XRay

Transparency for the data-driven Web.

Mathias Lécuyer
Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn
Augustin Chaintreau, *and* Roxana Geambasu

Columbia University

Why this ad?

Cedars Hotel Loughborough

36 Bedrooms, Restaurant; Bar Free WiFi, Parking, Best Rates

www.thecedarshotel.com

Why this ad?

Homosexuality

Cedars Hotel Loughborough

36 Bedrooms, Restaurant; Bar Free WiFi, Parking, Best Rates

www.thecedarshotel.com

Why this ad?

Ralph Lauren Online Shop

The official Site for Ralph Lauren Apparel, Accessories & More

www.ralphlauren.com

Why this ad?

Ralph Lauren Online Shop

The official Site for Ralph Lauren Apparel, Accessories & More

www.ralphlauren.com

Pregnancy

Did you know?

Did you know?

- Data Brokers can tell when you're sick, tired and depressed (and sell the information) [CNN ’14]
- Google Apps for Ed used institutional emails to target ads in personal accounts? [SafeGov’14]
- Credit companies use Facebook data to decide loans? [CNN’13]

Welcome to the big data world

- Myriad of web services parties collect **immense information** about us and use it for varied purposes
- Data has lots of **beneficial uses**
    - Useful recommendations
    - Powerful, predictive applications
    - Improve business with effective product placement
    - Improve public health, disaster response
    - …

Big data lacks transparency

- We have **no visibility** into what services do with our data:
  - What is the data used for exactly?
  - Is it being shared? With whom?
  - Can we delete it?
- Obscurity threatens to transform the data-driven web into a breeding ground for **data misuse**.
- **No robust tools** exist to reveal data (mis)uses, even **auditors** cannot find answers.

Question: can we build tools that reveal data misuse?

- Which emails trigger which ads?
- Which prior searches trigger which prices?
- Does Facebook share our data with third-parties?

Question: can we build tools that reveal data misuse?

- Which emails trigger which ads?
- Which prior searches trigger which prices?
- Does Facebook share our data with third-parties?

Can we do **taint tracking** systems?

- Lots of prior work, many successful systems (e.g., Taintdroid)
- Assume a controlled environment (runtime, language, OS)
- Need something for the complex and uncontrolled web

XRay

- First **generic data tracking system** for the Web
  - associate inputs (e.g., emails) outputs (e.g., ads)
- It is **accurate**, **scalable**, and **generic**
  - Works now on Gmail, Amazon and YouTube
- Provides **key building blocks** for a new ecosystem of tools to keep big data in check

Overview

Motivation

Design

Evaluation

Goals

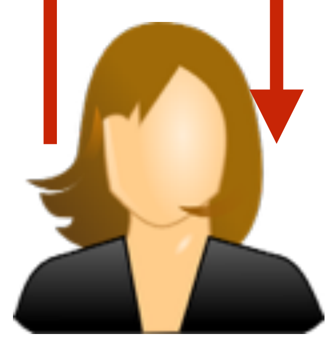
1. **Fine-grained**, **accurate** data use prediction
   - Predict use at individual input level (e.g., emails)
2. **Scalability**
   - Track many inputs (e.g., 100s of emails)
3. **Widely applicable** and **Self-Tuning**
   - Applies to many services (e.g., gmail, amazon…)

Web service model

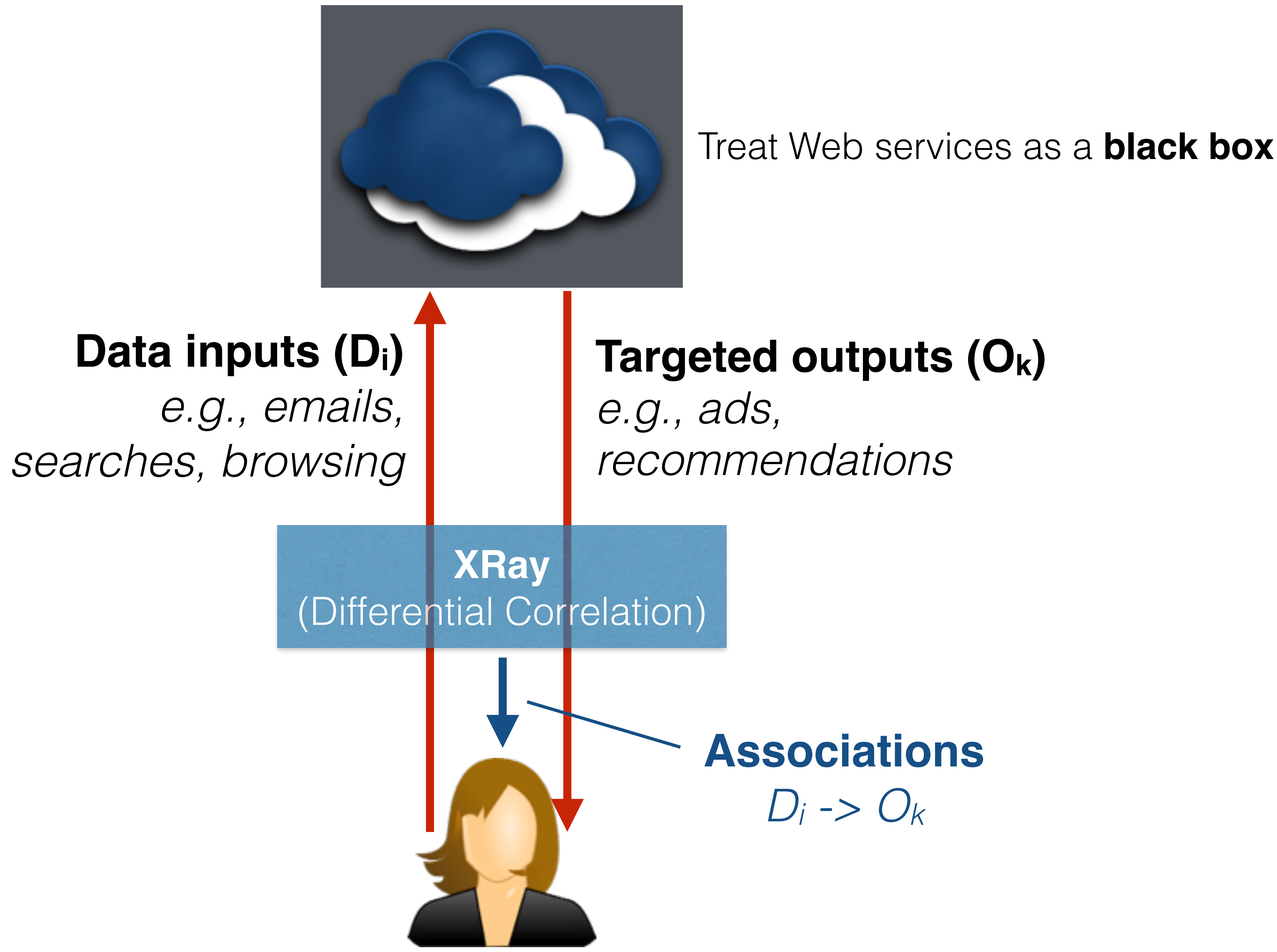
Treat Web services as a **black box**

Data inputs (D_i)
e.g., emails, searches, browsing

Targeted outputs (O_k)
e.g., ads, recommendations



Web service model

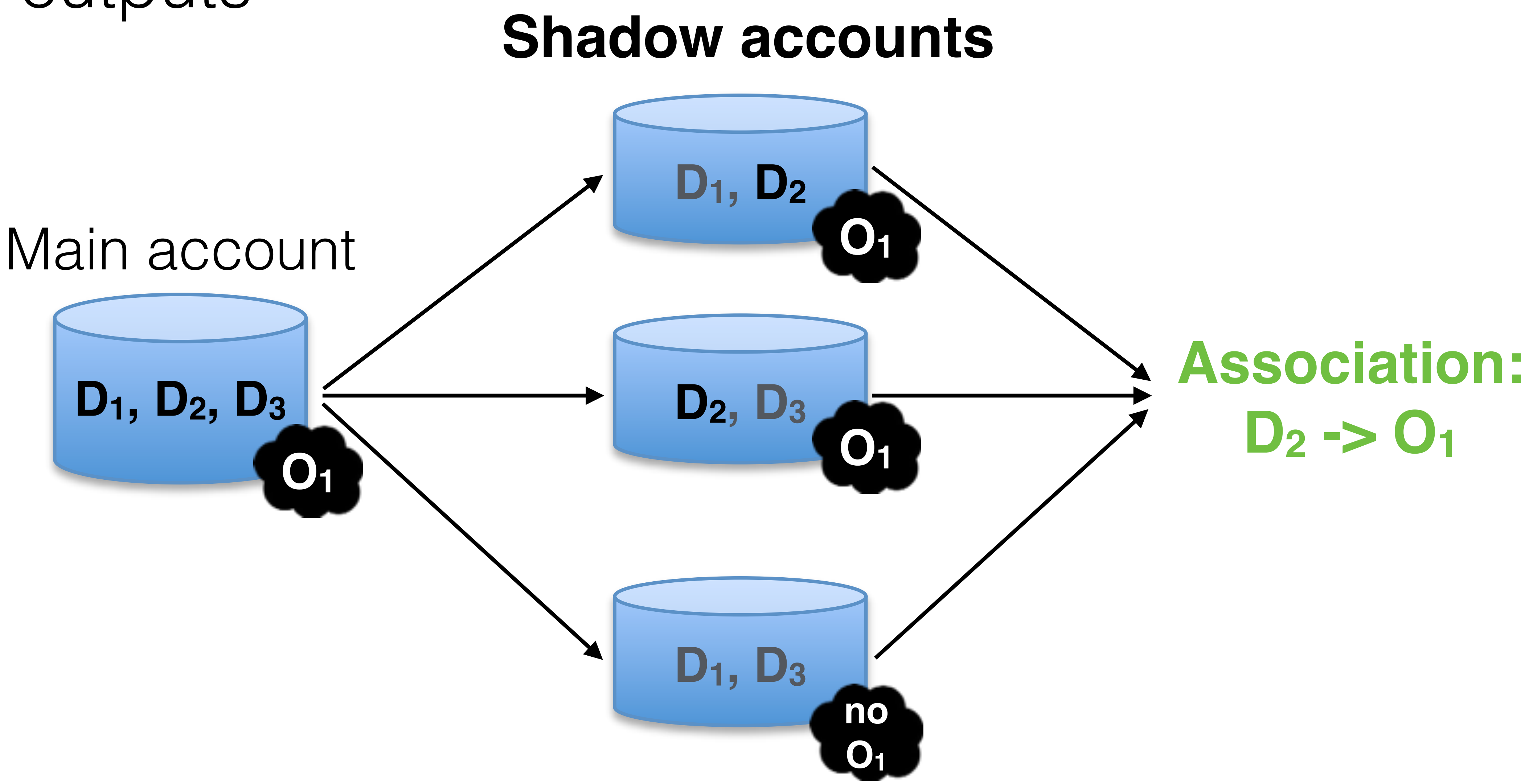


Differential Correlation

- Key idea: **correlate inputs with outputs**
    - Populate extra accounts with subsets of inputs
    - Use shadow account observations to relate inputs to outputs

Differential Correlation

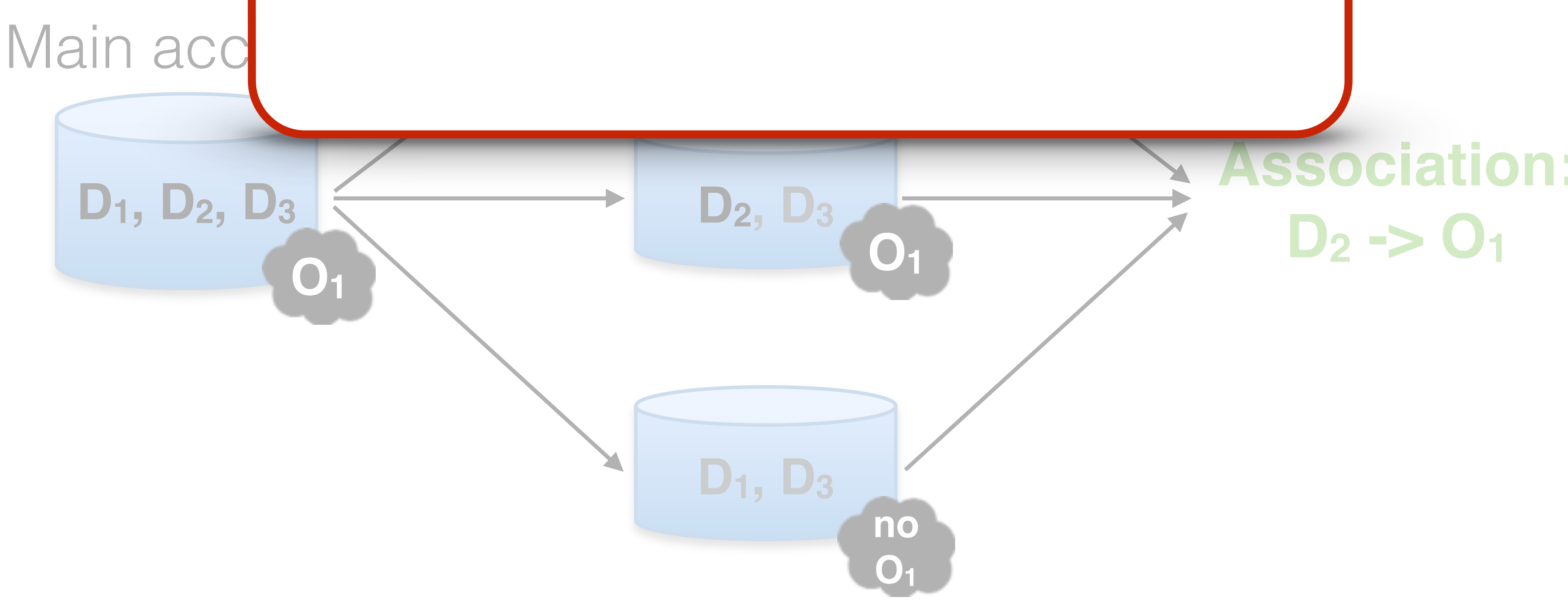
- Key idea: **correlate inputs with outputs**
  - Populate extra accounts with subsets of inputs
  - Use shadow account observations to relate inputs to outputs



Challenge: scaling

- Key idea: **correlate inputs with outputs**
  - Populate extra accounts with subsets of inputs
  - Use shadow account observations to relate inputs to outputs

It sounds like a lot of accounts…



Scalable algorithms

- **Theorem** *Under certain assumptions for any ε > 0 there exists an algorithm that requires* **C x log(N) accounts** *to correctly identify the inputs of a targeted ad with probability (1 – ε).*

- Algo1: **Set Intersection** (simple, not robust)

- Algo2: **Bayesian** (more robust)

Algo1: Set Intersection

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

Step1: **Randomly** assign emails to shadow accounts.

Step2: Take the sets of emails from accounts where the ad appeared.

Step3: Compute the **intersection** of these sets.

Step4: **if** the intersection is non empty**:**
it is the targeted emails
else there is no targeting.

Step 1

D_1, D_2	D_2, D_3	D_1, D_3

Step 2

D_1, D_2	D_2, D_3	~~D_1, D_3~~

Step 3

→ D_2

Algo1: Set Intersection

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

Step1: **Randomly** assign emails to shadow accounts.

Step2: Take the sets of emails from accounts where the ad appeared.

Step3: Compute the **intersection** of these sets.

Step4: **if** the intersection is non empty**:**
it is the targeted emails
else there is no targeting.

- We prove it needs a **logarithmic number of accounts** in number of inputs for high probability of detection

Challenge: it needs tuning

- Ads must **never appear in the wrong accounts**
  - Not true: email redundancy, cache
  - Need a manual threshold to detect emails in a significant number of accounts
- Doesn't take **low signal** into account
  - Need hard coded minimum number of accounts that see an ad
- Tuning is service specific and hard to do

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

  compute prob. $\mathbb{P}[\vec{x} | D_i]$

  $\mathbb{P}[D_i | \vec{x}]$= apply_bayes $\mathbb{P}[\vec{x} | D_i]$

end

compute prob. $\mathbb{P}[\vec{x} | D_\emptyset]$

$\mathbb{P}[D_\emptyset | \vec{x}]$ = apply_bayes $\mathbb{P}[\vec{x} | D_\emptyset]$

return D_i with max $\mathbb{P}[D_i | \vec{x}]$

- Bayes' rule:

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \times \mathbb{P}[A]}{\mathbb{P}[B]}$$

- Probability of observations:

With D_i targeted:

$$\mathbb{P}[\vec{x} | D_i] = (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}$$

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

compute prob. $\mathbb{P}\left[\vec{x} \mid D_i\right]$

$\mathbb{P}\left[D_i \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_i\right]$

end

compute prob. $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

$\mathbb{P}\left[D_\emptyset \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

return D_i with max $\mathbb{P}\left[D_i \mid \vec{x}\right]$

- Bayes' rule:

$$\mathbb{P}\left[A \mid B\right] = \frac{\mathbb{P}\left[B \mid A\right] \times \mathbb{P}\left[A\right]}{\mathbb{P}\left[B\right]}$$

- Probability of observations:

With D_i targeted:

$$\mathbb{P}\left[\vec{x} \mid D_i\right] = (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}$$

Algo2: Bayesian (XRay)

Input: Output O_K *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input

compute pro

$\mathbb{P}[D_i | \vec{x}]$ = ap

end

compute prob

$\mathbb{P}[D_\emptyset | \vec{x}]$ = ap

return D_i with

- Bayes' rule:

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \times \mathbb{P}[A]}{\mathbb{P}[B]}$$

- Probability of observations:

With D_i targeted:

$$\begin{aligned}\mathbb{P}[\vec{x} | D_i] = \quad & (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \\ & \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}\end{aligned}$$

- P_{in} : probability to see ad if targeted email in account

- P_{out} : probability to see ad if targeted email **not** in account

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

compute prob. $\mathbb{P}\left[\vec{x} \mid D_i\right]$

$\mathbb{P}\left[D_i \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_i\right]$

end

compute prob. $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

$\mathbb{P}\left[D_\emptyset \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

return D_i with max $\mathbb{P}\left[D_i \mid \vec{x}\right]$

- Bayes' rule:

$$\mathbb{P}\left[A \mid B\right] = \frac{\mathbb{P}\left[B \mid A\right] \times \mathbb{P}\left[A\right]}{\mathbb{P}\left[B\right]}$$

- Probability of observations:

With D_i targeted:

$$\mathbb{P}\left[\vec{x} \mid D_i\right] = (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}$$

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

compute prob. $\mathbb{P}\left[\vec{x} \mid D_i\right]$

$\mathbb{P}\left[D_i \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_i\right]$

end

compute prob. $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

$\mathbb{P}\left[D_\emptyset \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

return D_i with max $\mathbb{P}\left[D_i \mid \vec{x}\right]$

- Bayes' rule:

$$\mathbb{P}\left[A \mid B\right] = \frac{\mathbb{P}\left[B \mid A\right] \times \mathbb{P}\left[A\right]}{\mathbb{P}\left[B\right]}$$

- Probability of observations:

With D_i targeted:

$$\mathbb{P}\left[\vec{x} \mid D_i\right] = (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}$$

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

compute prob. $\mathbb{P}\left[\vec{x} \mid D_i\right]$

$\mathbb{P}\left[D_i \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_i\right]$

end

compute prob. $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

$\mathbb{P}\left[D_\emptyset \mid \vec{x}\right]$ = apply_bayes $\mathbb{P}\left[\vec{x} \mid D_\emptyset\right]$

return D_i with max $\mathbb{P}\left[D_i \mid \vec{x}\right]$

- If an email is targeted, we can tell which one.
- **Challenge**: what if no tracked input is targeted?

Algo2: Bayesian (XRay)

Input: Output O_k *(an ad)*, Inputs D_is *(emails)*, Observations x

Output: Targeted input

foreach input Di **do**

compute prob. $\mathbb{P}[\vec{x}|\, D_i]$

$\mathbb{P}[D_i|\, \vec{x}]$= apply_bayes $\mathbb{P}[\vec{x}|\, D_i]$

end

compute prob. $\mathbb{P}[\vec{x}|\, D_\emptyset]$

$\mathbb{P}[D_\emptyset|\, \vec{x}]$ = apply_bayes $\mathbb{P}[\vec{x}|\, D_\emptyset]$

return D_i with max $\mathbb{P}[D_i|\, \vec{x}]$

- Bayes' rule:

$$\mathbb{P}[A|\, B] = \frac{\mathbb{P}[B|\, A] \times \mathbb{P}[A]}{\mathbb{P}[B]}$$

- Probability of observations:

With D_i targeted:

$$\mathbb{P}[\vec{x}|\, D_i] = (p_{\text{in}})^{|A_i \cap A_k|} (1 - p_{\text{in}})^{|A_i \cap \bar{A}_k|} \times (p_{\text{out}})^{|\bar{A}_i \cap A_k|} (1 - p_{\text{out}})^{|\bar{A}_i \cap \bar{A}_k|}$$

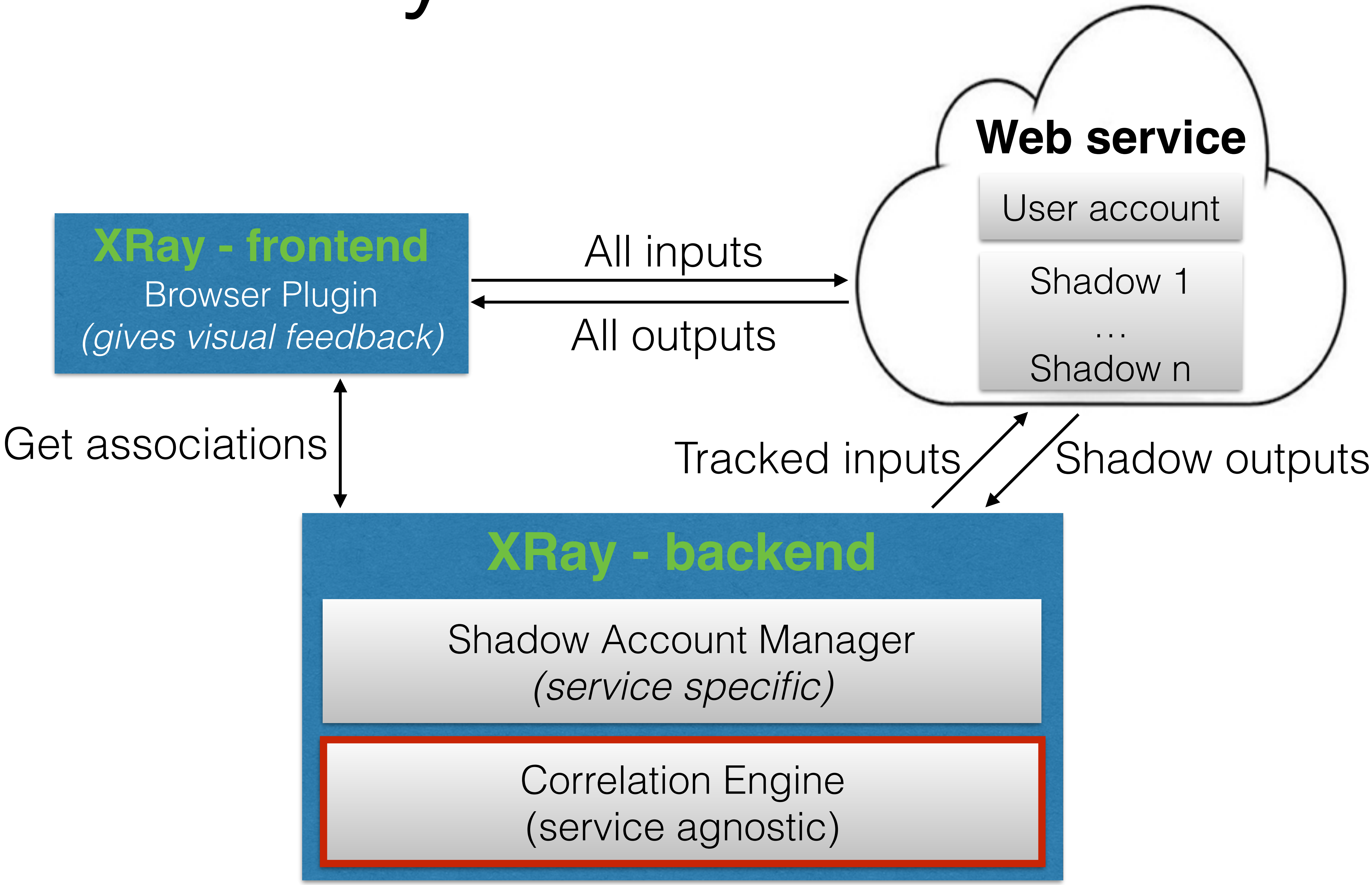
Without targeting:

$$\mathbb{P}[\vec{x}|\, D_\emptyset] = (p_\emptyset)^{|A_k|} (1 - p_\emptyset)^{|\bar{A}_k|}$$

Bayesian can self-tune

- **Automatic self-tuning** with classic iterative inference to learn the parameters
- **Many other challenges** (input overlap, different kind of targeting…).

XRay's architecture



Prototype

- We built the prototype for **Gmail**, to associate ads to the emails they target.
- Applied correlation engine **as-is** to **Amazon** product recommendations and **YouTube** video recommendations.
- **0 lines of code to change** to adapt the correlation mechanisms.

Talk overview

Motivation

Design

Evaluation

Evaluation questions

How accurate is XRay?

Is XRay general, extensible and self-tuning?

How does XRay scale with the number of inputs?

How can we manage input overlap?

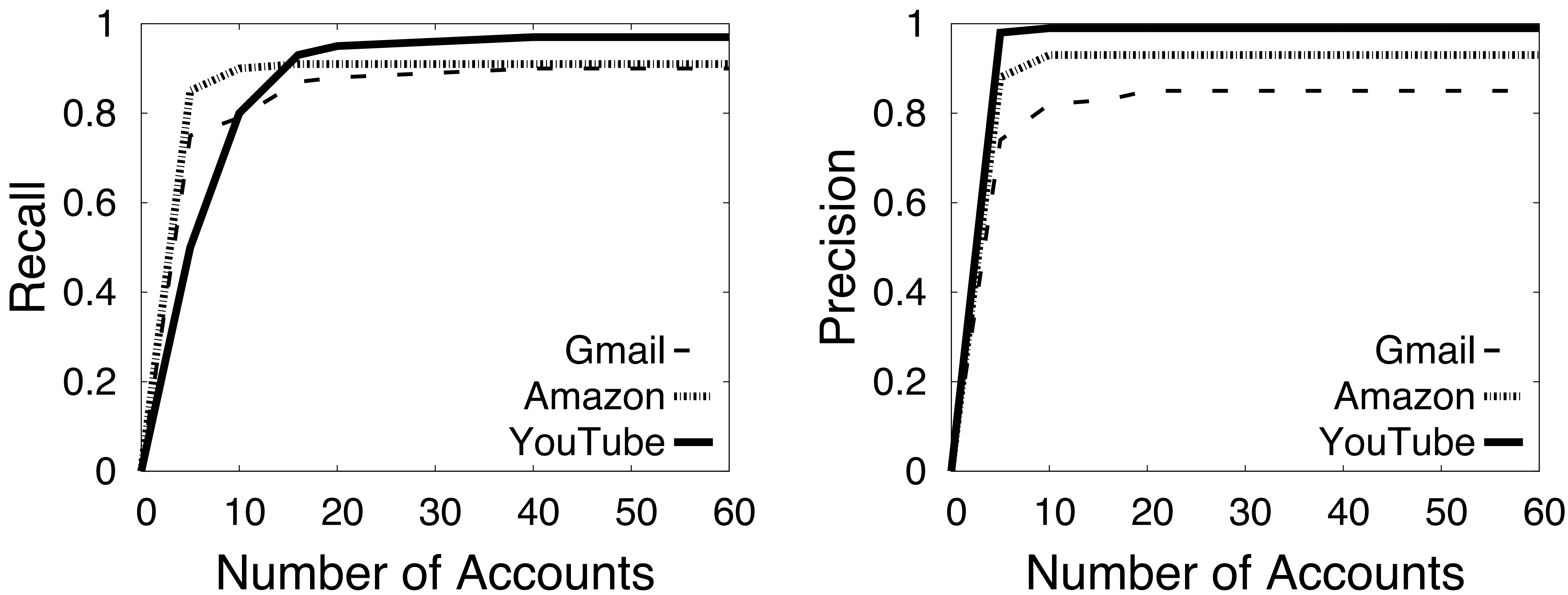
Is XRay useful?

#1 How accurate is XRay?

- We measured recall and precision for XRay's associations on **Gmail**, **YouTube** and **Amazon**
- We need Ground Truth:
  - Ground truth provided by Amazon and YouTube
  - Manual labeling and validation for Gmail

#1 How accurate is XRay?

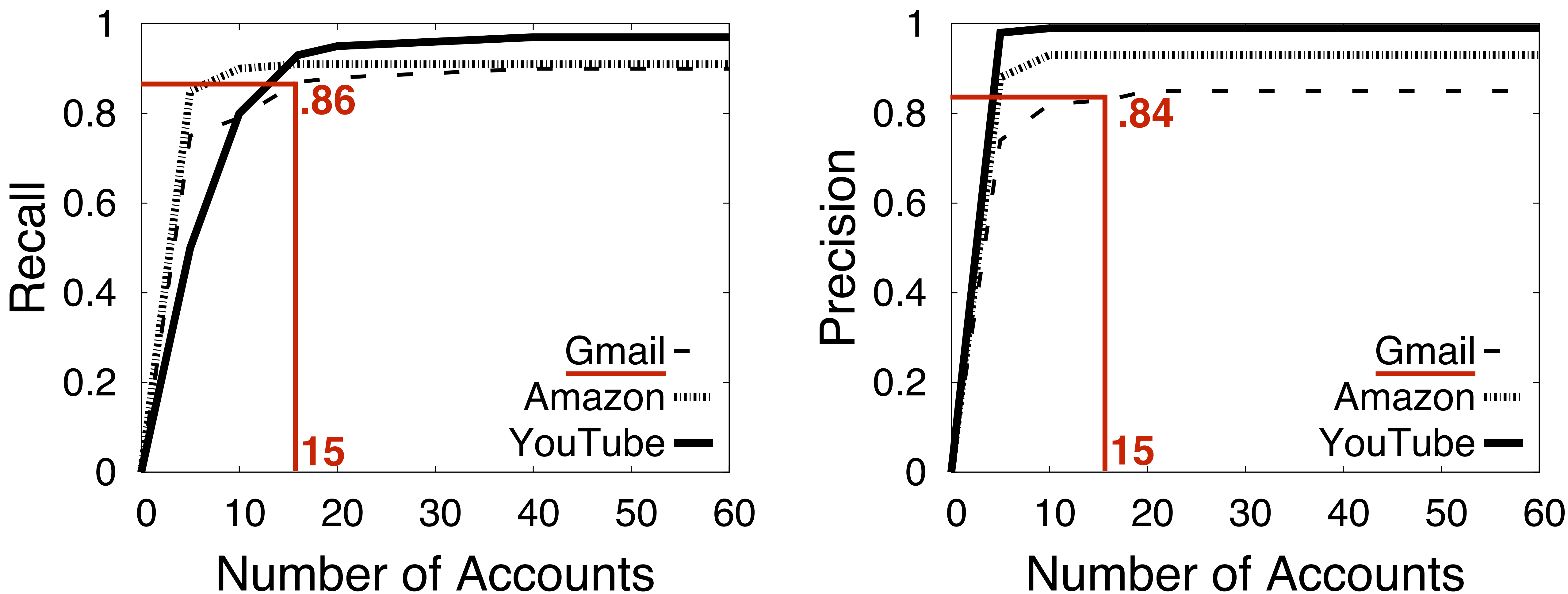
Recall and precision are 80-90% for Gmail, YouTube, and Amazon.



Number of inputs (e.g., emails): 16

#1 How accurate is XRay?

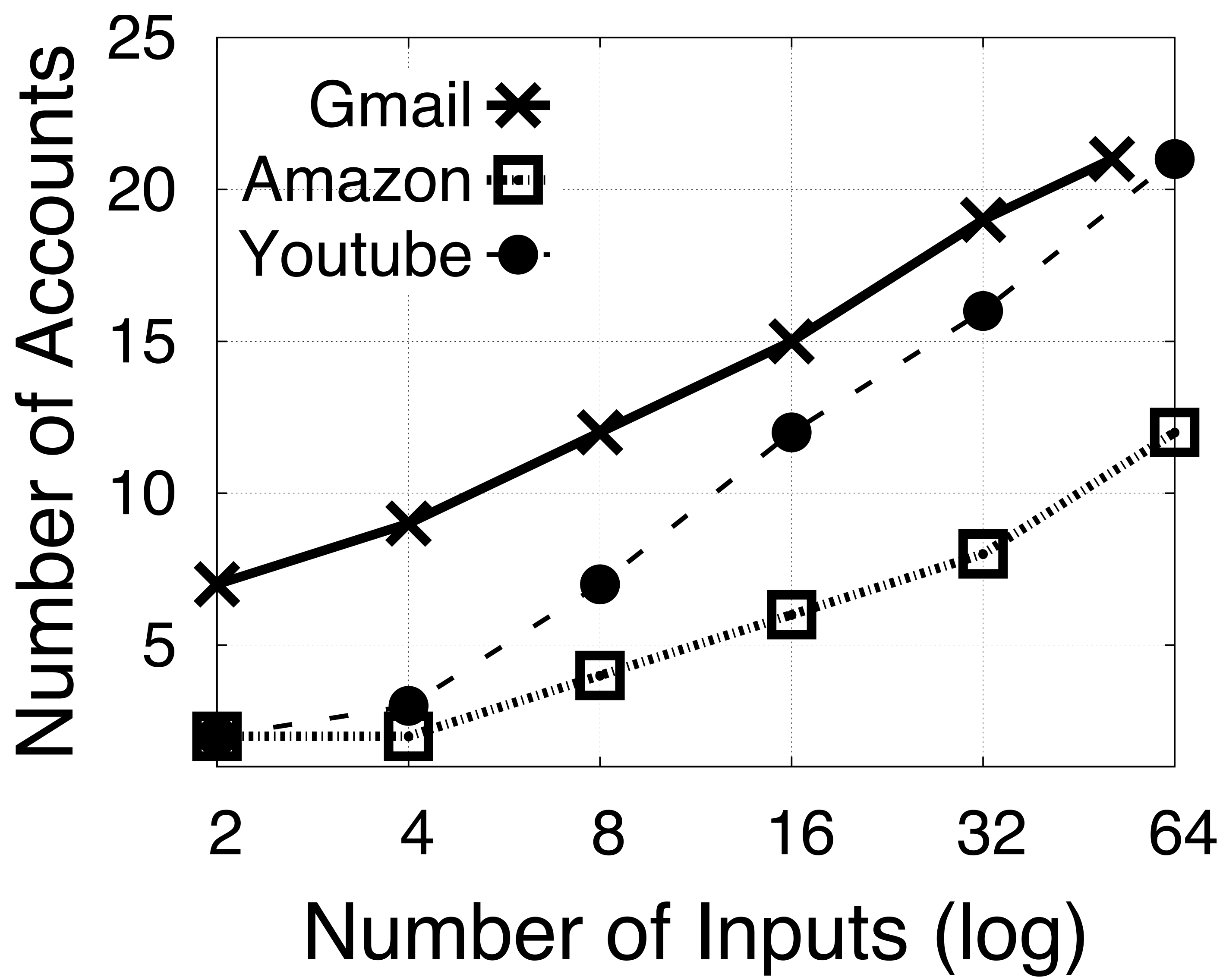
Recall and precision are 80-90% for Gmail, YouTube, and Amazon.



Number of inputs (e.g., emails): 16

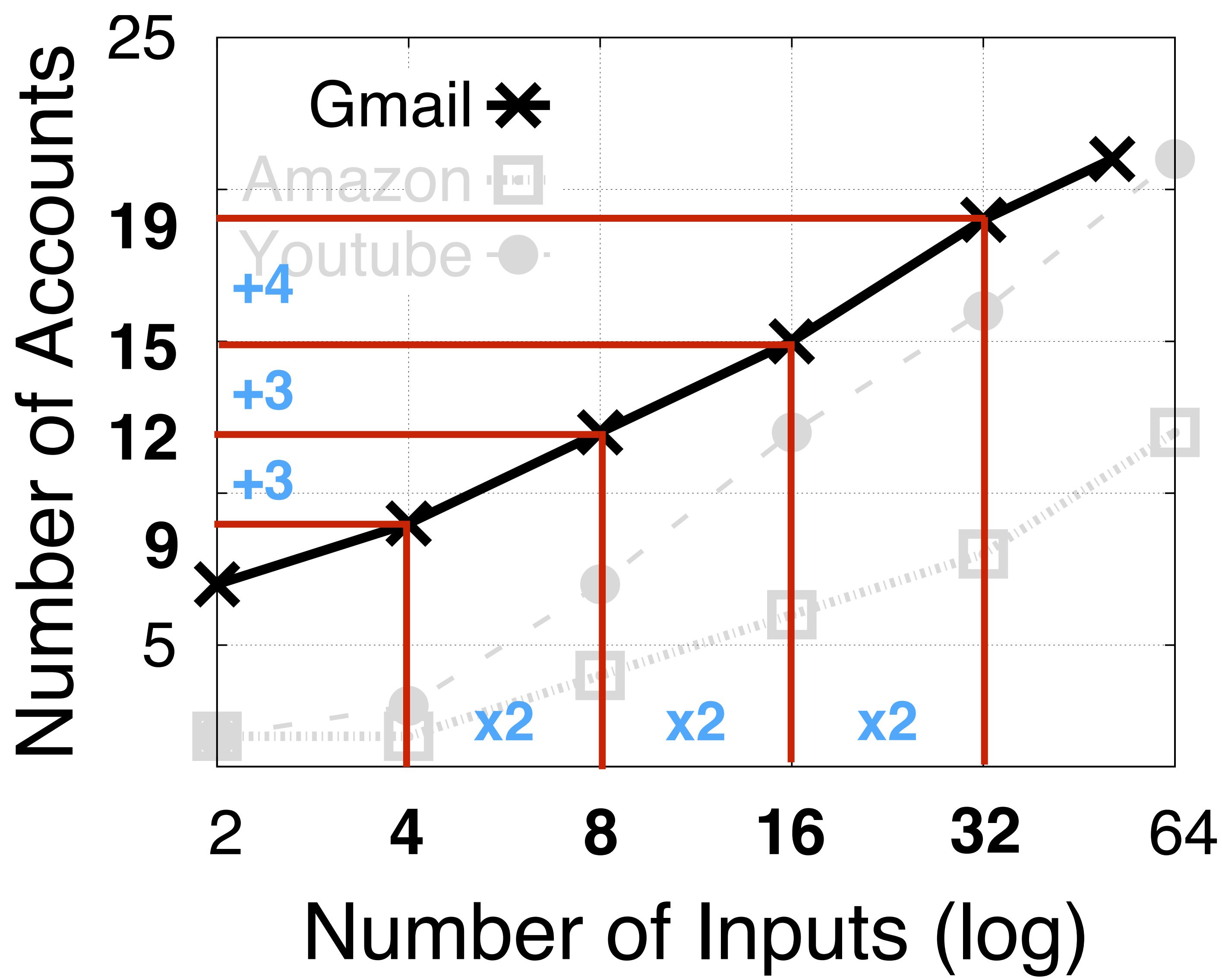
#2 How does XRay scale with the number of inputs?

Logarithmic dependency, as theory predicted.



#2 How does XRay scale with the number of inputs?

Gmail: 2x inputs, +3 accounts.



#3 Is XRay useful?

- We collected ads targeting a few sensitive topics
  - debt
  - pregnancy
  - race
  - various diseases
  - sexual orientation
  - divorce
- For each topics we have one email with relevant keywords
- We analyzed very strong associations detected by XRay

Example associations

Topic	Targeted ads	Score
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99
	Adult Assisted Living. Affordable Assisted Living.	0.99
Cancer	Ford Warriors in Pink. Join The Fight.	1.0
	Rosen Method Bodywork for physical or emotional pain.	0.98
Depression	Shamanic healing over the phone.	0.99
	Text Coach - Get the girl you want and Desire.	0.99
African American	Racial Harassment? Learn your rights now.	0.99
	Racial Harassment, Hearing racial slurs?	0.99
Homosexuality	SF Gay Pride Hotel. Luxury Waterfront.	0.99
	Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.	0.96
Pregnancy	Ralph Lauren Apparel. Official Online Store.	0.99
	Find Baby Shower Invitations. Get up to 60% off here!	0.99
Divorce	Law Attorneys specializing in special needs kids education.	0.99
	Cerbone Law Firm, Helping Good People Thru Bad Times.	0.99
Debt / Loan	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99
	Car Loan without Cosigner 100% Accepted. [...]	0.99
	Car Loans w/ Bad Credit 100% Acceptance! [...]	0.99

Example associations

Topic	Targeted ads	Score
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99
	Adult Assisted Living. Affordable Assisted Living.	0.99
Cancer	Ford Warriors in Pink. Join The Fight.	1.0
	Rosen Method Bodywork for physical or emotional pain.	0.98
Depression	Shamanic healing over the phone.	0.99
	Text Coach - Get the girl you want and Desire.	0.99
African American	Racial Harassment? Learn your rights now.	0.99
	Racial Harassment, Hearing racial slurs?	0.99
Homosexuality	SF Gay Pride Hotel. Luxury Waterfront.	0.99
	Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.	0.96
Pregnancy	Ralph Lauren Apparel. Official Online Store.	0.99
	Find Baby Shower Invitations. Get up to 60% off here!	0.99
Divorce	Law Attorneys specializing in special needs kids education.	0.99
	Cerbone Law Firm, Helping Good People Thru Bad Times.	0.99
Debt / Loan	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99
	Car Loan without Cosigner 100% Accepted. [...]	0.99
	Car Loans w/ Bad Credit 100% Acceptance! [...]	0.99

Example associations

Topic	Targeted ads	Score
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99
	Adult Assisted Living. Affordable Assisted Living.	0.99
Cancer	Ford Warriors in Pink. Join The Fight.	1.0
	Rosen Method Bodywork for physical or emotional pain.	0.98
Depression	Shamanic healing over the phone.	0.99
	Text Coach - Get the girl you want and Desire.	0.99
African American	Racial Harassment? Learn your rights now.	0.99
	Racial Harassment, Hearing racial slurs?	0.99
Homosexuality	SF Gay Pride Hotel. Luxury Waterfront.	0.99
	Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.	0.96
Pregnancy	Ralph Lauren Apparel. Official Online Store.	0.99
	Find Baby Shower Invitations. Get up to 60% off here!	0.99
Divorce	Law Attorneys specializing in special needs kids education.	0.99
	Cerbone Law Firm, Helping Good People Thru Bad Times.	0.99
Debt / Loan	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99
	Car Loan without Cosigner 100% Accepted. [...]	0.99
	Car Loans w/ Bad Credit 100% Acceptance! [...]	0.99

Example associations

Topic	Targeted ads	Score
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99
	Adult Assisted Living. Affordable Assisted Living.	0.99
Cancer	Ford Warriors in Pink. Join The Fight.	1.0
	Rosen Method Bodywork for physical or emotional pain.	0.98
Depression	Shamanic healing over the phone.	0.99
	Text Coach - Get the girl you want and Desire.	0.99
African American	Racial Harassment? Learn your rights now.	0.99
	Racial Harassment, Hearing racial slurs?	0.99
Homosexuality	SF Gay Pride Hotel. Luxury Waterfront.	0.99
	Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.	0.96
Pregnancy	Ralph Lauren Apparel. Official Online Store.	0.99
	Find Baby Shower Invitations. Get up to 60% off here!	0.99
Divorce	Law Attorneys specializing in special needs kids education.	0.99
	Cerbone Law Firm, Helping Good People Thru Bad Times.	0.99
Debt / Loan	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99
	Car Loan without Cosigner 100% Accepted. [...]	0.99
	Car Loans w/ Bad Credit 100% Acceptance! [...]	0.99

Example associations

Topic	Targeted ads	Score
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99
	Adult Assisted Living. Affordable Assisted Living.	0.99
Cancer	Ford Warriors in Pink. Join The Fight.	1.0
	Rosen Method Bodywork for physical or emotional pain.	0.98
Depression	Shamanic healing over the phone.	0.99
	Cerbone Law Firm, Helping Good People Thru Bad Times.	0.99
Debt / Loan	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99
	Car Loan without Cosigner 100% Accepted. [...]	0.99
	Car Loans w/ Bad Credit 100% Acceptance! [...]	0.99

The New York Times

In a Subprime Bubble for Used Cars, Borrowers Pay Sky-High Rates

By JESSICA SILVER-GREENBERG and MICHAEL CORKERY JULY 19, 2014 12:36 PM

Conclusion

- Without transparency, the wonderful big-data Web threatens to become a breeding place for **deceptive and unfair practices**
- **XRay**: the first **generic**, **scalable**, and **accurate** building block for revealing data targeting
  - Relies on differential correlation, which has provable properties
- We hope it will support the building of a **new generation of auditing tools** to keep the big-data Web in check

Code & Data Available

- http://xray.cs.columbia.edu
- Come to me for a demo