Monitoring the Data Tsunami

Johannes Gehrke

Cornell University

SFB 876; January 20, 2011.

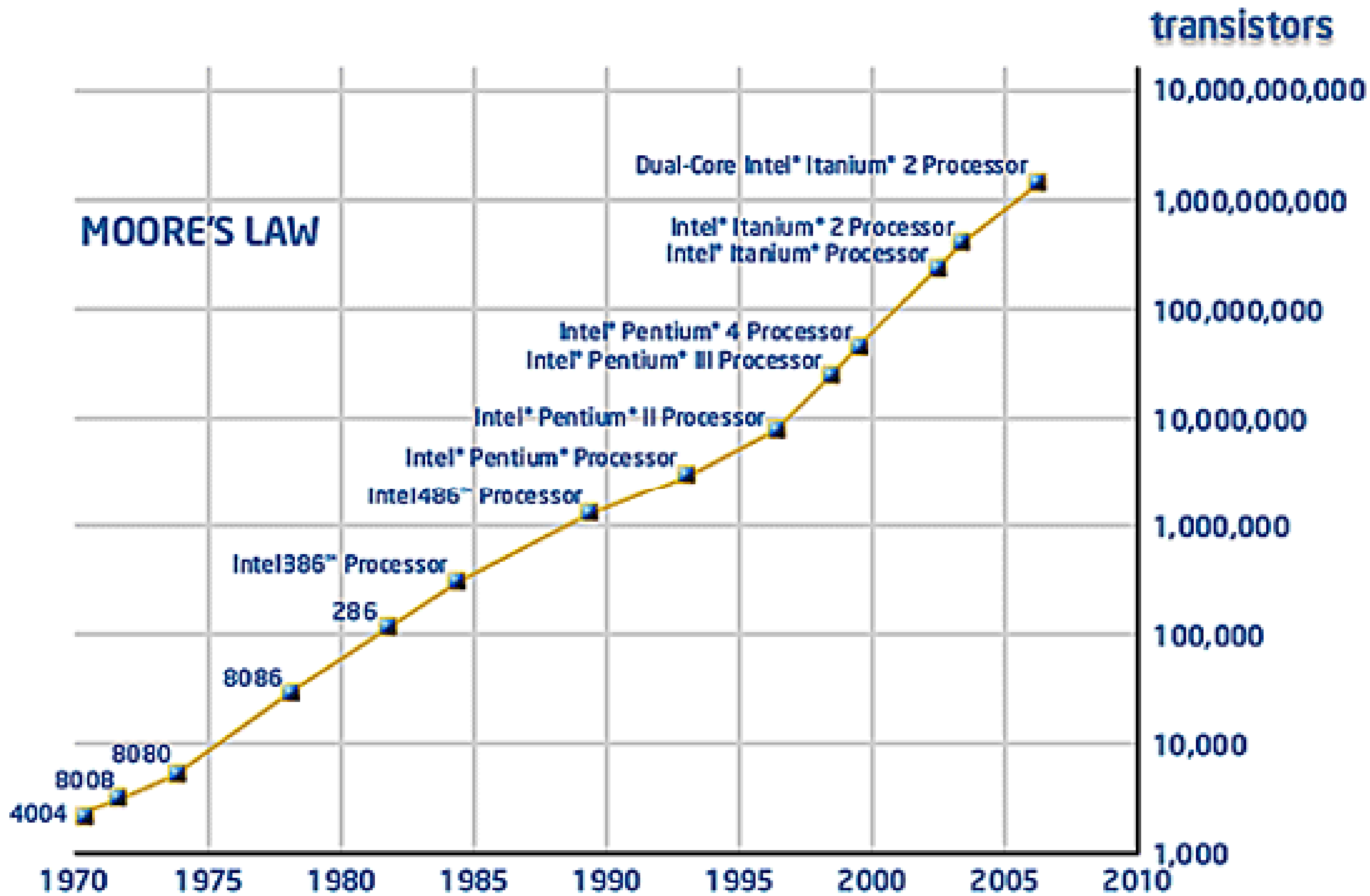
Cornell University

An Abundance of Data

- Supermarket scanners
- Credit card transactions
- Call center records
- ATM machines
- Web server logs
- Customer web site trails
- Podcasts
- Blogs
- Closed caption
- Scientific experiments
- Sensors
- Cameras
- Interactions in social networks
- Facebook, Myspace
- Twitter
- Speech-to-text translation
- Email

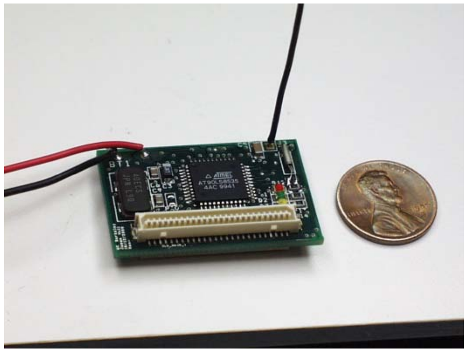
•Print, film, optical, and magnetic storage: 5 Exabytes (EB) of new information in 2002, doubled in the last three years
[How much Information 2003, UC Berkeley]

Driving Factors: A LARGE Hardware Revolution

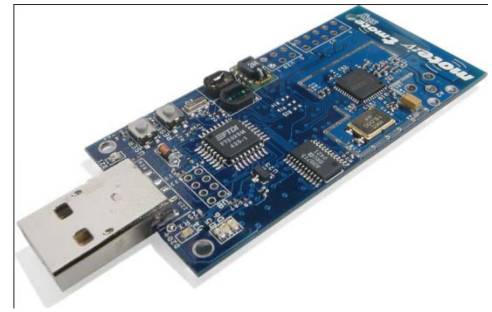


[Intel Corporation]

A small Hardware Revolution



http://www.snm.ethz.ch/Projects/MicaZ



http://www.snm.ethz.ch/Projects/TmoteSky

http://lecs.cs.ucla.edu/Resources/testbed/testbed-overview.html

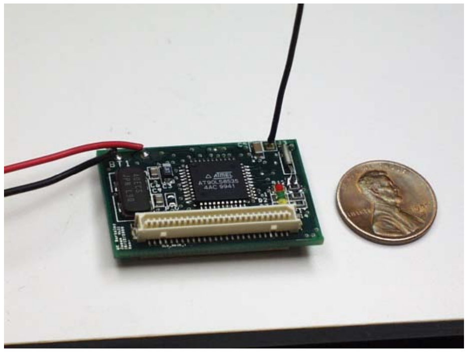
http://www.snm.ethz.ch/Projects/Telos

http://www.snm.ethz.ch/Projects/Mica2Dot

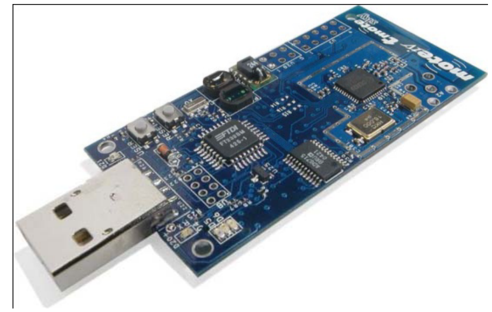
- Moore's Law
  - In 1965, Intel Corp. cofounder Gordon Moore predicted that the density of transistors in an integrated circuit would double every year.
  - Later changed to reflect 18 months progress.

Cornell University

Driving Factors: A small Hardware Revolution



http://www.snm.ethz.ch/Projects/MicaZ



http://www.snm.ethz.ch/Projects/TmoteSky

http://lecs.cs.ucla.edu/Resources/testbed/testbed-overview.html

http://www.snm.ethz.ch/Projects/Telos

http://www.snm.ethz.ch/Projects/Mica2Dot

- Experts on ants estimate that there are 10^{16} to 10^{17} ants on earth. In the year 1997, we produced one transistor per ant. [Gordon Moore]

Cornell University

Driving Factors: Connectivity and Bandwidth

- Metcalf’s law (network usefulness increases squared with the number of users)

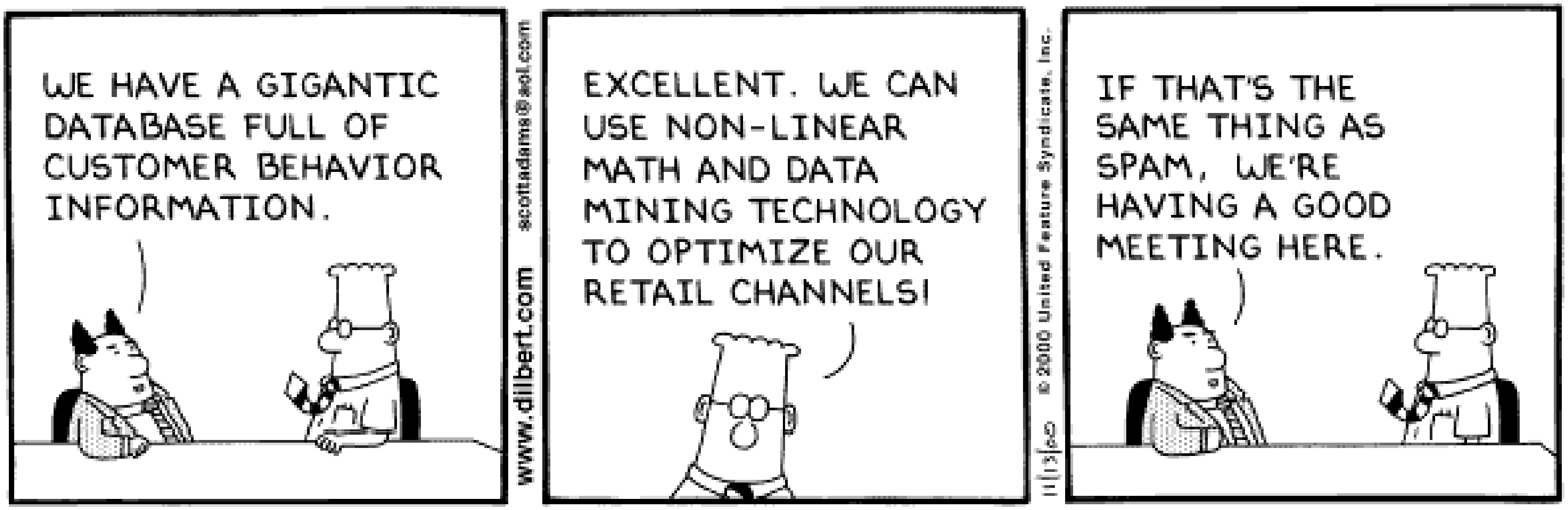
- Gilder’s law (bandwidth doubles every 6 months)

Definition

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Example pattern (Census Bureau Data):
If (relationship = husband), then (gender = male). 99.6%

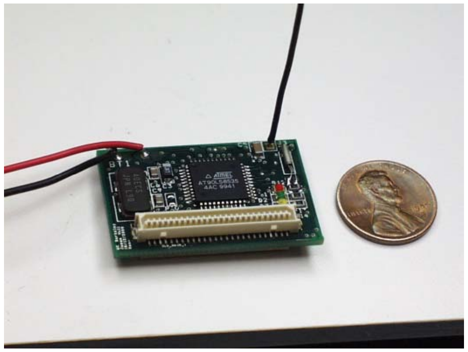
WHY?


WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
scottadams@aol.com
www.dilbert.com
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
© 2000 United Feature Syndicate, Inc.
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.

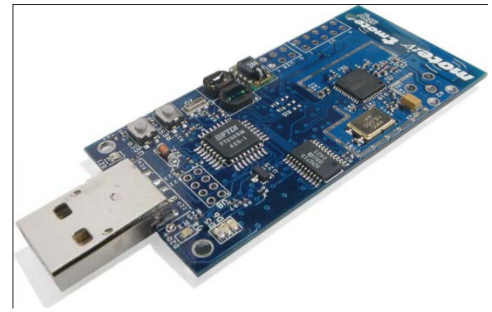
Why? Three Examples

- Sensor networks
- BIG Science Data
- Photos and videos

A small Hardware Revolution



http://www.snm.ethz.ch/Projects/MicaZ



http://www.snm.ethz.ch/Projects/TmoteSky

http://lecs.cs.ucla.edu/Resources/testbed/testbed-overview.html

http://www.snm.ethz.ch/Projects/Telos

http://www.snm.ethz.ch/Projects/Mica2Dot

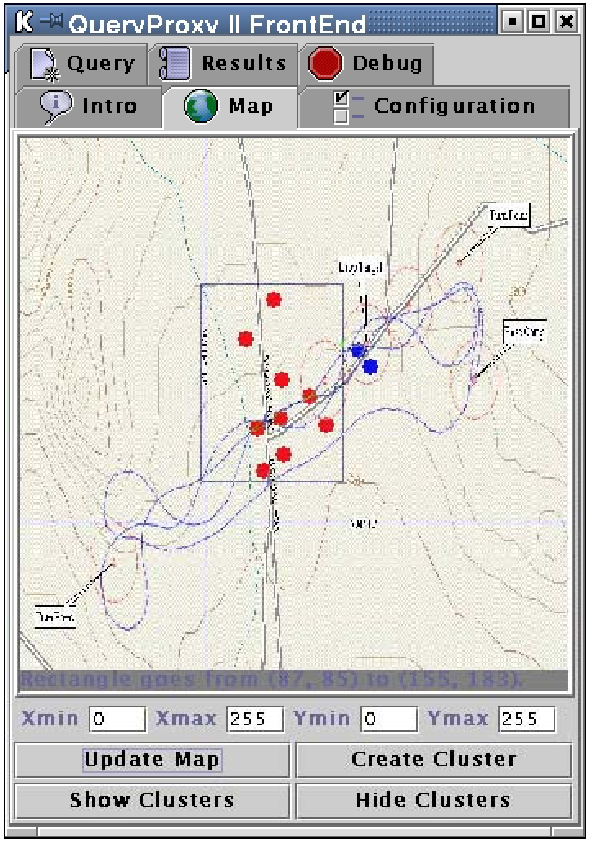
Flexible Decision Support

Traditional

Procedural addressing of individual sensor nodes; user specifies how task executes, data is processed centrally.

Today

Complex declarative querying and tasking. User isolated from "how the network works", in-network distributed processing.



http://www.cs.cornell.edu/bigreddata/cougar/

Cornell University

Querying: Model

Time	Value
12	82
13	83

Time	Value
13	82
15	83

Time	Value
13	82
15	84

Time	Value
14	79
15	83

Time	Value
13	82
15	83

Time	Value
13	80
16	83

Example Queries

- Snapshot queries:
  - What is the concentration of chemical X in the northeast quadrant?

```
SELECT AVG(R.sensor.concentration)
FROM Relation R
WHERE R.sensor.loc in (50,50,100,100)
```

  - In which area is the concentration of chemical X higher than the average concentration?

```
SELECT AVG(R.sensor.concentration)
FROM Relation R
GROUP BY R.area
HAVING AVG(R.sensor.concentration) >
               (SELECT AVG(R.sensor.concentration)
                FROM Relation R
                GROUP BY R.area)
```

Example Queries (Contd.)

- Long-running queries
  - Notify me over the next hour whenever the concentration of chemical X in an area is higher than my security threshold.

```
SELECT R.sensor.area, AVG(R.sensor.concentration)
FROM Relation R
WHERE R.sensor.loc in rectangle
GROUP BY R.sensor.area
DURATION (now,now+3600)
```

- Archival queries
  - Periodic data collection for offline analysis

Goals

- Declarative, high-level tasking
- User is shielded from network characteristics
  - Changes in network conditions
  - Changes in power availability
  - Node movement
- System optimizes resources
  - High-level optimization of multiple queries
  - Trade accuracy versus resource usage versus timeliness of query answer

Challenges

Technical:

- Scale of the system
- Constraints
  - Power, communication, computation
- Constant change, uncertainty from sensor measurements
- Distribution and decentralization

http://www.fatvat.co.uk/2010/07/stop-traffic.html

Application:

- Traffic monitoring
- Health Care
- Care for the elderly

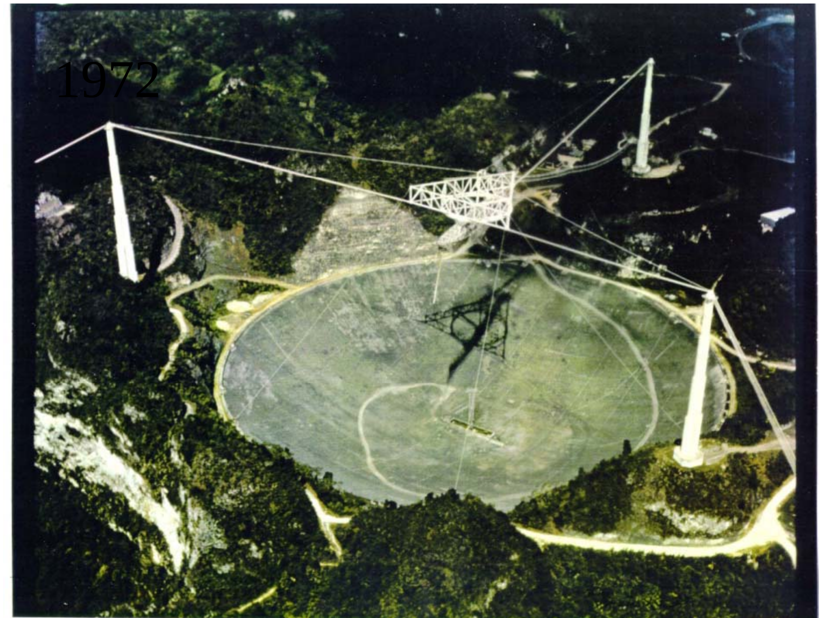
And of course the resulting data tsunami!

Cornell University

Three Examples

- Sensor networks
- BIG Science Data
- Photos and videos

http://www.geoeye.com/CorpSite/gallery/detail.aspx?iid=275&gid=51

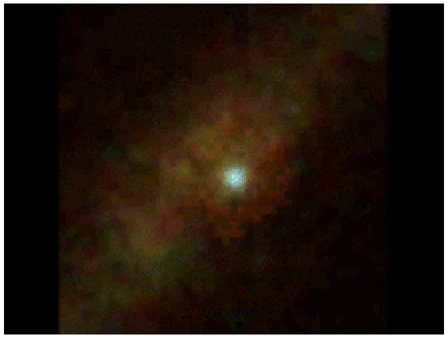


http://www.naic.edu/

Cornell University


http://www.naic.edu/

Pulsars



http://en.wikipedia.org/wiki/Pulsar

- Pulsars are rotating stars
- Of interest are
  - Millisecond pulsars
  - Compact binaries
- Example:
  - Hulse-Taylor binary
  - Used to infer gravitational waves in support of Einstein's General Theory of Relativity
  - Nobel price in physics in 1993

Cornell University

Pulsar Surveys

- Most demanding of the ALFA surveys
  - ~ 100 MB/s to disk
  - ~ 1 PB for entire survey (3-5 yr @ 6-10% duty cycle)
- Requires coarsely parallel processing of raw data in discrete, local data chunks
  - processing time ~ 50-200x data acquisition time on single processor (Intel 2.5 GHz 512k cached with 1GB ram)
  - depends on data set details, algorithms, code
  - Distributed initial processing (Cornell + 5 sites)
- Requires meta-analysis of data products of the initial analysis
  - Database and data mining research problems

Cornell University

Project Requirements

- Data
  - 14 TB every 2 weeks
  - Shipped on USB-2 disk drives
  - Need to archive raw data 5+ years
  - Need to make data products to the astronomy research community
- Processing
  - Extremely processor intensive
    - Currently just exhaustive search over a large parameter space (periodicity, dispersion, time)
  - Find new pulsars --- and other *interesting* phenomena
- More information: http://arecibo.tc.cornell.edu/hiarchive/

Cornell University

Three Examples

- Sensor networks
- BIG Science Data
- Photos and videos

The Need for Large-Scale Image Processing

Photos:

- **5 billion** – Photos hosted by Flickr
- **3000+** – Photos uploaded per minute to Flickr.
- **130 million** – At the above rate,
  the number of photos uploaded per month
- **3+ billion** – Photos uploaded per month to Facebook.

www.facebook.com

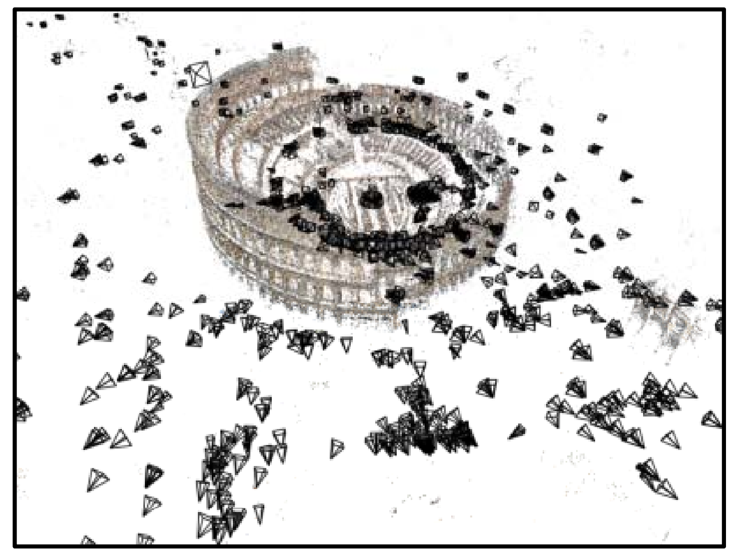
flickr™

www.flickr.com

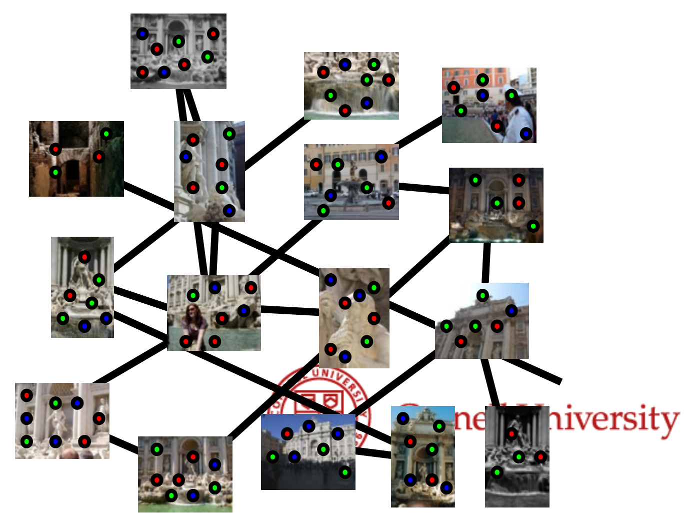
Video:

- **2 billion** – The number of videos watched per day on YouTube.
- **35** – Hours of video uploaded to YouTube every minute.
- **186** – The number of online videos the average Internet user watches in a month (USA).
- **2+ billion** – The number of videos watched per month on Facebook.
- **20 million** – Videos uploaded to Facebook per month.

The Power of a Data-Rich Environment



- Current System: 150,000 photos take 1 day on 500 cores
- Goal : Billions in days



Pictures courtesy of Noah Snavely
http://www.cs.cornell.edu/~snavely/

Statue of Liberty

7834 images registered (322 in skeletal set)

Picture courtesy of Noah Snavely
http://www.cs.cornell.edu/~snavely/

Cornell University

Summary: Why

- Sensor networks
- BIG Science Data
- Photos and videos

- Many others:
  - Cloud
  - Multi-core
  - Handheld devices

Talk Outline

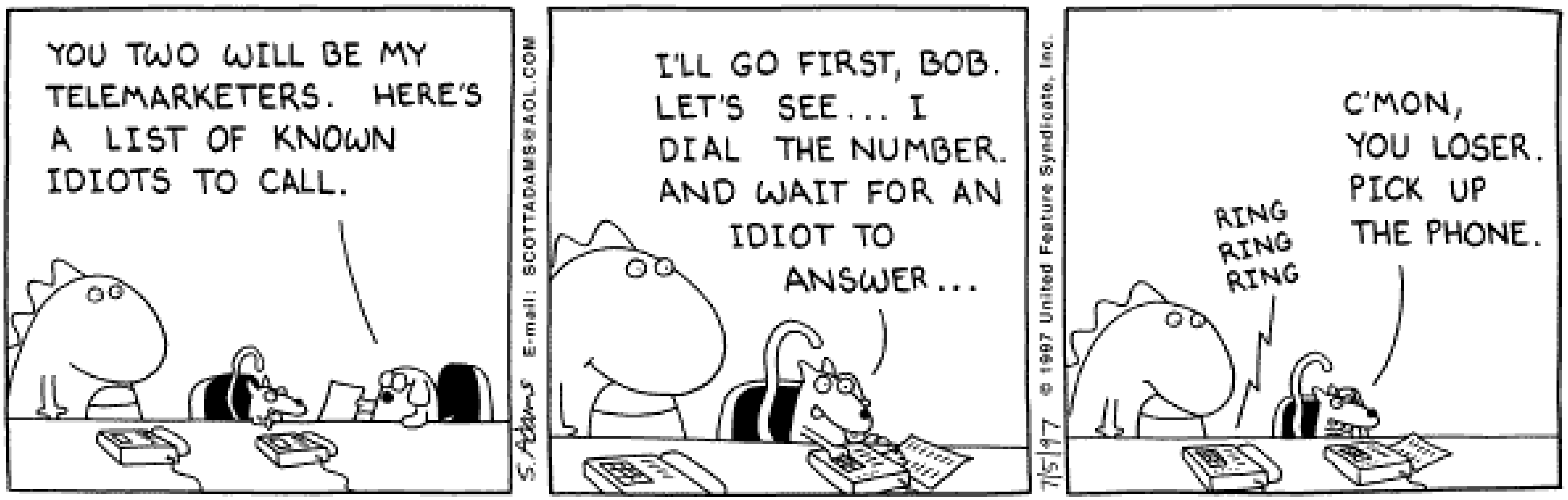
- Introduction
- Techniques for data stream processing
- Data privacy
- Conclusions

Cornell University

Talk Outline

- Introduction
- Techniques for data stream processing
- Data privacy
- Conclusions

HOW


YOU TWO WILL BE MY TELEMARKETERS. HERE'S A LIST OF KNOWN IDIOTS TO CALL.
I'LL GO FIRST, BOB. LET'S SEE... I DIAL THE NUMBER. AND WAIT FOR AN IDIOT TO ANSWER...
RING RING RING
C'MON, YOU LOSER. PICK UP THE PHONE.
S. Adams
E-mail: SCOTTADAMS@AOL.COM
© 1997 United Feature Syndicate, Inc.
7/5/97

Talk Outline

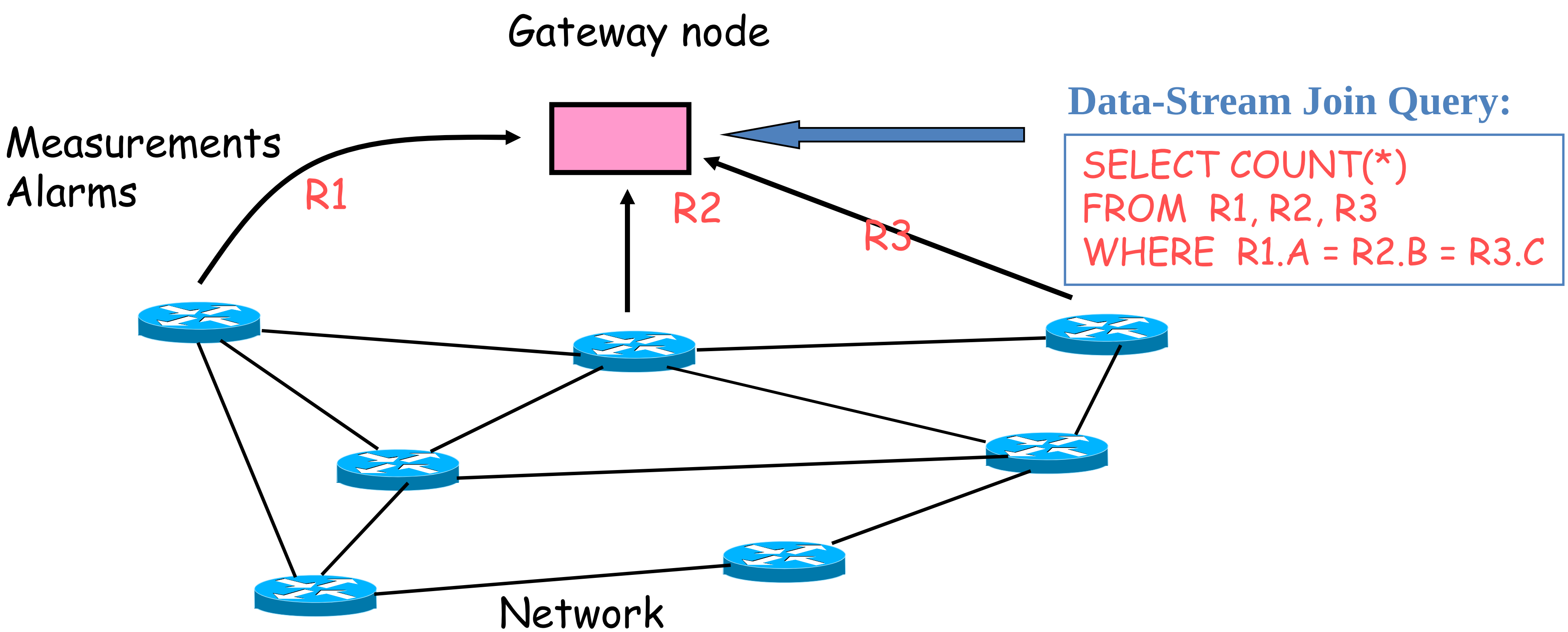
- Introduction
- Techniques for data stream processing
  - Stream summaries
  - Complex event processing
- Data privacy
- Conclusions

Talk Outline

- Introduction
- Techniques for data stream processing
  - Stream summaries
  - Complex event processing
- Data privacy
- Conclusions

Processing Network Data Streams

- Data-stream processing arises naturally in Network Management
  - Data tuples arrive continuously from different parts of the network
  - Archival storage is often off-site (expensive access)
  - Queries can only look at the tuples *once, in the fixed order of arrival* and with *limited available memory*



Minos N. Garofalakis, Johannes Gehrke, Rajeev Rastogi: Querying and mining data streams: you only get one look a tutorial. SIGMOD Conference 2002: 635

Cornell University

Data Stream Processing Model

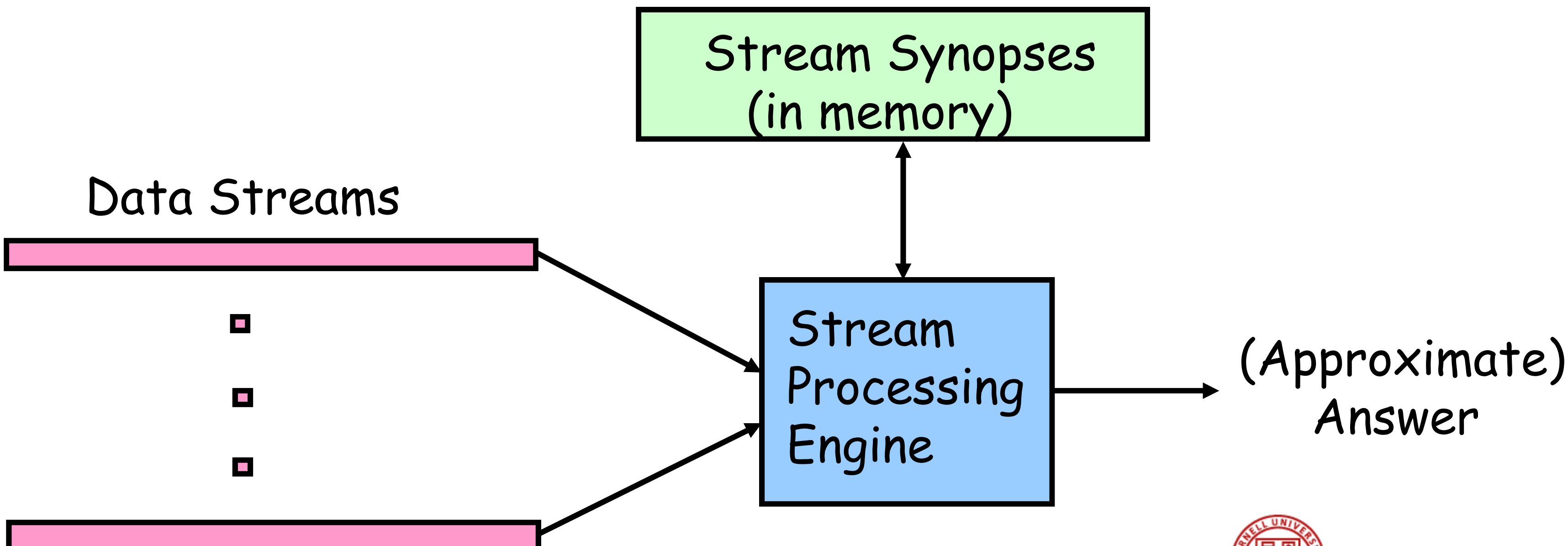
- Approximate query answers often suffice (e.g., trend/pattern analyses):
  - High-level analysis, then (expensive) retrieval and deep analysis of relevant data

<u>Approach:</u>

- Build small synopses of the data streams online
- Use synopses to provide (good-quality) approximate answers

Data Stream Processing Model

- Requirements for stream synopses
  - Single pass: Each tuple is examined at most once, in fixed (arrival) order
  - Bounded storage: Log or poly-log in data stream size
  - Real-time: Per-record processing time (to maintain synopsis) must be low



Cornell University

Sketches

- Summary structure which can be constructed in one pass
- Incrementally maintainable
- Provable performance guaratnees

- Example: AMS sketches [N. Alon, Y. Matias and M. Szegedy, The space complexity of approximating the frequency moments, STOC 1996]

Estimating Self-join Sizes

- Example scenario
  - Stream R: a b a c c a
  - Compute: SJ(R)
    - SJ(R) = COUNT(R $\bowtie_A$ R) = $\Sigma_i f(i)^2$
  - SJ(R)= $\Sigma_i f(i)^2 = 3^2+1^2+2^2=14$

R

i	f(i)
a	3
b	1
c	2

- Any *deterministic* algorithm to approximate SJ(R) needs at least Ω(|Dom(A)|) memory [AMS96]

Cornell University

AMS Sketches

- Main features
  - Randomized technique
  - Summarize information in the stream with a single number $\Rightarrow$ atomic sketch

Estimating Self-Join Size

- Method for estimating SJ(R):
  - Select a family of independent {+1,-1} random variables
    - $\{\xi_i$: i=1..|dom(A)|} with $P[\xi_i=+1]=P[\xi_i=-1]=\frac{1}{2}$
    - $E[\xi_i]=0$
  - Compute atomic sketch: $X=\Sigma_{i \in Dom(A)} f(i)\ \xi_i$
    - Stream R: a b a c c a
    - $X = \xi_a + \xi_b + \xi_a + \xi_c + \xi_c + \xi_a$
  - Claim: X^2 approximates SJ(R)

AMS Sketches: Analysis

– Compute: $X = \sum_i f(i)\,\xi_i$

Want: $SJ(R) = \sum_i f(i)^2$

– $X^2 = \sum_i f(i)^2 \xi_i^2 + \sum_{i \neq j} f(i)f(j)\,\xi_i\xi_j$

$= \sum_i f(i)^2 + \sum_{i \neq j} f(i)f(j)\,\xi_i\xi_j$

– $E[X^2] = \sum_i f(i)^2 + \sum_{i \neq j} f(i)f(j)\,E[\xi_i\xi_j]$

$= SJ(R) + 0$

Atomic Sketch Computation

Crucial point:

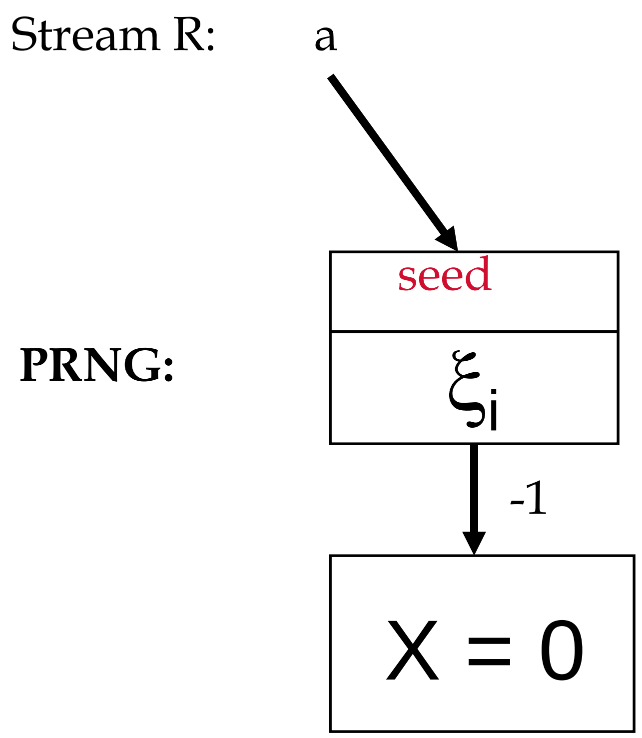
ξ_i values need not be fully independent Pairwise independence suffices

$\Rightarrow \xi_i$'s can be generated efficiently from small seeds [ABI86]

$\Rightarrow \xi$ vector is not stored. Required elements generated on the fly from seed of size O(log|Dom(A)|)

Cornell University

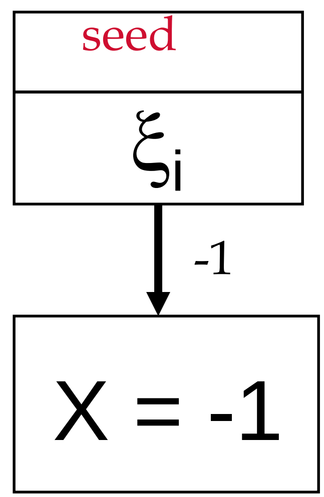
Example



Example

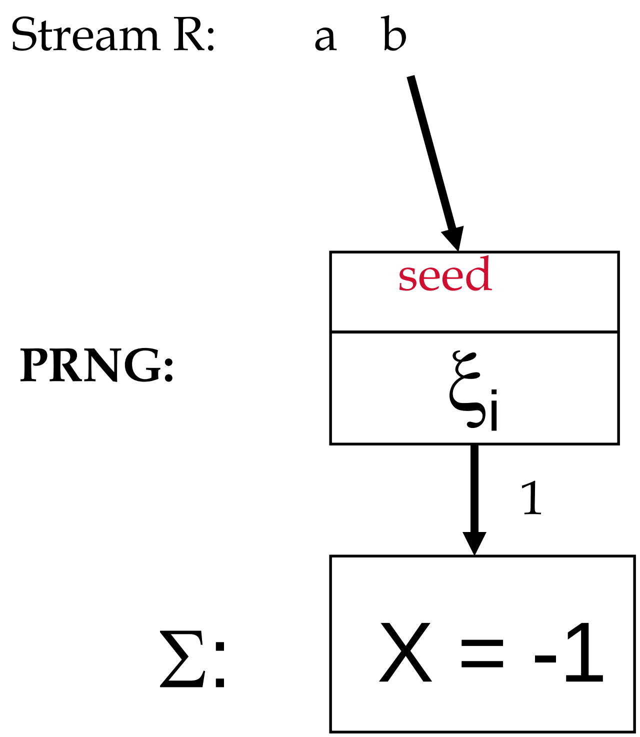
Stream R: a

PRNG:



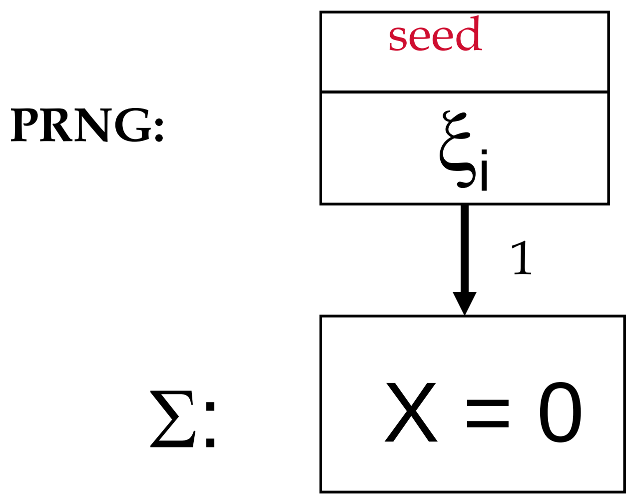
Σ:

Example

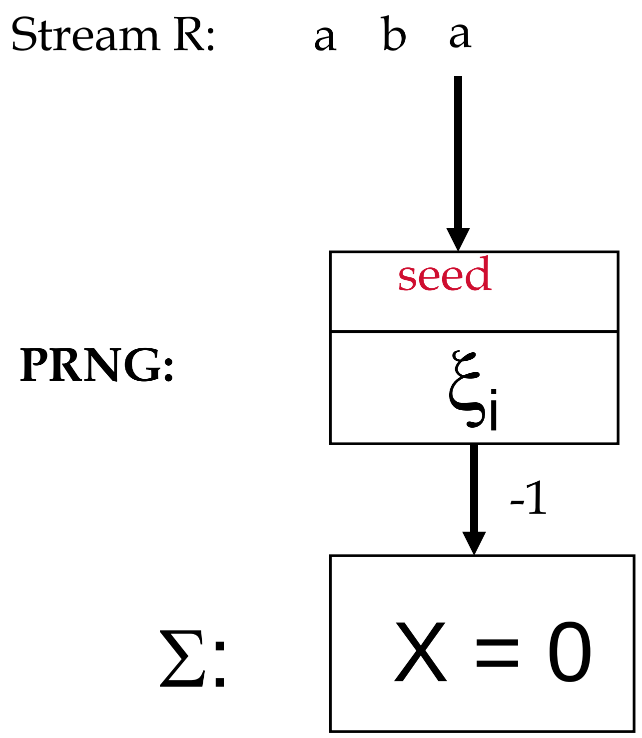


Example

Stream R: a b



Example

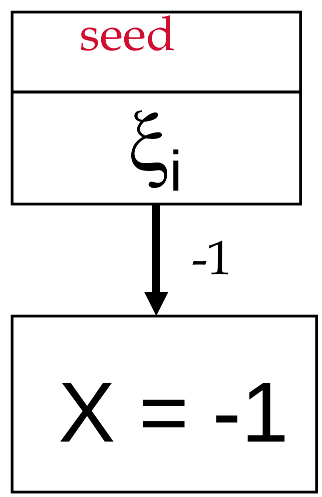


Example

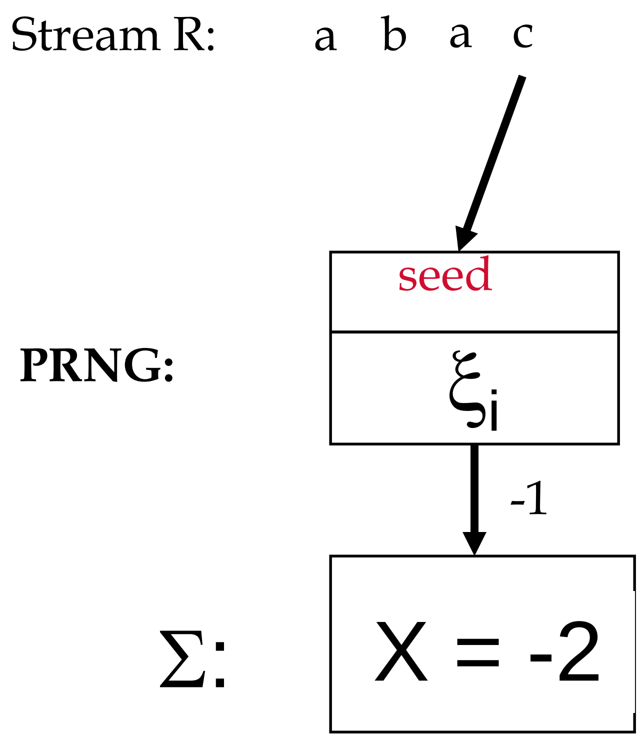
Stream R: a b a

PRNG:

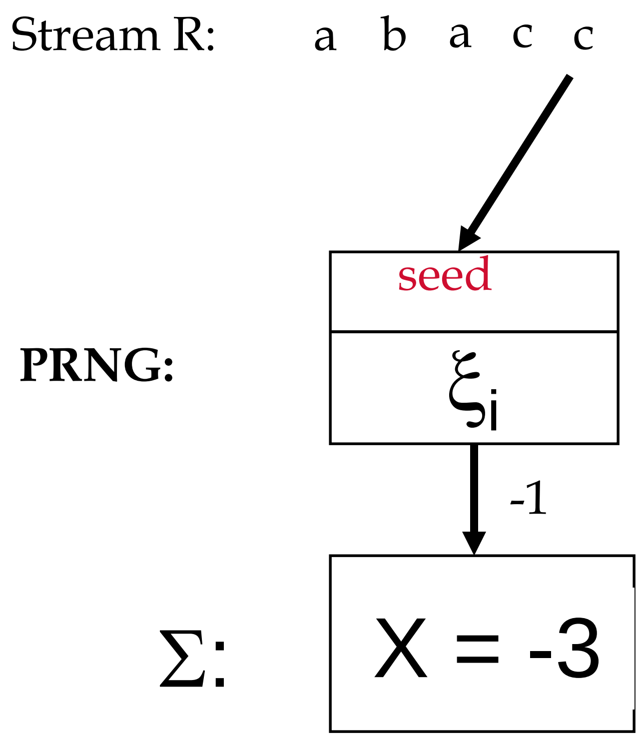
Σ:



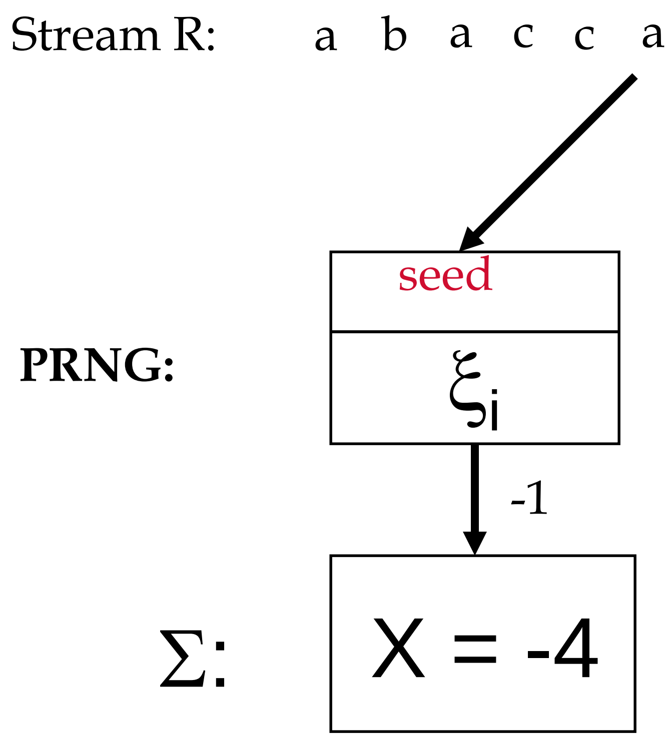
Example



Example



Example



Example

Stream R: a b a c c a

PRNG: seed

ξ_i

Σ: $X = -4$

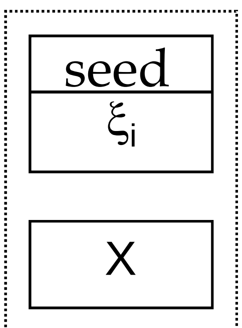
$Z = X^2$

Estimator Z=16 SJ(R)=14

Boosting

- Boosting: (ε,δ) guarantees

  Using $O(Var[Z] \log(1/\delta) \,/\, (\varepsilon^2 E^2[Z]))$ i.i.d. copies of Z, the computed estimate Z^* approximates E[Z] within (ε,δ)
  - $P(|Z^*-E[Z]| > \varepsilon E[Z]) \leq \delta$



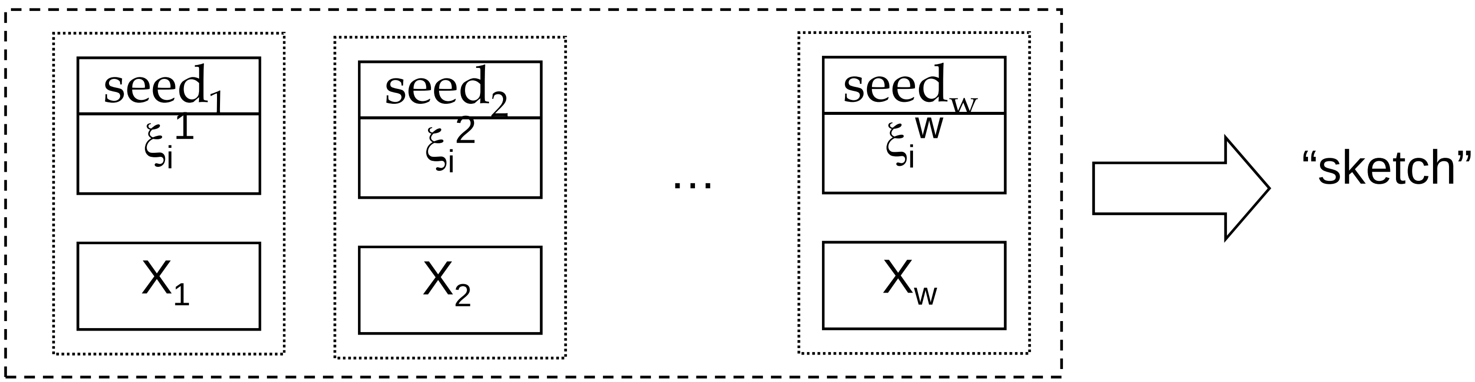
Cornell University

Boosting

- Boosting: (ε,δ) guarantees

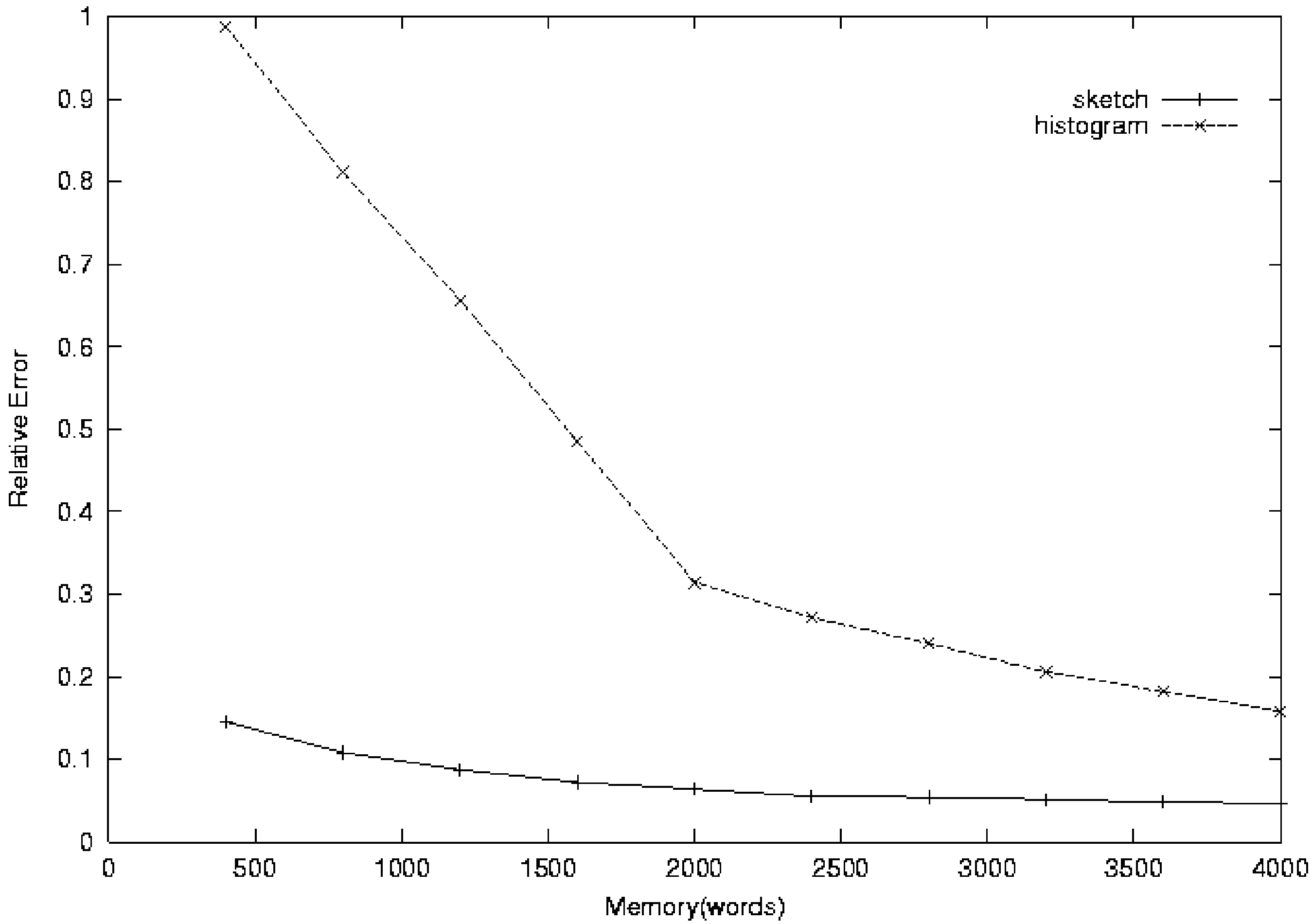
  Using $O(\mathrm{Var}[Z] \log (1/\delta) / (\varepsilon^2 E^2[Z]))$ i.i.d. copies of Z, the computed estimate Z^* approximates E[Z] within (ε,δ)

  – $P(|Z^*-E[Z]| > \varepsilon E[Z]) \leq \delta$



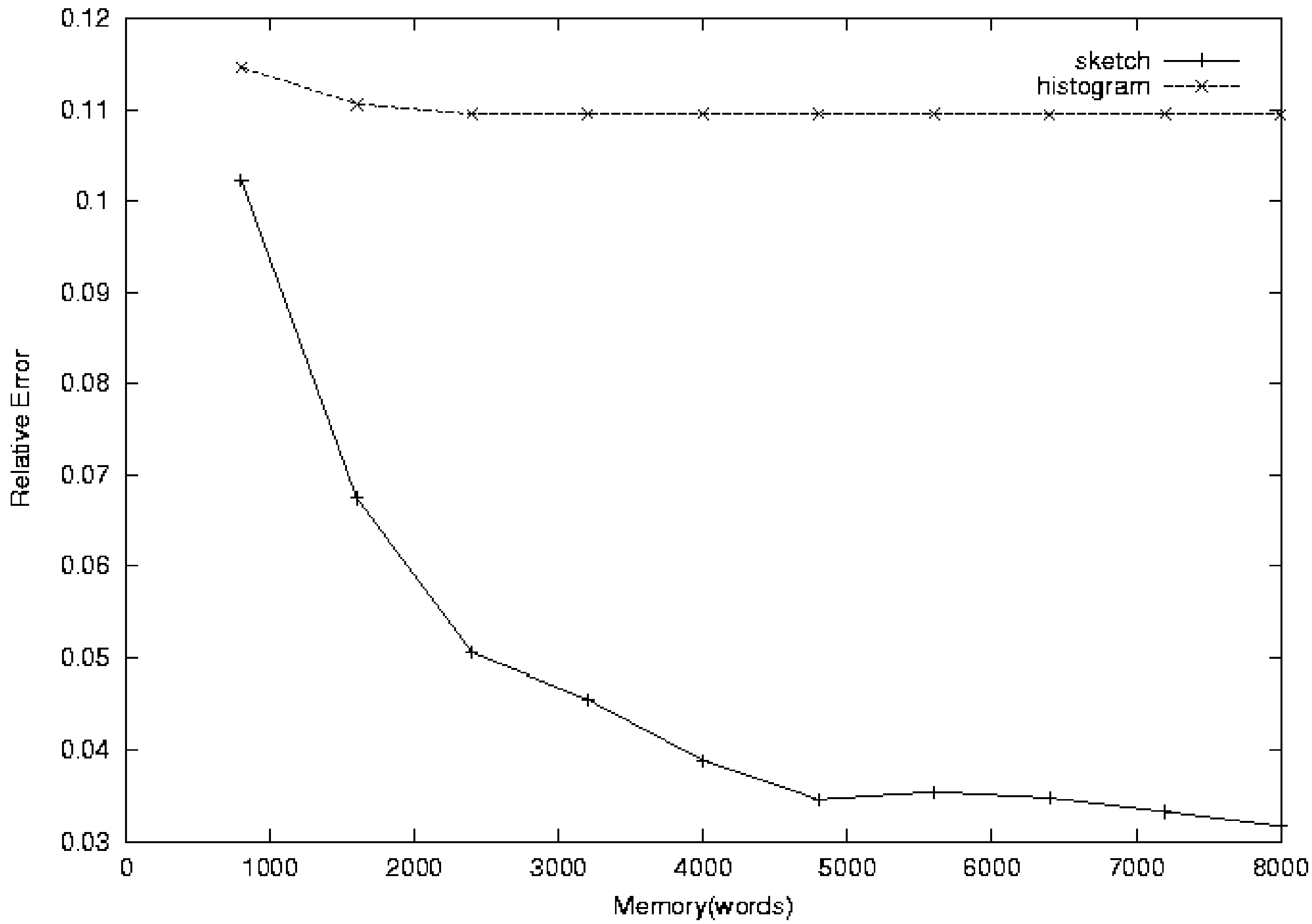
- Need ξ_i's to be 4-wise independent to get low variance

Performance: An Example



From: Alin Dobra, Minos N. Garofalakis, Johannes Gehrke, Rajeev Rastogi:
rocessing complex aggregate queries over data streams. SIGMOD Conference 2002: 61-72

Example: Two-Dimensional Join



From: Alin Dobra, Minos N. Garofalakis, Johannes Gehrke, Rajeev Rastogi:
rocessing complex aggregate queries over data streams. SIGMOD Conference 2002: 61-72

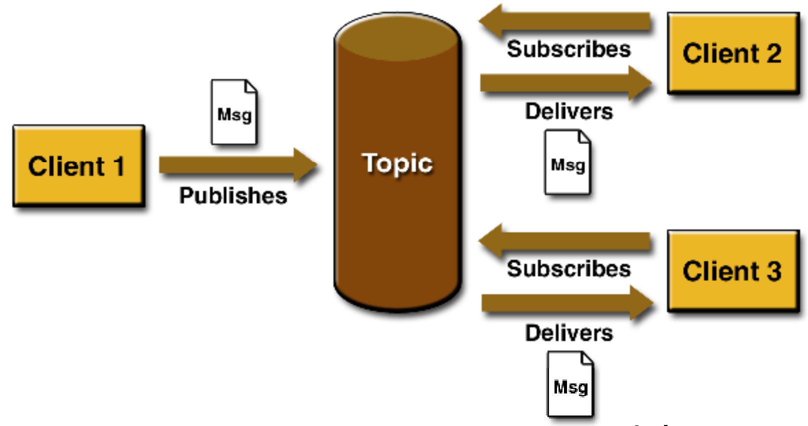
Cornell University

Talk Outline

- Introduction
- Techniques for data stream processing
  - Stream summaries
  - Complex event processing
- Data privacy
- Conclusions

Standard Pub/Sub

- Publishers generate data
  - Events, publications
- Subscribers describe interests in publications
  - Queries, subscriptions
- Asynchronous communication
  - Decoupling of publishers and subscribers
- Example: Tibco, Twitter



Source: JMS tutorial

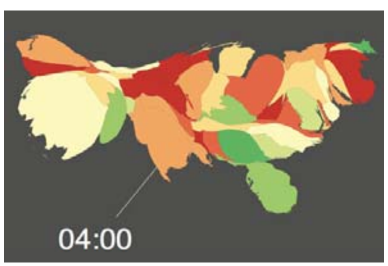
TIBCO® The Power of Now®

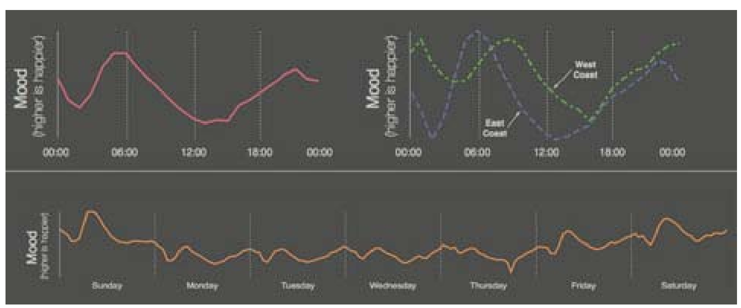
Cornell University

http://www.apple.com/downloads/dashboard/email_messaging/facebookalert.html, www.tibco.com, www.twitter.com

Limitation of Standard Pub/Sub

- Scalable implementations have very simple query languages
  - Simple predicates, comparing message attributes to constants
  - E.g., topic='politics' AND author='J. Doe'
- Many *monitoring applications* need sequence patterns





http://www.ccs.neu.edu/home/amislove/twittermood/, www.jodange.com

Examples

- Stock monitoring
  - Notify me when the price of IBM is above \$83, and the *first* MSFT price afterwards is below \$27.
  - Notify me when the price of any stock increases monotonically for ≥30 min.

Examples

- RSS feed monitoring
  - Once CNN.com posts an article on Technology, send me the first post referencing (i.e., containing a link to) this article from the blogs to which I subscribe

Examples

- System event log monitoring
  - In the past 60 seconds, has the number of failed logins (security logs) increased by more than 5? (break-in attempt)
  - Have there been any failed connections in the past 15 minutes? If yes, is the rate increasing?

Solutions?

- Traditional pub/sub
  - Scalable, but not expressive enough
- Database Management System (DBMS)
  - Static datasets, one-shot queries
- Data Stream Management Systems (DSMS)
  - Limited MQO work
- Active databases (triggers), event processing systems
  - None had all desired features: expressiveness, precise formal semantics, system implementation with scalability in event rate and number of queries

The Main Goal of Cayuga

- Language
  - Expressiveness
    - Filter, project, aggregate, join (correlate) events from multiple streams
  - Precise, formal semantics
    - Fully composable operators with formal semantics
- System
  - Scalability in event rate and number of queries

http://www.cs.cornell.edu/bigreddata/cayuga/

Cornell University

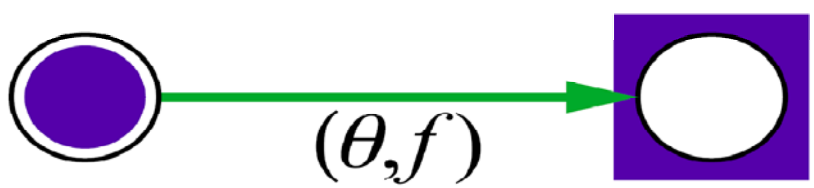
Cayuga Stream Algebra

- Compositional: operators produce new streams from existing ones

- Translation to generalized Nondeterministic Finite Automata
  – Edge transitions on input events
  – Automaton instances carry relevant data from matched events

Approach: Compose Queries Through Operators

- Relational operators (on non-temporal attributes)
  - Selection σ_θ
  - Projection π_X
  - Renaming ρ_f
  - Union $\bigcup$
- Together these give standard pub/sub



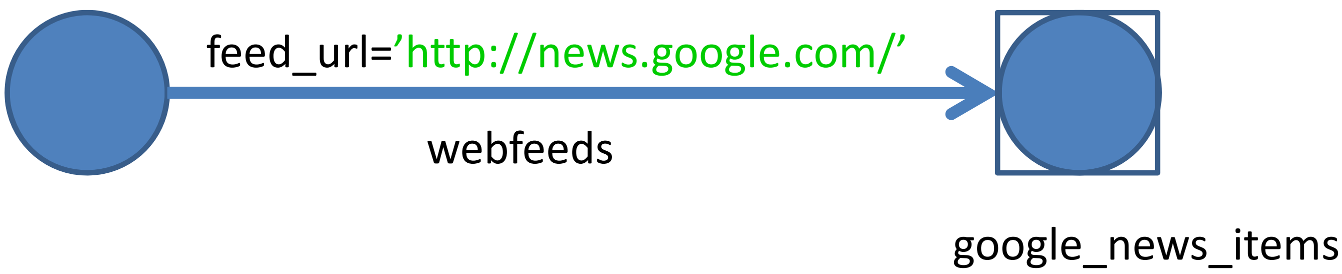
Automaton for $\rho_f \circ \sigma_\theta(S)$

Cornell University

Example Query Q1

- Q1: Find me all RSS items published by Google News

```
SELECT * FROM
    FILTER {feed_url='http://news.google.com/'}(webfeeds)
PUBLISH google_news_items
```



Cornell University

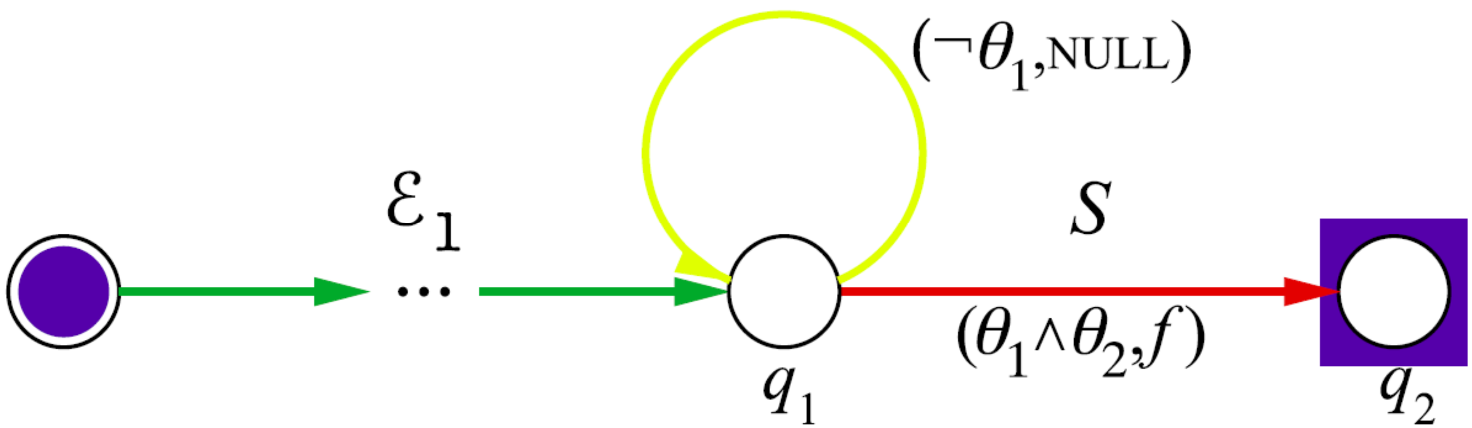
Sequence Operator

- *Sequence* operator $S_1;_\theta S_2$
- After an event from S_1 is detected, match the first event from S_2 that satisfies the condition

Sequence Operator (Contd.)

- Sequencing is a weak join on timestamps
  - Can join an event with one later in future...
  - Or with the immediate successor
    - Can be useful for queries about causal relationships

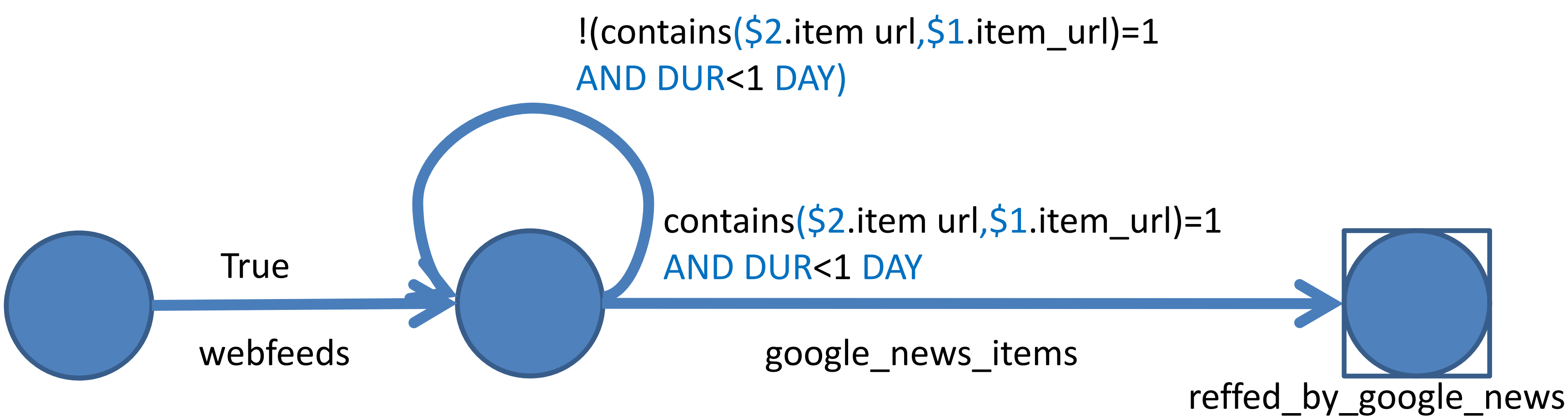
Automaton for $\rho_f \circ \sigma_{\theta_2}(\mathcal{E}_1 ;_{\theta_1} S)$



Example Query Q2

- Q2: Find me all news items that are published by some site, followed by an item from Google referring to it within 1 day.

```
SELECT $2.summary, $1.item_url FROM
    webfeeds
        NEXT {contains($2.item url,$1.item_url)=1 AND DUR<1 DAY}
    google_news_items
PUBLISH reffed_by_google_news
```



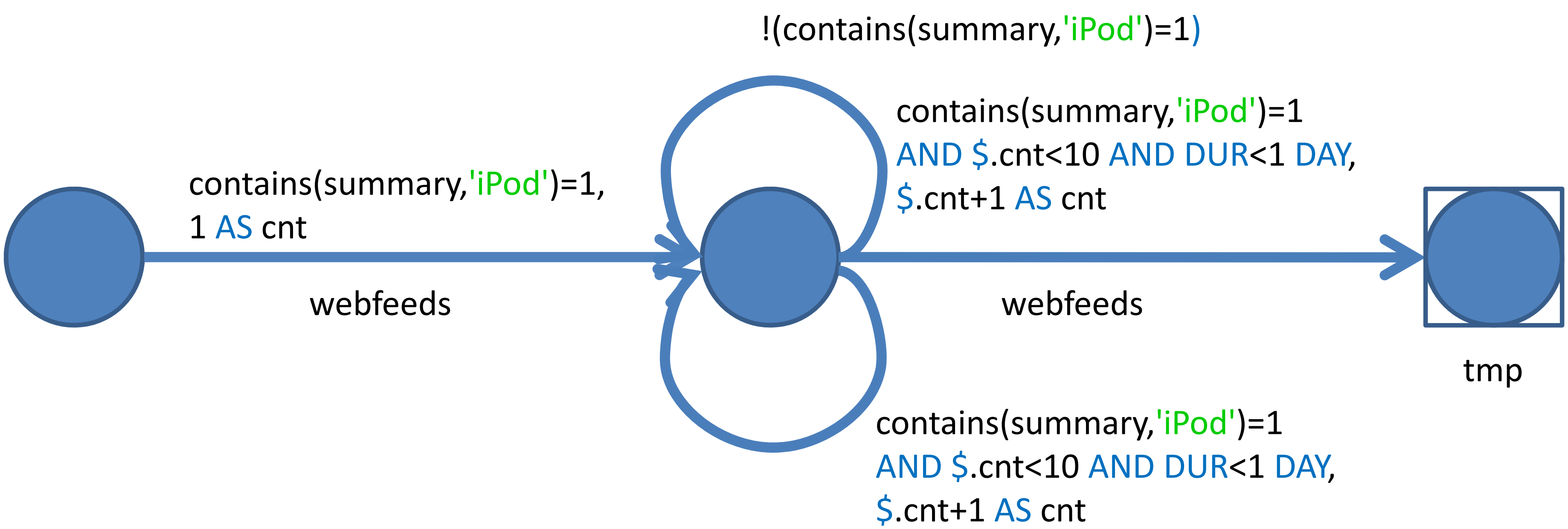
Cornell University

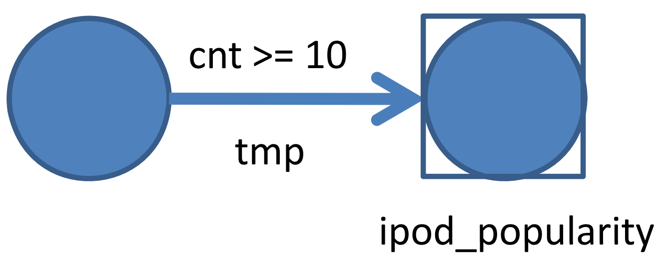
Example Query Q3

- Q3: Notify me when the word iPod has been mentioned by at least 10 articles in the last 1 day

```
SELECT * FROM
  FILTER {cnt >= 10}(
    (SELECT *, 1 AS cnt FROM FILTER{contains(summary,'iPod')=1}(webfeeds))
      FOLD {, $.cnt<10 AND DUR<1 DAY, $.cnt+1 AS cnt}
        (SELECT * FROM FILTER {contains(summary,'iPod')=1}(webfeeds))
  )
PUBLISH ipod_popularity
```

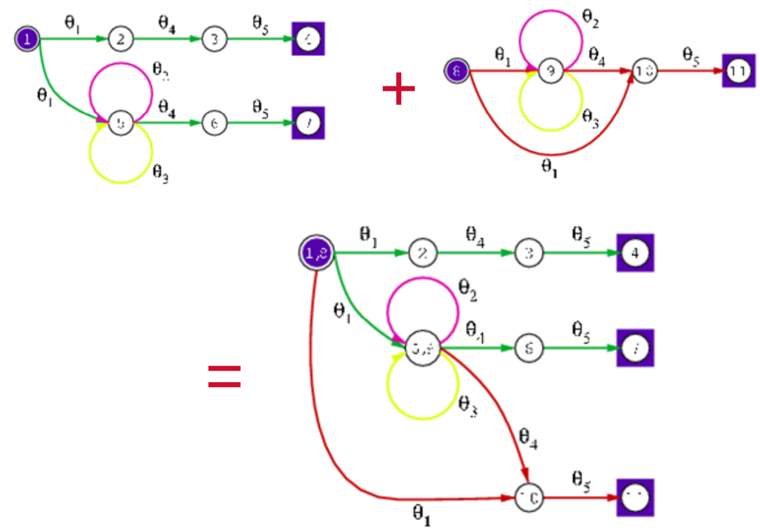
Automata for Q3



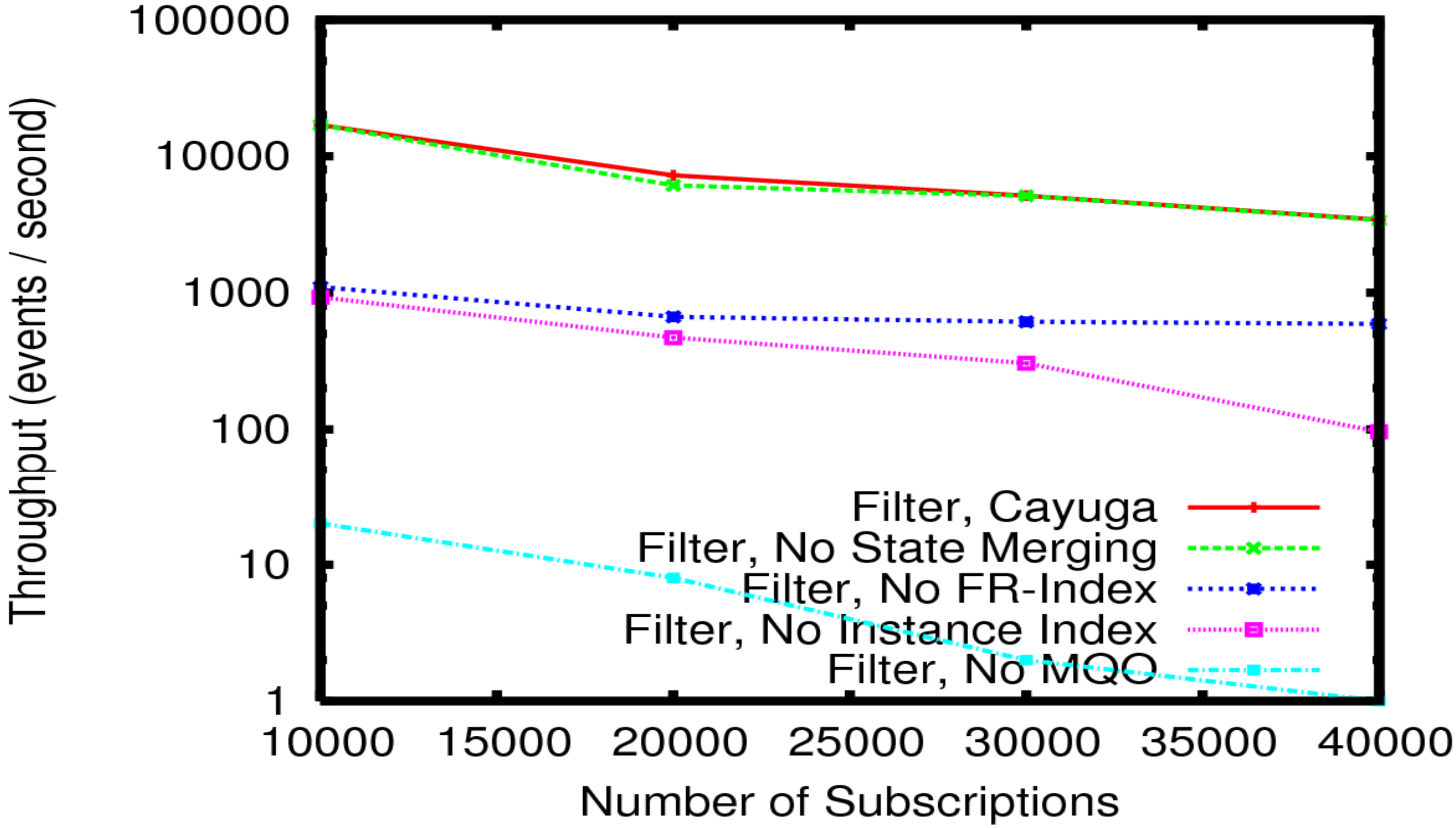


Other Techniques

- We saw: Selection, sequencing, iteration
- Algebra:
  - Aggregation
  - Re-subscription
- Implementation:
  - Automata merging for similar queries
  - Automatic indexing
- Extensions:
  - XML streams
  - Distribution



Sample Performance



Cornell University

More information: http://www.cs.cornell.edu/bigreddata/cayuga/

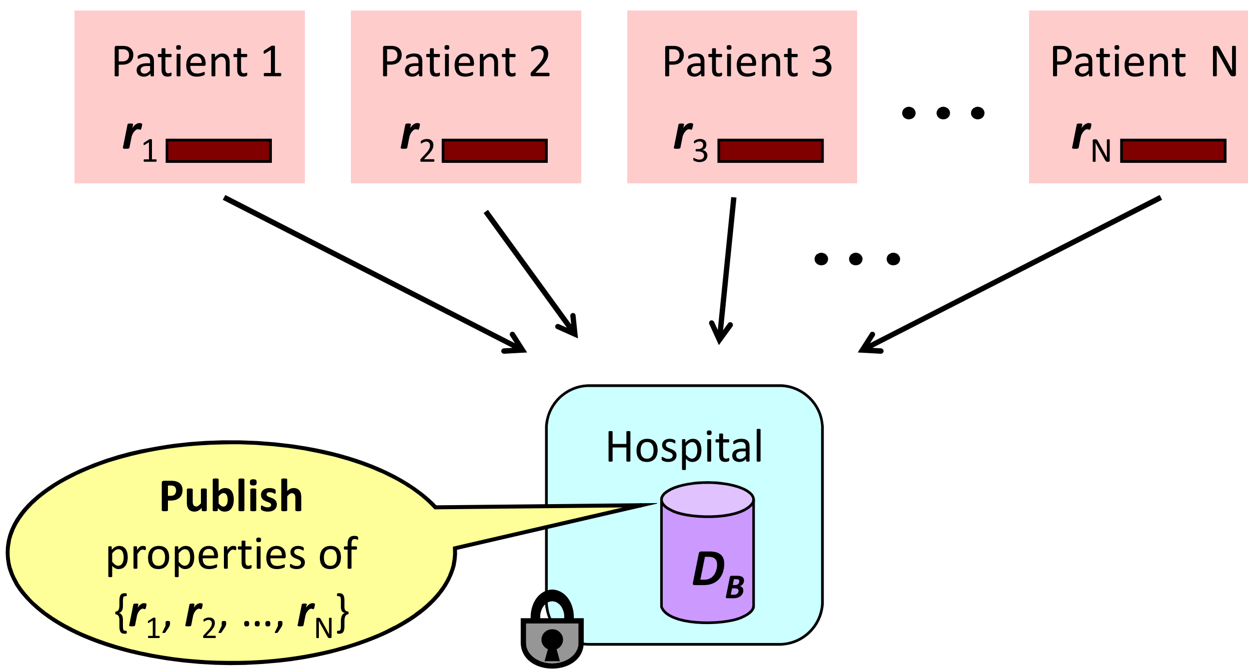
Talk Outline

- Introduction
- Techniques for data stream processing
  - Stream summaries
  - Complex event processing
- Data privacy
- Conclusions

Data Collection Agencies Publish Sensitive Information to Facilitate Research.

Publish information that:

- Discloses as much statistical information as possible.
- Preserves the privacy of the individuals contributing the data.

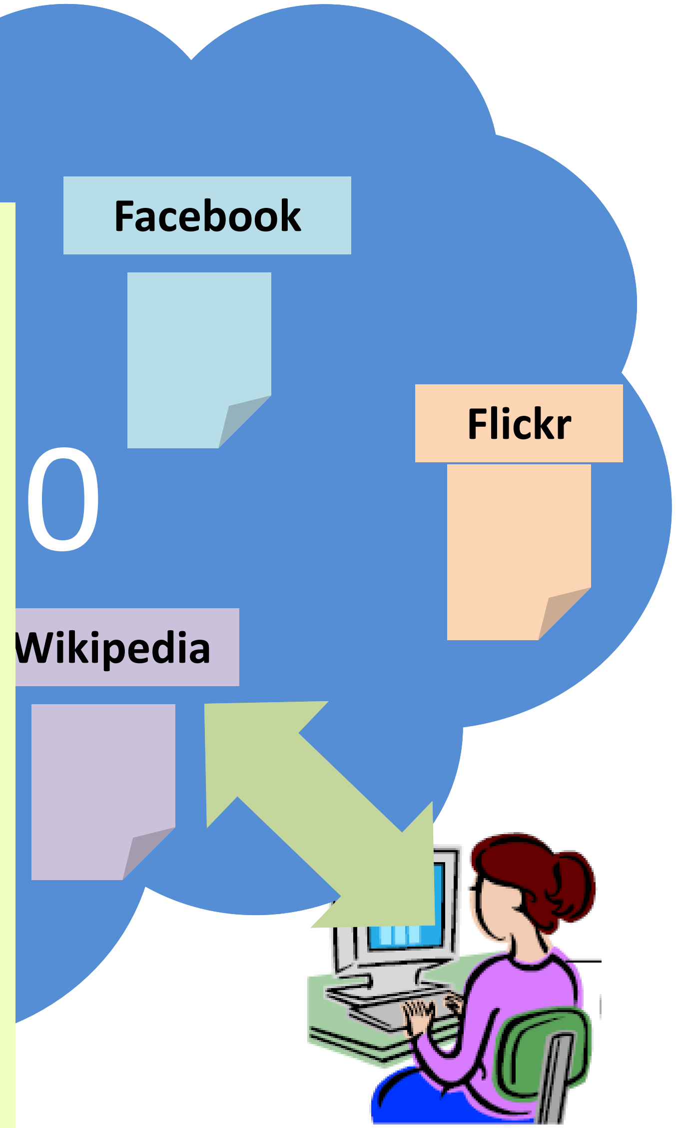


Johannes Gehrke, Daniel Kifer, Ashwin Machanavajjhala:
Privacy in data publishing. ICDE 2010: 1213

Estimated User Data Generated
Per Day:

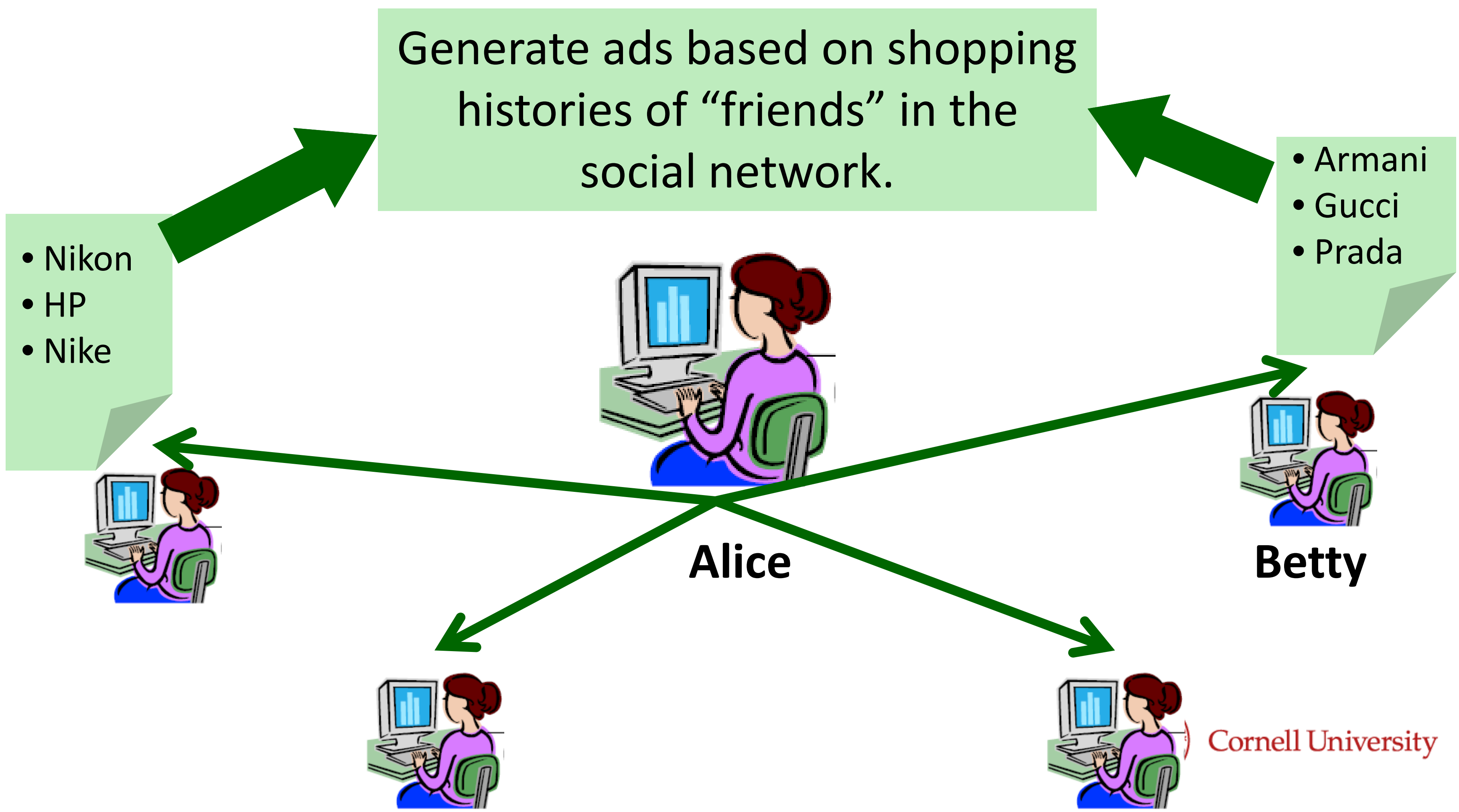
- 8-10 GB public content

- **~4 TB* private content**
  - **Emails**
  - **Instant messages**
  - **Tags/Page Views/Annotations**
  - **Browsing and Shopping histories**
  - **Social Networks …**



Improving Web Experience by Exploiting User Generated Content

Example 1: **Social Advertising**

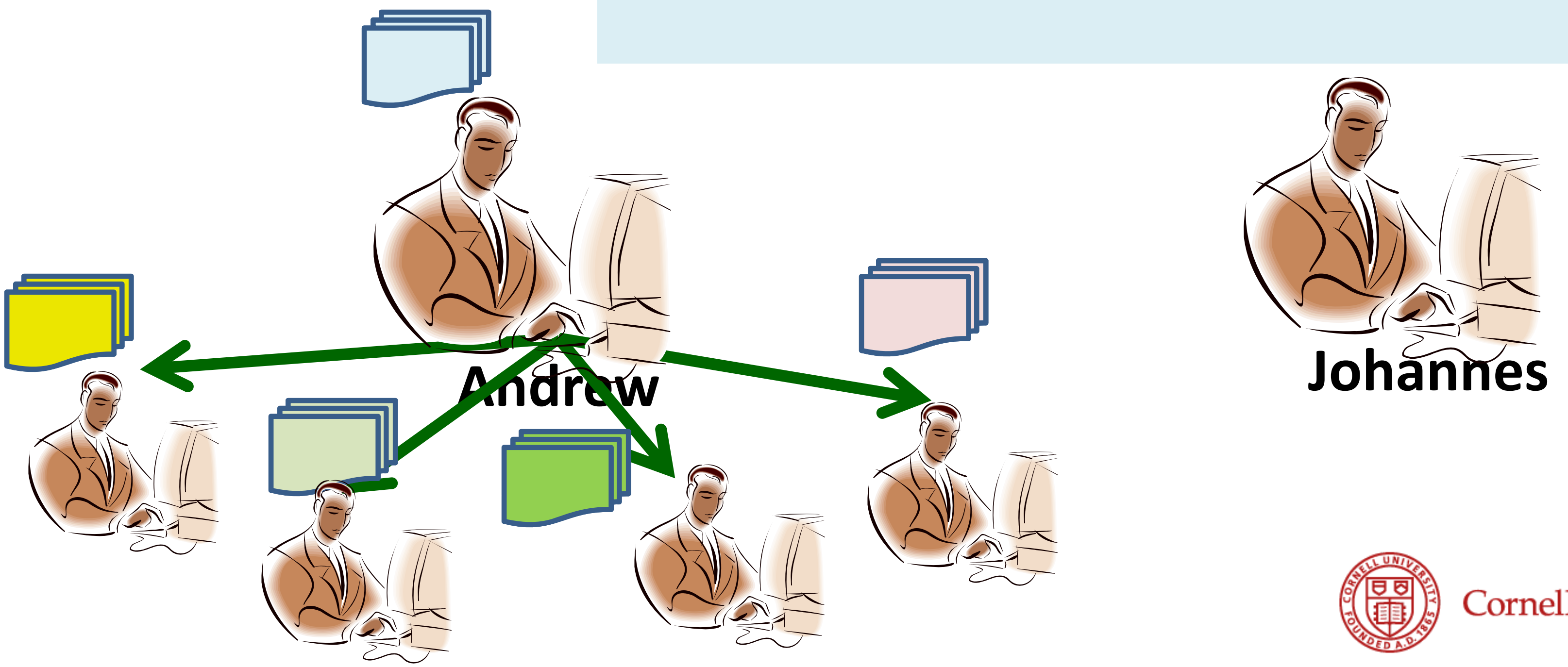


Cornell University

Improving Web Experience by Exploiting User Generated Content

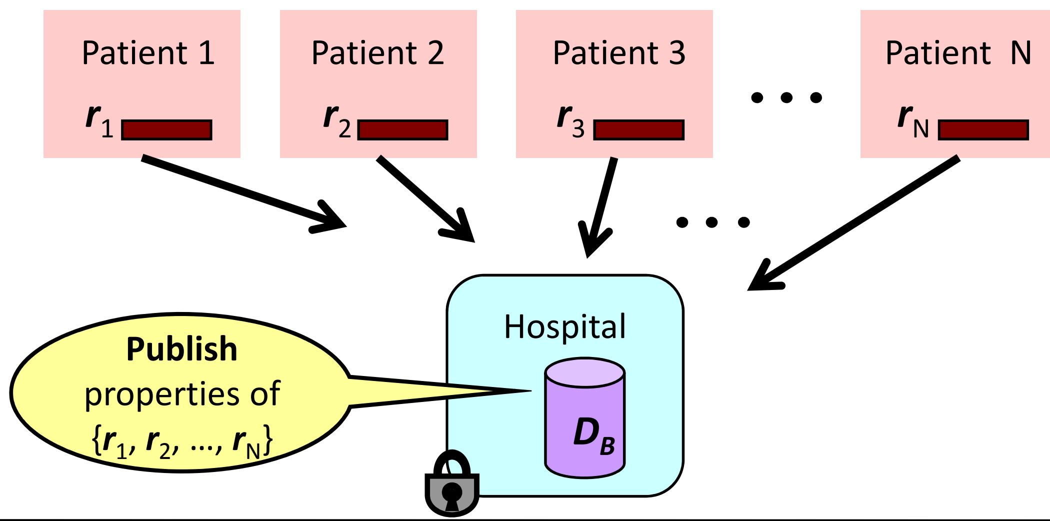
Example 2:
User Targeted Subscriptions

Recommend papers to Johannes based on the papers read by Andrew (and his collaborators/peers).

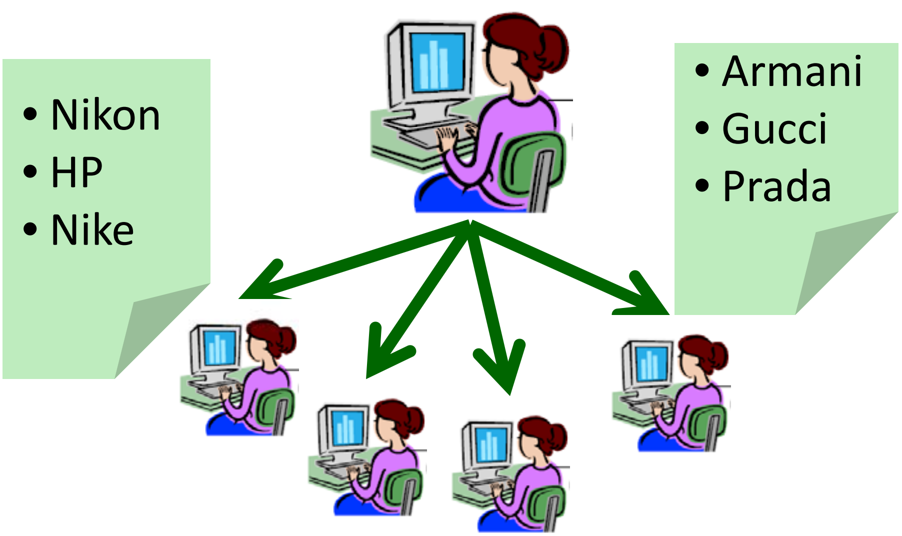


Valuable Information Can be Learned by Sharing Personal Data.

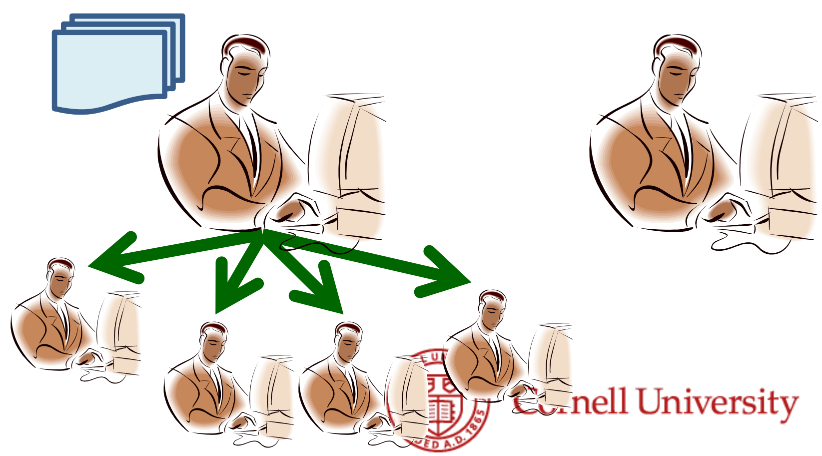
Data Publishing



Social Advertising



User Targeted Subscriptions



What about Privacy?

"… Last week AOL did another stupid thing …
… but, at least it was in the name of science…"

Alternet, August 2006

AOL Data Release …

AOL "anonymously" released a list of 21 million web search queries.

UserIDs were replaced by random numbers …

865712345	Uefa cup
865712345	Uefa champions league
865712345	Champions league
865712345	Champions league final
236712909	exchangeability
236712909	Proof of deFinitti's theorem
127652340	Zombie games
127652340	Warcraft
127652340	Beatles anthology
127652340	Ubuntu breeze
865712345	Grammy nominees
865712345	Amy Winehouse rehab

A Face Is Exposed for AOL Searcher No. 4417749 [New York Times, August 9, 2006]

…

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.

…

A Face Is Exposed for AOL Searcher No. 4417749 [New York Times, August 9, 2006]

Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. “We all have a right to privacy,” she said. “Nobody should have found this all out.”

[In response, she plans to drop her AOL subscription.]

New York Times

What is Privacy?

- *"The claim of individuals, groups, or institutions to determine for themselves* ***when, how and to what extent information about them is communicated to others****"*

  Westin, Privacy and Freedom, 1967

- But we need *quantifiable* notions of privacy …

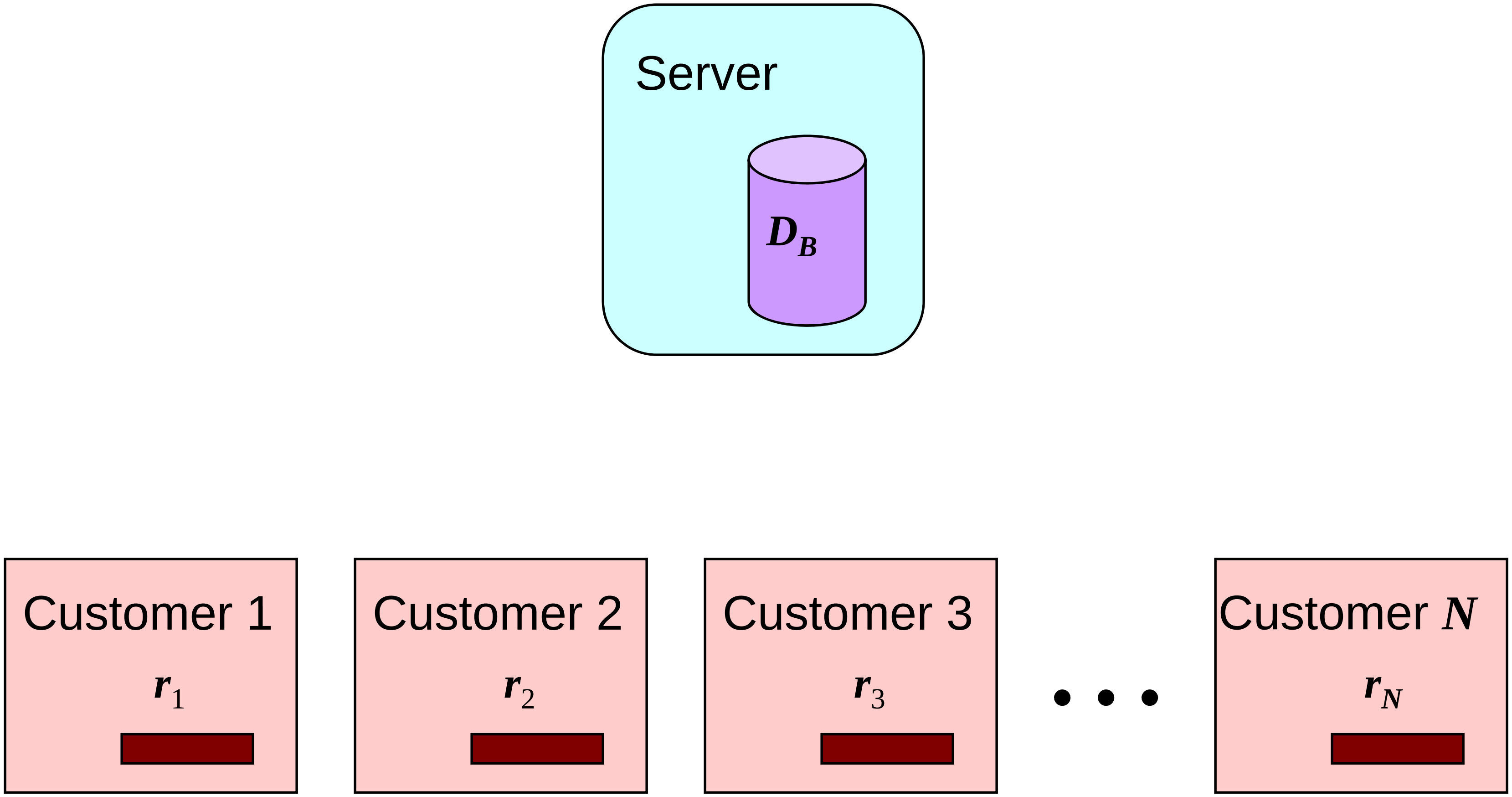
Cornell University

What is Privacy?

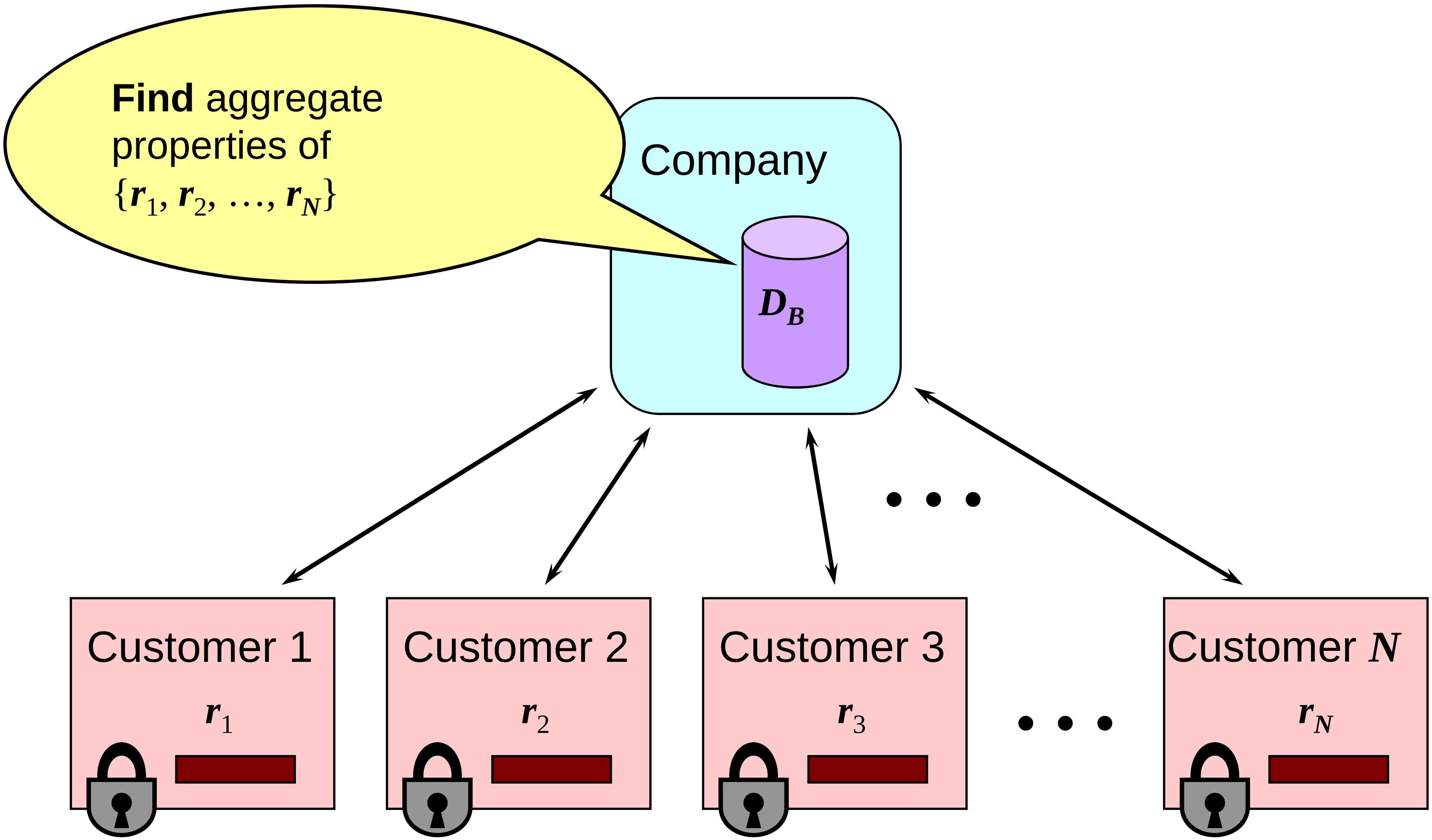
... nothing about an individual should be learnable from the database that cannot be learned without access to the database …

T. Dalenius, 1977

The Setup


Server
D_B
Customer 1
r_1
Customer 2
r_2
Customer 3
r_3
• • •
Customer N
r_N

Model I: Untrusted Data Collector



Cornell University

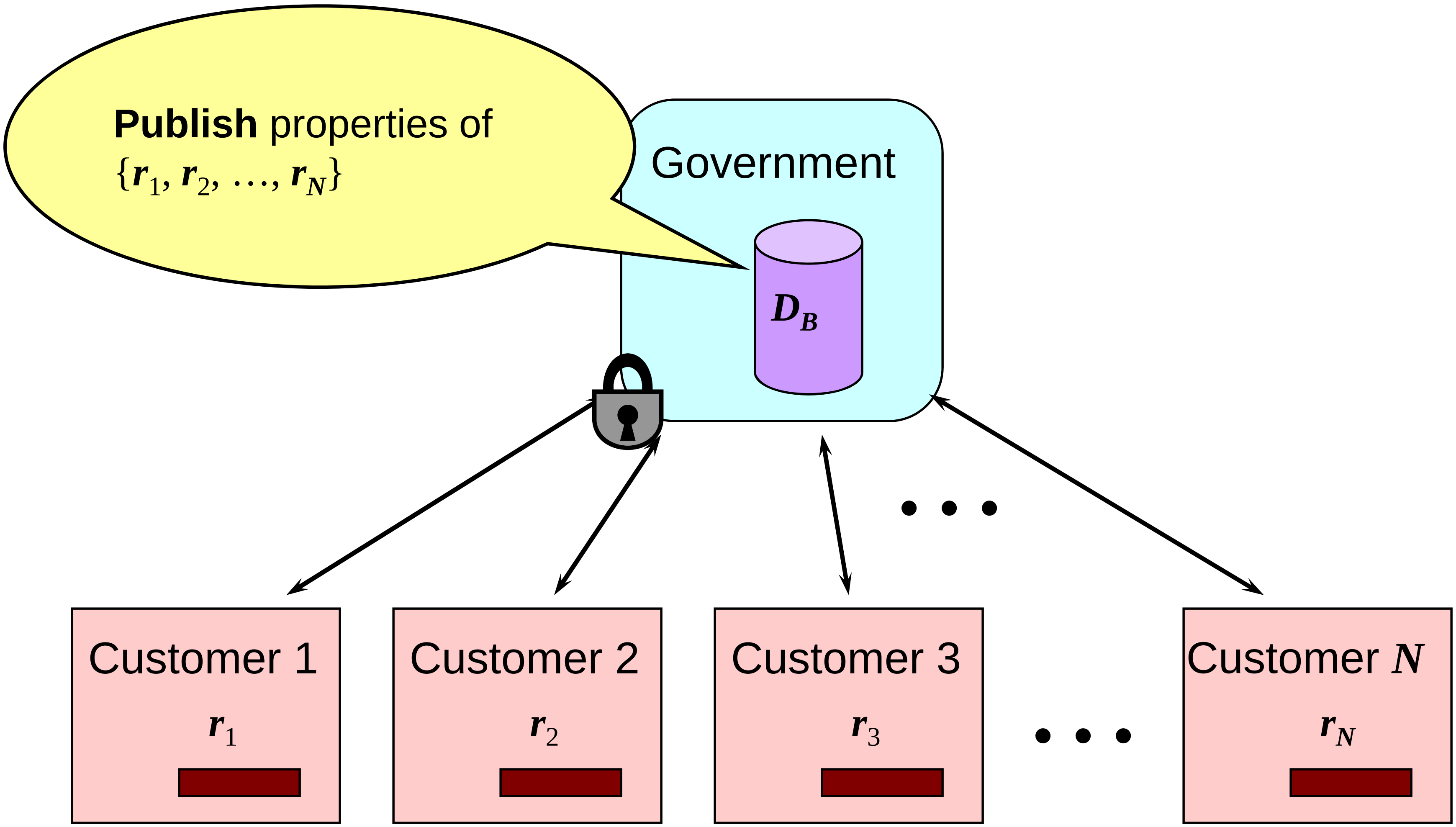
Minimal Information Sharing

- Ideally, we want an algorithm that discloses only the query result, and only to the requesting party. (In practice, we need some extra disclosure.)

- How do we design algorithms that compute queries while preserving data privacy?

- How do we measure privacy (this extra disclosure)?

Model II: Trusted Data Collector



Cornell University

Disclosure Limitations

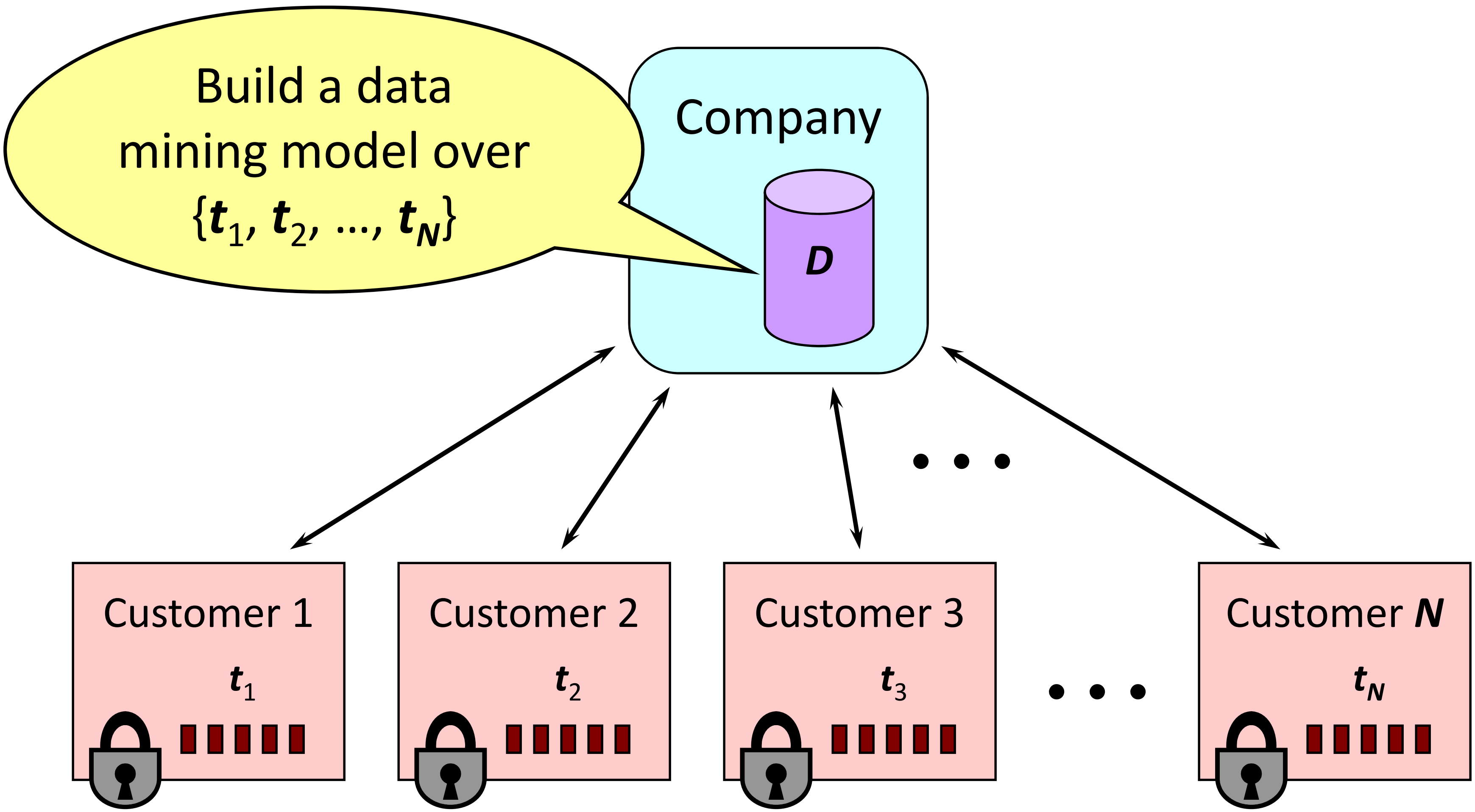
- Ideally, we want a solution that discloses as much statistical information as possible while preserving privacy of the individuals who contributed data.

- How do we design algorithms that allow the "largest" set of queries that can be disclosed while preserving data privacy?

- How do we measure disclosure?

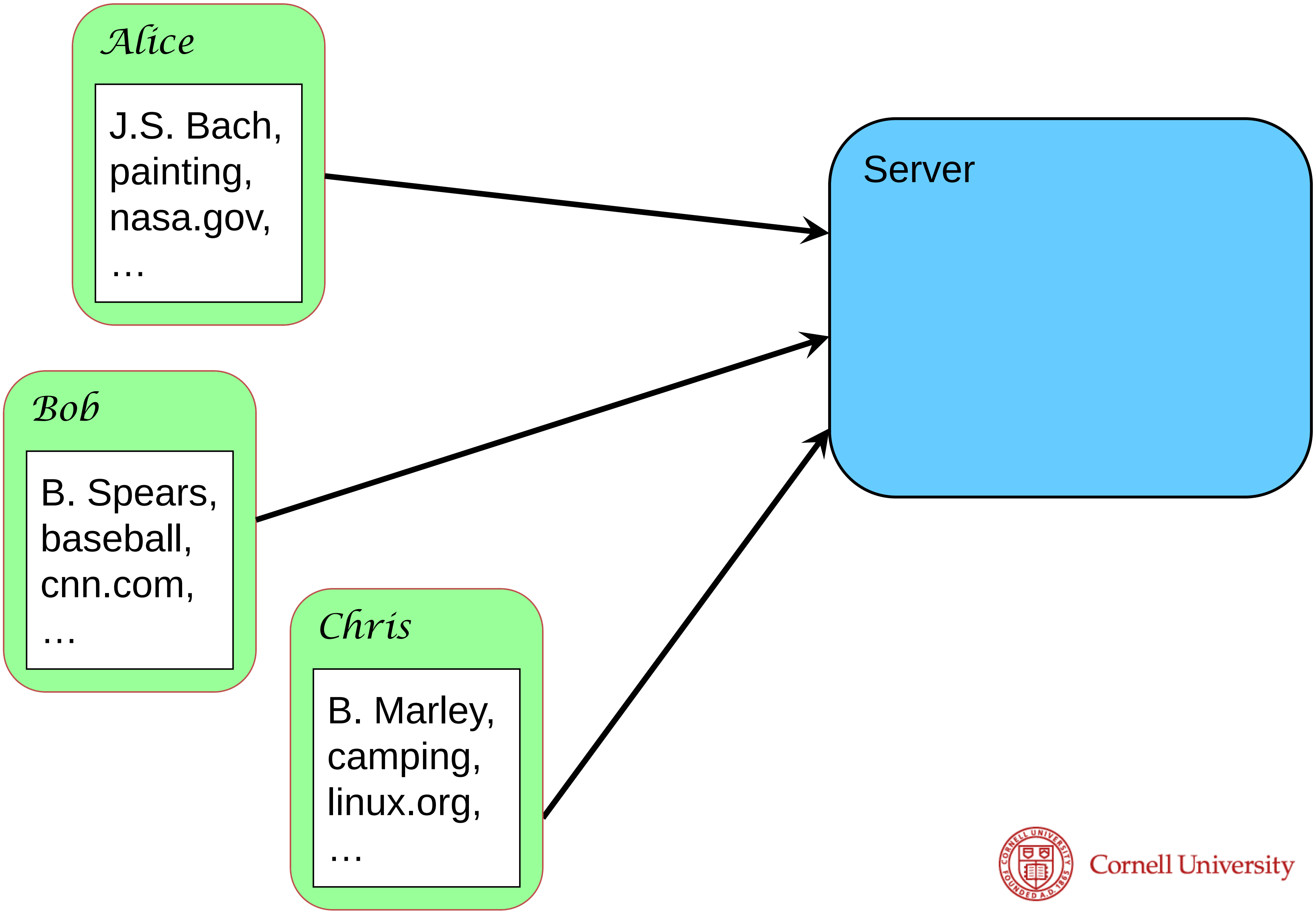
Cornell University

Untrusted Data Collector



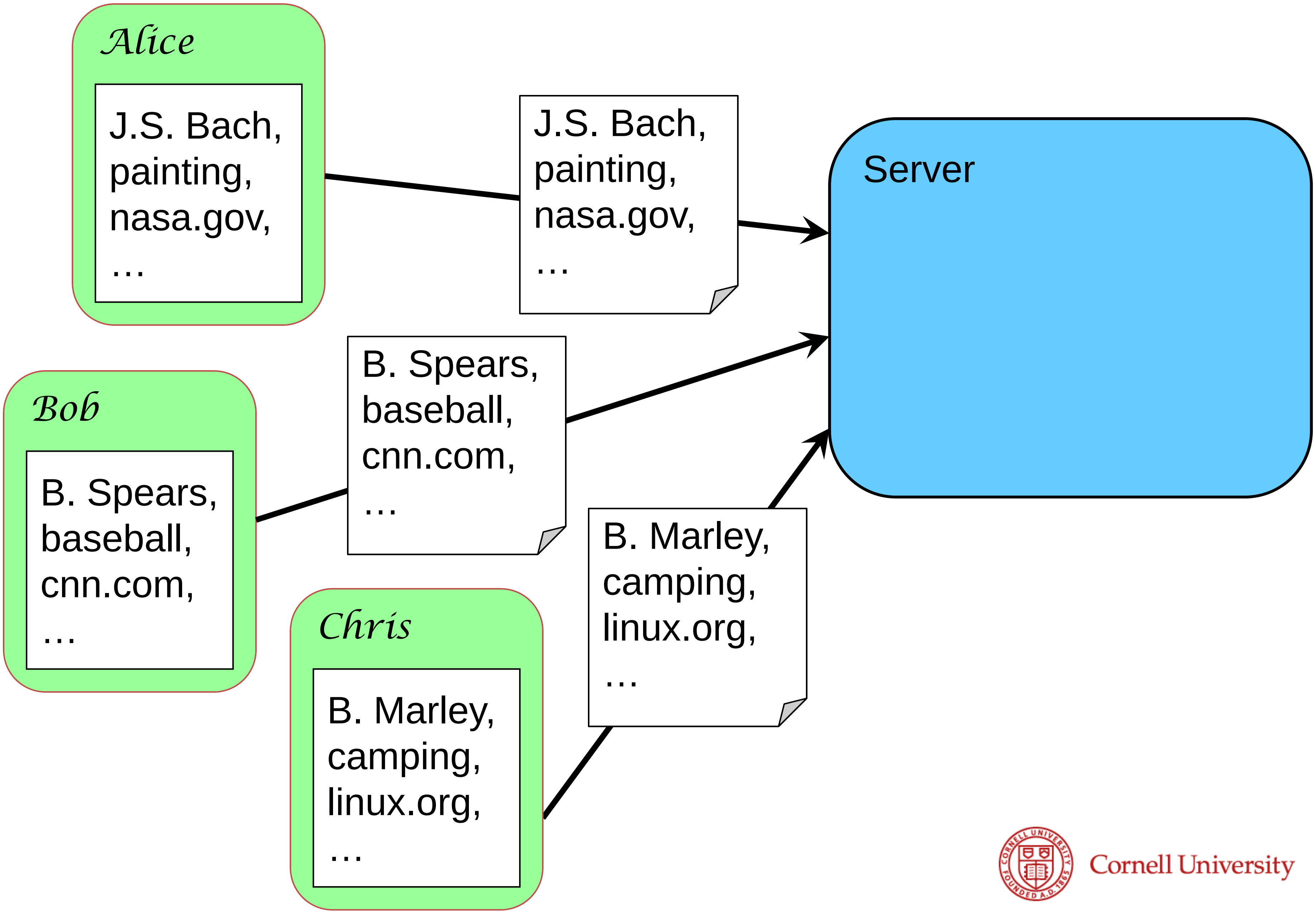
Cornell University

The Model



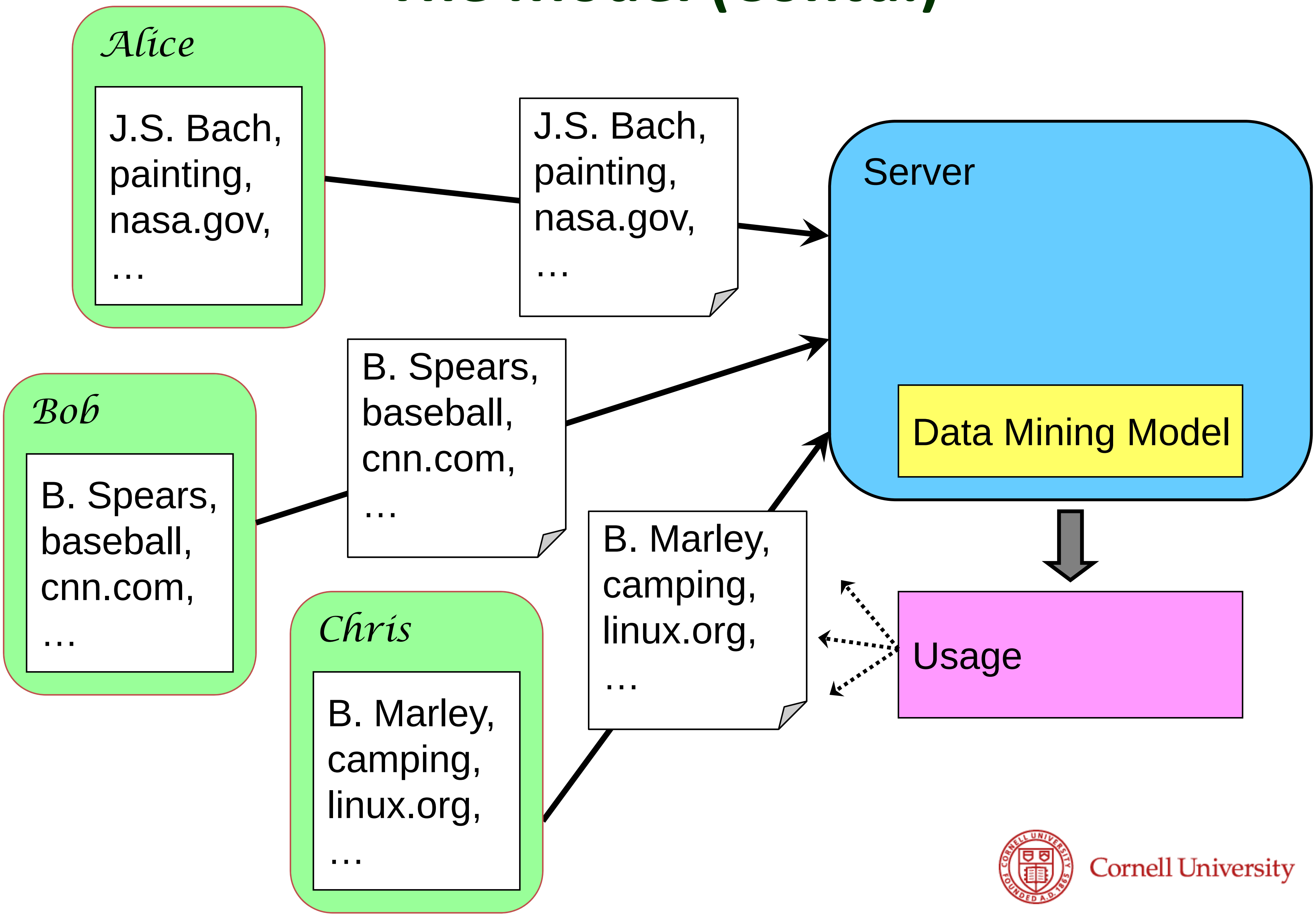
Cornell University

The Model

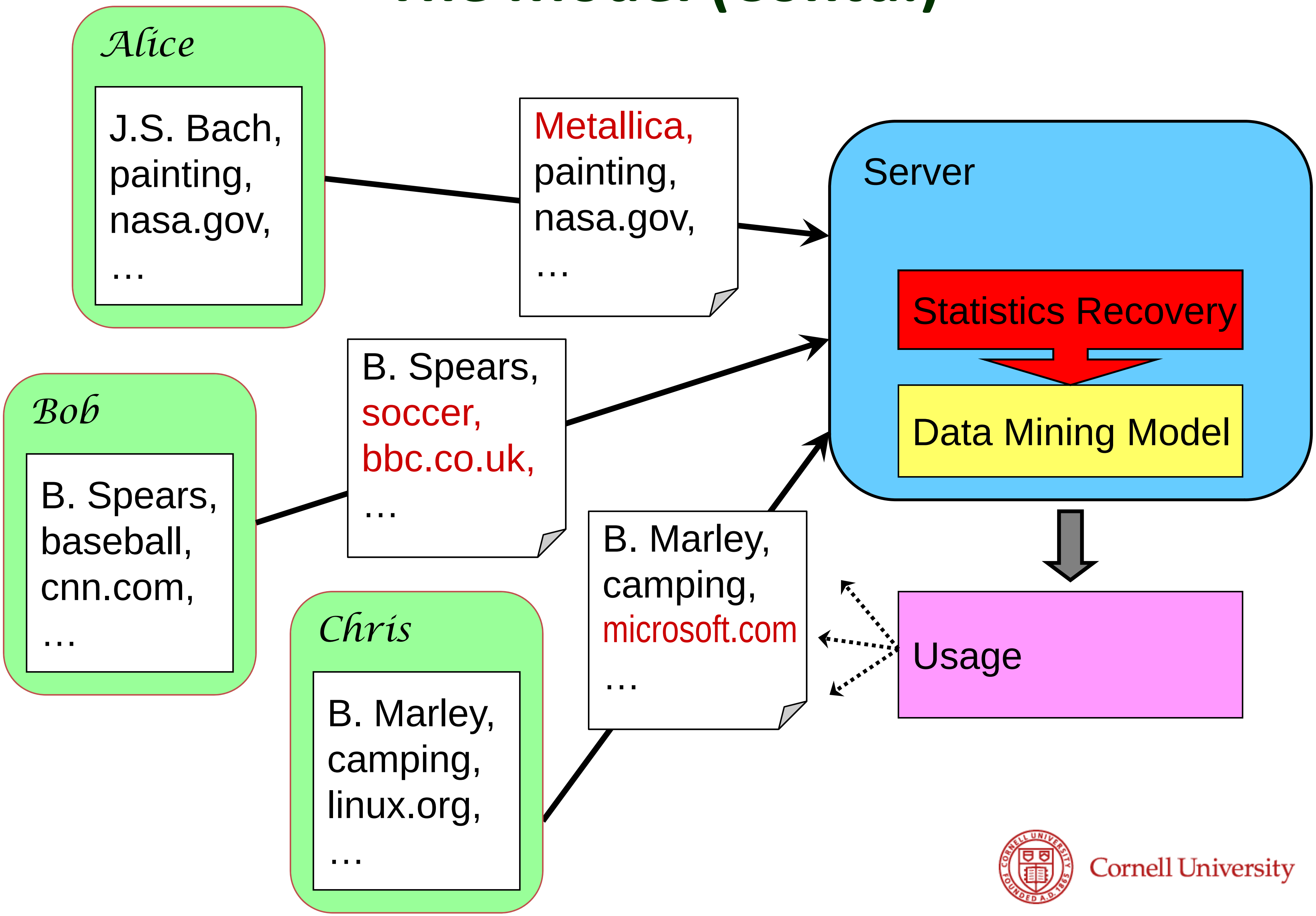


Cornell University

The Model (Contd.)



The Model (Contd.)



Cornell University

Problem

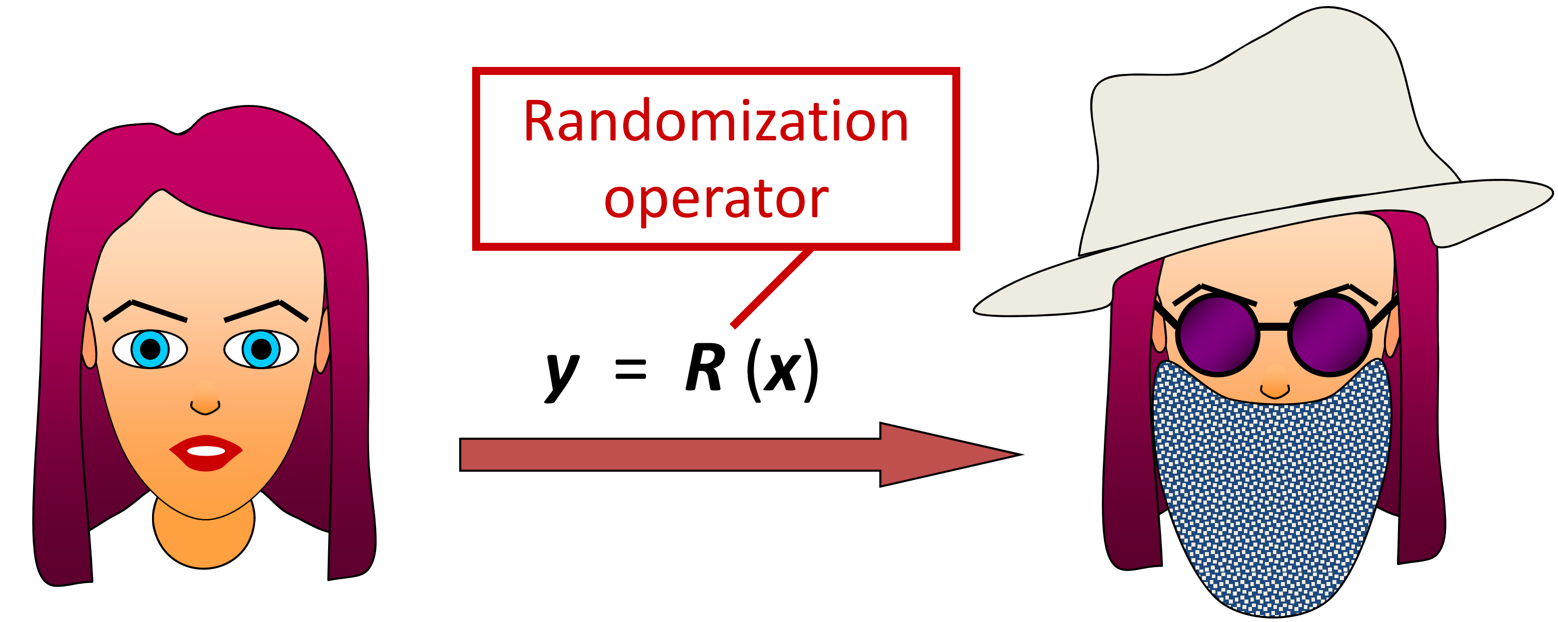
How to randomize the data such that

- We can build a good data mining model (utility)
  - Very simple model: Frequent itemsets (commonly occurring preferences)
- While preserving privacy at the record level (privacy)
  - What does privacy mean?

Motivation: A Social Survey

- Measures opinions, attitudes, behavior
- Problem: Questions of a sensitive nature
  - Examples: sexuality, incriminating questions, embarrassing questions, threatening questions, controversial issues, etc.
  - The “non-cooperative” group leads to errors in surveys and inaccurate data
  - Even though privacy is guaranteed, skepticism prevails

The Model



x

Original (private) data

y

Randomized data

Assumptions:

- Described by a random variable X.
- Each individual client is independent.

Described by a random variable $Y = R(X)$.

Cornell University

The Randomized Response Model

[Stanley Warner; JASA 1965]

- Respondents are given:
  1. A source of randomness (a biased coin)
  2. A statement: I am a member of the XYZ party.
- The procedure:
  – Flip the coin, associate Head with Yes, Tail with No
  – Answer YES if coin gives correct answer, answer NO otherwise

Randomized Response (Contd.)

- The procedure:
  - Flip the coin, associate Head with Yes, Tail with No
  - Answer YES if coin gives correct answer, Answer NO otherwise

	Yes	No
Head (Yes)	YES	NO
Tail (No)	NO	YES

Another View: Two Questions

- Respondents are given:
  1. A coin
  2. Two logically opposite statements:
     - S1: I am a member of the XYZ party.
     - S2: I am **not** a member of the XYZ party.

- The procedure:
  – Flip the coin
  – Answer either statement S1 or S2.

Randomized Response (Contd.)

- Version 1
  - Flip the coin, associate Head with Yes, Tail with No
  - Answer YES if coin gives correct answer, answer NO otherwise
- Version 2
  - Two logically opposite statements
  - Answers either statement S1 or S2.

	Yes	No
Head (Yes)	YES	NO
Tail (No)	NO	YES

	Yes	No
Head (S1)	YES	NO
Tail (S2)	NO	YES

Cornell University

Analysis

π = the true probability of property S in the population.

p = the probability that the coin says YES.

Y_i = 1 if the i^{th} respondent says 'yes'.

0 if the i^{th} respondent reports 'no'.

- $P(Y_i=1) = \pi p + (1-\pi)(1-p) = p_{YES}$
- $P(Y_i=0) = (1-\pi)p + \pi(1-p) = p_{NO}$

	Yes	No
Head	YES	NO
Tail	NO	YES

Cornell University

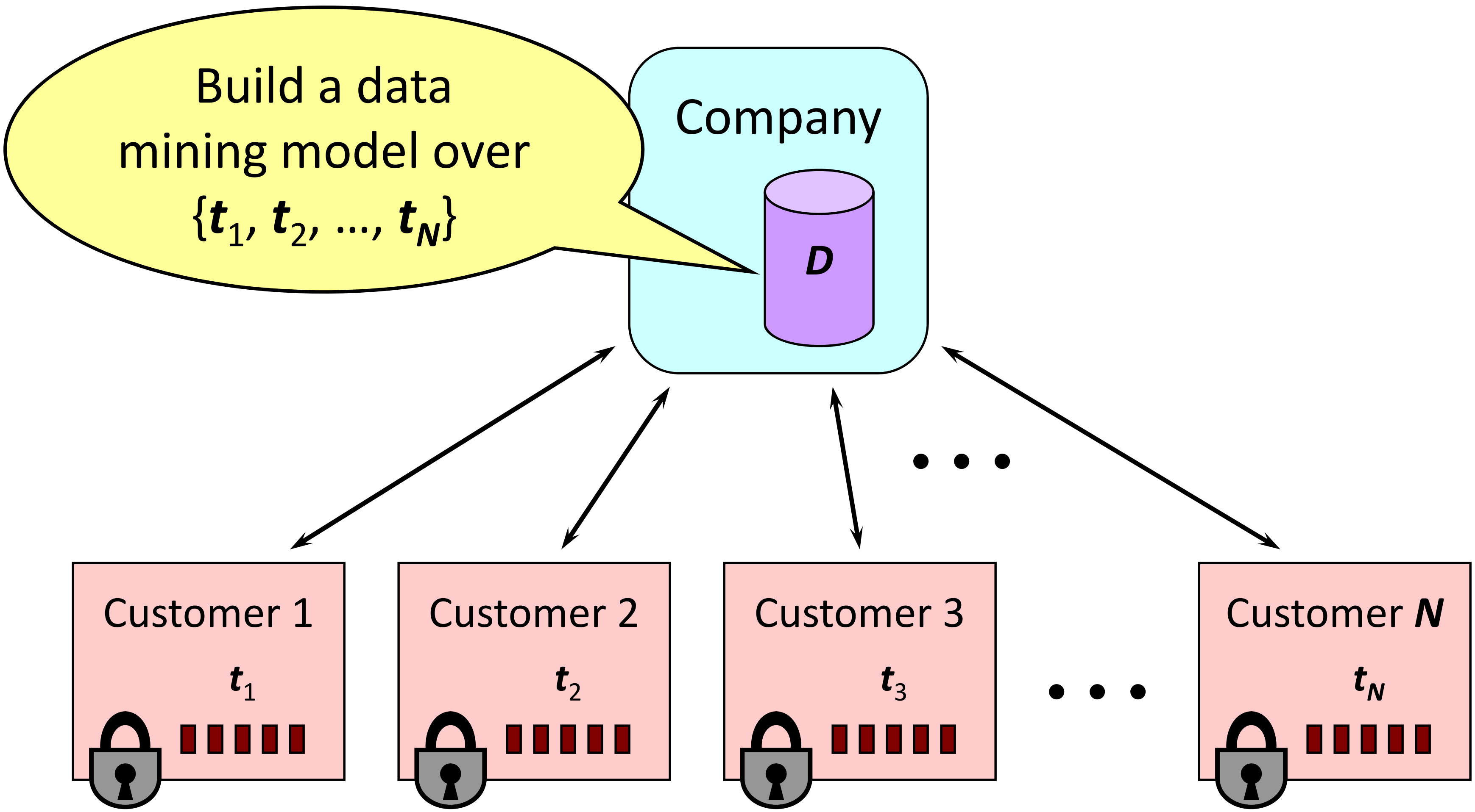
Analysis (Contd.)

- Assume a sample with n records
  - n1 say YES, (n-n1) say NO
- Likelihood of this sample:
  - $L = p_{YES}^{n1}\ p_{NO}^{(n-n1)}$
    (Note: L is a function of π, p, n, n1)
  - This gives a maximum likelihood estimate for π of
    $\pi^{hat} = (p-1)/(2p-1) + n1/n(2p-1)$
- Easy to show:
  - $E(\pi^{hat}) = \pi$
  - $Var(\pi^{hat}) = \pi(1-\pi)]/n + [1/[16(p-0.5)^2]-0.25]/n$

    Variance = Sampling + Coin Flips
- But what type of "privacy guarantees" does randomized response provide?

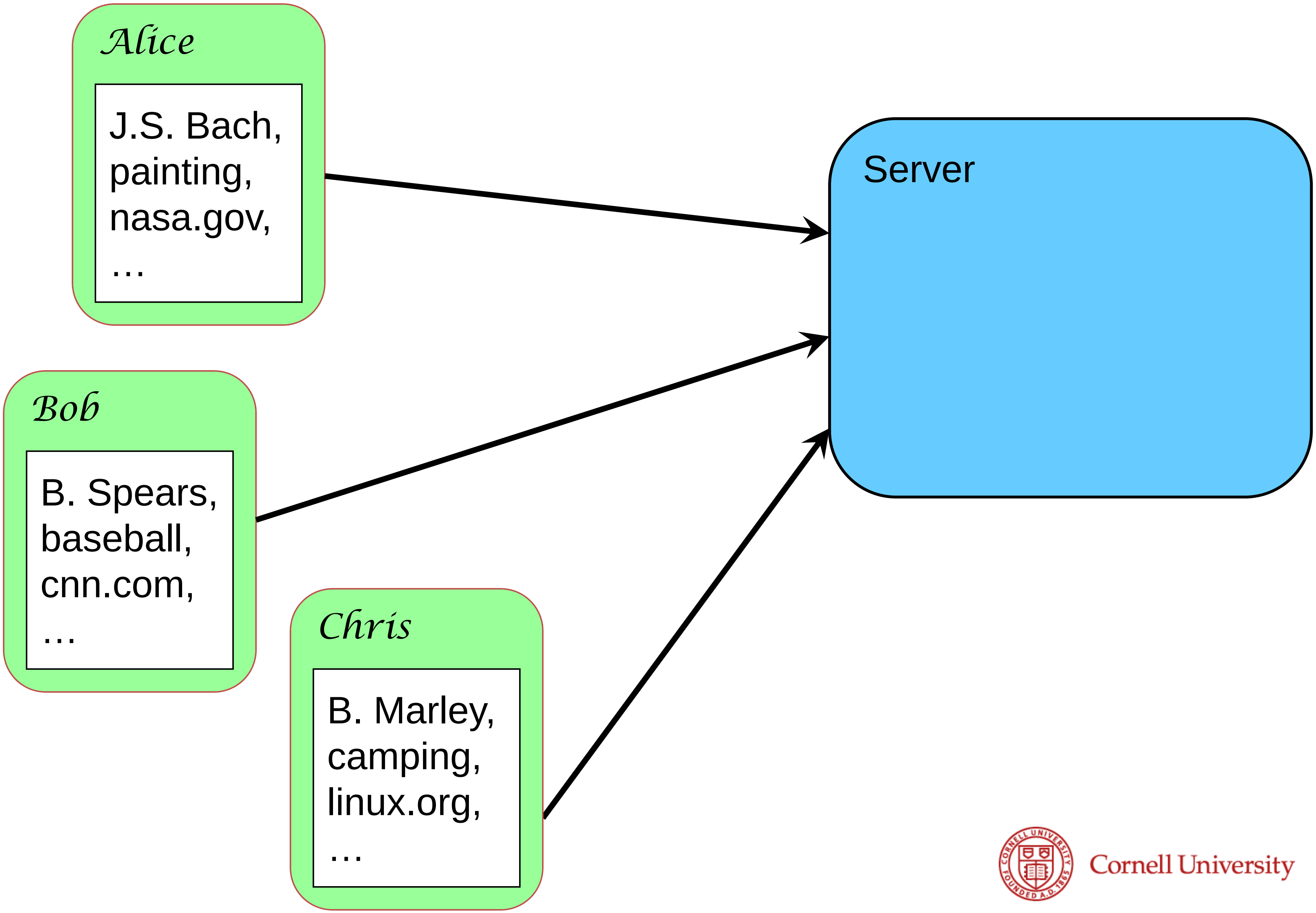
Cornell University

Untrusted Data Collector



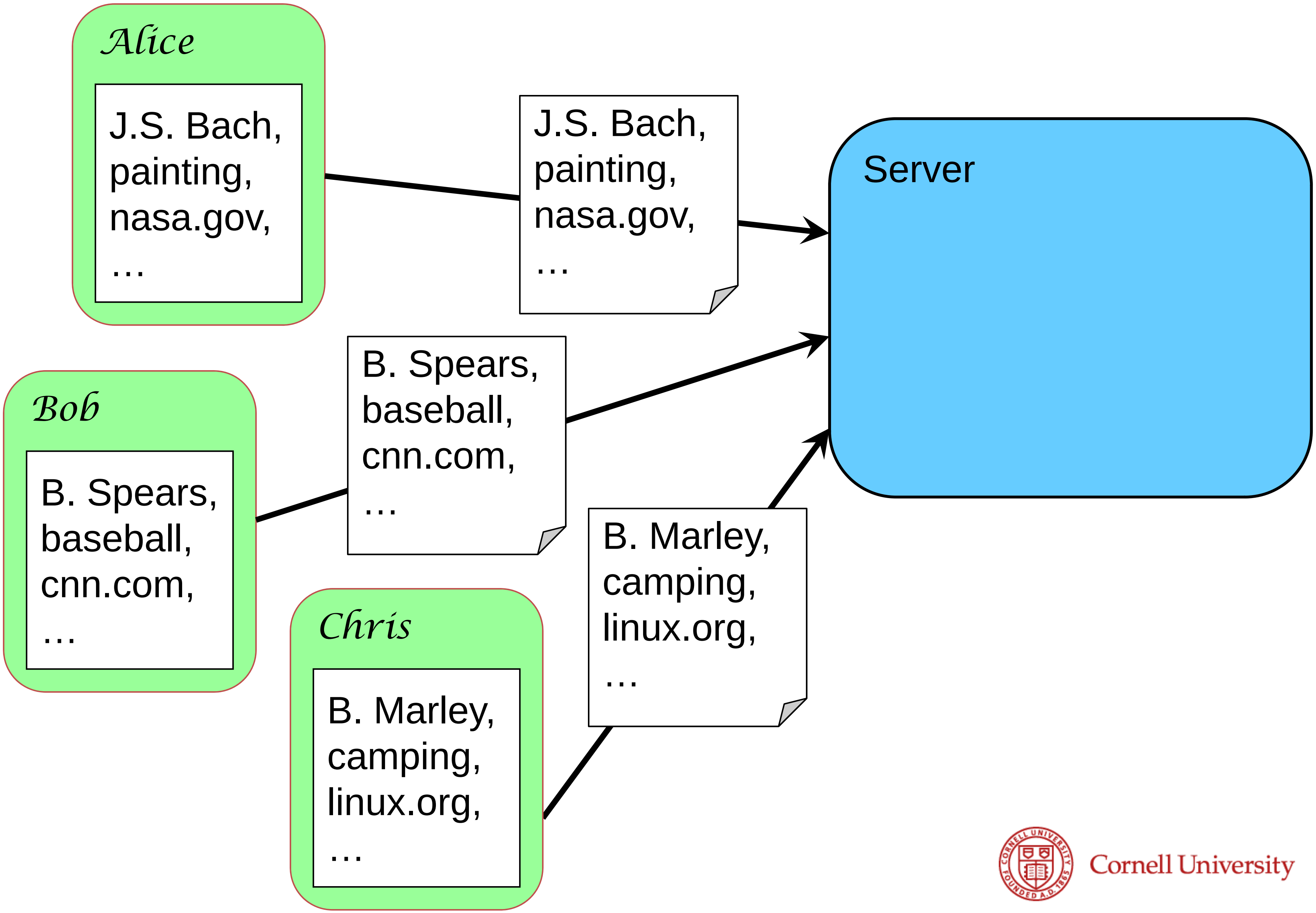
Cornell University

The Model

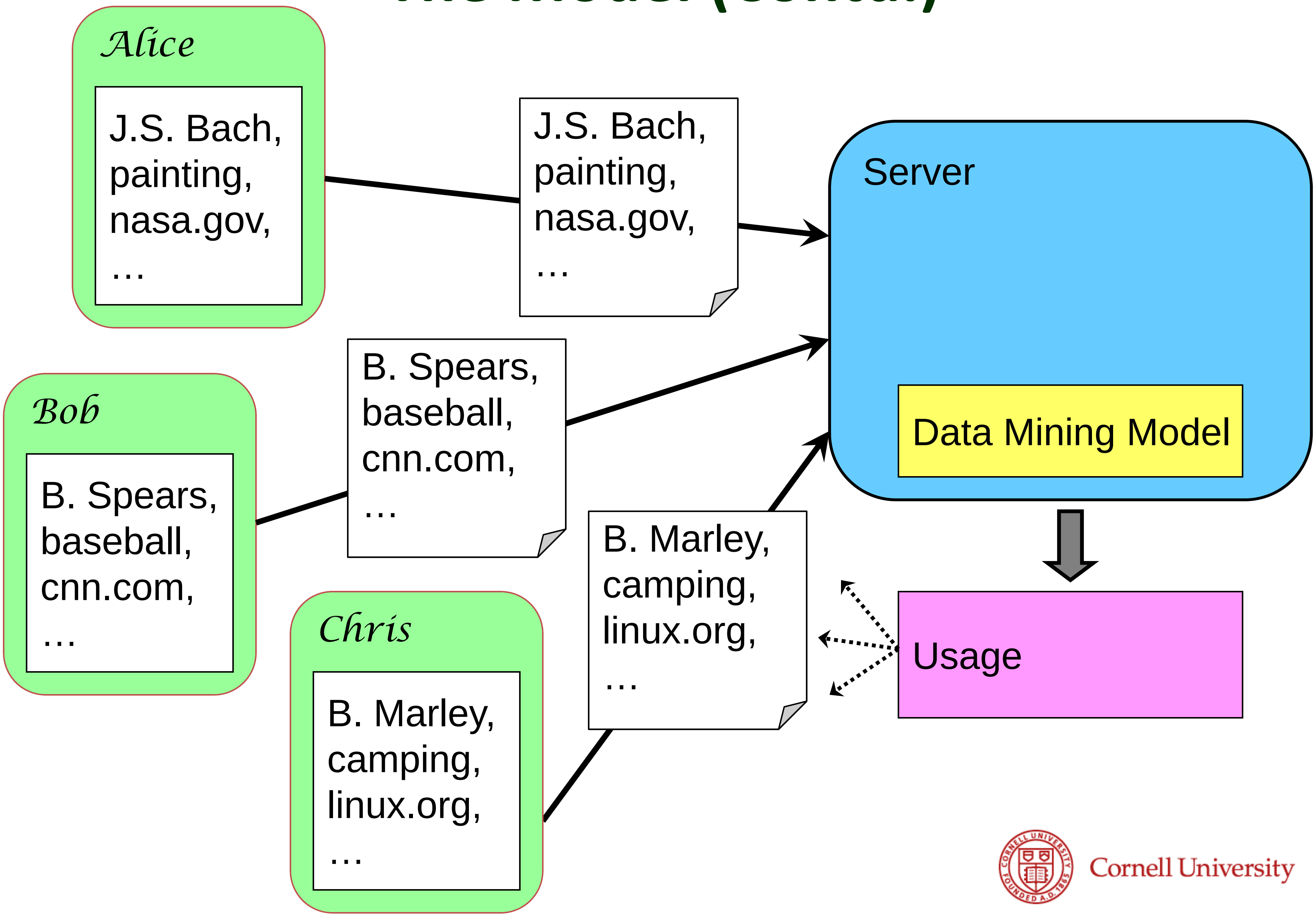


Cornell University

The Model

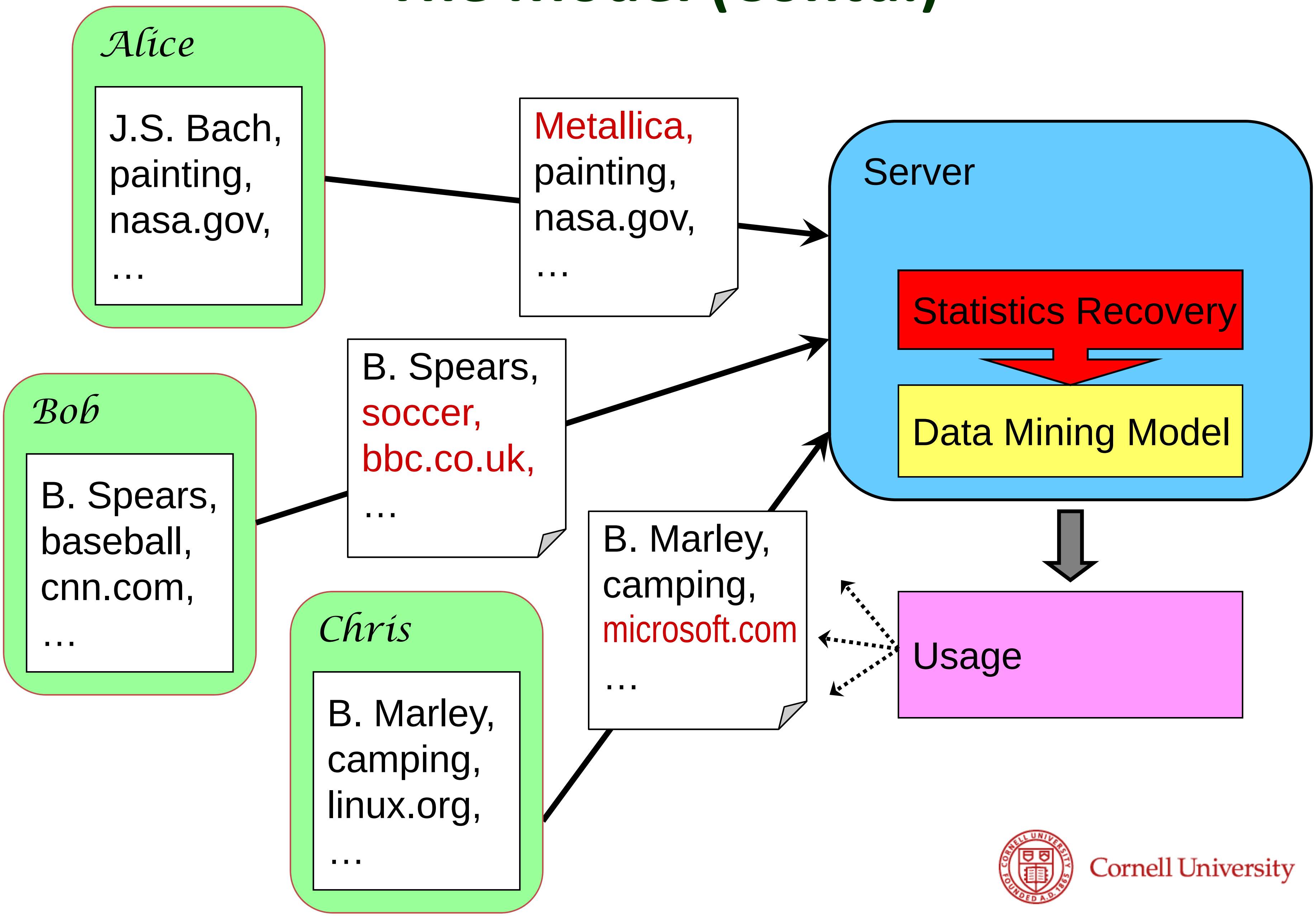

Alice
J.S. Bach,
painting,
nasa.gov,
…
J.S. Bach,
painting,
nasa.gov,
…
Server
Bob
B. Spears,
baseball,
cnn.com,
…
B. Spears,
baseball,
cnn.com,
…
Chris
B. Marley,
camping,
linux.org,
…
B. Marley,
camping,
linux.org,
…

The Model (Contd.)



Cornell University

The Model (Contd.)



Cornell University

Randomized Response Revisited

Return to our recommendation service. A "randomized response"-style algorithm:

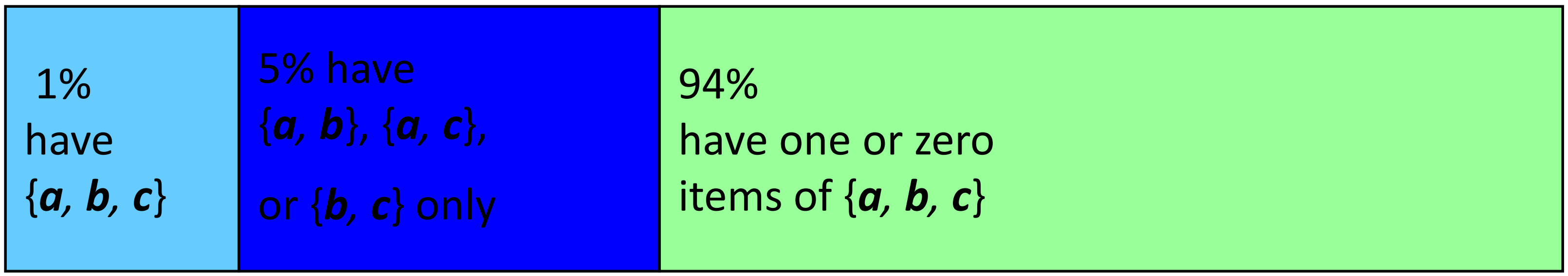
Given a set of preferences:

- Keep (preference) item with 20% probability,
- Replace with a new random item with 80% probability.

Cornell University

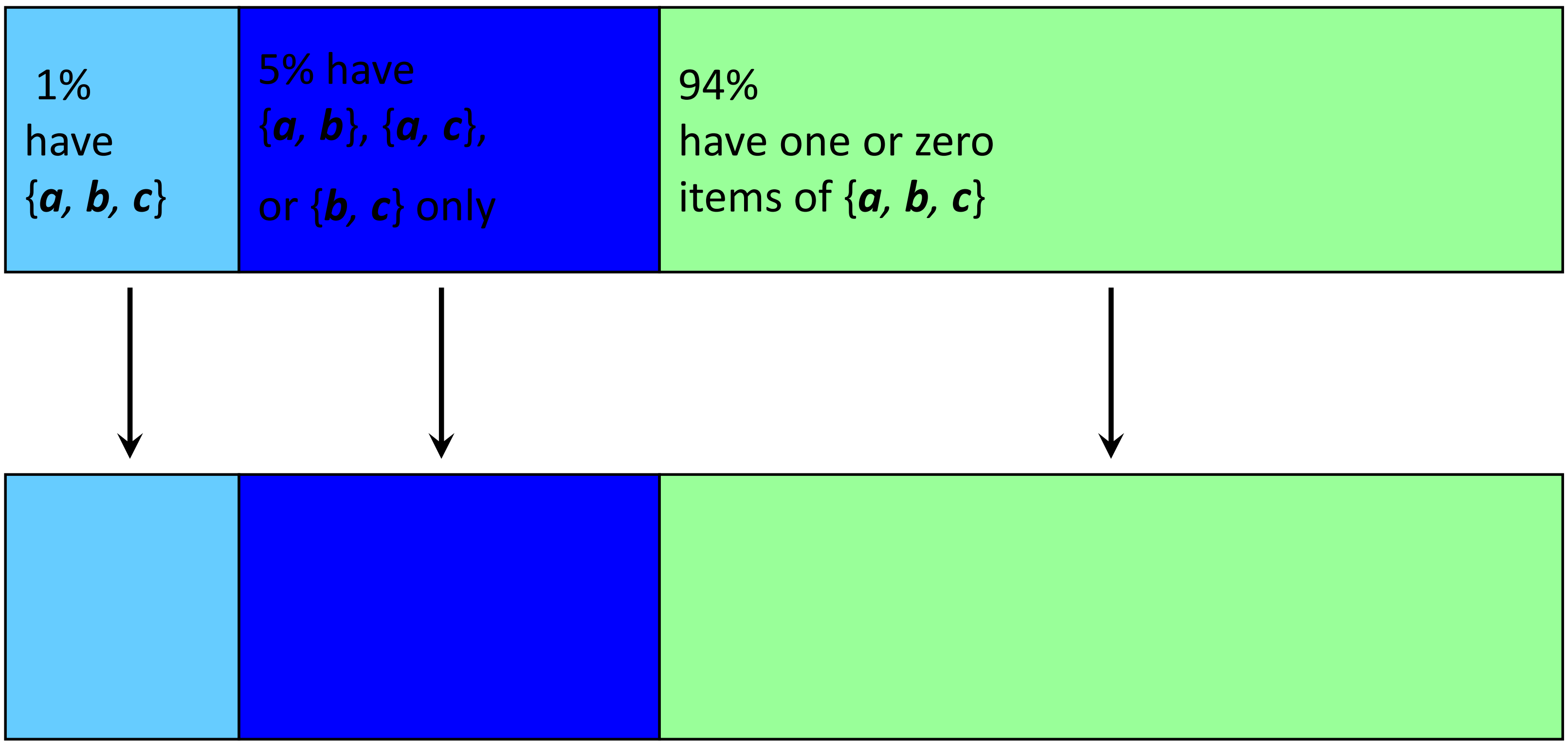
Example: {a, b, c}

10 M transactions of size 10 with 10 K items:



Example: {a, b, c}

10 M transactions of size 10 with 10 K items:



After randomization: How many have {*a, b, c*} ?

Cornell University

Example: {a, b, c}

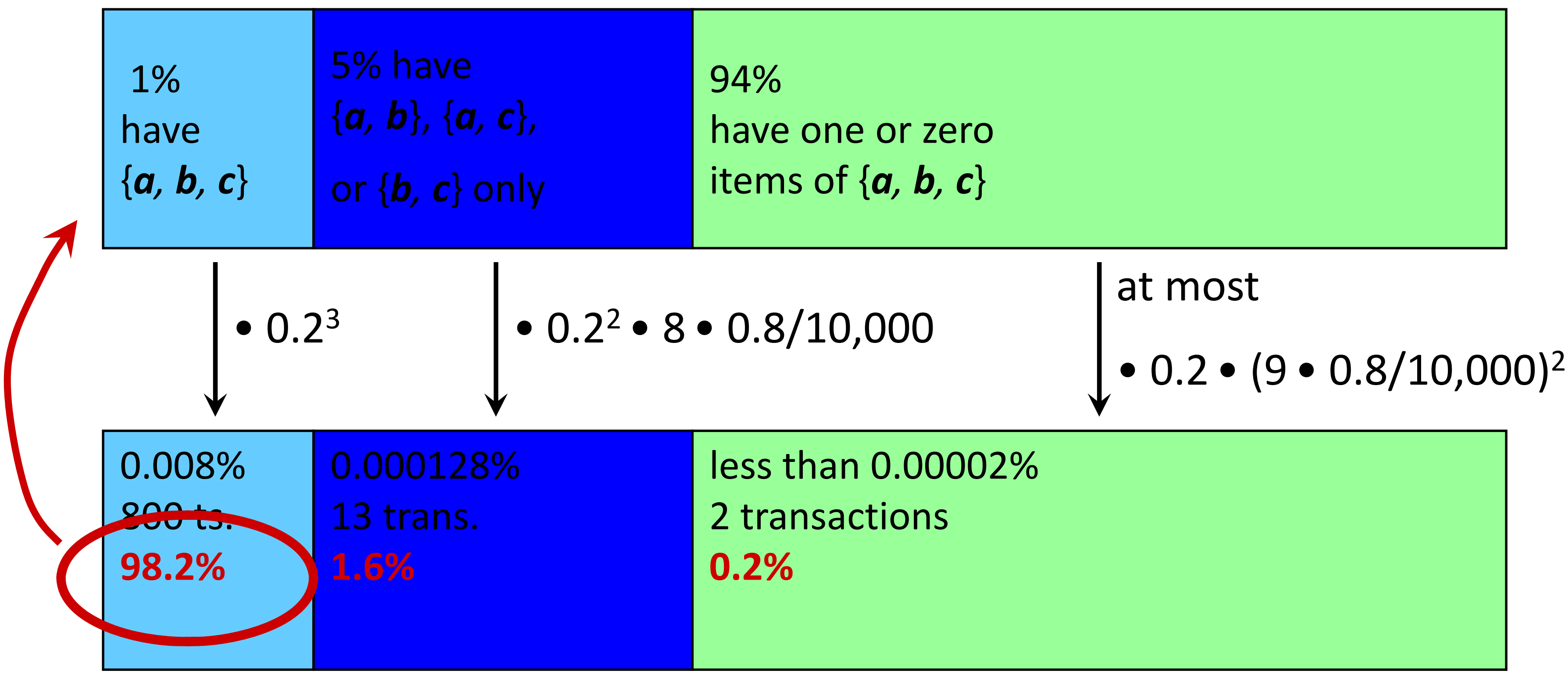
10 M transactions of size 10 with 10 K items:

1% have {***a, b, c***}	5% have {***a, b***}, {***a, c***}, or {***b, c***} only	94% have one or zero items of {***a, b, c***}
↓ $\bullet\ 0.2^3$	↓ $\bullet\ 0.2^2 \bullet 8 \bullet 0.8/10{,}000$	↓ at most $\bullet\ 0.2 \bullet (9 \bullet 0.8/10{,}000)^2$
0.008% 800 ts.	0.000128% 13 trans.	less than 0.00002% 2 transactions

After randomization: How many have {***a, b, c***} ?

Example: {a, b, c}

10 M transactions of size 10 with 10 K items:



After randomization: How many have {***a, b, c***} ?

Example: {a, b, c}

- A-priori, we only know with 1% probability that {a, b, c} occurs in the original transaction

- Given {a, b, c} in the randomized transaction, we have about 98% certainty that {a, b, c} occurred in the original transaction.

- This is called a privacy breach.

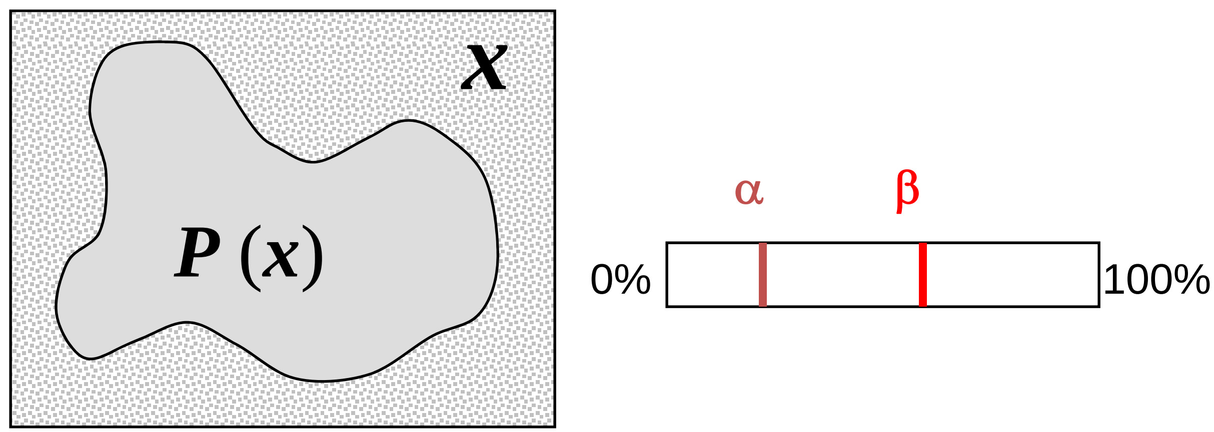
- The example randomization preserves privacy "on average," but not "in the worst case."

Cornell University

α-to-β Privacy Breach

Let ***P*** (***x***) be any property of client's private data;

Let $0 < \alpha < \beta < 1$ be two probability thresholds.



Example:

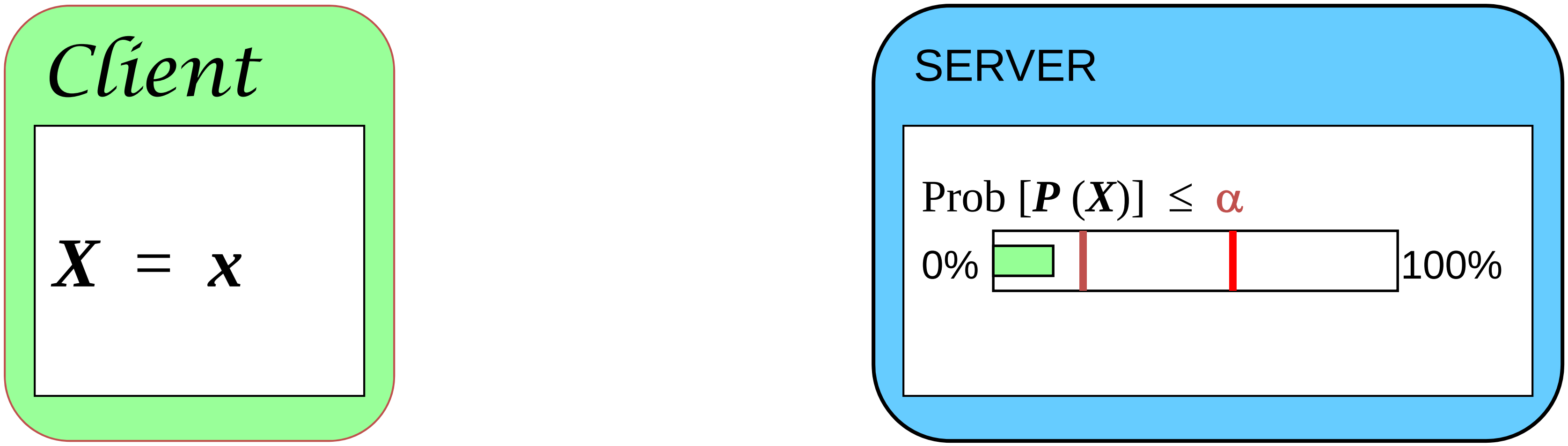
P (***x***) = "transaction ***x*** contains {***a***, ***b***, ***c***}"

α = 1% and β = 50%

α-to-β Privacy Breach

Let ***P*** (***x***) be any property of client's private data;

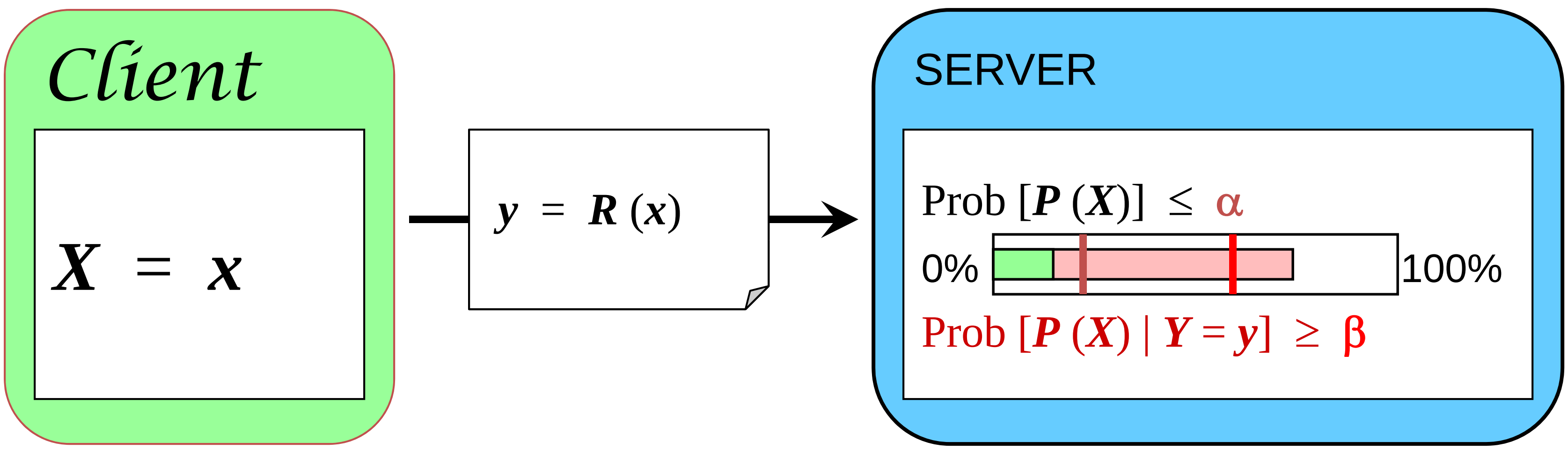
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



α-to-β Privacy Breach

Let $\boldsymbol{P}(\boldsymbol{x})$ be any property of client's private data;

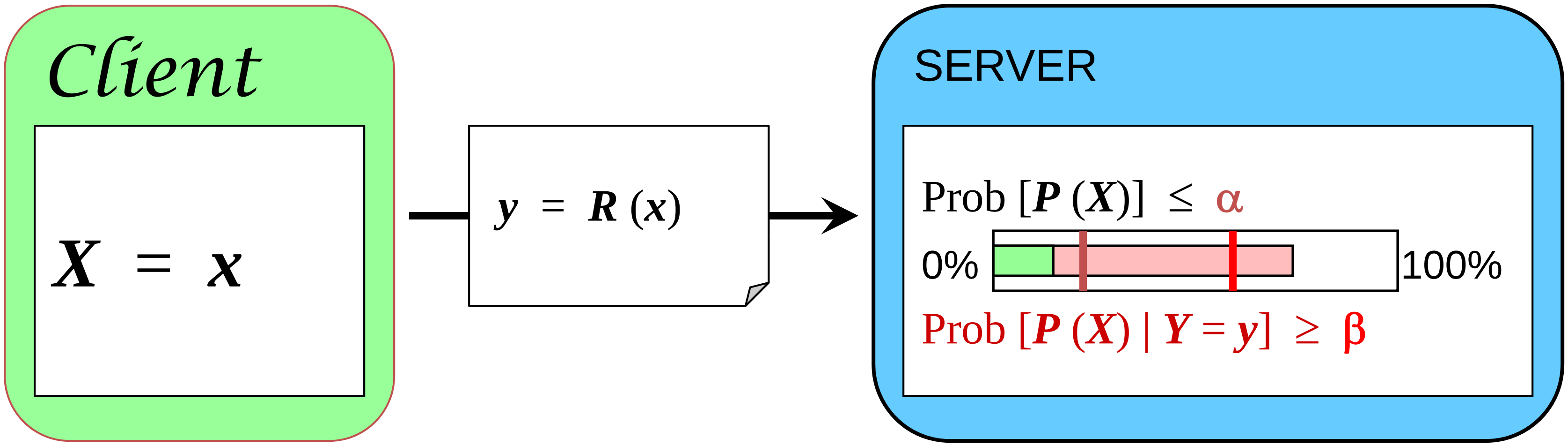
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;

Let $0 < \alpha < \beta < 1$ be two probability thresholds.



Disclosure of y causes an α-to-β privacy breach w.r.t. property $P(x)$.

α-to-β Privacy Breach

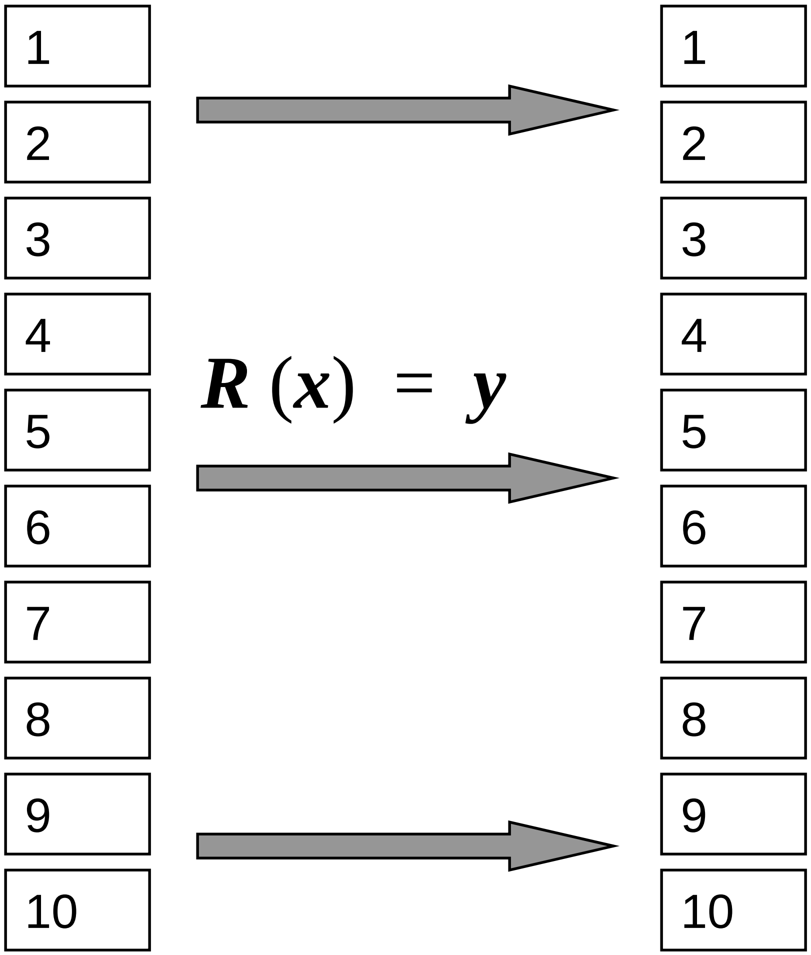
Checking for α-to-β privacy breaches:

- There are exponentially many properties ***P*** (***x***) ;
- We have to know the data distribution in advance in order to check whether
  Prob [***P*** (***X***)] $\leq \alpha$ and Prob [***P*** (***X***) | ***Y*** = ***y***] $\geq \beta$

Is there a simple property of randomization operator ***R*** that limits privacy breaches?

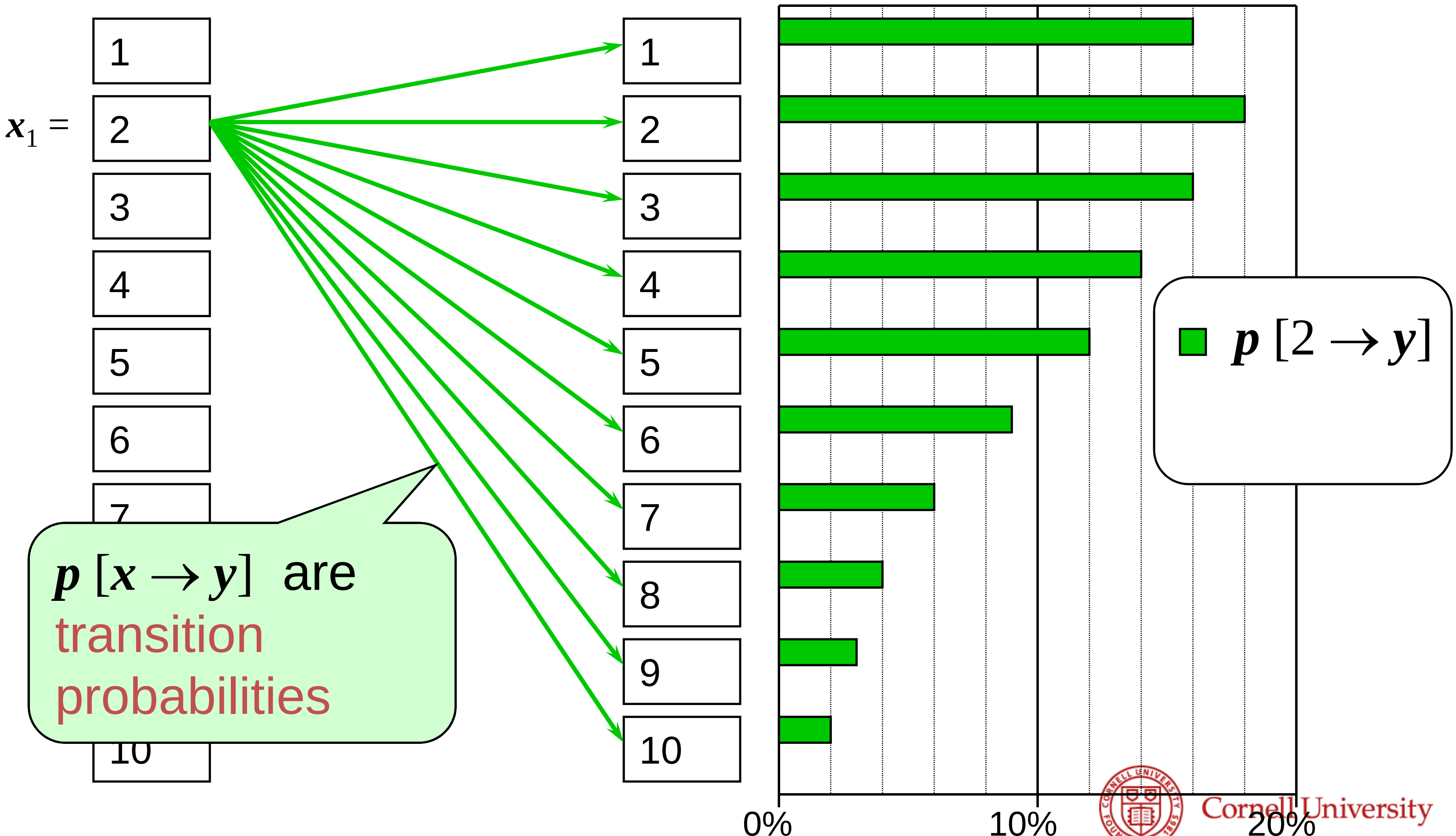
Cornell University

Amplification Condition

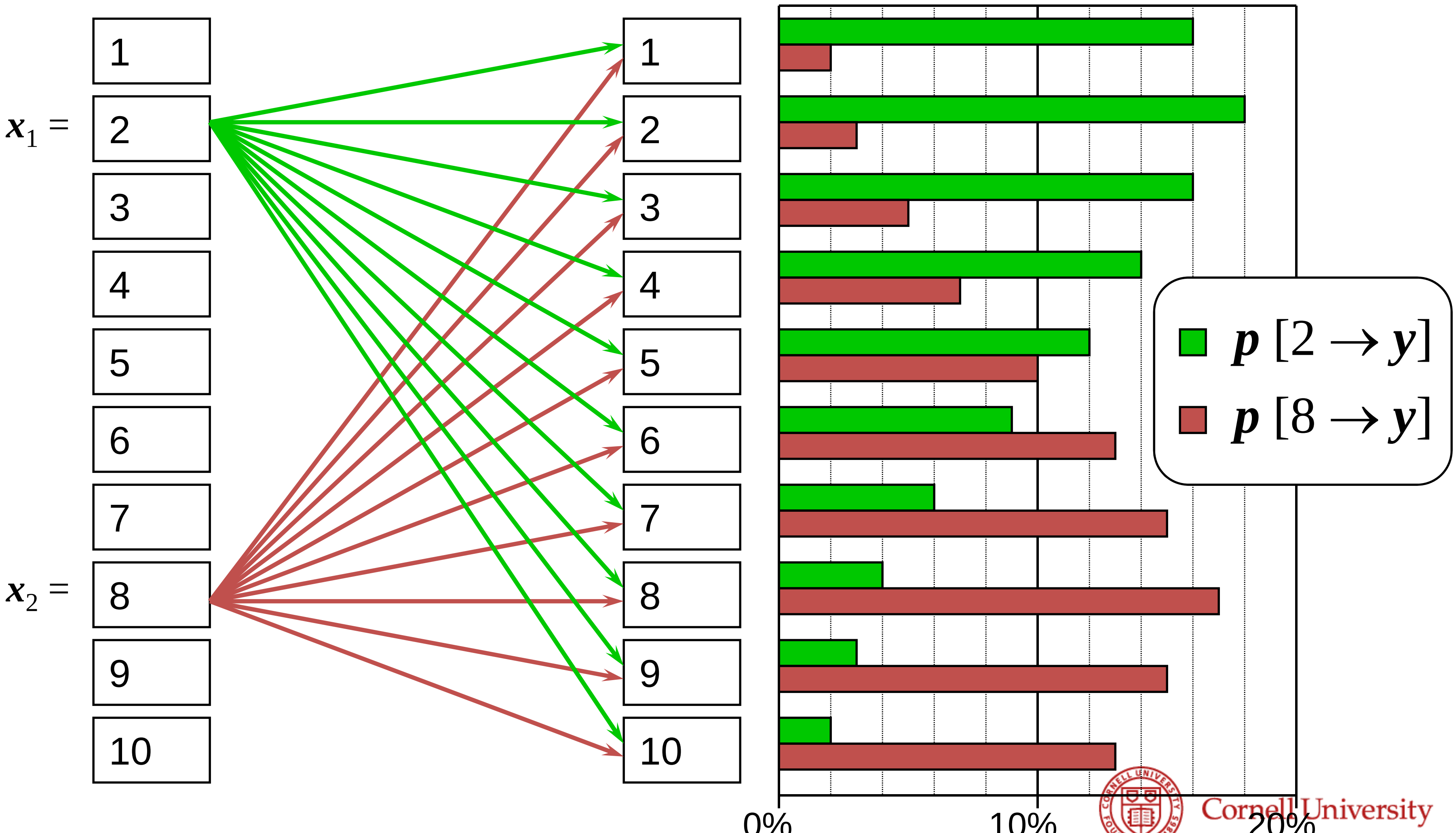


Cornell University

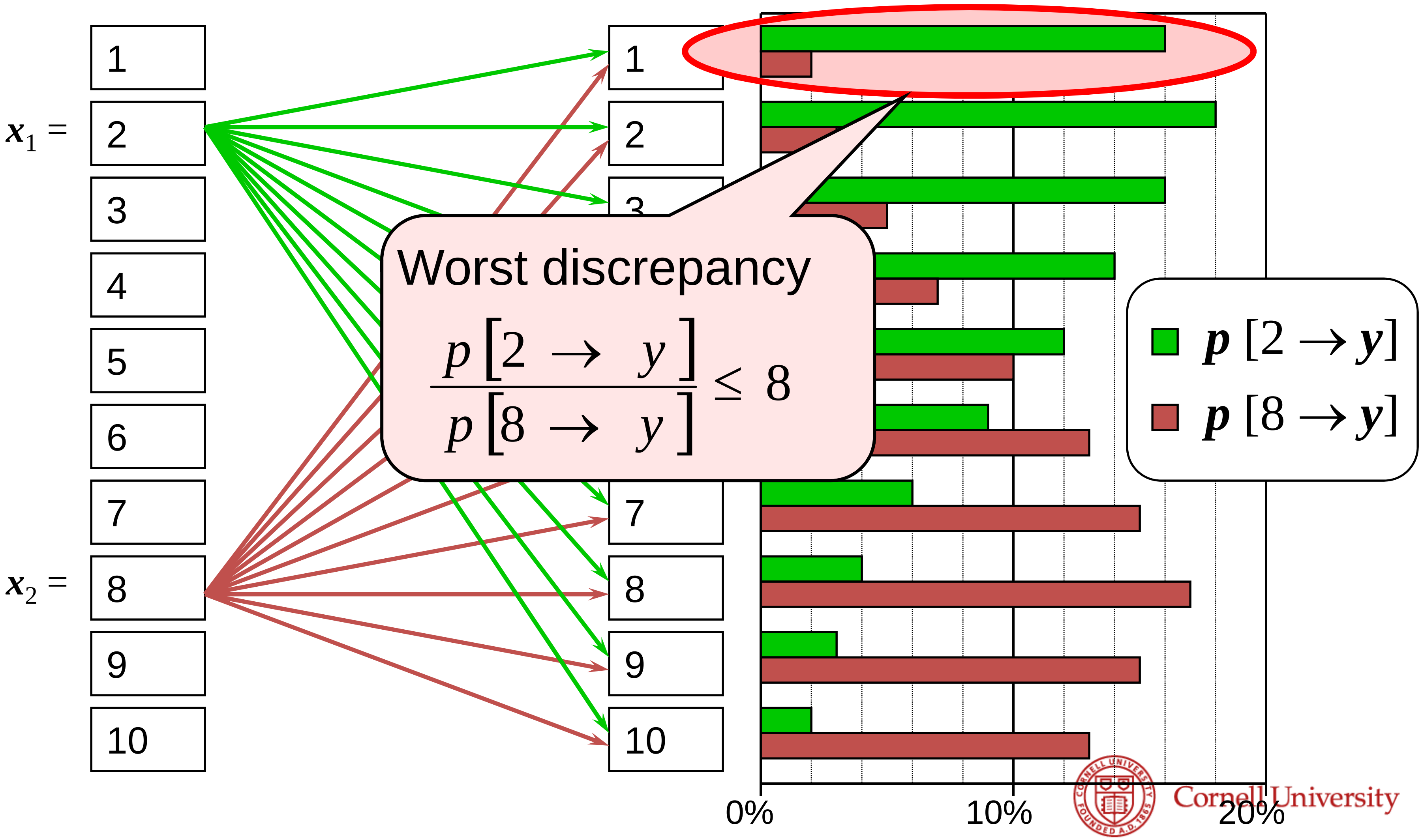
Amplification Condition



Amplification Condition



Amplification Condition



Amplification: Summary

- An α-to-β privacy breach w.r.t. property P(x) occurs when
  - Prob [P is true] $\leq \alpha$
  - Prob [P is true | Y = y] $\geq \beta$.

- Amplification methodology limits privacy breaches by just looking at transitional probabilities of randomization.
  - Does not use data distribution; only check:

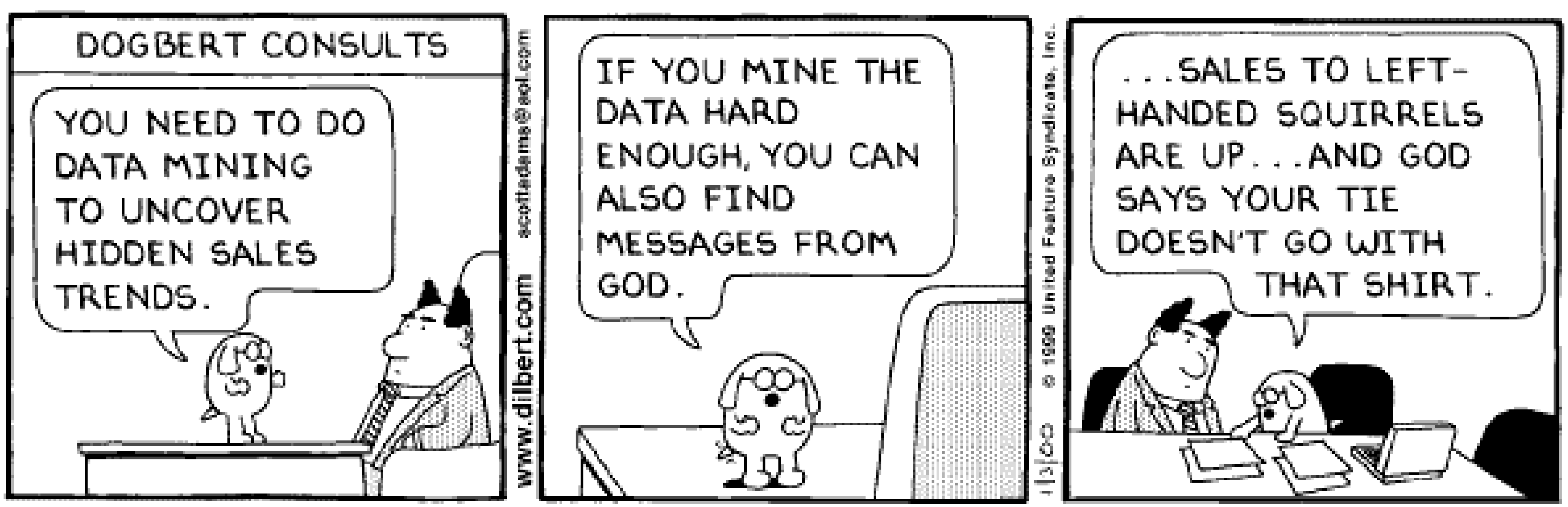
$$\max_{x_1, x_2} \max_{y} \frac{p[x_1 \to y]}{p[x_2 \to y]} \leq \gamma$$

Cornell University

Alexandre V. Evfimievski, Johannes Gehrke, Ramakrishnan Srikant:
Limiting privacy breaches in privacy preserving data mining. PODS 2003: 211-222

Privacy: The Floodgates are Open

- Formal notions of privacy: L-Diversity, t-closeness, differential privacy, zero-knowledge privacy
- Attacks: DeFinetti attack, re-identification attacks in graphs [Netflix]
- Applications: Privacy in social networks, location privacy



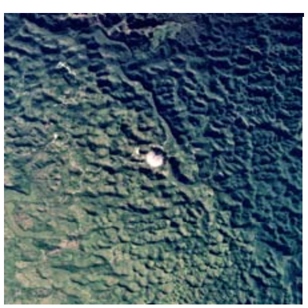
Summary

- Motivation: Large data
  - Many modalities
  - Many applications
  - Resource constraints are everywhere!
- Techniques:
  - Sketches
  - Automata-based complex event processing
- Data privacy as an emerging concern

CS and the Knowledge Economy

- Data and its connection to the real world motivate students to study computer science
- Programmers are creative!

http://scratch.mit.edu/



http://mindhacks.org/category/creativity/

Cornell University

Questions?

johannes@cs.cornell.edu

http://www.cs.cornell.edu/johannes

This talk discusses joint work with Rakesh Agrawal (Microsoft), Manuel Calimlim (Cornell), Alan Demers (Cornell), Abhinandan Das (Google), Alin Dobra (University of Florida), Alexandre Evfimievski (IBM), Minos Garofalakis (University of Crete), Mingsheng Hong (Vertica), Daniel Kifer (PSU), Ashwin Machanavajjhala (Yahoo!), Rajeev Rastogi (Yahoo!), Mirek Riedewald (Northeastern), Ramakrishnan Srikant (Google), Niki Trigoni (Oxford), Walker White, and Yong Yao (Google).

Cornell University

Picture from: http://diyblogger.net/is-it-the-business-of-creativity-or-creativity-of-business