Lecture 1: Preliminaries

Intro to Data Science for Public Policy
Spring 2017

Jeff Chen + Dan Hammer

Roadmap

- What is data science?
- Course Roadmap and Structure
- Hands-On Time
  - Getting familiar with R

An Opening Story


Credit: Nicholas Buer


Where is the best place to photograph the **Milky Way**?
Credit: Nicholas Buer


Death Valley
California


Death Valley at
night with clouds


Olympic National Park
Washington State


Hurricane Ridge
(so far, so good)


Hurricane Ridge
(not so good)

Re-Evaluated Strategy

~~Where is the best place to photograph the Milky Way?~~

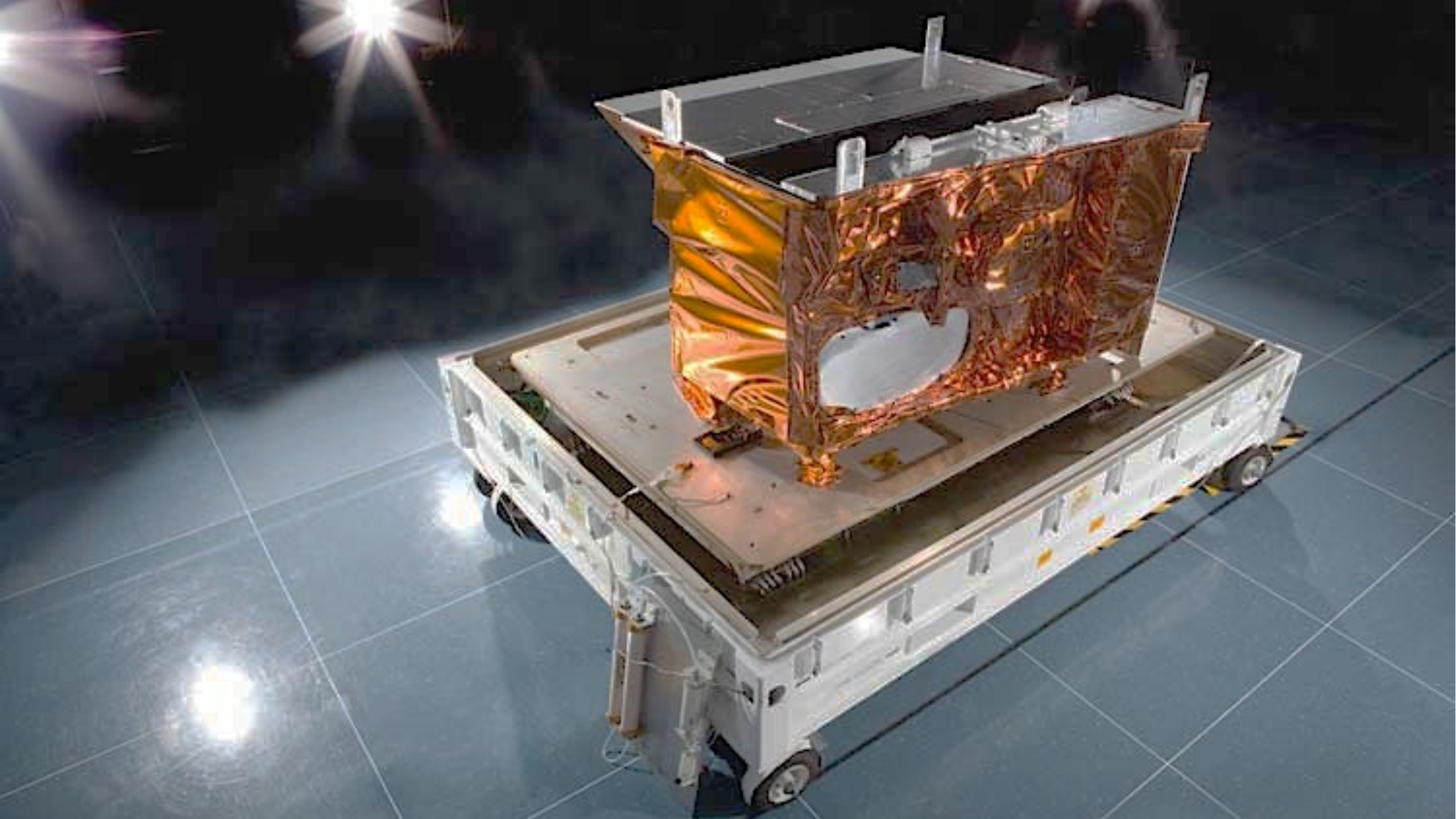
Where is an accessible place to photograph the Milky Way?

Re-Evaluated Strategy

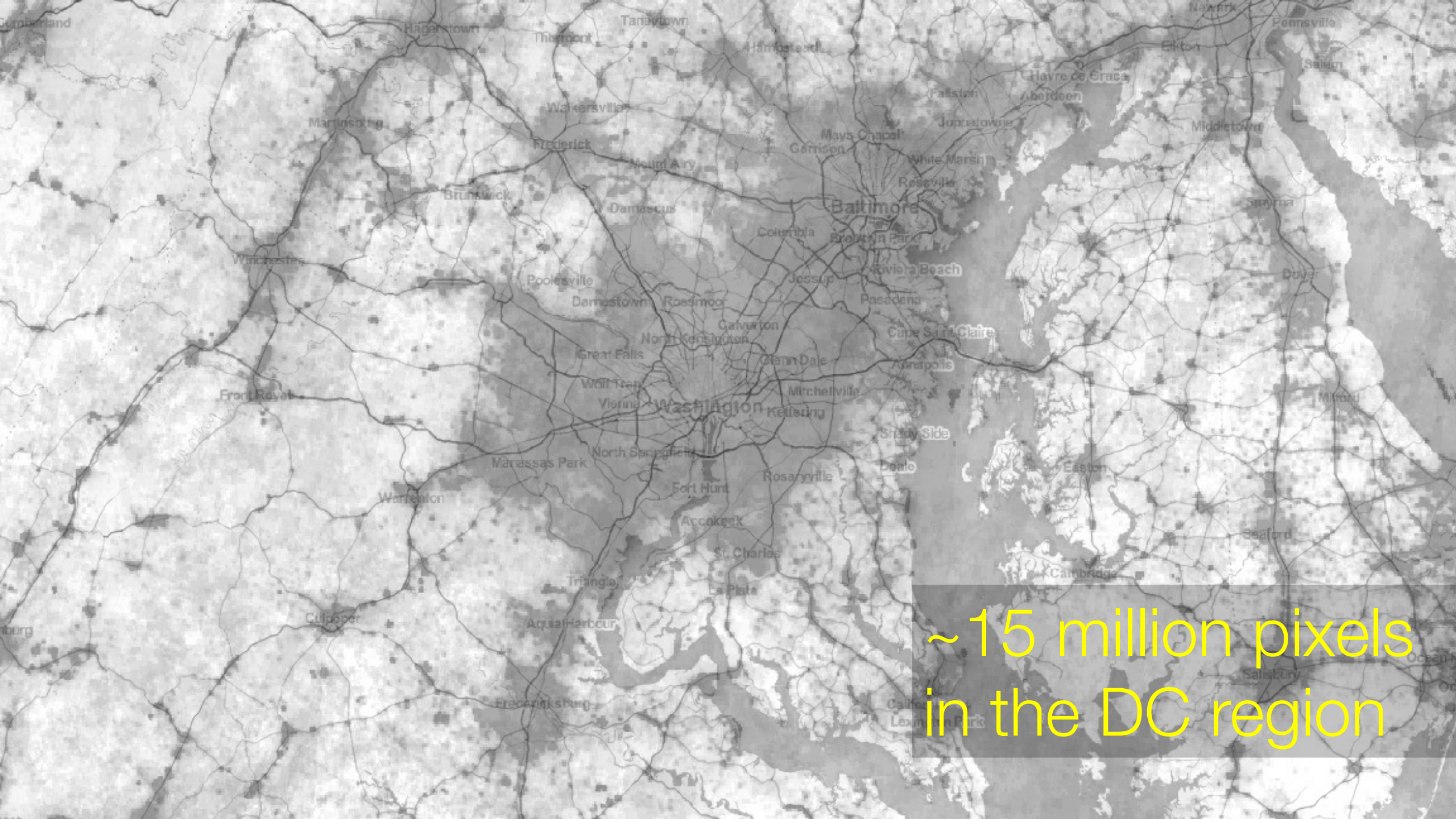
Pr(Minimal Freak Out) =
Pr(~~Poor~~ Good Life Decision)=

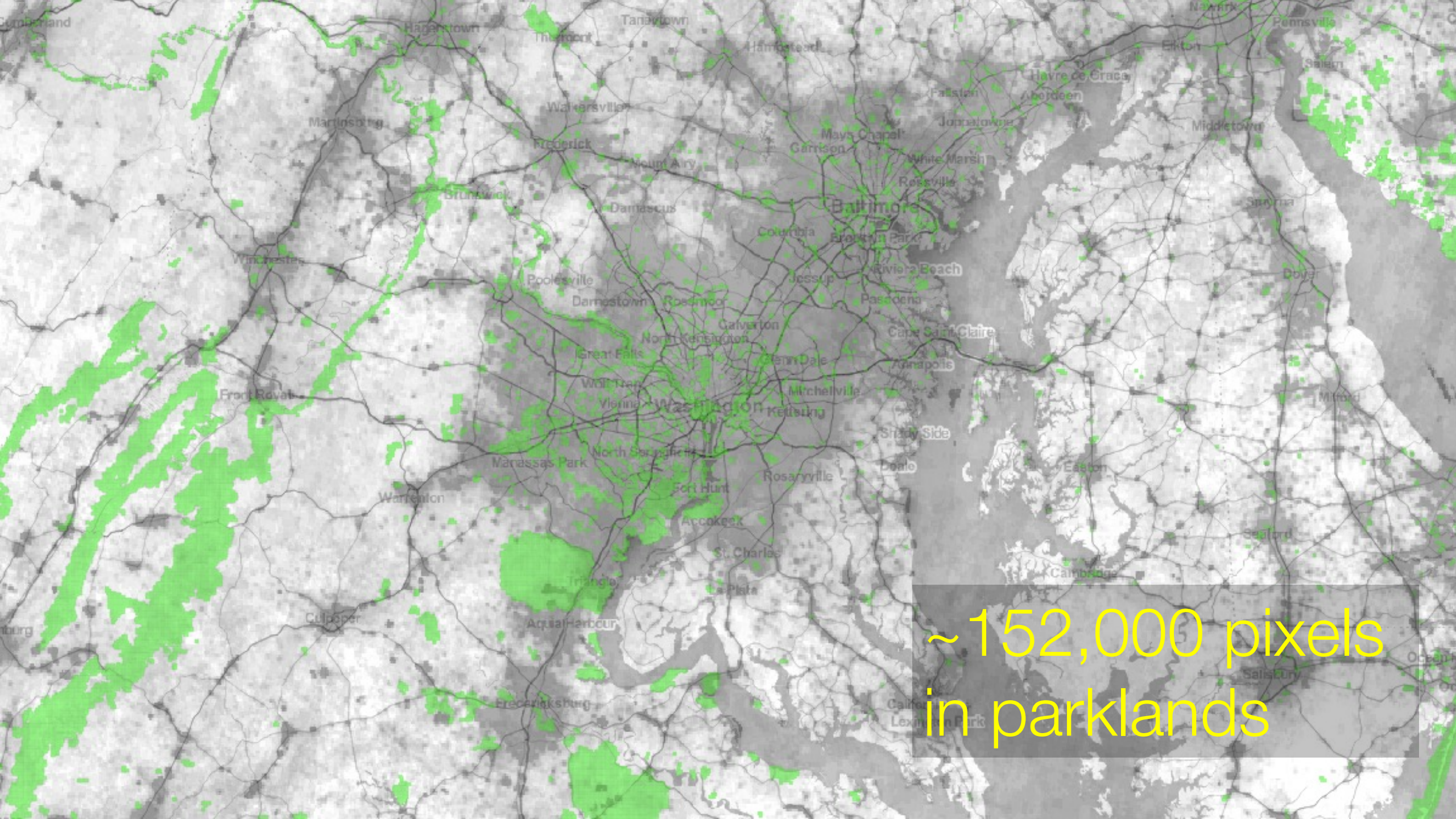
f(Dark enough,
Close by,
Clear skies)

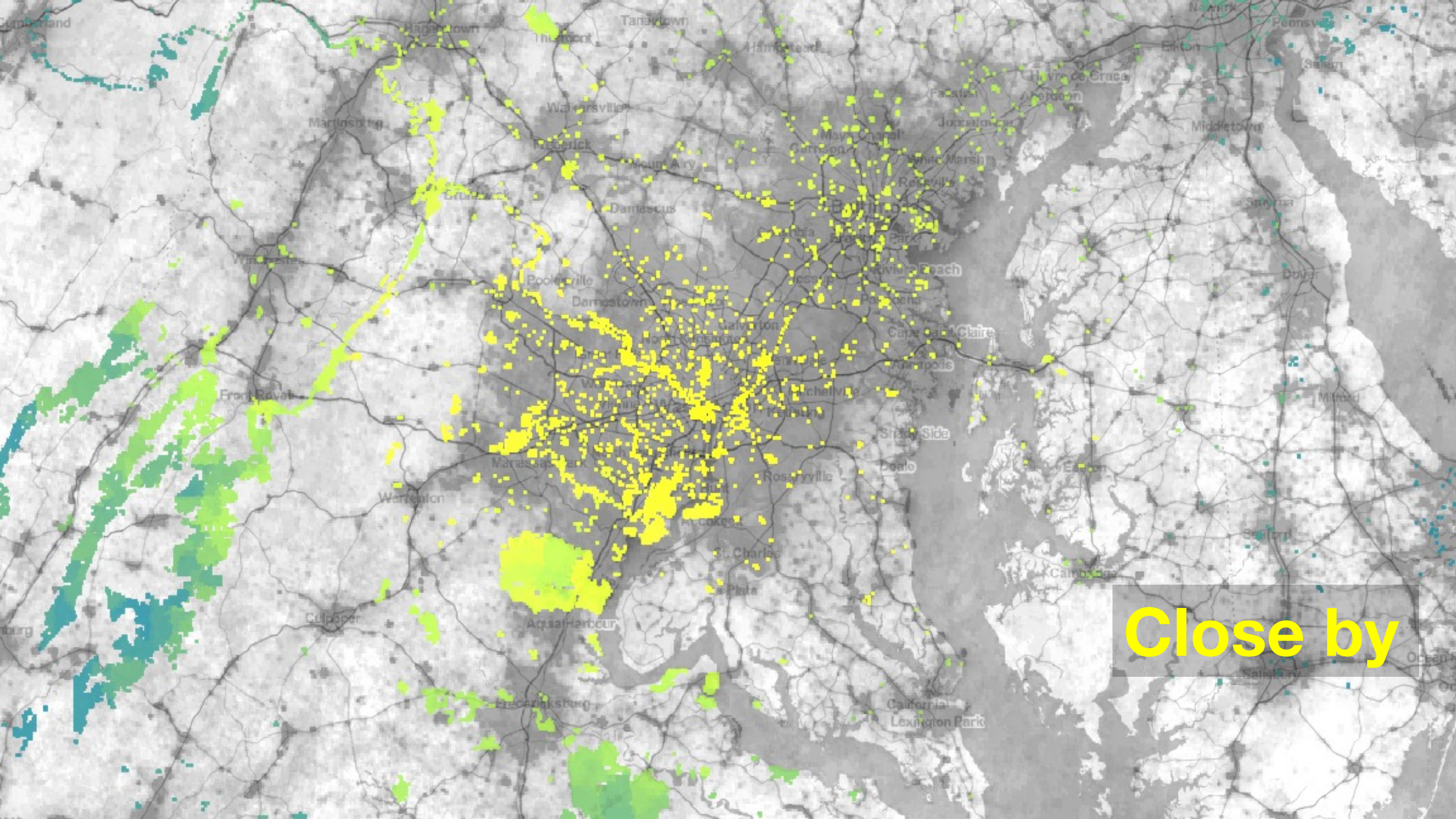
To be able to look **up**, we need to first look **down**.

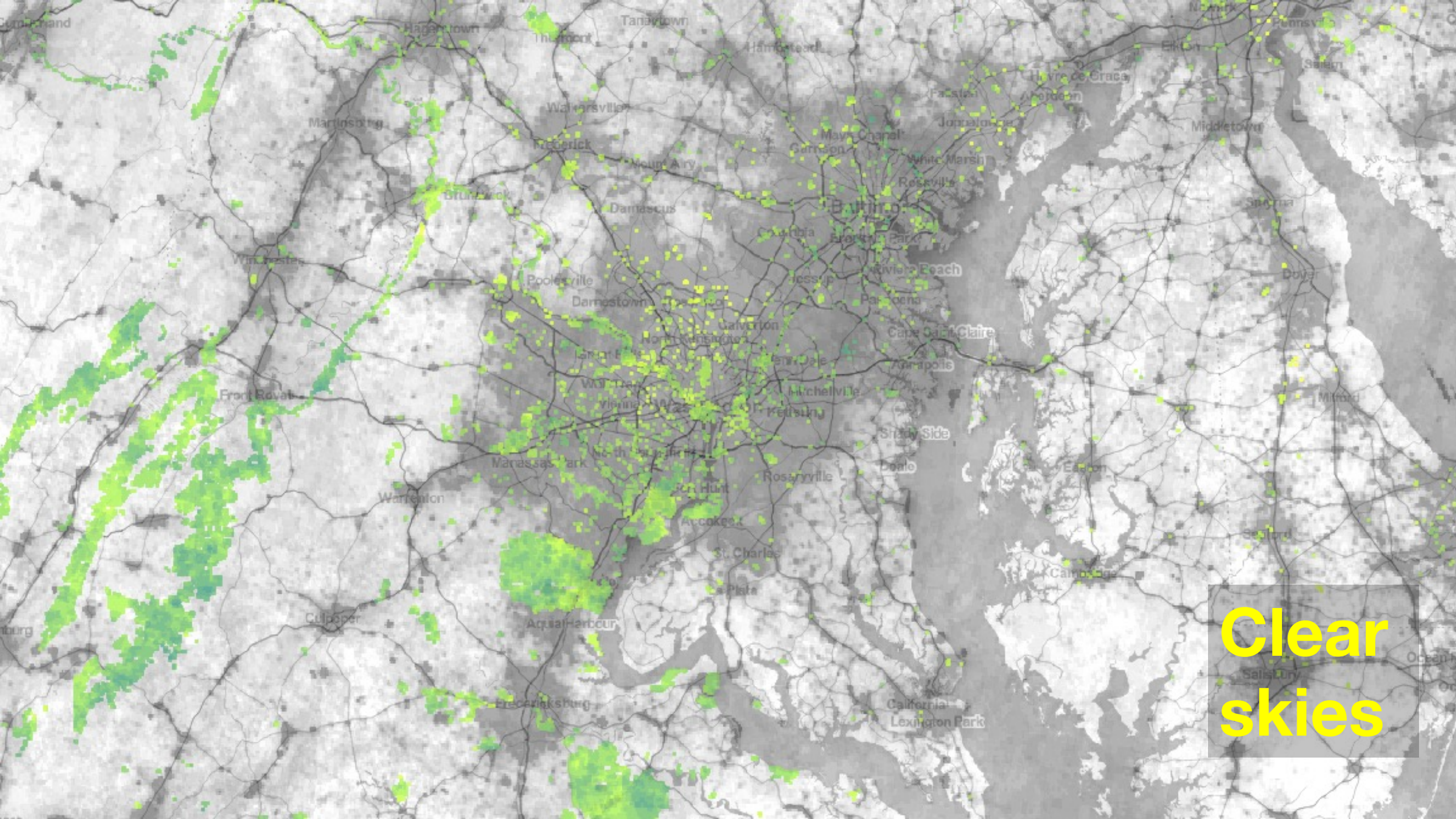


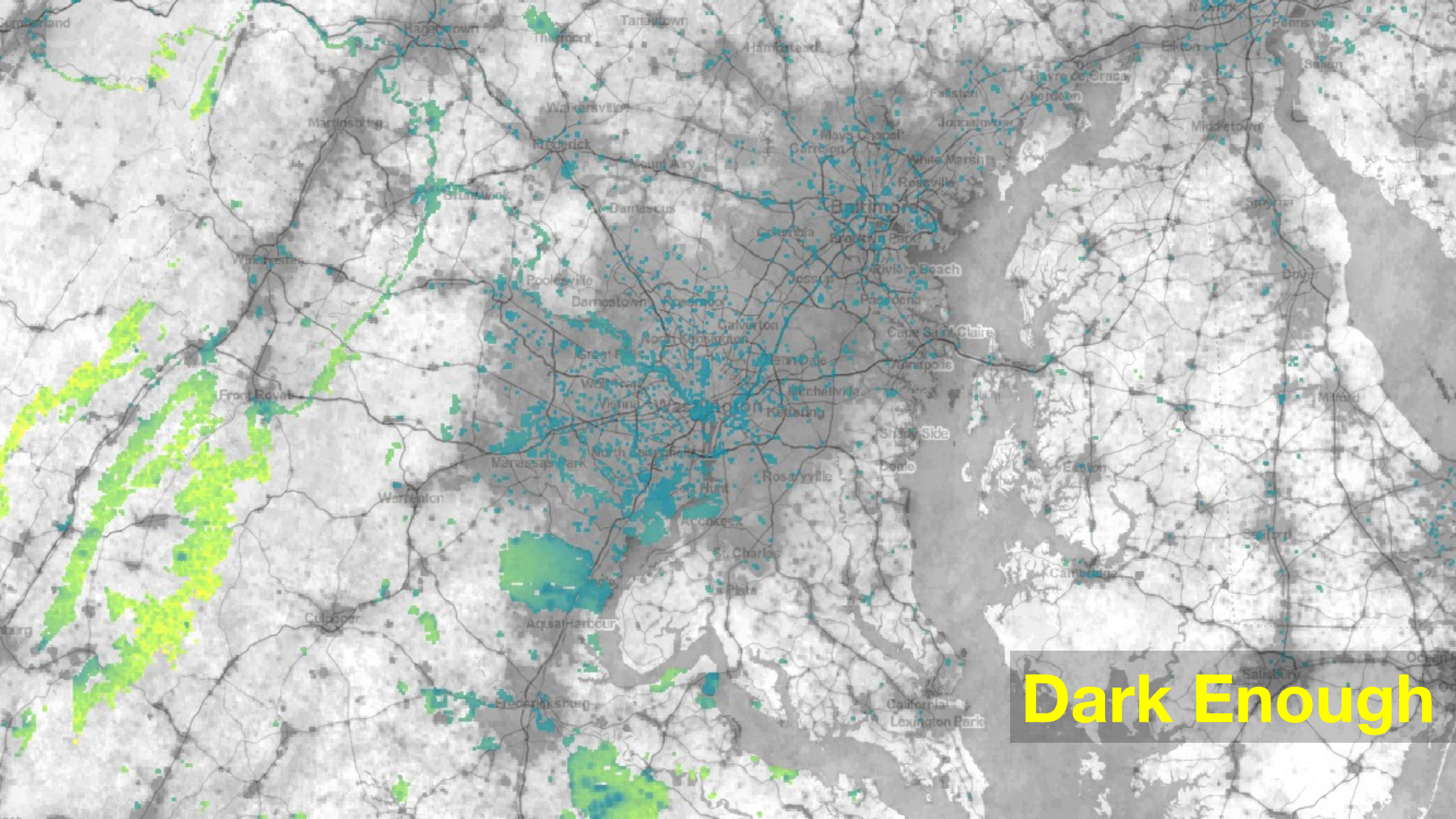

~518 million pixels
or potential locations

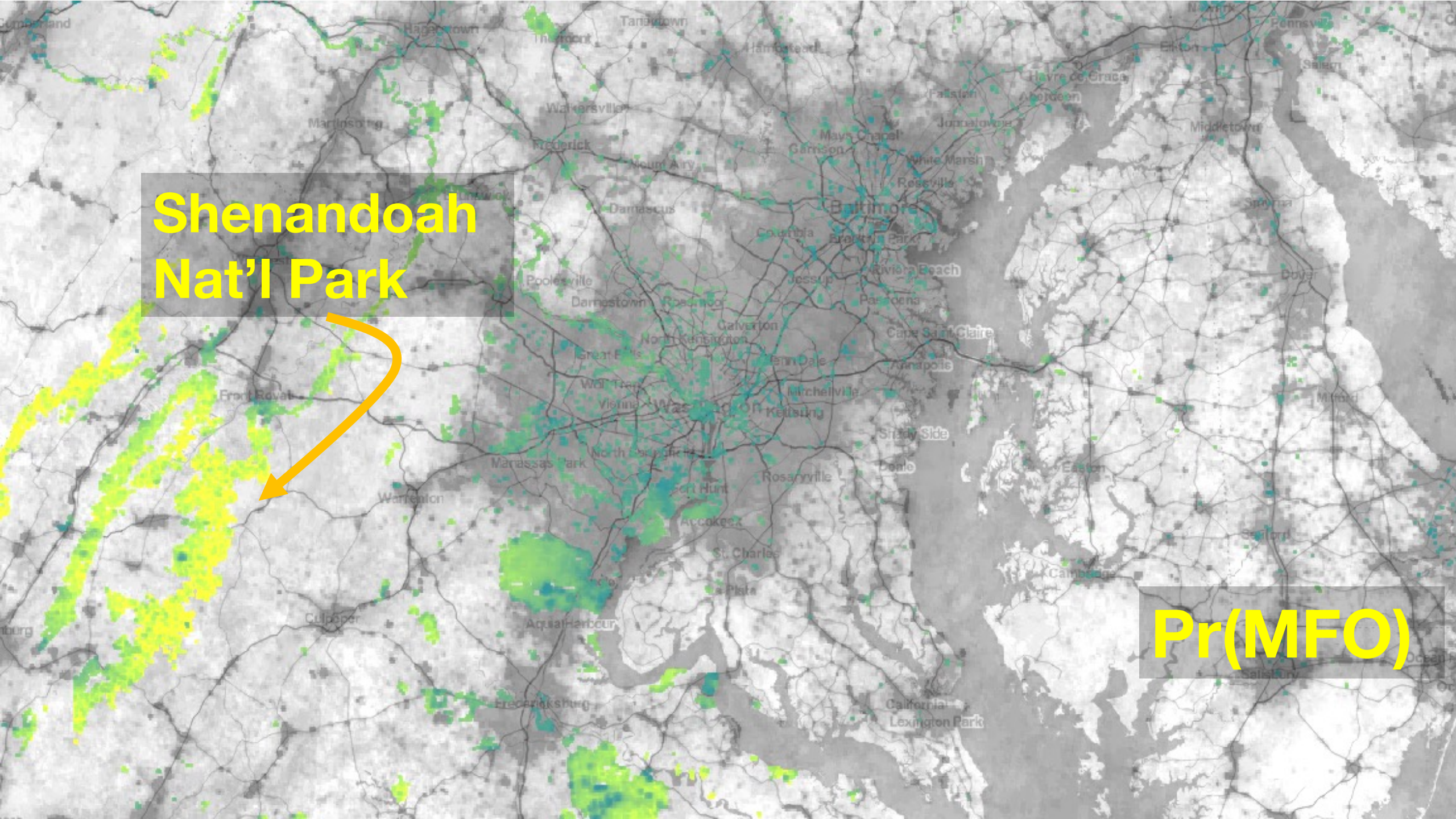

~15 million pixels
in the DC region

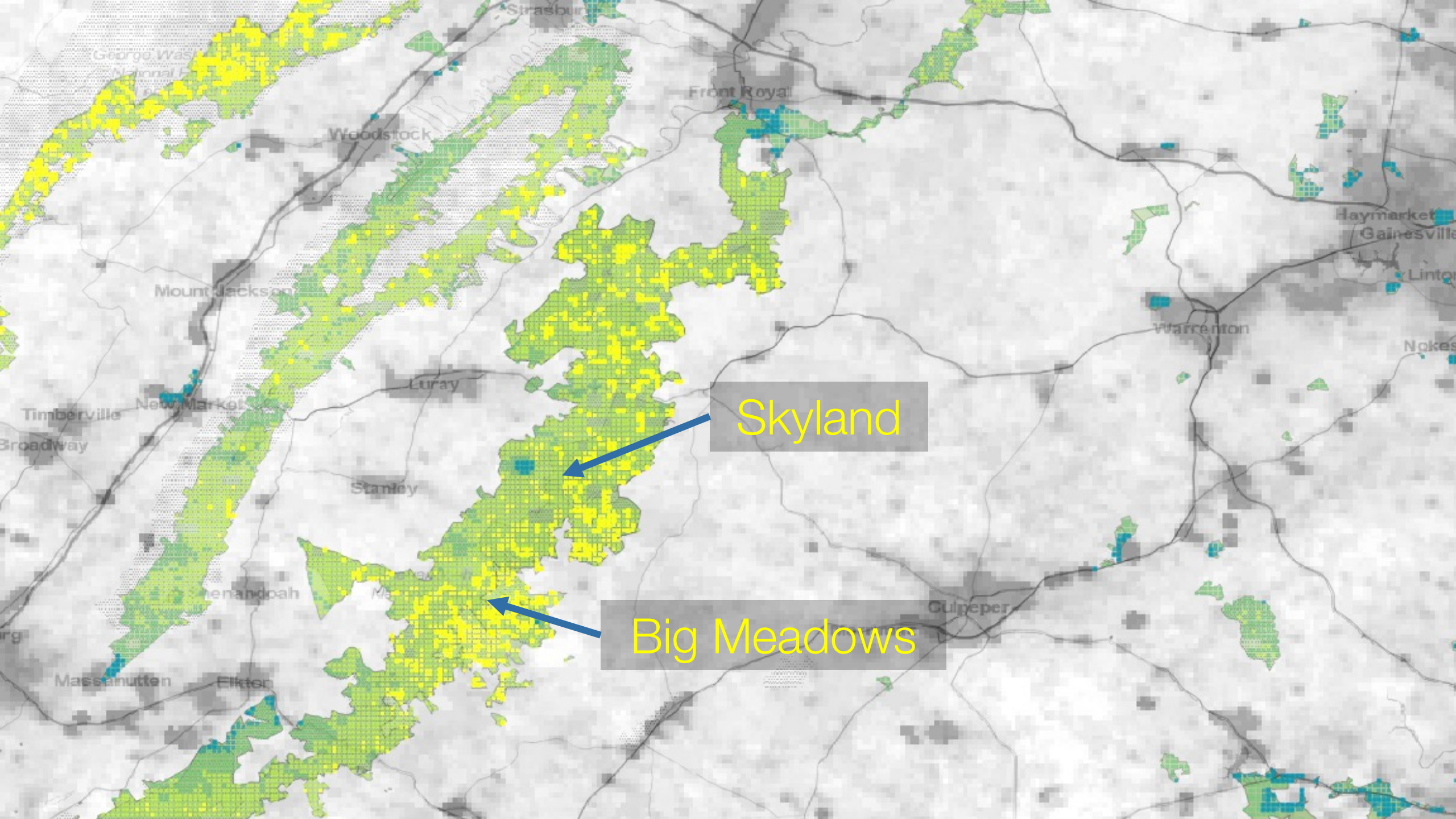

~152,000 pixels
in parklands


Close by

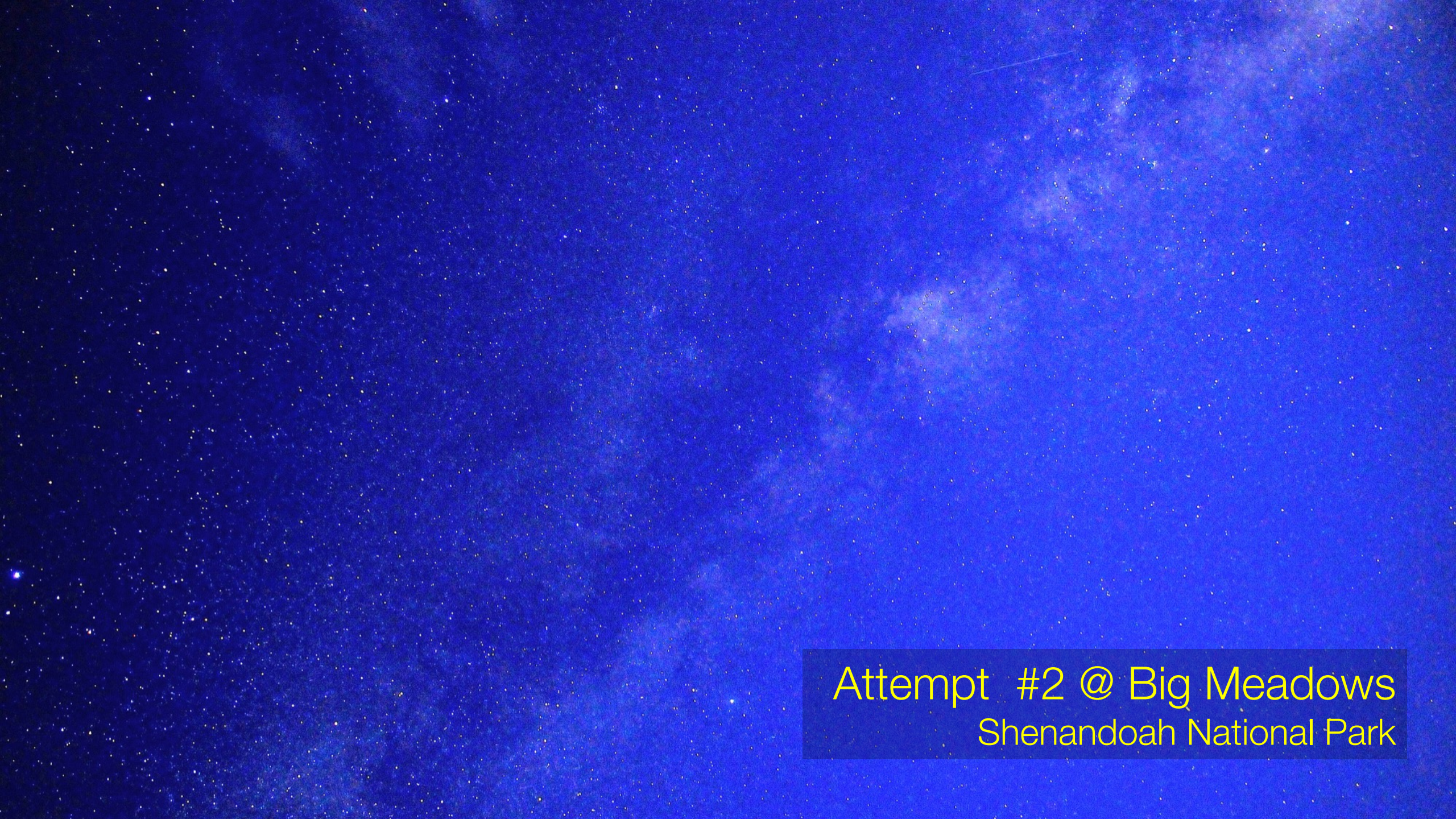

Clear skies


Dark Enough

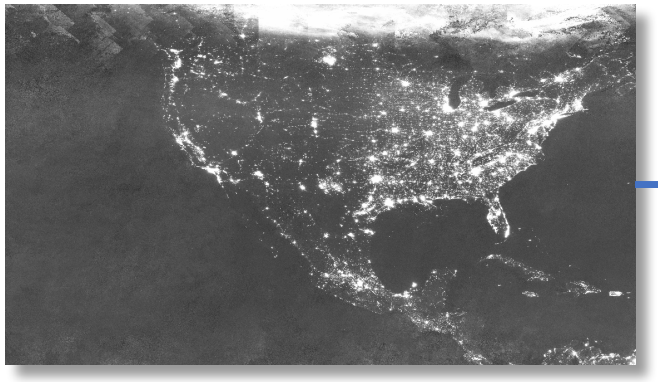

Shenandoah
Nat'l Park
Pr(MFO)


Skyland
Big Meadows
Front Royal
Woodstock
Mount Jackson
Luray
Stanley
Timberville
Broadway
Warrenton
Culpeper
Haymarket
Gainesville
Elkton


Attempt #1 @ Skyland
Shenandoah National Park


Attempt #2 @ Big Meadows
Shenandoah National Park

One dataset coupled with data science can enable dozens of new field-deployed capabilities beyond its original intended purpose.



VIIRS DNB Imagery

RECREATION. Function for finding dark sky locations for stargazing.

SOCIOECONOMIC. Estimating more timely population and poverty estimates for developing parts of the world (Chen & Norhaus, 2015), (Jean, Luo & Kim, 2016)

FISHERIES. Monitoring illegal nighttime fishing in Indonesia (Elvidge et al., 2015)

ENERGY. Monitoring natural gas flares from petrochemical extraction (Elvidge et al., 2016)

FIRES. Identifying wildfires as they start and spread (NASA, 2015)

STORMS. Spotting blackouts due to storm impacts (NASA, 2012)

What skills were required to make the Milky Way example a reality?

- Ask a data-enabled question
- Convert the question into quantifiable parameters (e.g. universe, variables)
- Obtain any form of data and understand what it is
- Store the data in a re-usable efficient form
- Transform any type of data from one form to another
- Relate multiple features to one another
- Optimize or prioritize values based on some function
- Provide human interpretable results that have a narrative that is approachable

What is data science?!

Many have tried to define it, but there isn't perfect consensus. But generally it involves the following:

- measurement + experimental design
- [computer] programming
- mathematics + statistics
- communications

Measurement + Experimental Design

Concepts and considerations that feed into this dimension of data science

- Data collection
- Estimation strategy
- Operations and process design
- Quantitative evaluation design
- Algorithm performance
- Train/Validate/Test, K-folds cross validation

Computer programming

Concepts and considerations that feed into this dimension of data science

- Languages: Python, R, Julia, SQL
- Control structures
- Feature engineering
- Algorithm design
- APIs and web services
- Database architecture
- High performance computation
- Extract-Transform-Load

Mathematics + Statistics

Concepts and considerations that feed into this dimension of data science

- Machine learning: supervised and unsupervised
- Error functions and objective functions
- Information theory
- Frequentist and Bayesian Statistical Theory
- Probability density functions
- Matrix and linear algebra
- Vector calculus
- Combinatorics

Communications

Concepts and considerations that feed into this dimension of data science

- Visualization
- Interactives
- Data storytelling
- Consensus building
- Productizing data

Roadmap

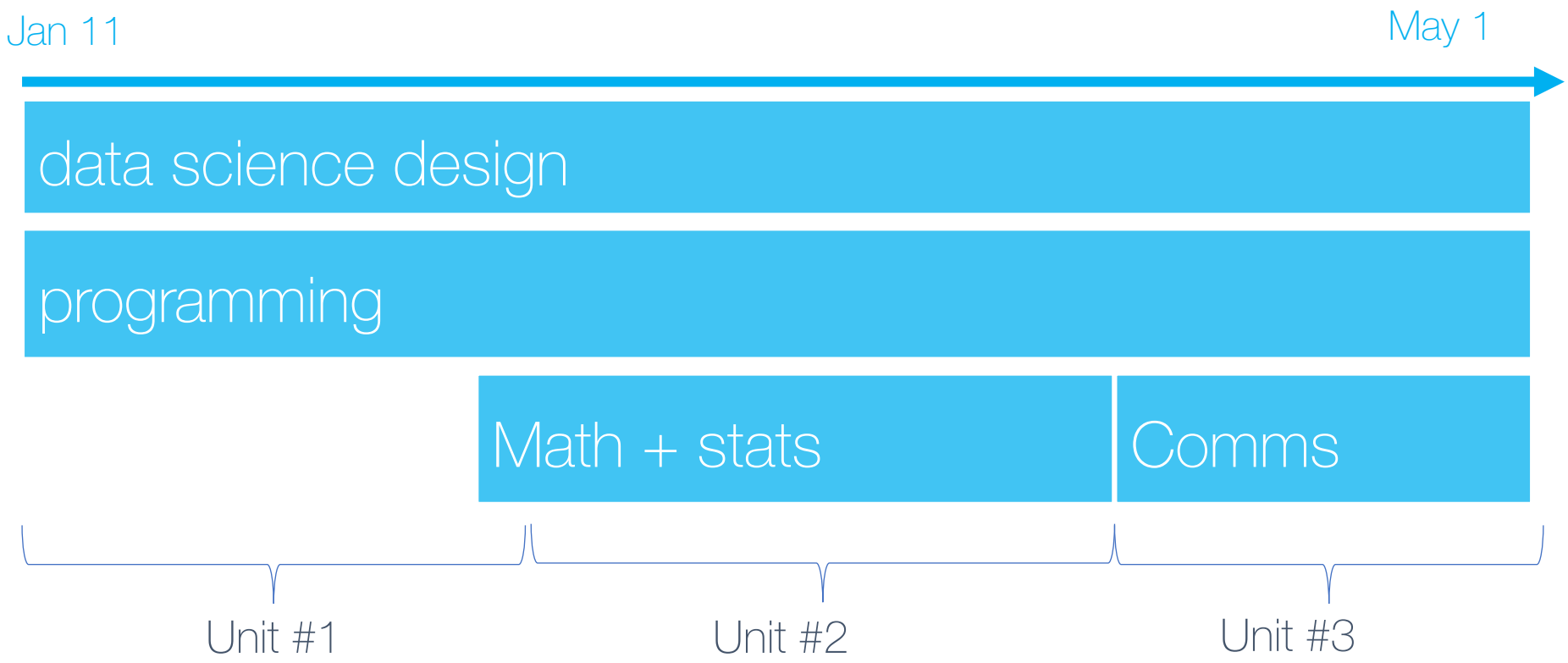
- What is data science?
- **Course Roadmap and Structure**
- Hands-On Time
  - Getting familiar with R

Course goals [tangibles]

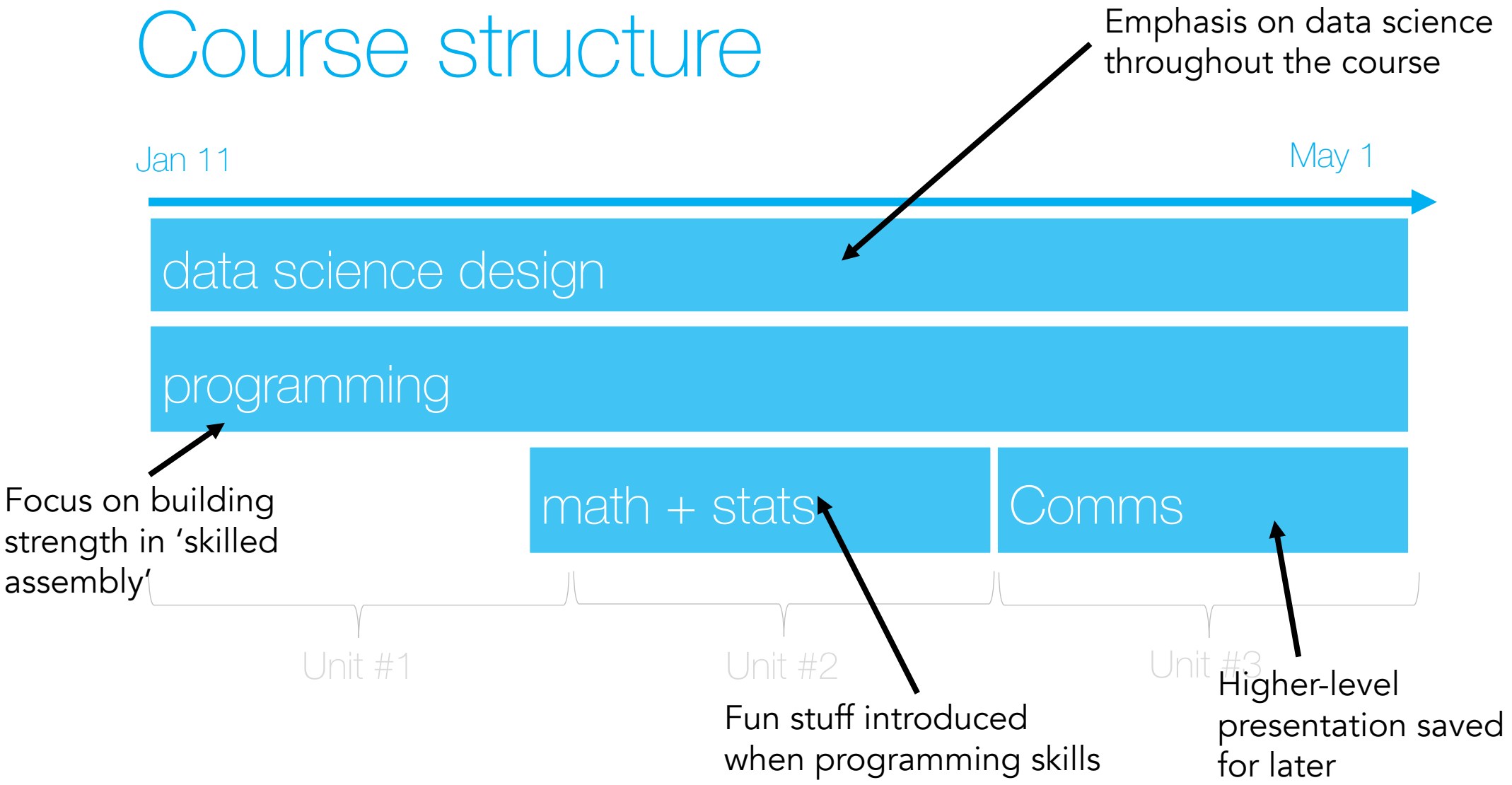
Great data-driven policy only is great when it is implemented.
Otherwise, it's still just an idea. This is a course about planning and executing, focusing on employable technical skills.

- Be able to write a predictive function from scratch
- Design and implement a small data science project, including sample design, ETL, and visualization
- Diagnose how a data science project can fit into a policy process
- Walk away with a project that is non-standard

Course structure



Course structure



Play-by-Play: Programming Basics

By Lecture

1. Intro to R
2. Data Manipulation
3. Functions and Control Structures
4. Exploratory Data Analysis

Play-by-Play: Data Analysis + Modeling

By Lecture

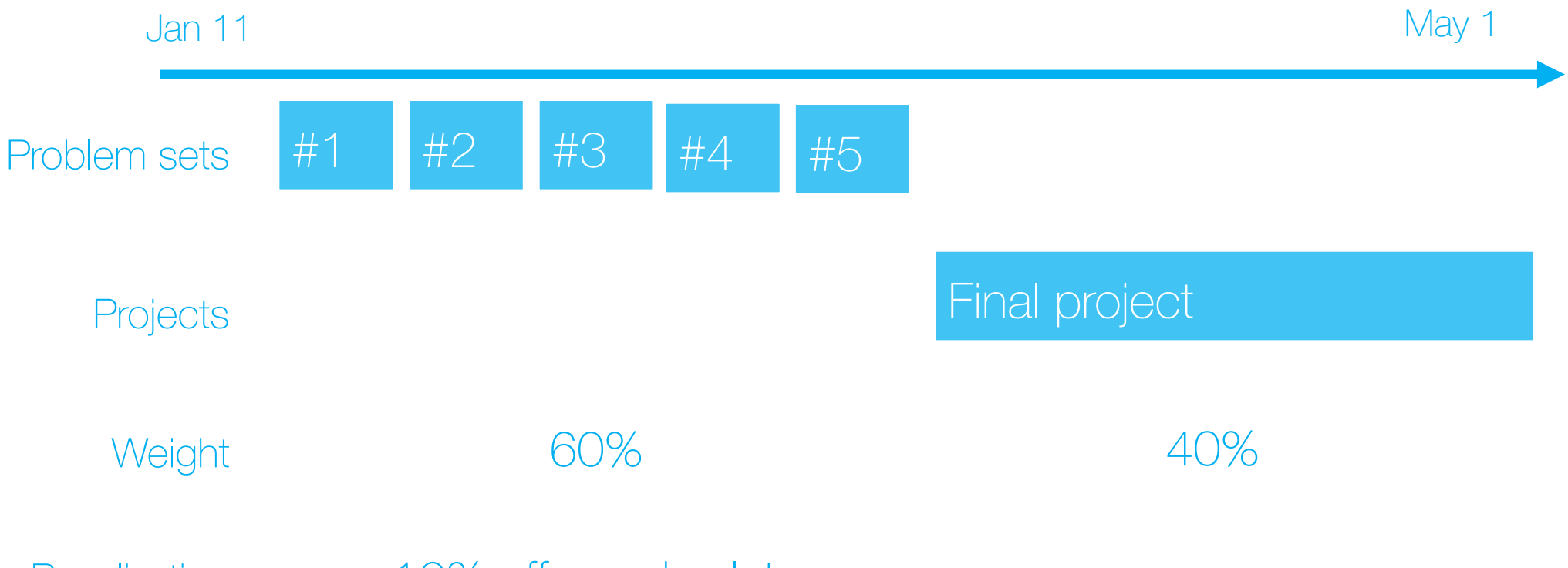
5. Supervised Learning + Regression
6. Simulations + Selection Bias
7. Regression Discontinuity + Diff-in-Diff
8. Classification Techniques: KNN, GLM, and Decision Trees
9. Unsupervised Learning: K-Means and PCA

Play-by-Play: Data Enhancement + Viz

By Lecture

10. Data Storytelling
11. Web Service APIs + Spatial Data
12. Spatial Data + Maps
13. Export Results + Interactivity (Intro to Shiny)

Assignments



Typical class structure

Part 0:

Review homework [if there was one]

Part 1:

The big picture, the technical theory, the use cases

Part 2:

Code-along – solidify

Roadmap

- What is data science?
- Course Roadmap and Structure
- **Hands-On Time**
  - Getting familiar with R

Coverage

- Rstudio console
- Rmarkdown
- Basic functions in R

To Do For Next Class

- Github.com/georgetownmccourt/data-science
- Review Lecture 01 Notes (here)
- Read Lecture 02 tutorial (here)

Last word of wisdom: Things I look for in a data scientist

Ability to code

- Can be demonstrated by having an original set of code projects on Github that involves original algorithms

Ability to communicate

- Can articulate the business case (visuals, documentation)

Ability to scope

- Can figure out the point of the problem (narratives)