Questify Data Collection

Tyler Venner

December 2025

1 Dataset Class Mapping and Taxonomy

The dataset consists of 12 classes targeting medically significant arthropods in North Carolina. Images are sourced from iNaturalist with a "Research Grade" quality filter. The taxonomy strategy varies by class, utilizing Family-level, Genus-level, or Order-level groupings to best capture the target vectors.

Table 1: Target Classes and Taxonomy Definitions

ID	Class Name	Target Taxon	Rank
0	Venomous Spiders	*Latrodectus*, *Sicariidae*	Genus/Fam
1	Ticks	*Ixodidae*	Family
2	Mosquitoes	*Culicidae*	Family
3	Stinging Wasps	*Vespidae*	Family
4	Bees	*Apidae*	Family
5	Fire Ants	*Solenopsis*	Genus
6	Assassin Bugs	*Reduviidae*	Family
7	Venomous Caterpillars	*Megalopygidae*, *Limacodidae*	Family
8	Blister Beetles	*Meloidae*	Family
9	Scorpions	*Vaejovidae*	Family
10	Horse/Deer Flies	*Tabanidae*	Family
11	Centipedes	*Chilopoda*	Order

2 Biological Context and Taxonomic Challenges

To effectively train the classification model, it is necessary to understand the biological hierarchy (taxonomy) from which the classes are derived. The dataset operates at varying levels of taxonomic specificity, primarily Family, but occasionally Genus or Order.

2.1 Taxonomic Levels of Specificity

Taxonomy classifies organisms into a hierarchy: *Kingdom* → *Phylum* → *Class* → *Order* → *Family* → *Genus* → *Species*. This project utilizes three distinct levels:

- **Family Level:** Most classes (e.g., *Culicidae* for Mosquitoes, *Vespidae* for Wasps) are defined at the Family level. Organisms within a Family typically share strong visual characteristics (morphology) that a Convolutional Neural Network (CNN) can easily generalize.
- **Genus Level (High Precision):** Class 5 (Fire Ants) targets the specific genus *Solenopsis*. This requires the model to distinguish minute features (such as head shape and coloration) to differentiate dangerous fire ants from harmless ants in the broader *Formicidae* family.

- **Order Level (Broad Categorization):** Class 11 (Centipedes) targets the Order *Chilopoda*. This is a high-level grouping containing diverse species, but they share a unique body plan (elongated, multi-segmented) distinct from all other classes.

2.2 Morphological and Biological Challenges

2.2.1 The "Stinging Trio" (Order Hymenoptera)

Classes 3 (Wasps), 4 (Bees), and 5 (Fire Ants) belong to the same Order, *Hymenoptera*, sharing genetic and visual similarities. The model must learn subtle distinctions:

- **Wasps (*Vespidae*):** Characterized by smooth, shiny bodies and a distinct "pinched" waist.
- **Bees (*Apidae*):** Distinguished by a hairy/fuzzy texture (for pollen collection) and stockier body shape.
- **Fire Ants (*Solenopsis*):** Must be distinguished from other ants by specific coloration (copper-brown head, darker abdomen) and lack of wings (in most castes).

2.2.2 Metamorphosis and Life Stages (Class 7)

Class 7 (Venomous Caterpillars) presents a unique challenge due to metamorphosis. The families *Megalopygidae* and *Limacodidae* include both the venomous larval stage (caterpillar) and the harmless adult stage (moth).

> **Data Requirement:** Training data must be strictly filtered for **"Life Stage: Larva"** to prevent the model from learning to flag harmless moths as dangerous.

2.2.3 Intra-Class Variance (Class 0)

Class 0 (Venomous Spiders) aggregates two visually distinct families:

1. **Theridiidae (We only use Latrodectus):** Bulbous, round abdomens with glossy textures.
2. **Sicariidae (Recluses):** Flatter bodies, longer legs, and matte textures.

The model must learn a multi-modal representation for this single class, as the two subgroups share few visual features despite both being "venomous spiders."

3 Data Collection and Balancing Strategy

We utilized the *pyinaturalist* API wrapper to query the iNaturalist database. An initial query of research-grade observations in North Carolina (NC) revealed significant disparities in species prevalence. While classes such as *Apidae* (Bees) and *Vespidae* (Wasps) yielded over 15,000 local observations, medically critical classes were dangerously scarce.

This scarcity was not uniform across entire classes, but often specific to genera within a class. For example, within the **Venomous Spiders** class, Widow spiders (*Latrodectus*) were abundant in NC ($N = 1,159$), while Recluse spiders (*Sicariidae*) were virtually absent ($N = 29$).

3.1 Hybrid Geographic Strategy

To address this volume disparity while preserving ecological relevance, we adopted a **Hybrid Geographic Strategy**. We established a minimum threshold of $N = 1,000$ images for model viability.

- **Tier 1: NC-Sourced (Local Context).** For taxa with $N > 1,000$ local observations, we restricted data collection to North Carolina (Place ID: 30). This ensures the model learns to distinguish targets against the specific background flora and lighting conditions of the target deployment region.

Table 2: **Initial Query:** Data Availability in North Carolina (Place ID: 30)

Class #	Class Group	Specific Taxon	Taxon ID	NC Count
0	Venomous Spiders	Widow Spiders	47370	1,159
0	Venomous Spiders	Sixeyed Sicariid Spiders	48140	**29**
1	Ticks	Hardbacked Ticks	51673	1,292
2	Mosquitoes	Mosquitoes	52134	1,486
3	Stinging Wasps	Hornets, Paper Wasps	52747	16,308
4	Bees	Apidae (Honey/Bumble)	47221	38,268
5	Fire Ants	Solenopsis Fire Ants	67597	1,536
6	Assassin Bugs	Assassin Bugs	48959	8,212
7	Venomous Caterpillars	New World Flannel Moths	84186	**295**
7	Venomous Caterpillars	Slug Caterpillar Moths	84165	1,683
8	Blister Beetles	Blister Beetles	59510	**755**
9	Scorpions	Devil Scorpions	52572	**26**
10	Horse/Deer Flies	Horse and Deer Flies	47821	**652**
11	Centipedes	Centipedes	49556	2,158

- **Tier 2: USA-Sourced (Morphological Learning).** For taxa with insufficient local data ($N < 1,000$), we expanded the geographic filter to the United States (Place ID: 1). This prioritizes learning the *morphology* of the insect (which is consistent across the continent) over local background context.

3.1.1 Intra-Class Source Splitting

A key idea in our dataset is the splitting of sources *within* a single class label.

- **Venomous Spiders:** We source *Latrodectus* (Widows) from NC to maintain local context, but expand *Sicariidae* (Recluses) to the USA ($N = 29 \rightarrow 7,109$). This prevents "Feature Collapse," where the model might otherwise ignore the minority Recluse class and incorrectly learn that "Venomous Spider" is synonymous only with the visual features of a Black Widow.

- **Venomous Caterpillars:** Similarly, we source *Limacodidae* (Slug Caterpillars) locally but expand *Megalopygidae* (Flannel Moths) to the USA. *Note: To ensure we capture only the relevant life stage, we explicitly filter for the 'Larva' annotation (Term ID: 1, Value: 6) via the API. The expanded geographic scope provided sufficient volume ($N = 6,376$) to support this strict filtering without data loss.*

Table 3: **Final Configuration:** Optimized Data Sources and Counts

Class #	Class Group	Specific Taxon	Source	ID	Final Count	Strategy
0	Venomous Spiders	Widow Spiders	NC	47370	1,159	Local Context
0	Venomous Spiders	*Sicariidae* (Recluses)	**USA**	48140	7,109	**Intra-Class Split**
1	Ticks	Hardbacked Ticks	NC	51673	1,292	Local Context
2	Mosquitoes	Mosquitoes	NC	52134	1,486	Local Context
3	Stinging Wasps	Vespidae	NC	52747	16,308	Abundant (Sampled)
4	Bees	Apidae	NC	47221	38,268	Abundant (Sampled)
5	Fire Ants	Solenopsis	NC	67597	1,536	Local Context
6	Assassin Bugs	Reduviidae	NC	48959	8,212	Local Context
7	Venomous Caterpillars	*Megalopygidae*	**USA**	84186	6,376	**Intra-Class Split**
7	Venomous Caterpillars	*Limacodidae*	NC	84165	1,683	Local Context
8	Blister Beetles	Meloidae	**USA**	59510	36,288	Morphological Expansion
9	Scorpions	Vaejovidae	**USA**	52572	26,090	Morphological Expansion
10	Horse/Deer Flies	Tabanidae	**USA**	47821	15,505	Morphological Expansion
11	Centipedes	Chilopoda	NC	49556	2,158	Local Context

3.2 Class Balancing

To prevent the model from biasing towards majority classes (e.g., Bees with 38k images vs. Widows with 1.1k), we implemented a **strict cap of 2,000 images per class** for the training set. For abundant classes, images are randomly sampled. For scarcity classes expanded to the USA (like Scorpions), we collect up to the cap to ensure a perfectly balanced training distribution, preventing the “long-tail” problem common in biological datasets.