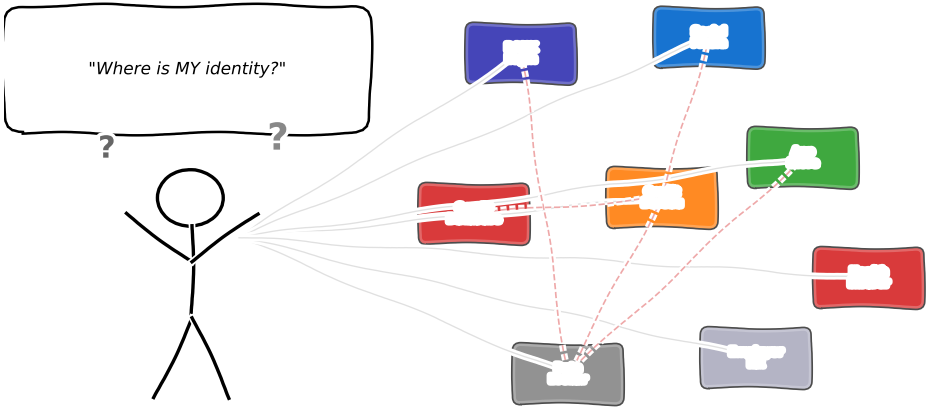
Lesson 8.1: Digital Identity and the Data Economy

Module 8: The Future Problem

Prof. Dr. Joerg Osterrieder

Digital Finance — BSc Course

Who Are You, Really?



Your identity: held by 8+ organizations

"I have 47 passwords and none of them are my identity"

Learning Objectives

After completing this lesson, you will be able to:

1. **Compare** centralized, federated, and self-sovereign identity models [Analyze]
2. **Explain** how verifiable credentials and zero-knowledge proofs preserve privacy [Understand]
3. **Describe** the structure of the data economy and alternative data marketplaces [Understand]
4. **Evaluate** the trade-off between data utility and individual privacy [Evaluate]
5. **Identify** privacy-preserving computation techniques relevant to finance [Understand]

Bloom's levels covered: Understand, Analyze, Evaluate

Objectives follow Bloom's taxonomy: Understand → Analyze → Evaluate.

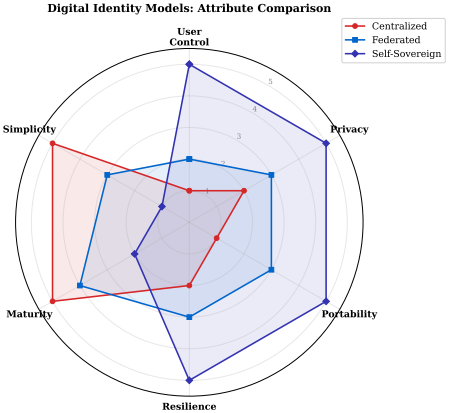
Bridge: From the Current System to What Comes Next

Modules 1–7 built the current system:

- Module 1: How money moves (cost)
- Module 2: Who can access finance (access)
- Module 3: Whom we trust (trust)
- Module 4: How we manage risk
- Module 5: What we automate
- Module 6: Infrastructure we rely on
- Module 7: Rules we follow (compliance)

Module 8 asks: what comes next?

- Who controls your digital identity?
- Who profits from your data?
- Can we have **both** utility and privacy?



Three paradigms of digital identity, each with different trust assumptions.

The future of finance depends on solving the identity and data problems simultaneously.

Definition: Digital Identity

A **digital identity** is the collection of electronically stored attributes, credentials, and behavioral data that uniquely represents an individual, organization, or device in digital systems.

In financial services, digital identity must answer three questions:

1. **Authentication:** Are you who you claim to be?
2. **Authorization:** What are you allowed to do?
3. **Attribution:** Who is responsible for this action?

Why this matters for finance:

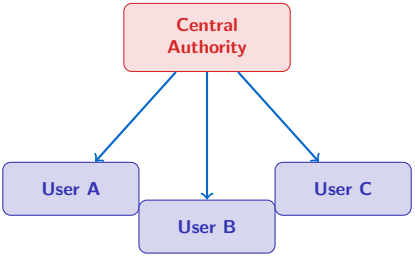
- KYC (Know Your Customer) is the gateway to every financial service
- Identity fraud costs the global financial industry an estimated $40+ billion per year (synthetic figure)
- 1.4 billion people worldwide lack formal identity documents — excluding them from finance

Digital identity is the foundation of financial inclusion, fraud prevention, and regulatory compliance.

Model 1: Centralized Identity

Centralized Identity

A single authority issues, stores, and verifies identity. The user has no control over how the identity data is used.



All identity flows through one authority.

Examples:

- Government-issued national ID numbers
- Bank-issued customer IDs
- Social media login ("Sign in with Platform X")

Strengths:

- Simple, well-understood model
- Single point of accountability

Weaknesses:

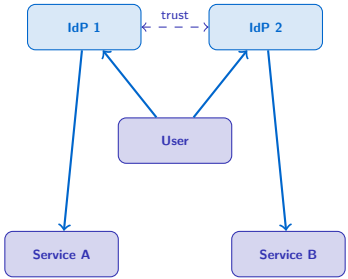
- Single point of failure — one breach exposes all users
- User has no control over their own data
- Creates data silos between institutions

Centralized identity dominates today but concentrates risk and power in a single entity.

Model 2: Federated Identity

Federated Identity

Multiple identity providers agree on standards so users can use one credential across many services. Trust is distributed among a **federation** of providers.



Multiple providers share trust through federation agreements.

Examples:

- European eIDAS framework (cross-border ID recognition)
- "Login with Google / Apple" across websites
- Nordic BankID systems (Sweden, Norway)

Strengths:

- Reduces duplication — one login, many services
- Interoperability across institutions

Weaknesses:

- User still depends on the identity provider
- Provider can track user activity across services
- Federation governance is complex

Model 3: Self-Sovereign Identity (SSI)

Self-Sovereign Identity (SSI)

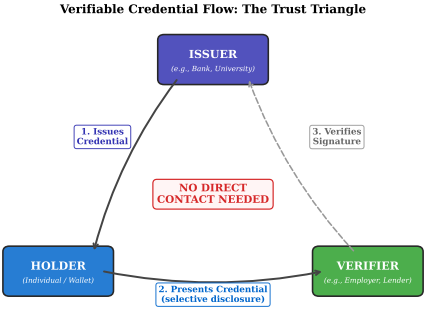
The **individual** holds and controls their own identity credentials in a digital wallet. No central authority is required to verify identity at the time of use.

Core principles:

- **User-controlled:** You hold your own credentials
- **Portable:** Credentials work across any service
- **Minimal disclosure:** Share only what is needed
- **Verifiable:** Any party can cryptographically verify authenticity

Key technologies:

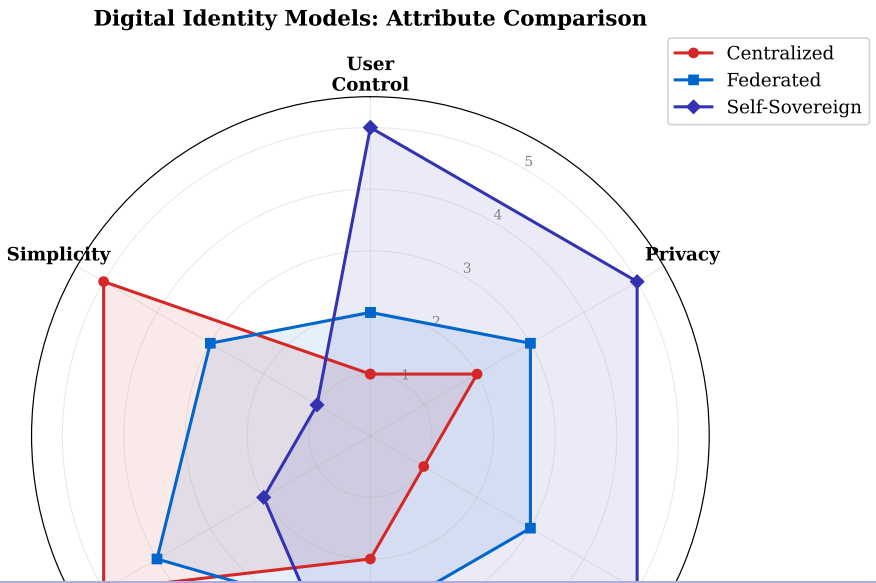
- Decentralized Identifiers (DIDs)
- Verifiable Credentials (VCs)
- Zero-Knowledge Proofs (ZKPs)



The "trust triangle" of SSI: Issuer, Holder, Verifier.

SSI shifts power from institutions to individuals — the most radical identity paradigm shift.

Comparing Identity Models



Verifiable Credentials: The Building Block of SSI

Definition: Verifiable Credential (VC)

A **verifiable credential** is a tamper-proof, cryptographically signed digital document that contains claims about a subject (e.g., "this person is over 18") and can be verified by any party without contacting the issuer.

The trust triangle:

1. **Issuer** — Creates and signs the credential (e.g., a university issues a diploma)
2. **Holder** — Stores the credential in a digital wallet (e.g., the graduate)
3. **Verifier** — Checks the credential's authenticity (e.g., an employer)

Financial use cases:

- KYC credential: "Bank X verified this person's identity on date Y"
- Credit credential: "This person has a credit score above 700"
- Accreditation: "This person is a certified financial advisor"

Verifiable credentials decouple the act of issuing identity from the act of verifying it.

Zero-Knowledge Proofs: Proving Without Revealing

Zero-Knowledge Proof (ZKP) – Conceptual

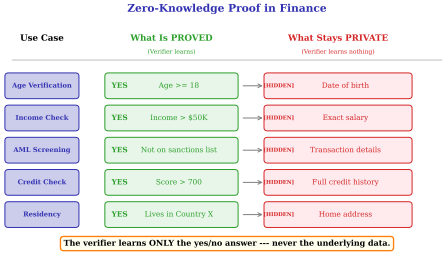
A **zero-knowledge proof** allows one party (the prover) to convince another party (the verifier) that a statement is true **without revealing any information** beyond the truth of the statement itself.

Everyday analogy:

- You prove you are over 18 **without showing your date of birth**
- You prove you have sufficient funds **without revealing your balance**
- You prove you are a citizen **without revealing your passport number**

Key properties:

- **Completeness:** If the statement is true, the verifier will be convinced
- **Soundness:** If the statement is false, no cheating prover can convince the verifier



ZKP Applications in Finance

Where zero-knowledge proofs solve real financial problems:

Use Case	What Is Proved	What Stays Private
Age verification	Customer is ≥ 18	Exact date of birth
Income verification	Income exceeds threshold	Exact salary amount
AML compliance	Funds are not on sanctions list	Transaction details
Credit check	Score above minimum	Exact score, debt history
Accredited investor	Net worth $\geq \$1M$	Full financial statement
Tax compliance	Tax paid matches obligation	Detailed income breakdown

Key insight: Financial regulation often requires proving a *predicate* (yes/no condition), not revealing the *underlying data*. ZKPs make this possible.

ZKPs enable regulatory compliance with minimal data exposure — the best of both worlds.

The Data Economy: Your Data Has a Price

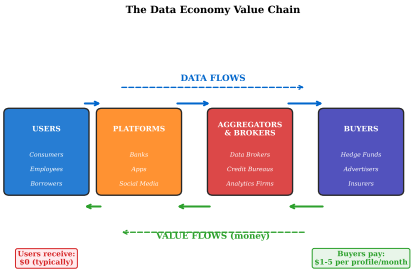
Definition: Data Economy

The **data economy** is the ecosystem in which personal, transactional, and behavioral data is collected, aggregated, analyzed, sold, and used as an economic input — often without the knowledge or meaningful consent of the individuals who generated it.

Scale of the data economy:

- Global data market estimated at $200+ billion (synthetic estimate)
- Financial data is among the **most valuable** categories
- Average consumer's data generates $50–$200/year for data brokers (illustrative)

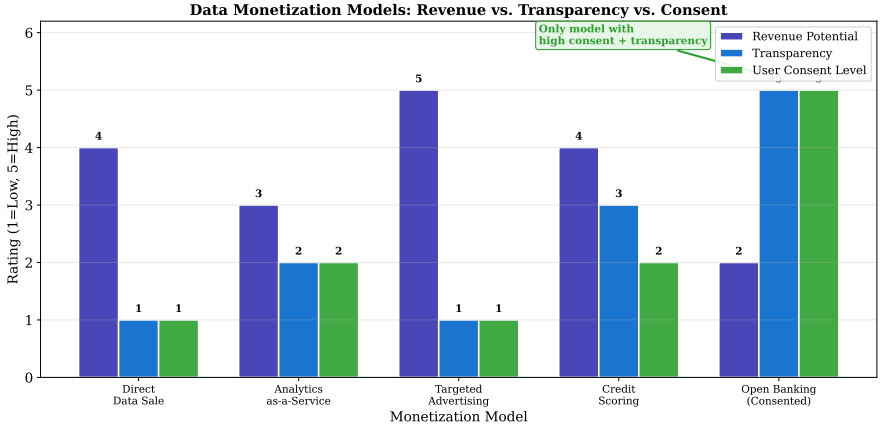
The paradox: Users generate the data, but platforms and brokers capture most of the value.



Data flows from users to brokers to buyers — value flows the other way.

The data economy creates enormous value but distributes it unevenly between data generators and data exploiters.

How Financial Data Is Monetized

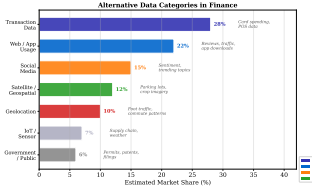


- **Direct sale:** Data brokers sell raw consumer financial profiles
- **Analytics-as-a-service:** Aggregated insights sold without raw data
- **Advertising:** Targeted financial product ads based on spending patterns
- **Credit scoring:** Bureau scores derived from borrower payment data
- **Open Banking APIs:** Consented data sharing (e.g., PSD2 in Europe)

Alternative Data: Beyond Traditional Financial Records

Definition: Alternative Data

Alternative data refers to non-traditional data sources used in financial analysis — including satellite imagery, social media sentiment, web traffic, geolocation, and IoT sensor data — that supplement conventional financial statements and market data.



Growth drivers:

- Traditional data (earnings, filings) is available to *everyone*
- Alternative data provides **informational edge**
- Hedge funds spend an estimated $1–2 billion/year on alternative data (illustrative)

Alternative data is transforming investment analysis but raises serious privacy and consent questions.

Alternative Data: A Taxonomy

Category	Examples	Financial Use
Satellite / Geospatial	Parking lot traffic, crop growth	Retail revenue estimation
Social media	Sentiment, trending topics	Market sentiment analysis
Web / App usage	Product reviews, web traffic	Consumer demand forecasting
Transaction data	Aggregated card spending	Revenue "nowcasting"
Geolocation	Foot traffic, commute patterns	Commercial real estate valuation
IoT / Sensor	Supply chain sensors, weather	Commodity price prediction
Government / Public	Permits, filings, patents	Competitive intelligence

Key concern: Much alternative data is derived from individuals who did **not consent** to its use in financial analysis.

The value of alternative data lies precisely in its asymmetry — those who have it can outperform those who do not.

Data Marketplaces: Buying and Selling Information

A growing ecosystem connects data providers with financial buyers:

Type	How It Works	Financial Relevance
Data exchanges	Structured marketplace for curated datasets	Hedge funds buy alternative data feeds
Data-as-a-Service	API access to real-time data streams	Banks integrate risk signals
Data cooperatives	Members pool data for mutual benefit	Credit unions share anonymized data
Personal data markets	Individuals sell their own data	Experimental (e.g., data dividends)

Quality challenges:

- **Provenance:** Where did the data come from? Was it ethically collected?
- **Freshness:** How current is the dataset?
- **Coverage:** Does it represent the population, or is it biased?
- **Compliance:** Does using this data comply with GDPR, CCPA, etc.?

Data marketplaces are the "stock exchanges" of the data economy — but regulation is still catching up.

The Consent Problem

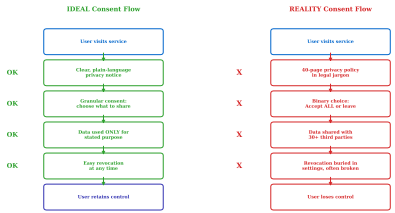
How consent should work:

1. User is informed about what data is collected
2. User understands how it will be used
3. User freely agrees (or declines)
4. User can revoke consent at any time

How consent actually works:

1. User is shown a 40-page privacy policy
2. User clicks “I Agree” without reading
3. Data is shared with dozens of third parties
4. Revoking consent is technically difficult or impossible

Average person would need 76 working days per year to read all the privacy policies they encounter (illustrative estimate).



“Informed consent” in the data economy is largely a fiction — the information asymmetry is too great.

Privacy Regulation: The Global Response

Regulation	Jurisdiction	Key Provisions
GDPR (2018)	European Union	Right to erasure, data portability, consent requirements, €20M or 4% revenue fines
CCPA / CPRA	California, USA	Right to know, delete, opt-out of data sale
LGPD (2020)	Brazil	Modeled on GDPR, covers all sectors
PIPL (2021)	China	Consent-based, strict cross-border transfer rules
DPDPA (2023)	India	Consent-based, government exemptions
PSD2 / Open Banking	EU / UK	Customer-consented data sharing via APIs

Trend: Privacy regulation is expanding globally, but enforcement and scope vary widely.

Key tension: Regulators want both **data protection** (GDPR) and **data sharing** (Open Banking) simultaneously. Reconciling these goals is a central challenge.

Financial institutions must navigate overlapping and sometimes contradictory privacy regulations.

The Fundamental Trade-off: Data Utility vs. Privacy

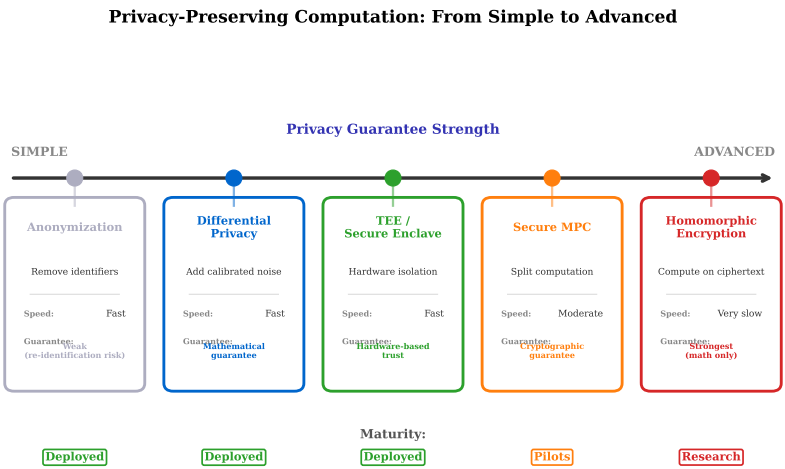
More data sharing creates more value — but also more risk:

Benefits of Data Sharing	**Risks of Data Sharing**
Better credit decisions for underserved populations	Identity theft and fraud
Faster, cheaper KYC processes	Discrimination through profiling
Improved fraud detection	Surveillance and loss of autonomy
More accurate risk models	Data breaches with cascading harm
Financial inclusion for the unbanked	Manipulation (e.g., predatory lending targets)

Central question of this lesson: Can technology allow us to extract data's utility *without* sacrificing individual privacy?

The utility–privacy trade-off is not binary — emerging technologies aim to maximize both.

Privacy-Preserving Computation: An Overview



- **What you see:** Five techniques arranged from simple (anonymization, differential privacy) to advanced (homomorphic encryption), with trade-offs in speed, privacy strength, and maturity
- **Key pattern:** Three techniques are already deployed (anonymization, differential privacy, TEEs); SMPC and homomorphic

Definition: Differential Privacy

Differential privacy adds carefully calibrated noise to data or query results so that no individual's data can be identified, while aggregate statistics remain accurate.

How it works (conceptual):

- A query asks: "What is the average income in this ZIP code?"
- The system adds random noise to the answer (e.g., $\pm$\$500)
- The aggregate statistic is approximately correct
- But no individual's income can be reverse-engineered

Financial applications:

- Aggregate credit risk analytics without exposing individual borrowers
- Market research on spending patterns
- Regulatory reporting with privacy guarantees

Trade-off: More noise = more privacy, but less accuracy. The "privacy budget" (ϵ) controls this balance.

Differential privacy provides a mathematically rigorous guarantee — unlike ad-hoc anonymization.

Homomorphic Encryption (Conceptual)

Definition: Homomorphic Encryption

Homomorphic encryption allows computations to be performed on **encrypted data** without ever decrypting it. The result, when decrypted, is the same as if the computation had been performed on the plaintext.

Conceptual example:

1. Bank A encrypts its fraud data and sends it to a shared analytics platform
2. The platform runs a fraud detection model **on the encrypted data**
3. The result is returned, encrypted, to Bank A
4. Bank A decrypts the result — the analytics platform **never saw the raw data**

Current limitations:

- Extremely computationally expensive (100–10,000× slower than plaintext)
- Practical for simple operations; complex ML models remain challenging
- Active research area — performance improving yearly

Homomorphic encryption is the "holy grail" of privacy computation but remains too slow for most real-time financial applications.

Secure Multi-Party Computation (SMPC)

Definition: Secure Multi-Party Computation

SMPC allows multiple parties to jointly compute a function over their combined data without any party revealing its private input to the others.

Financial use case — Fraud detection consortium:

- Five banks each hold their own transaction data
- They want to build a shared fraud detection model
- **Problem:** Sharing raw data violates privacy law and competitive interests
- **SMPC solution:** Each bank contributes encrypted data shares; the joint computation produces a fraud model without any bank seeing another's data

Other applications:

- Collaborative credit scoring across lenders
- Regulatory stress tests using combined industry data
- AML watchlist matching without revealing customer lists

SMPC enables industry collaboration that is currently impossible due to privacy and competitive barriers.

Trusted Execution Environments (Secure Enclaves)

Definition: Trusted Execution Environment (TEE)

A **TEE** (or "secure enclave") is a hardware-isolated area within a processor where code and data are protected from the rest of the system — even from the operating system and system administrators.

How it works:

- Sensitive data is loaded into the enclave
- Computation happens inside the enclave
- Results are exported, but raw data **never leaves** the enclave
- The hardware provides cryptographic proof ("attestation") that the correct code ran

Financial applications:

- Cloud-based financial analytics on confidential data
- Cross-border regulatory data sharing
- Key management for digital asset custody

Limitation: TEEs depend on trusting the hardware manufacturer. Hardware vulnerabilities (e.g., side-channel attacks) have been demonstrated.

TEEs offer practical privacy today but require trust in hardware, not just mathematics.

Comparing Privacy-Preserving Techniques

Property	Diff. Privacy	Hom. Encrypt.	SMPC	TEE
Data accuracy	Approximate	Exact	Exact	Exact
Performance	Fast	Very slow	Moderate	Fast
Trust model	Math	Math	Math	Hardware
Multi-party	No	No	Yes	With setup
Complexity	Low	Very high	High	Medium
Maturity	Deployed	Research	Pilots	Deployed

Key insight: No single technique solves all problems. Real-world systems often combine multiple approaches:

- TEE for fast execution + differential privacy for output protection
- SMPC for multi-bank collaboration + homomorphic encryption for specific sub-computations

The privacy toolbox is growing — future financial infrastructure will layer multiple techniques.

Putting It Together: A Future Financial Identity System

Imagine a financial identity system built on these technologies:

1. **SSI wallet** holds your verifiable credentials (KYC, income proof, credit score)
2. **ZKP** lets you prove "income $>$ \$50K" without revealing the amount
3. **Differential privacy** ensures aggregate analytics cannot identify individuals
4. **SMPC** allows banks to collaborate on fraud detection without sharing customer data
5. **TEE** runs risk models on your encrypted data in the cloud

What this enables:

- Instant, portable KYC — open a bank account in minutes, anywhere
- Financial inclusion — prove creditworthiness without a formal credit history
- Regulatory compliance without mass surveillance
- Competitive collaboration — banks cooperate on fraud while competing on products

---

The future financial system can be both more open and more private — if we build it right.

Challenges and Barriers to Adoption

Why aren't we there yet?

Barrier	Description
Performance	Homomorphic encryption and SMPC are still too slow for high-throughput finance
Standards	SSI and VC standards are still evolving (W3C DID, ISO 18013-5)
Incentives	Incumbents profit from data asymmetry — why give it up?
Regulation	Privacy laws vary by jurisdiction; cross-border identity is unsolved
User experience	Cryptographic wallets are still too complex for average consumers
Governance	Who resolves disputes in a decentralized identity system?
Key management	Losing a private key could mean losing your entire identity

Adoption will be gradual: Expect hybrid systems — centralized identity with SSI extensions — before full decentralization.

Technical feasibility is necessary but not sufficient — adoption requires solving governance, UX, and incentive problems.

Case Study: EU Digital Identity Wallet (eIDAS 2.0)

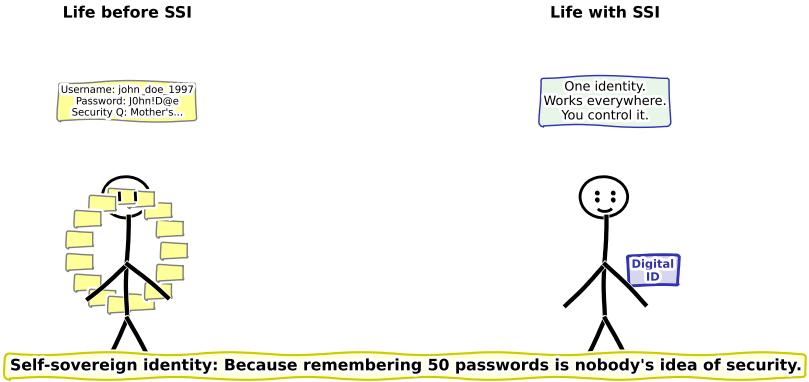
The EU is building the world's first large-scale government-backed digital identity wallet:

- **Regulation:** eIDAS 2.0 (revised 2024) mandates EU member states to offer a Digital Identity Wallet to all citizens
- **Features:**
  - Store national ID, driver's license, diplomas, health data
  - Selective disclosure — share only what is needed
  - Cross-border recognition across all 27 EU member states
  - Legally equivalent to in-person identification
- **Financial implications:**
  - Banks can onboard customers using the wallet (instant KYC)
  - Reduces duplicate identity verification across institutions
  - May disrupt the €10+ billion identity verification industry

Timeline: Full deployment expected by 2026–2027.

The EU Digital Identity Wallet is the most ambitious government-led digital identity project in the world.

One Last Thought...



Sometimes the best way to remember a concept is to laugh about it.

Key Takeaways

1. Digital identity has evolved through three paradigms: **centralized**, **federated**, and **self-sovereign**
2. **Verifiable credentials** and **zero-knowledge proofs** enable privacy-preserving identity verification
3. The **data economy** generates enormous value from personal data, but users capture very little of it
4. **Alternative data** provides informational edges in finance but raises consent and privacy concerns
5. **Privacy-preserving computation** (differential privacy, homomorphic encryption, SMPC, TEEs) enables data utility without exposing raw data
6. The fundamental trade-off is **data utility vs. privacy** — emerging technologies aim to maximize both
7. Adoption requires solving not just technology, but **governance, incentives, and user experience**

The future of finance is a system that is simultaneously more open, more private, and more user-controlled.

Summary / Next Lesson Preview

This lesson: We explored how digital identity is evolving from centralized to self-sovereign models, examined the data economy and alternative data, and surveyed privacy-preserving computation techniques.

Key vocabulary:

- Self-sovereign identity (SSI)
- Verifiable credentials (VCs)
- Zero-knowledge proofs (ZKPs)
- Data economy
- Alternative data
- Differential privacy
- Homomorphic encryption
- Secure multi-party computation
- Trusted execution environments
- electronic Identification, Authentication and trust Services (eIDAS) 2.0

Next lesson (M8L2): *Decentralized Finance 2.0* — We will examine how DeFi protocols are evolving beyond simple token swaps toward programmable financial infrastructure, institutional adoption, and the convergence of traditional and decentralized finance.

Review: Can you explain how a zero-knowledge proof allows KYC verification without exposing personal data?