Propuesta de un IDAC para Rubin en Argentina. talk @ JACA 2021

Mariano Domínguez @ IATE-OAC-UNC

CONICET

UNC Universidad Nacional de Córdoba

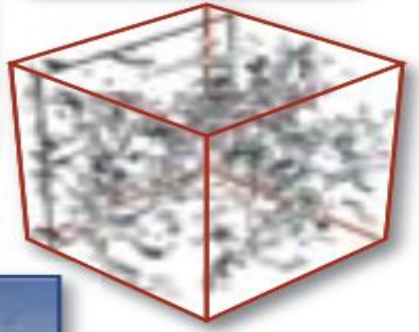
I A T E

Science Paradigms

IAU Astroinformatics and Astrostatistics

- Thousand years ago: science was **empirical**
  describing natural phenomena
- Last few hundred years: **theoretical** branch
  using models, generalizations
- Last few decades: a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

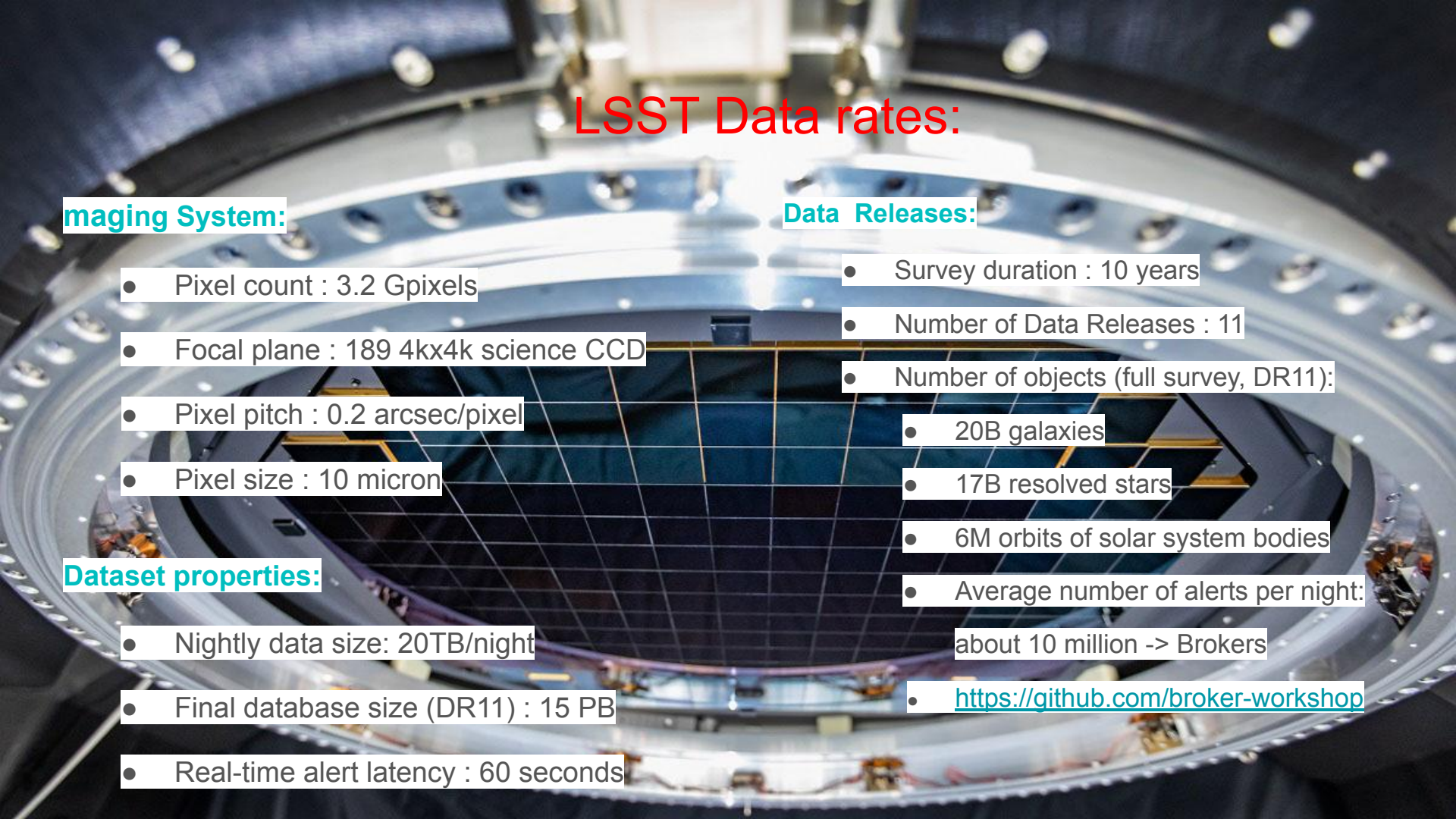
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



LSST Argentina Group: involves 65 researchers

3 PIs DESC 2+1 PIs Gal 2 PIs SMWLV

1 PI AGNs, 1PI TVS, 1PI Inf 1PI SS 1PI SL



LSST Data rates:

maging System:

- Pixel count : 3.2 Gpixels
- Focal plane : 189 4kx4k science CCD
- Pixel pitch : 0.2 arcsec/pixel
- Pixel size : 10 micron

Dataset properties:

- Nightly data size: 20TB/night
- Final database size (DR11) : 15 PB
- Real-time alert latency : 60 seconds

Data Releases:

- Survey duration : 10 years
- Number of Data Releases : 11
- Number of objects (full survey, DR11):
  - 20B galaxies
  - 17B resolved stars
  - 6M orbits of solar system bodies
- Average number of alerts per night:

  about 10 million -> Brokers
- https://github.com/broker-workshop

Prompt data products (for transients, variables, and moving objects):

- Alert packets for sources detected in difference images (difference = direct-template), delivered to brokers (60 seconds).
- Catalogs of sources (associated by location into "objects") detected in difference images (24 hours).
- Catalogs of moving objects with orbital parameters from the MPC (a result of daytime processing).
- Direct and difference images (24 hours).

Data Releases data products (for static sky sources like stars and galaxies):

- A global and uniform processing of all the data taken from the start of the survey until a given date (typically 6 months before the release).
- Raw and calibrated images, the calibration data, and deeply stacked coadded images.
- Source and object catalogs with measurements (centroids, fluxes, magnitudes, shape and size parameters, and more).
- Includes re-processed Prompt data products and direct image light curves.

For more information please see "The Rubin Data Products, Abridged" or the full Rubin "Data Products Definitions Document".

How can I access and process Rubin Observatory data and simulations?

Data

The whole Rubin Observatory data for the LSST (raw and calibrated images, yearly data releases) will be archived in its entirety in two facilities on different continents: NCSA in the US and CC-IN2P3 in France. Users with data rights will access the survey releases and data products through tools provided by the Project, from a Data Access Center (DAC). Currently planned DACs are at NCSA, at CC-IN2P3, and in Chile; additional DACs are under consideration in other participating countries.

Database queries for catalogs

Many users will only desire the output catalogues provided in each data release.

DACs will provide access via QServ, which allows efficient querying of the petabyte-size catalogs.

http://slac.stanford.edu/exp/lsst/qserv/

https://ieeexplore.ieee.org/document/6114487/

S3.1 TITLE: Serving LSST Catalogs from the IATE Lite IDAC

S3.2.1 Background: Description

IATE is a founder member of the UNC supercomputing centre (CCAD: https://ccad.unc.edu.ar/) where its computing hardware is hosted and also where many of our computing staff works. This is a modern, fast and efficient supercomputer centre. CCAD has provided computing resources for IATE astronomers for over a decade, hosting the data, developing simulations and supporting its analysis.

S3.3 PLANNED ACTIVITIES

S3.3.1 Activity: Description

If selected, IATE would secure funding to purchase additional hardware necessary to store the LSST data. Since this would involve only catalogue data, we believe that we would need only approx. 200 Terabytes of storage. We also believe that LSST users will need to do some basic data reduction and correlation with other data sets, and we intend to provide additional storage 400TB and 128 + CPU cores, guaranteed for LSST researchers' use.

we do intend to attempt to run and support the LSST Science Platform, while fully conforming to all other Guidelines for Rubin Independent Data Access Centers as specified in the https://rtn-003.lsst.io/ .

Workspaces Applications Nov 29 9:53 AM

data.lsst.cloud

VERA C. RUBIN OBSERVATORY

Portal Notebooks APIs Documentation Support Community Log in

RSP not necessary available, i.e Dask or Spark same as LINCC.

Rubin Science Platform

Portal

Discover data in the browser

Learn more about the portal.

Notebooks

Process and analyze LSST data with Jupyter notebooks in the cloud

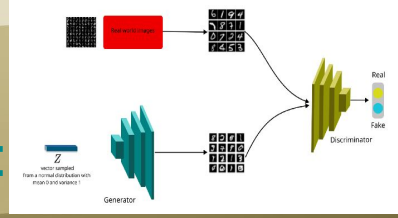
Learn more about notebooks.

APIs

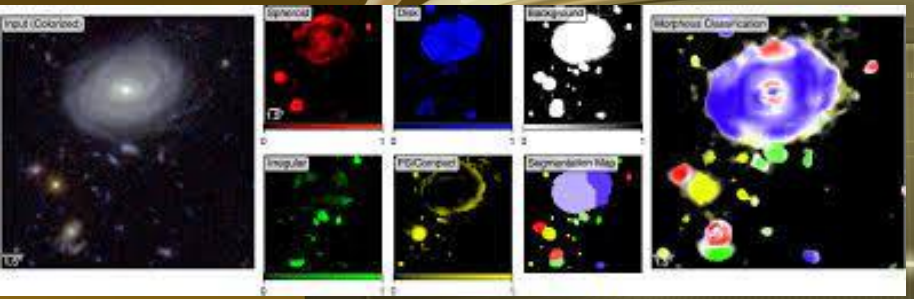
Learn how to programatically access data with Virtual Observatory interfaces

camels_fig_high....png images (2).jpeg BernardiniEtAl-2....jpg GAN-DM.png FFWkZu4WUAM....jpeg FFWkZt_XIAULD....jpeg FFWkXa_XIAEPY....jpeg Show all

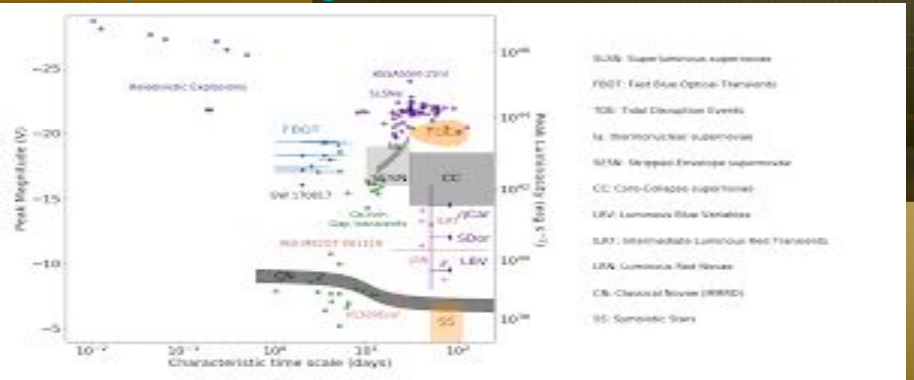
Possible AI Specialization? Data reasons:



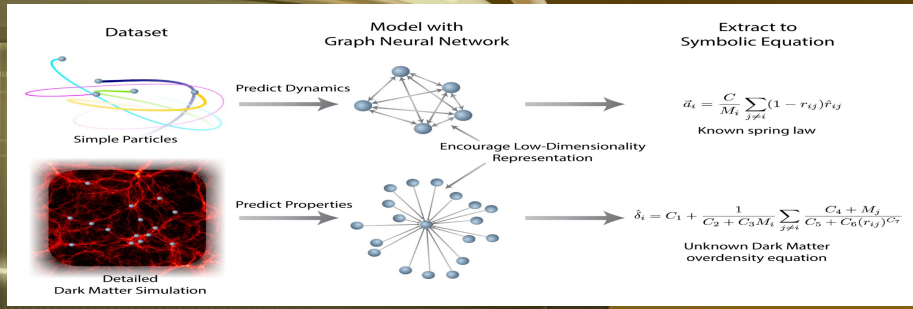
Important to object Detection/Clasification
https://iopscience.iop.org/article/10.3847/1538-4365/ab8868



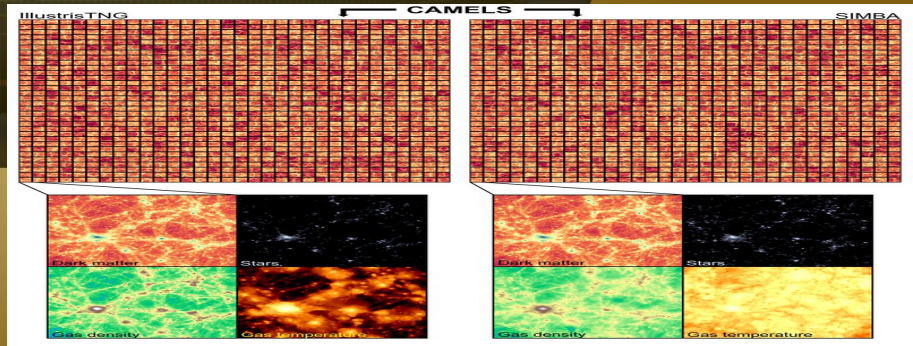
https://arxiv.org/abs/2107.10281



Important to Inference in Cosmology:
https://arxiv.org/abs/2111.08683



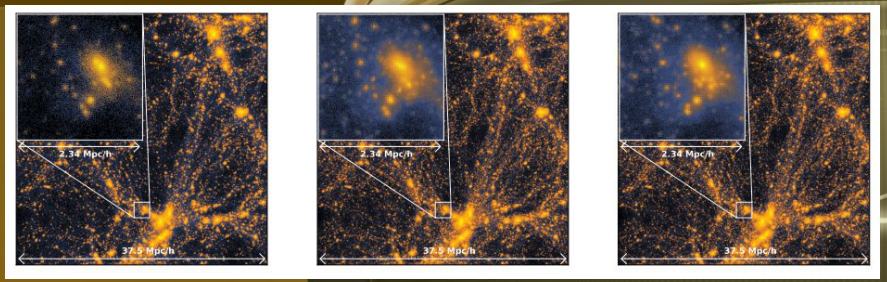
https://www.simonsfoundation.org/2021/10/24/astrophysicists-reveal-largest-ever-suite-of-universe-simulations/?



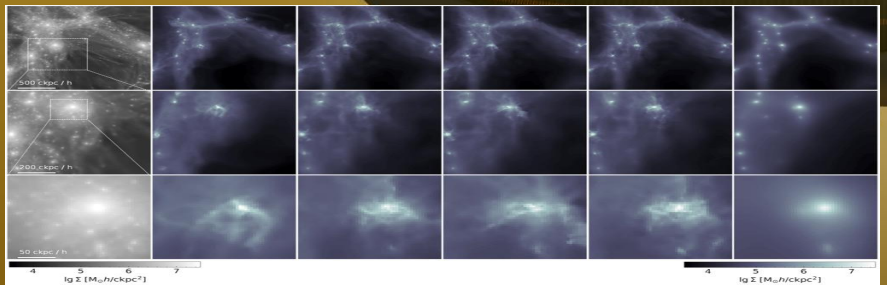
Possible AI Specialization? Simulations reasons:

Strong impact on Simulations improvement:

https://arxiv.org/abs/2111.06393

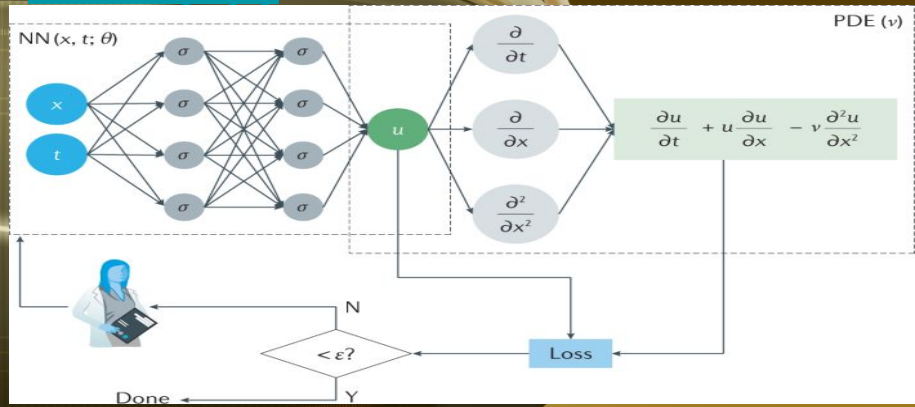


https://arxiv.org/abs/2110.11970



Physics-informed neural networks

https://www.nature.com/articles/s42254-021-00314-5



Differentiable programming has been used within backpropagation of neural networks, probabilistic programming, and Bayesian inference.

See PyMC3, JAX, FLUX

Nodo IA UNC+Gob Cdba + Cluster Technology Cordoba

Nodo de colaboracion Cientifico Industrial para Investigacion y Desarrollo de la Inteligencia Artificial.

Es una iniciativa para construir conocimiento en Ciencia de Datos y desarrollar soluciones innovadoras basadas en la inteligencia artificial

Multiples actividades de vinculacion en desarrollo.

Diplodatos FaMAF-UNC y Matematica Aplicada

Nabucodonosor 1, 2 -> https://ccad.unc.edu.ar/2021/09/04/el-ccad-obtuvo-un-pfi-por-6m-para-actualizar-mendieta/

Nuevos stakeholders: 500 K U$S por firmarse

En desarrollo: 1.5 M U$$ bajo coordinacion del MinCyT Cdba en colaboracion con UNRC, UTN, CONAE, UCC, US21, UBP, +INTI, +INTA para trabajos de vinculacion en IoT, DS, ML, Robotica!

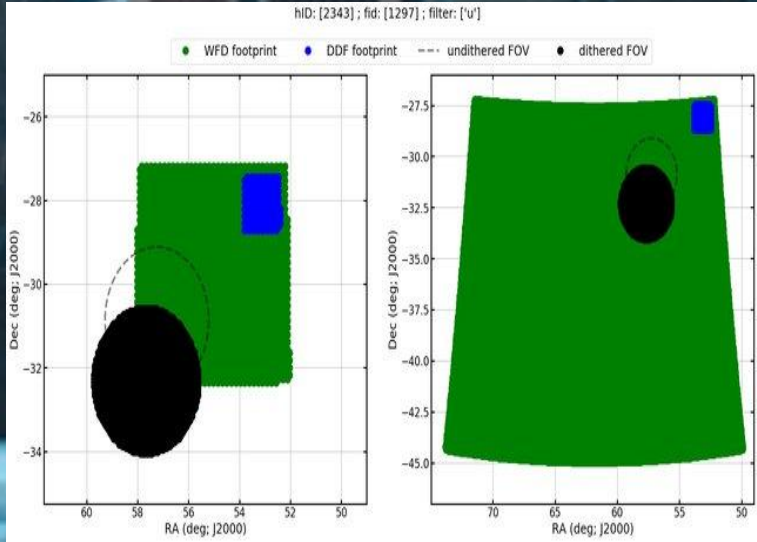
50K for

FireML

& LSST

Gracias por su atencion!

https://www.conicet.gov.ar/el-conicet-y-aws-anuncian-una-convocatoria-para-acelerar-e-impulsar-la-innovacion-de-proyectos-en-la-nube/ We plan to put DC2 simulation on a **Dask cluster using** Amazon SageMaker and AWS Fargate



Why you should understand ML:
https://www.youtube.com/watch?v=SGUCcjHTmGY