



Privacy Engineering

Lesson 1 - Overview

- **Privacy Engineering**
 - The field of research and practice that designs, implements, adapts and evaluates theories, methods, techniques and tools to systematically capture and address privacy issues when developing socio-technical systems.
 - Data Protection by Design and by Default
 - Appropriateness of Cost and Efforts
- **Privacy as**
 - Right to be alone
 - Controlling the Public Perception of Oneself
 - Not being subject to uncontrolled / excessive / secret collection or analysis
- **Abstract goals**
 - **Lawfulness**
 - Any processing of personal data must be explicitly legitimised – either by legal admission or by individually given consent
 - In the case of consent-based legitimisation, consent is required to be sufficiently specific in matters of data, purpose(s) and controller, to be given freely and unambiguously, and to be based on a well-informed decision
 - Everything not explicitly legitimised is forbidden
 - **Purpose Limitation**
 - Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.
 - Personal data may only be processed for explicit and well-specified purposes it was originally collected for. (or compatible purposes)
 - **Data Minimization**

- The amount of personal data processed shall always be limited to the absolute minimum necessary for fulfilling the intended purpose.
 - This comprises not only limiting the amount of data itself but also its “person-relatedness” and the number of parties able to access it.
- Transparency
 - Data subjects must be provided with easily intelligible information about the conditions framing the processing of personal data.
 - In particular, this includes information about :
 - Data Categories
 - Purposes
 - Storage durations
 - Safeguards
 - and the functioning of automated decision-making and profiling
- Accuracy
 - Personal data must be accurate and kept up to date.
 - Data subjects shall be provided with means for letting inaccurate data about them be rectified, completed, deleted, or - in case of accuracy being contested - temporarily blocked.
 - Only means that are reasonable with regard to the purpose of processing must actually be implemented
- Security
 - Personal data must be subject to appropriate technical and organisational security measures directed at established security goals like confidentiality, integrity, or availability.
 - The appropriateness of a certain security measure, in turn, depends on multiple further factors, particularly including the costs of implementation
- Accountability
 - The controller is responsible for implementing required privacy measures and must be able to demonstrate having done so.
 - The ability to demonstrate again requires dedicated technical and non-technical measures as long as they are “appropriate” for the individual case



- Data Portability
 - Data subjects must be able to receive the personal data relating to them and provided by them from the controller in a “structured, commonly used and machine- readable format”
- Enforcement
 - Regulations are enforced by independent supervisory authorities, order corrective measures, and/ or impose serious administrative fines.

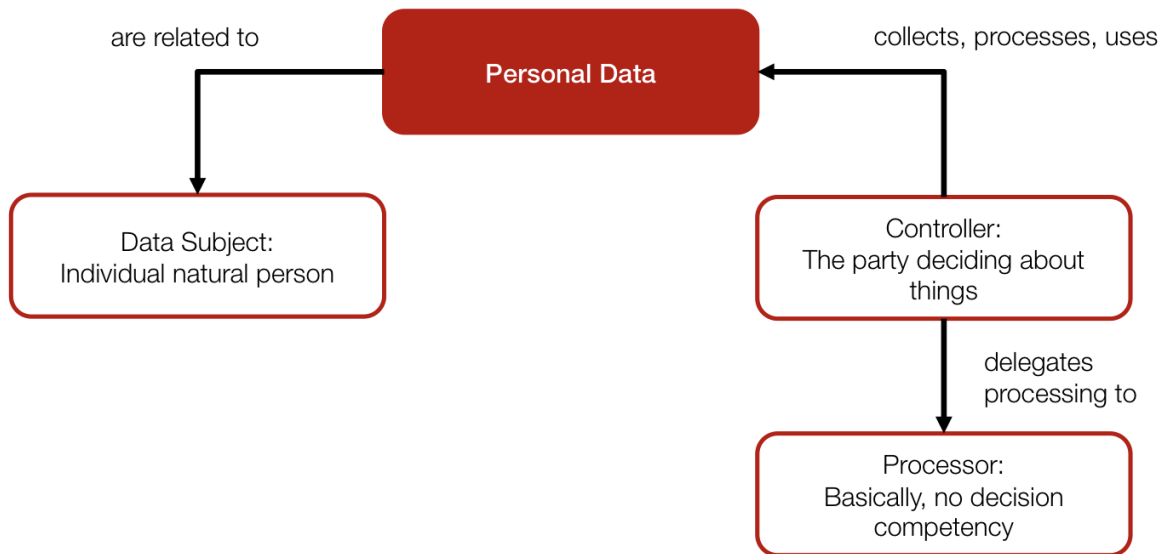
Lesson 2 - Foundation

- **Continental Privacy**
 - Right to informational self-determination
 - Capacity of individual to determine in principle the disclosure and use of personal data
- **American Privacy**
 - Right to freedom from intrusions by the state, especially in one’s own home reasonable expectations of privacy
- Privacy - No single definition of Truth, but family members: “Right to be alone”
- Two cultures : Dignity vs Liberty:
 - Control ones public image
 - Reasonable Expectations of privacy
 - Contextual integrity
 - Informational Self-Determination

Lesson 3 - Privacy Law

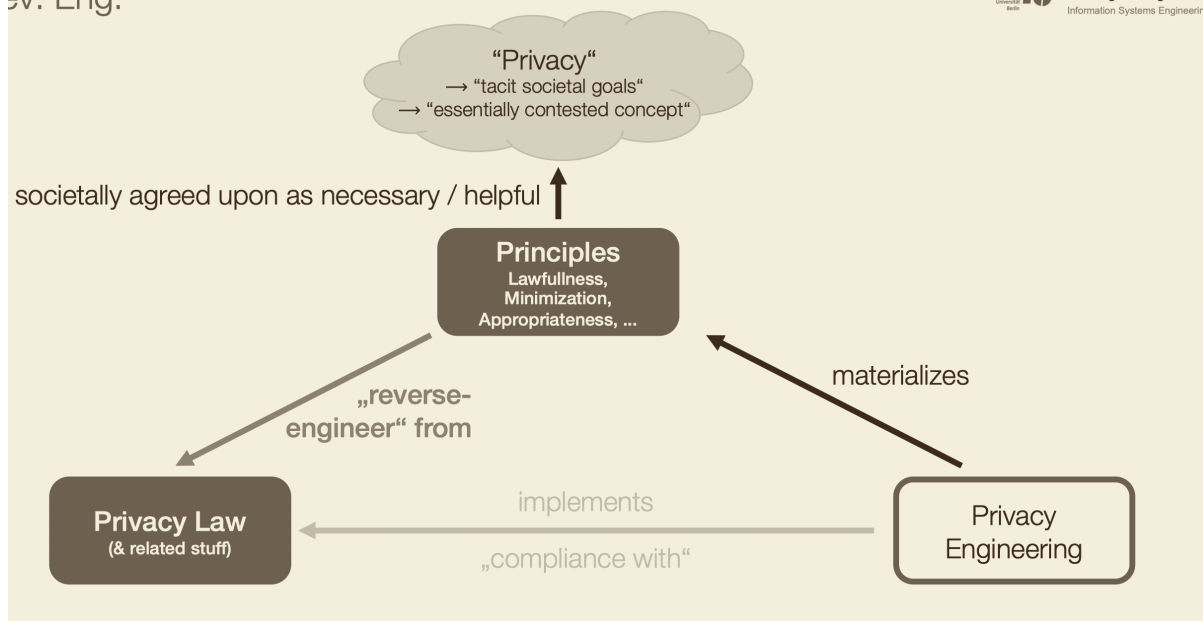
- Privacy Engineering
 - → (materialize) → “Privacy”: tacit society goals, essentially contested concepts

- Privacy Engineering → (implements / compliance with) → Privacy Law
- **GDPR**
 - **General Data Protection Regulation**
 - GDPR is aimed at establishing digital single market with similar rules.
 - When cross-country cooperation, it only allowed if appropriate level of protection present. Whenever “personal data” about EU citizens are collected, processed, stored, or used in an information system, the GDPR must be complied with (leaving aside some exceptions in the international context)
 - No applicability of GDPR **without personal data**.
- **Personal Data**
 - Any information relating to an identified or identifiable natural person(data subject)
 -  Amazon Orders + Name, Utility bills
 -  Fingerprints, IP address
- **Identifiable natural person**
 - is one who can be identified, directly or indirectly, in particular by reference to an identifier or to one or more factors specific to the physical, physiological, identity of natural person.
- **Data Subject:** Individual natural person
- **Controller:** The party deciding about things
- **Processor:** Basically, no decision competency



Pallas, | Privacy Engineering
19

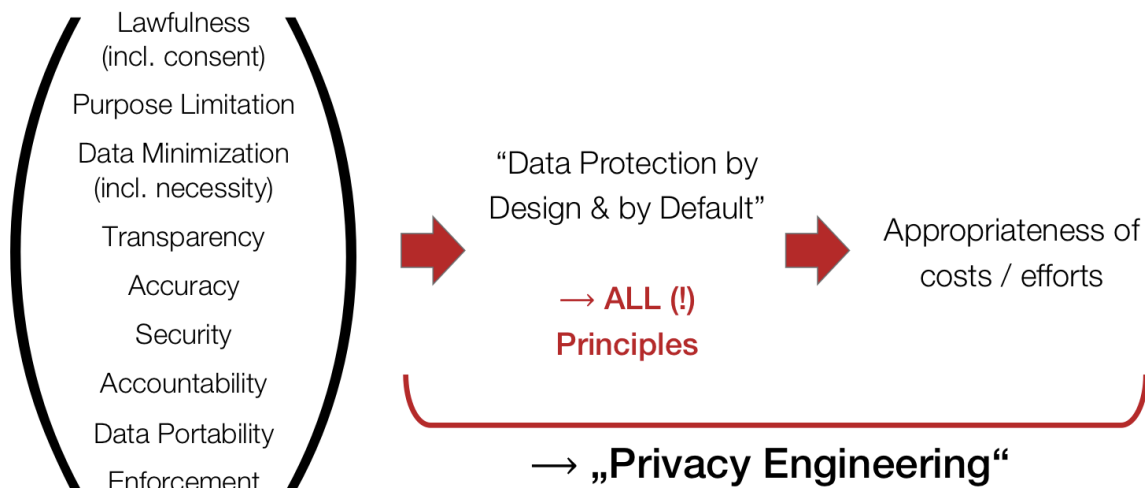
- **Controller** : The party deciding about things
 - The **controller** shall be responsible for, and be able to demonstrate compliance with rules set out in the GDPR
 - **Controller** shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures [...]"
- **Processor** : Basically, no decision competency
 - contracts ensuring that **processor** processes the personal data only on documented instructions from the **controller**
- **GDPR(General Data Protection Regulation)** applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.
- Principles → "reverse-engineer" → from Privacy Law



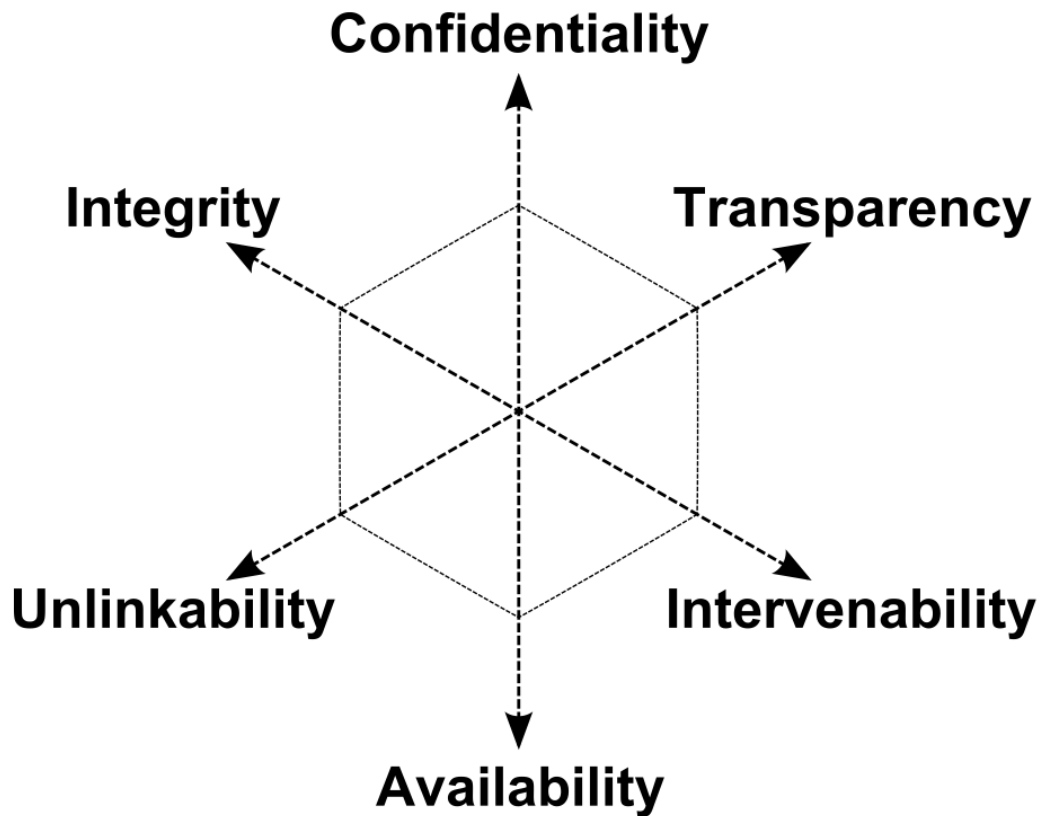
- **Processing** shall be **lawful** only if and to the extent that **at least one** of the following applies:
 - the data subject has given **consent** to the processing of his or her personal data for one or more specific purposes...
- **Consent** of the data subject means: freely given, specific, informed consent.
- consent → specific → "Utiliser" + "Purpose"
- **Purpose Limitation:** Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

Lesson 4 : Privacy by Design

- **Privacy by Design**
 - One of the core concepts of modern privacy legislation
 - Aims to ensure the effective implementation of privacy principles through specific technologies and their design.
- **"by Design & by Default"**
- Taking into account the state of the art, the cost of implementation and the controller shall implement appropriate technical and organisational measures which are designed to implement data- protection principles in an effective manner.



- Privacy by Design - Principles:
 - Proactive, not reactive
 - Privacy as Default setting
 - Privacy Embedded into Design
 - Full functionality - Positive Sum, not Zero Sum
 - End-to-End Security - Full lifecycle Protection
 - Visibility and Transparency
 - Respect for User Privacy - Keep it User-Centric
- To propose a framework that integrates existing research to provide engineers a clear roadmap for building privacy-friendly information systems.



Hansen, Jensen, and Rost (2015), Protection Goals for Privacy Engineering

Privacy Engineering Based on Protection Goals:

- Core Ideas:
 - The concept of protection goals, which has proven to be valuable in security engineering, can also be harnessed with regard to privacy
 - Protection goals are related to legal requirements - without resembling them in a 1-to-1 fashion
- Aim:
 - Better address technical audience
 - Transpose legal requirements into a representation more familiar for engineers
 - Concrete technological implementation guidance less in focus
- Confidentiality, Integrity, and Availability are well-known from security domain, however for privacy:

- Availability as “Accessibility by all parties”
- Integrity including : “non-repudiation” and “Authenticity”
- Protection Goals: **Unlink-ability**
 - Privacy-relevant data cannot be linked across domains. This implies that privacy-relevant data are not linkable to any privacy-relevant information outside of the domain.

Connected to: **Data Minimization, Purpose Specification, Purpose Limitation, Contextual Integrity**

- Protection Goals: **Transparency**
 - All privacy-relevant data processing - including the legal, technical, and organisational setting - can be understood and reconstructed at any time.
 - The information has to be available before, during and after the processing takes place

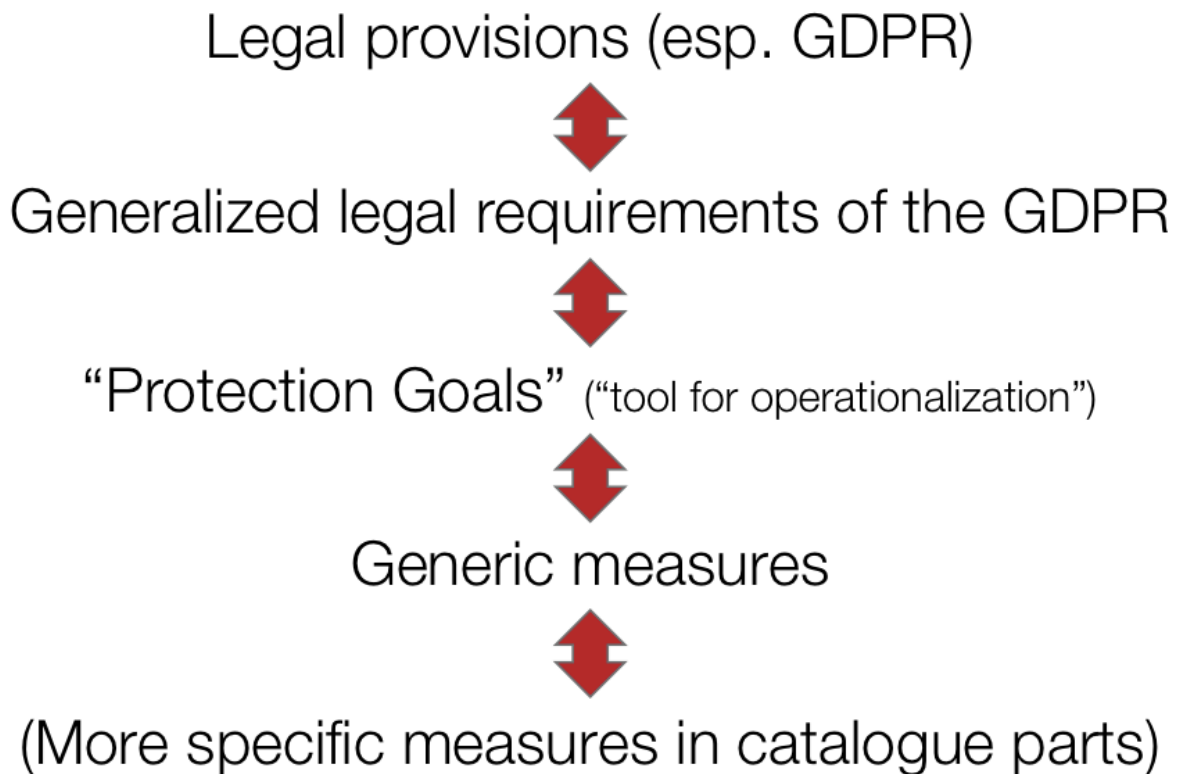
Connect to: **Transparency rights, Accountability**

- Protection Goals: **Intervenability**
 - intervention is possible especially for the data subject and authorities concerning all ongoing or planned privacy-relevant data processing.
 - The objective of intervenability consists of the effective enforcement of changes and corrective measures.

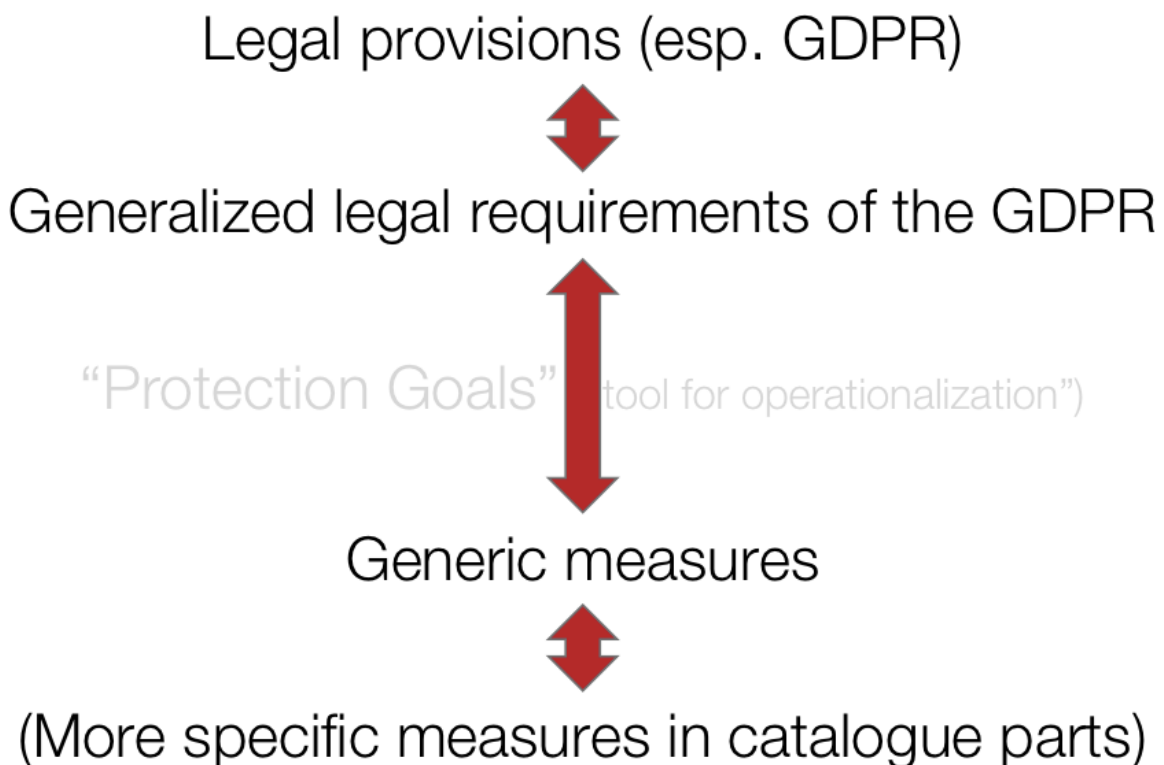
Connect to : **Right to Rectification(Accuracy), Consent Withdrawal(Lawfulness), Authorities Enforcement capabilities.**

- **Confidentiality vs. Availability:**
 - If a system provides confidentiality, this implies that access to certain data is restricted for certain entities - thereby violating availability
- **Intervenability vs. Integrity:**

- Integrity disallows subsequent changes to the integrity-critical data and processes which intervenability requires
- **Transparency vs. Unlinkability:**
 - Transparency intends to increase an understanding of the actual data processing, eg. by logging the actions of user and administrators, and Unlinkability tries to avoid such knowledge.
- **SDM: Standard Data Protection Model**
 - Protection goals:
 - Additional goal of data minimization
 - typical measures
 - protection levels
 - procedural aspects
- SDM will play important role in practice
- DPAs must check whether legal provisions of GDPR are complied with
- PPGs(Privacy Protection Goals) are rather self-defined or self-interpreted
- The logic behind SDM's Protection Goals:



- The Logic behind SDM w/o Protection Goals:



- Strategies vs Patterns:
 - A **design strategy describes a fundamental approach** to achieve a certain design goal.
 - It has certain properties that allow it to be distinguished from other fundamental approaches that achieve the same goal.
 - In principle, privacy enhancing technologies(PET) are used to implement a certain privacy **design pattern** with concrete technology
- Privacy Design Strategies - DB Metaphor:
 - Minimise
 - The amount of personal data that is processed should be restricted to the minimal amount possible.
 - Associated Tactics/Patterns: Exclude, Carefully select, strip, destroy data, anonymization / pseudonymization
 - Hide
 - Any personal data, and their interrelationships, should be hidden from plain view.
 - Encryption, mix networks, anonymization / pseudonymization, adding noise, differential privacy
 - Separate
 - Personal data should be processed in a distributed fashion, in separate compartments whenever possible
 - Client-side processing, peer-to-peer approaches, multi-party-computation
 - Aggregate
 - Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is still useful
 - Dynamic granularity(eg. location data), K-Anonymity, I-Diversity
 - Inform
 - Data subjects should be adequately informed whenever personal data is processed

- P3P, Privicons, data breach notifications, particular focus on HCI, algorithmic transparency
- Control
 - Data subjects should be provided agency over the processing of their personal data
 - Technical consent mechanisms, Privacy Dashboards
- Enforce
 - A privacy policy compatible with legal requirements should be in place and should be enforced
 - Purpose-based Access Control, Sticky Policies
- Demonstrate
 - Be able to demonstrate compliance with the privacy policy and any applicable legal requirements
 - Privacy Management systems, Logging, auditing

	Purpose limitation	Data minimisation	Data quality	Transparency	Data subject rights	The right to be forgotten	Adequate protection	Data portability	Data breach notification	(Provable) Compliance
MINIMISE	○	+								
HIDE		+					○			
SEPARATE	○						○			
AGGREGATE	○	+								
INFORM				+	+				+	
CONTROL			○		+			+		
ENFORCE	+		+			+	+			○
DEMONSTRATE										+

Lesson 5 : Advanced Concepts

- Design Choices significantly influence outcomes
 - First-mentioned Option
 - Default Settings
- Core goal of privacy is Self-Determination
 - individual autonomy
 - Being able to achieving own preferences in actual practice of systems usage
 - Being able to understand “what is going on”
 - Being able to reasonably exert influence in practice
- Goals of Usable Privacy(Policies)
 - Transparency: Making clear the data being collected and why
 - Legibility: Making data practices **comprehensible** to potential users
 - Relevance: Making data practices relevant to those it concerns in the context of service sue
 - Choice: Providing **understandable choices** regarding access and use of data to help user make an informed decision
 - Agency: **Providing visible controls that are easy to locate, understand and action** - to support decision-making and setting preferences
- Compliance Budget
 - Security incidents often rooted in employee “misbehavior”
 - Given that they should know better, why don’t users comply with security policies?
 - Core insight from empirical research: Security is “secondary task”, stepping back behind “getting work done”

- Users comply up to a certain “budget”, circumvent security measures afterwards, letting overall security fall down
- Summary
 - Privacy typically involves trade-offs
 - Economic Experiments to understand individuals’ privacy behavior
 - Bounded rationality: First options, defaults → Nudging
 - Usable Privacy: Comprehensibility, understandable choices, easily available controls → Interface Design
 - Usable Security/Privacy: Conceptual models, efforts, ...

Lesson 6 : Toolkit 1 - Crypt, Hashing, Anonymisation

- **Encryption**
 - converts the original representation of the information, known as plaintext, into an alternative form known as cipher text
 - Ideally only authorised parties can decipher a cipher a cipher text back to plaintext and access the original information.
- **Cipher**: is an algorithm for performing encryption or decryption
- Core property of **secure encryption schemes**
 - Kerckhoff’s Principle: A crypto-system should be secure even if everything about the system, except the key, is public knowledge.
 - Claude Shannon: the enemy knows the system
- **Symmetric Encryption** : Block Ciphers
 - Content to be encrypted is split-up into multiple blocks of fixed length
 - Block-wise encryption following more sophisticated operations, typically with multiple rounds and changing(derived) round-keys to harden cryptanalysis.
 - Decrypt: Inverse operations with same key → **Symmetric Encryption**

- Asymmetric Encryption
 - Based on public-private key-pairs
 - Encrypt_Public → Decrypt_Private
 - A sends message to B encrypted with B's Public Key
 - Only B can decrypt message using Private Key
 - Encrypt_Private → Decrypt_Public
 - Hashing can basically be understood as a more sophisticated form of a "checksum"
 - Used to check for integrity(non-alteration) of covered contents
 - Core property: Different content must lead to different hash
 - Signature:
 - A hashes message, encrypts hash with A's Private Key, sends result along with message
 - By Decrypting hash with A's Public Key, B can validate that message originates from A
- Cipher-Suites
 - It defines how encryption is to be carried out which algorithms are to be used between participants
 - Suite is in TLS negotiated between participants during handshake, based on configuration / capabilities
 - Consists of different components
- Hashing
 - Hashing can basically be understood as a more sophisticated form of a "checksum"
 - Used to check for integrity(non-alteration) of covered contents
 - Core property: Different content must lead to different hash
 - Ideally, changing 1 bit of content leads to 50% changed bits in the hash
- Https
 - Cipher suite negotiated during Handshake

- ECDHE or others for Key negotiation
- RSA for authenticating identity
- AES for content encryption (symmetric) - Different operation modes(GCM, CBC) of secondary relevance
- SHA384 etc, for hashing → integrity
- Symmetric Encryption
 - Whenever high amounts of data need to be protected against eavesdropping or disclosure
- Asymmetric Encryption
 - Whenever (comparably small amounts of) data need to be encrypted without having or being able to use an agreed-upon key
 - When signature functionality is needed
- Hashing
 - When the integrity or non-alteration of data needs to be ensured
 - Where 2 pieces of data are to be compared especially remotely.
 - In special case of password hashing: Salt & Pepper, special algorithms
- Certificates
 - When the identity of a communication partner needs to be proven uses hashing and asymmetric encryption

Privacy stages	identifiability	Approach to privacy protection	Linkability of data to personal identifiers	System Characteristics
0	identified	privacy by policy	linked	<ul style="list-style-type: none"> • unique identifiers across databases • contact information stored with profile information
1	pseudonymous	(notice and choice)	linkable with reasonable & automatable effort	<ul style="list-style-type: none"> • no unique identifies across databases • common attributes across databases • contact information stored separately from profile or transaction information
2		privacy by architecture	not linkable with reasonable effort	<ul style="list-style-type: none"> • no unique identifiers across databases • no common attributes across databases • random identifiers • contact information stored separately from profile or transaction information • collection of long term person characteristics on a low level of granularity • technically enforced deletion of profile details at regular intervals
3			unlinkable	<ul style="list-style-type: none"> • no collection of contact information • no collection of long term person characteristics • k-anonymity with large value of k

Spiekermann and Cranor (2009), Engineering Privacy

- **Anonymisation** : Data rendered anonymous in such a manner that the data subject is not or no longer identifiable
- **Pseudonymization** : the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provide that such additional information is kept separately
- **k-Anonymity**
 - Generalise data so that risk of re-identification is soften while data still providing value
 - Table X satisfies k-Anonymity if and only if each sequence of identifying values appears with at least k occurrences in X
- **L-Diversity**

- Data are depersonalized or aggregated in a way ensuring that each set/group contains at least L different “sensitive” values

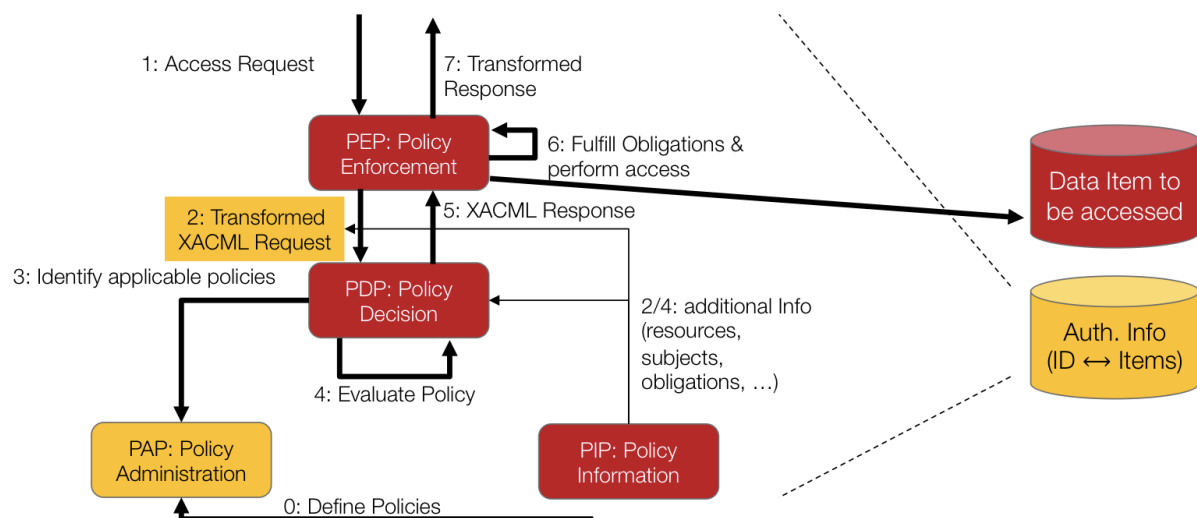
- **K-Anonymity vs. L-Diversity**

-

Lesson 7 : Toolkit 2 - Identities, ID-Mgm, ABC, XACML, P3P

- **XACML:**

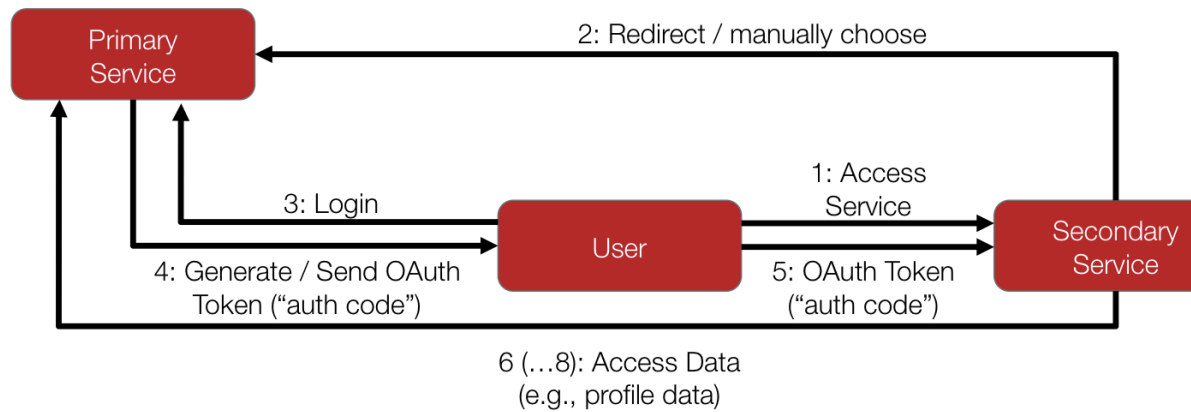
- Extensible Access Control Markup Language
- Defines specification language and architecture / components for implementing fine-grained access control
- In straightforward model, service-provider gets “certified” ID with several attributes of this ID



See: El-Aziz and Kannan (2013): A Comprehensive Presentation...

- **OAuth-Based Access**

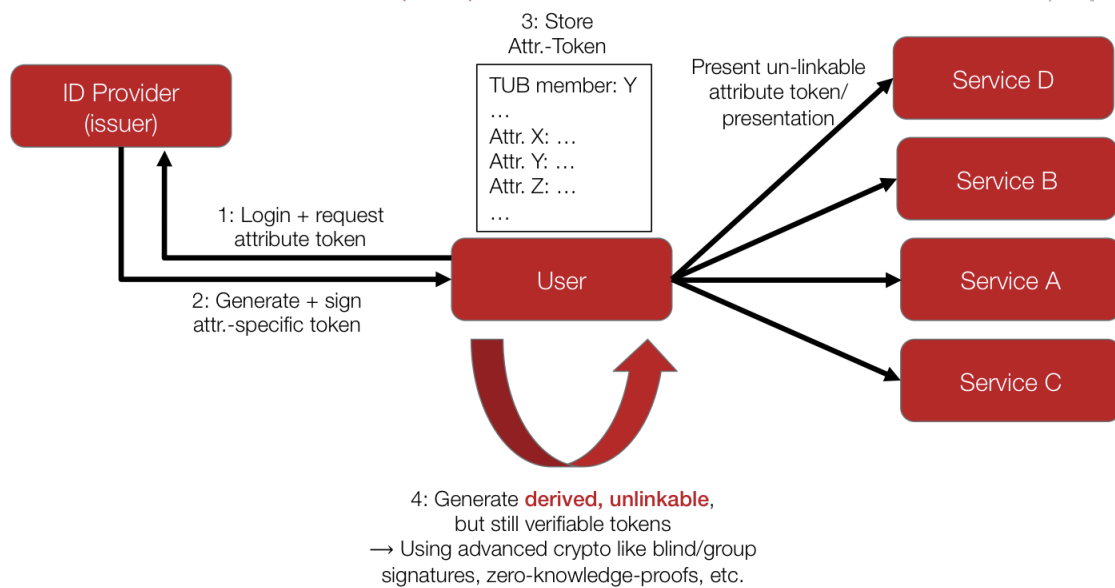
- OAuth somewhat comparable, providing ID + access to (certain, definable) data



Authorization (allowed to access what),
not **Authentication** (Who is this)!

- ABC
 - Attribute-Based Credentials
 - Revocation Authority
 - Allow for revocation of previously issued credentials
 - Inspector
 - Additional possibility to specify credentials which can only be “accessed or seen” by specified inspectors
 - Restrict access of verifier (service provider) even further
 - Whenever only a certain attribute (age, affiliation) needs to be provided in proven form, ABC are the more privacy-friendly alternative to approaches based on identities containing excessive information

Attribute-Based Credentials (ABC): Idea



Pallas | Privacy Engineering

- P3P
 - Platform for Privacy Preferences Project
 - Motivation
 - Privacy policies are not readable / understandable / comparable
 - Codify and publish website's privacy policies in machine-readable format
 - Read out in user agent: Browser
 - Present through appropriate UI
 - Behave depending on policy and configuration
- Summary
 - Individual accounts → (Federated) Identity Management
 - XACML, esp. general components
 - Privacy-friendly options: Attribute-based approaches, especially ABC
 - P3P as early approach for managing “privacy preferences” on the Web – with more recent successors (IAB TCF, GPC, ...)
 - More advanced approaches needed for technically mediated privacy preference handling / “consent management”
 - Indispensable in context of IoT, service orchestration, etc.

Lesson 8 : Performance Evaluation

- Aim
 - Appropriateness weighings in Privacy Engineering
 - Foundations of Benchmarks
 - Conducting (Privacy-related) Benchmarks in practice (+ examples)
- **Privacy Engineering**
 - Data Protection by Design and by Default
 - Appropriateness of cost and efforts
- Progress in State of Art influences what measures are deemed appropriate
- As soon as the achievable risk-reduction ‘outweighs’ cost/effort, measure is to be taken
- Benchmark vs Monitoring

Category	Benchmarking	Monitoring
Motivation	Answer a specific question	Detect undesirable system states
System under test (SUT)	Test deployment	Production system
Point in time	Repeated short test runs	Permanently running process
Level of control	Complete control	Passive observation, no control
Stress on system	Artificially created	Actual production load
Observable metrics	Arbitrary	OS level and custom, discrete events
Impact on SUT	Strong	Negligible
Visibility delay	Often offline analysis	Near real-time

- Benchmarking and Privacy Engineering
 - For Benchmarking in the context of Privacy Engineering, our focus typically is on the impact of certain privacy or security mechanism / configuration on other qualities of the overall system:

- Performance (throughput, latency)
 - Cost (fixed, variable)
 - Indirectly: Consistency of data (staleness, ordering)
 - Sometimes: Availability (MTBF, MTTR, ...)
- In the context of Privacy, cost / effort is the primary factor to weigh a certain technology / configuration against
 - Good Benchmarks:
 - Relevance
 - A reader of the result believes the benchmark reflects something important
 - Impact on similar resources easier to benchmark, but impact on required resources often more interesting
 -
 - Repeatability
 - confidence that the benchmark can be run a second time with the same result
 - Use existing benchmarking tools (like YCSB, ApacheBench, ...) whenever available and applicable to intended setting
 - When existing tools are not appropriate or feasible, implement and make available own automated benchmark + load generator
 - Cloud-based benchmarking as well-established practice
 - Do multiple benchmark runs at different time of the day
 - Fairness
 - All systems and/or software being compared can participate equally
 - When multiple architecture / design / configuration options seem feasible, benchmark all of them
 - However, sometimes well-considered reduction of parameter design space necessary
 - Verifiability

- confidence that the documented result is real
- Economical Feasibility
- 6-Step approach applied to security but also valid for principle of privacy
 - Identify relevant parameters & trade-offs
 - Reduction of parameter design space
 - Preference ordering
 - Experiment planning
 - Experiment execution
 - Result analysis
- Trade-off Example : Apache Cassandra
 - Distributed, highly scalable, ring-structured datastore
 - Used by Apple, Netflix , ebay ...
 - Typical usage: Self-deployed installation comprising n public cloud instances
 - Data in transit encryption
 - TLS
 - Full set of config options as provided by TLS implementation
- Apache Hbase
 - Core component of Apache Hadoop Big Data ecosystem
 - Use : Self-deployed installation comprising n public cloud instances
 - Data in transit encryption
 - Native encryption based on Kerberos, Java SASL
 - Basically only on/off on 2 layers
 - Performance impact to be expected ~ 10%
- Amazon DynamoDB

- Hosted managed Cloud Storage Service
- Procured by throughput units
- Data in transit encryption
 - TLS-based HTTPS
 - Strongly limited subset of TLS configurations

Lesson 9 : Advanced Tools

- Property-Preserving Encryption
 - Order-Preserving Encryption
 - Range-queries on encrypted data
 - Max, Min, Median... on encryption data
 - Order-based indexing on encrypted data
 - Partial Order-Preserving Encoding
 - Employ usual semantically secure encryption like AES
 - Organise data and ordering information in tree structure
 - One tree per searchable dimension (DB Column)
 - Upon write, put encrypted item to root buffer
 - Upon first query, sort all unsorted into tree, based on comm. protocol with client-side oracle, answer based on tree-stored order info
 - Fully Homomorphic Encryption
 - Calculations on encrypted values
 - Allows to calculate any mathematical function based on emulated logic gates
- **Sticky Policies**
 - Personal Data communicated across many organisational boundaries
 - Data subject should be able to specify who may use data or what is done with data

- Core idea: Data itself carries policies sticked to it
- **Policies attached to data can specify**
 - Proposed use of the data → purpose
 - Restrictions on the set of platforms, networks, part of enterprise that the data may be used / processed in → security requirements
 - Specific obligations to be complied with (allowed people, processes, third parties)
 - Blacklists, obligation for notifications about disclosure, deletion/anonymisation obligation
 - List of Trusted Authorities providing assurance about requirements being met during the process of granting access
- **Sticky Policies : Intended Applications**
 - Actual possession of data becomes irrelevant
 - Data itself describes use-cases and conditions for usage
 - Enforcement through Trusted Authorities
 - 100% confidence would require totally secure environments
 - Concept of attaching policy to data itself - with or without complex crypto - can prove valuable in distributed multi-party settings