BIOINFORMATICS AND DATA SCIENCE PART II

WEEK 1

WHO AM I?

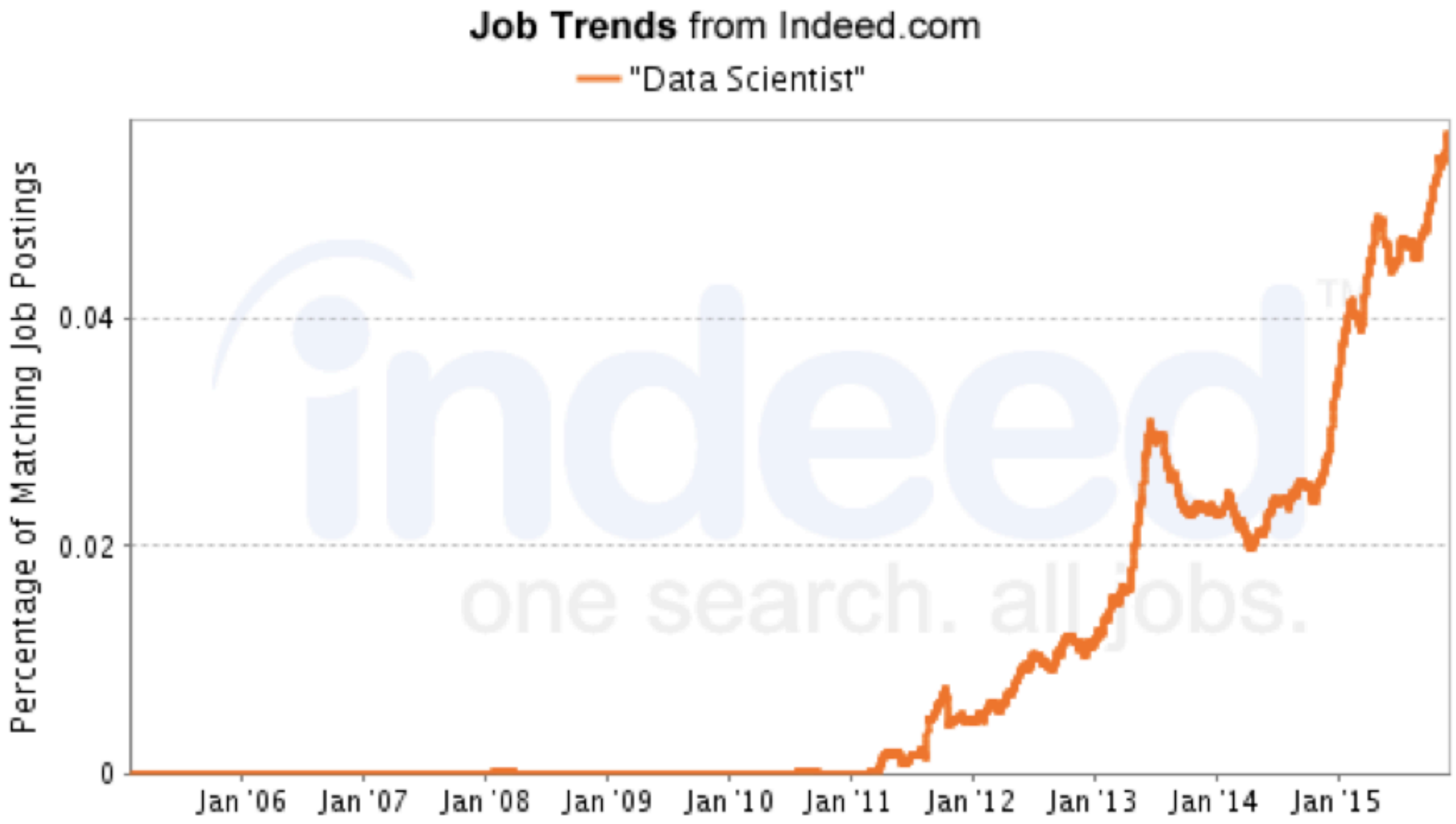
WHO AM I?

- I have been working in bioinformatics for 10 years
- 3 postdocs in bioinformatics
- Develop software and pipelines for handling large biological data sets - genomics, machine learning..
  - www.github.com/juliema
- This course is an accumulation of many years working in the field.

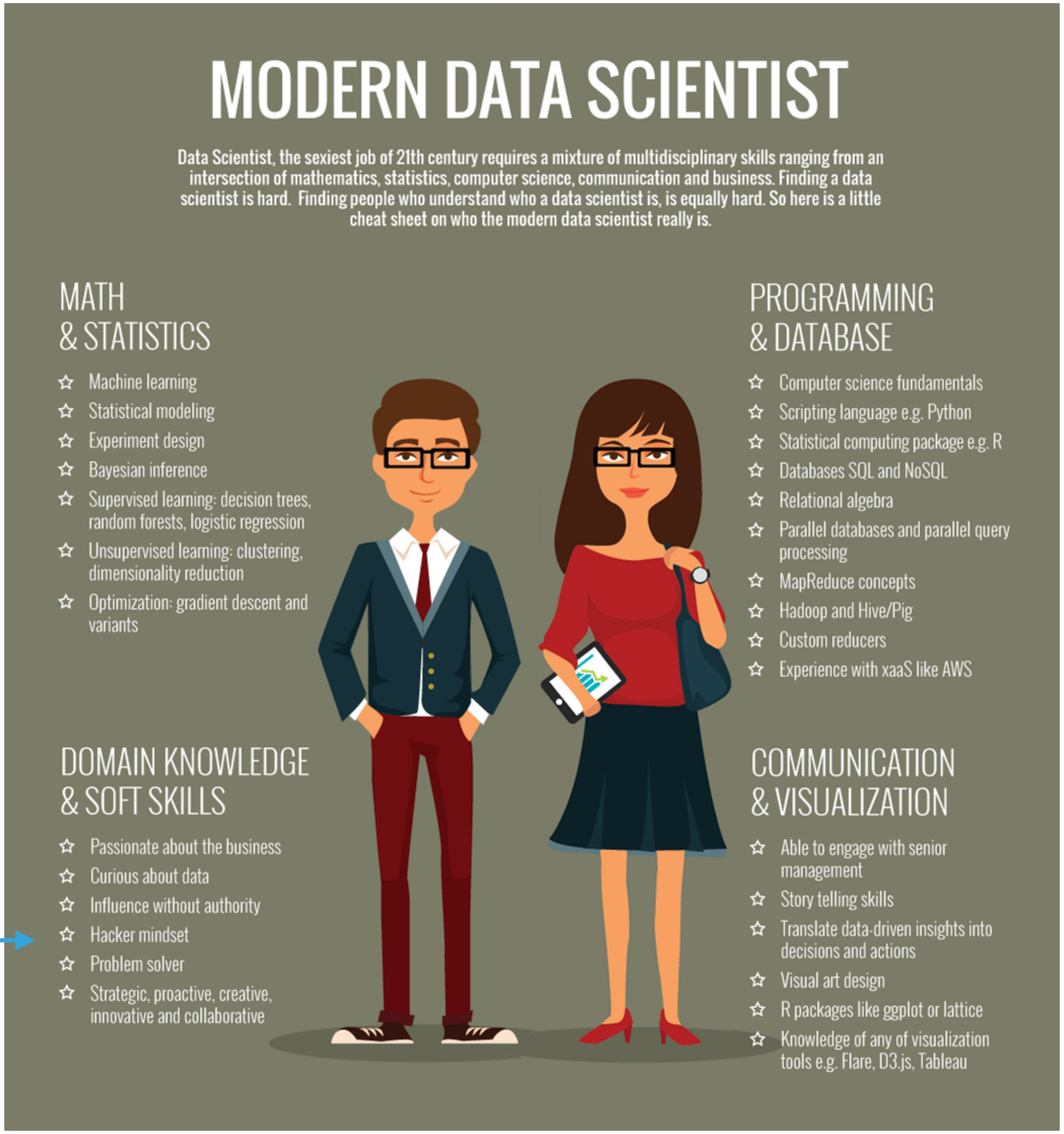
WHAT IS DATA SCIENCE? WHAT IS BIOINFORMATCS?

- Data science: techniques used when trying to extract insights and information from data
  - Collect
  - Store, Transform, Clean
  - Analyze, model
  - Visualize
- Bioinformatics: combines the principles of biology, computer science mathematics and statistics to understand biological data

CAREERS IN DATA SCIENCE - 95K - 165K



CAREERS IN DATA SCIENCE – 95K – 165K



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Biological Background

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand

Marketing DISTILLERY

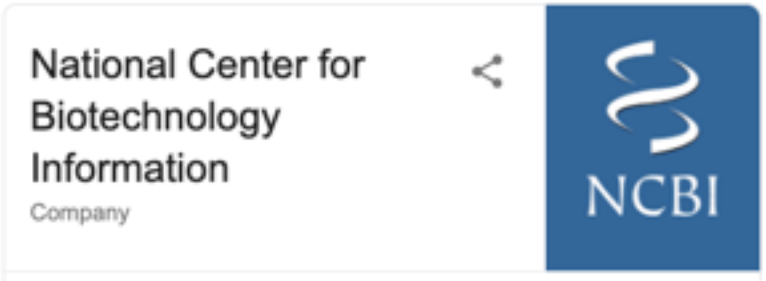
A DATA SCIENTIST USES THE OPTIMUM TOOLS AVAILABLE

- Coding
- Machine learning
- Relational databases
- Hacking
- High performance clusters
- ** pillars of data science expertise
  - Domain expertise - Biology
  - Mathematics - statistics
  - Computer science - software architecture -at minimum scripting skills
  - Communication expertise

IN BIOLOGY

- Machine learning techniques to identify new species
- Scripting to clean, and organize data
- Develop software and code to handle large genomic datasets
- Cloud computing for large - scale analysis
  - Phylogenomics
  - Species distribution models
- Data integration - combining data from different databases to more completely understand how and why species are where they are

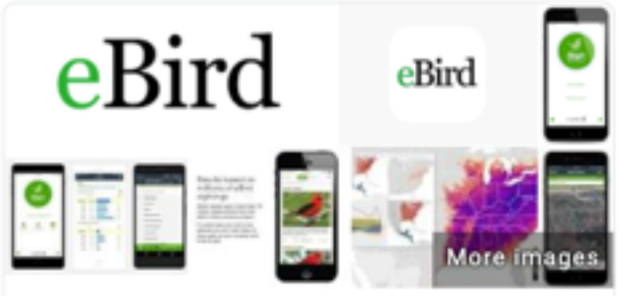
IN BIOLOGY WE ARE IN MORE NEED OF DATA SCIENTISTS THAN EVER BEFORE



National Center for Biotechnology Information

Company

IN BIOLOGY WE ARE IN MORE NEED OF DATA SCIENTISTS THAN EVER BEFORE



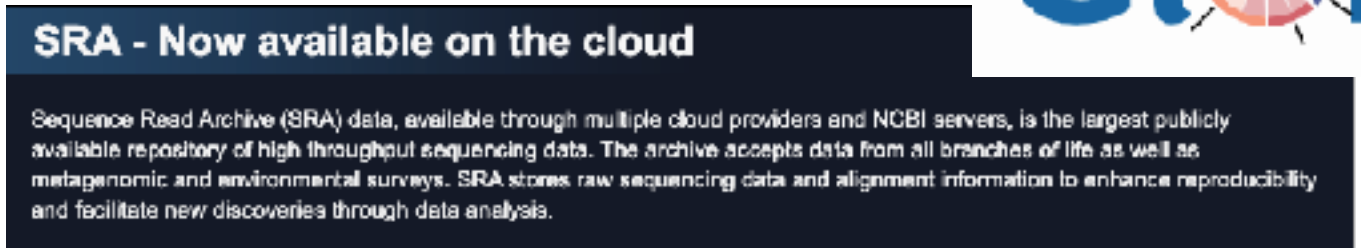
iNaturalist

National Center for Biotechnology Information

Company

NCBI

GloBI



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

iDigBio

iDigBio Integrated Digitized Biocollections

RED LIST

THE IUCN RED LIST OF THREATENED SPECIES™

Global Biodiversity Information Facility

GBIF Global Biodiversity Information Facility

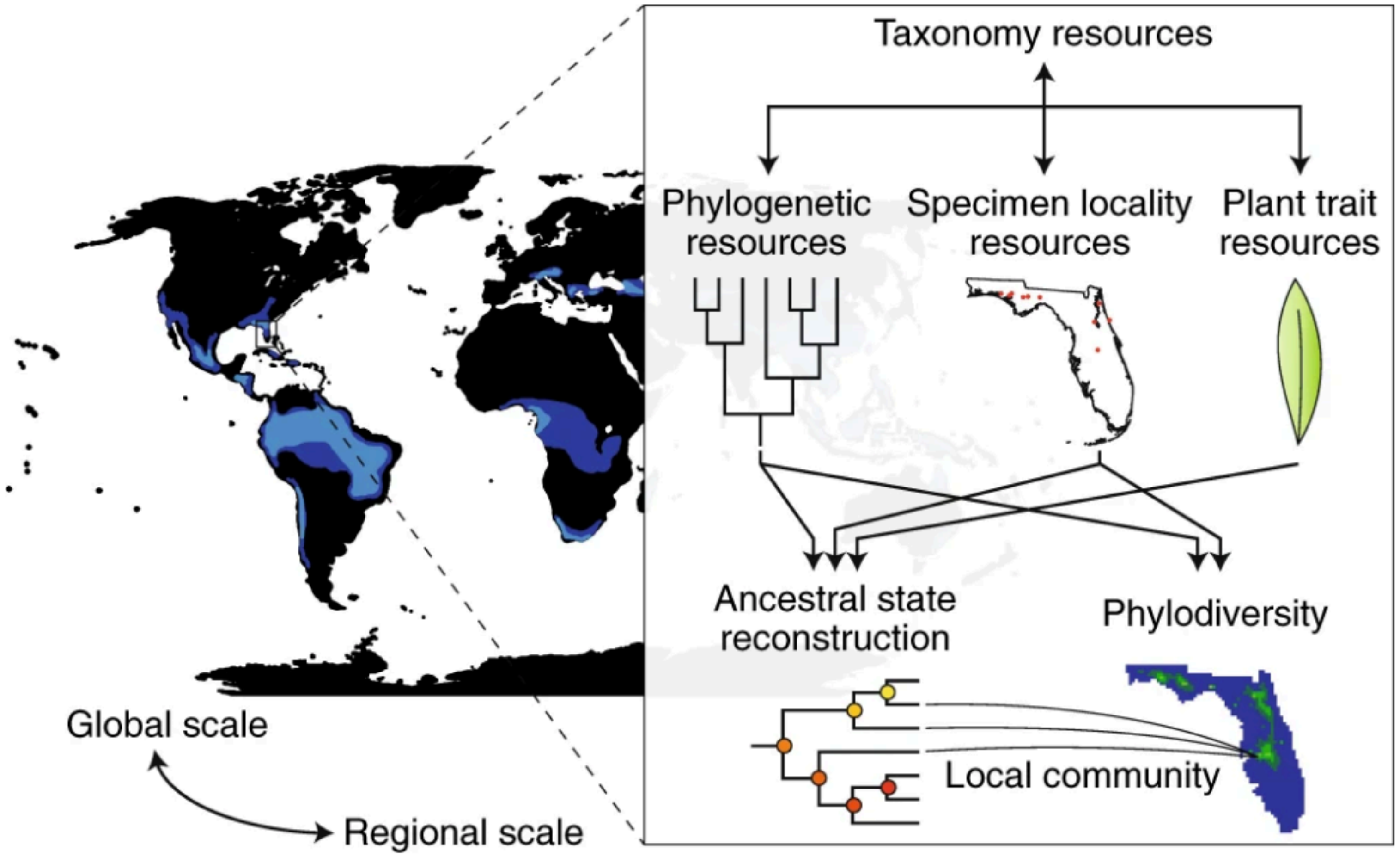
WORLDWIDE PDB PROTEIN DATA BANK

Open Tree of Life

Hundreds of years of research publications with data -s species distributions etc..

WE ARE AT THE POINT WHERE WE CAN START ASKING BIG PICTURE QUESTIONS

From: Biodiversity synthesis across the green branches of the tree of life



Global resources can be spatially, taxonomically or temporally subset to serve focused needs, as shown. The top rows in the inset show needed data resources, and the bottom row shows synthetic products, such as regional assessments of evolutionary diversity that inform about community processes and conservation priorities.

Allen et al., 2019 Nature Plants

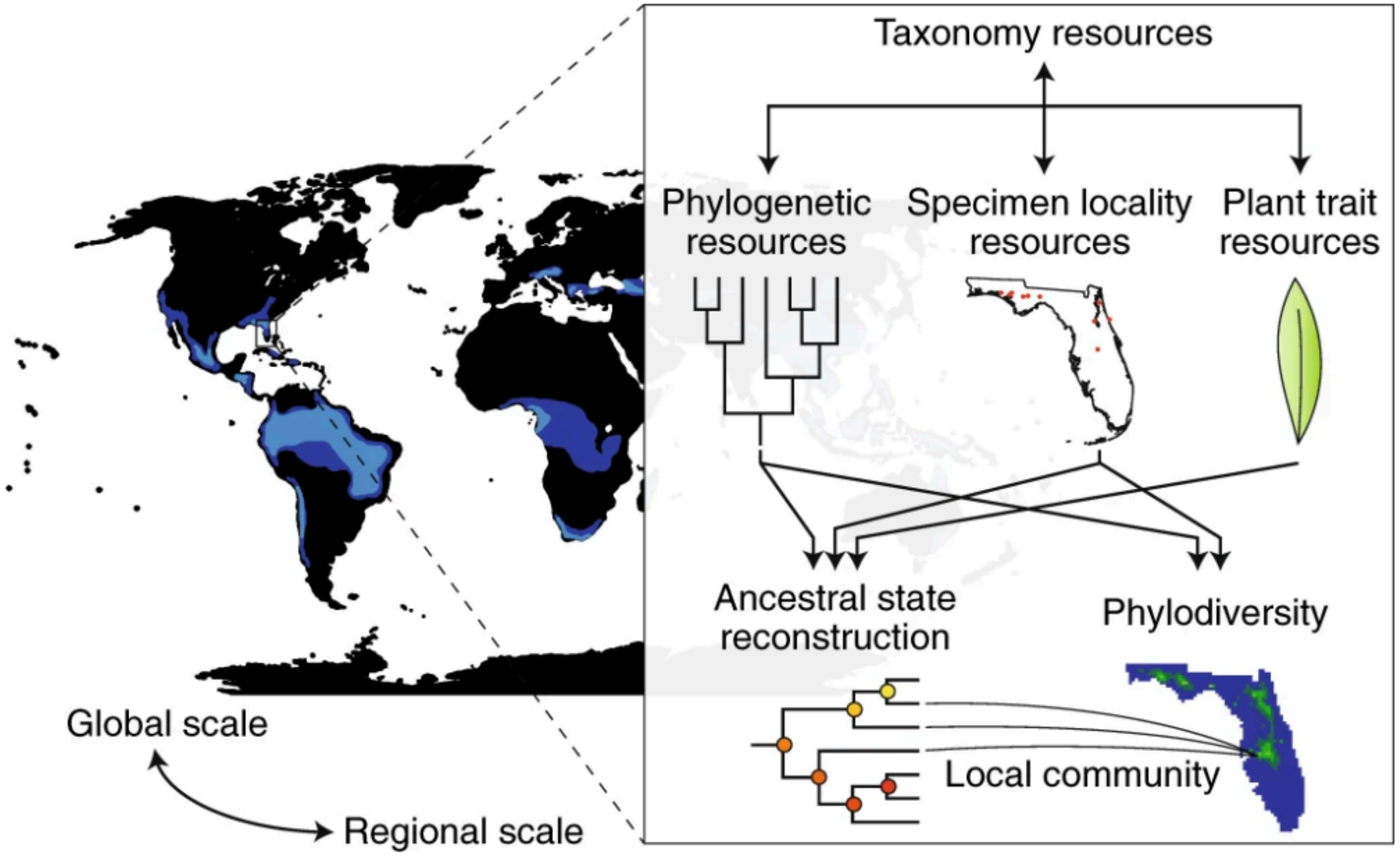
What areas host the largest portions of the tree of life and hence merit species conservation concern?

How have past and present ecological interactions affected species distributions, movements and evolutionary history?

What morphological and physiological traits have enabled colonization into and diversification within novel habitats?

WE ARE AT THE POINT WHERE WE CAN START ASKING BIG PICTURE QUESTIONS

From: Biodiversity synthesis across the green branches of the tree of life



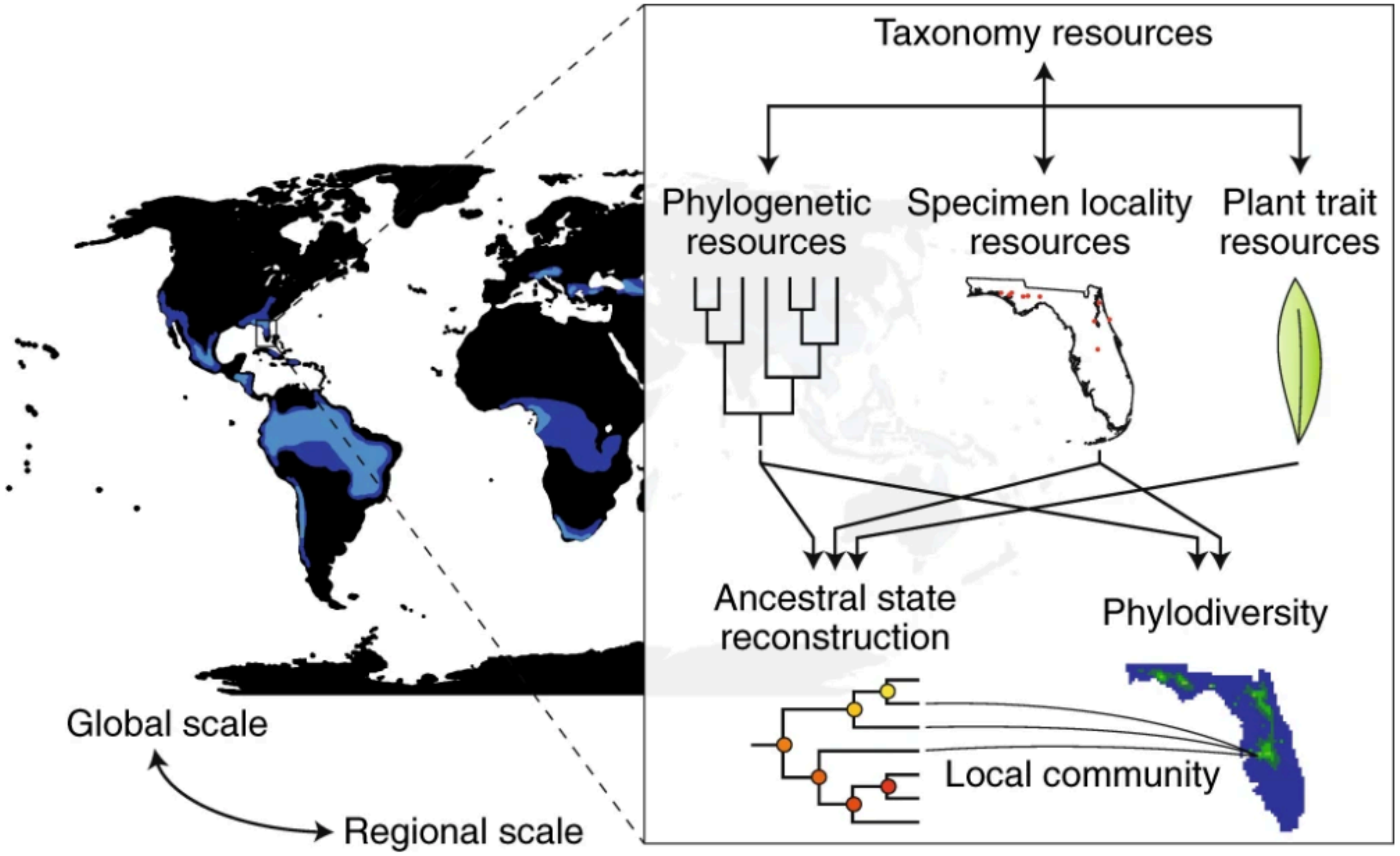
Global resources can be spatially, taxonomically or temporally subset to serve focused needs, as shown. The top rows in the inset show needed data resources, and the bottom row shows synthetic products, such as regional assessments of evolutionary diversity that inform about community processes and conservation priorities.

Allen et al., 2019 Nature Plants

What genetic changes resulted in new traits that enabled colonization and diversification?

WE ARE AT THE POINT WHERE WE CAN START ASKING BIG PICTURE QUESTIONS

From: Biodiversity synthesis across the green branches of the tree of life



Global resources can be spatially, taxonomically or temporally subset to serve focused needs, as shown. The top rows in the inset show needed data resources, and the bottom row shows synthetic products, such as regional assessments of evolutionary diversity that inform about community processes and conservation priorities.

Allen et al., 2019 Nature Plants

What genetic changes resulted in new traits that enabled colonization and diversification?

How can improved understanding of past speciation- extinction dynamics facilitated a better understanding of how biodiversity will respond to a changing environment?

CODE OF CONDUCT

CODE OF CONDUCT

- Recognize people have all different skill levels
- Relearning things forgotten
- Remove and change misconceptions
- Affects Diversity in Bioinformatics
- Part of your grade!

WHAT ARE WE GOING TO DO?

- Give you the skills!
- Go over common skills/tools and we use in bioinformatics
- Integrate data science principles with biological datasets and how we handle them
- Class has 5 modules - building on the skills you learned last semester
  - Part I Unix - version control - GitHub
  - Python - pandas, notebooks
  - Data Visualization
  - Data Cleaning - Relational Databases
  - Clusters

WHO ARE YOU?

- Who are you?
- What are you working on?
- What kind of operating system do you have?

RECIPROCAL INTERVIEW

- What excites you about this class?
- What frightens you about this class?
- What questions do you have?

SYLLABUS/COURSE MATERIAL

- https://github.com/juliema/Data_Science_For_Biology_II
-

LETS GET SET UP

- Mac - utilities/terminal
- Windows - https://gitforwindows.org/
  - More advanced cygwin, secure shell (SSH) clients (Putty)

DOWNLOAD DATA FOR TODAY

- www.github.com/juliema/bioinformatics_and_data_science
- Download data for today:
- http://swcarpentry.github.io/shell-novice/data/data-shell.zip
    - Put it on your desktop and unzip/extract it
    - Open a terminal and type: `cd`
        - hit enter