



WP3: Computing Requirements

Anna Scaife (UMAN) & Rosie Bolton (UCAM)



The objective of the AENEAS project is to develop a concept and design for a distributed, federated European Science Data Centre (ESDC) to support the astronomical community in achieving the scientific goals of the Square Kilometre Array (SKA)



OBJECTIVES

WP3 will identify and assess the components necessary to bring about a European Science Data Centre, both in hardware and software, from a **total science delivery perspective**.

The focal questions are:

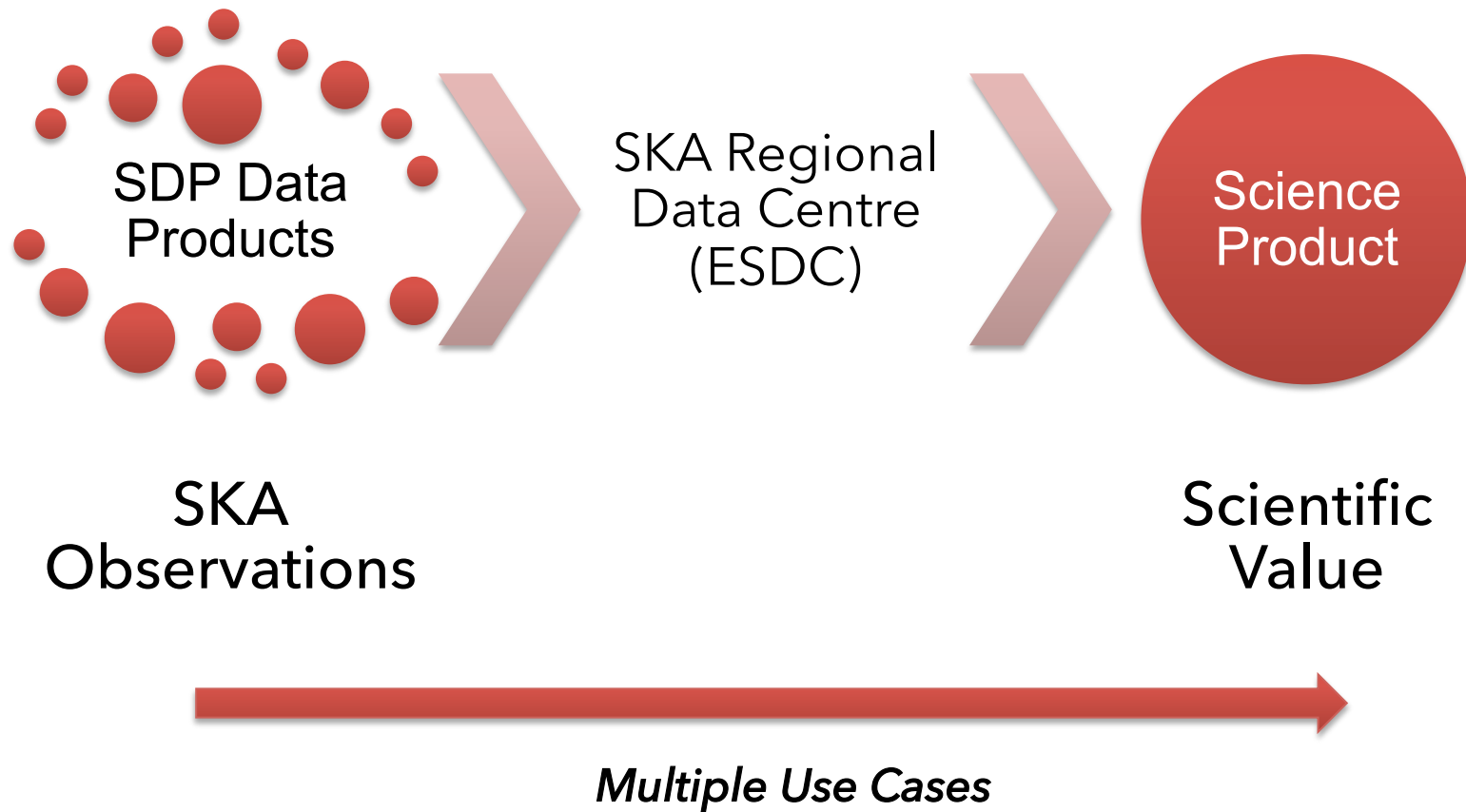
"What does the ESDC need to do to maximise European science delivery from the SKA?"

"How can we build such a science data centre and at what cost?"



OBJECTIVES

- Develop a set of **design recommendations** for the ESDC pertinent to (1) data handling strategy, (2) scientific functionality and (3) software environment.
- Produce a **high level architectural design** for the ESDC with a **sizing** and **costing** estimate.
- Provide **supporting verification work**, including both theoretical analyses and direct prototyping of critical elements.
- Identify **gaps**, highlight **risks** and make recommendations with respect to **mitigation**.

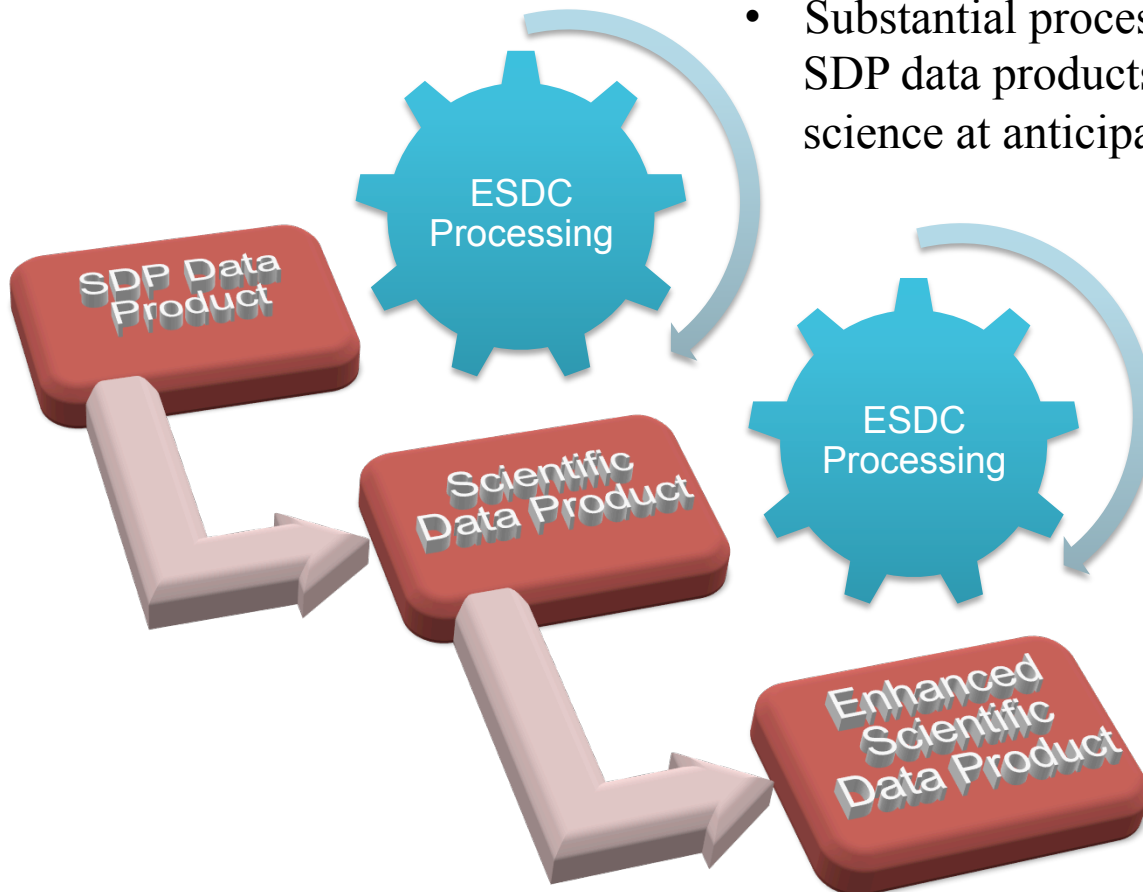


WORK BREAKDOWN

- T3.1** ESDC Processing: Inventory of SKA science cases and post-SDP computing requirements
- T3.2** ESDC Data storage: Inventory and sizing of SKA science data products and ESDC user-derived products
- T3.3** Evaluation of existing HPC, cloud and distributed computing technologies
- T3.4** Design and costing for distributed ESDC computing architecture
- T3.5** Requirements for interfaces to SKA Science Archives & Other Repositories
- T3.6** Validation, Verification & Proof of concept activities utilizing SKA pathfinder and pre-cursor facilities

T3.1 ESDC Processing

- Substantial processing and manipulation of SDP data products will be to deliver survey science at anticipated fidelity.



- Combining with data products from other observatories will be necessary to maximise scientific return.

T3.1 ESDC Processing

The SKA has a list of **13 High-Priority Science Objectives (HPSOs)**.

This task will **focus specifically on the delivery of these key experiments** and their compute processing requirements, to provide a basis for sizing and costing efforts.

Scientific return will be enhanced by combining SKA outputs with data from other observatories. Using results and insights from the ASTERICS program, WP3 will **make estimates for ESDC resources needed to deliver data combining** and maximize scientific return on the ESFRI astronomy projects.

Additional work: whether specific Science Use Cases (**open time programs**) could have significantly different compute requirements. WP3 will also consider the options for “**Discovery Products**” not covered by specific experiments, but piggy-backing on observing time.

T3.1 ESDC Processing

Partners: UCAM (lead), INAF, CSIC

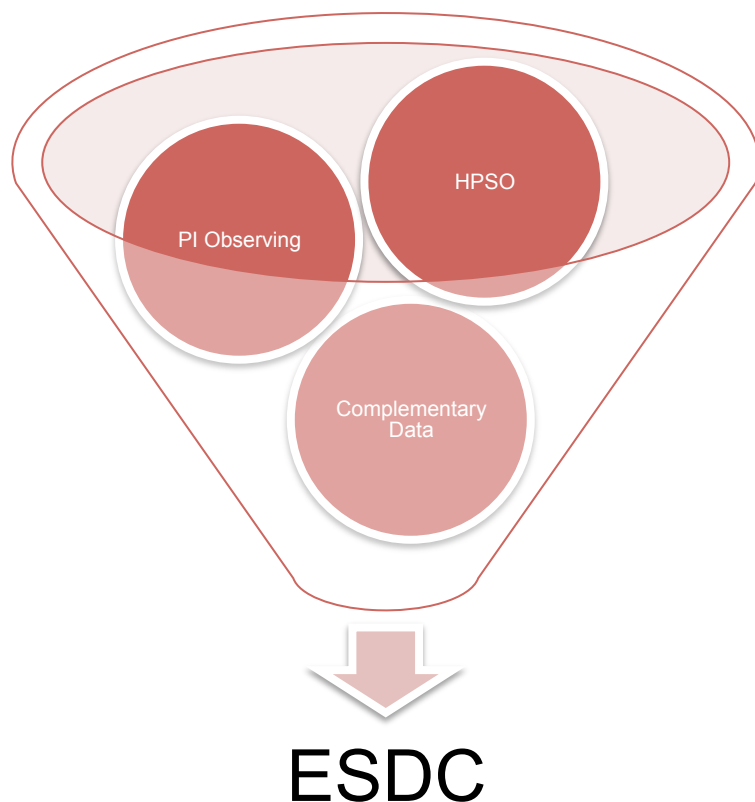
Stakeholders: VUW

Output: System sizing and functional requirements for Deliverable 3.1

Deliverable

D3.1 Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science **T0+21months**

T3.2 ESDC Data storage



- Volume of data arriving will depend on observing strategy. Need to be able to accommodate required data.
- Data access patterns will vary depending on processing needs. Need to be able to access data efficiently within ESDC.
- Final science products will need extraction.



T3.2 ESDC Data storage

SKA will provide a well-defined set of data products, but the total volume will depend greatly on the details of both the key science programs and open time programs undertaken, as well as the observing strategy for those programs determined by the observatory.

This task will **assess the data storage requirements** from a European perspective, considering the type, size and volume of SKA (SDP) and ESDC user-derived data products.

This task will also **analyse data access patterns** that might be expected for the ESDC.

T3.2 ESDC Storage

Partners: UCAM (lead), ASTRON, INAF, CSIC

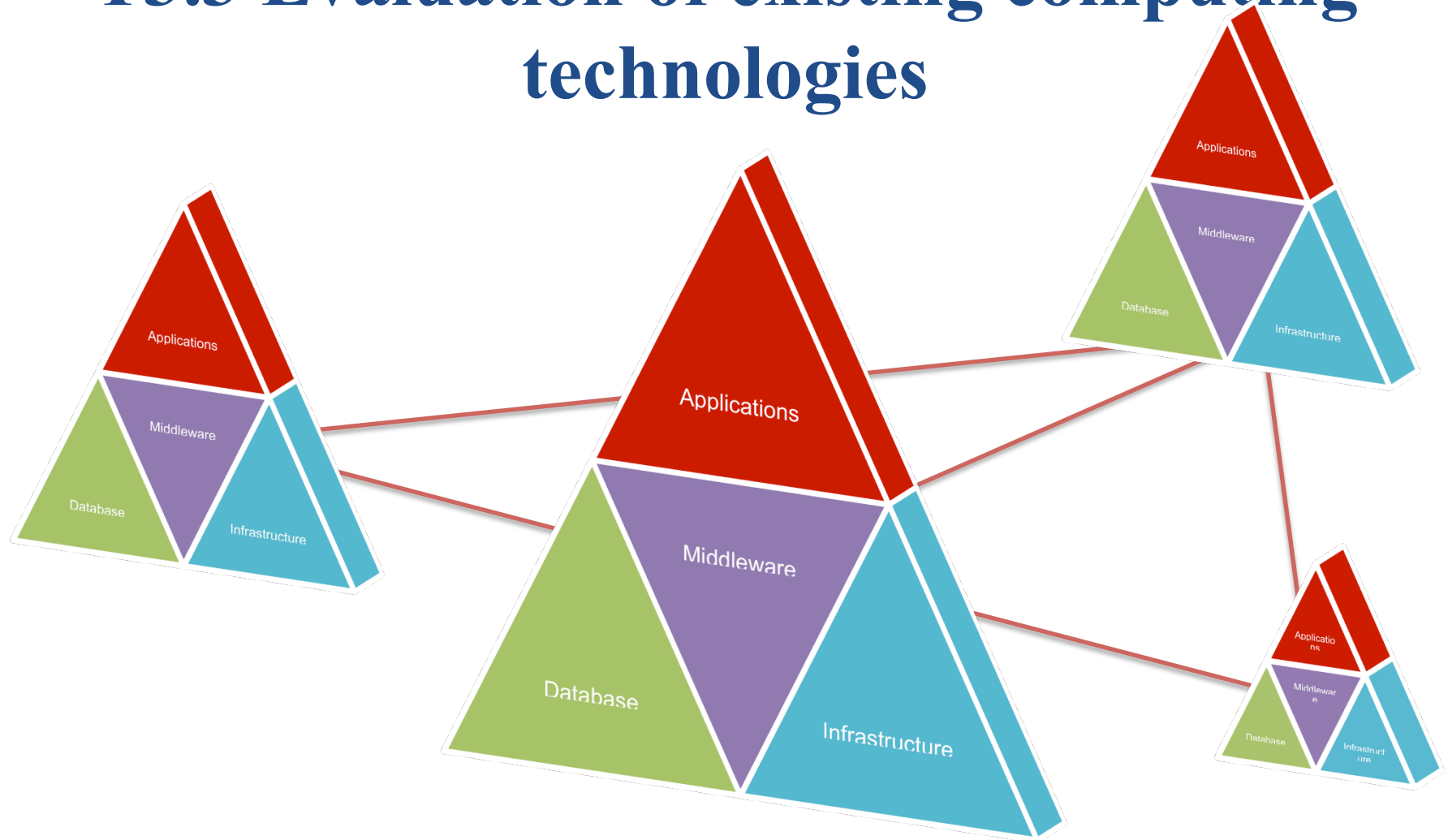
Stakeholders: VUW

Output: Data storage and functional requirements for Deliverable 3.1

Deliverable

D3.1 Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science **T0+21months**

T3.3 Evaluation of existing computing technologies



T3.3 Evaluation of existing computing technologies

Middleware capabilities will be assessed with the aim of **ensuring portability of data and applications** in the distributed environment. It will also include an analysis of available **replica management** and data transport organisation tools.

Additional services for federated compute such as **Authentication and Authorization Infrastructure (AAI)**, efficient movement of data based on policies; integration with HPC software stacks; **accounting elements** and ensuring proper and fair use of resources.

WP3 will look at the possibility of building on top of **industry-standard Big Data / data- science top level software stacks**.

T3.3 Evaluation of existing computing technologies

Partners: UCAM (lead), STFC, EGI, EPFL, ASTRON, Chalmers, CSIC, IT

Stakeholders: SNIC

Output: Report on suggested solutions for each key area including assessed list of options and recommendations for Deliverable 3.2.

Deliverable

D3.2 Report on suggested solutions to address each of the key software areas associated with running a distributed ESDC. **T0+24months**

T3.4 Design and costing for distributed ESDC computing architecture

Based on input from the evaluations in D3.1 and D3.2 this task will provide a **top-level architecture and functional design** for the ESDC.

To proceed we will make use of the inventory of national roadmaps from WP2 (D 2.1) and **determine the potential for incorporation / co-use of existing or planned facilities** to achieve economies of scale.

WP3 will develop a **costing of additional resources** needed (over and above existing facilities) to bring about a functioning ESDC, considering the full SKA observatory lifecycle from commissioning as the SKA is built and well into full operations as the SKA observatory develops and undergoes upgrade cycles.

T3.4 Design and costing for distributed ESDC computing architecture

Partners: UCAM (lead), STFC, EGI, EPFL, ASTRON, CSIC, IT

Stakeholders: SNIC

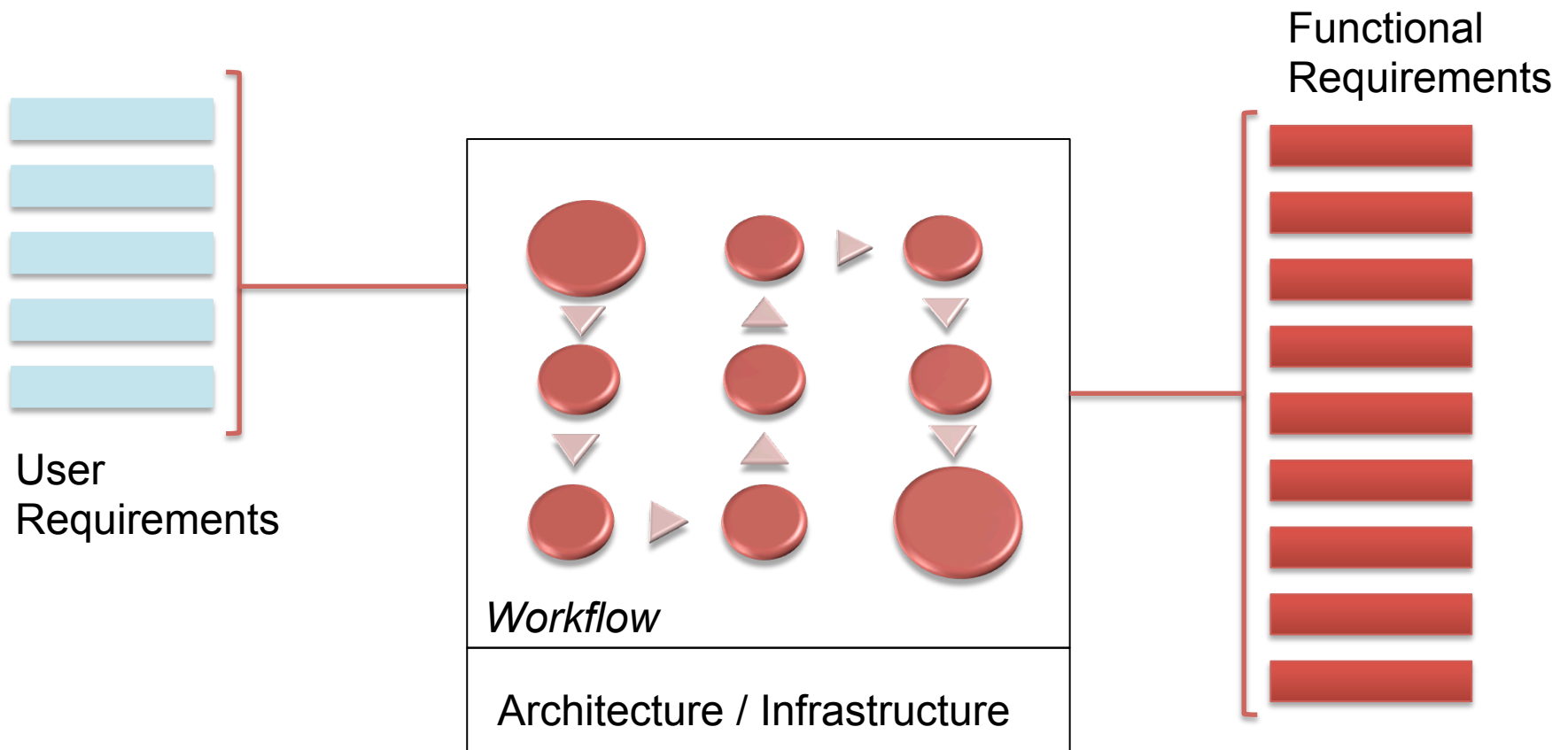
Output: A preliminary system sizing estimate for Deliverable 3.3 and a documented design for a ESDC model to appear in Deliverable 3.4.

Deliverable

D3.3 Preliminary System sizing report **T0+24**

D3.4 Report on design & costing for ESDC **T0+36**

T3.5 Requirements for interfaces to SKA Science Archives & Other Repositories



T3.5 Requirements for interfaces to SKA Science Archives & Other Repositories

This task performs **technical evaluation of interface requirements**, including:

- **evaluation of existing policies** for scientific data centre interactions
- **gap analysis** for SKA-specific needs
- **verified assessments of data ingest** from global sources including the SDP, other nodes within the ESDC, and external archives (e.g. LSST, EUCLID, JWST etc).

Recommendations for ensuring compatibility with **VO standards** and **minimum meta-data requirements** will be made.

This task will also inform **policy decisions governing the persistence of user work-flows** to enable reproducibility of results or regeneration of data.

T3.5 Requirements for interfaces to SKA Science Archives & Other Repositories

Partners: UMAN (lead), INAF, STFC, CSIC

Stakeholders: VUW

Output: Report on suggested solutions to interface requirements for a distributed ESDC as Deliverable 3.5.

Deliverable

D3.5 Report on suggested solutions to interface requirements for a distributed ESDC **T0+36**

T3.6 Validation, Verification & Proof of concept activities

The work in this task includes the

- **provision of a standardized set of appropriate test data.**
- **prototype software** blocks to verify and validate the functional requirements
- **pilot workflows to verify recommendations** on applicability of different middleware environments.

It will also provide verification activities:

- **required data access patterns**
- **user interface requirements can be mapped effectively to workflow models**
- **evaluating the ingest requirements for data transfers**

This task will provide technical work to **verify the scaling of critical elements for the system sizing.**

T3.6 Validation, Verification & Proof of concept activities

Partners: UMAN (lead), UCAM, Chalmers STFC

Stakeholders: VUW

Output: Inputs to multiple deliverables.

Deliverable

D3.3 Preliminary System sizing report **T0+24**

D3.4 Report on design & costing for ESDC **T0+36**

D3.5 Report on suggested solutions to interface requirements for a distributed ESDC **T0+36**

WP3 Summary

