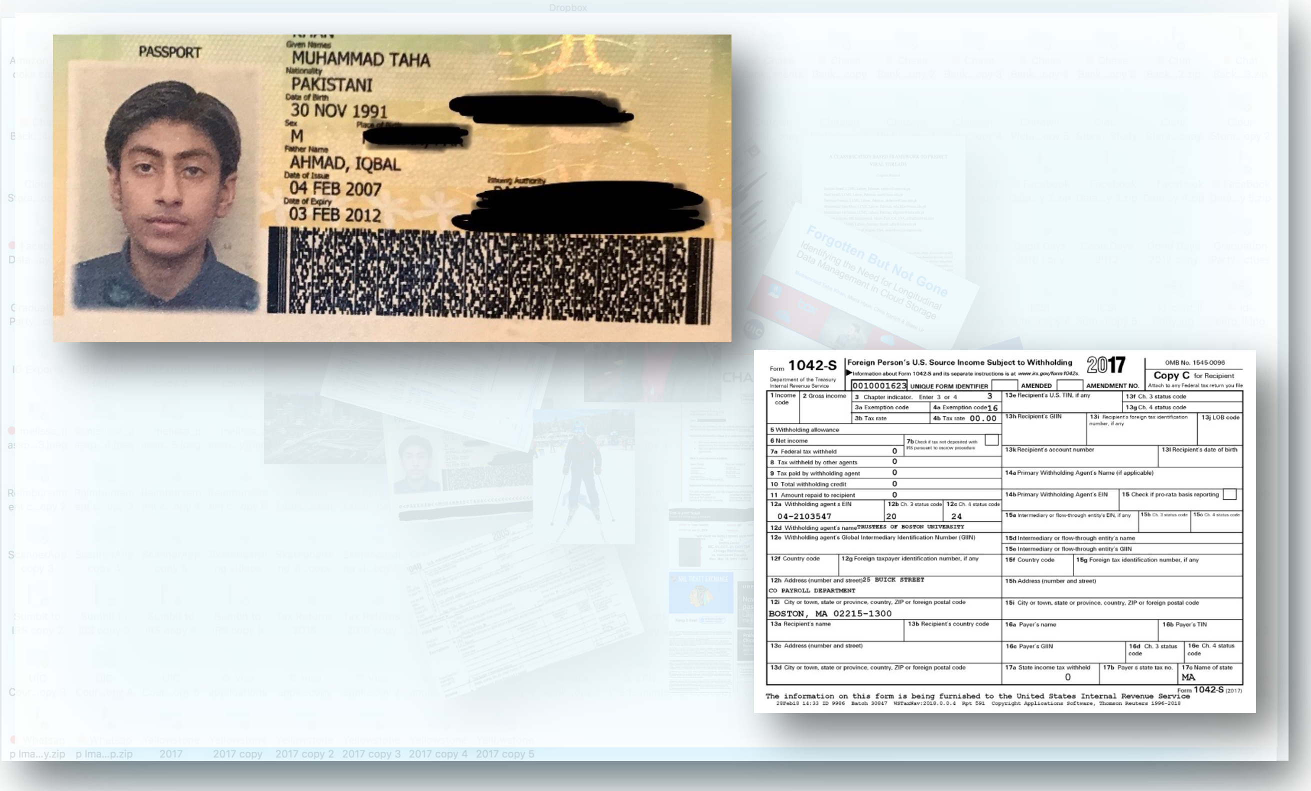
Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage

Mohammad Taha Khan, Christopher Tran,
Shubham Singh, Dimitri Vasilkov,
Chris Kanich, Blase Ur, Elena Zheleva

My Personal Dropbox



PASSPORT

Given Names
MUHAMMAD TAHA
Nationality
PAKISTANI
Date of Birth
30 NOV 1991 [redacted]
Sex
M
[redacted]
Father Name
AHMAD, IQBAL
Date of Issue
04 FEB 2007
Issuing Authority
[redacted]
Date of Expiry
03 FEB 2012

Form **1042-S** Department of the Treasury Internal Revenue Service

Foreign Person's U.S. Source Income Subject to Withholding **2017**
▶ Information about Form 1042-S and its separate instructions is at *www.irs.gov/form1042s.*

OMB No. 1545-0096
Copy C for Recipient
Attach to any Federal tax return you file

0010001623 UNIQUE FORM IDENTIFIER □ AMENDED □ AMENDMENT NO.

1 Income code
2 Gross income
3 Chapter indicator. Enter 3 or 4 3
3a Exemption code
4a Exemption code 16
3b Tax rate
4b Tax rate 00.00
5 Withholding allowance
6 Net income
7a Federal tax withheld 0
7b Check if tax not deposited with IRS pursuant to escrow procedure □
8 Tax withheld by other agents 0
9 Tax paid by withholding agent 0
10 Total withholding credit 0
11 Amount repaid to recipient 0
12a Withholding agent s EIN 04-2103547
12b Ch. 3 status code 20
12c Ch. 4 status code 24
12d Withholding agent's name TRUSTEES OF BOSTON UNIVERSITY
12e Withholding agent's Global Intermediary Identification Number (GIIN)
12f Country code
12g Foreign taxpayer identification number, if any
12h Address (number and street) 25 BUICK STREET
CO PAYROLL DEPARTMENT
12i City or town, state or province, country, ZIP or foreign postal code
BOSTON, MA 02215-1300
13a Recipient's name
13b Recipient's country code
13c Address (number and street)
13d City or town, state or province, country, ZIP or foreign postal code
13e Recipient's U.S. TIN, if any
13f Ch. 3 status code
13g Ch. 4 status code
13h Recipient's GIIN
13i Recipient's foreign tax identification number, if any
13j LOB code
13k Recipient's account number
13l Recipient's date of birth
14a Primary Withholding Agent's Name (if applicable)
14b Primary Withholding Agent's EIN
15 Check if pro-rata basis reporting □
15a Intermediary or flow-through entity's EIN, if any
15b Ch. 3 status code
15c Ch. 4 status code
15d Intermediary or flow-through entity's name
15e Intermediary or flow-through entity's GIIN
15f Country code
15g Foreign tax identification number, if any
15h Address (number and street)
15i City or town, state or province, country, ZIP or foreign postal code
16a Payer's name
16b Payer's TIN
16c Payer's GIIN
16d Ch. 3 status code
16e Ch. 4 status code
17a State income tax withheld 0
17b Payer s state tax no.
17c Name of state MA

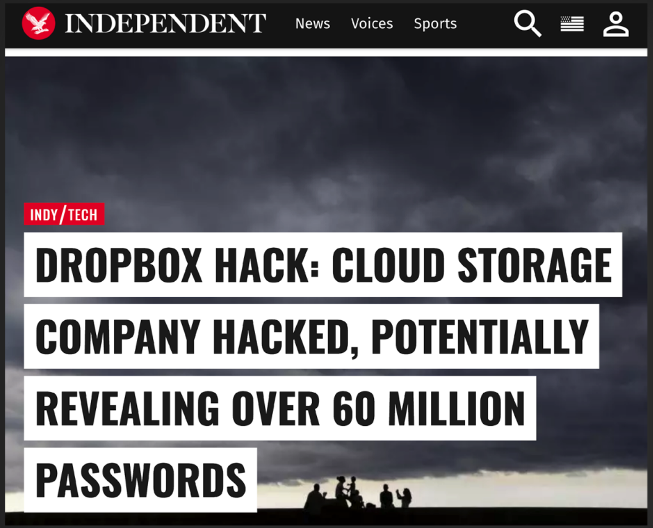
Form **1042-S** (2017)

The information on this form is being furnished to the United States Internal Revenue Service
28Feb18 14:33 ID 9986 Batch 30847 WSTaxNav:2018.0.0.4 Rpt 591 Copyright Applications Software, Thomson Reuters 1996-2018

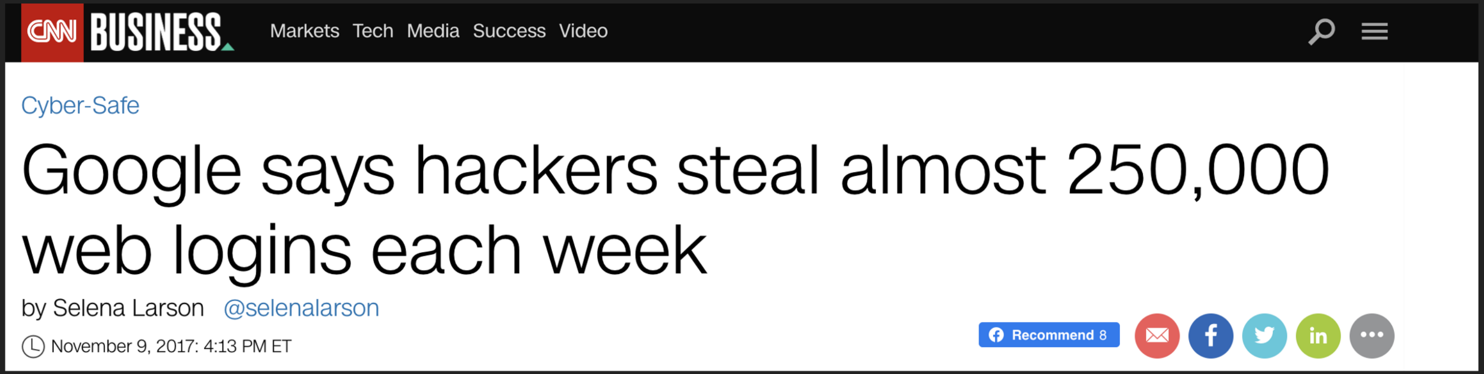
Taha's Dropbox

User since 2009
8.7GB, 19,245 Files

The Risk of Breaches



DROPBOX HACK: CLOUD STORAGE COMPANY HACKED, POTENTIALLY REVEALING OVER 60 MILLION PASSWORDS

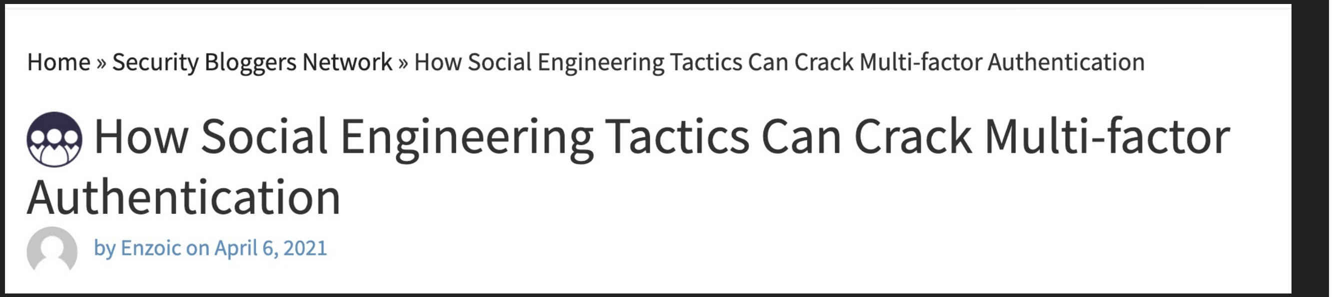


Cyber-Safe

Google says hackers steal almost 250,000 web logins each week

by Selena Larson @selenalarson

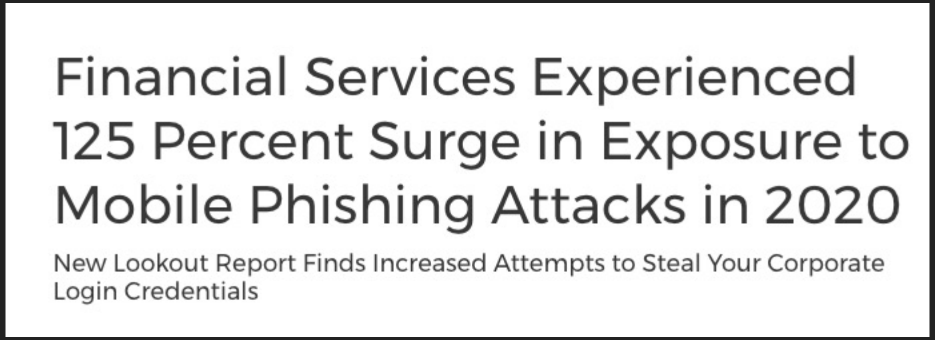
November 9, 2017: 4:13 PM ET



Home » Security Bloggers Network » How Social Engineering Tactics Can Crack Multi-factor Authentication

How Social Engineering Tactics Can Crack Multi-factor Authentication

by Enzoic on April 6, 2021



Financial Services Experienced 125 Percent Surge in Exposure to Mobile Phishing Attacks in 2020

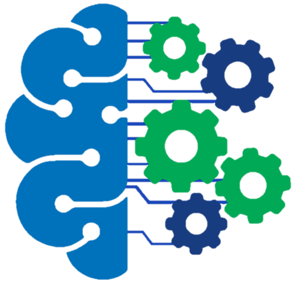
New Lookout Report Finds Increased Attempts to Steal Your Corporate Login Credentials

Goal: Semi-Automated Cloud Management

Machine learning can help users manage large, long-lived archives

Challenges:

- No existing datasets
- File management is subjective and personal
- What features are predictive?



Contribution: User-centered design of semi-automated classifiers

Research Goals and Approach

Goals:

1. Identify relevant file characteristics
2. Collect features and labels
3. Develop semi-automated management tools

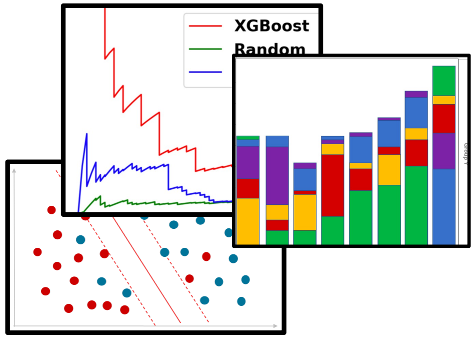
Approach:

Qualitative Interviews

→

Data-Collection Survey Study

→



Classifier Design and Evaluation

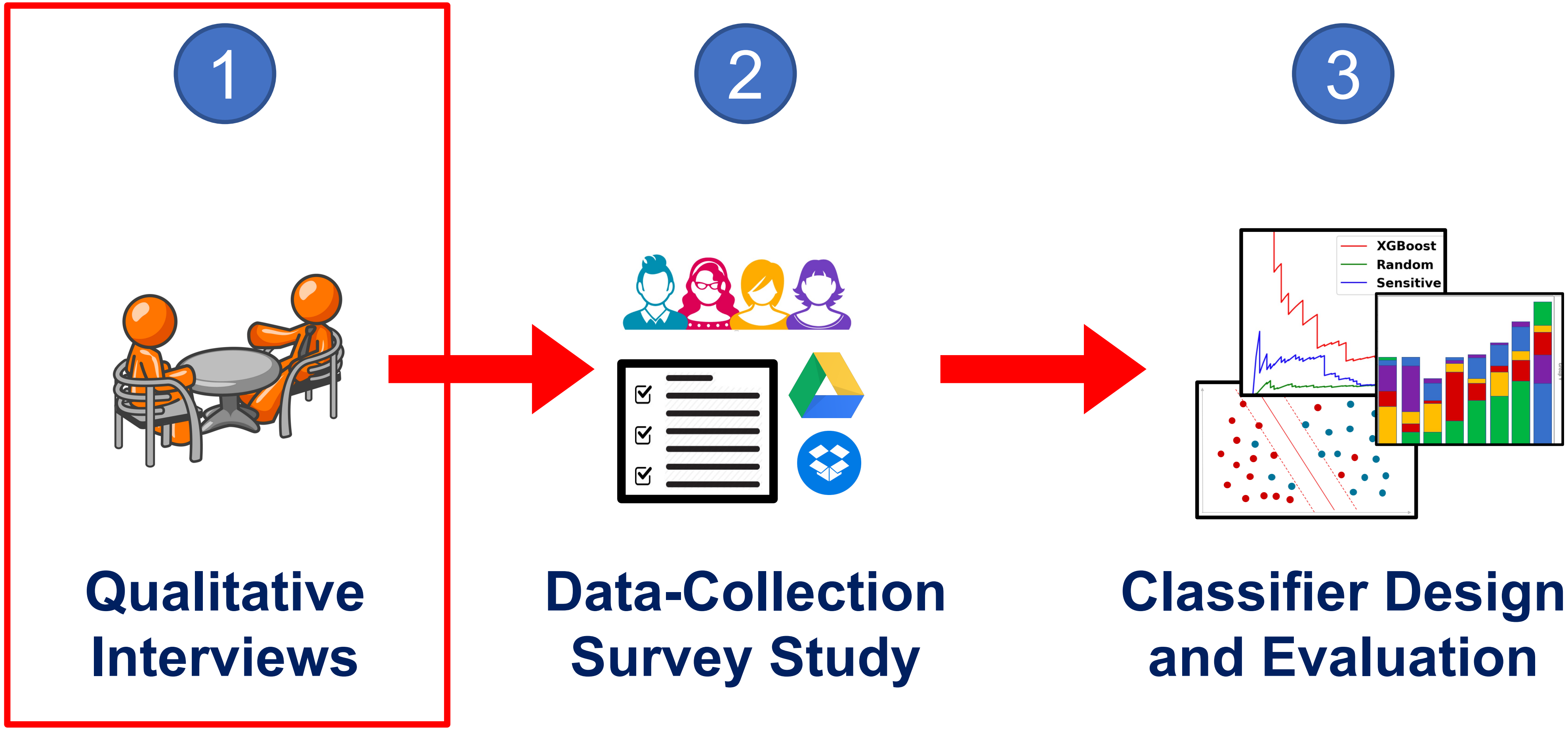
Sensitivity and Usefulness

Devised metrics of **sensitivity** and **usefulness** to capture subjectivity of file management

Hypothesized file management based on **sensitivity** and **usefulness**

Sensitivity	Usefulness	Management
✗	✗	Delete 🗑
✓	✗	
✗	✓	Keep as-is
✓	✓	Protect 🔒

Approach



1

Qualitative Interviews

2

Data-Collection Survey Study

3

Classifier Design and Evaluation

Qualitative Interviews

Goal: Understand subjective opinions of file sensitivity and usefulness

Interviewed 17 participants from diverse backgrounds

Explored mental-models of participants

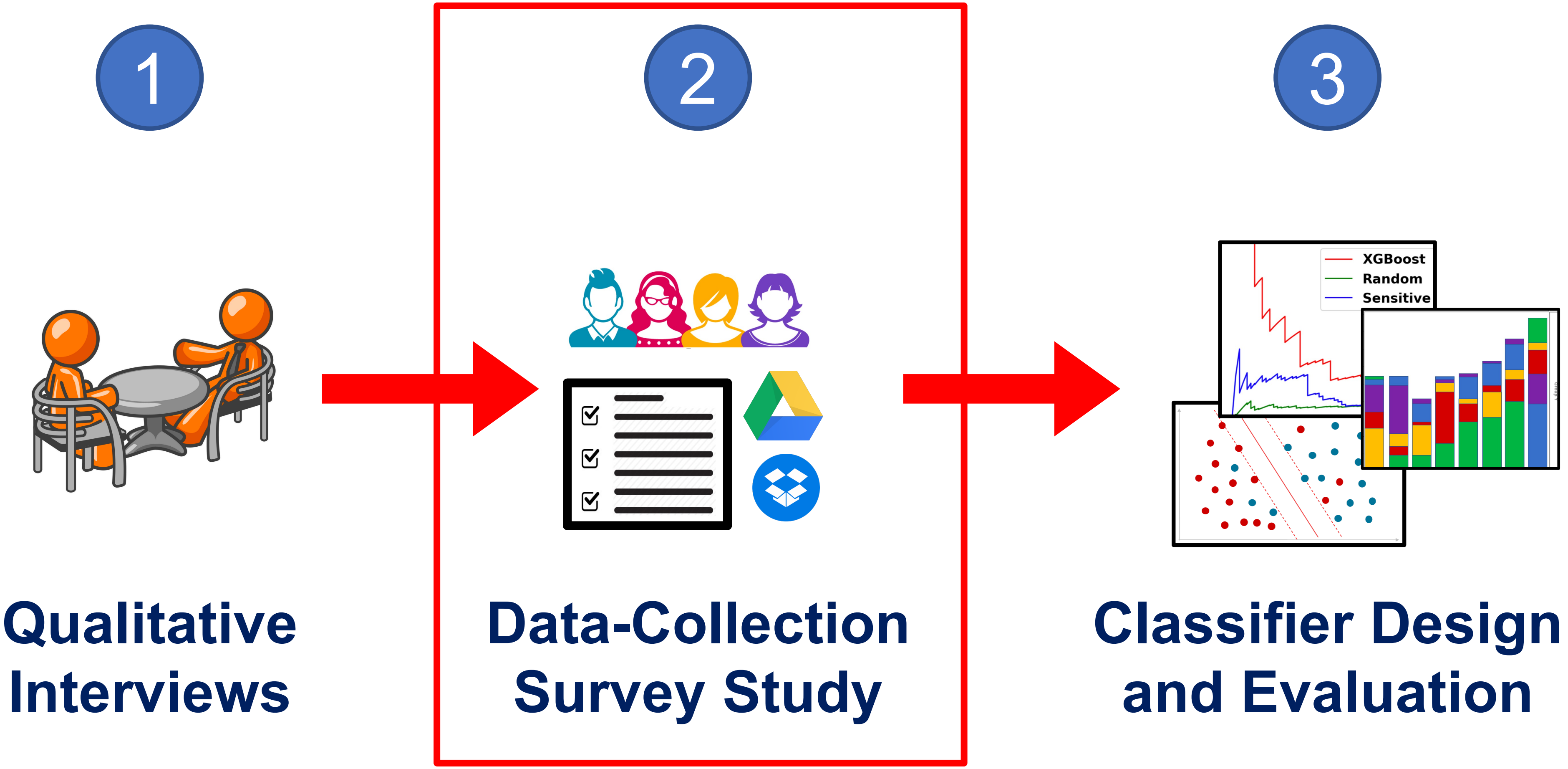
Characteristics of Potentially *Sensitive* Files

Personally identifiable or financial content
Nude, intimate, or embarrassing content
Content concerned with self-presentation
Proprietary and confidential information

Characteristics of Potentially *Useful* Files

Files for future reference
Files regularly accessed or shared
Memories and files with sentimental value
Backup archives

Approach

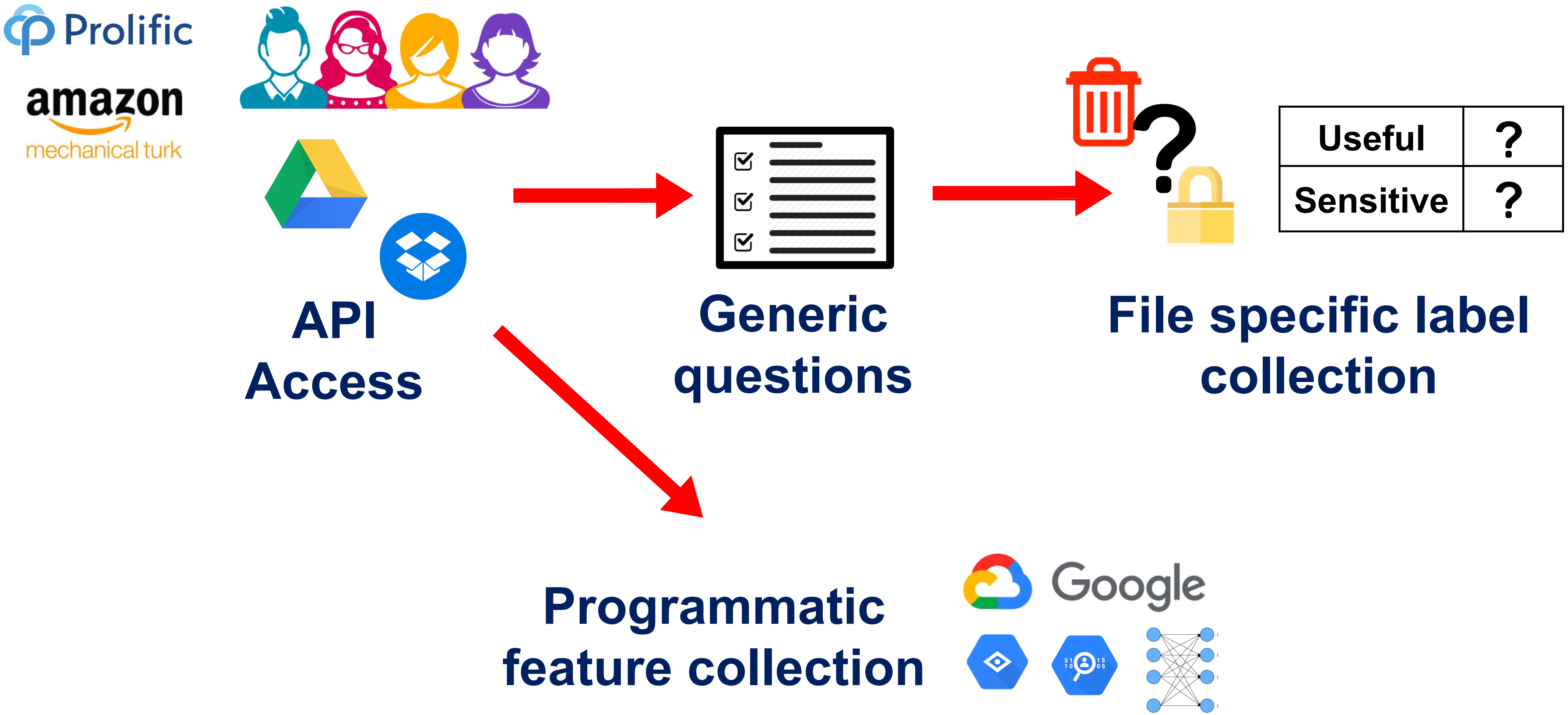


1 Qualitative Interviews

2 Data-Collection Survey Study

3 Classifier Design and Evaluation

Data-Collection Framework



API Access

Generic questions

File specific label collection

Useful	?
Sensitive	?

Programmatic feature collection

Features Collected

Dropbox and GDrive API

- Account age
- File name
- File size
- **Access details**
- **Sharing status**

.
.
.

Google Vision

- Image objects
- **Adult**
- **Racy**
- **Violent**
- Spoof

.
.
.

Google DLP

- **Name**
- **SSN,**
- **Email**
- **License #**
- **Credit card**
- **Bank Info**

.
.
.

Local text processing

- Doc topics
- Bag of words
- Word2vec

.
.
.
.

Approach

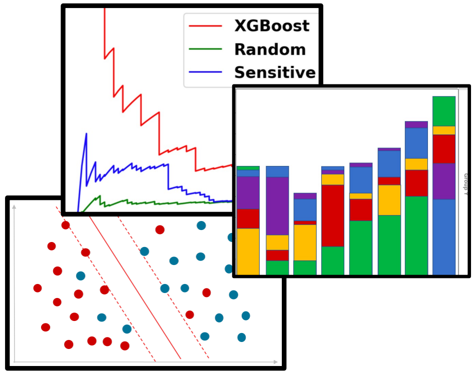
1

Qualitative Interviews

2

Data-Collection Survey Study

3



Classifier Design and Evaluation

Classifier Design and Evaluation

Goal: Partially automate file management via machine learning

Classifier	Prediction Class
Sensitivity	Sensitive, Not Sensitive
Usefulness	Useful, Not Useful
Management	Keep as is, Delete, Protect

Performance of the *Sensitivity* Classifier

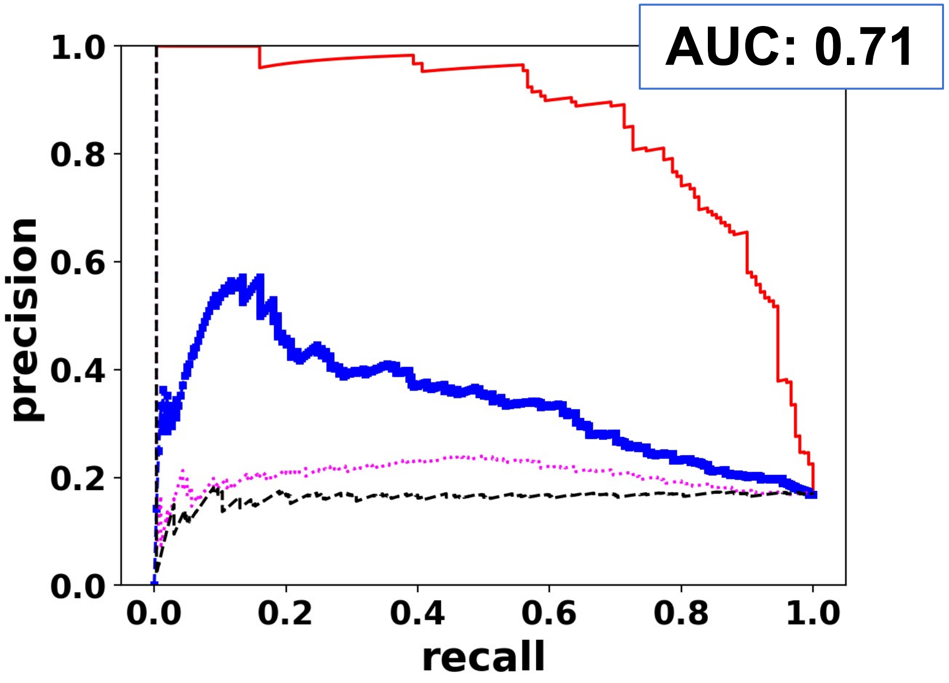
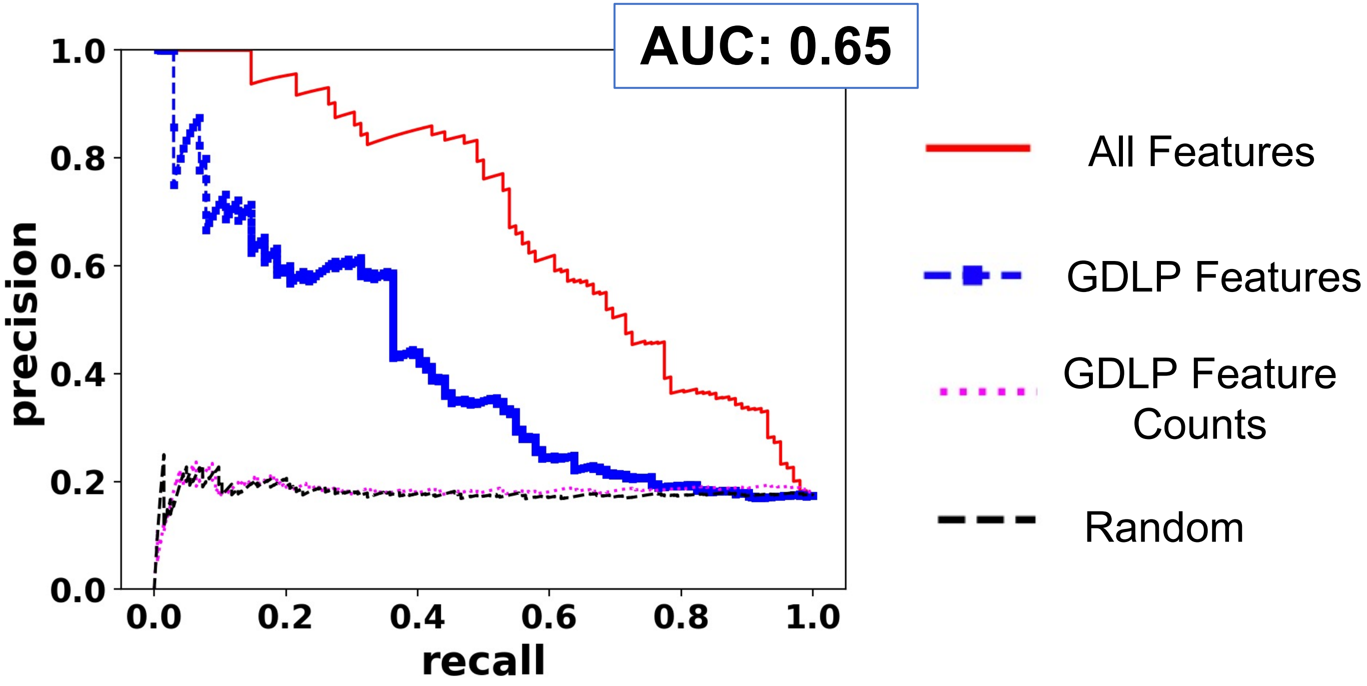
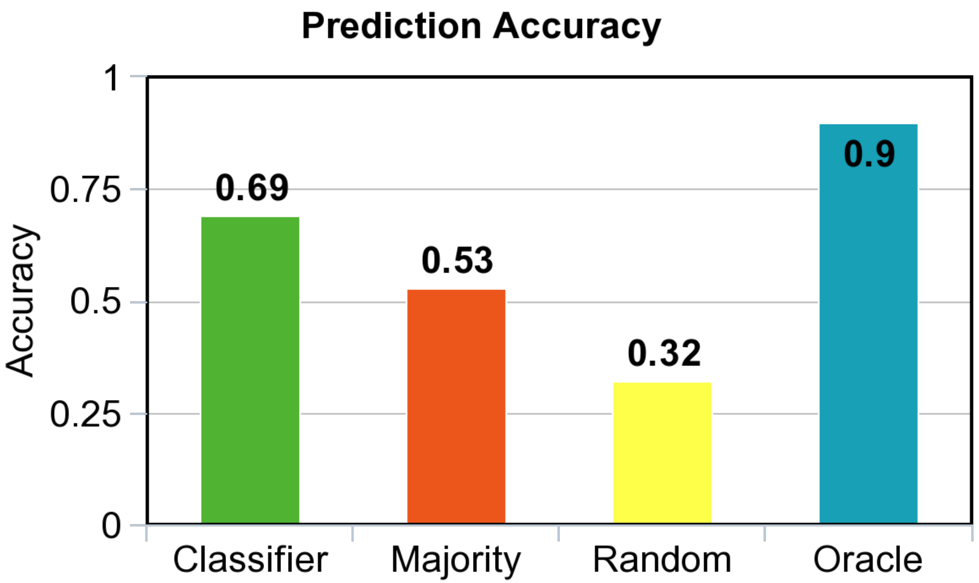


Image Sensitivity Classifier



Document Sensitivity Classifier

Accuracy of the *Management* Classifier

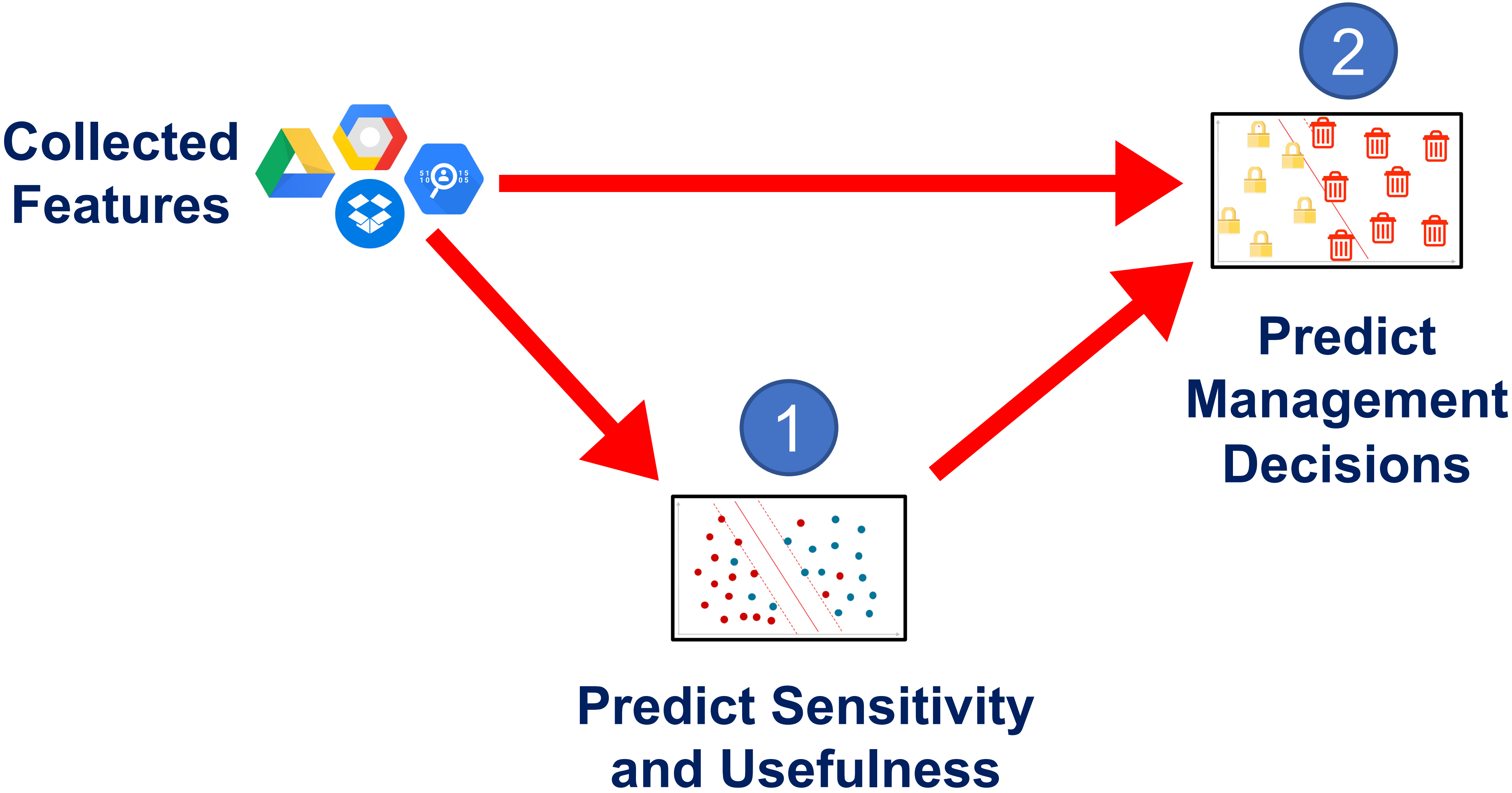


(Features only)

(Features + user reported sensitivity and usefulness labels)

69% management prediction accuracy by just using collected features

Two-Step Classification



Collected Features

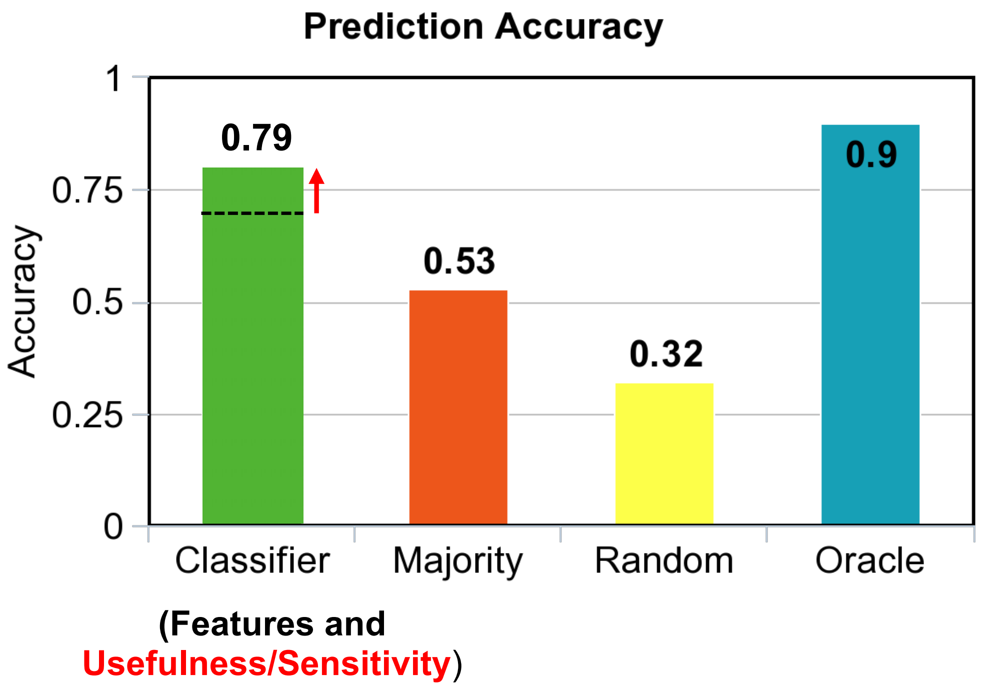
1

Predict Sensitivity and Usefulness

2

Predict Management Decisions

Increased Accuracy of the *Management* Classifier



10% increase in accuracy when training includes sensitivity and usefulness

Conclusion

Qualitative interviews → characteristics of sensitive and useful files

Quantitative user study → labeled dataset

Predicting file sensitivity and usefulness helps predict file management

Our prototype web app: https://cloudsweeper.app

Conclusion

- Qualitative interviews → characteristics of sensitive and useful files
- Quantitative user study → labeled dataset
- Predicting file sensitivity and usefulness helps predict file management
- Our prototype web app: https://cloudsweeper.app

Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage

Mohammad Taha Khan (tkhan@wlu.edu)**,** Christopher Tran, Shubham Singh, Dimitri Vasilkov, Chris Kanich, Blase Ur, Elena Zheleva