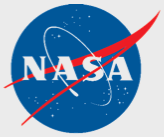
Amy Braverman

(Jet Propulsion Laboratory)

Inference and Uncertainty

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Before we begin, the goal of these modules is to:

- Present statistical inference and its role in data analytics in the simplest terms possible.
  - Fewest number of topics.
  - Least amount of math (but assuming knowledge of some calculus and limits).
  - Not intended to be comprehensive or thorough.
- Three broad topic areas:
  - Review of basic probability.
  - Basic concepts of inference.
  - Introduction to two popular resampling (non-parametric) procedures for inference.

Outline

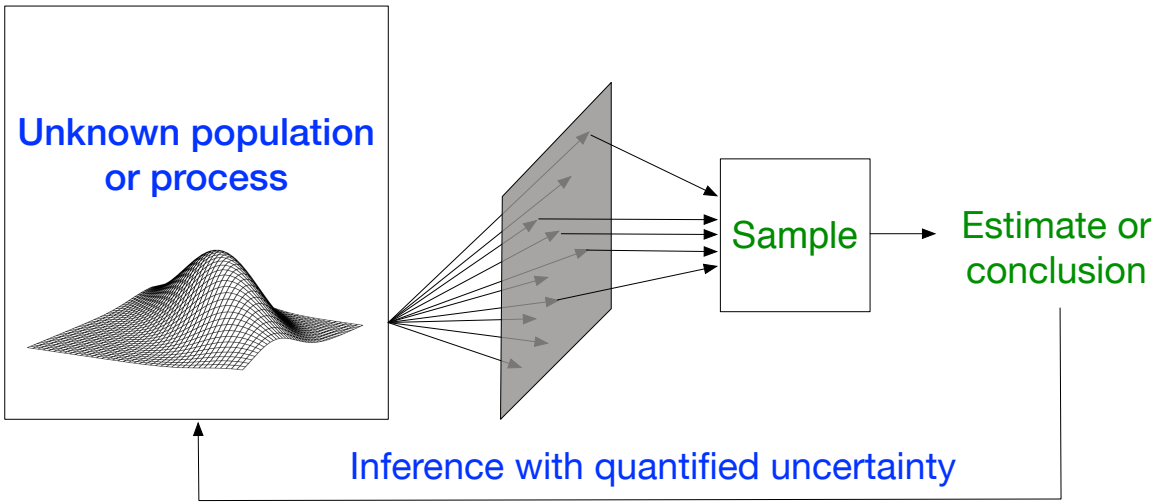
Goal of this part is to introduce some basic definitions:

- Inference
- Populations and samples.
- Exploratory and confirmatory analysis.
- Massive data sets: populations or samples?

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Inference

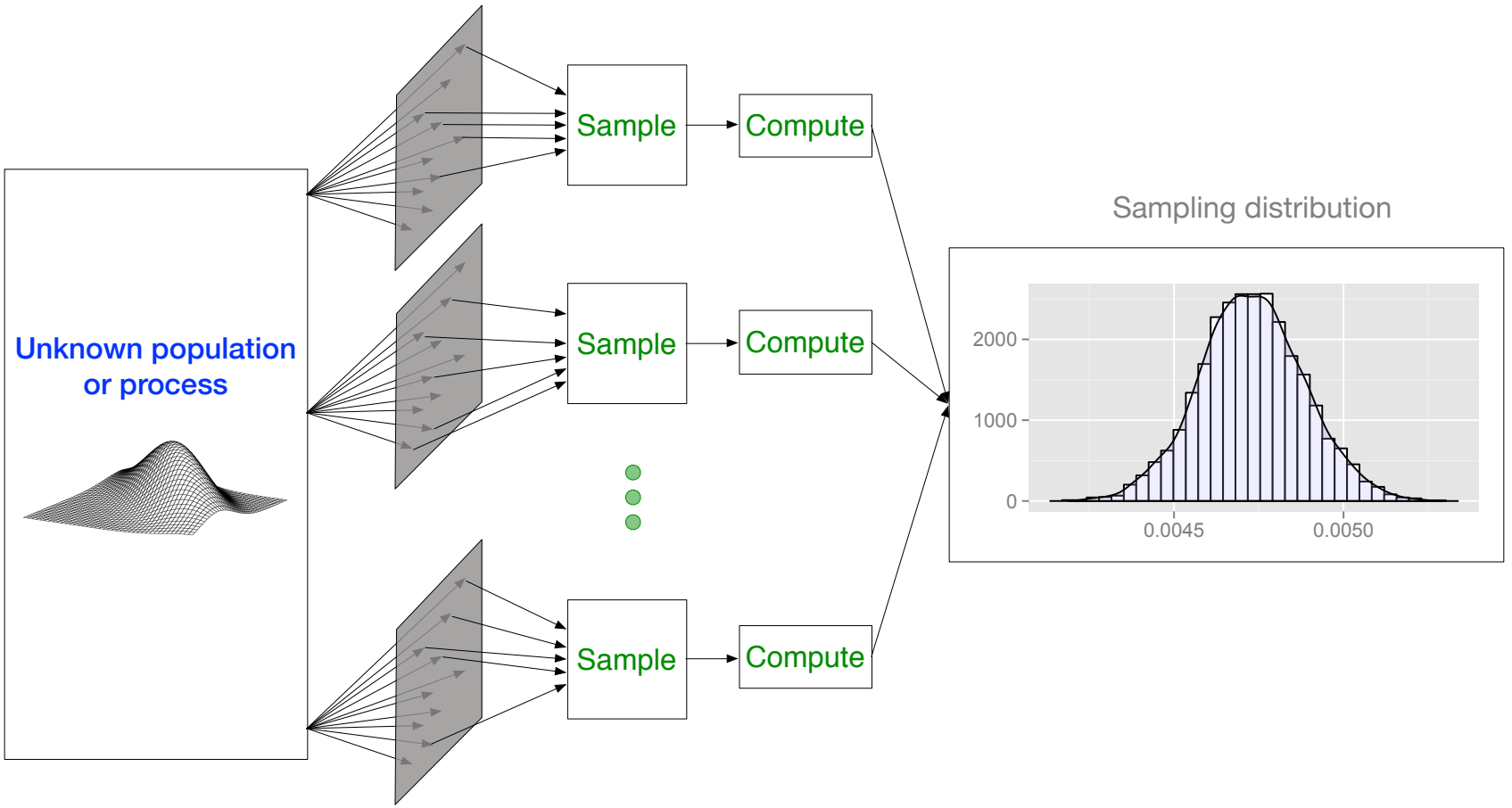


- True population or process is modeled probabilistically.
- Sampling supplies us with realizations from the probability model.
- Compute something, but recognize that *we could have just as easily gotten a different set of realizations*.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

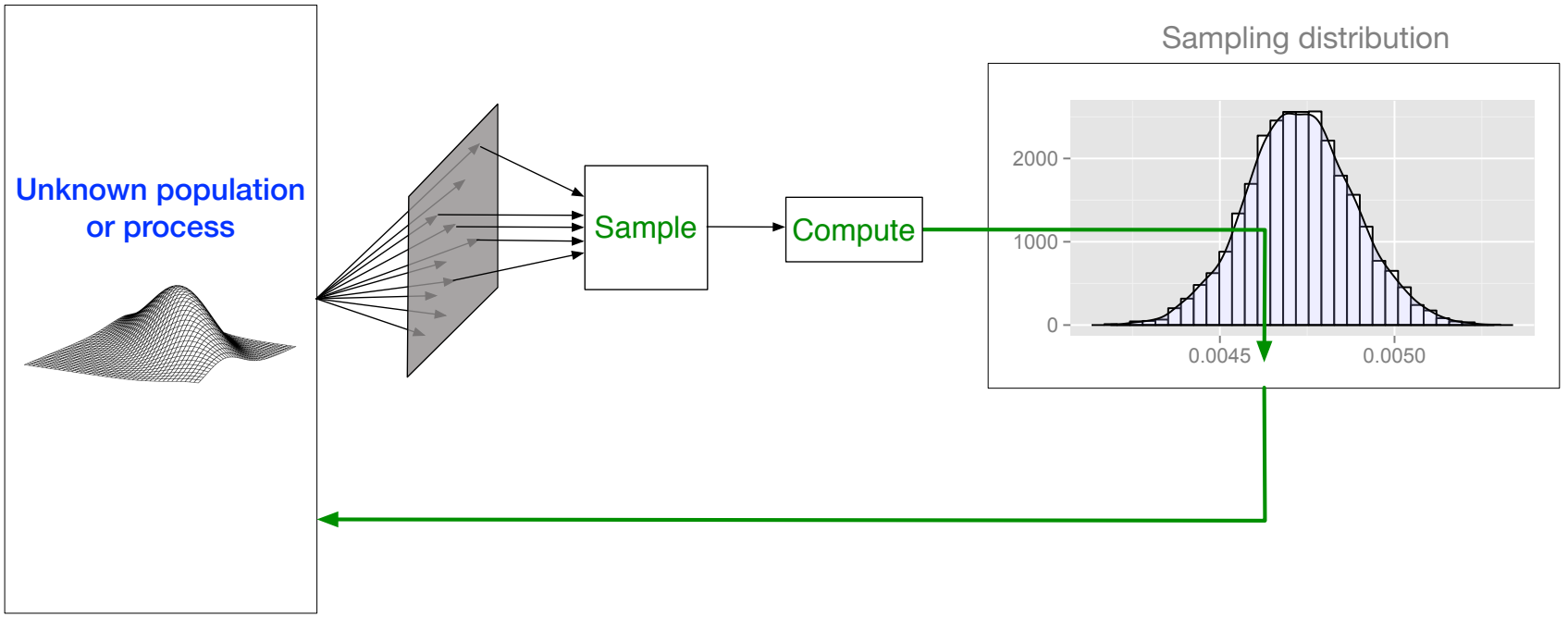
Inference



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Inference



- We want to infer the characteristics of the true probability model from our *one* sample.
- There are generally two approaches: Frequentist and Bayesian. We'll come back to that.

Populations and samples

- An *observational unit* is an object about which we want to know something.
- A *variable* is a quantity we measure on an observational unit.
- A *population* is a collection of observational units.
  - Populations can be finite (e.g., all citizens of the US that are alive today) or infinite (all citizens of the US that were ever or will ever be alive).
  - A physical process or mechanism can be thought of as generating an infinite population (temperature at every moment in time at some location).
- A *parameter* is a quantity computed from all units in the population.

Populations and samples

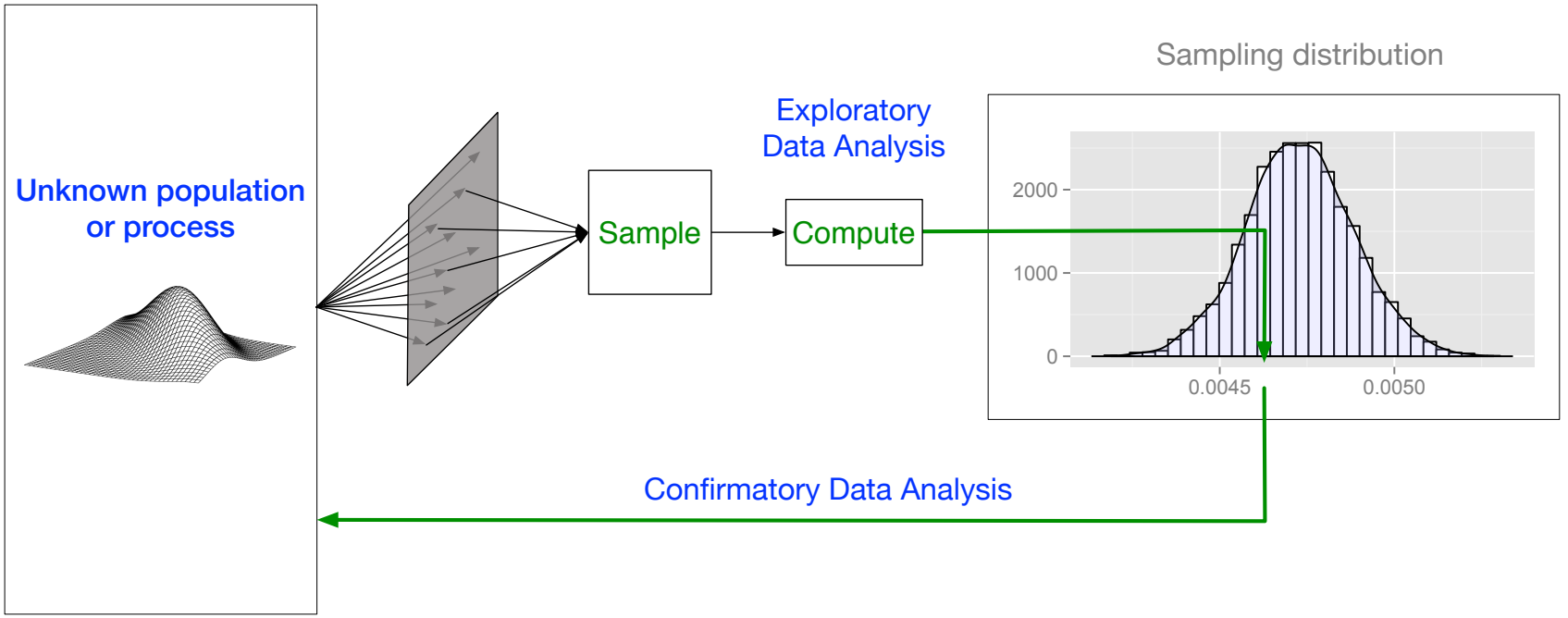
- An *sample* is a subset of the population.
- A *statistic* is a quantity computed from a sample.

Sidebar: observational study vs. experiment.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Exploratory vs confirmatory analysis



Exploratory vs confirmatory analysis

- EDA illuminates structures (patterns, relationships, etc.) in the sample. EDA is often necessary to formulate hypotheses about the unknown population. These hypotheses will be tested using confirmatory data analysis.

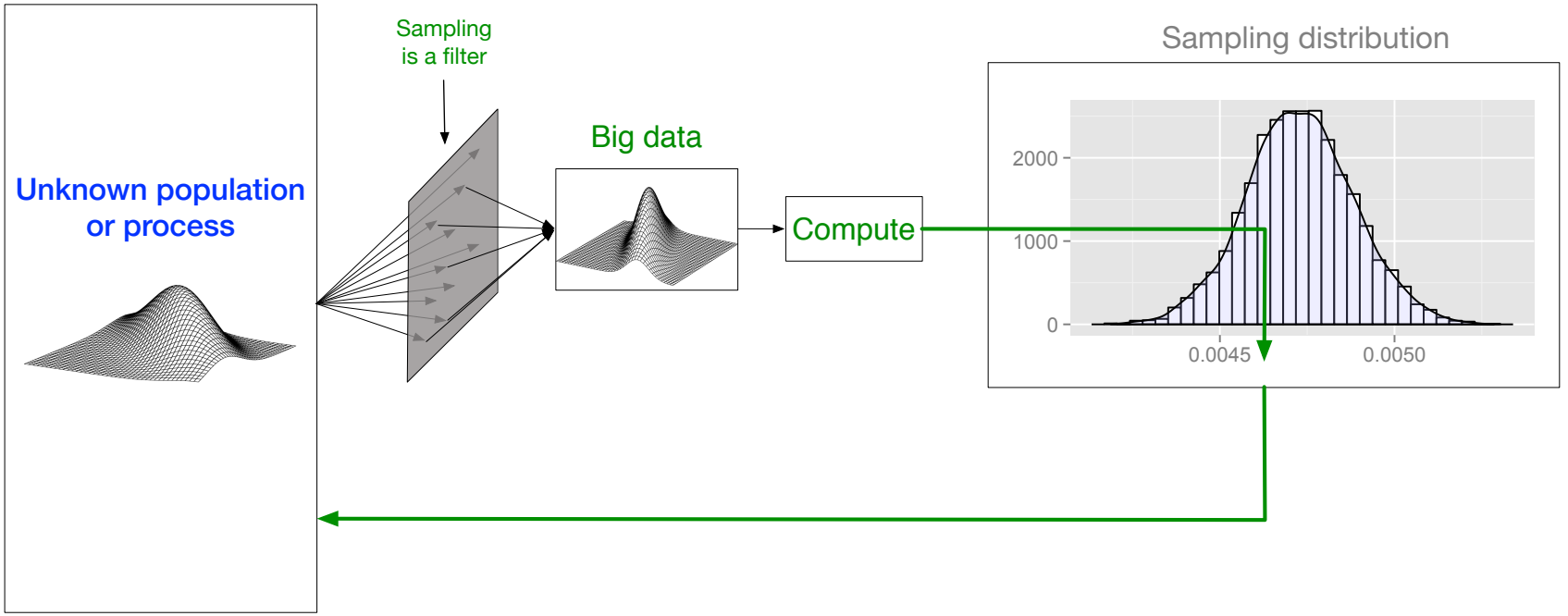
- In confirmatory data analysis (CDA), we use the tools of statistical inference to make definitive probabilistic statements about the population based on the sample.

- Tools of statistical inference: *hypothesis testing* and *estimation*.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Massive data sets: populations or samples?

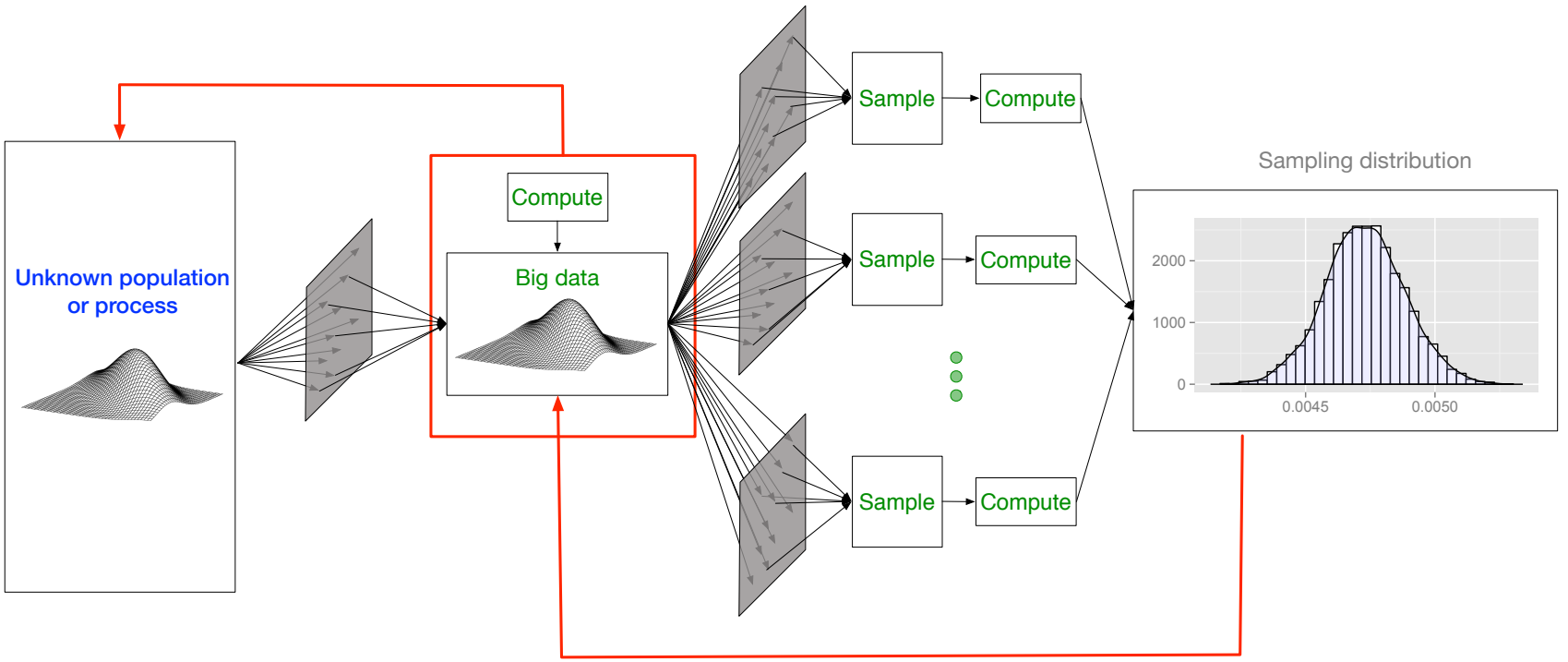


- These days the sample can be “big"; too big to compute with using traditional methods.
- Two strategies: bigger, better, faster algorithms (machine learning) or make the data smaller (another stage of sampling, data reduction).

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Massive data sets: populations or samples?



- Both. Surely we should be able to exploit this to provide probabilistic uncertainties for things we compute from big data.

Next

In the next module, we start the discussion with a brief review of basic probability.