U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND ARMY RESEARCH LABORATORY

HTMDEC Data Science Overview

B. Christopher Rinderspacher

Research Chemist, Ph.D.

FCDD-RLW-MMC

ASPIRATIONS

Holy grail:

Enter research problem → suitable analyses and workflows are suggested → robot executes

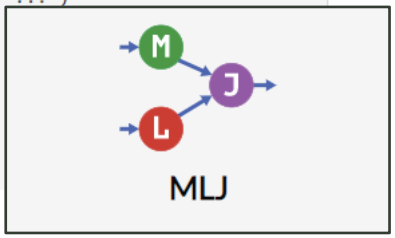
Anybody can do it:

- Compose research workflow from components for data capture, analysis, and **decision** pipelines
- Low-code, no-code approach
- Lower barrier to program planning and execution
- **High reusability**

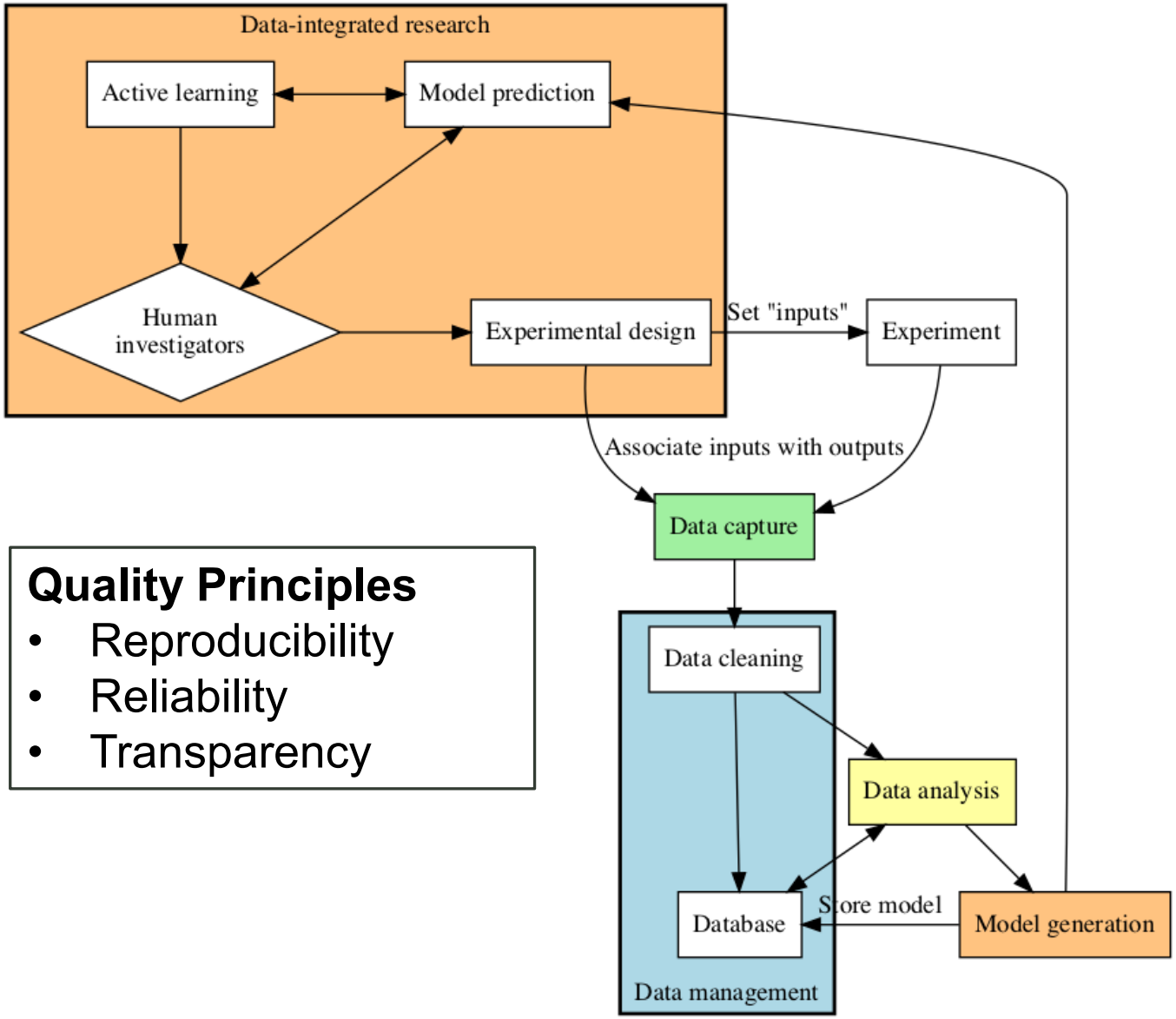
Also see: AutoML, autoconfig, Orange Data Mining, BPMS, cookiecutter, DrWatson.jl, or Data Versioning Control frameworks

To list *all* models available in MLJ's model registry do `models()`. Listing the models compatible with the present data:

```
julia> models(matching(X,y))
42-element Array{NamedTuple{(:name, :package_name, :is_supervised, :docstring, :hyperparameter_
 (name = AdaBoostClassifier, package_name = ScikitLearn, ... )
 (name = AdaBoostStumpClassifier, package_name = DecisionTree, ... )
 (name = BaggingClassifier, package_name = ScikitLearn, ... )
 (name = BayesianLDA, package_name = MultivariateStats, ... )
 (name = BayesianLDA, package_name = ScikitLearn, ... )
 (name = BayesianQDA, package_name = ScikitLearn, ... )
 (name = BayesianSubspaceLDA, package_name = MultivariateStats, ... )
 (name = ConstantClassifier, package_name = MLJModels, ... )
 (name = DecisionTreeClassifier, package_name = DecisionTree, ... )
 (name = DeterministicConstantClassifier, package_name = MLJModels, ... )
 ⋮
 (name = RidgeCVClassifier, package_name = ScikitLearn, ... )
 (name = RidgeClassifier, package_name = ScikitLearn, ... )
 (name = SGDClassifier, package_name = ScikitLearn, ... )
 (name = SVC, package_name = LIBSVM, ... )
 (name = SVMClassifier, package_name = ScikitLearn, ... )
```



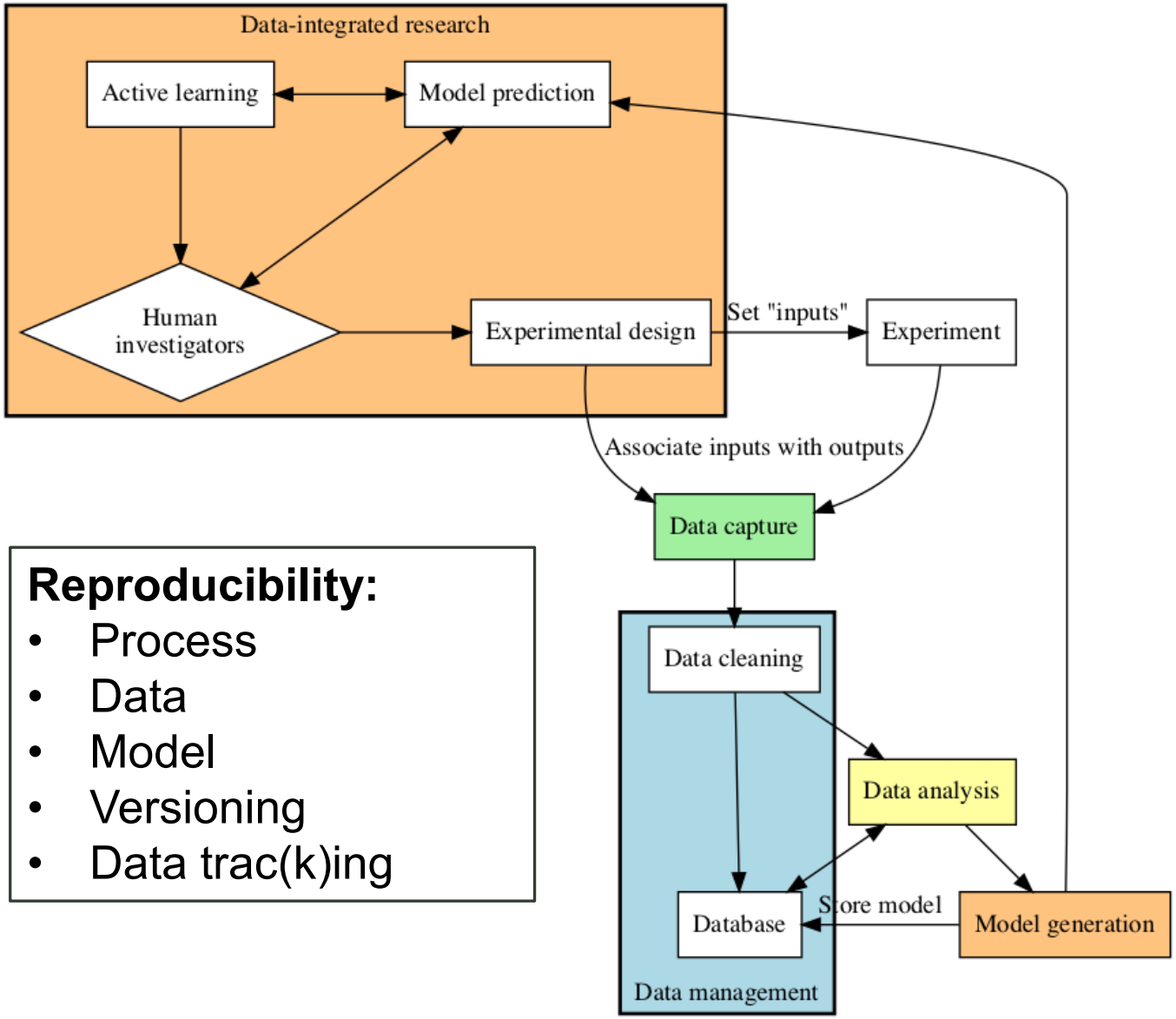
DATA INFORMED RESEARCH FLOW



Quality Principles

- Reproducibility
- Reliability
- Transparency

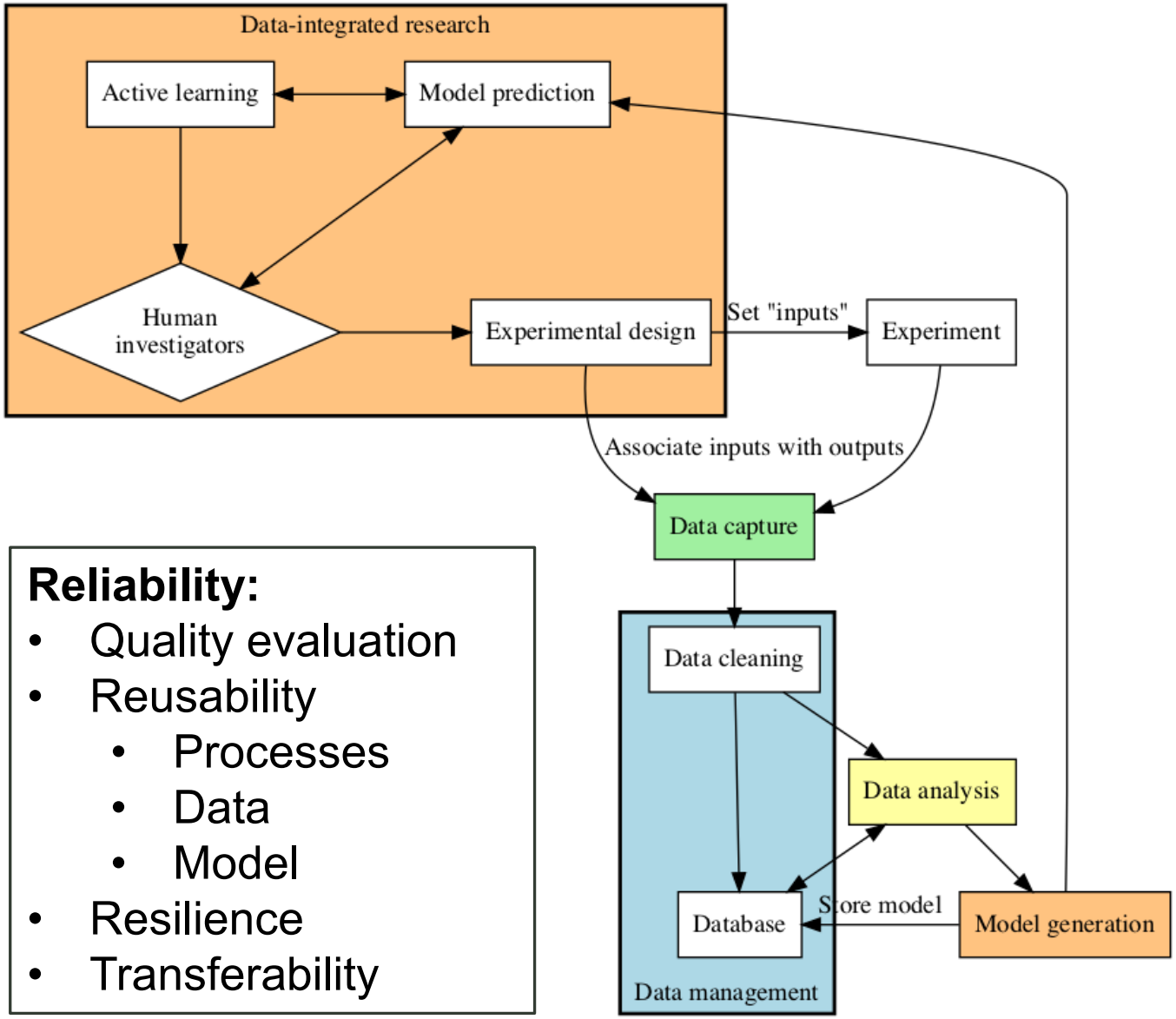
DATA INFORMED RESEARCH FLOW



Reproducibility:

- Process
- Data
- Model
- Versioning
- Data trac(k)ing

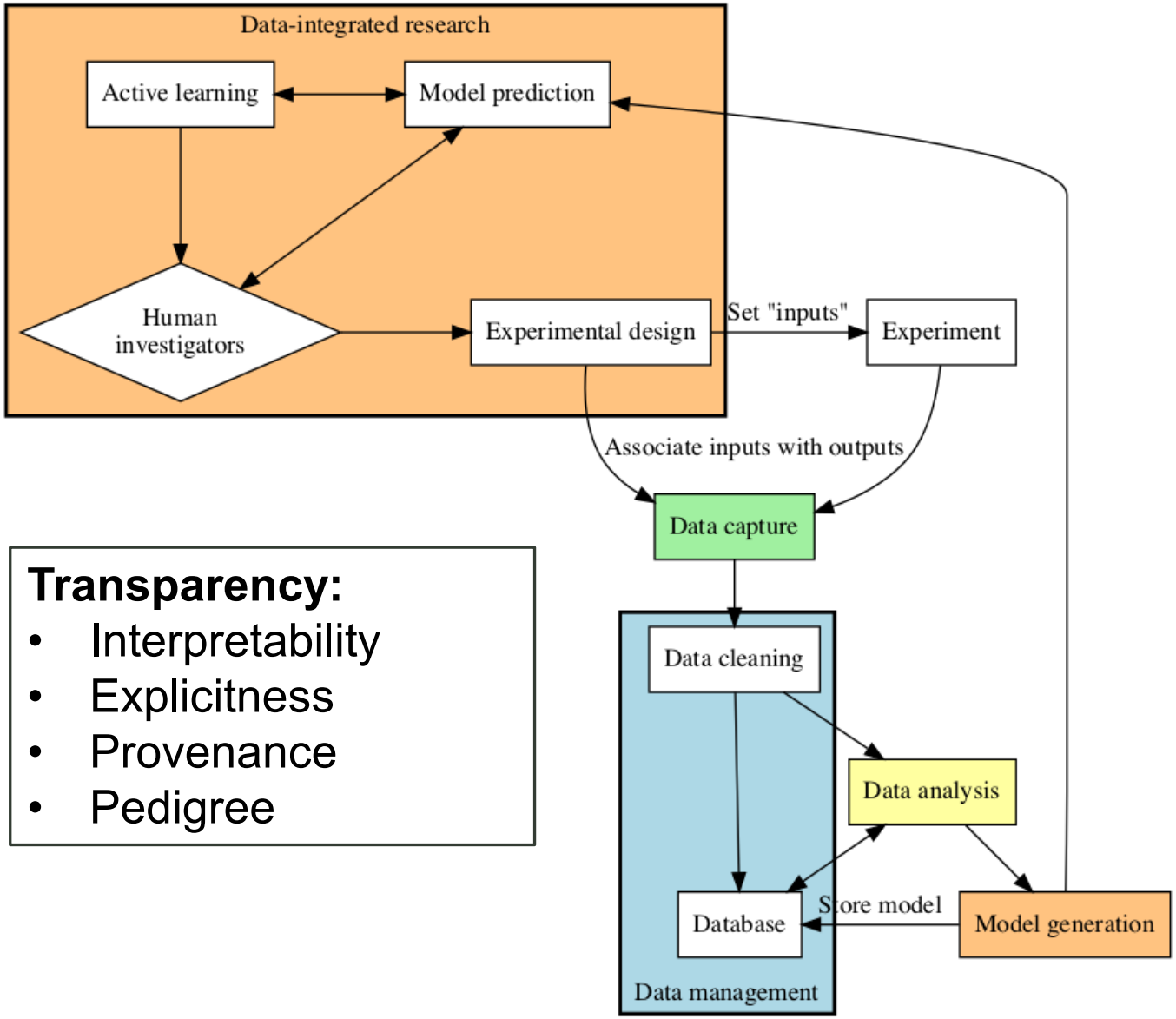
DATA INFORMED RESEARCH FLOW



Reliability:

- Quality evaluation
- Reusability
  - Processes
  - Data
  - Model
- Resilience
- Transferability

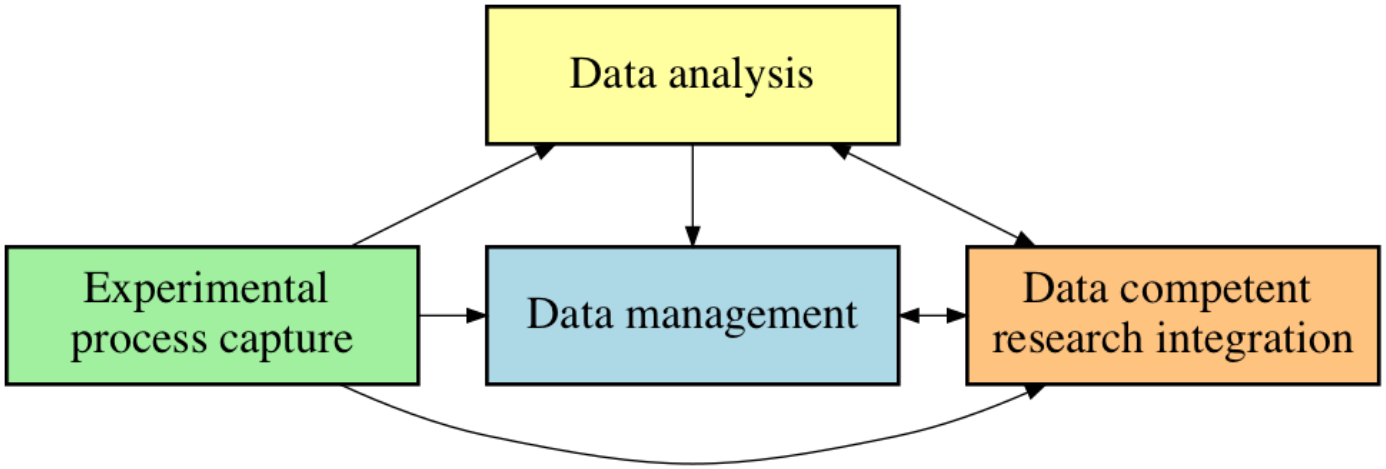
DATA INFORMED RESEARCH FLOW



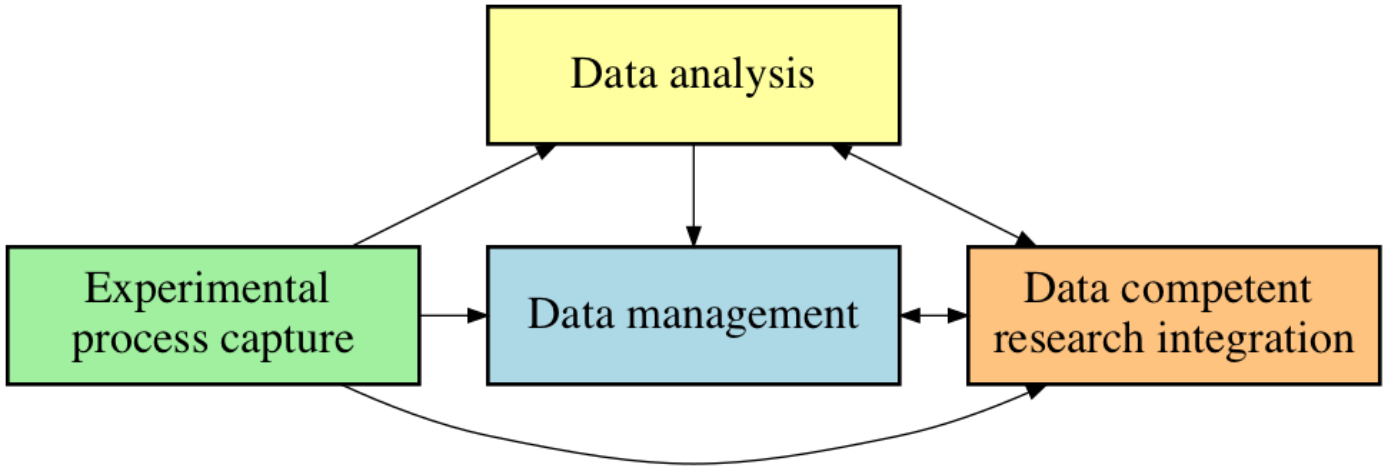
Transparency:

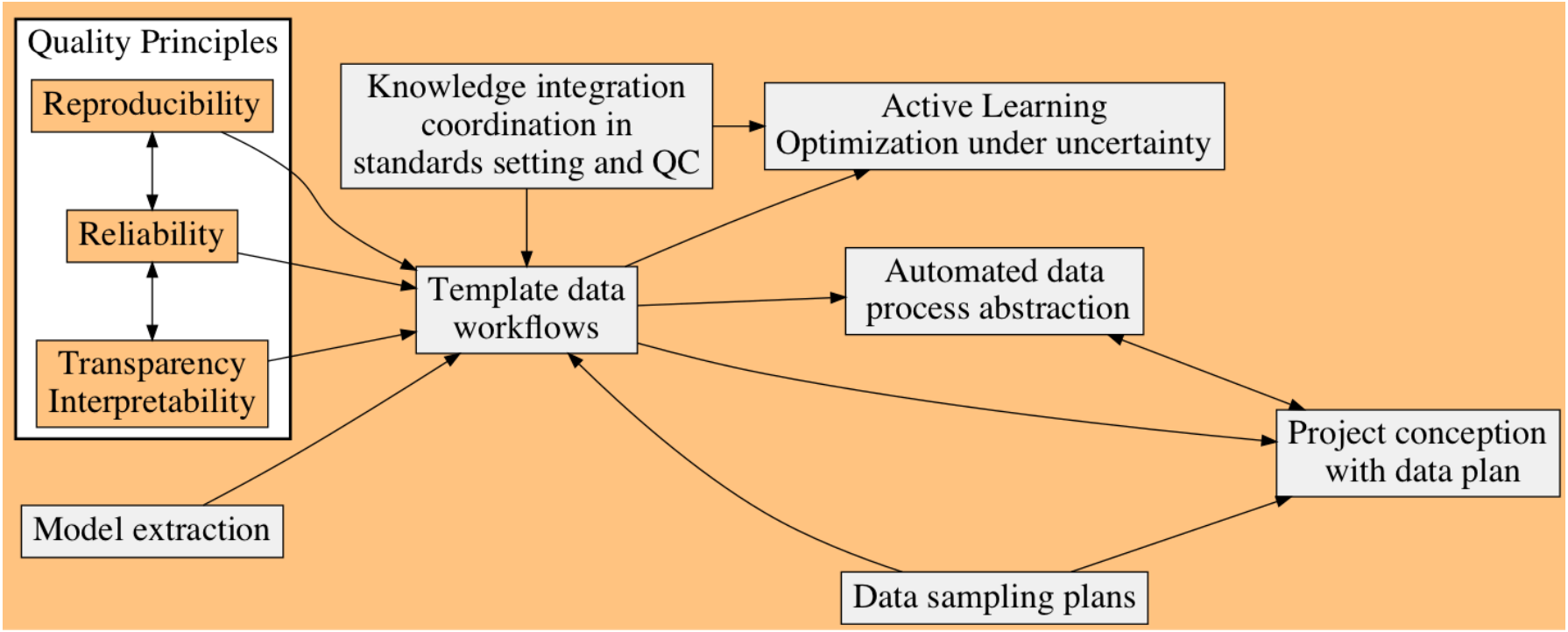
- Interpretability
- Explicitness
- Provenance
- Pedigree

DATA INFORMED RESEARCH REQUIREMENTS

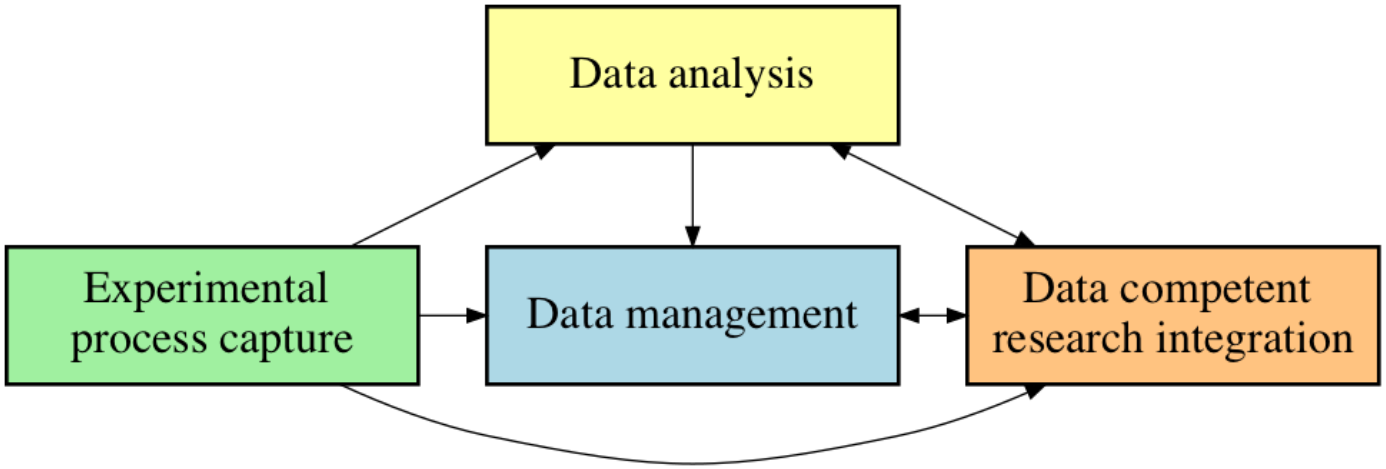


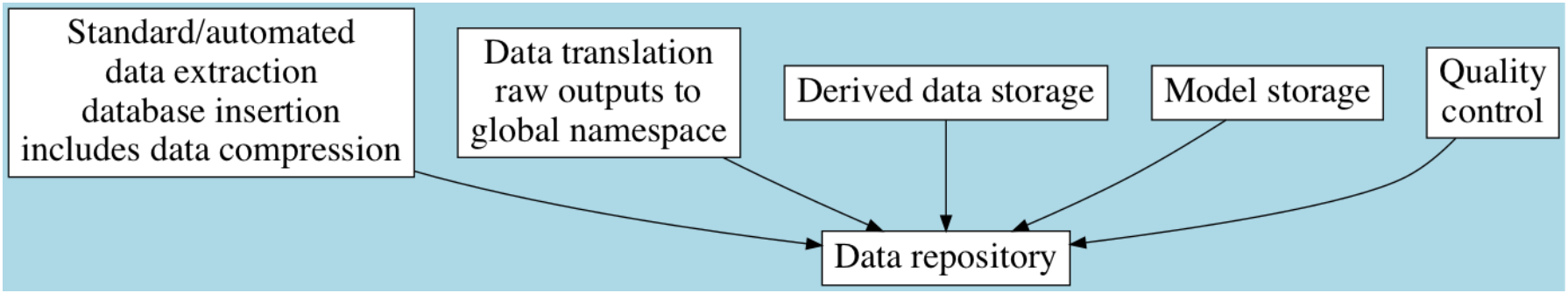
DATA COMPETENT RESEARCH INTEGRATION



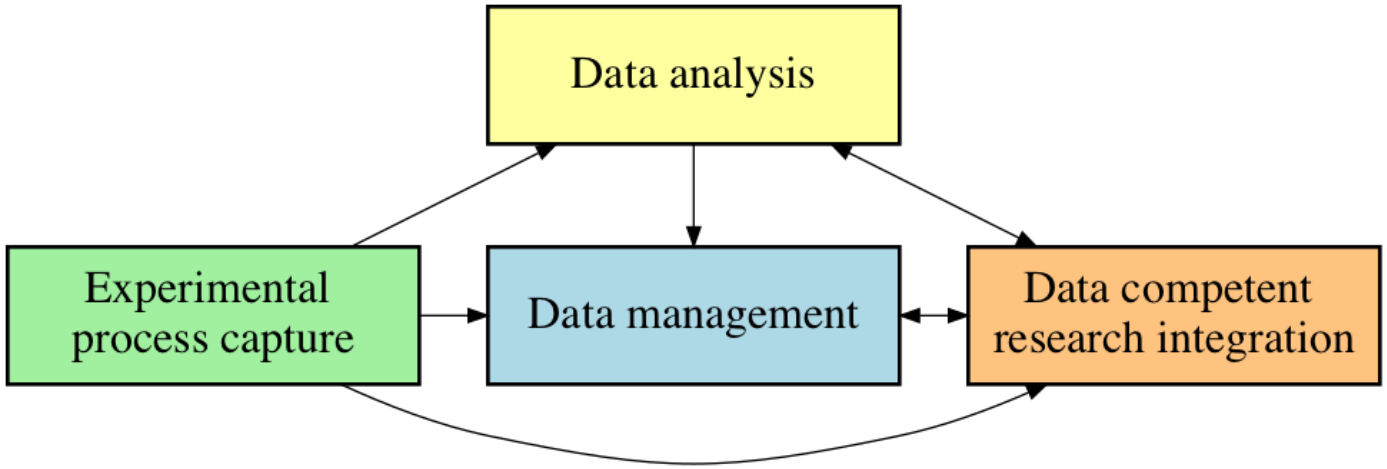


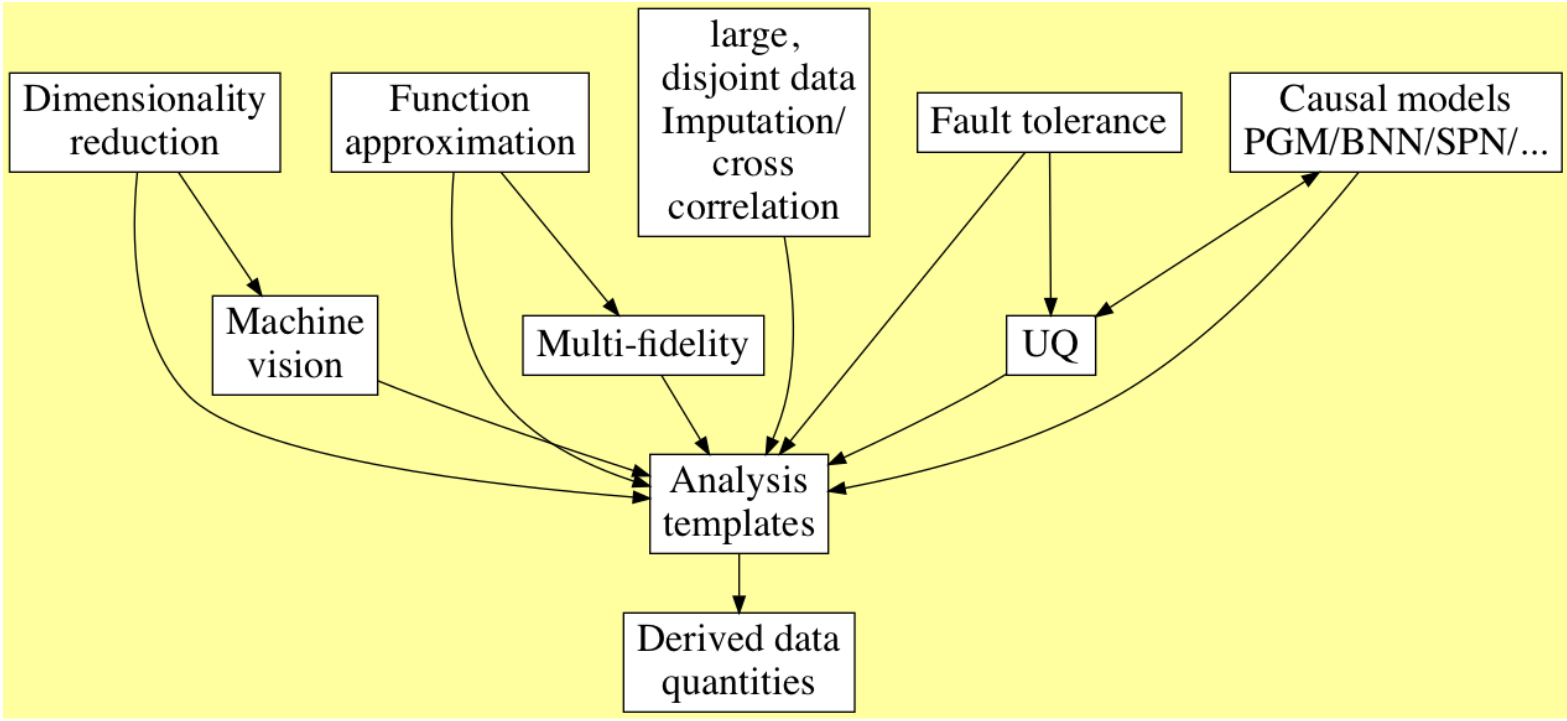
DATA MANAGEMENT COMPONENTS



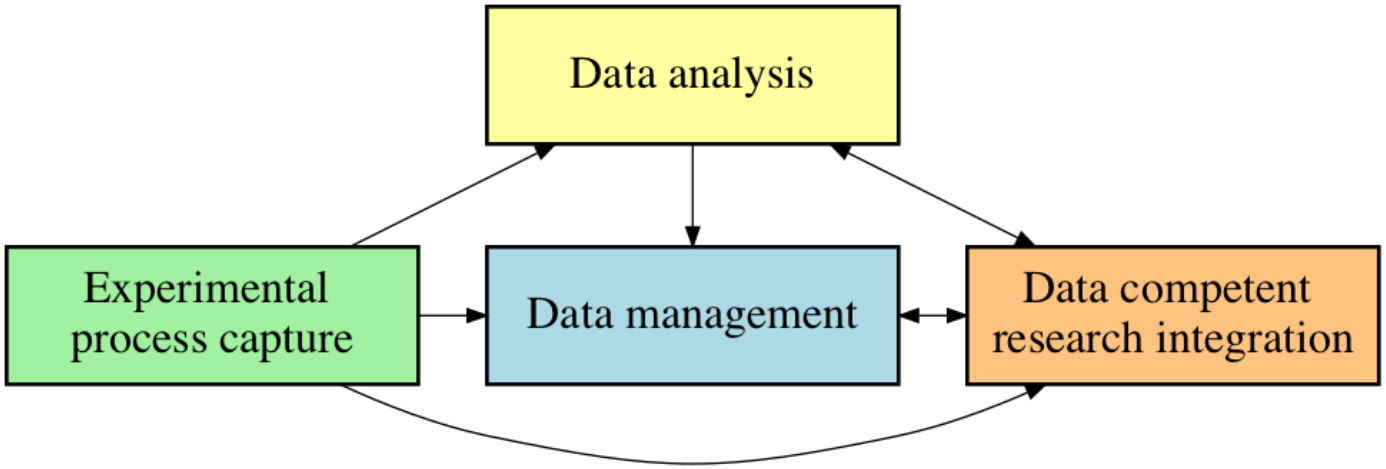


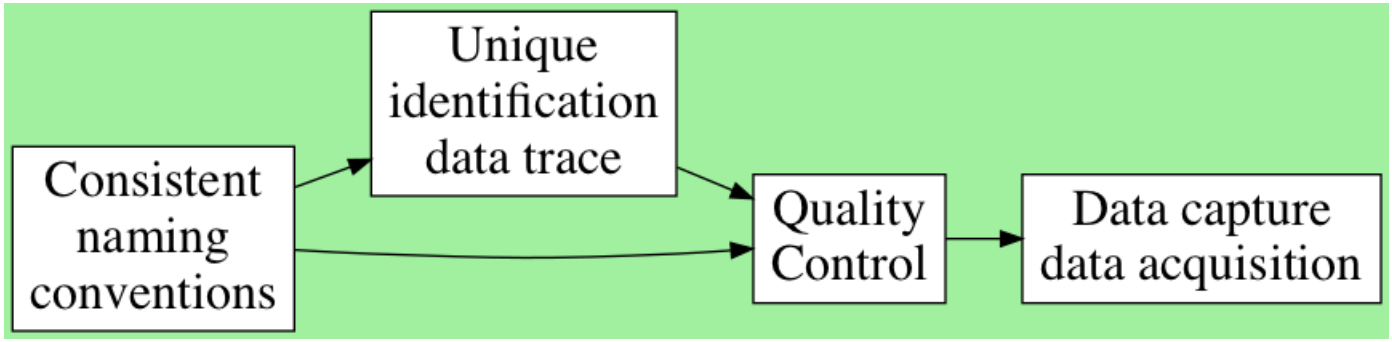
DATA ANALYSIS COMPONENTS



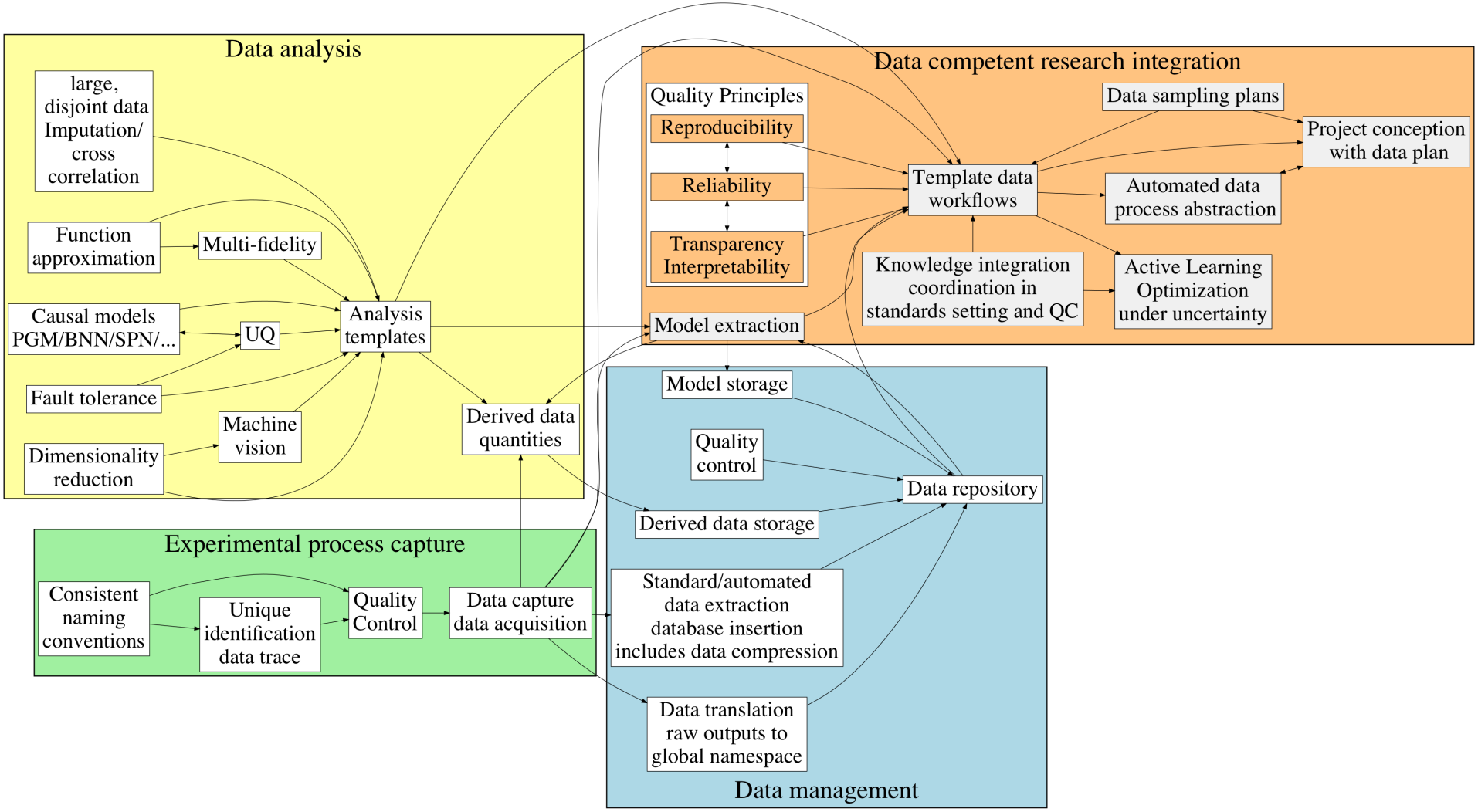


DATA CAPTURE/ACQUISITION

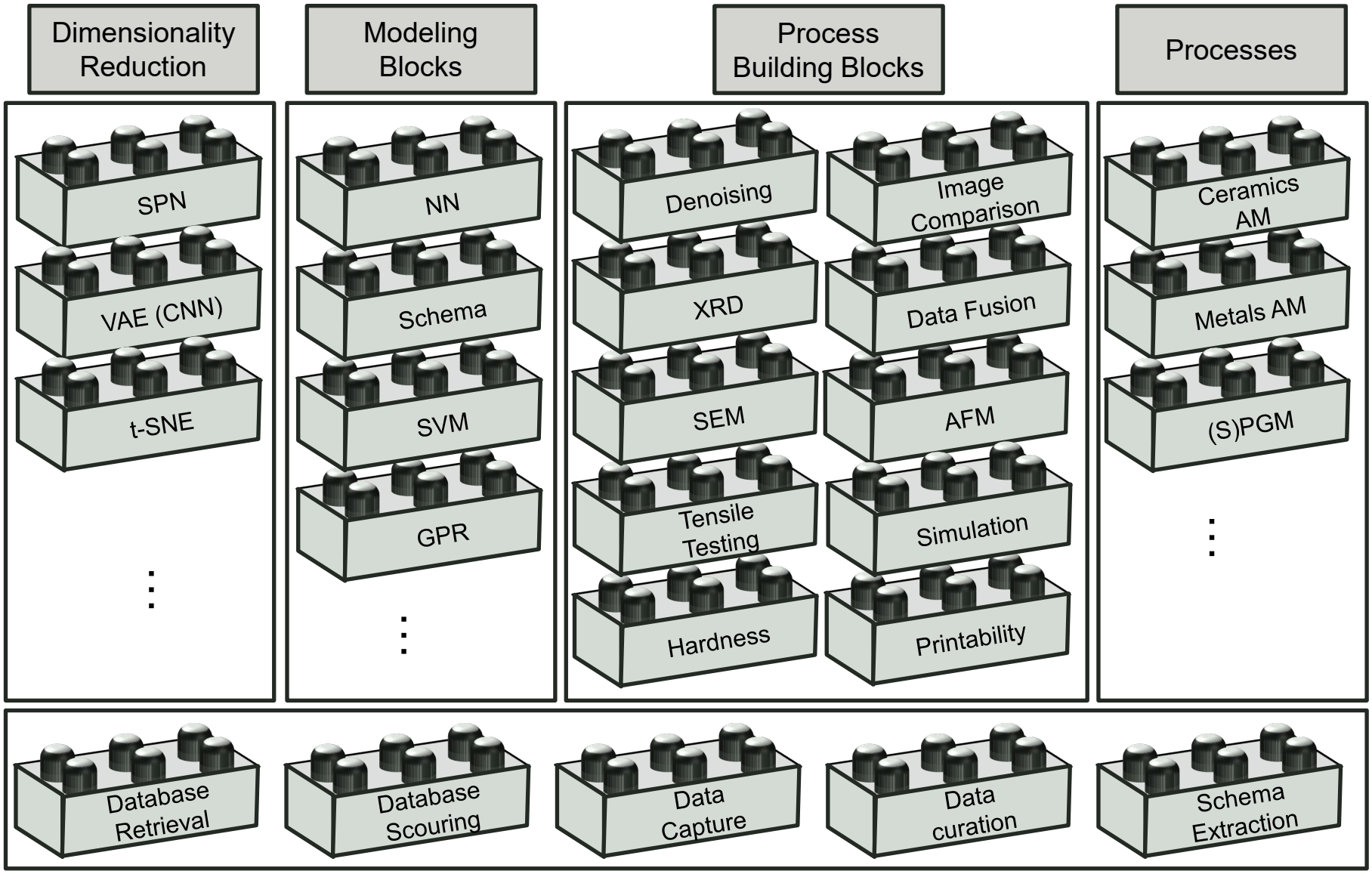




FULL DATA DEPENDENCY GRAPH



COMPOSABLE BUILDING BLOCKS



CHALLENGES

Reproducibility Challenges

- Data drift
- Concept drift
- UX/UI
- Equivalence problem
- User error detection

Transparency Challenges

- UX/automation balance
- Data/Process Visualization
- Documentation overhead
- Process development
- Findability/Access control/Information security

Reliability Challenges

- Interoperability
- Schema consistency
- Domain specific conventions

- User error mitigation
- Problem-specific UQ
- Applicability assessment

Backup Slides