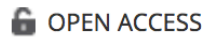


Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing


Mary Regina Boland,
Konrad J. Karczewski,
Nicholas P. Tatonetti

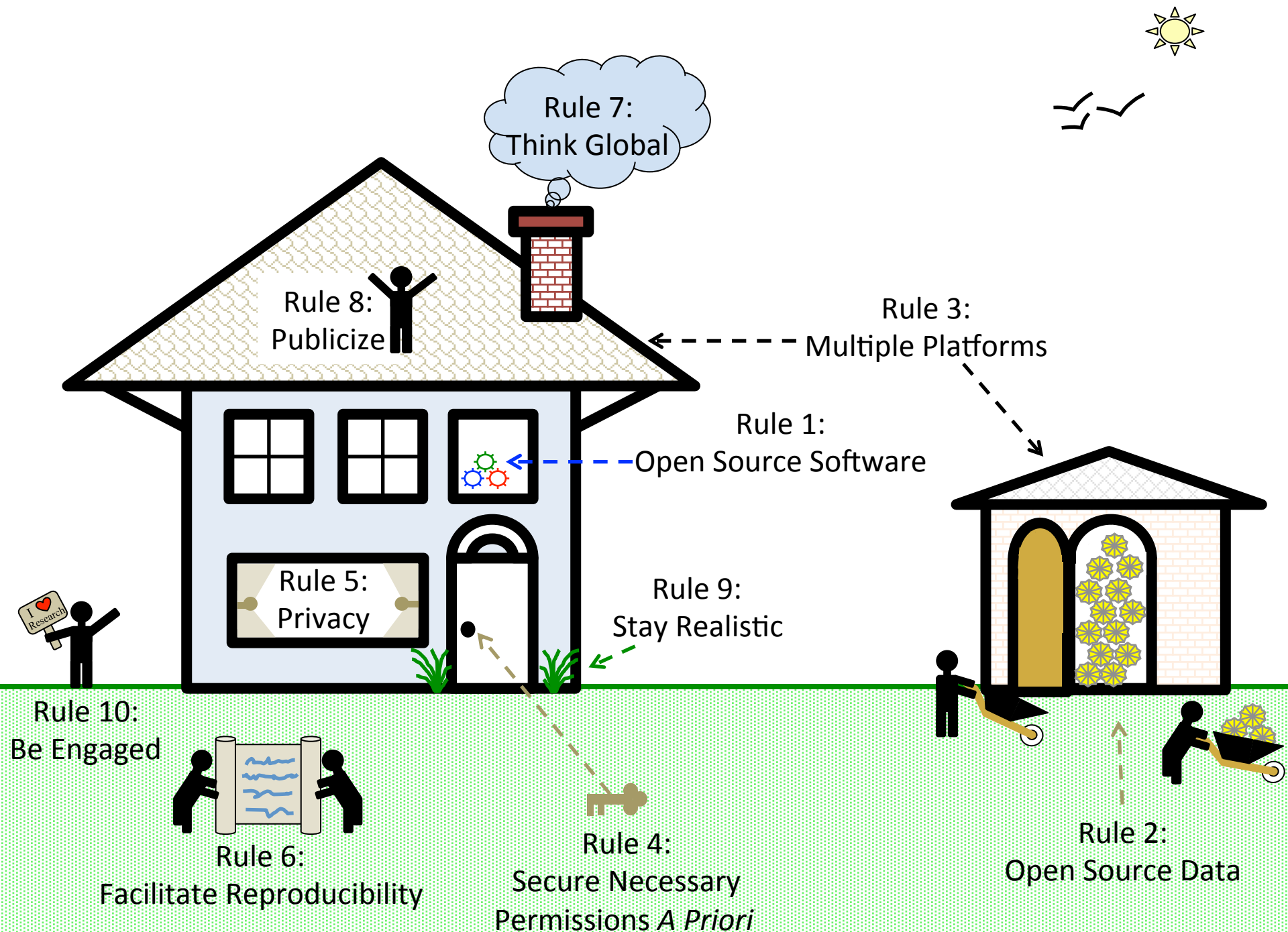
OHDSI Meeting - February 21, 2017

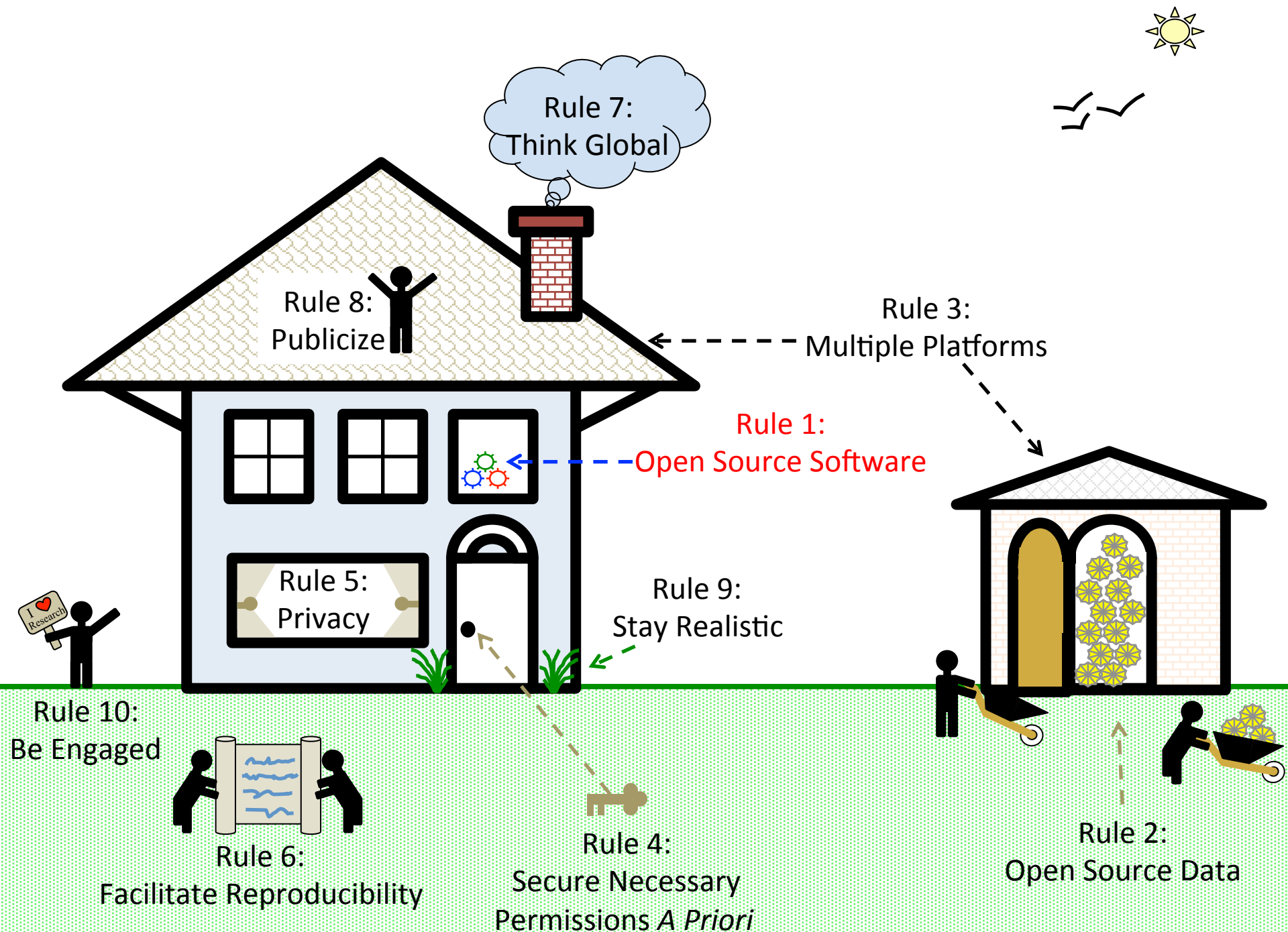


EDITORIAL

Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing

Mary Regina Boland , Konrad J. Karczewski, Nicholas P. TatonettiPublished: January 19, 2017 • <http://dx.doi.org/10.1371/journal.pcbi.1005278>**Article****Authors****Metrics****Comments****Related Content**URL: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005278>





Rule 1: Make Software Open-Source

- Cornerstone of effective collaborations
- It is necessary for collaborators to have access to code in a repository that is shared among collaborators (although, this could be private and not open to the general public)
- Masum *et al.* advocate the reuse of existing code in their Ten Simple Rules for cultivating open science. However, this is often easier said than done. As long as the backend algorithms remain hidden, open science will not be possible

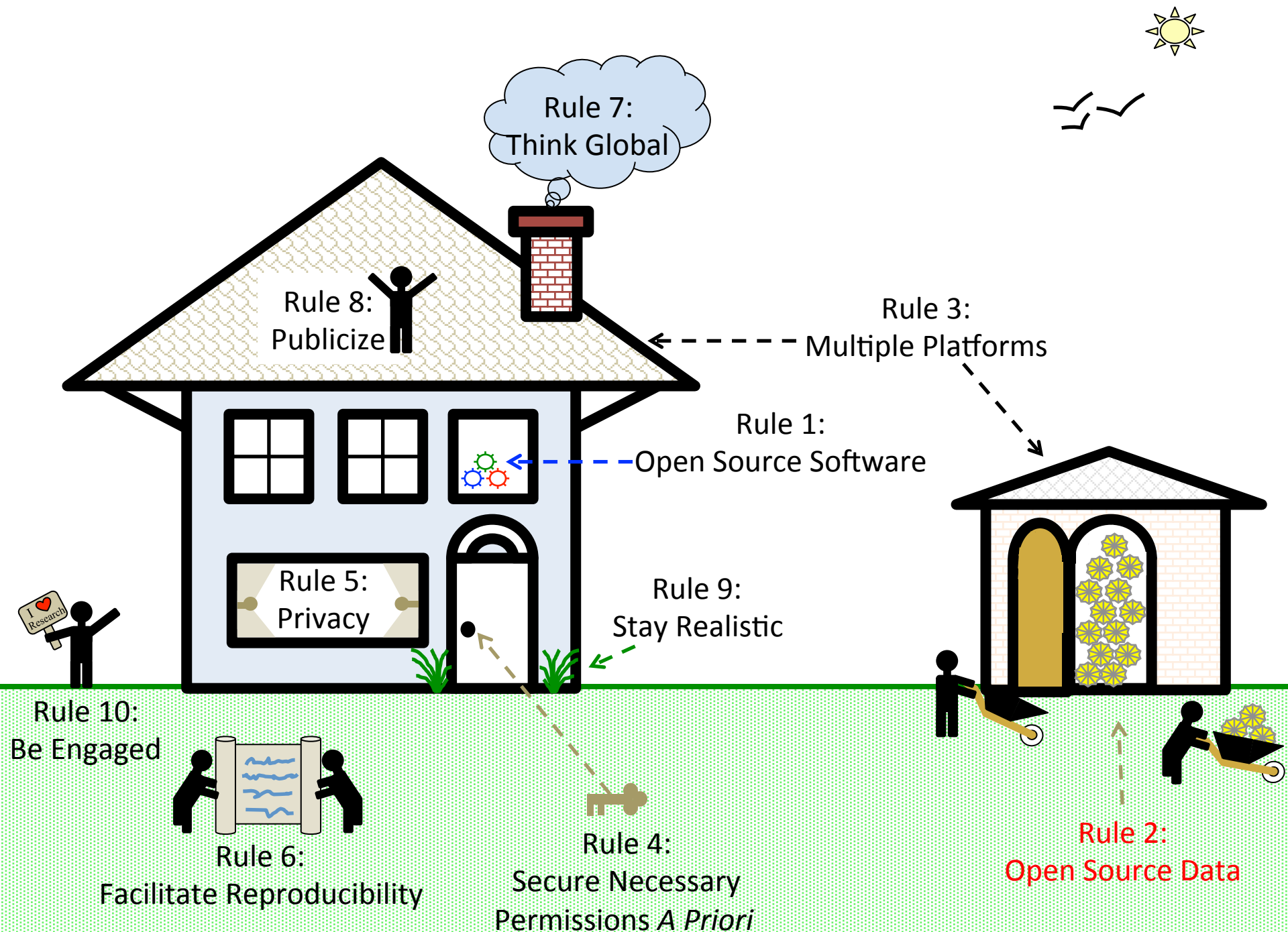
Useful Refs: Prlić A, Procter JB. Ten Simple Rules for the Open Development of Scientific Software. PLoS Comput Biol. 2012; 8(12):e1002802. doi: 10.1371/journal.pcbi.1002802 PMID: 23236269

Masum H, Rao A, Good BM, Todd MH, Edwards AM, Chan L, et al. Ten Simple Rules for Cultivating Open Science and Collaborative R&D. PLoS Comput Biol. 2013; 9(9):e1003244. doi: 10.1371/journal.pcbi.1003244 PMID: 24086123

Rule 1: Make Software Open-Source

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 1: Make Software Open-Source		
	Github	https://github.com
	CRAN	https://cran.r-project.org
	Bioconductor	https://www.bioconductor.org



Rule 2: Provide Open-Source Data

- Deposit Source Data in Appropriate Repositories
 - Source data could include not only processed or cleaned data used in algorithms but also raw data files. These files can often be very large; therefore, they are often stored in some external site or data warehouse
 - It is also helpful to provide intermediate data files at various stages of processing
 - While a multi-site research project is still ongoing, data can be shared in a private shared space until all necessary data quality checks have been conducted and the findings have been published. After publication, data can be deposited

Rule 2: Provide Open-Source Data

- Consider Middle-Ground Data Sharing Approaches for Sensitive Data
 - For example, the database for Genotypes and Phenotypes or dbGaP (<https://www.ncbi.nlm.nih.gov/gap>) provides data holders with two levels of access:
 - Open: allows for broad release of nonsensitive data online
 - Controlled: allows for controlled release allows sensitive datasets to be shared with other investigators, provided certain restrictions are met
 - Increases the ability for researchers to share portions of their data that would not be shareable otherwise

Rule 2: Provide Open-Source Data

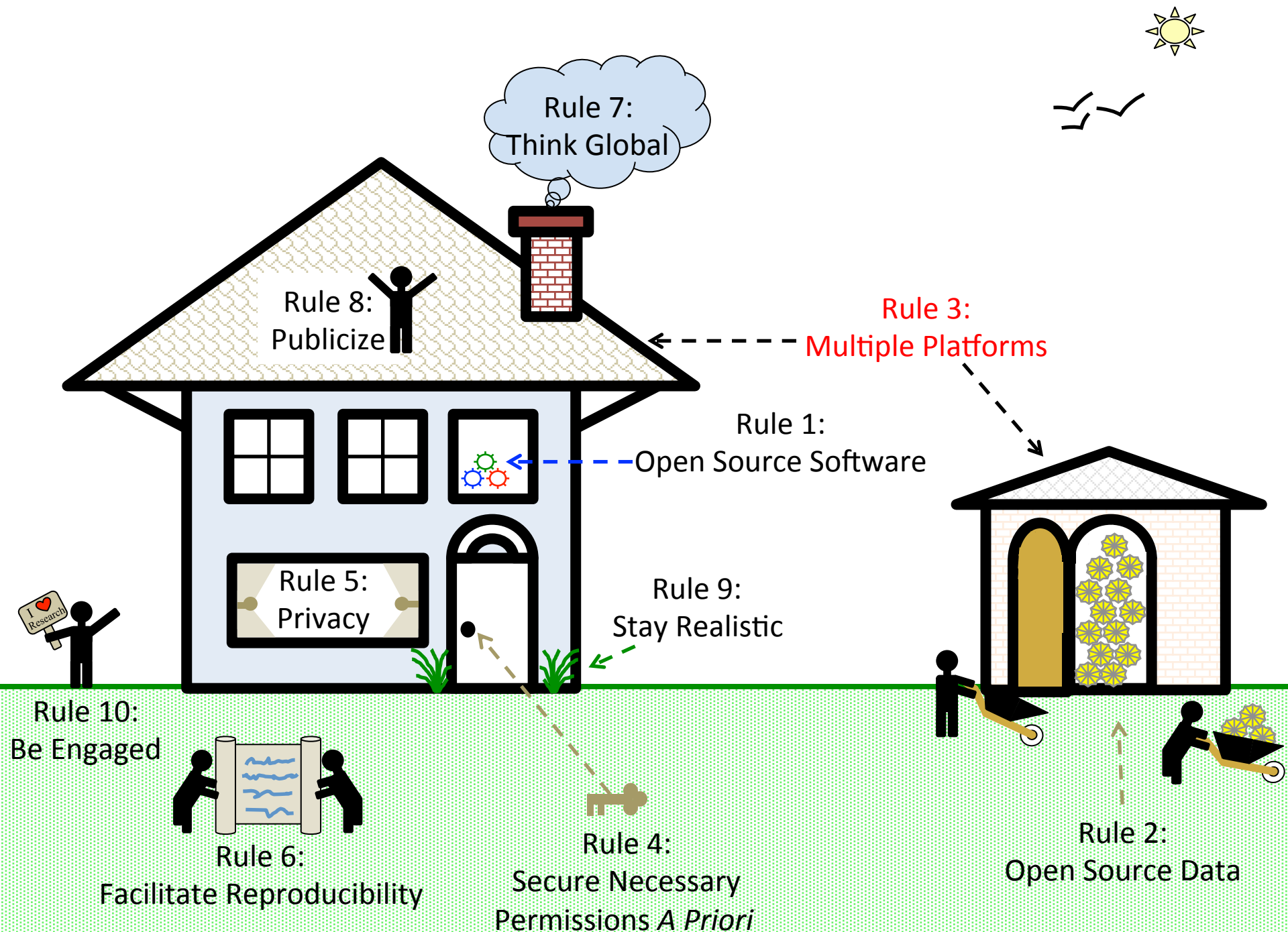
- Consider Middle-Ground Data Sharing Approaches for Sensitive Data
 - Federated Access systems that allow researchers to query databases containing sensitive data while preventing direct access to the data itself
 - Examples:
 - Shared Health Research Information Network (SHRINE), is a USA Federated system that is HIPAA compliant.
 - BioGrid Australia (<https://www.biogrid.org.au/>) allows researchers to access hundreds of thousands of health records through a linked data platform where individual data holders maintain control of their data
 - Other groups have approached the problem from another angle:
 - Provide summary statistics computed over large cohorts (e.g., ExAC browser/database)
 - Maintains privacy while providing others with important information about the populations that can be used in subsequent analyses and comparisons

Rule 2: Provide Open-Source Data

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 2: Provide Open-Source Data (When Possible)		
<i>Deposit Source Data in Appropriate Repositories</i>		
	Sequence Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sra
	Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo
	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar
<i>Consider Middle-Ground Data Sharing Approaches for Sensitive Data</i>		
	dbGaP	https://www.ncbi.nlm.nih.gov/gap
	Shared Health Research Information Network (SHRINE)	https://catalyst.harvard.edu/services/shrine
	BioGrid Australia	https://www.biogrid.org.au





Rule 3: Use Multiple Platforms to Share Research Products

- Research products take many different forms, including:
 - 1) **raw source** data regardless of collection type (e.g., health data, genomic data, survey data, and epidemiological data)
 - 2) **software code** (mentioned in rule 1)
 - 3) **metadata elements** and **results of computations used to generate figures** published in scientific research. Some data types cannot be fully shared (e.g., EHR data), but most algorithms and summary results/statistics are shareable.

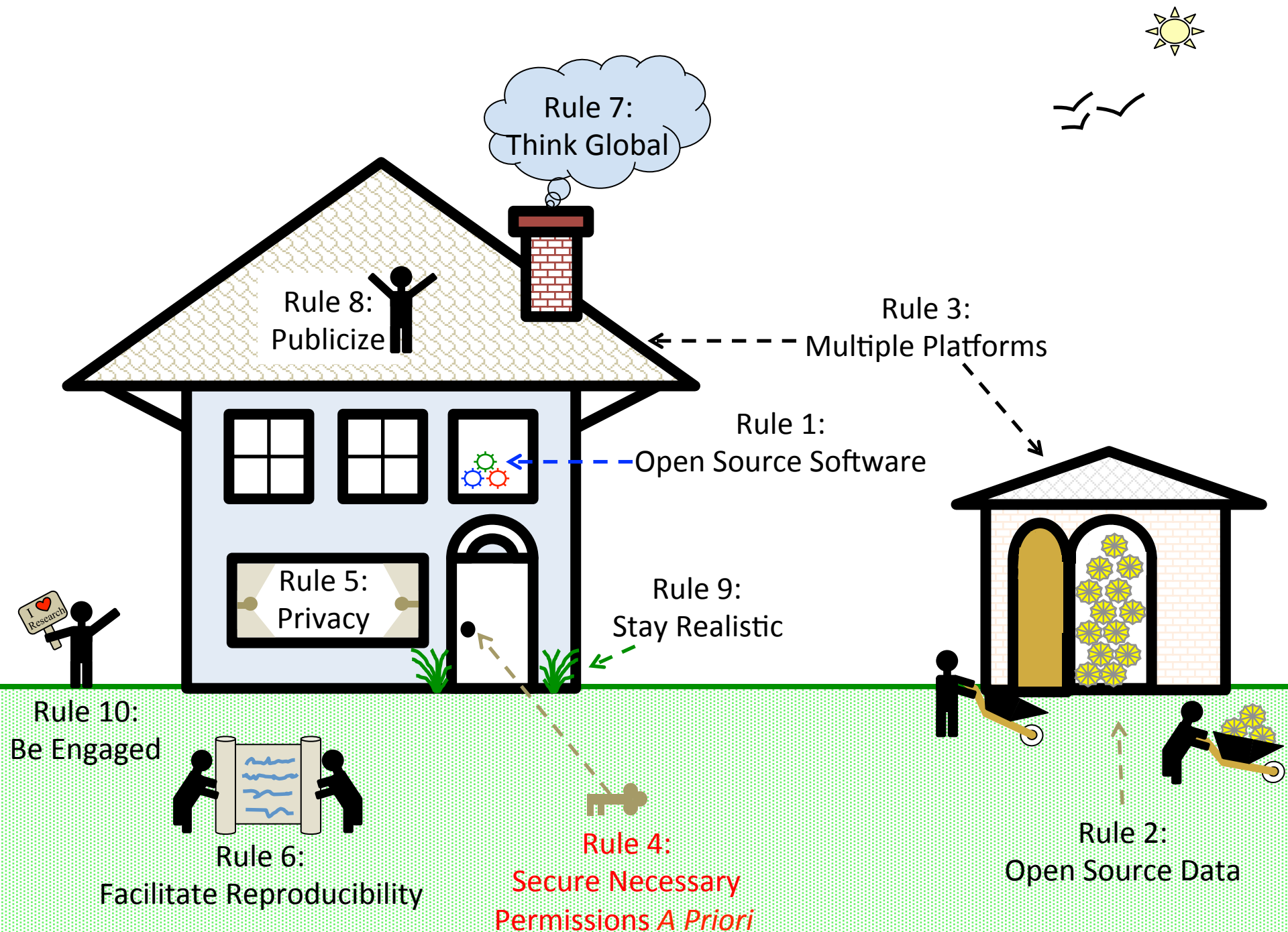
Rule 3: Use Multiple Platforms to Share Research Products

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 3: Use Multiple Platforms to Share Research Products		
	Figshare	https://figshare.com
	Github	https://github.com
	ExAC Browser	http://exac.broadinstitute.org
	Google Forums	



- Different Platforms are Optimal for Different Data Types:
 - 1) **raw source** data regardless of collection type (e.g., health data, genomic data, survey data, and epidemiological data): GEO, dbGaP, etc.
 - 2) **software code**: **GitHub**
 - 3) **metadata elements** and **results of computations used to generate figures**: **Figshare** is great for data and results related to figures, ExAC browser is good for high-level exploration of data



Rule 4: Secure Necessary Permissions/Data Use Agreements

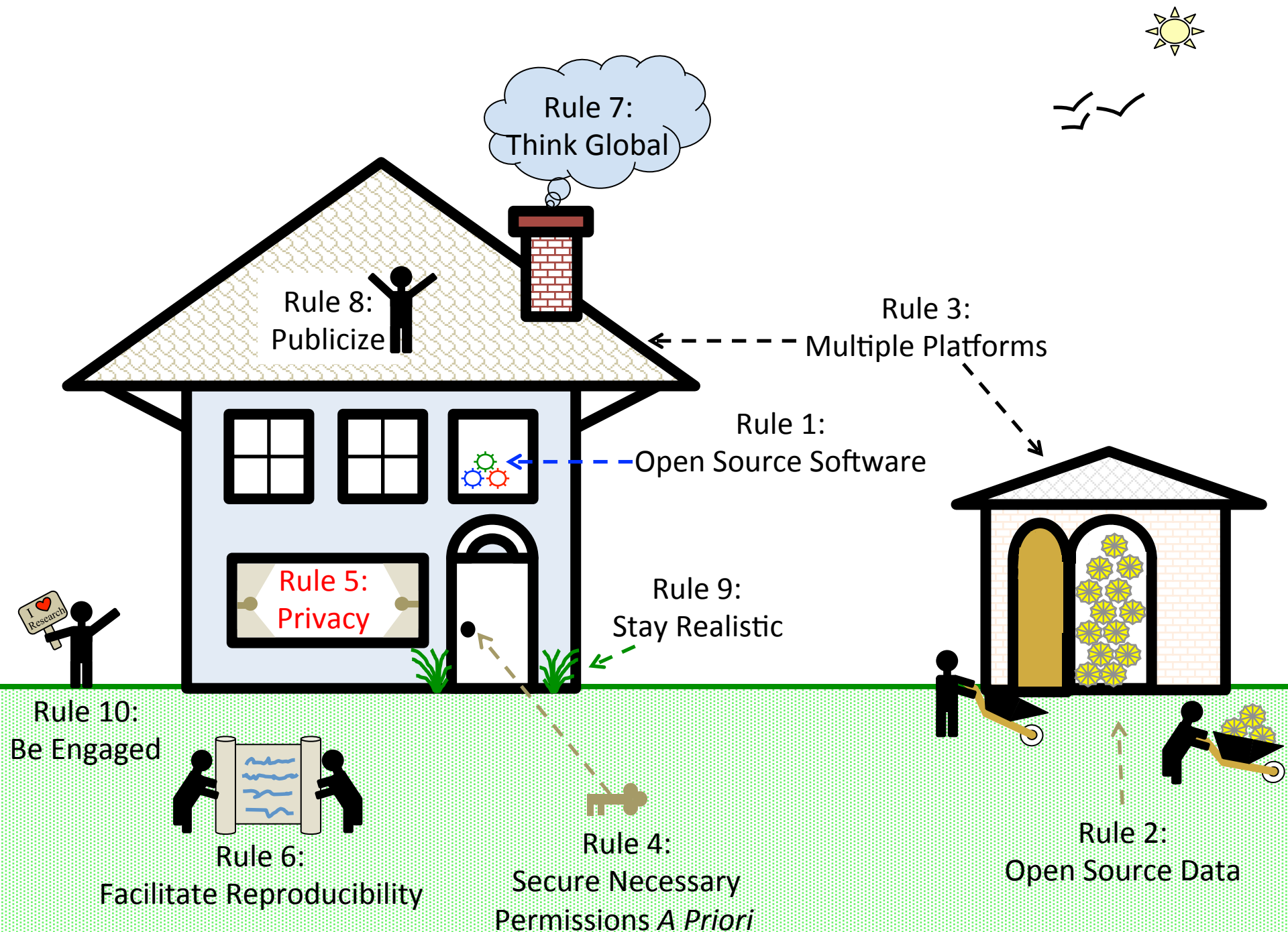
A Priori

- Some datasets have stipulations that affect publication
 - E.g., NASA
- Others have DUAs to ensure that patient privacy is protected
 - E.g., SEER and SEER-MediCare
- Data that is not sensitive may have restrictive DUAs for other reasons (e.g., data from a collaborator in industry)
- Some countries have different legal constraints that can affect DUAs
 - Important to investigate *a priori*

Rule 4: Secure Necessary Permissions/Data Use Agreements *A Priori*

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 4: Secure Necessary Permissions/Data Use Agreements A Priori		
<i>Guides for Creating a DUA</i>		
	Department of Health and Human Services Best Practice Guide for DUA	http://www.hhs.gov/ocio/eplc/EPLC%20Archive%20Documents/55-Data%20Use%20Agreement%20(DUA)/eplc_dua_practices_guide.pdf
	Health Care Systems Research Network DUA Toolkit	http://www.hcsrn.org/en/Tools%20&%20Materials/GrantsContracting/HCSRNDUAToolkit.pdf
<i>Example DUAs</i>		
	NASA DUA	http://above.nasa.gov/Documents/NGA_Data_Access_Agreement_new.pdf
	SEER-MEDICARE DUA	https://healthcaredelivery.cancer.gov/seermedicare/obtain/seerdua.docx



Rule 5: Know the Privacy Rules for Your Data

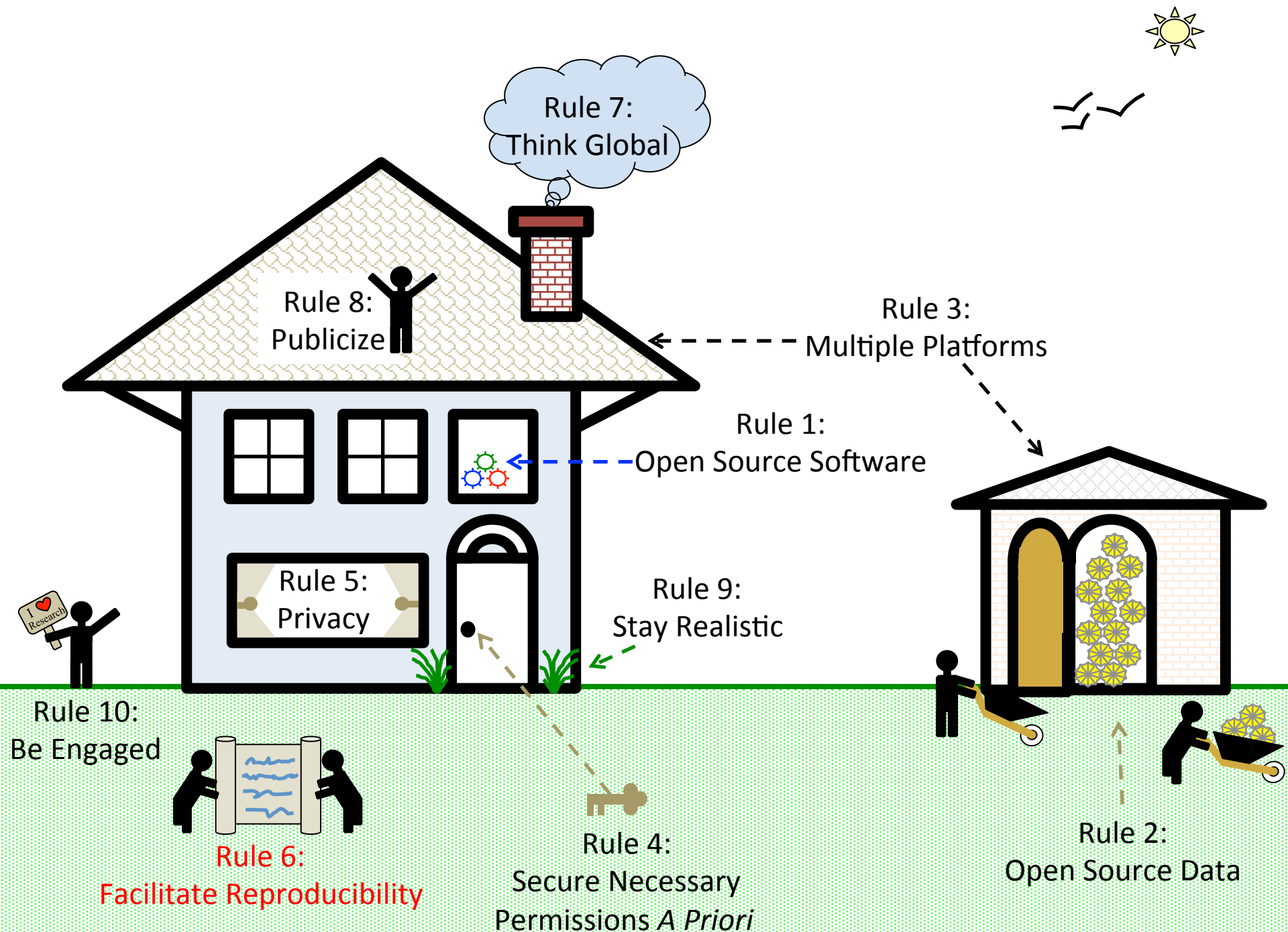
- Certain datasets, e.g., genomic and EHR data, may be impossible to fully publish on an open platform due to the Health Insurance Portability and Accountability Act (HIPAA) privacy rules and other privacy concerns
- Methods that anonymize patient information while allowing patient-level data sharing may be the way of the future (El Emam et al.)
- However, institutional-specific policies and/or country-specific laws can limit or prevent usage of such methods. This is an important item to consider and discuss with all collaborators at the outset of any collaboration.

El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. *bmj*. 2015; 350:h1139. doi: 10.1136/bmj.h1139 PMID: 25794882

Rule 5: Know the Privacy Rules for Your Data

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 5: Know the Privacy Rules for Your Data		
	Health Insurance Portability and Accountability Act (HIPAA)	http://www.hhs.gov/hipaa/for-professionals/privacy



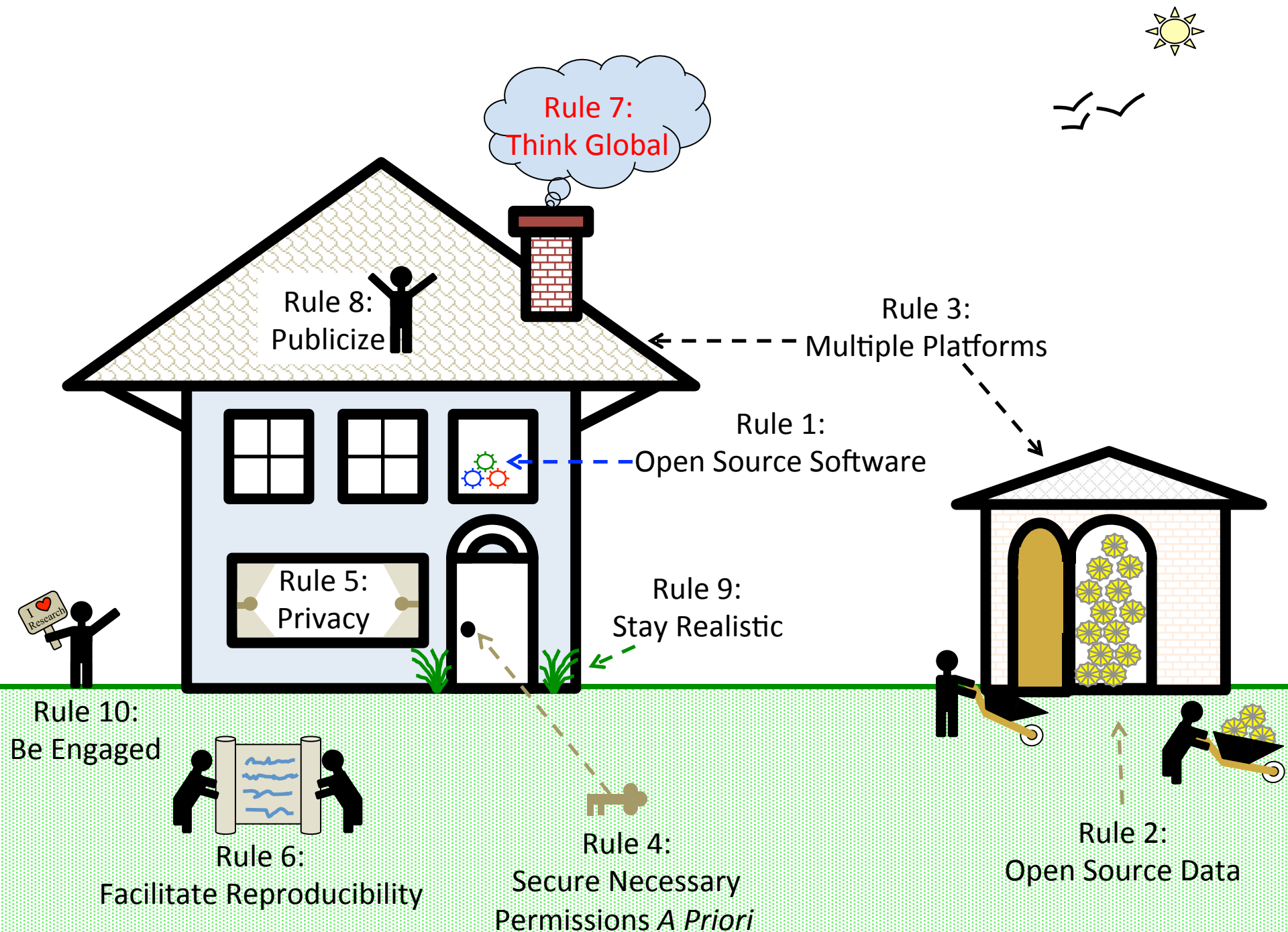
Rule 6: Facilitate Reproducibility

- **Keeping track of** research results and **how data were generated** is **vital** for reproducibility
- Site-level record keeping is essential when engaging in multi-site collaborations. If one aspect of a methodology is not conducted in the same way at one site, the overall results can be affected in drastic ways
- Researchers that depend on local EHR terminology systems for identifying patient populations must standardize and harmonize phenotype definitions across multiple sites
- Platforms are needed to provide links to all necessary documentation, code, and data schemas to help facilitate this process

Rule 6: Facilitate Reproducibility

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 6: Facilitate Reproducibility		
<i>Resources for Increasing Research Reproducibility</i>		
	MetaSub Research Integrity and Reproducibility	http://metasub.org/research-integrity-and-reproducibility/
	Reproducibility and Open Science Working Group—GitHub	http://uwescience.github.io/reproducible/guidelines.html https://github.com/uwescience/reproducible
<i>Example Projects with Assessed Reproducibility</i>		
	eMERGE PheKB	https://phekb.org/network-associations/emerge



Rule 7: Think Global

- **Mechanical differences** (i.e., the software language and documentation)
 - Software documentation (especially important for open-source languages) not available in many languages (often English only)
 - R a popular open-source language yet has official documented translations in only four languages: English, Russian, German, and Chinese (<https://www.r-project.org/other-docs.html>)
- **Conceptual differences** (i.e., country or region-specific medical definitions)
 - World Health Organization works tirelessly to integrate different conceptual interpretations of diseases into a standard guideline

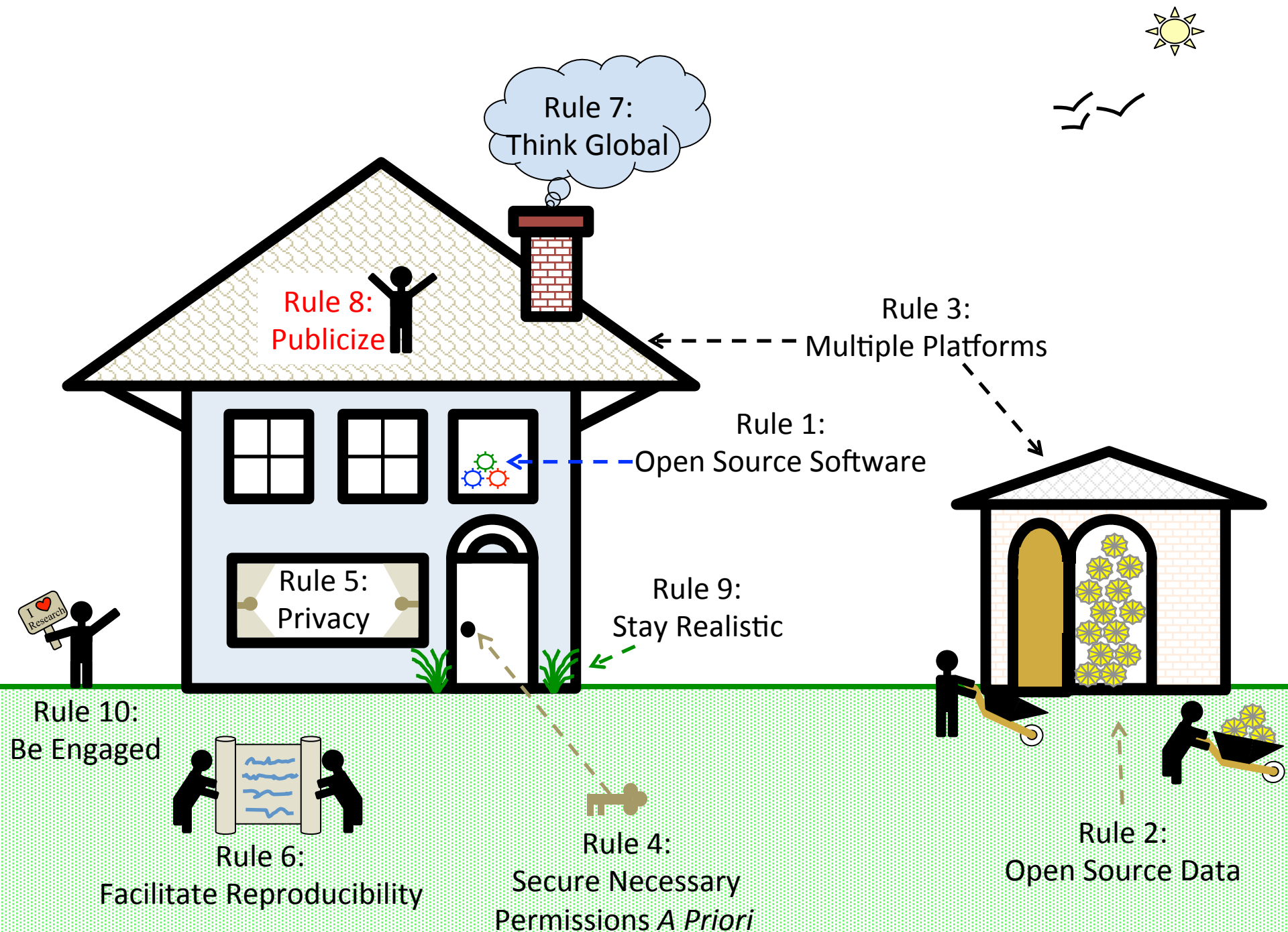
Rule 7: Think Global

- **Conceptual differences** (i.e., country or region-specific medical definitions)
 - The Max Planck Institute for Demographic Research (MPIDR) in Germany collaborated with two separate groups to produce two databases containing international data
 - Required definitional harmonization and cleaning
 - Made available to users in an open format via two specially designed databases:
 - Human Fertility Database (<http://www.humanfertility.org/cgi-bin/main.php>)
 - Human Mortality Database (<http://www.mortality.org/>)
 - Only cleaned data are returned to users in a standardized format, allowing users to easily compare countries with one another
 - Provide detailed descriptions of how they harmonized various timescales across countries in a methods document (<http://www.humanfertility.org/Docs/methods.pdf>) that they could have submitted as a research report paper

Rule 7: Think Global

Table 1. Example sources and sites for each of the ten simple rules.

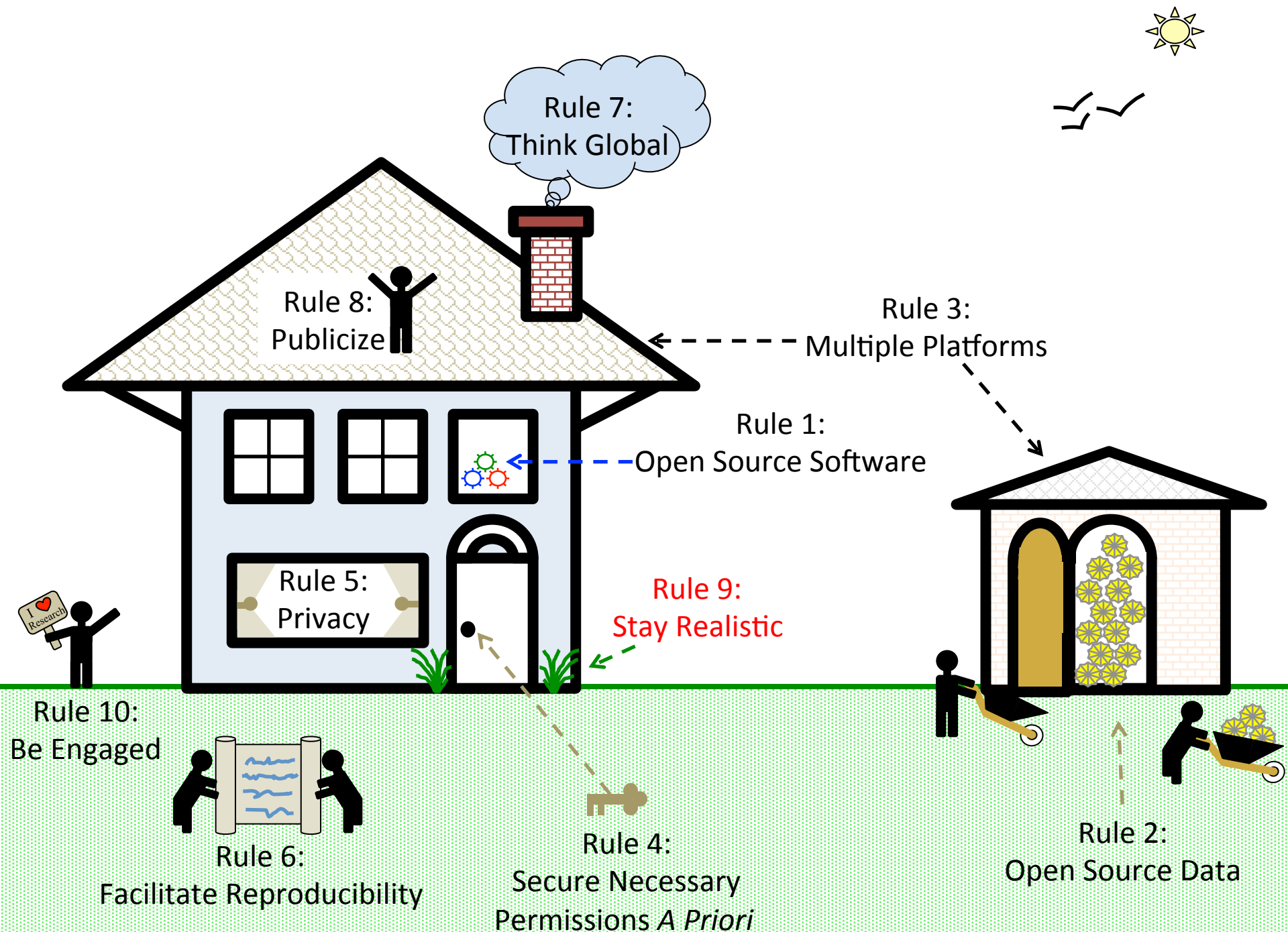
Rule	Example	Site
Rule 7: Think Global		
<i>Guides for Collaborating Globally</i>		
	National Academies “Collaborating with Foreign Partners to Meet Global Challenges” Resources	http://sites.nationalacademies.org/PGA/PGA_041691
	Global Alliance for Genomics and Health	http://genomicsandhealth.org/work-products-demonstration-projects/catalogue-global-activities-international-genomic-data-initiati
	The Global Strategy of the US Department of Health and Human Services	http://www.hhs.gov/sites/default/files/hhs-global-strategy.pdf
<i>Examples of Successful International Projects</i>		
	Human Fertility Database	http://www.humanfertility.org/cgi-bin/main.php
	Human Mortality Database	http://www.mortality.org



Rule 8: Publicize Your Work

Table 1. Example sources and sites for each of the ten simple rules.

Rule	Example	Site
Rule 8: Publicize Your Work		
<i>Research Without Novelty Requirement</i>		
	<i>PLOS ONE</i>	http://journals.plos.org/plosone
	<i>Scientific Reports</i>	http://www.nature.com/srep
	<i>Cell Reports</i>	http://www.cell.com/cell-reports/home
<i>Data Resources (Web Browsers, Databases)</i>		
	<i>Scientific Data</i>	http://www.nature.com/sdata
	<i>Database</i>	https://database.oxfordjournals.org
<i>Pure Open Science Research (all data must be open)</i>		
	F1000	https://f1000research.com

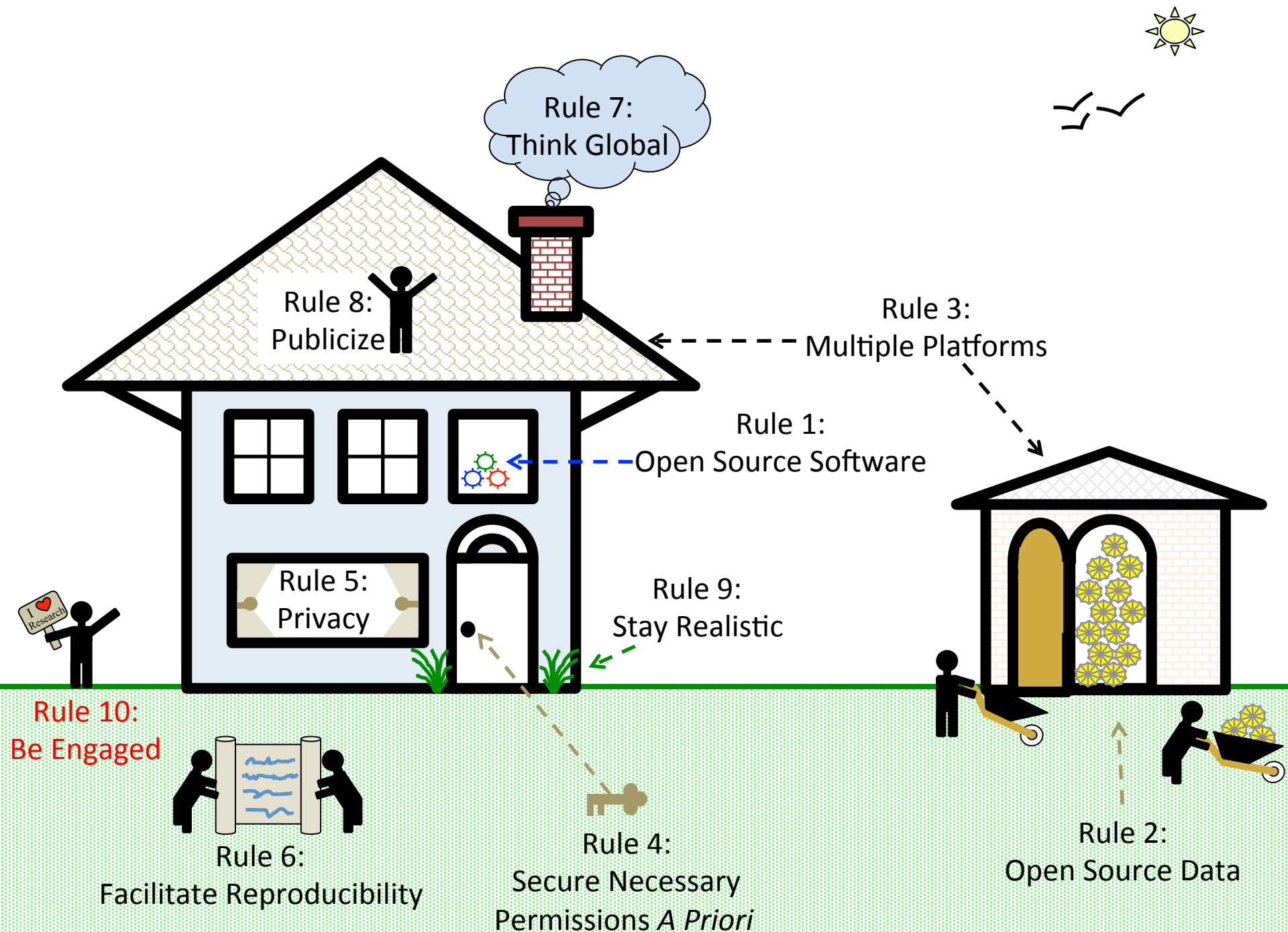


Rule 9: Stay Realistic, but Aim High

- Resist the urge to overstate the claims of your research
 - Stay humble and grounded in reality
- Share as much as possible with the research community to prevent any potential issues
 - The more data you share – the better it will be to avoid any issues after publication
- Don't be afraid to challenge the status quo
 - Pangea

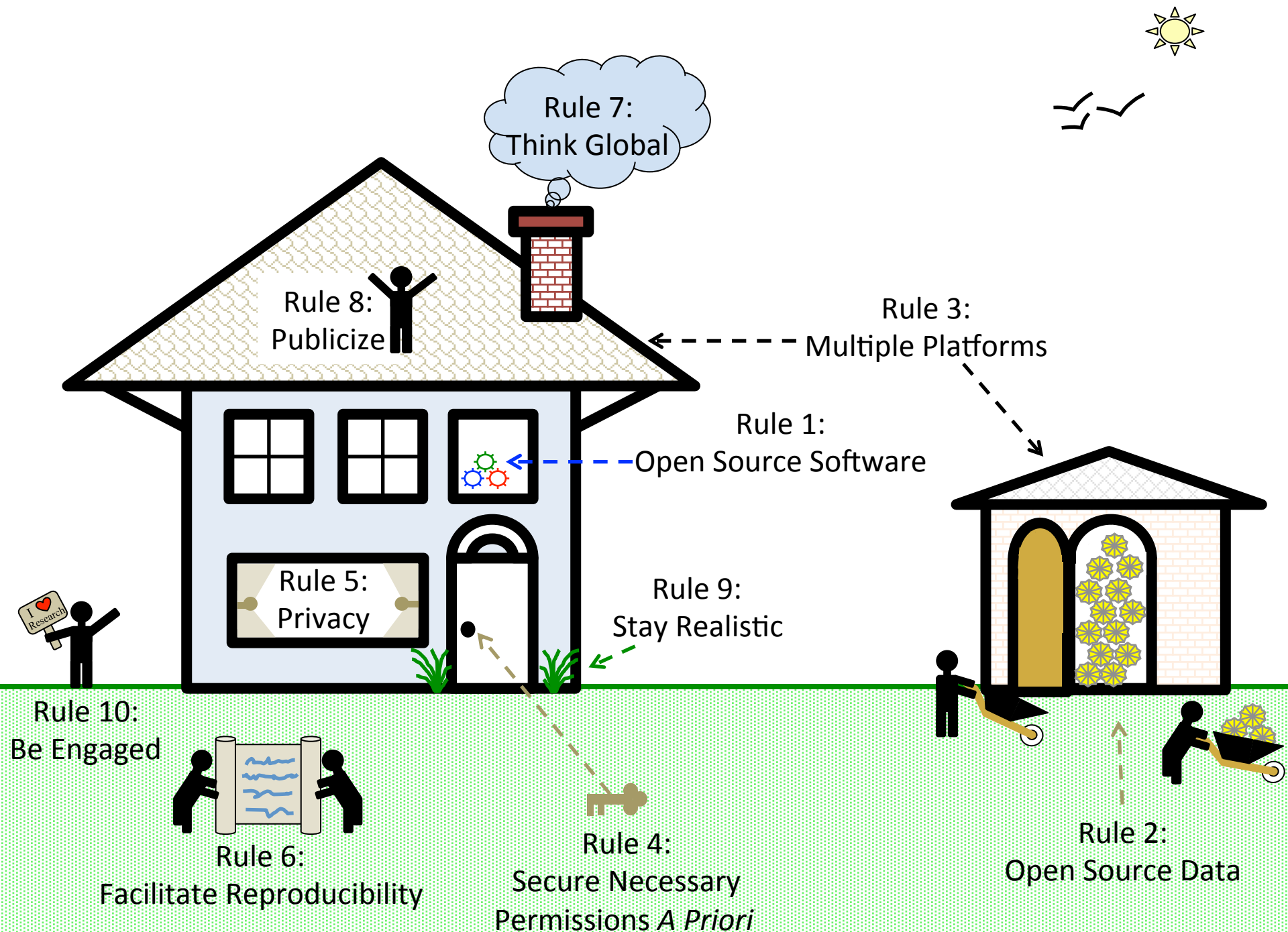
Rule 9: Stay Realistic, but Aim High

Rule	Example	Site
Rule 9: Stay Realistic		
	Retraction Watch	retractionwatch.com



Rule 10: Be Engaged

- Communicate with them using various software social platforms: Github, figshare, and so forth. Respond readily when users have questions and concerns
- Engage with researchers in non-traditional ways
 - Gear, “swag”
- Bottom line:
 - Care deeply about your research
 - If you care and you make it known that you care deeply about the problem, then it becomes possible to convince others that your research is important



Questions?

Mary Regina Boland
mary.boland@columbia.edu