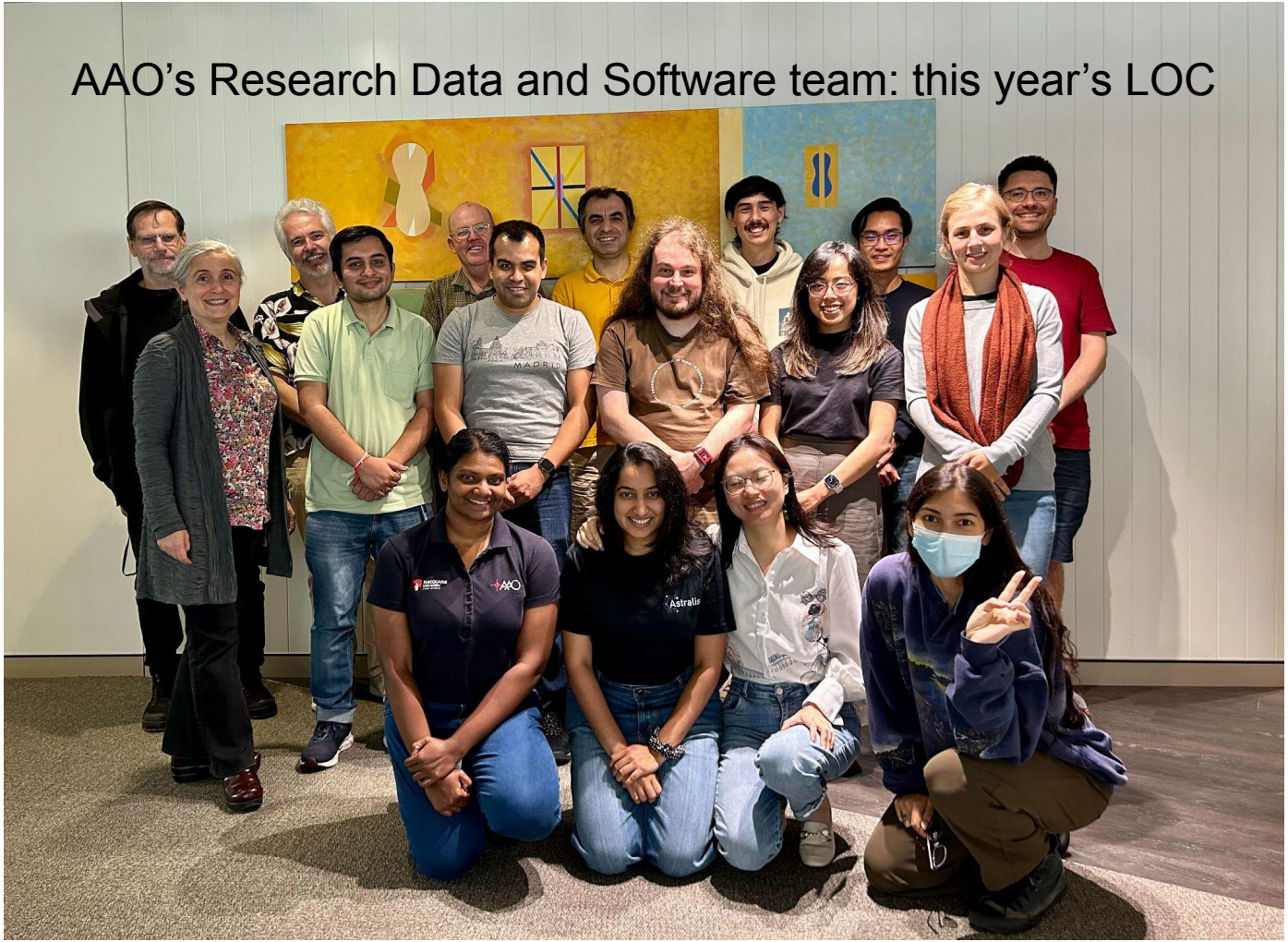
Data Central: current status and future plans

James Tocknell
(on behalf on the Data Central team)

data central

AAO's Research Data and Software team: this year's LOC



LOC has RED on the name badges

Ask them about Data Central during the breaks

data central

What is Data Central — Not Just an Archive!

- Central science platform for astronomy research teams to collaborate
  - Survey planning/management tools (target-selector and observation-log webapps)
  - Large servers for running reduction pipelines/post-processing of results
  - Team data stores and collaboration tools
- Hosts Australia's optical survey data, plus other data that lacks a home
  - Accessible via Web Portal, VO, custom APIs and Jupyterhub (code to the data)
- Creates reusable-but-bespoke applications for astronomers
  - Data Aggregation Service and 4HS Optical Data Inspector
  - Astronomer-friendly wrappers around databases, including ongoing maintenance
  - Tools to compare theory with observation
  - Provides ADACS Software Support (proposal-base support system)
- Applies astronomy tools and knowledge to other fields
  - Archaeology, biology, early childhood education and more

data central

What Data Central hosts

- Host of Australian optical data and as backstop for data that lacks a home
  - Previously no central place for optical data
  - CASDA (*see talk by Minh Huynh*) is funded to support archives and surveys for CSIRO-managed facilities, but this only covers some of Australia's radio data
  - The MWA archive (*see talk by Mouriyan Rajendran*) does not have capacity to host survey-team science products (e.g. GLEAM, GLEAM-X)
- Hosts publicly 19 surveys with 26 data releases, with additional not-yet-public surveys and data releases in various stages of readiness
  - Primarily Australian data, but also mirrors of large object catalogues like 2MASS and WISE to enable effective crossmatching
  - Public data volumes are dwarfed by the amount of internal or in-prep data we are hosting for teams

data central

Our current VO services

- SCS: Oldest of our VO services (pure python + django)
  - Previously was somewhat bespoke using djangorestframework with PrestoDB being the backend; now wraps our containerised TAP service, and provides more details in the VOTable (e.g. the underlying query run)
- SIA (v2) + datalink cutouts (pure python + django)
  - Most of our optical data from survey teams is given as large swarps, so we do cutouts rather than serving out the whole swarp
- SSA + datalink for original and “simple 1D fits” spectra (pure python + django)
  - Survey teams provide spectra in various formats, we serve out both the original files as well as simple 1D fits files for quick-look purposes in splat (or similar tool), and provide specutils loaders for the original files
- TAP: Uses the excellent vollt codebase by Grégory Mantelet with a PostgresSQL cluster as the database
  - Citus is the PostgreSQL cluster tool we use for increased performance

data central

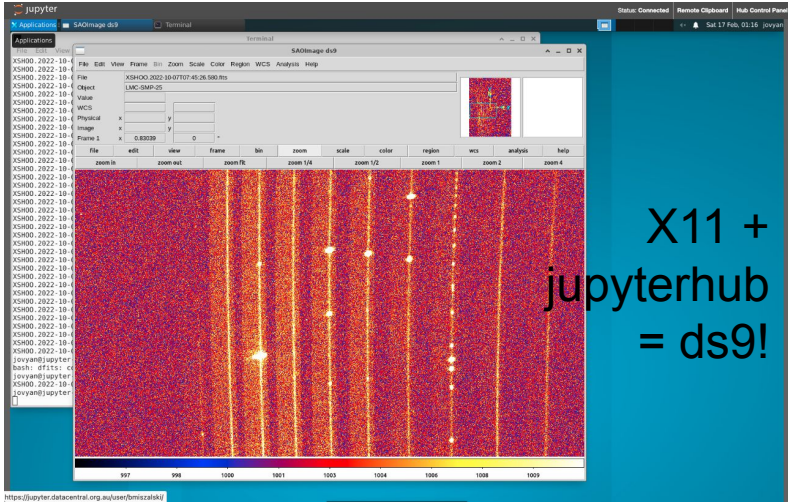
New, In development and future VO services

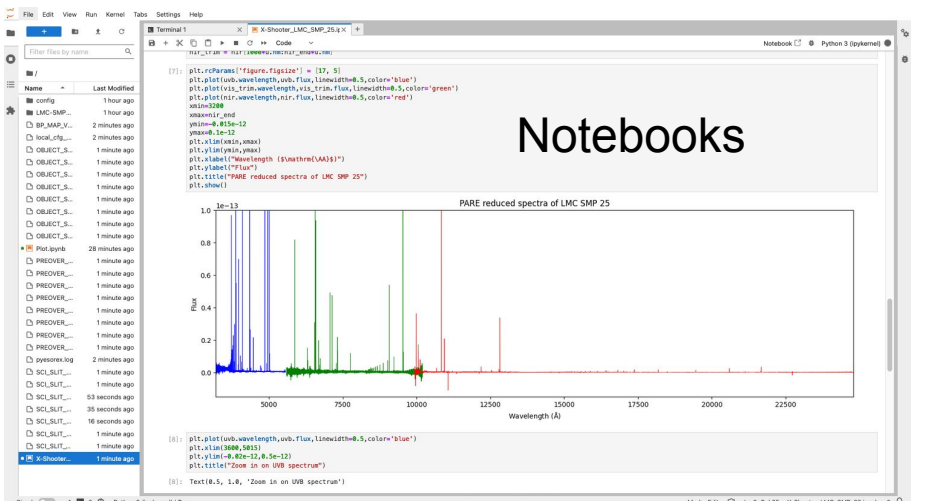
- SSA for Integral Field Spectra (IFS)
  - Our current SSA implementation was designed around 1D spectra
  - We have implemented a new SSA service (which reuses parts of our existing service) which is designed to work with IFS products
- “Simple Data Access”/datalink for other products
  - We’re also going to be hosting data products which are the outputs of post-processing (best fits, models, plots etc. that are in either jpg/png or bespoke formats), and these need to be queryable also
- Datalink all the things!
  - Each of our current datalink services do not link to the other services, introducing crosslinks between them would allow for finding more related data

data central

Data Central: Science Platform

- Providing compute + data for both individuals and teams
  - Built on k3s, with Ceph and Proxmox coming in to replace legacy systems
- Jupyterhub: code to all the data!
  - **Direct access to all public data**, plus (in dev) private data releases and personal and team spaces via custom FUSE file system
  - Specific images for targeted reduction environments for ESO and the AAT
- VMs/dedicated hardware for large scale reduction for ESO large programmes
  - Accessible via VPN or via browser (using Apache Guacamole)
  - We may look at providing LXC containers to allow for a more “unix-server-like” setup





data central

Data Central: Science Platform

- Large scale data upload/download/sync
  - WebDAV (OwnCloud/NextCloud)
  - RSync (+ ContainerSSH in the future)
  - Possibly replace/complement WebDAV with VOSpace, given WebDAV's issues with large file trees?
- Infrastructure for SSO/access control and team management
  - Spin up new instances of services via containers, enabling us to port in team-specific services/tools that need a new home (e.g. GLEAM-X R-shiny team tool, Smart Pulsar database/data viewer)
- DCvalidate: saving time for survey teams and our engineers!
  - Scalable solution to validating survey team data and metadata for ingestion to VO+web portal backend

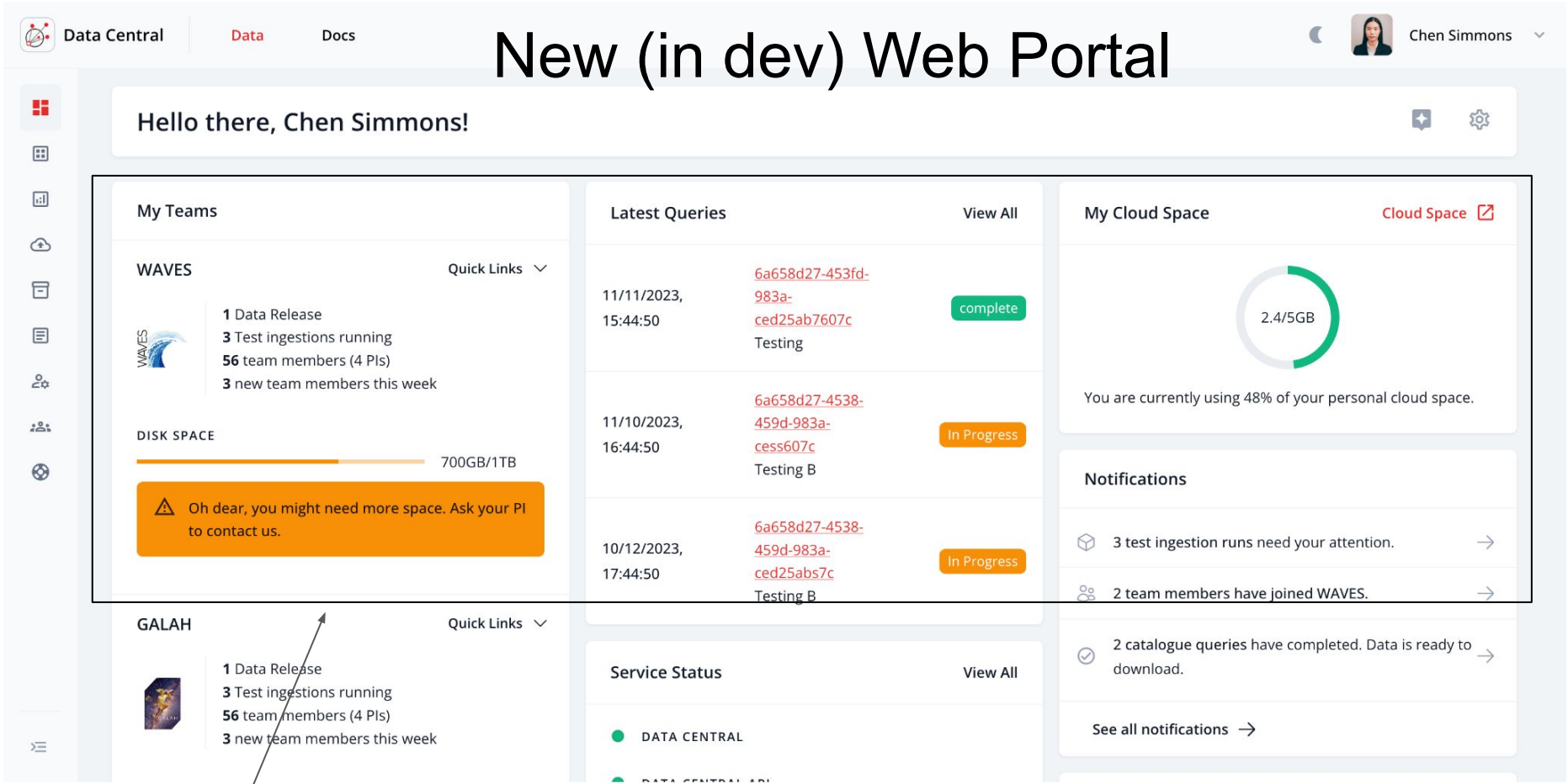
data central

Reusable-but-Bespoke Science Applications

- Data Aggregation Service (DAS) developed for CRAFT (Fast Radio Bursts)
  - Builds on VO services with a more efficient and scalable async client architecture
  - Made general purpose for any transient research, and tech stack informed many of the other apps we've built
- Hector: IFU-based galaxy survey using on AAT
  - Target selector app built on **Aladin-lite** to validate input catalogues
  - Galaxy Morphology app built from components of target selector app
  - Observations log app to track details of how survey is progressing (e.g. does a field need to be re-observed)
- 4HS: 4MOST survey that is using reprocessed VISTA data for input catalogue
  - Optical Data Inspector for target selection built on **Aladin-lite+hipsgen+mocpy**
- Tools to compare theory with observations
  - Simspin: generate mock galaxy observations
  - CosmoDragon: generate synthetic radio jet emission maps

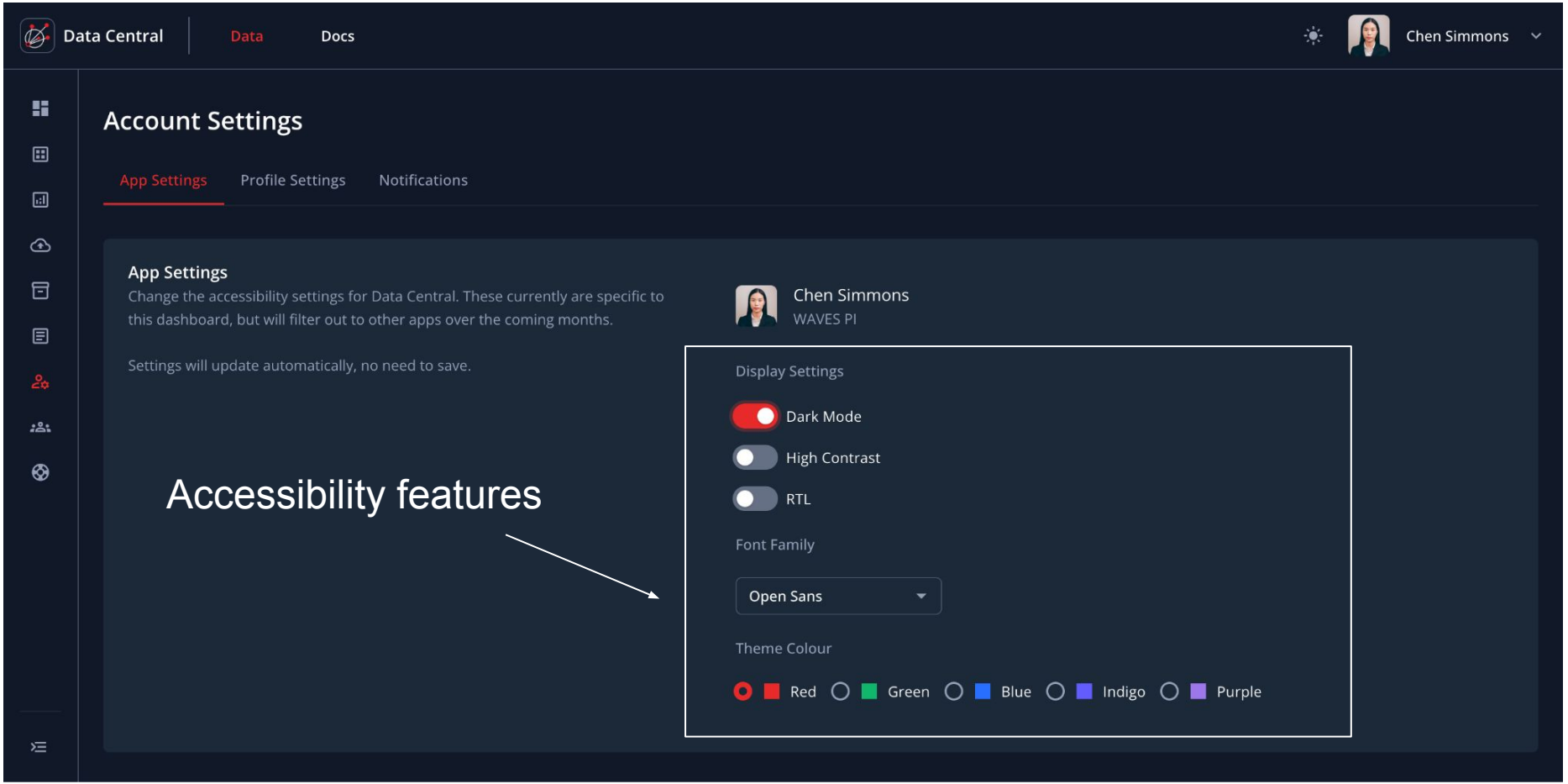
data central

New (in dev) Web Portal



Integrate all services in one portal

data central



data central

Questions?

https://datacentral.org.au/

Tech Stack and tools

- Python, mostly django (and djangorestframework), but with some use of starlette for async/web-socket usage, and celery for worker jobs
- JS/TS with react for frontend (new web portal will use next.js)
- PostgresSQL as database (though some apps use MySQL/MariaDB, MongoDB or InfluxDB as needed)
- Apereo CAS + Active Directory for users/groups
- Team Collaboration tools: Wikis (Wiki.js), Mailing Lists (mailman), Collaborative Document Editors (OwnCloud/NextCloud+OnlyOffice)
- Deploying either through docker-compose or k8s (using k3s as k8s distro)
- Debian-based Linux Distros (traditionally Ubuntu, moving away due to issues with snaps)
- On-Prem hardware (mostly supermicro servers and disc arrays, moving to Ceph for scalable storage and Proxmox for VMs)
    - Cloud in AU is either the most expensive region (or second most, depending on the product), not including exchange rate

data central