Fairness, Transparency, and Privacy in AI @ LinkedIn

Krishnaram Kenthapadi

AI @ LinkedIn

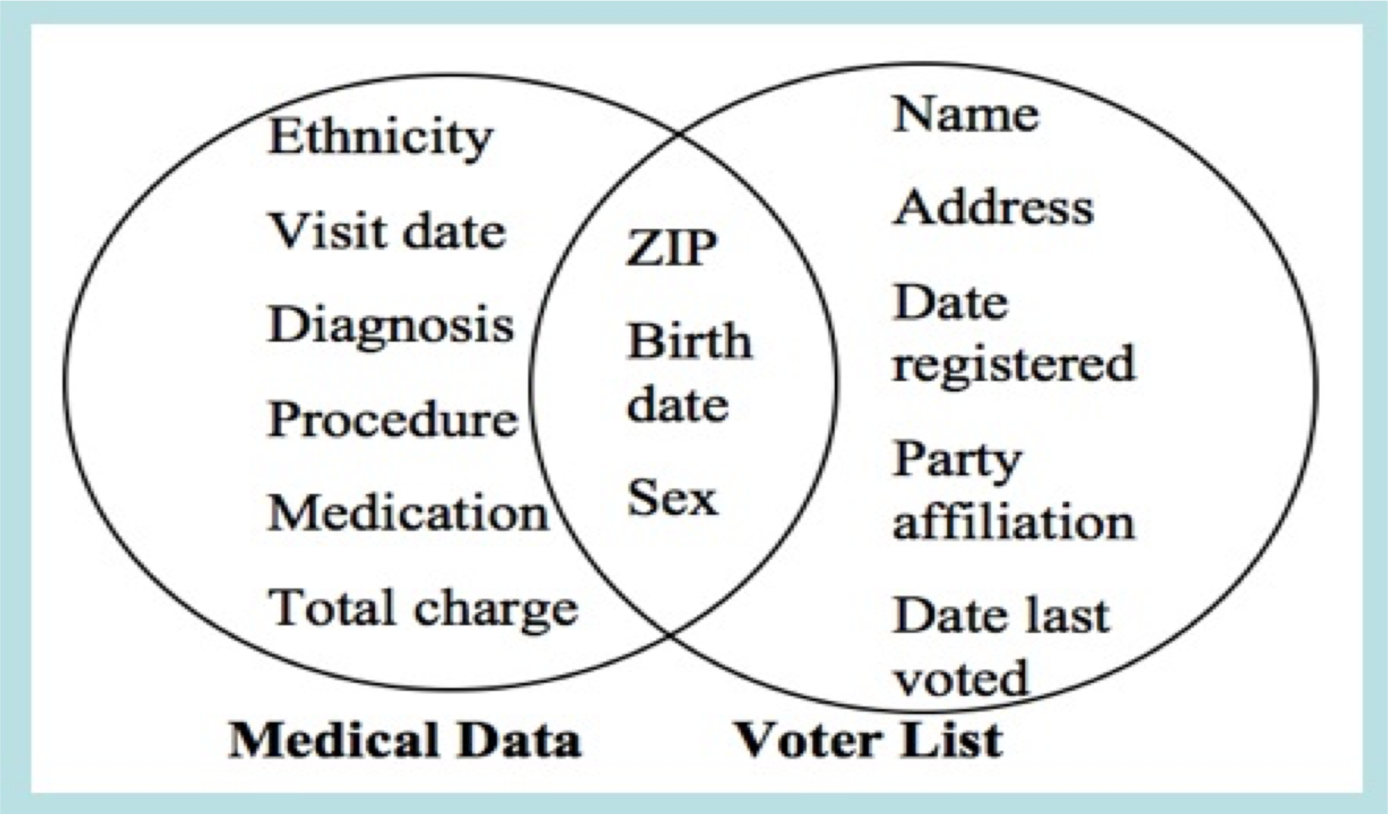
QConSF Talk, November 2018

William Weld vs Latanya Sweeney

Massachusetts Group Insurance Commission (1997):

Anonymized medical history of state employees (all hospital visits, diagnosis, prescriptions)

Latanya Sweeney (MIT grad student): $20 – Cambridge voter roll



born July 31, 1945
resident of 02138

64%

uniquely identifiable with
ZIP + birth date + gender
(in the US population)

Golle, “Revisiting the Uniqueness of Simple Demographics in the US Population”, WPES 2006

Netflix Prize

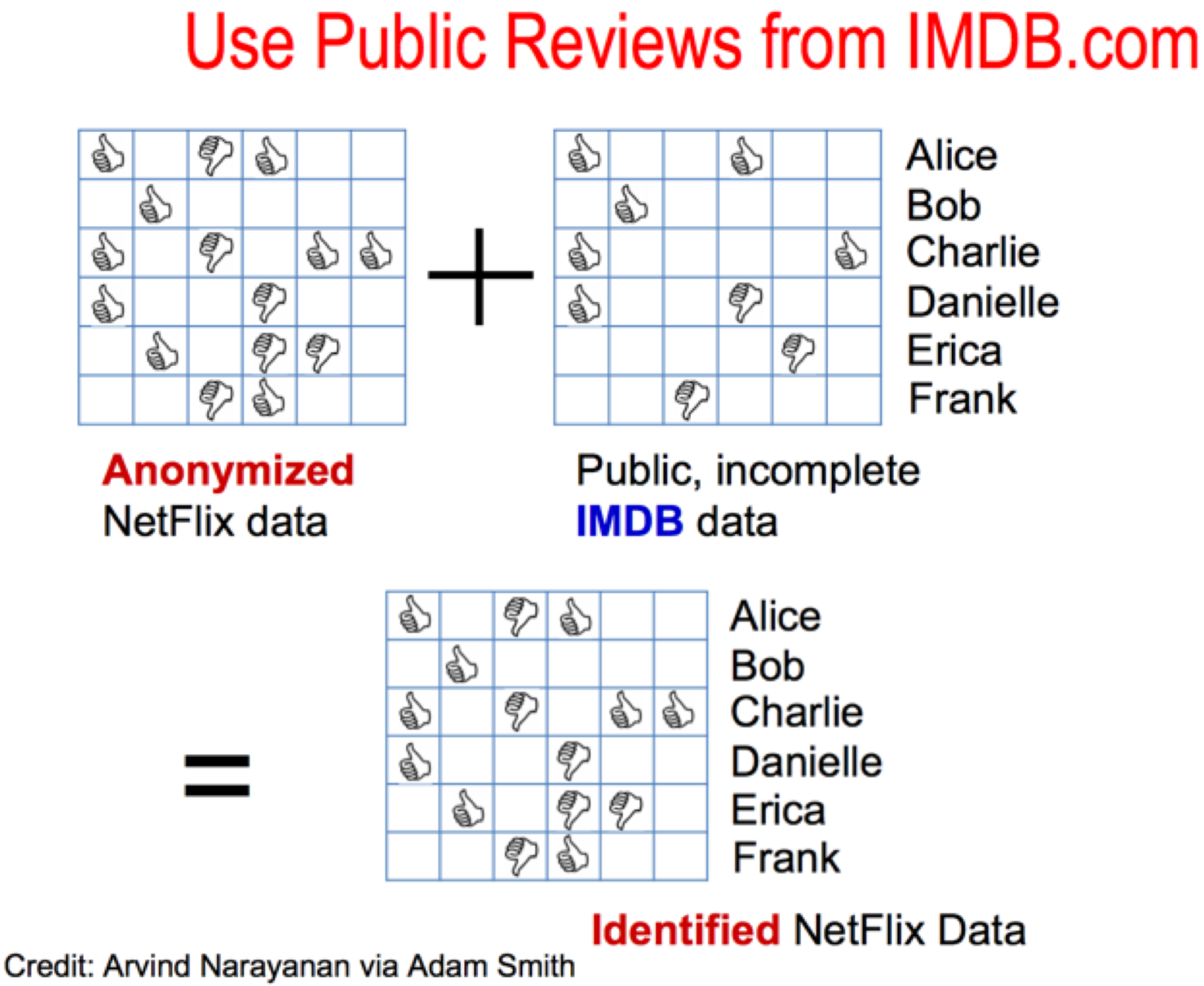
Netflix Prize

Oct 2006: Netflix announces Netflix Prize

- 10% of their users
- average 200 ratings per user

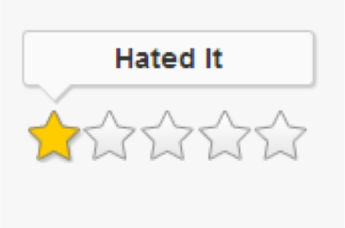
Narayanan, Shmatikov (2006):

Deanonymizing Netflix Data

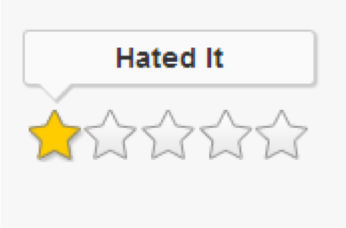


Narayanan, Shmatikov, [Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)](), 2008

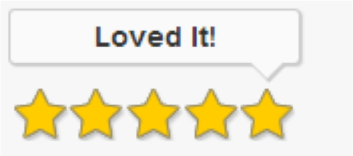
- Noam Chomsky in Our Times



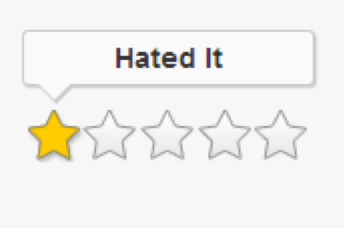
- Farenheit 9/11



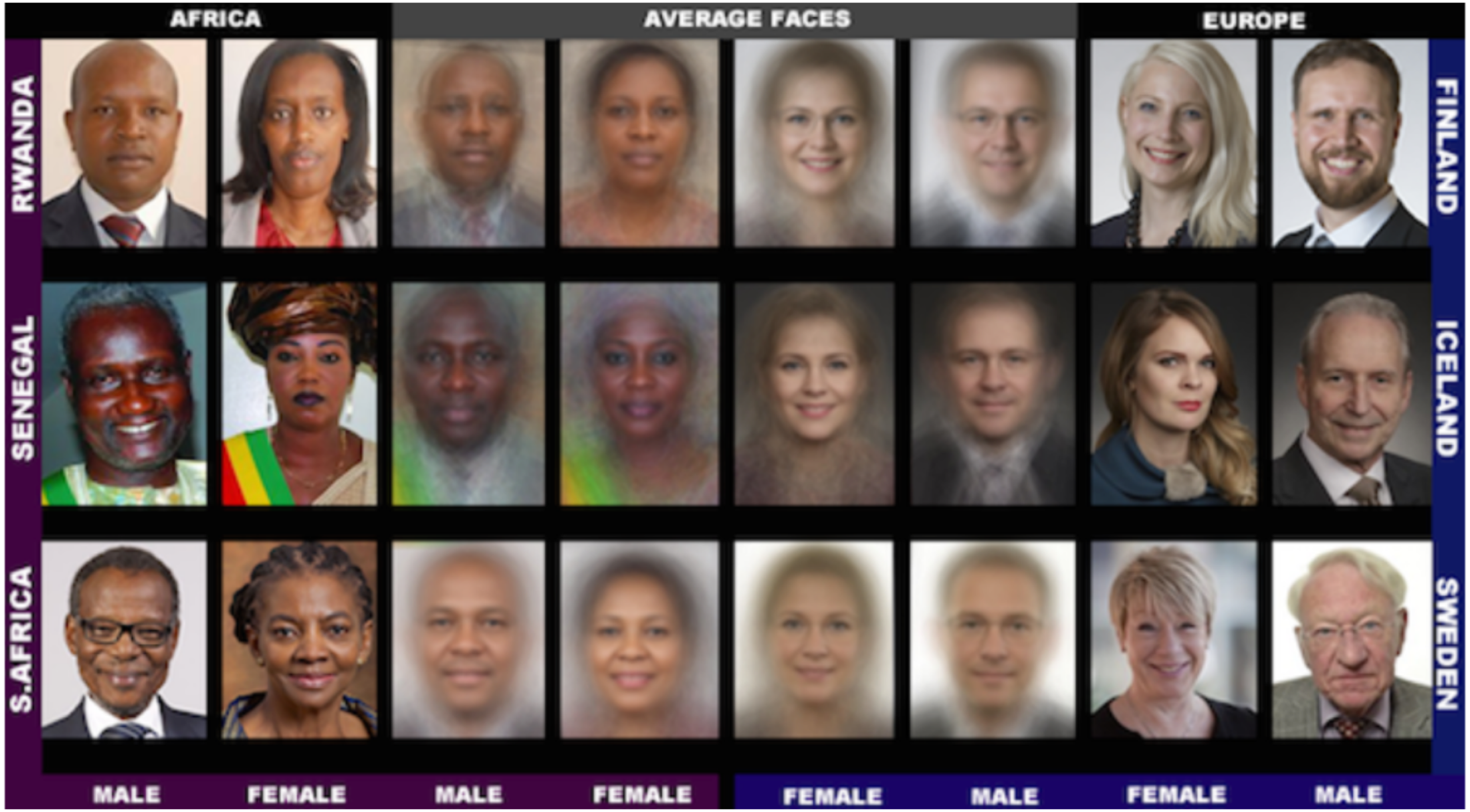
- Jesus of Nazareth



- Queer as Folk



Gender Shades [Joy Buolamwini & Timnit Gebru, 2018]



- Facial recognition software:
  Higher accuracy for light skinned men
- Error rates for dark skinned women: 20% - 34%

Algorithmic Bias

- Ethical challenges posed by AI systems

- Inherent biases present in society
  - Reflected in training data

  - AI/ML models prone to amplifying such biases
    - ACM FAT* conference / KDD’16 & NIPS’17 Tutorials

“Privacy and Fairness by Design” for AI products

AI @ LinkedIn

Case Studies @ LinkedIn

- Privacy
- Fairness

LinkedIn's Vision

Create economic opportunity for every member of the global workforce

LinkedIn's Mission

Connect the world's professionals to make them more productive and successful

LinkedIn Economic Graph

575M
Members

26M
Companies

15M
Jobs

50K
Skills

60K
Schools

109B
Updates viewed

AI @LinkedIn

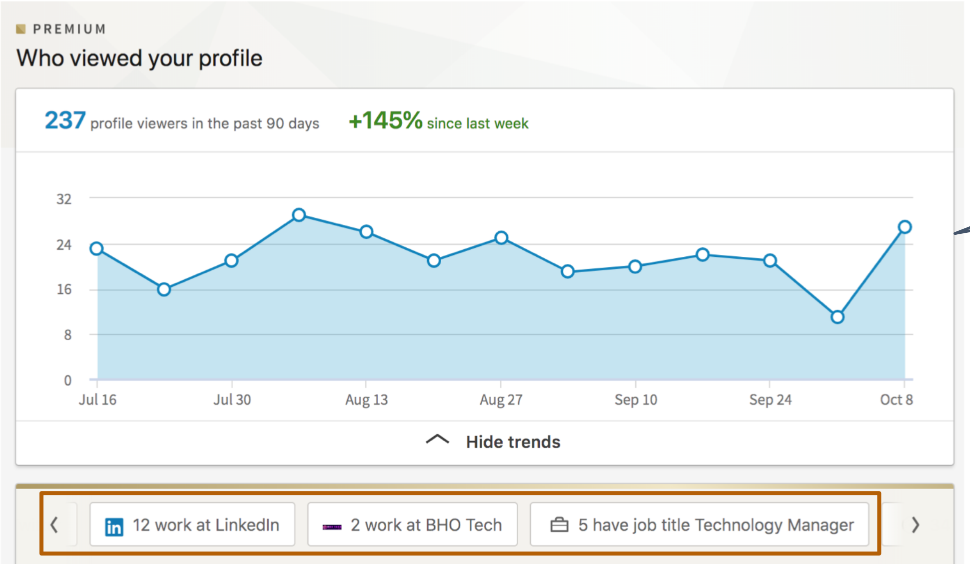
Scale

2 PB	2.15 PB	25 B	200	53 B
data processed offline per day	data processed nearline per day	parameters in ML models	ML A/B experiments per week	graph edges with 1B nodes

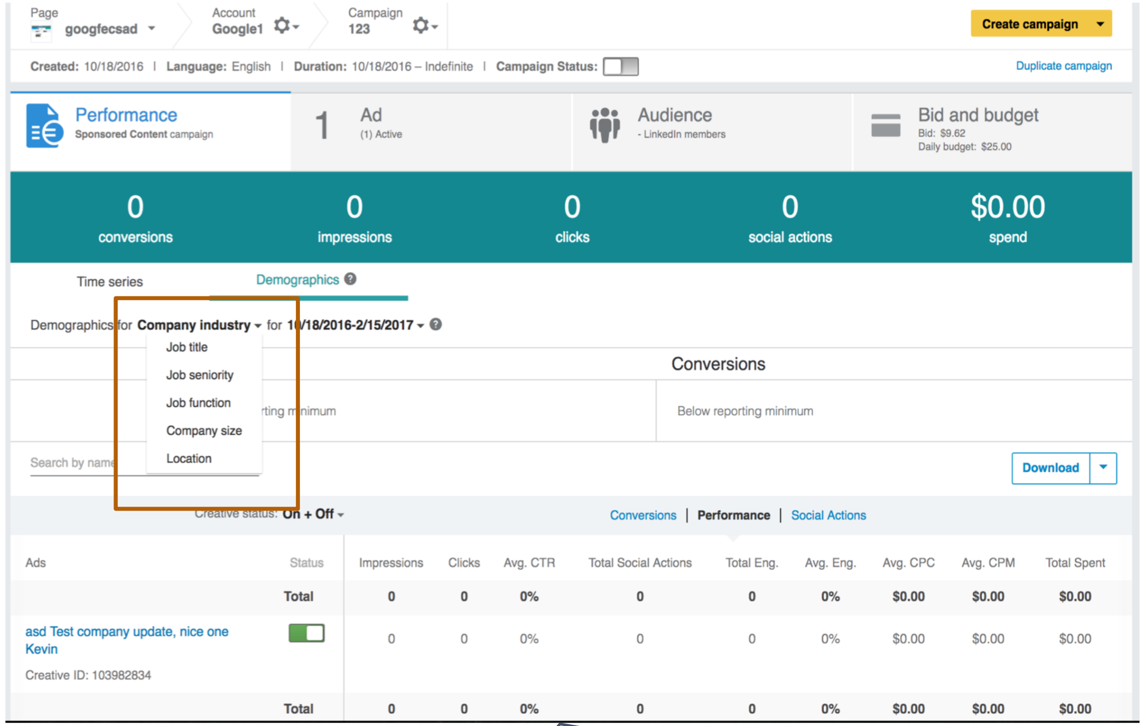
Privacy in AI @ LinkedIn

- **Framework to compute robust, privacy-preserving analytics**
- **Privacy challenges/design for a large crowdsourced system (LinkedIn Salary)**

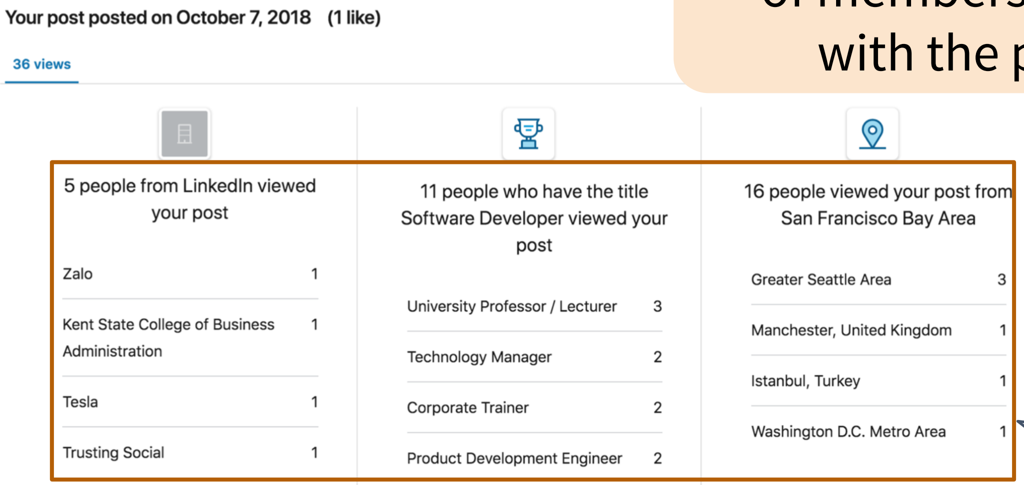
Analytics & Reporting Products at LinkedIn



Profile View Analytics



All showing demographics of members engaging with the product



Content Analytics

Ad Campaign Analytics

Analytics & Reporting Products at LinkedIn

- Admit only a small # of predetermined query types
  - Querying for the number of member actions, for a specified time period, together with the top demographic breakdowns

"SELECT COUNT(*) FROM table(statType, entity) WHERE timeStamp $\geq$ startTime AND timeStamp $\leq$ endTime AND $d_{attr} = d_{val}$"

Analytics & Reporting Products at LinkedIn

- Admit only a small # of predetermined query types
  - Querying for the number of member actions, for a specified time period, together with the top demographic breakdowns

E.g., *Clicks on a given ad*

E.g., Title = “Senior Director”

“SELECT COUNT(*) FROM table(statType, entity) WHERE timeStamp $\geq$ startTime AND timeStamp $\leq$ endTime AND $d_{attr} = d_{val}$”

Privacy Requirements

- Attacker cannot infer whether a member performed an action
  - E.g., click on an article or an ad
- Attacker may use auxiliary knowledge
  - E.g., knowledge of attributes associated with the target member (say, obtained from this member’s LinkedIn profile)
  - E.g., knowledge of all other members that performed similar action (say, by creating fake accounts)

Possible Privacy Attacks

Targeting: Senior directors in US, who studied at Cornell	Matches ~16k LinkedIn members → over minimum targeting threshold ✔
Demographic breakdown: Company = X	May match exactly one person → can determine whether the person clicks on the ad or not ✘
Require minimum reporting threshold	Attacker could create fake profiles! E.g. if threshold is 10, create 9 fake profiles that all click. ✘
Rounding mechanism E.g., report incremental of 10	Still amenable to attacks E.g. using incremental counts over time to infer individuals' actions ✘

Need rigorous techniques to preserve member privacy
(not reveal exact aggregate counts)

Key Product Desiderata

- Coverage & Utility

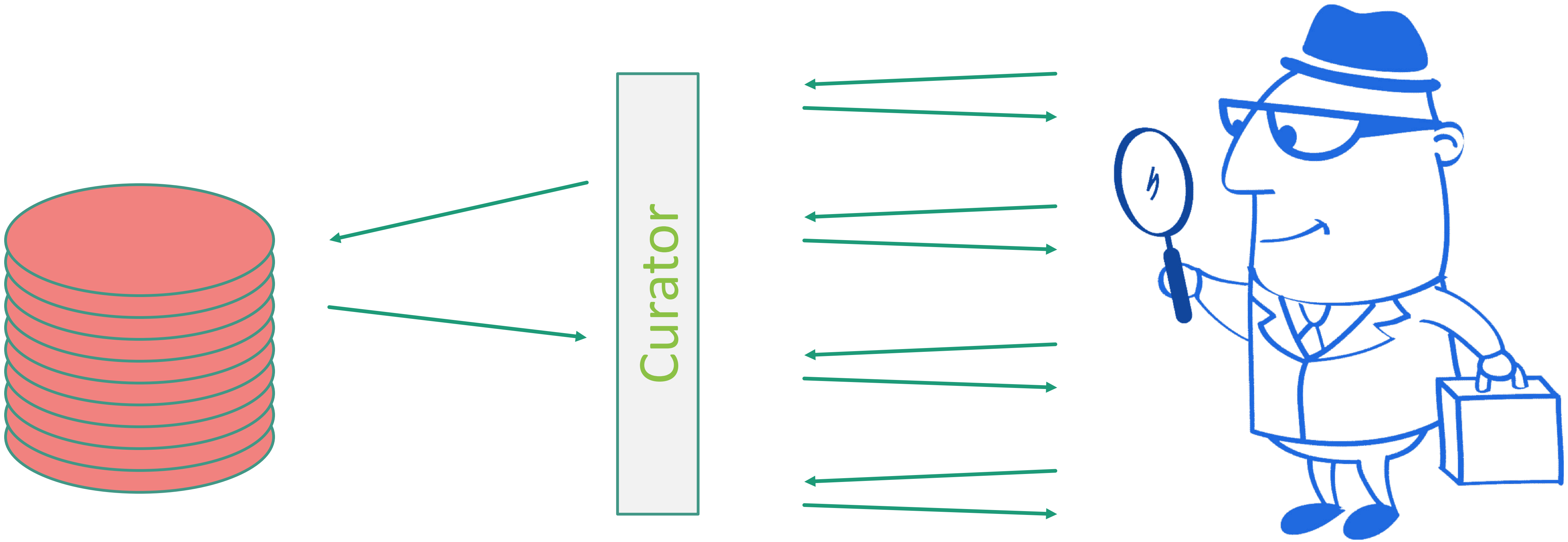
- Data Consistency
  - for repeated queries
  - over time
  - between total and breakdowns
  - across entity/action hierarchy
  - for top k queries

Problem Statement

Compute robust, reliable analytics in a privacy-preserving manner, while addressing the product desiderata such as coverage, utility, and consistency.

Differential Privacy

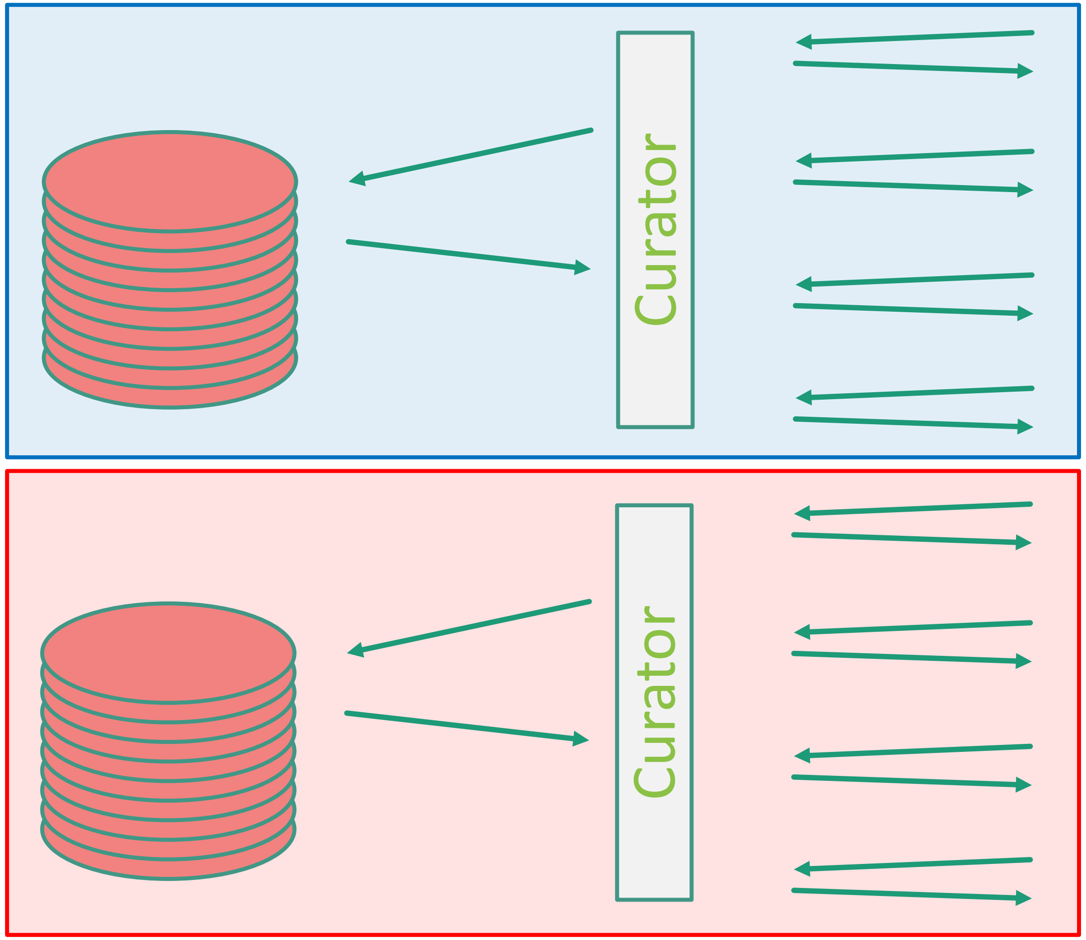
Defining Privacy



Defining Privacy

Your data in
the database

~~Your data in
the database~~

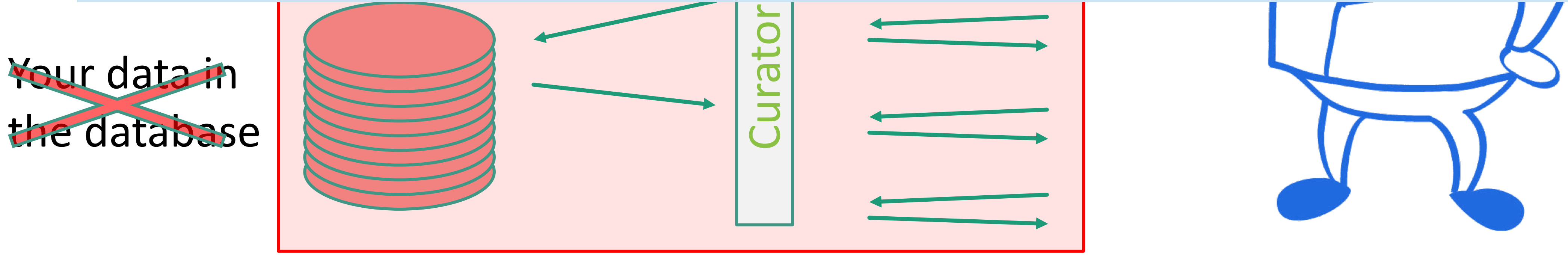


Defining Privacy

Intuition:

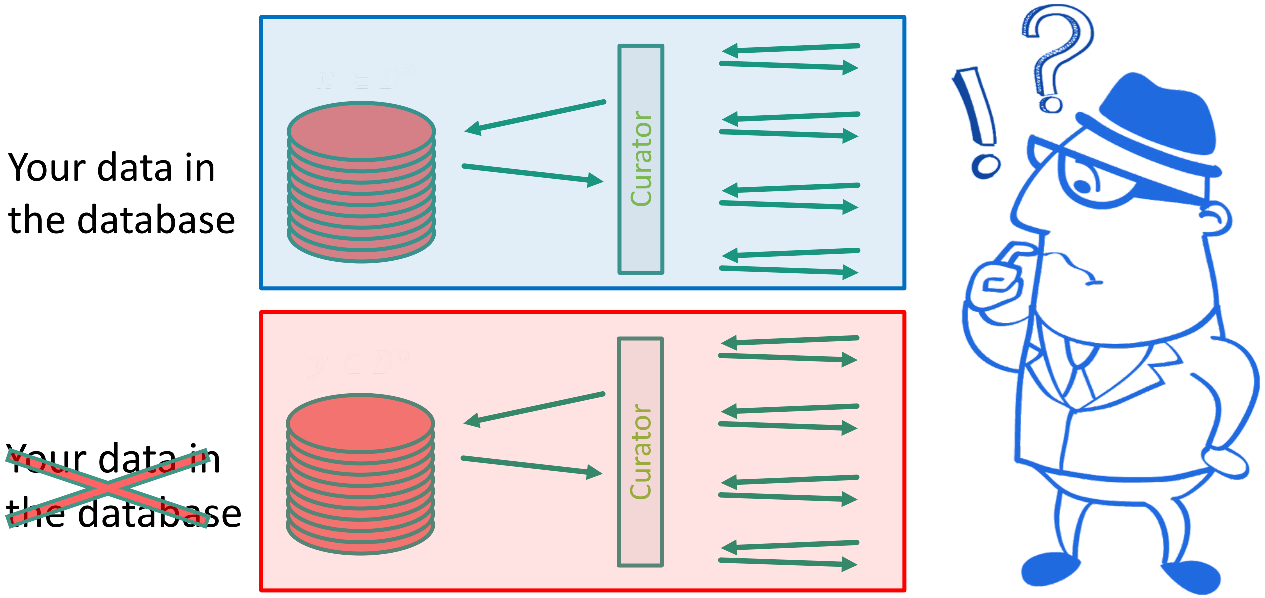
- A member's privacy is preserved if …
  - *"The released result would nearly be the same, whether or not the user's information is taken into account"*

- An attacker gains very little additional knowledge about any specific member from the published result



Differential Privacy

ε-Differential Privacy: For neighboring databases D and D' (differ by one record), the distribution of the curator's output $f(D)$ on database D is (nearly) the same as $f(D')$. $\forall S$: $\Pr[f(D) \in S] \leq \exp(\varepsilon) \cdot \Pr[f(D') \in S]$



Parameter ε quantifies information leakage
(smaller ε, more private)

Dwork, McSherry, Nissim, Smith, "Calibrating Noise to Sensitivity in Private Data Analysis", TCC 2006

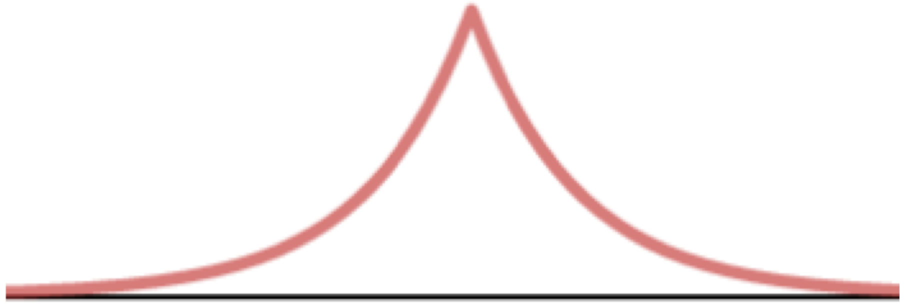
Differential Privacy: Random Noise Addition

If ℓ_1-sensitivity of $f : \mathrm{D} \to \mathbb{R}^n$:

$$\max_{D,D'} \|f(D) - f(D')\|_1 = \mathrm{s},$$

then adding Laplacian noise to true output

$$f(D) + \mathrm{Laplace}^n(\mathrm{s}/\varepsilon)$$



offers ε-differential privacy.

Dwork, McSherry, Nissim, Smith, “Calibrating Noise to Sensitivity in Private Data Analysis”, TCC 2006

PriPeARL: A Framework for Privacy-Preserving Analytics

K. Kenthapadi, T. T. L. Tran, ACM CIKM 2018

Pseudo-random noise generation, inspired by differential privacy

- Entity id (e.g., ad creative/campaign/account)
- Demographic dimension
- Stat type (impressions, clicks)
- Time range
- Fixed secret seed

→

Uniformly Random Fraction

- Cryptographic hash
- Normalize to (0,1)

→

Laplace Noise

- Fixed ε

→

True Count

↓

Random Noise

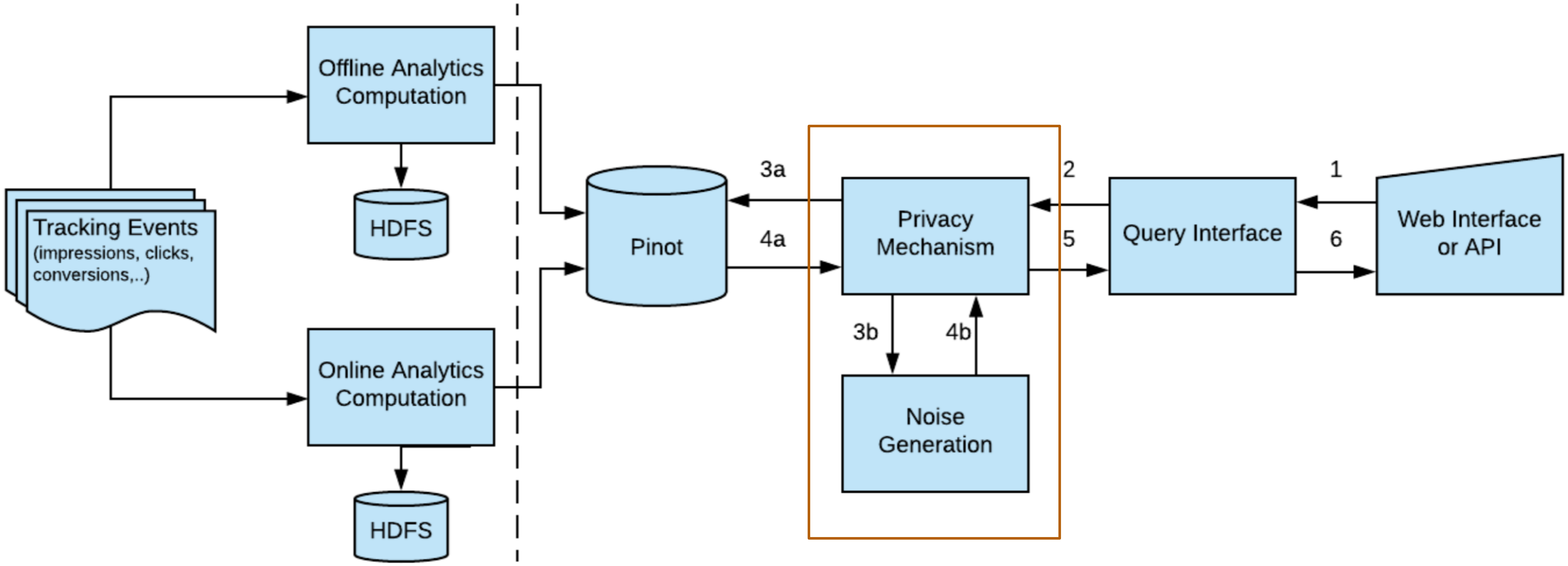
↓

Noisy Count

- Pseudo-random noise → same query has same result over time, avoid averaging attack.
- For non-canonical queries (e.g., time ranges, aggregate multiple entities)
  - Use the hierarchy and partition into canonical queries
  - Compute noise for each canonical queries and sum up the noisy counts

To satisfy consistency requirements

System Architecture



Lessons Learned from Deployment (> 1 year)

- Semantic consistency vs. unbiased, unrounded noise
- Suppression of small counts
- Online computation and performance requirements
- Scaling across analytics applications
  - Tools for ease of adoption (code/API library, hands-on how-to tutorial) help!

Summary

- Framework to compute robust, privacy-preserving analytics
  - Addressing challenges such as preserving member privacy, product coverage, utility, and data consistency

- Future
  - Utility maximization problem given constraints on the 'privacy loss budget' per user
    - E.g., noise with larger variance to impressions but less noise to clicks (or conversions)
    - E.g., more noise to broader time range sub-queries and less noise to granular time range sub-queries

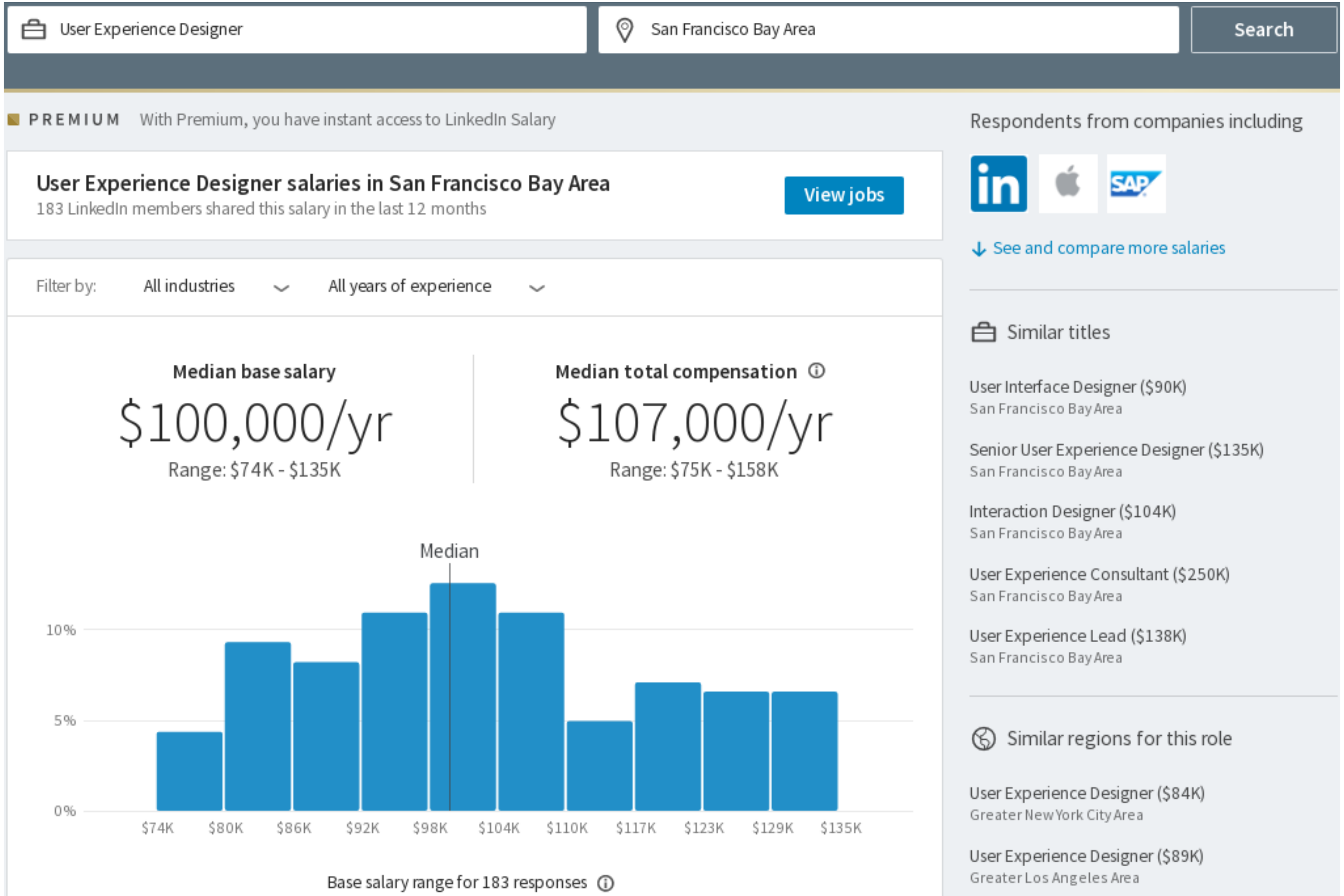
- Reference: K. Kenthapadi, T. Tran, [PriPeARL: A Framework for Privacy-Preserving Analytics and Reporting at LinkedIn](), ACM CIKM 2018.

Acknowledgements

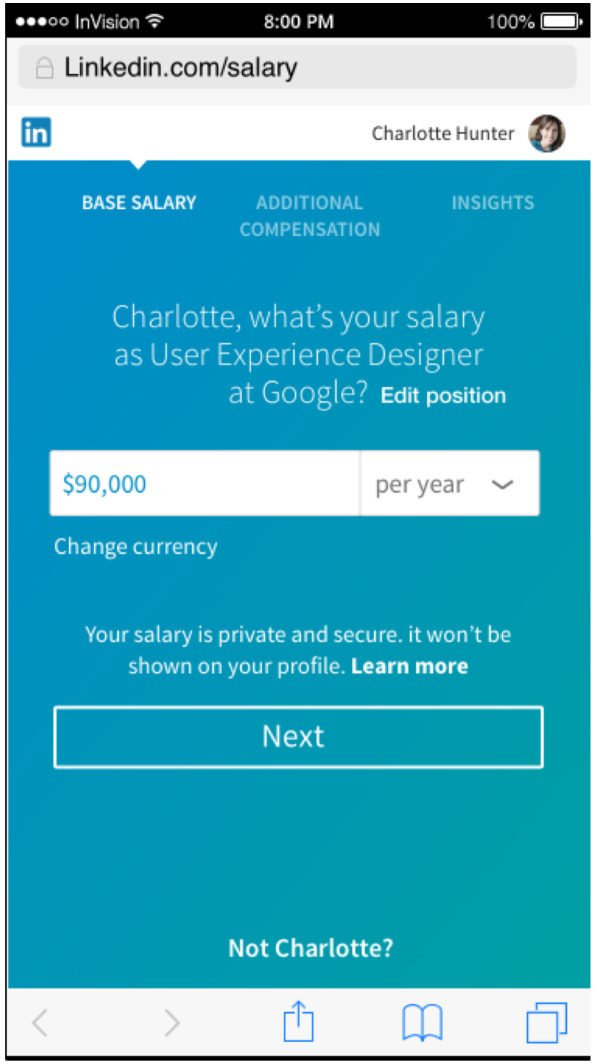
- Team:
  - AI/ML: Krishnaram Kenthapadi, Thanh T. L. Tran
  - Ad Analytics Product & Engineering: Mark Dietz, Taylor Greason, Ian Koeppe
  - Legal / Security: Sara Harrington, Sharon Lee, Rohit Pitke
- Acknowledgements (in alphabetical order)
  - Deepak Agarwal, Igor Perisic, Arun Swami

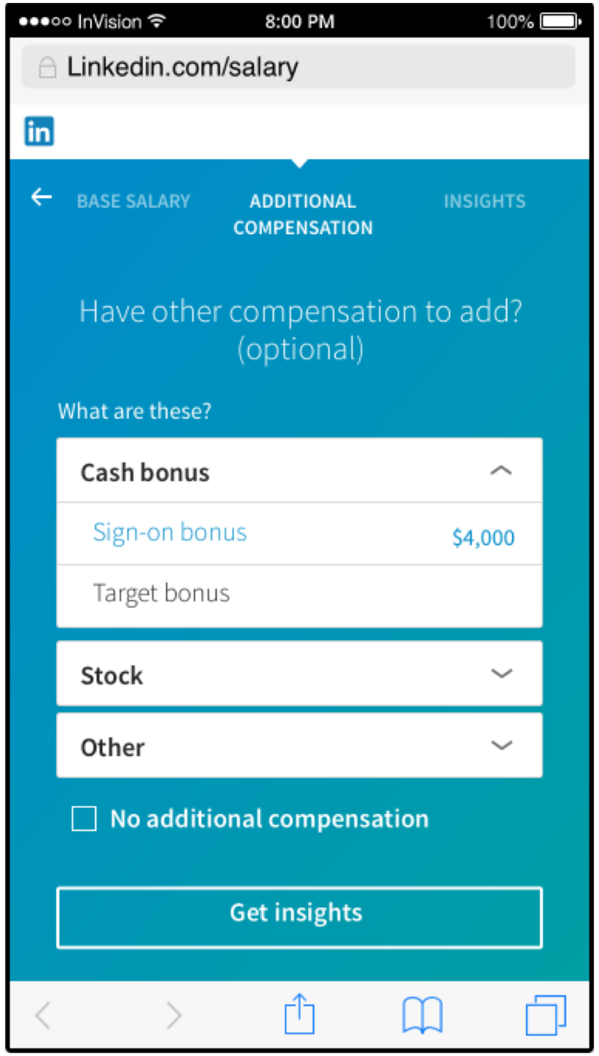
LinkedIn Salary

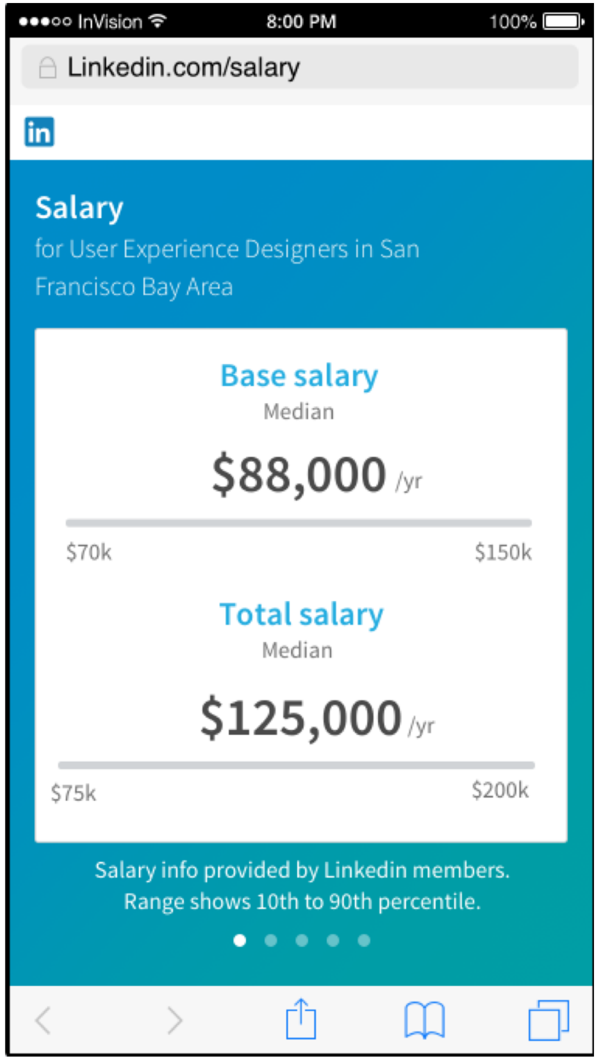
LinkedIn Salary (launched in Nov, 2016)



Salary Collection Flow via Email Targeting







Data Privacy Challenges

- Minimize the risk of inferring any one individual's compensation data
- Protection against data breach
  - No single point of failure

Achieved by a combination of techniques: encryption, access control, de-identification, aggregation, thresholding

K. Kenthapadi, A. Chudhary, and S. Ambler, LinkedIn Salary: A System for Secure Collection and Presentation of Structured Compensation Insights to Job Seekers, IEEE PAC 2017 (arxiv.org/abs/1705.06976)

Problem Statement

- *How do we design LinkedIn Salary system taking into account the unique privacy and security challenges, while addressing the product requirements?*

De-identification Example

Title	Region	Company	Industry	Years of exp	Degree	FoS	Skills	$$
User Exp Designer	SF Bay Area	Google	Internet	12	BS	Interactive Media	UX, Graphics, …	100K

Title	Region	$$
User Exp Designer	SF Bay Area	100K
User Exp Designer	SF Bay Area	115K
…	…	…

Title	Region	Industry	$$
User Exp Designer	SF Bay Area	Internet	100K

Title	Region	Years of exp	$$
User Exp Designer	SF Bay Area	10+	100K

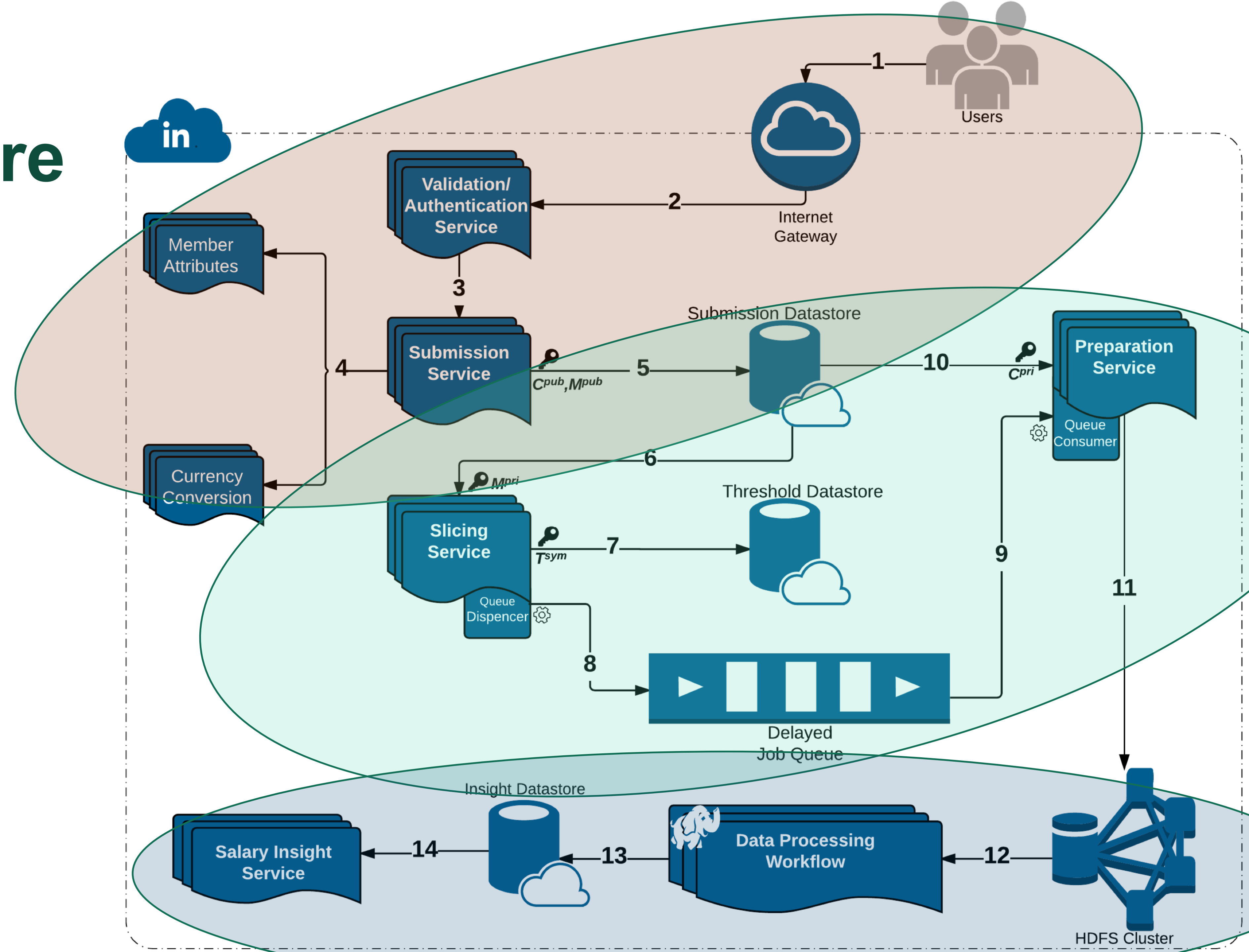
Title	Region	Company	Years of exp	$$
User Exp Designer	SF Bay Area	Google	10+	100K

#data points > threshold?

Yes ⇒ Copy to Hadoop (HDFS)

Note: Original submission stored as encrypted objects.

System Architecture



Summary

- LinkedIn Salary: a new internet application, with unique privacy/modeling challenges
- Privacy Design & Architecture
- Provably private submission of compensation entries?

Acknowledgements

- Team:
  - AI/ML: Krishnaram Kenthapadi, Stuart Ambler, Xi Chen, Yiqun Liu, Parul Jain, Liang Zhang, Ganesh Venkataraman, Tim Converse, Deepak Agarwal
  - Application Engineering: Ahsan Chudhary, Alan Yang, Alex Navasardyan, Brandyn Bennett, Hrishikesh S, Jim Tao, Juan Pablo Lomeli Diaz, Patrick Schutz, Ricky Yan, Lu Zheng, Stephanie Chou, Joseph Florencio, Santosh Kumar Kancha, Anthony Duerr
  - Product: Ryan Sandler, Keren Baruch
  - Other teams (UED, Marketing, BizOps, Analytics, Testing, Voice of Members, Security, …): Julie Kuang, Phil Bunge, Prateek Janardhan, Fiona Li, Bharath Shetty, Sunil Mahadeshwar, Cory Scott, Tushar Dalvi, and team
- Acknowledgements (in alphabetical order)
  - David Freeman, Ashish Gupta, David Hardtke, Rong Rong, Ram Swaminathan

What's Next: Privacy for ML / Data Applications

- Hard open questions
  - Can we simultaneously develop highly personalized models and ensure that the models do not encode private information of members?
  - How do we guarantee member privacy over time without exhausting the "privacy loss budget"?
  - How do we enable privacy-preserving mechanisms for data marketplaces?

Fairness in AI @ LinkedIn

Guiding Principle:
“Diversity by Design”

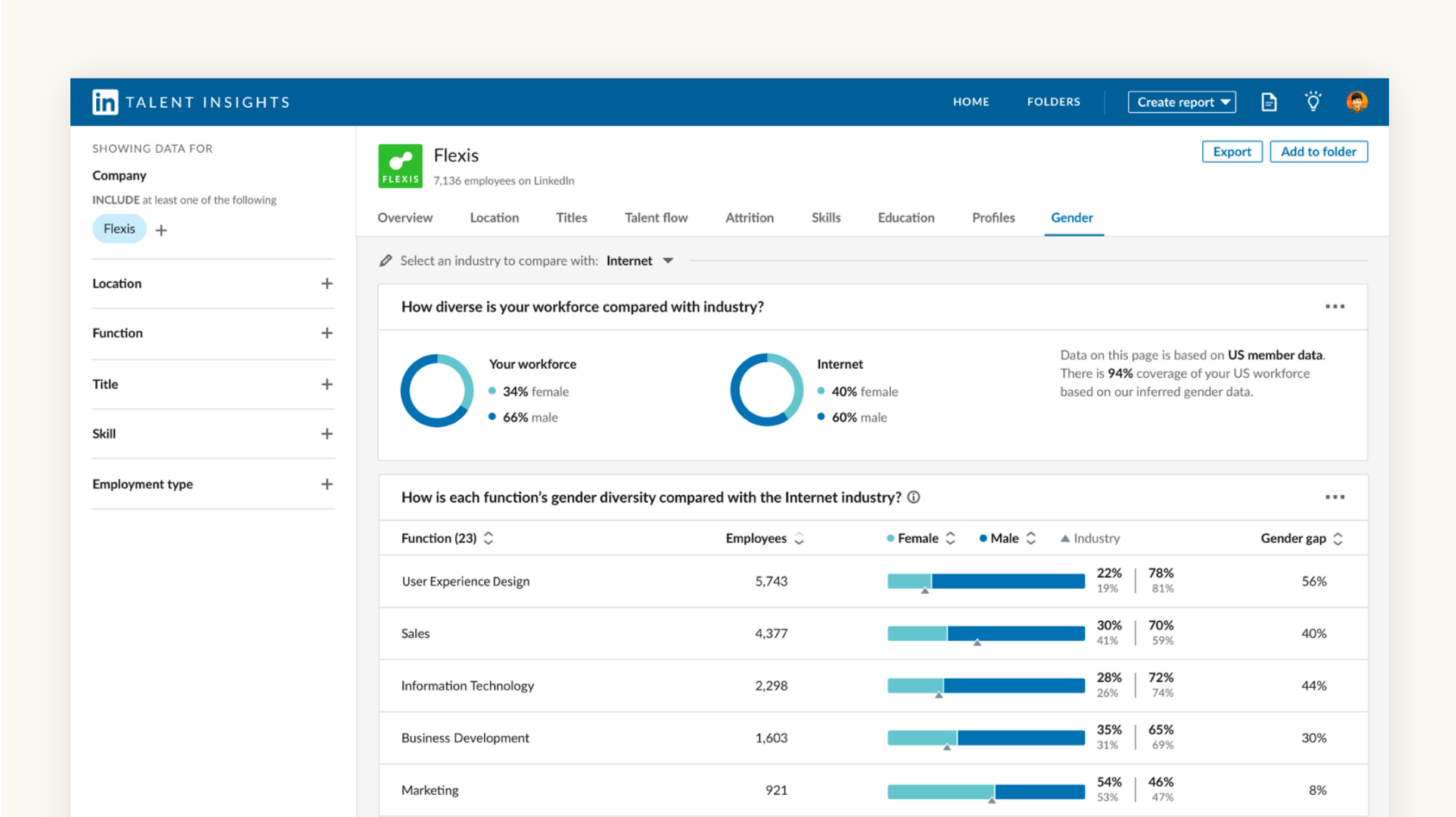
“Diversity by Design” in LinkedIn’s Talent Solutions

Insights to Identify Diverse Talent Pools

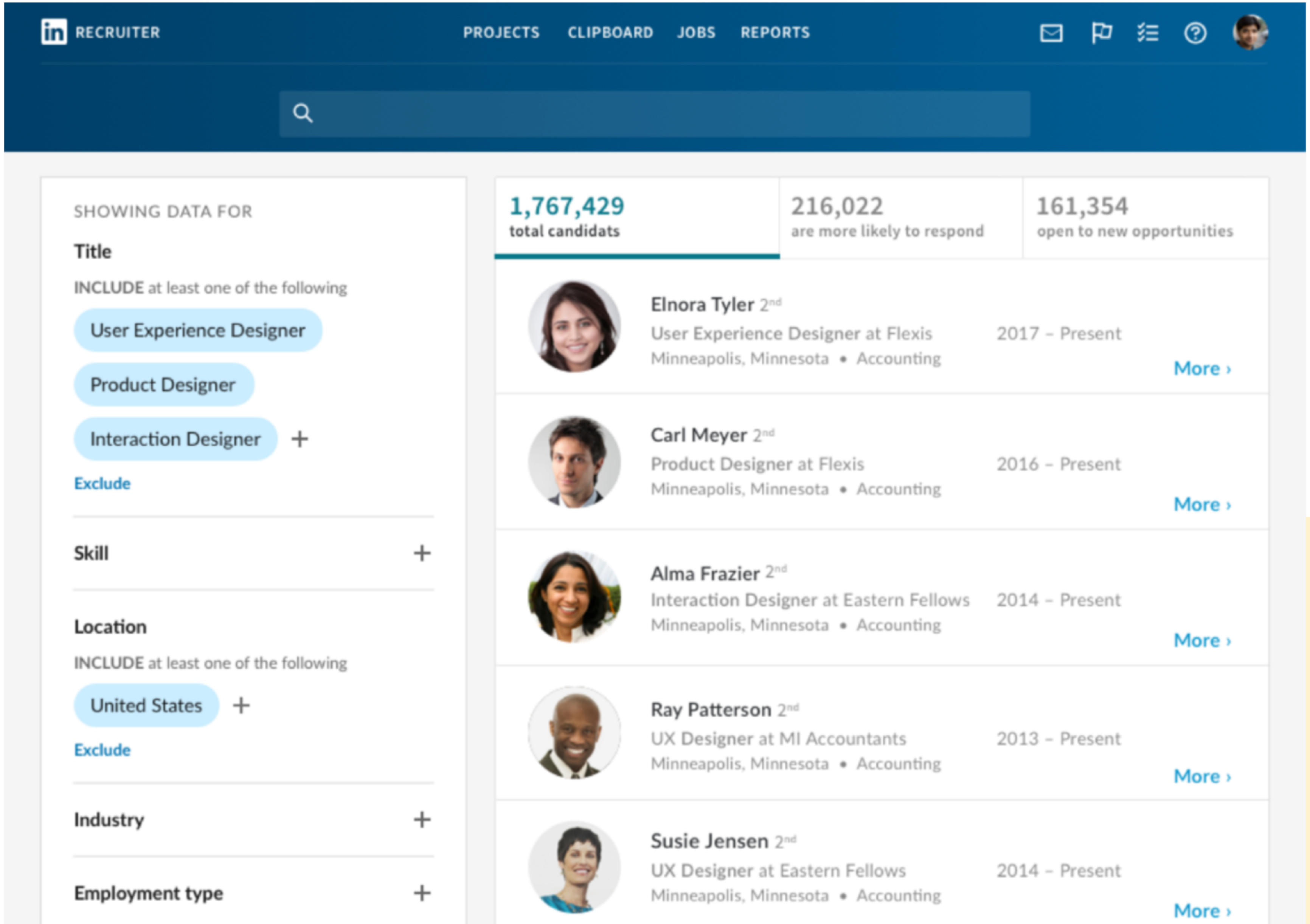
Representative Talent Search Results

Diversity Learning Curriculum

Plan for Diversity



Representative Ranking for Talent Search



S. C. Geyik, K. Kenthapadi, [Building Representative Talent Search at LinkedIn], LinkedIn engineering blog post, October’18.

Intuition for Measuring Representativeness

- Ideal: same distribution on gender/age/… for
  - Top ranked results and
  - Qualified candidates for a search request
    - LinkedIn members matching the search criteria

- Same proportion of members with each given attribute value across both these sets

- "Equal opportunity" definition [Hardt et al, NIPS'16]

Reranking Algorithm for Representativeness

- Determine the target proportions within the attribute of interest, corresponding to a search request

- Compute a fairness-aware ranking of size k

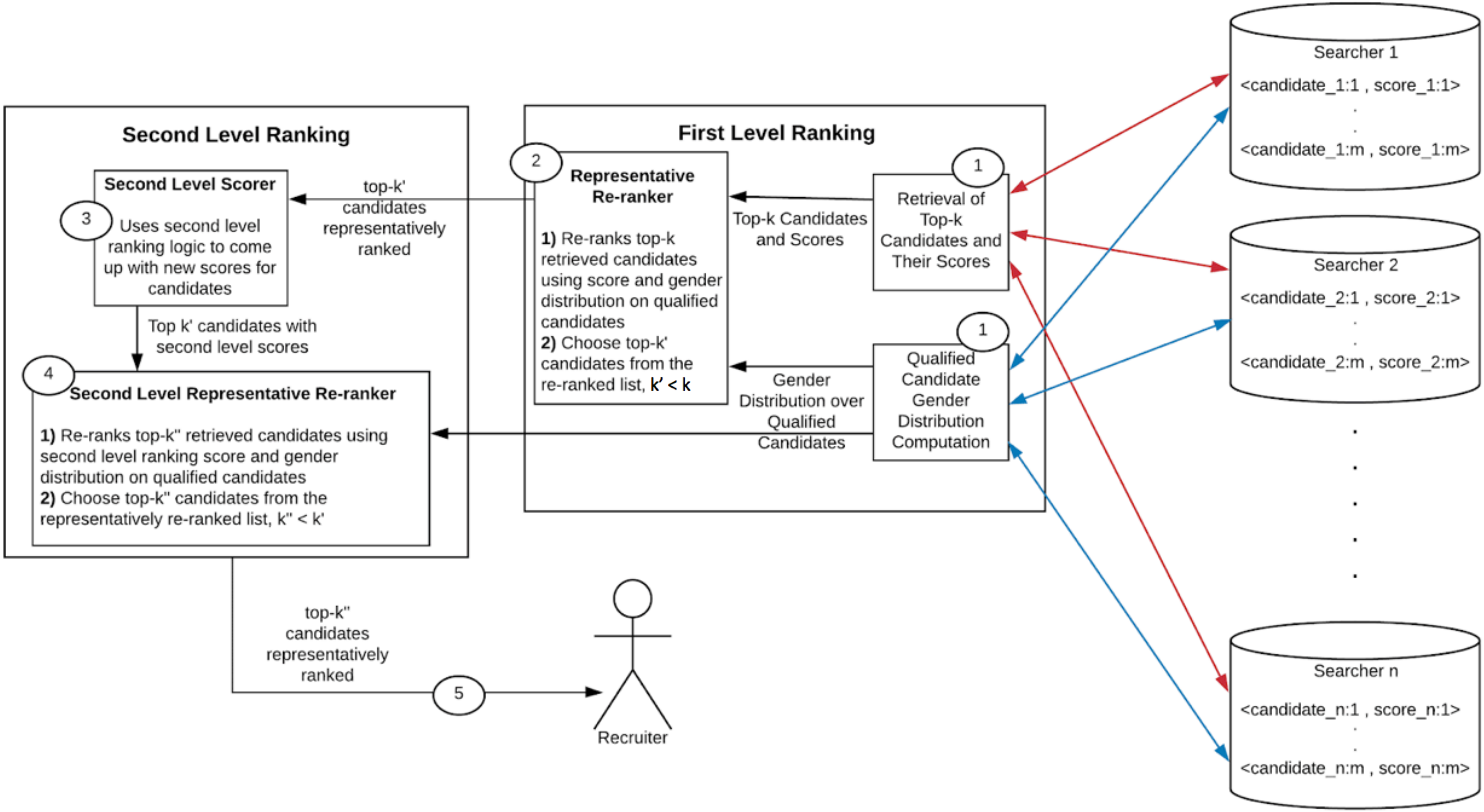
Target Proportions within the Attribute of Interest

- Compute the proportions of the values of the attribute (e.g., gender, gender-age combination) amongst the set of qualified candidates
  - "Qualified candidates" = Set of candidates that match the search query criteria
    - Retrieved by LinkedIn's Galene search engine

- Target proportions could also be obtained based on legal mandate / voluntary commitment

Fairness-aware Reranking Algorithm

- Partition the set of potential candidates into different buckets for each attribute value
- Rank the candidates in each bucket according to the scores assigned by the machine-learned model
- Merge the ranked lists, balancing the representation requirements and the selection of highest scored candidates



Second Level Ranking
Second Level Scorer
3
Uses second level ranking logic to come up with new scores for candidates
Top k' candidates with second level scores
4
Second Level Representative Re-ranker
1) Re-ranks top-k'' retrieved candidates using second level ranking score and gender distribution on qualified candidates
2) Choose top-k'' candidates from the representatively re-ranked list, k'' < k'
top-k' candidates representatively ranked
First Level Ranking
2
Representative Re-ranker
1) Re-ranks top-k retrieved candidates using score and gender distribution on qualified candidates
2) Choose top-k' candidates from the re-ranked list, k' < k
Top-k Candidates and Scores
1
Retrieval of Top-k Candidates and Their Scores
Gender Distribution over Qualified Candidates
1
Qualified Candidate Gender Distribution Computation
Searcher 1
<candidate_1:1 , score_1:1>
<candidate_1:m , score_1:m>
Searcher 2
<candidate_2:1 , score_2:1>
<candidate_2:m , score_2:m>
Searcher n
<candidate_n:1 , score_n:1>
<candidate_n:m , score_n:m>
top-k'' candidates representatively ranked
5
Recruiter

Validating Our Approach

- Gender Representativeness
  - Over 95% of all searches are representative compared to the qualified population of the search
- Business Metrics
  - A/B test over LinkedIn Recruiter users for two weeks
  - No significant change in business metrics (e.g., # InMails sent or accepted)
- Ramped to 100% of LinkedIn Recruiter users worldwide

Lessons Learned in Practice

- Collaboration/consensus across key stakeholders
  - product, legal, PR, engineering, AI, …
- Post-processing approach desirable
  - Agnostic to the specifics of each model
    - Scalable across different model choices for our application
  - Easier to incorporate as part of existing systems
    - Build a stand-alone service or component for post-processing
    - No significant modifications to the existing components

Acknowledgements

- Team:
  - AI/ML: Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi
  - Application Engineering: Gurwinder Gulati, Chenhui Zhai
  - Analytics: Patrick Driscoll, Divyakumar Menghani
  - Product: Rachel Kumar

- Acknowledgements (in alphabetical order)
  - Deepak Agarwal, Erik Buchanan, Patrick Cheung, Gil Cottle, Nadia Fawaz, Rob Hallman, Joshua Hartman, Sara Harrington, Heloise Logan, Stephen Lynch, Lei Ni, Igor Perisic, Ram Swaminathan, Ketan Thakkar, Janardhanan Vembunarayanan, Hinkmond Wong, Lin Yang, Liang Zhang, Yani Zhang

Reflections

- Lessons from privacy & fairness challenges ➔ Need "Privacy and Fairness by Design" approach when building AI products

- Case studies on privacy & fairness @ LinkedIn
  - Collaboration/consensus across key stakeholders (product, legal, PR, engineering, AI, …)

Thanks! Questions?

- References
  - K. Kenthapadi, I. Mironov, A. G. Thakurta, Privacy-preserving Data Mining in Industry: Practical Challenges and Lessons Learned, ACM KDD 2018 Tutorial (Slide deck)
  - K. Kenthapadi, T. T. L. Tran, PriPeARL: A Framework for Privacy-Preserving Analytics and Reporting at LinkedIn, ACM CIKM 2018
  - K. Kenthapadi, A. Chudhary, S. Ambler, LinkedIn Salary: A System for Secure Collection and Presentation of Structured Compensation Insights to Job Seekers, IEEE Symposium on Privacy-Aware Computing (PAC), 2017
  - S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness Aware Talent Search Ranking at LinkedIn, Microsoft's AI/ML conference (MLADS Spring 2018). **Distinguished Contribution Award**
  - S. C. Geyik, K. Kenthapadi, Building Representative Talent Search at LinkedIn, LinkedIn engineering blog post, October 2018

kkenthapadi@linkedin.com

Backup

Privacy: A Historical Perspective

Privacy Breaches and Lessons Learned

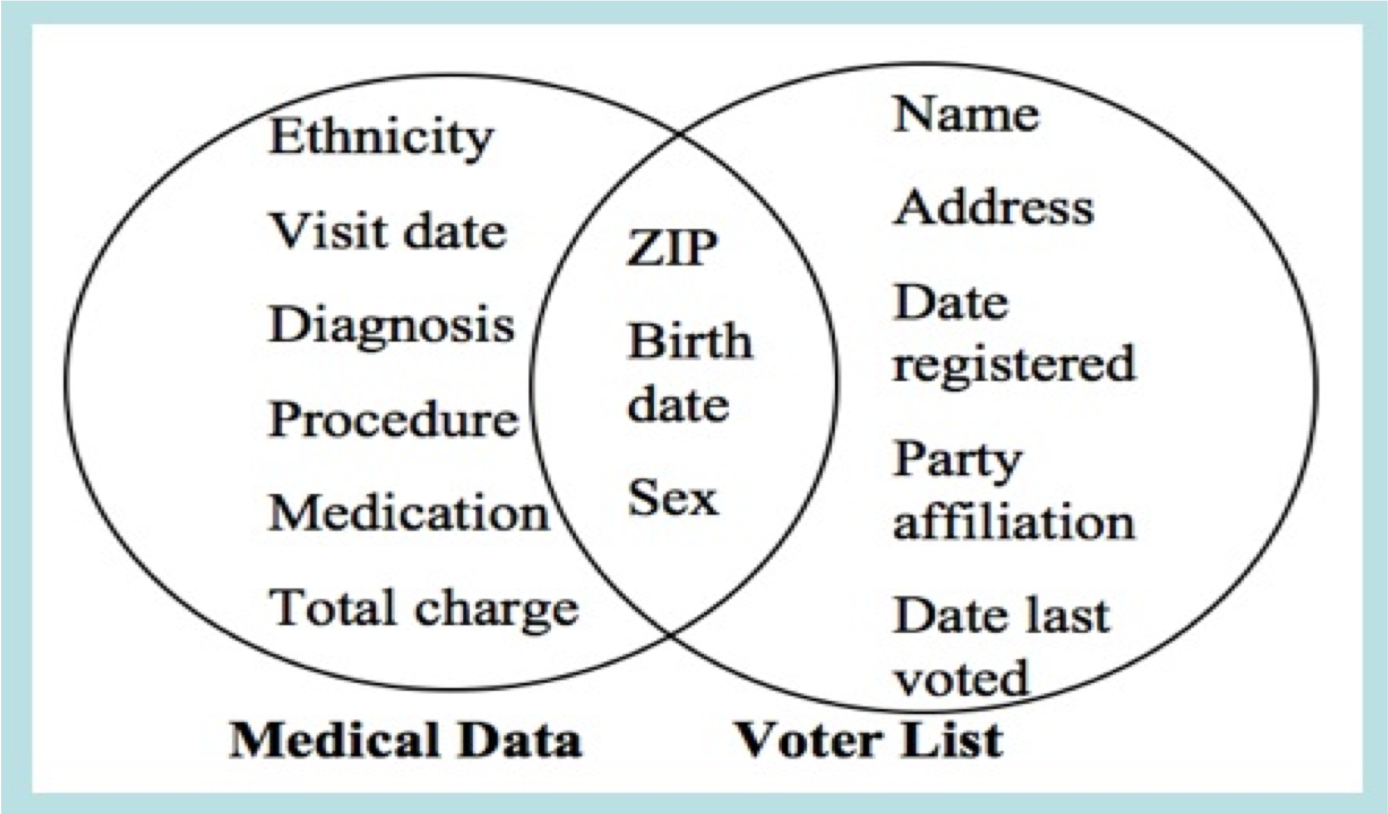
Attacks on privacy

- Governor of Massachusetts
- AOL
- Netflix
- Web browsing data
- Facebook
- Amazon
- Genomic data

William Weld vs Latanya Sweeney

Massachusetts Group Insurance Commission (1997):
    Anonymized medical history of state employees (all hospital visits, diagnosis, prescriptions)

Latanya Sweeney (MIT grad student): $20 – Cambridge voter roll



born July 31, 1945
resident of 02138

Attacker's Advantage

- Auxiliary information

AOL Data Release

August 4, 2006: AOL Research publishes anonymized search logs of 650,000 users

August 9:
New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

Attacker's Advantage

- Auxiliary information
- Enough to succeed on a small fraction of inputs

De-anonymizing Web Browsing Data with Social Networks

Key idea:

- Similar intuition as the attack on medical records
- Medical records: Each person can be identified based on a combination of a few attributes
- Web browsing history: Browsing history is unique for each person
- Each person has a distinctive social network ➔ links appearing in one's feed is unique
- Users likely to visit links in their feed with higher probability than a random user

    - "Browsing histories contain tell-tale marks of identity"

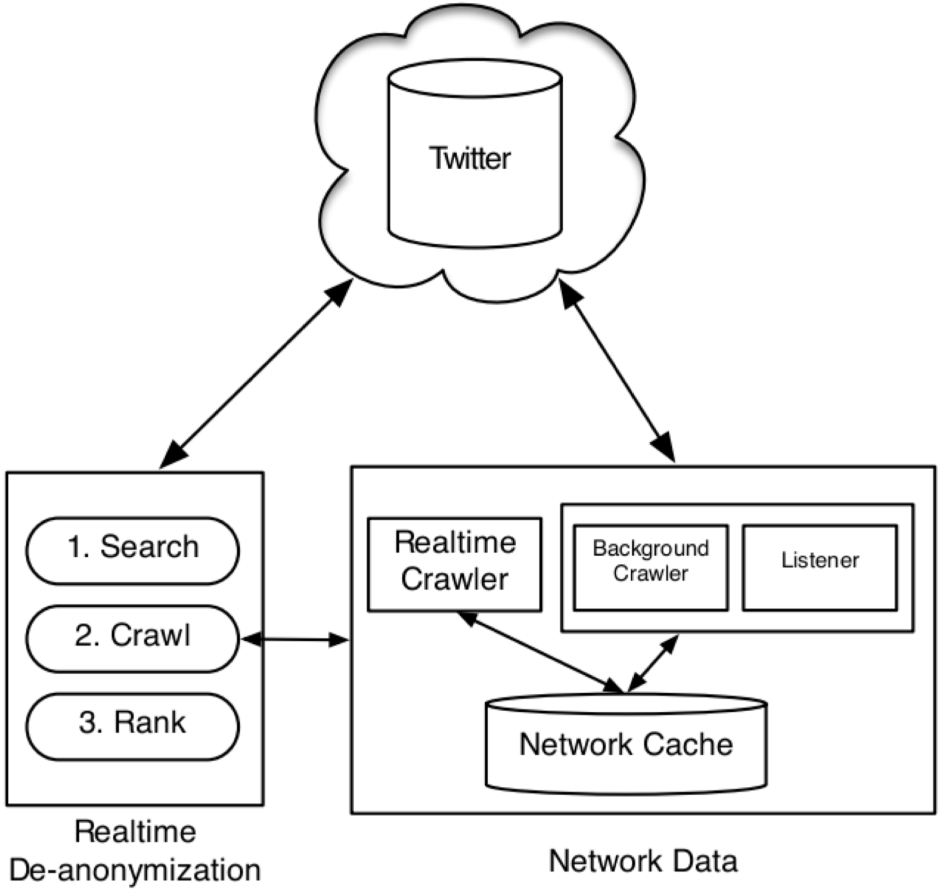


Figure 2: System architecture for real-time de-anonymization of web browsing histories.

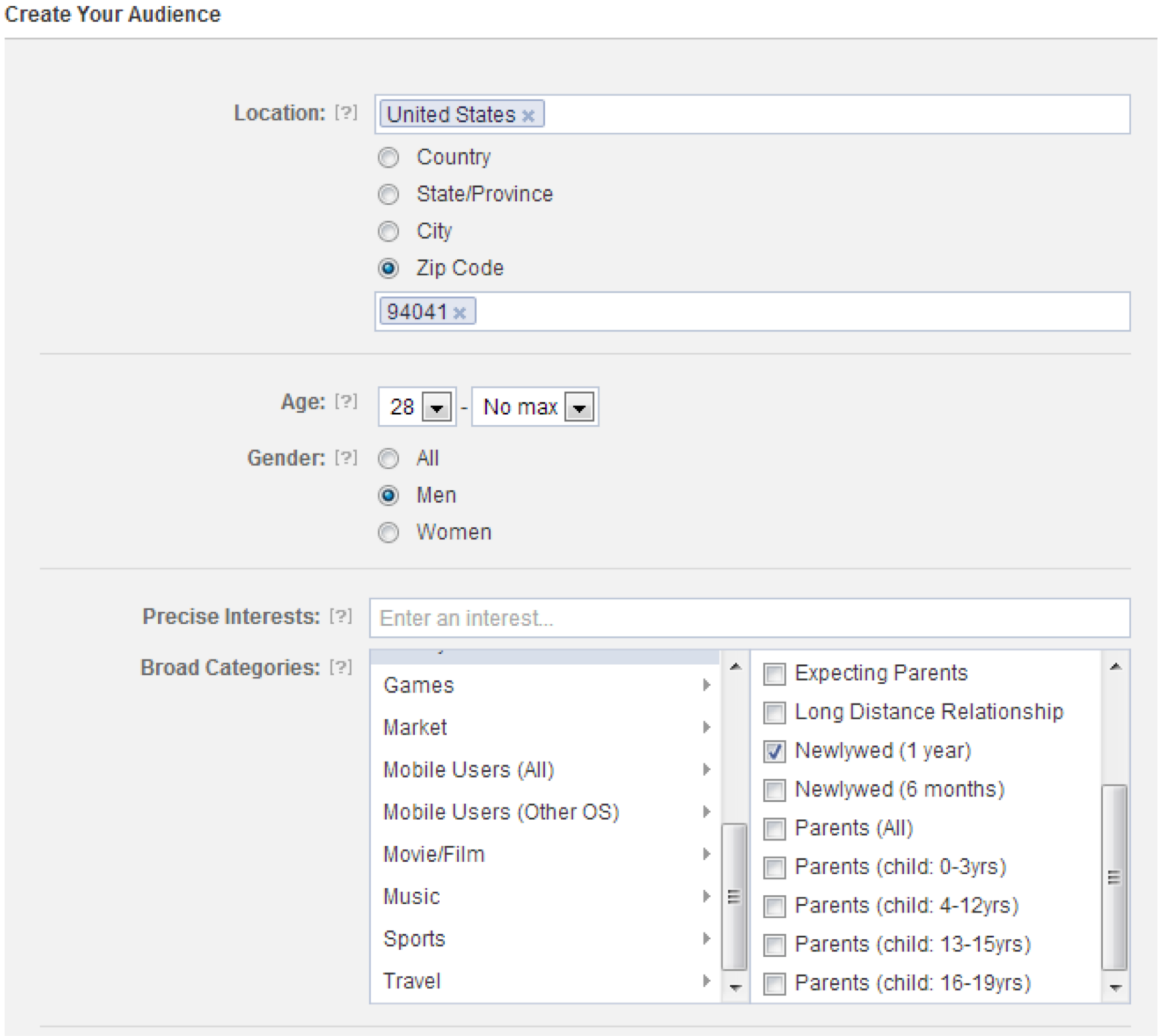
Su et al, [De-anonymizing Web Browsing Data with Social Networks](), 2017

Attacker's Advantage

- Auxiliary information
- Enough to succeed on a small fraction of inputs
- High dimensionality

Privacy Attacks On Ad Targeting

Ad targeting:



Korolova, "Privacy Violations Using Microtargeted Ads: A Case Study", PADM 2010

Facebook vs Korolova

10 campaigns targeting 1 person (zip code, gender, workplace, alma mater)

Age	Ad Impressions in a week
21	0
22	0
23	8
…	…
30	0

Korolova, "Privacy Violations Using Microtargeted Ads: A Case Study", PADM 2010

Facebook vs Korolova

10 campaigns targeting 1 person (zip code, gender, workplace, alma mater)

Interest	Ad Impressions in a week
A	0
B	0
C	8
…	…
Z	0

Korolova, “Privacy Violations Using Microtargeted Ads: A Case Study”, PADM 2010

Facebook vs Korolova: Recap

- Context: Microtargeted Ads
- Takeaway: Attackers can instrument ad campaigns to identify individual users.
- Two types of attacks:
  - Inference from Impressions
  - Inference from Clicks

Privacy Violations Using Microtargeted Ads: A Case Study

Aleksandra Korolova*

Abstract. We propose a new class of attacks that breach user privacy by exploiting advertising systems offering microtargeting capabilities. We study the advertising system of the largest online social network, Facebook, and the risks that the design of the system poses to the privacy of its users. We propose, describe, and provide experimental evidence of several novel approaches to exploiting the advertising system in order to obtain private user information.

The work illustrates how a real-world system designed with an intention to protect privacy but without rigorous privacy guarantees can leak private information, and motivates the need for further research on the design of microtargeted advertising systems with provable privacy guarantees. Furthermore, it shows that user privacy may be breached not only as a result of data publishing using improper anonymization techniques, but also as a result of internal data-mining of that data.

We communicated our findings to Facebook on July 13, 2010, and received a very prompt response. On July 20, 2010, Facebook launched a change to their advertising system that made the kind of attacks we describe much more difficult to implement in practice, even though, as we discuss, they remain possible in principle. We conclude by discussing the broader challenge of designing privacy-preserving microtargeted advertising systems.

Keywords: Facebook, social networks, targeted advertising, privacy breaches

Attacker's Advantage

- Auxiliary information
- Enough to succeed on a small fraction of inputs
- High dimensionality
- Active

Attacking Amazon.com

Items frequently bought together

Bought: A B C D E

Z: **Customers Who Bought This Item Also Bought**

A C D E

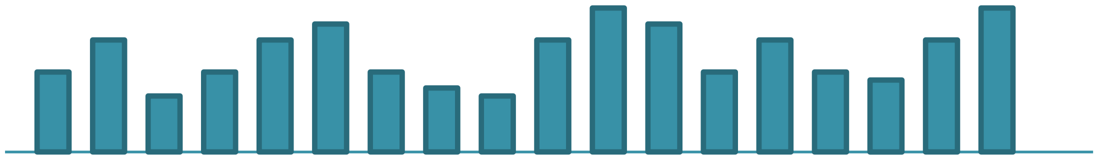
Calandrino, Kilzer, Narayanan, Felten, Shmatikov, “You Might Also Like: Privacy Risks of Collaborative Filtering” (2009)

Attacker's Advantage

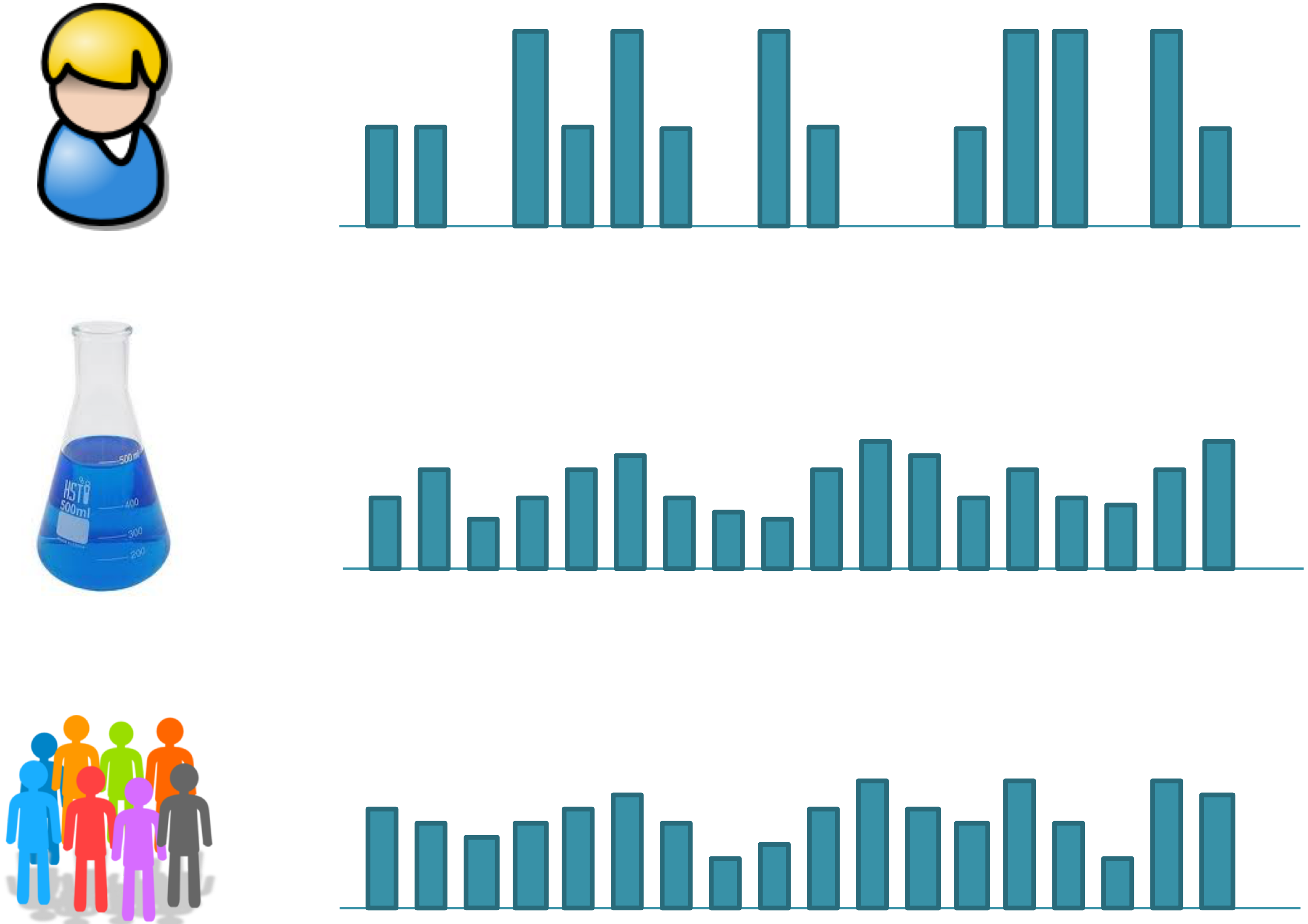
- Auxiliary information
- Enough to succeed on a small fraction of inputs
- High dimensionality
- Active
- Observant

Genetic data

Homer et al., “Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays”, PLoS Genetics, 2008



Bayesian Analysis



Reference population

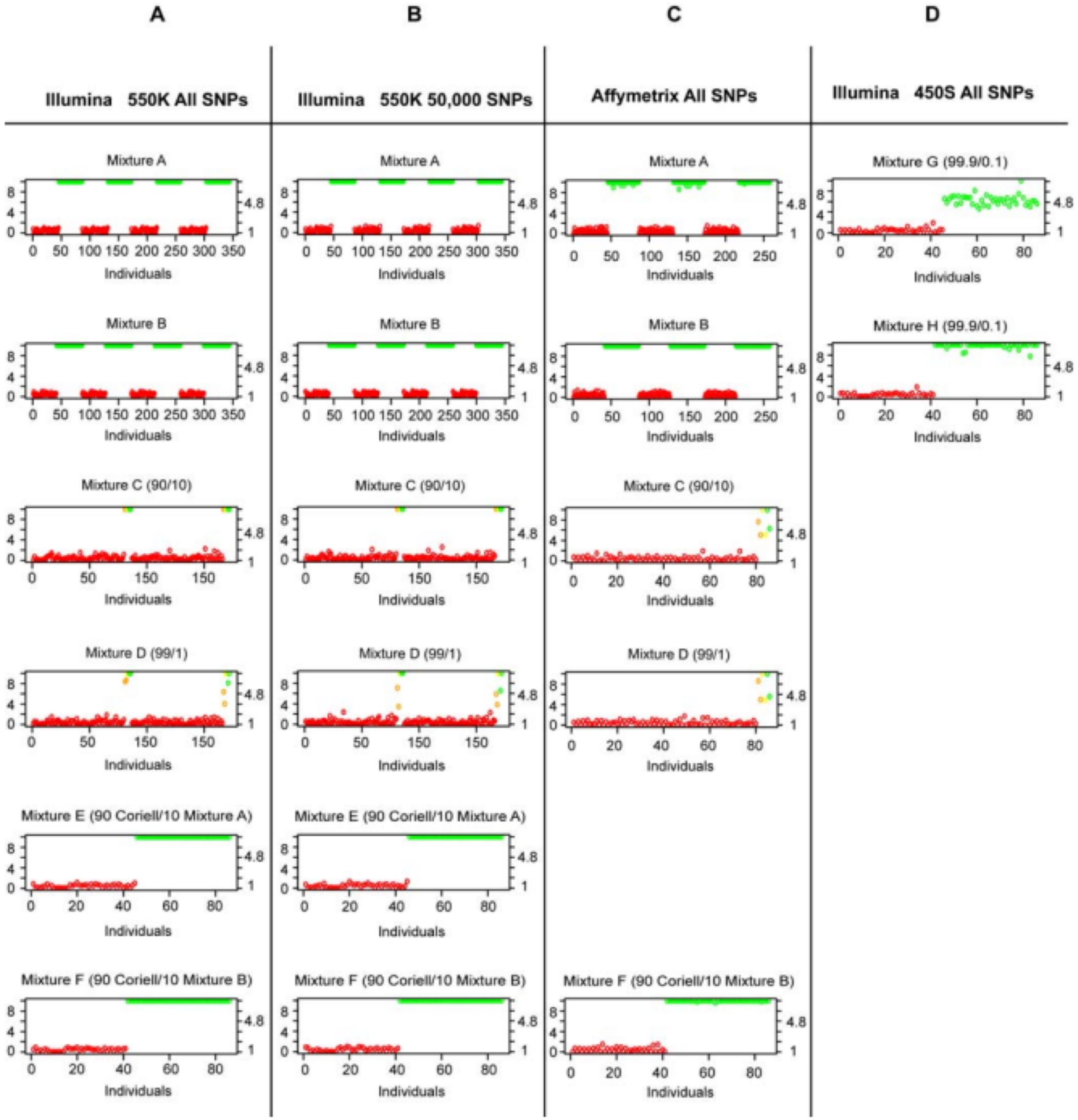
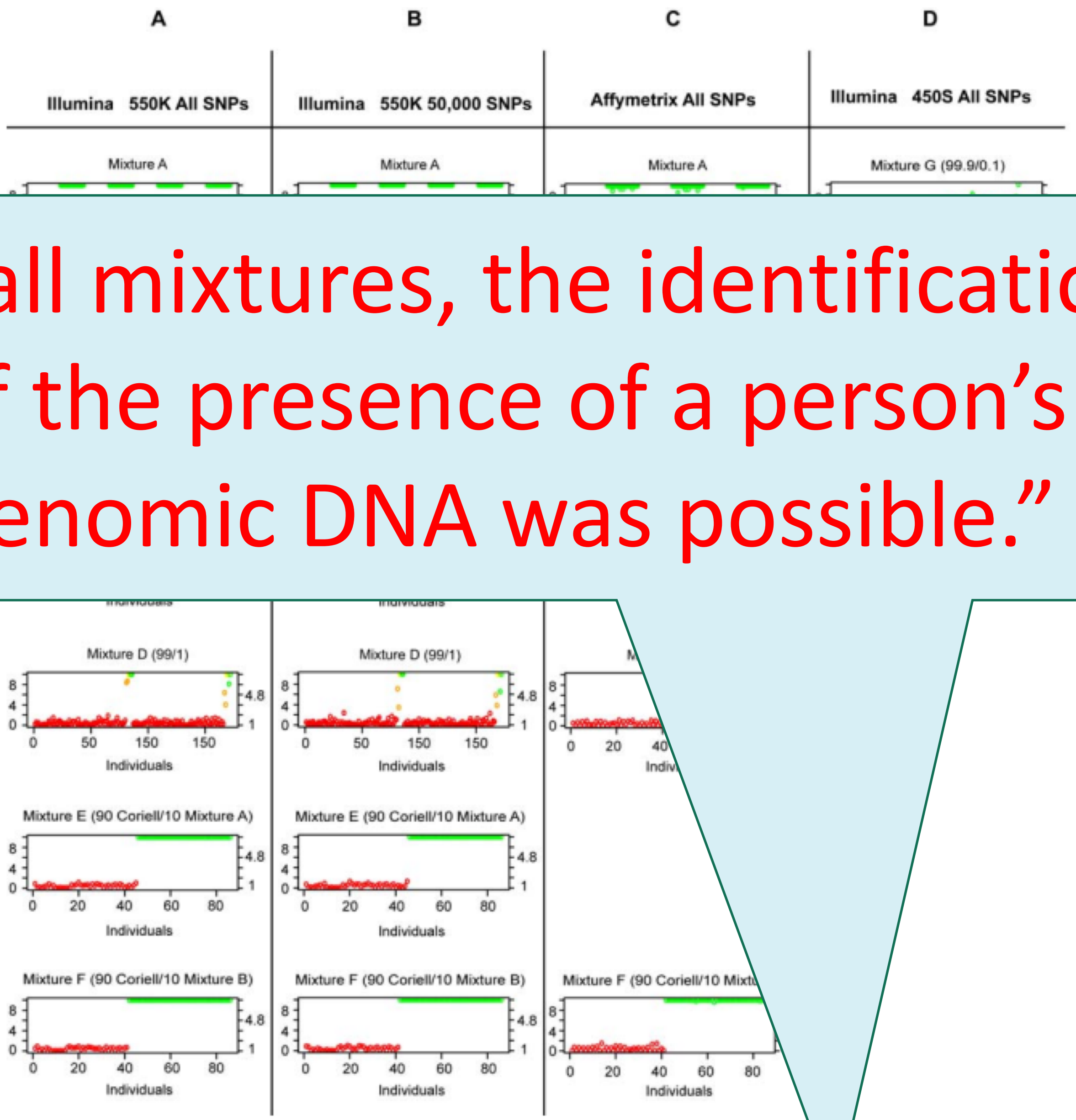


Figure 3. Experimental validation using a series of mixtures (see Methods A–F) assayed on the Affymetrix GeneChip 5.0, Illumina BeadArray 550 and the Illumina 450S Duo Human BeadChip. The x-axis shows each individual in the CEU HapMap population, the left y-axis shows the p-value (log scaled), and the right y-axis shows the value of the test statistic. For mixtures A, B, E and F those in the mixture are colored green and those not in the mixture are colored red. For mixtures C and D those individuals who are not in the mixtures are colored red, those individuals who are related to the 1% or 10% individuals in the mixtures are colored orange, those individuals who are related to the 90% or 99% are colored yellow, and those people in the mixture are colored green. In all mixtures, the identification of the presence of a person's genomic DNA was possible.

doi:10.1371/journal.pgen.1000167.g003



"In all mixtures, the identification of the presence of a person's genomic DNA was possible."

Figure 3. Experimental validation using a series of mixtures (see Methods A–F) assayed on the Affymetrix GeneChip 5.0, Illumina BeadArray 550 and the Illumina 450S Duo Human BeadChip. The x-axis shows each individual in the CEU HapMap population, the left y-axis shows the p-value (log scaled), and the right y-axis shows the value of the test statistic. For mixtures A, B, E and F those in the mixture are colored green and those not in the mixture are colored red. For mixtures C and D those individuals who are not in the mixtures are colored red, those individuals who are related to the 1% or 10% individuals in the mixtures are colored orange, those individuals who are related to the 90% or 99% are colored yellow, and those people in the mixture are colored green. In all mixtures, the identification of the presence of a person's genomic DNA was possible.

doi:10.1371/journal.pgen.1000167.g003

… one week later

Zerhouni, NIH Director:

“As a result, the NIH has removed from open-access databases the aggregate results (including P values and genotype counts) for all the GWAS that had been available on NIH sites”

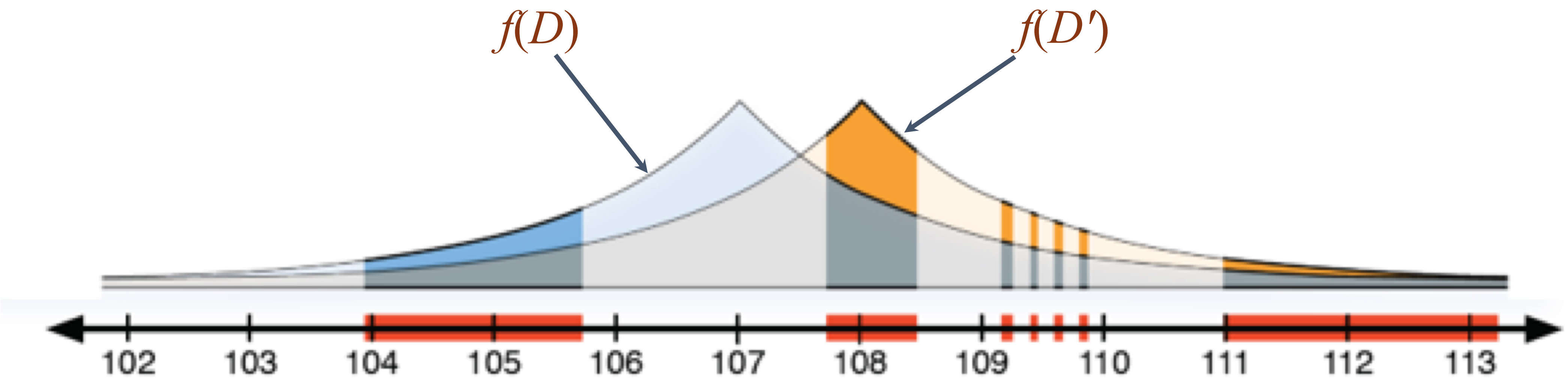
Attacker's Advantage

- Auxiliary information
- Enough to succeed on a small fraction of inputs
- High dimensionality
- Active
- Observant
- Clever

Differential Privacy: Takeaway points

- Privacy as a notion of stability of randomized algorithms in respect to small perturbations in their input
  - Worst-case definition
  - Robust (to auxiliary data, correlated inputs)
  - Composable
  - Quantifiable
- Concept of a privacy loss budget
- Noise injection

“Bad Outcomes” Interpretation



— bad outcomes

— probability with record x

— probability without record x

Bayesian Interpretation

- Prior on databases p
- Observed output O
- Does the database contain record x?

$$\frac{p(D|O)}{p(D'|O)} = \frac{p(D)}{p(D')} \frac{p(O|D)}{p(O|D')}$$

$$\exp(-\varepsilon) \le \frac{p(O|D)}{p(O|D')} \le \exp(\varepsilon)$$

Differential Privacy

- Robustness to auxiliary data

- Post-processing:
  If $M(D)$ is differentially private, so is $f(M(D))$.

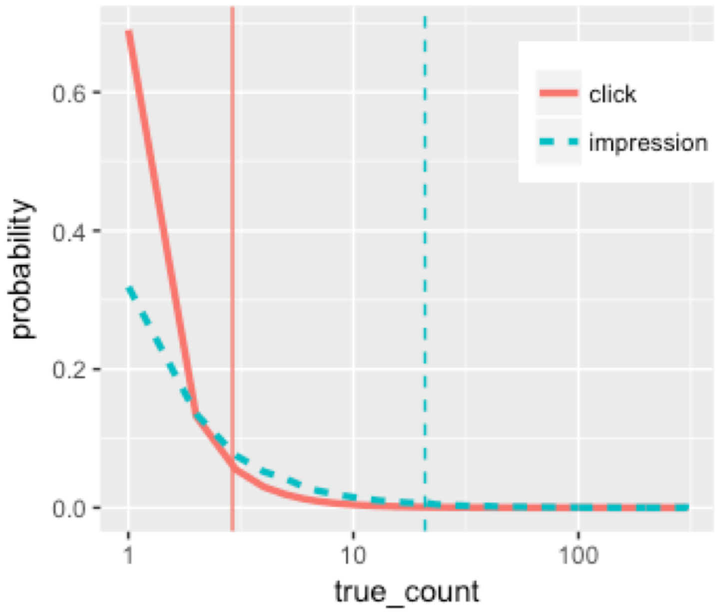
- Composability:
  Run two ε-DP mechanisms. Full interaction is 2ε-DP.

- Group privacy:
  Graceful degradation in the presence of correlated inputs.

Performance Evaluation: Setup

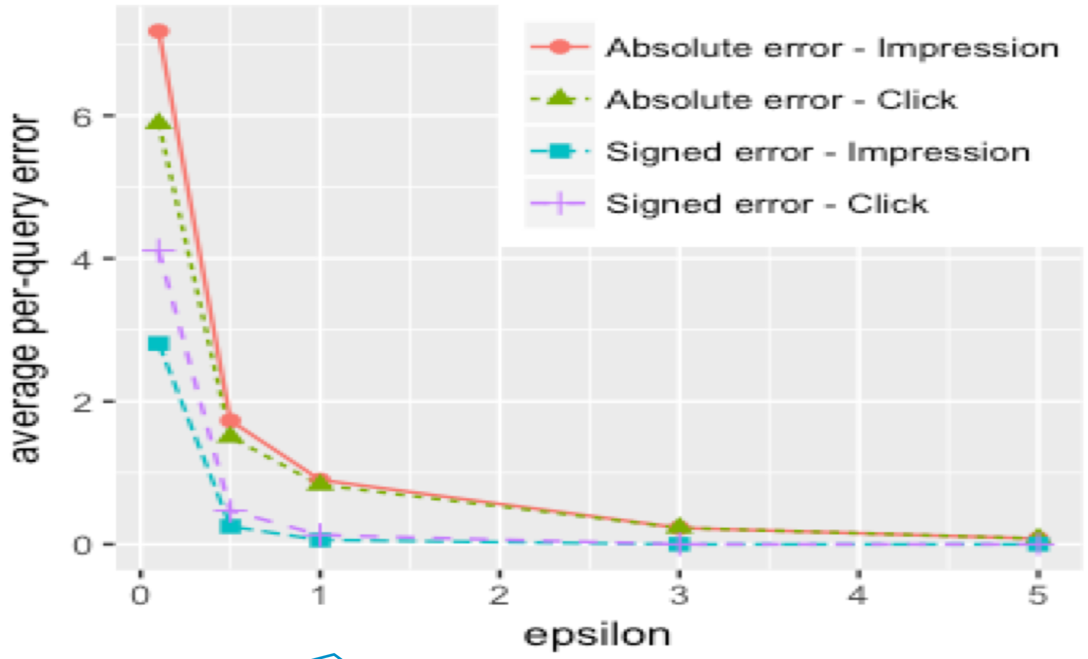
- Experiments using LinkedIn ad analytics data
  - Consider distribution of impression and click queries across (account, ad campaign) and demographic breakdowns.

- Examine
  - Tradeoff between privacy and utility
  - Effect of varying minimum threshold (non-negative)
  - Top-n queries

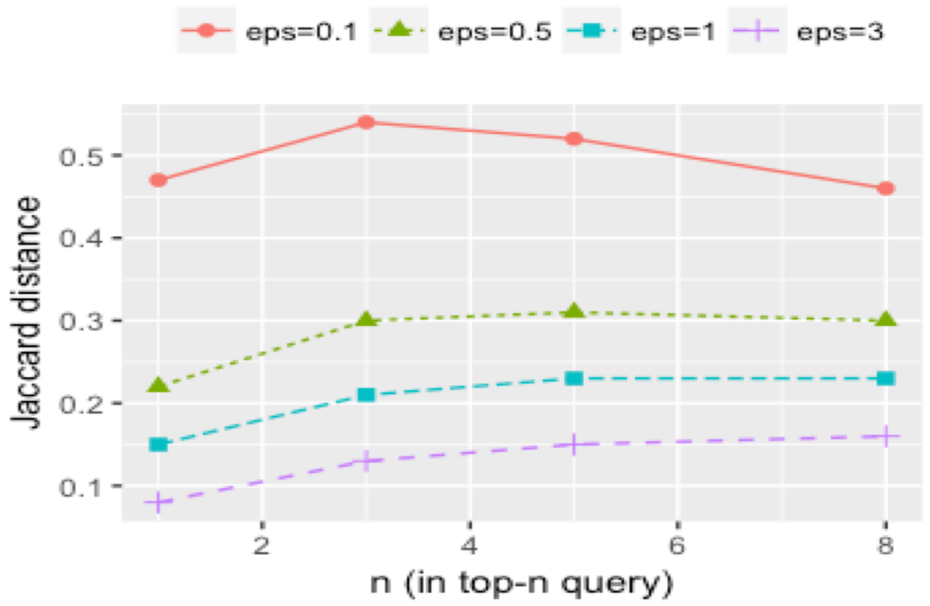


Performance Evaluation: Results

Privacy and Utility Tradeoff



Top-N Queries



For $\varepsilon = 1$, average absolute and signed errors are small for both queries (impression and click)
Variance is also small, ~95% of queries have error of at most 2.

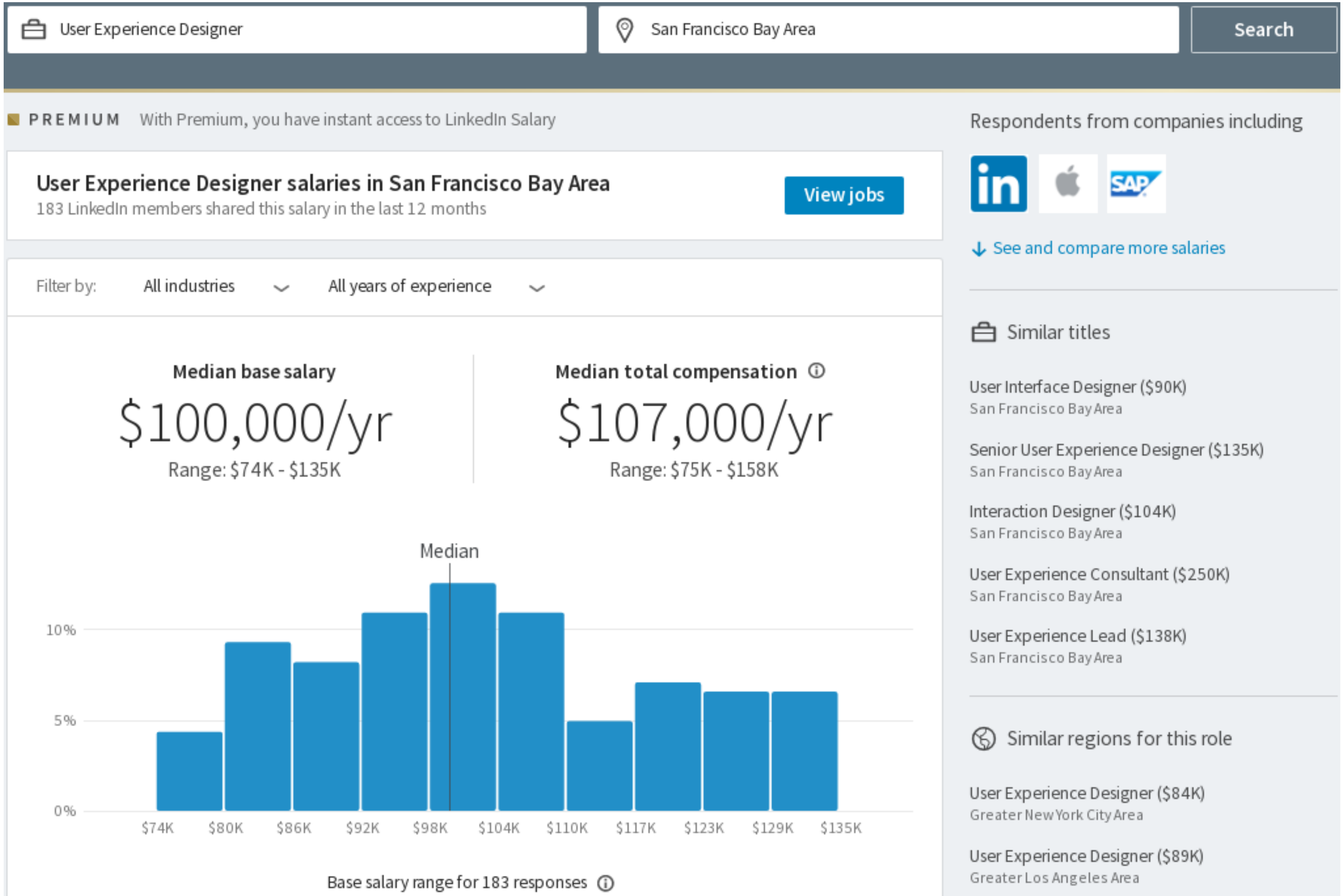
Common use case in LinkedIn application.
Jaccard distance as a function of ε and n.
This shows the worst case since queries with return sets $\leq n$ and error of 0 were omitted.

LinkedIn Salary

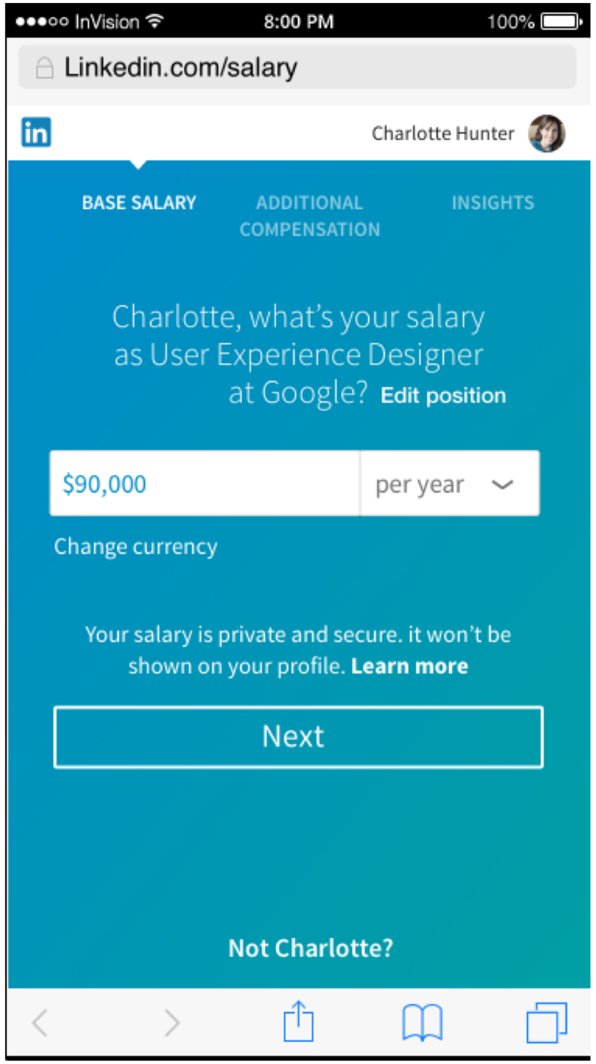
Outline

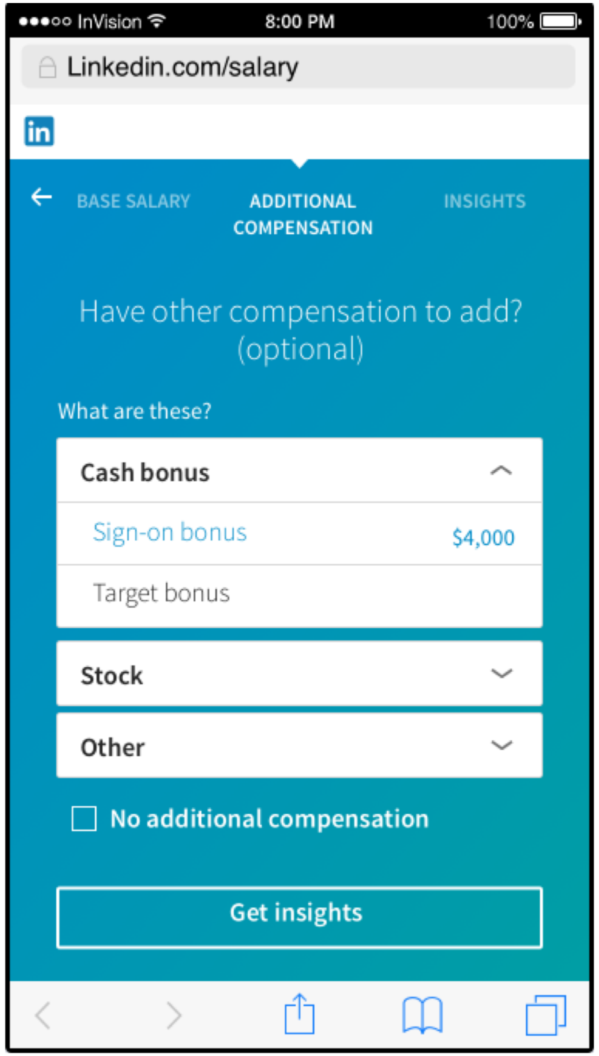
- LinkedIn Salary Overview
- Challenges: Privacy, Modeling
- System Design & Architecture
- Privacy vs. Modeling Tradeoffs

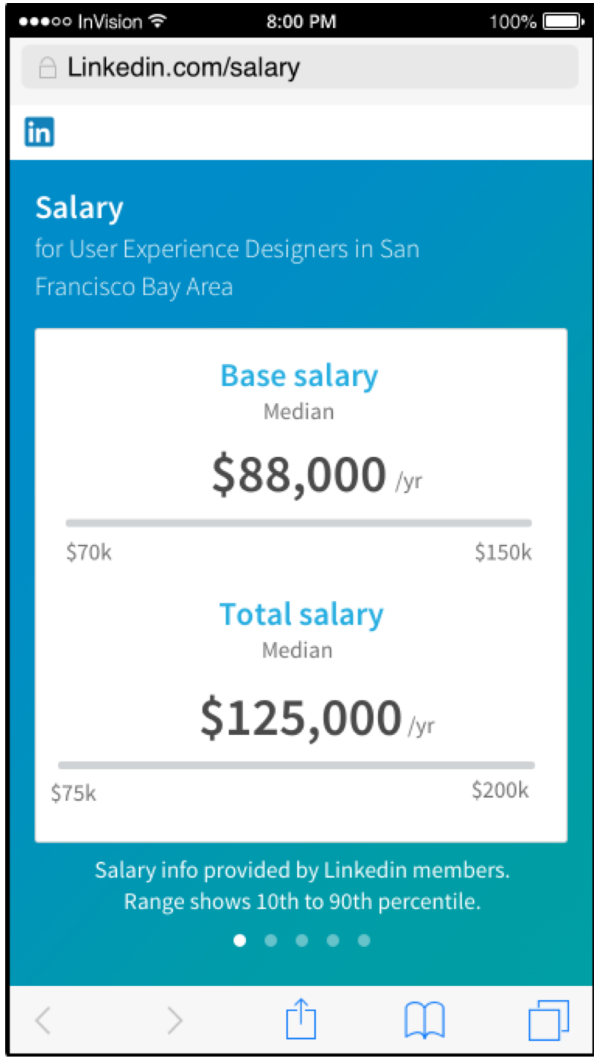
LinkedIn Salary (launched in Nov, 2016)



Salary Collection Flow via Email Targeting







Current Reach (November 2018)

- A few million responses out of several millions of members targeted
  - Targeted via emails since early 2016
- Countries: US, CA, UK, DE, IN, …
- Insights available for a large fraction of US monthly active users

Data Privacy Challenges

- Minimize the risk of inferring any one individual's compensation data

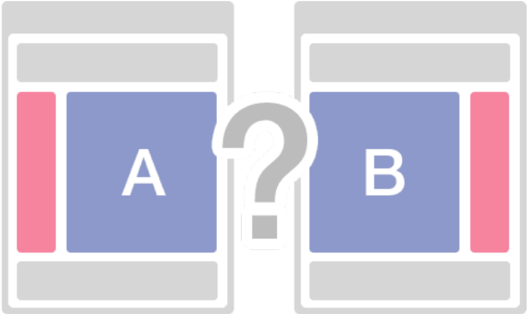
- Protection against data breach
  - No single point of failure

Achieved by a combination of techniques: encryption, access control, de-identification, aggregation, thresholding

K. Kenthapadi, A. Chudhary, and S. Ambler, [LinkedIn Salary: A System for Secure Collection and Presentation of Structured Compensation Insights to Job Seekers](), IEEE PAC 2017 (arxiv.org/abs/1705.06976)

Modeling Challenges

- Evaluation



- Modeling on de-identified data

- Robustness and stability

- Outlier detection

K. Kenthapadi, S. Ambler, L. Zhang, and D. Agarwal, Bringing salary transparency to the world: Computing robust compensation insights via LinkedIn Salary, CIKM 2017 (arxiv.org/abs/1703.09845)

X. Chen, Y. Liu, L. Zhang, and K. Kenthapadi, How LinkedIn Economic Graph Bonds Information and Product: Applications in LinkedIn Salary, KDD 2018 (arxiv.org/abs/1806.09063)

Problem Statement

- *How do we design LinkedIn Salary system taking into account the unique privacy and security challenges, while addressing the product requirements?*

Open Question

- Can we apply rigorous approaches such as differential privacy in such a setting?
  - While meeting reliability / product coverage needs

- Worst case sensitivity of quantiles to any one user's compensation data is large
  - ➔ Large noise may need to be added, depriving reliability/usefulness

- Need compensation insights on a continual basis
  - Theoretical work on applying differential privacy under continual observations
    - No practical implementations / applications

  - Local differential privacy / Randomized response based approaches (Google's RAPPOR; Apple's iOS differential privacy; Microsoft's telemetry collection) don't seem applicable

De-identification Example

Title	Region	Company	Industry	Years of exp	Degree	FoS	Skills	$$
User Exp Designer	SF Bay Area	Google	Internet	12	BS	Interactive Media	UX, Graphics, …	100K

Title	Region	$$
User Exp Designer	SF Bay Area	100K
User Exp Designer	SF Bay Area	115K
…	…	…

Title	Region	Industry	$$
User Exp Designer	SF Bay Area	Internet	100K

Title	Region	Years of exp	$$
User Exp Designer	SF Bay Area	10+	100K

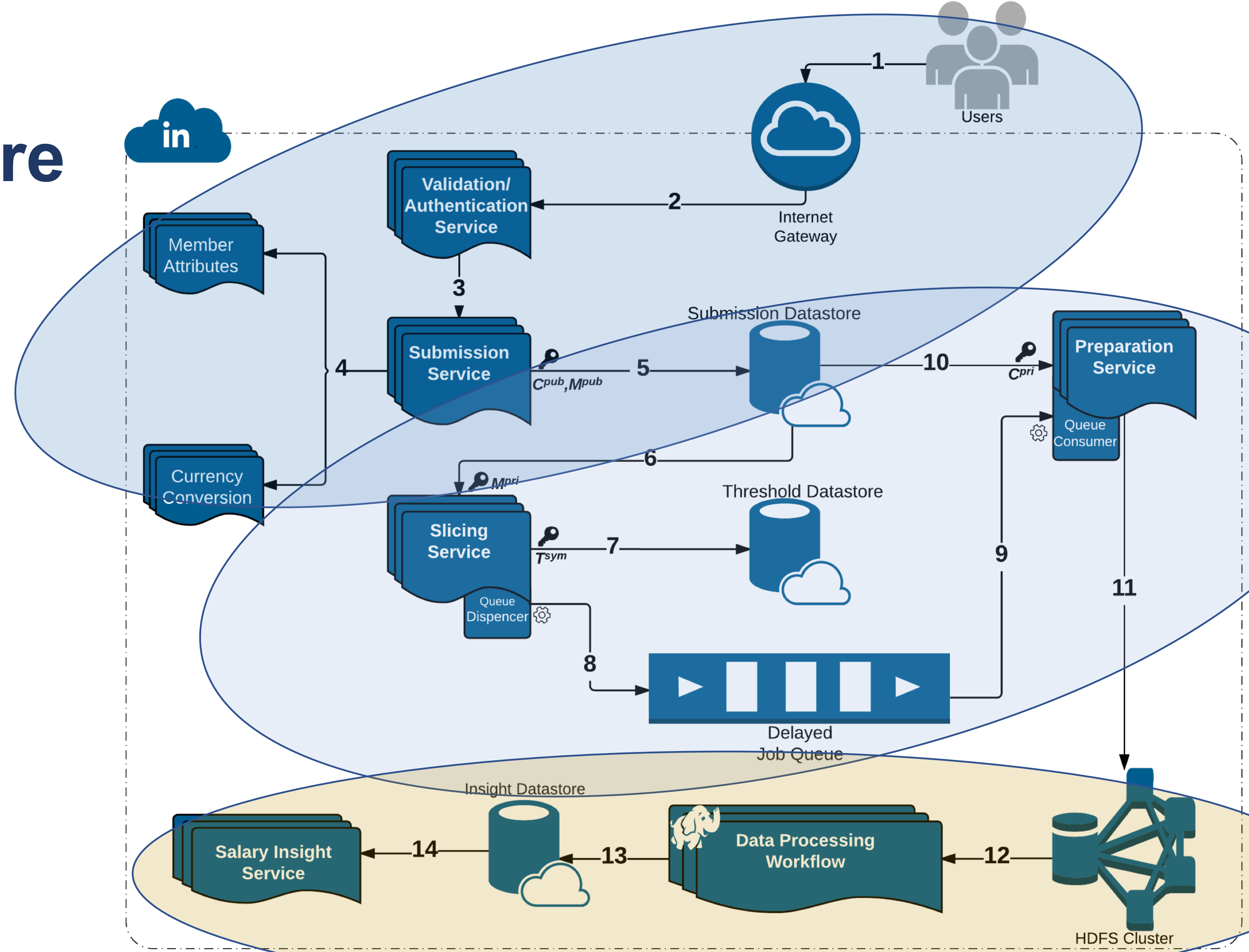
Title	Region	Company	Years of exp	$$
User Exp Designer	SF Bay Area	Google	10+	100K

#data points > threshold?

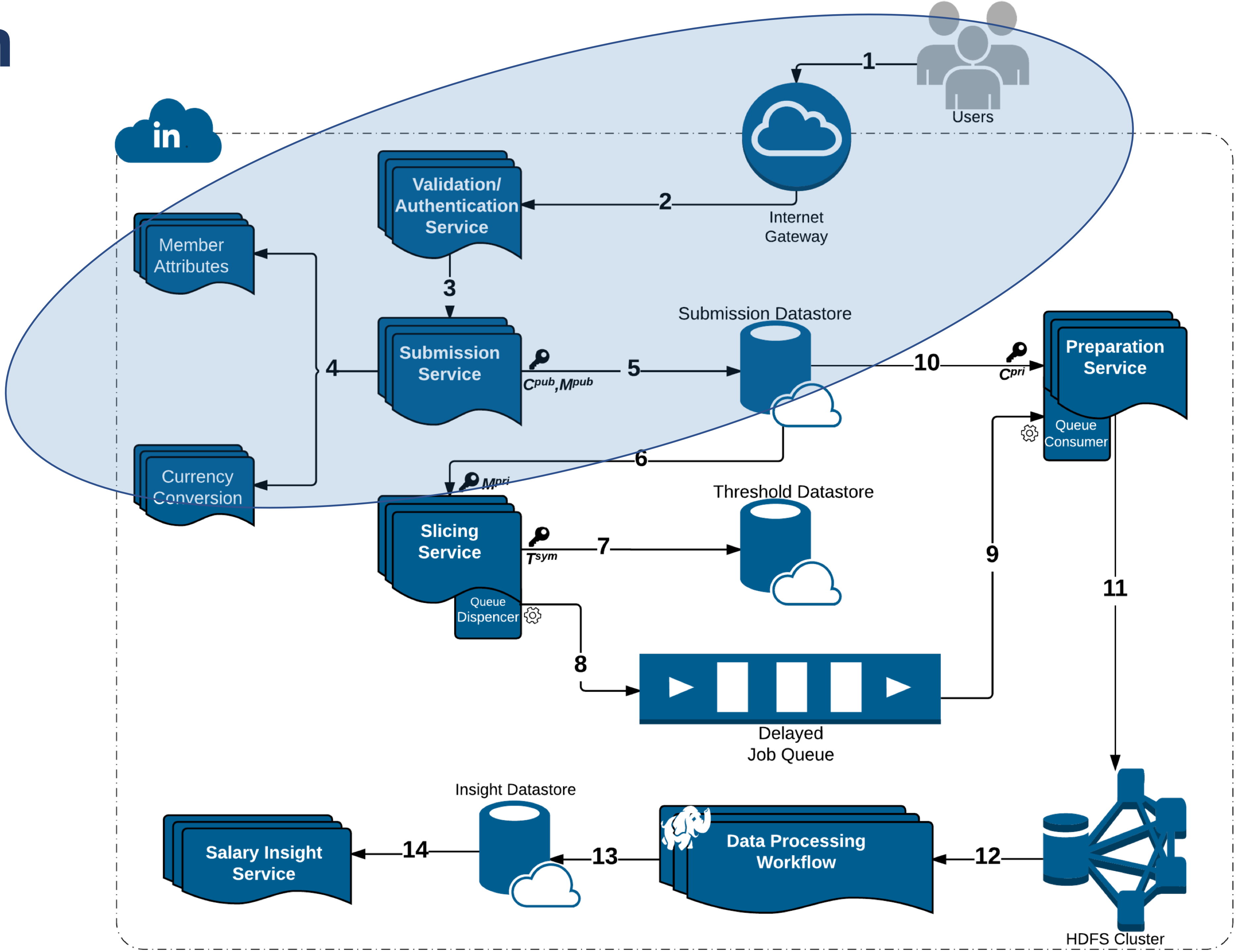
Yes ⇒ Copy to Hadoop (HDFS)

Note: Original submission stored as encrypted objects.

System Architecture

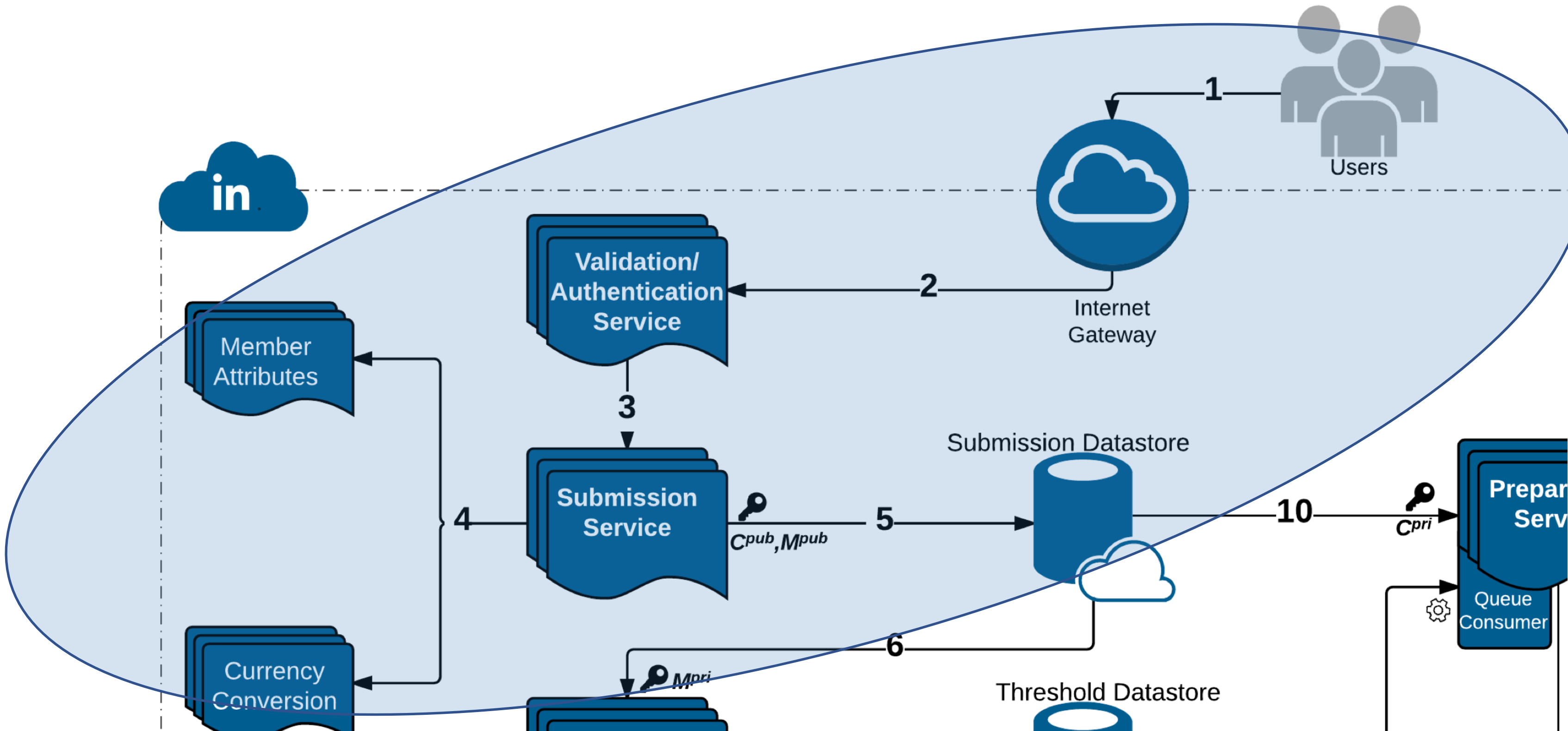


Collection & Storage

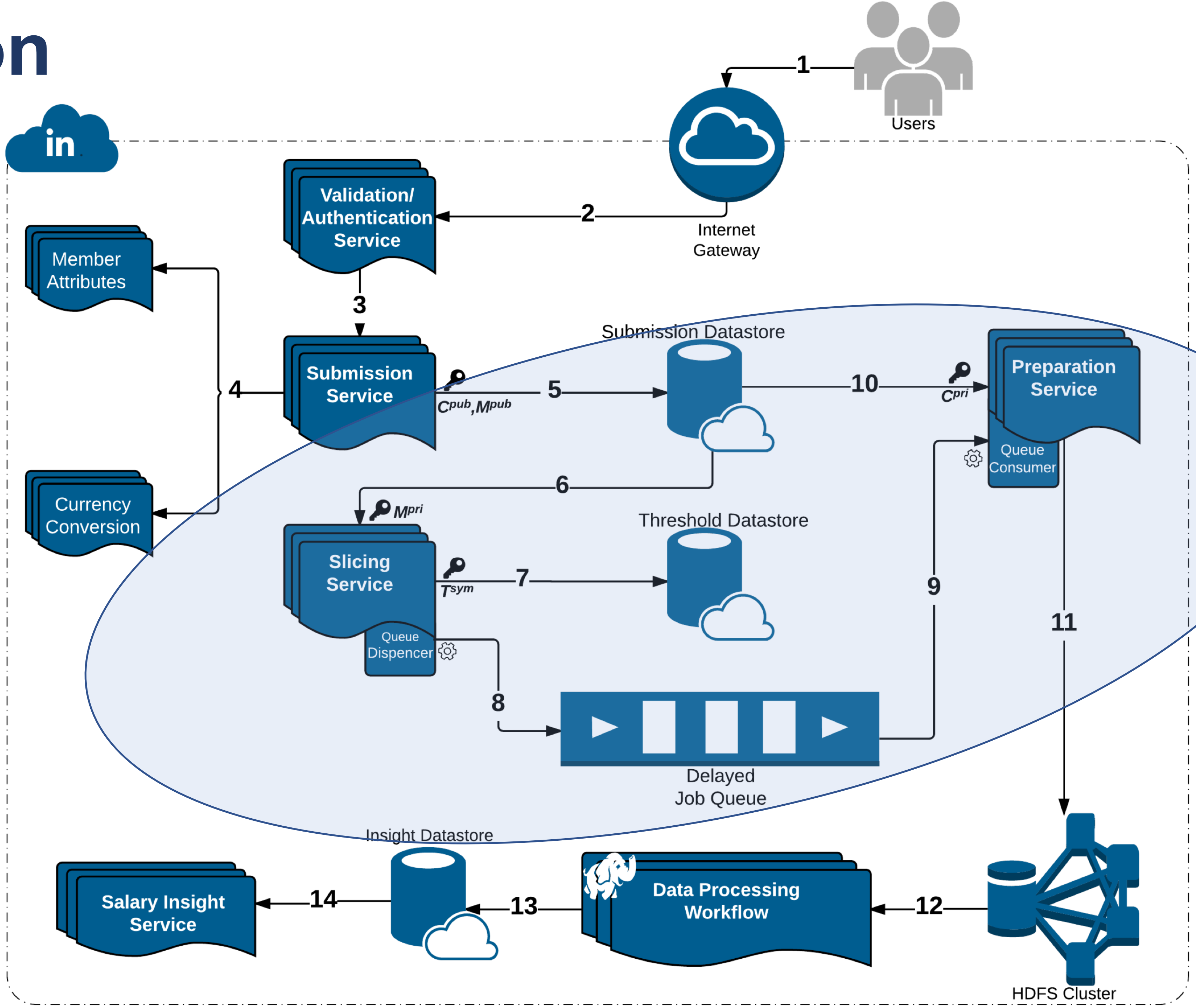


Collection & Storage

- Allow members to submit their compensation info
- Extract member attributes
  - E.g., canonical job title, company, region, by invoking LinkedIn standardization services
- Securely store member attributes & compensation data



De-identification & Grouping



De-identification & Grouping

- Approach inspired by k-Anonymity [Samarati-Sweeney]

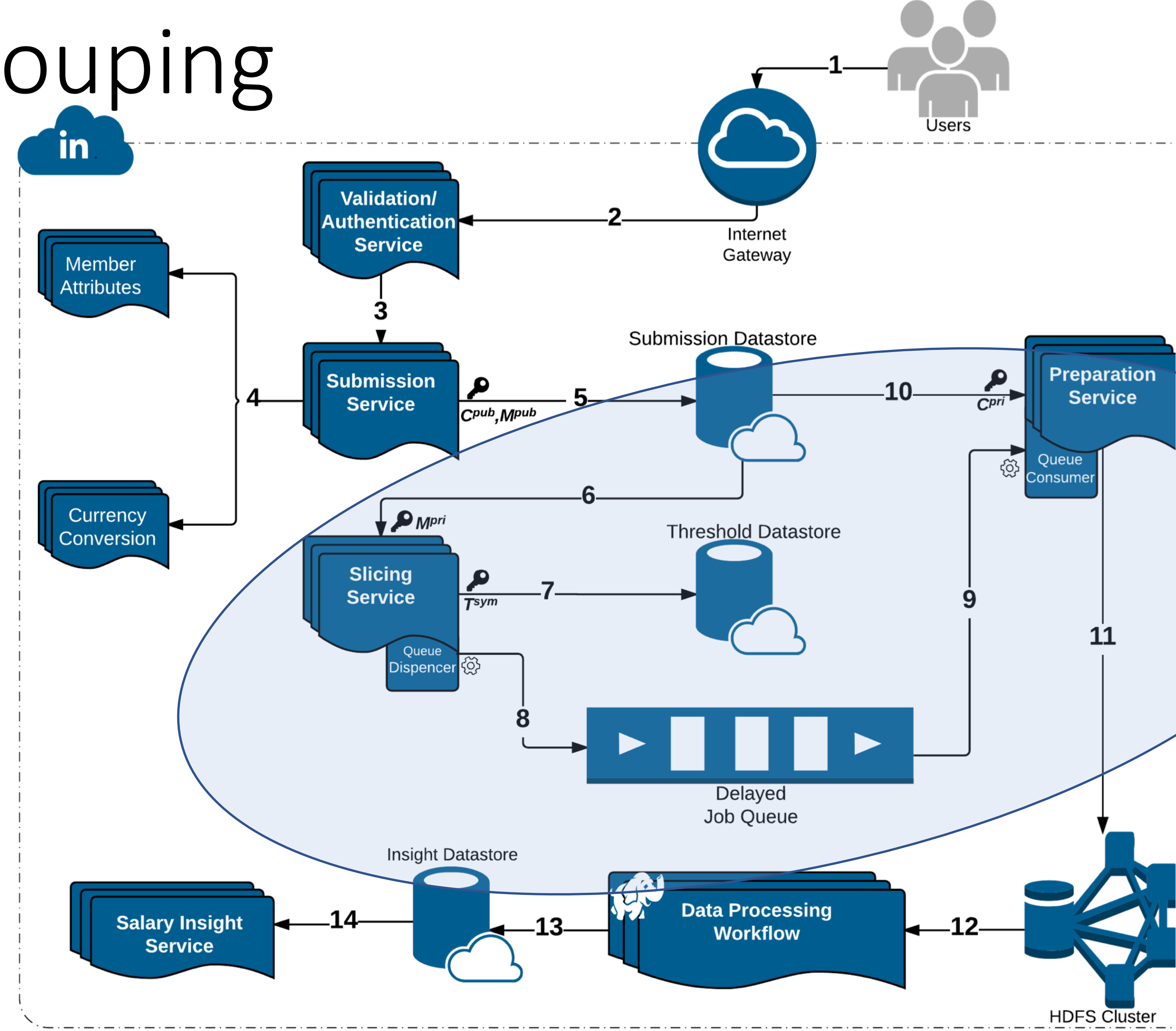
- "Cohort" or "Slice"
  - Defined by a combination of attributes
    - E.g, "User experience designers in SF Bay Area"
  - Contains aggregated compensation entries from corresponding individuals
  - No user name, id or any attributes other than those that define the cohort

  - A cohort available for offline processing only if it has at least k entries

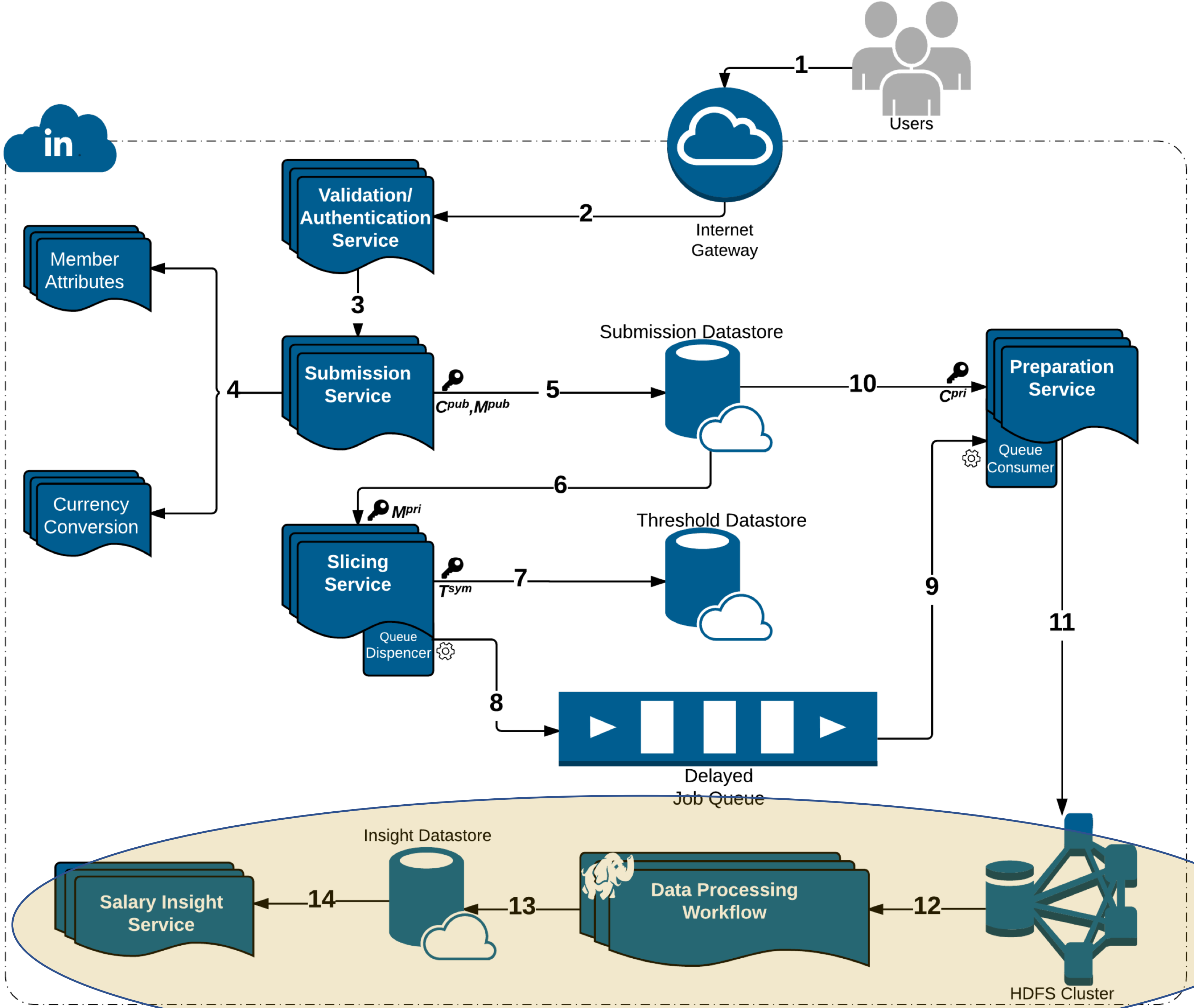
  - Apply LinkedIn standardization software (free-form attribute → canonical version) before grouping
    - Analogous to the generalization step in k-Anonymity

De-identification & Grouping

- Slicing service
  - Access member attribute info & submission identifiers (no compensation data)
  - Generate slices & track # submissions for each slice
- Preparation service
  - Fetch compensation data (using submission identifiers), associate with the slice data, copy to HDFS

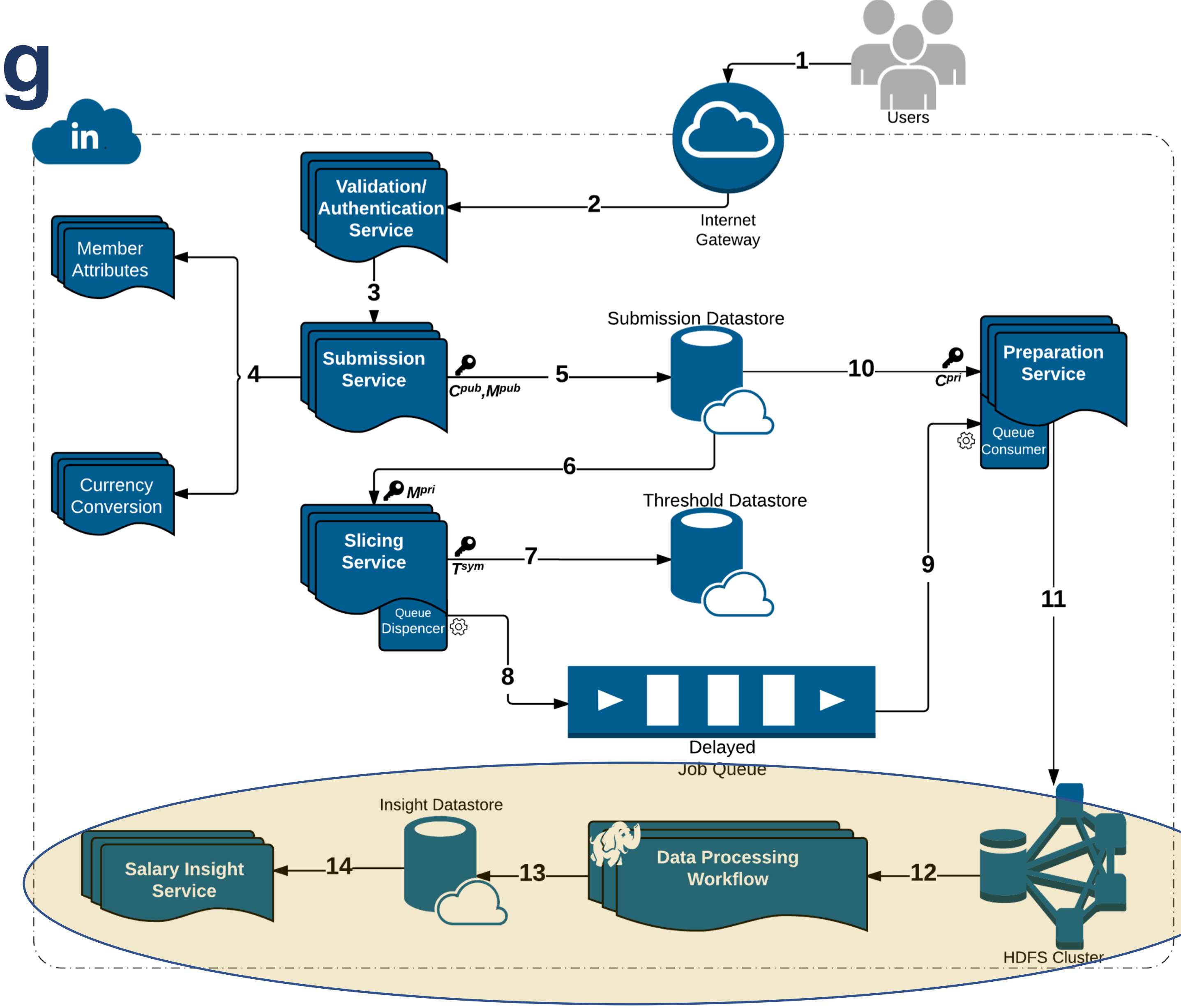


Insights & Modeling

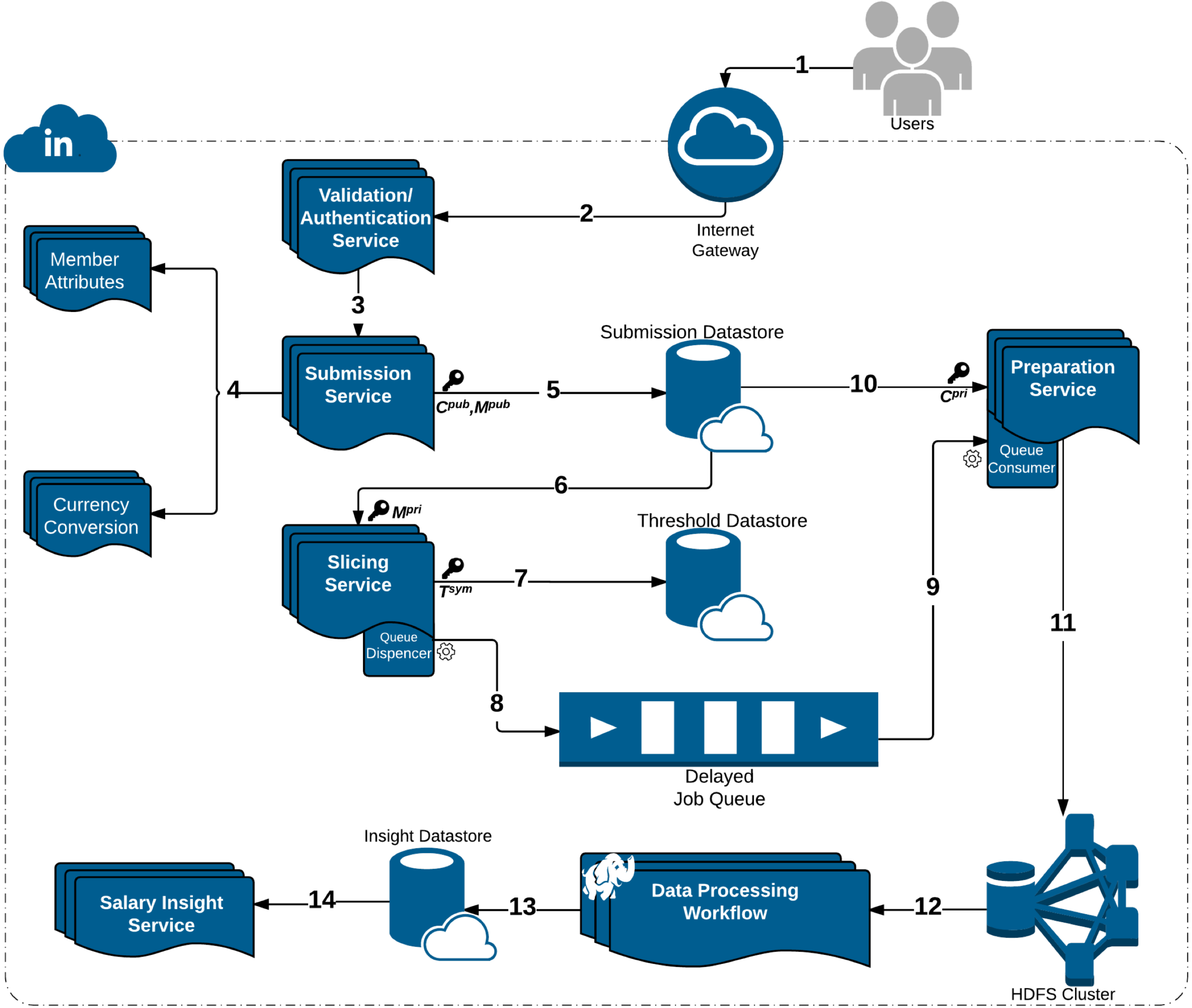


Insights & Modeling

- Salary insight service
  - Check whether the member is eligible
    - Give-to-get model
  - If yes, show the insights
- Offline workflow
  - Consume de-identified HDFS dataset
  - Compute robust compensation insights
    - Outlier detection
    - Bayesian smoothing/inference
  - Populate the insight key-value stores

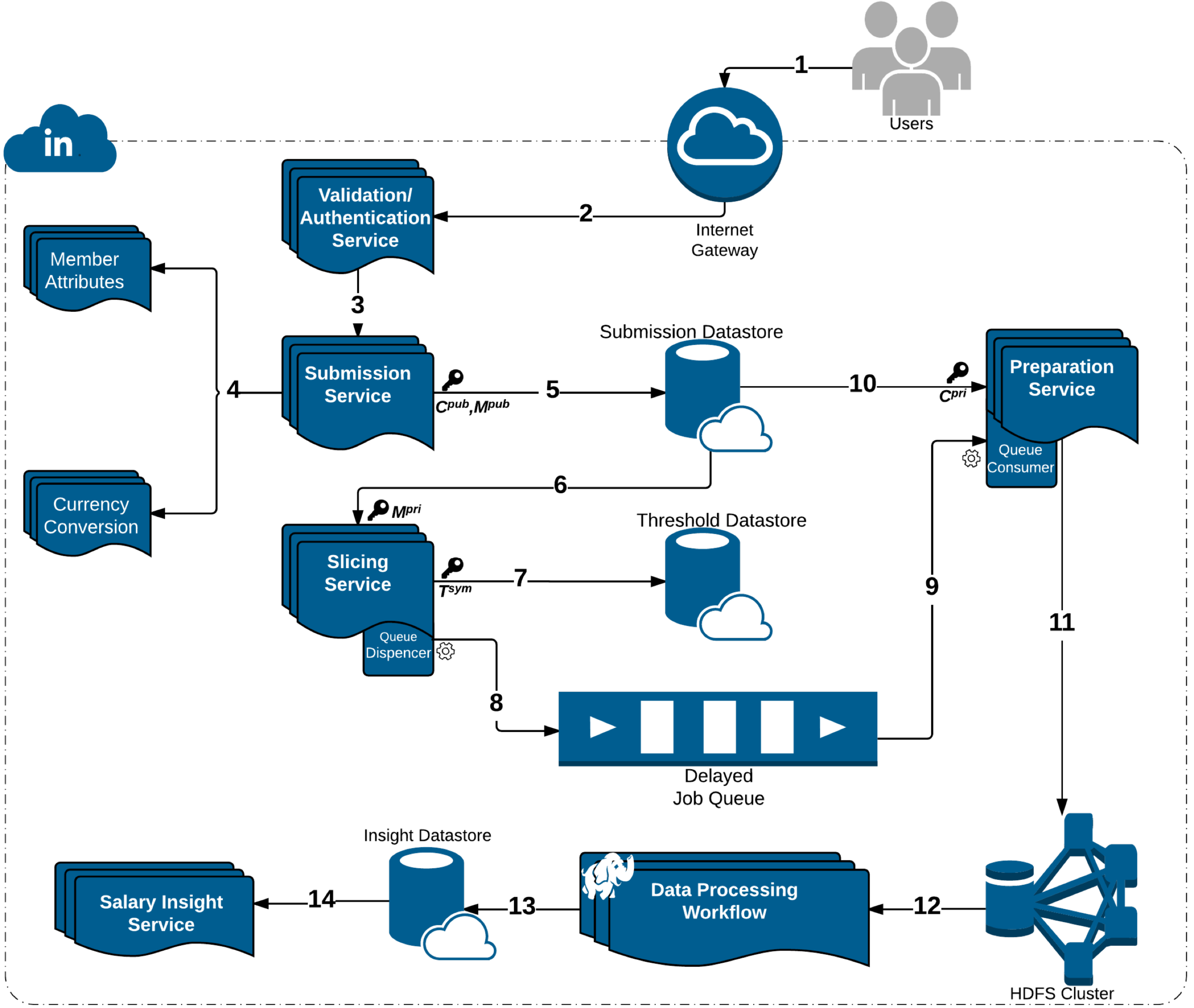


Security Mechanisms



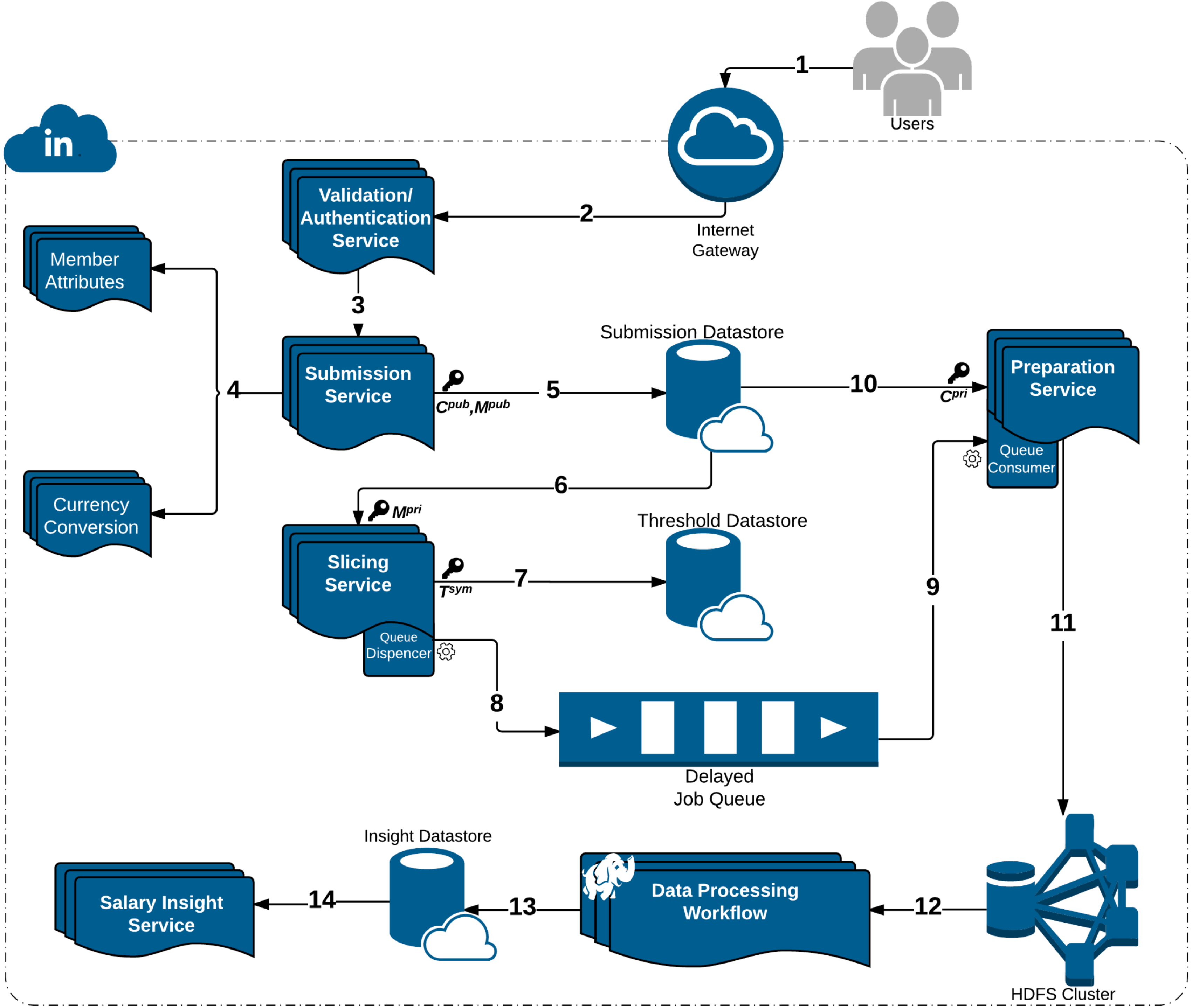
Security Mechanisms

- Encryption of member attributes & compensation data using different sets of keys
  - Separation of processing
  - Limiting access to the keys



Security Mechanisms

- Key rotation
- No single point of failure
- Infra security



Preventing Timestamp Join based Attacks

- Inference attack by joining these on timestamp
  - De-identified compensation data
  - Page view logs (when a member accessed compensation collection web interface)
  - ➔ Not desirable to retain the exact timestamp

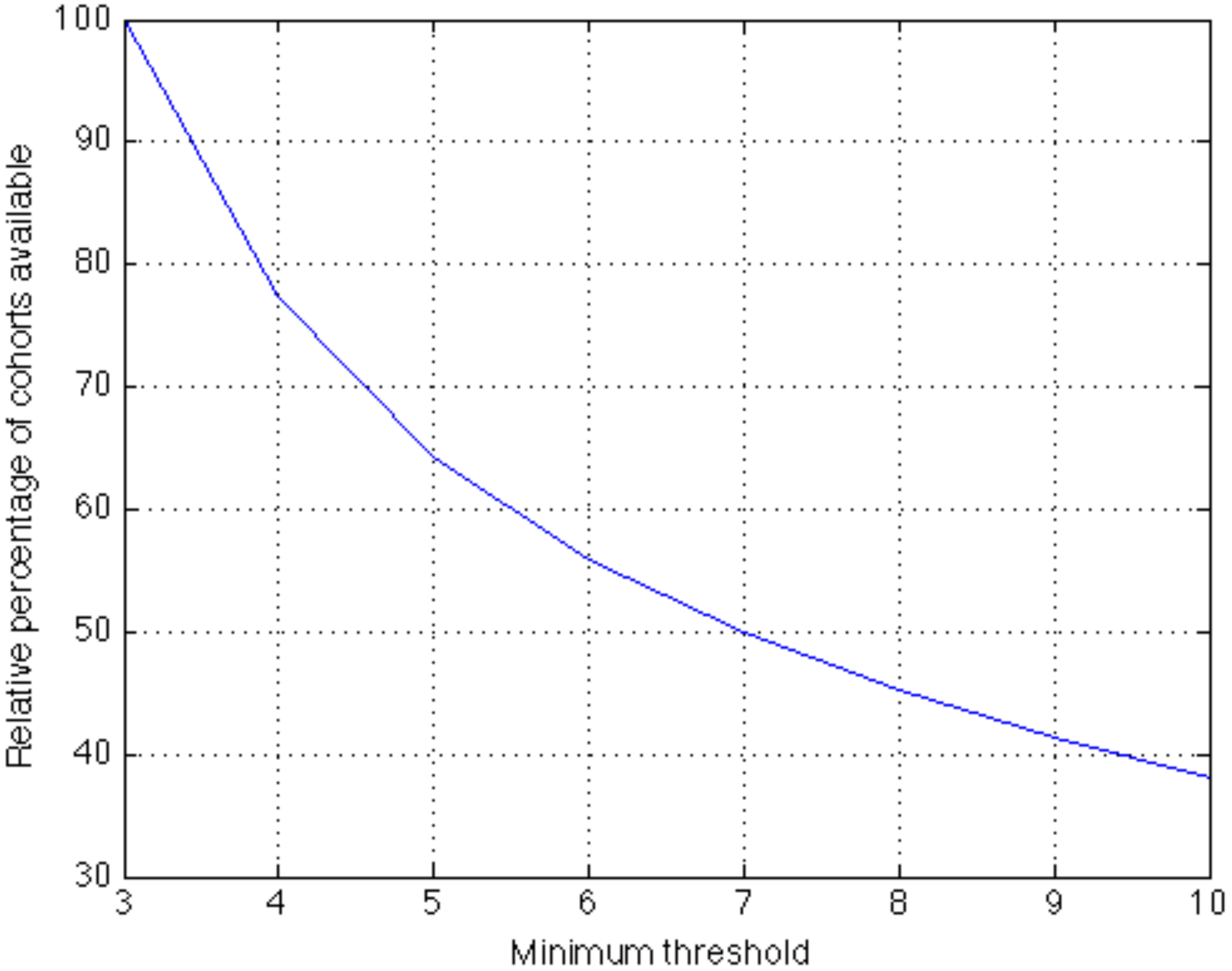
- Perturb by adding random delay (say, up to 48 hours)

- Modification based on k-Anonymity
  - Generalization using a hierarchy of timestamps
  - But, need to be incremental
  - ➔ Process entries within a cohort in batches of size k
    - Generalize to a common timestamp
    - Make additional data available only in such incremental batches

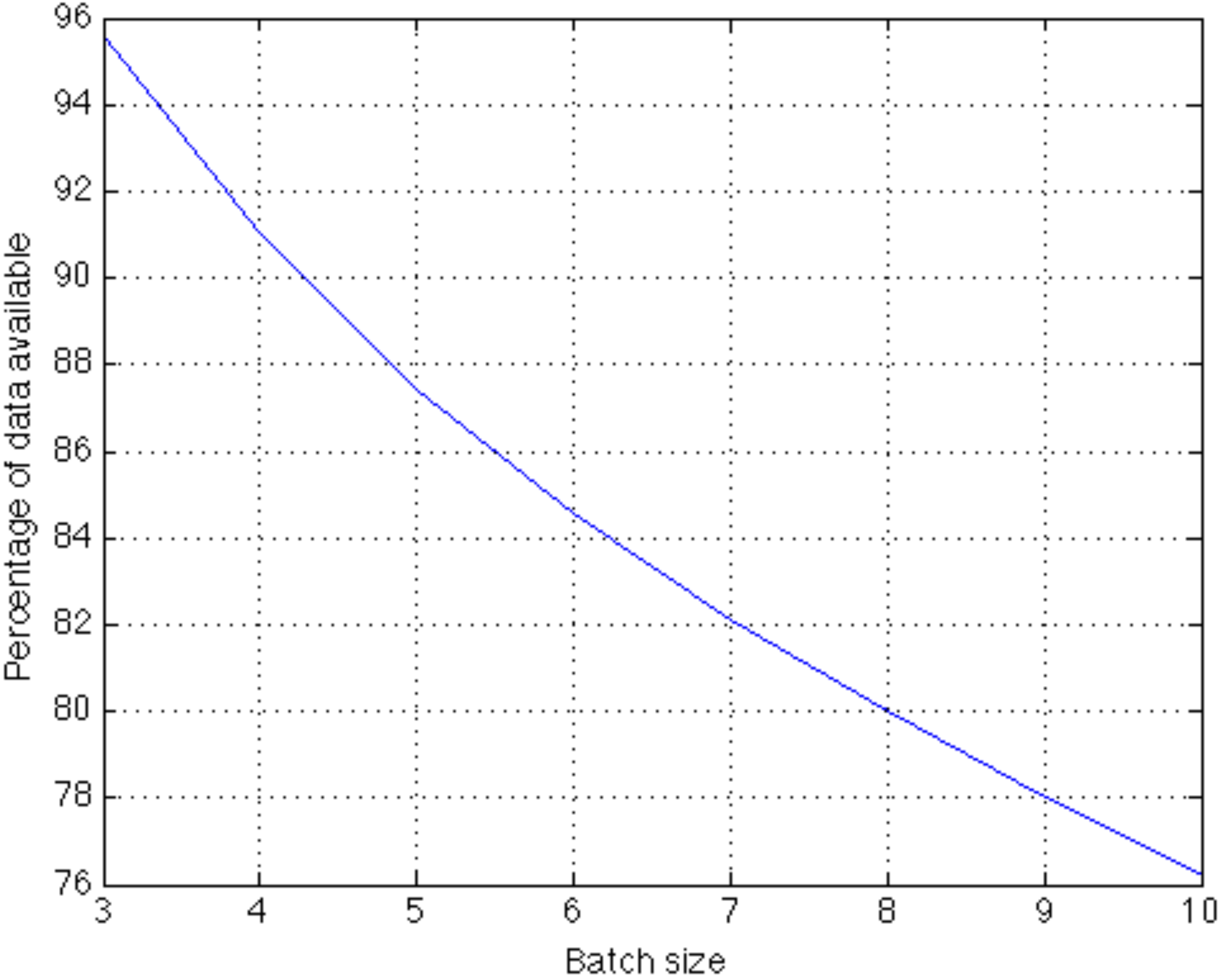
Privacy vs Modeling Tradeoffs

- LinkedIn Salary system deployed in production for ~2.5 years
- Study tradeoffs between privacy guarantees (‘k’) and data available for computing insights
  - Dataset: Compensation submission history from 1.5M LinkedIn members
  - Amount of data available vs. minimum threshold, k
  - Effect of processing entries in batches of size, k

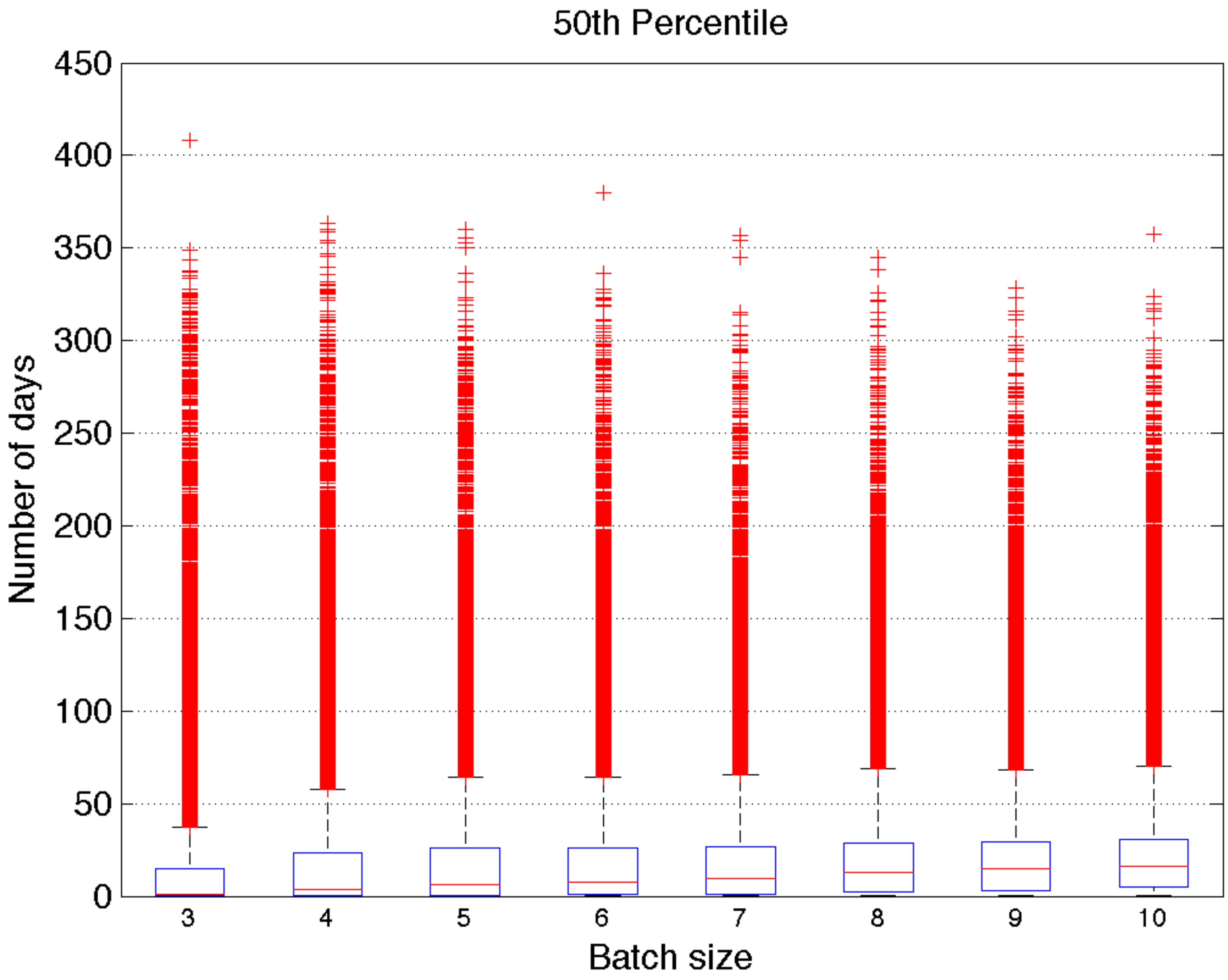
Amount of data available vs. threshold, k



Percent of data available vs. batch size, k



Median delay due to batching vs. batch size, k



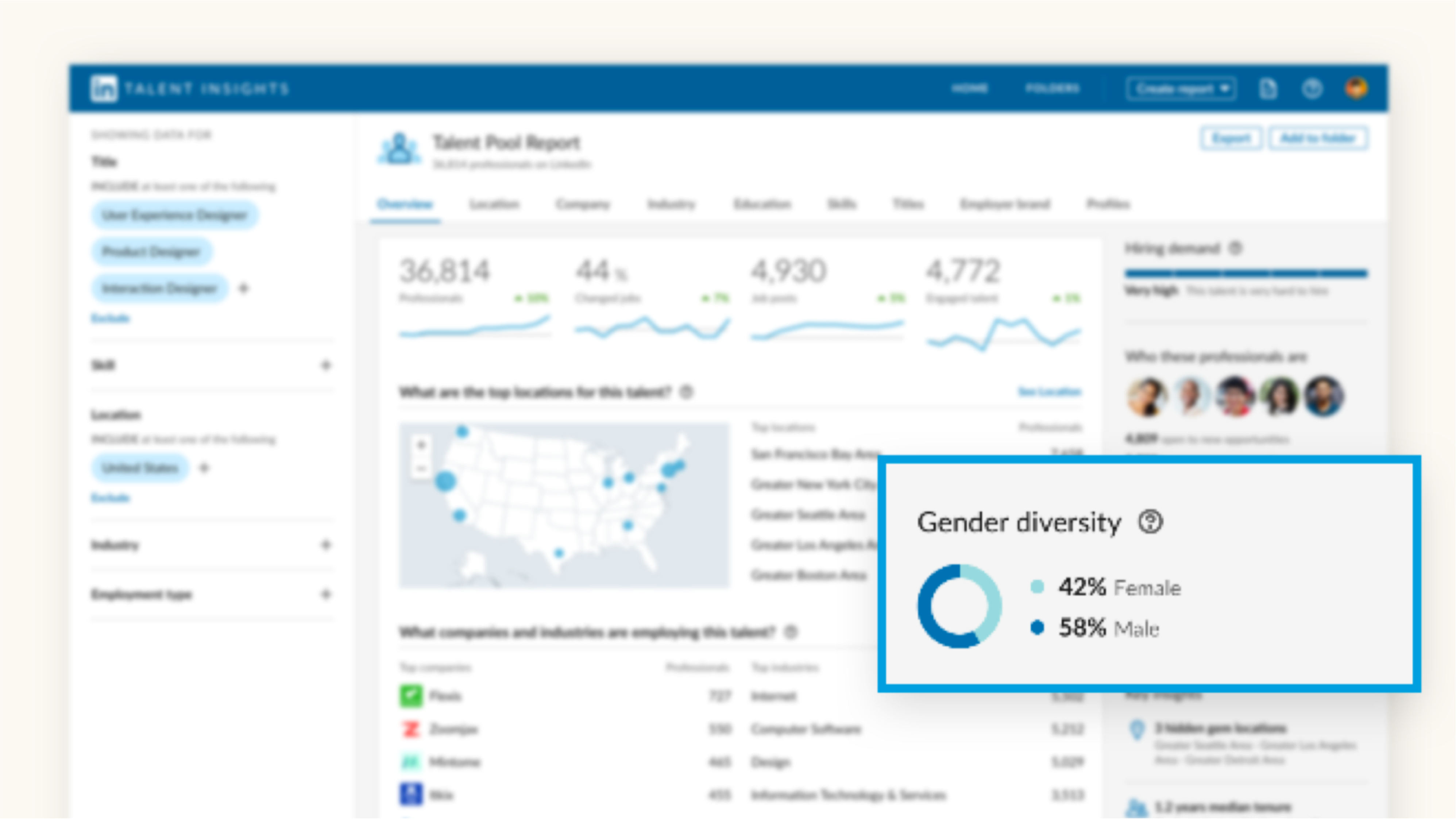
Key takeaway points

- LinkedIn Salary: a new internet application, with unique privacy/modeling challenges

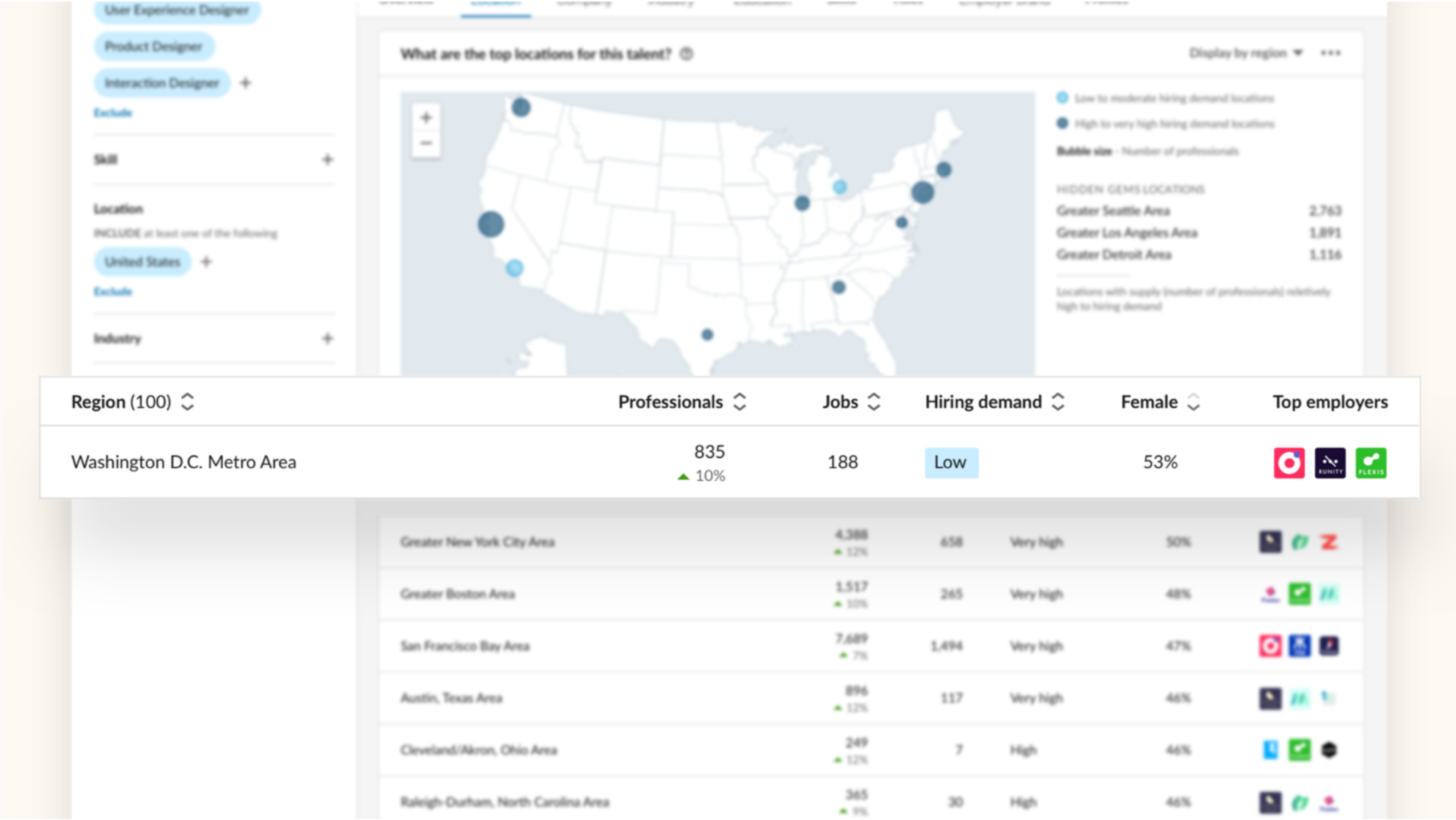
- Privacy vs. Modeling Tradeoffs

- Potential directions
  - Privacy-preserving machine learning models in a practical setting [e.g., Chaudhuri et al, JMLR 2011; Papernot et al, ICLR 2017]
  - Provably private submission of compensation entries?

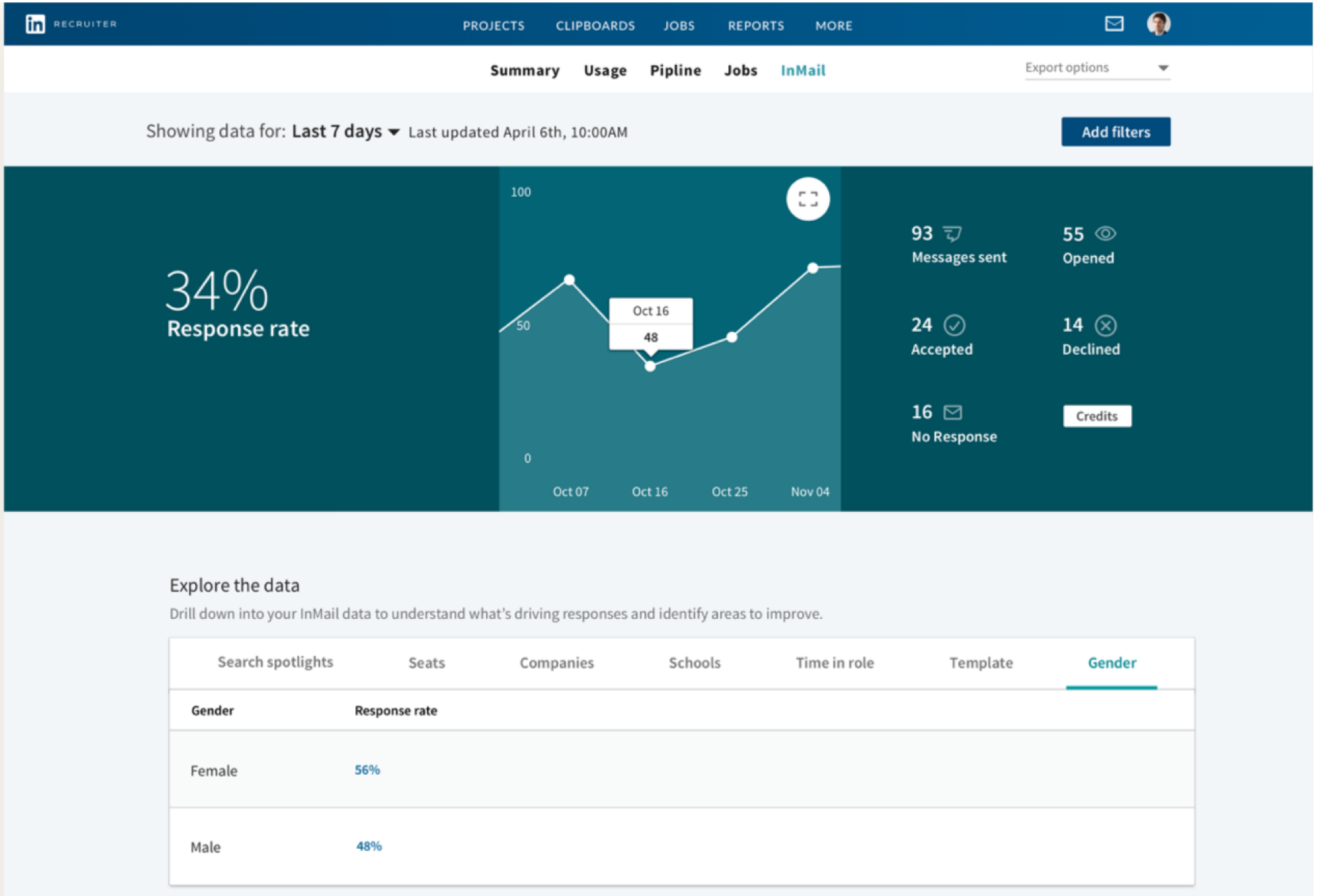
Plan for Diversity



Identify Diverse Talent Pools



Inclusive Job Descriptions / Recruiter Outreach



Measuring (Lack of) Representativeness

- **Skew@k**
  - (Logarithmic) ratio of the proportion of candidates having a given attribute value among the top k ranked results to the corresponding proportion among the set of qualified candidates

$$Skew_v@k(\tau_r) = \log_e\left(\frac{p_{\tau_r^k,r,v}}{p_{q,r,v}}\right)$$

  - **Minimum Discrete Skew:** Minimum over all attribute values genders (e.g., the most underrepresented gender's skew value).
    - Skew = 0 if we have $\lfloor p_{q,r,v} * k \rfloor$ candidates from value v in the top k results