

Unit 7: Data Annotation

Annotation & Typology

Lesson 48

Derek Ruths

Overview of unit

Objectives:

- Understand how to approach data annotation (whether automated or manual)
- Know how to run a small human annotation task

1. Typologies
2. Building a typology
3. Applying a typology – human coding
4. Confirming an annotation
5. Applying a typology – classifier

Lesson overview

Objectives

- Understand what a typology is
- Understand why typologies are at the heart of annotation

Outline

- Review of annotation
- What is a typology?
- Examples of typologies in annotation

Data science project phases

Annotation is about assigning properties that are not (reliably) self-reported.



Fruit, vegetable, or animal?

Fruit, vegetable, or animal?



Fruit

Fruit, vegetable, or animal?



animal

Fruit, vegetable, or animal?



ambiguous

Fruit, vegetable, or animal?



The goal of annotation

↙ according to designer

Accurately and reproducibly assign labels to a set of data objects

determine labels
we need

The goal of annotation

Tweet	Type of weather
It's pouring outside.	rain
Just came inside soaking wet.	rain
Blizzard conditions out there! #hotchocolatetime	snow
Going to get a bit wet catching the bus today!	rain
Sunglasses weather. Can't wait to take a walk.	sun
Pouring myself some cereal this morning.	n/a

gold standard questions → questions you know
the answer to

The two annotation tasks

1. Devising the labels to use → typology
2. Assigning the labels → reliability

Types of properties

Real-valued:

- Location in the world
- Temperature
- Color

} you could use a number

Categorical:

- Type of daily activity
- Type of food
- Is married

} no ways to order

← annotation almost always talks about this

↙ We need because of categorical properties

What is **typology**?

↙ clear and explicit

A **comprehensive**, **sharply-defined** **categorization system**

↑ anything you encounter must be labeled

Examples:

- BAD typology for books: cookbooks, sci-fi novels, biographies
- GOOD (albeit, limited) typology for books: fiction, non-fiction

(bad because this is the complete list)

Everything fits in here (comprehensive)

Properties of a **good typology**

- **Comprehensive**: any object in the space will fit one (or more) types
- **Well-defined types**: each type has a definition that allows systemic evaluation of whether an object is of that type
- **Objective-ish**: definitions should strive to be as objective as possible... objective often cannot be achieved, but there are strategies to make subjective categories more objective. ←

→ eg whether or not something
is cozy
→ very subjective

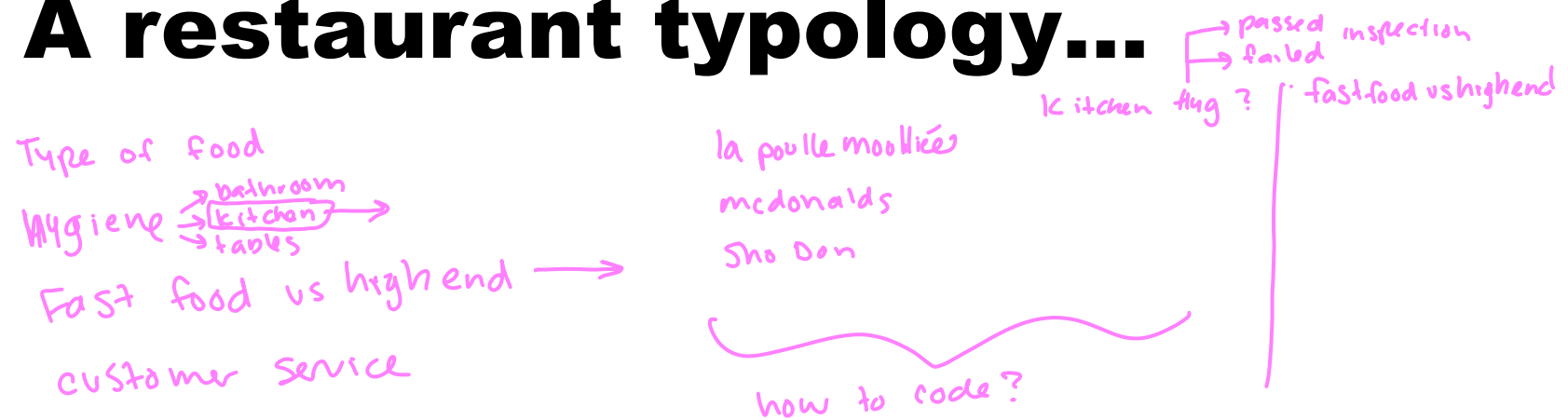
Annotation & Typology

Any categorical property requires a typology

Some might seem easy on the surface, but can be

- **Hard to define** (even given all information about an object, it's hard to know what label to give it)
 - Hate speech
 - Pornography
 - Product quality
- **Hard to discern** (hard to determine from available information)
 - Home location
 - Marital status

A restaurant typology...



ask what defines fast food?

typology

categories: Fast food
high end

} need clear, objective criteria

Lesson wrap-up

Takeaways

- In order to annotate, we often require typologies
- Typologies take effort to get right

Up next

- Building a typology