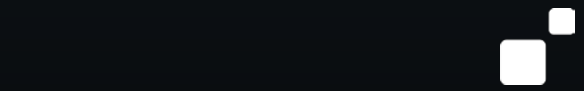
MARKLOGIC WORLD 19

Search and the MarkLogic Data Hub

Mary Holstege
Distinguished Engineer

Innovation



Unknown, taken in 1862, public domain

A journey of discovery

Innovation comes from change

- Change begets innovation
  - Response to disruptions of environmental conditions or relationships
- Innovation begets change
  - Creates disruptions to environmental conditions or relationships

John Gerrard Keulemans, public domain

The power of small iterated changes:

w∆z

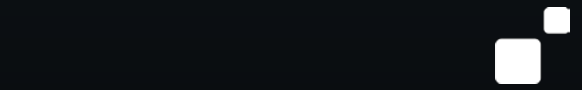
Uwe Gille, https://commons.wikimedia.org/w/index.php?curid=48277827

It is easier to survive small failed experiments

Aaron Logan http://www.lightmatter.net/gallery/Animals/panda

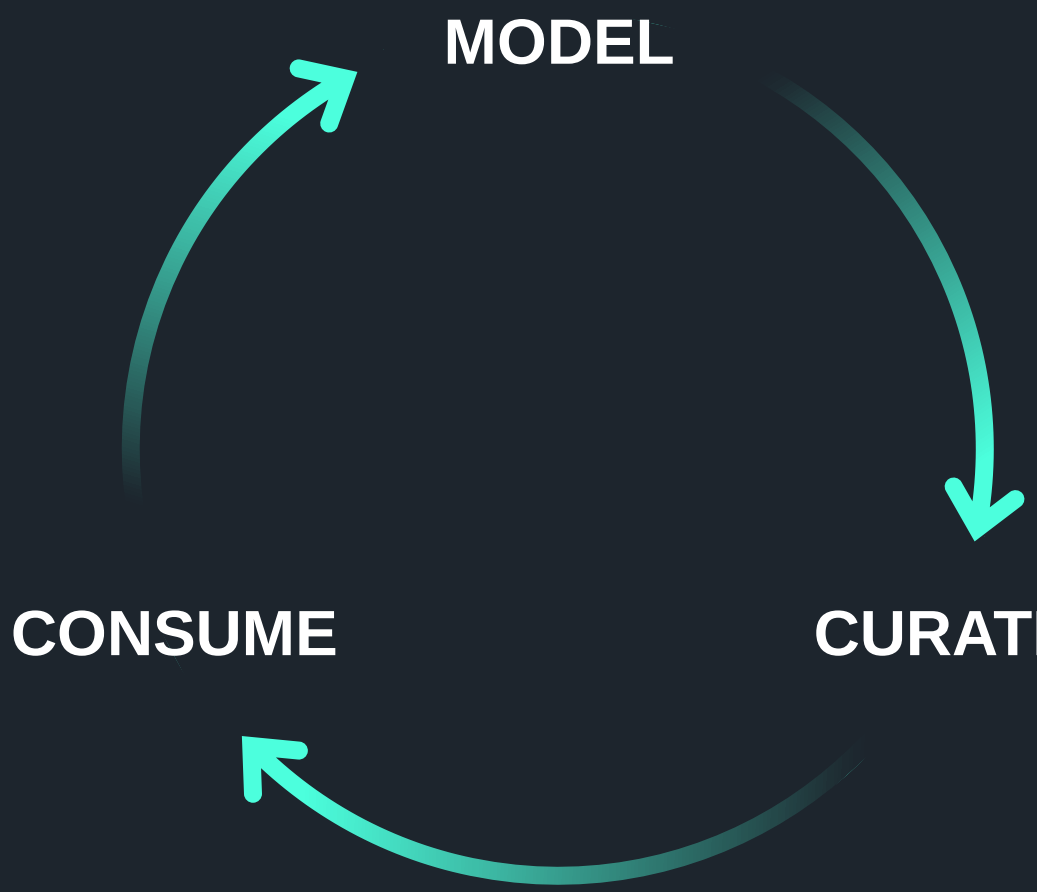
Data Hubs and Search

What does search give a Data Hub?



Platform for innovation

Cycle of innovation



The search perspective

M.C. Escher

Universal

- The data defines the index
- The data defines what questions can be asked

Natural History Museum, London

Chip Clark, SI, NMNH - Smithsonian Institution, public domain

Selective

- Examine just the data you need
- Specialized indexes

Ranked

- Shades of grey: ordered
- Some matches are better
- Measurement of match

Natural History Museum, London

finitely, occur in all classes of the higher animals, and very much in proportion as their mode of life requires them. When concealment is of more importance, then the recognition is made effective by differences of shape or of motions and attitudes, or by special cries, as in the cuckoo. Among the birds of the tropical forests, while the ground colour is often protective, as in the green of parrots, the smaller fruit-

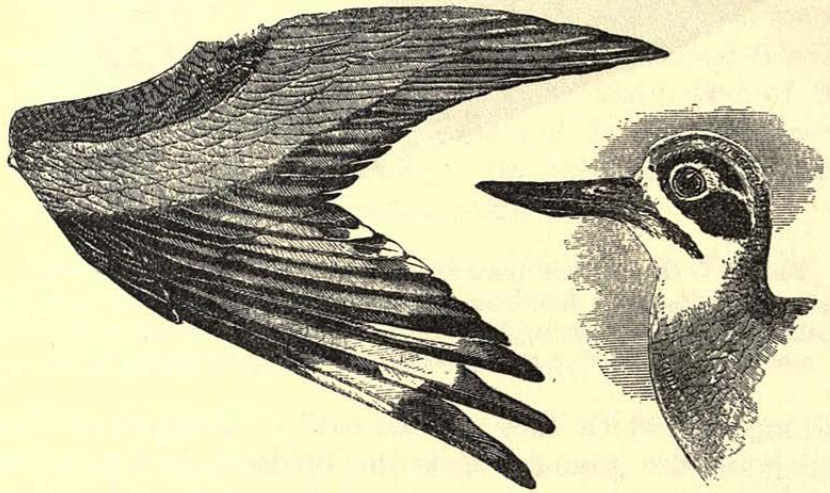


FIG. 38.—*ŒDICNEMUS RECURVIROSTRIS* (Great Indian Stone-Curlew). This species is found all over India, and also in Ceylon and Burma. This species is clearly defined by the upturned bill and the compact black mark around the eye.

pigeons of the Malay Archipelago, many of the barbets, and hosts of other birds, yet the different species will be almost always characterised by spots or bands, or caps of brilliant or contrasted colours. But as these usually break up the green body into irregular portions, and as flowers of equally varied hues are common on trees, or on the orchids and other epiphytes that grow upon their branches, the general effect is by no means conspicuous.

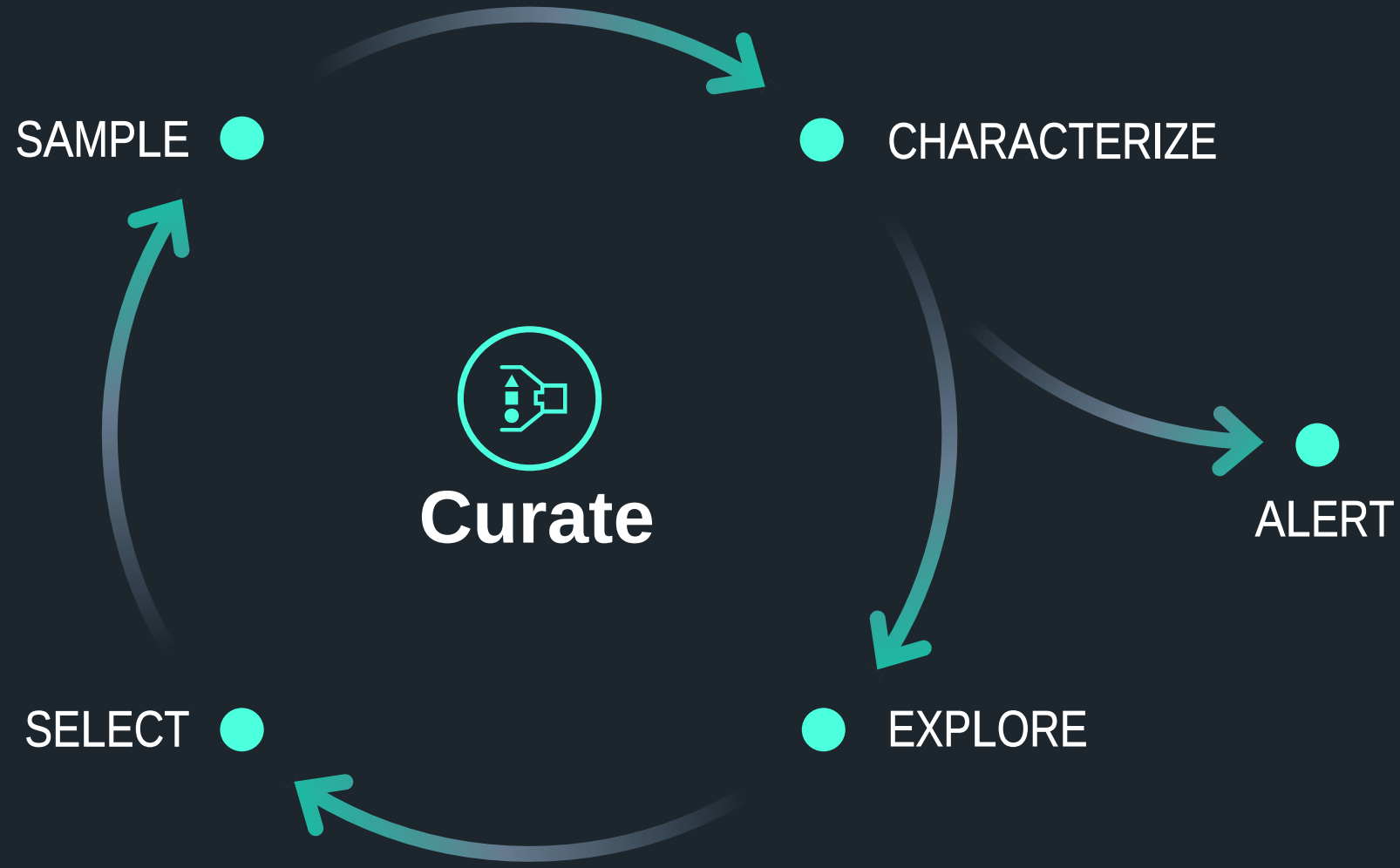
Now, without this principle of the necessity for external differences for purposes of recognition of each species by their own kind, and especially of the sexes by each other,

Wallace, Aflred Russel (public domain)

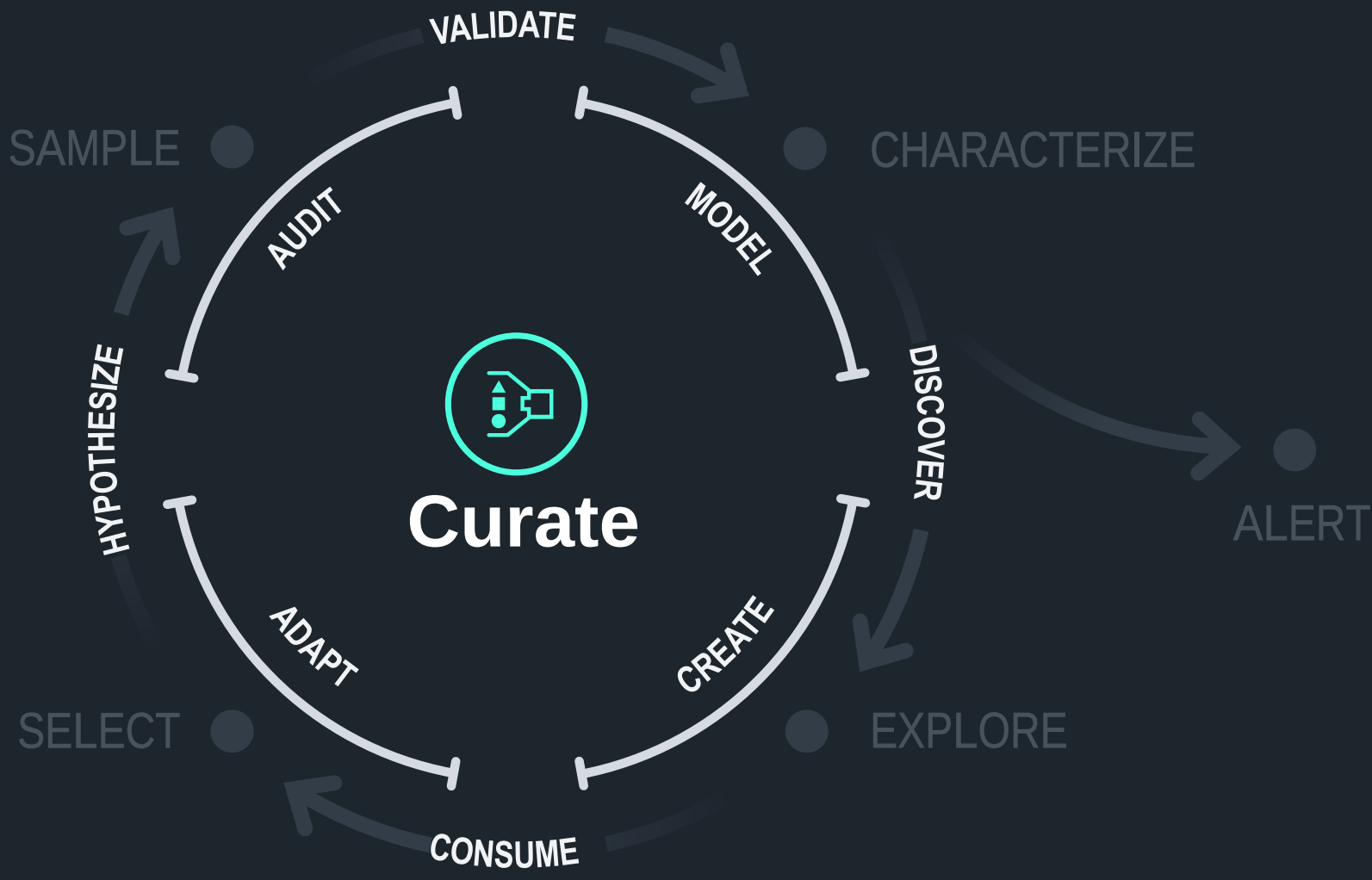
Contextual

- Business records are business documents
- Your next innovation is in the comments
- Narrative, relationships, metadata

Cycle of search activities



Cycle of search activities



Select: Obtain targeted data

- Operational access

- Locate by universal characteristics
- Locate by context
- Ranked matches

Chip Clark, SI, NMNH - Smithsonian Institution, public domain

Sample: Create slices of data

- Modeling and machine learning
- Analytics and data characterization
- Data quality verification
- Auditing

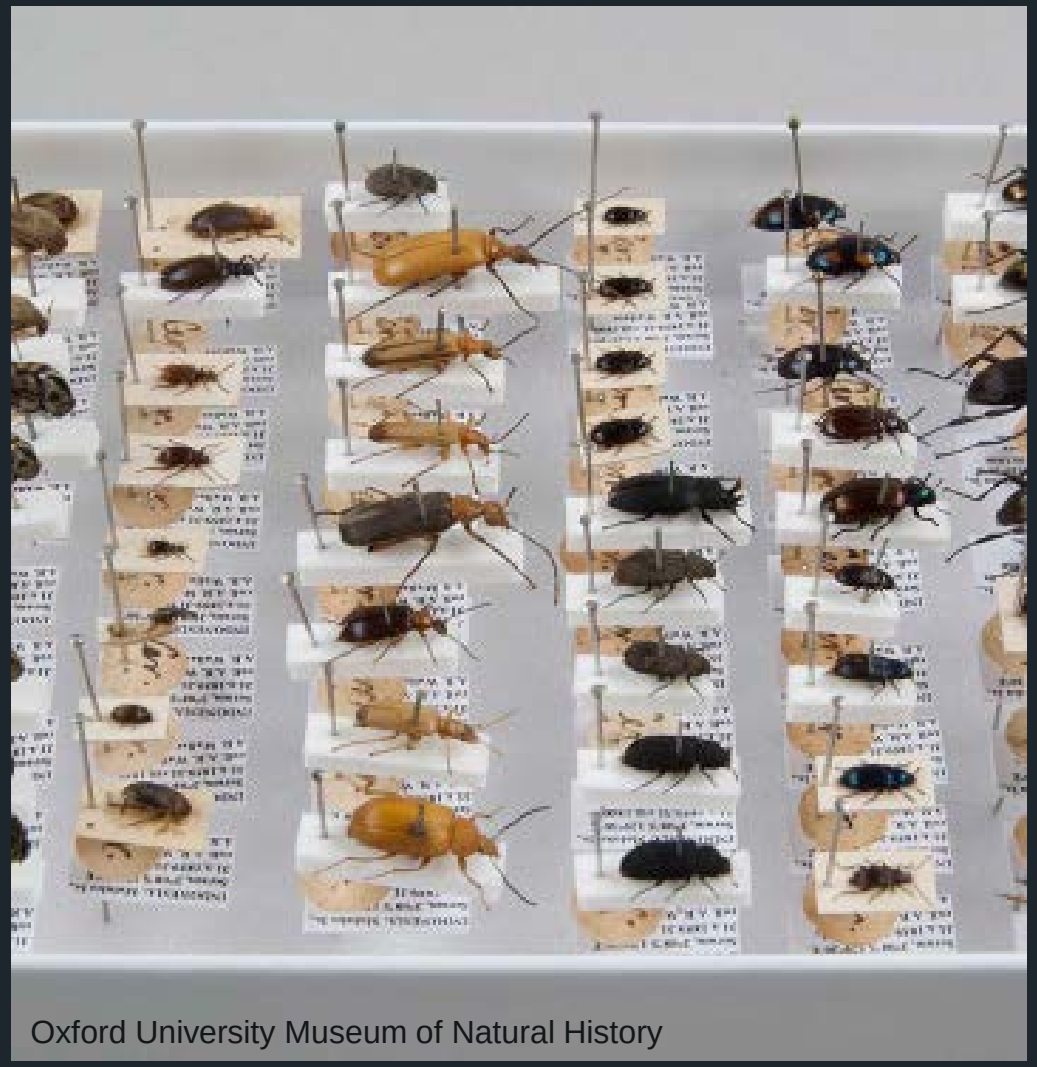
- Select by universal characteristics
- Rank by random score to get random sample

Natural History Museum, London

Characterize: Analyze data slices

- Discovery
- Data quality verification
- Group, cluster, compare
- Data distributions: variability, outliers

- Turn metadata into data



Oxford University Museum of Natural History

Explore: Navigate and investigate data

- Discovery
- Understanding data in context

- Select starting points
- Navigate contextual relationships

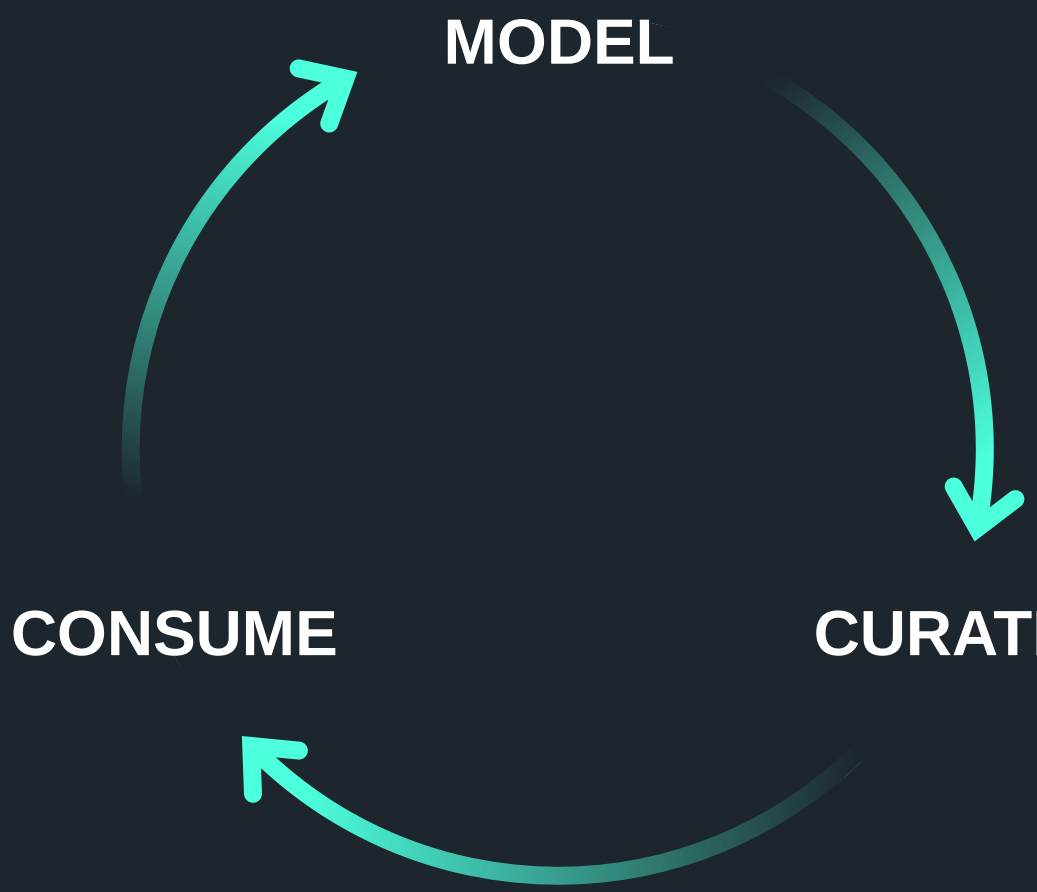
P.F. Siebold – Seibolldcollectie Naturalis (public domain)

Alert

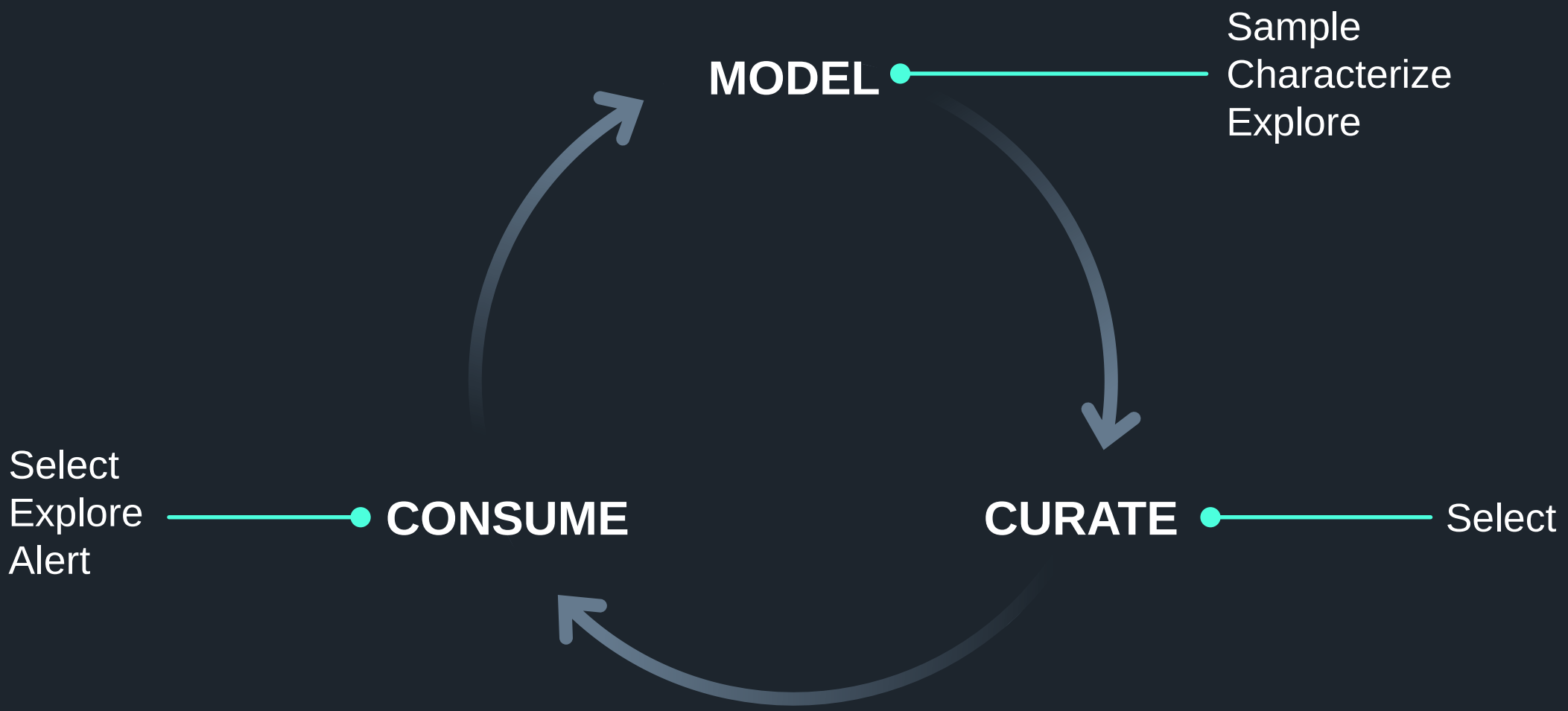
Get notified about interesting data

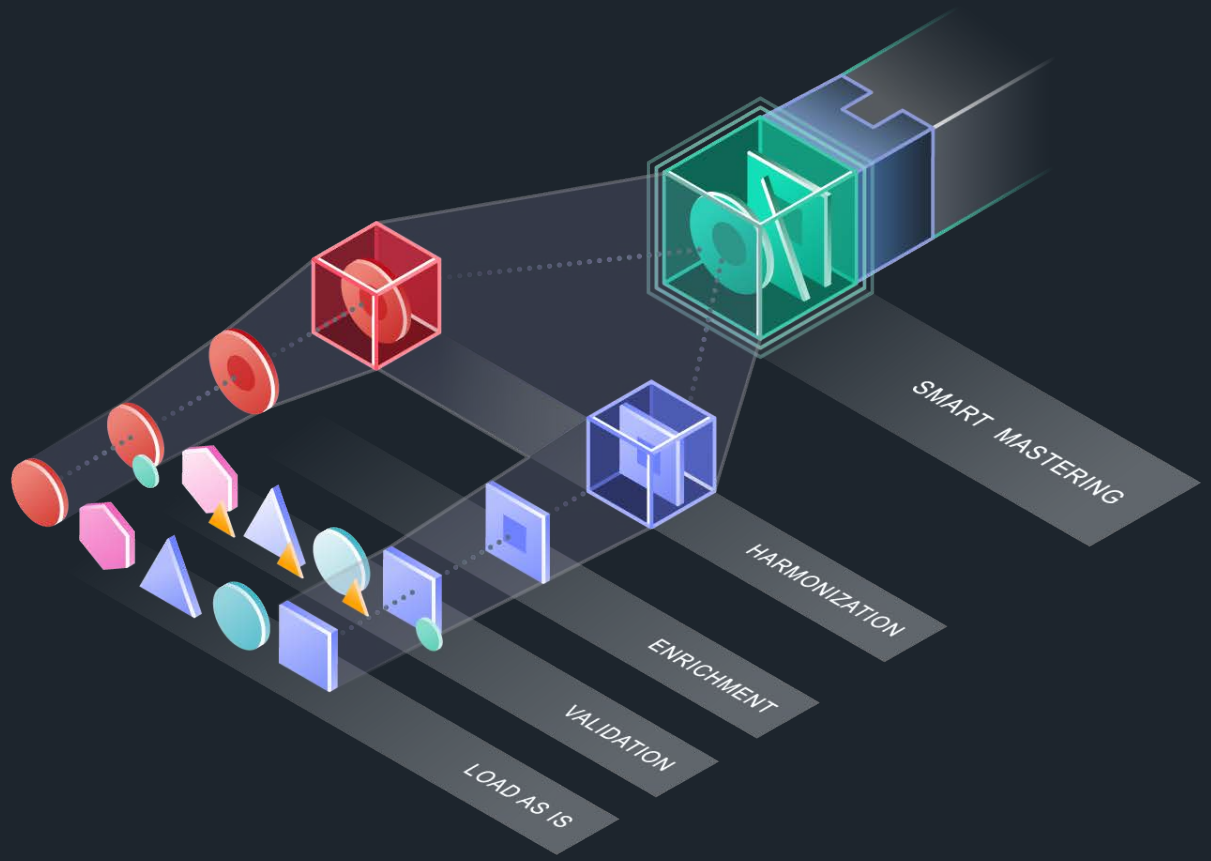
- Get relevant data without asking
- Latest matching any criteria
- Any complex query

Cycle of innovation



Cycle of innovation





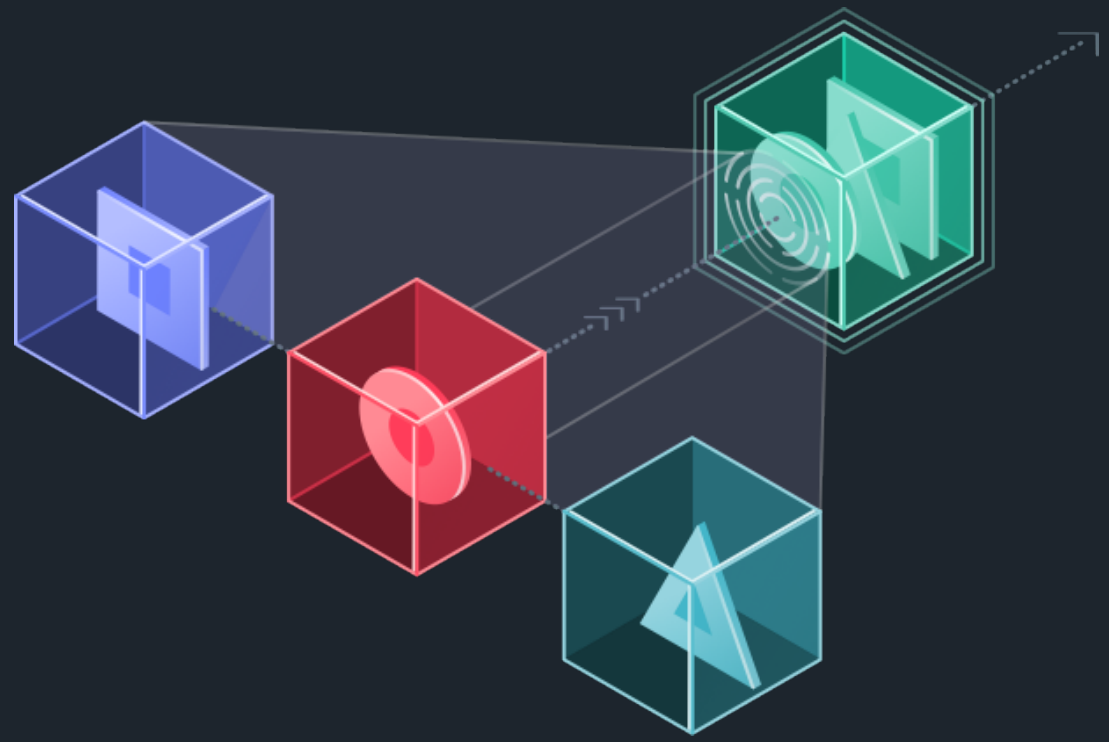
Curated Data

- Harmonized
- Mastered
- Reference data
- Enrichment
- Provenance and lineage

Smart Mastering

- Characterize/explore candidate matches
- Sample training and test sets

- Selection of candidate matches
  - Selective, based on universal criteria
  - Ranked
- Linked to sources, preserving context



Reference data

- Select reference data to consolidate based on universal criteria
  - Maybe ranked matches
- Link to reference data to provide context

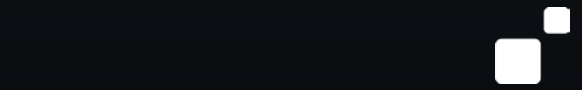
Enrichments

- Classifications, keywords, entity recognition, semantic relations, …
  - Sample, characterize, explore training and test data
  - Select and rank data to apply

- Universal: Make context explicit to enhance selectiveness and ranking

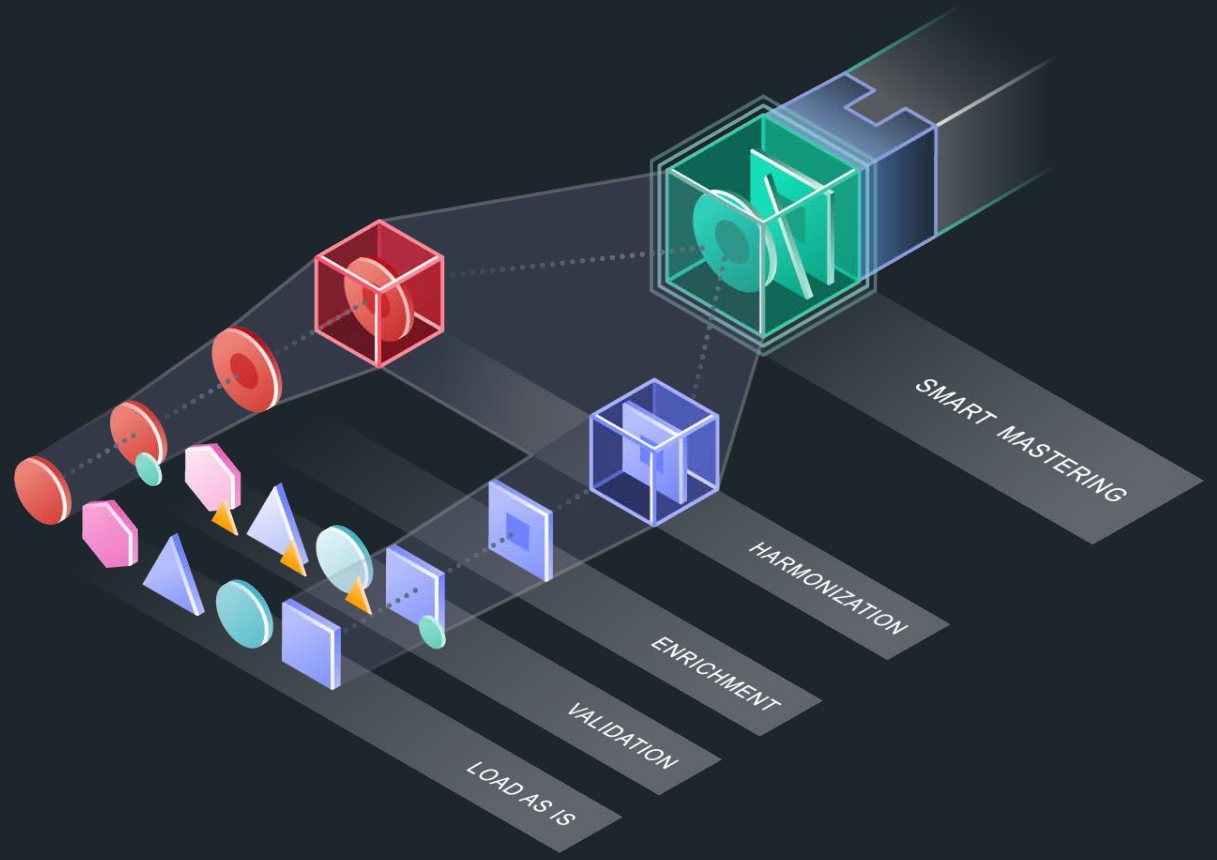
Search and Data Hubs

What does a Data Hub give search?



Why search needs a Data Hub

- Curated data
  - Applying principal of universality
  - Better context
  - Better selectiveness
  - Better ranking



- More relevant results

Curation gives context

Curation means knowing about your data and what it means

Institute for Biodiversity Science and Sustainability

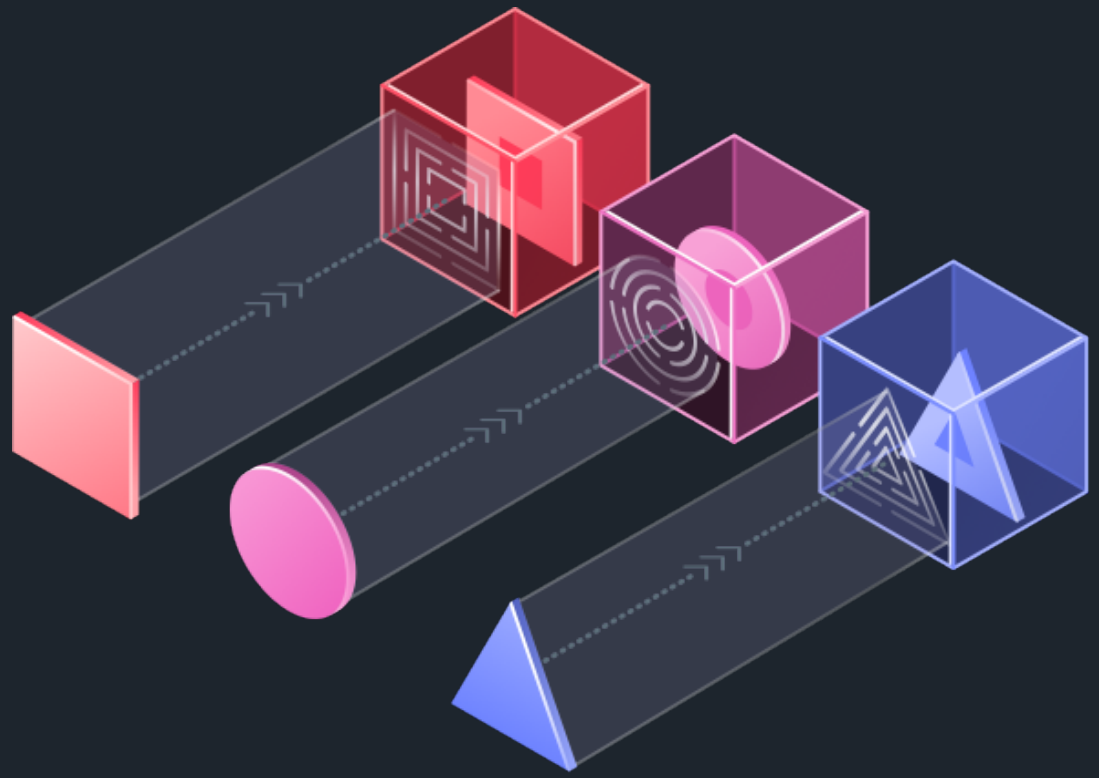
CALIFORNIA ACADEMY OF SCIENCES

CAS » IBSS (Research) » Invertebrate Zoology & Geology » **Search the Collections Catalog**

CAS Invertebrate Zoology Collection Database

CAS-IZ 10264.00	***Anoplodactylus digitatus*** (Boehm, 1879) ARTHROPODA: PYCNOGONIDA: PANTAPODA: Phoxichilidiidae						
Identified By	C.A. Child Nov 1979	Specimen Count	5	Original Fixative	100% alcohol	Preservative	75% EtOH
Acc. Num.	10615	Field Number		Expedition Name			
Collector	Dave Coleman	Collection Date	25 Jan 1975	Collecting Method		Collection Name	
Locality	INDONESIA: Sumatra Island: North Sumatra: Strait of Malacca, near Lhokseumawe, steel pertamina fuel dock pilings, 5 10' N, 97 08' W. DEPTH: 0-12 ft					Lat.,Long.	5°10'N, 97°8'W
Depth	12 ft	Elevation		Substratum		Intertidal	No
EM Mounts		Photos		Micro. Slides			
Publications							

Close and Return



Harmonized Data

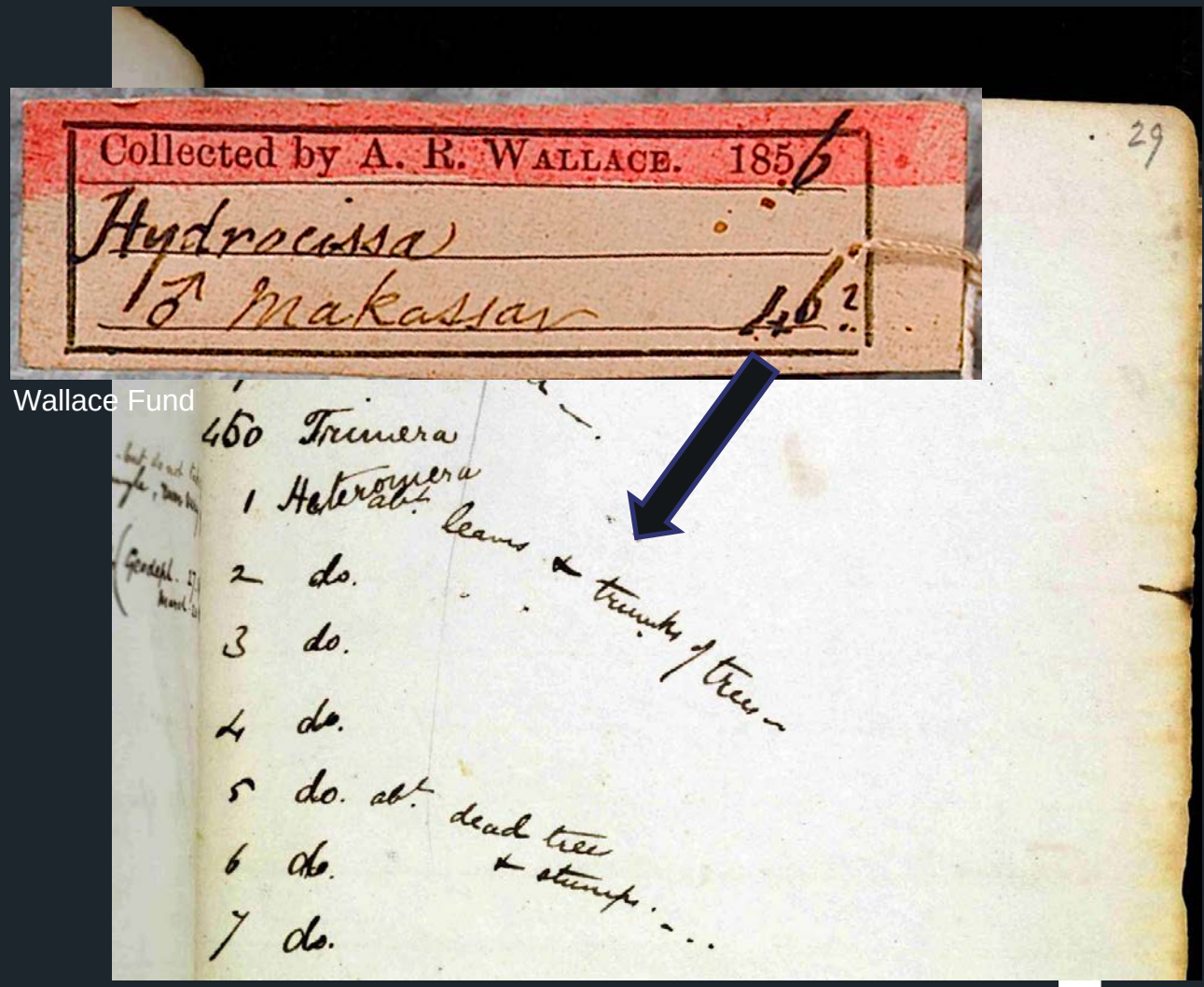
- Well-defined facets
- Simpler search interface

- More relevant results

Mastered data

- Consolidated entities
  - Full context
- Preconsolidated duplicates

- More relevant results



Wallace Fund

Natural History Museum, London

Enrichments

- Add contextual information
- Turn metadata into data

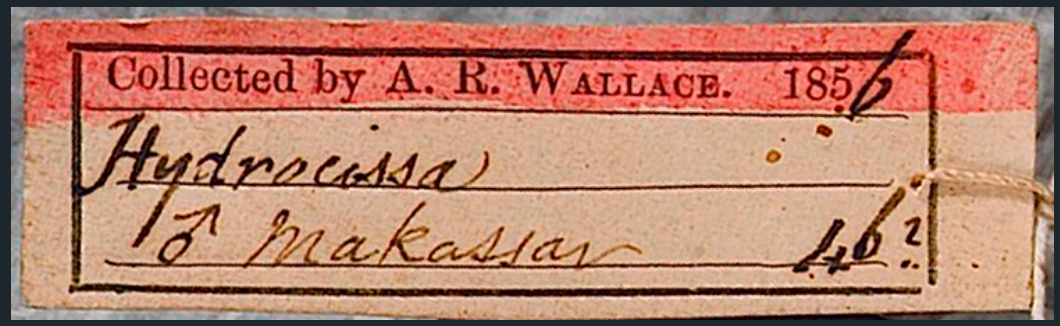
- More relevant results

Natural History Museum, London

Provenance and lineage

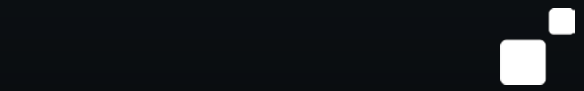
- Context of data
- Quality of data

- More relevant results



Wallace Fund

Recap



Data Hub as innovation engine

- Change → innovation → change
- Small iterated incremental change
- A small failed experiment is easier to survive than a large failed experiment

Search as a way of thinking

- What is the searchly way to solve this?
  - Universal: turn metadata into data
  - Selective: only look at the relevant slice
  - Ranked: measures of relevance of matches, ordered
  - Contextual: not factoids, information

Search drives the innovation engine

- Search tasks powered by built-in MarkLogic advanced search
  - Sample
  - Characterize
  - Explore
  - Select
  - Alert

Search for curation

- Search activities:
  - Prepare for curation
  - Perform curation
  - Consume curation

Curation for search

- Apply principle of universality to get better selectiveness, ranking, context
- Search gets better

Be a better panda