Lecture 1: introduction and administrative details

Lecture PETs4DS: Privacy Enhancing Technologies for Data Science

Dr. Benjamin Heitmann and Prof. Dr. Stefan Decker
Informatik 5
Lehrstuhl Prof. Decker

RWTH AACHEN UNIVERSITY

Outline

- Examples
- Motivation for lecture
- Administrative details
- Overview of topics

High-impact example 1: Email hacks during run-up to US election

Security or privacy?

- At least 20.000 emails where leaked by a hacker called “Guccifer 2.0” in June 2016
- The emails where from the Democratic National Committee
- They showed that Democratic party favored Hillary Clinton over Bernie Sanders long before primaries started
- Multiple dumps of emails
- Potential connection to Russia
- Clear political impact

High-impact example 2: DDoS attacks triggered by IoT Botnets

Definitively security related! But also no privacy without security!

- On 21st October 2016 several major web sites like Amazon, Twitter, Spotify, Netflix and Paypal where knocked of the web
- Reason: Distributed Denial of Service attack on DNS provider Dyn
- Majority of traffic generated by Mirai botnet
- 10s of millions of IPs involved
- Most from Internet of Things (IoT) devices like cameras and routers using XiongMai hardware.
- IoT devices are hard to patch: old exploits and default passwords are easy to use.

Details at http://arstechnica.com/information-technology/2016/10/inside-the-machine-uprising-how-cameras-dvrs-took-down-parts-of-the-internet/

High-impact example 3: CIA versus Apple iPhone unlock protection

Security and privacy related?

- In March 2016, Apple refused to help FBI access data from a locked iPhone.
- The iPhone belongs to the deceased shooter of the San Bernardino shooting
- Many big players in silicon valley have written letters of support: Amazon, Google, Facebook, Snapchat, Twitter, Microsoft

- **Apple's main argument:** if Apple helps the US government today in overcoming the built-in security measures of the iPhone, this will set a precedent in the US and internationally.

Security versus Privacy

- Another take on Apple
- Shows how intertwined security and privacy sometimes is.

https://youtu.be/zsjZ2r9Ygzw?t=921

High-impact example 4: Netflix data de-anonymisation law suit

Only privacy related?

- Netflix data set was most important data set in recommender systems research.
  - 100m ratings, 500k users, 17k movies.
  - Anonymised by removing personally identifiable information
- Netflix released the data for a 1 million USD prize contest over 3 years starting in 2006.
- In 2008, researchers showed they could de-anonymise the data and identify themselves.
- This resulted in an expensive law-suit and attention from Federal Trade Commision (FTC).
- Netflix shut down the sequel of the contest and never again released data for research.

Details in Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). 2008." *University of Texas at Austin* (2008).

High-impact example 5: Target pregnancy advertisements

Related to privacy and data science

- Target is a large US chain of supermarkets
- Target creates a profile for every user based on credit card and shopping history

- Pregnancy prediction score based on 25 products, e.g.: scent-free soap, bags of cotton balls, supplements, hand-sanitizer, wash cloth.
- In 2012, in Minneapolis, one angry father of a high-school girl, demanded to talk to store manager.
- His daughter got coupons for baby clothes.
- She was pregnant, but had not told her father.
- Target new about it before the father was told.

Details in http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

Intuition about difference between security and privacy

- **Security is about protecting data from unauthorised access**
- **Privacy is about how you can use data after you acquire it**
  - **Acquiring the data can happen from open sources as well!**
- We will look at this difference in more detail later

Brainstorming session

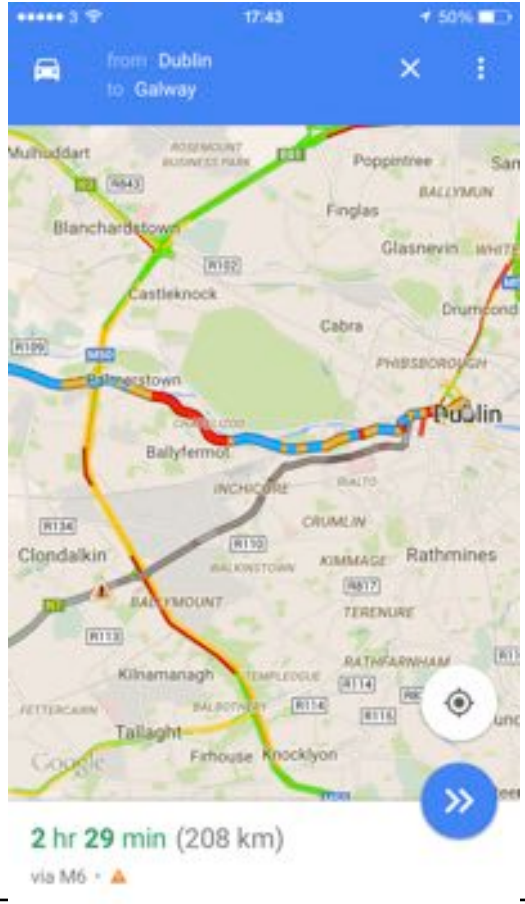
In your own words: why is privacy important today? What is the connection to data science?

Motivation

The elephant in the room: everybody in computer science / data science is doing surveillance today.

- Data science and privacy are fundamentally opposed
- “Surveillance is the business model of the internet” – Bruce Schneier
- There is always a trade-off: utility versus privacy
- Companies have **intimate knowledge** about users.
- Users are worried about privacy but feel powerless

Examples from Facebook and Google

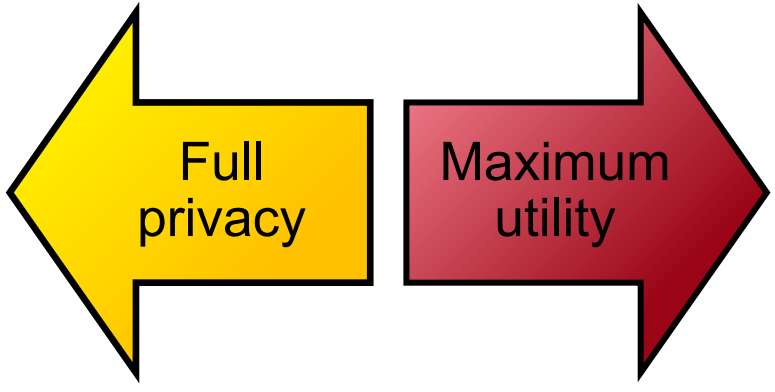


Motivating problem for the lecture

How to open up the spectrum between full privacy and maximum utility?

- Today consumers have two choices
  1. **Maximise utility:** give all their data to the service
  2. **Maximise privacy:** do NOT use the service

- **We need PETs: Privacy Enhancing Technologies**
  - PETs open up the spectrum between the extremes
  - Allow services to select level of trade-off
  - Develop business models incorporating all parts of the spectrum
  - Ultimately provide more choices to consumers



Focus of lecture: PETs for Data Science

- Great time for research on PETs, as demand is rising.
- Overview of approaches which “add privacy to big data”.
- Lecture provides foundation for research and development of privacy-enabled alternatives.

- **Good solutions exist for:**
  - Secure channels
  - Anonymisation of data
  - **Data mining:** using aggregated and anonymised data to create insights

- **Active research topics:**
  - PETs for machine learning
  - PETs for personalisation
  - Requires using personal profile in addition to aggregated data

What is Data Science?

- Data Science is an umbrella term.
- **Processes and systems to extract knowledge or insights from data.**
- This includes approaches such as:
  - Data mining: discover patterns in large data sets.
  - Machine Learning: give computers ability to learn without being explicitly programmed".
  - Personalisation: adapt a service or product to the preferences of a specific individual or group.
- Scalability for "big data" is usually a requirement.
- **You can think of data science as "the science behind big data".**

Data Science

Privacy definitions: hard versus soft privacy

- **Soft privacy:**
  - Main idea: user has lost control of personal data already.
  - User has to trust honesty and competence of data controller.
  - Data controller has to protect the data.
- **Hard privacy:**
  - Main idea: data minimization
  - User provides as little data as possible.
  - Reduce the need to trust other entities in the system.
  - Empower user to protect this data.

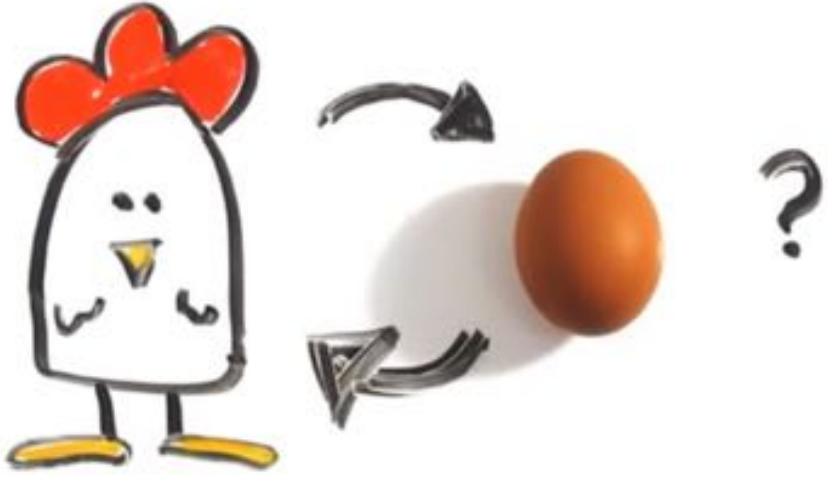
Use cases for Big Data in which privacy is relevant today

- Internet of Things (IoT) and cyber-physical systems
  - Home automation is entering consumer market: light switches, heating control.
  - Smart cities: same idea, bigger scale.

- Advances in the life sciences:
  - Personalized medicine requires patient data
  - Patient data is subject to very strict legislation, usually can't leave country.
- eLearning:
  - Parents expect data about learning behavior of children to be protected.
- Smart consumer experiences:
  - "Safe Harbour" agreement will be replaced by "Privacy Shield" agreement

Business models based on PETs in Data Science

Where is the money?



- Majority of current business based on free services and maximalist data collection
- Research on alternatives has been missing until recently
- Hen and egg problem

- New incentives for privacy-enabled alternatives are emerging:
  - Customer demand
  - Legislation
  - Use cases with strict privacy requirement

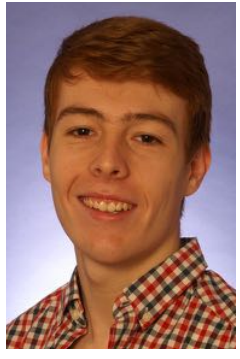
Administrative details

The team behind the lecture

Prof. Dr. Stefan Decker

Dr. Benjamin Heitmann

Felix Hermsen



Carsten Stoffels

Contact

- Location:
  - Chair Informatik 5
  - Ahornstraße 55, Building extension E2, 2nd floor
- Web page of the chair:
  - http://dbis.rwth-aachen.de/
- Contact via email: heitmann@dbis.rwth-aachen.de
- L2P:
  - https://www3.elearning.rwth-aachen.de/ws16/16ws-51481/
  - Registration for lecture via campusOffice
  - All materials and information about lecture will be provided via L2P

Details about lectures and recitations (“Übung”)

- Always in room 5053.1
- Two weekly time slots:
  - Tuesday, 16:15 to 17:45
  - Thursday, 16:15 to 17:45

- 12 Lectures and 6 recitations
- Irregular schedule, usually 2 lectures followed by one recitation.
- Please always check L2P for current schedule!
- Modifications to schedule can happen and will be announced on L2P.

- First lecture on 27.10.2016
- Last event currently scheduled for 9.2.2017

Recitation (“Übung”) and assignments (“Übungsblätter”)

- Assignments are usually published one week before the associated recitation
- Assignments do not have to be submitted.
- Assignments are not graded.
- Dr. Heitmann will show how to solve the assignments in the recitation
- In addition, the recitation often provides the opportunity for students to show how they solved the exercises.
  - Exercises will indicate if students can present their solutions.

- Assignment types will include:
  - Analyse and discuss
  - Perform algorithm on paper
  - Programming exercises

Preliminary schedule for October and November

- Thu, 27.10.2016: lecture
- Thu, 3.11.2016: lecture
- Thu, 10.11.2016: recitation
- Tue, 15.11.2016: lecture
- Thu, 17.11.2016: lecture
- Thu, 24.11.2016: recitation
- Tue, 29.11.2016: lecture

- Please check L2P for up to date schedule and further dates.

Exam and presence exercise (“Präsenzübung”)

- Written exam on two dates. Dates have to be decided.
- **In order to be eligible for the exam, you have to pass the presence exercise.**
  - In order pass presence exercise more than 50% of points are required.
  - The presence exercise will cover the contents of the first half of the lecture.
  - Style of presence exercise is mostly multiple-choice and exercises based on assignments.
- The presence exercise will have two dates, in December and/or Januar. Dates will be announced soon.
  - In case of fail / illness / time conflict, retest at second date is possible.
- Both the presence exercise and the exam follow the style of the assignments.

Other details

- The language of the lecture is English.
  - This includes the assignments, the recitation, the exam and all interactions between the team and the students.
- The lecture and the recitations will not be recorded.
- Literature:
  - We use a mix of recent book chapters and papers from conferences and journals of the last few years as primary source material.
  - Sources for each lecture will be given as references.

Your contribution / What we expect from you

- The lecture gives a snapshot of an active area of research
  - That is why there is no single book covering the lecture
- We expect you to actively engage with the topic:
  - Take a look at the related sources
  - Work on the assignments
  - Optionally: present your solutions during the recitation
- Show that you understand the big picture and the high level connections
- Show that you can solve the exercises from the assignments
- Develop your own opinion based on your understanding of the area

Open questions:

- Should there always be at least one week to work on one assignment?

- When should the presence exercise be?
- How long between dates?
  - last Thursday before Christmas?
  - first week after Christmas
  - late January?

Overview of topics

Overview of contents of the lecture

1. Definitions of privacy and security.
2. Modelling of privacy threats.
3. Achieving and measuring privacy through statistics.
4. Approaches to anonymization and de-anonymization of data.
5. Computation on encrypted data.
6. Approaches for using encrypted data for data science.
7. Hiding of user queries and data in cloud computing.
8. Compromises between full surveillance and full privacy.
9. Privacy by design as a cross-cutting software design approach.

Depending on progress of lecture, we might look at additional topics as well, such as societal issues (law, business models and ethics).

Topic details: Privacy, security, threat models

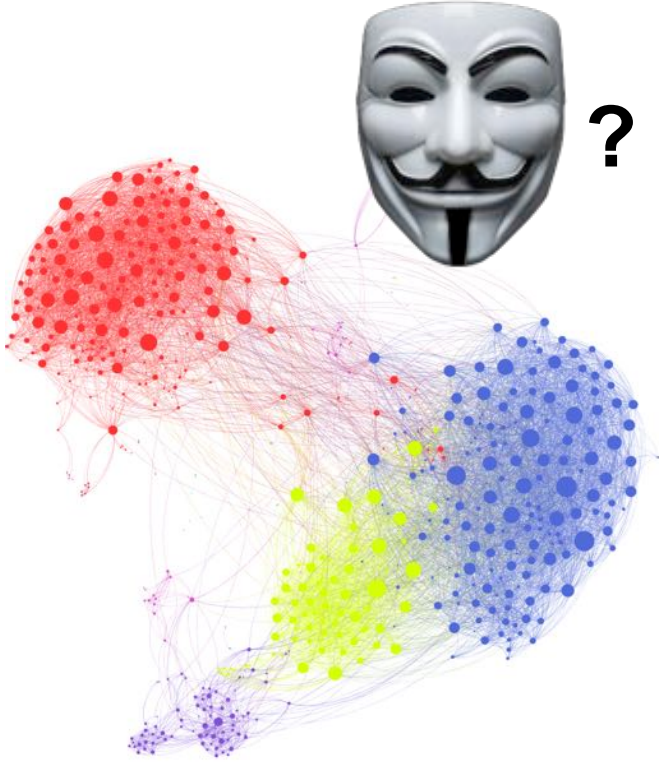
- Privacy and security are different. Sometimes even opposed.
- Definitions from inside and outside computer science.

VS.

- Threat models:
  - Provide common vocabulary to talk about problems / leaks / breaches
  - Security threat model: STRIDE
  - Privacy threat model: LINDDUN
  - Learn to use threat models to analyse existing apps / web sites.

Achieving and measuring privacy through statistics

- What are the state of the art approaches to anonymise data?
  - Without using any kind of encryption!
- Anonymity metrics like k-anonymity
- Learn to apply this to examples.
- How does this apply to different data models?
  - Relational data
  - Statistical data
  - Graph data



Computation using encrypted data for Data Science

- Idea: work with encrypted data without decrypting it first.
- This could allow e.g. user profiles to be encrypted.
- Basics of the encrypted cloud.
- Differences between existing approaches:
  - Architecture requirements
  - Complexity
  - Generality
  - Maturity
- Learn how to apply a simple baseline approach to example data.
- Understand when to apply these approaches and when not.

Perspective of this lecture on cryptography

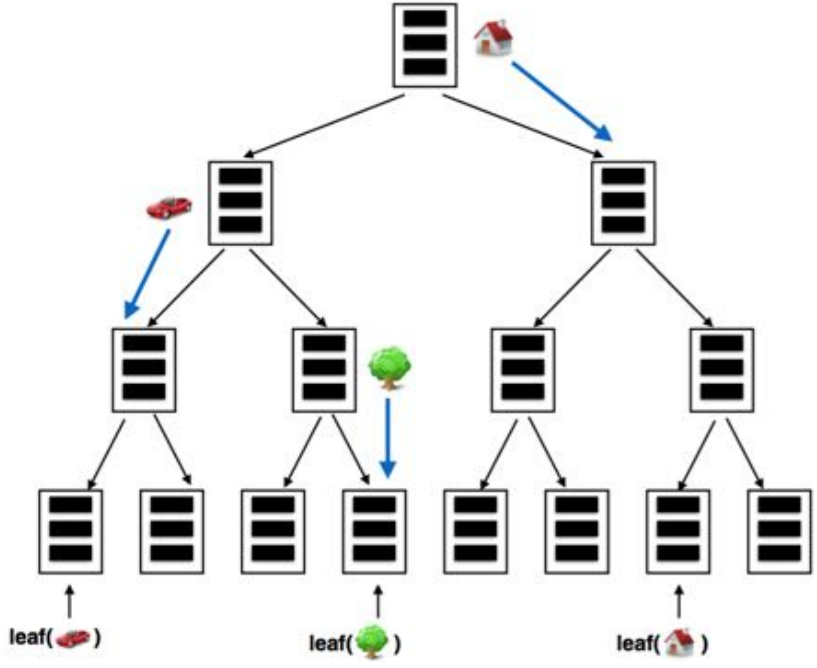
Our view: use cryptography as a black box

How we analyze an approach:

1. Identify threat model addressed by approach
2. Identify requirements for approach,
3. Identify inputs and outputs
4. Identify computational complexity of approach
5. Identify maturity of approach and of software tools implementing approach
6. Discuss suitability of approach for Data Science

Hiding of user queries and data in cloud computing.

- The cloud introduces new adversaries.
  - Cloud operator could run your code in debugger.
- Is it possible to hide user queries from a database?
  - Can you still have access rights enforcement?
- Is it possible to hide content of data structures like binary trees?



Compromises between full transparency and full privacy

- Encrypting everything has the potential to lead to an unaccountable society without trust between anybody.
- If “everybody” agrees that something should be decrypted, what should happen?
  - Phones of dead terrorists?
  - Data storage of convicted criminals?

- Are backdoors a good idea?
- Are there better alternatives which can restore trust and mitigate abuse?

Privacy by design

- Introduces privacy as a cross-cutting concern for the whole application.
- Addresses the user experience.
- Can be used as a guideline when re-engineering existing applications or designing new ones.
- Shows how the different topics of the lecture fit together.

Other opportunities for you to engage with the topic

- **Thesis topics** available upon request:
  - implement some form of data mining, machine learning or personalisation
  - Use privacy by design to re-engineer existing state-of-the-art approaches
  - Implement PETs such as SMPC, Homomorphic encryption, Oblivious Data Structures, Blockchain, Private Information Retrieval
  - Use cases: medicine and IoT
- **Seminar** “Privacy and Big Data” in Summer Semester
- **Hiwi Jobs:** C++ and Java.