Optics

Participatory activities for purposeful, transparent, and people-centric dataset documentation

Optics
A Guiding Metaphor

Agent Information Journeys =

Perspective —————○ Lens —————○ **Scopes**
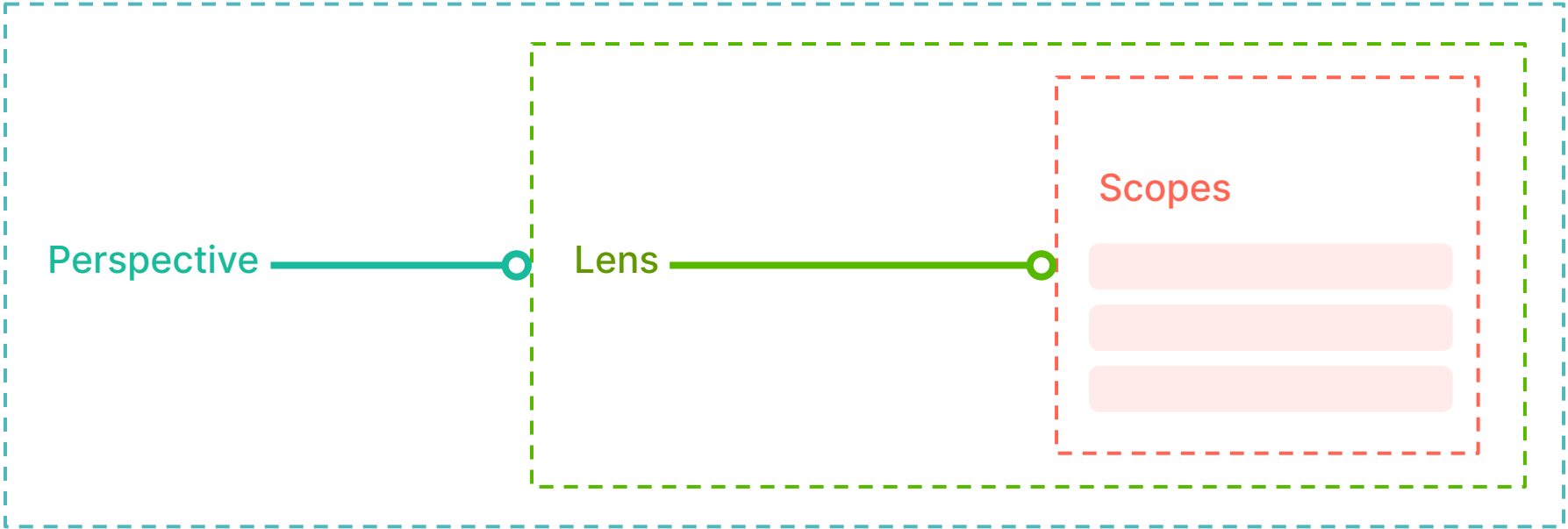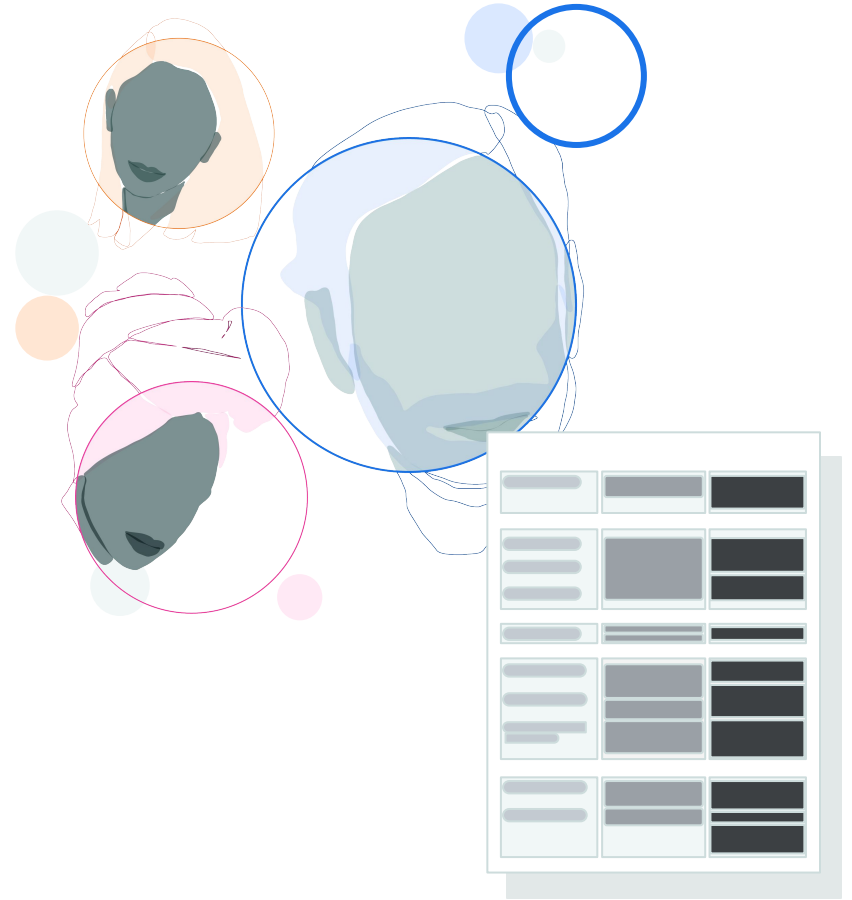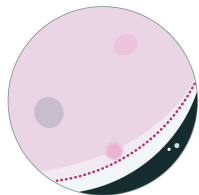
Agent Information Journeys =
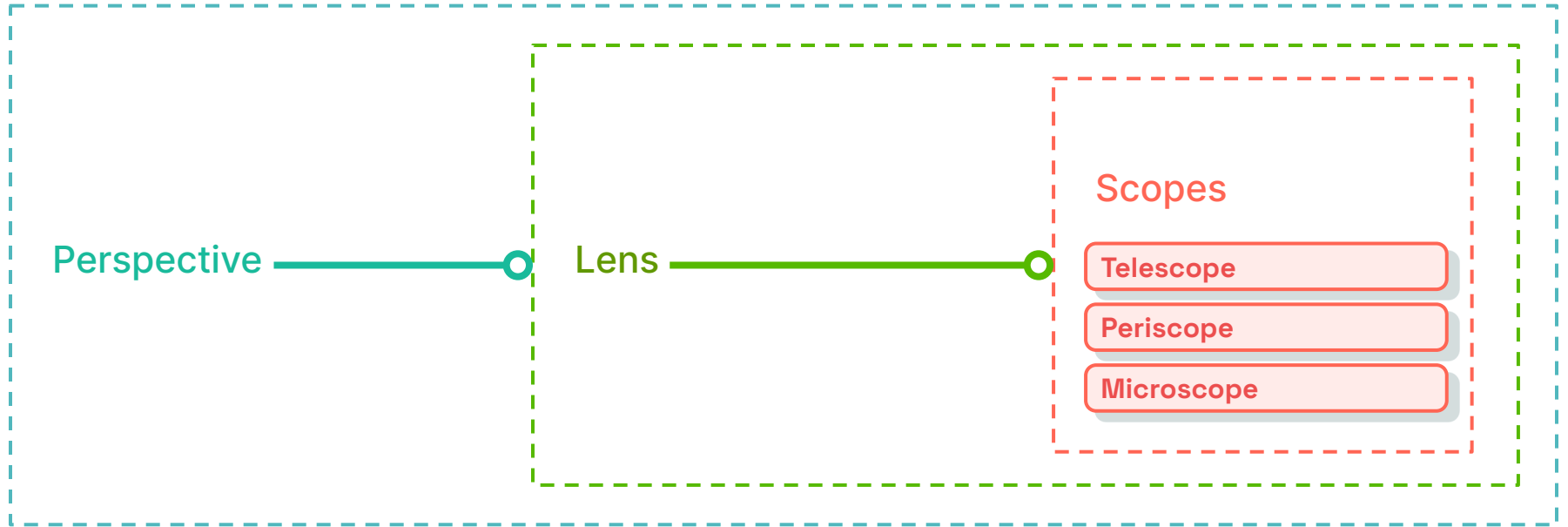
People often ask a series of questions in quick succession to make sense of the world around them.

We call these scopes.

Scopes are a metaphor to focus and frame questions and reveal obvious, non-obvious, visible, and invisible aspects of datasets.
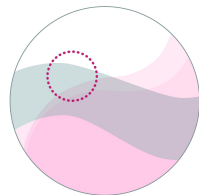
Types of scopes

Telescopic

Questions about attributes commonly found across multiple datasets.

When thinking of telescopic scopes, think of questions that tag "characteristics."

For example:

Does this dataset contain PII?

Periscopic

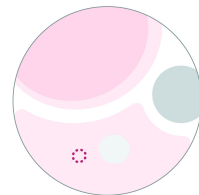**Questions about attributes specific to the producer's dataset.**

When thinking of periscopic scopes, think of questions that describe "observations."

For example:

How many features contain PII?

Microscopic

Questions about unobservable aspects of the dataset.

When thinking of telescopic scopes, think of questions that demand "explanations."

For example:

How was PII anonymized in this dataset?

Scopes in Data Cards

TELESCOPIC
(global attributes)

PERISCOPIC
(local observations)

MICROSCOPIC
(unobservable explanations)

PRIMARY DATA TYPE
Image Data

DATASET FUNCTION(S)
Training
Testing

LICENSE TYPE(S)
CC-BY-4.0

DATA COLLECTION METHOD(S)
Crowdsourced

SAMPLING METHOD(S)
Unsampled

LABELING METHOD(S)

DATASET CHARACTERISTICS
(All numbers are approximate)

Total Instances	478k+
Total Classes	6k+
Total Labels	1.27m+
Algorithmically Generated Labels	1.11m+
User Contributed Labels	505k+
Human Verified Labels	All labels verified

LAST UPDATED   Oct 2018
VERSION   1.0
STATUS   Actively Maintained

DATA SOURCE(S)
• Contributions by global users
  of the Crowdsource app
• Vendor data collection efforts

GEOGRAPHIC DISTRIBUTION

83%	India
2%	Vietnam
2%	Brazil
1%	Israel
1%	Nigeria
1%	Thailand
1%	Colombia
1%	UAE
8%	Others (each less than 1%)

LABEL TYPE(S)

NATURE OF CONTENT
Labeled images of objects (household goods, commercial products), vehicles, plants, animals and people (faces blurred).

EXCLUDED DATA
All EXIF data including location has been removed

PRIVACY   PII associated with human subjects removed

SUMMARY OF LICENSE PERMISSIONS (CC-BY-4.0)
• You are free to share and adapt
• Attribution required
• You cannot apply any additional restrictions

ACCESS COST   Open Access

DATA SELECTION
All images are opted-in for open-sourcing by Crowdsource app contributors

FILTERING CRITERIA
• PII: Name tags, Unblurred faces, etc.
• Inappropriate Content
• Unusable Imagery

LABELING PROCEDURE - HUMAN

TELESCOPIC
(global attributes)

Primary Data Type
Tabular, Image, Audio, Speech, Time-series, Network, Graph, etc.
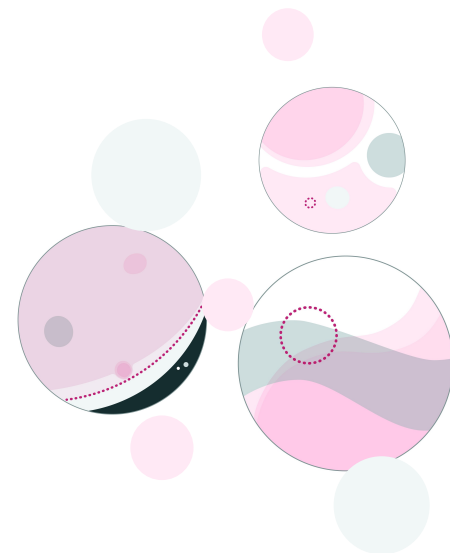
PERISCOPIC
(local observations, demands evidence)

Total Instances — 478k+
Total Classes — 6k+
Total Labels
1.27m+
Algorithmically Generated Labels — 1.11m+
User Contributed Labels — 505k+
Human Verified Labels — All labels
verified
...
...

MICROSCOPIC
### (unobservable explanations, demands rationale)

Nature Of Content Labeled images of objects (household goods, commercial products), vehicles, plants, animals and people (faces blurred).

Excluded Data All EXIF data including location has been removed.

Privacy PII associated with human subjects removed where possible. To preserve demographic context, features are aggregated using ...

Many scopes help maintain balance
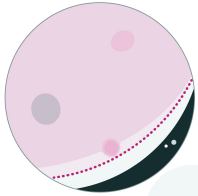
If we only use telescopes, Data Cards will only **describe the obvious** – but without telescopes, agents won't know where to begin.
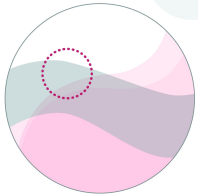
If we only use periscopes, Data Cards can quickly **get overly technical**, without providing any details on how we got there, or why it matters.

If we only use microscopes, we can easily get **lost in the details**, and lose sight of the bigger picture.

	TELESCOPIC	PERISCOPIC	MICROSCOPIC
Description	Attributes	Observations	Explanations
Input Type	Yes/No or multiple choice responses	Key-Value pairs or short descriptions	Detailed, long-form descriptions
Format	Tags, Chips	Text, Tables, and Visualizations w/ Links	Paragraphs, Tables, Visualizations, Links
Content	Universal attributes of datasets	Unique, observable characteristics of the dataset	Rationales, decisions, and policies that shape datasets
Utility	Find information quickly; Indexical– search or scan Keywords, descriptors	Facilitate quick assessments Tactical – what statements	Provide context and links to additional information Rationale and process – why and how statements
Value	Provide an overview and help habituate agents in the Data Card. Make Data Cards easy to index in repositories	Often reproducible from the dataset, prevent re-work, and add more context for agents when stacked	Unobservable attributes and considerations important to agents, but accessible only to producers
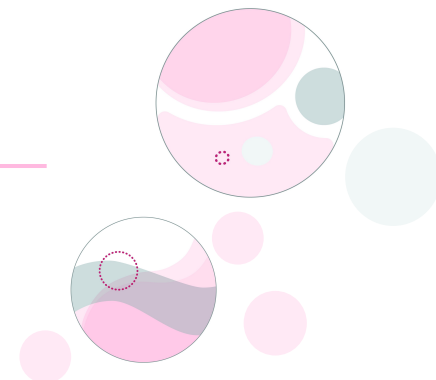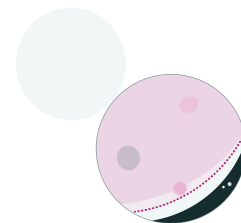
As a data scientist, I want to know
about the structure of the dataset, so I ask...

... what is the data format?
... are there any media in the dataset?

... how many features are in the dataset?
... how many features are engineered?

... which features are strongly correlated
    and why? Was this intentional?

`telescopic`

`periscopic`

`microscopic`

As a **healthcare analyst**, I want to know
about the data collection protocol, so I ask...

... was data collected through a health survey?
... what was the duration of the pilot? Short? Long?

`telescopic`

... what race/ethnicity/language was represented?
... were other data sources combined with health survey?

`periscopic`

... how was the data collection process established for
first-time parents from non-traditional situations?

`microscopic`

Try it yourself

As a _____,   perspective
I want to know _____,   lenses
so I ask...

... _____   telescopic question   **?**   **Ask for attributes**

                                       Follow-up
                                       Ask for evidence

... _____   periscopic question   **?**

                                       Follow-up
                                       Ask for rationale

... _____   microscopic question   **?**