Privacy vs. Data: Business Models in the digital, mobile Economy

Lecture 7 + 8
Personal Data Collection & Usage

WS 2011/2012

Dr. Andreas Albers
www.m-chair.net

Agenda

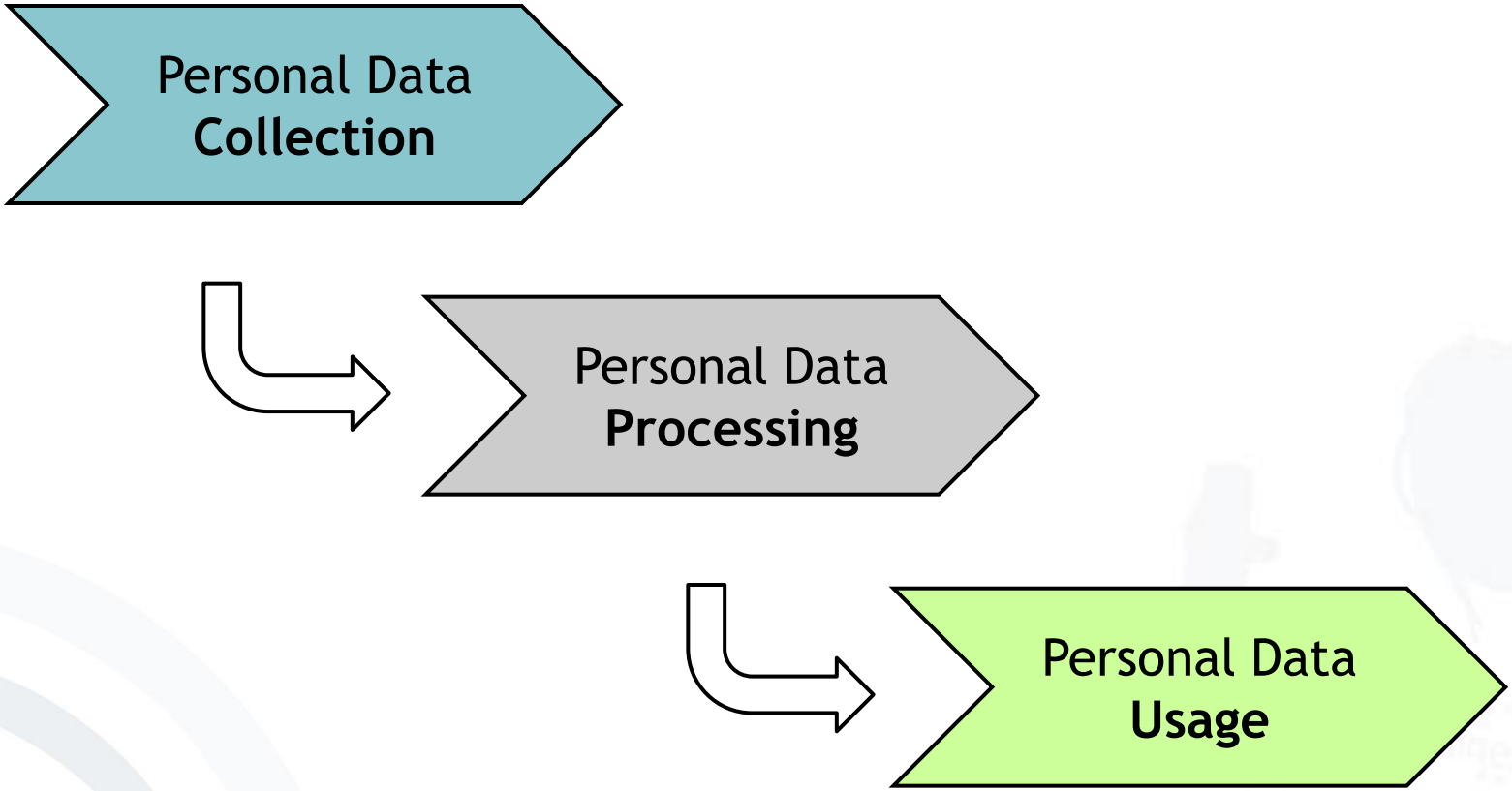
- Introduction
- Means of User Identification
- Personal Data Collection
- Personal Data Processing
- Personal Data Usage

Personal Data in the European Union

Personal data (EU Directive 95/46/EC)

- *'Personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;*

Steps towards Commercial Personal Data Usage



Requirements for Personal Data Collection

- Availability of a *unique identifier* for an individual (e.g. IP-address)

- Access to personal data by
  - observation of the behaviour of individuals (i.e. “Tracking”)
    → *Implicit Data*
  - deliberate disclosure of personal data through individuals (i.e. “User-generated Personal Data”)
    → *Explicit Data*

Technological Enablers for Personal Data Collection

- Online/Mobile Web Logs/Trackers
  - e.g. Google Analytics, Yahoo Tracker
- Online/Mobile In-Services/Apps Logging
  - e.g. creation of log files during service usage
- OS-Build-in Logging functionality (Desktop/Mobile OS)
  - e.g. log files generation during service usage
- Spyware (excluded from course)
  - e.g. unwanted software secretly installed along a regular software installation process
- …

Parties involved in Personal Data Collection

- First Party (owner/operator of a service or software)
  - e.g. Amazon tracking visited products on their website by individuals

- Third Party (organisation unrelated to a service, but enabled to collect personal data)
  - e.g. Google Analytics tracking visited news articles on CNN.com

Commercialisation Areas of Personal Data

- Targeting of Advertisements

- Product Recommendations

- Service/Content Personalisation

- …

Agenda

- Introduction
- Means of User Identification
- Personal Data Collection
- Personal Data Processing
- Personal Data Usage

Identifying Individuals Online: Cookies

- Cookies
  - are small text files stored on Desktop PCs and mobile devices
  - have multiple purposes (e.g. session creation, storing of user configuration or personalisation information),
  - and allow user identification

- For *user identification,* a globally unique identifier (GUI) (e.g. number) is stored in the cookie
  - If a user accesses a website, the GUI is collected by the server and referenced to a (maybe) existing user profile stored on the provider's server

- Special types of cookies
  - Flash or Silverlight cookies
  - Used for Flash (Adobe) or Silverlight (Microsoft) applications, but much more persistent that regular browser cookies

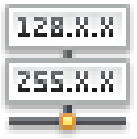
- EverCookie (very, very special type of cookie)
  - See for yourself: http://samy.pl/evercookie/ ;-)

Identifying Individuals Online: Web Storage

- Web Storage (as part of HTML 5)
  - Similar to cookies, but with more storage capabilities and a more sophisticated interface
  - Up to 10 MB storage space per domain

- The storage of (personal) user data is actively supported by the W3C

Identifying Individuals Online: IPv6 Addresses

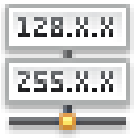


- Internet Protocol (IP) based addressing
  - Every host and router on the internet has an IP address.
  - An IP address is unambiguous. Two computers cannot use the same (public) IP address at the same time.

- IPv4 (currently used)
  - Example: 192.168.133.47
  - IPv4 supports 2^{32} = 4,294,967,296 addresses,
  - But: There are no more unallocated IPv4 internet addresses left!
  - Solution so far: On reconnect to the service provider, a new (dynamic) IP-address is assigned to the client (exception: static IP-addresses)

Identifying Individuals Online: IPv6 Addresses (2)

- IPv6
  - An IPv6 address is consists of 128 bits (instead of 32-bit).
  - The new IPv6 address space supports 2^{128} addresses = 340,282,366,920,938,463,463,374,607,431,768,211,456



- Problem? Let's see …
  - *"So we could assign an IPV6 address to EVERY ATOM ON THE SURFACE OF THE EARTH, and still have enough addresses left to do another 100+ earths. It isn't remotely likely that we'll run out of IPV6 addresses at any time in the future."* (Steve Leibson)
  - →Every individual could theoretically get a unique IP-address at birth for life.

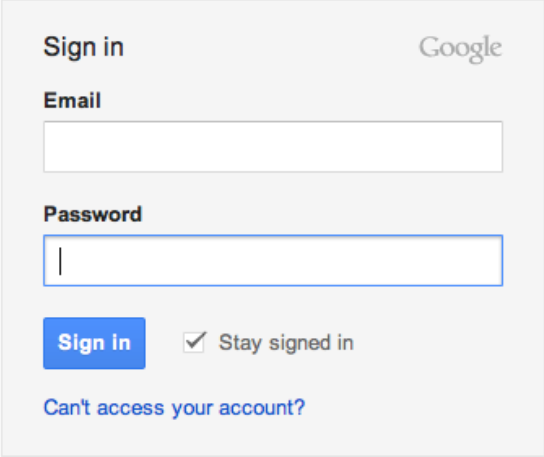
- Solution approach towards more privacy for IPv6 addressing
  - IPv6 "Privacy Extensions" allow randomisation of IP address assignments
  - Problem: Not perfectly working / not activated by default by some OSs

Identifying Individuals Online: Misused Mechanisms

- HTTP/HTML Mechanisms *misused* for user identification
  - Window.Name caching
  - Storing cookies in RGB values of auto-generated, force-cached PNGs
  - Using HTML5 canvas tag to read pixels (cookies) back out
  - HTTP authentication caching
  - …

Identifying Individuals Online: (Single) Sign On

- Sign-On
  - Actively logging into Websites (e.g. Google)



- Single-Sign-On
  - Logging on to multiple sites using a single ID provider (e.g. Facebook Connect)

Identifying Individuals Online: Browser Fingerprinting

- Identification of users based on numerous available browser information
  - Browser version
  - OS version
  - OS language
  - Screen resolution
  - Installed fonts
  - Installed plug-ins
  - …

- For more information,
  - visit http://browserspy.dk/ or
  - check your browser fingerprint for yourself: http://panopticlick.eff.org/

Identifying Individuals Mobile: Mobile Device / App ID

- Mobile Device ID
  - Globally unique proprietary identifier for a mobile device assigned by device manufacturers
  - Accessible for application developers via designated API (Application Programming Interface)
  - Not to be confused with the IMEI (International Mobile Equipment Identity) of the GSM (Global System for Mobile Communications) Standard

- Mobile Application ID
  - (Globally) unique proprietary identifier for a mobile apps assigned by application developers or providers
  - Online/Mobile browsers allow only cookies / web storage as means for user identification (which can be deleted at any time by a user)
  - Mobile Apps can freely store any kind of information persistently on the mobile device

Identifying Individuals Mobile: Mobile Device Fingerprinting

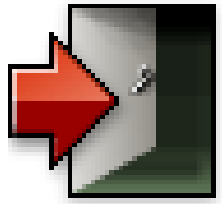
- Collection of *software* and *hardware* settings acquired from a remote mobile device
  - The summary of these setting provides a pretty unique fingerprint
  - In comparison to the browser fingerprint, a device fingerprint can hardly be changed by a user
    - Changing one setting (e.g. OS version) is often not enough

- Start-ups with a business model based on fingerprinting
  - User identification for targeting of mobile advertisements
  - Mobile access control
  - Attaching user reputation to mobile devices
  - Blue Cava: www.bluecava.com

bluecava
where devices rule.

Identifying Individuals Mobile: Apps Auto Sign-In on Mobile Devices

- Mobile apps and devices allow automatic sign-in without user interaction



- Example
  - Desktop Browser: User has to enter credentials manually and start the sign-in process
  - Mobile App: Credentials are stored within the app and sign-in automatically happens on application launch

Identifying Individuals Online/Mobile: User Profiling

- Similar to browser or device fingerprinting
- Collecting as much as possible information about a user
  - Information are non-identifying
  - e.g. interests, age, gender, home town, habits, etc.
- At some point in time, the user profile becomes highly distinct compared to other user profiles and can be used as identifier.

Upcoming Means of Identitication

- Biometric Identification
  - Face Recognition (already used by Facebook)
  - Voice Recognition (Apple's Siri never forgets a nice chat with its owner)
  - Implanted ID chips
  - Gait recognition
  - Electronic Fingerprint scan
  - Iris Scan
  - …

Source: blog.bufferapp.com

Agenda

- Introduction
- Means of User Identification
- Personal Data Collection
- Personal Data Processing
- Personal Data Usage

Means of Personal Data Collection

- Online Tracking
- Mobile Tracking
- Mobile-Offline Tracking
- Closed Loop Tracking
- User-generated Personal Data

Raw Online Tracking Data about Individuals

- Online Tracking on Websites
  - User-Accessed websites across different providers (e.g. via Google Analytics)
  - Access time, IP-address, geographic region of access, visiting time, clicked links, viewed content, entered text, submitted requests
  - Browser information (Apple or PC computer, language, OS, etc.)

- Desktop Tracking Software
  - Virtually any data, content or user behaviour is theoretically available for Trackers

Online Tracking Possibilities

- Online Tracking on Websites
  - Limited data collection possibilities due to browser as “sandbox” runtime for web applications

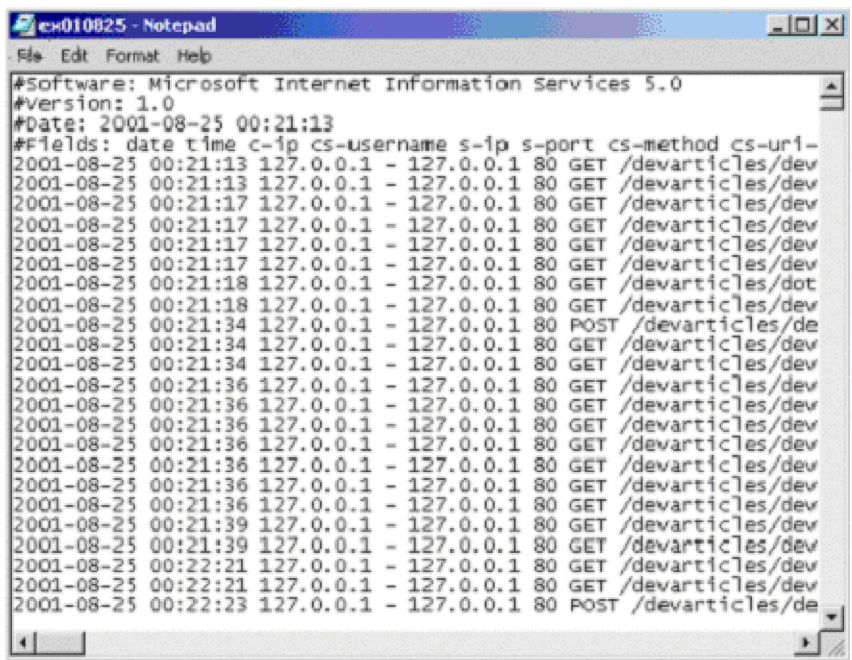
VS.

- Desktop Tracking Software
  - Depending to the access and execution rights of the software, there are no limitations for data collection efforts

- Web Server Log File
  - A **server log** is a file (or several files) automatically created and maintained by a server of activity performed by it.

- Log Entry Example

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326



Data Sources for Online Web Tracking

- Browser Information
  - Browser version,
  - OS Version,
  - OS language,
  - Screen resolution,
  - installed fonts,
  - installed plug-ins,
  - …
- For more information visit http://browserspy.dk/



Data Sources for Online Web Tracking

- Browser History Stealing
  - Most web browsers keep record of visited URLs by a users
  - These visited URLs are coloured differently when displayed on a website
  - This colouring can be checked using JavaScript

- Browser History Stealers
  - Check if certain URLs have been visited by a user
  - BUT: Stealers cannot iterate through the browser history
  - INSTEAD: Stealers have to check for URL individually if it has been visited by a user (e.g. spiegel.de, tagesschau.de, heise.de, etc.)

Data Sources for Online Web Tracking

- JavaScript-based Analytics
  - Real-time tracking of users visiting a website via JavaScript
  - Analytics often provided by Third Parties (e.g. Google) in exchange for access to analytics data for the First Party website
  - Client-based approach, which allows a more detailed collection of user behaviour compared to server logs

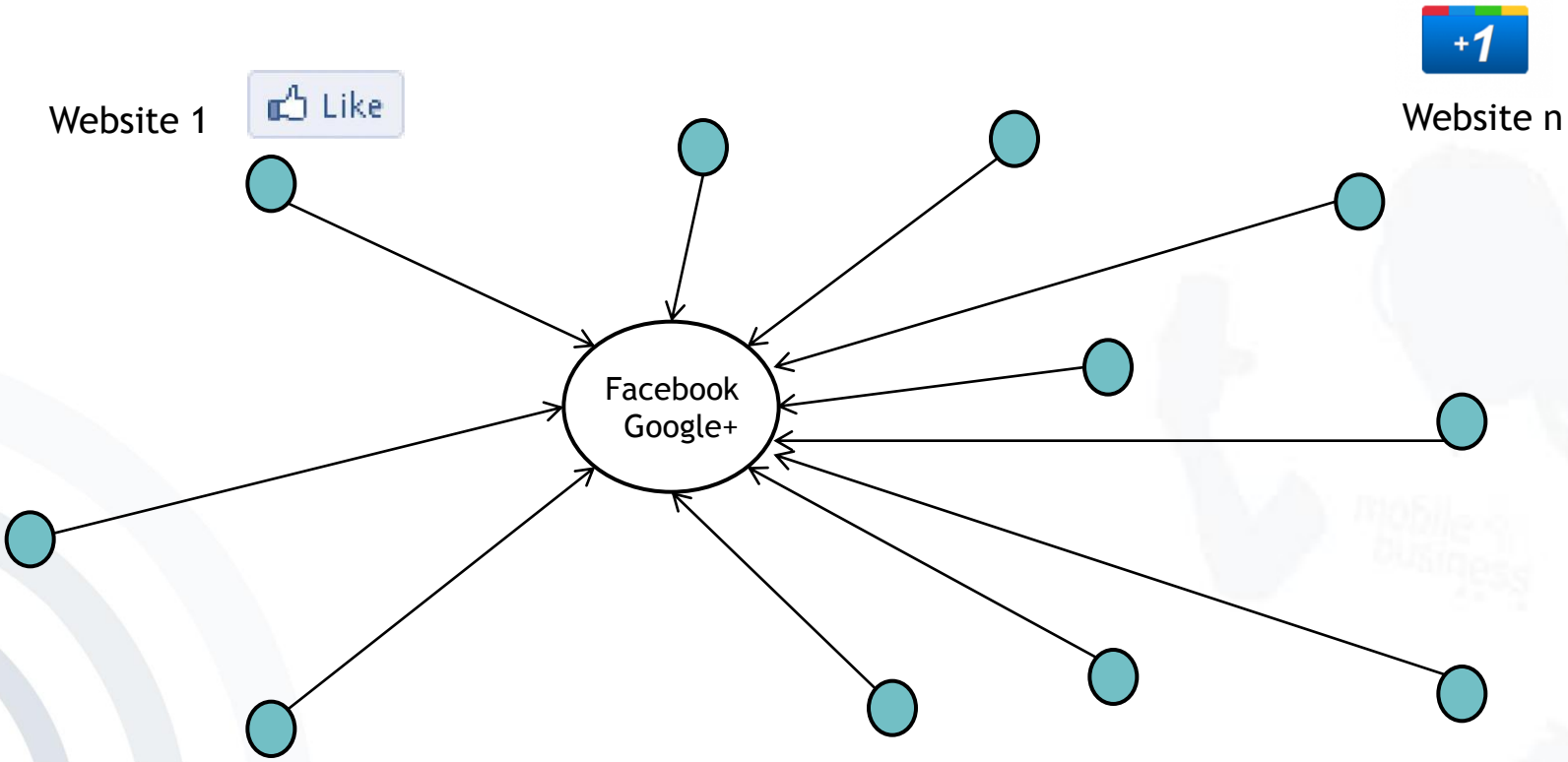
Data Sources for Online Web Tracking

- Web Bugs
  - 1x1 pixel size images on websites typically invisible to users and loaded from Third Party servers
  - Allow Third Parties to track page visits of users and place cookies
  - Works in HTML-eMails as well (allows to determine if and when email was read)

- URL Referrers
  - Webservers are able to determine from which prior website a user is coming from
  - Example
    1. User searches on Google for a tech online site
    2. Click on first search results (e.g. arstechnica.com)
    3. ArsTechnica receives the URL referrer: www.google.de/search?q=tech site

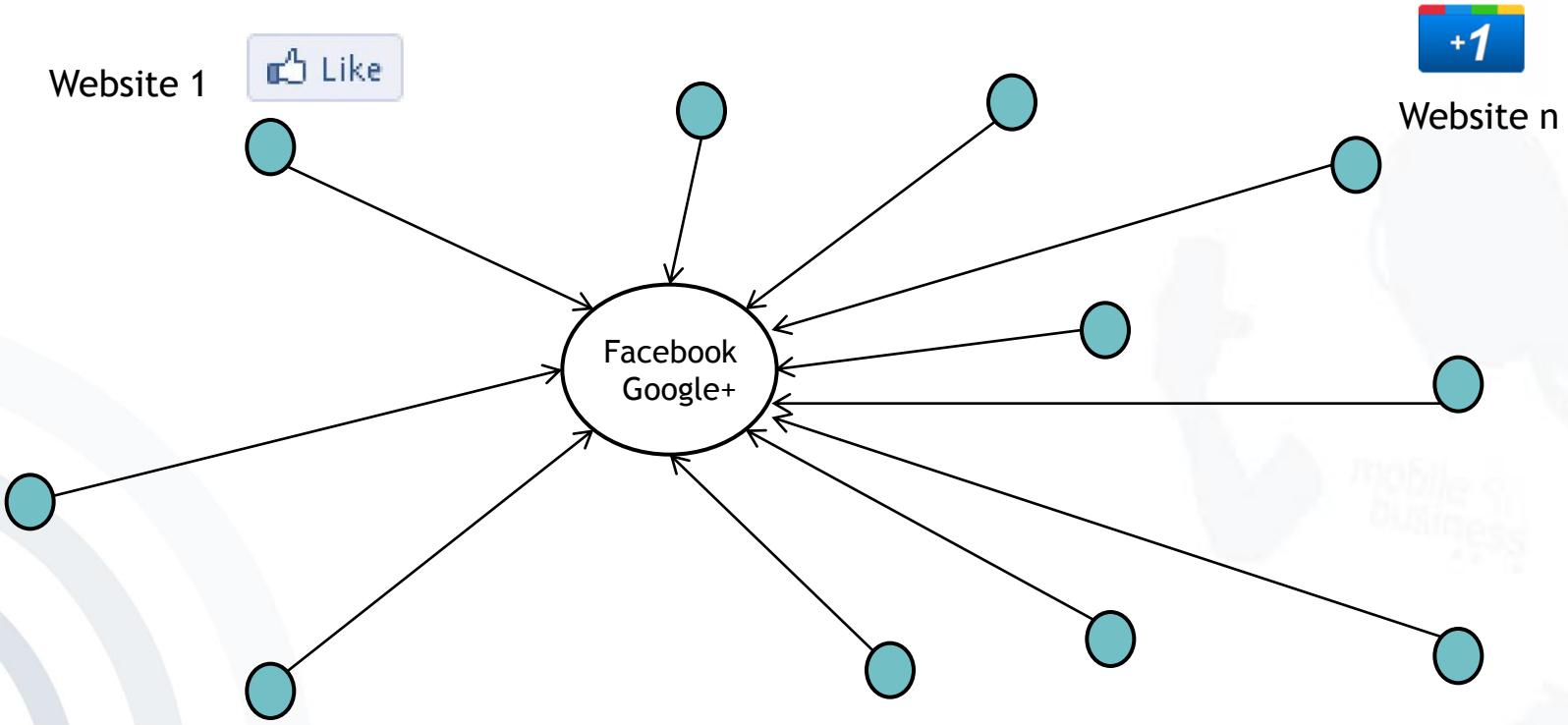
Data Sources for Online Web Tracking

- Like/Google+ Button (**without** FB/Google account)
- Websites with Like/Google+ Button report visitors to Facebook/Google
- Like/Google+ Button works similar to Web Bugs for users without FB/Google account



Data Sources for Online Web Tracking

- Like/Google+ Button (user **with** FB/Google account)
- Websites with Like/Google+ Button report the visited websites of an individual user to Facebook/Google
- This data persistently stored with the account of this user (even if the user is not logged in)



Data Sources for Online Transaction Tracking

- Tracking of conducted transactions at Online Market Places
  - Apple iTunes
  - Google Android Market
  - Amazon Market Place
  - …

- Tracking of Online payments
  - Google Checkout
  - PayPal
  - …

- Knowledge about confirmed payments for a conducted transaction are more valuable for an advertiser than any knowledge about clicked or viewed online advertisements.

Online Gaming Tracking

- World of Warcraft (Example)

- Every activity (e.g. step, click, interaction, fight, communication, etc.) is observed and analytically processed: 24/7 virtual surveillance!
- Can this data be linked to other personal data in the real world?

Online Communication Tracking

- Communication Tracking (e.g. GMail)
  - Google offers free eMail Service
  - Almost „unlimited“ storage space
  - Sophisticated Web Interface

- What gets Google in exchange?
  - eMails are stored „forever“
  - eMails are scanned in order to build a user profile
  - Advertisements are displayed and targeted based on existing eMail communications

Tracking in Cloud-based Data Storage Services

- Cloud-based Data Storage Services provide data about
  - browser configuration synchronisation
  - document synchronisation (e.g. iCloud)
  - cloud-based music & video streaming services
  - cloud-based photo collections
  - used cloud-enabled desktop or mobile cloud applications
  - ….

Mobile Tracking

Raw Mobile Tracking Data about Individuals

- Tracked Data on Mobile Websites
  - User-accessed websites across different providers (e.g. via Google Analytics)
  - Access time, IP-address, geographic region of access, visiting time, clicked links, viewed content, entered text, submitted requests
  - Browser information (computer type, language, OS, etc.)

- Data collected via Device/App Tracking Software
  - Virtually any data, content or user behaviour is theoretically available
  - Unique, personal data on mobile devices: location data and mobile device ID
  - In the near future more sensor data will become available as well: e.g. light, noise, temperature, polls, etc.

- Data tracking of device and usage by manufacturers is quasi "standard".

Mobile Tracking Possibilities

- Mobile Tracking on Websites
  - Limited data collection possibilities due to browsers or web applications captured in a “sandbox” runtime

VS.

- Mobile App/Device Tracking Software
  - Device manufacturers (e.g. Apple) are more restrictive about what kind of apps are allowed on a mobile device compared to desktop PC applications
    - For instance, all iOS Apps run in a sandbox and have only a limited set of APIs to access data from other Apps or from the OS
    - Apple “handpicks” all apps to be offered in the AppStore
  - Tracking of device and app usage is often “build-in”
  - For users, mobile apps more important than Mobile Web Apps

Data Sources for Mobile Web Tracking

- With some limitations, basically the same data sources for mobile web tracking as for online web tracking exist
  - Web server logs
  - Analytics scripts
  - Browser information
  - History stealing
  - URL referrer
  - …

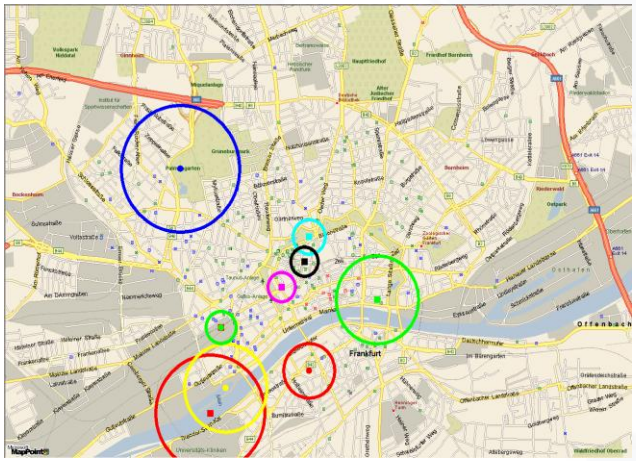
Data Sources for Mobile Usage Tracking

- Mobile Device Usage Trackers
  - Registering mobile device for first time use
  - Tracking of mobile apps launches
  - Tracking of installed applications
  - …

- Mobile In-App Tracker
  - Analytics software (First or Third Party) to track user behaviour
  - First Party is typically mobile app provider
  - Third party is typically an advertising network

Data Sources for Mobile Location Tracking

- Location Tracking
  - Tracking the different locations every time a user launches a location-based service app on his mobile device
- Localisation Technologies
  - WiFi (the mostly frequently used)
  - (Assisted)-GPS
  - Mobile Network: COO (Cell of Origin)
  - Bluetooth, IrDA, NFC
  - Barcode Scanning
  - Manual user entry

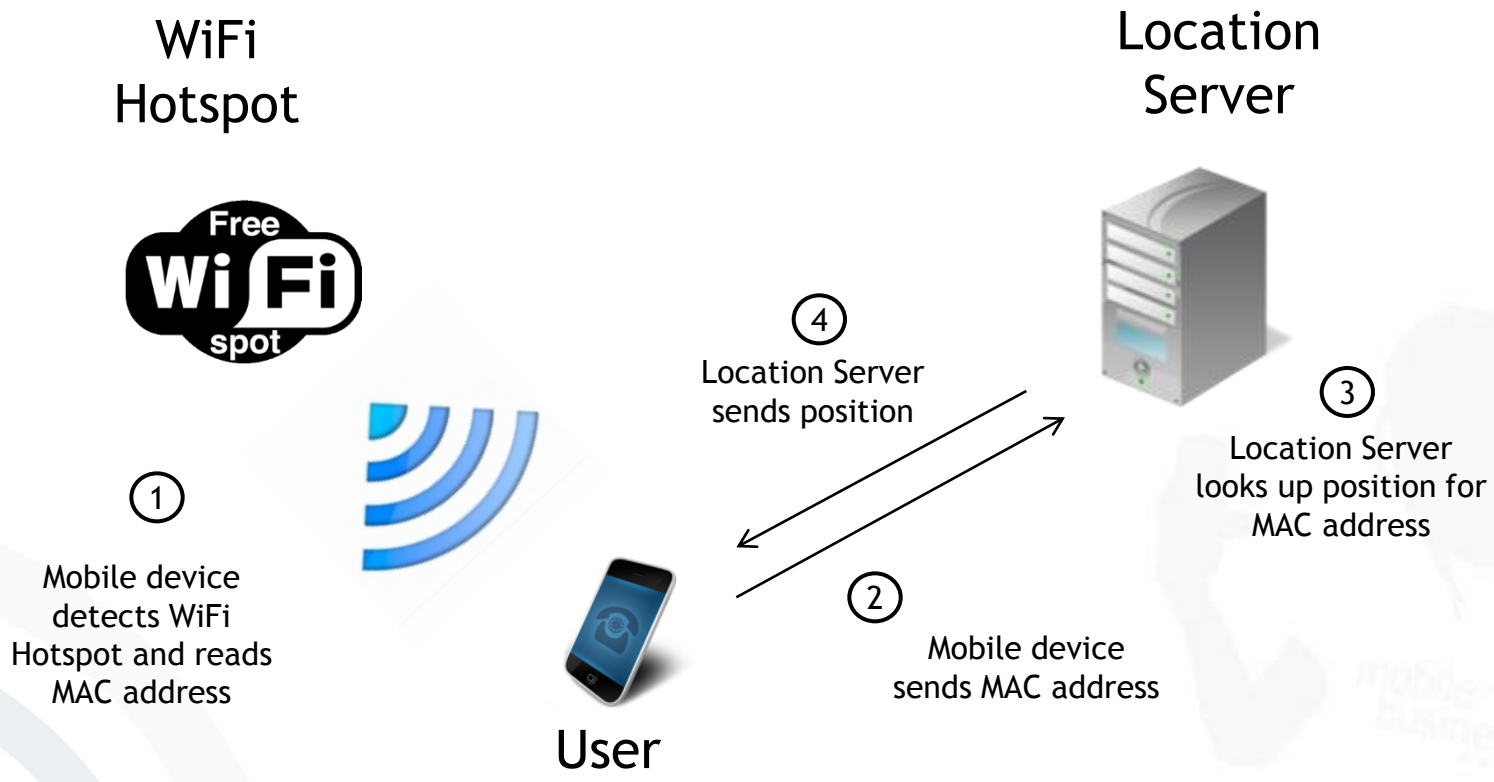


WiFi Localisation

Preparation

- WiFi Hotspots in a city are collected and stored in a central database
  - Collection via plane or by car (e.g. Google Street View or Microsoft Bing Maps)
  - WiFi Hotspot data record contains geo-position and MAC-Address of WiFi Router
  - Two location providers world-wide so far
    - Google and Skyhook

WiFi Localisation Process



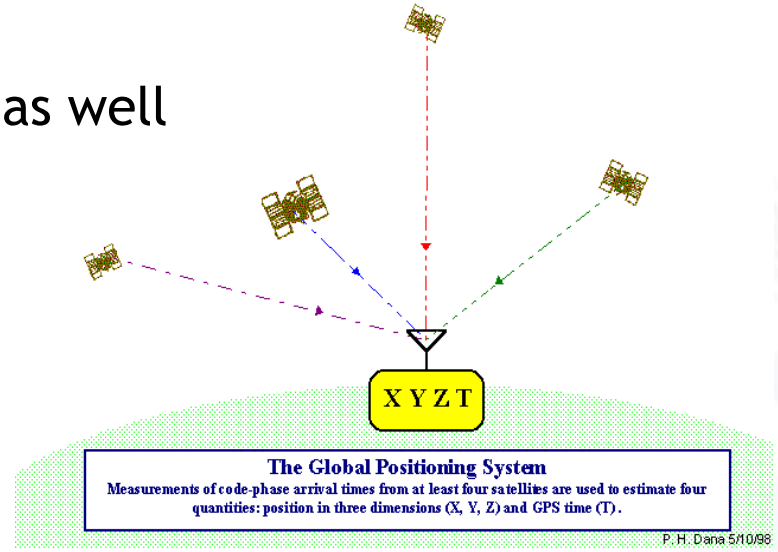
WiFi Localisation

- Advantages
  - Quite exact: approx. 20m,
  - Very fast localisation
  - WiFi available in many mobile devices

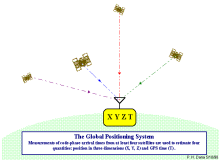
- Disadvantages
  - Privacy issues due to location submitted to location providers during positing process
  - Only two location providers existing (location profiles about individuals become possible)
  - Increased power consumption in non-stop operation.

GPS Localisation

- GPS – Global Position System is operated by US-Department of Defence
- Four Satellites needed for exact positioning (X,Y,Z + Time)
- Assisted GPS calculates the rough location based on Cell-of-Orgin first in order to then speed up the GPS based positioning process
- Other systems will become available as well
  - Galileo in 2013 (European Union)
  - GLONASS in 2013 (Russian Space Forces)



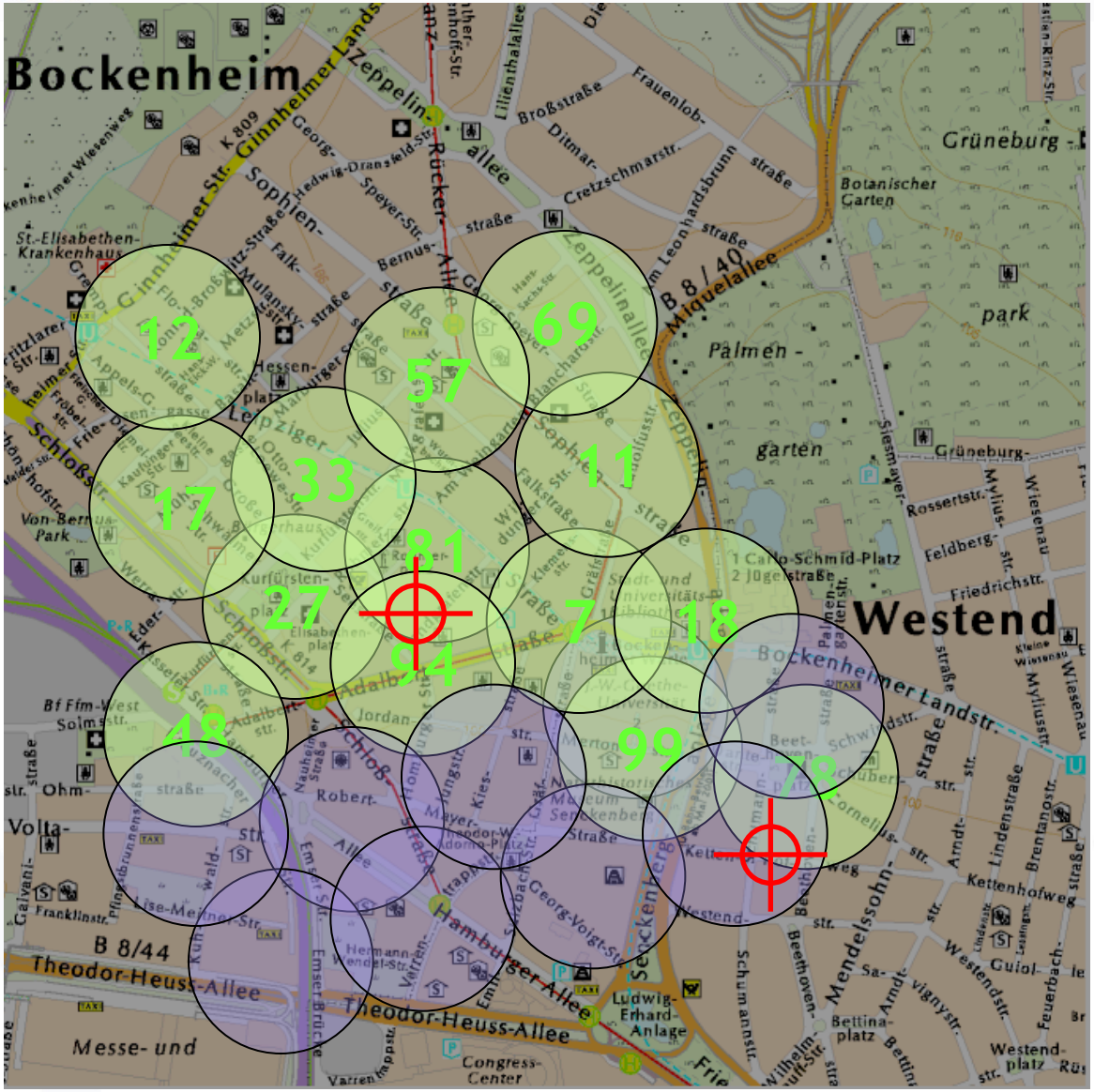
GPS



- Advantages
  - Quite exact: 5-15m,
  - Low cost chip sets, embeddable in terminals,
  - Large choice of standard software for applications available.

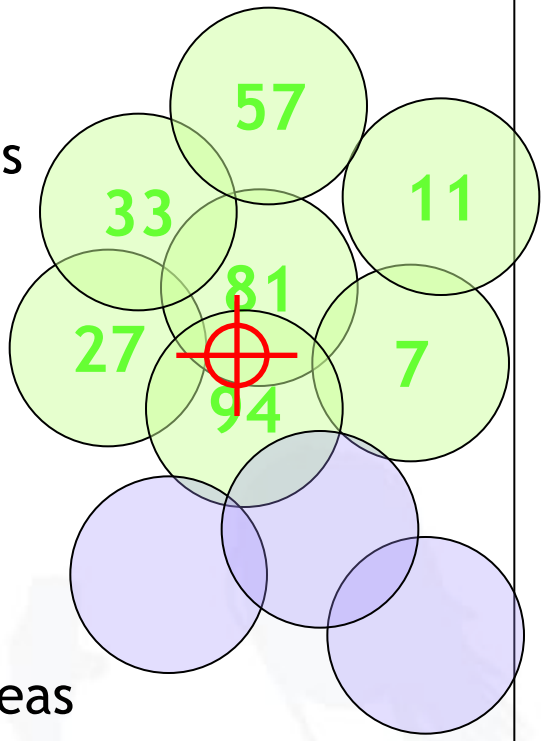
- Disadvantages
  - Works only outdoors
  - USA can manipulate or disconnect the signals whenever they want
  - Long initialisation time (up to 3 minutes)
  - High power consumption in non-stop operation.

Cell of Origin (COO) Localisation



COO Localisation

- COO Advantages
  - Works with every mobile device because it is mobile network-based
  - Very fast localisation



- COO Disadvantages
  - Accuracy depends in the size of the cell
  - 300m in city centers, up to 30km in rural areas
  - Privacy issues due to the fact that the mobile operator always knows the location of the user
  - Costs may occur for localisation requests

Cross-Channel Data Collection
Online – Mobile - Offline

Online – Mobile - Offline Tracking at the example of Google

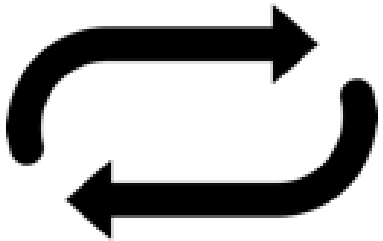
1. At lunch time, Frank searches on his office desktop PC via **Google Search** for “tasty sandwiches”
2. **Google Search** personalises his search results based on his search history and displays **Google AdWords ads** based on the user’s search query as well as his Google+ profile
3. Frank is displayed four **Google AdWords** ads on the Google search result page and Frank clicks on the second ad “Subways”
4. Frank is transferred to the ad’s landing page (i.e. Subways website). There, **Google Analytics** checks what kind of food Frank is interested in – by observing his browsing behaviour
5. Based on Frank’s derived tastes, **Google Offers** sends him a coupon for the Subway close to his office directly on his phone. It is stored in the **Google Wallet** application of his phone
6. Frank goes to Subways, pays his lunch and redeems his coupon both via the **Google Wallet** application. He receives a discount in exchange from Subways on his ordered food.

Online – Mobile - Offline Tracking at the example of Google

- Benefits for Frank
  - Highly personalised support service for satifying his needs while receiving a discount
- Involved Google Services
  - Google Search, Google+, Google AdWords, Google Analytics, Google Offers, Google Wallet

Online – Mobile - Offline Tracking at the example of Google

- Benefits for Google
  - Closed Loop Marketing



- Google knows
  - how Franks search relates to the clicked advertisement (Google Search)
  - how Frank interacts with advertiser (Google Analytics)
  - that Frank actually bought a product m advertised via Google at the store (Google Wallet & Offers)

- The Closed Loop
  - The information about every phase of Frank's transaction is eventually fed back into Google Search and Google AdWords in order to improve the future personalisation of the service and in particular the targeting of the ads.

Online – Mobile - Offline Tracking? This is just the beginning!

- Internet of Things
  - Increasingly “things” rather than “people” get connected to the Internet
  - Example: The in(famous) refrigerator, which orders food from the grocery store whenever its about to run out of the latter
    - Serious examples, please …

SmartHome

- Using Internet-enabled devices to manage and check home appliances such as
  - Heating
  - Windows shutters
  - Lights in the house/apartment
  - Locking of doors
  - Devices (e.g. TV, washing machine, etc.)
  - …
- Problem: This could enable your service provider to get insights about what your are doing at home
  - No problem?

winfwiki.wi-fom.de

Smart Meter

- A **smart meter** is usually an electrical meter that records consumption of electric energy in intervals of an hour or less and communicates that information at least daily back to the utility for monitoring and billing purposes

  [Federal Energy Regulatory Commission Assessment of Demand Response & Advanced Metering, 2011]

Smart Meter

- Advantages
  - No more service people at home required in order to check the current reading of a meter– everything can be done remotely
  - A user can also check the power consumption of every single device in their household (if connected to a single power point)

- Disadvantages
  - Service Providers are able to derive certain information about users, which may be considered private, e.g.
    - What kind of devices a user owns
    - When a user is at home
    - How many people are currently in the house/apartment
    - In which room of his house/apartment a user currently is
    - What movie he is currently watching (based on the power consumption pattern of his TV)
    - …

User-generated Personal Data

User-generated Personal Data

- Typical Categories in Social Media
  - Links
  - Pictures
  - Videos
  - Status Updates
  - Comments about other individuals
  - Profile Information
  - Messages
  - …

User-generated Personal Data

- Additional “Secondary” Data
  - Social graph
    - Relationships to other individuals
  - User Groups
    - User recipient groups for user-generated personal data (content)
  - Privacy Settings
    - Privacy settings towards other individuals or groups of individuals

Agenda

- Introduction
- Means of User Identification
- Personal Data Collection
- Personal Data Processing
- Personal Data Usage

Personal Data Processing

Knowledge Discovery in Databases (KDD) can be applied to personal data

- **Step 1:** data collection; first level: physical
- **Step 2:** data preparation; second level: empirical
- **Step 3:** data mining; third level: syntactical
- **Step 4:** interpretation; fourth level: semantic
- **Step 5:** Determine actions; fifth level: pragmatic

[Fidis 2005]

User-generated Personal Data and Tracking combined

- Google+ / Facebook knows about
  - all posted content (e.g. profile information, posted pictures, comments, etc.) and
  - all "liked/+1" content

  → **User-generated (explicit) personal data**

- Google+ / Facebook knows about
  - (many) visited websites through Like/+1 button

  → **Tracked (implicit) personal data**

→ Towards Predictive Behavioral Targeting

Data Aggregation and Attribute Derivation

From raw personal data records can be derived …

- geographic movement profiles (from tracked locations)
- personal users interests (from visited websites)
- education (from visited websites)
- spoken language (from browser / devices information)
- used technology to access the Internet (from OS information)
- estimation of age and gender (from viewed online content)
- online/offline shopping habits (from transaction history)
- mobile & online & offline habits (mobile wallet as link)
- relationships between users (from social networks)
- …

External Data Cross-Referencing

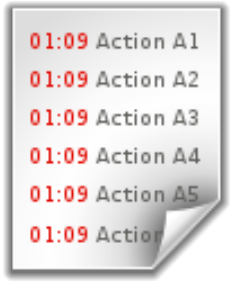
- Cross-referencing of individual user profiles with
  - socio-demographic data
    - Results in statistical information about age, gender, education, income, occupation, geographic area of an individual
  - personal social data
    - Results in statistical information about general living habits, related persons, degree of relation, relation between persons of an individual
  - business data
    - Results in in statistical information about partnerships, supervisor-employee relationship, education, etc. about an individual
- Foundation of cross referencing results are commercial statistical databases such as of Nielsen & Co.

Agenda

- Introduction
- Means of User Identification
- Personal Data Collection
- Personal Data Processing
- Personal Data Usage

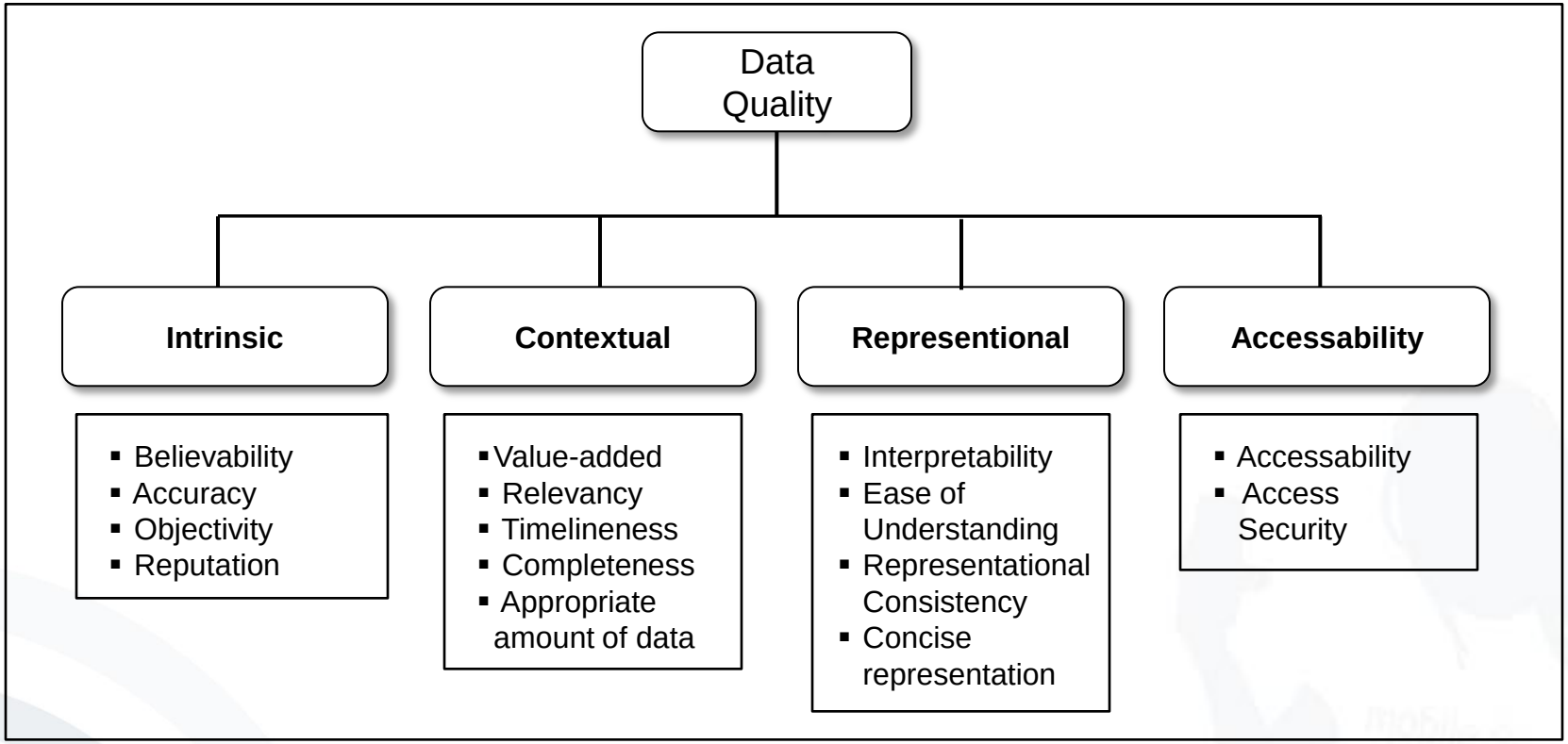
Using Explicit vs. Implicit Personal Data

- Explicit “Personal Data”
  - is explicitly expressed by individuals
  - can deliberately be wrong
  - can be misunderstand
  - can be outdated
  - can be incomplete
    - on purpose
    - without an individual knowing



- Implicit “Personal Data”
  - is derived from the observation of individuals’ behavior
  - can be derived wrong
  - can be interpreted wrong

Quality of Personal Data



Source: Wang and Strong (1996)

Personal Data Usage Application

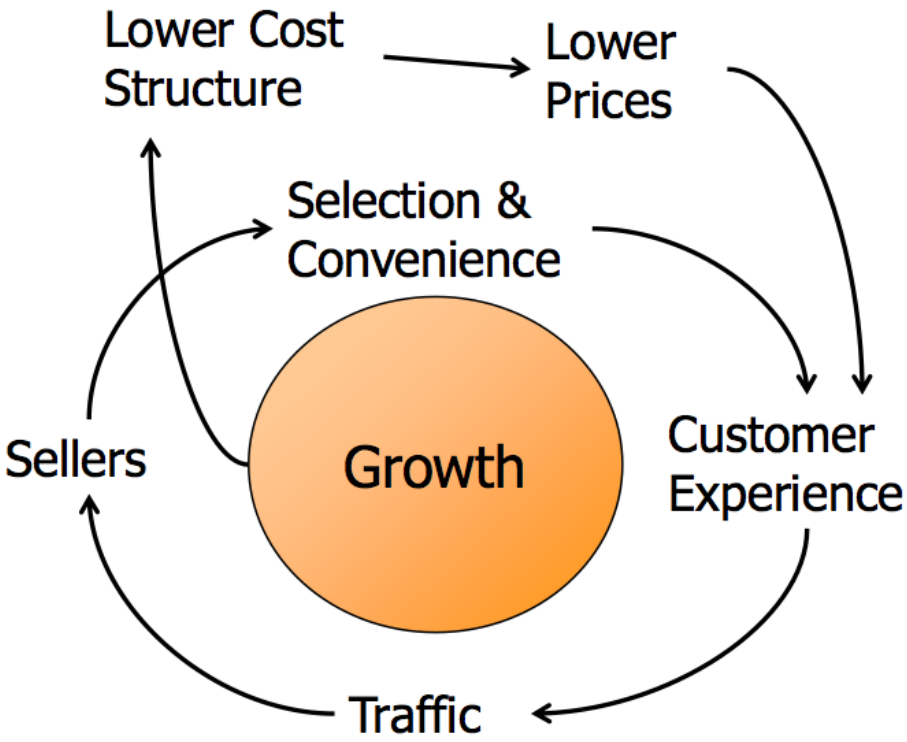
- Production Recommendation
- Personalisation of Content & Services
- Targeting of Advertisements
- Price Discrimination
- …

Recommender Systems

- Example Amazon‘s Value Creation
  - Product Recommendation drives „Convenience“ and „Customer Experience“

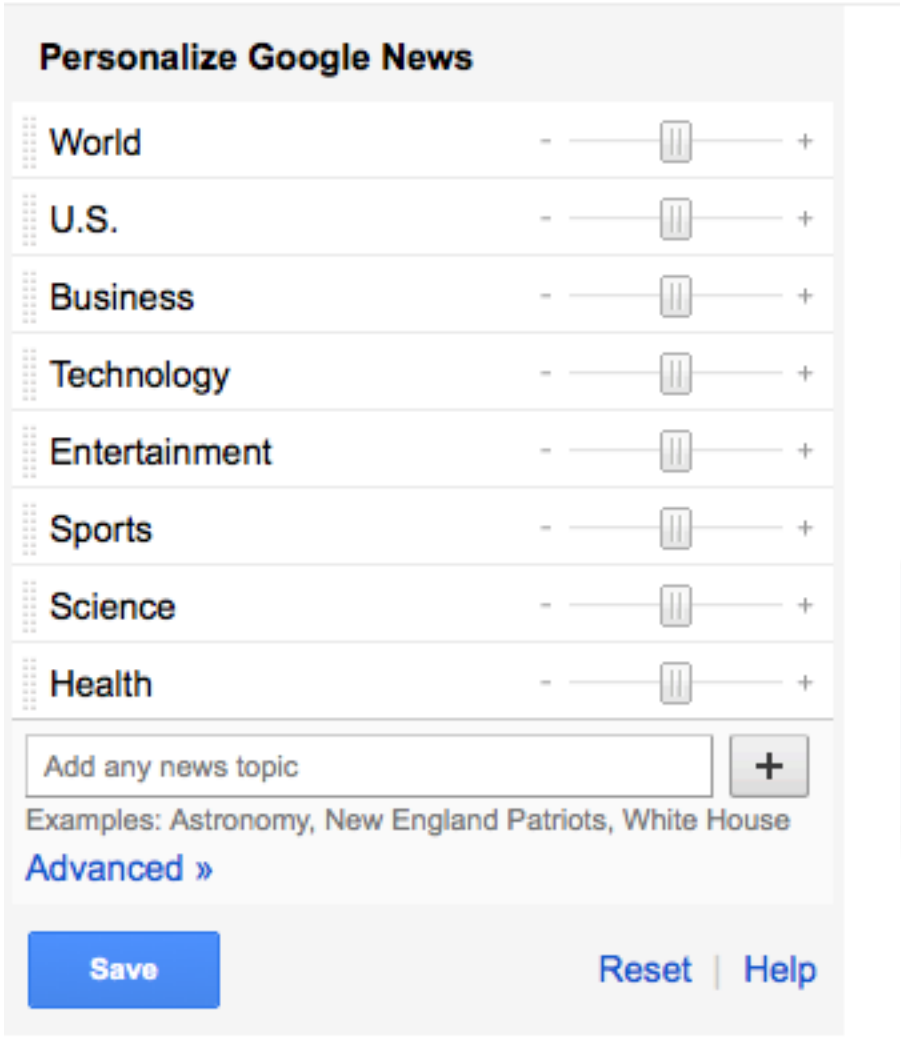
Recommender Systems

- Amazon‘s Value Creation



[seoromeo.com, 2011]

Personalisation based on explicit User Preferences



Targeting of Online/Mobile Advertisements

- **Technical Targeting**
  - Using, among others, IP address, browser model, or operating system for targeting purposes
- **Contextual Targeting**
  - Mainly text-based contents of an accessed website are analysed and Online Marketing campaigns are targeted accordingly.
- **Keyword Targeting**
  - According to the keyword or query provided by an online user, the respective Online Marketing campaigns are displayed.
- **Behavioural Targeting**
  - This approach observes the behaviour of users. It attempts to derive their potential needs statistically as a targeting foundation from the acquired *click stream*.
- **Profile-based Targeting**
  - Profile-based targeting requires that online users voluntarily disclose personal information (e.g. age, gender, personal interests).

Price Discrimination

A simple “theoretical” example …

- **Assumption:** Individuals, who own a Apple computer have in general a higher willingness to pay for certain kind of products
- Internet Browsers transfer the OS version of a client to the webserver of a merchant (MacOS or Windows)
- Possible resulting price discrimination on the merchant website

PC User: 450€

Mac User: 499€

Literature

- **Albers, A.; Figge, S.; Radmacher, M. (2005)**
  LOC3 - Architecture Proposal for Efficient Subscriber Localisation in Mobile Commerce Infrastructures, in: Proceedings of 2nd IEEE International Workshop on Mobile Commerce and Services (WMCS'05); München

- **Fritsch, L. and Muntermann, J. (2005)**
  Aktuelle Hinderungsgründe für den kommerziellen Erfolg von Location-based Service-Angeboten,*Proceedings der Konferenz Mobile Commerce Technologien und Anwendungen (MCTA)*, Bonn, Gesellschaft für Informatik

- **Schiller, J.; Voisard, A. (2004)**
  Location Based Services, Morgan Kaufmann, ISBN 1-55860-929-6

- **Wang, R. and Strong, D. (1996)**
  Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems* (12:4), pp. 5-34.