# Statistical Core of AI

Tian Zheng

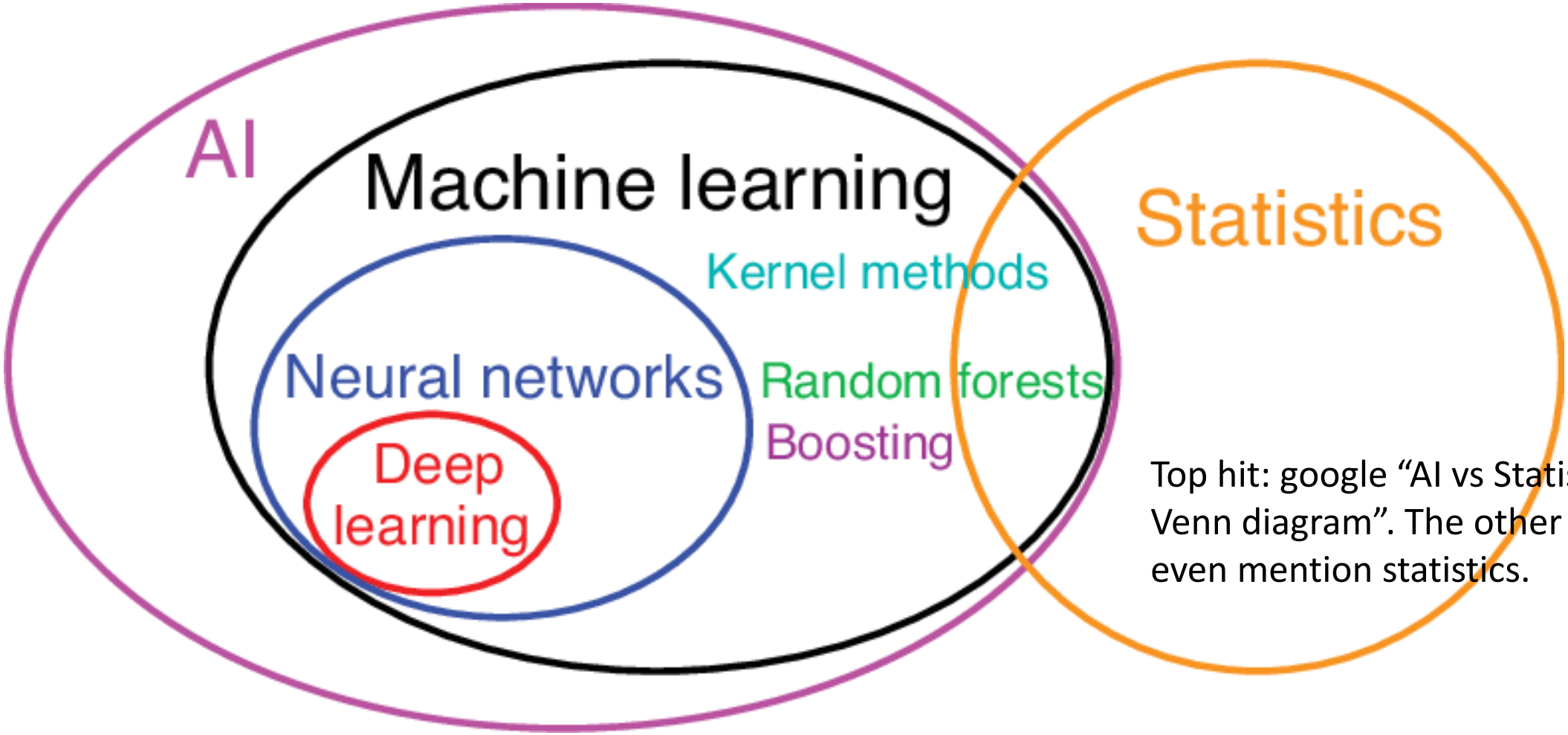
Department of Statistics, Columbia University

AI Day for Federal Statistics: CNSTAT Public Event

May 2, 2024

National Academy of Sciences, Washington, DC

# Repositioning Statistics as a Core of AI?




Top hit: google “AI vs Statistics Venn diagram”. The other don’t even mention statistics.

Hsieh, William W. "Evolution of machine learning in environmental science—A perspective." *Environmental Data Science* 1 (2022): e3.

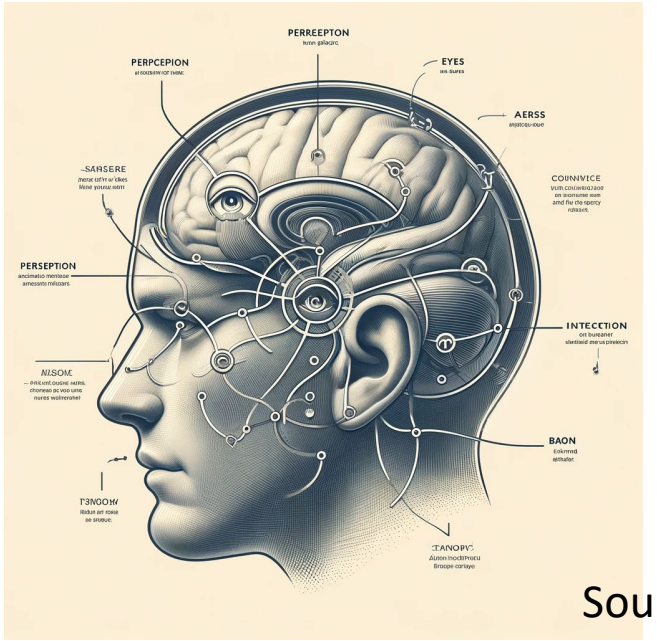
# Definition: Intelligence, Statistics, and Artificial Intelligence

## Intelligence

(Wikipedia) “the ability to perceive or infer information; and to retain it as knowledge to be applied to adaptive behaviors within an environment or context.”

## Statistics

(Wikipedia) a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data … in the context of uncertainty and decision-making in the face of uncertainty.



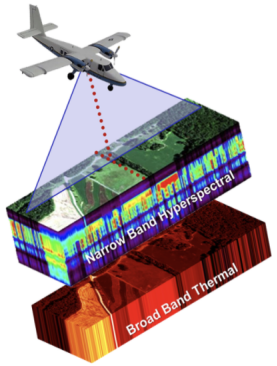


Source: Dall-E. Typos as expected.

# Ecological remote sensing data for studying forest structure

- Traditional approach: ground based observations
  - Expensive, hard to scale, dangerous, require expert knowledge
- New technology: remote sensing techniques
  - High-resolution forest data
  - Automatic, large-scale, no expert labels

**G-LiHT v.2: Instruments & Specifications**



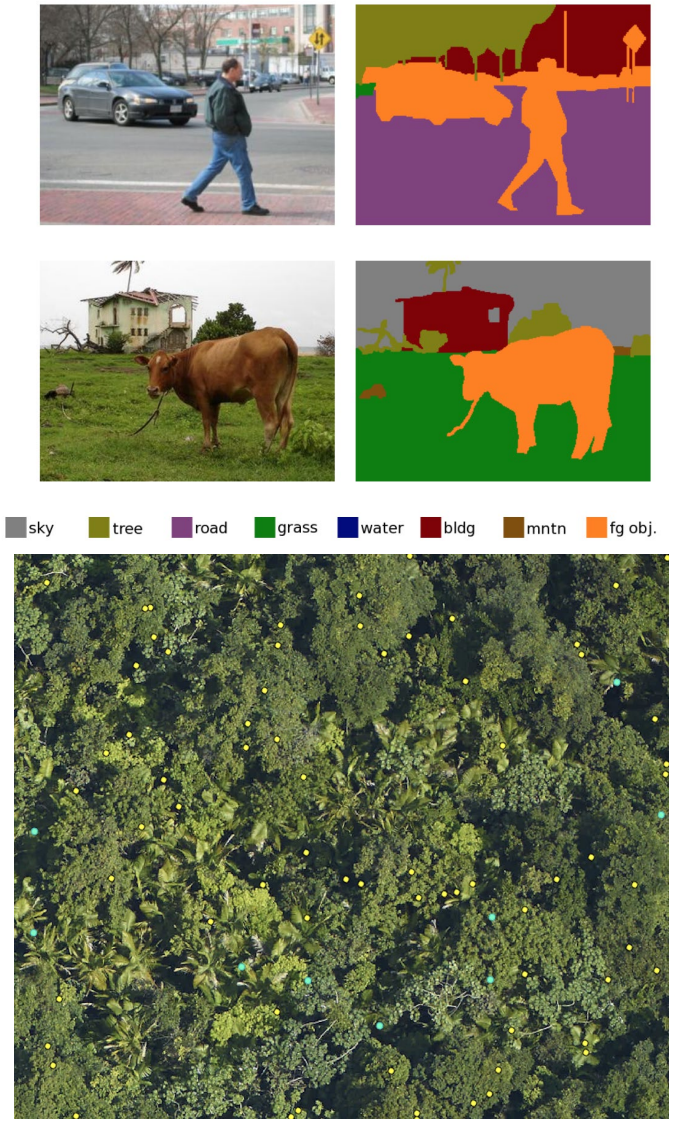

Designed to study composition, structure, and function of terrestrial surfaces using scanning LiDAR, Hyperspectral, and Thermal infrared imaging.

1. Dual Scanning LiDARs
2. VNIR Imaging Spectroscopy
3. Broad Band Thermal Imaging
4. High Resolution Aerial Photos
5. Precision GPS-INS

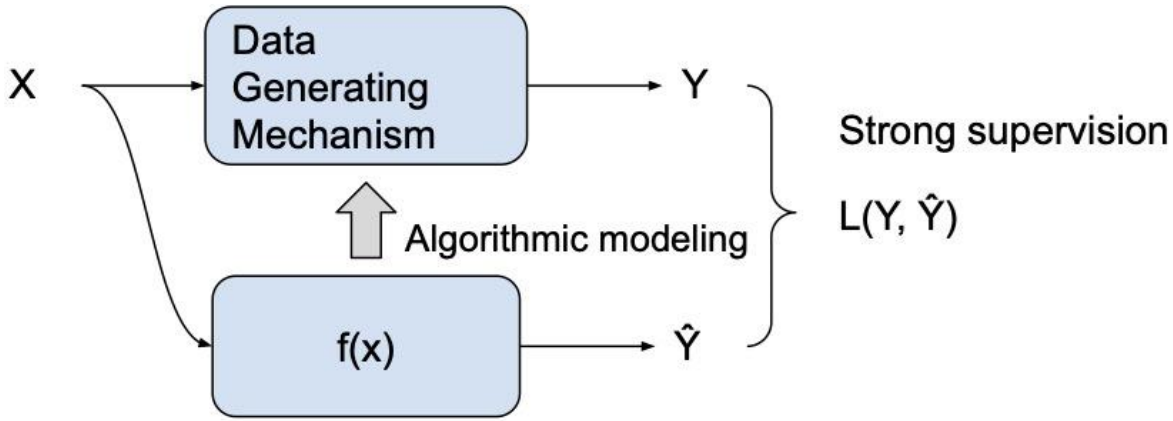
**Specific trade names are for informational purposes only and do not constitute an endorsement by NASA.*

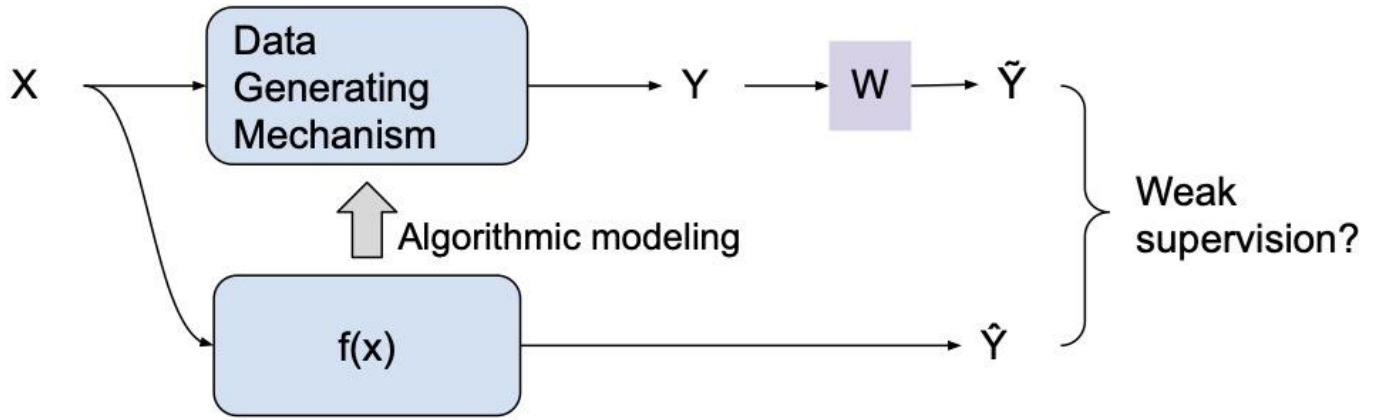
# Seeing the trees from the rainforest?

- Unstructured data
- Lack of foreground and background
- Limited spatial coverage (<1%) of area has ground-based labels
- Substantial displacement between tree root locations and canopies

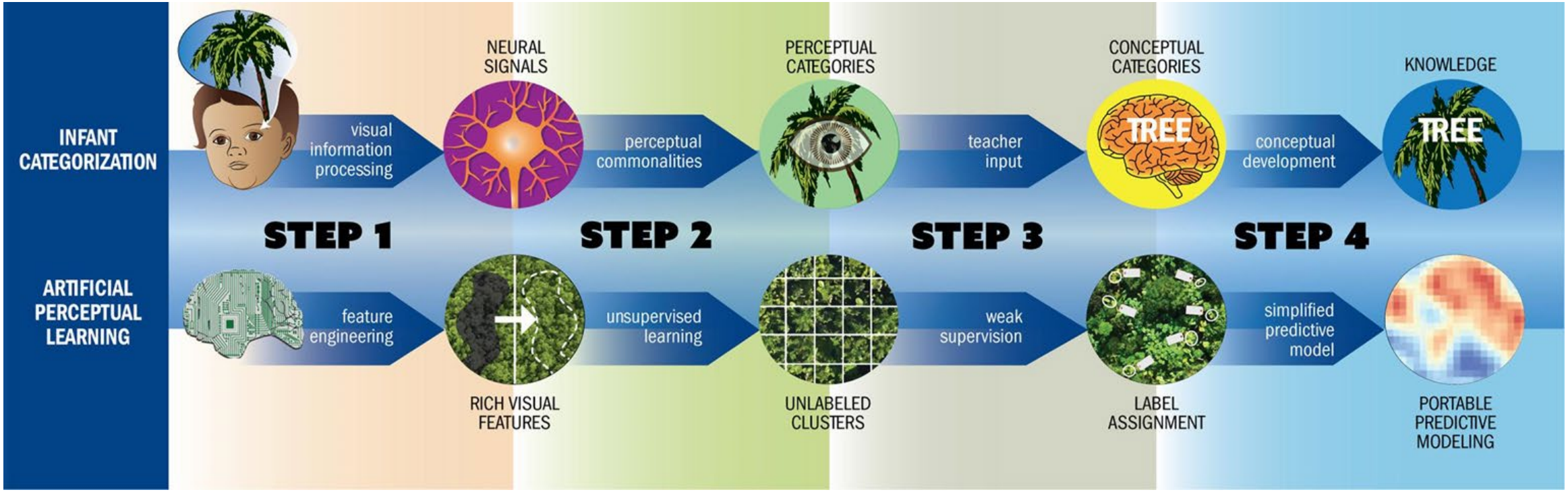


# Learning *without* strong supervision



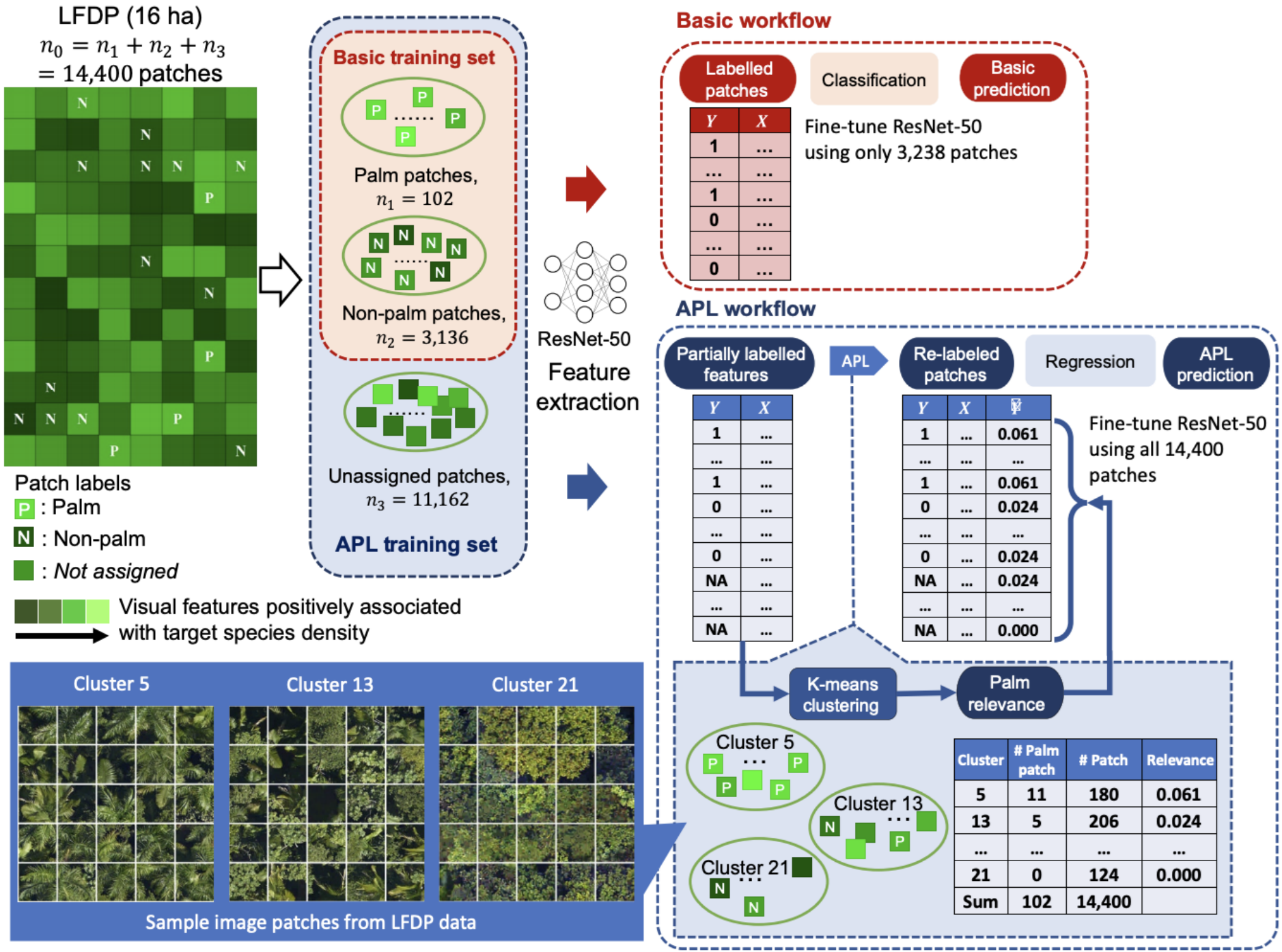



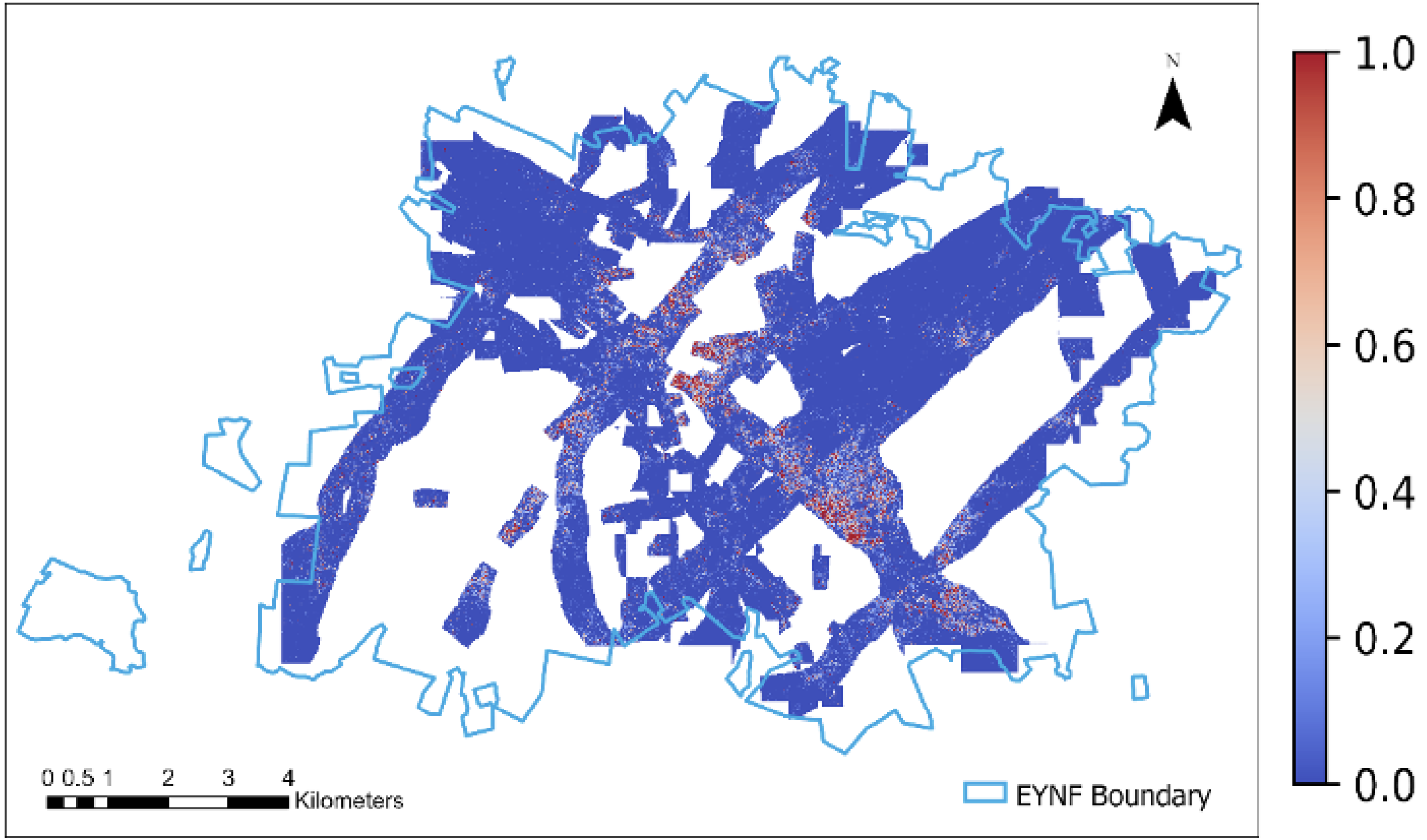
# Artificial Conceptual Learning




Tang, C., Uriarte, M., Jin, H., C Morton, D., & Zheng, T. (2021). Large-scale, image-based tree species mapping in a tropical forest using artificial perceptual learning. *Methods in Ecology and Evolution*, *12*(4), 608-618.

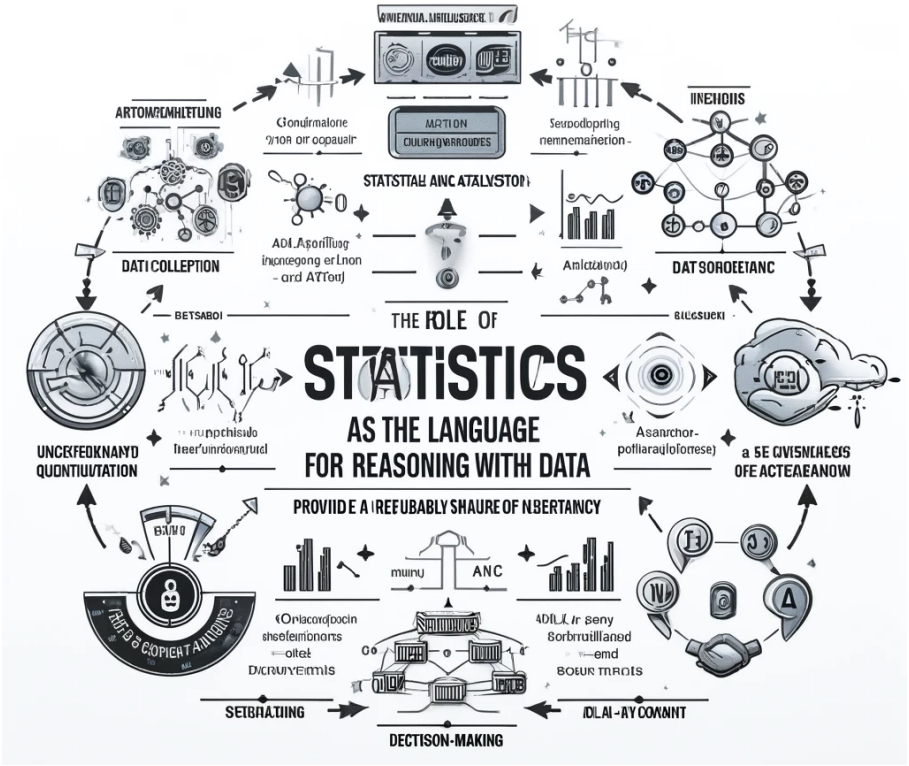
# Artificial Conceptual Learning




N
1.0
0.8
0.6
0.4
0.2
0.0
0 0.5 1 2 3 4
Kilometers
EYNF Boundary

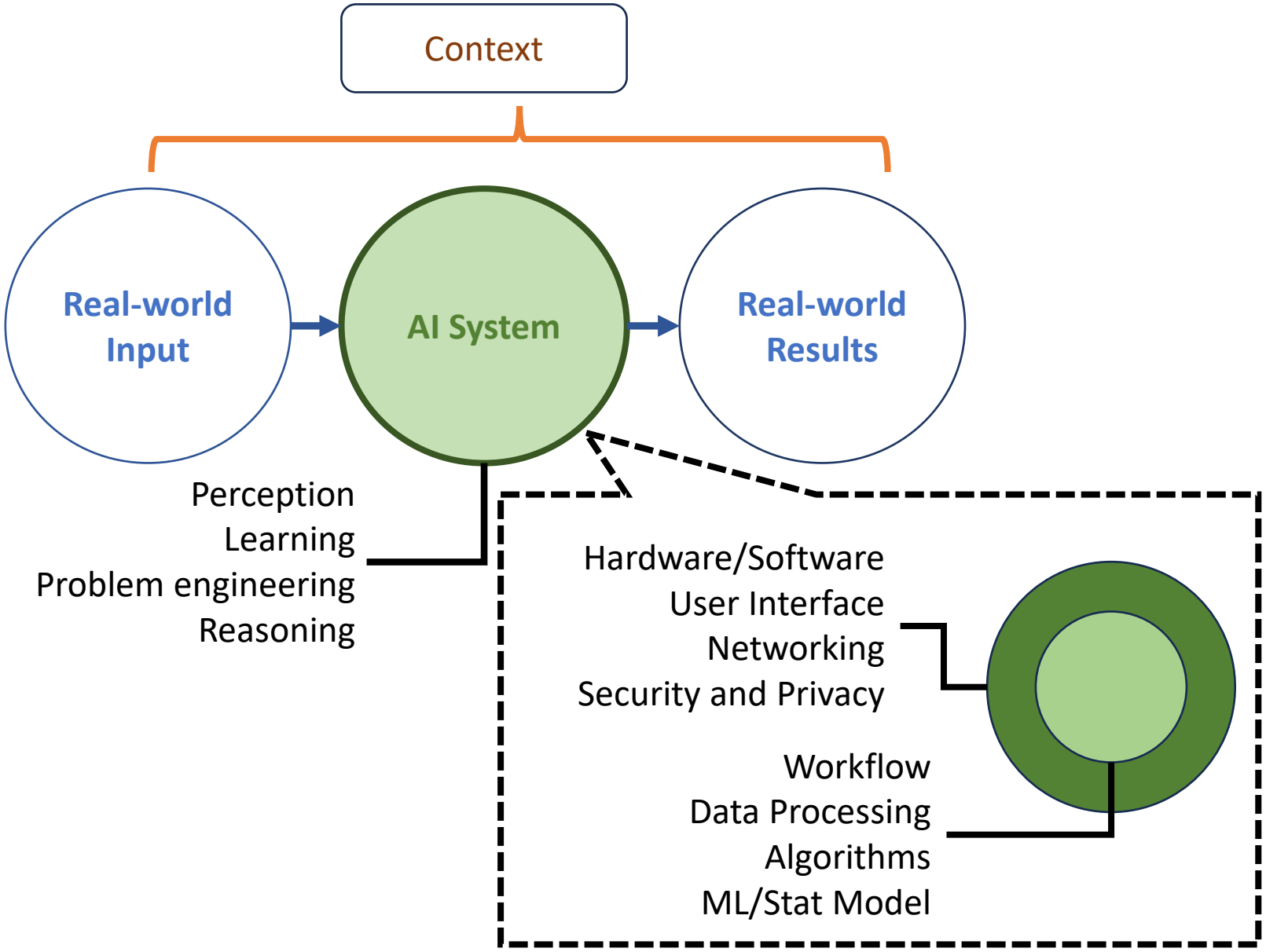
# Statistics for AI: how to reason with data

- Move from the accuracy vs interpretability dichotomy towards pursuing *reliability* and *utility*
  - Uncertainty quantification
  - Knowledge-informed learning
  - Multimodal data and multiscale systems
  - Design of verification, evaluation, and interpretation experiments
- Statistics is good at *being reliable*. We could be more proactive on *being useful*.




Source: Dall-E. Typos as expected.

# Where does statistics fit in the core of AI




**Development Tasks:**

- **Problem identification and set up**
- Problem modulization
- Metric development
- Workflow development
  - Training
  - Deployment
- Data Engineering
- Model development
- Model evaluation
- System development and deployment
- System evaluation and testing

**Development resources:**

- **Training data**
- Computing infrastructure
- Engineering resources
- **Domain knowledge**
- Data science expertise

**Domain Science** — Statistics skills; Computer/Data science skills

# Repositioning Statistics as a Core of AI! How?

## -- From the 2019 “Statistics at Crossroads” report

- The field of Statistics is at a crossroads: we either flourish by embracing and leading Data Science or we decline and become irrelevant.
- In the long run, to thrive, we must redefine, broaden, and transform the field of Statistics.
- We must evolve and grow to be the transdisciplinary science that collects and extracts useful information from data.

https://www.nsf.gov/mps/dms/documents/Statistics_at_a_Crossroads_Workshop_Report_2019.pdf

# Repositioning Statistics as a Core of AI!

- Statistics discipline
  - Advocate for openness - we are what we do
  - Move from interdisciplinary mindset to a transdisciplinary mindset
    - Proactively work on foundational problems that are important for other disciplines.
- Statistical community
  - Build up community-level initiatives for impact
    - (Low-hanging fruits) Publishing open problems and datasets with starter codes, run bootcamps and hack events to get interested statisticians started, incentives, training, promote best practices (research, collaboration, credit-attribution)
    - (Hard but necessary) Community-scale actions on infrastructures and resources
    - (Short-term PR actions) Rapid-action campaigns, journal special issues, communication and outreach efforts, conference workshops (both ML/DS/AI ones and domain science ones), collaboration in both research and **education**.
    - (Long-term planning) Networks of resourceful stakeholders

# Repositioning Statistics as a Core of AI!

- Statistical community (cont'd)
  - Support next-generation statisticians who are motivated to have impacts.
    - Training skills for impact
      - Engineering skills; Collaboration and communication
      - Train the trainers, shared curricula
    - Recognize the need to admit and train statisticians who are future AI leaders
      - Training grants, targeted fellowships, program tracks
    - Community building
      - Skill workshops and hack events within a meeting; challenges, pooling resources and efforts
    - Mentoring, support, and promotion.

Stats Up AI Alliance

# Stats Up AI – https://statsupai.org/

Be a scientist
Team up
Education reform
Leadership
Collaboration