Enabling Retrospective Management of Data in The Cloud

Mohammad Taha Khan

PhD Defense
May 11, 2020

Committee Members

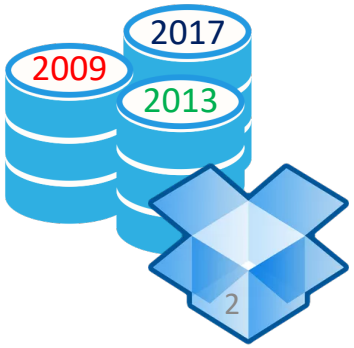
Chris Kanich (Chair & Adviser)
Robert Sloan
Ajay Kshemkalyani
Blase Ur
Narseo Vallina-Rodriguez

Cloud Storage Services

Cloud storage has becoming increasingly popular

Cloud storage has various use cases

Individuals have accumulated years of data



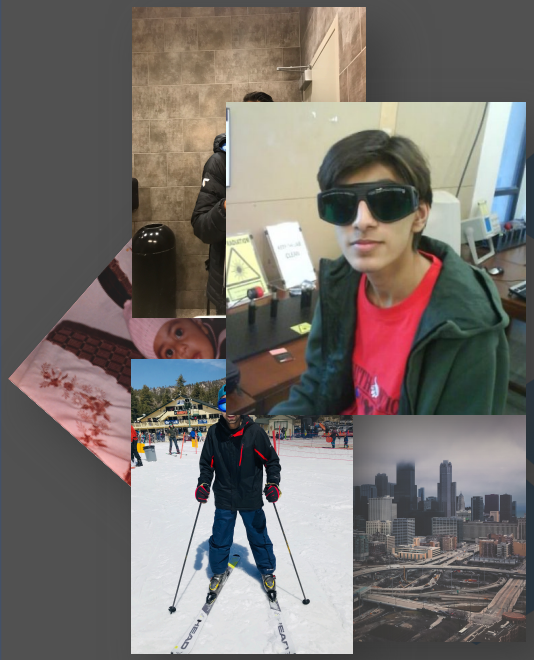
Issues With Long Term Data

1. Evolving and social and personal contexts of data

2. Increased latent risk from sensitive information in files

3. Manual management is infeasible due to scale of data

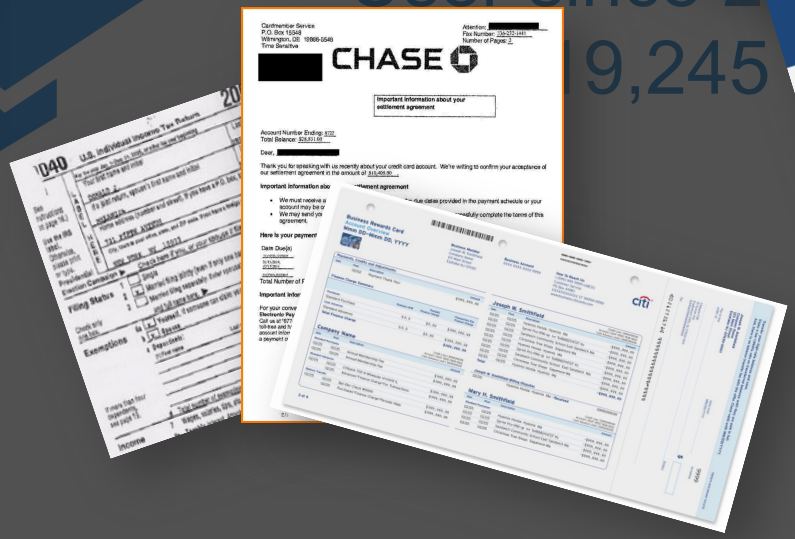
A Personal Example



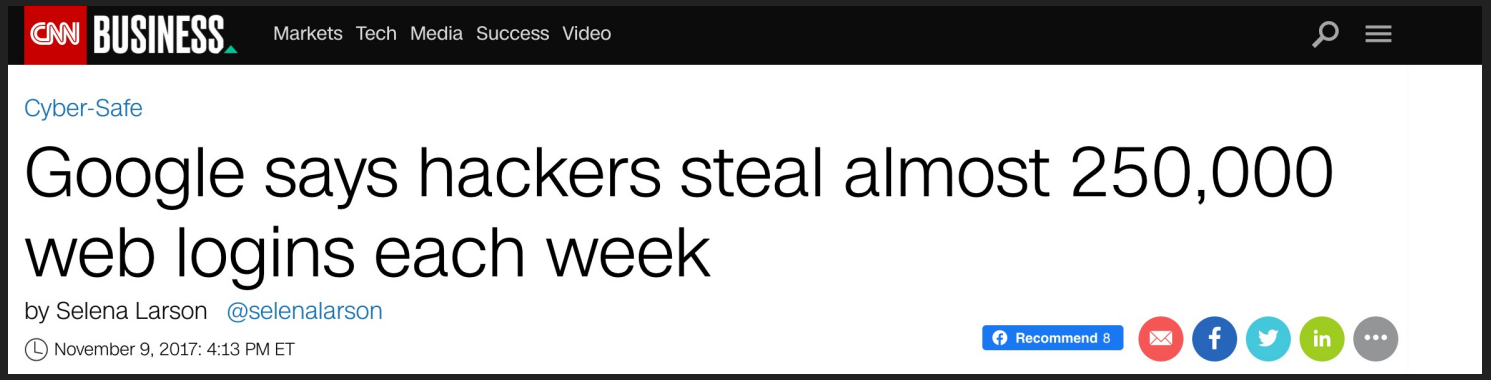
Taha's D

User since 2

9,245



Account Breaches Do Happen



Related Work

Cloud storage usage and privacy

- Evaluation of data integrity practices
- Understanding storage perceptions in the cloud

Risk of storing online data

- Study of online privacy perceptions
- Evaluation of cybercrimes (doxing, stalking)

Related Work

Retrospective management

- Management of cloud/social media
- Understanding of data significance/temporality

Management interfaces

- Privacy interfaces for file systems and emails
- Learning based privacy management in social media

Research Hypothesis

I hypothesize that over the years, cloud storage services have evolved into sophisticated and versatile data-stores that contain information that is stale and even poses a privacy risk to users, and this necessitates the development of methods that are specialized in accurately determining the extent of this risk and delivering precise retrospective remediation through automated management.

Thesis Statement

Through the process of empirical user studies, I first assess the extent of sensitive and expendable data in the cloud, and after determining that the volume of data is infeasible for manual management, I next explore users' interpretation on the kinds of management they intend, and integrate them into developing a learning based mechanism for the protection and management of their cloud accounts.

Research Goals (RG)

RG-I: Quantifiably establish the need for retrospective management among users

Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. "Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage". In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

RG-II: Effectively identify target files and manage them through automated techniques

Mohammad Taha Khan, Chris Tran, Shubham Singh, Dimitri Vasilkov, Will Brackenbury, Chris Kanich, Blase Ur, Elena Zheleva , . "Alethia: Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage". ***Under Submission***

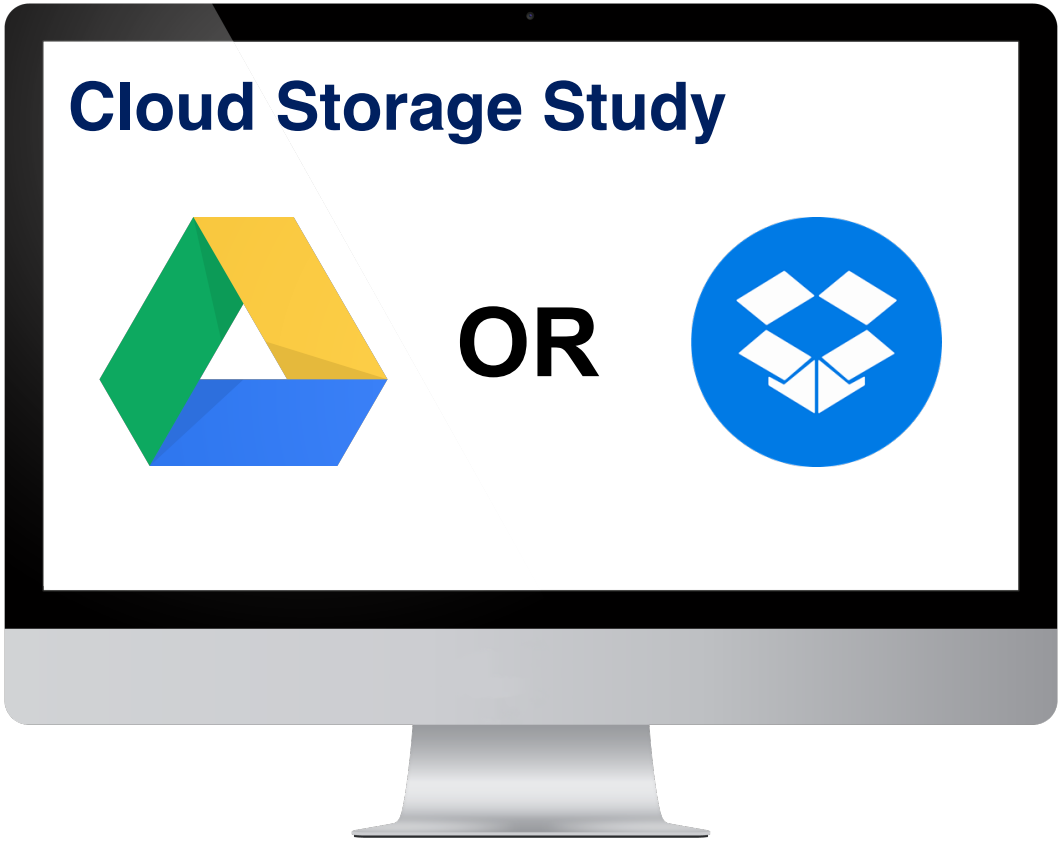
Primary Study Overview

API-based File Access

100 Participants

Survey-based Study

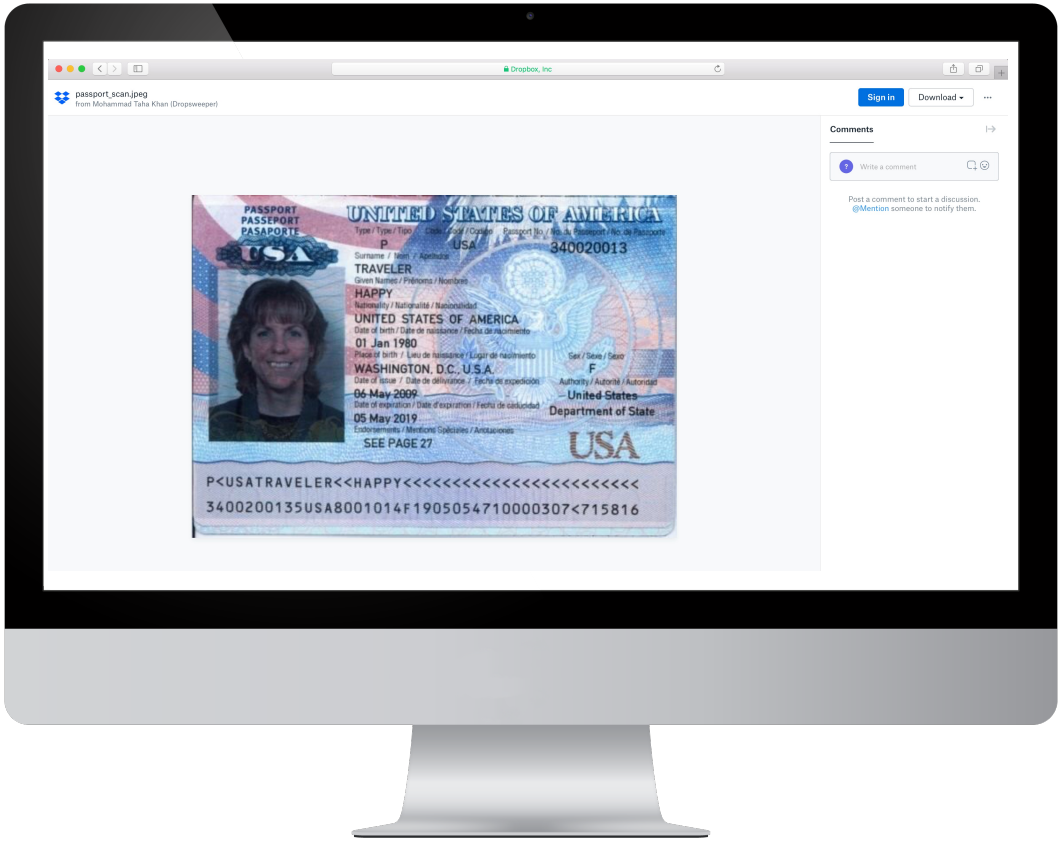
Primary Study Overview



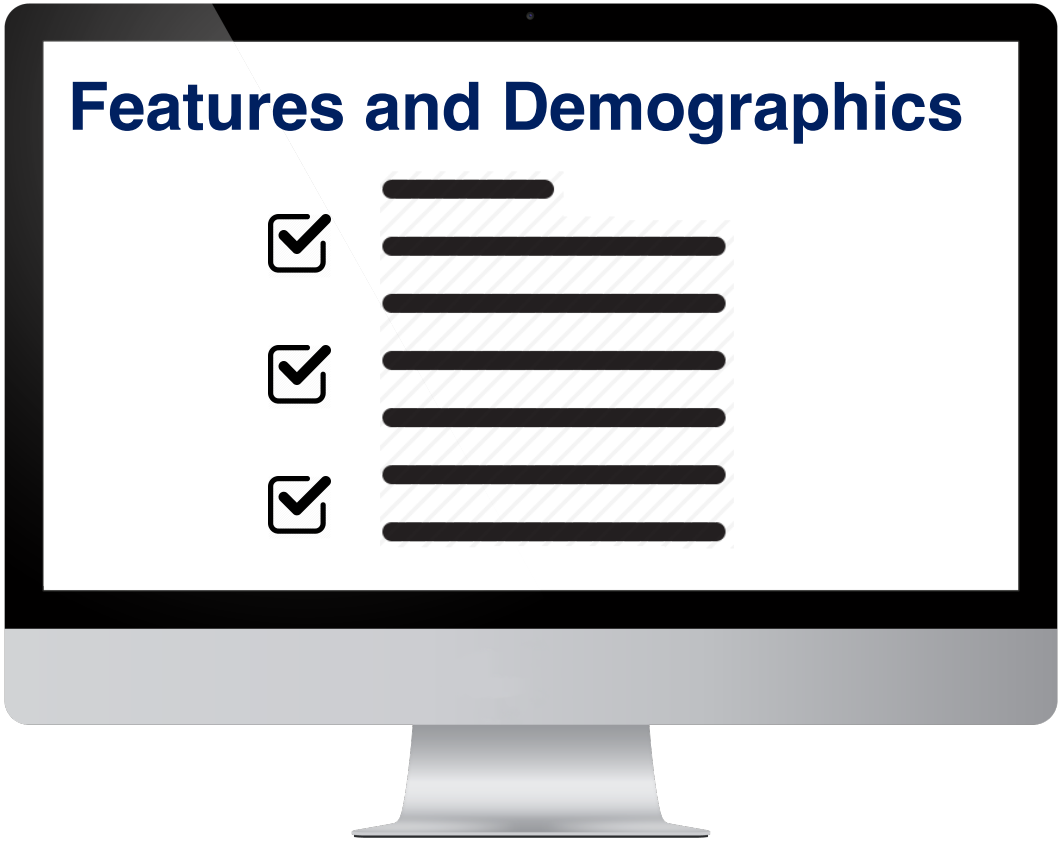
Primary Study Overview

Primary Study Overview

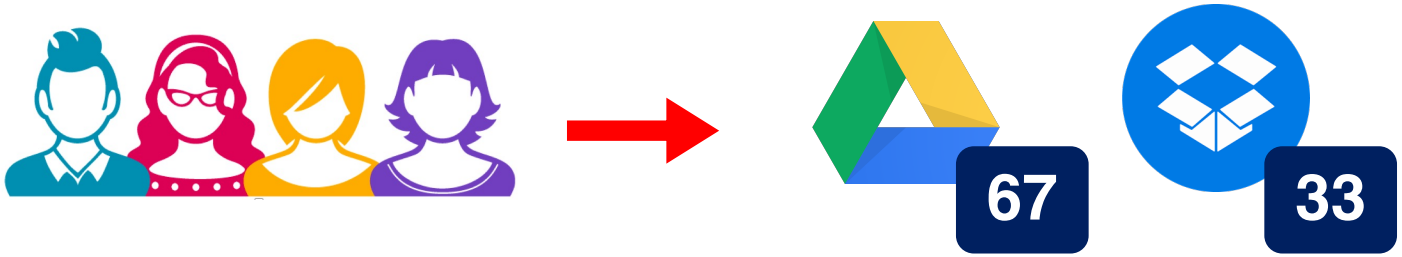
Primary Study Overview



Primary Study Overview



Survey Participants



88% of accounts > 3 years old

80% used for both professional and personal

Media and documents were most common

Some Files Will Never be Accessed

Never Access In Future:

23%

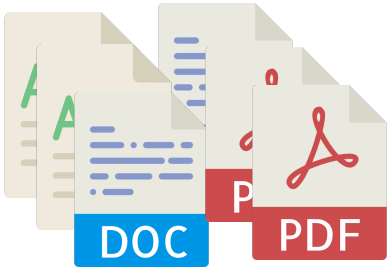
30%

Desired Management Decisions

Keep As Is	Delete	Protect
		
59%	34%	7%

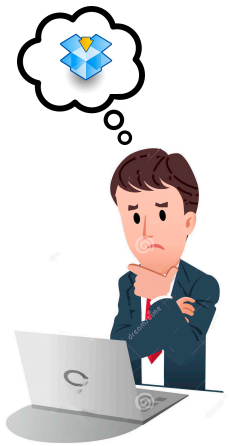
Generalizing Decisions

Generalized decision to similar files



Delete other “not useful” files

Not All Files Were Remembered



Not Remembered:

34%

39%

File Sharing

Keep Sharing

34%

Stop Sharing

11%

Key Findings

Many flies in the cloud…

- have been forgotten
- are no longer useful
- contain sensitive information

Disconnect between desire and ability

Need for tools to manage large archives over time

Follow Up Study

RG-II: Effectively identifying target files and managing them through automated techniques

Responses of primary study had insights regarding sensitivity and usefulness

Accomplished next study in three-steps

Follow Up Study: Three Part Approach

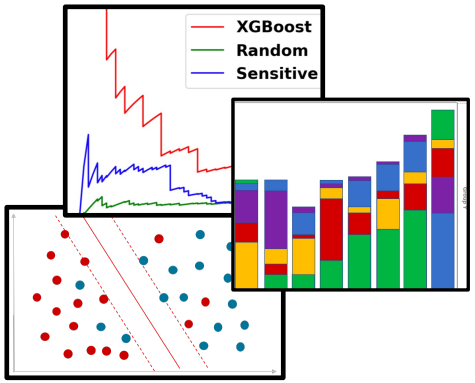
1

Qualitative Interviews

2

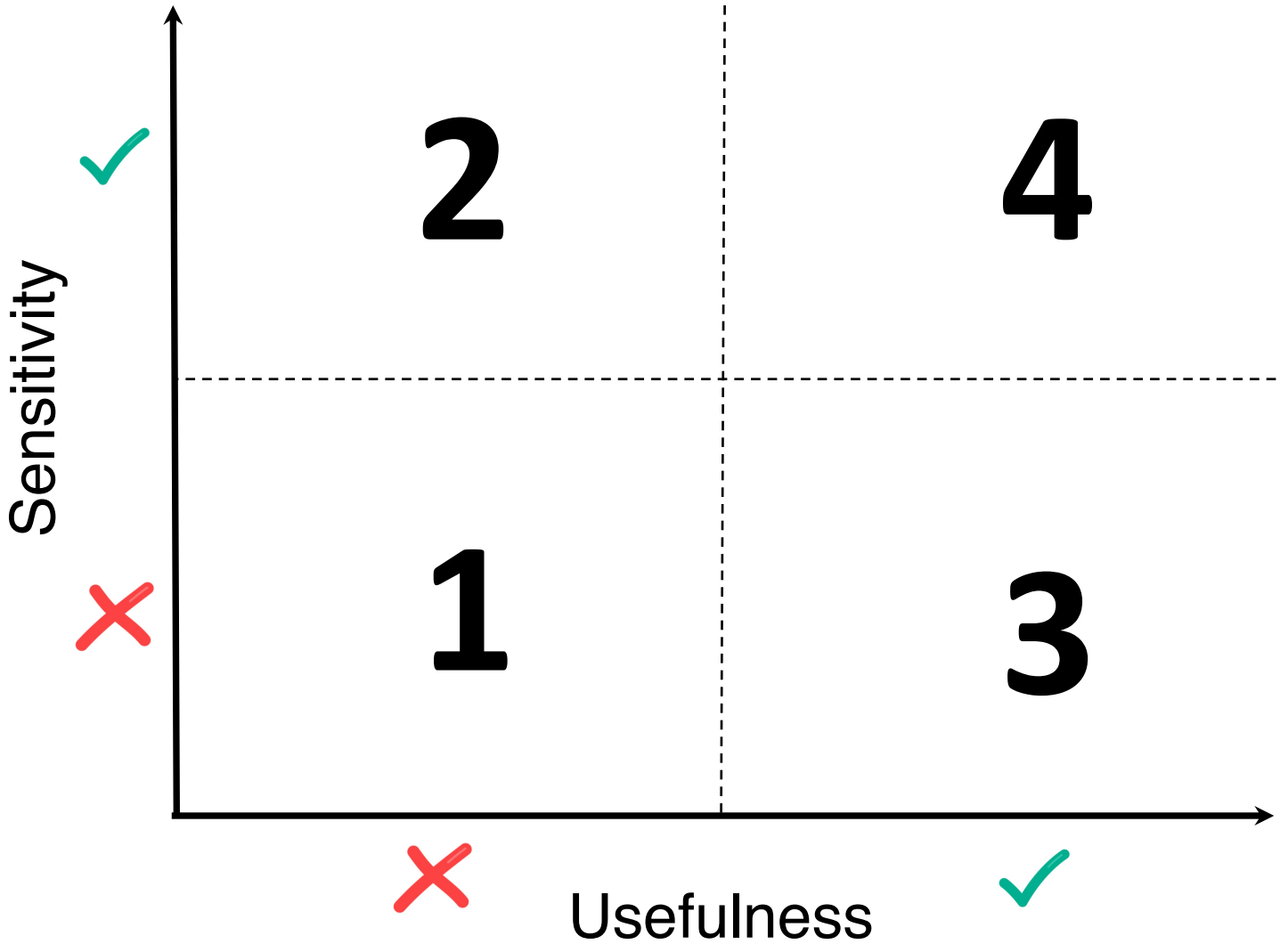
Data Collection Survey Study

3



Model Design and Evaluation

Sensitivity and Usefulness



Sensitivity and Usefulness

Sensitivity

2

*"It's **not essential**, but it's also related to a **secret activity** so I wouldn't want it to be available just in case."*

4

*"It has my **name on it**, it has money that I've earned, it has **where** I've **earned that money**. That's not **information** the **world needs to know**."*

1

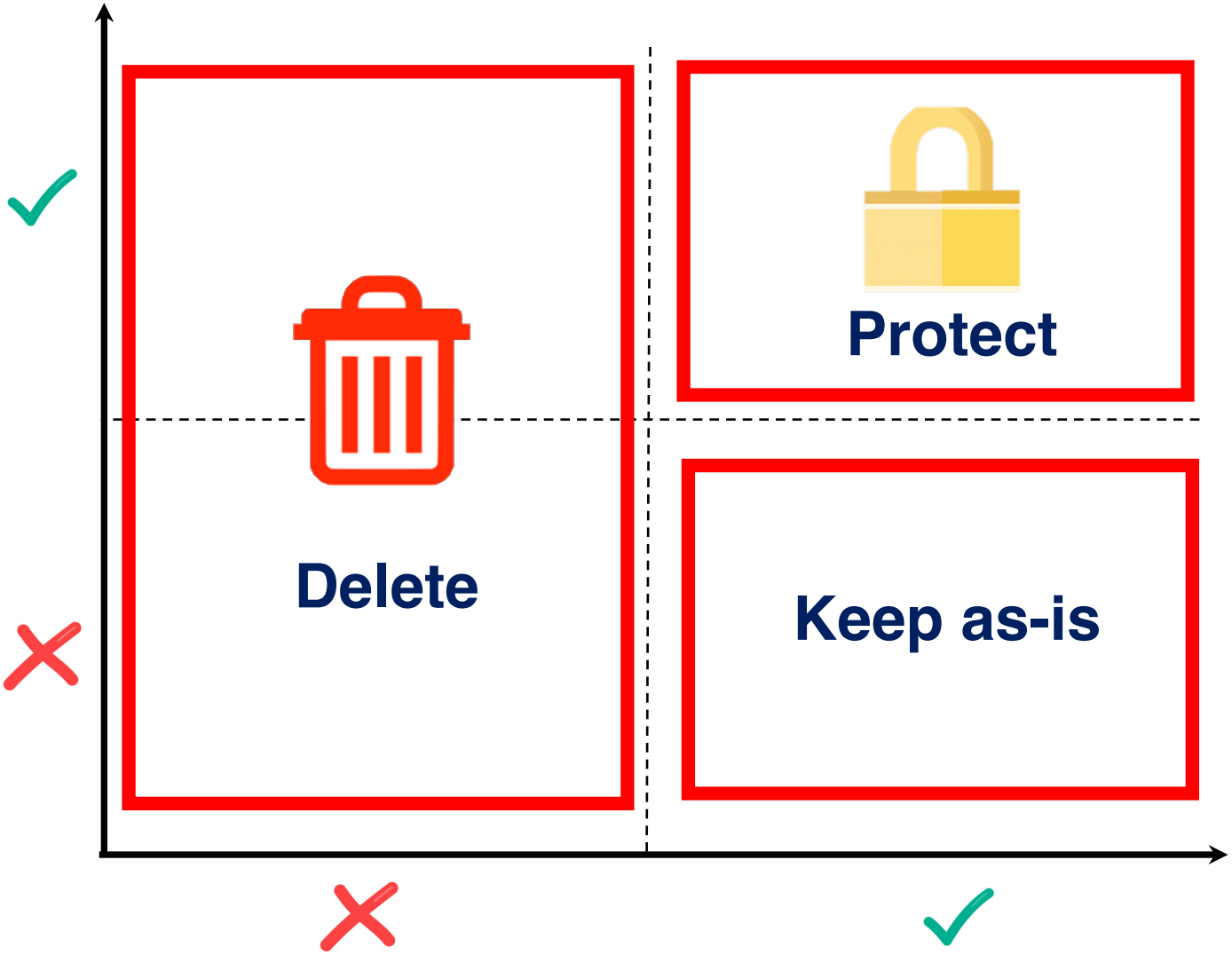
*"It seems **pointless** to have, and I have **no attachment** to it."*

3

*"It is a **research paper**, so it might still have **some value to me,** especially if a couple of years down the line I end up working in the field"*

Usefulness

Sensitivity and Usefulness



Follow Up Study: Three Part Approach

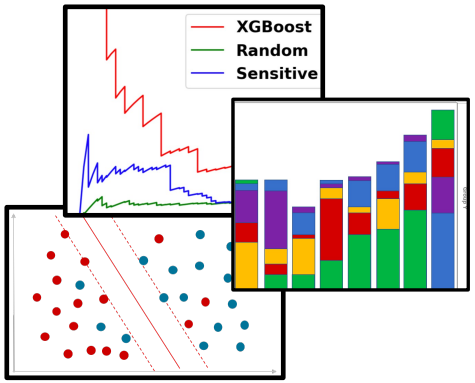
1

Qualitative Interviews

2

Data Collection Survey Study

3



Model Design and Evaluation

Qualitative Interviews

Sensitivity and usefulness and can be subjective

Performed qualitative interviews

Explored mental-models of participants

Qualitative Interviews

17 Participants from Craigslist

Two part discussion of sensitivity and usefulness

Transcription of responses

Sensitivity Categories

Personally identifiable or financial details
Intimate or embarrassing content
Content concerned with self image
Proprietary and confidential information

Usefulness Categories

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

Follow Up Study: Three Part Approach

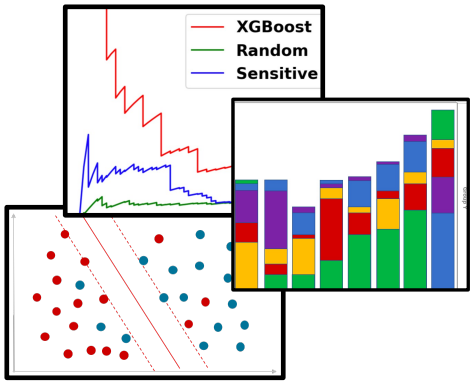
1

Qualitative Interviews

2

Data Collection Survey Study

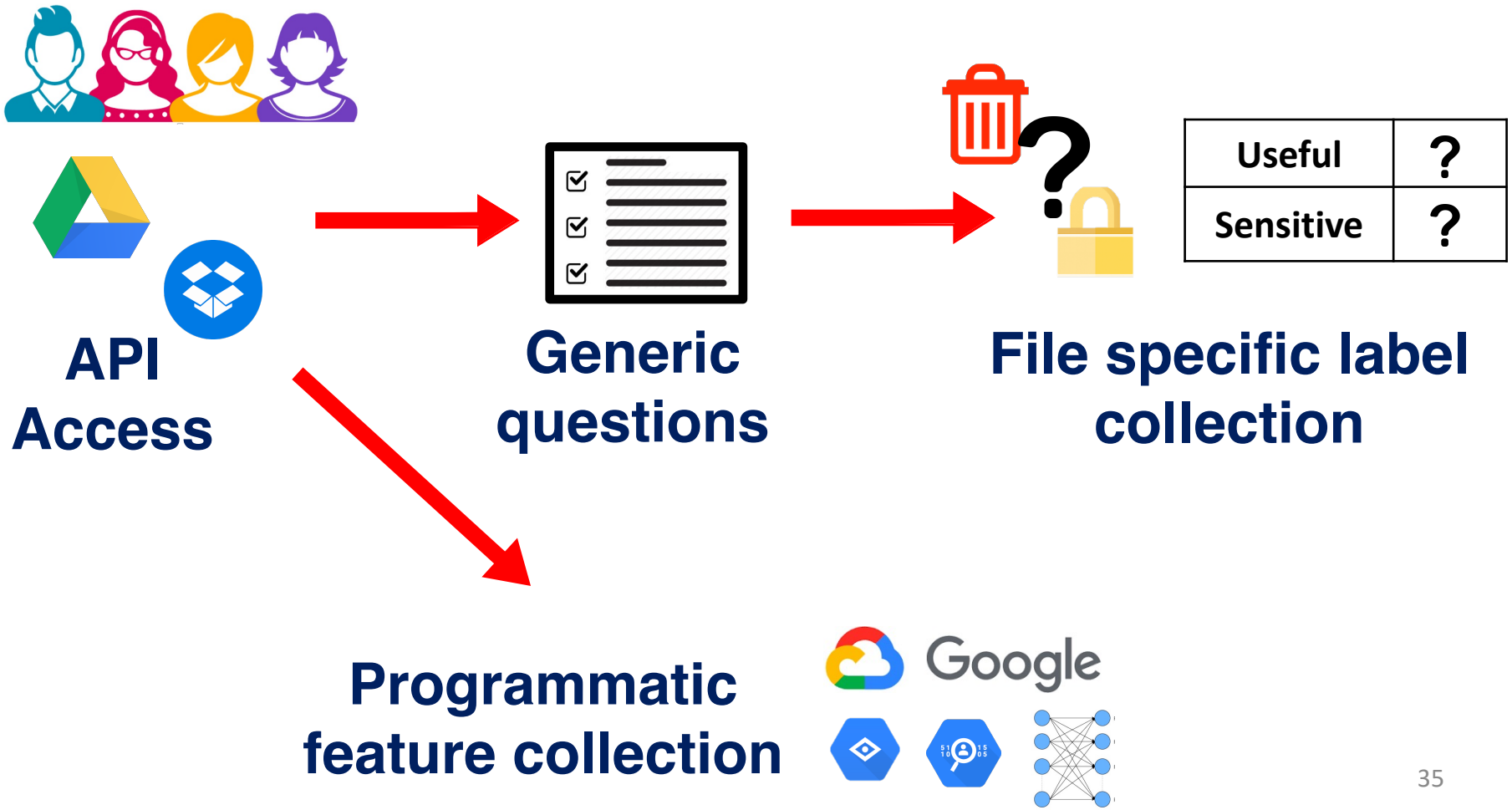
3



Model Design and Evaluation

Data Collection Framework

Data collection for a supervised learning model



Data Collection Framework

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
- TF-IDF

.
.
.

Data Collection Framework

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
- TF-IDF

.
.
.

Data Collection Framework

Dropbox and GDrive API

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

.
.
.

Google Vision

- **Image objects**
- Adult
- Racy
- Violent
- Medical

.
.
.

Google DLP

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

.
.

Local text processing

- **Doc topics**
- **Bag of words**
- Word2vec
- TF-IDF

.
.
.

2 Rounds of Data Collection

Challenge: Sensitive files are sparse in the cloud

Category	% Files
Sensitive, Useful	10%
Sensitive, Not useful	3%
Not Sensitive , Not Useful	35%
Not Sensitive, Useful	52%

3.5% increase in sensitive files for round 2

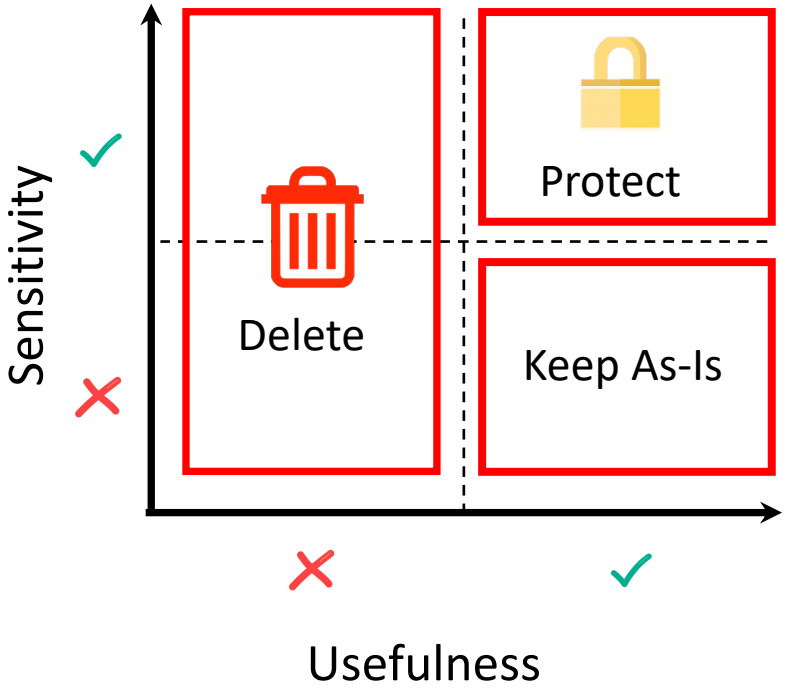
3525 file labels collected for 108 participants

Participants Provided Rich Data

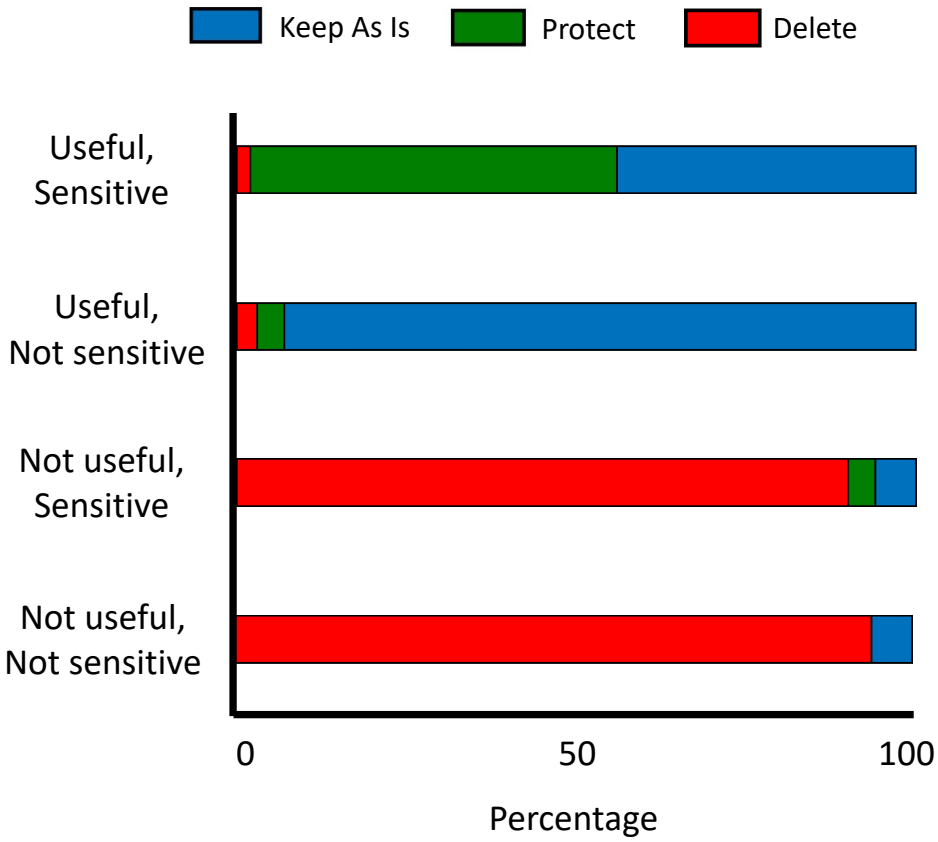
Description	% of Participants
Sensitivity File Categories	
PII of participant	62%
PII of other than the participant	31%
Intimate or embarrassing content	30%
Confidential/proprietary information	23%
Usefulness File Categories	
Future reference	96%
Sentimental value	87%
Backup and archives	91%

Sensitivity and Usefulness Evaluation

Initial Hypothesis



Empirical Evaluation



Follow Up Study: Three Part Approach

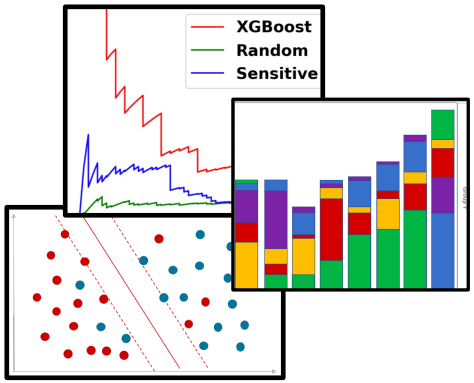
1

Qualitative Interviews

2

Data Collection Survey Study

3



Model Design and Evaluation

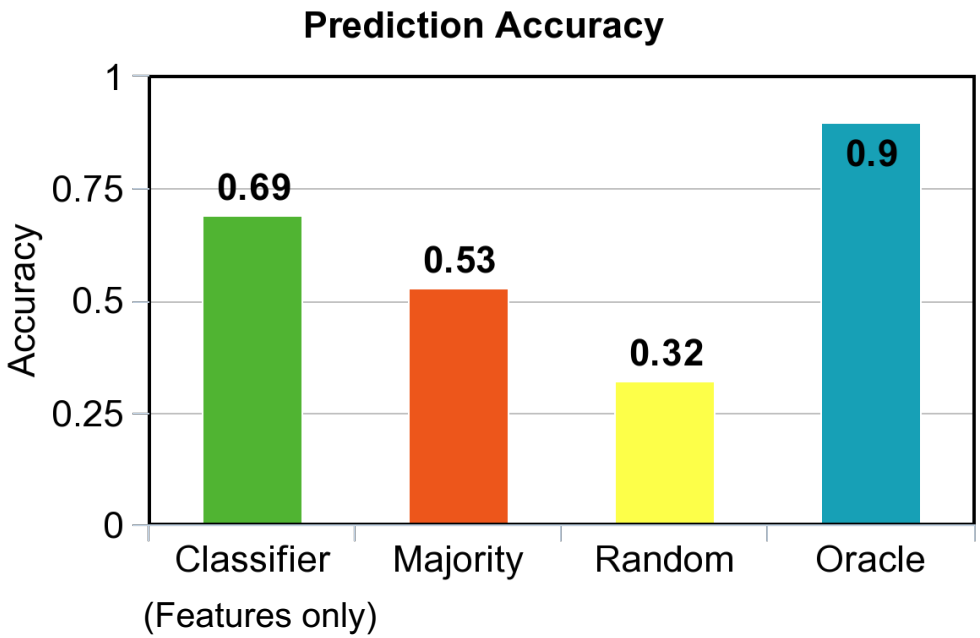
Management Through Classification

Final Goal: automate the management decisions via learning

3 classifiers to achieve learning-based management

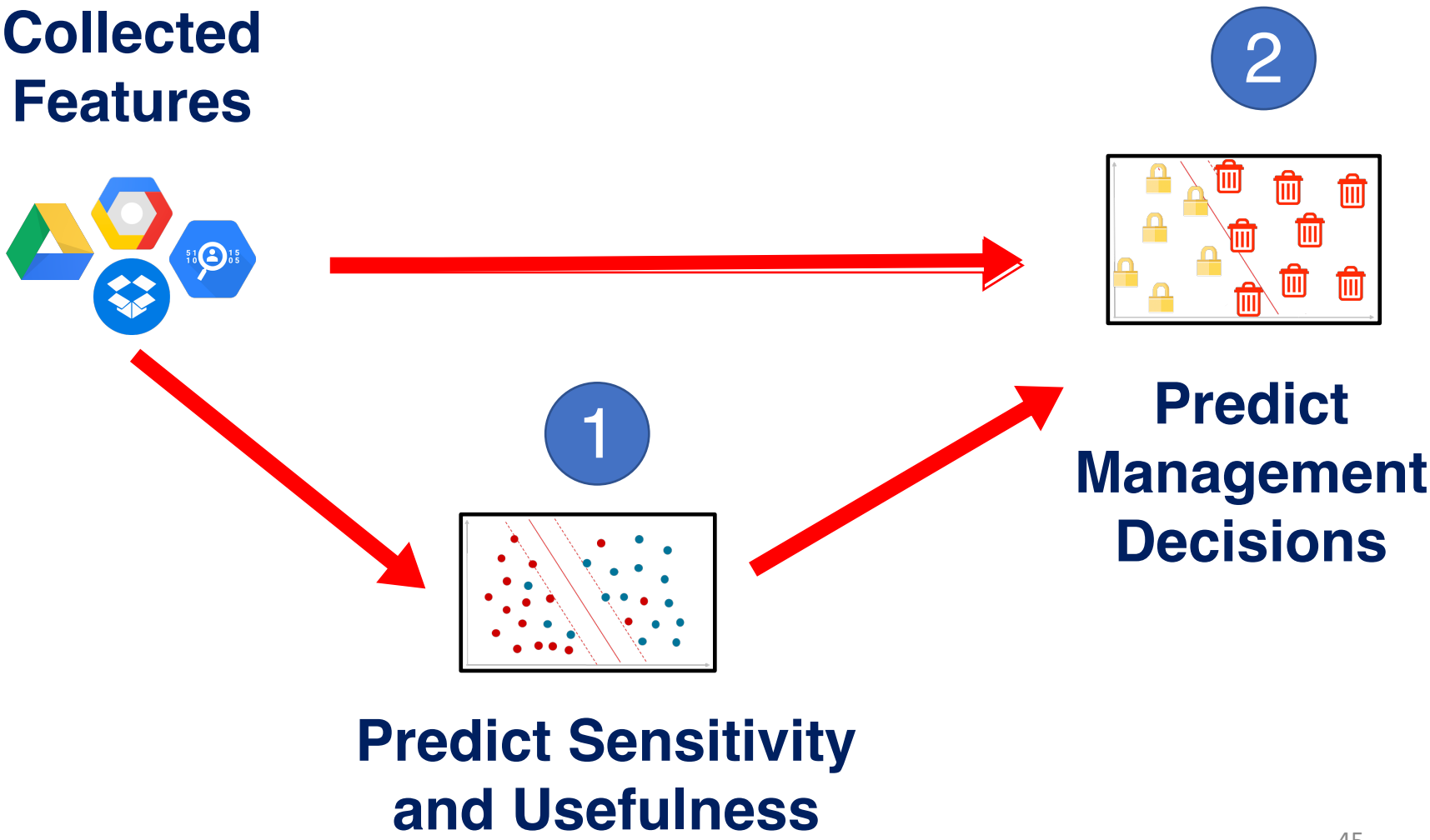
Classifier	Prediction Class
Primary: Management	Keep, Delete, Encrypt
Sensitivity	Sensitive, Not Sensitive
Usefulness	Useful, Not Useful

Accuracy of Management

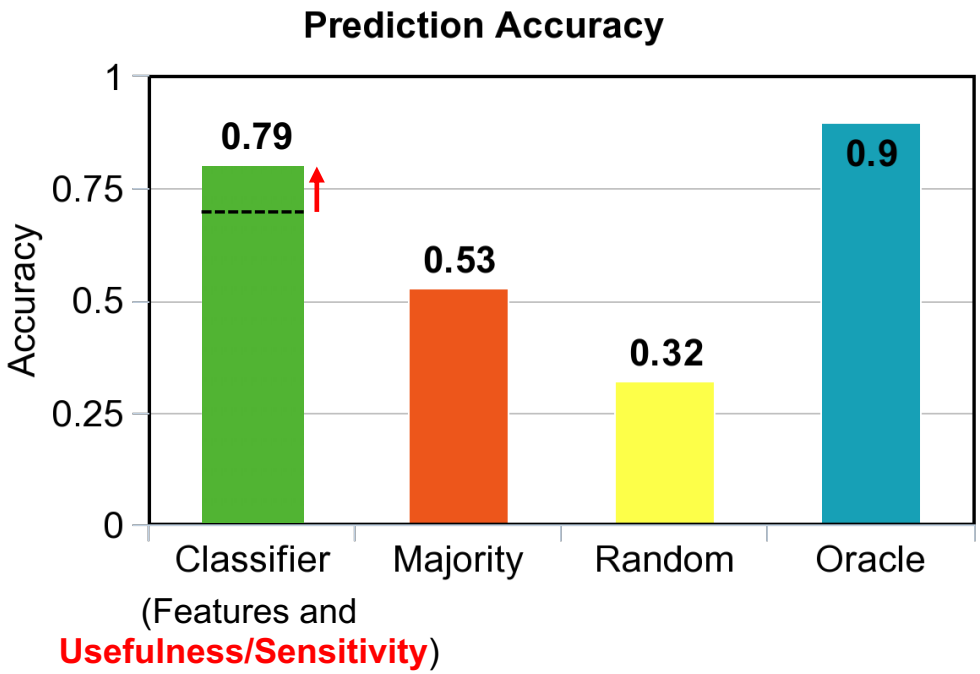


69% management prediction accuracy by just using collected features

Two Step Classification



Accuracy of Management



10% increase with a two-step classification involving sensitivity and usefulness

Conclusion

Evident need for retrospective management in the cloud

File usefulness and sensitivity are important characteristics

Qualitative insights play a key role in effective management

Cloud management requires a human centered approach

Future Work

Extend the current framework into production

Explore additional learning techniques

Incorporating personalization into the classifiers

Mapping HITL approach to other online platforms

Additional Research

Understanding and Measuring Cybercrime

Quantifying the Negative Externalities of Typosquatting – *IEEE S&P 2015*

A Comparison of Cyber and Regular Fraud in the US– *IEEE ConPro 2017*

Investigating Online Privacy Tools and Practices

An Empirical Analysis of the Commercial VPN Ecosystem– *ACM IMC 2018*

Moving Beyond Set-It-And-Forget-It Privacy Settings on Social Media – *ACM CCS 2019*

A Special Thanks!

Especially to my adviser, **Chris**, my committee members and all my collaborators!