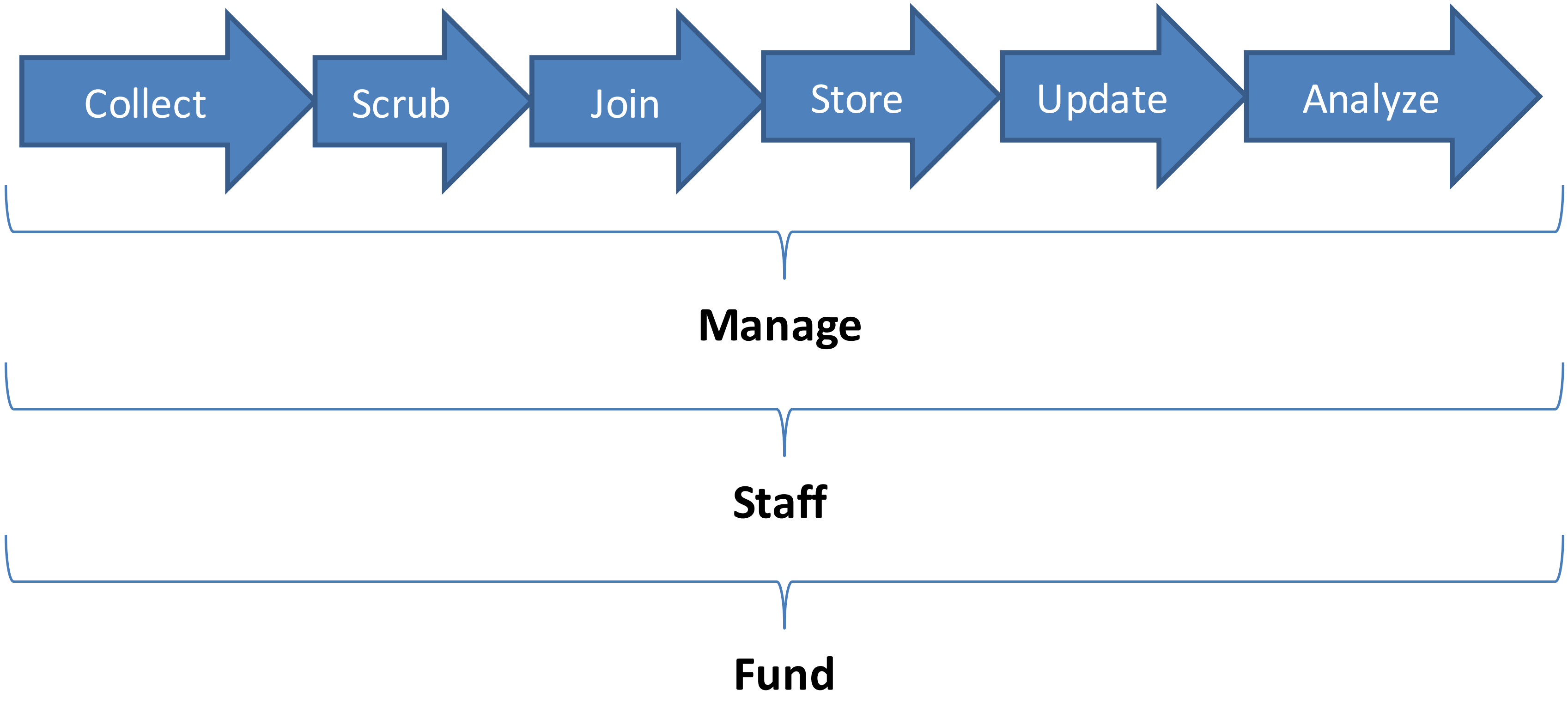
The joy of data cleaning

McGill’s 9 easy steps of ecoinformatics



Collect

Scrub

Join

Store

Update

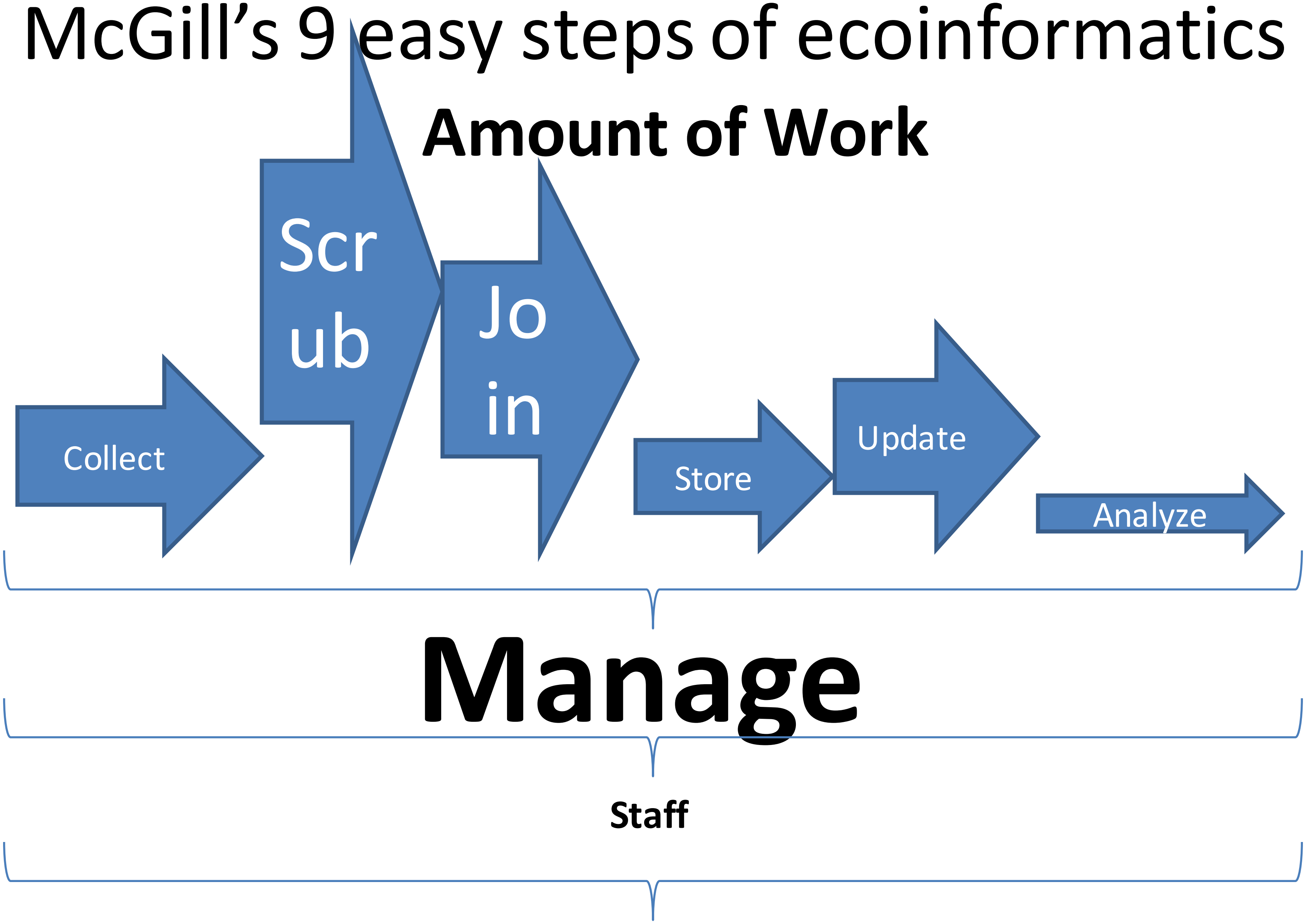
Analyze

Manage

Staff

Fund

McGill’s 9 easy steps of ecoinformatics



Amount of Work

Collect

Scrub

Join

Store

Update

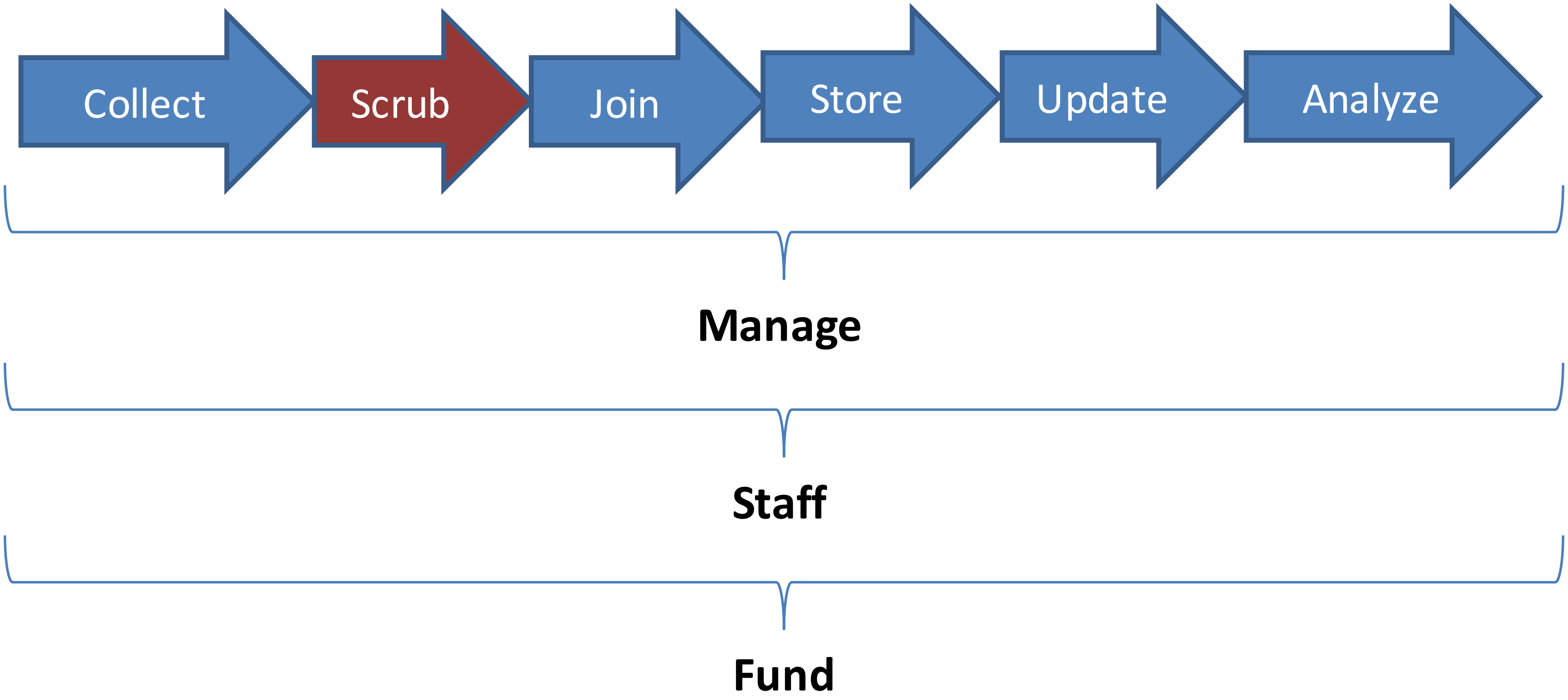
Analyze

Manage

Staff

Fund

McGill’s 9 easy steps of ecoinformatics



Gartner Group

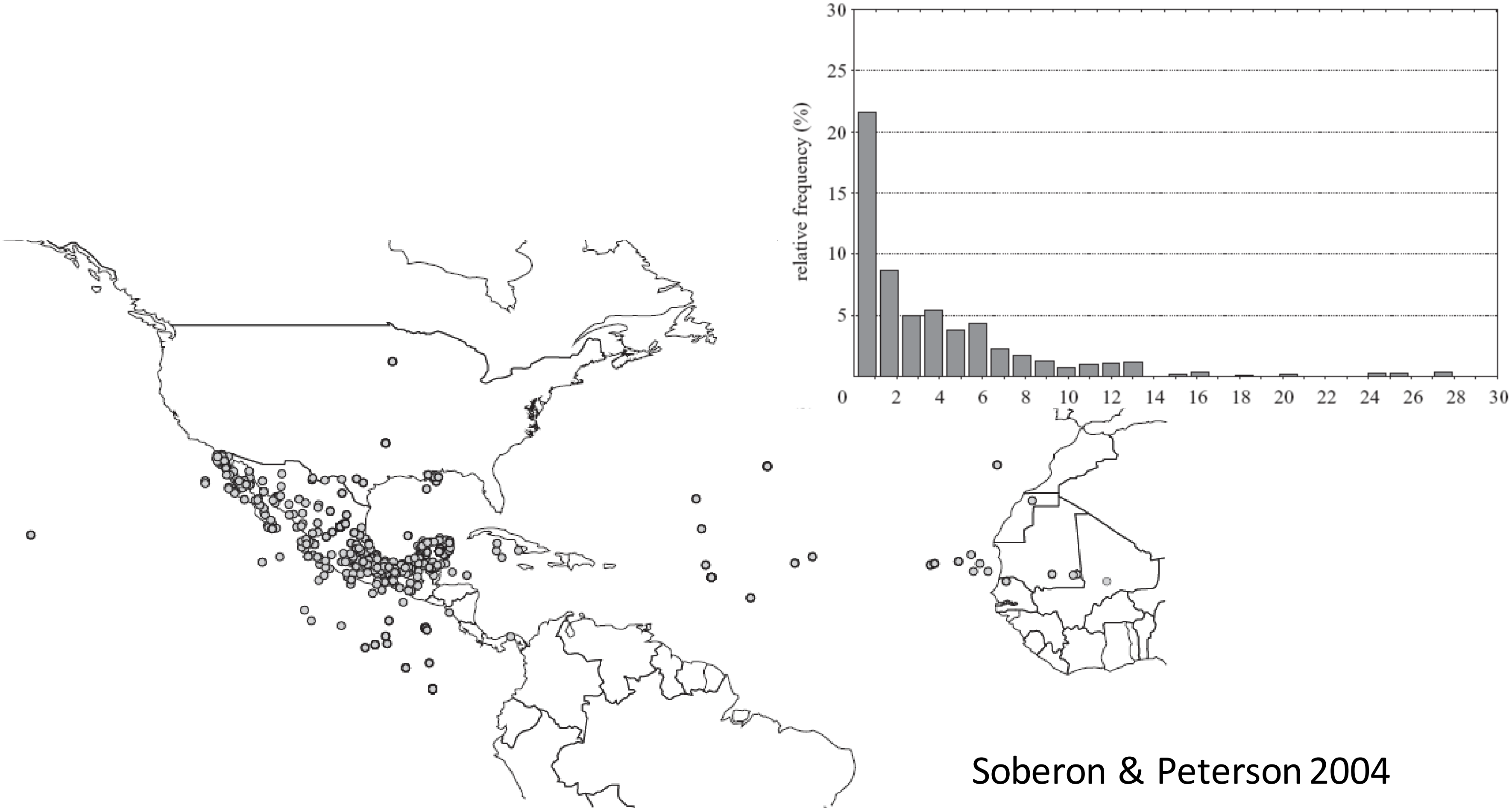
70% of datawarehousing

is in data preparation

4 dimensions

- Values
  - 100 cm of rainfall yesterday
  - 0 for NA
  - 1.00 vs 10.0 (transcription errors)
  - Instrument errors
  - Data filling?
- Space
  - Geocoding (Convention center Baltimore→39.2883N, 76.6181W)
  - Geoscrubbing
    - 42,100 for North America
    - 100, 42 for North America
    - 0,0
    - State centers
- Time
  - Best tools, but amazing how often 6/14/2015 vs 2015/6/14
- Taxonomy
  - Misspellings
  - Synonymy

Synonyms and errors



Soberon & Peterson 2004

Taxonomic Scrubbing in BIEN

- 2.5M records→ 600,000 “species” in New World!
- 600,000 names→300,000 standardized names after synonymy and misspelling (fuzzy matching)

- TNRS service

  Boyle et al 2013

Geoscrubbing - Synonymy

MEX	ISO 3166-1 alpha-3
MX	ISO 3166-1 alpha-2
Mexico	“official” geonames.org (and gadm.org) name
MEXICO	capitalization insensitive
México	geonames.org alternate name
MÃ©xico	recognizable misencoding of México
M&#233;xico	translatable HTML character code
Mexi	not matched

439 country “names”

62 (14%) unrecognizably misspelled

377 recognized

→ 193 recognized countries (49% synonyms)

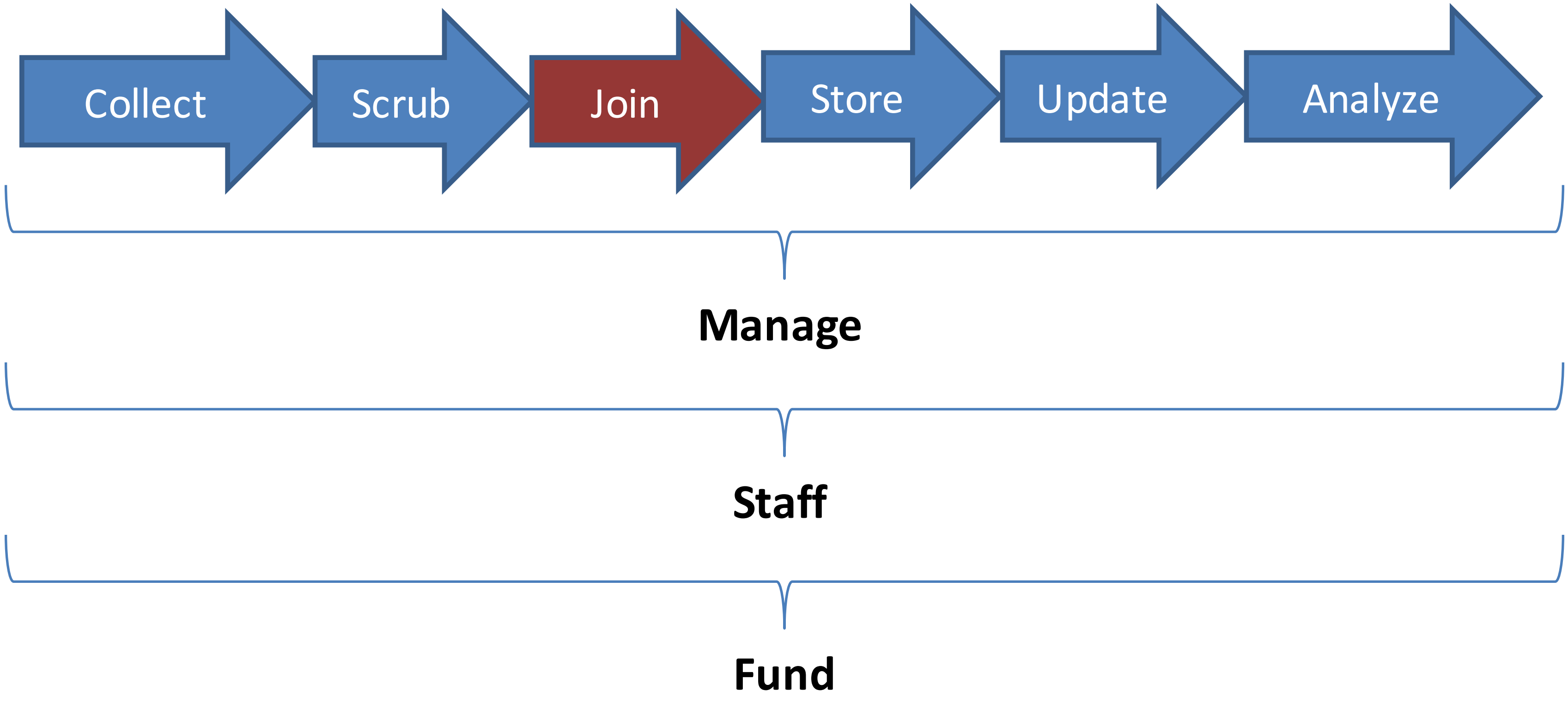
43% canonical names

GeoScrubbing in BIEN

- ~ 1/3 of records had no lat/lon (or 0/0)
- >1/2 had + longitude
- About 1% had other obvious lat/lon errors
- 15% not in the right country
- 25% not in the right state/province

- 67%*85%*75%=41% correct!

McGill’s 9 easy steps of ecoinformatics



Collect

Scrub

Join

Store

Update

Analyze

Manage

Staff

Fund

Joining data

- Cleaning & synonyms
  - *Pinus strobus* in USFIA vs *Pinus strobus L.* in MOBOT
- Semantic joining
  - USFIA has # stems per 0.04 ha plot
  - MOBOT has a specimen card/occurrence
- Record connecting
  - The easy part – databases do this well

Two laws of scrubbing & joining

- Gartner’s law #1 – expect it to be 70% of your work
- McGill’s law #2 – expect ~50% of the data to be wrong