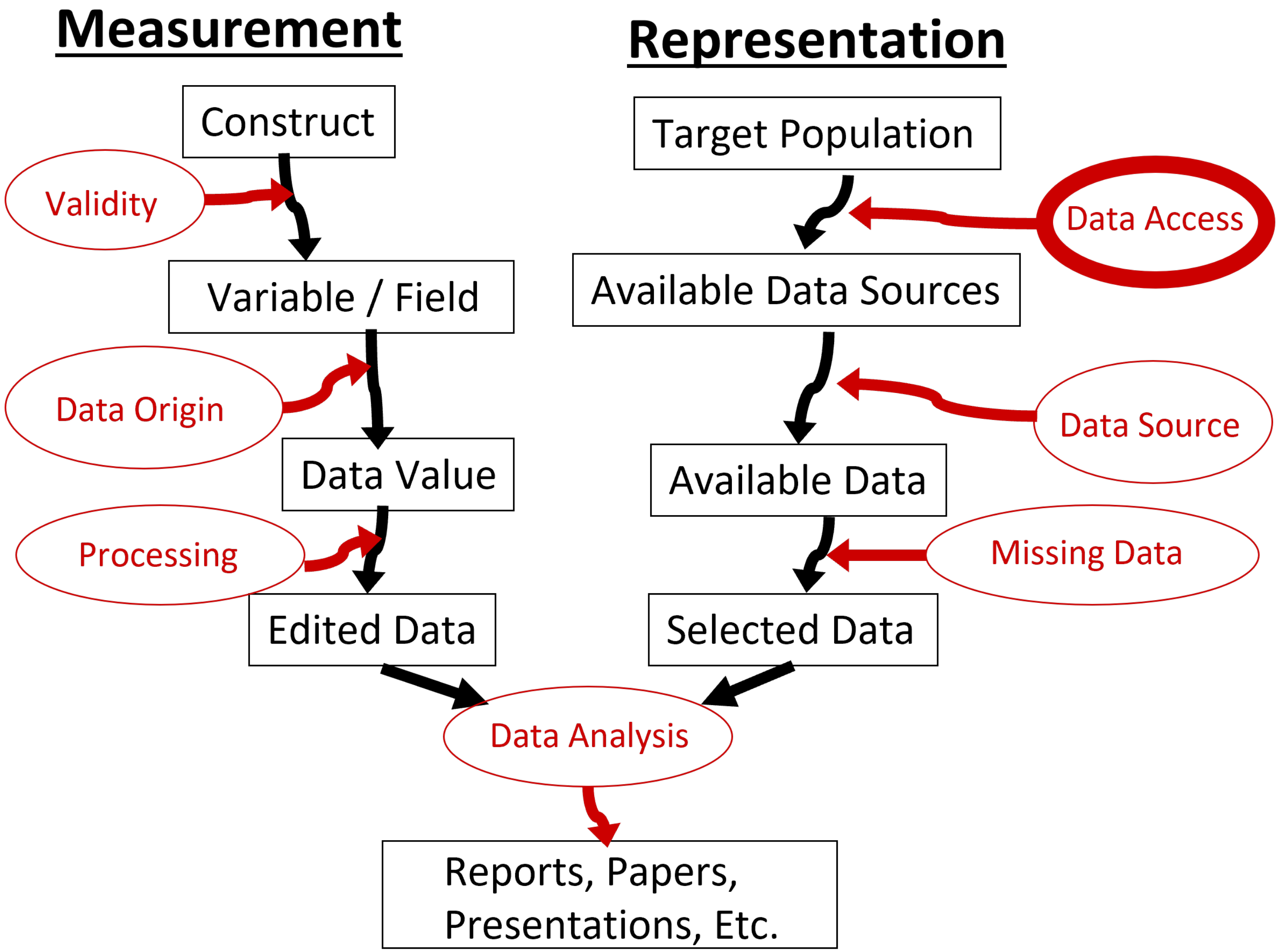
Data Access Threats for Gathered Data | Part 1

By Trent D. Buskirk

Dimensions of TDQ



Threats to Quality Related to Data Access: Web Sourced Data (1)

- Architecture of Web Pages
  - Web scraping methods can be used to obtain information from the internet from various Open Data Sources such as Wikipedia…
  - Extensible Markup Language (**XML**) is a markup language that defines a set of rules for encoding documents to be both human-readable and machine-readable and an estimated 60% of webpages are created using XML (Grijzenhout and Marx, 2013).

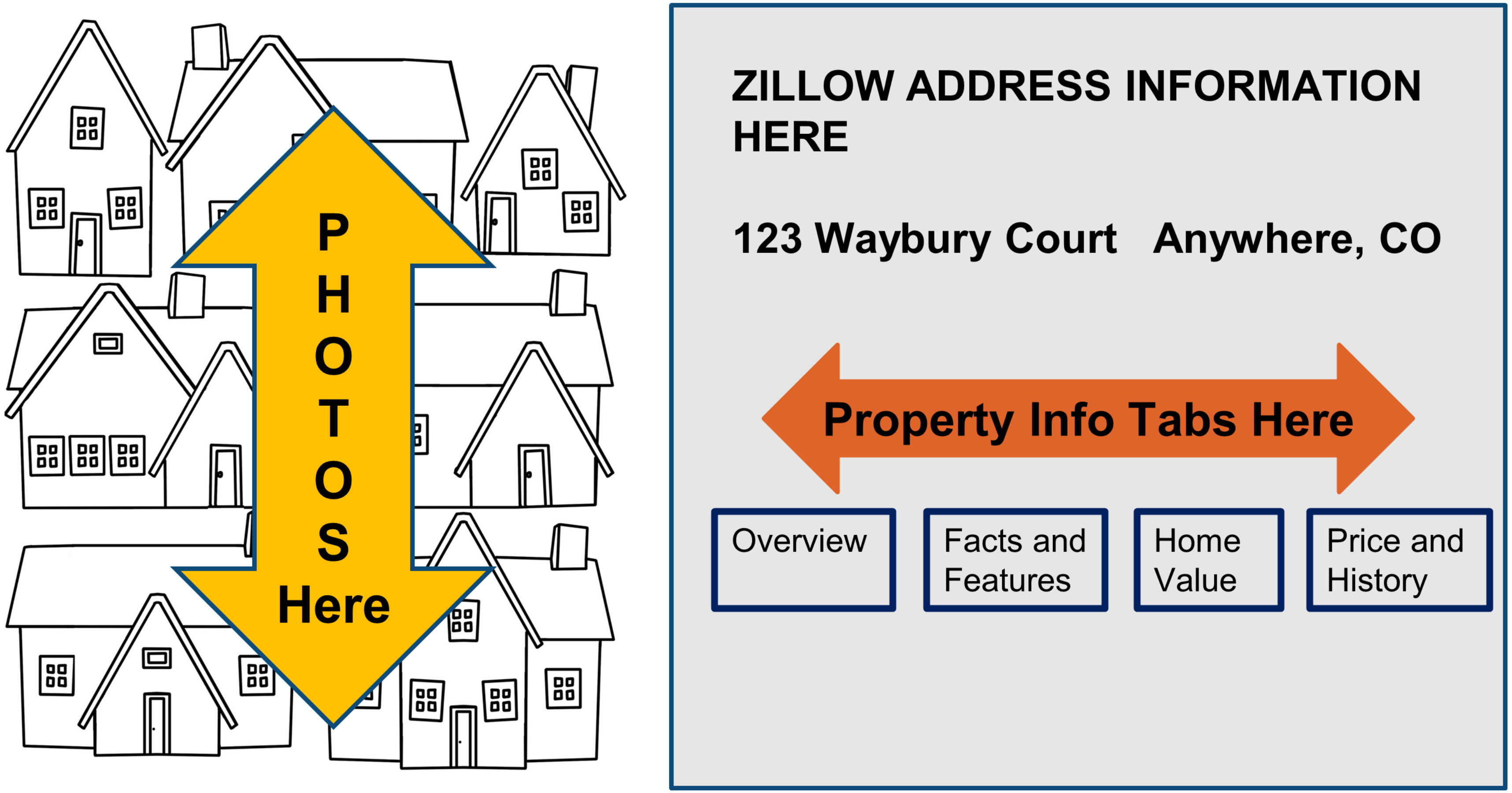
Threats to Quality Related to Data Access: Web Sourced Data (2)

- Architecture of Web Pages…
  - An estimated 15% of XML documents lack well-formedness, a quality indicator representing how “grammatically correct” the document is in terms of the XML language.
  - A bulk of these errors come from a mismatch or missing tags in the document.
  - Errors such as these in XML files render them useless for applying XML-based web-scraping queries to them to extract information such as XPath.

Threats to Quality Related to Data Access: Web Sourced Data (3)

- Web Page Content and Formatting
  - Web sources may display similar information in different formats depending on variables that may not be known to researchers at the time a web scraper is being designed.
  - For example, Zillow.com optimizes display of properties depending on their status:
    - Active For Sale/Rent properties to see photos on the left pane and other information by tab on the right.
    - Inactive properties have a different layout with photos on the top and no tabs for features.

Zillow Active For Sale Listing Layout



PHOTOS Here

ZILLOW ADDRESS INFORMATION HERE

123 Waybury Court   Anywhere, CO

Property Info Tabs Here

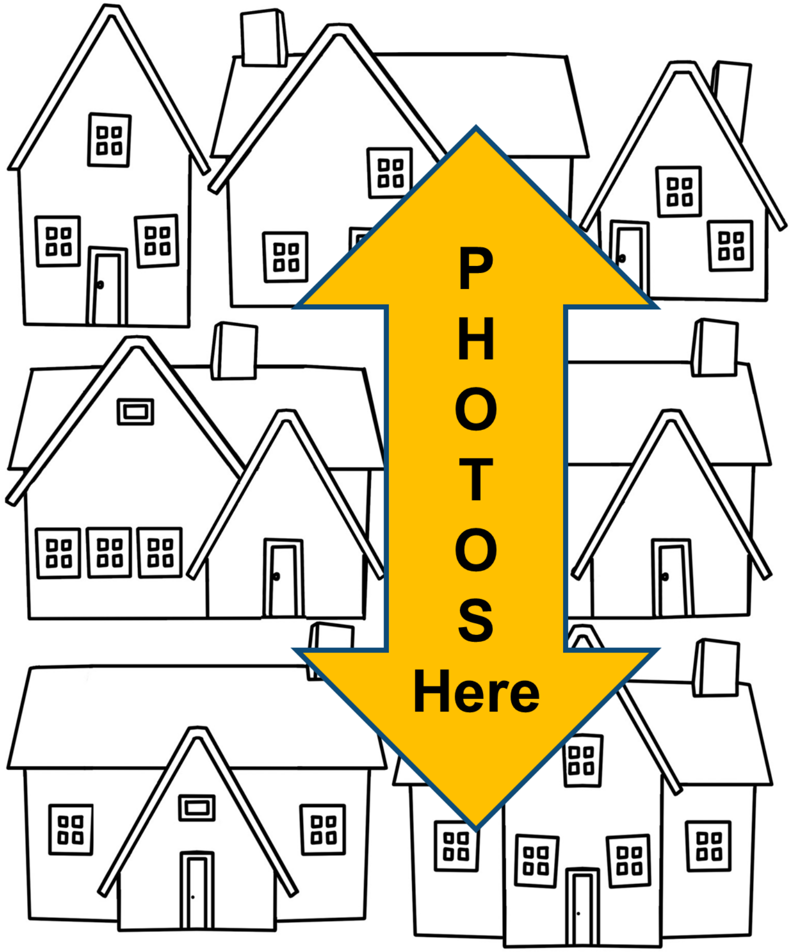
Overview

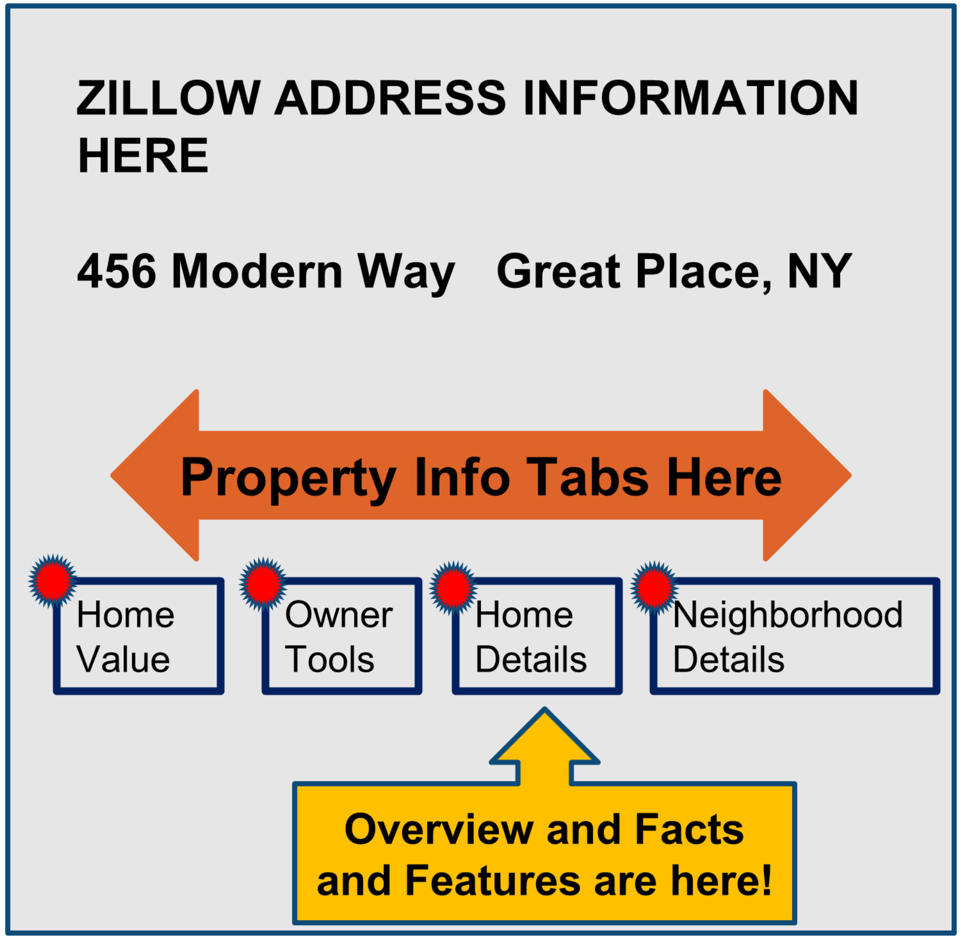
Facts and Features

Home Value

Price and History

Zillow Inactive For Sale Listing Layout





ZILLOW ADDRESS INFORMATION HERE

456 Modern Way   Great Place, NY

Property Info Tabs Here

Home Value

Owner Tools

Home Details

Neighborhood Details

Overview and Facts and Features are here!

Implications of Zillow Layout on Data Access

So a programmed web scraper trained on a list of primarily non-active listings and applied to a random sample of addresses may code several property-specific variables missing because the layout and page structure (underlying XML) is different.

While it is clear that Active Status explains differences in layout and two different web scrapers could be developed and applied accordingly, the status information may not be known a priori.

What’s Next?

In the next Lecture Video we will conclude our discussion of Threats related to Data Access for Gathered Data as we discuss using APIs to gather data and the growing digital divide for access in social media data.

UNIVERSITY OF MICHIGAN