Frontiers of

Biodiversity Data Science:

What can/can't we learn

from Natural History Collections Data?

Suggested Plan for This Session

Three broad topics, each with two variants

- **Natural History Collections Knowledge**
  - **Do know**
  - **Don't know**
- **Natural History Collections Data**
  - **Is digitized**
  - **Isn't digitized**
- **Machine Learning from Natural History Collections Data**
  - **Can do**
  - **Can't do**

Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity

J. Hortal et al. 2015

Table 1 Definitions (and original references) for the seven main shortfalls of biodiversity knowledge

Shortfall	Aspect of biodiversity	Definition
Linnean	**Species**	**Most of the species on Earth have not been described and cataloged (Brown & Lomolino 1998); this concept can be extended to extinct species (Hortal et al. 2015)**
Wallacean	**Geographic distribution**	**Knowledge about the geographic distribution of most species is incomplete; it is inadequate at all scales most of the time (Lomolino 2004)**
Prestonian	**Populations**	**Data on species abundance and population dynamics in space and time are often scarce (Cardoso et al. 2011)**
Darwinian	**Evolution**	**Lack of knowledge about the tree of life and the evolution of species and their traits (Diniz-Filho et al. 2013)**
Raunkiæran	**Functional traits and ecological functions**	**Lack of knowledge about species' traits and their ecological functions (Hortal et al. 2015)**
Hutchinsonian	**Abiotic tolerances**	**Lack of knowledge about the responses and tolerances of species to abiotic conditions (i.e., their scenopoetic niche; Hortal et al. 2015, redefined from Cardoso et al. 2011)**
Eltonian	**Ecological interactions**	**Lack of knowledge on species' interactions and these interactions' effects on individual survival and fitness (Hortal et al. 2015)**

The Seven Impediments in Invertebrate Conservation and How to Overcome Them

P. Cardoso et al. 2011

Three Societal Dilemmas

- Public Dilemma
  - People throughout the world do not recognize invertebrates or their roles in the ecosystem.
  - In consequence, the public has the tendency to disregard invertebrate species as in need of protection.
- Political Dilemma
  - Many policymakers and stakeholders see invertebrates as species that, if needed,
  - are indirectly protected by "umbrella" vertebrate species.
  - In consequence, protection measures and funding are limited.
- Scientific Dilemma
  - The discovery and description of new species and the collecting of spatial and temporal data on known species
  - are increasingly regarded as dated science.
  - In consequence, taxonomy and classical ecology are underfunded.

Four Scientific Shortfalls

- Linnean
- Wallacean
- Prestonian
- Hutchinsonian (initial formulation)

Biodiversity Data Science

Machine Learning on Big Data

Volume

Velocity

Variety

Variability

Validity

Veracity

Vagueness

Vocabulary

Venue

Value

Viability

Volatility

Virtue

Vulnerability

Visualization

Machine Learning Challenges

Unsupervised Learning

Natural Language Processing

Ontologies

Transparency

Explanations

Success / Failure Rates and Weights

Tuning