# Virtual Observations 2016

- Caput college – special
- From Virtual Observations -> Data Science 16/17
- Data science – data science – data science
  - future, jobs in and outside astronomy
- EV Astronomical Information Technology
  - many ways of being an astronomer- and beyond

# Virtual Observations 2016 is about

- BigData
- Data bases
- Information Systems in Astronomy
- Data mining
- Data processing
- Data  federations

# Big Data Machines

- Lofar 2010 www.lofar.org
- VST-OMEGACAM 2011 www.astro.rug.nl/~omegacam
- ALMA 2013 www.eso.org/sci/facilities/alma
- GAIA 2014 www.sci.esa.int/gaia/
- EUCLID 2020 www.euclid-ec.org
- LSST 30Tb/night www.lsst.org
- SKA 2022 www.skatelescope.org
- The virtual observatory, Euro-VO, IVOA

# Virtual Observations 2016 languages & standards data base

- SQL & relational DBMS - query language
- XML & XSD - data modeling
- UML & SADT - data modeling
- R - compute/ db binding
- Python - compute/ db binding

# Virtual Observations 2016

- April - June 2016
- 2h lecture + 2h lecture + 2h werkcollege / week
- 7 June 14:00 – 15:40  Euro-Vis conference visit
- Exams: June 2016

# Virtual Observations 2016 schedule

- Monday 13:00 ZG 257
- Tuesday 17:00  PC room  ZG 142
- Thursday 15:00 PC room  ZG 142
- EV – data science- information systems
- Dr Andrey Belikov - techniques
- Dr Gijs Verdoes Kleijn – astrometry, photometry

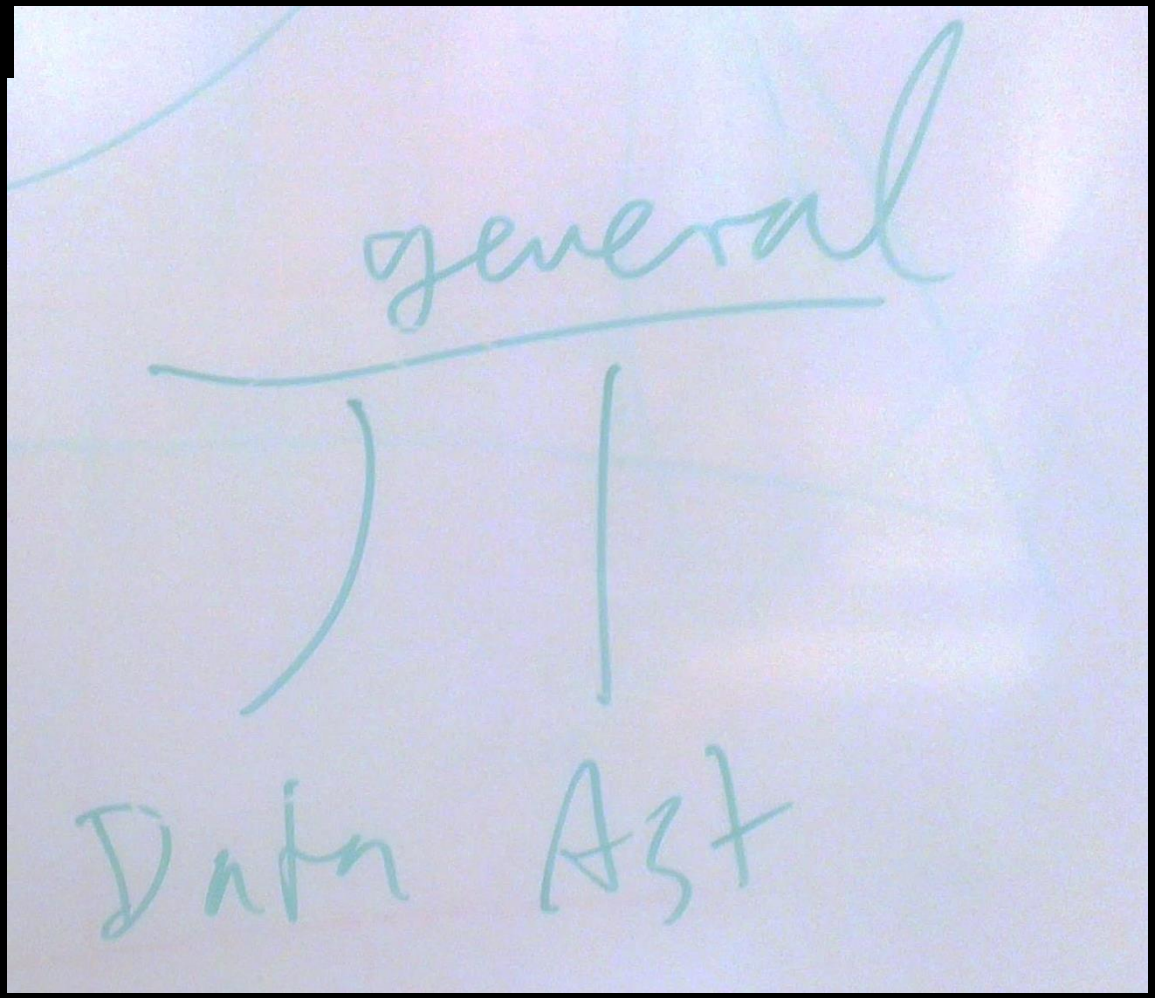
# Virtual Observations - exams

- 1 mandatory werkcollege tasks
- 1  examination task (free)  4-5 pages -> exam

# Virtual Observations 2016 contacts


- Prof Dr Edwin A. Valentijn
- valentyn@astro.rug.nl
- Andrey Belikov
- belikov@astro.rug.nl
- r. 127
- www.astro.rug.nl/~belikov/VO2016/

# DS & CS - data scientist



# Data federations

Edwin A. Valentijn

Prof Astronomical Information Technology

Target- OmegaCEN – Kapteyn

University of Groningen

Data Science & Complex Systems Symposium

# Compute centric 1970's main frames

- 1 -100 Mbyte
- Compute – main frame
- Data Store - local
- Data manage - by hand

User = programmer

# Compute centric 1980-90's workstations

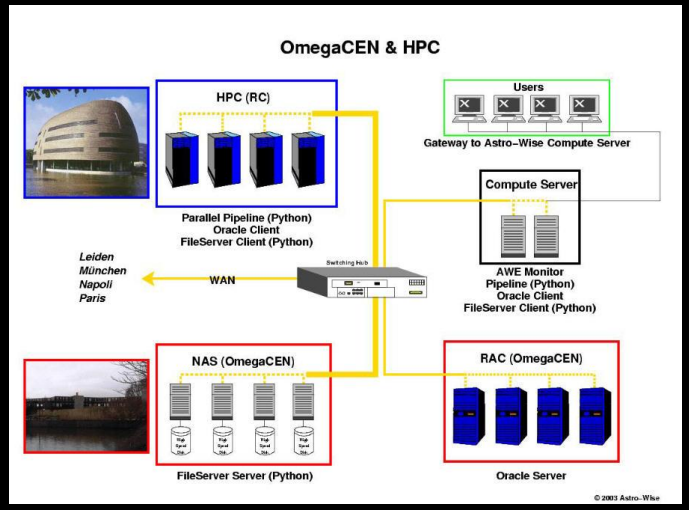
- 1 -100 Gbyte
- Compute – workstation
- Data Store - local
- Data manage – filesystem /tables

User = programmer

# e-science

- Beyond “workstation science” of the 80-90’s
- Distributed services
- Distributed communities
- Distributed archives
- p2p networks – KAZAA- NAPSTAR
  - Share cpu
  - Share storage
  - Share info / meta data /knowledge

# Datacentric 2000-2010's local networks and internet




2003 Rug-CIT

- 1-100 Tbyte – Pbytes
- Compute – local-grid
- Data Store - distributed
- Data manage - database

} User

# Datacentric > 2015's living archives communities - data federations


- 1-100 Pbyte
- Compute – local– grid -local
- Data Store - distributed
- Data manage - database

} data scientist

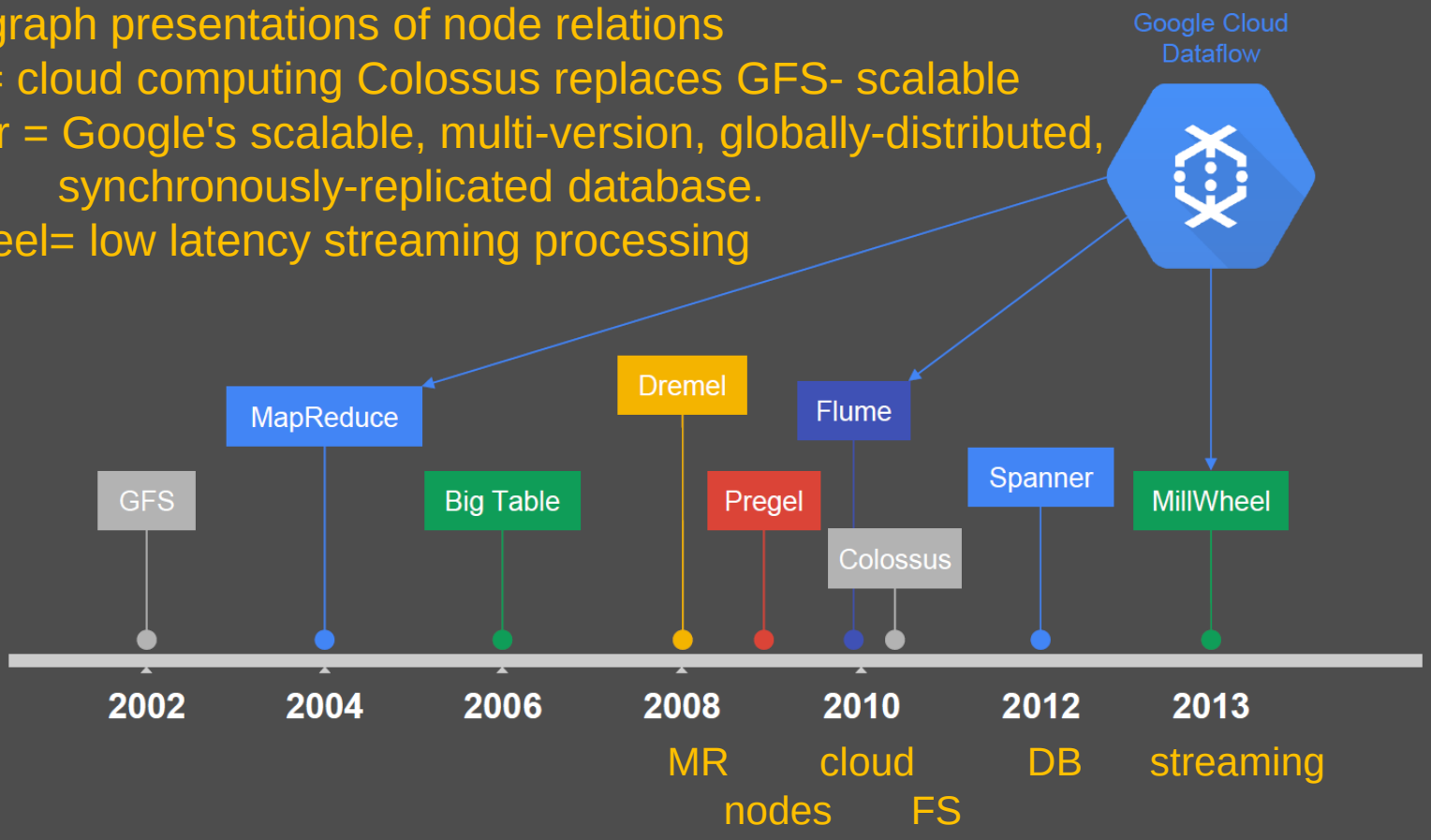
# why?

- Moore's law N -> 2 N in 2 years
- both cpu and data
- But N data point have N**2 connections in 2 D
- Cpu's can't cope with increasing dataconnections
- We have to be be smart and write N Log N code
- we have to smartly index data, eg in trees
- -> i.e. data about data - Metadata
- The programmer becomes a data scientist

# Big five of Big data

- Microsoft
- Amazone
- Google
- Yahoo
- Facebook

Dremel = scalable Map reduce 1000's cpus
Pregel graph presentations of node relations
Flume = cloud computing Colossus replaces GFS- scalable
Spanner = Google's scalable, multi-version, globally-distributed, synchronously-replicated database.
Millewheel= low latency streaming processing





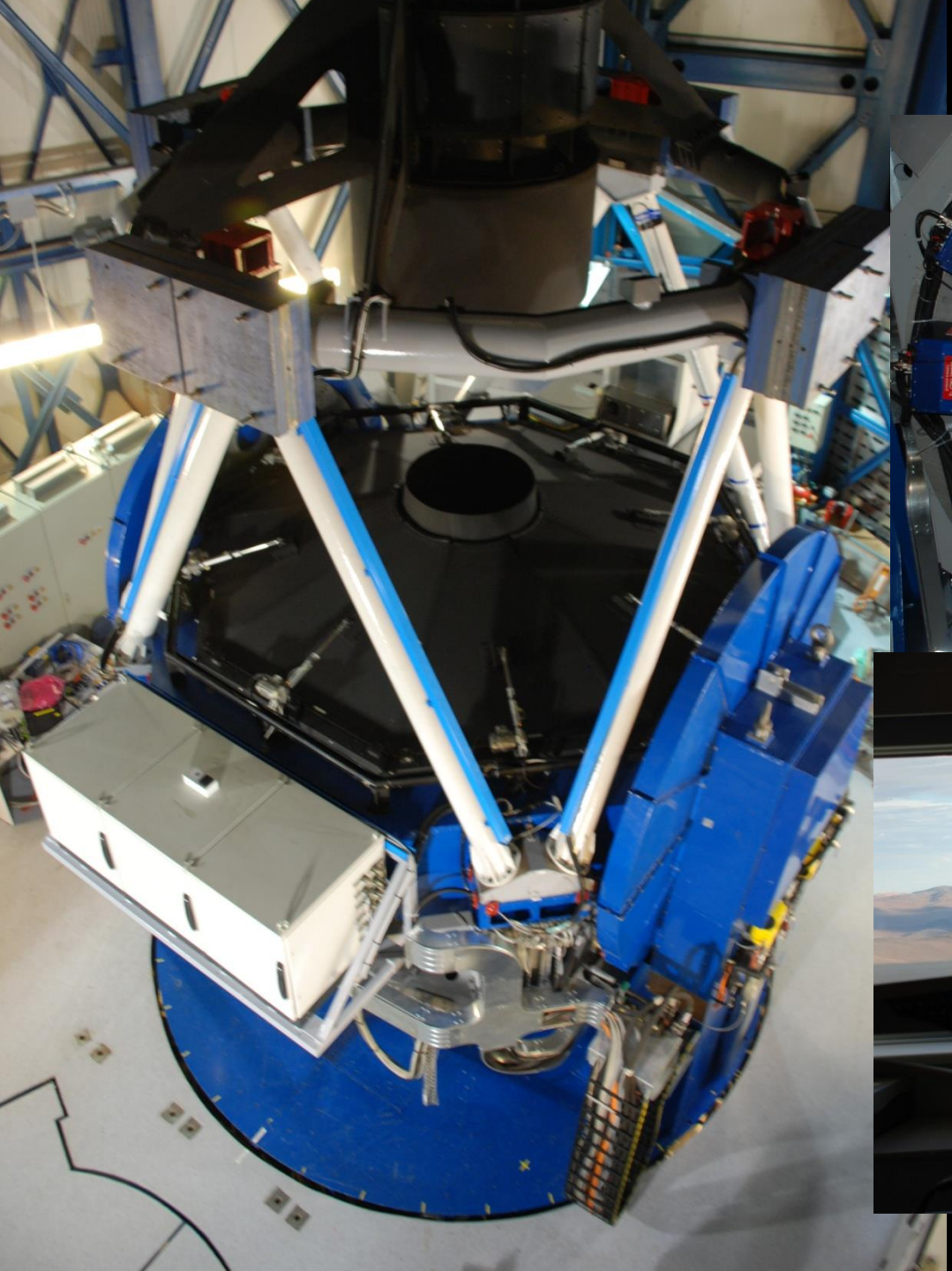

Target

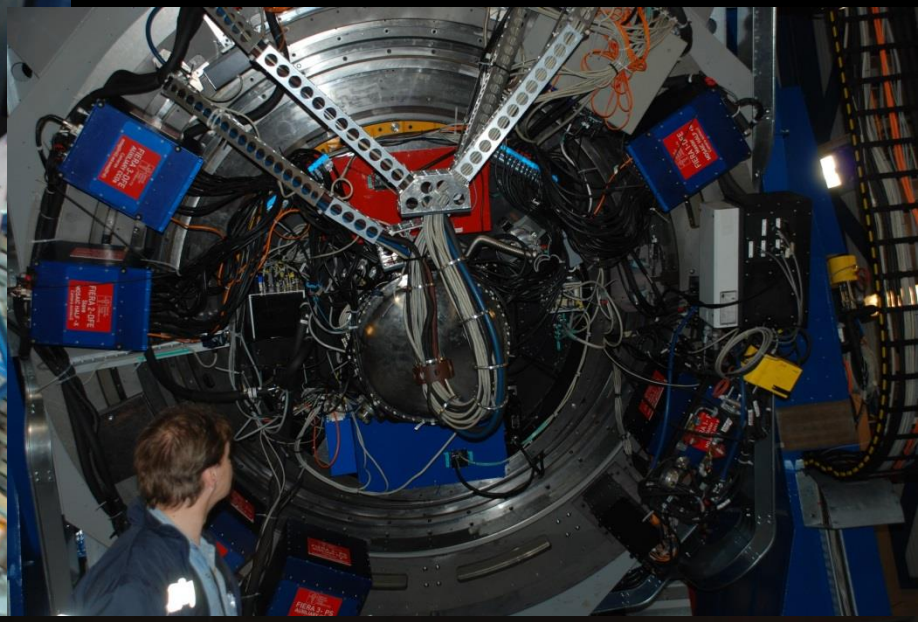
# Petabyte Scale Data at Facebook

Dhruba Borthakur ,
Engineer at Facebook,
XLDB Conference at Stanford University, Sept 2012

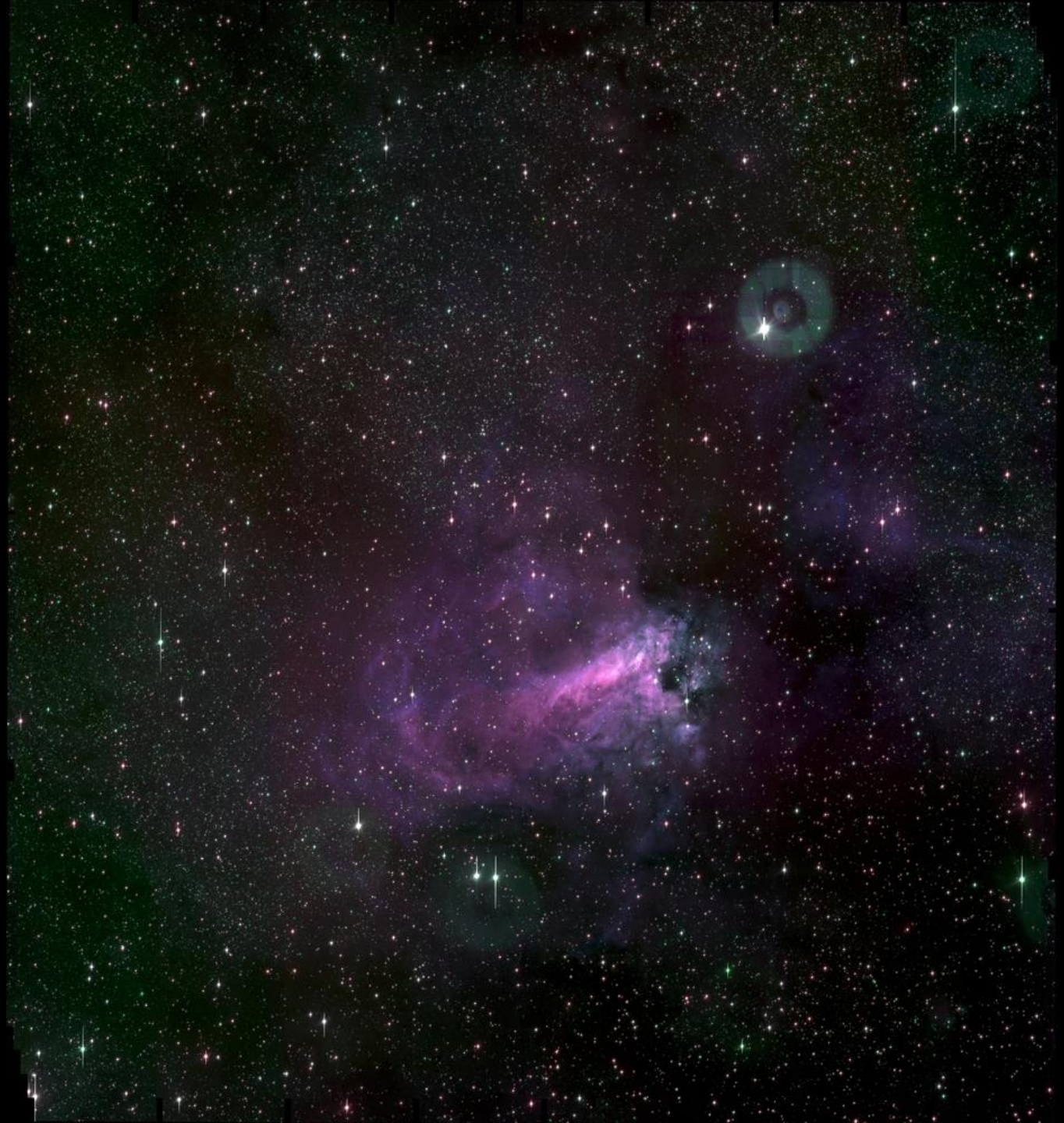
| | Total Size | Technology | Bottlenecks |
|---|---|---|---|
| **Facebook Graph** | Single digit petabytes | MySQL and TAO | Random read IOPS |
| **Facebook Messages and Time Series Data** | Tens of petabytes | HBase and HDFS | Write IOPS and storage capacity |
| **Facebook Photos** | High tens of petabytes | Haystack | storage capacity |
| **Data Warehouse** | Hundreds of petabytes | Hive, HDFS and Hadoop | storage capacity |


KiDS
VISTA

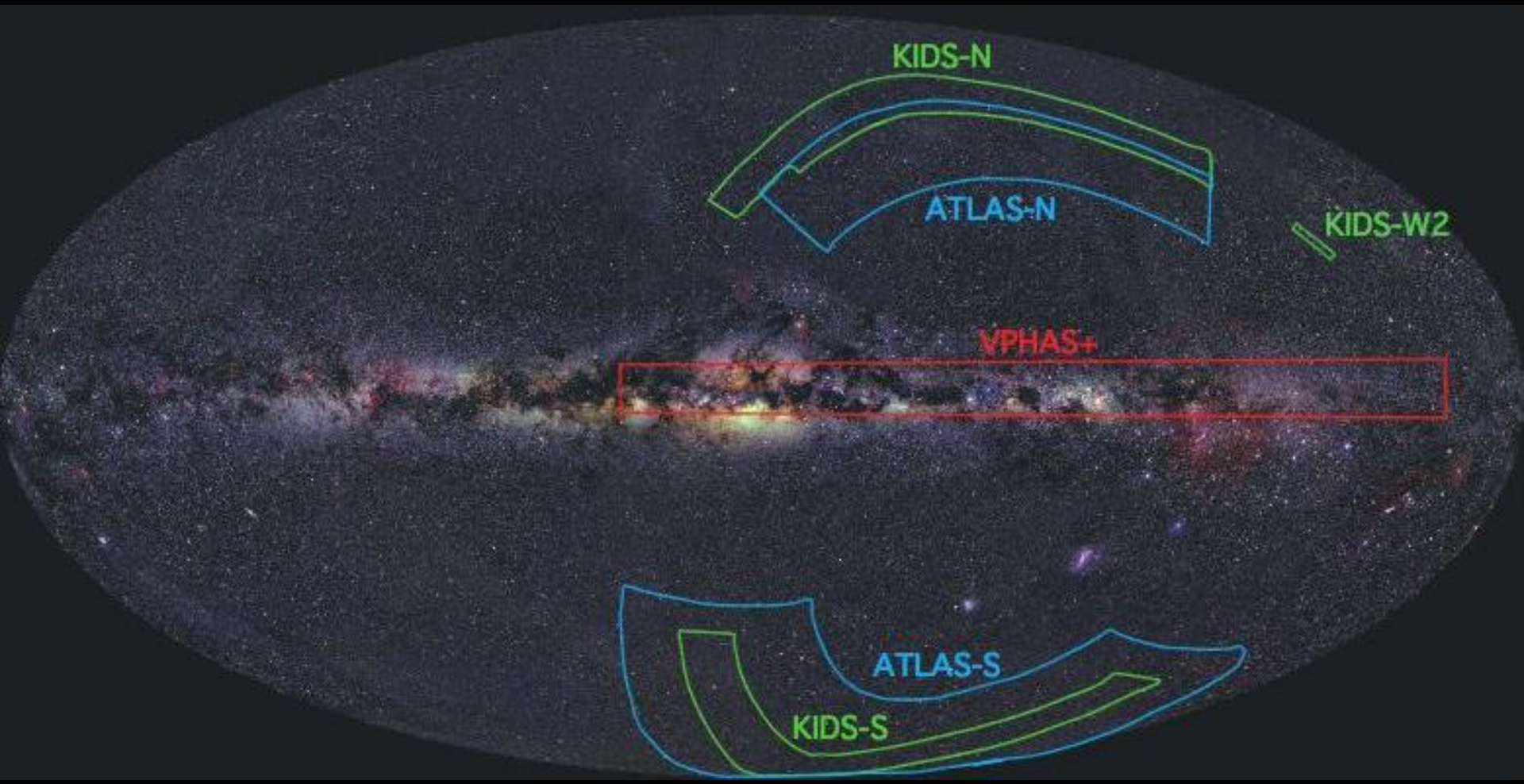




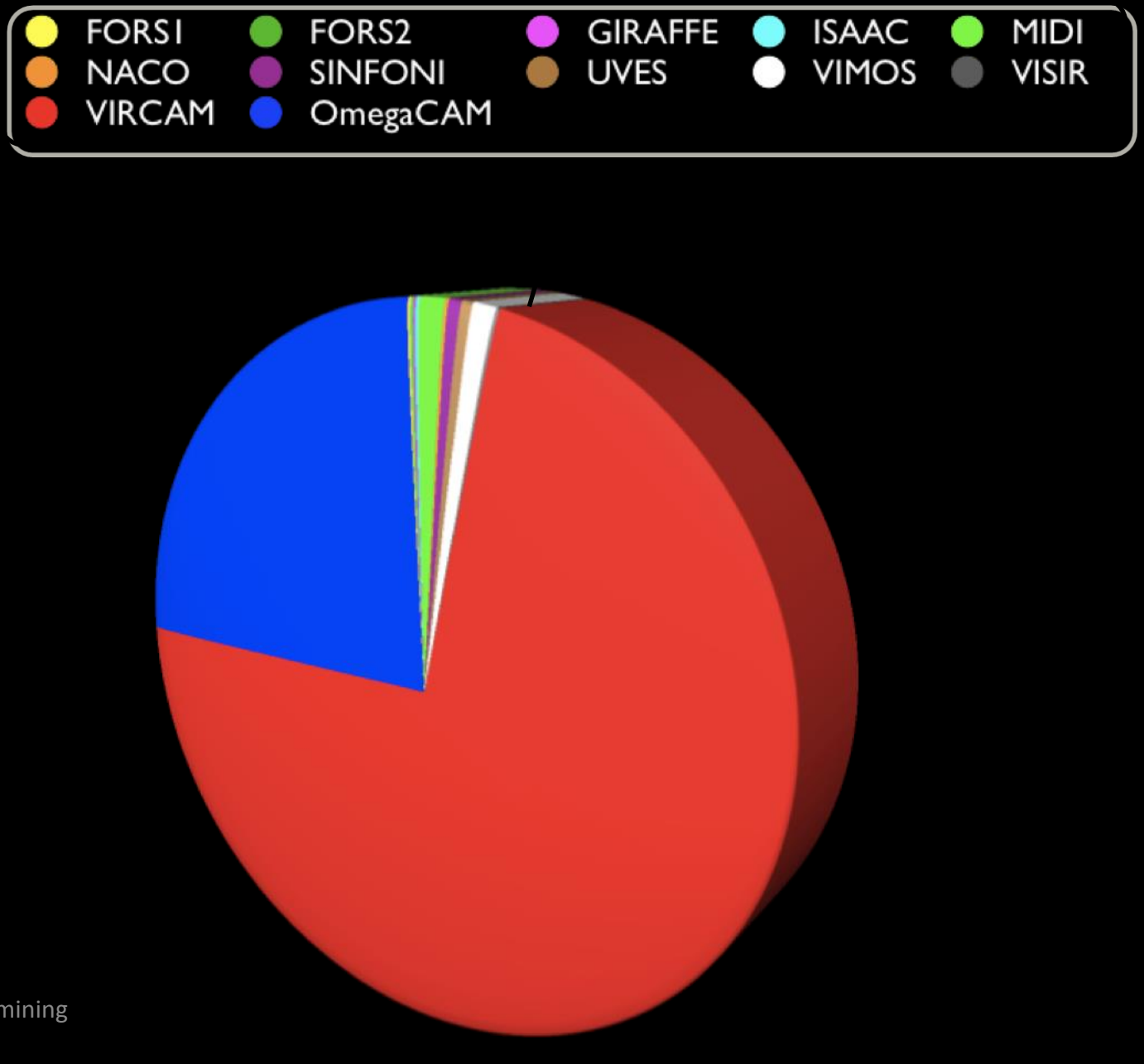

Leo triplet



# ESO public surveys



# Paranal Monthly Data Rates 2007 statistics

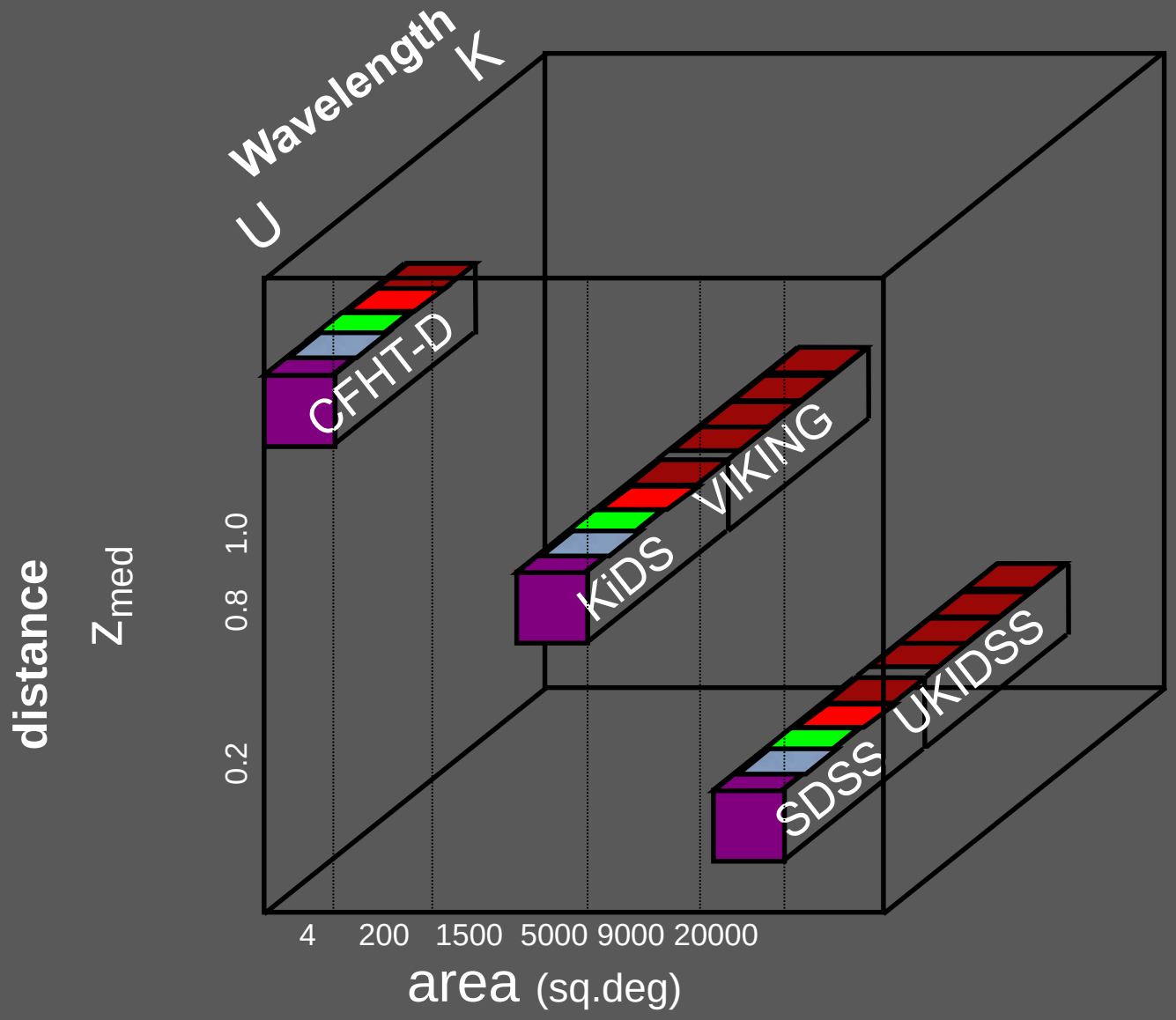


# KiDS vs. other surveys

Earlier:

SDSS

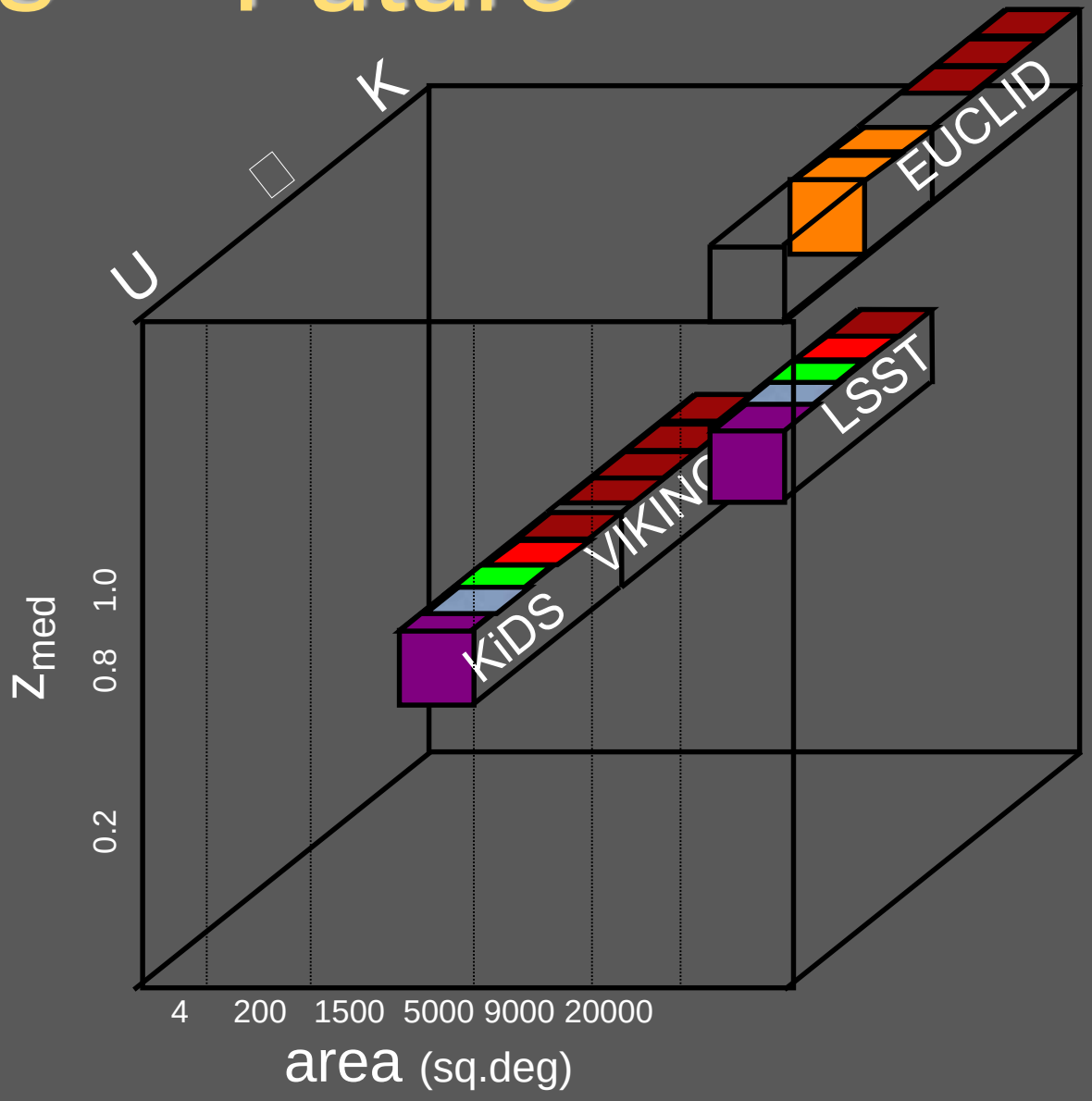
CFHT-D




Target

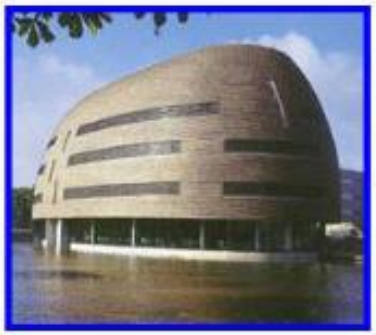
# Kids - Future

**Euclid**
**Target 2013-2018**
**Science Ground Segment**
**Datacenter**
**Datacentric- WISE**
**ESA**




Target

# VST - Virtual Survey Telescope



HPC (RC)

Parallel Pipeline (Python)
Oracle Client
FileServer Client (Python)

Users

Gateway to Astro–Wise Compute Server

Compute Server

AWE Monitor
Pipeline (Python)
Oracle Client
FileServer Client (Python)

*Leiden*
*München*
*Napoli*
*Paris*

WAN

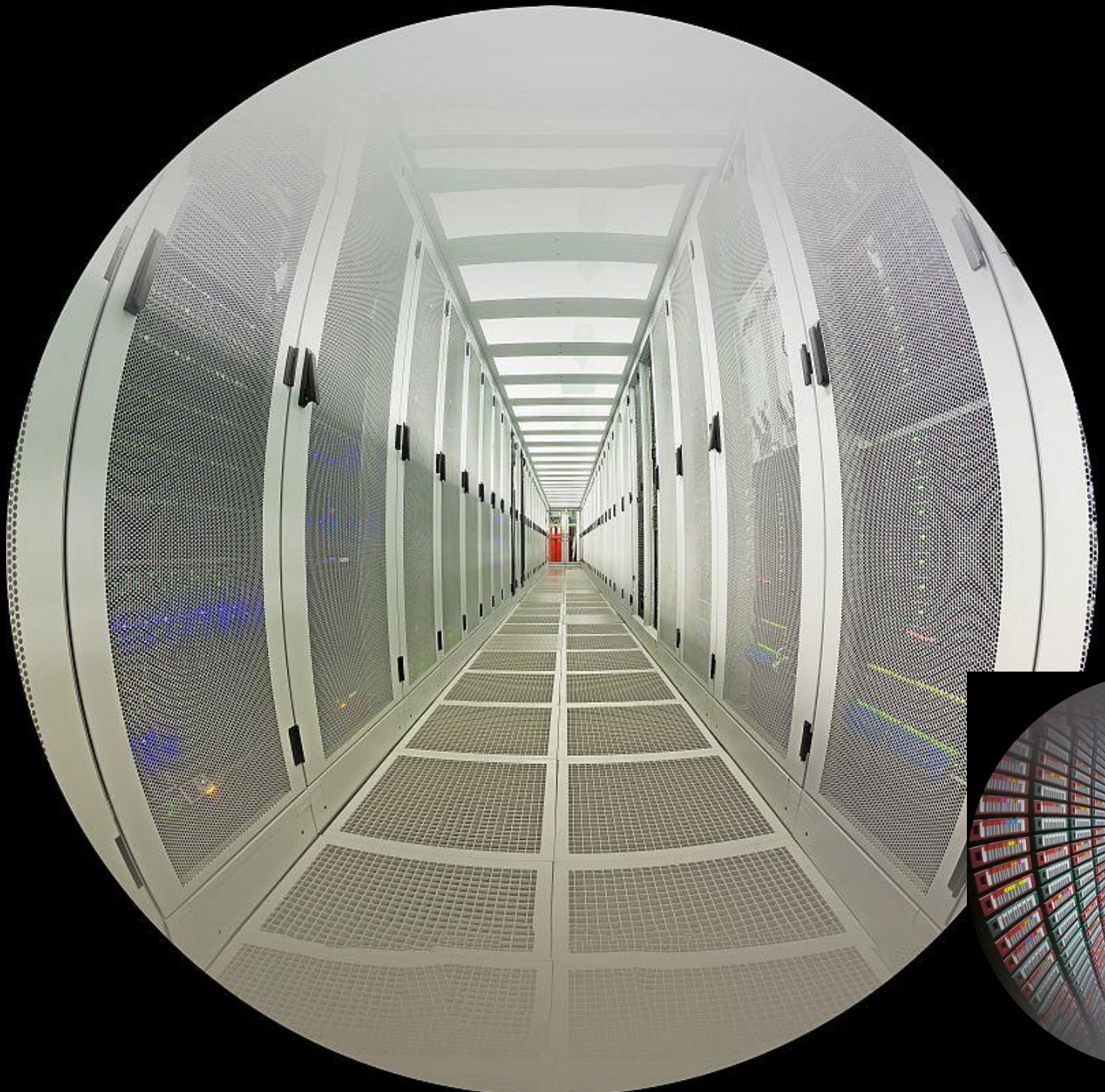
Switching Hub

NAS (OmegaCEN)

FileServer Server (Python)

RAC (OmegaCEN)

Oracle Server



# control

- Limited control
- States run behind
- Commercial companies in the lead

  *.doc,  *.pdf ,  Google, *png, *.jpeg

- Powers of two:
  - -Information Universe
  - -Infoversum movie

# 7 bit

LYNDON B. JOHNSON

XXXVI *President of the United States: 1963-1969*

**127 - Memorandum Approving the Adoption by the Federal Government of a Standard Code for Information Interchange.**

*March 11, 1968*

# 7 bit ASCII

| bits | #states | Byte |
|---|---|---|
| 0 | 1e+00 1 | pre Big Bang |
| 1 | 2e+00 2 | Big Bang |
| 7 | 1e+02 128 | ASCII |

## 7 bits: 0110110

- Our in vitro Information Universe is fully due to agreements between people
  - 7 bit ASCII
  - standards -bitstreams

# Basics- Surveys

- Defined area on sky
- Homogeneous

  Survey limit

    Flux (magnitude)

    Size

    Surface brightness

    distance
- Quality control

# Basics - time

Everything changes in time

- Physical changes
- Our inside in modeling
- Methods, code, bugs

# basics- pipelines

- Workflow
- What triggers a pipeline?
  - Data items
  - Operators
  - users

# Basics - Information systems

- Pipeline design
- Standards: Fits, table format- VO Standards
- Protocols
- Project management- sociology
- Data model
- Data base
- Integrated/ distributed File systems
  - Grid FTP, AstroWise,  Hadoop, Cloud, Dropbox
  - Distributed computing

  Grids, cloud computing

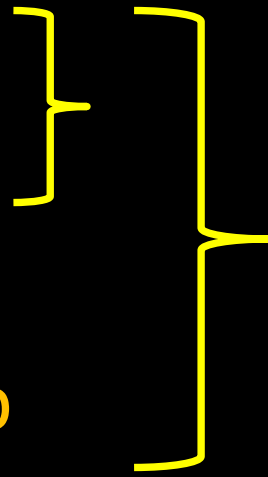
# Astro-WISE information system – fully datacentric

All data beyond pixel data is Metadata

all pixel data <–>data servers

all Metadata <–> database

compute clusters / GRIDs all I/O to db



- all components scalable
- all components EU distributed

# N params N data
# back to basics

- Joins – links
- ++ Inheritance – dependencies

- Everything in cs is adresses
  memory, ASCII, namespaces, registry
- Optimize , organize, index
- management

# Peta -100 Peta

| bits | #states | | Byte |
|---|---|---|---|
| 0 | 1e+00 | 1 | pre Big Bang |
| 1 | 2e+00 | 2 | Big Bang |
| 8 | 3e+02 | 256 | Machu Pichu |
| 16 | 7e+04 | 65536 | |
| 24 | 2e+07 | 16777216 | Mega |
| 32 | 4e+09 | 4294967296 | Giga |
| 40 | 1e+12 | 1099511627776 | Tera |
| 48 | 3e+14 | 281474976710656 | |
| 56 | 7e+16 | 72057594037927936 | Peta |
| 64 | 2e+19 | 18446744073709551616 | 100 Peta |

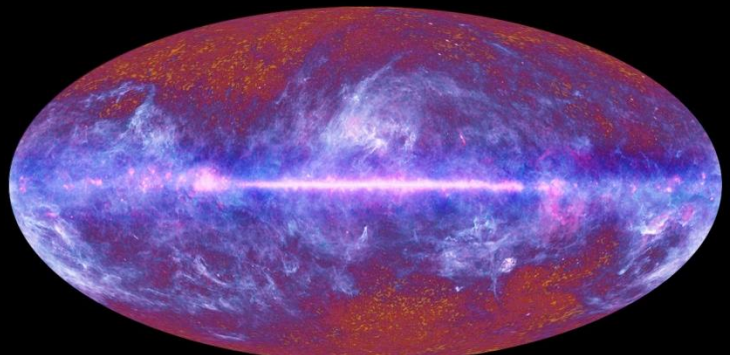
Data about

Big Data

Metadata

# The Universe - 256 bit

| bits | #states | | Byte |
|---|---|---|---|
| 0 | 1e+00 | 1 | pre Big Bang |
| 1 | 2e+00 | 2 | Big Bang |
| 8 | 3e+02 | 256 | Machu Pichu |
| 16 | 7e+04 | 65536 | |
| 24 | 2e+07 | 16777216 | Mega |
| 32 | 4e+09 | 4294967296 | Giga |
| 40 | 1e+12 | 1099511627776 | Tera |
| 48 | 3e+14 | 281474976710656 | |
| 56 | 7e+16 | 72057594037927936 | Peta |
| 64 | 2e+19 | 18446744073709551616 | 100 Peta |
| 128 | 3e+38 | 340282366920938463463374607431768211456 | |
| 256 | 1e+77 | 115792089237316195423570985008687907853269984665640564039457584007913129639936 | |



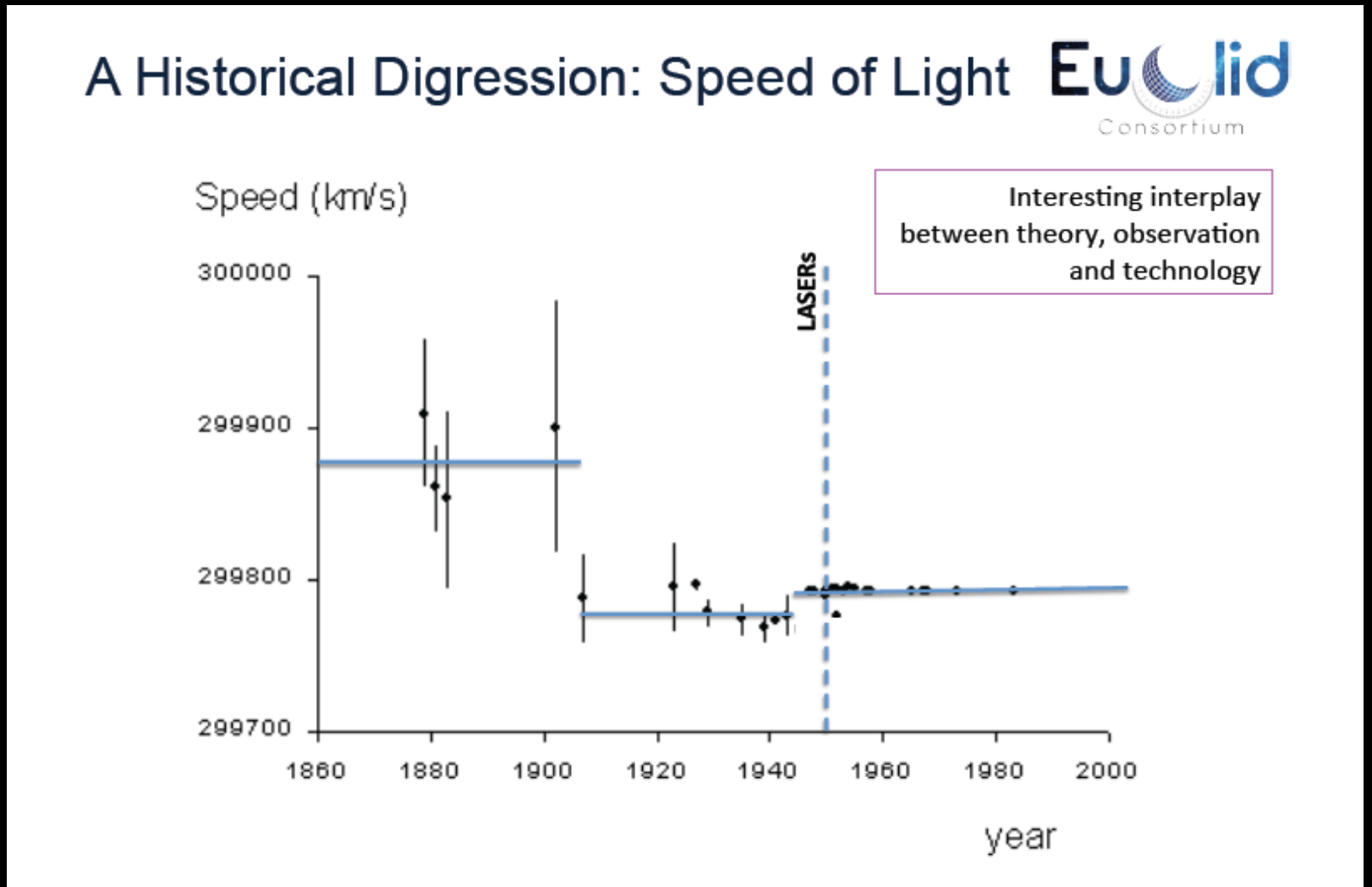
Target

# Users – two flavours

- Finding extremely rare objects  e.g.
  - Near earth objects
  - High redshift Quasars
- Statistics  on all data
  - Dark matter mapping
  - Dark energy origins

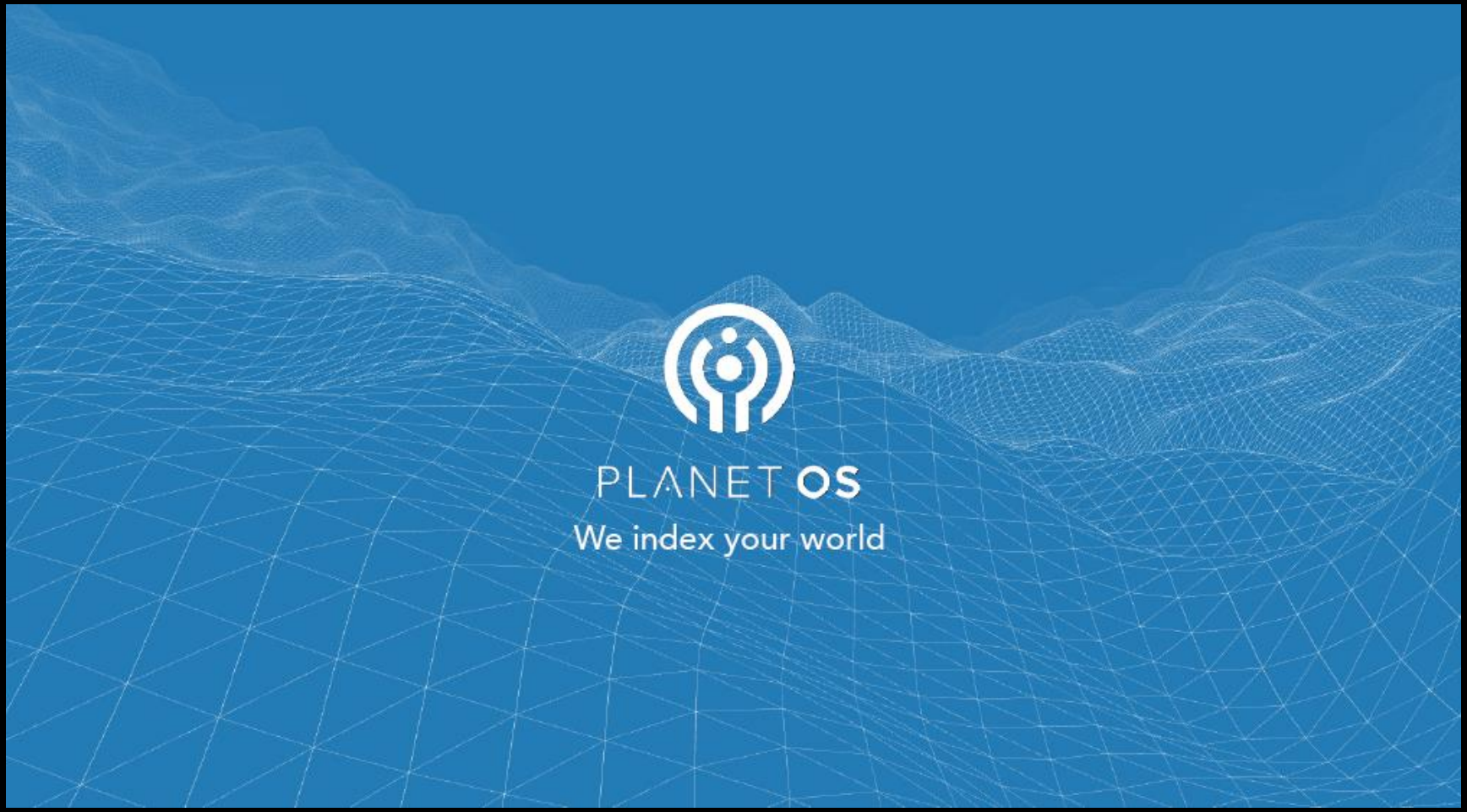
# SQRT (N) OK but Systematics in Big Data…..

## Euclid –backward chaining – re-process



# Quality control

- Distributed
- Shared over the whole community
- web based
- OmegaCAM calibration plan
- OmegaCAM observing strategies




Target