



Privacy: more than meets the eye

Daniel Kifer (Penn State University) 

Data

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer

Big Data?

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer
123-456-7890	12345	5	French	Viral
135-711-1317	11220	98	German	Heart
...	...	...	...	...

Big Data?

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0899	14853	50	Indian	Cancer
265-04-1275	14853	50	Russian	Heart
574-22-0242	11350	47	American	Viral
388-32-1539	14850	37	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer
123-456-7890	12345	5	French	Viral
135-711-1317	11220	98	German	Heart
...	...	...	...	...

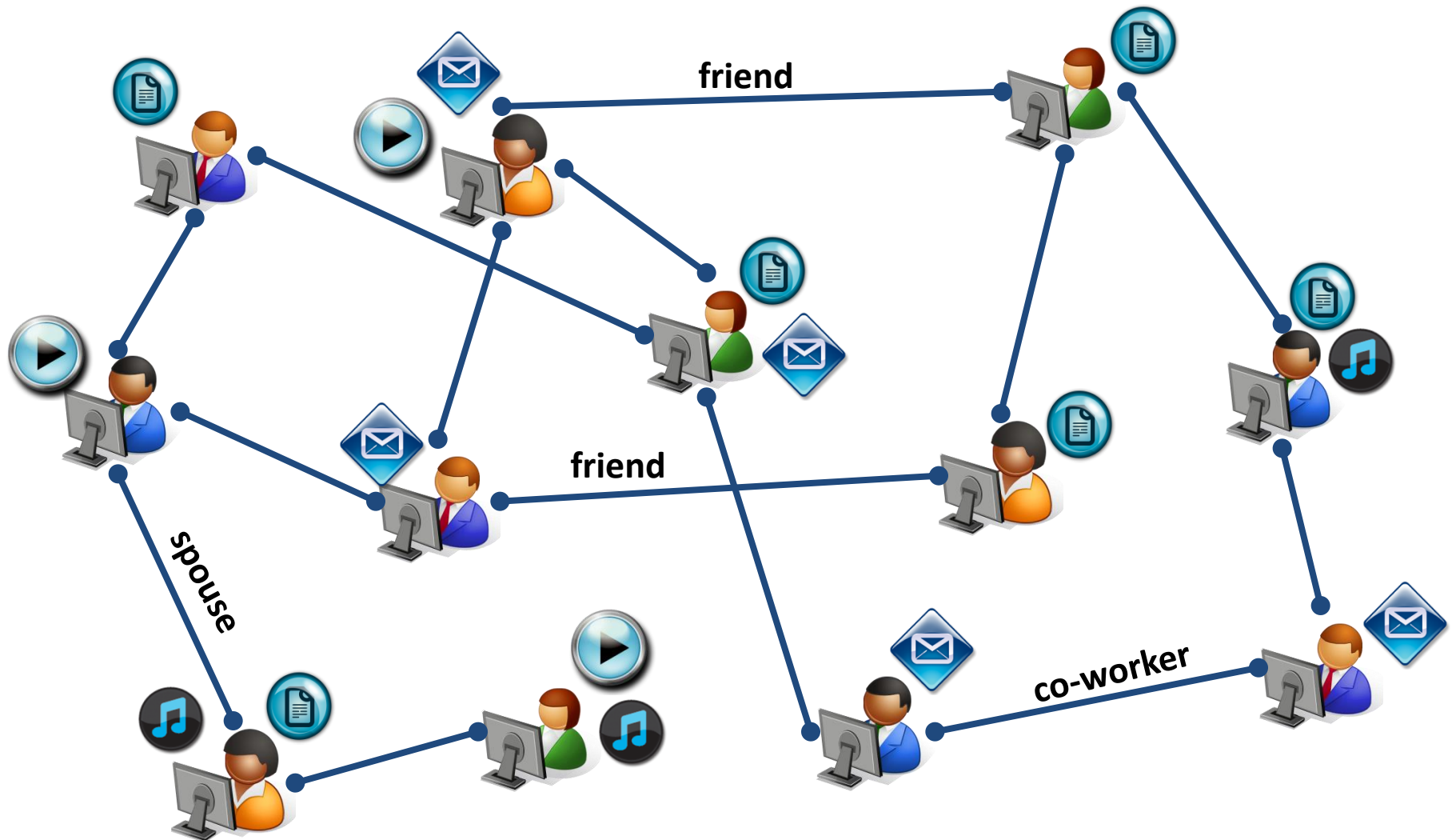
Big Data?

[illegible]

Big Data?

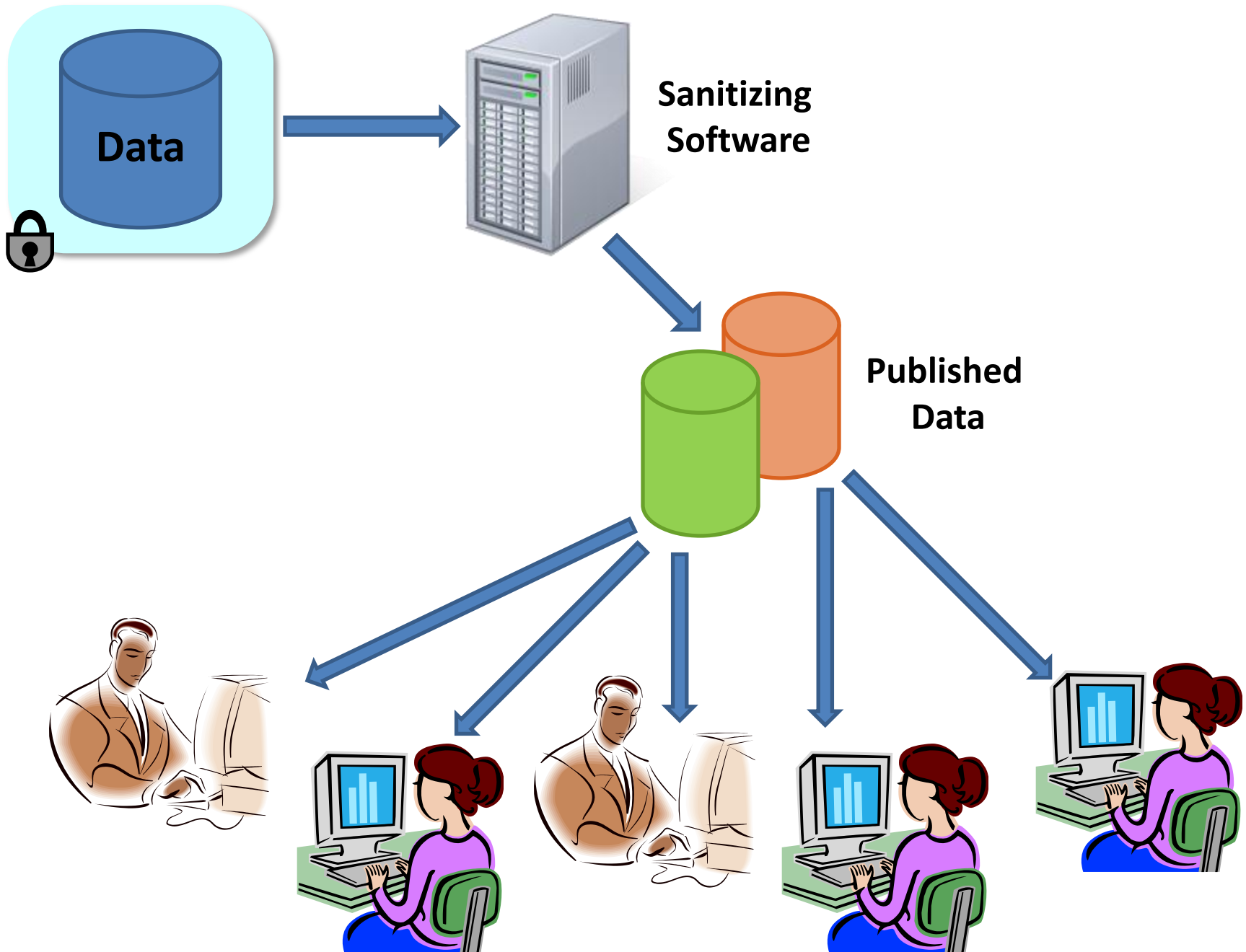
[illegible]

Big Data



Big Data





Lesson 1

- Privacy is **not** about what is in/out of sanitized data.

Lesson 1

Ashwin222	Uefa cup
Ashwin222	Uefa champions league
Ashwin222	Champions league final
Ashwin222	Champions league final 2008
Dkifer1567	exchangeability
Dkifer1567	Proof of deFinetti's theorem
David1234	Zombie games
David1234	Warcraft
David1234	Beatles anthology
David1234	Ubuntu breeze
Ashwin222	Grammy 2008 nominees
Ashwin222	Amy Winehouse rehab

Lesson 1 (this doesn't work)

865712345	Uefa cup
865712345	Uefa champions league
865712345	Champions league final
865712345	Champions league final 2008
236712909	exchangeability
236712909	Proof of deFinetti's theorem
112765410	Zombie games
112765410	Warcraft
112765410	Beatles anthology
112765410	Ubuntu breeze
865712345	Grammy 2008 nominees
865712345	Amy Winehouse rehab

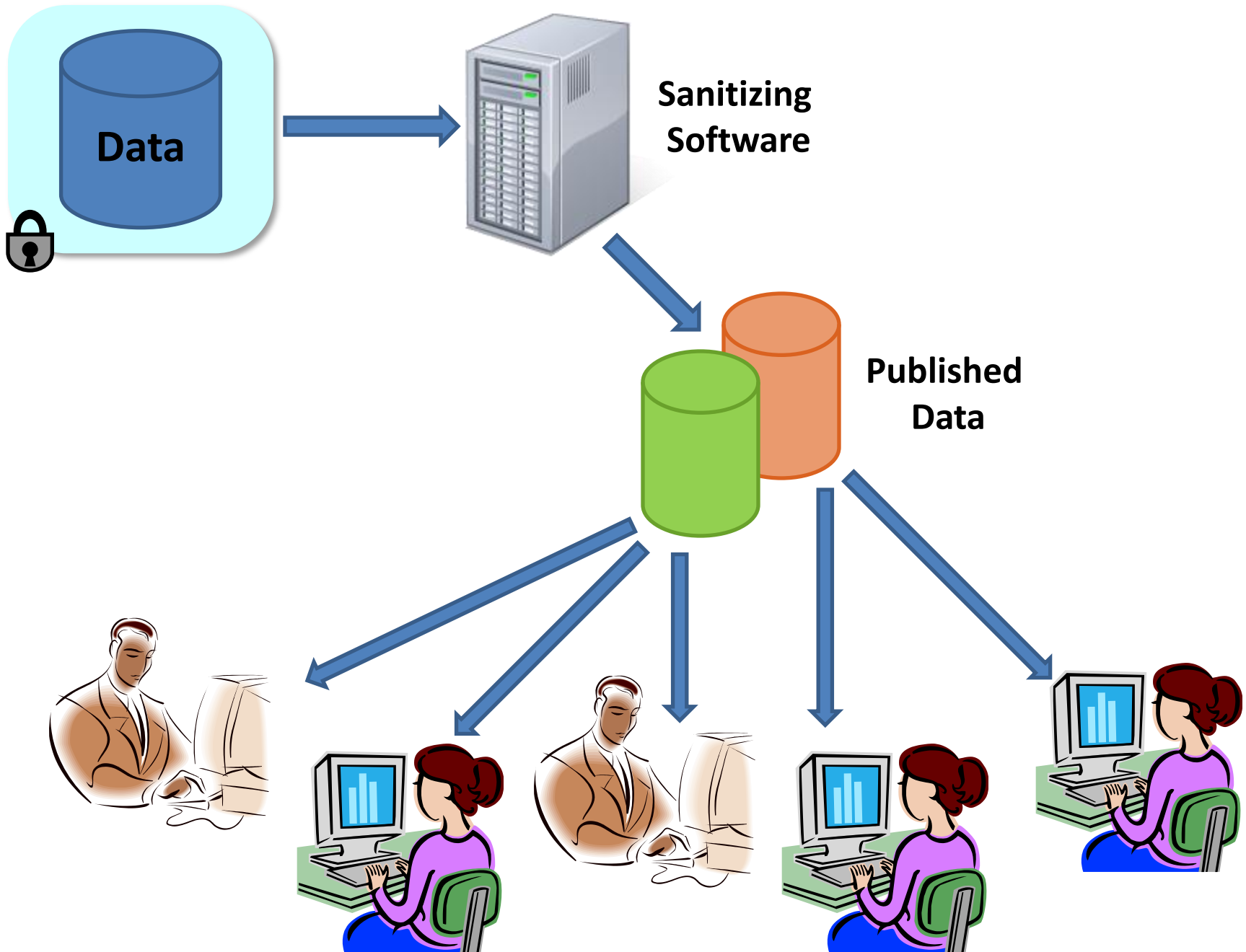
Lesson 1 (neither does this)

865712345	8722	7801		
865712345	8722	5426	8838	
865712345	5426	8838	9199	
865712345	5426	8838	9199	9626
236712909	4903			
236712909	4238	1915	4903	5769
112765410	2505	4109		
112765410	8435			
112765410	9273	7149		
112765410	3982	1713		
865712345	7236	9626	2760	
865712345	1343	1579	3493	

Kumar, Novak, Pang, Tomkins. "On anonymizing query logs via token based hashing." ACM International conference on the World Wide Web (WWW) ,2007.

Lesson 1

- Privacy is **not** about what is in/out of sanitized data.
 - It is about (statistical) inference



Lesson 2

- Privacy is about **behavior** of sanitizing software (i.e. the algorithm).
 - But **not** about **what** it outputs.

- Who stole the cookie from the cookie jar?



I can neither
confirm nor
deny that I stole
the cookie.



I can neither
confirm nor
deny that I stole
the cookie.

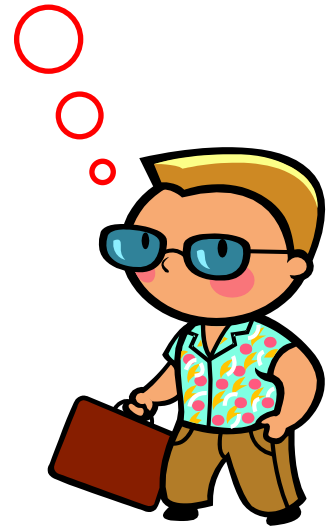
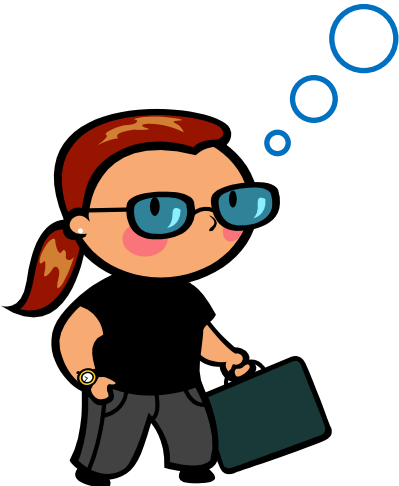


- Who stole the cookie from the cookie jar?



If steal: "cannot confirm
nor deny"
If not steal: "not I"

Always say "cannot
confirm nor deny"



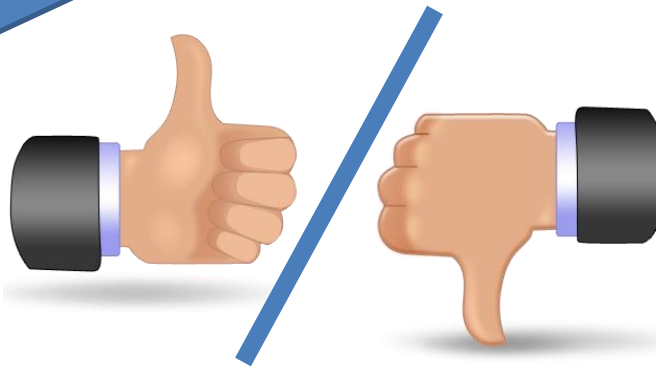
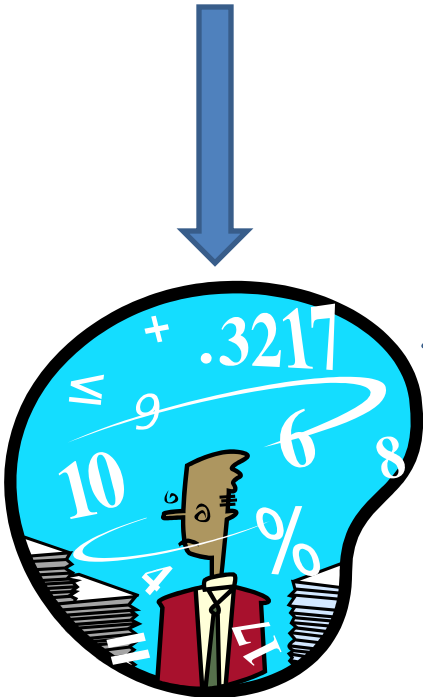
Lesson 2

- Privacy is about **behavior** of sanitizing software (i.e. the algorithm).
 - But **not** about **what** it outputs.
- Algorithm must be revealed (for end-user utility)
- How likely is output if secret is true?
- How likely is output if secret is false?
- $P(\text{output} \mid \text{secret}=\text{true})$ vs. $P(\text{output} \mid \text{secret}=\text{false})$
 - e.g., Differential privacy, Pufferfish, Adversarial Privacy, etc.

Lesson 3

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer
123-456-7890	12345	5	French	Viral
135-711-1317	11220	98	German	Heart
...	...	...	...	...

Based on the data,
 $P(\text{Bob has cancer})=0.9$



Lesson 3

- Statistical inference:



Learning something about the population
(underlying phenomenon that generated data)

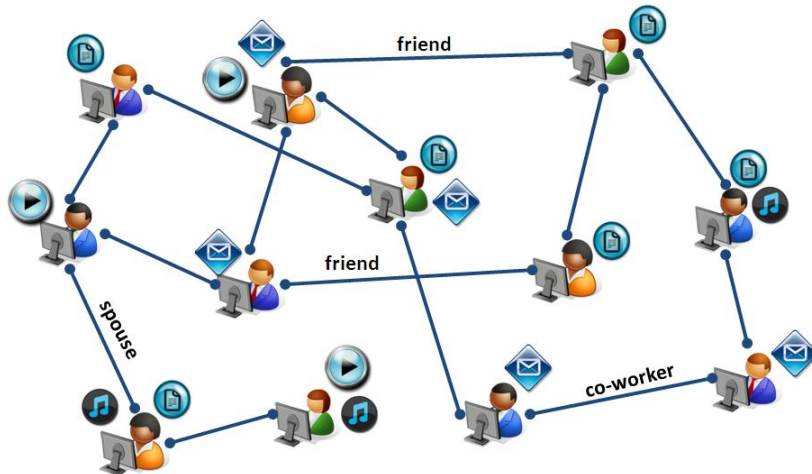


Learning how Bob differs from the population

Population inference

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer
123-456-7890	12345	5	French	Viral
135-711-1317	11220	98	German	Heart
...	...	...	...	...

Well understood
(and differential privacy
separates population
inference from individual
variations)



More work needed
(what is a "property of the
population"? what is an
"individual variation"?
what is the underlying
phenomenon?)



