Data Privacy

Boston University CS 506 - Lance Galletti

Disclaimer

A large part of this talk is my opinion. I encourage you to disagree, discuss, and debate throughout.

Why I care

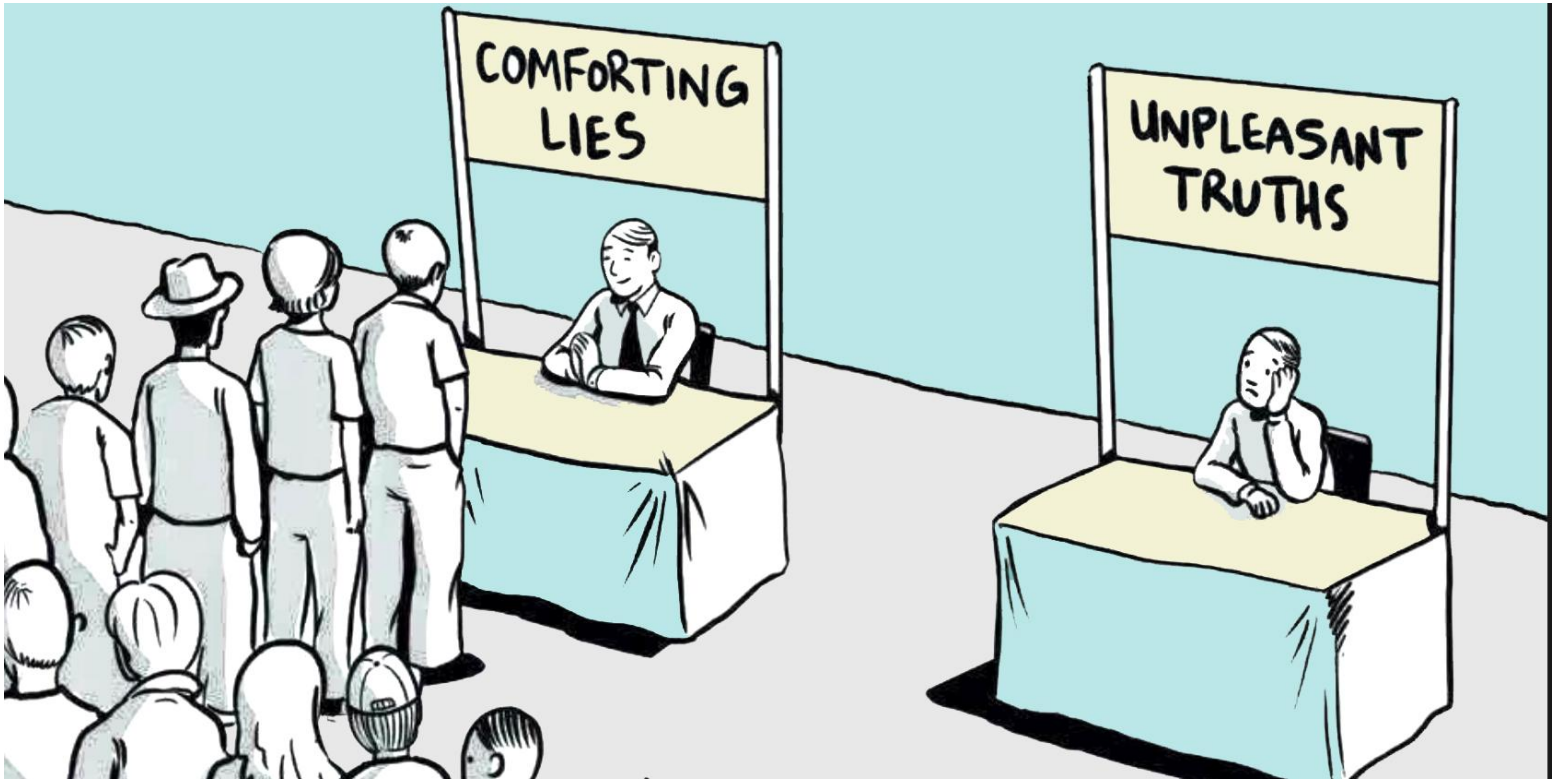
To me, privacy is **not about hiding illegal activity.**

It’s about:

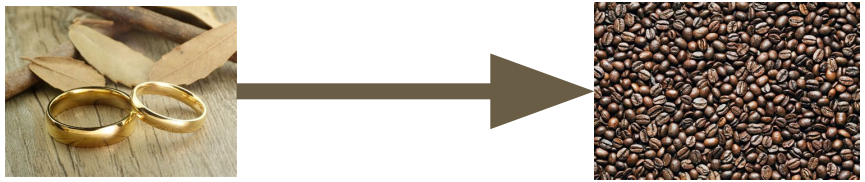
- Freedom of choice
- Equal opportunity / non-discrimination
- Accountability and transparency

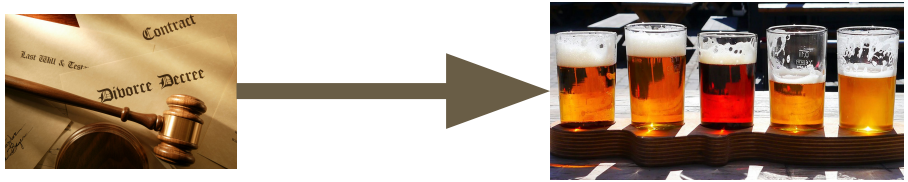
User - Product Relationship Today

User - Data-Driven Product Relationship Today



User - Data-Driven Product Relationship Today





User - Data-Driven Product Relationship Today

TARGET

Freedom of Choice

Freedom of Choice

Q: Do you think you are in control of your decisions or do you think your actions / decisions are determined by forces external to your will?

Q: What affects your decision process?

Q: When making decisions, is it better to be aware of the factors that influence these decisions?

https://en.wikipedia.org/wiki/Groundhog_Day_(film)

Data Determines

- Access to:
  - Health care
  - Insurance
  - Loans
  - Information
  - Jobs
- Distribution of funds
- Release from prison
- And more!

Bias

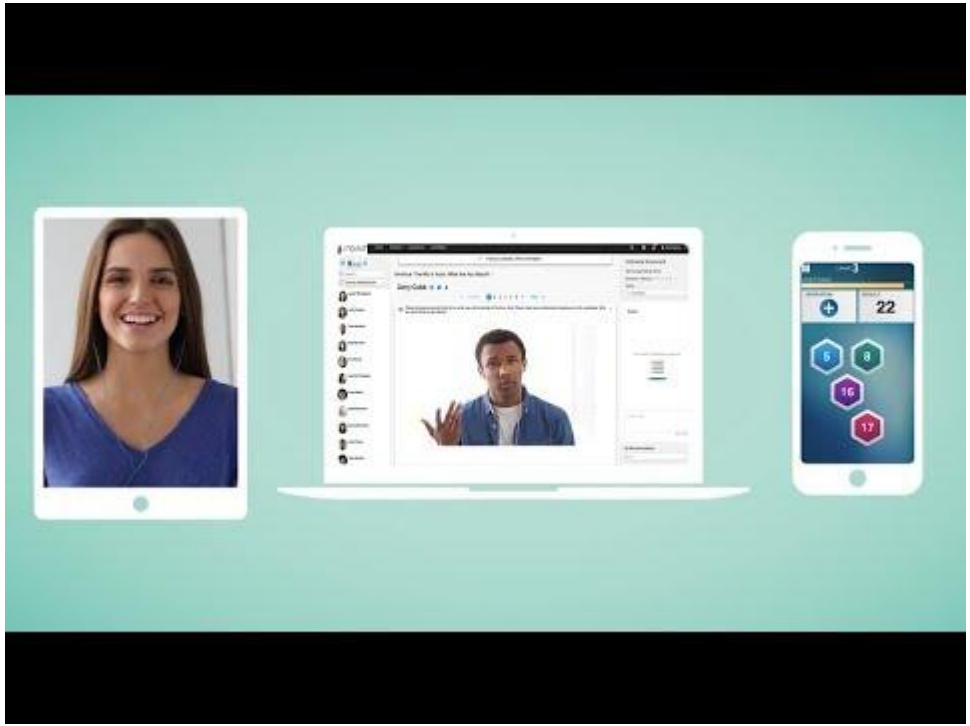
Lack of transparency and thus accountability.

These patterns are learned on historical data: reflecting society's past and existing biases and inequalities [4] which are then perpetuated by that same lack of visibility and accountability.

Garbage in - Garbage out: training state of the art models on garbage data can only produce garbage results

Bias

Bias



Facial recognition being used more and more:
https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Bias

Predicting criminality from faces: https://arxiv.org/pdf/1611.04135.pdf

Sexist AI: https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/

Sometimes adding more tech to a problem can make new problems...

Correlation vs Causation

Data used for risk assessment in releasing prisoners

https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Again black box algorithms

Critical Algorithms Studies

https://socialmediacollective.org/reading-lists/critical-algorithm-studies/

Subject to Future Scrutiny

With data being stored potentially indefinitely, data shared today will be subject to the latest cutting-edge information-extraction techniques of years and decades from now. This **far exceeds the timescale we are used to engaging with** and we cannot predict what information future tools will be able to extract from the data we submit today.

Users cannot then reasonably consent to all the ways in which their data, whose richness is undisclosed and will only increase with the pace of technology, is used or will be used.

Challenge of Regulations

What control users do have over their data is entirely at the whim of the UI exposing and accurately acting upon that control.

Regulations [7] have mandated UI changes that have made it possible, for example, for some users to delete their data from platforms. Such regulations are difficult to enforce because it ultimately boils down to our ability to verify software correctness.

What's worse, users can follow the UI steps to delete their data but the information extracted from this data may have already been incorporated into the models at large — a change that cannot be undone.

PII - Personally Identifiable Information

Most Data Privacy laws are based on hiding / removing PII. But PII is contextual.

Uniquely identified by zip, birthday, and sex:
https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

De-Anonymization of Netflix dataset using Amazon Reviews
https://courses.csail.mit.edu/6.857/2018/project/Archie-Gershon-Katchoff-Zeng-Netflix.pdf

Data Brokers

One argument for privacy has been that Data Brokers and companies make money off of your data and individuals deserve to get a slice of the pie.

https://clearcode.cc/blog/what-is-data-broker/

Q: if you can sell your data, how much is your identity worth?

What can you do now?

- Ask if giving your data is required
- Read the terms and conditions to understand what is done with your data
- Clear cookies
- Reflect on what activity patterns define you (PII)
- Think about what data you leak across software - try to compartmentalize, make it difficult to join your data from different sources
- Reflect on the decisions you make

Research I have been doing

1. Find incentives for companies to move to privacy conscious tools
2. Create tools for users to regain control over their data and agency over their lives

Incentivizing the use of privacy preserving tools

1. How do we define privacy? - differential privacy makes an attempt
2. What are the drawbacks? What must a company give up to make the transition? What types of applications would be unaffected today?

Differential Privacy - How it works

Key idea: learn about the population not about any specific individual.

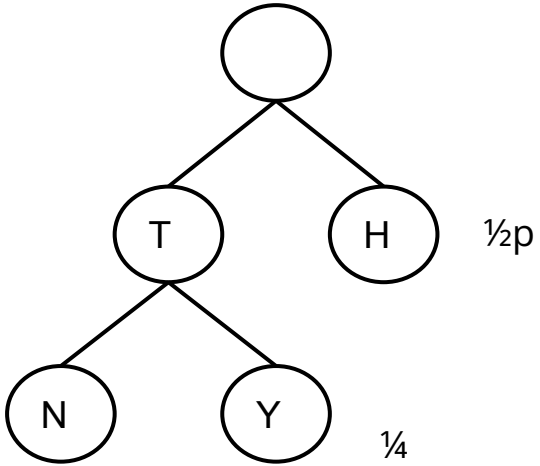
Many ways to do this. One way is to store data in the following way:

1. Ask user A if they have attribute B (example: do you smoke etc.)
2. Toss a coin.
3. If Heads then answer honestly.
4. If Tails, then flip again.
5. If Heads then answer “Yes”
6. If Tails again then answer “No”.

Differential Privacy - How it works

Assume true proportion of Attribute B is p. How do we estimate p?

What proportion of Attribute B (call it p_{obs}) do we expect?



So $p_{obs} = ½p + ¼$

Meaning

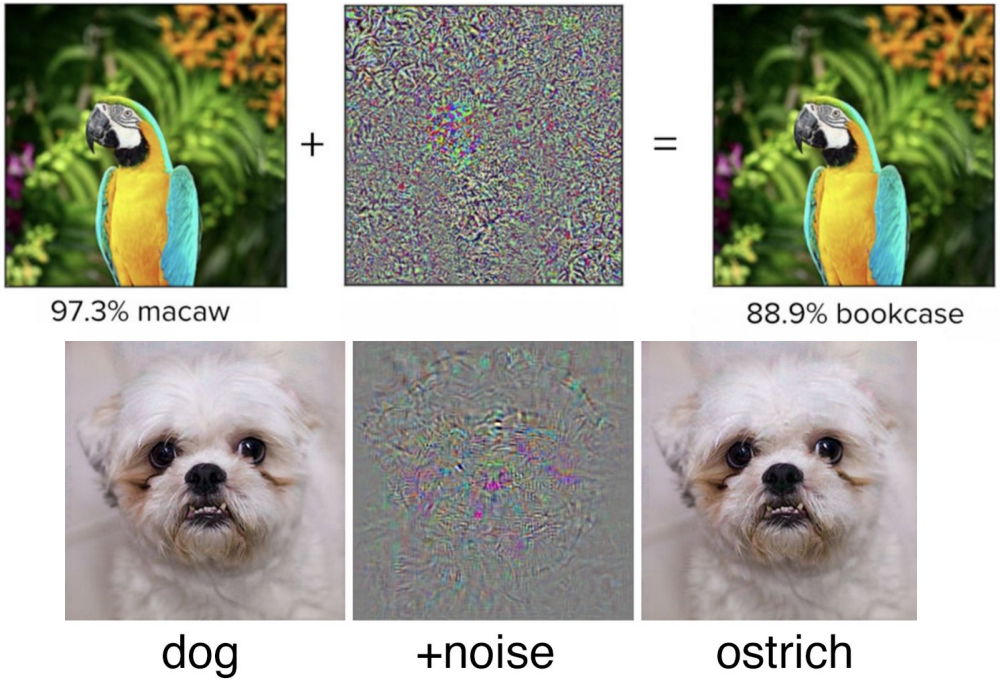
$P = 2 * (p_{obs} - ¼)$

Differential Privacy - Limitations

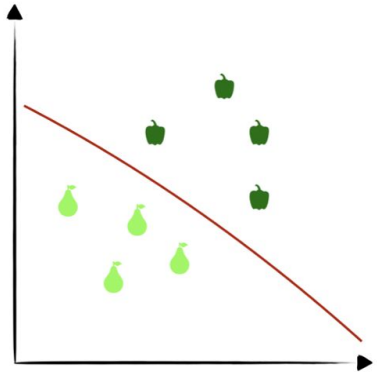
1. Generally complex: needs more simple examples to explain how it works (this is where you can contribute today!)
2. No consensus on how to calibrate the odds of the coin flip in the previous example (every company uses a different value and doesn't usually disclose it - ex: apple)
3. Does not work if you're looking to analyze outliers or do anomaly detection
4. Does not work well on small datasets

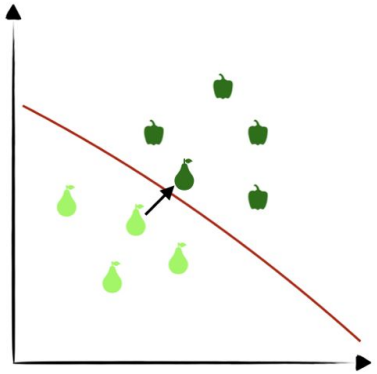
Privacy Tools for the Average User

Fooling Classifiers by creating adversarial examples.



Privacy Tools for the Average User





In order to fool a given model, we need a process by which we can move data slightly passed the decision boundary

More Links

- https://www.heinz.cmu.edu/~acquisti/papers/AcquistiGrossStutzman-JPC-2014.pdf
- https://www.heinz.cmu.edu/~acquisti/papers/Acquisti_Welfare_Impact_of_Targeted_Advertising_WP.pdf

Some Advice For Job Searching

1. What are your principles? What will you be uncompromising about? What gets you fired up? What do you believe in?
2. Don’t change yourself to meet what you think the company wants to see
3. When you interview you are also interviewing them
4. Interviews are often more about “fit” than they are about technical competency (hence the importance of 1 - 3)
5. For your first job it’s good to have great mentors to help you learn
    a. How senior is the team you’ll be working with?
6. You’re not an imposter - no one knows everything. Be transparent about what you do and don’t know. Be willing to learn.