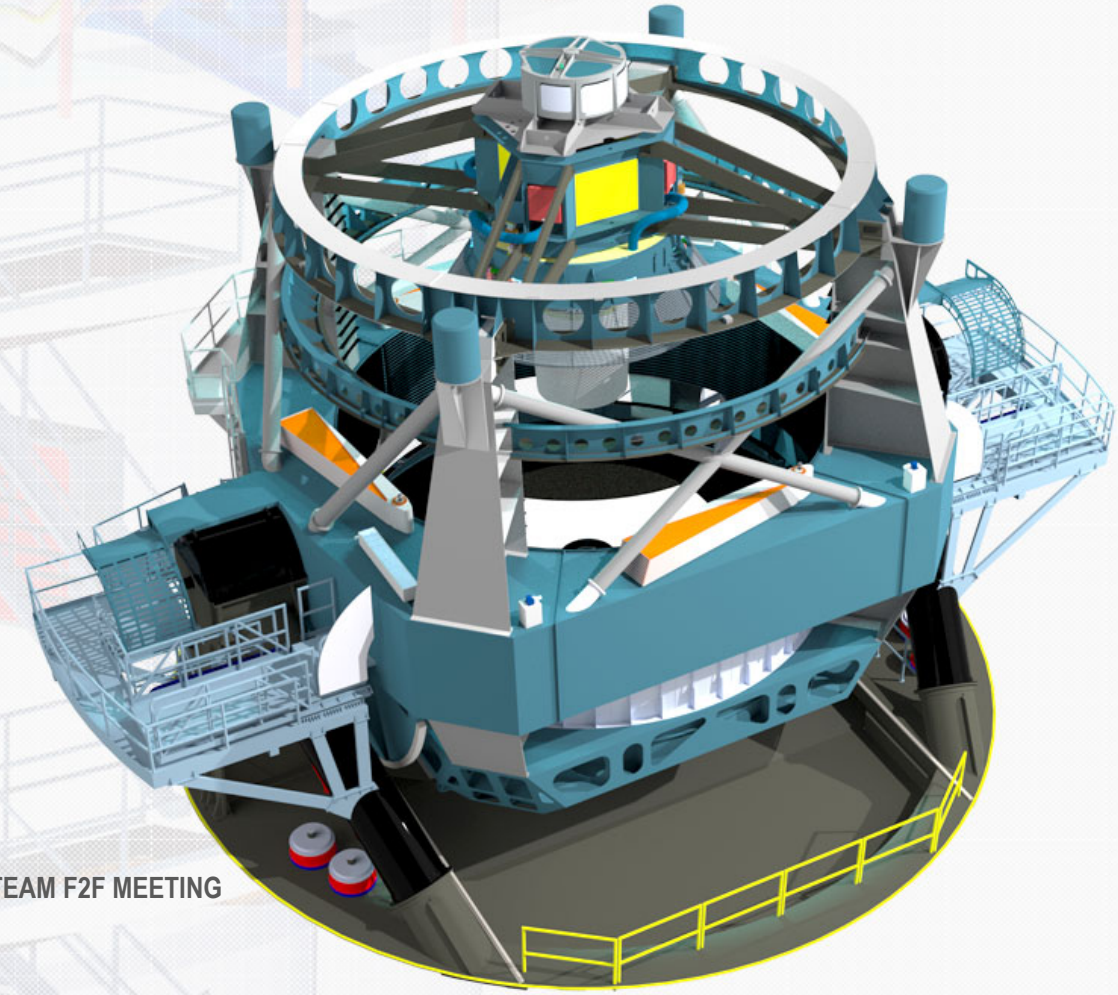
Replan Overview and Status

Mario Juric,
University of Washington



DATA MANAGEMENT LEADERSHIP TEAM F2F MEETING
January 10-12th, 2017

Note

- Note: this presentation will be as much about *collection* of information as it will be about presenting what's known.

Replan Goals

1. **Improve the design baseline and concepts of operation to enable work package estimation at the required level of accuracy**
   - This includes developing the concepts of operation for the data processing system, and updates to the definition of deliverables (data products and services, LSE-163).
   - Also includes adding a significant level of detail to the design documents for the science pipelines, middleware, the science user interface, and infrastructure.

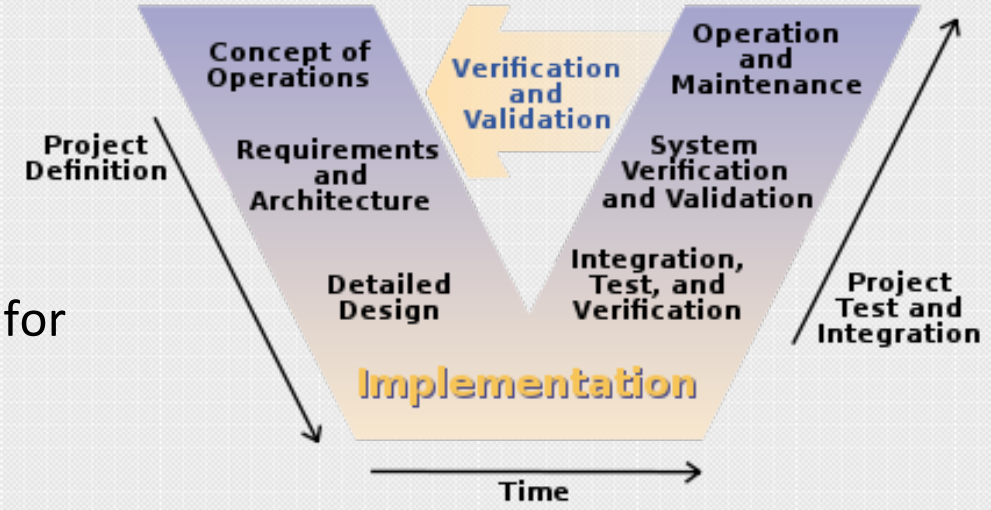
2. **Improve the project management structure and processes for planning, tracking, and earned value accounting**
   - Ensure all planned deliverables are clearly defined and valued appropriately.
   - Ensure the long-term estimation methodology incorporates best practices gathered in Constr. Year 1.
   - Jacek Becla's presentation will address the current status and efforts in this area

3. **Ensure we have a clear, fully resourced plan, consistent with DM team performance, budget, and schedule**
   - Incorporate lessons learned from the 1st year of construction, and ensure we're confident that all DM-related scope is accounted for, planned for, and adequately resourced. Make early recommendations on contingency draw or scope adjustments, if needed.

4. **Focus the subsystem culture towards deliverables and accountability**
   - Reorganize the subsystem to clearly define areas of authority and lines of responsibility
   - Work against bottom-up scope creep and non-essential development

Replan Strategy

- Identify and write missing concepts of operations
- Update the requirements
- Update/define the system architecture
- Update the designs to a level need for long-term planning and estimation



- Develop implementation plans
- Update processes for agile execution of plans
- Define the verification and validation strategy
- (post-replan: continuously maintain and adjust the resulting body of work)

Challenges: the need for parallel development in many of these areas; desire to minimize impact on ongoing construction.

Concepts of Operations

- **Top-level documents:**
  - **Operations plan** (LPM-181; first draft November 2016, not CCB approved)
    - DM contributions by Petravick, Johnson, Ivezic, Connolly, Lim, et al.
  - **Commissioning plan** (current draft January 2016, not CCB approved)
    - DM contributions by Lim, Ivezic, Lupton, Connolly, et al.
  - **Data Products Definition Document** (updated and CCB approved in September 2016)
    - Result of the scipi-wg work (co-chaired by Ivezic & Lupton)

- **DM-level ConOps / Vision documents:**
  - Data Processing System ConOps and Architecture
    - Will result in an update to LDM-230 (?, KTL & DP?)
  - SUIT vision documents
    - Should it be incorporated into an overall "services vision document"
  - Overall system vision (PowerPoint)
    - Needs to be written, approved (MJ?)

LSST Mission Goals and System Vision

The LSST will be a facility whose primary mission is to acquire, process, and make available to the data-rights holders the data collected by its telescope and camera. Our primary products are the stream of events alerts (Level 1) and Data Release data products (Level 2).

To make those products available and useful to the community, we're building **Data Access Centers**. These will expose the LSST data to the data rights holders through a number of data access center services.

LSST Portal: The Web Window into the LSST Dataset

The Web Portal to the archive will enable browsing and visualization of the available datasets in ways the users are accustomed to at archives such as IRSA, MAST, or the SDSS archive, with an added level of interactivity.

Through the Portal, the users will be able to view the LSST images, request subsets of data (via simple forms or SQL queries), construct simple plots, and generally explore the LSST dataset in a way that allows them to identify and access (subsets of) data required by their science case.

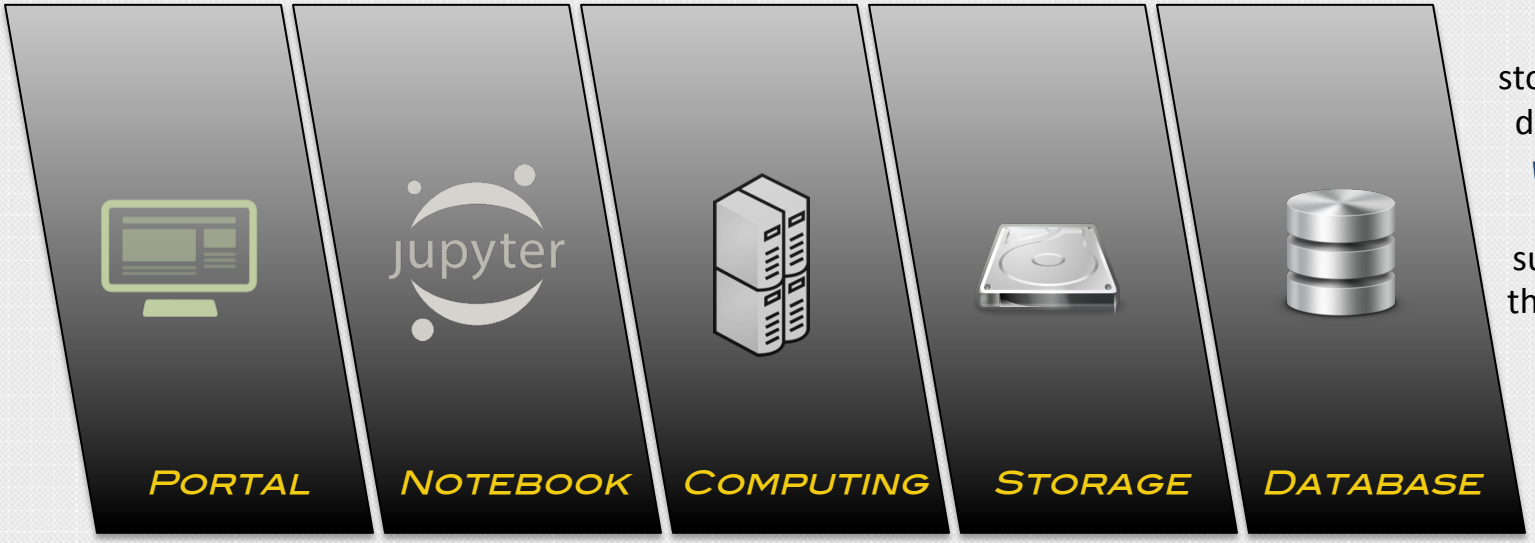
Next-to-the-data Analysis: Jupyter Notebooks

The tools exposed through the Web Portal will permit simple exploration, subsetting, and visualization LSST data. They may not, however, be suitable for more complex data selection or analysis tasks.

To enable that next level of next-to-the-data work, we plan to enable the users to launch their own Jupyter notebooks at our computing resources at the DAC. These will have fast access to the LSST database and files. They will come with commonly used and useful tools preinstalled (e.g., AstroPy, LSST data processing software stack).

This service is similar in nature to efforts such as SciServer at JHU, or the JupyterHub deployment for DES at NCSA.

Computing, Storage, and Database Resources



Computing, file storage, and personal databases (the *"user workspace"*) will be made available to support the work via the Portal and within the Notebooks.

An important feature is that no matter how the user accesses the DAC (Portal, Notebook, or VO APIs) they always "see" the same workspace.

How big is the "LSST Science Cloud" (@ DR2)?

- **Computing:**
  - ~2,400 cores
  - ~18 TFLOPs
- **File storage:**
  - ~4 PB
- **Database storage**
  - ~3 PB

This is shared by all users. We're estimating the number of potential DAC users to be in the low 1000s (relevant for file and database storage).

Not all users will be accessing the computing cluster concurrently. A reasonable guess may be in 10-100 range.

Though this is a relatively small cluster by 2020-era standards, it will be **sufficient to enable preliminary end-user science analyses** (working on catalogs, smaller number of images) and creation of some added-value (Level 3) data products.

Think of this as having your own server with a few TB of disk and database storage, right next to the LSST data, with a chance to use tens to hundreds of cores for analysis.

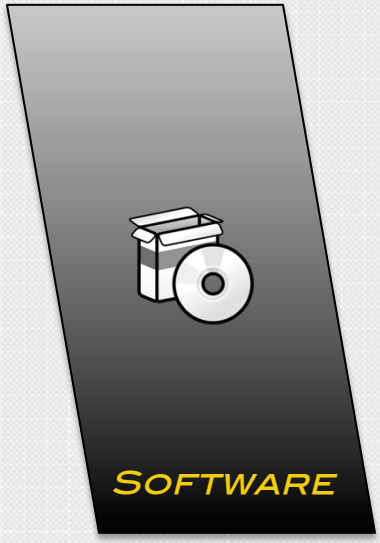
For larger endeavors (e.g., pixel-level reprocessing of the entire LSST dataset), the users will want to use resources beyond the LSST DAC (more later).

Open Sourcing the Software: Reproducibility and Algorithmic Insight

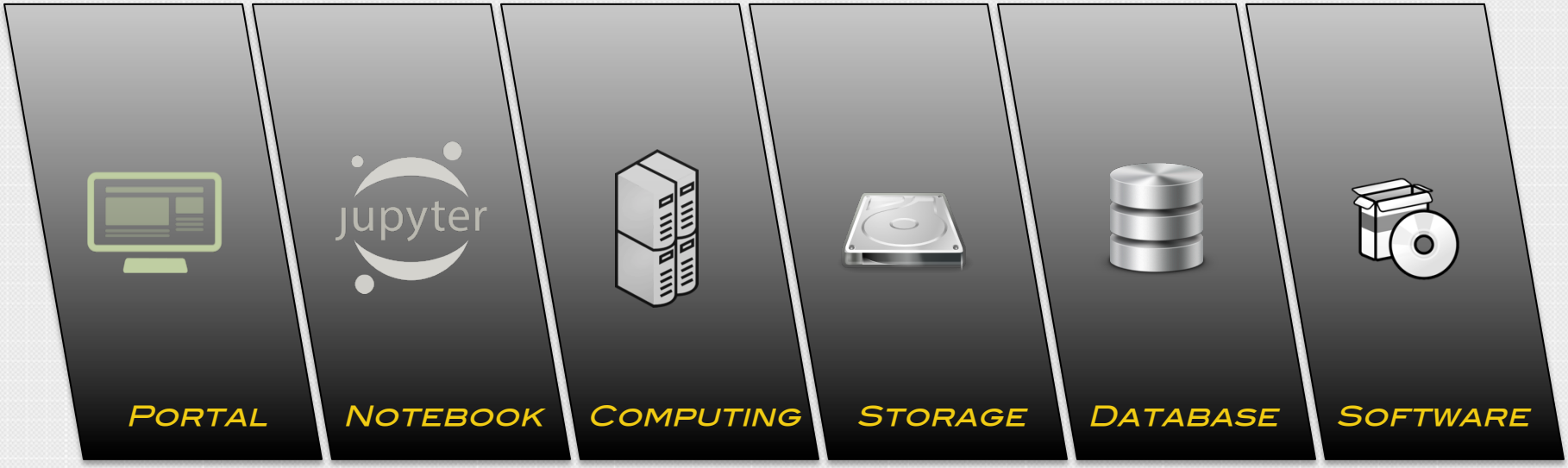
As the final “piece of the puzzle”, we’re also be making available the source code of the LSST data processing software (and configurations used in processing).

This will enhance reproducibility of the LSST data products, as well as provide source-code level of insight into algorithms utilized by LSST data processing.

Having the source code may also enable community efforts extend and apply the LSST codes to projects beyond the LSST. Some efforts, such as processing of HSC Survey data (Miyazaki et al.) or of CFHT-LS (Boutigny et al.), are already under way.



Putting it all together: the LSST Science Platform



How we (think) we will work with LSST data?

- Most users are likely to begin with the Web Portal, to become familiar with the LSST data set and query smaller subsets of data for "at home" analysis. Some may use the tools they're accustomed to (e.g., TOPCAT, Aladin, AstroPy, etc.) to grab the data using LSST's VO-compatible APIs.

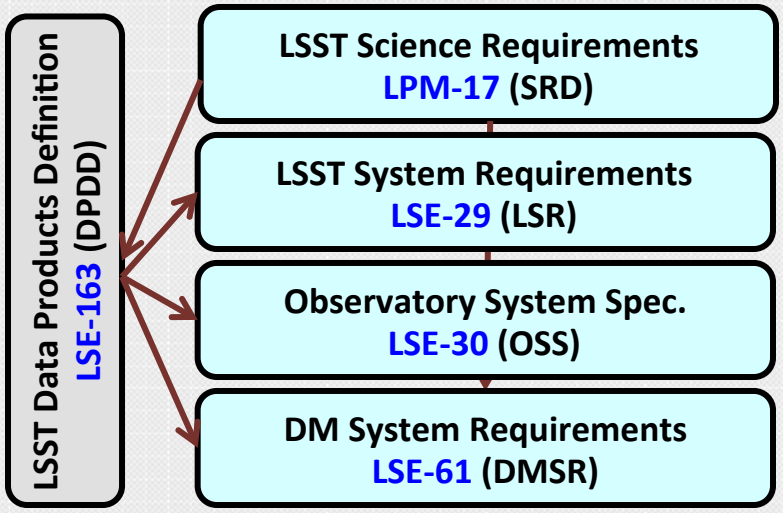
- Some users may choose to continue their analysis by utilizing resources available to them at the DAC. They'll access these through Jupyter notebook-type remote interfaces, with access to a mid-sized computing cluster. **It's quite possible that a large fraction of end-user ("single PI") science may be achievable this way**.

- For users who need larger resources, they may be able to apply for more resources at adjacent computing facilities. For example, U.S. computing is located in the National Petascale Computing Facility at the National Center for Supercomputing Applications (NCSA). Significant additional supercomputing is expected to be available at the same site (e.g., NPCF currently hosts the Blue Waters supercomputer).

- Finally, rights-holders may utilize their own computing facilities to support larger-scale processing. As they're open source, they may re-use our software (pipelines, middleware, databases) to the extent possible.

Vision and ConOps' -> Requirements

- Status:
  - Major DMSR update under way
    - Consistency with DPDD/OSS/LSR
    - Updates to OSS and LSR likely to be needed
  - SUIT requirements
    - Delivered in July
    - Need to incorporate in DMSR
  - ICDs are another source of requirements.



- Verification:
  - Verification matrices being constructed with guidance and oversight by Systems Engineering

Design Documents

- **LDM-148: The DM System Design Document**
  - Needs to be updated with the information from overall system vision
  - Should the definition of a minimum viable system reside here (more later)
- **LDM-151: Science Pipelines Design Document**
  - Major update, nearly completed (Ivezic & Lupton WG)
- **LDM-135: Database design**
  - In good shape overall, may need updates later
- **LDM-230: Automated Operations Design**
  - Needs updating using the series of documents developed by NCSA as a basis
- **LDM-129: Infrastructure Design Document**
  - Needs to be updated to reflect the current thinking
  - System architecture model in EA
- **LDM-152: Middleware design**
  - Lim & Petravic WG (status?)
  - Butler, Task Framework, Orchestration system design, etc.?
- **LDM-131: SUIT (Conceptual) Design Document**
  - Needs update to make it consistent with SUIT vision, requirements, and plans

Developing the long-term plan

Constructing LSST DM: Strategy

1. **Define a "minimum viable system"**
   - This is not the same as satisfying minimum requirements; this is preventing "catastrophic" failure (i.e., inability to get an NCR approved)

2. **Phase requirements accordingly**
   - In other words, prioritize features.
   - Part of requirements document update work.

3. **Deploy the minimum viable system (DAC, L1, DRP, L3) as early as possible**
   - **First prototype: the Prototype DAC (December 2016; more later)**
   - Target for MVS: ~mid 2018, then continue incrementally integrating feature improvements through 2022 (details of timing TBD)
   - Will allows us to enter "soft" failure mode, the point where data collection (operations) can start even if the DM system does not meet all requirements
   - Enables practicing of ops and commissioning processes, enables us to think about serving the community in commissioning

4. **Sequence development so the system being built can support commissioning**
   - Linking our milestones to commissioning task force milestones

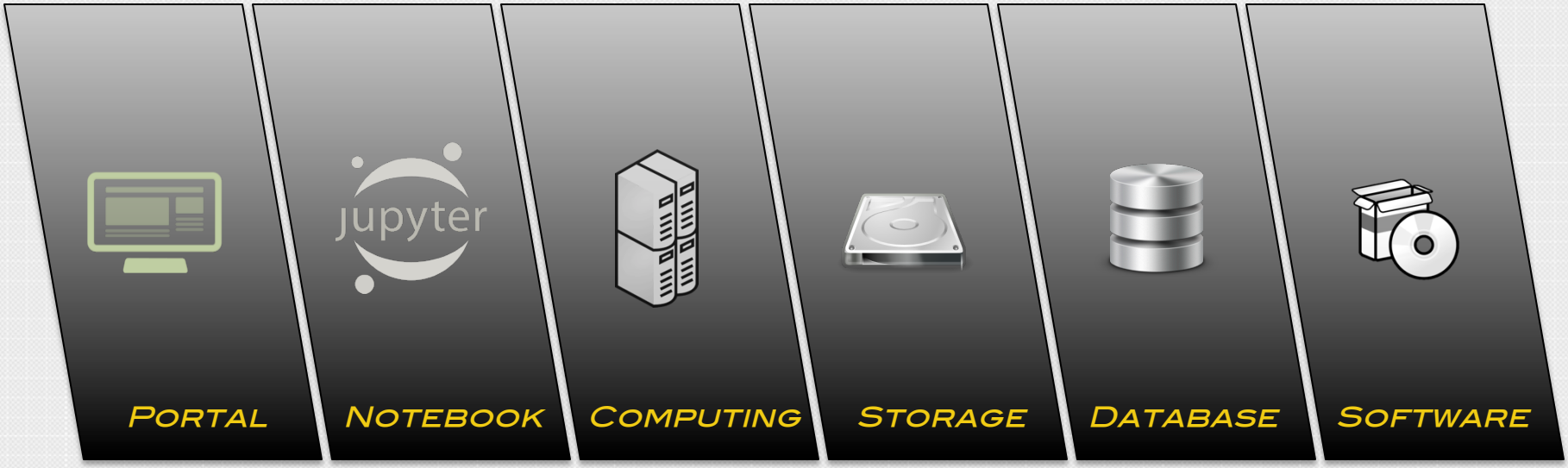
Plan Development Guidelines

- **General principles**
  - **Build what's needed to support commissioning**
  - **Try to keep pipelines on the critical path:** the pipelines should be pacing the overall development; w/o them there's no scientific deliverable, and are a unique (and risky) product we cannot get off-the-shelf.
  - **Incremental I&T:** work towards a minimum functioning system along all axes (L1, L2, all DAC services)

- Directions given to managers: **Reorient development to support getting a minimal DAC, Level 1 and Level 2 system running by mid 2018 (the entire system, not just the pipelines; the complexity is also in I&T and processes)**
  - SUIT development to support pipeline QA over the short term (~year)
  - SQuaRE development to support QC tools and deliver initial Jupyter images (~within a year)
  - DB development to produce minimum usable L3 database support early (~within a year)
  - Middleware development to support DRP & Level 1 execution
  - Infrastructure development to support PDAC deployments and batch system for L2/L1

- The above changes will align DM development groups better, with externally visible progress (new features and processed datasets accessible through our DAC) occurring on regular intervals (every ~6 months).

The End Goal: “LSST Science Platform” (2021)

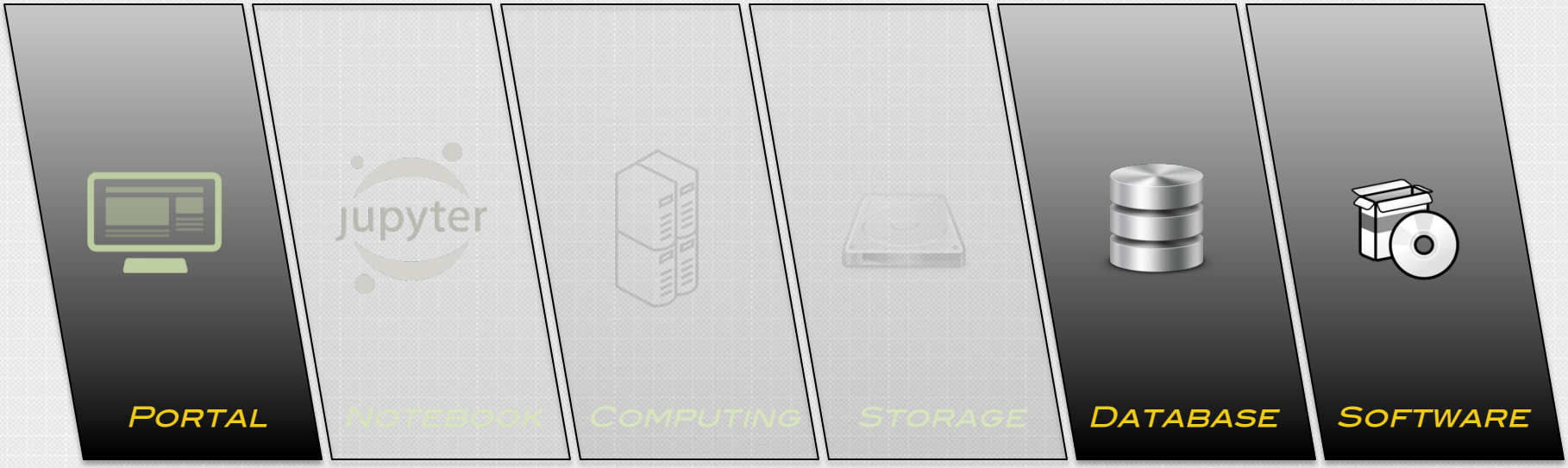


Integrating the Data Center: December 2016

The focus through the end of this development cycle is on standing up and integrating the Prototype Data Access Center at NCSA.

This effort will mate NCSA infrastructure, SLAC's qserv database, and IPAC's Science User Interface portal. We will load LSST-reprocessed SDSS Stripe 82 data. This will be opened for testing to a selected subset of the community.

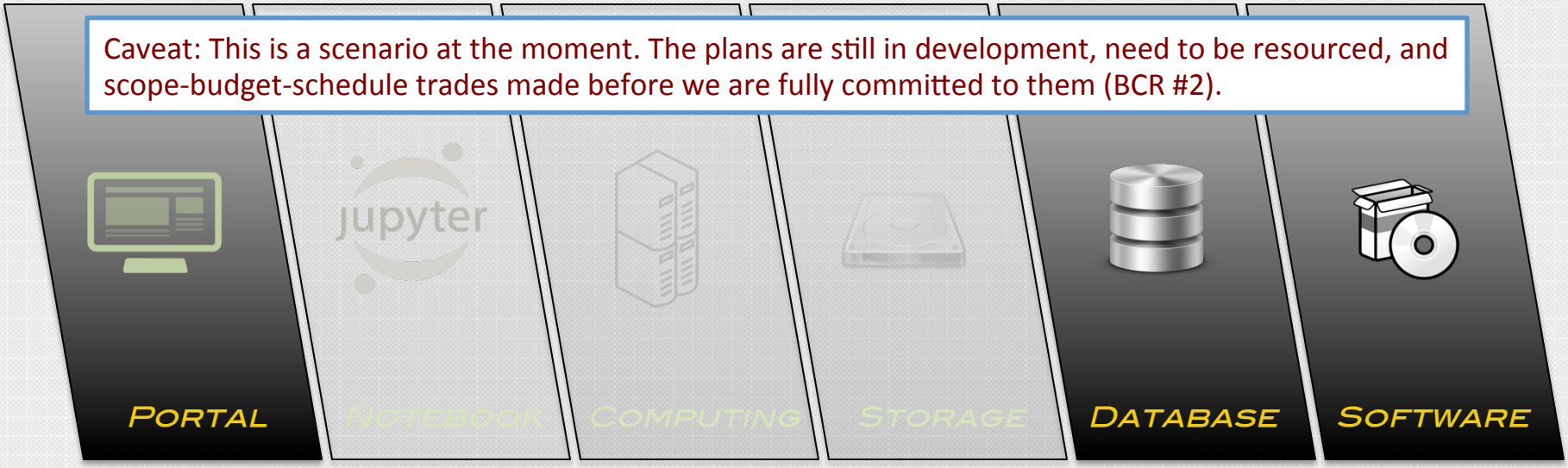
This is the beginning of I&T of user-visible DM services.



Integrating the Data Center: June 2017

For the following 6 months (Dec 2016-May 2017), we plan to add the WISE/NEOWISE catalog to the PDAC. Exercises the scale and time-domain capabilities, ability for multi-catalog science. Also deploying infrastructure to perform regular data release processing at scale

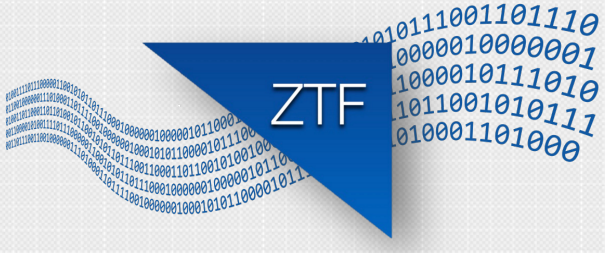
Science pipelines will deliver initial Level 2 and Level 1 end-to-end system prototypes (incl. alert production). Key features: deblending, multifit, DCR-templates.



Integrating the Data Center: December 2017

Initial PDAC integration in all feature areas

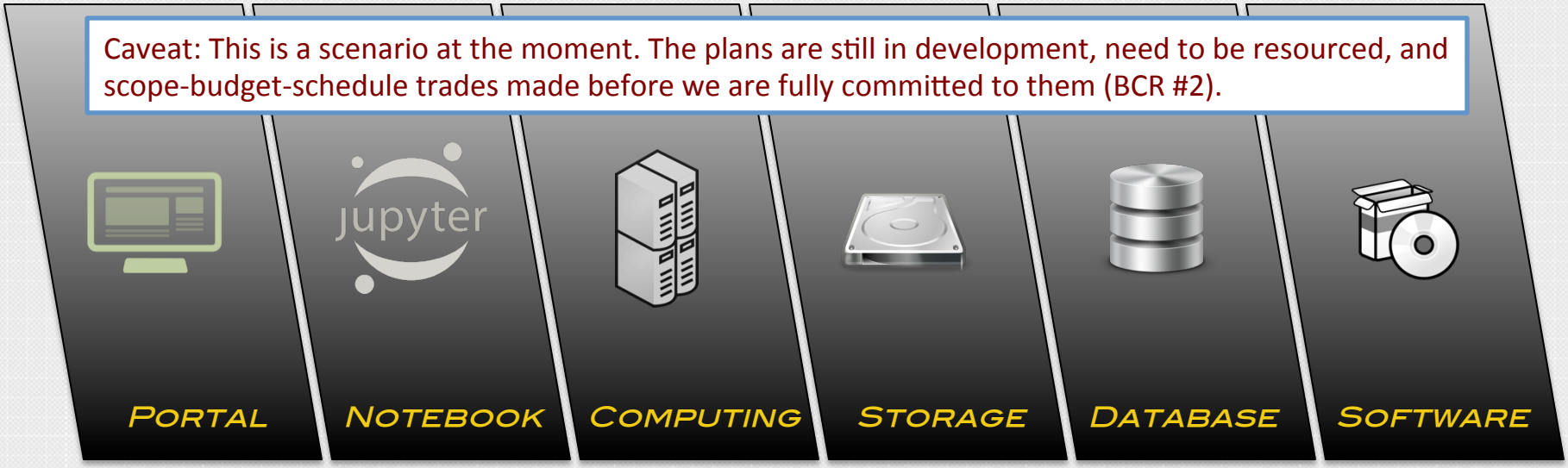
- Deploy initial Notebook capability
- Deploy initial PDAC computing
- Deploy initial PDAC storage capability
- Deploy initial "user tables" capability
- Established service processing



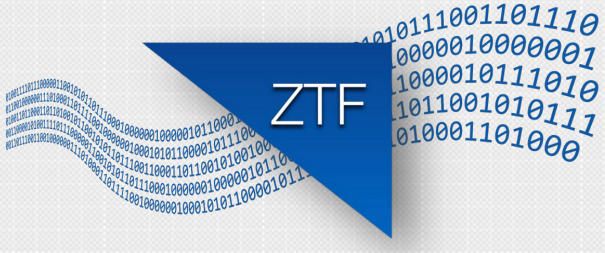
Science pipelines and middleware

- Run a DRP using HSC-Survey data
- Generate alerts off of ZTF commissioning data

Caveat: This is a scenario at the moment. The plans are still in development, need to be resourced, and scope-budget-schedule trades made before we are fully committed to them (BCR #2).

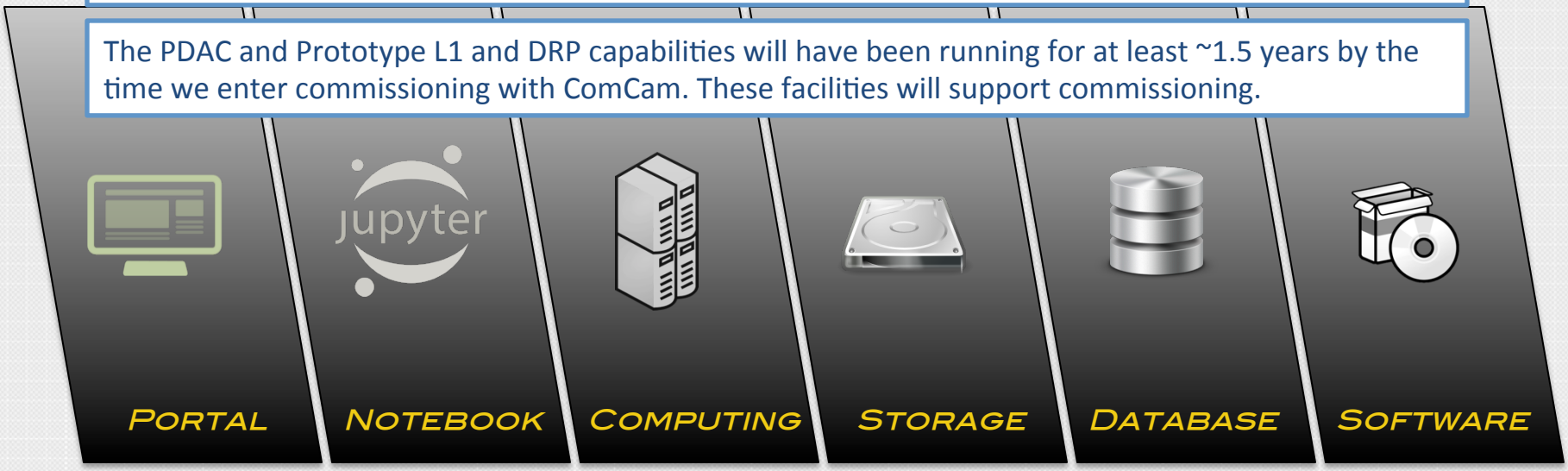


Integrating the Data Center: September 2019



Caveat: This is a scenario at the moment. The plans are still in development, need to be resourced, and scope-budget-schedule trades made before we are fully committed to them (BCR #2).

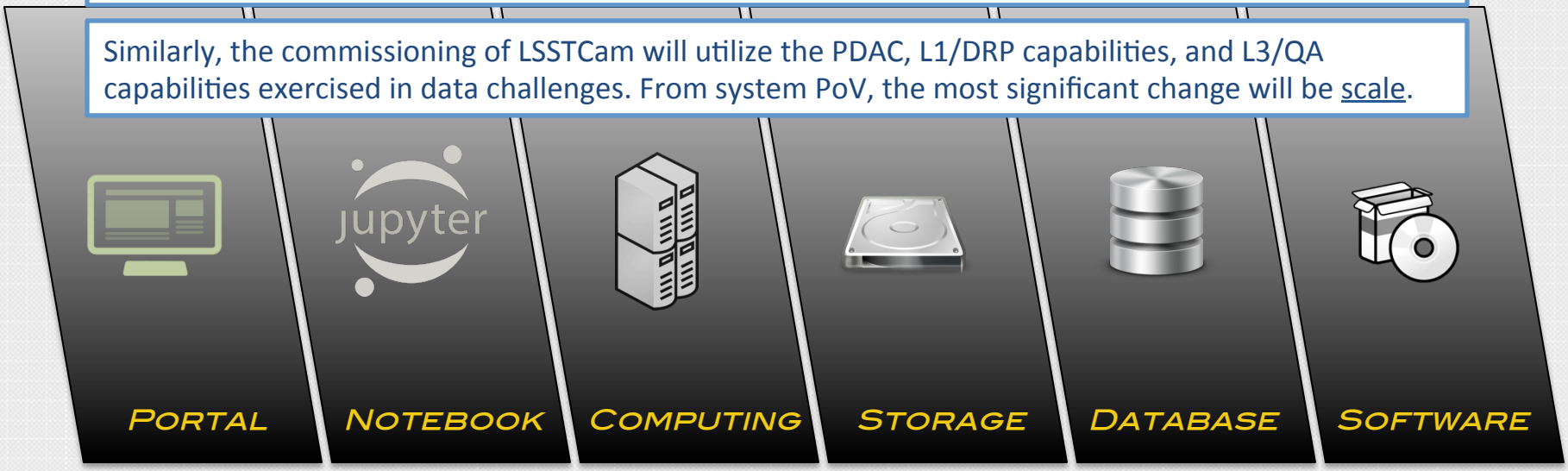
The PDAC and Prototype L1 and DRP capabilities will have been running for at least ~1.5 years by the time we enter commissioning with ComCam. These facilities will support commissioning.



Integrating the Data Center: October 2020

Caveat: This is a scenario at the moment. The plans are still in development, need to be resourced, and scope-budget-schedule trades made before we are fully committed to them (BCR #2).

Similarly, the commissioning of LSSTCam will utilize the PDAC, L1/DRP capabilities, and L3/QA capabilities exercised in data challenges. From system PoV, the most significant change will be scale.



Plan Development

- The teams, led by the TCAMs, have been developing the first version of the plan following the guidelines given (see preceding slides).

- They've attempted to satisfy the timeline defined by operations, commissioning, and construction-time milestones
  - n.b.: we haven't yet defined in detail the "minimum viable system"; for the first iteration, we wanted each group to show through prioritization what they feel the minimum in their area is.

- If faced with budget/schedule overrun, the instruction was to allow the plan to run over into commissioning/operations.
  - Of course, we will "pull things back" through value engineering and descopes once we understand the overall plan

- For subsequent iterations, we will consider descopes (including architecture changes) and firmly define the minimum viable system given the architecture that emerges.

Current status

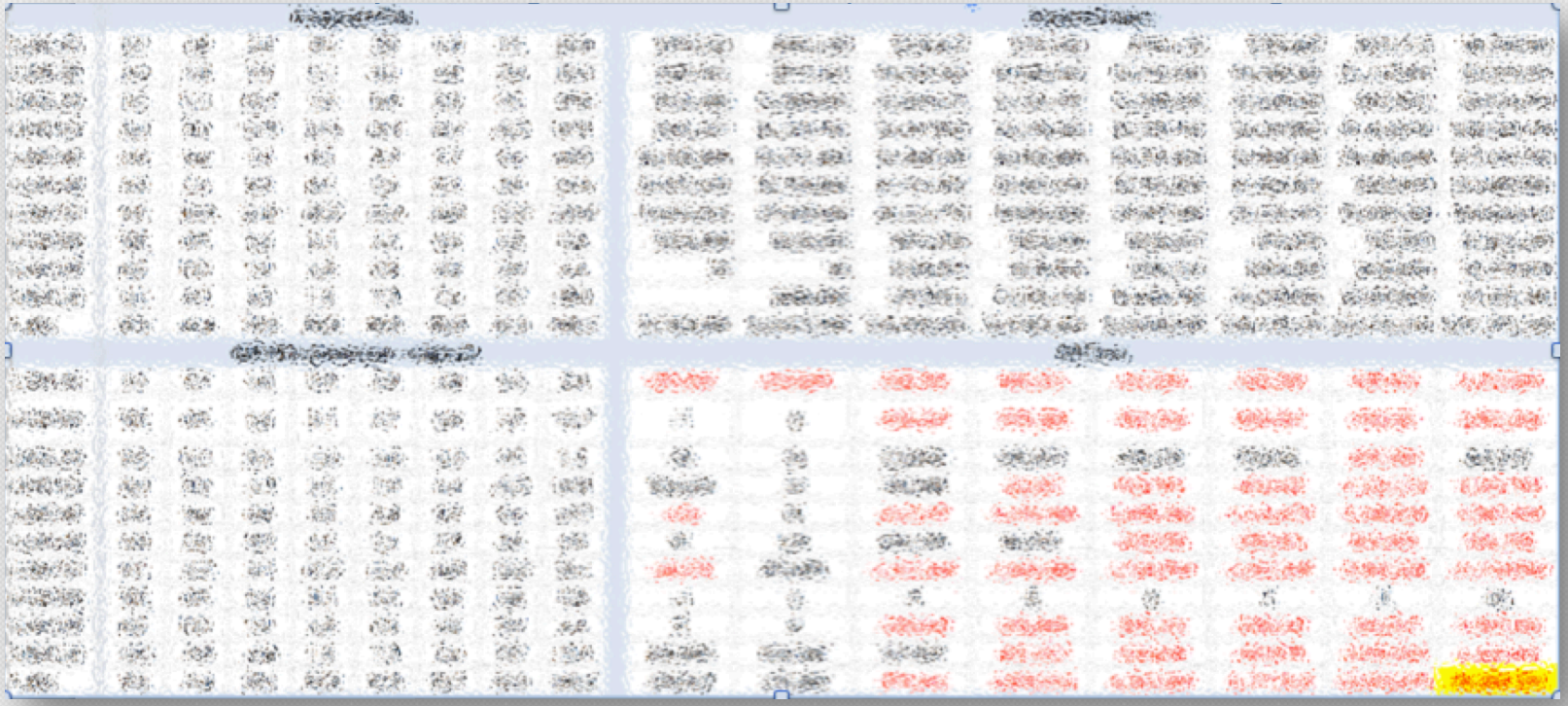
- **Software payloads**
  - Each group has developed the plans for construction of their deliverables. The planning packages are in spreadsheets at https://confluence.lsstcorp.org/x/XIzYAg
  - Exposed the needs for functionality to be delivered from other groups in the "DM L3 Milestones" spreadsheet https://docs.google.com/spreadsheets/d/1Z0Atx7Wn5BlbmFhtLjl9qahSVcLfGXwhvq59is8rZvE/edit#gid=0
  - TBD:
    - Link deliveries to commissioning milestones
    - Make explicit the link to milestones from the guidelines (i.e., the construction era goals)
    - Expose milestones for deliveries of major functionality (i.e., "multifit delivered")
      - DM Project Science needs to define these
    - Link L3 milestones to activities; we can't yet integrate the plan w/o this linkage
      - Some teams have it in internal versions (Princeton, NCSA, others?)
- **Data center**
  - Developed work packages tracing back from operational and commissioning business needs; work packages in OmniPlan
  - Not linked to construction-era business needs; therefore not fully scheduled yet.
  - TBD:
    - Link to software payload milestones
    - Schedule the work packages

Integrated Milestone Chart

OmniPlan file and PDF attached to the meeting confluence page

Cost projections

- Costing currently in spreadsheet that Victor has access to.
- Last updated mid-December; needs further update
- Currently showing ~$31M shortfall (but does not include funding for commissioning).



What could to be done next (proposal)

- Collect and finalize document updates
  - Execute DM-level change requests wherever possible
  - Open an LSST-level LCR for the coming change request

- Move the information from various spreadsheets and JIRA into a project management tool
  - Moving back/forth between spreadsheets and a PMCS tool is labor intensive
  - Project standard is P6, but OmniPlan or MS Project may work better in this phase
- Link software milestones and activities
- Link activities to commissioning and early operations milestones
- Reconcile data center and payload development plans
- Re-cost

- Begin considering value engineering options and possible descopes