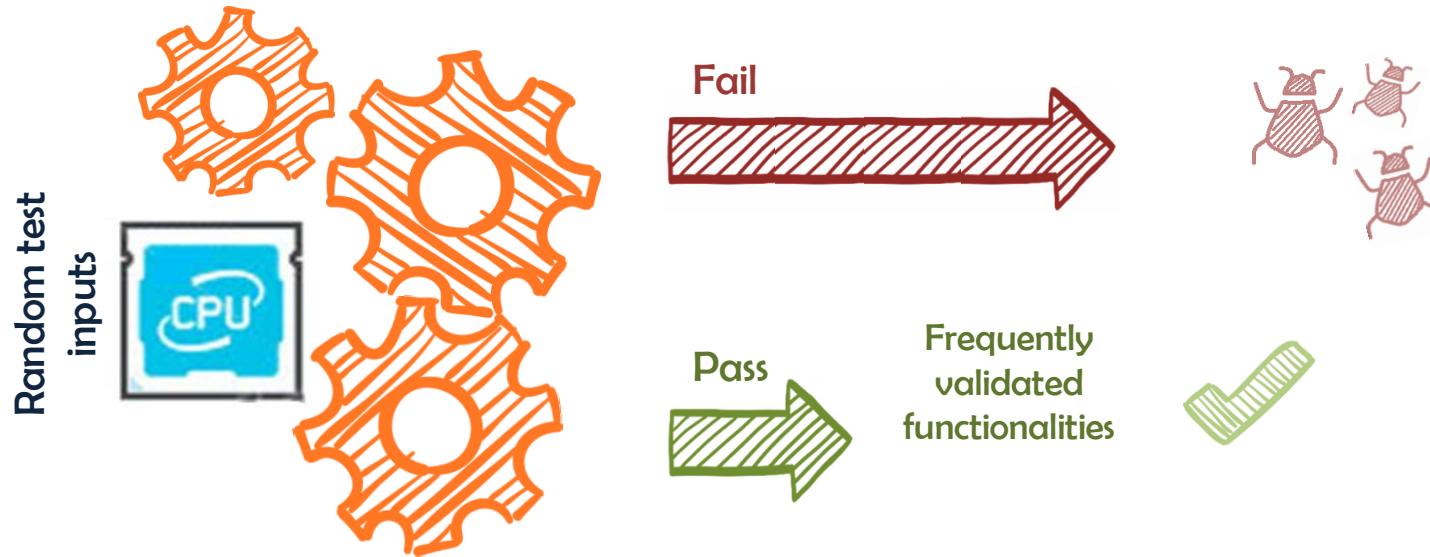


Random test inputs

Fail

Pass

Frequently validated functionalities

**Many thousands of random tests**

# PROBLEM STATEMENT: HIDDEN BUGS FOR LACK OF VALIDATION

**CPU Validation Process:**

Random test inputs

Fail
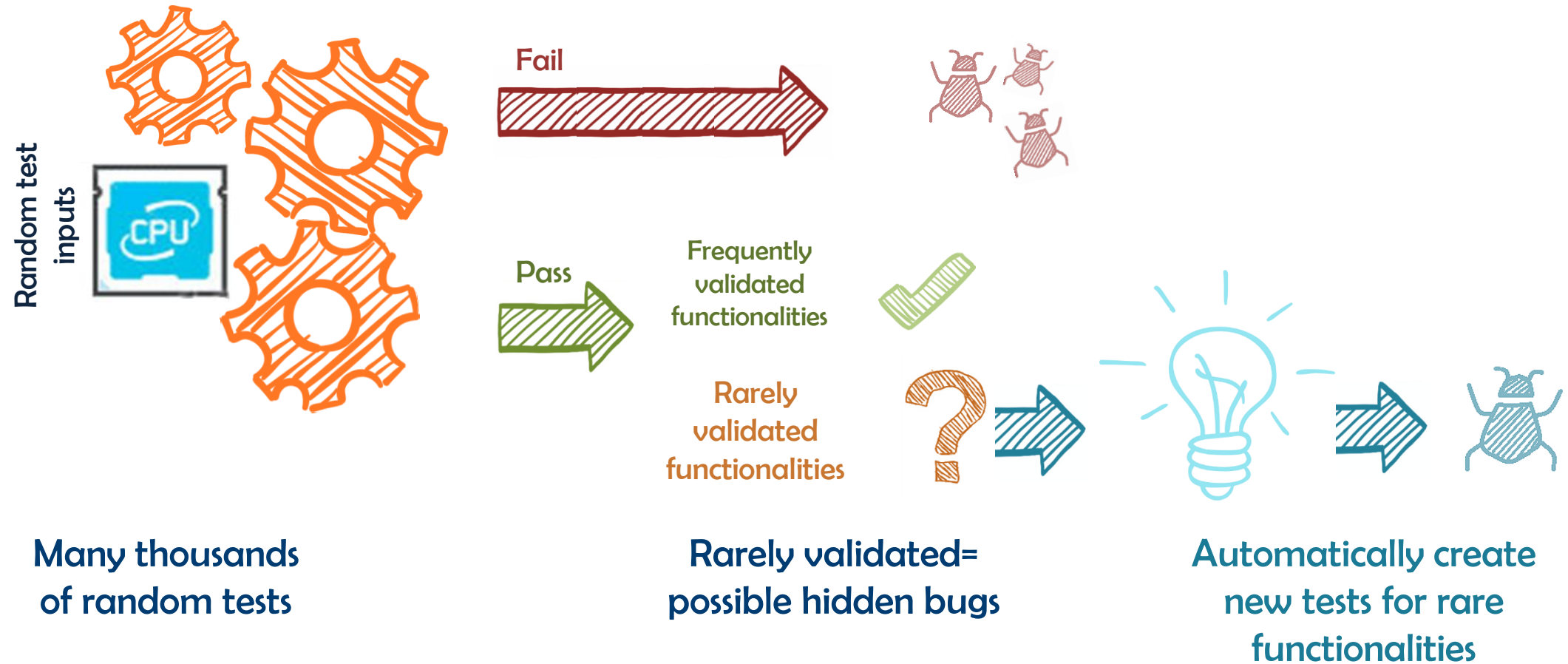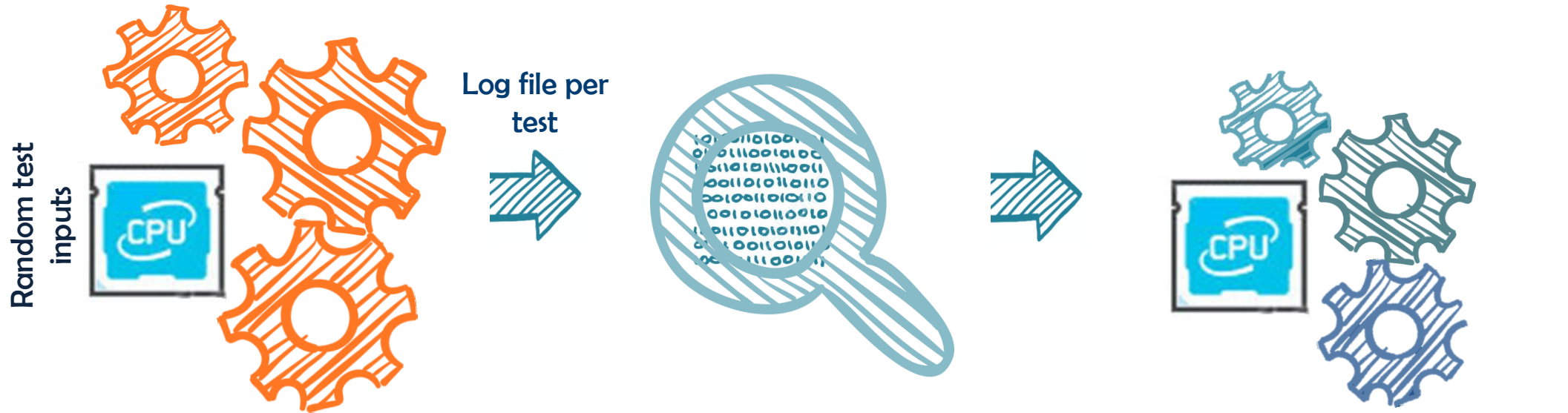
Pass

Frequently validated functionalities

Rarely validated functionalities

**Many thousands of random tests**

**Rarely validated= possible hidden bugs**
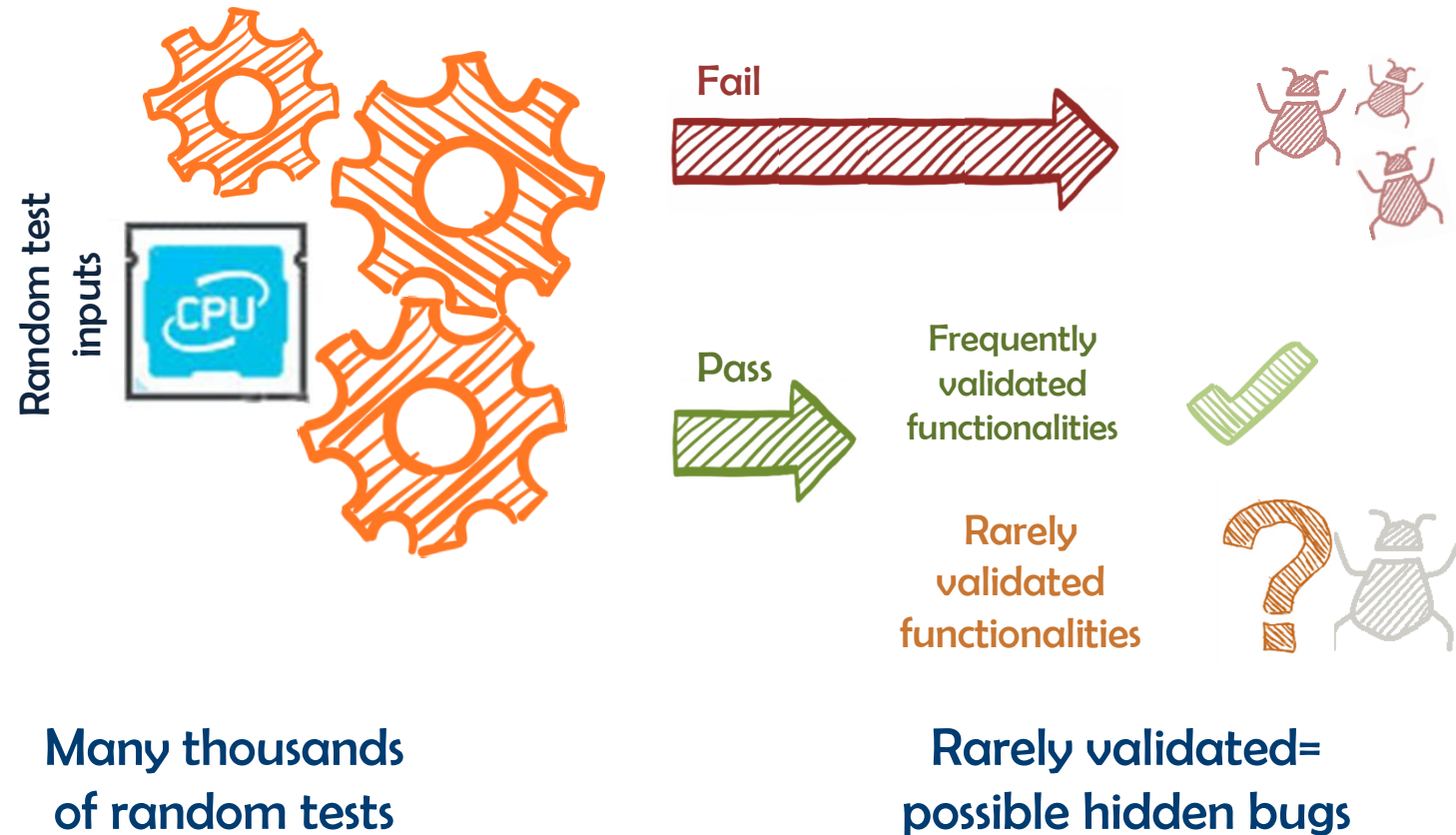
# SOLUTION: AUTOMATIC TEST GENERATION

**CPU Validation Process:**



Random test inputs

Fail

Pass

Frequently validated functionalities

Rarely validated functionalities

**Many thousands of random tests**

**Rarely validated= possible hidden bugs**

**Automatically create new tests for rare functionalities**

# SOLUTION: AUTOMATIC TEST GENERATION

**Fully Automated Process:**

Random test inputs

Log file per test

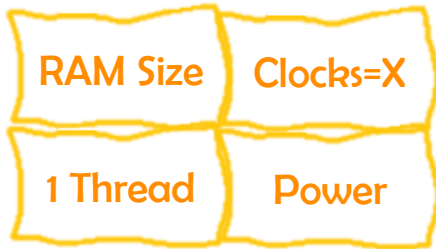**Many thousands of random tests**

**Pattern Recognition Across Thousands of Tests Inputs & outcomes**
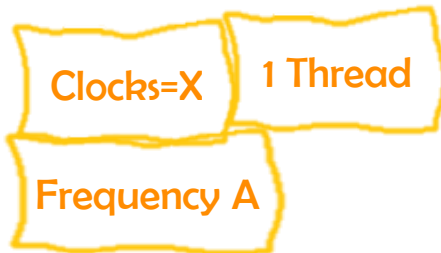
**Run New Tests for Rare Functionalities**

# DATA: TEST RECORDINGS – INPUTS AND OUTCOMES

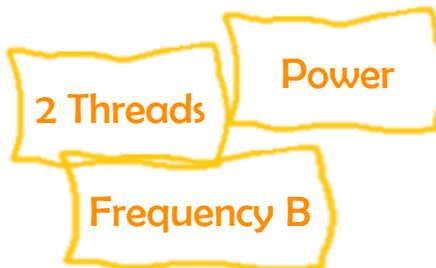

**Test Inputs**

**Test Recordings**

Test 1:
- RAM Size
- Clocks=X
- 1 Thread
- Power

A B F G G G A ⭐ A F B R C

10-200 records

Test 2:
- Clocks=X
- 1 Thread
- Frequency A

C B A C G F A R F ⭐ G G A

Test 3:
- 2 Threads
- Power
- Frequency B

G A C B G A F R F R F B G

Many Thousands Records

Test...

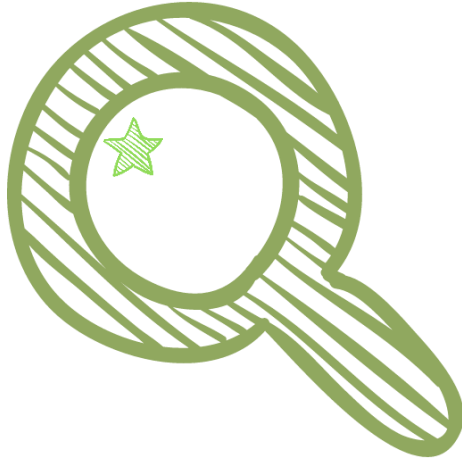⭐ Functionality Validated
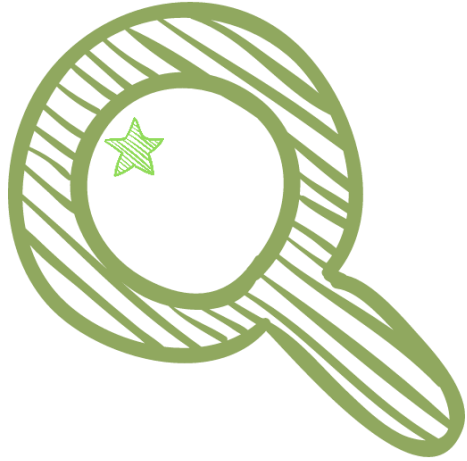
# CHALLENGES



Highly Imbalanced Data



Magnitude: Many thousands of tests, thousands of tests inputs



Evaluation is not straight forward (hidden data)

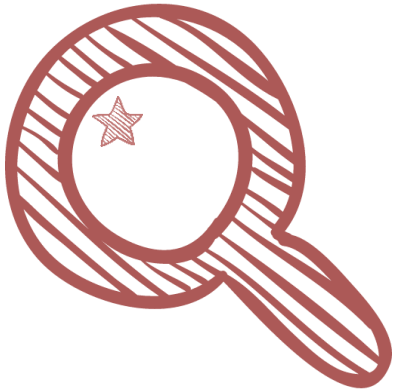# CHALLENGES SOLUTIONS

**Highly Imbalanced Data**

Magnitude: Many thousands of tests, thousands of tests inputs
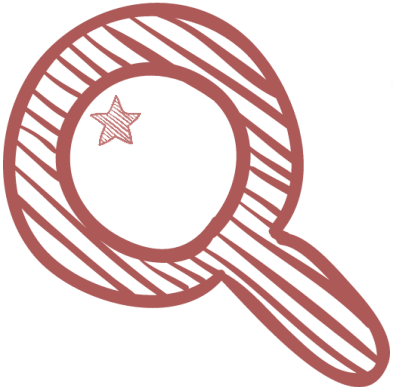
Evaluation is not straight forward (hidden data)

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE



**Highly Imbalanced:**
0.02% Positive Examples

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE



Highly Imbalanced:
0.02% Positive Examples

Massive Down Sampling

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE

Classifier Inference Approach:

Random Forest

Inference: Feature Importance

Highly Imbalanced:
0.02% Positive Examples

Massive Down Sampling

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE

**Classifier Inference Approach:**

Random Forest

Inference: Feature Importance

**Filter Approach:**

Correlation

Inference: Positive Connection Only
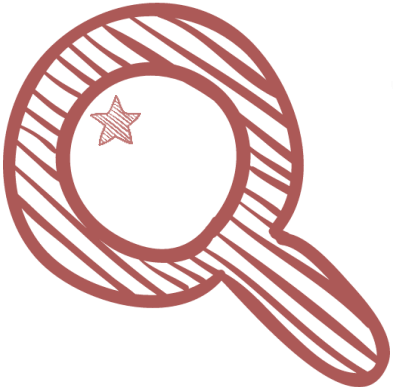
Massive Down Sampling

Highly Imbalanced:
0.02% Positive Examples

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE

**Classifier Inference Approach:**

Random Forest

Inference: Feature Importance

**Highly Imbalanced:**
0.02% Positive Examples

Massive Down Sampling

**Filter Approach:**

Correlation

Inference: Positive Connection Only

**Frequent Item-Set Approach :**

Apriori/ FP-Growth Algorithms

Recall as Obj. Function

# IMBALANCED DATA SOLUTION: DOWN SAMPLING + INFERENCE

**Classifier Inference Approach:**

Random Forest + Inference: Feature Importance

Highly Imbalanced: 0.02% Positive Examples

Massive Down Sampling +

**Filter Approach:**

Correlation + Inference: Positive Connection Only

Ensemble of Algorithms

**Frequent Item-Set Approach :**

Apriori/ FP-Growth Algorithms + Recall as Obj. Function
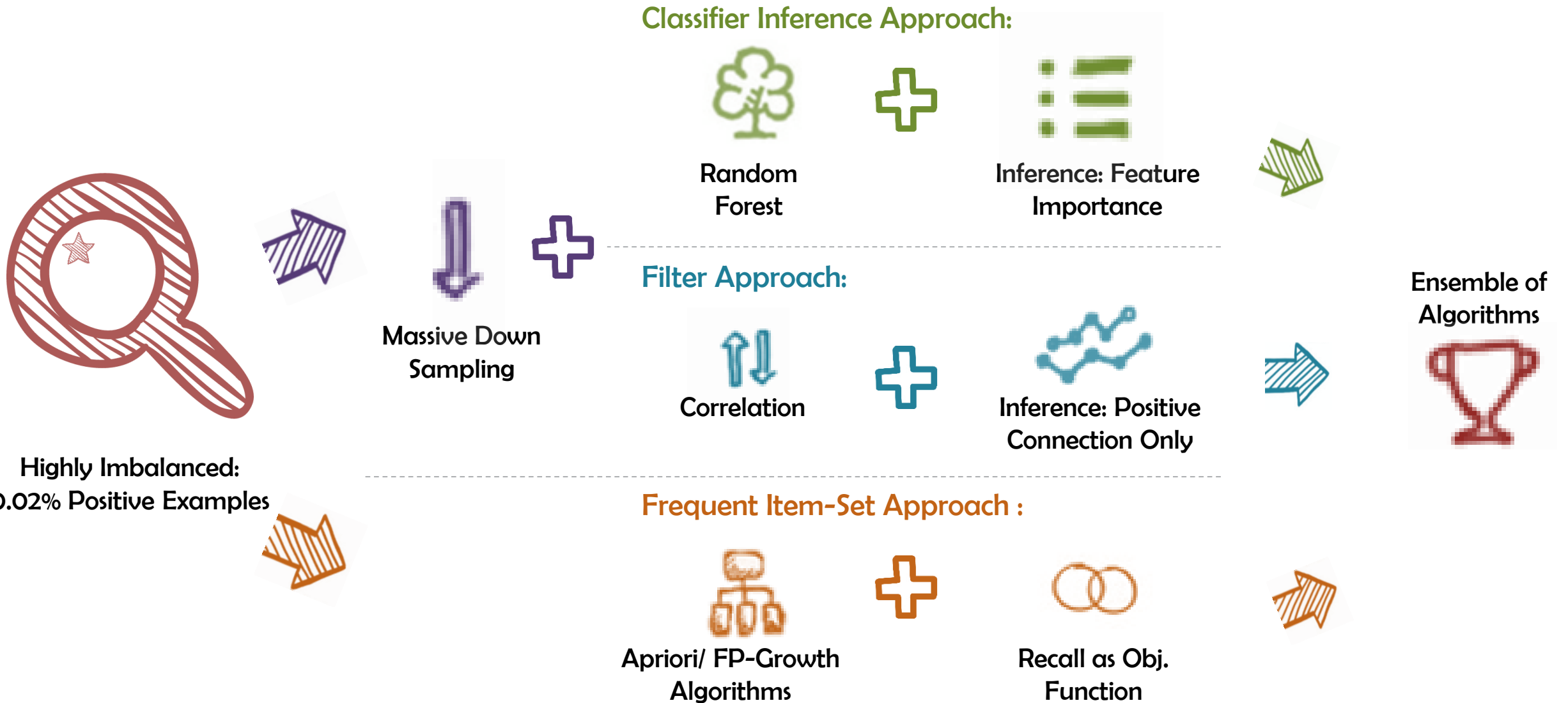
# CHALLENGES SOLUTIONS
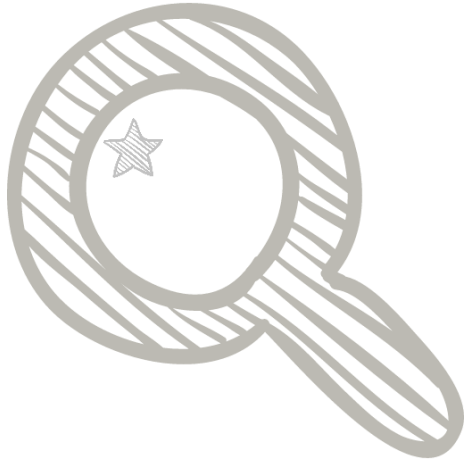
Highly Imbalanced Data

**Magnitude: Many thousands of tests, thousands of tests inputs**

Evaluation is not straight forward (hidden data)

# MAGNITUDE SOLUTION: SPARK BASED BIG DATA PLATFORM
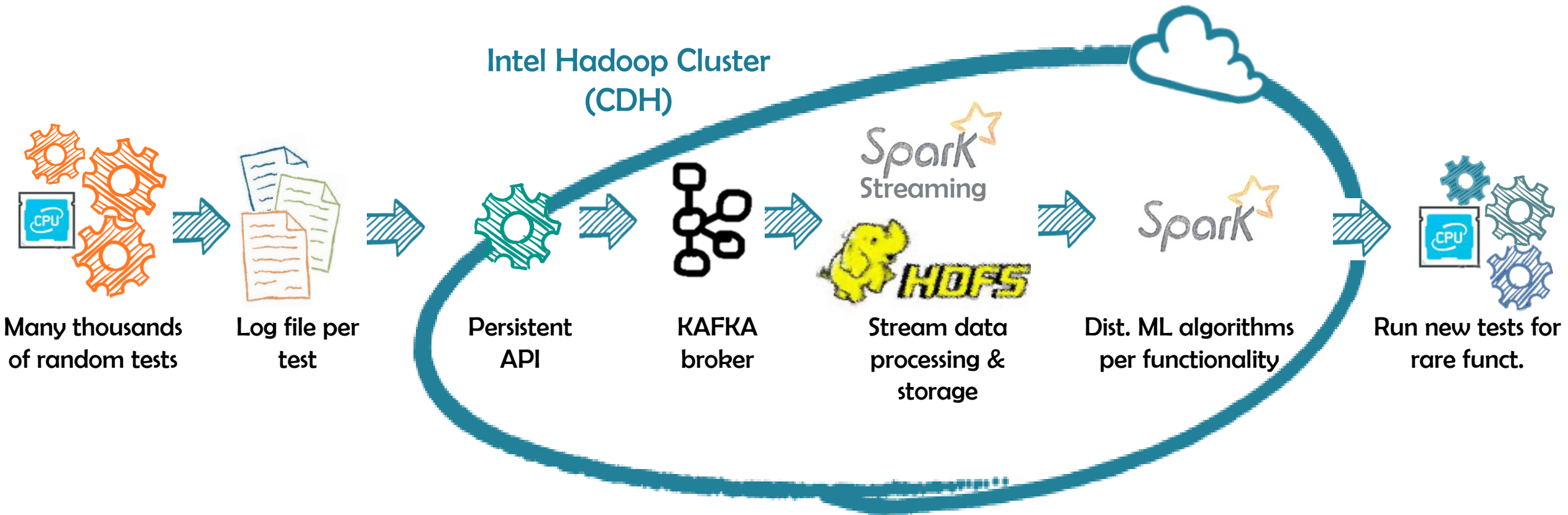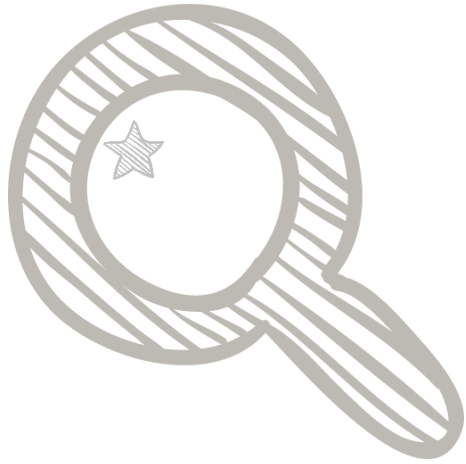


Intel Hadoop Cluster (CDH)

Many thousands of random tests → Log file per test → Persistent API → KAFKA broker → Stream data processing & storage → Dist. ML algorithms per functionality → Run new tests for rare funct.

# CHALLENGES SOLUTIONS

Highly Imbalanced
Data

Magnitude: Many
thousands of tests,
thousands of tests inputs

Evaluation is not
straight forward
(hidden data)

# EVALUATION SOLUTION: CLOSED LOOP & MEASUREMENT SEARCH

Evaluation is not
straight forward
(hidden data)

# EVALUATION SOLUTION: CLOSED LOOP & MEASUREMENT SEARCH

**Evaluation is not straight forward (hidden data)**

**Ensemble of Algorithms**

**Multiple Results**

# EVALUATION SOLUTION: CLOSED LOOP & MEASUREMENT SEARCH

**Evaluation is not straight forward (hidden data)**

**Ensemble of Algorithms**

**Multiple Results**

**Run Evaluation Loop**

# EVALUATION SOLUTION: CLOSED LOOP & MEASUREMENT SEARCH



Evaluation is not straight forward (hidden data)

Ensemble of Algorithms

Multiple Results

Run Evaluation Loop

Identify a Correlative measurement & threshold

# EVALUATION SOLUTION: CLOSED LOOP & MEASUREMENT SEARCH

Evaluation is not straight forward (hidden data) → Ensemble of Algorithms + Multiple Results → Run Evaluation Loop → Identify a Correlative measurement & threshold

F measure

## Results are very good (exact numbers confidential)
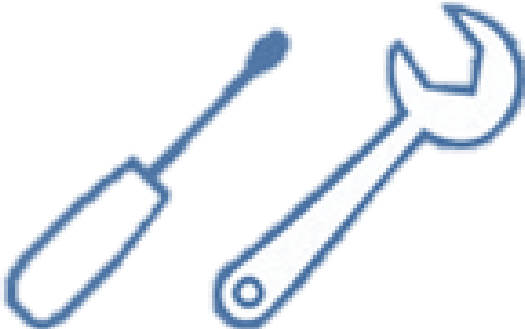
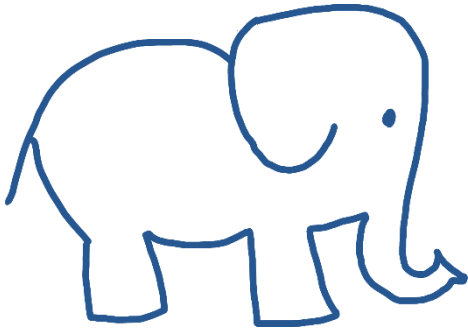Significant More Validation for Rare Func.

New Bugs are Being Found Earlier

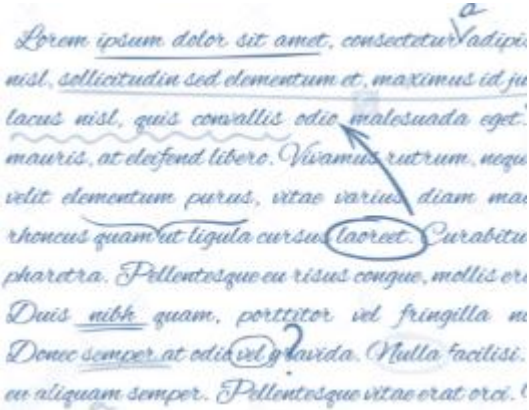# ADDITIONAL DIRECTIONS & NEXT STEPS

**Improve Existing Models and Post-Processing**

**Harness the Power of Big Data for Exhaustive Search**

**Approach as a Text Mining Problem**

THANK YOU

intel®