Big Data and Big Privacy

Finding an equilibrium

Francesco Picciotti

28th November 2017

Computer Ethics (2017-18), Politecnico di Milano

Central idea

The aim of the presentation is to show that:

- Privacy cannot be **downplayed** or traded for having **better services**
- Existing solutions are **not enough**
- Propose **a new perspective** for finding new solutions, as a constructive tradeoff in which both sides **collaborates** to achieve **convenient results** for both

Outline

1. Introduction to Big Data
2. Case study: The Big G services
3. The Big Issue
4. The Big Challenge

Introduction to Big Data

The reason of the rise

The Big Data happening is a **combination** of two technologies growth during the last 30 years:

- A significant paradigm-shift of AI
- Development of a big and unified data infrastructure

As a matter of fact, the huge amount of data daily produced by us is now easily **stored** and **available for other purposes**.

The Devil's in the... numbers

- Amazon.com handles millions of back-end operations every day, as well as queries from **more than half a million third-party sellers**. [Layton 2013]
- Facebook handles **50 billion photos** from its user base. [Johnson 2010]
- Google was handling roughly **100 billion searches per month** as of August 2012. [Sullivan 2015]

Case study: The Big G services

What is happening?

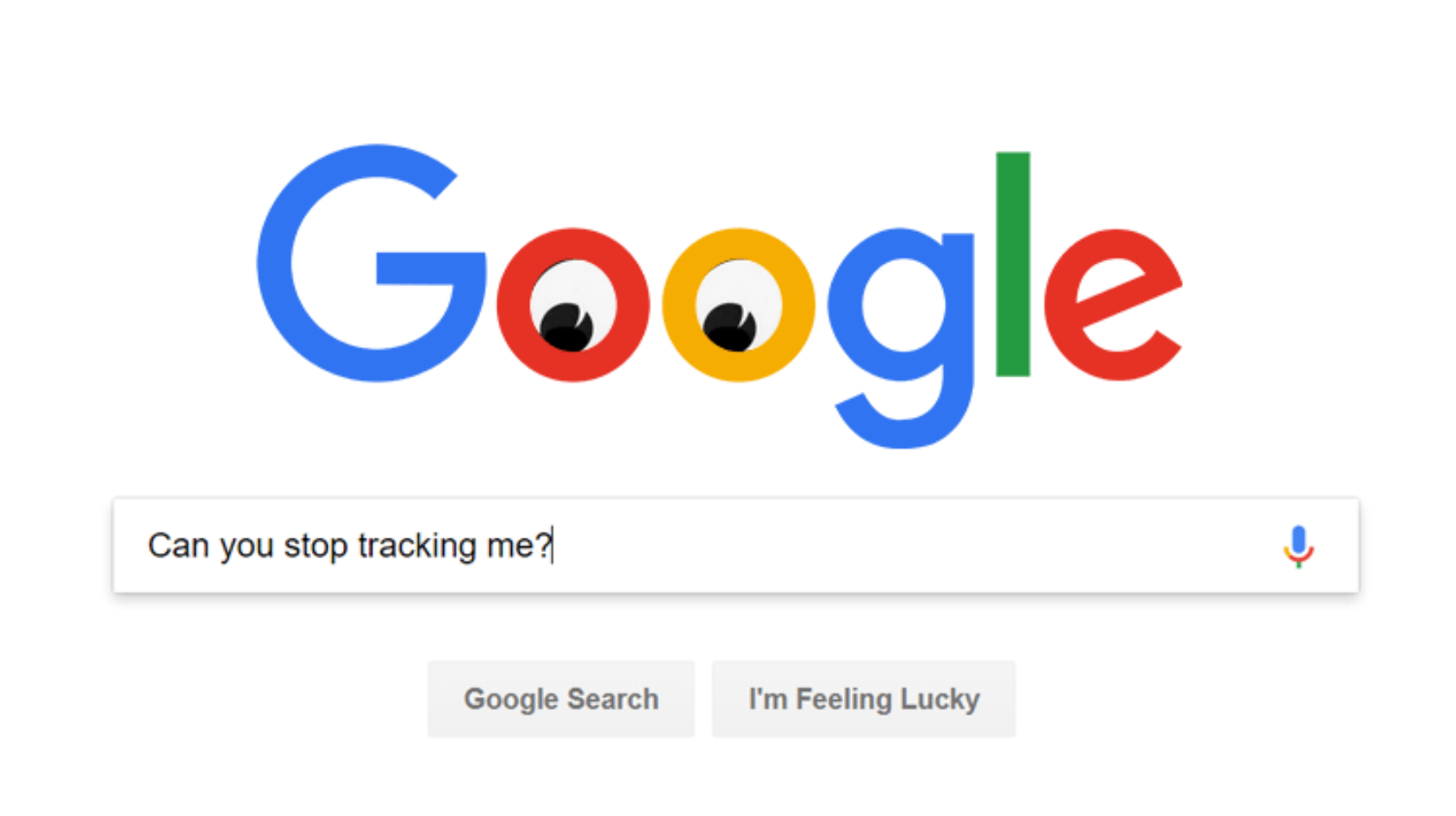


Figure 1: From [Sotiri 2017]

Behind the daily interaction with Google

Without any doubt Google is the most pervasive company, just think about of the **daily usage of services** in our **smartphones**.

Search Engine

The most known SEO stores permanently our **web searches**, monitoring our **interests** and **behaviors**.

Gmail

When we use the company's email services, Google **scans our emails**, as well as the **recipients**.

Behind the daily interaction with Google (cont'd)

Google Maps

If you have location history enabled then Google knows the **places that you hang out** or where you **travel to**.

Google Photos

When you upload your photos, you are giving the tech giant license to *"host, store, reproduce, modify, create derivative works, communicate, publish, publicly perform, publicly display and distribute"* [Google's Terms of Service] those photos

...and there are **plenty other services** and they are all for free.

The Big Issue

What is the problem?

It looks like we have to allow the **collection of our personal informations**.
One may say:

- The Google's data gathering implies a **significant improvement** of service
- The disclosure of personal information is **restricted**
- **Overconcern** about privacy

The problem is the Price

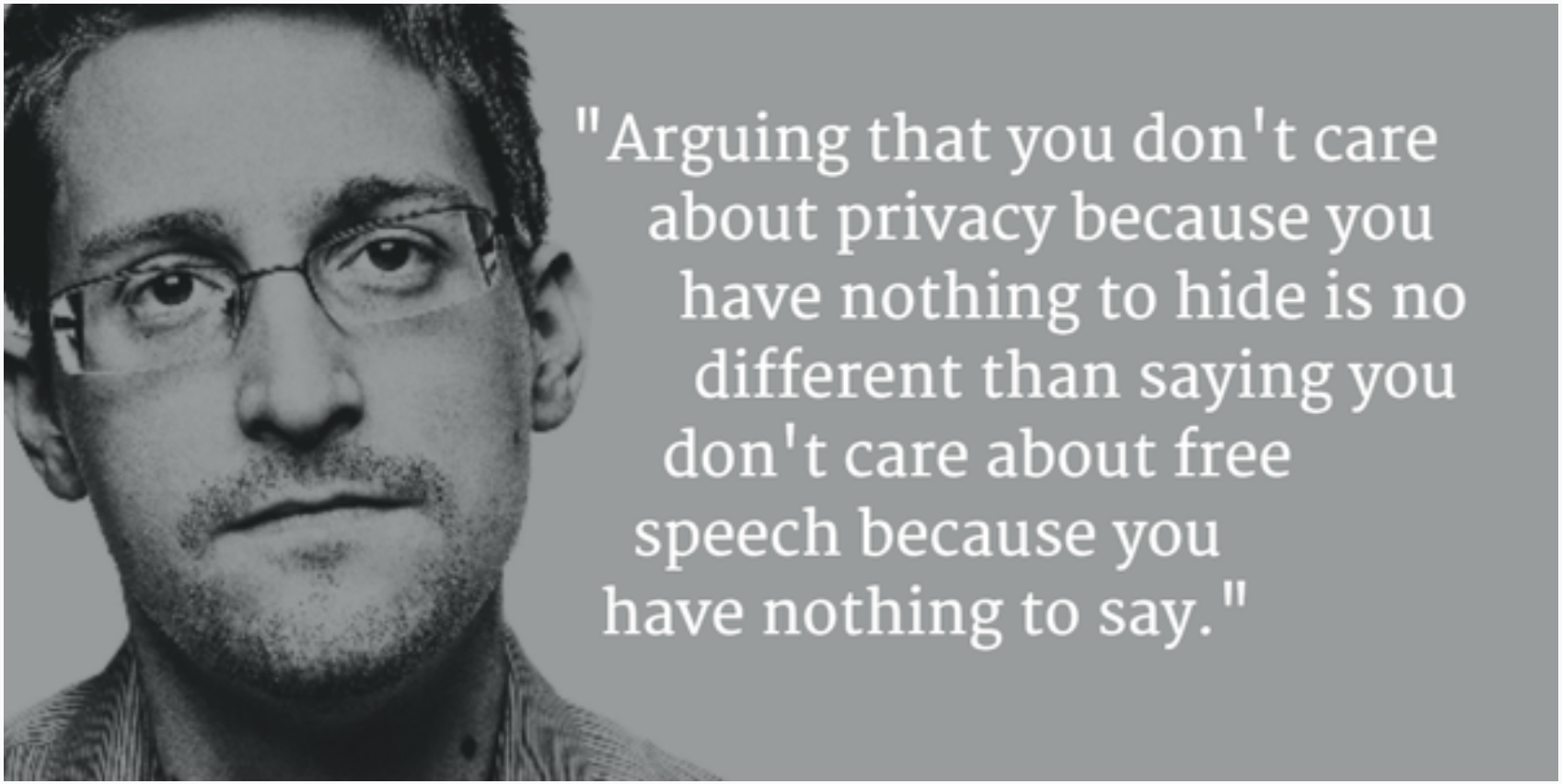


Figure 2: Edward Snowden's quote

The Price

The main reasons about the **preciousness** of our personal informations are conveyed by [Rachels 1975] :

- Protect people's interest
- Hide aspect of life or behavior that would be **embarrassing** to disclose
- Confidentiality
- Avoid to be judged with **unrelated** facts
- Control of the access to **personal information**

The social valence

The relevance as **social good** depicted by the several aspects [Regan 2015] :

- Common value, reframing privacy towards an **utilitarian view** (i.e. Privacy vs Security)
- Public value, necessary to **support democratic processes** and to the **forming of a body politic** or public (i.e. Ads targeting during Trump presidential campaign)
- Collective value, privacy **protecting the common pool** of personal informations (i.e. Data Breach)

The Contextual integrity valence

It is even clearer once considered as Contextual Integrity [Nissenbaum 2004], information flows according to:

- **Key actors**
- **Types of information**
- Constraints under which flow occurs (**Transmission**)

The miracle cure?

The existing ways to shield the privacy of users, according to [Barocas and Nissenbaum 2013], are:

- Anonymity: **Hide identities** (**PII**, thus Personally Identifiable Informations) from the records using a **unique persistent identifier** (i.e. Google's anonymous identifier is AdID)
- Informed Consent: Make users informed **who is collecting data**, if the **data collection is compulsory**, how **information** is **used** and **shared**

Limits

Anonymity

- Linkage Attack: An attacker can retrieve the identifying information **joining** the anonymized dataset with another one with identities
- Differencing Attack: Performing a sequence of queries on anonymized dataset, the attacker **can deduce** a certain person **is** in the dataset
- Pseudonymity: Even without knowing the **person's name**, company knows the user's behavior and tastes

Limits cont'd

Informed Consent

- Transparency Paradox: Simplicity and clarity results in **losses of fidelity**, therefore they have some degree of **obscurity** to hide violations
- Unpredictable: There are no **guarantees** on how much **time** the information will flow, who will **use** and work on these data
- Opt-out conditions, in order to use the service the user must **accept** the terms

The Big Challenge

The situation

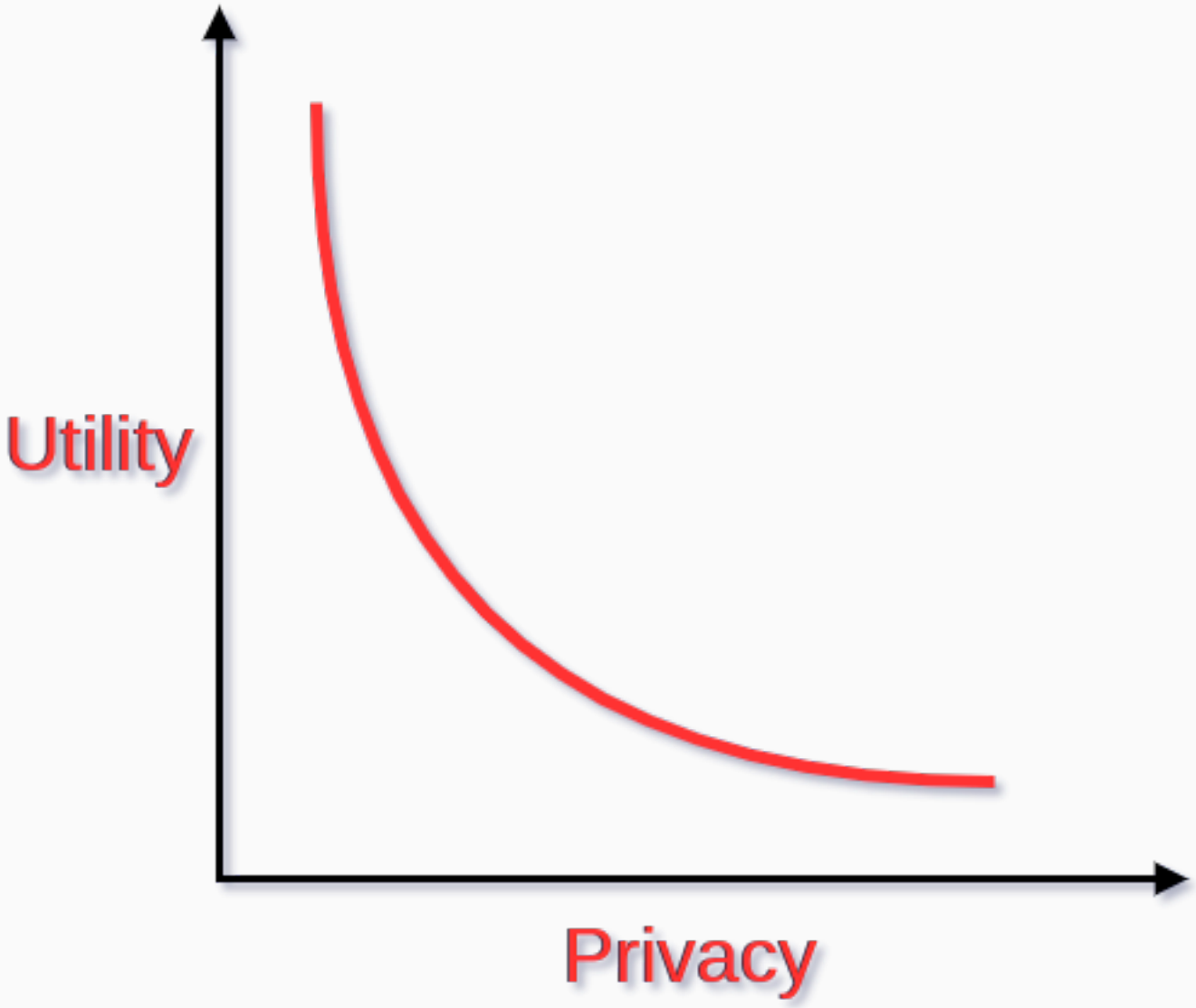


Figure 3: Inspired from [Preneel 2015]

Pushing the boundary up

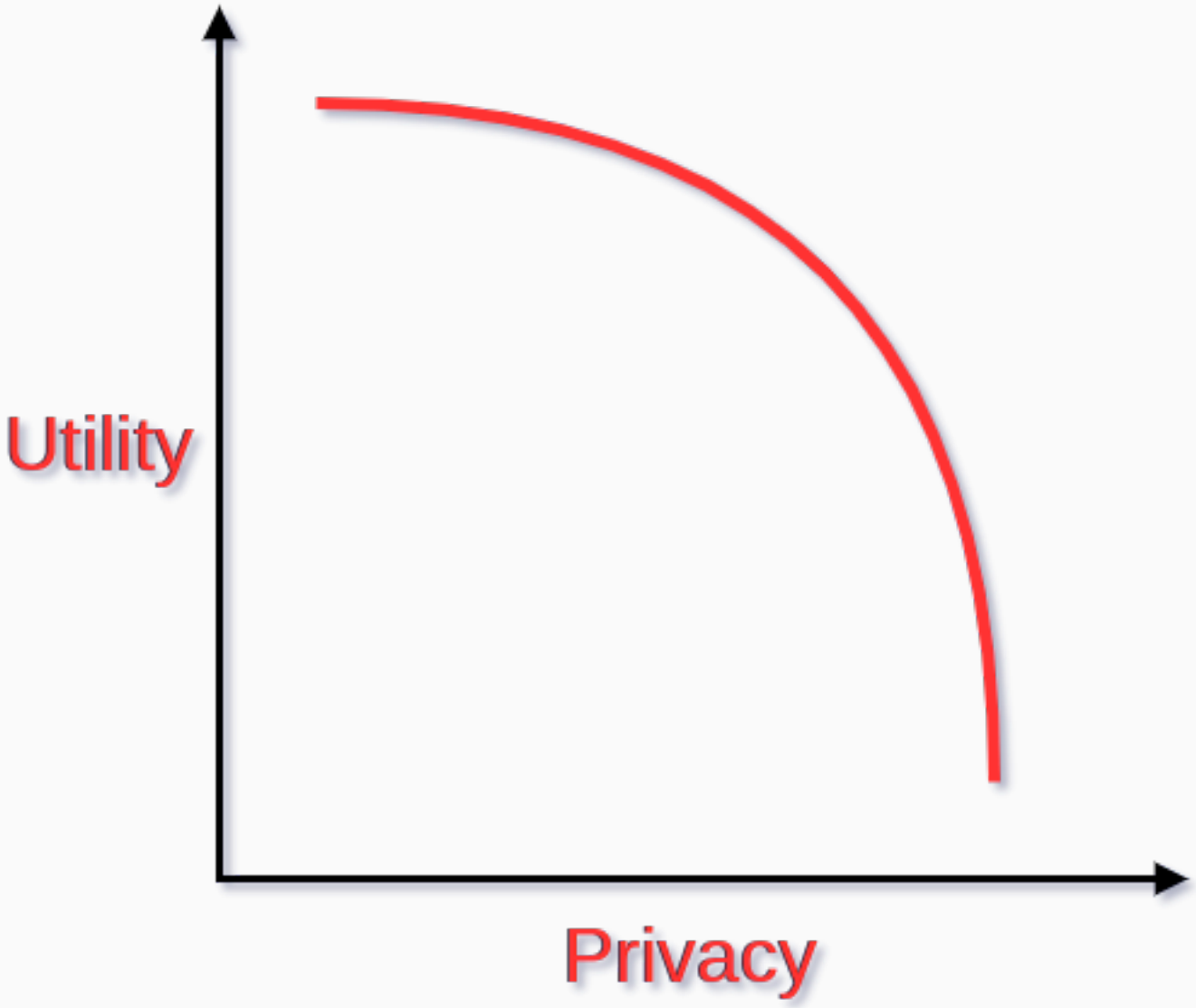


Figure 4: Inspired from [Preneel 2015]

Est modus in rebus: From the company's perspective

- Anonymity implementations has to be **improved** and also apply **data perturbation**
- The Informed Consent has to be at least clearer, but generally it is **inadequate**
- Reduce the damages of data breaches and leaks avoiding a **single point of failure**, using:
  - **Segmentation**, switching from big data to **small local data**
  - **Encryption** of data
- The willingness to give **some profit** up and put more effort on finding **concrete** and **adequate** solution

Est modus in rebus: From the user's perspective

- Basic notions of privacy and **how to protect it** (i.e. How to prevent Google tracking on any browser [Sotiri 2017])
- Increased **responsibility** about the usage of some tools or services
- More **awareness** on the existence of other **alternatives**
- **Consciousness** of the informed consent in which we "agree" on
- **Limiting the usage** of some technologies depending on the purpose (i.e. GPS tracking activities)

The Devil's in the… details

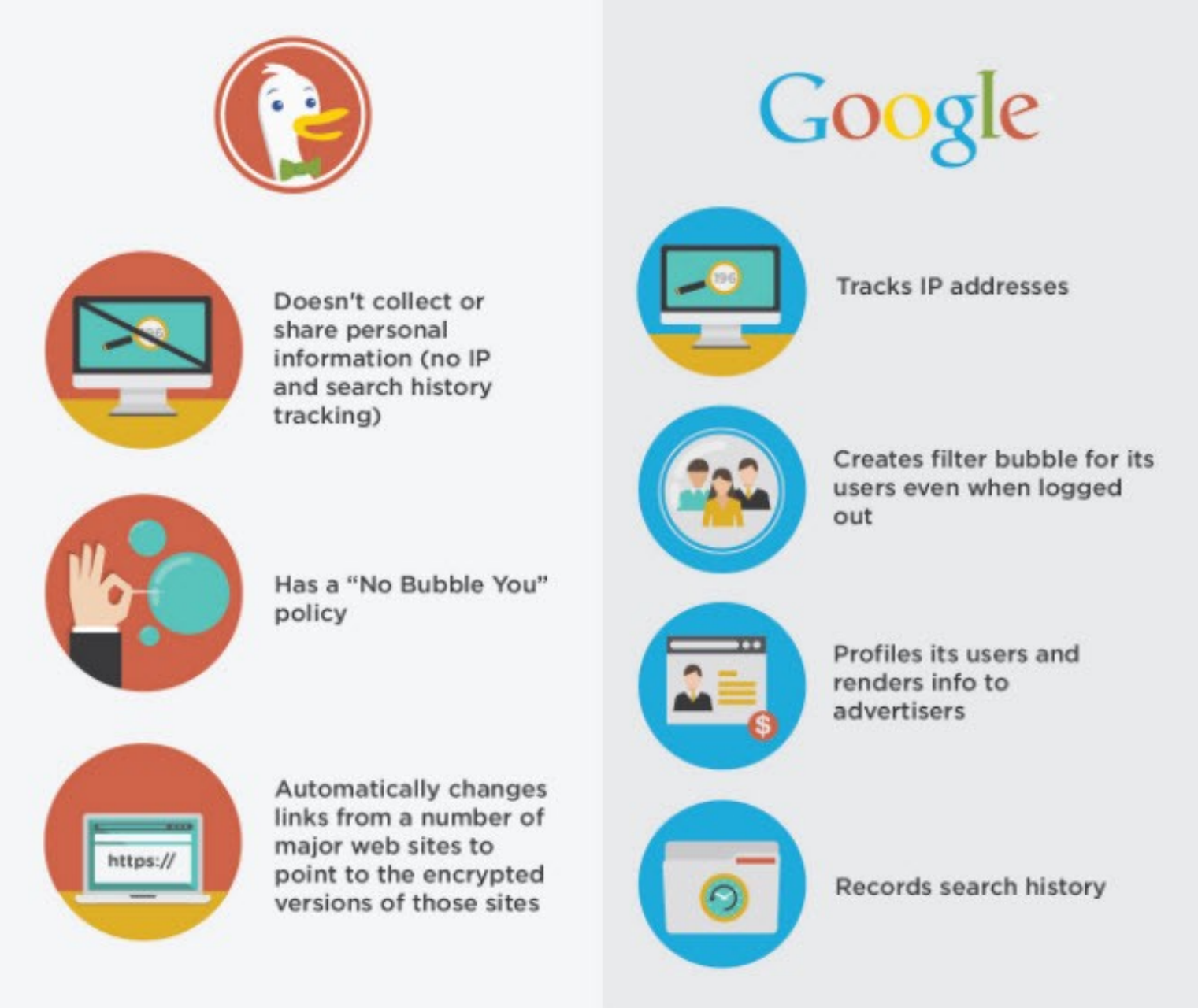
4. Personalise your Google experience

From better commute options in Maps to quicker results in Search, Google tools and services get faster and more useful with the activity data you let us save with your Google Account.

Google currently saves the following information (which is visible only to you):

- ✓ Web & App Activity
- ✓ Location History
- ✓ Device Information
- ✓ Voice & Audio Activity
- ✓ YouTube Search History
- ✓ YouTube Watch History

The need for an alternative



”Be the change that you wish to see in the world.”

Mahatma Gandhi

Thanks for your attention

Questions?

References

- Barocas, Solon and Helen Nissenbaum (2013). “Big data’s end run around anonymity and consent”. In: pp. 44–75.
- Johnson, Robert (2010). *Scaling Facebook to 500 Million Users and Beyond*.
- Layton, Julia (2013). *Amazon Technology*.
- Nissenbaum, Helen (2004). “Privacy as Contextual Integrity”. In: *Washington Law Review*.
- Preneel, Bart (2015). “Big data and little privacy: there is no alternative?” In: URL: `https://www.youtube.com/watch?v=uYk6yN9eNfc`.

- Rachels, James (1975). “Why Privacy is Important”. In: *Philosophy and Public Affairs*.
- Regan, Priscilla M. (2015). “Privacy and the common good: revisited”. In: *Social Dimensions of Privacy: Interdisciplinary Perspectives*.
- Sotiri, Elena (2017). *How To Prevent Google Tracking On Any Browser*.
- Sullivan, Danny (2015). *Google Still Doing At Least 1 Trillion Searches Per Year*.