Intro

INFX 551
Foundation of Data Curation

Information School
UNIVERSITY of WASHINGTON

Agenda

- Introductions
- Taking a class with me
- Assignments / Grades
- Topics
- Activity
- Lecture
- Discussion

Introduction

- Who are you?
- What are you interested in?
- What’s your career goal / dream job?
- What’s your grad school pet peeve?

Some important things about taking a class with me….

- I try to create a ‘choose your own adventure’ style to each topic. I encourage you to embrace this.

- I like to experiment. Sometimes experiments blow up.

- I am a very easy grader, but a stickler for details.

Class pact

- You have to try.
- You have to be curious
- You have to be comfortable with (or at least tolerate) uncertainty
- You have to embrace the emergent properties of learning with other human beings
- You have to agree to never ever under any circumstance show someone a data lifecycle diagram.

Topics

1.1 Introduction
1.2 Open & FAIR Data

2.1 Repositories
2.2 Metadata + Documentation
2.3 Preservation

3.1 Transparency
3.2 Privacy & Ethics
3.3 Quality Assurance

4.1 Sustainability
4.2 Wrap-up

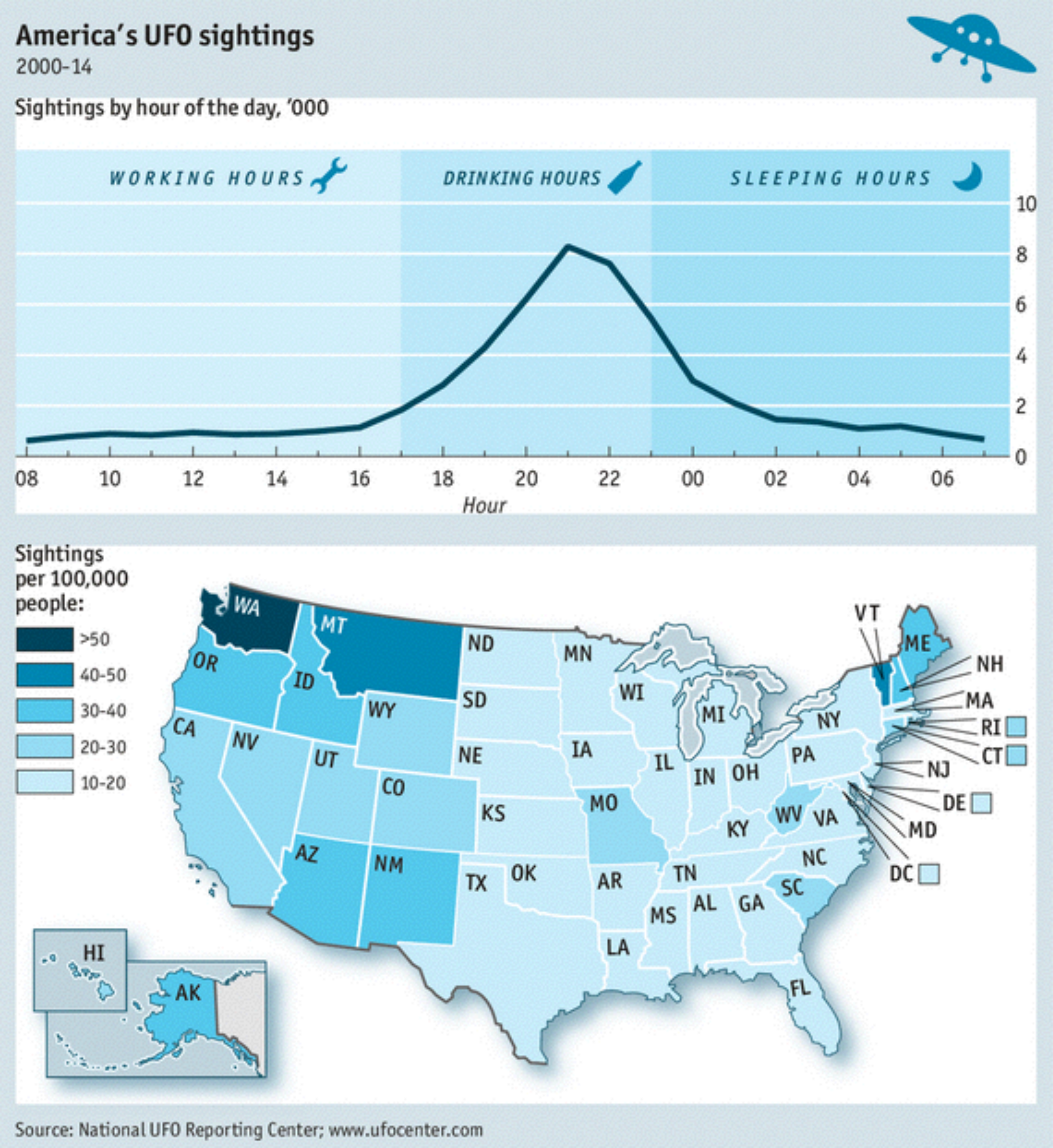
Assignments

- Data Paper (25%) - April 27
- Use Case (25%) - May 11
- Curation Protocol (40%) - June 8

Enough of me talking.
Let’s do an activity.

UFO Data…

1. What are the data?
2. How are data collected?
3. What interesting questions might you ask of this kind of data?
4. What needs to be done to the data in order to answer those questions?
5. How might you communicate your findings?



America's UFO sightings

2000-14

Source: National UFO Reporting Center; www.ufocenter.com

https://www.theatlantic.com/technology/archive/2014/06/the-hard-data-on-ufo-sightings-its-mostly-drunk-people-in-the-west/373645/

What do you mean by ‘data’ ?

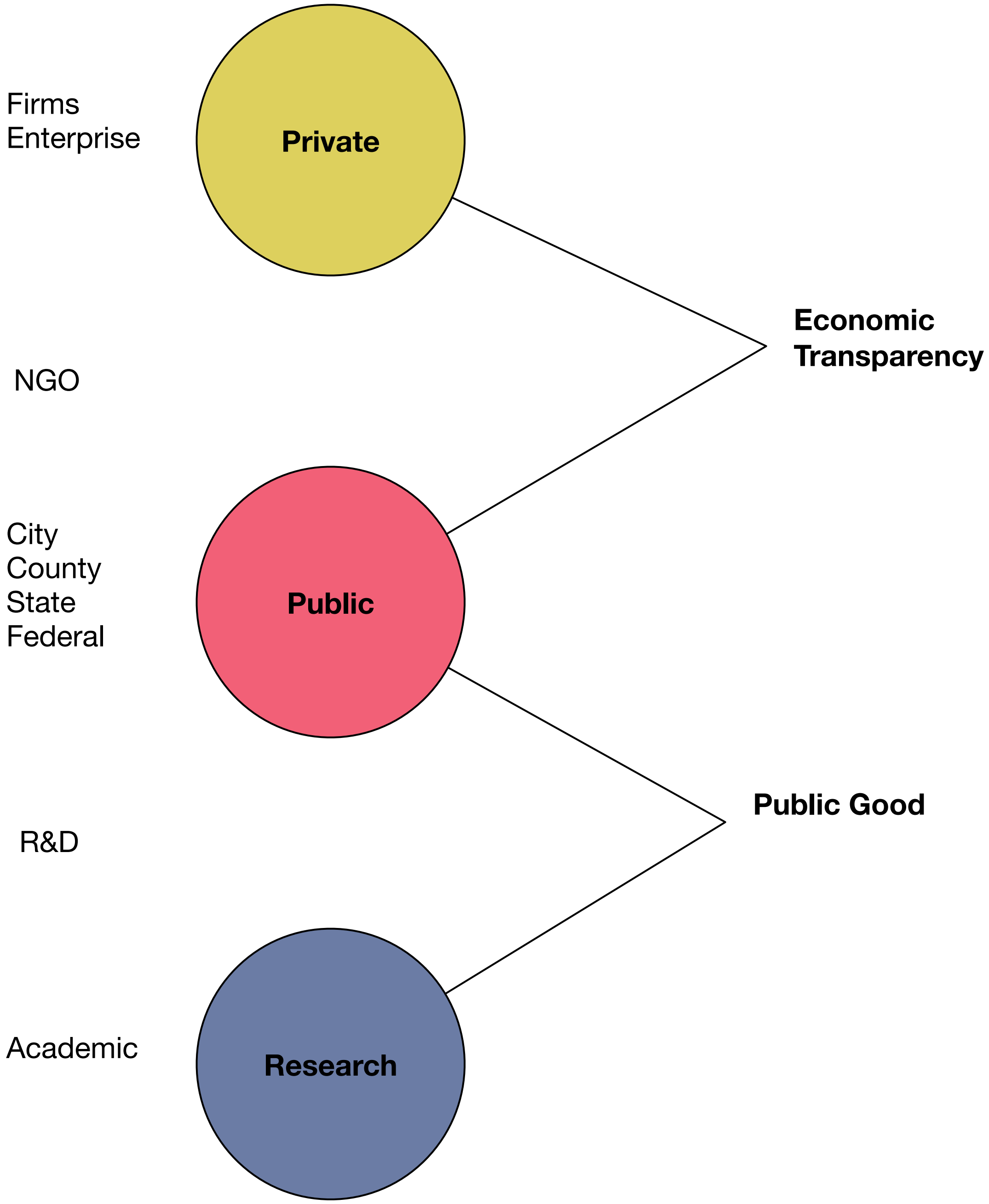
(easily one of the least interesting questions we can ask)

Type vs Role
distinctions

Type:
Donald Trump is a person.

Role:
Donald Trump is POTUS

Types of Data
(by sector)



Firms
Enterprise

Private

NGO

City
County
State
Federal

Public

R&D

Academic

Research

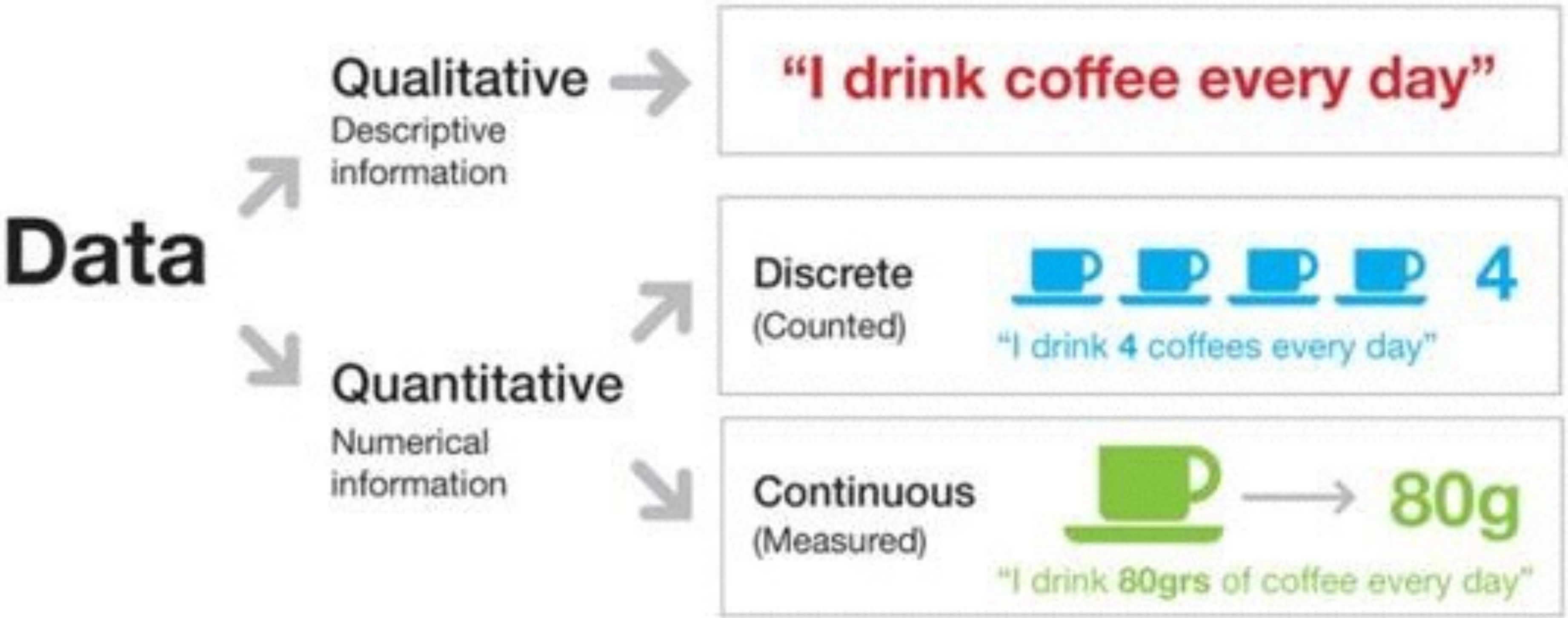
Economic
Transparency

Public Good

Types of Data

(by file format)

XML
XML
Databases
Flat files
Flat Files
EDI
EDI
Excel
XBRL
XBRL
JSON
JSON
WS
Web Services



Data

Qualitative

Descriptive information

"I drink coffee every day"

Quantitative

Numerical information

Discrete

(Counted)

4

"I drink **4** coffees every day"

Continuous

(Measured)

80g

"I drink **80grs** of coffee every day"

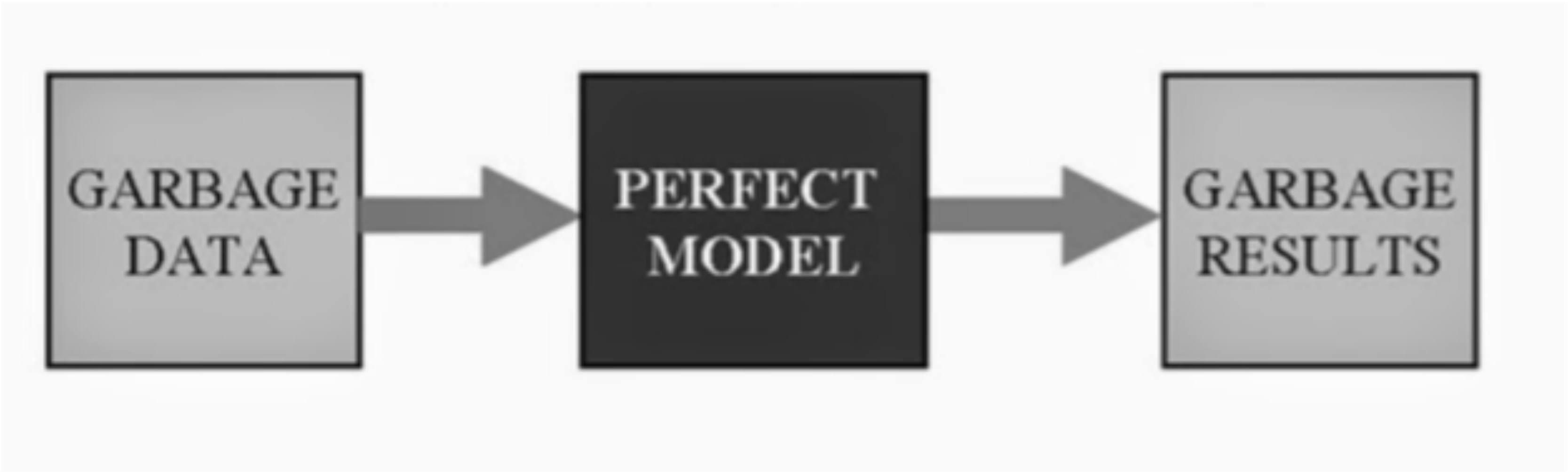
Data Roles

“What do you mean by curation”

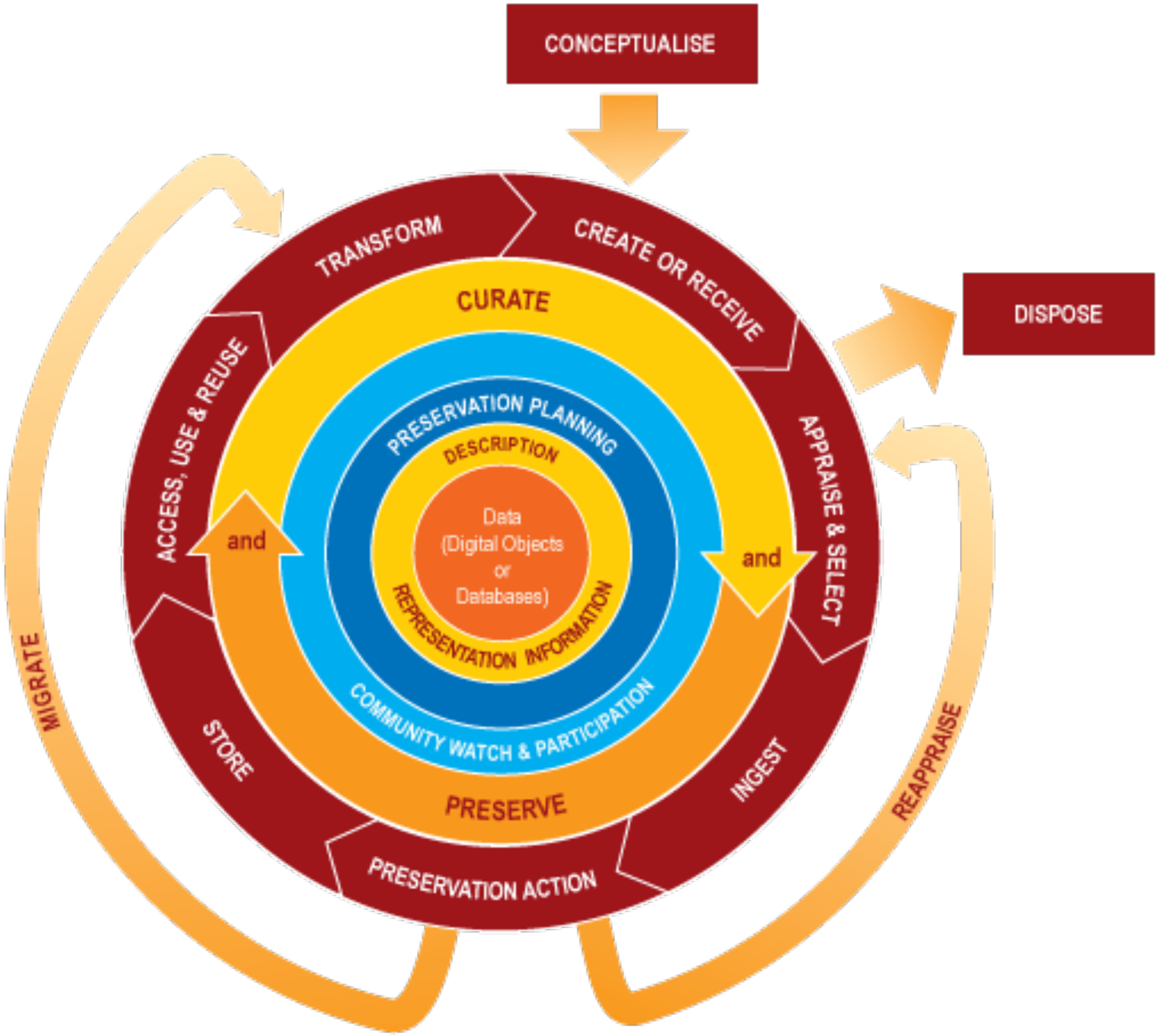
(a more interesting question!)

Data Curation

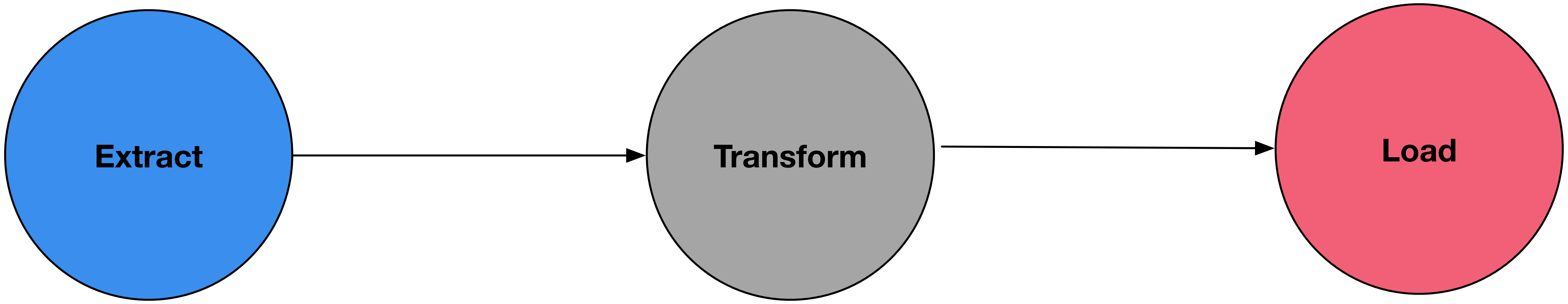
Old computer science saying ”Garbage in = Garbage out”

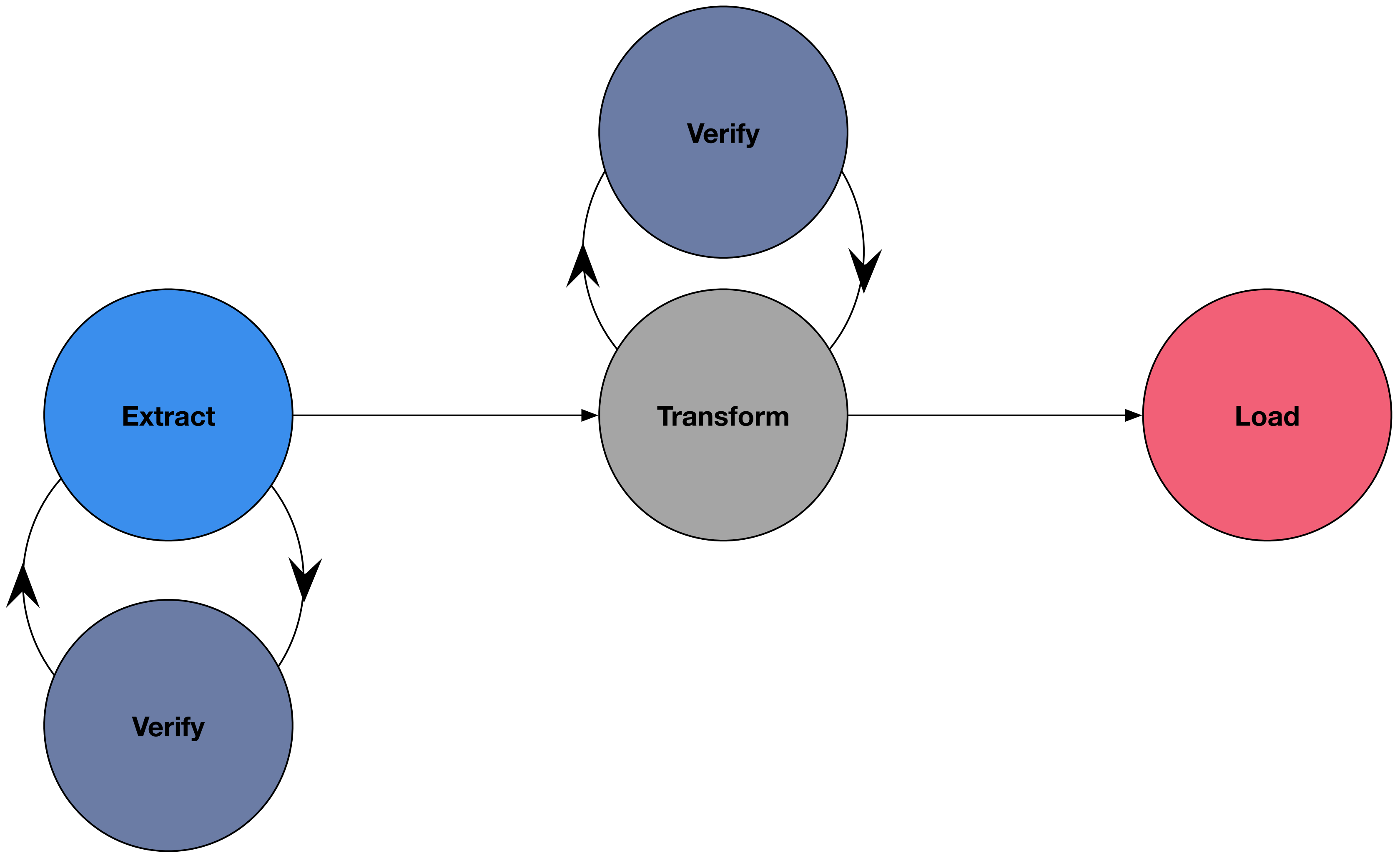


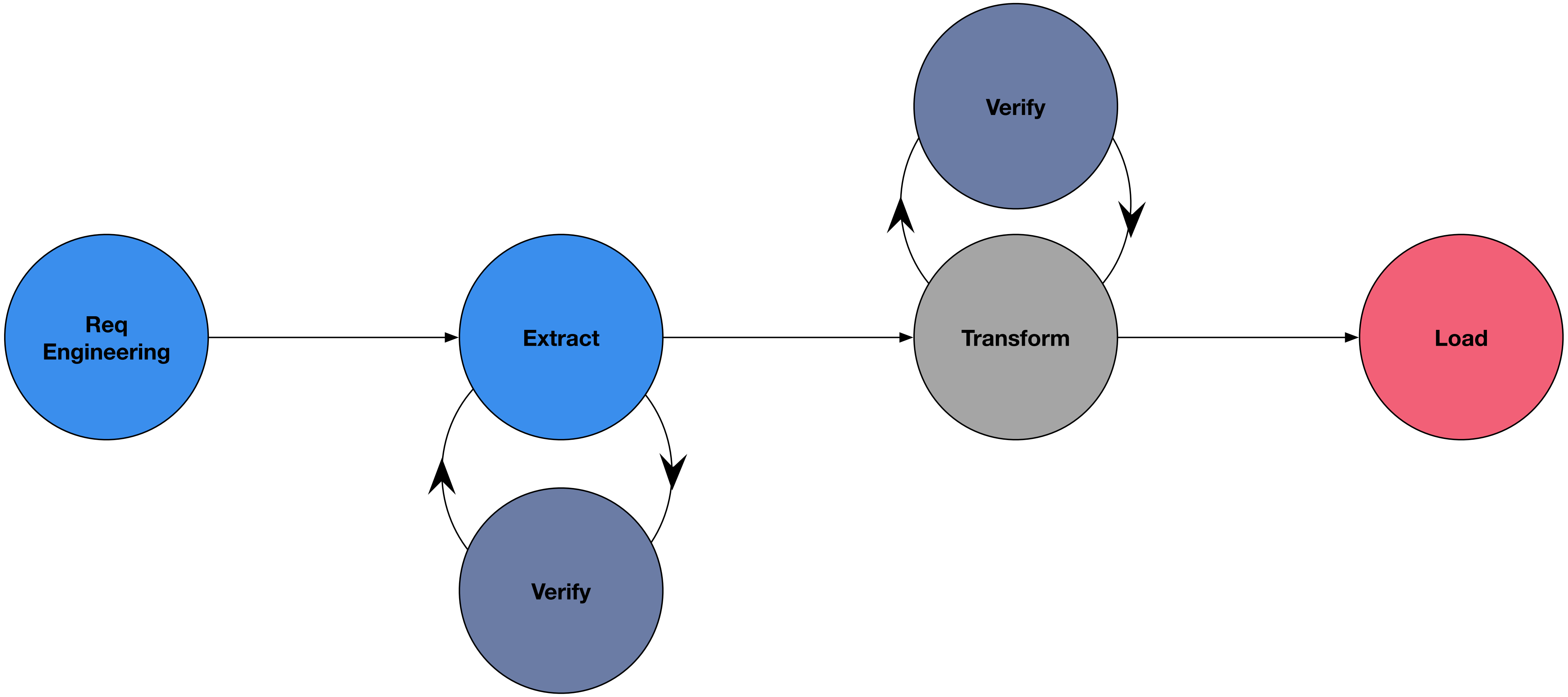
Data curation says “Quality in = Quality out”

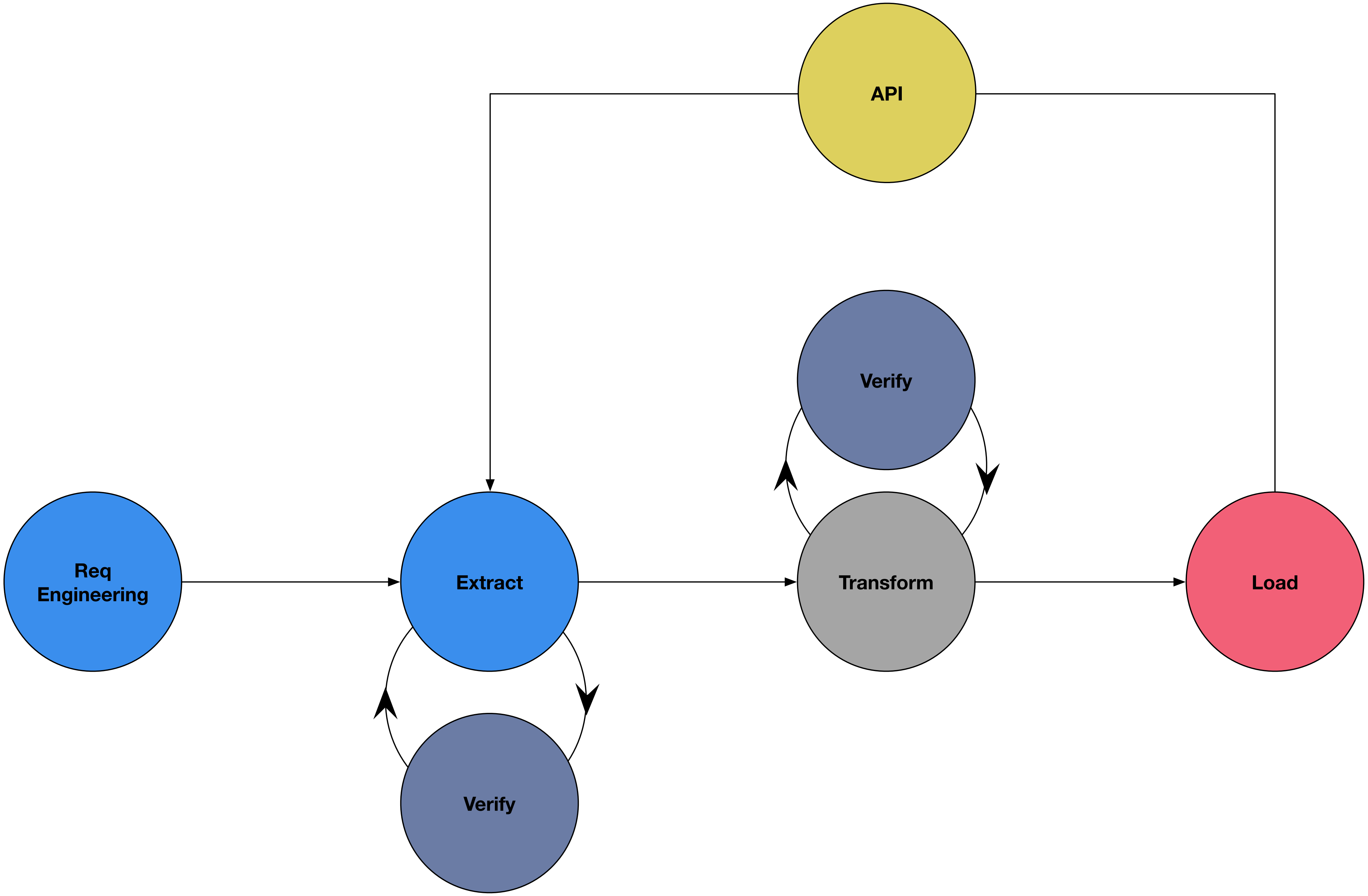

CONCEPTUALISE
DISPOSE
TRANSFORM
CREATE OR RECEIVE
APPRAISE & SELECT
INGEST
PRESERVATION ACTION
STORE
ACCESS, USE & REUSE
CURATE
PRESERVE
and
and
PRESERVATION PLANNING
COMMUNITY WATCH & PARTICIPATION
DESCRIPTION
REPRESENTATION INFORMATION
Data
(Digital Objects
or
Databases)
MIGRATE
REAPPRAISE

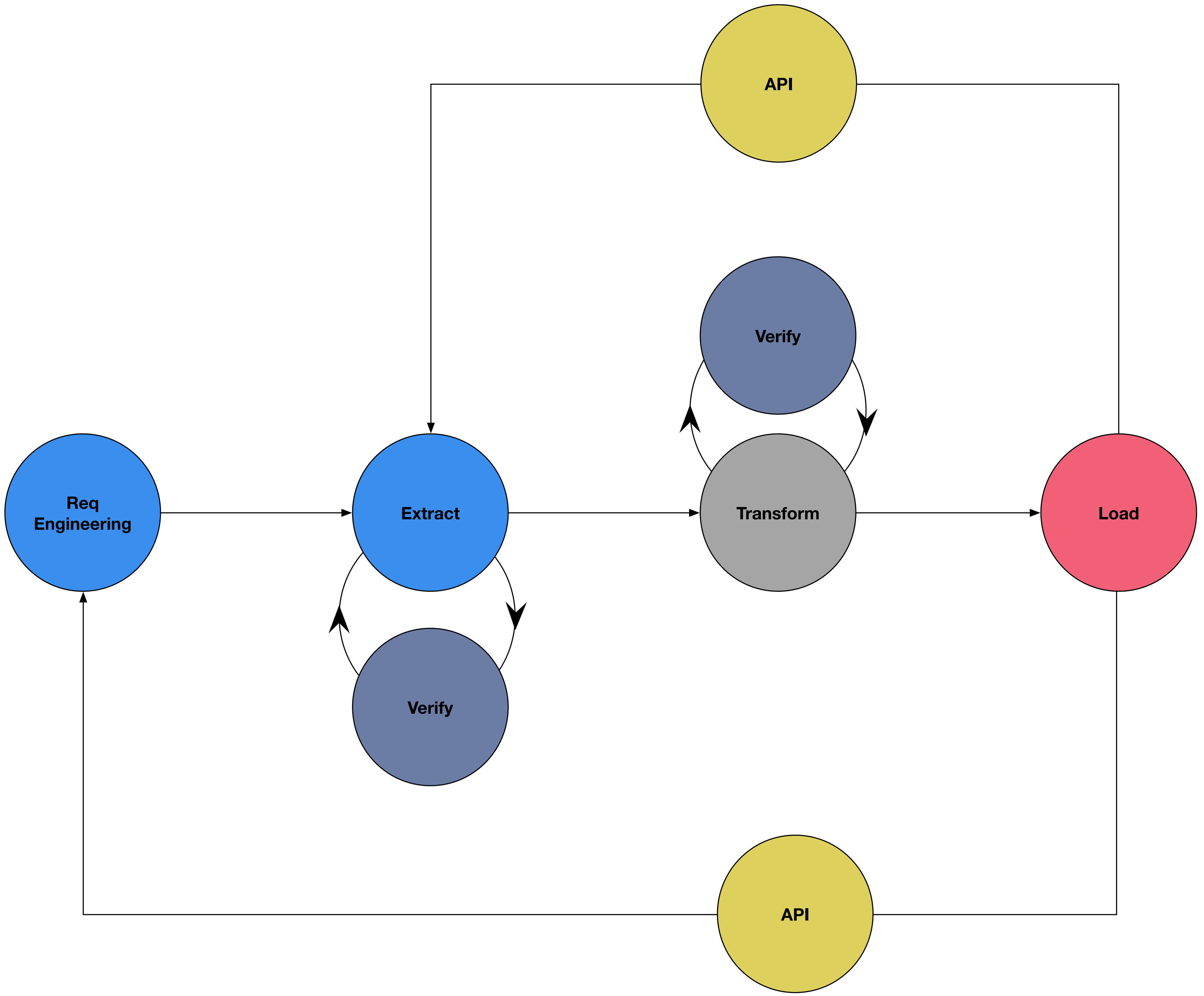
ETL: Simplest form of Curation Workflow

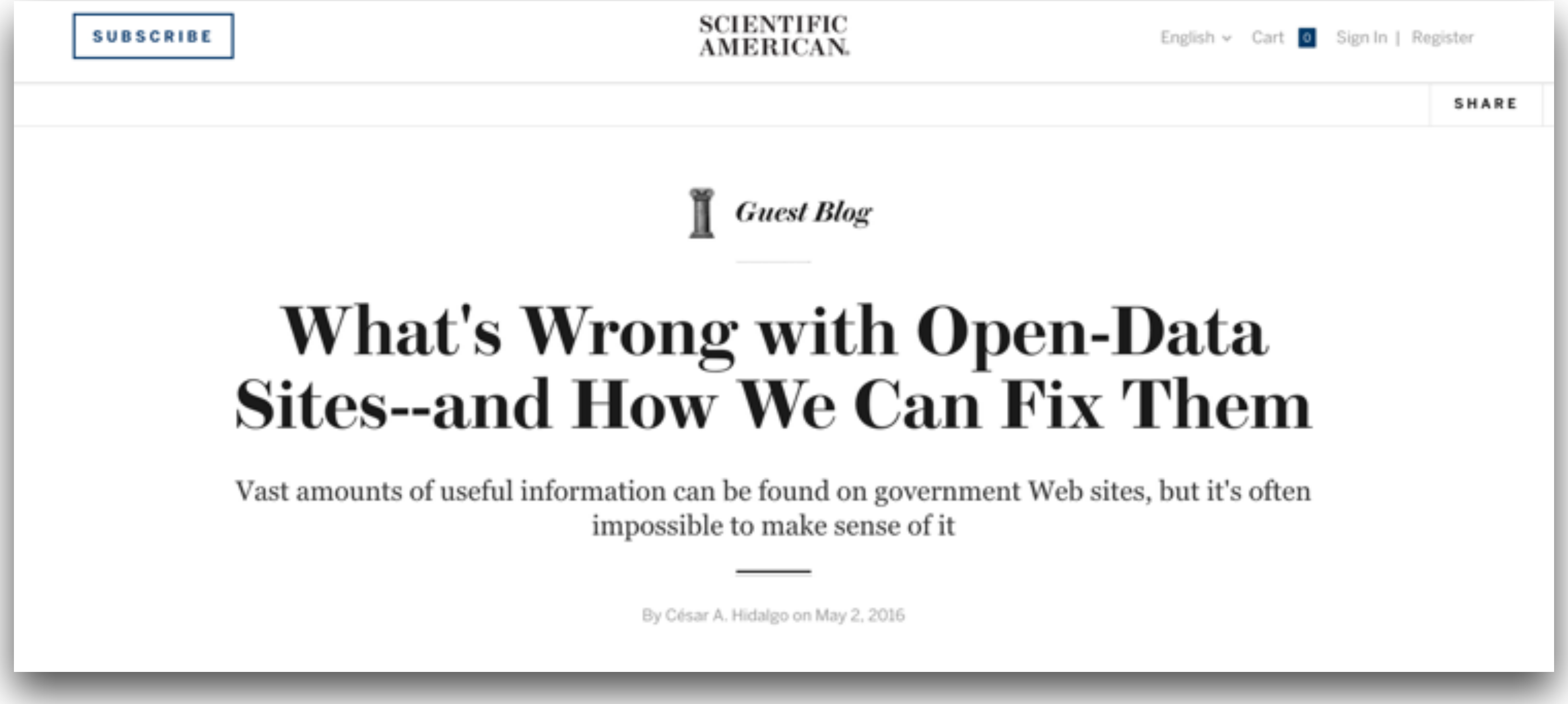



Verify
Extract
Transform
Load
Verify


Verify
Req
Engineering
Extract
Transform
Load
Verify


API
Verify
Req
Engineering
Extract
Transform
Load
Verify


API
Verify
Req Engineering
Extract
Transform
Load
Verify
API



Guest Blog

What's Wrong with Open-Data Sites--and How We Can Fix Them

Vast amounts of useful information can be found on government Web sites, but it's often impossible to make sense of it

By César A. Hidalgo on May 2, 2016

Next Week:

Open + Fair Data

Guest Talk: David Doyle

Data Pitch

Git + Github Lab Activity