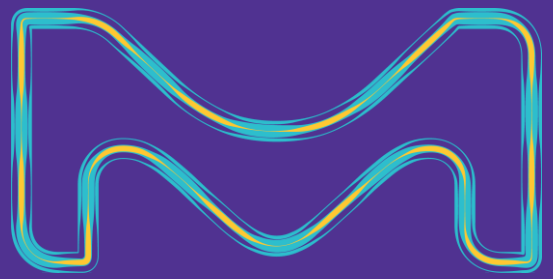
X-omics platform

A validated platform for multi-omics data analysis

Stefan Pinkert on behalf of the XOP team
R/Pharma 2020, 15th OCT 2020 1:40PM



MERCK

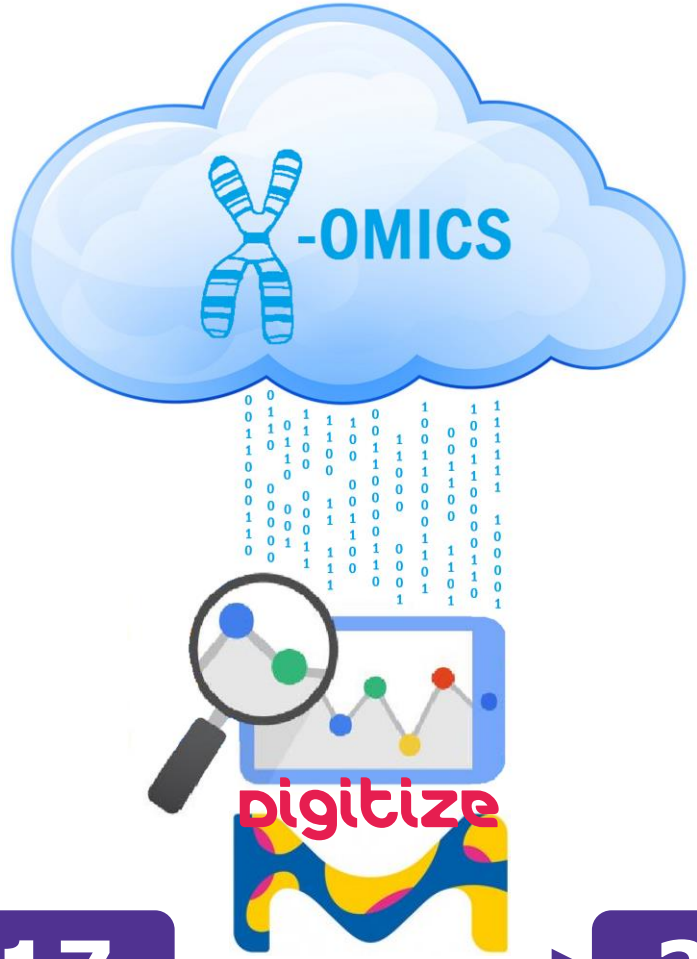
X-OMICS Platform to Boost Computational Capabilities in Biomarker R&D

Scientists: poor access to clinical biomarker data & computational tools

Bioinformatics: Insufficient automation of standard analyses

Little use of scalable & fast cloud infrastructure

No platform to manage & process all raw clinical omics data



One place to integrate omics data **from all sources** (clinical, public, consortia)

Self-service tools for broad user groups

Comprehensive management of **access rights**

Fully traceable & reproducible data processing

Set up to leverage emerging **opportunities by AI/ML** (digital pathology)

Cutting-edge analysis pipelines for genomic data

Validated

2017 ⇢ **2019**

MERCK

X-OMICS Platform successfully co-developed with Genedata

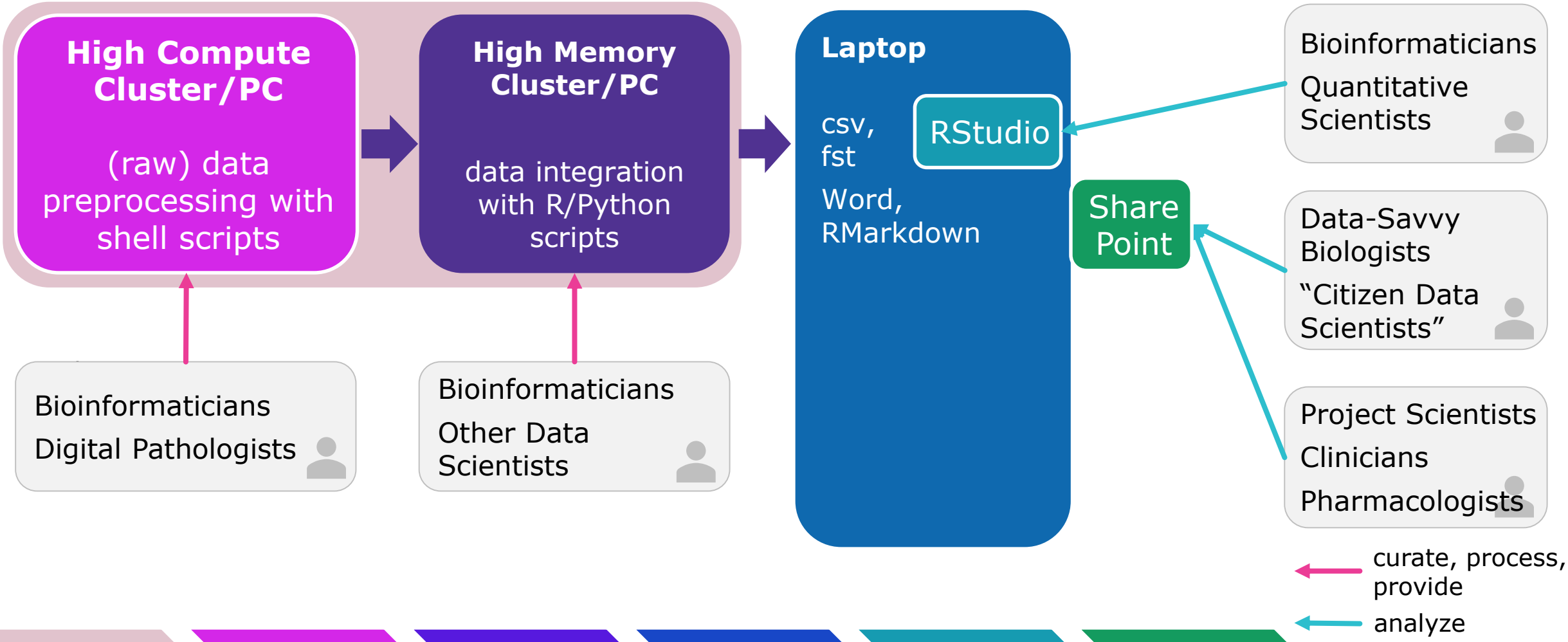
Source: https://www.genedata.com/products/profiler

exploratory research with R

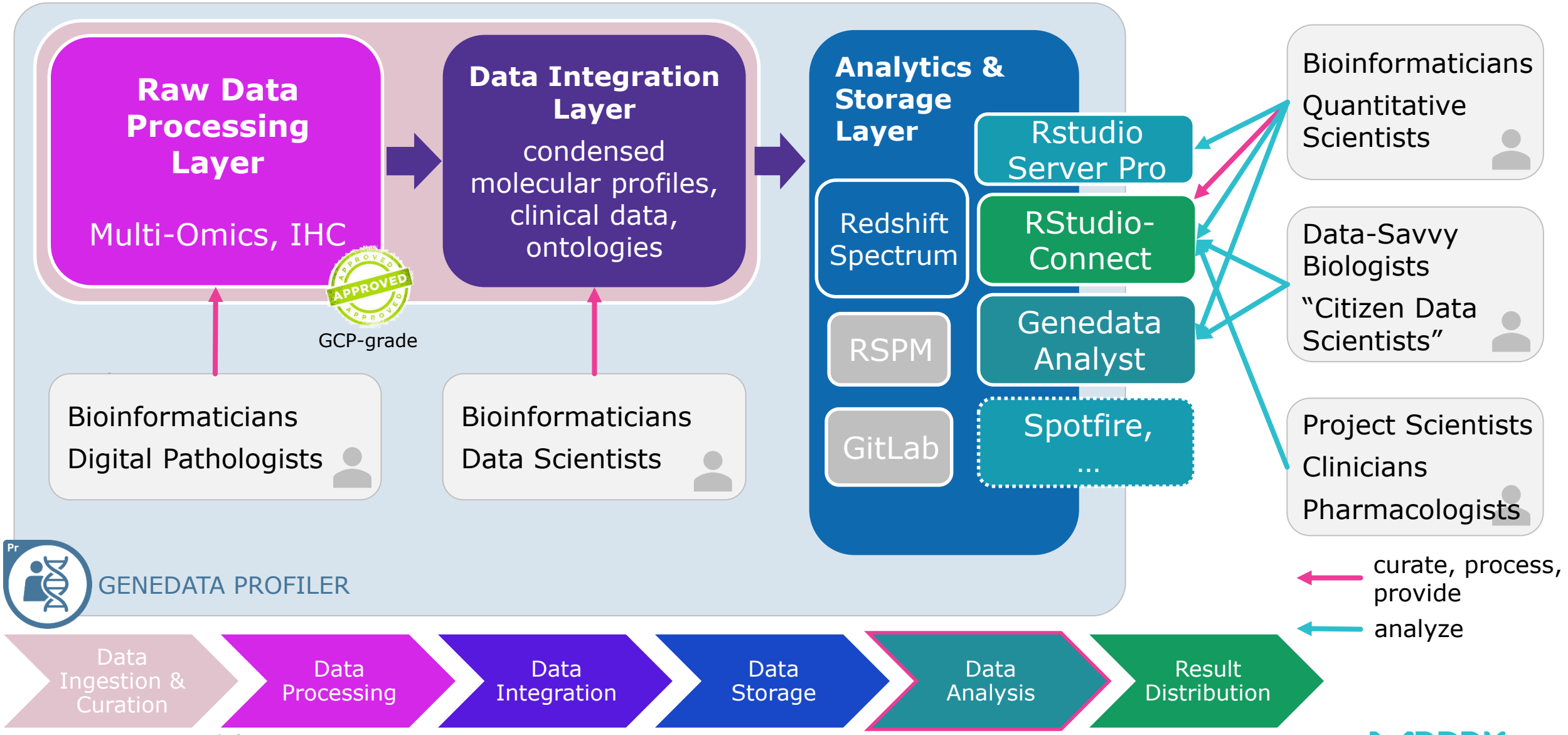
01

MERCK

How do exploratory researchers work?



How do researchers work with the X-OMICS Platform!



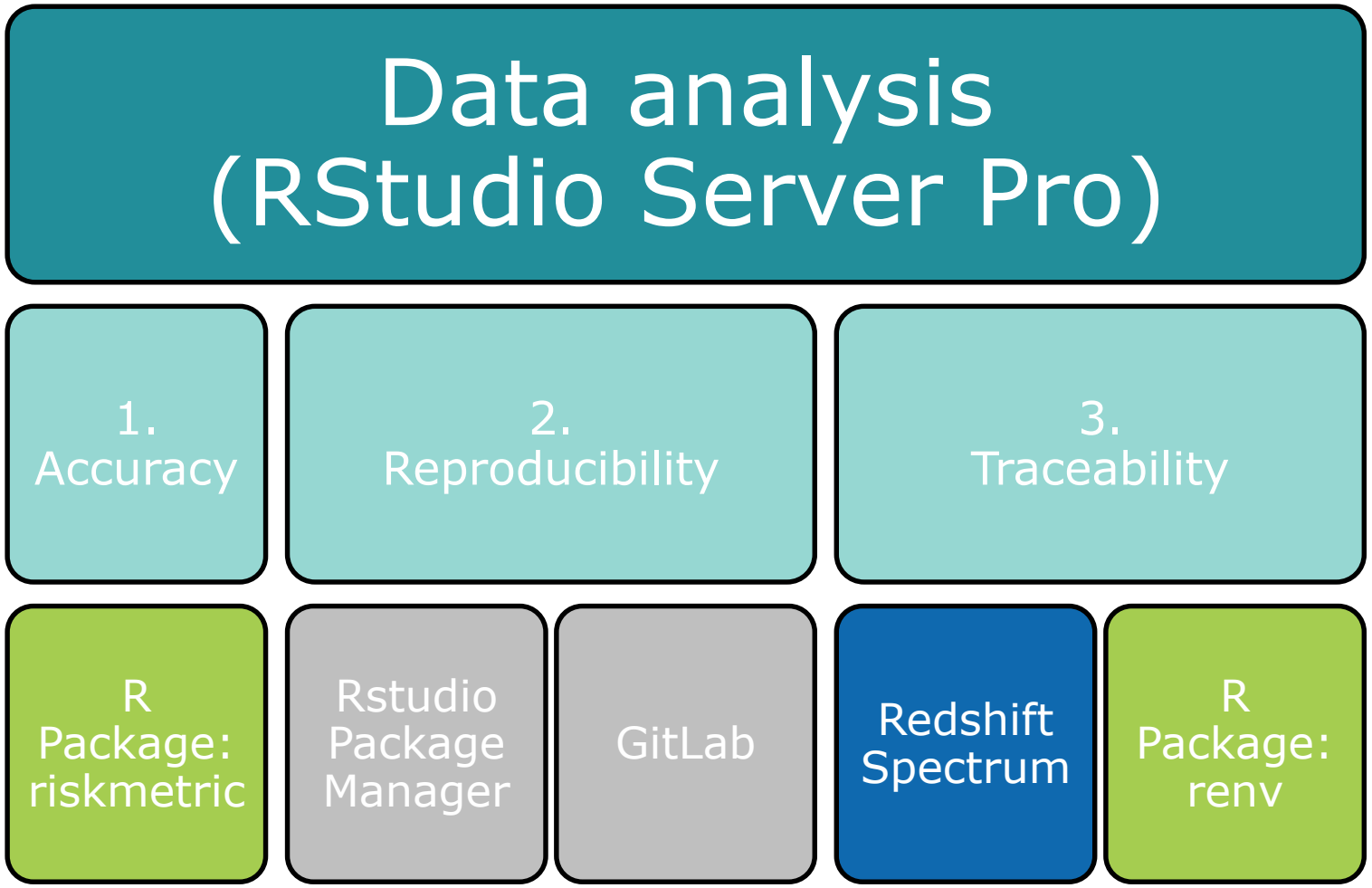
R Validation Hub
Which elements are required for a validated system?

Data Analysis

Since the FDA does not require use of any specific software for statistical analyses, the programming language R can be used if the R installation incorporates all of the following elements:

1. Accuracy
2. Reproducibility
3. Traceability

Source: https://www.pharmar.org/presentations/r_packages-white_paper.pdf



The XDP supports reproducibility

02

Reproducibility

X-OMICS Platform uses RSPM and GitLab to support Reproducibility

It is important to acknowledge that R (like other open source languages) presents additional challenges with respect to the reproducibility of an environment. The **evolution of R packages is effectively continuous** and thus maintaining a stable R installation that allows for the addition of new and/or updated packages can be a challenge.

Source: https://www.pharmar.org/presentations/r_packages-white_paper.pdf

RStudio Package Manager

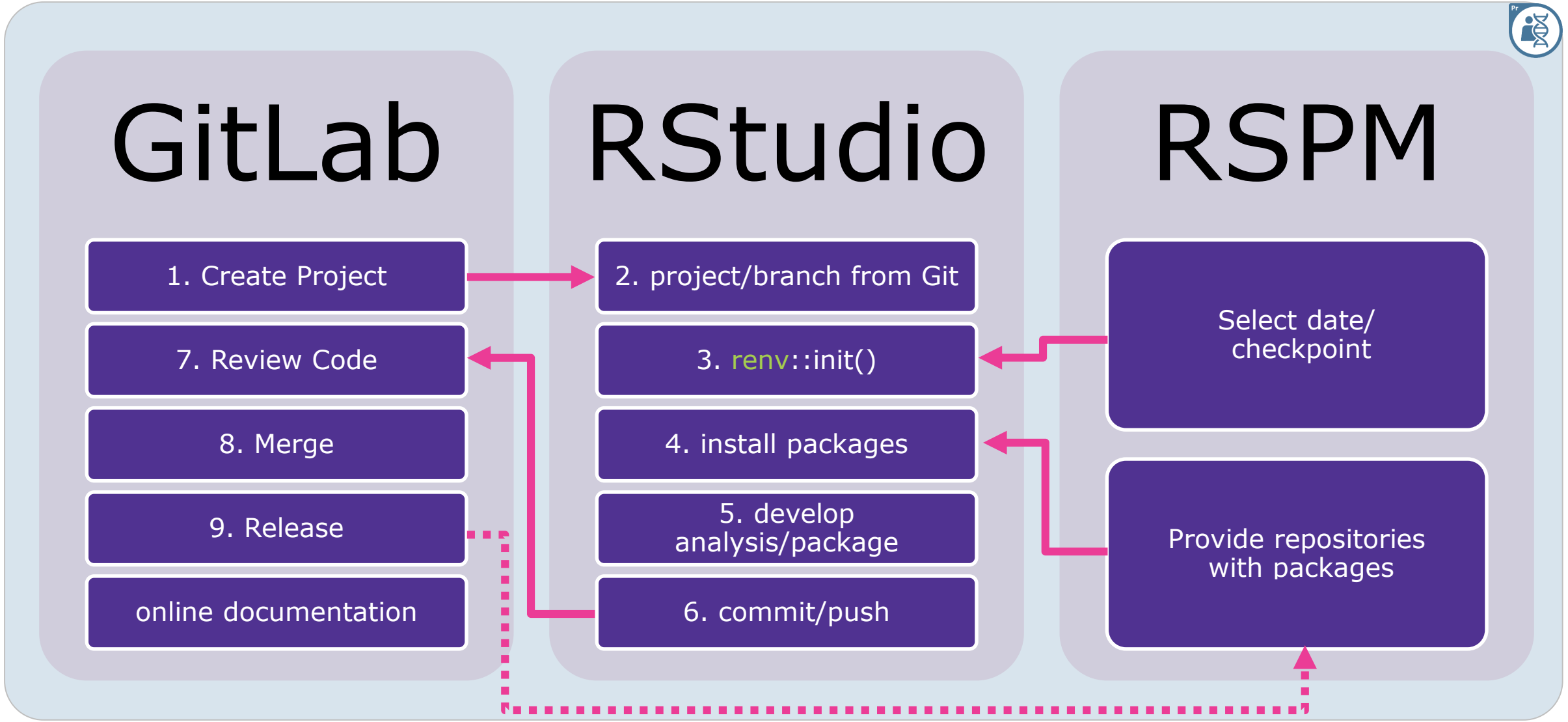
- Local copies of public packages
- Internal repositories / packages
- **Date based checkpoints**
- Hierarchical order of repositories

GitLab

- Analysis / package version control
- Role based code review & release
- Documentation
- Collaborative guiding documents

Reproducibility

The preconfigured X-OMICS platform supports reproducible workflows



03

The XOP supports traceability

Traceability
X-OMICS Platform uses renv and Redshift to support Traceability

Traceability of R installations

Develop system and process controls to automatically **document** the R **packages** and **installation dependencies** that are used in R analyses.

Source:
https://www.pharmar.org/overview

R package: renv

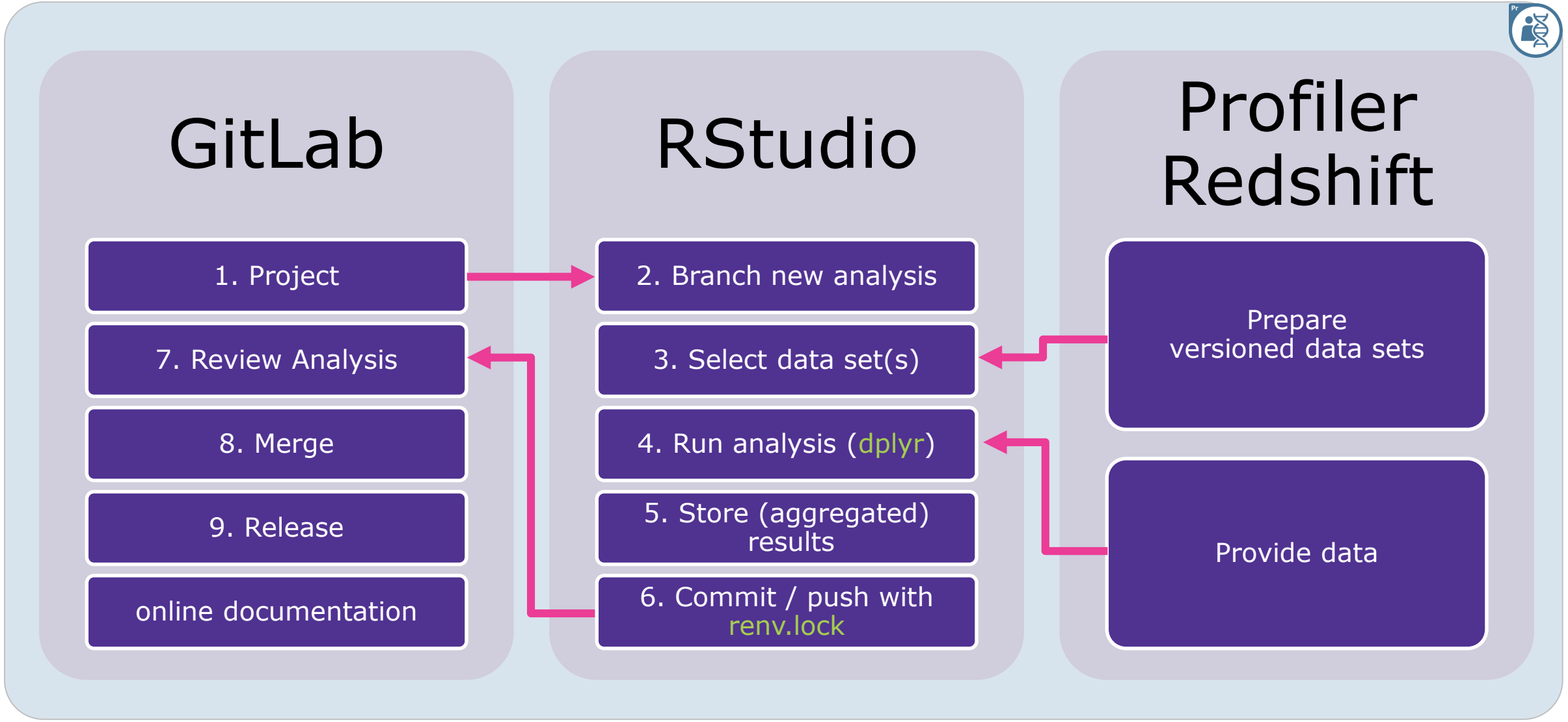
- Project specific package management tool
- References all packages & repositories with versions in a file **renv.lock**
- (soft)links all used packages

Redshift Spectrum (Profiler)

- S3 (parquet) file based; "unlimited" parallel versions
- Access with: SQL or dplyr
- Elastic scalability of storage and compute
- Caching

Traceability

The preconfigured X-OMICS platform supports traceable workflows



Traceability

Interactive Apps guide the traceable naming of data sets

GitLab

1.3 Controlled vocabulary for View Names

↑ go back to overview

`Genedata Profiler (GP)` *Views* are the most important data entities stored and shared by XOP and the names of these views have to follow a specific standard. Each view name on XOP has to consists of a concatenation of the following nine lower-case, alphanumeric, underscore (`_`) separated fields of the form:

`source_dataset_analyte_assay_analysis_release_maturity_curator_version`, with a specific example being:

```
----1-----_--2-_-3-_-4-_-5-_--6-_--7-_--8--_9
cbioportal_tcga_dna_wgs_cnv_20q1_prod_kreis_0
```

Since the last field, `version` is automatically assigned by `Genedata Profiler (GP)`, only the first eight fields have to be assigned by the user. Note that only lower-case alphanumeric characters and `_` are allowed in the final View name. Since `_` is the concatenation symbol of the view name it must not be used within any of the individual fields of the View name.

In detail, the View name fields are defined as follows:

Table 3: Field composition of the `Genedata Profiler (GP)` View name

Position	Field name	Example	Description
1	source	depmap	Lower-case shorthand name of the source of the data, most often an external organization/consortium such as `depmap`, `cbioportal`, `publication`, or the name of an internal unit or function at Merck if the data ha been internally generated
2	dataset	ccle	Lower-case shorthand name of the dataset, often defined by the originating organization. Examples are `tcga`, `mskimpact`, `pdxdatabase` and so forth.
3	analyte	dna	Lower-case shorthand name for the (usually biological) analyte from which the data has been derived by measurement. Examples are `rna` / `dna` (for sequencing measurements from nucleotides) `img` (for imaging), and `phe` for phenotypic data
4	assay	wgs	Lower-case shorthand name for the experimental assay that has been applied on the sample, such as whole-genome sequencing (`wgs`), whole-slide imaging (`wsi`), or extracts from a hospital's clinical data repository (`cdr`)
5	analysis	snv	Lower-case shorthand name for the in-silico analysis that has been applied on the assay data, such as short nucleotide variant calling (`snv`), processing of patient dempgraphics data (`dem`), or computing cell counts based on imaging (`cnt`)
6	release	20q1	Lower-case, lexicographically sortable release date of the data as relating to the original data source. Two formats are foreseen and supported by `xopdata yyqq (19q4)` and `yyyymmdd (20191231)`. The selected format should be kept constant, at least per dataset, but ideally per study.
7	maturity	prod	Lower-case term that captures the maturity of the data in the View; the only allowed names are `src` (for raw data originating from a data source), `dev` (for data of intermediary quality used for software development purposes), and `prod` (for data of production quality, i.e., the highest level of accuracy and robustness)
8	curator	kreis	Lower-case last name of the person within Merck who curated the data for import into XOP. Note that this role may imply initially ownership of the data within XOP; however, since ownership is transferable and View names are immutable, the owner of a View is instead captured in the View's metadata.
9	version	3	Numerical version identifier; this field is *automatically* added by `Genedata Profiler (GP)` during View generation by incrementing on existing versions of the same view. Versions start with `0`.

For the fields `analyte`, `assay`, and `analysis`, multiple examples are defined in a controlled vocabulary as follows; users are encouraged to mix and match to add new permutations whgile striving for re-use of existing identifiers and consistency within their own Studies:

Connect

Naming Guide: Datasets

Please select all fields:

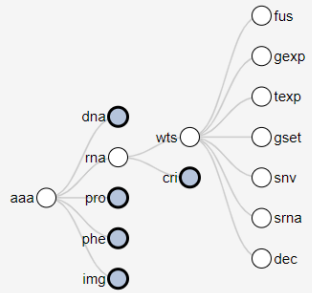
1. Source

gdc

2. Dataset

tcga

3. Analyte, 4. Assay, 5. Analysis



6. Release

20q1

7. Maturity

dev

8. Curator

pinkert

Recommended name:

gdc_tcga_rna_wts_gexp_20q1_dev_pinkert Copy

Description

Position	Field.name	Example	Description
1	source	depmap	Lower-case shorthand name of the source of the data, most ofter as depmap, cbioportal, publication, or the name of an internal uni internally generated
2	dataset	ccle	Lower-case shorthand name of the dataset, often defined by the tcga, mskimpact, pdxdatabase and so forth.
3	analyte	dna	Lower-case shorthand name for the (usually biological) analyte fr measurement. Examples are rna/dna (for sequencing measurem and phe for phenotypic data
4	assay	wgs	Lower-case shorthand name for the experimental assay that has whole-genome sequencing (wgs), whole-slide imaging (wsi), or e repository (cdr)
5	analysis	snv	Lower-case shorthand name for the in-silico analysis that has be nucleotide variant calling (snv), processing of patient dempgraphi based on imaging (cnt)
6	release	20q1	Lower-case, lexicographically sortable release date of the data a formats are foreseen and supported by xopdata yyqq (19q4) and format should be kept constant, at least per dataset, but ideally p
7	maturity	prod	Lower-case term that captures the maturity of the data in the Viev data originating from a data source), dev (for data of intermediary purposes), and prod (for data of production quality, i.e., the highe
8	curator	kreis	Lower-case last name of the person within Merck who curated th role may imply initially ownership of the data within XOP; howeve names are immutable, the owner of a View is instead captured in
9	version	3	Numerical version identifier; this field is automatically added by G generation by incrementing on existing versions of the same view

Traceability
X-OMICS Platform supports data traceability on multiple levels

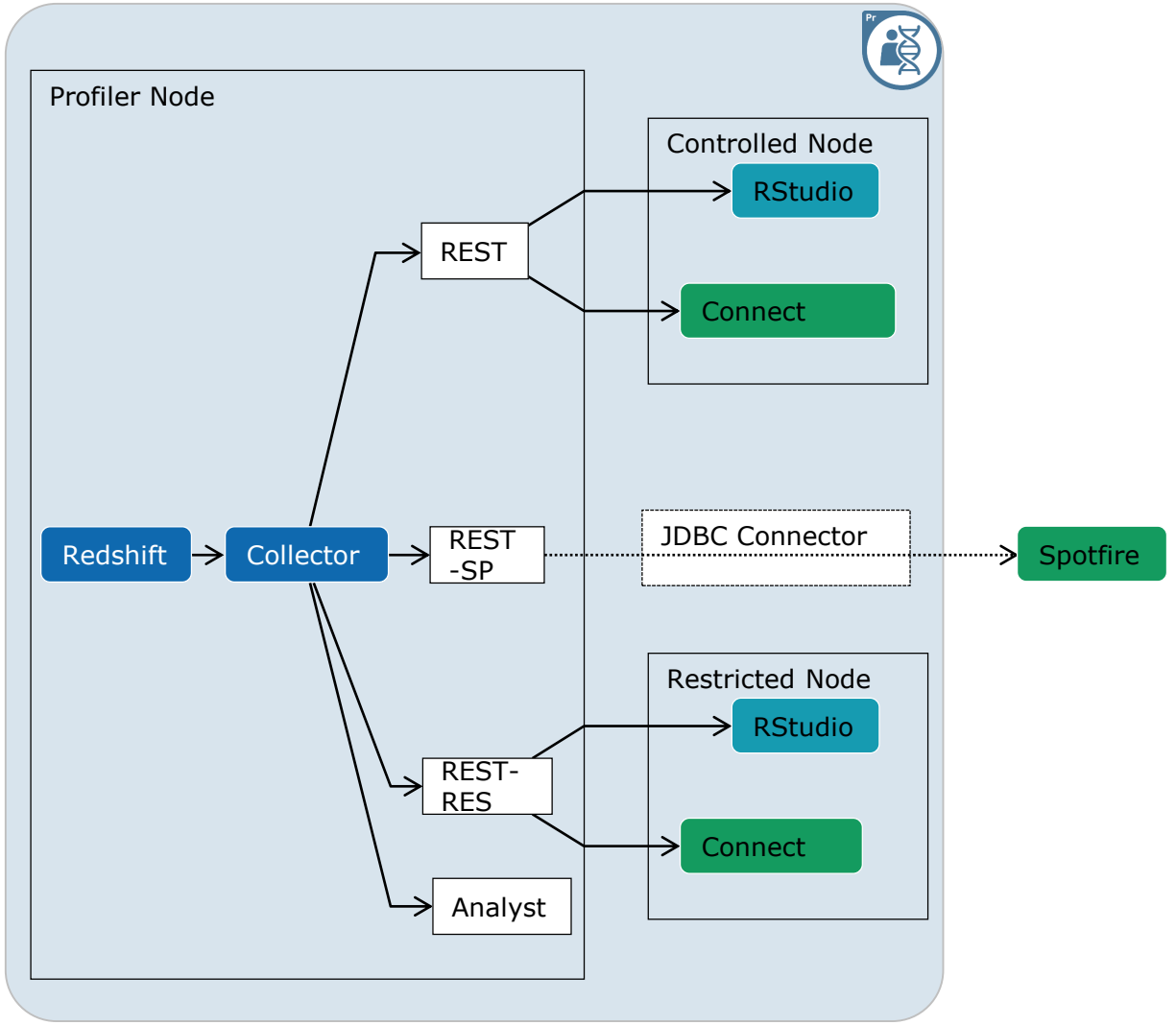
The Analytics DB (Redshift) contains the processed data.

The access from RStudio and RS-Connect is governed by REST interfaces.

The REST interfaces are unique for each requesting service (e.g. restricted RStudio) and controlled by Permission Tags.

Additionally access to data is restricted by the user having the appropriate role in the study containing the data.

Each data access is noted in the audit trail.



It’s a teams effort!
Acknowledgments