BEACON's Data Management Plan

David M. Bryson, Ian Dworkin, and Matt Rupp

Why plan for Data Management?

- Consider data handling, metadata, preservation, and analysis before the project begins
  - Ensures data are correct format, organized, and better annotated
  - Standards enable collaboration efficiency
  - Future submission to a database easier
  - Facilitates future data re-use

BEACON's Plan

- Three major sections:
  - Philosophy
  - Policy
  - Data Handling Guidelines

Data Management Plan

Philosophy

BEACON finds and facilitates a wealth of research and educational activities that produce information and data valuable to the scientific community. BEACON is committed to promoting, educating, and facilitating the responsible preservation and open access of these data. At the same time, we recognize the importance of preserving the professional interests of students, postdoctoral researchers, and faculty, who have worked to fund, organize, and complete complex projects and experiments and must be given reasonable opportunity to analyze and publish results before data are made available to the community at large. The BEACON Data Management Policy strives to balance these competing interests and to provide guidelines for handling the many disparate types of data.

Policy

The BEACON Data Management Policy applies to all researchers supported by BEACON funding, facilities, and resources. The word "researcher" is used to refer to all scientists, students, engineers, educators, staff, and any other persons who develop data, analyses, and/or ideas associated with BEACON projects and research. Research data and information are to be handled and made available in a manner consistent with the guidelines detailed below. Public release of data should be as soon as possible following the first publication based upon or utilizing those data. Other data should be released in a reasonable time frame following development.

Internal BEACON project budget requests should include statements regarding the specifics of how project data will be managed, including details for handling the privacy, confidentiality, security, and intellectual property of research products and their intended distribution at project completion. Project reporting must include statements detailing adherence to that plan and this policy. Adherence to these requirements will be overseen by the BEACON Managing Director.

Data Handling Guidelines

BEACON projects produce and make use of **five general types of data**: *collected data sets and experimental results* generated by research; the *configuration files* and scripts used to run computational analysis and experiments; the *source code* of software developed to enable those experiments; *information resources*, such as curriculum materials and documentation; and *administrative materials* such as policies or other reported information. Each type requires special care to ensure long term preservation in a useful and functional state. Training, being necessary to ensure compliance and facilitate successful implementation of the data management guidelines, will be provided during appropriate BEACON activities, such as the annual congress and weekly seminars. Researchers must make reasonable efforts to ensure that data subject to these disclosure guidelines is preserved prior to release.

Since each individual research and administrative community within BEACON has its own community standards for data storage and accessibility, the goal of this plan is to provide both flexibility and guidance for its members. Data released in accordance with this plan and associated guidelines must be available for a minimum of 3 years following release. This plan is intended to cover all researchers funded through BEACON as well as technical and administrative employees. Material that is proprietary, under copyright protection, patented, or confidential is not covered under this plan.

These guidelines are not intended to limit the scope of data that should be disclosed.

- **Collected Data Sets and Experimental Results**

  Research data and analysis material necessary to replicate, support, and validate research findings must be published to an appropriate third party archiver such as Dryad, GeneBank, or other

1 of 2

institutionally supported repositories. Data collected from experiments involving human subjects must be made anonymous and handled with appropriate care and in compliance with regulations. When possible, raw data should be published; otherwise the lowest possible level of aggregated data along with the material required to fully replicate experiments may be substituted. Data file formats should be suitable for long-term archiving, such as human readable formats (e.g. txt, csv, xml), well documented open source formats (e.g. hdf5), or where appropriate, industry standard binary formats (e.g. pdf). BEACON strongly encourages researchers to use open standard file formats where practical. Metadata including the name(s) and contact(s) of the creator(s) and PI(s) along with a description of contents and purpose should be included as part of the publication record of all materials.

- **Configuration Files**

  Complete configuration settings used in published experiments and analysis pipelines should be made available alongside archived data. Detailed setting listings and configuration files must clearly identify the version(s) of the software with which they were designed to be used, and provide basic instructions for their use as appropriate. Non-trivial experiment and analysis scripts should provide documentation and/or comments that explain their operation.

- **Source Code**

  The full source code of software developed or modified in the course of BEACON projects should be made publicly available under an OSI-approved open source license, unless existing license or intellectual property protections prevent such distribution. BEACON projects are encouraged to utilize public version control repository hosting services, such as GitHub and SourceForge to host the source code. Developed software should include at least the minimum documentation to build and execute it, along with metadata about authorship, purpose, and licensing.

- **Information Resources**

  Education and outreach materials produced and vetted during BEACON-funded projects should be made available to the public via the BEACON website. The materials should include descriptions of their intended use and links to relevant studies.

- **Administrative Materials**

  Administrative documents, including official reports, should be posted to either the center's website or wiki. Other administrative data should be kept in secure locations. Data used to generate administrative reports are to be maintained in a manner identical to research data but may be kept in a private location.

- **Other Materials**

  Other materials, such as samples, apparatuses, and other materials not explicitly mentioned above that are necessary for the reproduction of the research must be handled according to the standards of the field. Specific plans regarding such materials must be detailed in project budget requests and adherence to those plans detailed in project reports.

2 of 2

Philosophy

- BEACON's DMP strives to lay out a 'Meta' plan
- Two Primary Goals:
  - Promote responsible data handling and preservation
    - Open access, repeatability, and reuse
  - Preserve professional interests

Policy

- Applies to all individual and groups supported by BEACON funding, facilities, and resources
- Data should be handled and release according to guidelines
  - Must be made publicly available as soon as practical following first publication based upon those data
- Future budget requests must include project specific DMPs
- Project reports must include details about DMP adherence

Guidelines:
Collected Data Sets and Results

- Must be posted to a public archive
- Long-term archive formats
  - Human Readable (text)
  - Open Source
  - Industry Standard
- Metadata
  - Creator and PI details
  - Contents and Purpose

Icon Credit: Barry Mieny (http://barrymieny.deviantart.com)

Guidelines: Configuration Files

- Full, detailed settings of tools and protocols used in experiments and analysis
- Version information of software
- Basic instructions for use
- Documentation and comments explaining non-trivial aspects configuration and scripts

Guidelines: Source Code

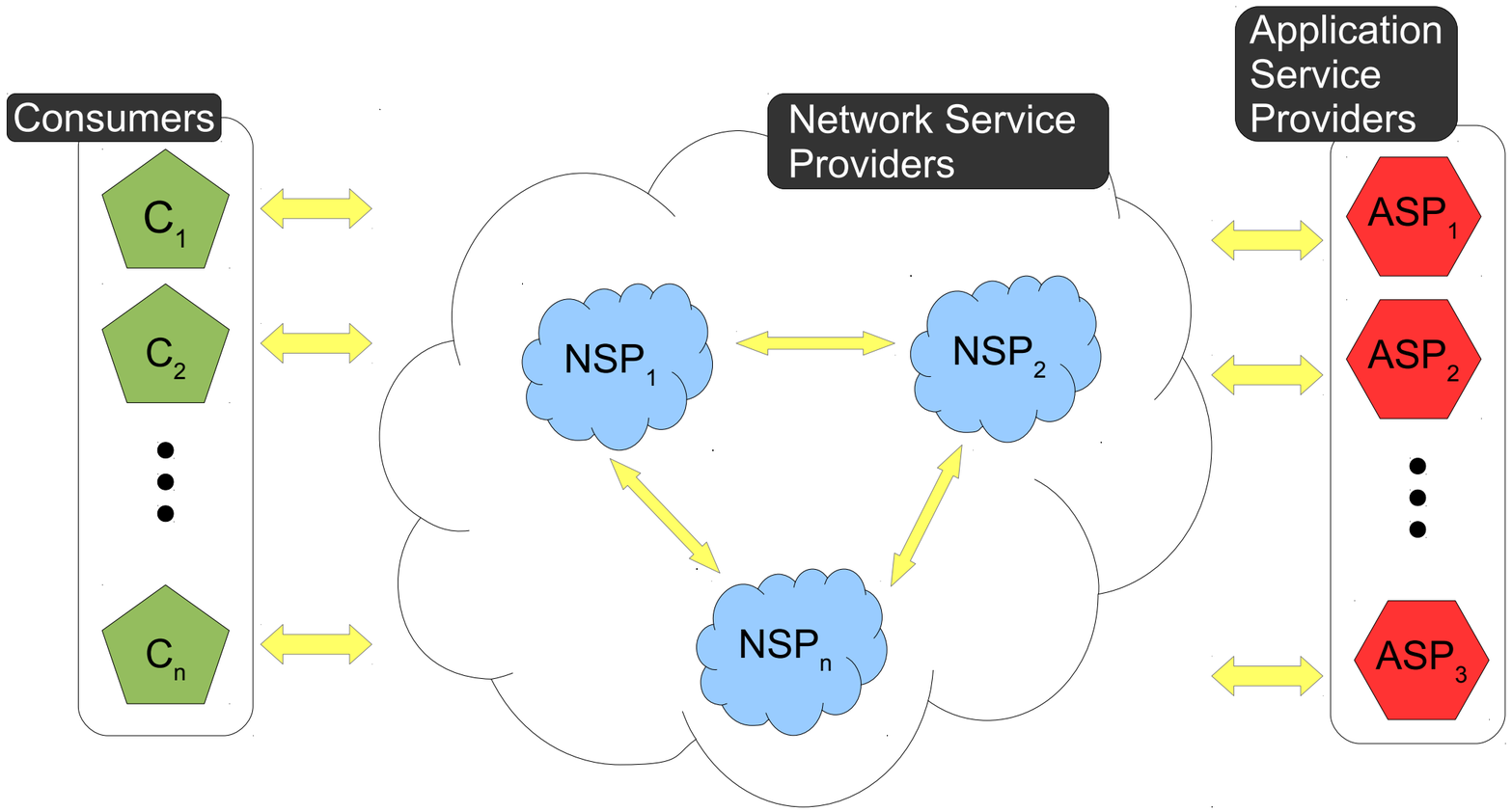
- BEACON developed code and software projects should be open source
- Public source code repositories encouraged
- Include basic instructions for building and use
- Metadata about authorship, purpose, and licensing

Guidelines: Other Resources and Materials

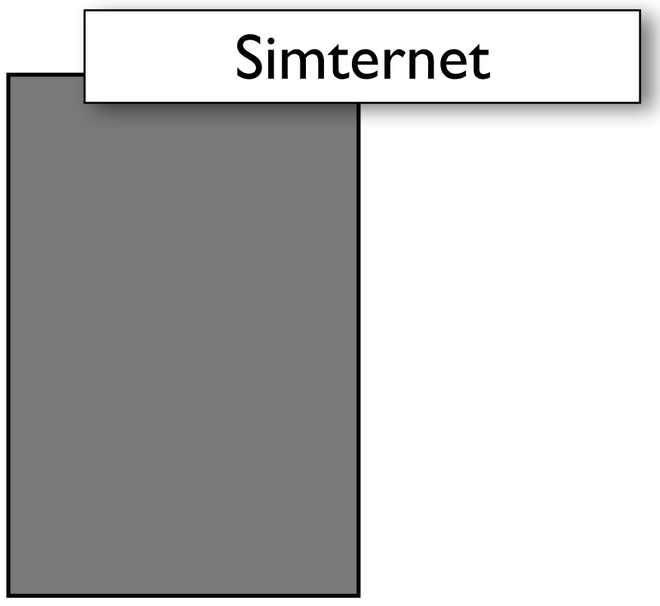
- Education, outreach, and administrative information should be posted to the BEACON website
- Other materials: refer to standards of the field
  - Must include details of those standards and handling plans in budget requests

An Example Project

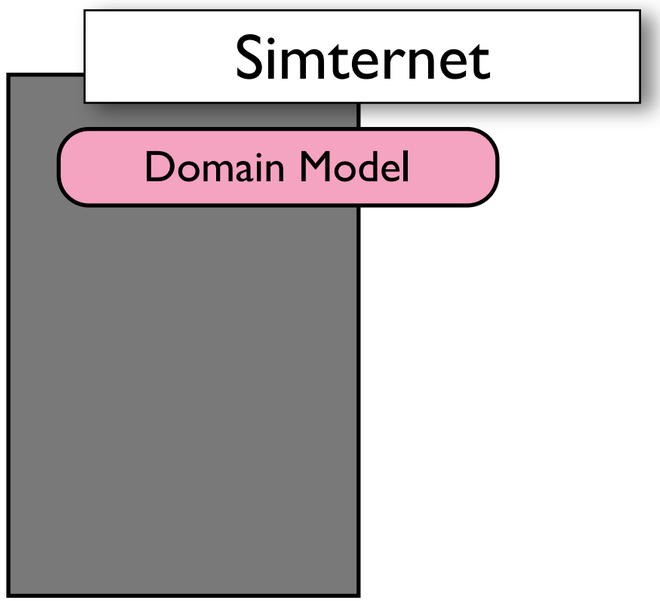
An Example Project



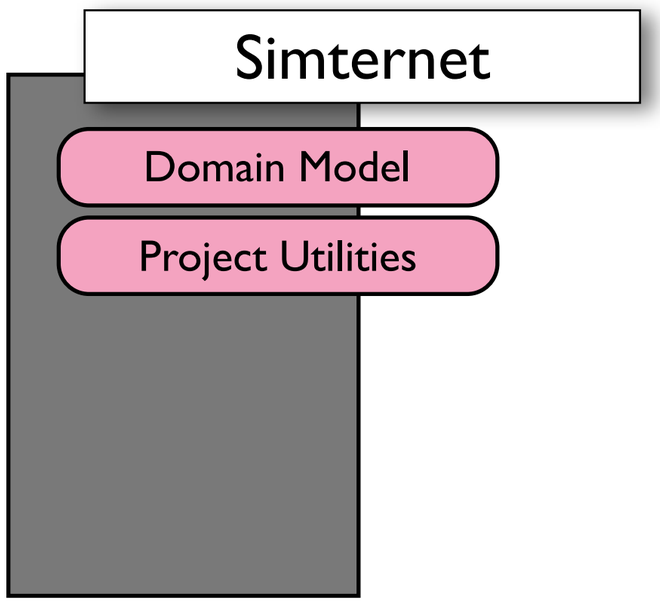
Project Data



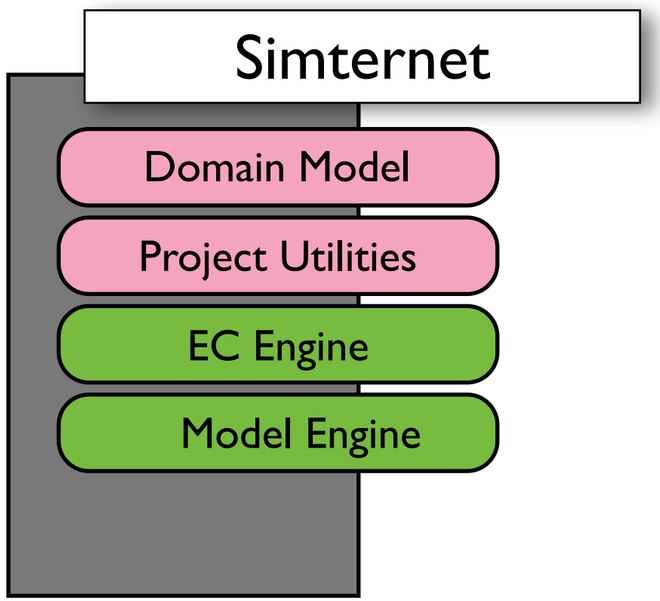
Project Data



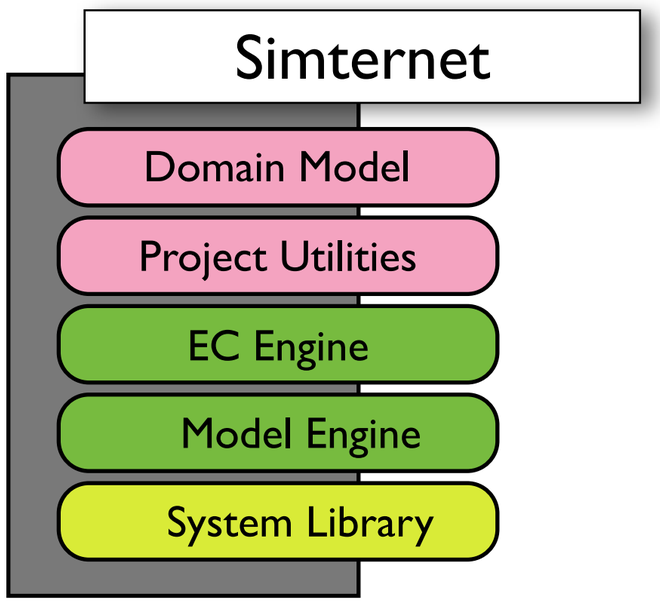
Project Data



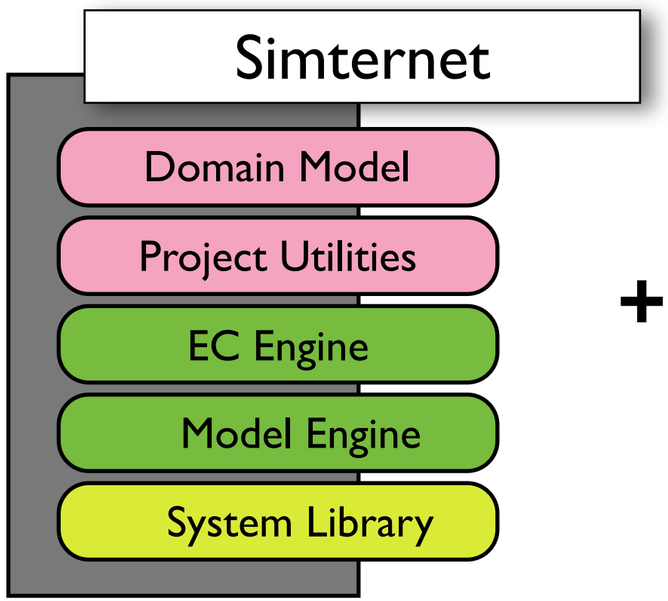
Project Data



Project Data

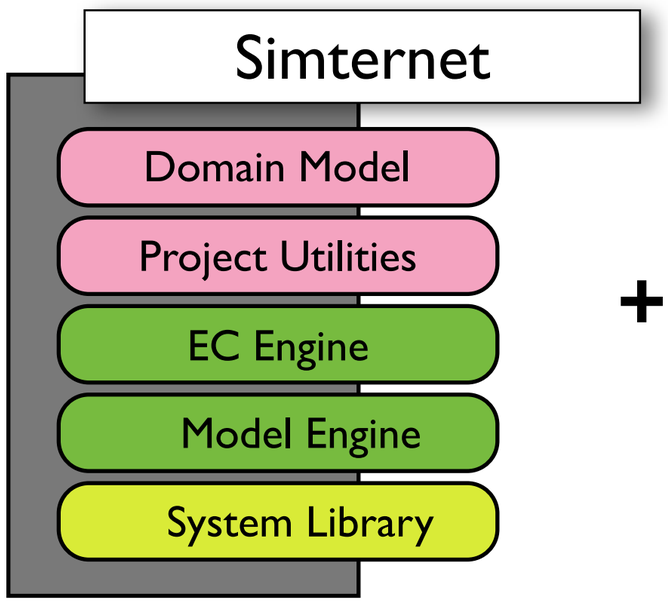


Project Data

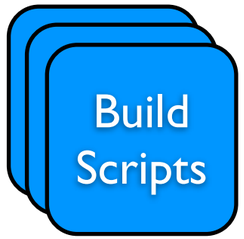


+

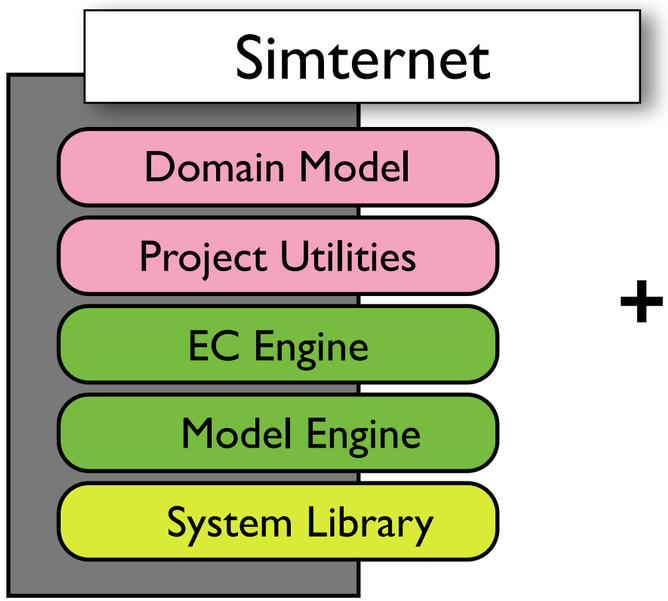
Project Data



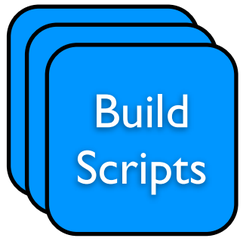
+



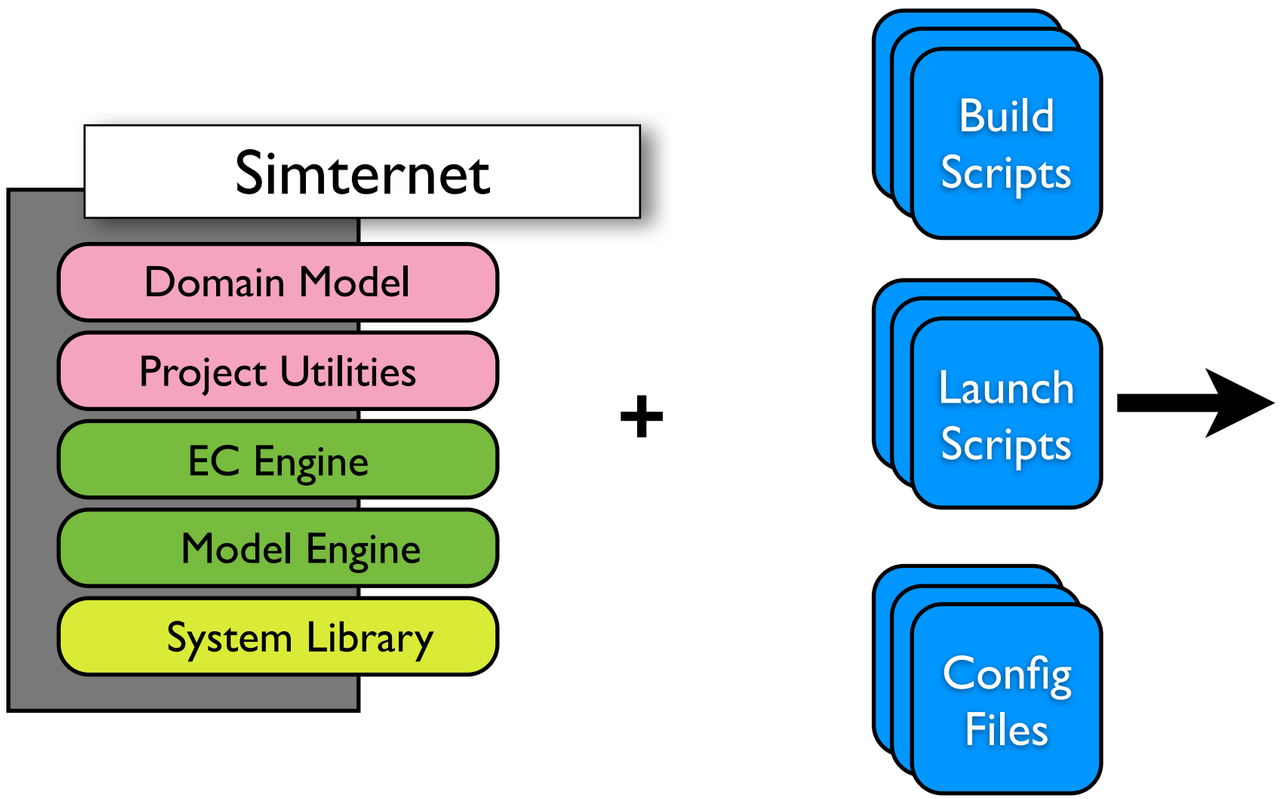
Project Data



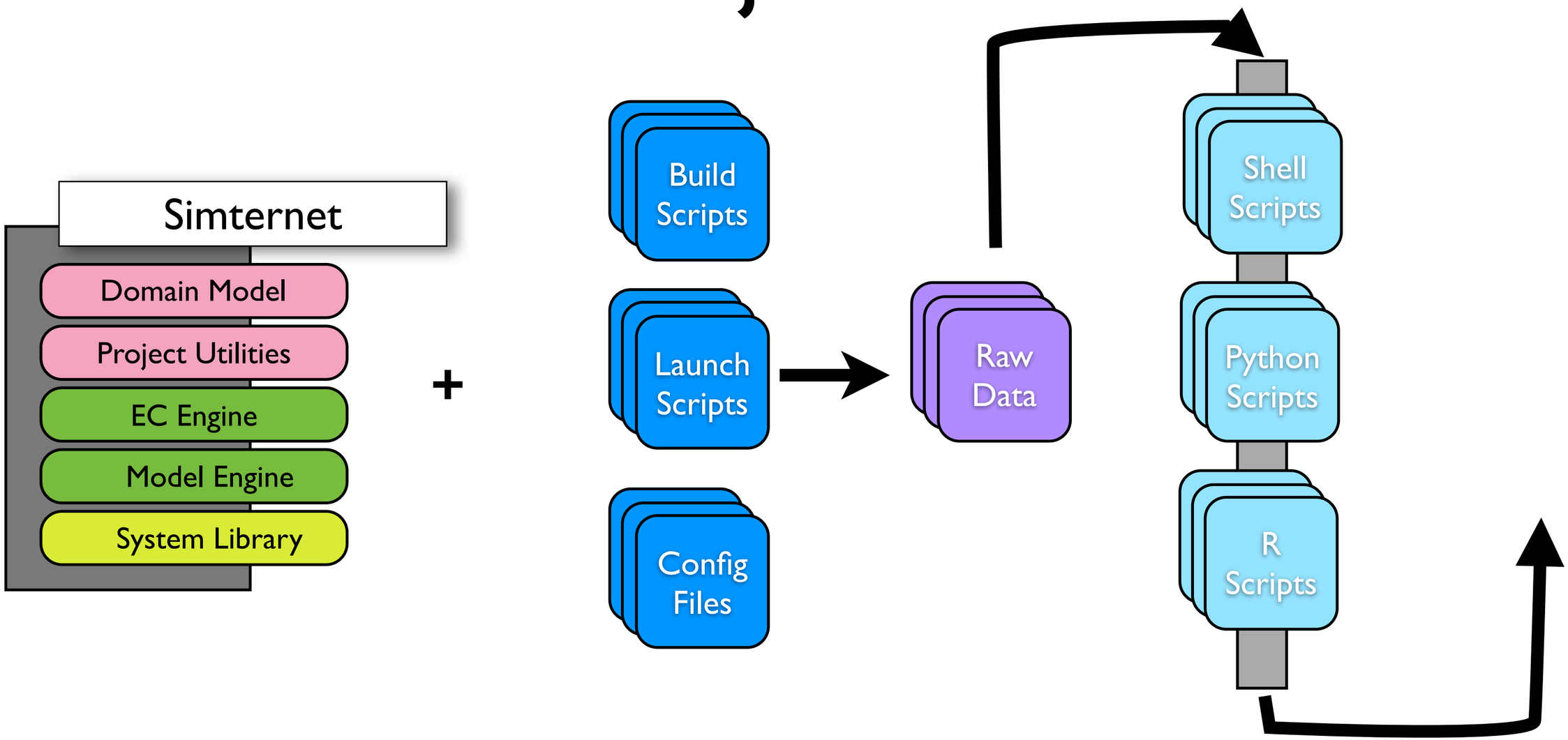
+



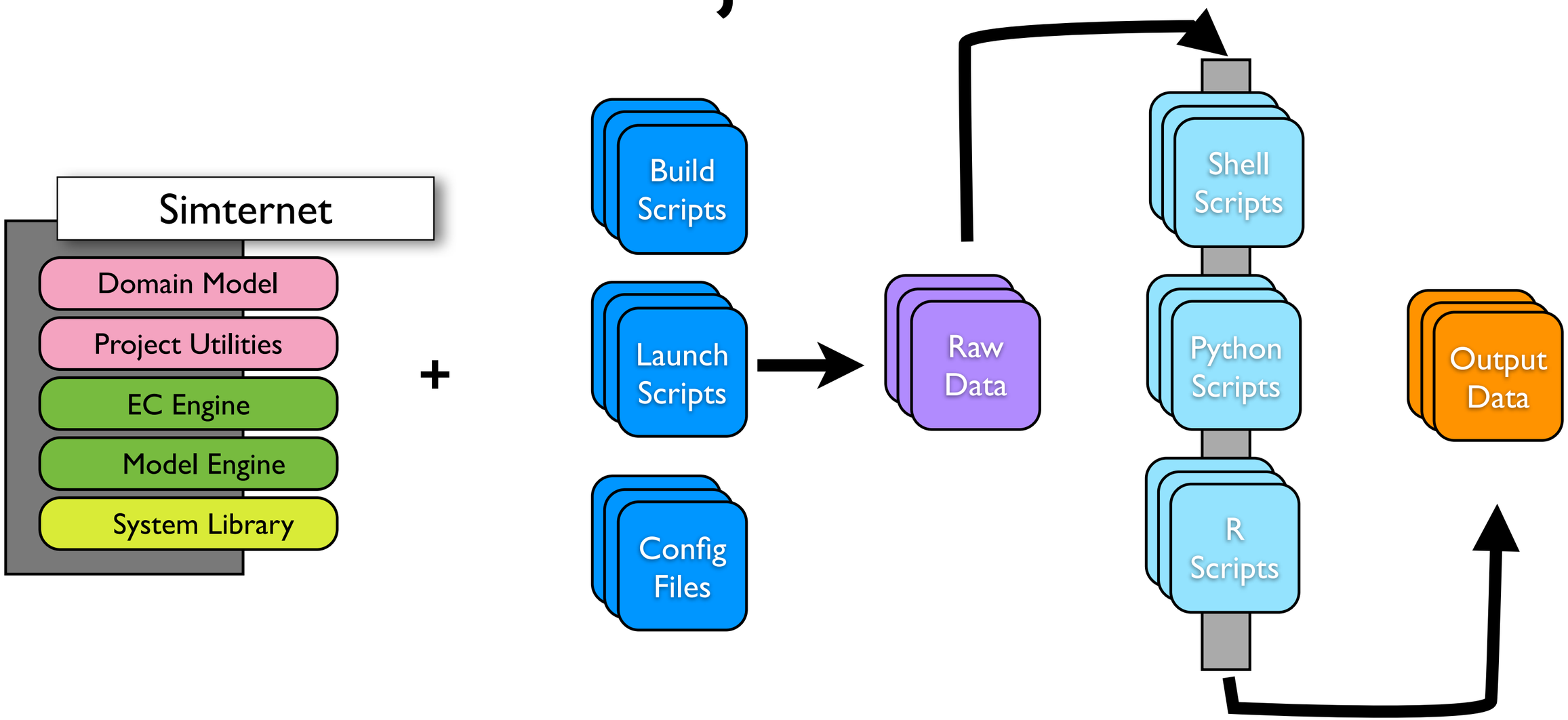
Project Data



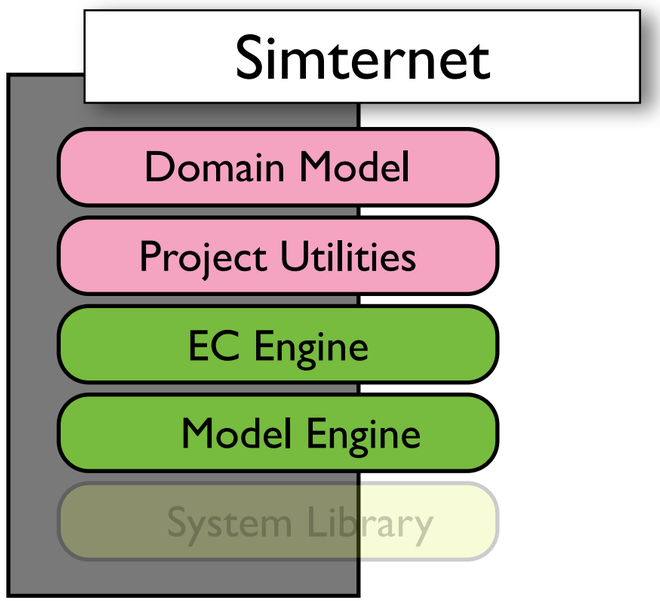
Project Data



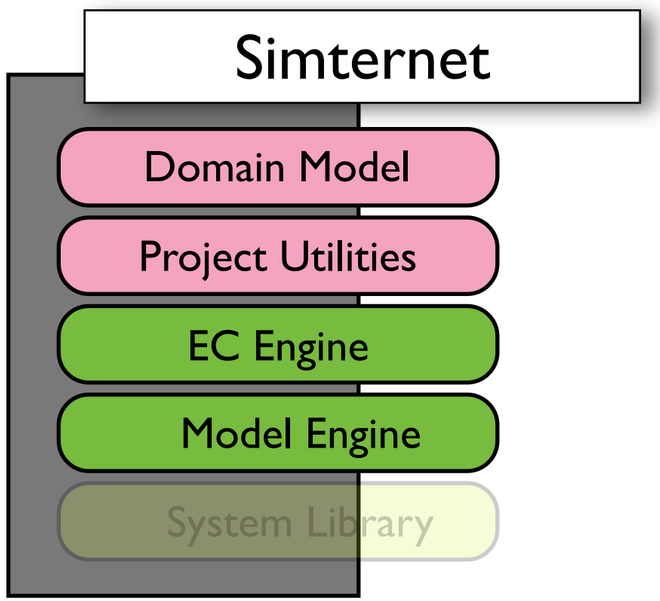
Project Data



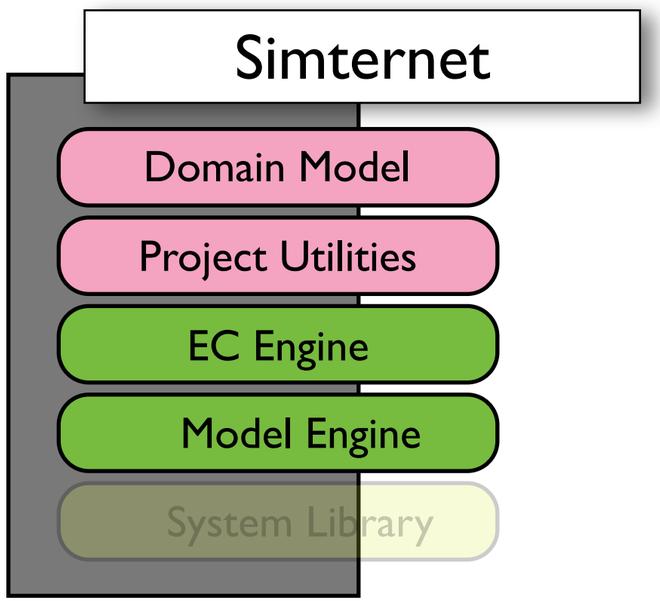
Source Code Data



Source Code Data



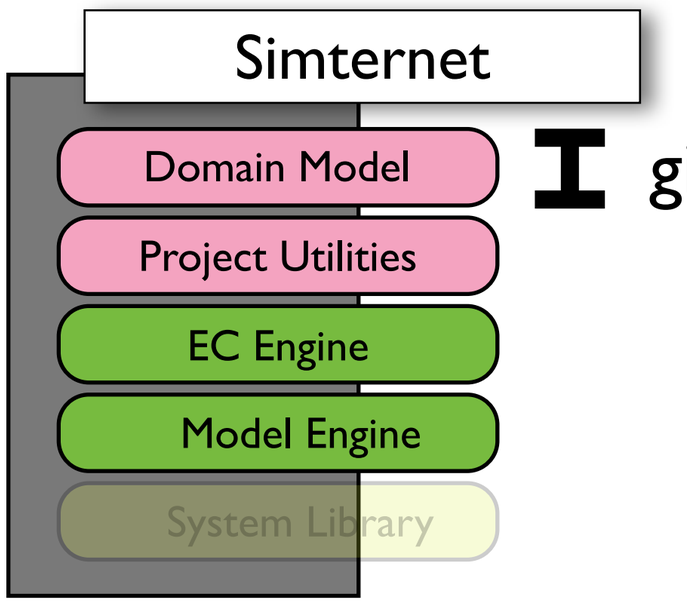
Source Code Data



git

GitHub

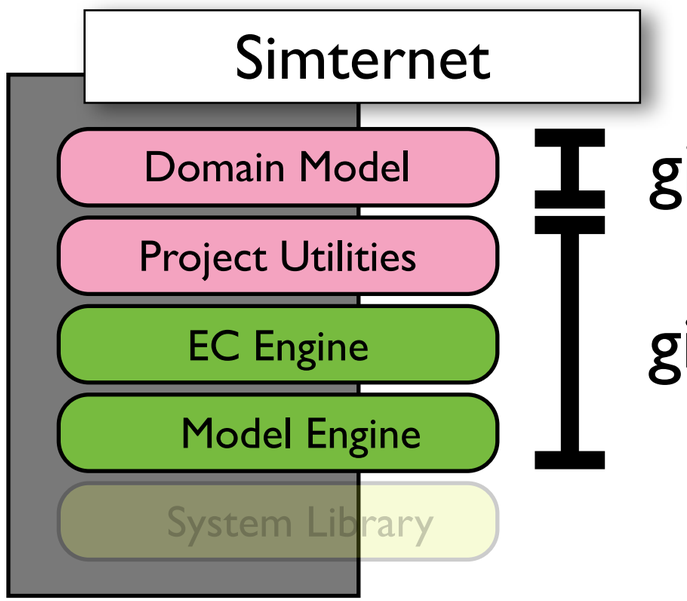
Source Code Data



git://github.com/kkoning/Simternet

GitHub

Source Code Data

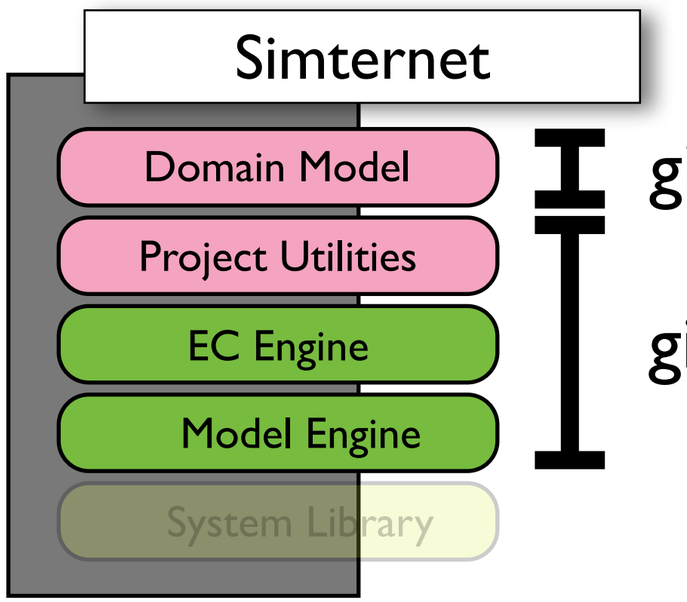


git://github.com/kkoning/Simternet

git://github.com/kkoning/Agency

GitHub

Source Code Data

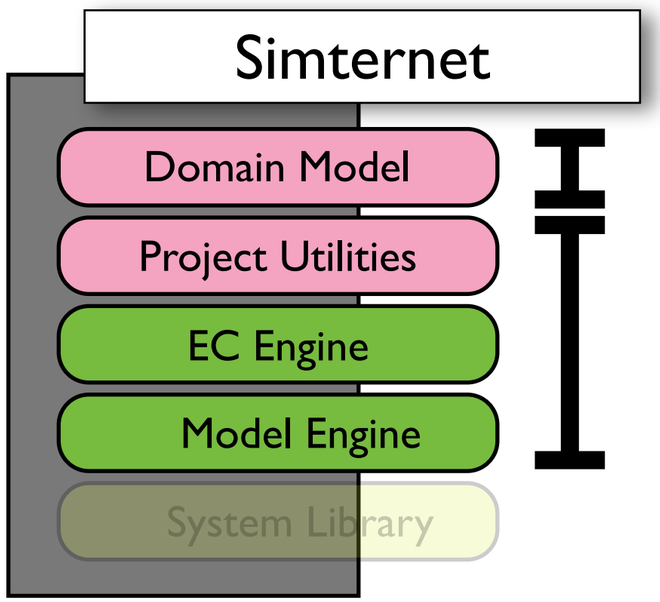


git://github.com/kkoning/Simternet

git://github.com/kkoning/Agency

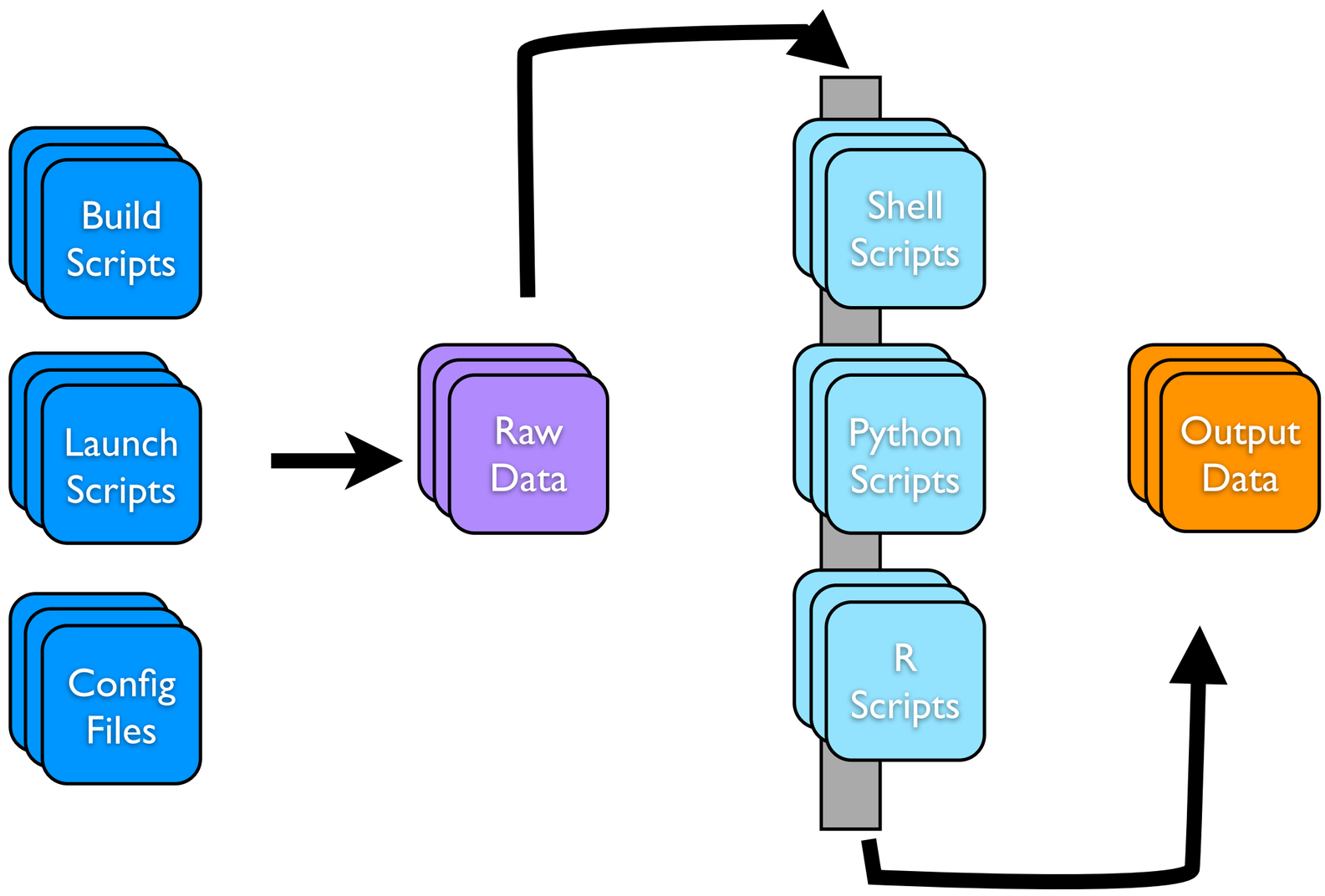
GitHub

Source Code Meta Data

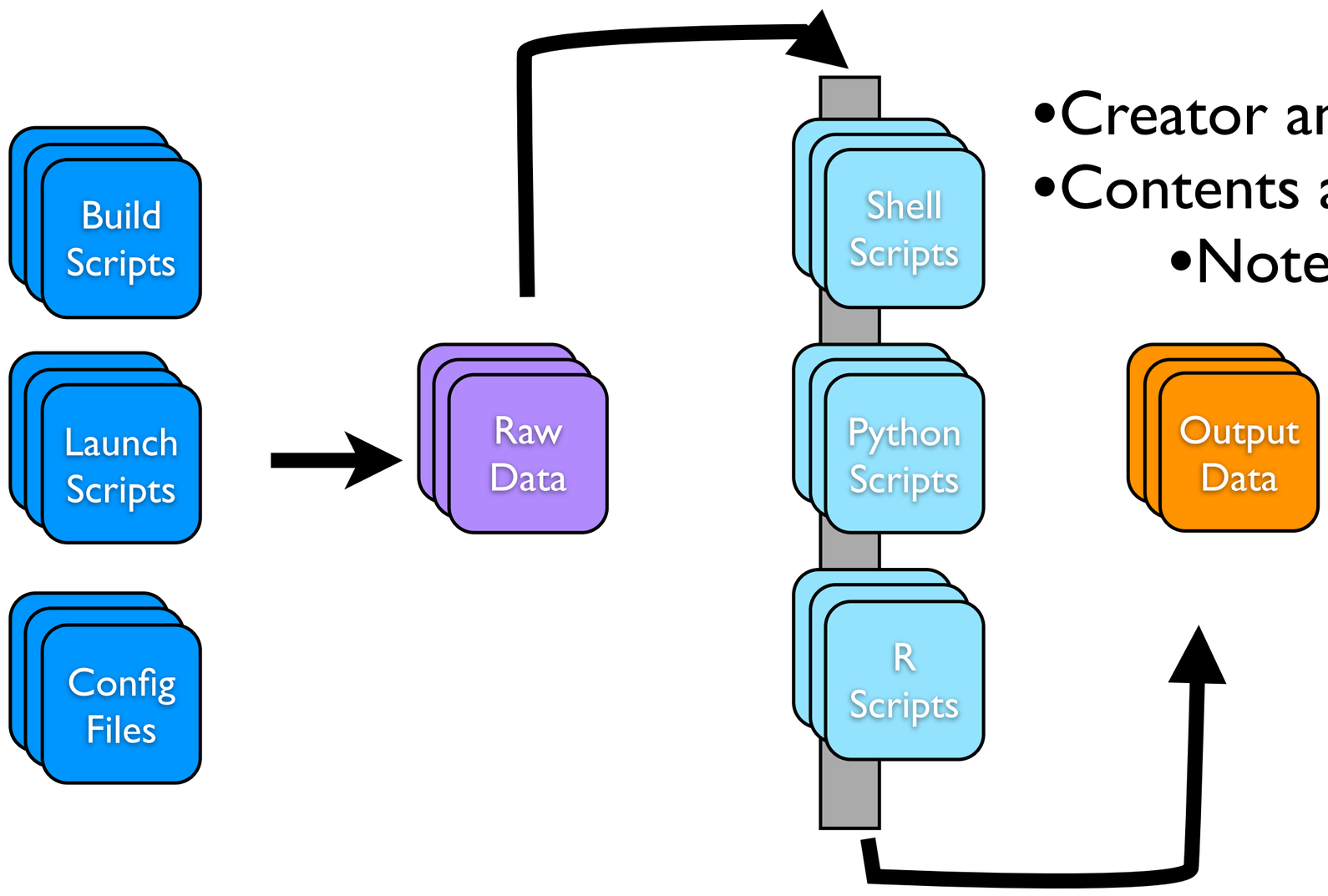


- Authors
- Purpose
- Licensing
- Instructions for use

Project Data

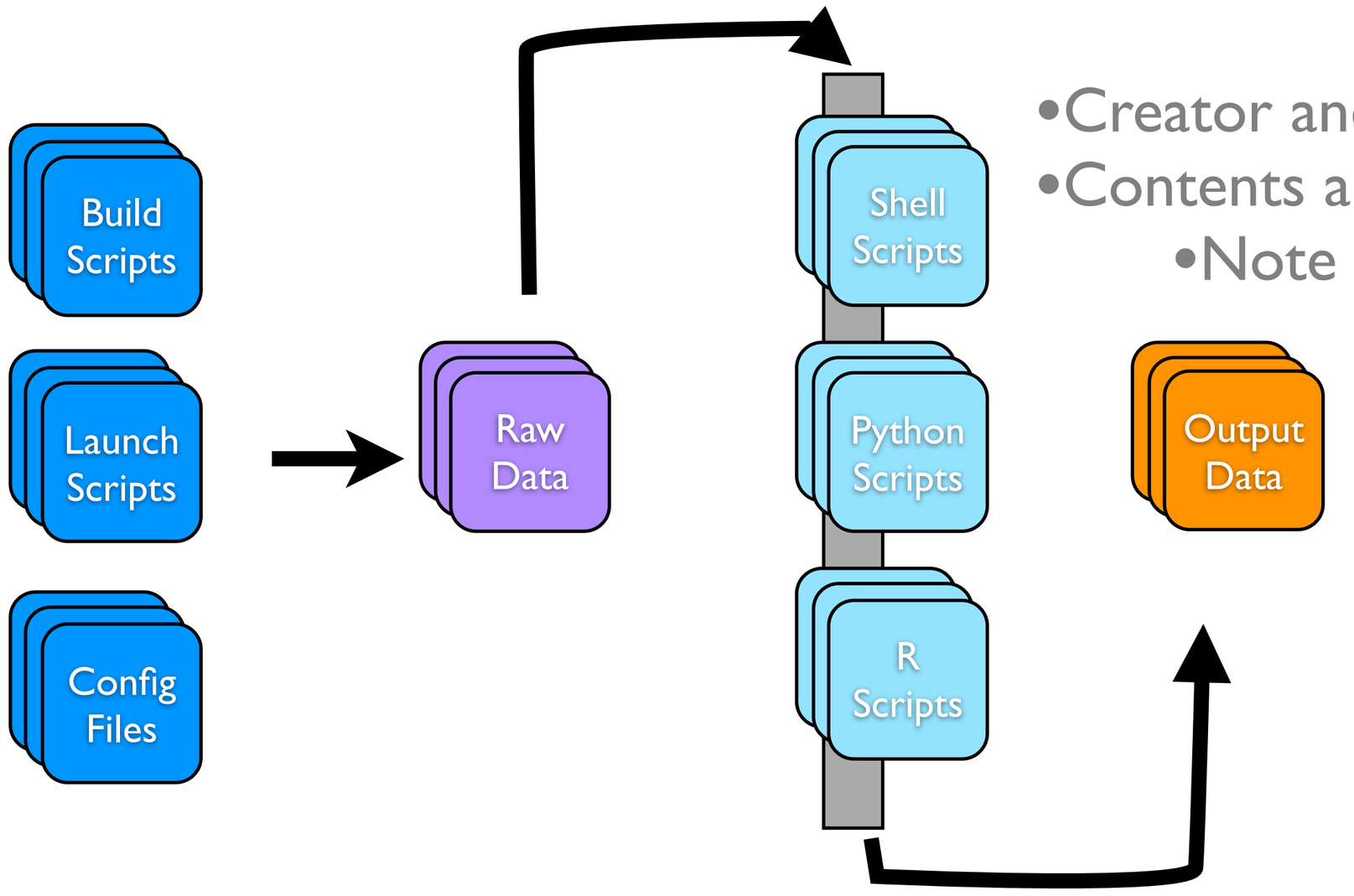


Project Meta Data



- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

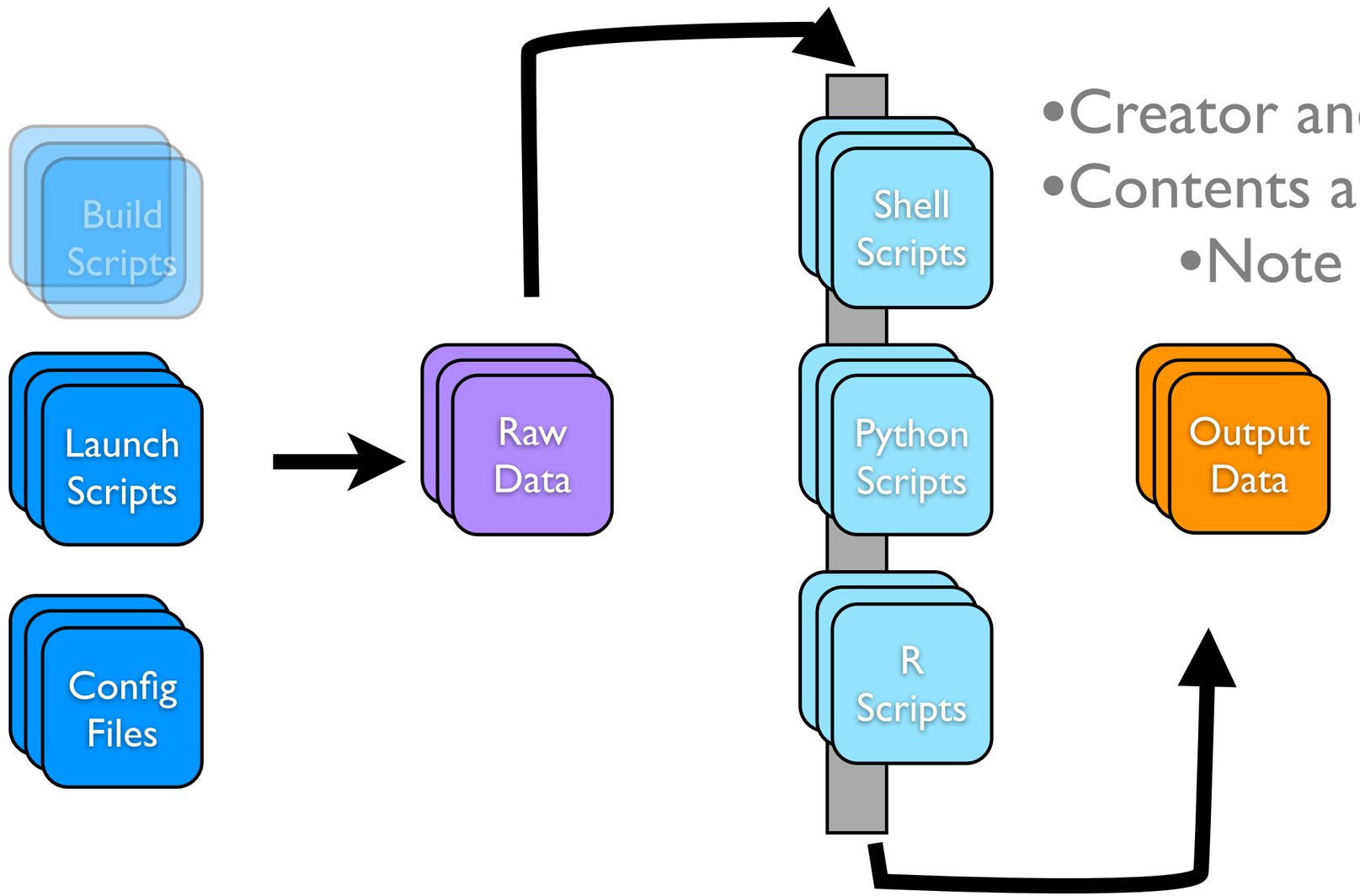
Project Data



- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

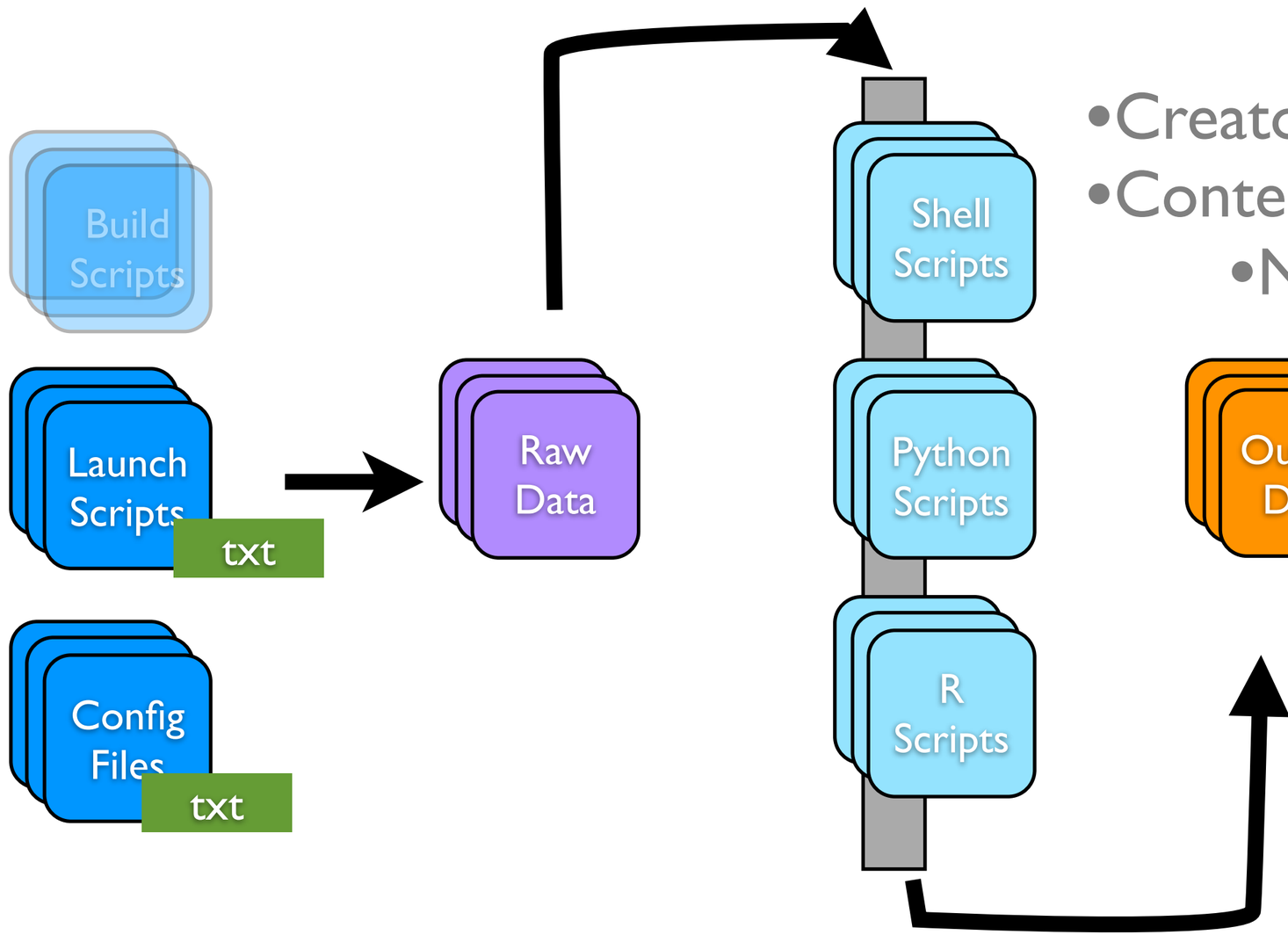
DRYAD

Project Data



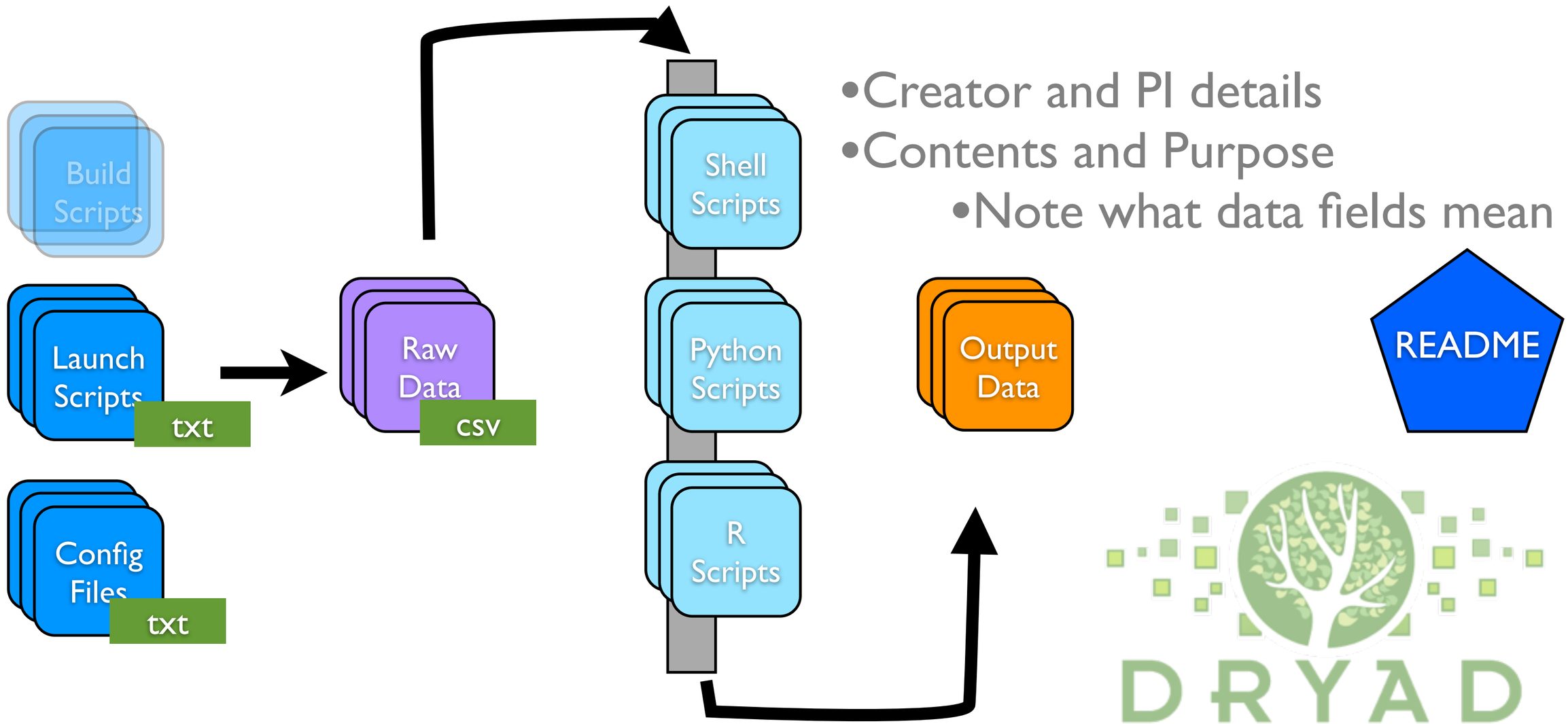
- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

Project Data



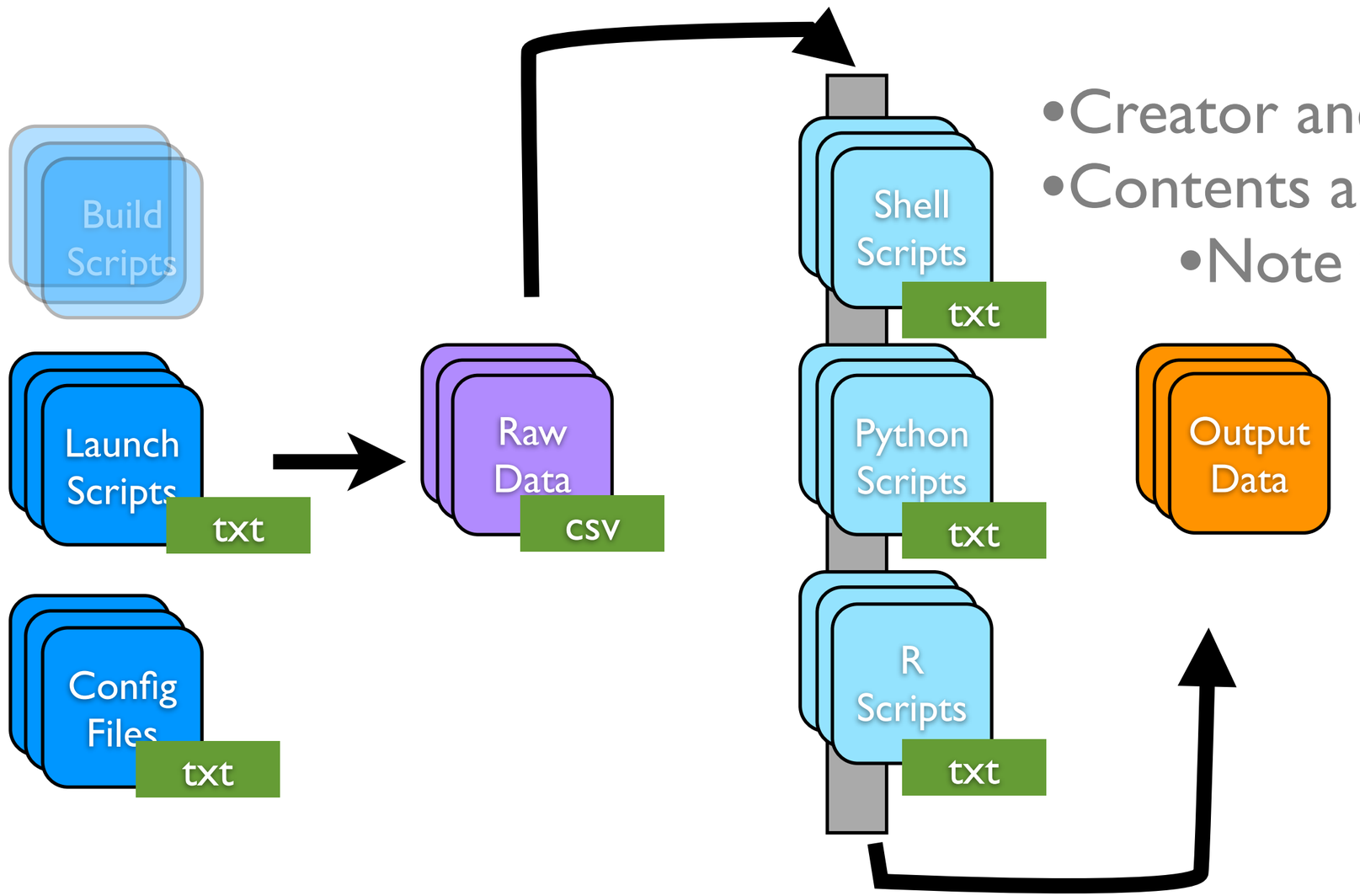
- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

Project Data



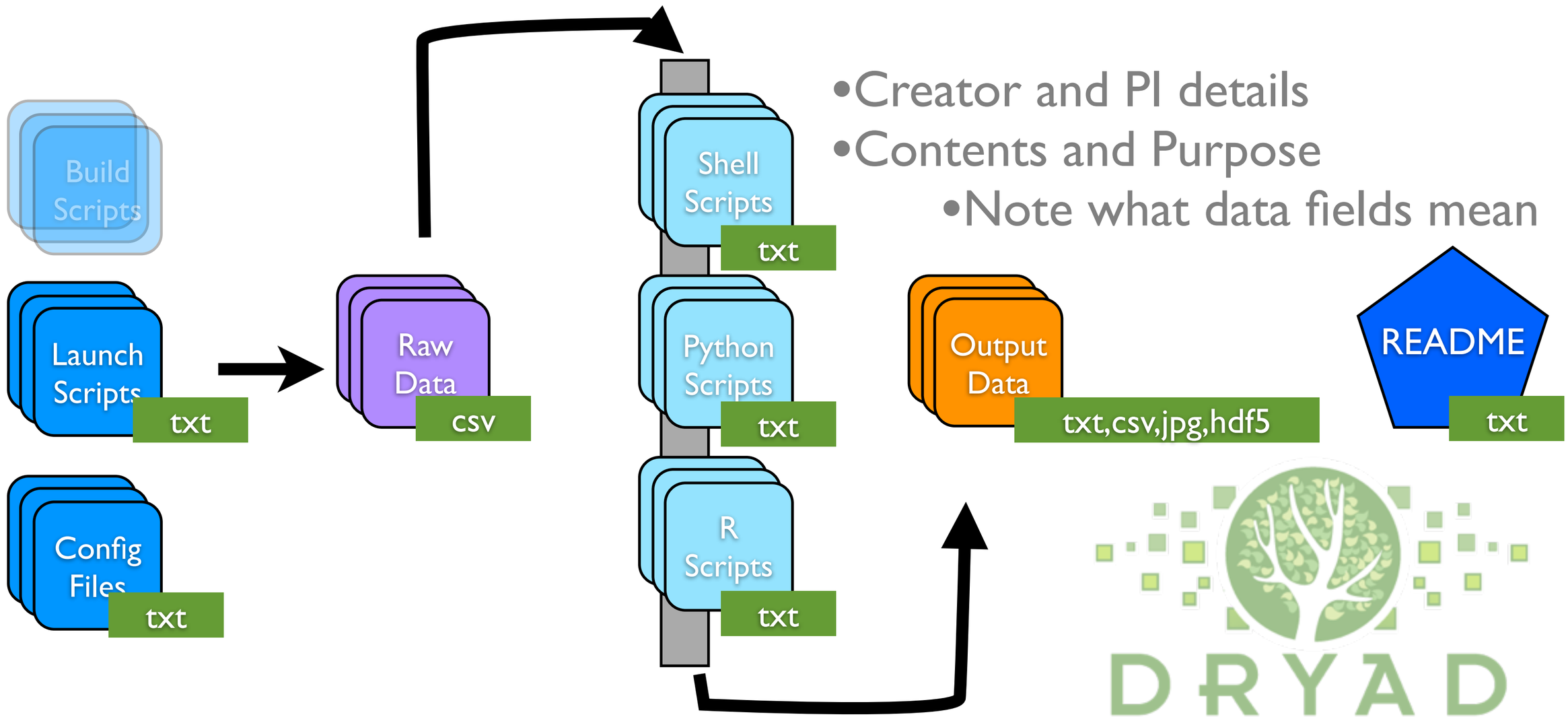
- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

Project Data



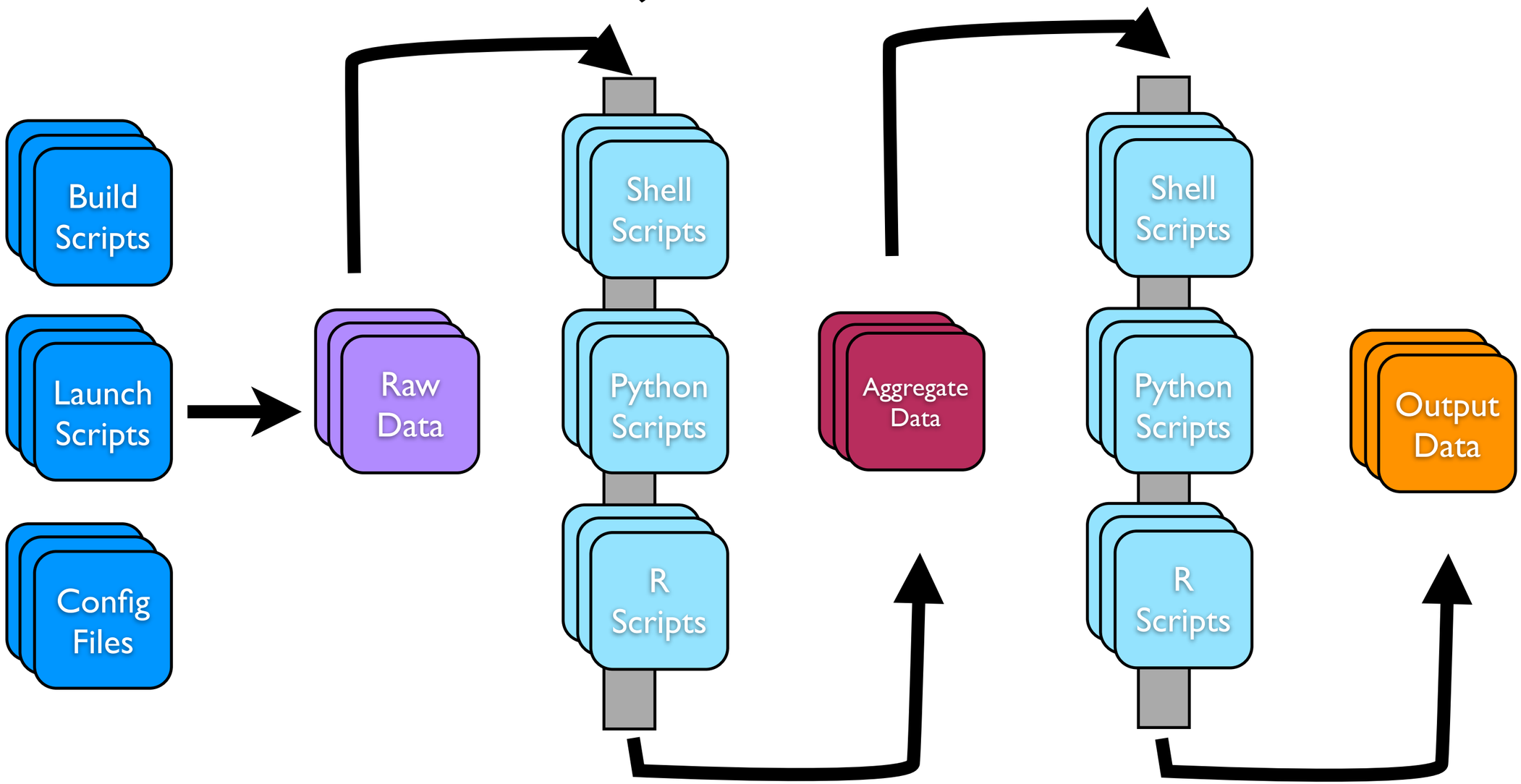
- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

Project Data

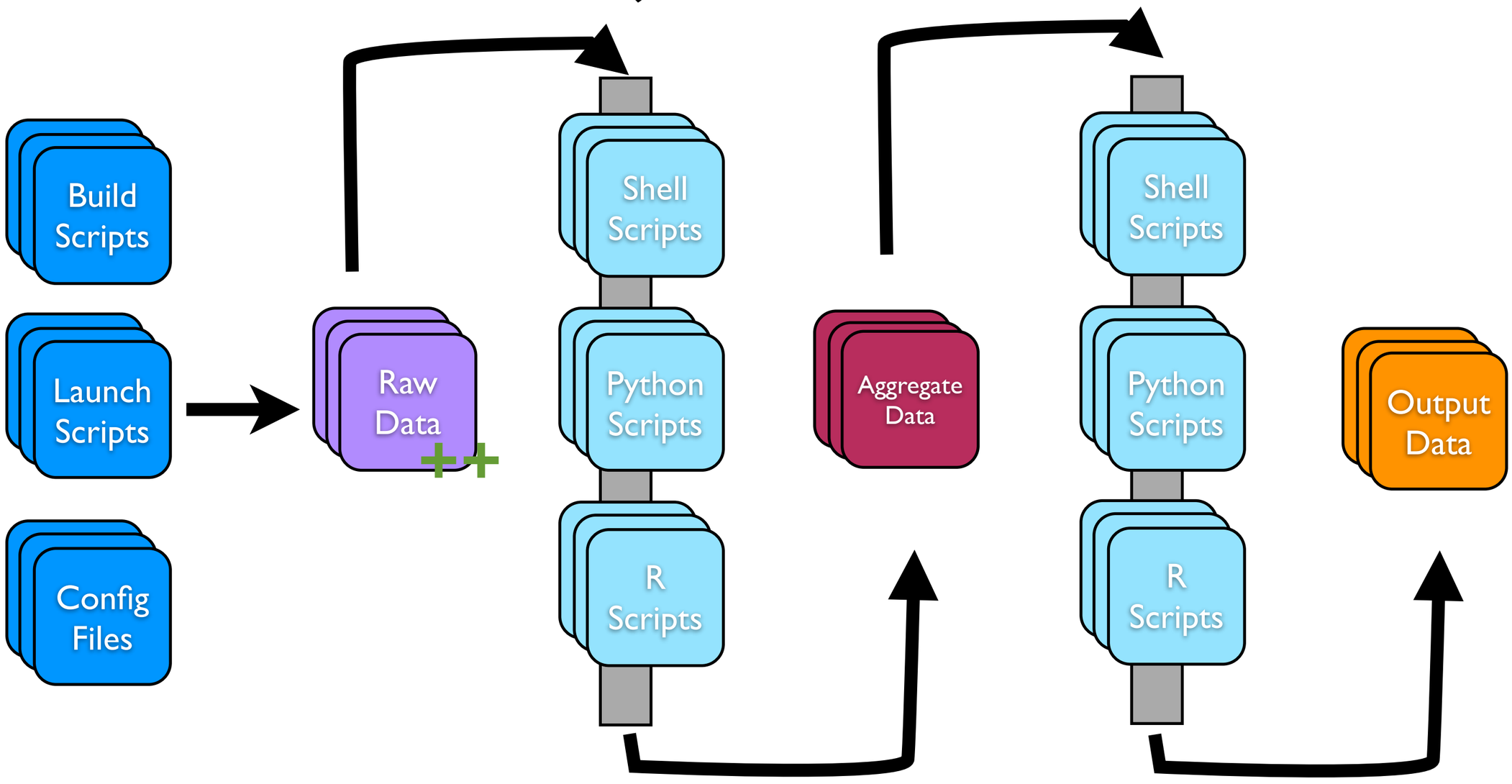


- Creator and PI details
- Contents and Purpose
  - Note what data fields mean

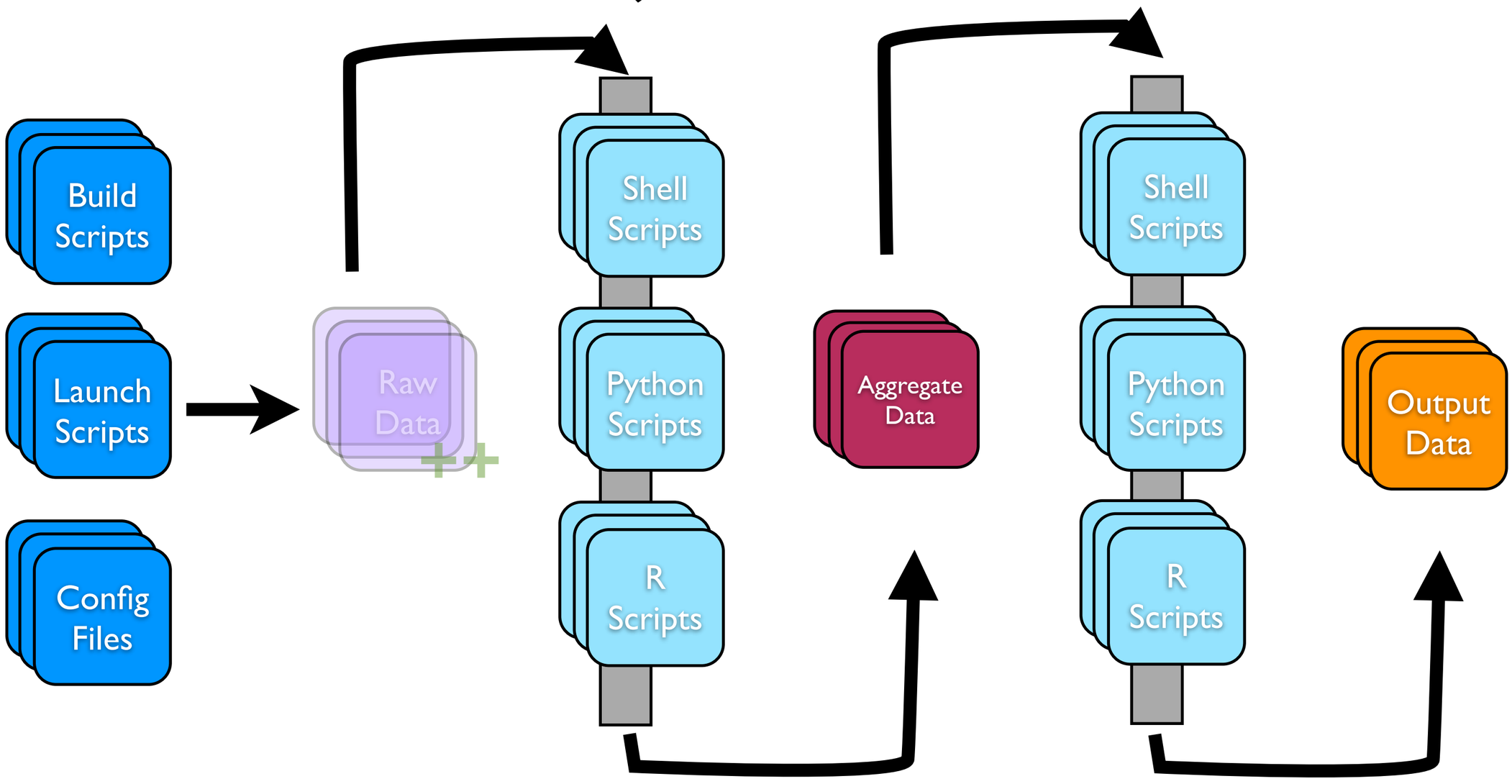
Project Data



Project Data



Project Data



Project Metadata

Project Metadata

- Comments about the experiment

Project Metadata

- Comments about the experiment
- PI and Author information

Project Metadata

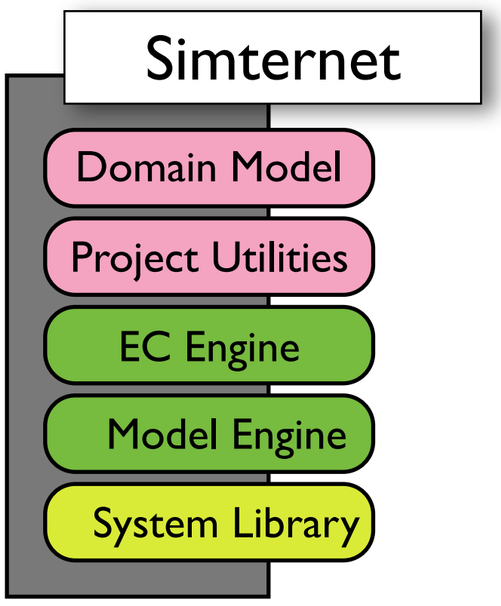
- Comments about the experiment
- PI and Author information
- Directions for use

Project Metadata

- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information

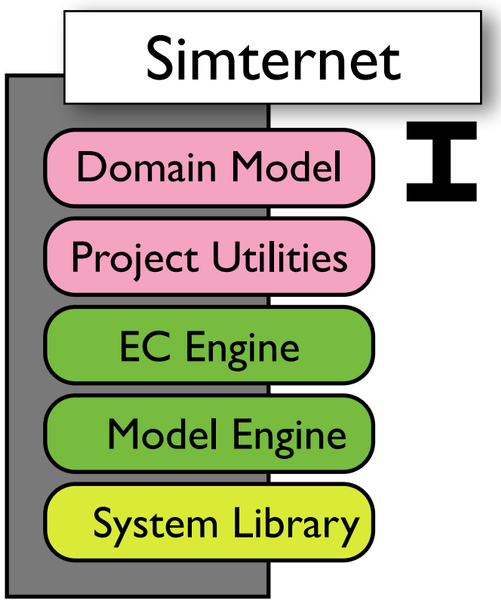
Project Metadata

- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information



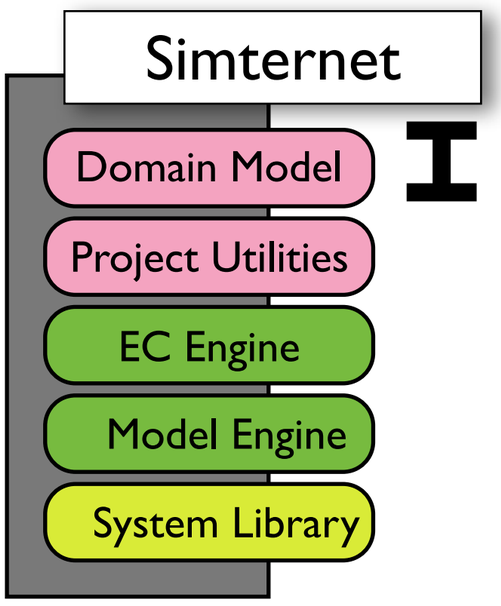
Project Metadata

- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information



Project Metadata

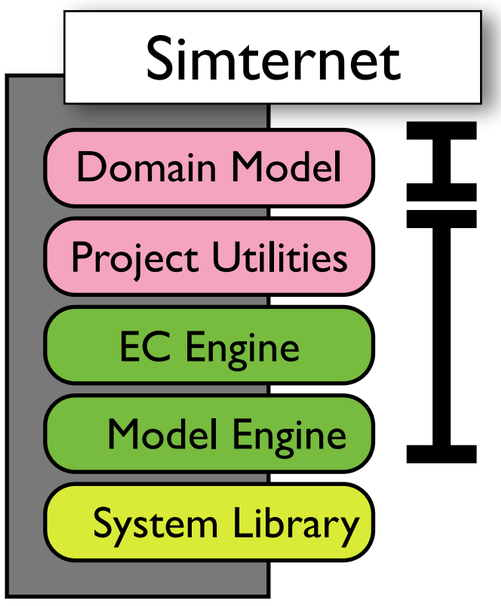
- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information



commit id: `7a62665e5411bfed1e0761708b1a6b6aa6746c75`

Project Metadata

- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information



commit id: `7a62665e5411bfed1e0761708b1a6b6aa6746c75`

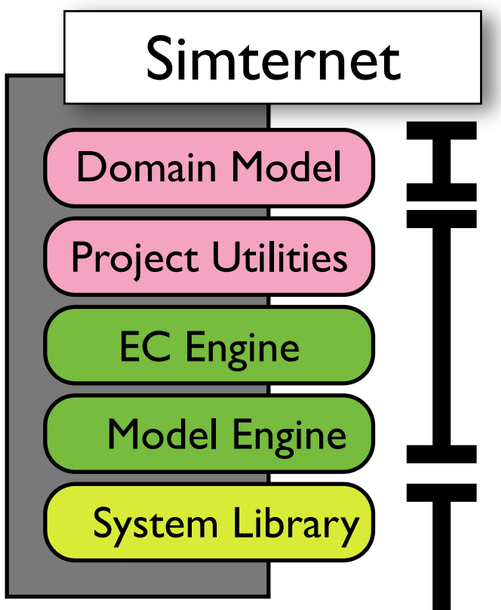
commit id: `f0fba2dbbff2cdd000eb7fa653b2b5194a70f041`

ECJ: `version 21`

MASON: `version 16`

Project Metadata

- Comments about the experiment
- PI and Author information
- Directions for use
- Software Version information



commit id: `7a62665e5411bfed1e0761708b1a6b6aa6746c75`

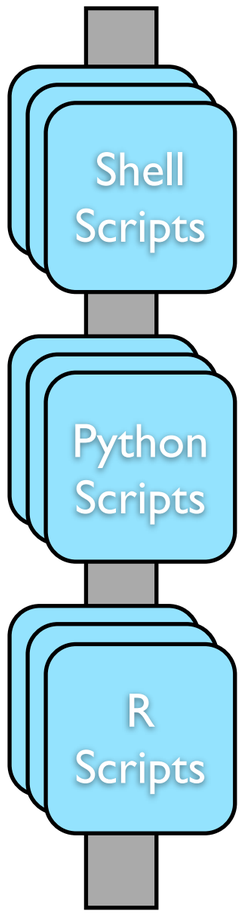
commit id: `f0fba2dbbff2cdd000eb7fa653b2b5194a70f041`

ECJ: `version 21`

MASON: `version 16`

```
java version "1.6.0_24"
OpenJDK Runtime Environment (IcedTea6 1.11.4) (rhel-1.49.1.11.4.el6_3-x86_64)
OpenJDK 64-Bit Server VM (build 20.0-b12, mixed mode)
```

Analysis Scripts

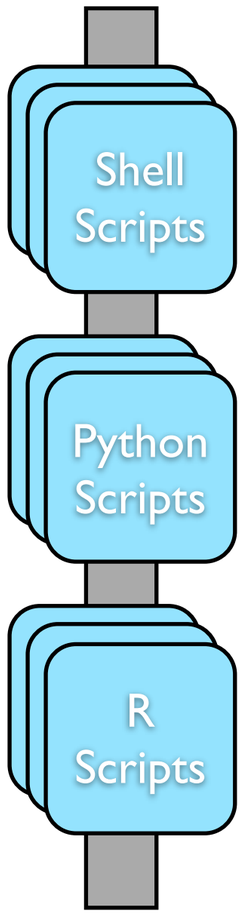


- Creator and PI details
- Contents and Purpose
- Instructions for Use

- Version information

```
Python 2.7.3 (default, Jun 27 2012, 13:11:23)
[GCC 4.4.5 20110214 (Red Hat 4.4.5-6)] on linux2

e.g.  numpy.__version__

numpy v1.6.2
scipy v0.10.1
matplotlib v1.3.0
```

Analysis Scripts



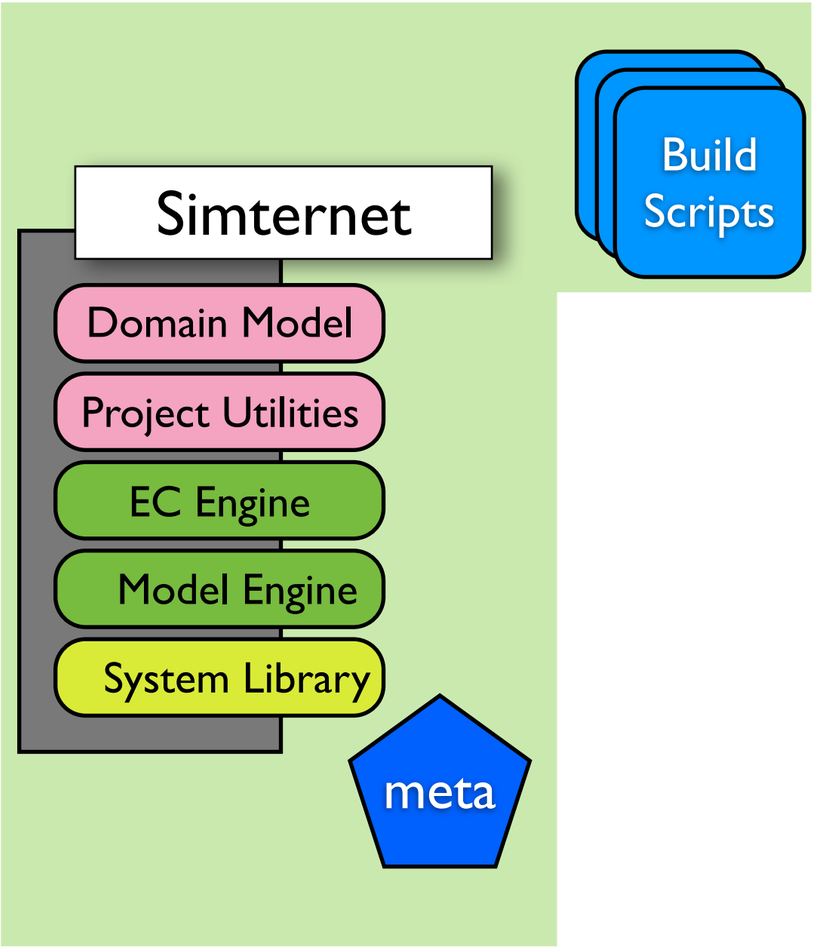
- Creator and PI details
- Contents and Purpose
- Instructions for Use

- Version information

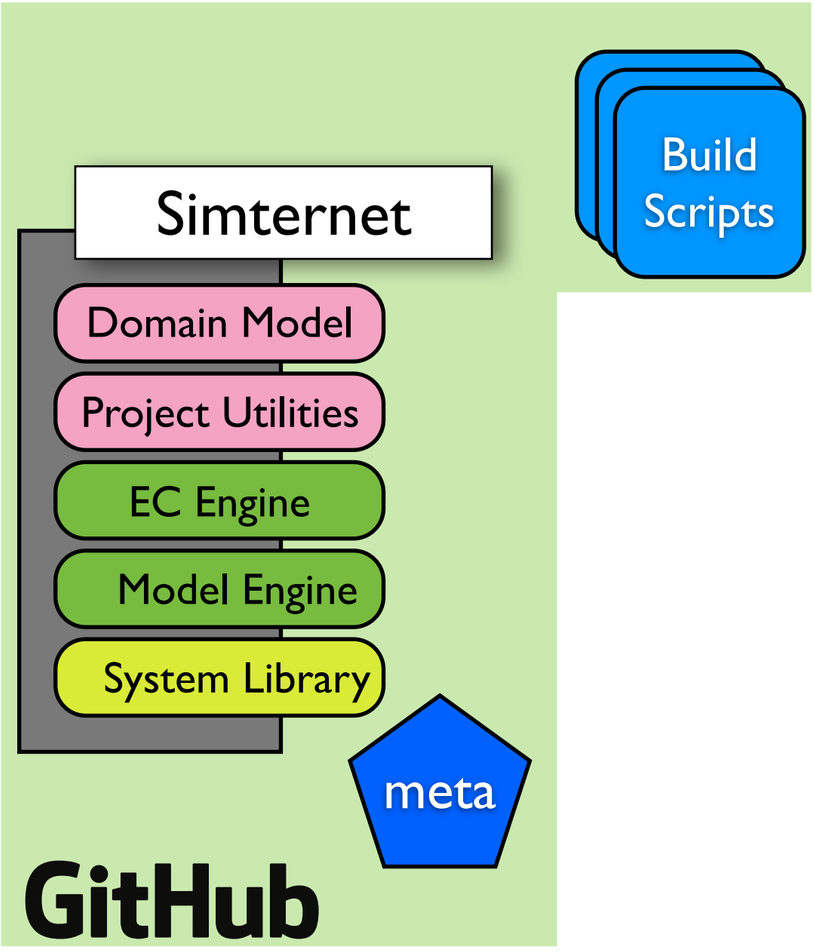
```
Python 2.7.3 (default, Jun 27 2012, 13:11:23)
[GCC 4.4.5 20110214 (Red Hat 4.4.5-6)] on linux2

e.g.  numpy.__version__

numpy v1.6.2
scipy v0.10.1
matplotlib v1.3.0
```

Project Data

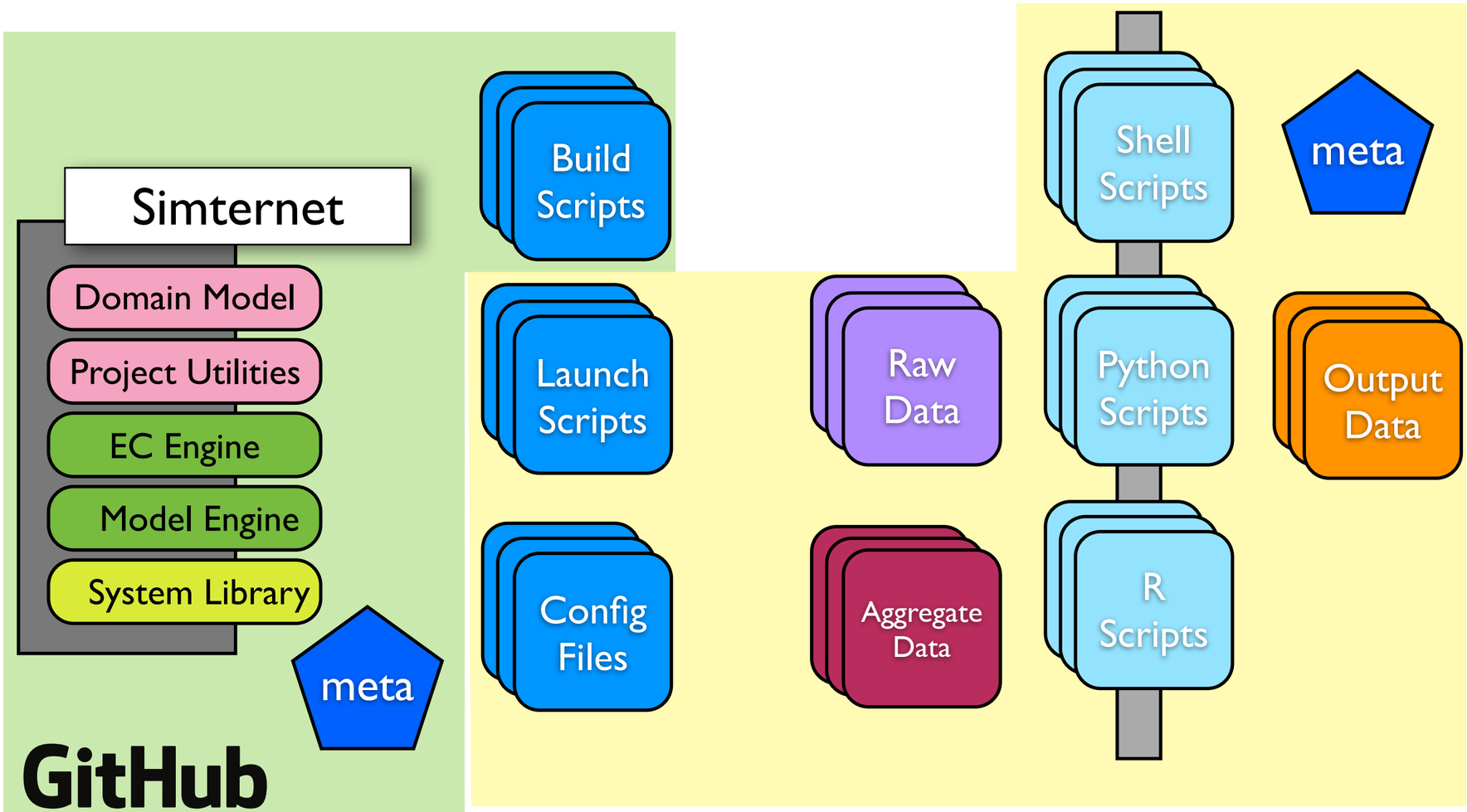
Project Data



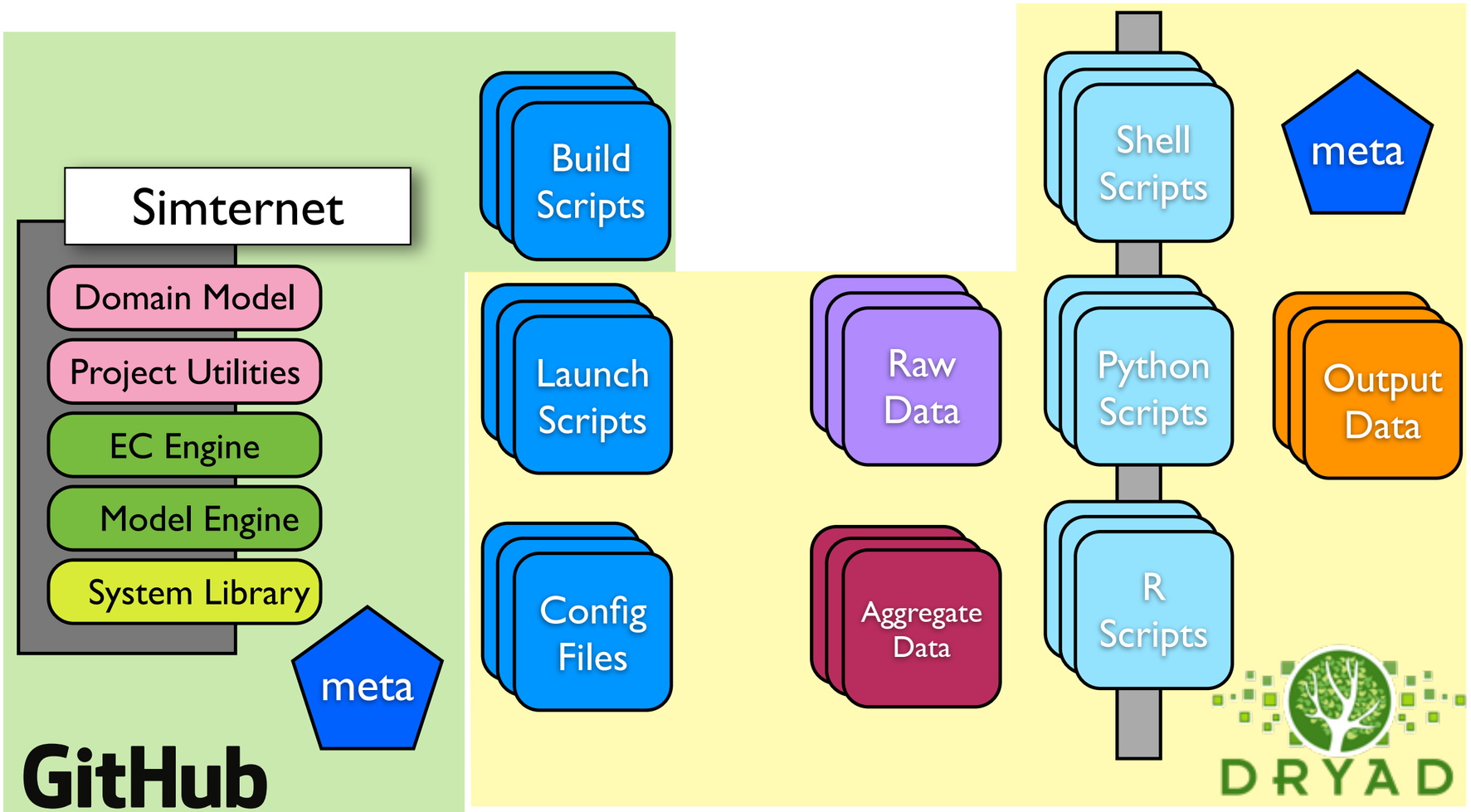
Project Data

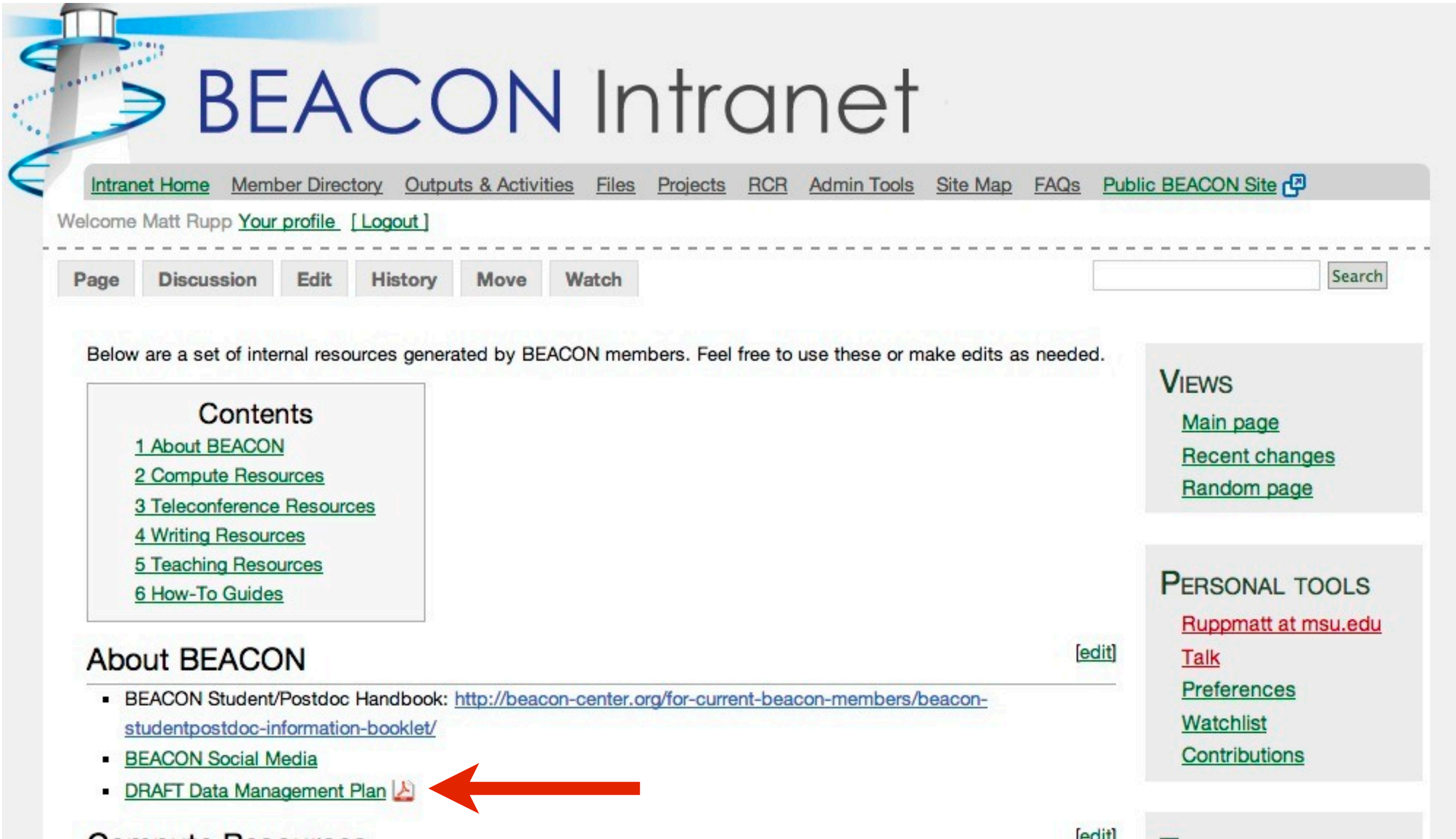


Project Data



Project Data



Backup Your Active Data

Preview the Data Management Plan