Quantitative Methods to Measure the Risk of Re-identification: Methodology Review

Bradley Malin, Ph.D.

November 30, 2017

Vanderbilt University, Professor (Biomedical Informatics, Biostatistics, & Computer Science)

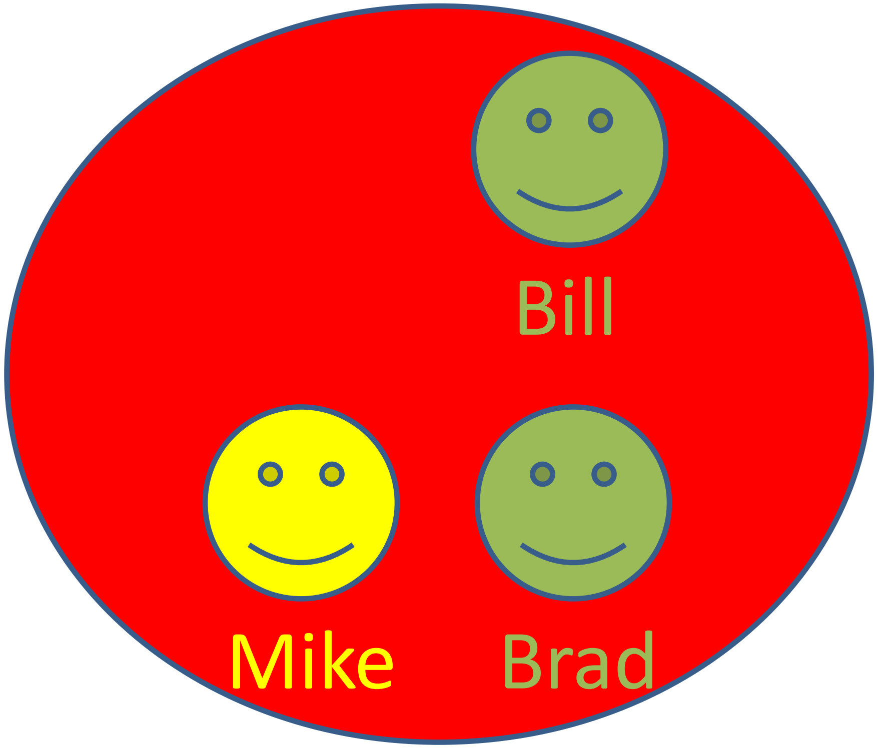
A Very Simplified View on Risk

$\mathrm{P}(reid) \approx$

- ***Distinguishability***
- Replicability
- Availability

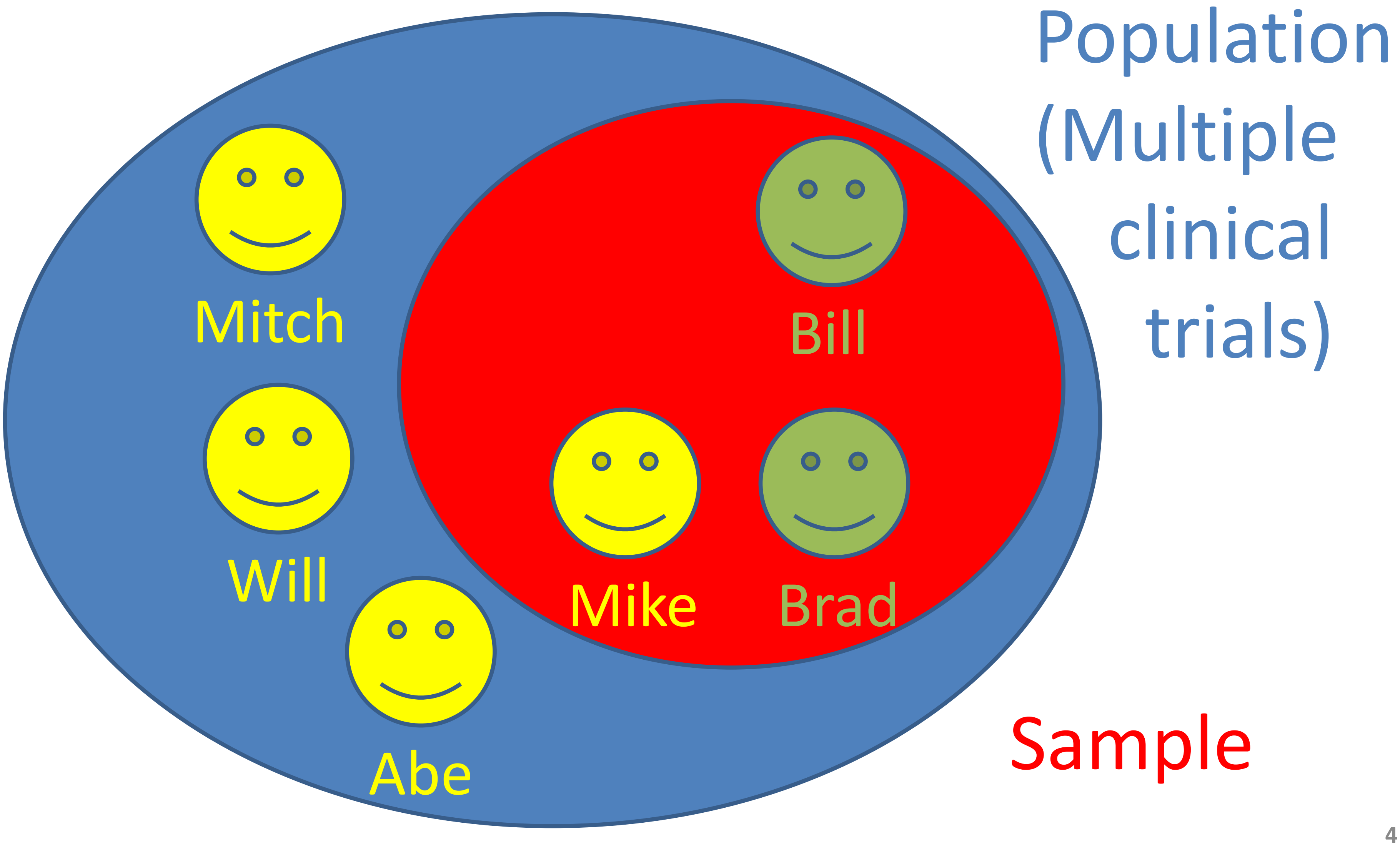
Data

Risk is Contextual

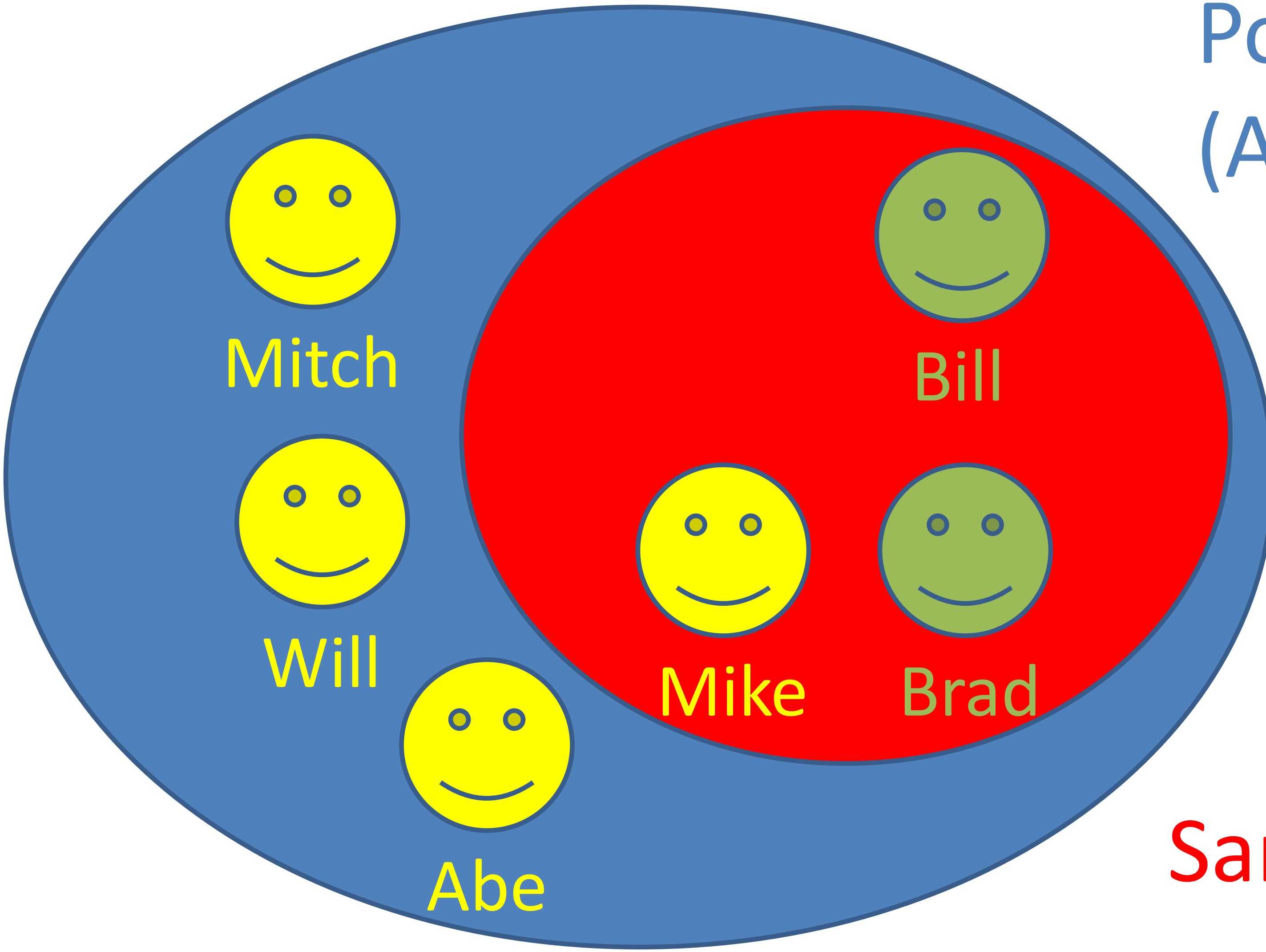


Sample

Risk is Contextual



Risk is Contextual



Population (All Eligible People)

Sample

Risk Measures

- Worst Case Risk Measures
  - Prosecutor: Most risky record in the *sample*
  - Journalist: Most risky record in the population
- Amortized (or Average) Measure
  - Marketer: Expected risk of an arbitrary record

A Very Simplified View on Risk

$$P(\mathit{reid}) \approx$$

$$P(\mathit{knowledge}) \, *$$

$$P(\mathit{reid} \mid \mathit{knowledge})$$

Prior

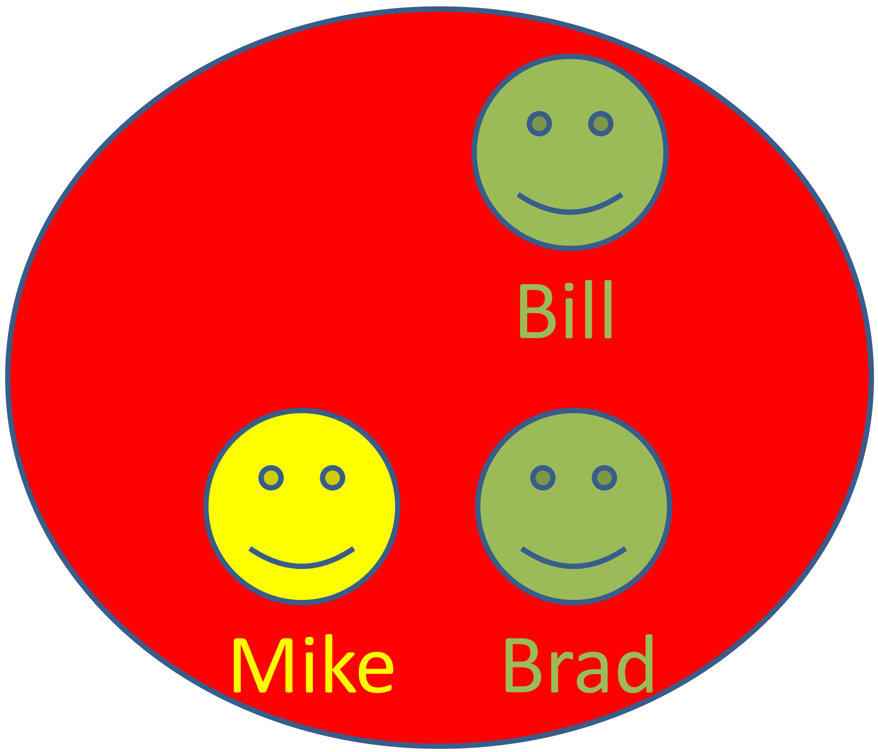
- Neighbor (friend)
- Prosecutor
- Journalist
- … pick a framework

- ***Distinguishability***
- Replicability
- Availability

Data

Prosecutor Risk

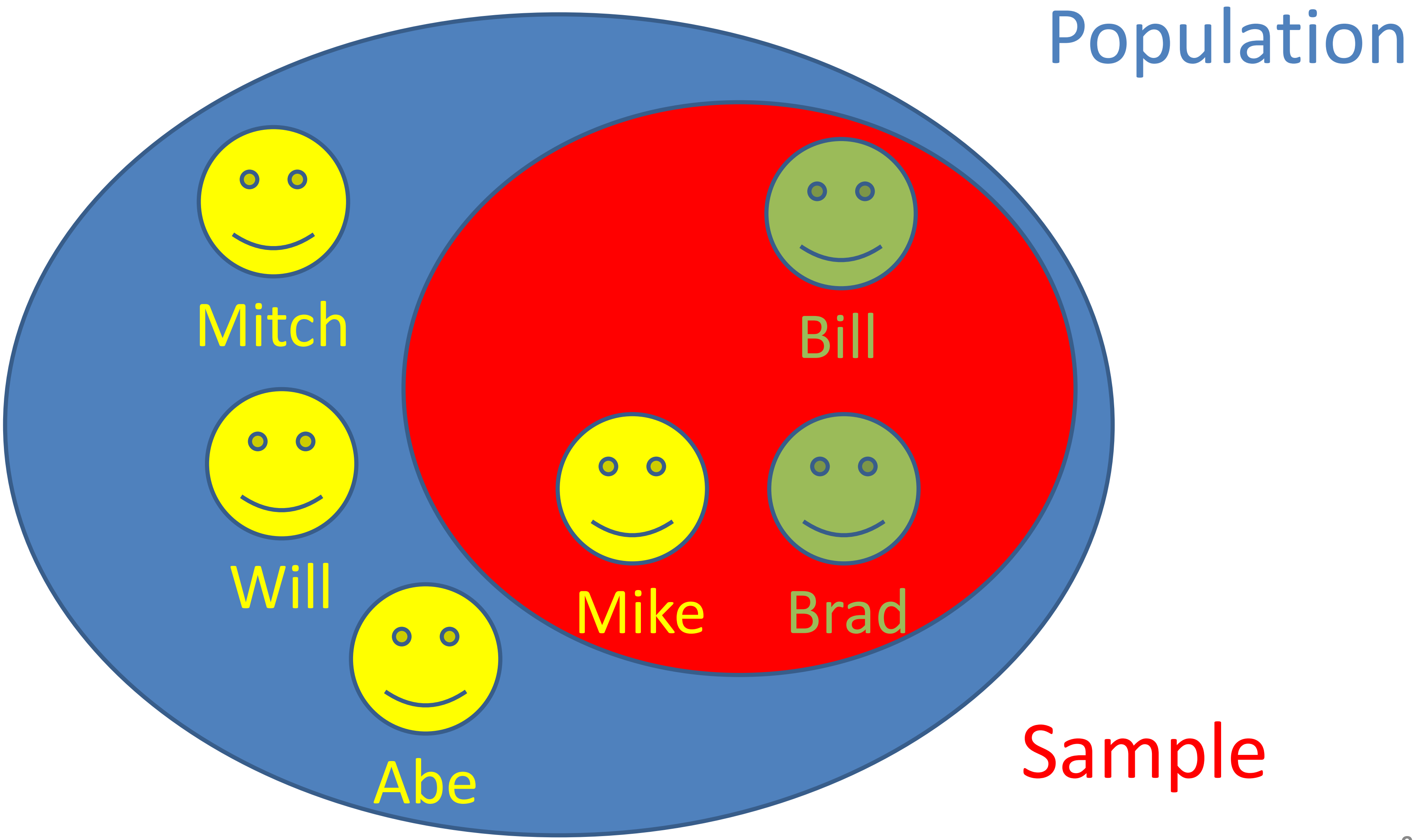
1 (Because of Mike)



Sample

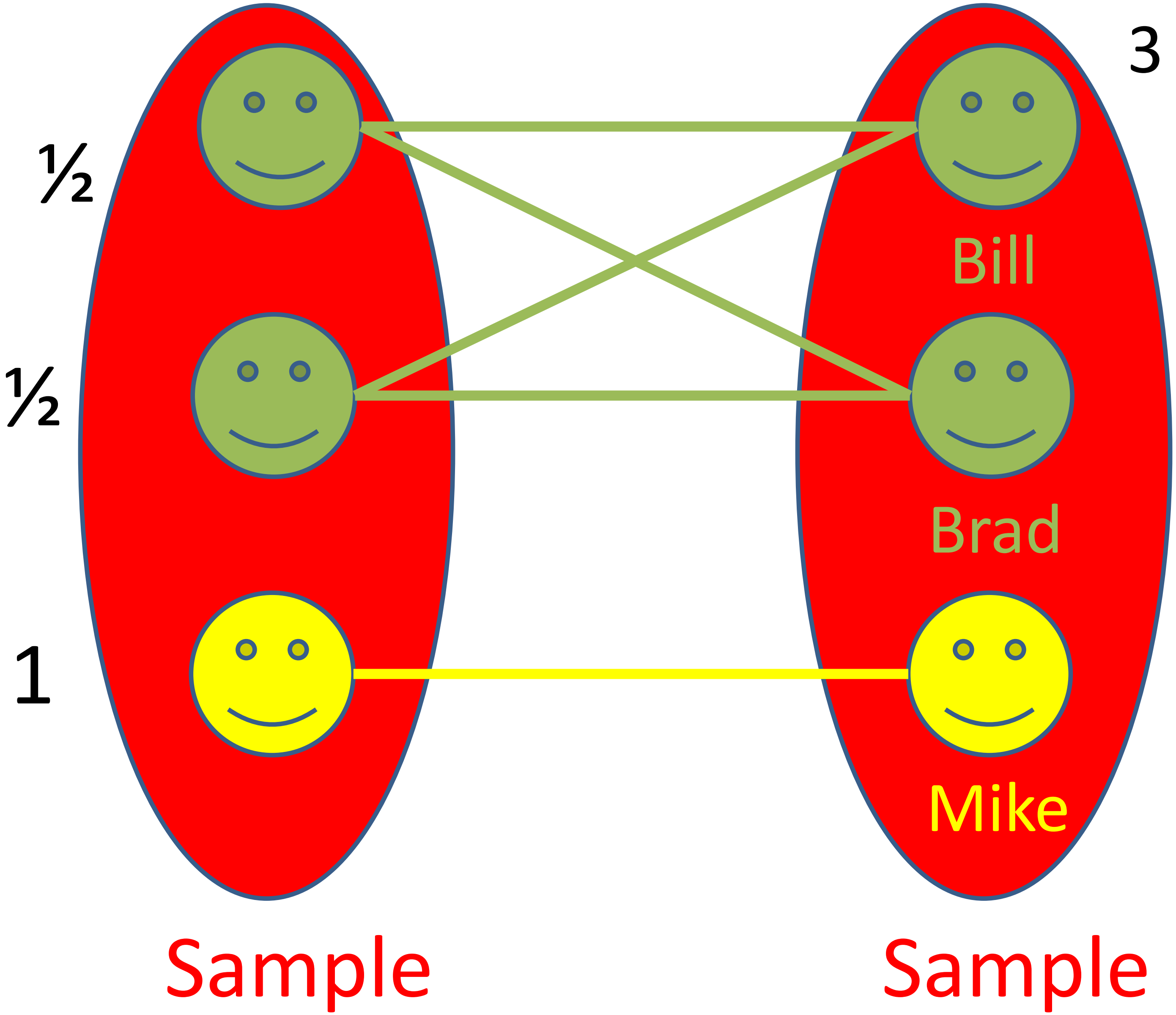
Journalist Risk

½ = 0.5 (Because of Bill & Brad)



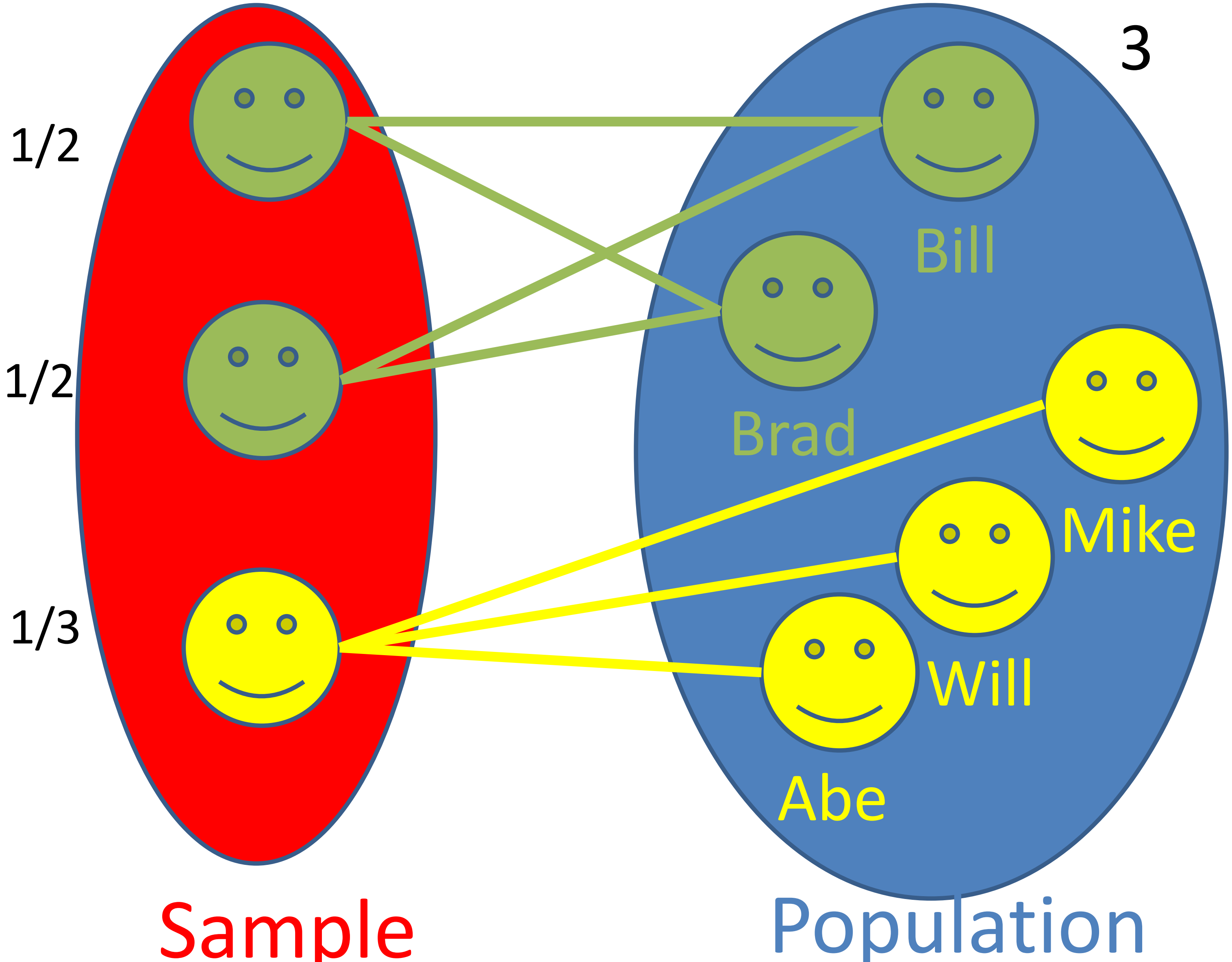
Marketer Risk (Sample)

$$\frac{\frac{1}{2} + \frac{1}{2} + 1}{3} = 2/3$$

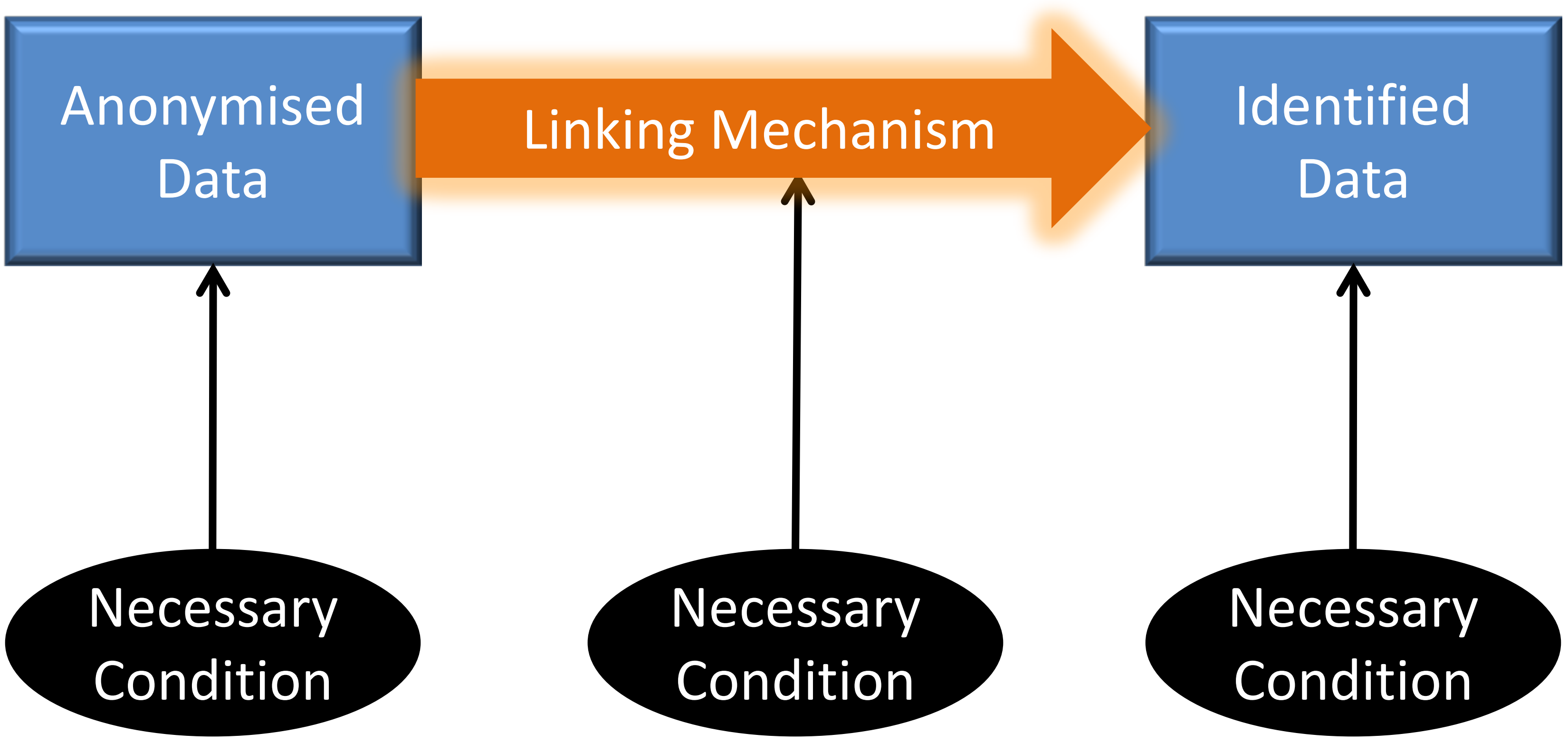


Marketer Risk (Population)

$$\frac{.5 + .5 + .33}{3} = .44$$

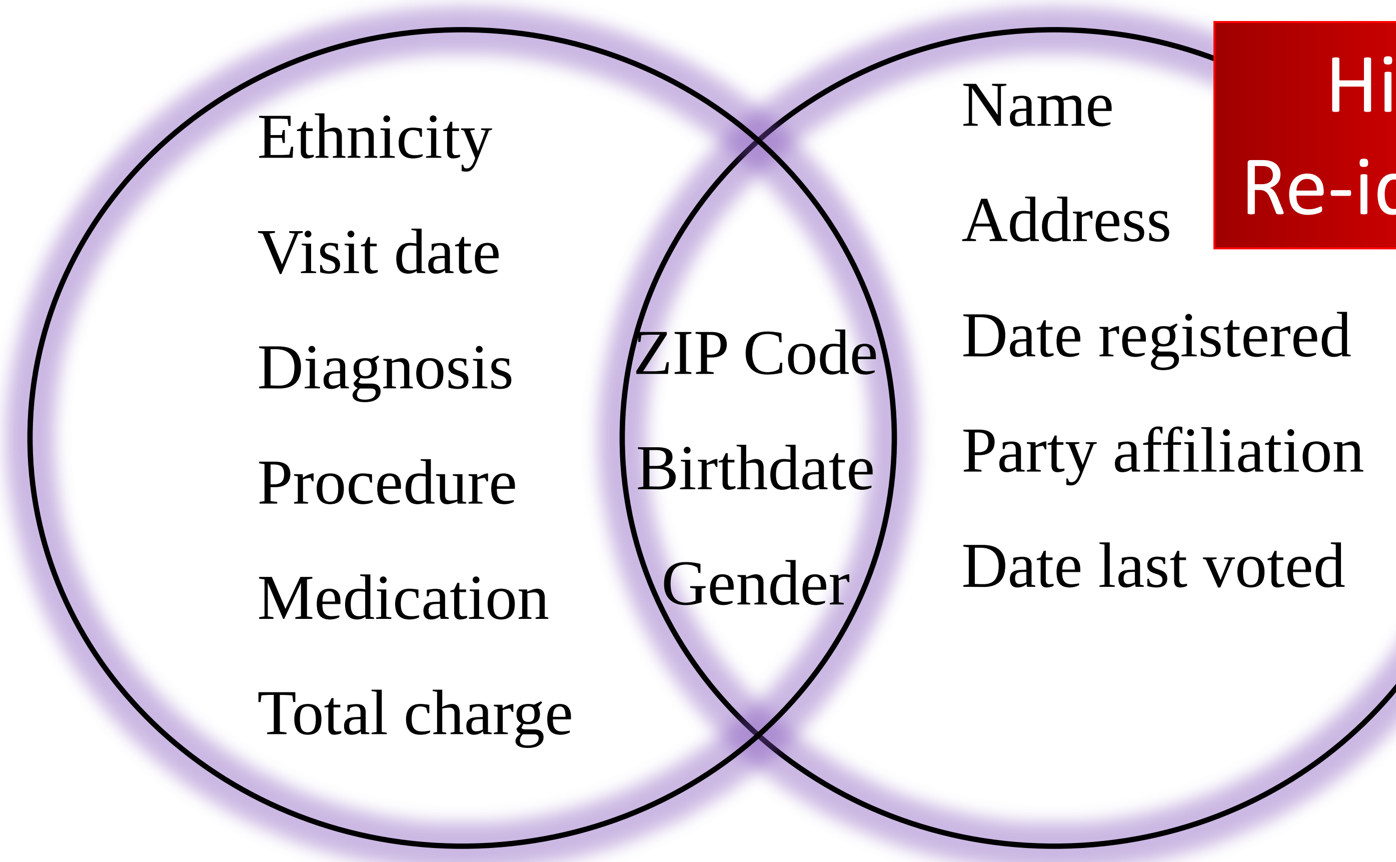


Central Dogma of Re-identification



Malin, Benitez, Loukides, & Clayton. Human Genetics. 2011.

A Famous Linkage Attack



U.S. Hospital Discharge Data

U.S. Voter List

High Profile Re-identification

Sweeney, JLME. 1997

But availability of Demographics Varies…

	IL	MN	TN	WA	WI
WHO	Registered Political Committees (ANYONE – In Person)	MN Voters	Anyone	Anyone	Anyone
Format	Disk	Disk	Disk	Disk	Disk
Cost	$500	$46; “use ONLY for elections, political activities, or law enforcement”	$2500	$30	$12,500
Name	●	●	●	●	●
Address	●	●	●	●	●
Date of Birth	●	○	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

Benitez & Malin. JAMIA. 2010.

Adversaries are Not All Knowing

- Research subject demographics
  - Unique (potential)
  - Replicable
  - Available

- Series of drug doses administered
  - Unique (potential)
  - Not replicable
  - Not available

Adversaries are Not All Knowing

(Assume knowledge between x and y features)

Feature	Observation
Date of Birth	1/1/1970
Gender	Male
Race	White
Country	France
Death	Yes
Occupation	Teacher
Married	Yes

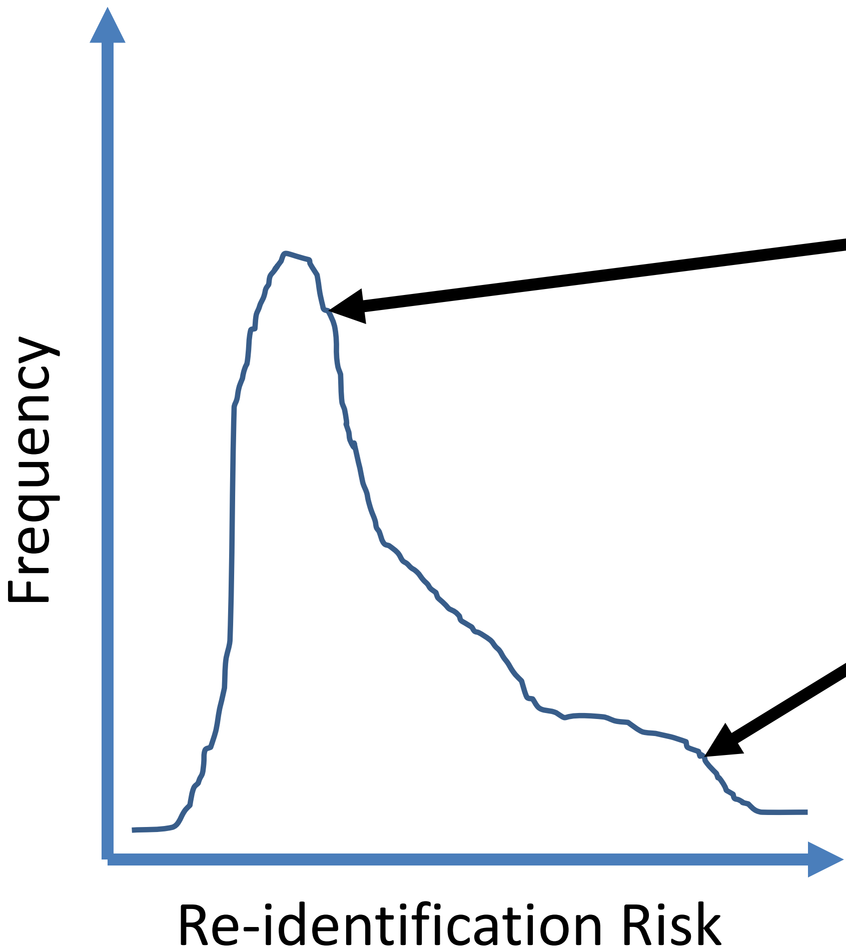
Feature	Observation
[redacted]	[redacted]
Gender	Male
[redacted]	[redacted]
Country	France
Death	Yes
Occupation	Teacher
[redacted]	[redacted]

Feature	Observation
Date of Birth	1/1/1970
[redacted]	[redacted]
[redacted]	[redacted]
Country	France
Death	Yes
Occupation	Teacher
Married	Yes

Feature	Observation
Date of Birth	1/1/1970
Gender	Male
Race	White
[redacted]	[redacted]
Death	Yes
Occupation	Teacher
Married	Yes

Feature	Observation
Date of Birth	1/1/1970
Gender	Male
Race	White
Country	France
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

… Which Means No Single Risk Score



Feature	Observation
[redacted]	[redacted]
Gender	Male
[redacted]	[redacted]
Country	France
Death	Yes
Occupation	Teacher
[redacted]	[redacted]

Feature	Observation
Date of Birth	1/1/1970
[redacted]	[redacted]
[redacted]	[redacted]
Country	France
Death	Yes
Occupation	Teacher
Married	Yes

Feature	Observation
Date of Birth	1/1/1970
Gender	Male
Race	White
[redacted]	[redacted]
Death	Yes
Occupation	Teacher
Married	Yes

Feature	Observation
Date of Birth	1/1/1970
Gender	Male
Race	White
Country	France
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

You May Not See Everything

- Field structured databases – you can issue exact queries!

- Semi-structured reports and narratives – you must rely on a mix of
  - Artificial intelligence
  - Human review
  - Sampling

- Problem has become a bit more tricky because leaks can now include *explicitly identifying* values

A Natural Language Setting

Original PHI

```
Smith, 61 yo ...
daughter, Lynn, to ...
oncologist Dr. White ...
5/13/10 to consider ...
SWOG protocol 1811, ...
was randomized 5/10 ...
to call Mr. Smith on ...
PLAN:Dr White and I ...
```

- Create a Gold Standard Dataset
  - Ask X humans to read and label a selection of records
  - Ensure concordance between human annotations (e.g., via a Cohen's Kappa)
- Apply (manual or automated) identifier detection strategy
- Compute performance in terms of:
  - (R)ecall – rate at which real identifier instances were detected
  - (P)recision – rate at which claimed identifiers are in fact real
  - F-measure – weighted average of R and P

A Natural Language Setting

- A leak is not necessarily a re-identification
- Need to assess the potential given the leak rates
- Can assess on a per research subject level (if labels are person specific)
- Alternatively, may assume each feature leaks at random

Original PHI

```
Smith, 61 yo ...
daughter, Lynn, to ...
oncologist Dr. White ...
5/13/10 to consider ...
SWOG protocol 1811, ...
was randomized 5/10 ...
to call Mr. Smith on ...
PLAN:Dr White and I ...
```

****Redacted PHI & Leaked PHI**

```
**pt_name<A>, **age<60s> yo ...
daughter, Lynn, to ...
oncologist Dr. **MD_name<C> ...
**date<5/28/10> to consider ...
SWOG protocol **other_id, ...
was randomized 5/10 ...
to call Mr. **pt_name<A> on ...
PLAN:Dr White and I ...
```

P(date of birth leak, date of death leak)

≈ P(date of birth leak) * P(date of death leak)

Hiding in Plain Sight (Carrell et al, 2013)

- Must be careful – this is a relatively new technique
- Also evidence to suggest a computer may mimic the initial detection strategy… and redact the fakes situated in a pattern (Li et al, 2016)

Original PHI	**Redacted PHI & Leaked PHI	Surrogate PHI & Hidden PHI

Original PHI

```
Smith, 61 yo ...
daughter, Lynn, to ...
oncologist Dr. White ...
5/13/10 to consider ...
SWOG protocol 1811, ...
was randomized 5/10 ...
to call Mr. Smith on ...
PLAN:Dr White and I ...
```

****Redacted PHI & Leaked PHI**

```
**pt_name<A>, **age<60s> yo ...
daughter, Lynn, to ...
oncologist Dr. **MD_name<C> ...
**date<5/28/10> to consider ...
SWOG protocol **other_id, ...
was randomized 5/10 ...
to call Mr. **pt_name<A> on ...
PLAN:Dr White and I ...
```

Surrogate PHI & Hidden PHI

```
Jones, a 64 yo ...
daughter, Lynn, for ...
oncologist Dr. Howe ...
5/28/10 to consider ...
SWOG protocol 1798, ...
was randomized 5/10 ...
to call Mr. Jones on ...
PLAN:Dr White and I ...
```

- In this case, we would need two risk measures
  - Human recognition risk
  - Computer-assisted recognition risk
- Recent approach to prevent this problem, but comes at a massive loss in precision (Li et al, 2017)

If Time Permits…

If Not…

Latest Development

- Methods presented model data sharer and adversary separately

- New approaches use game theory to consider their interactions (Wan et al 2015)

- Game theory requires robust estimates of many parameters, such as
  - benefit the sharer gets in providing data
  - benefit the attacker gets in re-identifying the data

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk ???

Strategies:
- Generalize Age
- Suppress Dates

…
- Perturb Geography

Publisher

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk ???

Attack Strategy A
Utility A
Risk A

Recipient's Best Strategy

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Publisher

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

Recipient's Best Strategy

⋮

Attack Strategy C
Utility C
Risk C

Publisher

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk ???

Publisher

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk A

Recipient's Best Strategy

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Publisher

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk A

⋮

Sharing Strategy Z
Utility Z
Risk Z

Publisher

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk A

⋮

Sharing Strategy Z
Utility Z
Risk Z

Publisher

Choose strategy that maximizes overall benefit

Optimizes the Risk-Utility tradeoff

Demographic Case Study

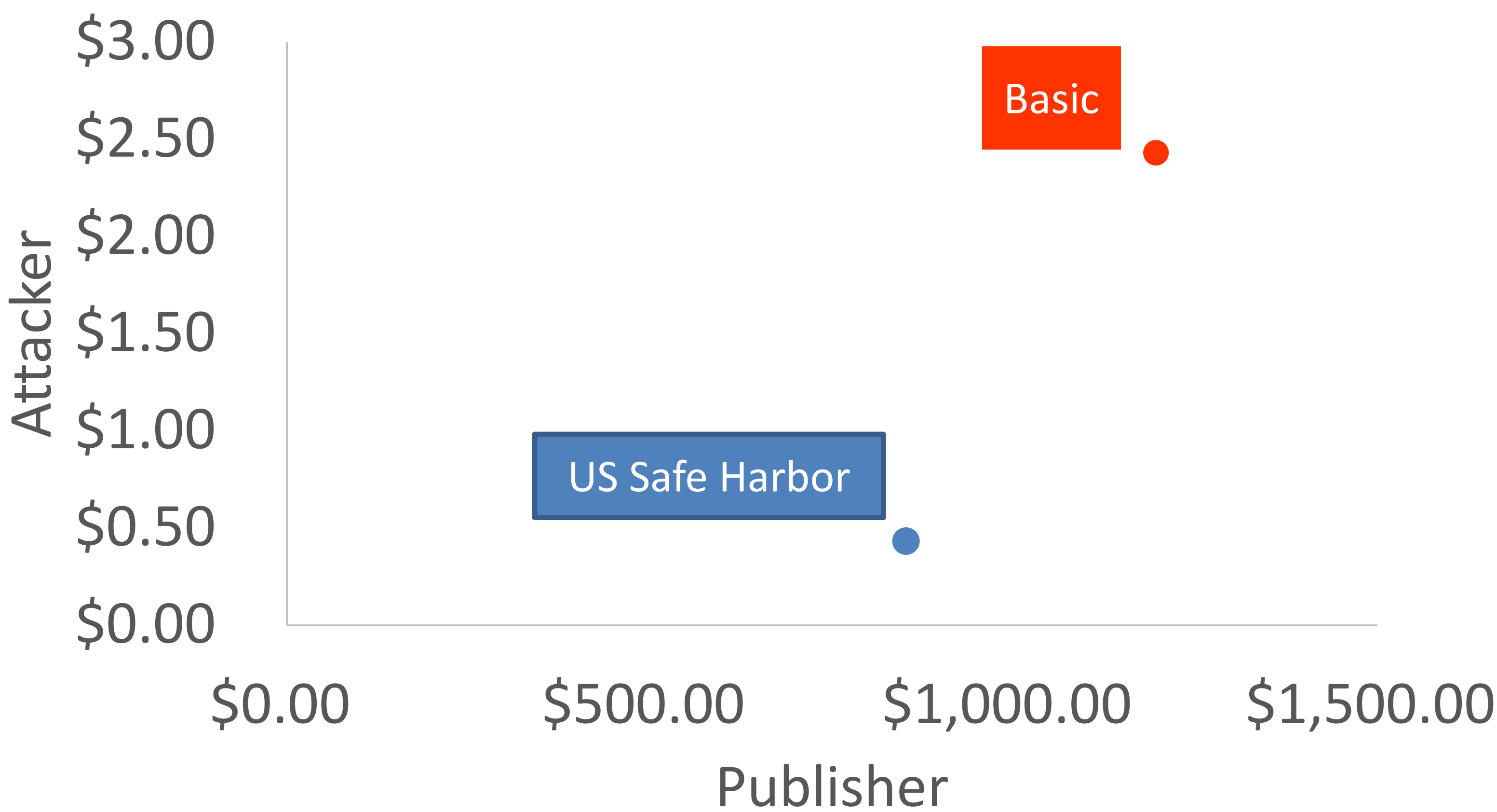
- ~30,000 Census records
- Average Payoff Per Record

- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record

Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record

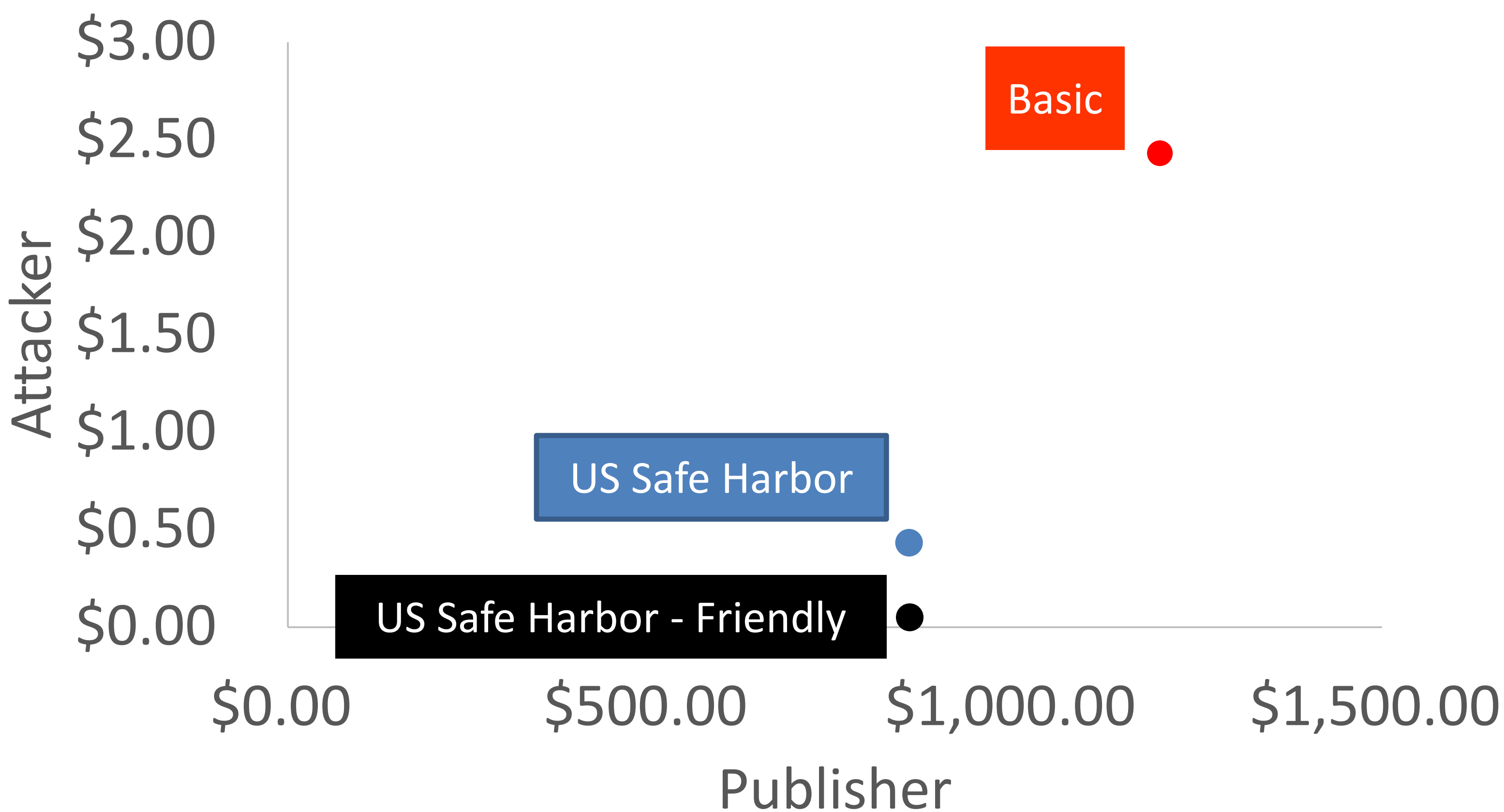
- $1200: Benefit per record
- $300: Cost per violation
- $4: Access cost per record



Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record

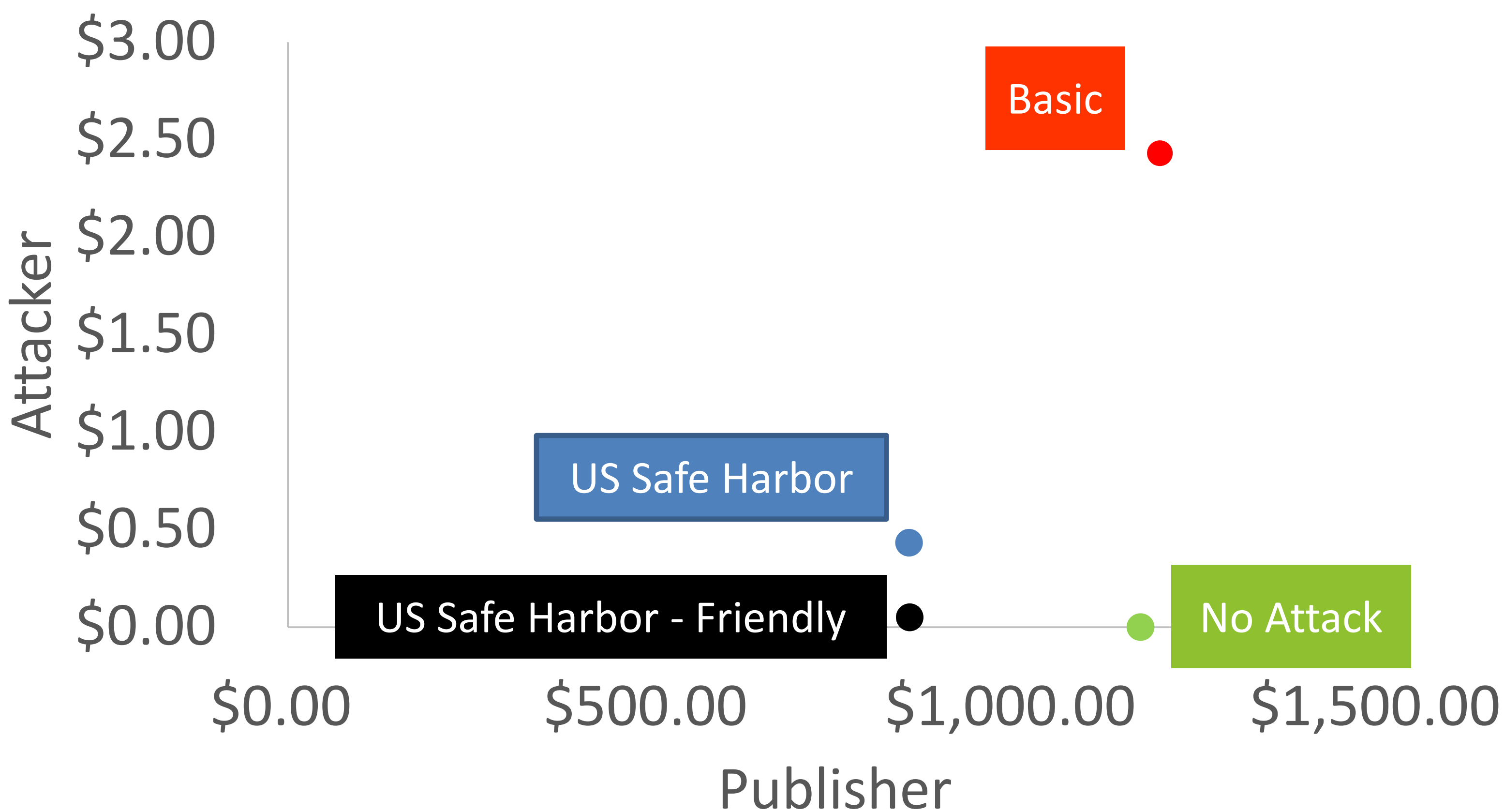
- $1200: Benefit per record
- $300: Cost per violation
- $4: Access cost per record



Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record

- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record



Some Relevant References

- Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. Journal of the American Medical Informatics Association. 2010; 17: 169-177.

- Carrell D, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association. 2013; 20: 342-348.

- El Emam, Malin B. Concepts and methods for de-identifying clinical trial data. In: Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. National Academies Press. 2015.

- Li B, et al. Scalable iterative classification for sanitizing large-scale datasets. IEEE Transactions on Knowledge and Data Engineering. 2017; 29: 698-711.

- Li M, et al. Optimizing annotation resources for natural language de-identification via a game theoretic framework. Journal of Biomedical Informatics. 2016; 61: 97-109.

- Wan Z, et al. A game theoretic technique for analyzing re-identification risk. PLoS One. 2015; 10: e0120592.