RSAConference2022
San Francisco & Digital | June 6 – 9

TRANSFORM

SESSION ID: **OST-R05**

xGitGuard: ML-Based Secret Scanner for GitHub

Bahman Rashidi, Ph.D.

Director , Cybersecurity & Privacy Engineering Research
Comcast

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA® Conference, RSA Security LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

Overview

COMCAST CYBERSECURITY

All images in this slide are from : https://unsplash.com/

Existing Approaches

- **TruffleHog**[1]
  - Detect high-entropy text
- **Gitguardian**[1]
  - Detects secrets using Regex classifiers
- **EarlyBird**
  - Detects specific patterns in files
- **NightWatch**
  - Entropy-based detection

Limitations with existing works:

- User decides the repos
- Regex-based detection
- Relying on entropy
- Exclusively cover API keys
- Scalability

1. S. Lounici, "Optimizing Leak Detection in Open-source Platforms with Machine Learning Techniques," in 7th International Conference on Information Systems Security and Privacy, Vienna, Austria, 2021.

RSAConference2022

xGitGuard

Introduction

xGitGuard

DETECT EXPOSED SECRETS

GITHUB

AI ENGINE

ACCURATE & SCALABLE

COMCAST CYBERSECURITY

What Secrets?

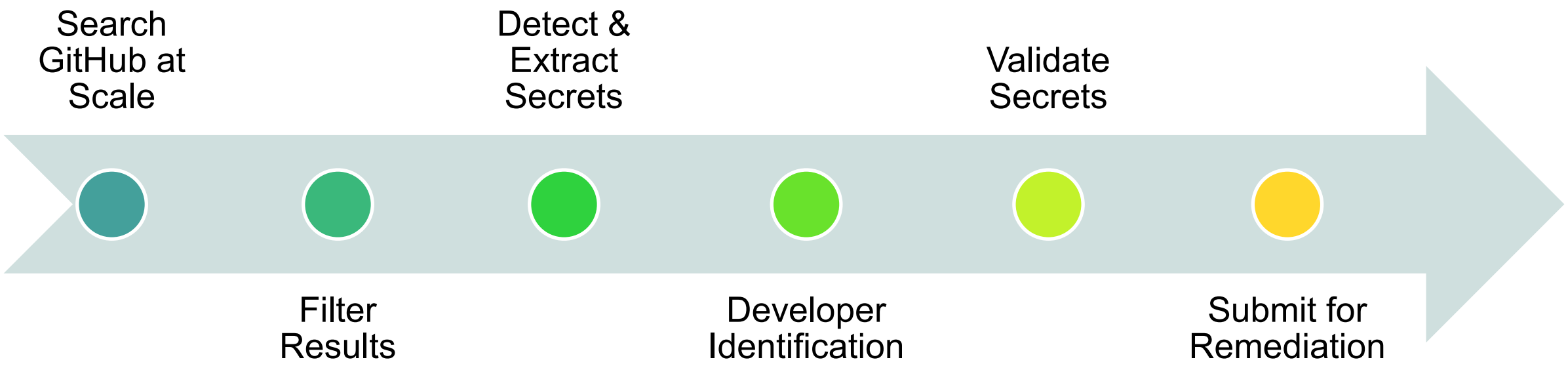
CREDENTIALS

- Username & passwords
- Server credentials
- Account credentials

TOKENS / KEYS

- Service API tokens (AWS, Azure, etc)
- Encryption keys

Workflow

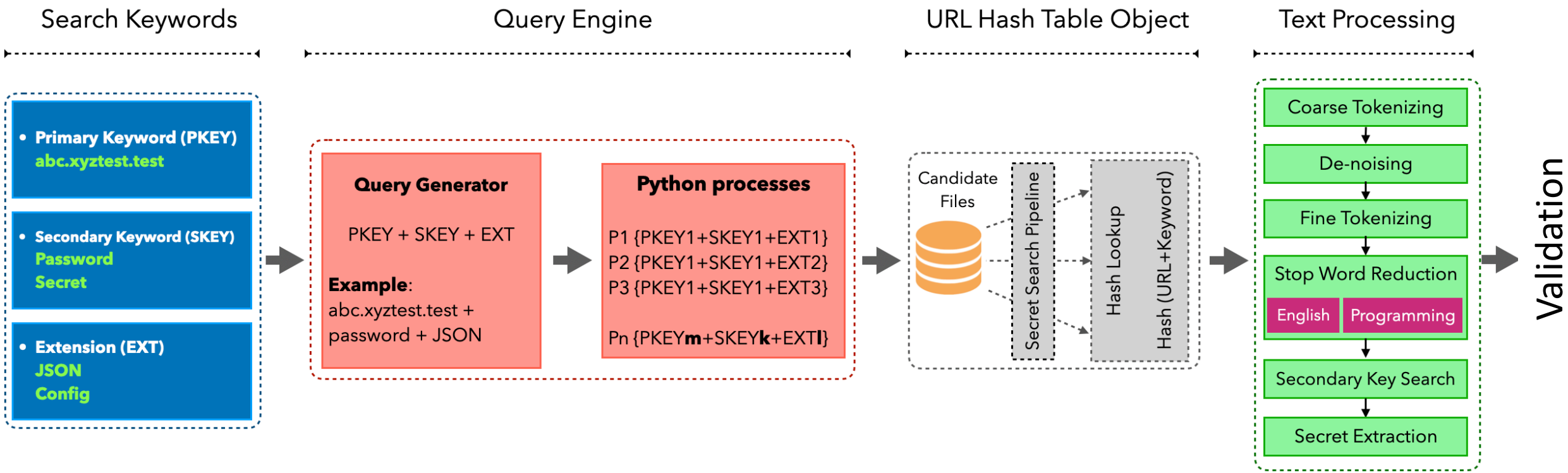

Search GitHub at Scale
Filter Results
Detect & Extract Secrets
Developer Identification
Validate Secrets
Submit for Remediation

RSAConference2022

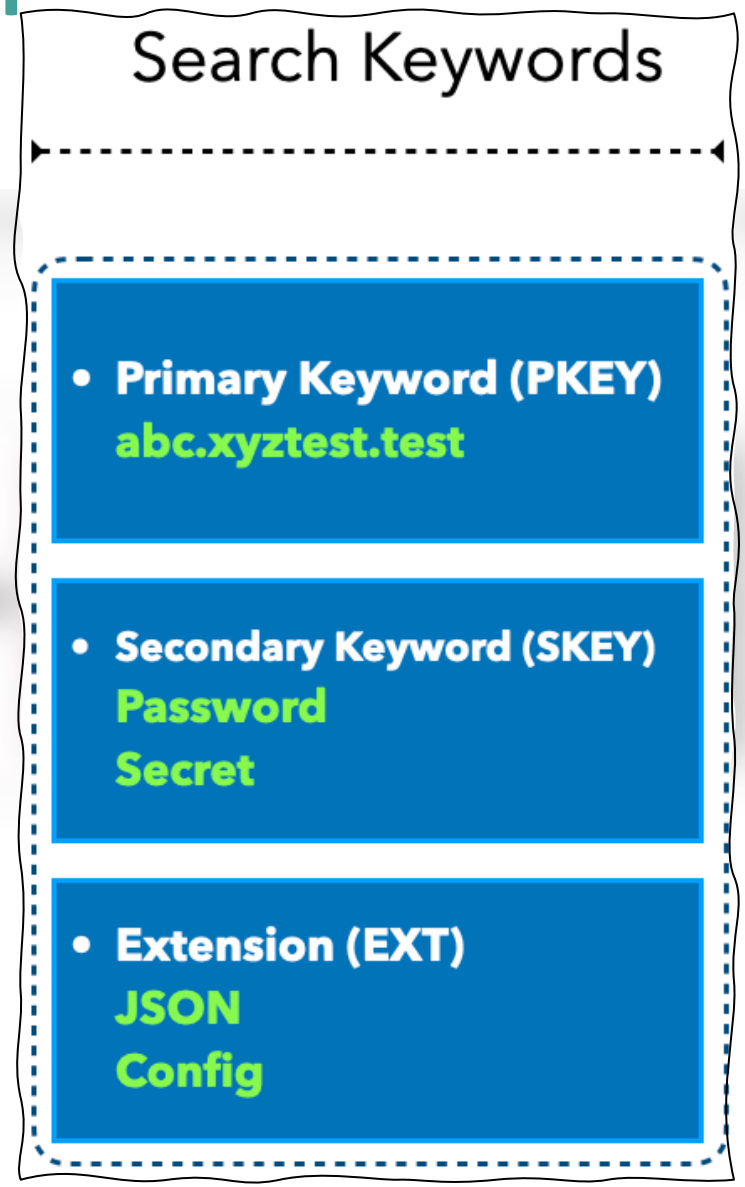
Architecture

Main components

Credential Detection



Credential Detection



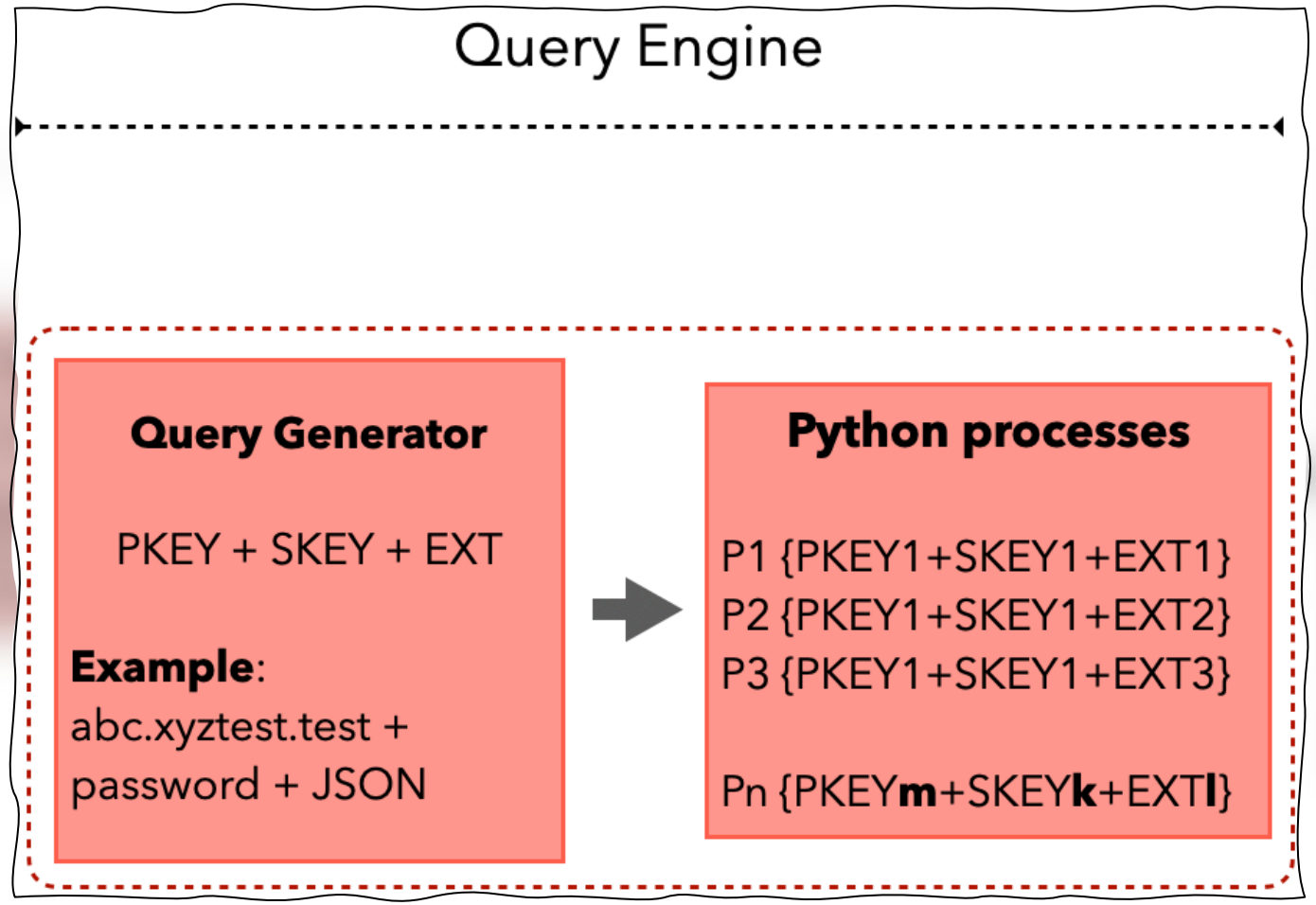
Credential Detection

```
19  Copyright 2021 abc.xyztest.test
20  import os
21  import sys
22
23  import numpy as np
24  from sklearn.feature_extraction.text import CountVectorizer
25
26  MODULE_DIR = os.path.dirname(os.path.realpath(__file__))
27  parent_dir = os.path.dirname(MODULE_DIR)
28  sys.path.append(parent_dir)
29
30  from utilities.file_utilities import read_yaml_file, read_csv_file
31  password = “abcd123”
32
33  logger = logging.getLogger("xgg_logger")
```

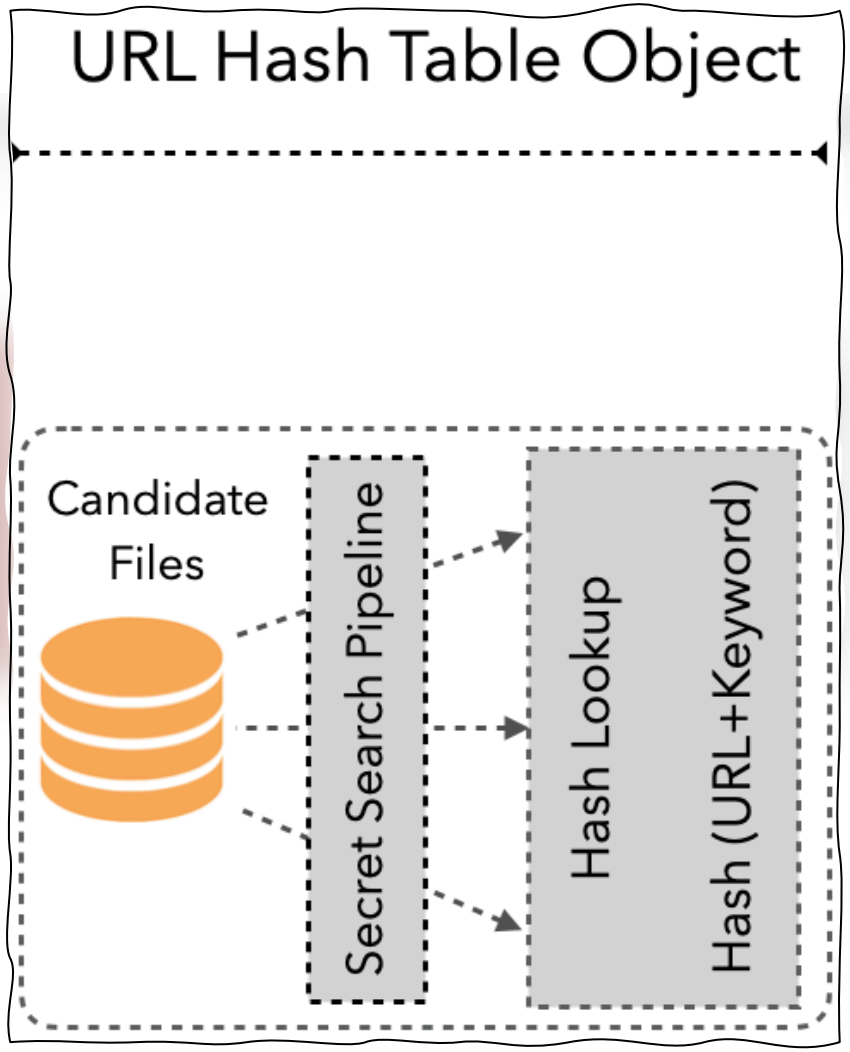
Primary Keyword

Secondary Keyword

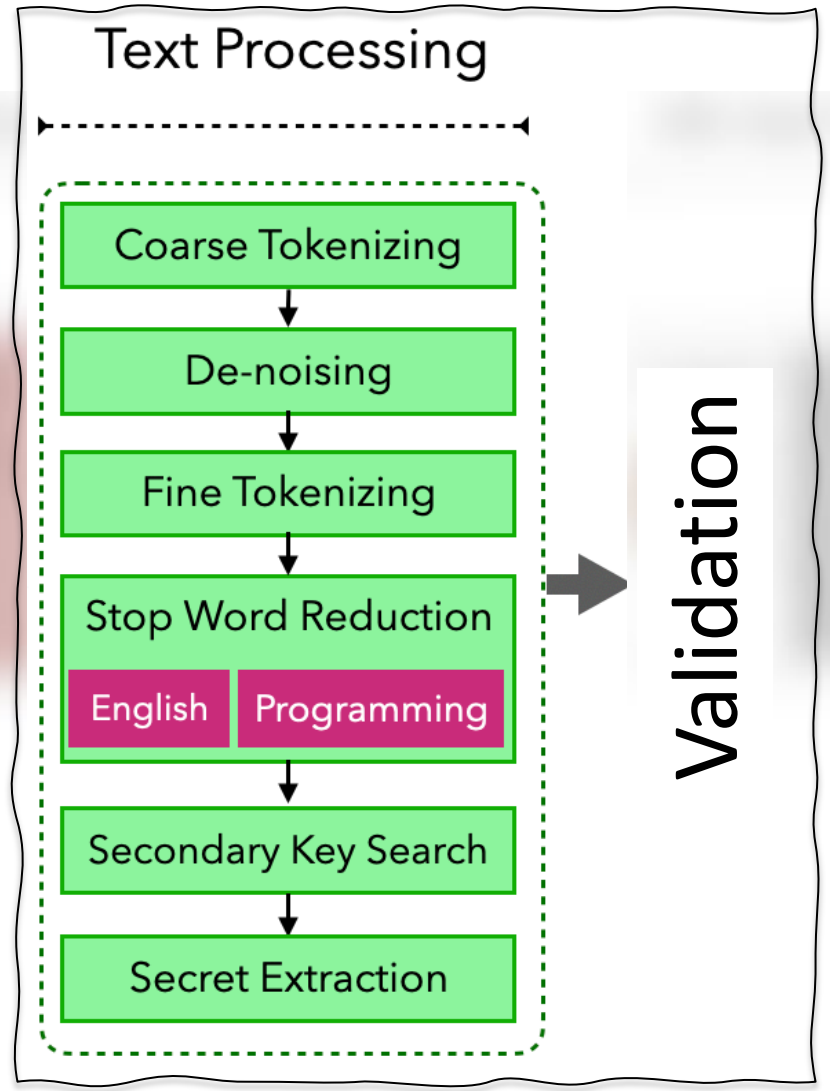
Credential Detection



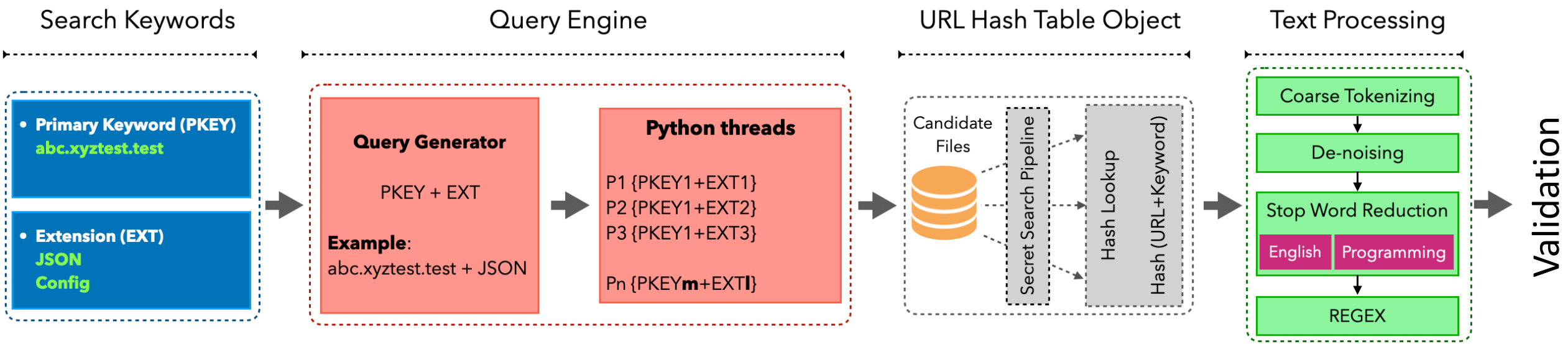
Credential Detection



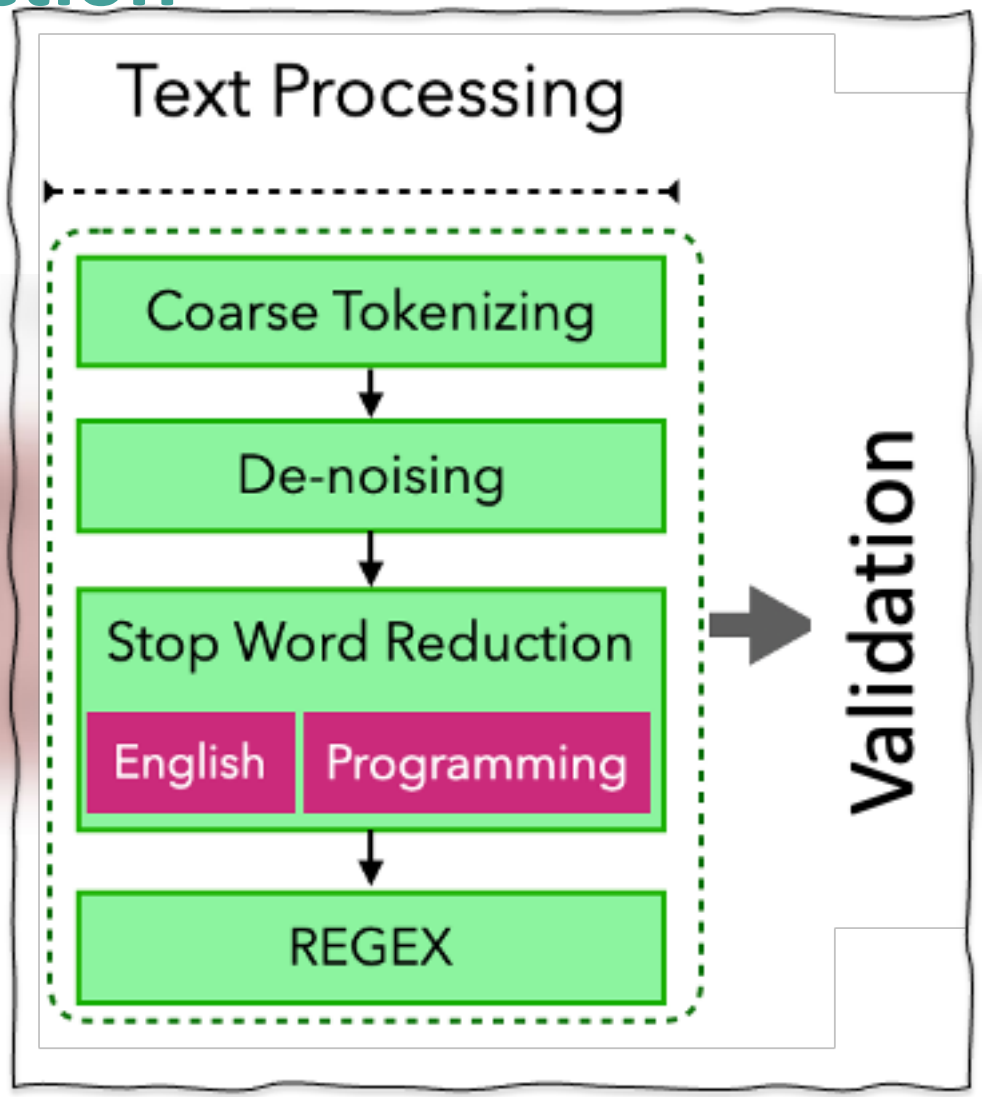
Credential Detection



Token/Key Detection



Token/Key Detection

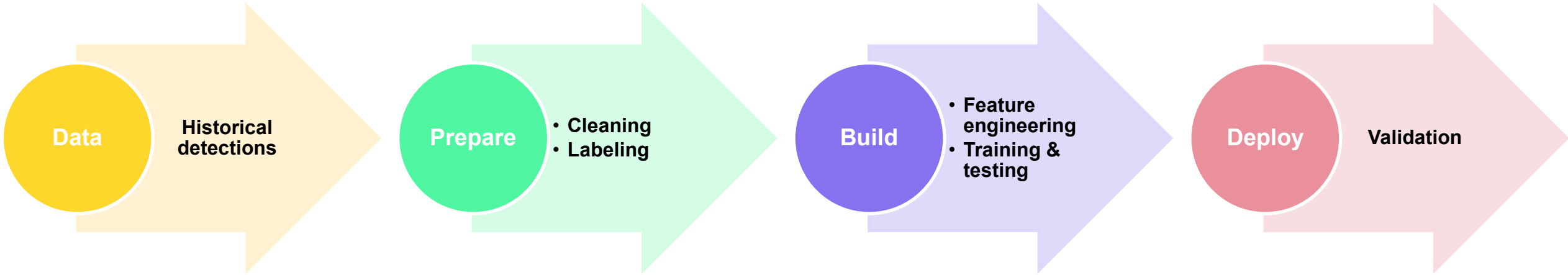


RSAConference2022

ML-based Validation Model

Validation model details

Validation Model



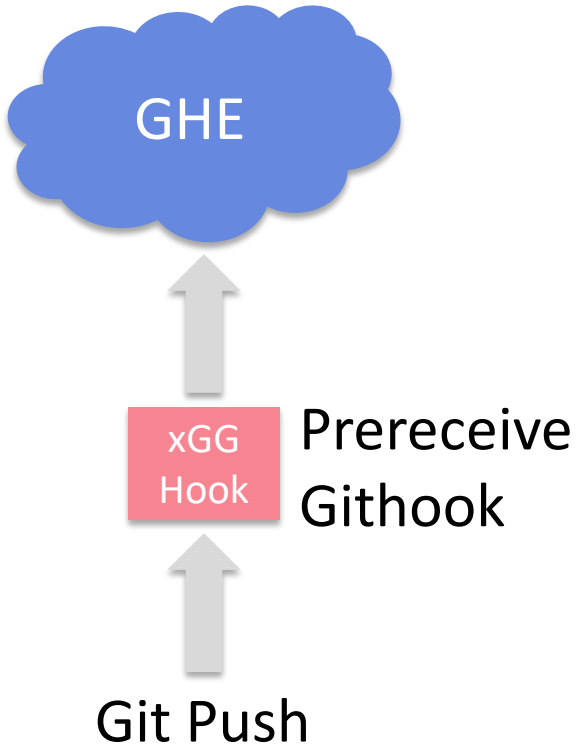
- Ngrams similarity to TPs
- Ngrams similarity to FPs
- Context-based
  - Line contains secondary keyword
  - Extension
  - Language
- Textual features

Use and Deployments

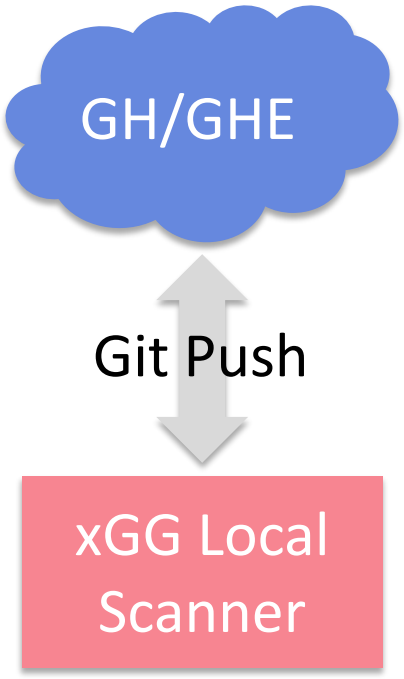
- xGitGuard is fully open-sourced:
- https://github.com/Comcast/xGitGuard
- Documented with details:
  - Installation
  - Configuration
  - Different deployments are covered

Use and Deployments

(a) Live scanner



(b) Deploy as a Githook



(c) Locally scan repos and files

Apply

- Train your developers
- Pro-actively detect secret exposures
  - Internally and externally
- Have a clear remediation plan
- Equip development teams with usable secret management services
- Mandate reviews on code version control platforms
- Rotate secrets periodically

COMCAST CYBERSECURITY

Thank you!

COMCAST
CYBERSECURITY