Data Quality in Citizen Science:

Overcoming the Participation - Data Quality Tradeoff

Roman Lukyanenko, Jeffrey Parsons, Yolanda Wiersma

Memorial University of Newfoundland, Canada

Thesis

- Data quality and level of participation in active citizen science projects can be increased simultaneously
- Requires relaxing assumptions about the kinds of entities about which information is collected

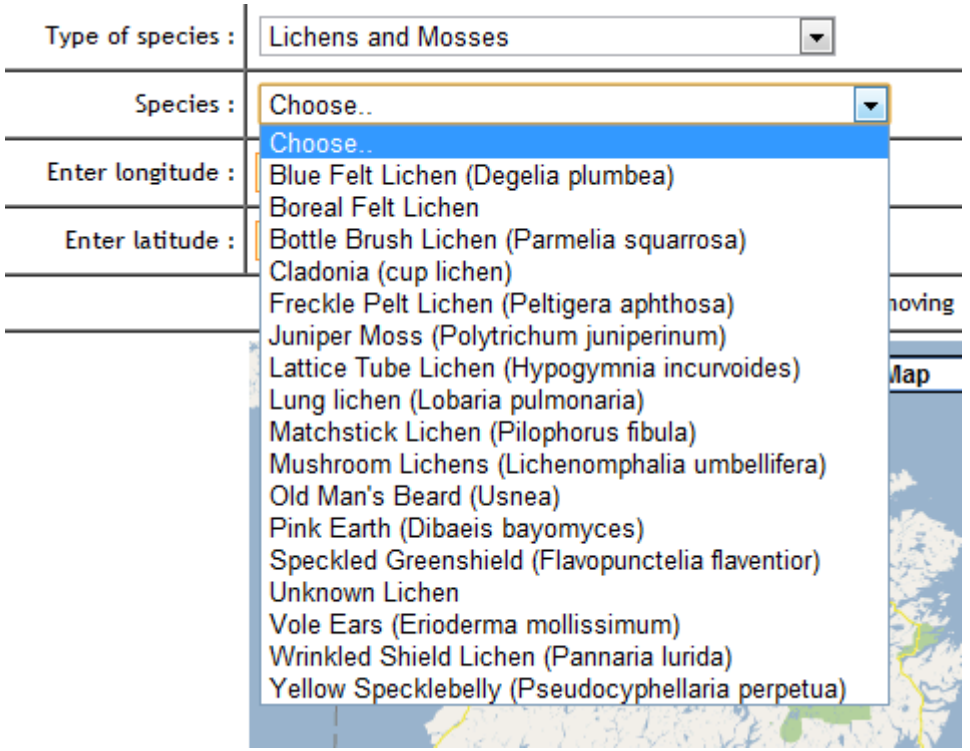
Outline

- What is CS?
- Participation
- Data Quality Implications
- Evidence of a Problem
- Toward a Solution

What is Citizen Science?

- Knowledge production by, and for, non-scientists (Ottinger 2010)
- Voluntary participation of amateur scientists in scientific endeavors (Silvertown 2009)
  - Passive (SETI@Home)
  - Intermediate (GalaxyZoo, FoldIt)
  - Active (eBird, iSpot, NL Nature)
- Users have become information producers
  - Crowdsourcing, social networking, citizen science
- Harnessing this potential is a key challenge

Participation and Data Quality



- Some lichen
- Observer and experts do not know exactly what it is

- Option: Unknown
- Option: Best guess
- Option: Abandon data entry

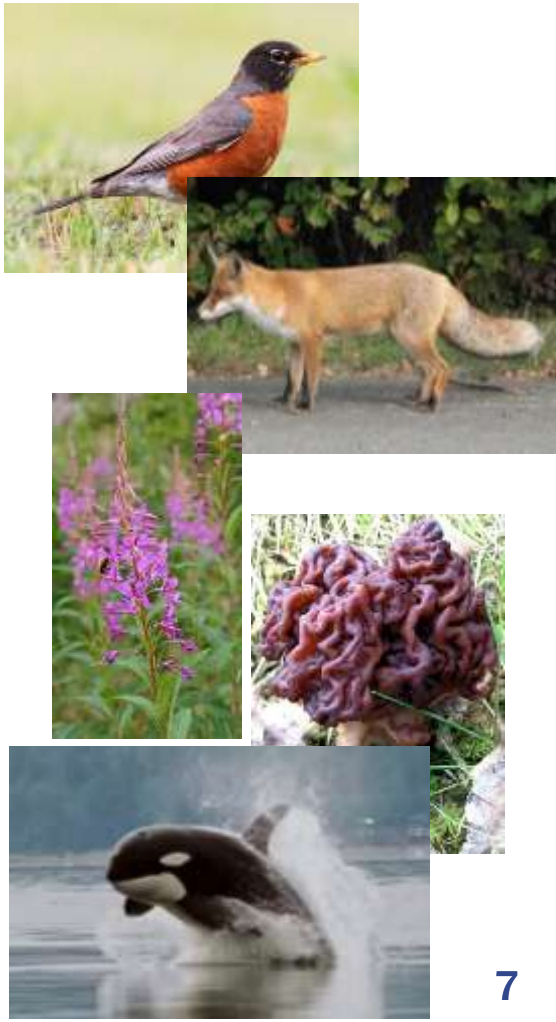
Data Quality (DQ) Limitations

- Ordinary citizens often unable to provide information to meet requirements of scientific projects
  - Arises from the requirement of positive identification
- Traditional solutions
  - Training and better instructions
  - Social networking
  - Expert judgement
- Both experts and novices provide data that is deficient in some ways!
  - Data quality is a function of conceptual modeling choices (of the way data is stored)

Evidence of a Problem

- Empirical study to
  - Measure the impact of categorization on data quality
  - Examine the role of “basic level” of categorization
- Stimuli: 24 images of plants and animals

Birds	American Robin, Blue Jay, Mallard Duck, Blue Winged Teal, Caspian Tern, Common Tern, Greater Yellowlegs, Spotted Sandpiper
Land Mammals	Eastern Chipmunk, Eastern Coyote, Moose, Red Fox, Woodland Caribou, Red Squirrel
Plants	Bog Labrador Tea, Calypso Orchid, Fireweed, Indian Pipe, Sheep Laurel
Mushrooms and Lichens	False Morel, Lung Lichen, Old Man's Beard
Marine Animals	Atlantic Salmon, Killer Whale

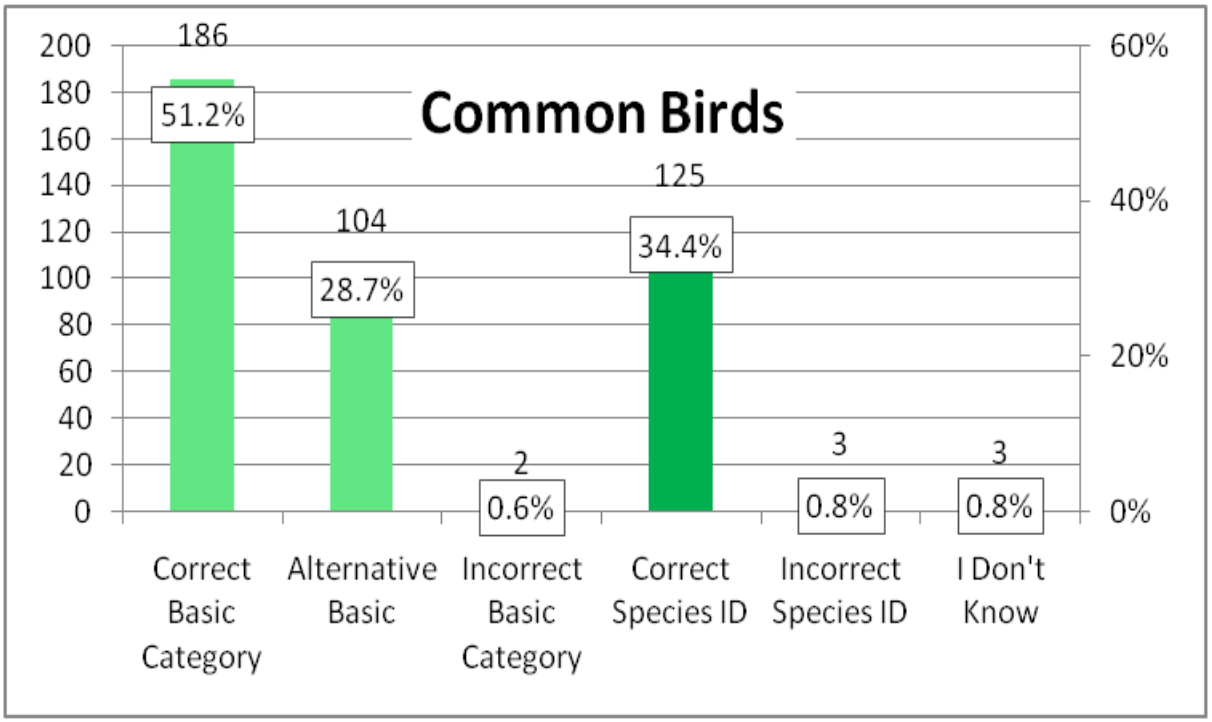


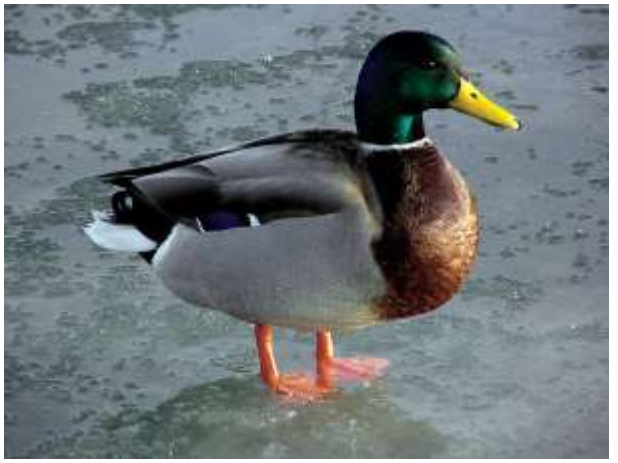
Procedure

- Participants (247) were organized in groups
  - Not domain experts
- Each group was shown a randomized sequence of 24 full-color images
- Each image was displayed for 50 seconds, followed by a short audio signal to herald stimulus change
- There were two study conditions (randomly assigned):
  - Condition 1: Q1. List features that describe what you see best
  - Condition 2: Q1. Indicate what it is (in one or more words)

    Q2. List features that describe what you see best

Preliminary findings

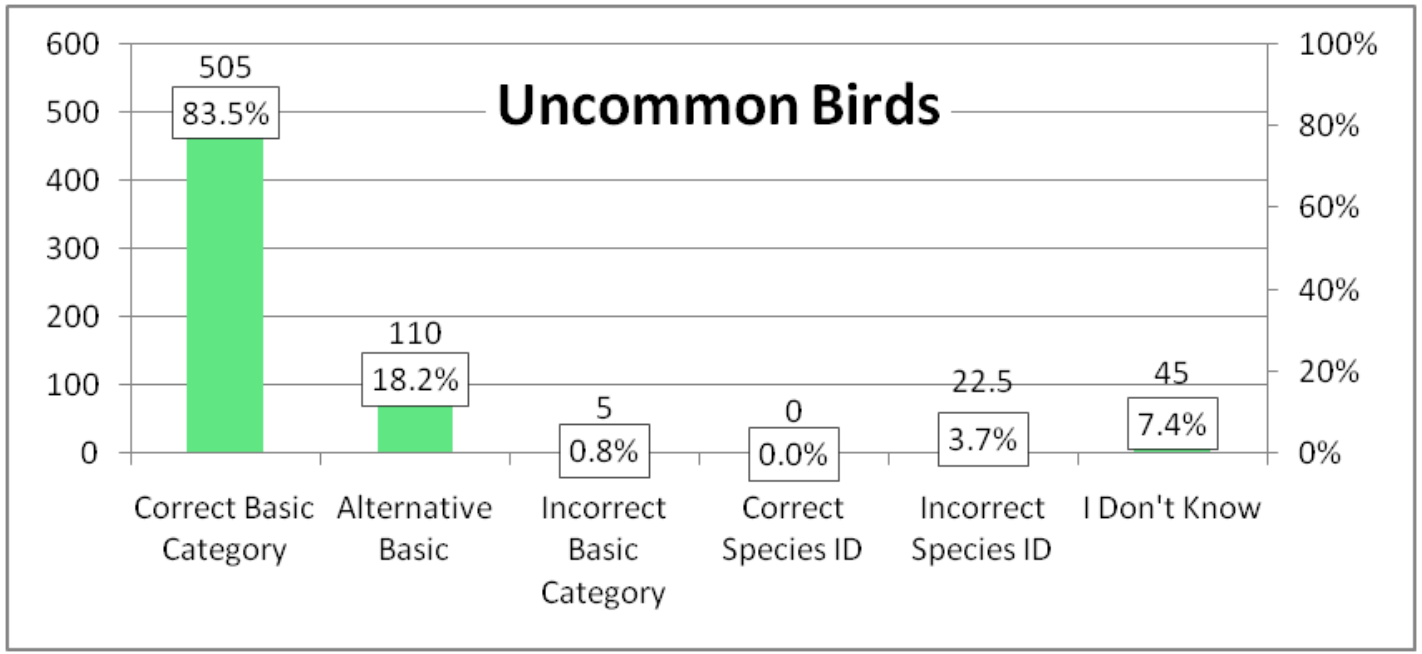




- Basic level: BIRD, DUCK (alt)
- Species:
  - American Robin (Turdus migratorius)
  - Blue jay (Cyanocitta cristata)
  - Mallard or Wild duck (Anas platyrhynchos)

Participation – Data Quality Tradeoff

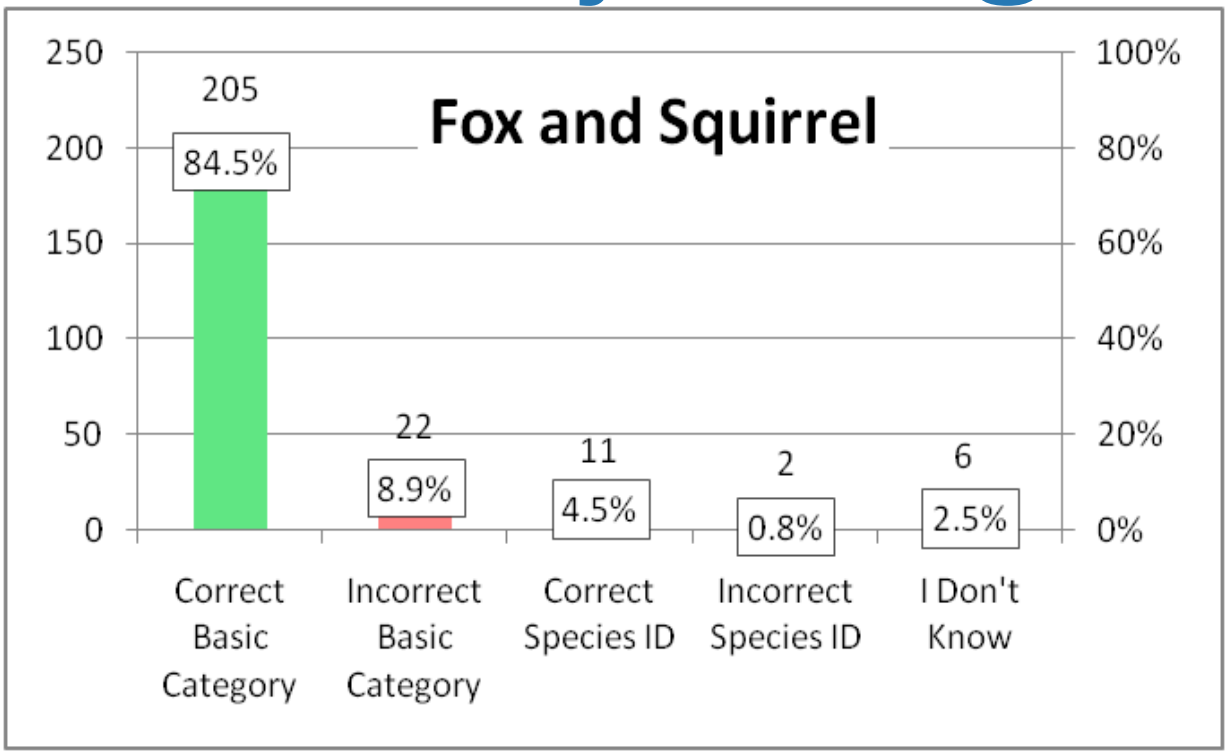
Preliminary findings



- Basic level: BIRD; DUCK (alt)
- Species:
  - Blue Winged Teal (Anas discors)
  - Caspian Tern (Hydroprogne caspia)
  - Common Tern (Sterna hirundo)
  - Greater Yellowlegs (Tringa melanoleuca)
  - Spotted Sandpiper (Actitis macularius)

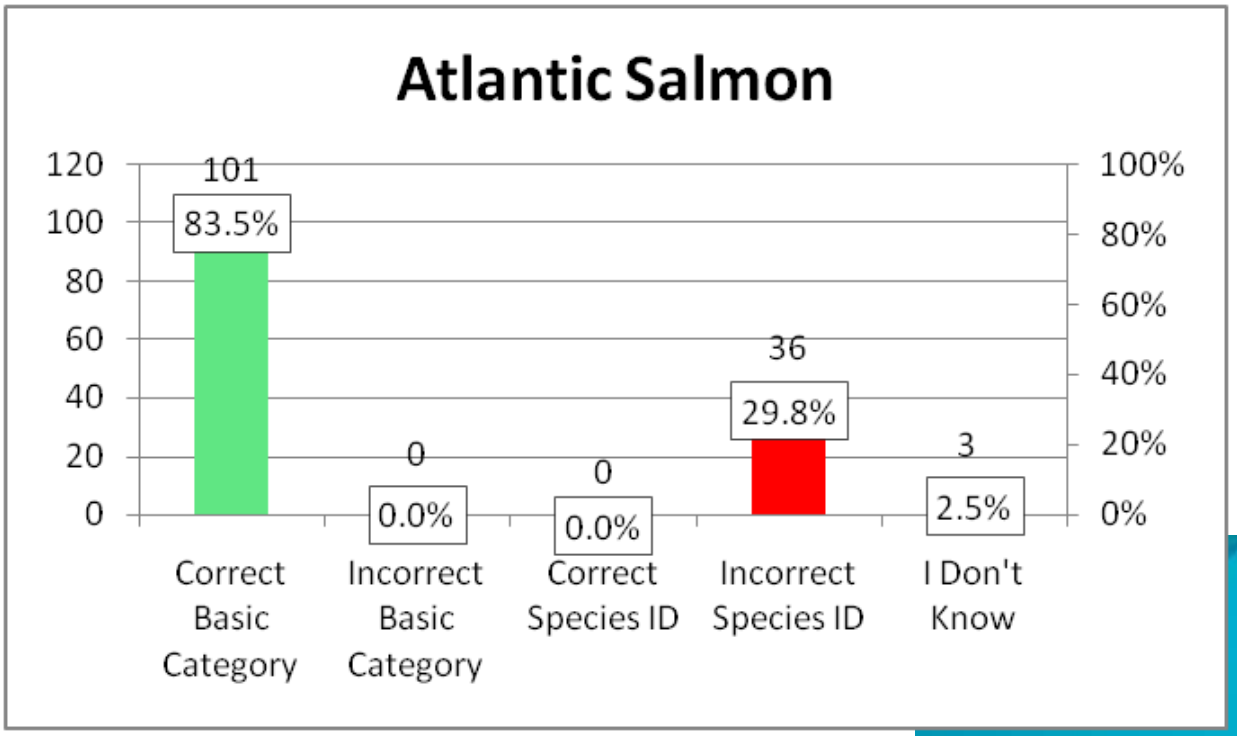
Participation – Data Quality Tradeoff

Preliminary findings

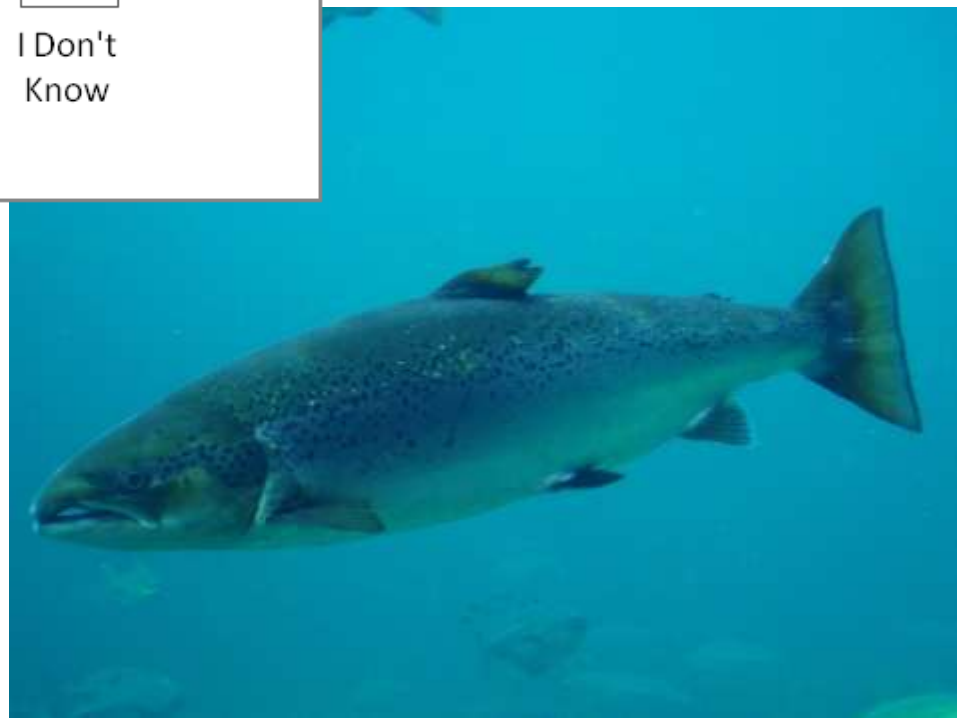


- Basic level: FOX, SQUIRREL
- Species:
  - Red Fox (Vulpes vulpes)
  - American Red Squirrel (Tamiasciurus hudsonicus)

Preliminary findings



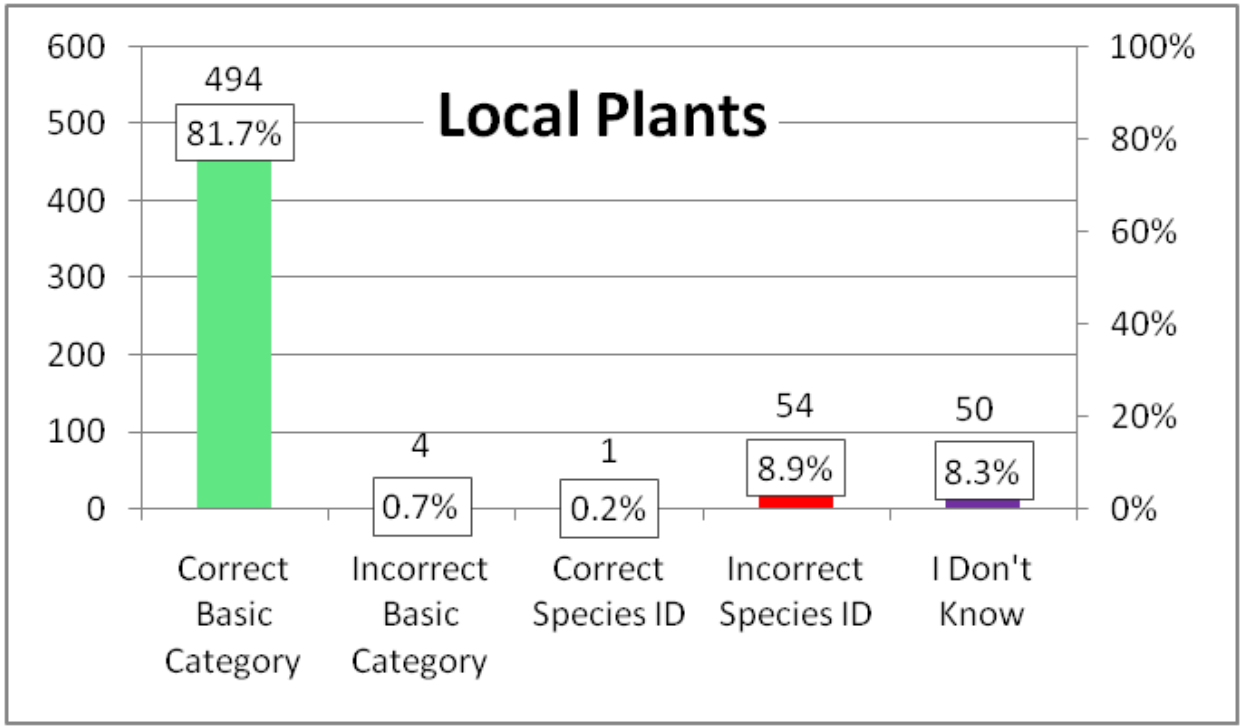
- Basic level: FISH
- Species:
  - Atlantic Salmon (Salmo salar)



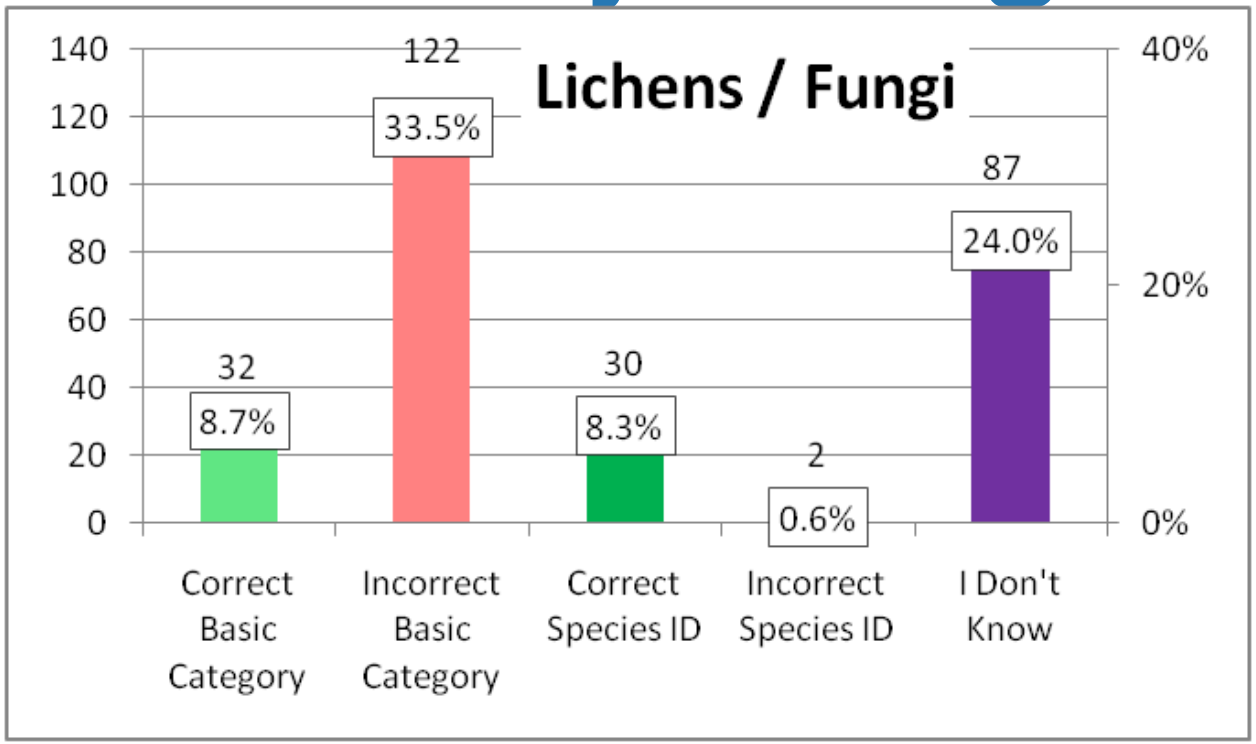
Participation – Data Quality Tradeoff

Preliminary findings

- Basic level: FLOWER
- Species:
  - Bog Labrador tea (Ledum groenlandicum)
  - Calypso Orchid (Calypso bulbosa)
  - Fireweed (Epilobium angustifolium)
  - Indian Pipe (Monotropa uniflora)
  - Sheep Laurel (Kalmia angustifolia)



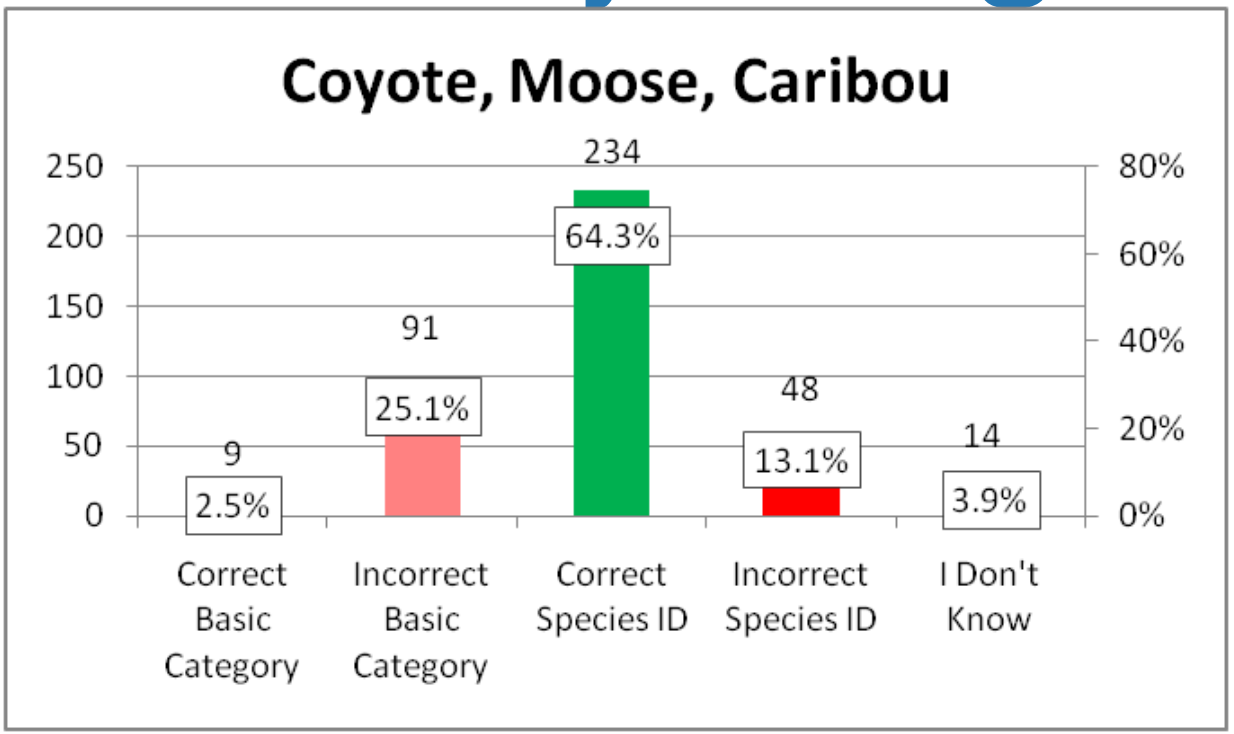
Preliminary findings



- Basic level: MUSHROOM, LICHEN?
- Incorrect basic: MOSS, GRASS, LEAF
- Species:
  - False morel (Gyromitra esculenta)
  - Lung lichen (Lobaria pulmonaria)
  - Old Man's Beard (Usnea)

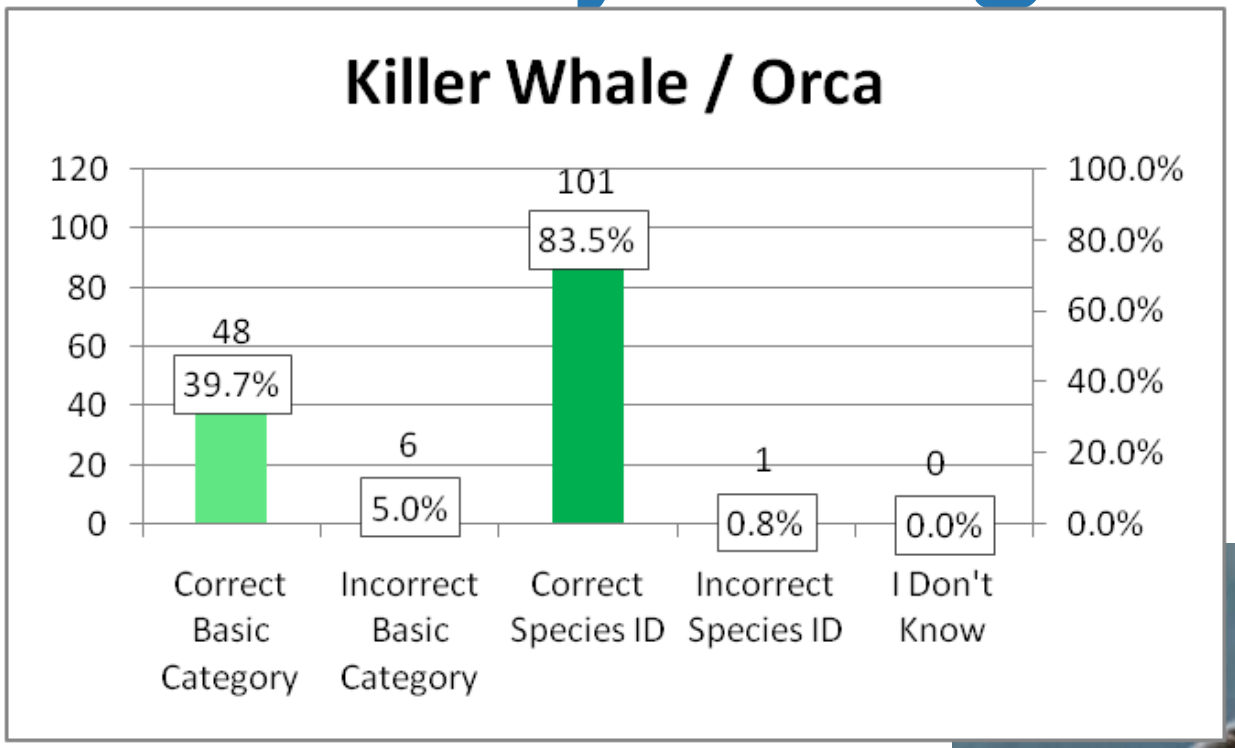
Participation – Data Quality Tradeoff

Preliminary findings



- Correct Basic: DEER
- Incorrect Basic: WOLF, FOX
- Species:
  - Eastern Coyote (Canis latrans)
  - Moose (Alces alces)
  - Woodland Caribou (Rangifer tarandus caribou)

Preliminary findings



- Basic level: WHALE
- Incorrect basic: FISH, DOLPHIN
- Species:
  - Killer Whale / Orca (Orcinus orca)

Summary

- Ordinary citizens are (often) unable to provide information to the degree desired by scientists (e.g., species level in natural history projects)
- Requirement of positive identification leads (often) to poor data quality
  - Guessing
  - Abandonment
- Solution may lie in attribute-based data collection and management
  - Basic categories provide a way to manage attributes
  - Useful information at a more abstract level

Toward a Solution: Instance-based data model

Goal: max quality, min participatory constraints; Assumption: low level of expertise

Long beak

White breast

Black wings

Blue feet

Webbed feet

Stands on rock

Bird

Sea bird?

Duck?

Instance

Objective

Objective

Low level

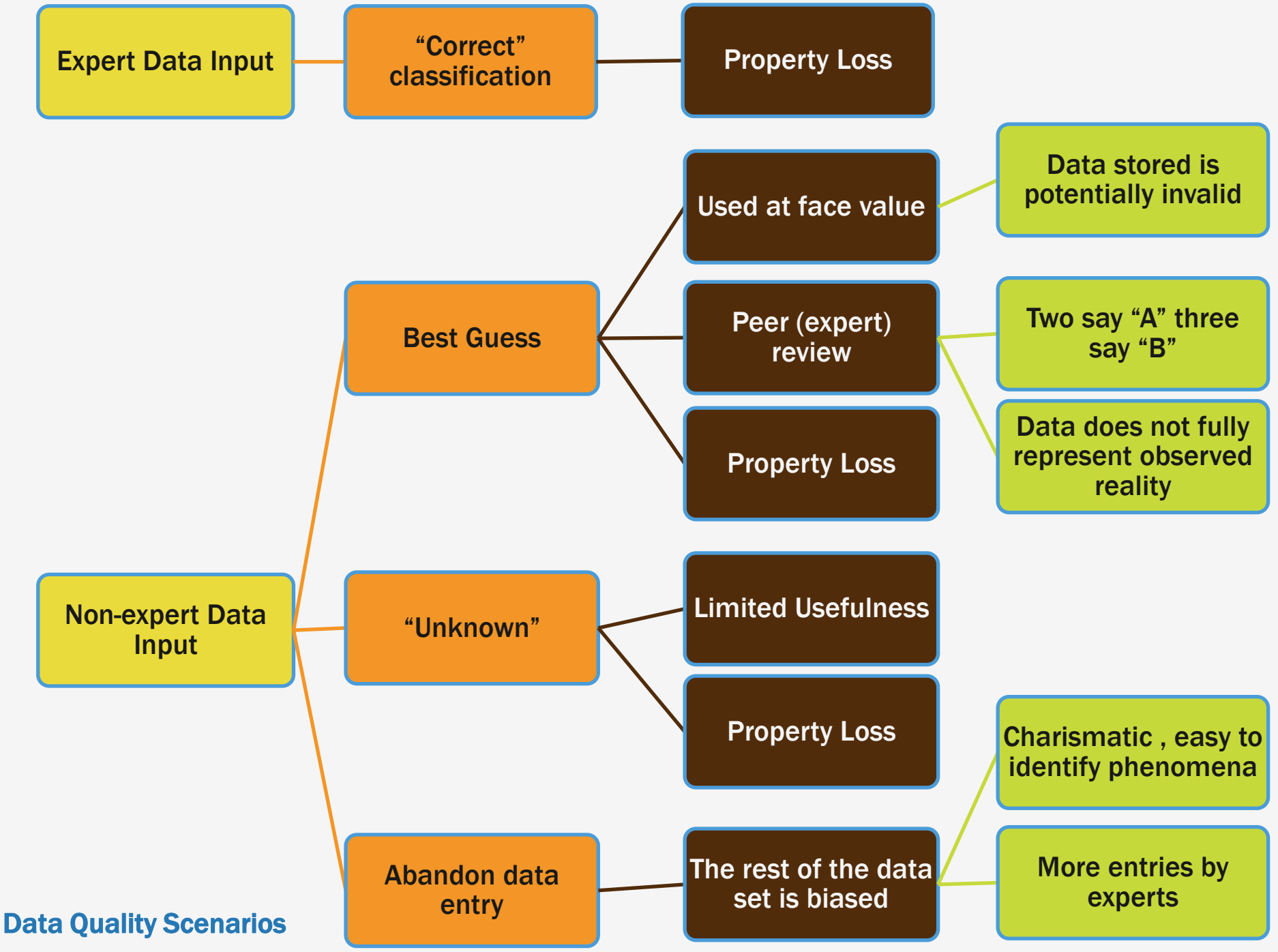
Individualistic

Context

Efficiency

Soft Constraints

Quality varies



Data Quality Scenarios