

Personal Online Data Stores

A Paradigm Shift for Data Collection

Prof Graham Williams, ANU

Chief Scientist, Software Innovation Institute

Cluster Lead, Data Science and Analytics

School of Computing, College of Systems and Society

19 August 2025



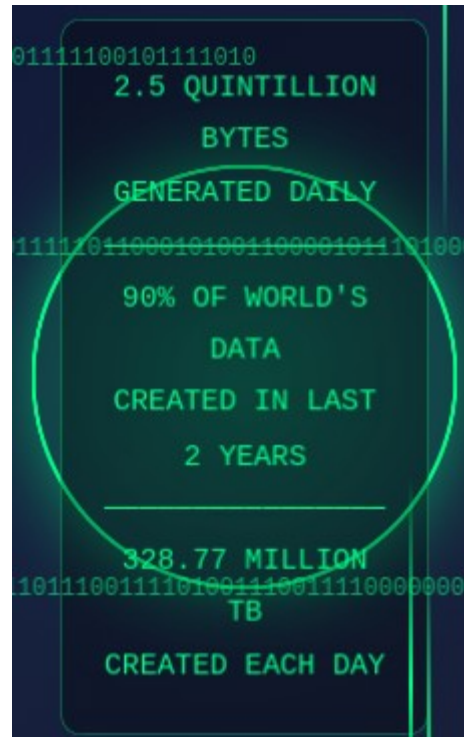
Australian
National
University

Software
Innovation
Institute

Today's World – Digitizing Data

The era of the computer has allowed us to digitize data and in this digital form to capture and share massive amounts of data globally.

Massive personal data (including from the dark web) is readily accessible to us today.

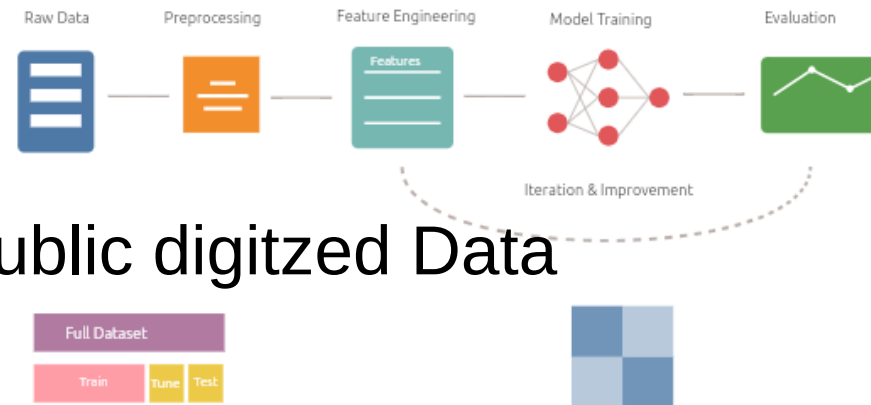


Machine Learning: Data to Knowledge

- We began to build tools to help us turn data into knowledge 1950's AI
- AI and Machine Learning – 1950's to 2025 – 3 or 4 Boom Bust Recycles
- Today's Generative AI and Large Language Models

Massive Data + Massive Compute

= AI Models Encompassing **All** public digitized Data



Data Science – The Joy of Data

- Data science over four decades

AI, ML, Knowledge Discovery, Statistics, Data Mining, Analytics, Data Science

- Cattle Grazing, Bush Fires, Car Loans, ... 1980's
- Qld Health, NRMA Insurance, Home Loans, ... 1990's
- Taxation, Compliance, Health Fraud, Govt, ... 2000's
- Climate, Chip Manufacturing, Recommendations, ... 2010's
- Indigenous Health Care, Clinical Trials, Privacy First Apps, ... 2020's
- Increasingly more data – the era of Massive Compute with Massive Data
- All of **your** data available for “**us**” to support, to monitor, to profit from, **you**, ...

The Data and Compute Boom

- From about 2015 we saw the emergence of massive data centres

Amazon, **Microsoft**, Google, ...

- We started deploying massive amounts of compute on massive data
- Every business needed to become a data based business
- In AI we threw compute at data like never before



Today's LLMs – ChatGPT +++

Data builds LLMs, but whose data, and whose models?

- Data is the foundation of today's AI, Machine Learning, Data Science
- Web has evolved to centralize and monetize everyone's data
- Who's data is it that has driven the massive concentration of wealth
- Large Language Models use massive data
- without permission/credit/sharing

Original Vision to Financial Reality

- Tim Berners-Lee's original vision was overwhelmed by profits
Originally the WWW would be an equaliser – access to all
- Business saw the opportunity for unimaginable profit
And the result is history.
- We have happily populated the commercial and public space with our own and often private data to support, as it turns out, building massive AI models that we might benefit from.



Today's Centralised Data Stores

Why is it a problem?

- Single point of failure
- ...
- **Data privacy and security risks:**
- Centralised data stores as attractive targets,
- compromising the entire data store, ...
- Lack of Transparency and Trust:
- Relying on some central authority to do the right thing.
- Data Ownership and Vendor Lock-In
- Who owns and controls **your** data?



Today's Paradigm Shift

- Indigenous (Māori) Sovereignty

"Power should be with those who the data is about."

- Tim Berners-Lee's original vision for his World Wide Web:

"Your personal data in one place, not spread everywhere."

- **From** centralized data collection and analysis
- **To** extremely distributed data stores.

Personal Online Data Stores – Pods

Solid (Socially Linked Data) Pods (Personal Online Data Stores)

- Identity, authentication, and authorisation through **OpenID**
- Data storage using **RDF** serialisation with **Turtle**
Semantic triples and knowledge graphs
- Granular access control with **Access Control Lists**
- Data encryption and sharing of encrypted data using public key cryptography

Trust No One

Solid Pods

❖ A Personal Online Datastore (Pod) is a place where you can store your own personal data and have complete control over that data.

❖ Individual as the first data user

- Individuals manage their own data in their Pod
- Individuals are better informed, and are better able to engage
- It is the only location for all their data – financial, health, personal, ...
- Multiple Data Sources can be ingested into Pods

❖ Not just a datastore – an ecosystem for innovation

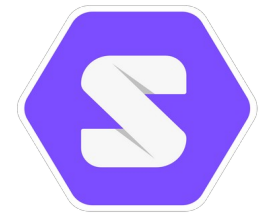
- Innovators can quickly deliver new apps
- Apps to add innovative value to our personal data
- Trusted data stores to improve lives with trusted local apps/agents for my interests
 - Eg: Personal Health Apps supporting individuals and health professionals



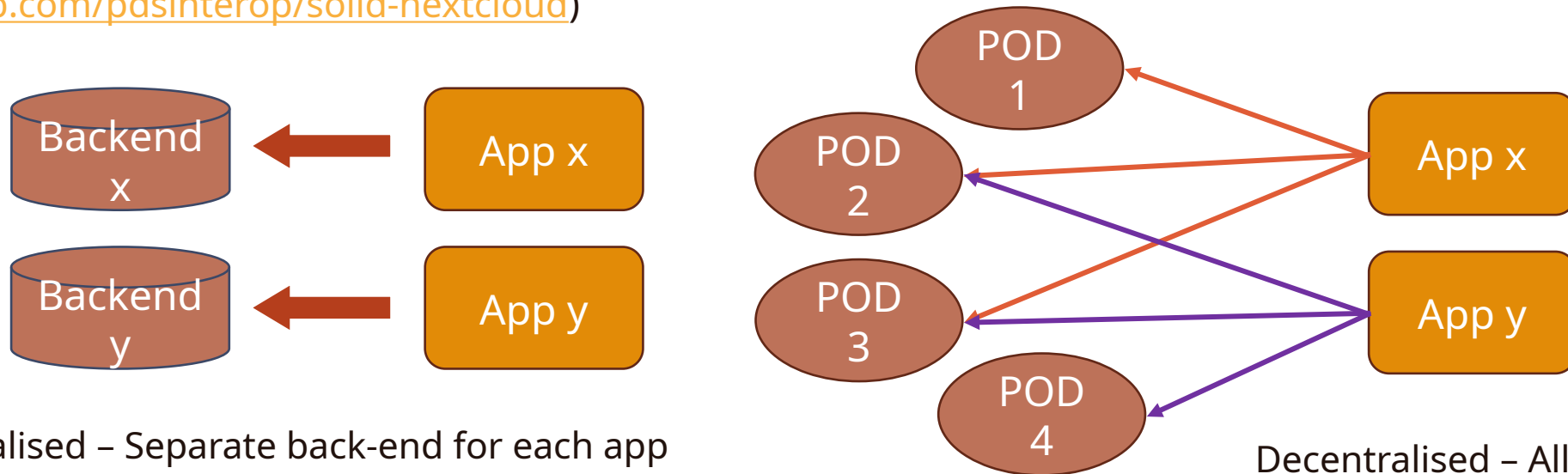
Source: <https://solidproject.org/>

Solid Server

Open Standard for Storing PODS



- ❖ Solid (**S**ocial **l**inked **d**ata) is a web decentralisation project led by Sir Tim Berners-Lee
- ❖ (<https://solidproject.org/> and <https://solidproject.au>)
- ❖ It contains a specification that lets individuals store their data securely in **open-standard** decentralised data stores called Pods, that can host any type of data
- ❖ Multiple **open source** Solid servers are available
 - ❖ Community Solid Server (<https://github.com/CommunitySolidServer/CommunitySolidServer/>), PHP Solid Server (<https://github.com/pdsinterop/php-solid-server>), or a Solid server as a plugin for Nextcloud (<https://github.com/pdsinterop/solid-nextcloud>)



Centralised – Separate back-end for each app

Decentralised – All apps same Pods

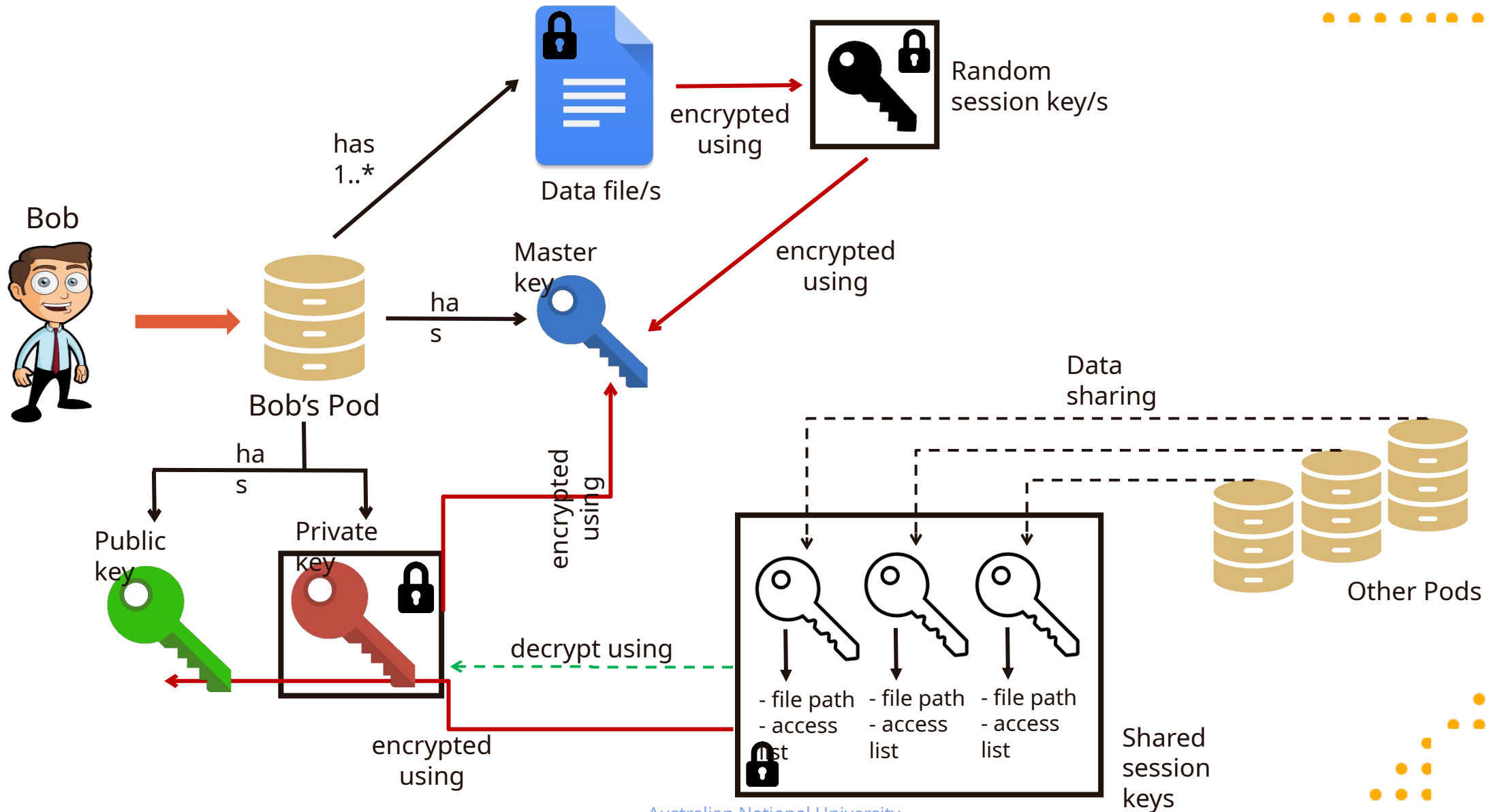
Earning Trust – Encrypted by Default

- ❖ **A zero trust security model.**
- ❖ A privacy and a security architecture where a system does not automatically trust anyone or anything inside or outside its perimeters and instead must verify everything constantly.
- ❖ Microsoft defined three Zero trust principles
 - ❖ **<https://www.microsoft.com/en-us/security/business/zero-trust>**
 - ❖ **Verify explicitly**
 - ❖ **Use least-privilege access**
 - ❖ **Assume breach**
- ❖ *So, how do we ensure a trusted environment in Solid PODs?*

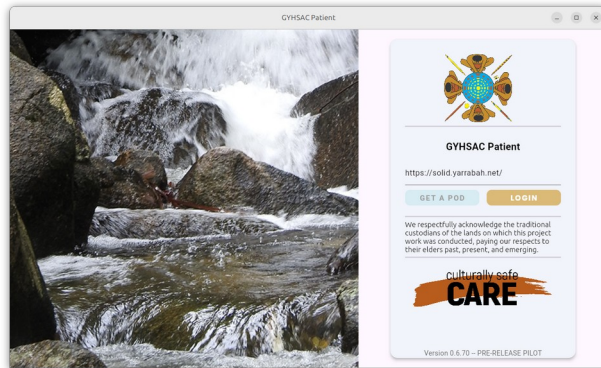
Trust No One – Data in Pods

- ❖ Data stored in Pods can be encrypted.
- ❖ Data are decrypted on device locally. No master key in Pods.
- ❖ Access to data can be controlled by Access Control Lists/Policies (ACLs).
- ❖ Encrypted data are shared with others using public-key infrastructure.

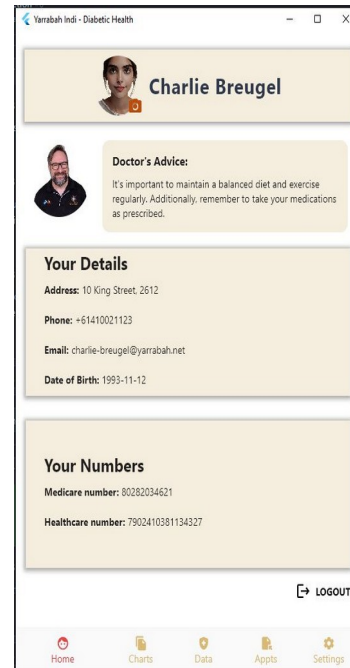
An Architecture for TNO in Pods



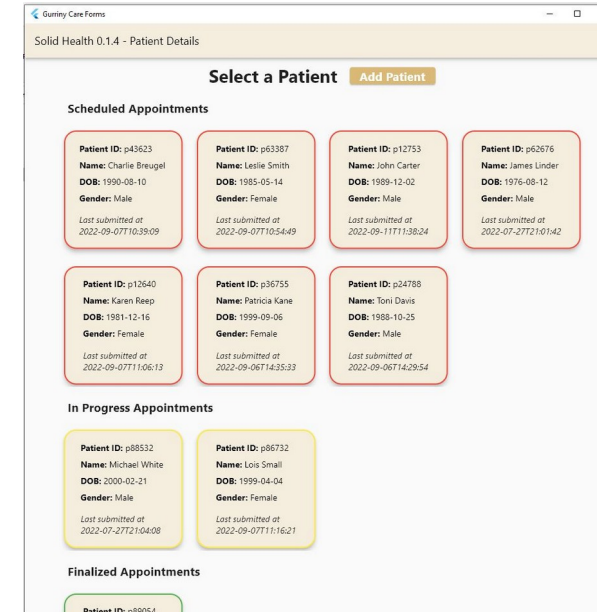
Gurriny Indigenous Health Care



Clinic App to provide interface for doctors to analytics, insights, data for evaluation



Patient App to provide patients with diabetes info support tools from their GP, health diary to record well being, data access and control to support strength-based approach to diabetes management



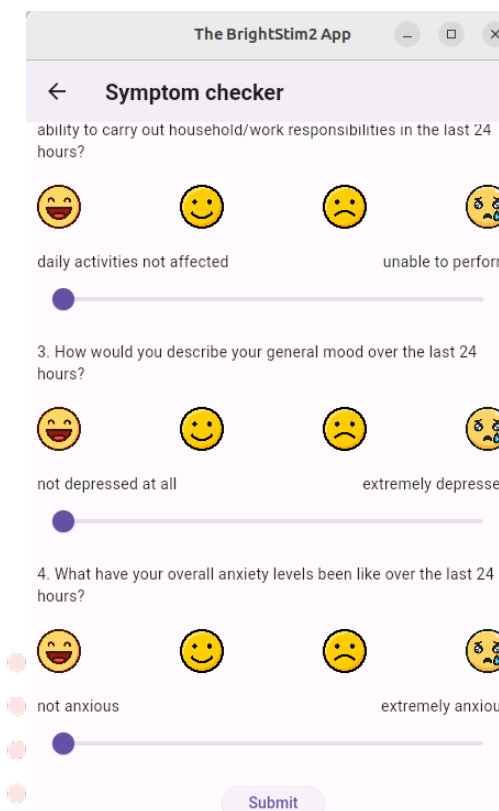
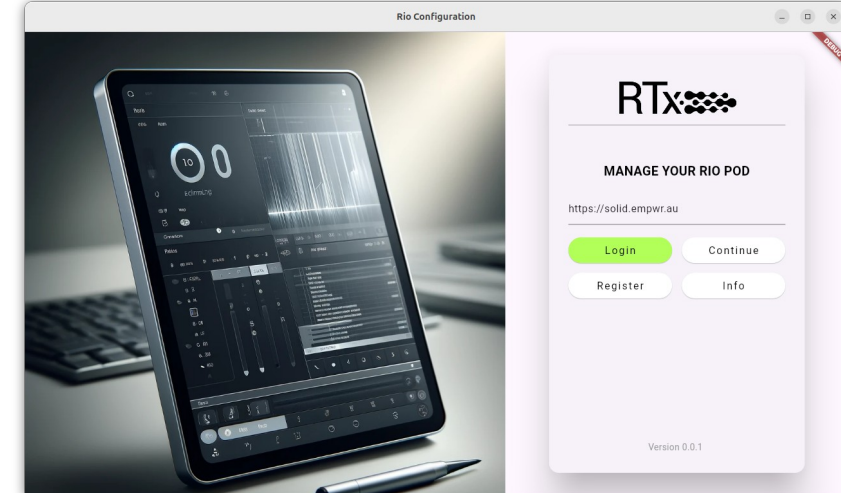
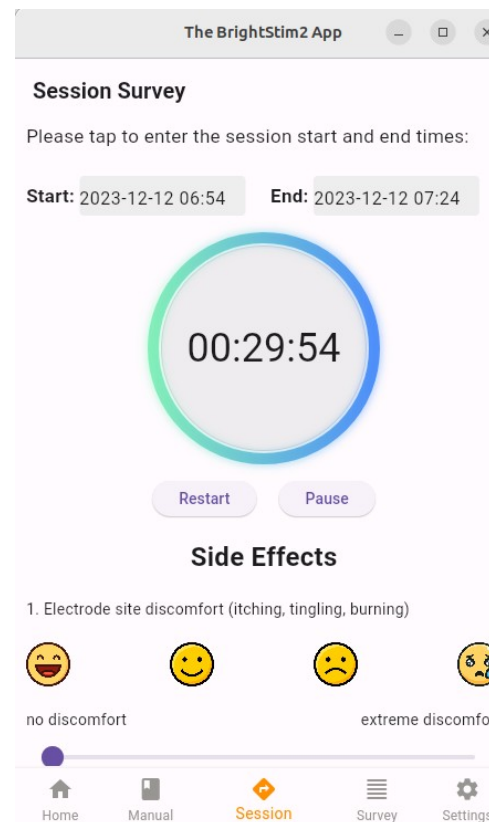
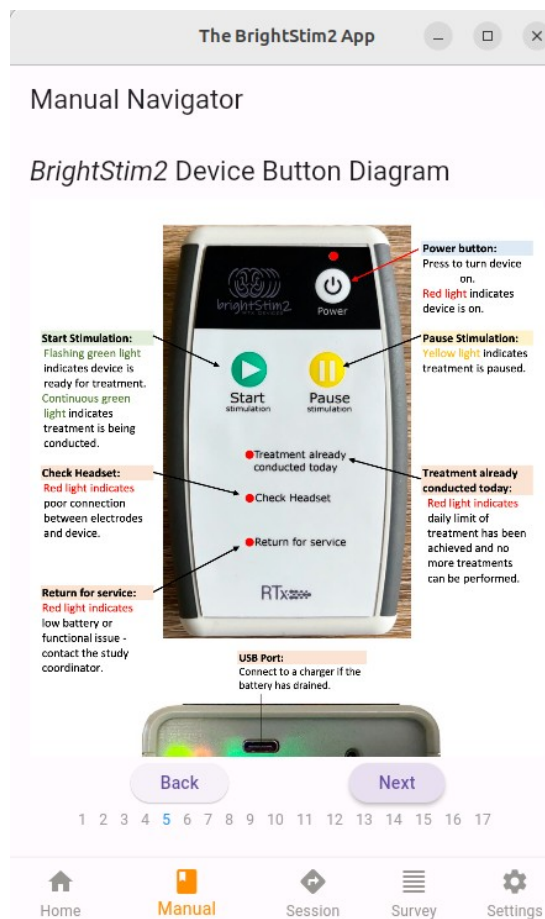
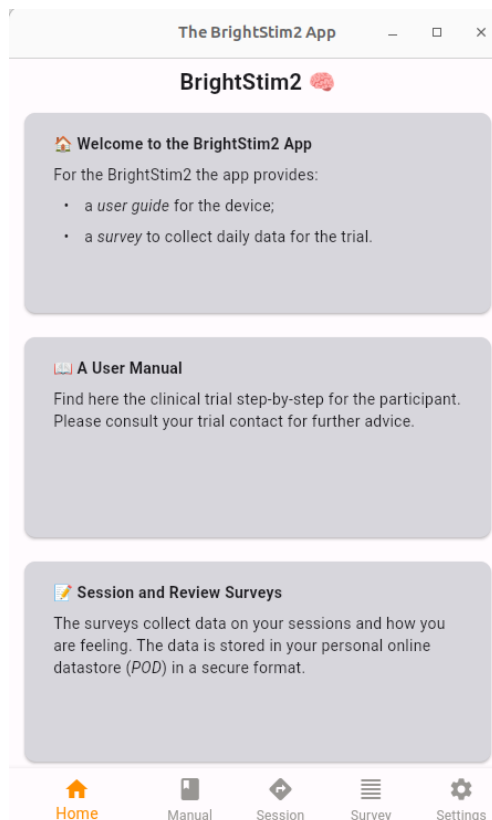
Care Coordination Team App for patient record access and data collection during home care visits to patients



Individuals have control over all their medical and health data.

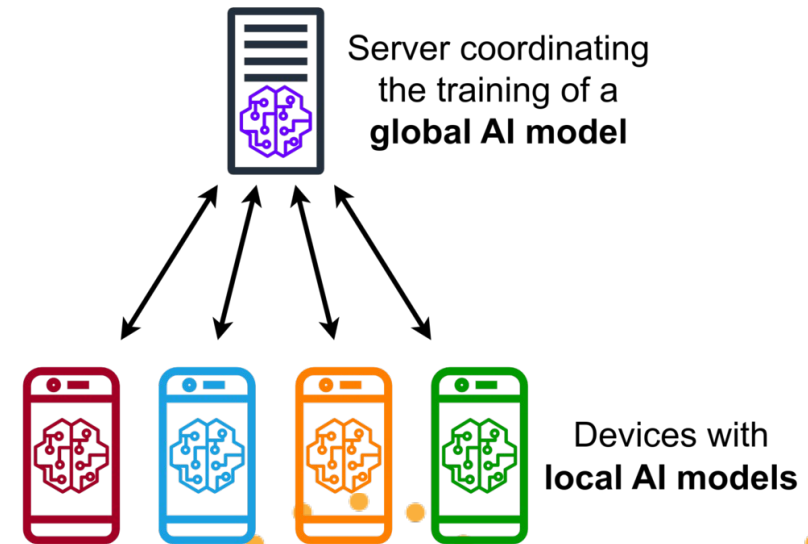
culturally safe
CARE

Empower Medical Devices Supporting Clinical Trials



Federated Learning

- ❖ Personalised services → medical, financial, educational
- ❖ Extremely distributed ML models → With dynamic and fast changing data
- ❖ Collaborations → More collaborative real-time data sharing + analytics, less data breaches (reduced centralised data storage)



SolidCommunity.au

SOFTWARE SHOWCASE

We have many Solid apps under development. Give them a go.

Our apps are written in [Flutter](#) and are open source, and run on any platform. You can try them out in the browser here or visit their github homes. We are also publishing them on the [Google Play Store](#). They are not all there yet, but keep an eye out for them. We are creating the future, so please bear with us as we bring it all together.

Tap on the link icon for any of the below apps to explore the web version of the app.

Template/Demo apps

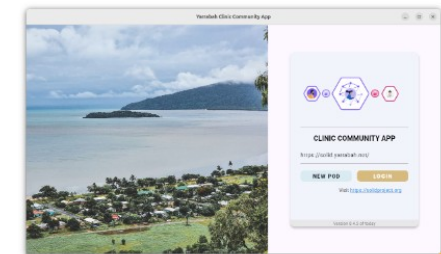
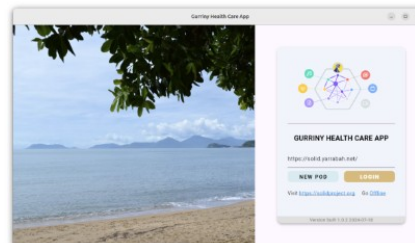
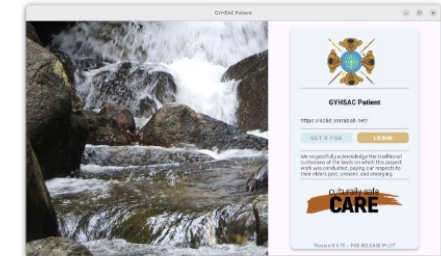
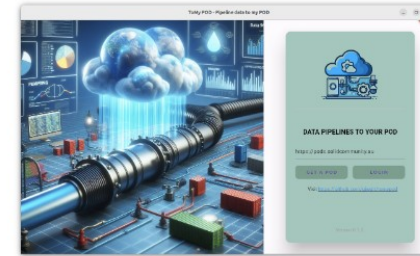
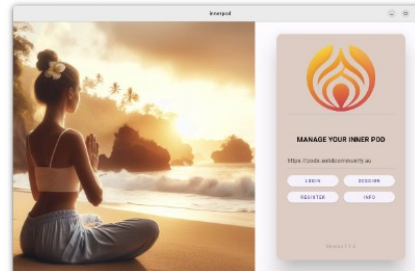
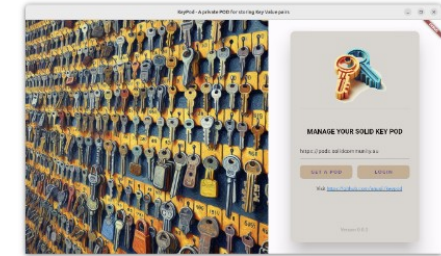
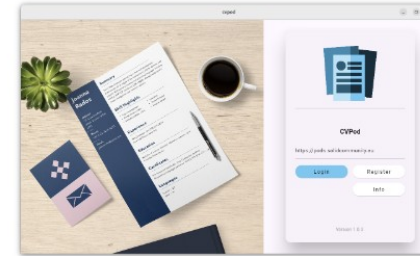
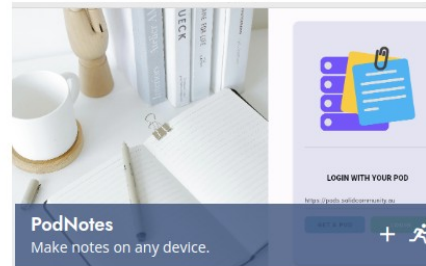
- NotePod – sharing
- HealthPod – personal ML
- MoviStar – recommender

Developer Libraries

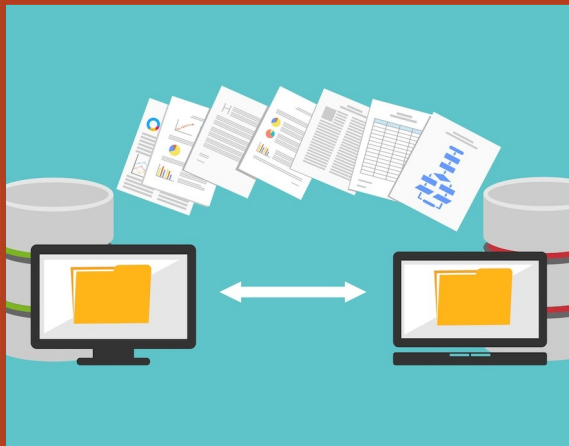
- Solidpod for Flutter

Deploy

- Android
- Windows
- Linux
- iOS
- macOS
- Web



AI Research Challenges



Data migration

Migrating data from other platforms or data sources to PODs can be complicated and time-consuming.



Data security

Data is decrypted on-device and so can be susceptible to **back doors**.



Trust

Once shared a POD owner can stop sharing but data already shared cannot be reliably unshared. Right to be **forgotten**.



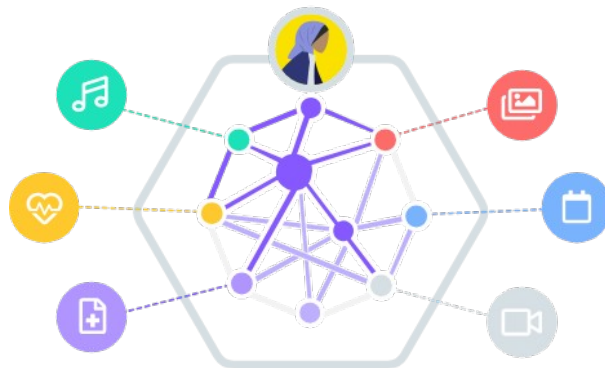
Federated DS + ML

The actual data sharing can be complicated and complex. Developing sophisticated privacy preserving approaches.

Thank You



- ❧ <https://sii.anu.edu.au>
- ❧ <https://solidcommunity.au>
- ❧ <https://solidproject.au>
- ❧ <https://solidproject.org>



Source: <https://solidproject.org/>