Recent Advances in General Phylogenetic Inference and Energy-efficient Computing

Alexandros Stamatakis[1,2,3]

1. Institute of Computer Science, Foundation for Research and Technology - Hellas

2. Heidelberg Institute for Theoretical Studies

3. Dept. of Informatics, Karlsruhe Institute of Technology

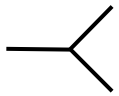
www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Outline

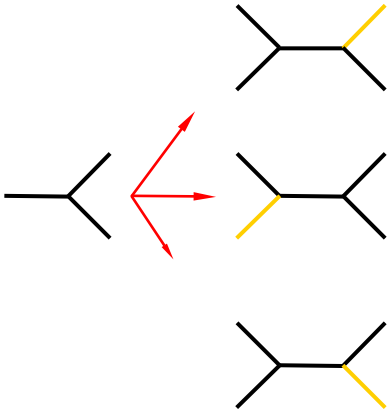
- **Recent Advances in Phylogenetic Inference**
- Energy-efficient Computing

The number of trees



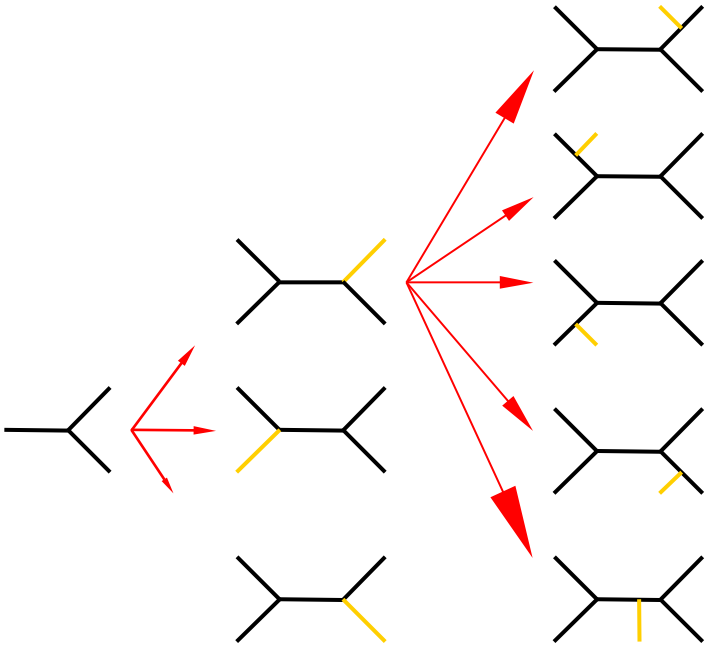
3 taxa → *1 tree*

The number of trees



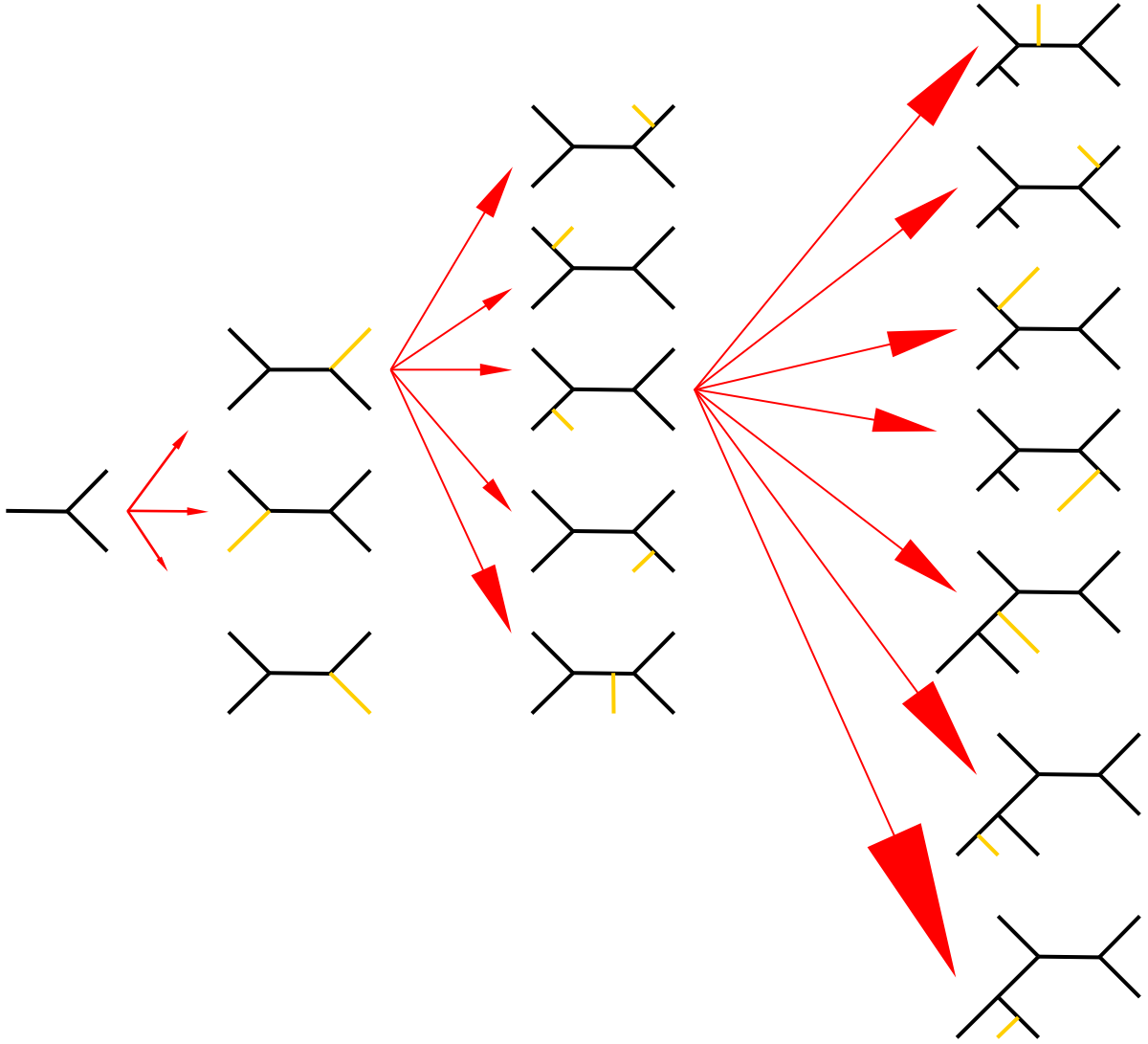
4 taxa → *3 trees*

The number of trees



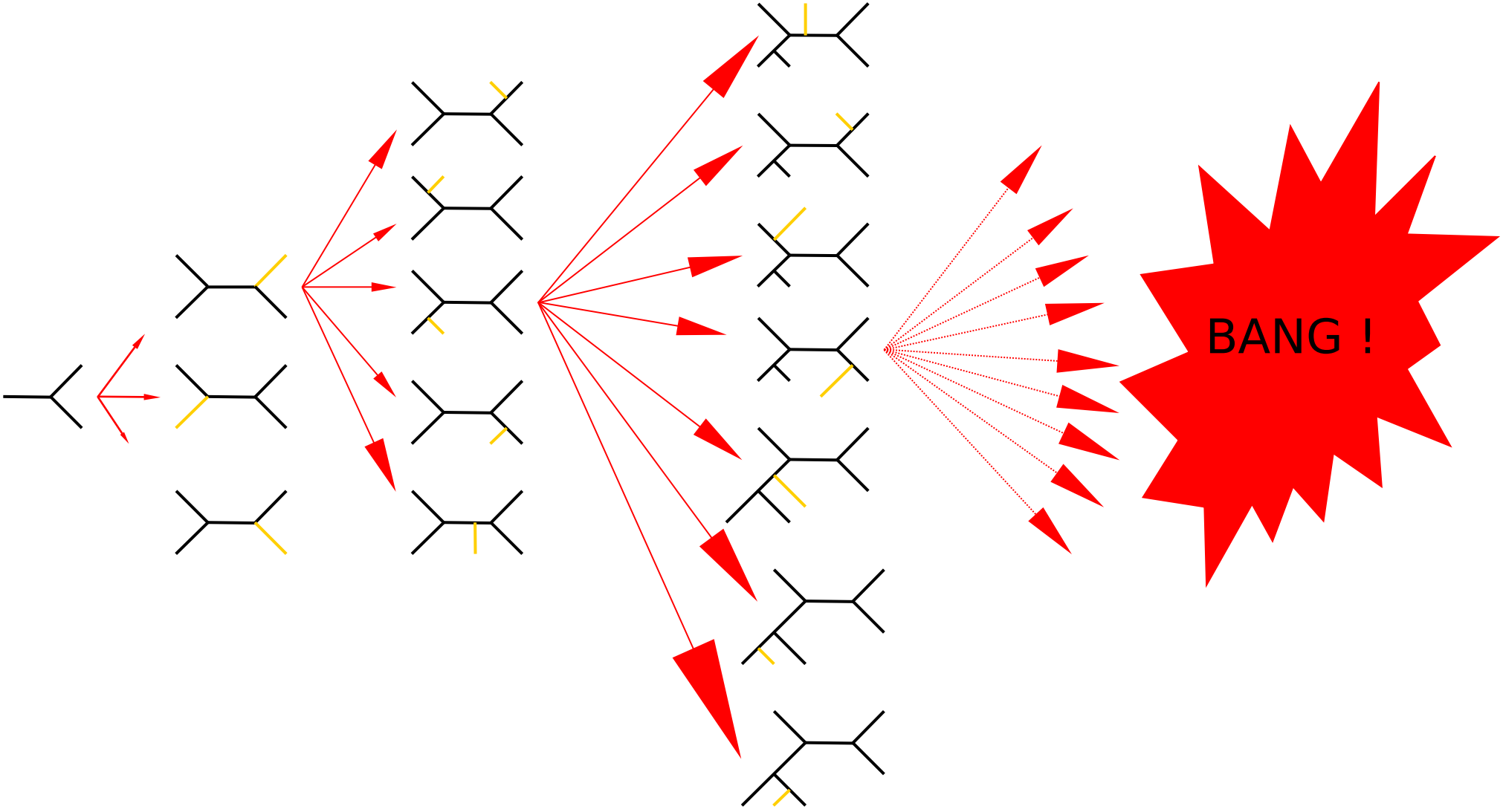
5 taxa → *15 trees*

The number of trees



6 taxa → *105 trees*

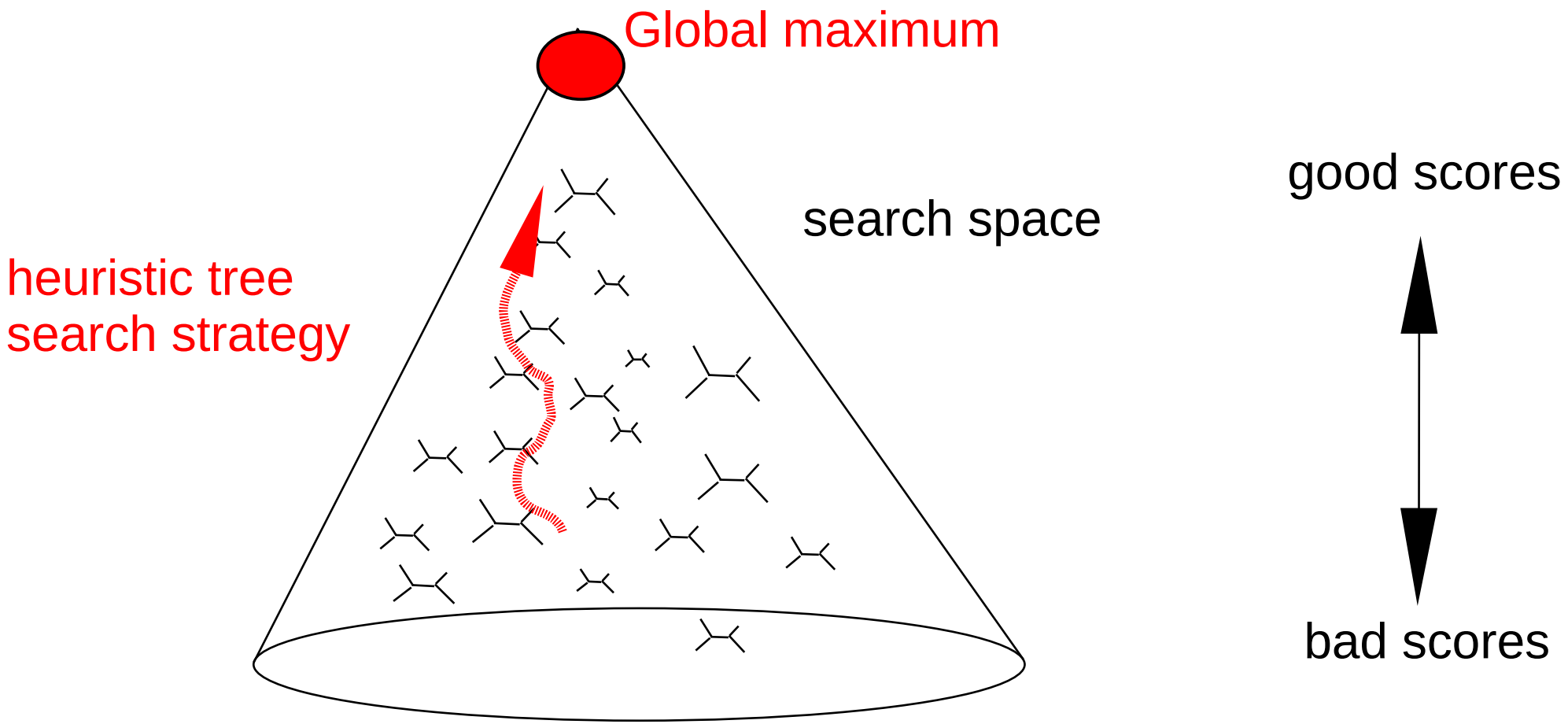
The number of trees explodes!



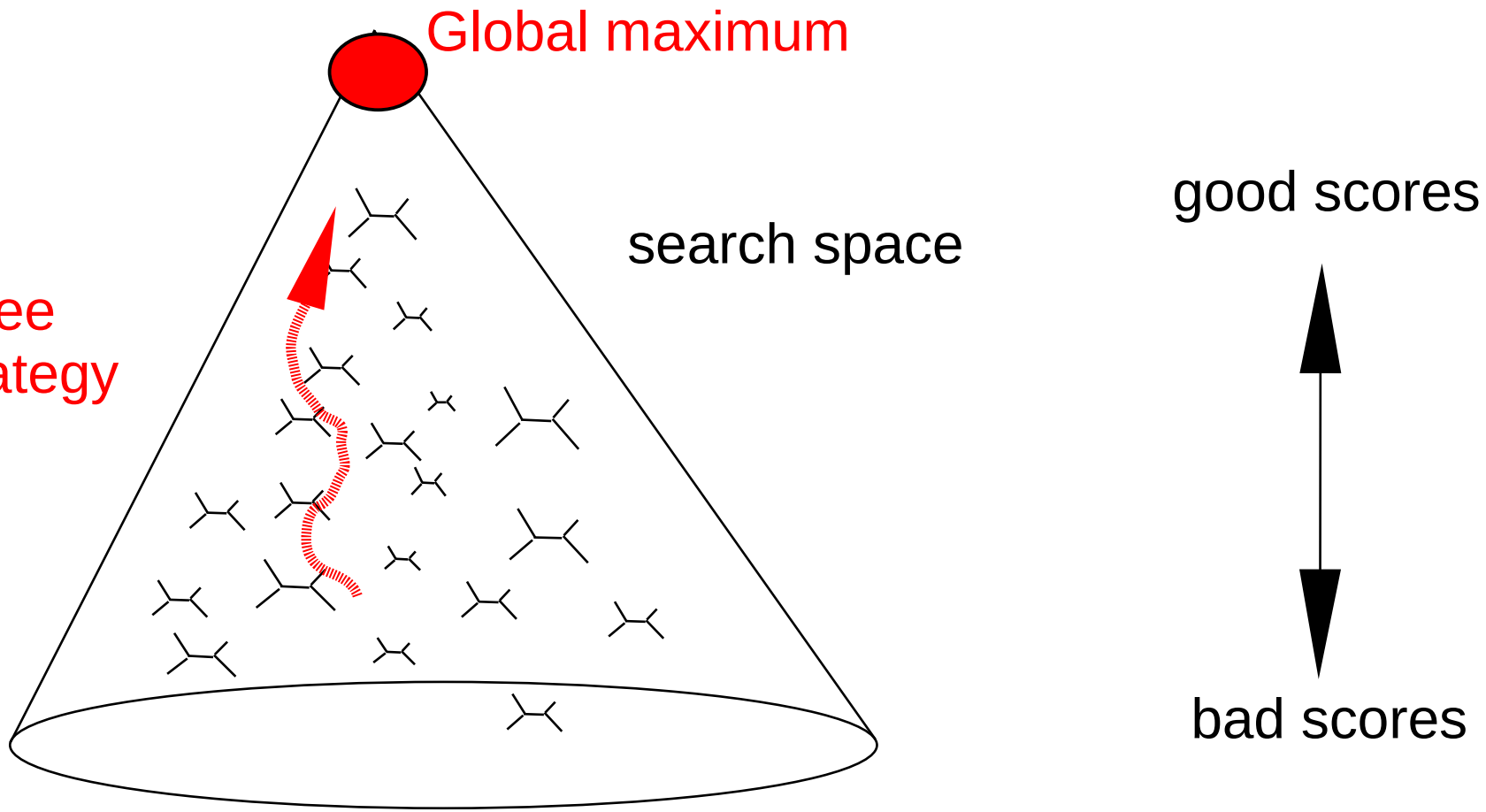
possible trees with 2000 taxa

```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000


GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

Number of unrooted binary trees for 2000 taxa: 3004963817421165615163291006568181498137723207423701308950495404301263652525830821082768599668824700046435273521426563428829589150234460006314939691306329704360561848618774654822779912235368092334555631999108345976931267565250128998674331877528114019609916315223670306091217357097623798477054676677795324797182614385273338226727784250737252849916669687584403510579587020686505817687044663181237429010214385064324713609344916670211359697569403006662526464792691245510314942366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124010661315694886196094119564673620034257524133527757508582916109642257572769976799140828334321016132740165283099380390459232769069003597291970994073934956348620389901074268728229759746553771022576726768428580118772249501062181173405232082653973429622273525365905158656313832720311198419874675997386463182903203832523085979979922161012272157808052481458312068440167606239306009711616729715504728487799634337531348994230372437347879131989085953764070134849446113877572576952408702461720107874297380462275052545706689372319418206440706891884003870590289772197516454495975821662130620506461776109948566373416818358498932907699338206780105243728461492403422961155182609778228619192672071295189589360099591309742330723163825184281103305710174411568843051318658775443763085003114511107238370397074651822320404061547082730786299575493310312752086167006607912980142622300565123522718063819509335872651728623589020520016144361756075654286471422126613004434807084067501589247673166341539540575074474994909831496473031080411401891849735912811228378774049884834056210242056642446386009389965085742961947269054301528123752651096581528469979703679217112903556809818079169587951614159281049528179855847292534447864424435998085315372047968149694659917686145337010519859285771574824559433772423695825762426630169463204824951822559392874031776234338810486046309751915569238711675130952134150988167154643078623526062377864068386804246902527491139319276802611515990582603886733172930713673903403618637463980605764836474670274466727880885337074254421922726677747003329403201038288035112689026255183096791948358678929370163768175304820633894387149793115235369822962511116307148294599211620803302684762013335690441089668145436150905155877581679770012563912151116237444170497371704604029481104114822286466131918821997571383368352072526055202769823974613218495249264897050790398360256255606289852288839561357874156576648899926087328661263064254326024897922911356007164057398451637524524337694375585738472554556439759960425591464011222114475523557317623997305774718395653121741653229598667590129411612392407220932503696731248844915537592106506560154167207741592362408686676753482865129648887390597075788024733934634708481590116397727977474804173162687009167287356121642268468160683198959801260376485615312781611689587215123123308760063473381097253118423339640390937378395066835578735307886358646400563299499490631187424029092779272693300322445377595797224873456891511458557078385054168166766742581130195806362190750079029503108820909727174813643698947397107993277770067630173061756653873972603777717300844134394051236690555449324861650825399577950363267049478442934988531727973481777971465671751511788763964340693324580763461107342143281950499096808740273976889147045174720555436969366874260147724189469321290245331733418828677319465354411330210086657508171324034264758048921862366346160795594372051639515409694980648624230979694721116966596158004178835992232266262849894235227692639112436076750899767178968345933789074234635455719353066561537990027791626563636186197404859093823406223545976973313721365934371758559066443964613283001136726019340687064423394891992153043852815416596301185494236348635245857466428346090629162795649265847230036085555989897619162932481400945924898994684688623225860170551468905649837276003927487069550463788819741699429049102853180410776516007263384721638903782270013843599599730265727198804324664312359758555217196960513921010226596324788783097740533313155176297815288071865260317632726482809449937045625809993805305849769957008028937980801490290100529389847227994716780482168942415911828425769646478650573153251783023363072982516922103465842658944474649161238546897185079681729091390321828341111848213847677283165486532123173820041319905105189670222018870495856871805095907303606934029372160389689176055876769553823180937058262570838983874090984686566342713975000132918351059433217298798252437075082720879598594371576676601557826996603431977526233088989962587800628009560944416932377944955441033696586261556256010669390303203878970983673786087056641433585106111658314520424513208508589994932364831689671194951671619567622707090697388958885557956246664153656172354930180739400476052980172177139168678800027785196617307006128451730758250373564310206511244373082522962504045316059074134388187256347791383066059093188025223100853401768402614015396169891920751471080337577088497401418345997539720598786820648791160649698581776011539720584982226989071813494326918018211733188063653910893689811714891357456680542807485170175858266639633570189354449832669762835092657922201746372190273119641751489944010079636876017826747107019945473218887832742608896672437157471342060009370425130989363053745978427998040313298941726649229042573095836853441621564055729028206622400386323752638091023326989783886042375962560156797526269507986398681042948323331602672165551781208992646778049357413263871374084238855465383361586434513054396242813972795597259951107063143059926154956229583202327080576811566906895866105220300573725298472118747827136713666058669271094875563974858494759108197270338782844398644867434562009581619303147273459619004993184243379752436624893633212448505971992523668529249305346252764137853413208943128901523738092556045987090912766662329678703328882059134949580074074473143388800724532321747309659741967114441453127132790205101004767101435063885795347844725538980154192331702751989618063515268254317319383292589193153016413054897231112866646549297193047929643282955671909288169209104233412200745424204990087258504620805110487588305949599031118873666850941488217257345763552339640384813182131674083590069164000532622581847837650678044511777173286581898992153583094477653505341796875

Approximately 3.00 times 10^6328
```

Problem Complexity

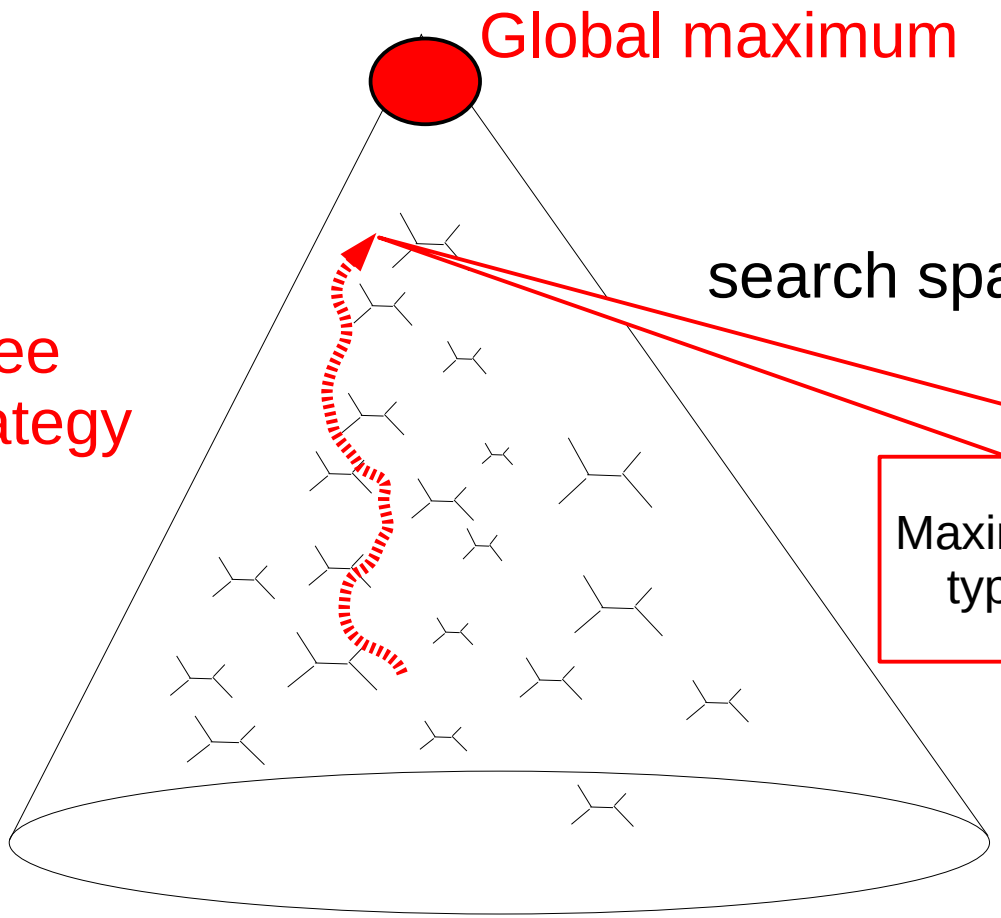


Problem Complexity

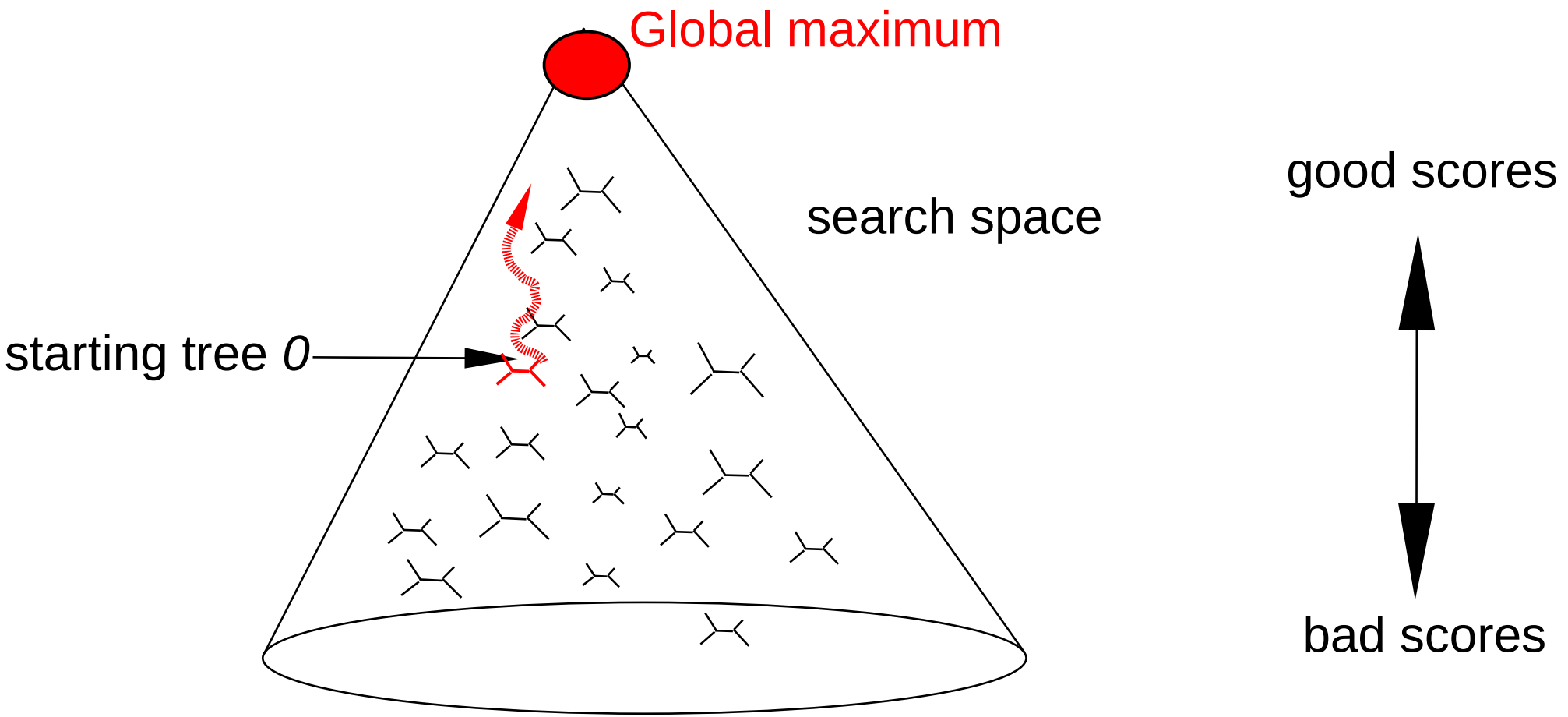


Finding the best tree under Maximum Likelihood is **NP-hard**!

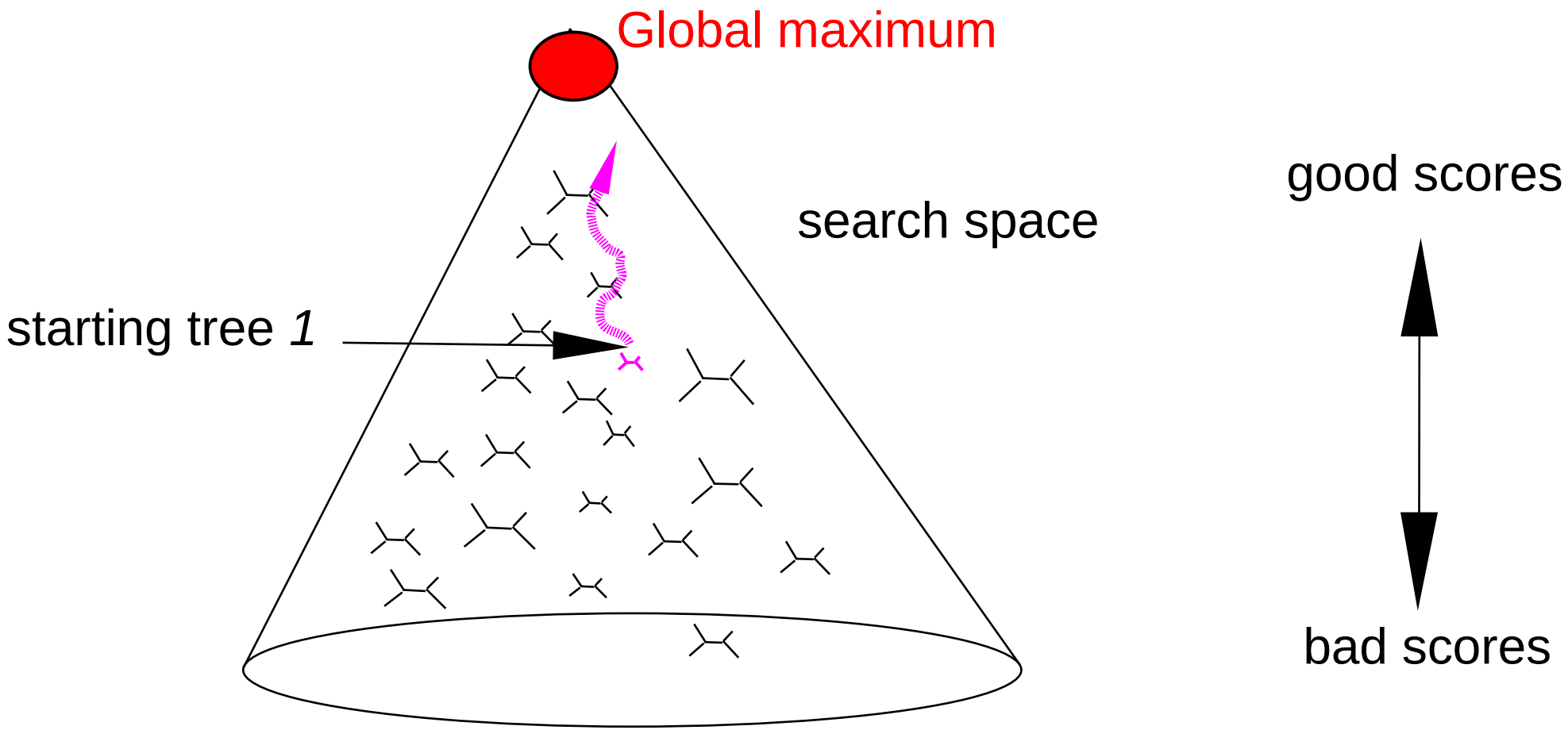
Problem Complexity



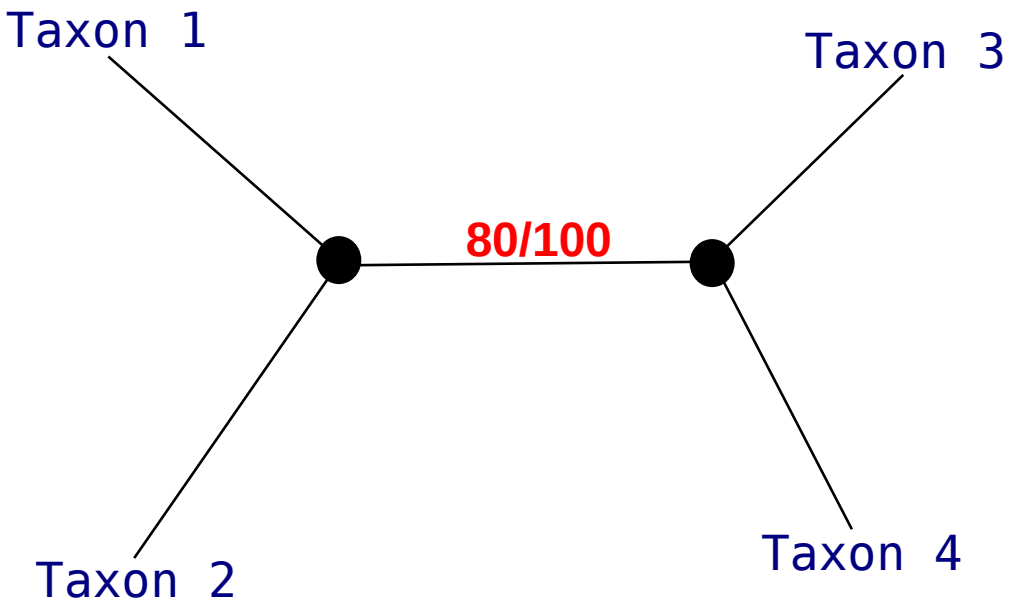
Starting Trees



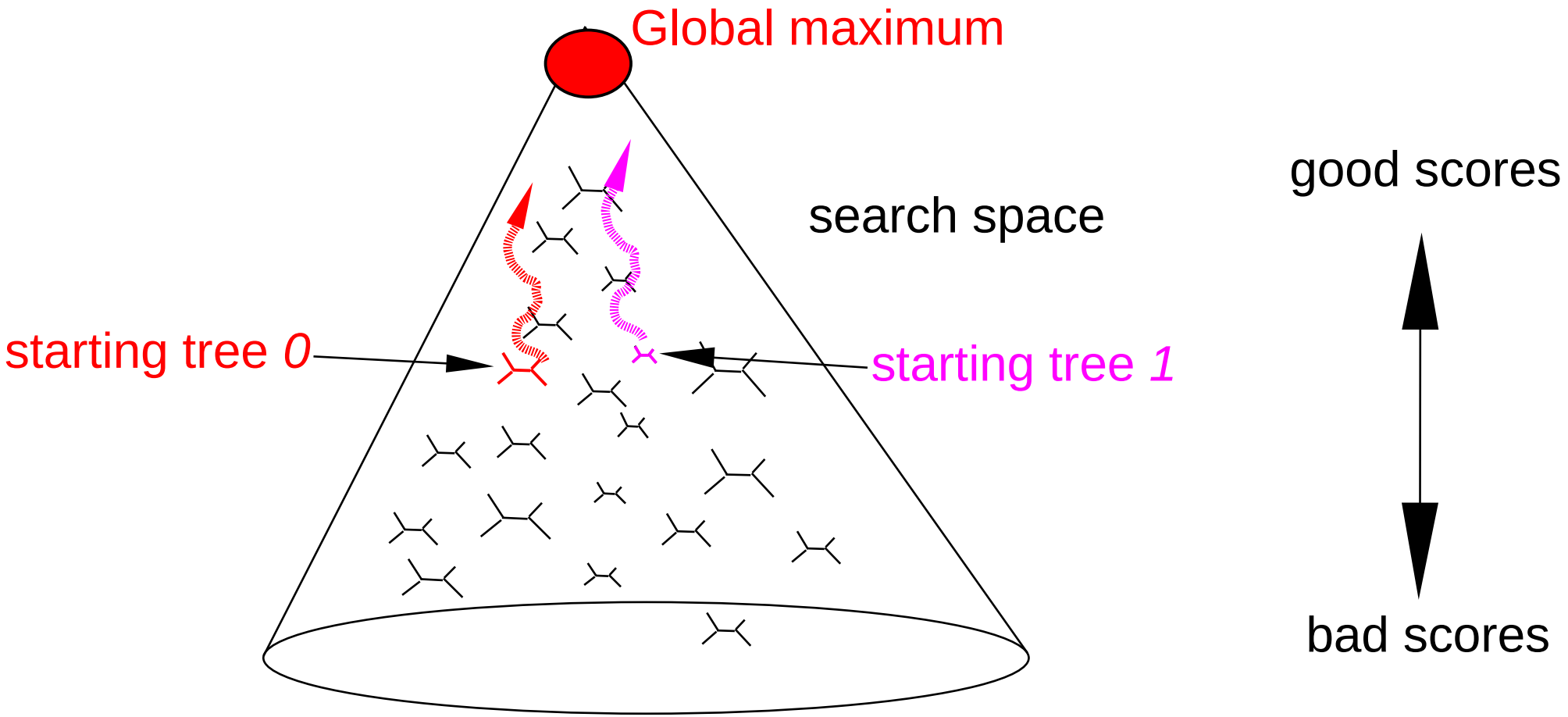
Starting Trees



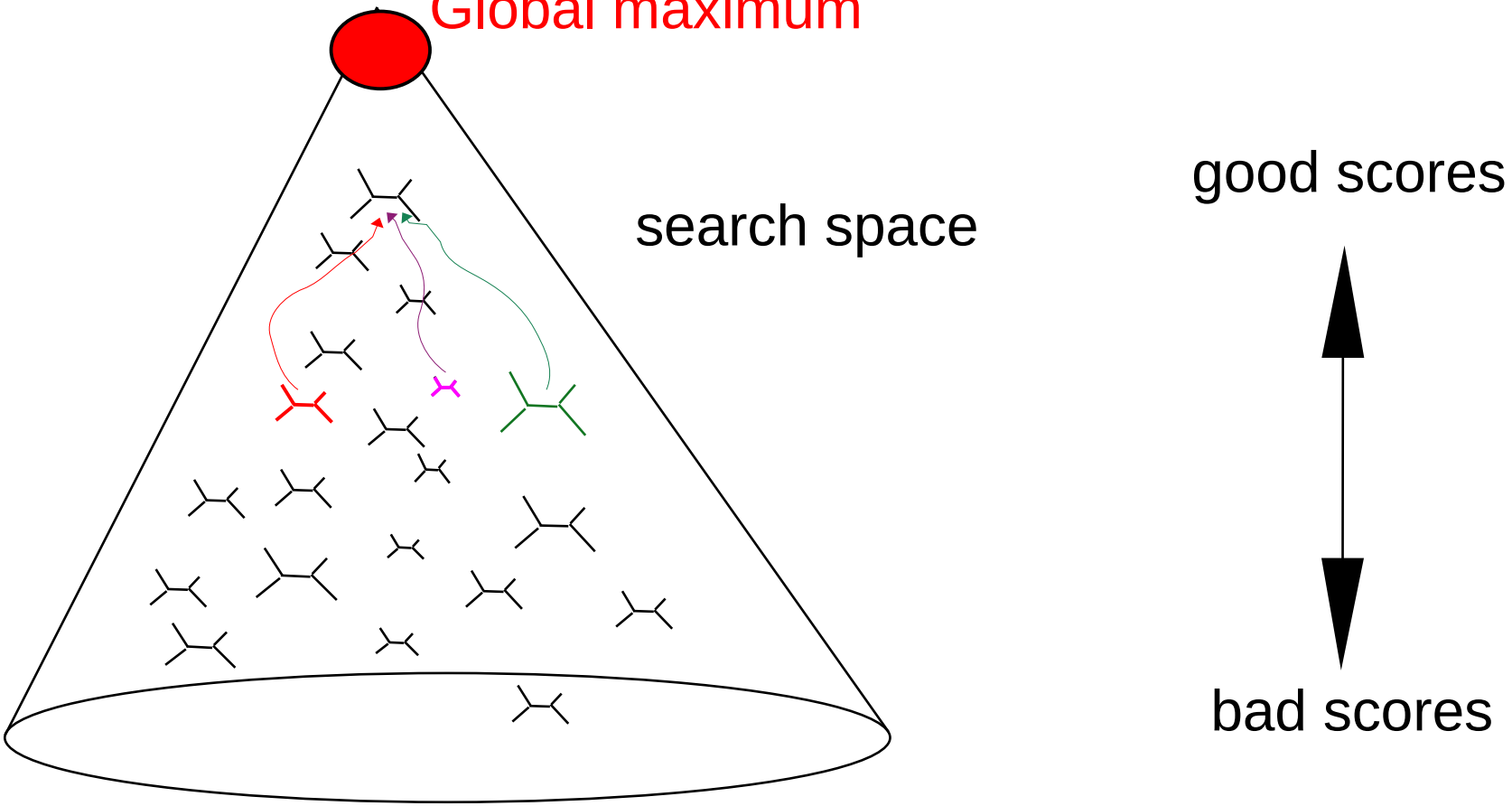
A Tree with Support Values



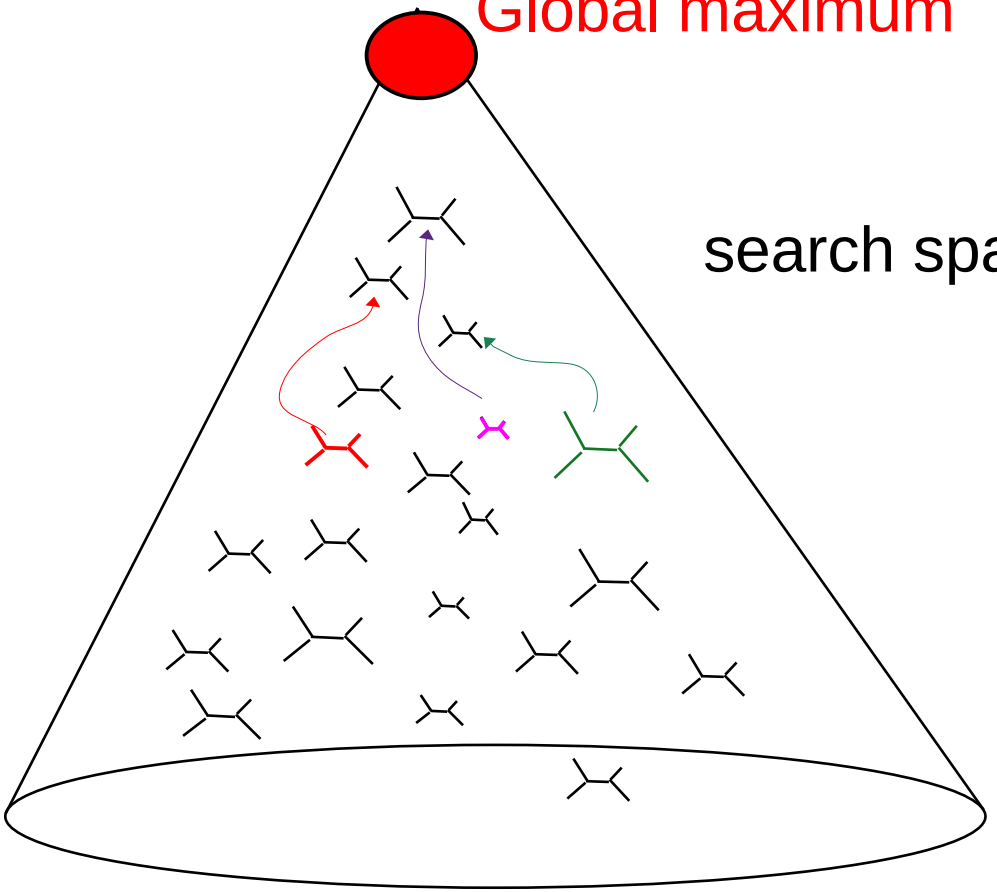
Can we predict how difficult a phylogenetic analysis will be?



Easy Dataset



Difficult Dataset



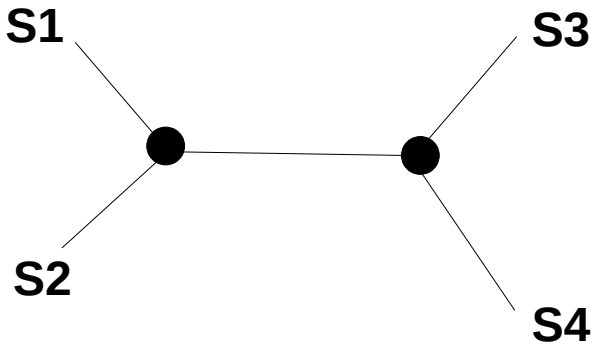
good scores

bad scores

Phylogenetic Inference

DNA input data

S1	ACGTT
S2	ACCGG
S3	TGGAG
S4	GGCTT

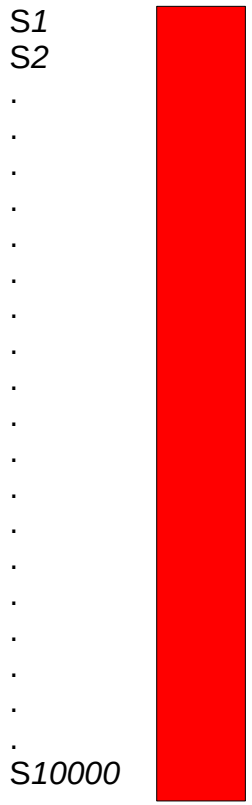


The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

Dataset Shapes

This?

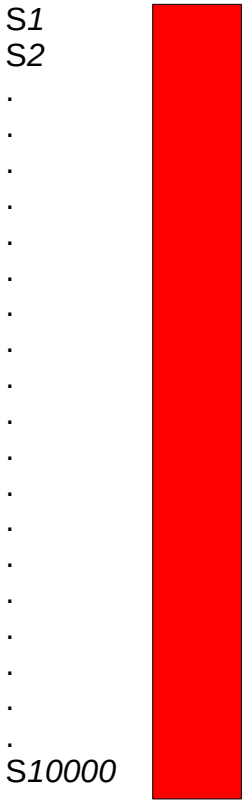
Which data is more difficult to analyze?



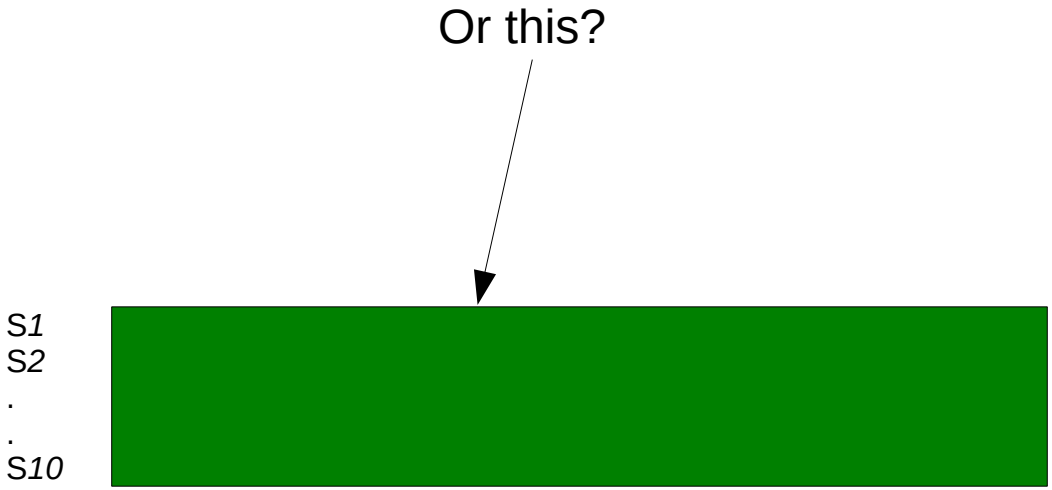
Thousands of sequences, short sequence length

Dataset Shapes

Which data is more difficult to analyze?

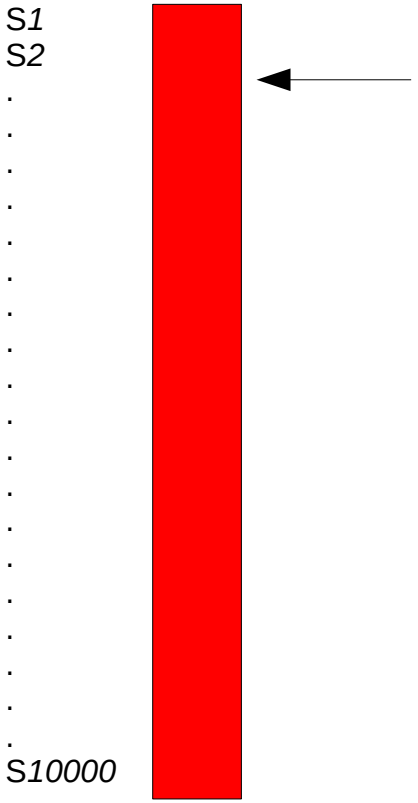


Or this?



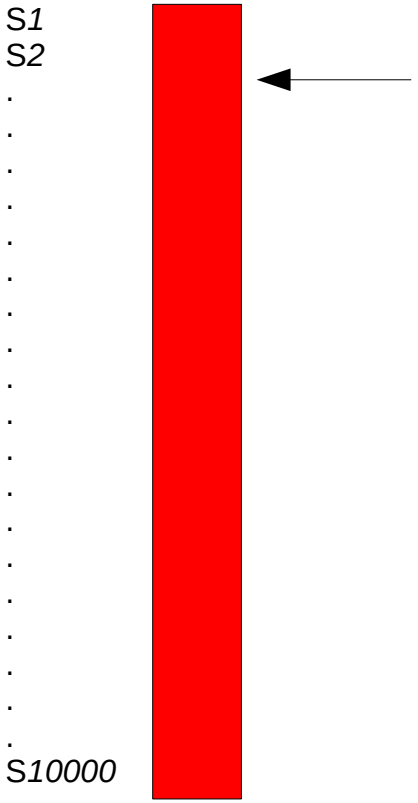
Few sequences, long sequence length

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

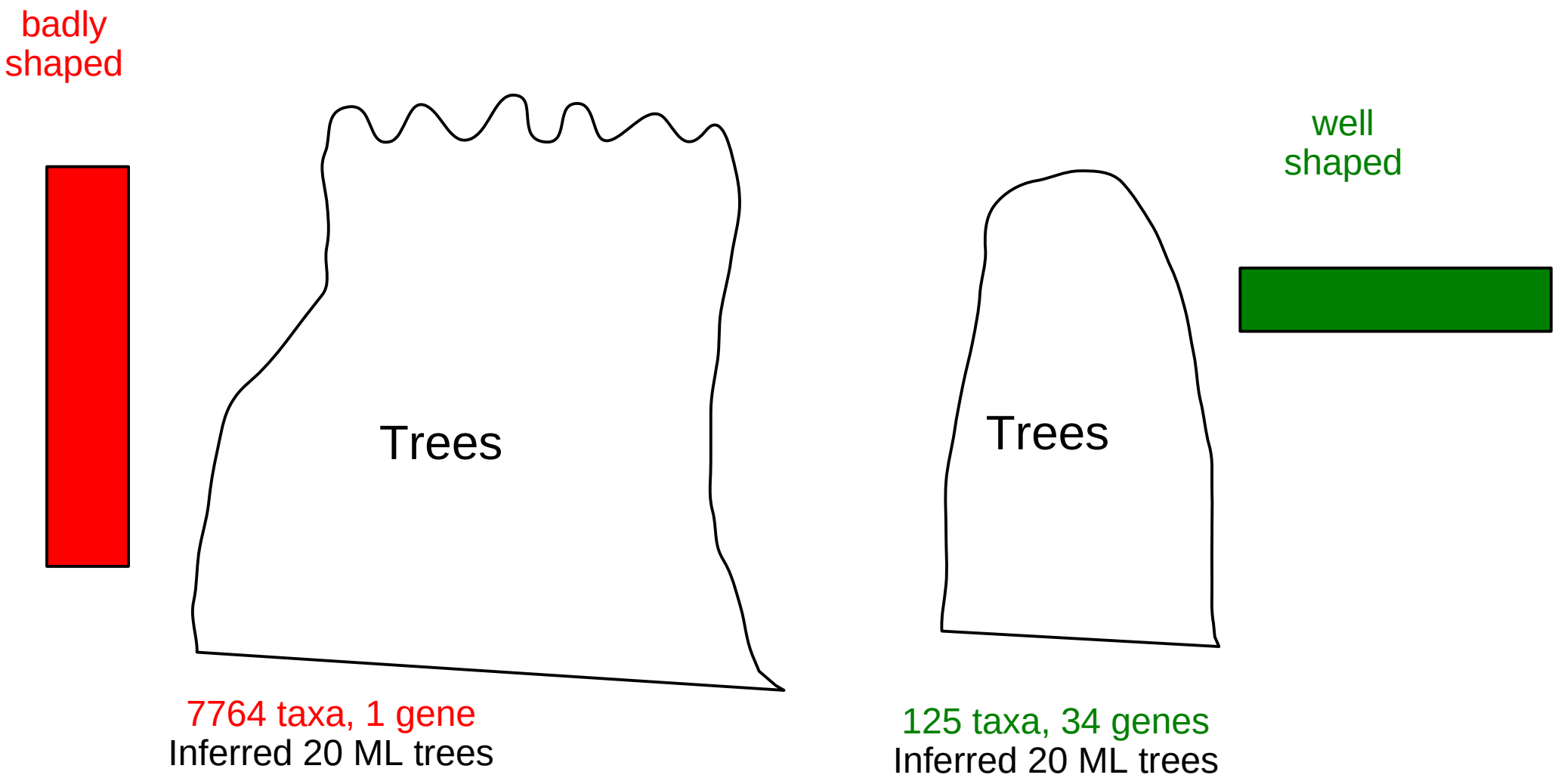
JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

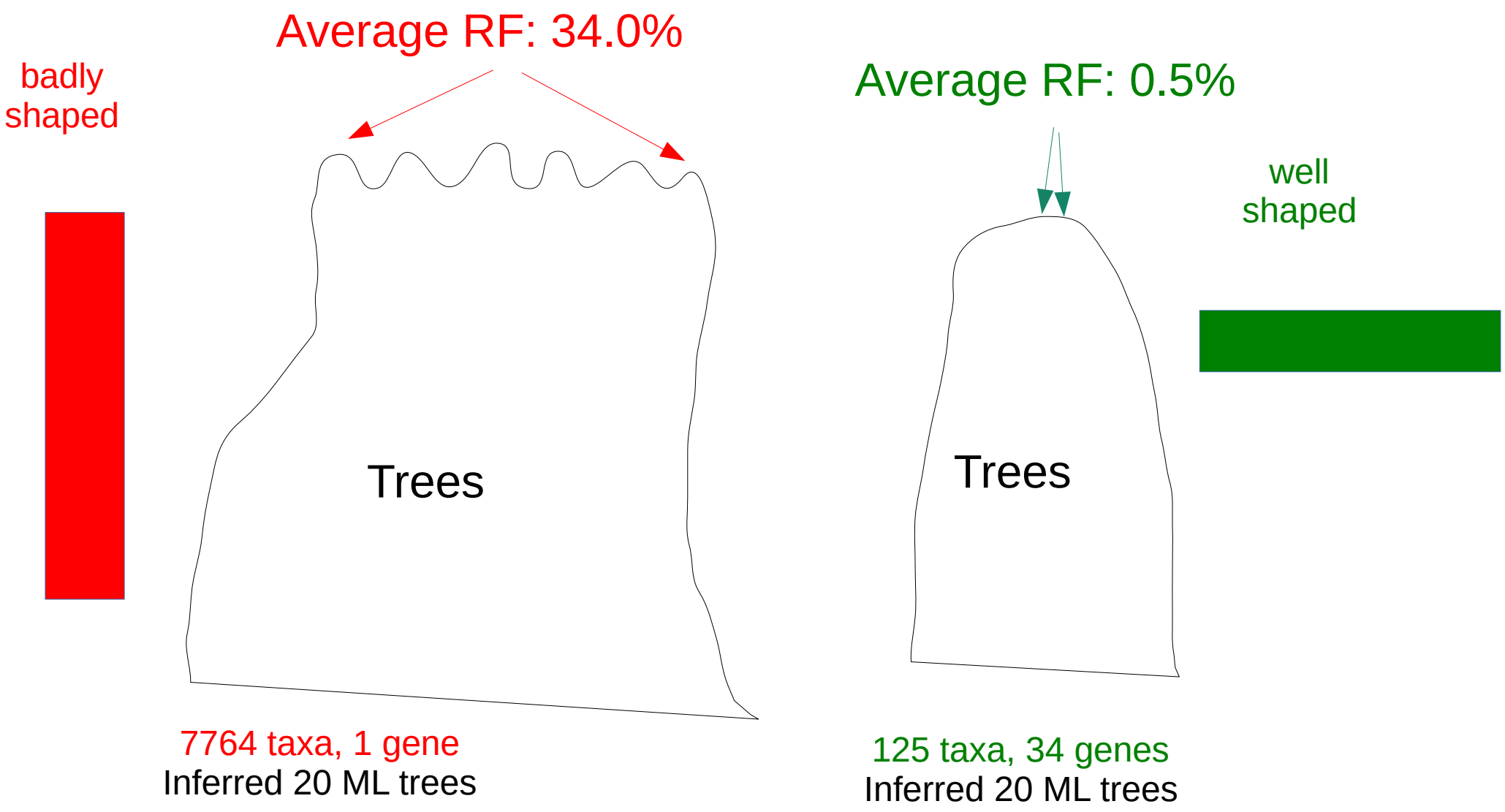
Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis
Author Notes

Molecular Biology and Evolution, Volume 38, Issue 5, May 2021, Pages 1777–1791,
https://doi.org/10.1093/molbev/msaa314
Published: 15 December 2020

Easy & Difficult Likelihood Surfaces



Easy & Difficult Likelihood Surfaces



Now we can quantify this

- In past years these slides about easy and hard datasets were very hand-wavy
- Since 2022 we can quantify & predict difficulty

JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

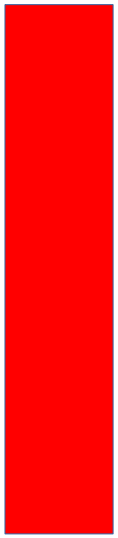
Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

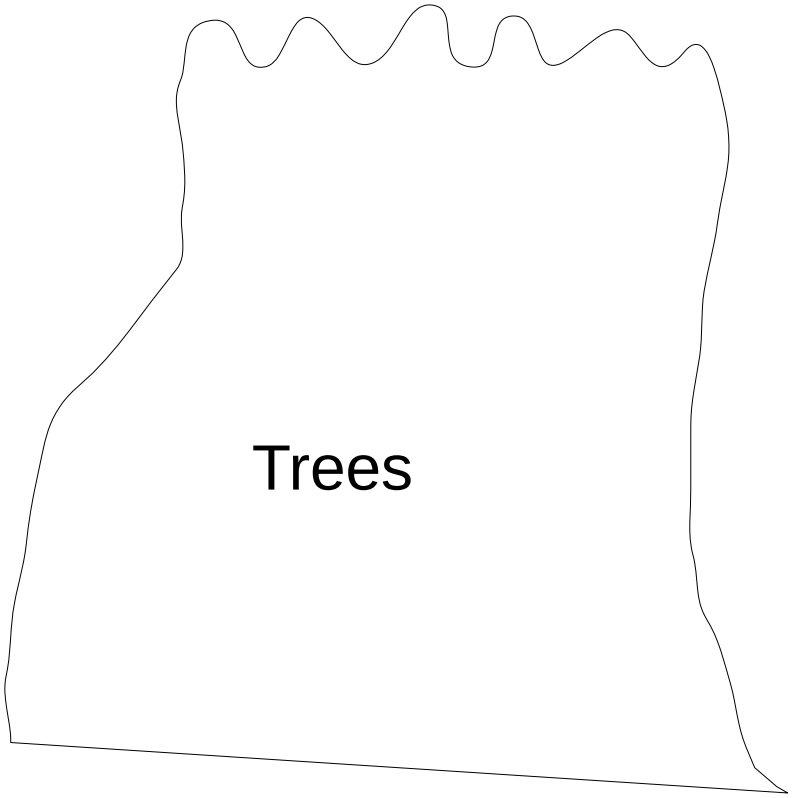
Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac254,
https://doi.org/10.1093/molbev/msac254
Published: 17 November 2022

Easy & Difficult Likelihood Surfaces

Difficulty: 0.63

badly shaped

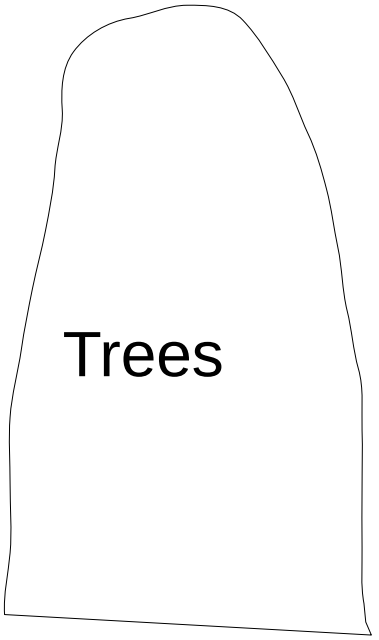




7764 taxa, 1 gene

Difficulty: 0.14

well shaped



125 taxa, 34 genes

`Pythia` Usage

- `Pythia` predicts difficulty of phylogenetic analysis via boosted tree regressor
- **Input:** molecular multiple sequence alignment or binary dataset
- **Output:** difficulty between **0.0** (easy) and **1.0** (hopeless)
- Invocations for our example datasets:

```
pythia --msa 125.phy --raxmlng ~/bin/raxml-ng
pythia --msa 7764.phy --raxmlng ~/bin/raxml-ng
```

- **Anecdotal Observations**
  - Good correlation between the difficulty score and the average bootstrap support values
  - “Apparent convergence” speed of MCMC analyses can be predicted

Small SARS-CoV-2 dataset

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" (https://doi.org/10.1093/molbev/msaa314)

```
The predicted difficulty for MSA examples/covid.fasta is: 0.84.

FEATURES:

num_taxa: 4869

num_sites: 28361

[ ... ]

num_sites/num_taxa: 5.82

[ ... ]

avg_rfdist_parsimony: 0.79

proportion_unique_topos_parsimony: 1.0

Feature computation runtime:    1830.182 seconds

[ ... ]
```

PYTHIA Features

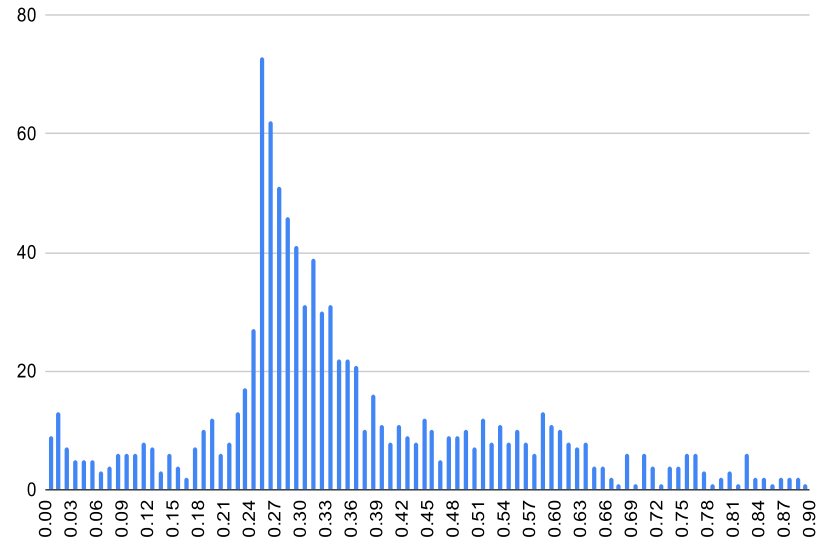
Table 1. Importance of the Subset of Features we use to Train Pythia.

Feature	Impurity Importance
% Unique topologies parsimony trees	42.9%
RF-distance parsimony trees	33.2%
Entropy	17.0%
Patterns-over-taxa	13.6%
% Gaps	2.5%
Bollback	2.3%
Sites-over-taxa	1.5%
% Invariant	0.6%

Parsimony = 76%

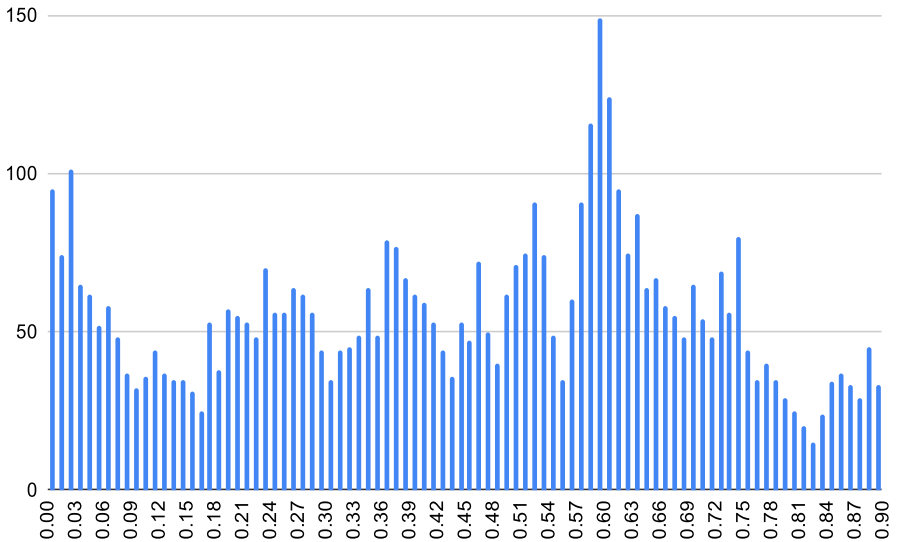
Empirical Difficulty Distributions

#MSAs



`RAxML-Grove` Database

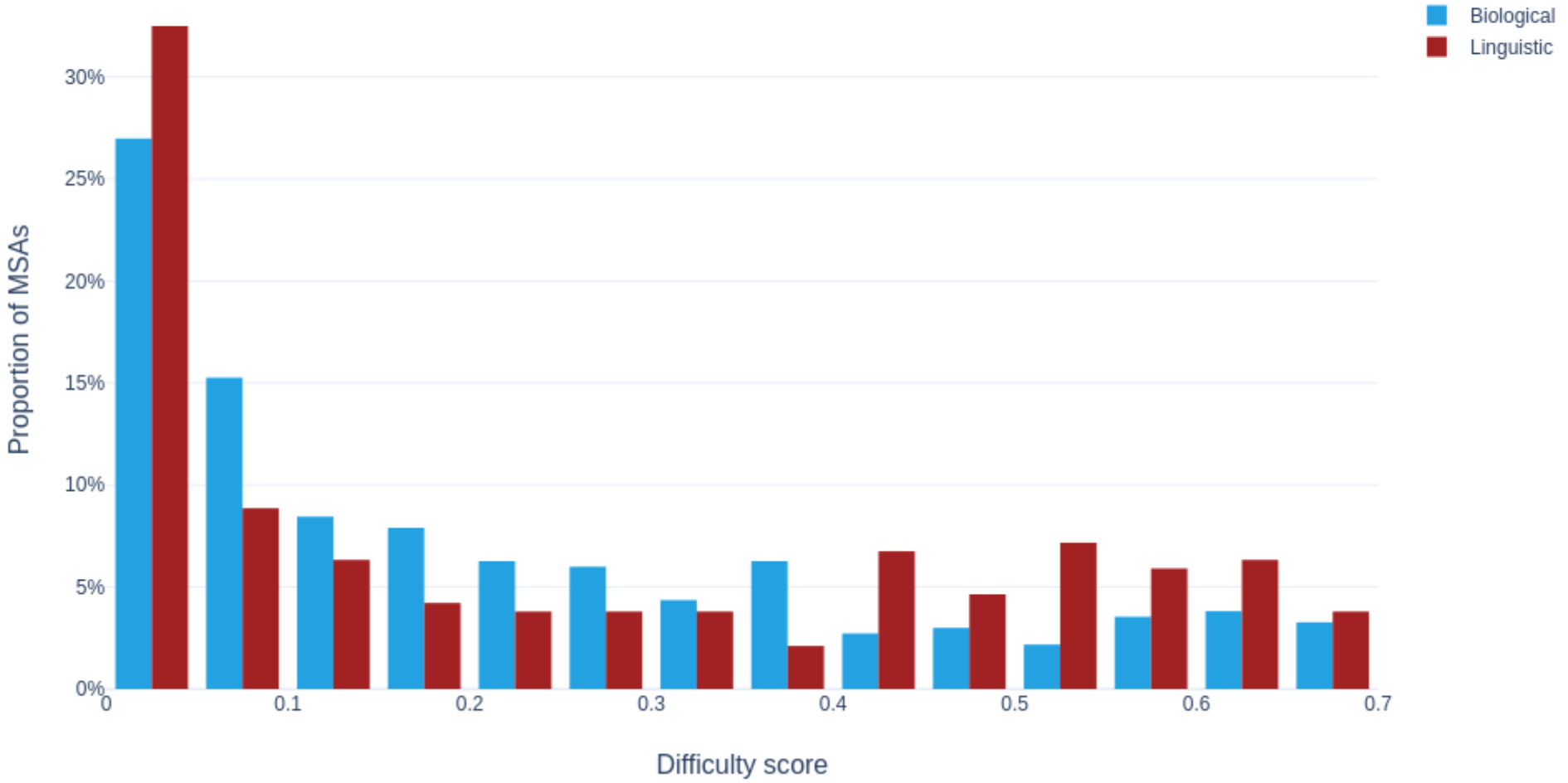
Easy ⟷ Hopeless



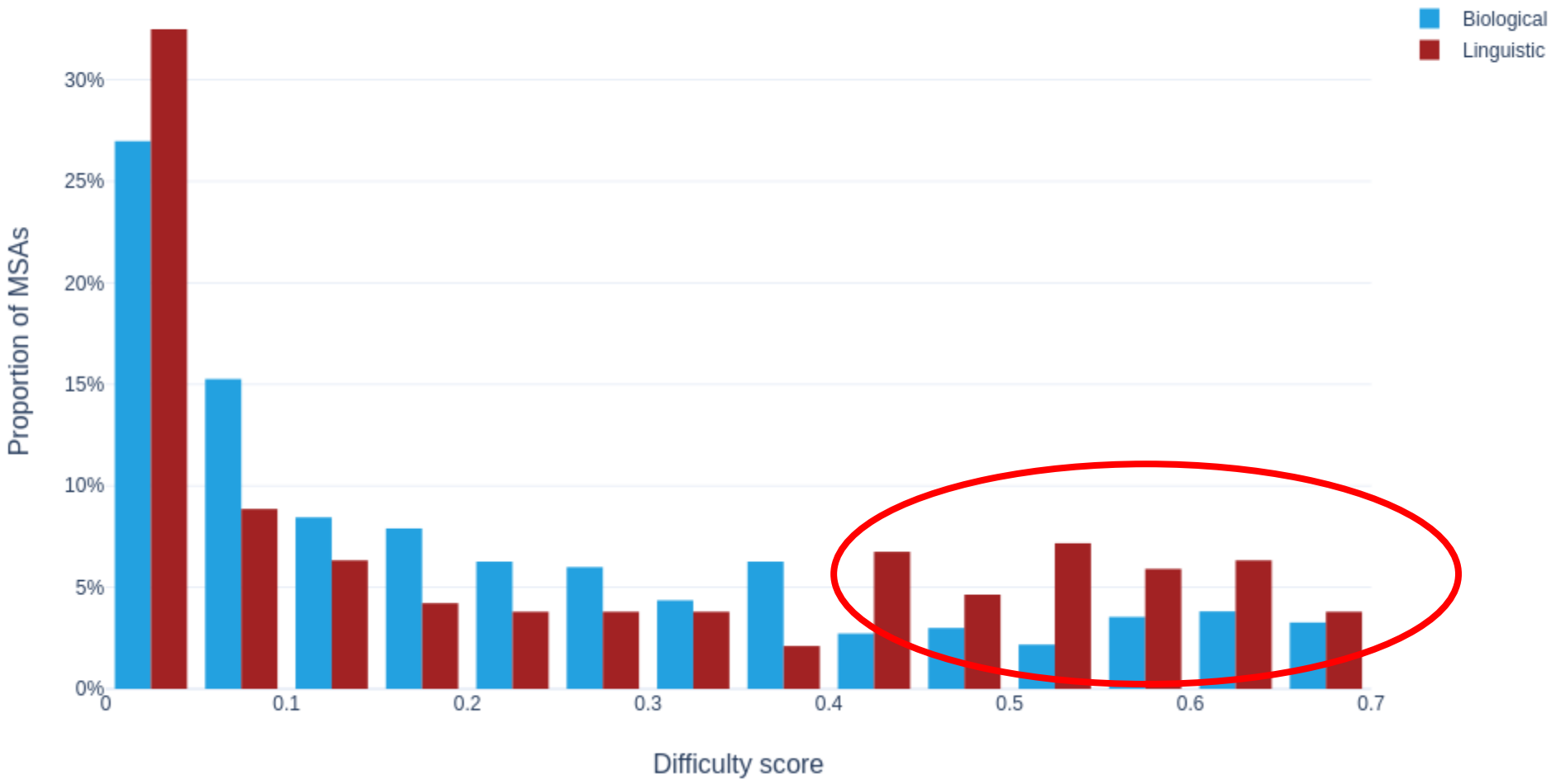
`TreeBase` Database

Easy ⟷ Hopeless

Difficulty: Binary Biological & Language Data

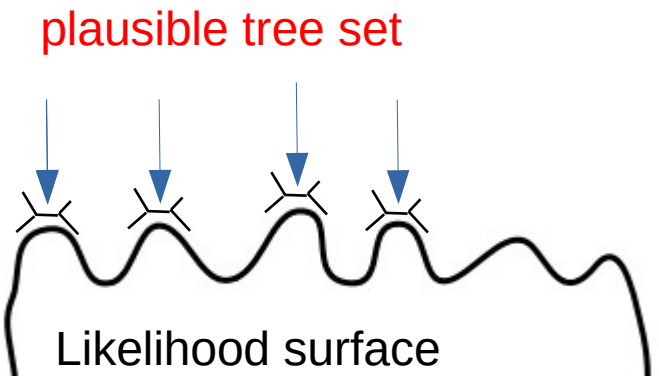


Difficulty: Binary Biological & Language Data

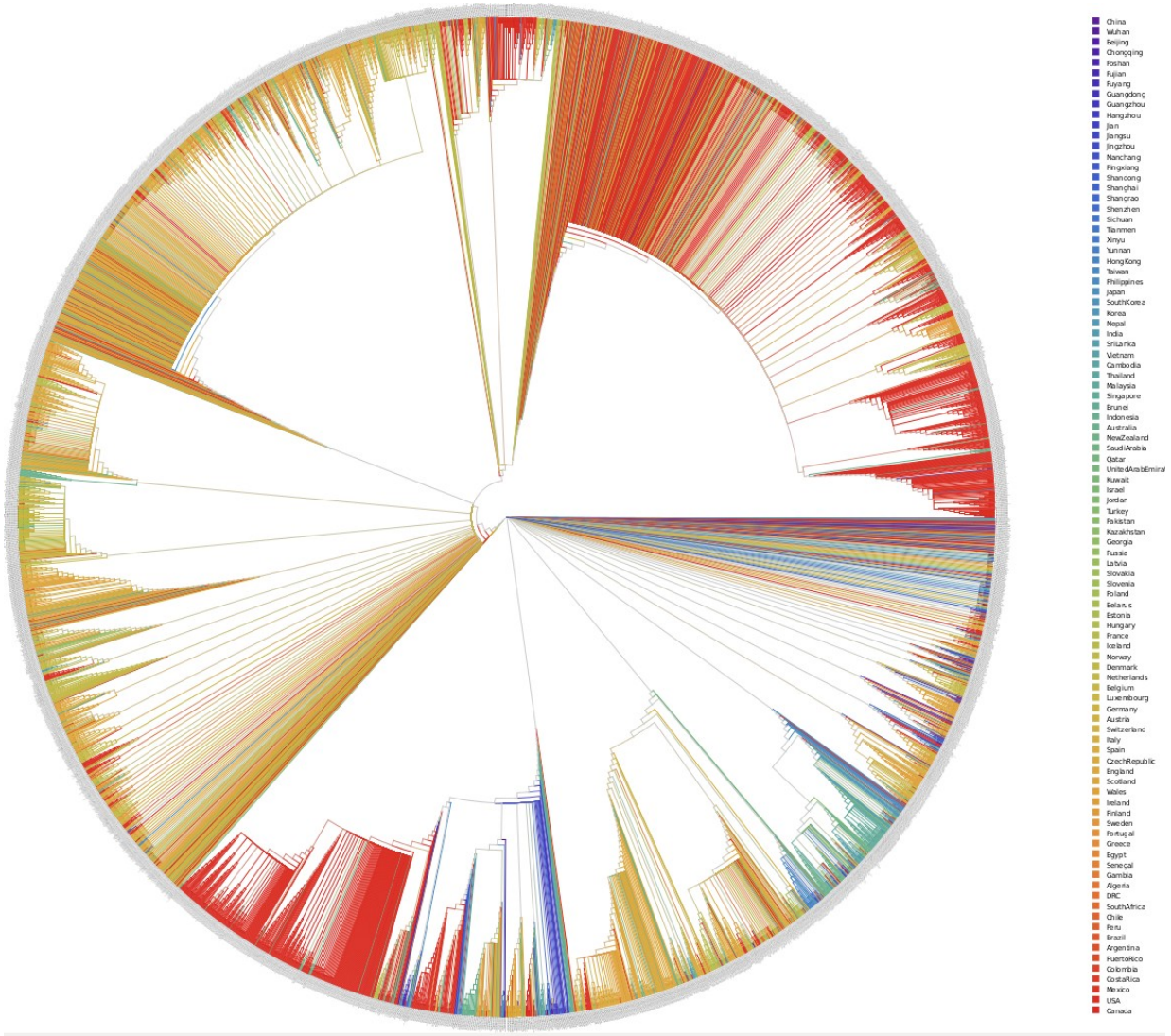


Difficult Datasets

- On difficult datasets
  - Infer a plausible tree set
  - And summarize it → summary statistics



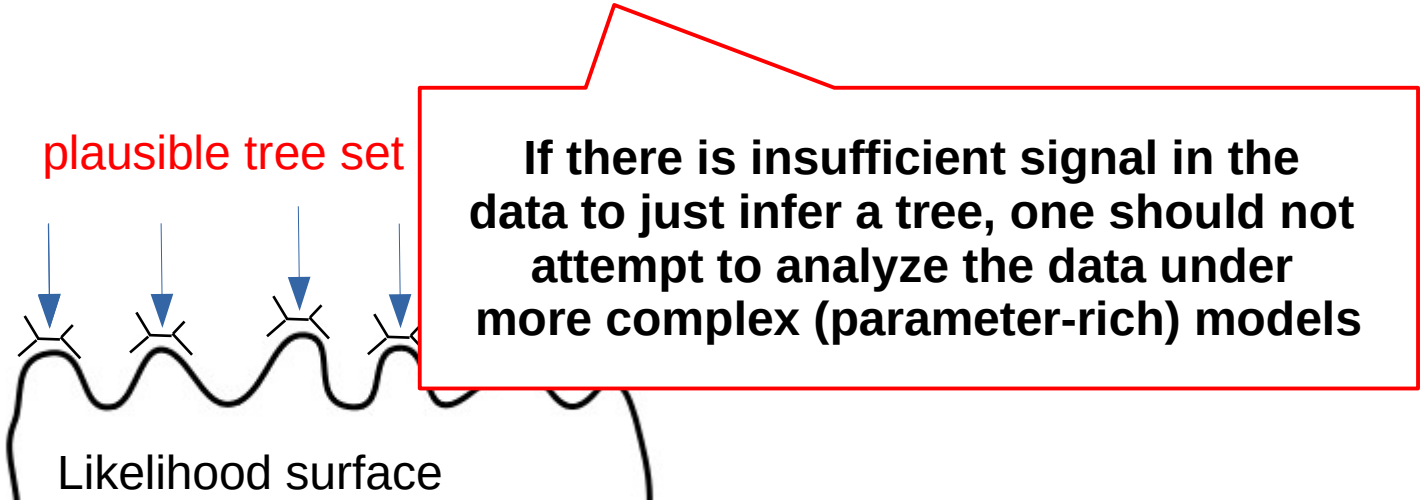
Summarized Trees



SARS-CoV-2 consensus tree colored by country

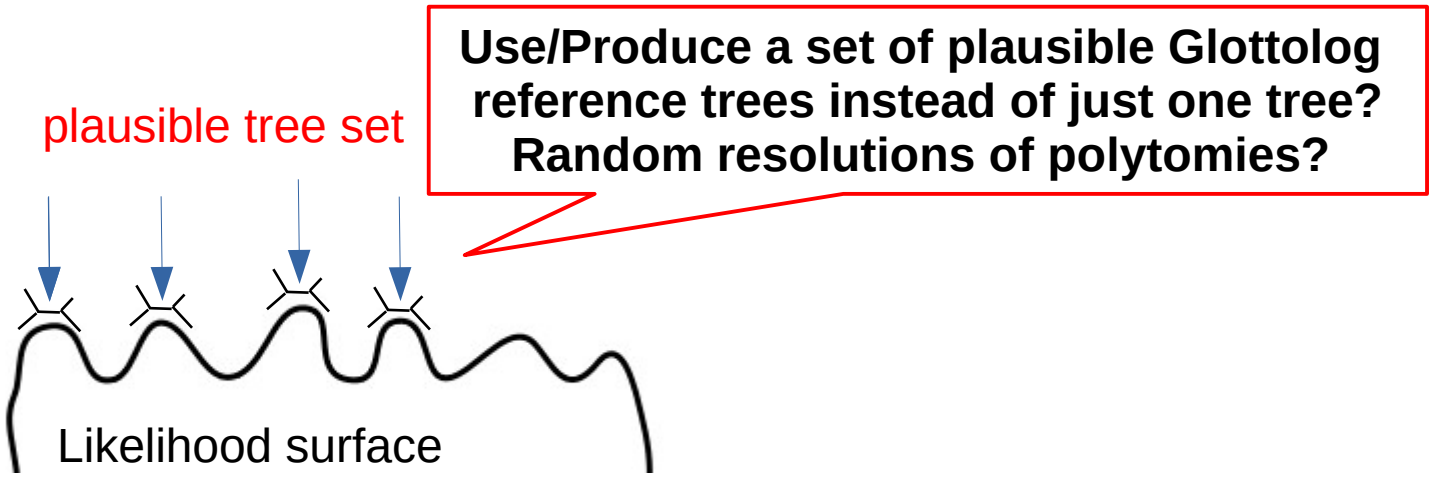
Difficult Datasets

- On difficult datasets
  - Infer a plausible tree set
  - And summarize it → summary statistics



Difficult Datasets

- On difficult datasets
  - Infer a plausible tree set
  - And summarize it → summary statistics



Conflicting Signal

- Quantify strength of conflicting signal?

JOURNAL ARTICLE

Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees FREE

Leonidas Salichos, Alexandros Stamatakis, Antonis Rokas ✉ Author Notes

Molecular Biology and Evolution, Volume 31, Issue 5, May 2014, Pages 1261–1271,
https://doi.org/10.1093/molbev/msu061
Published: 07 February 2014

Educated Bootstrap Guesser (EBG)

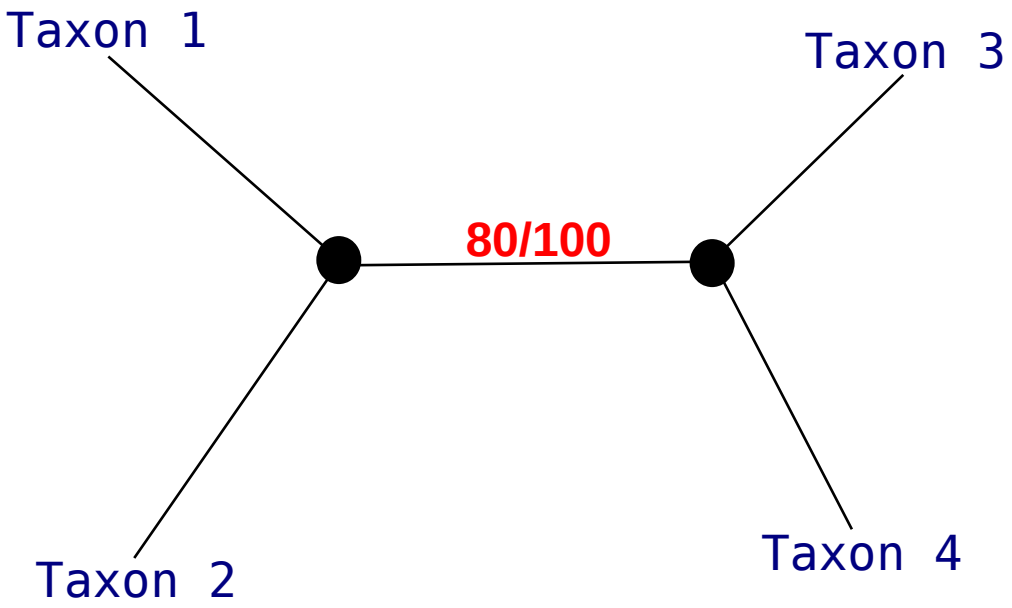
New Results

Predicting Phylogenetic Bootstrap Values via Machine Learning

Julius Wiegert, Dimitri Höhler, Julia Haag, Alexandros Stamatakis

doi: https://doi.org/10.1101/2024.03.04.583288

A Tree with computationally expensive Standard Bootstrap Values



Educated Bootstrap Guesser (EBG)

- One order of magnitude faster than existing fast methods (UFBoot2: UltraFast Bootstrap version 2)
- Median error of 5 when predicting bootstrap values between 0-100
- 1654 SARS-CoV2 sequences
  - Bootstrap prediction in 3 hours on mid-class laptop

The Renaissance of Parsimony for Machine Learning

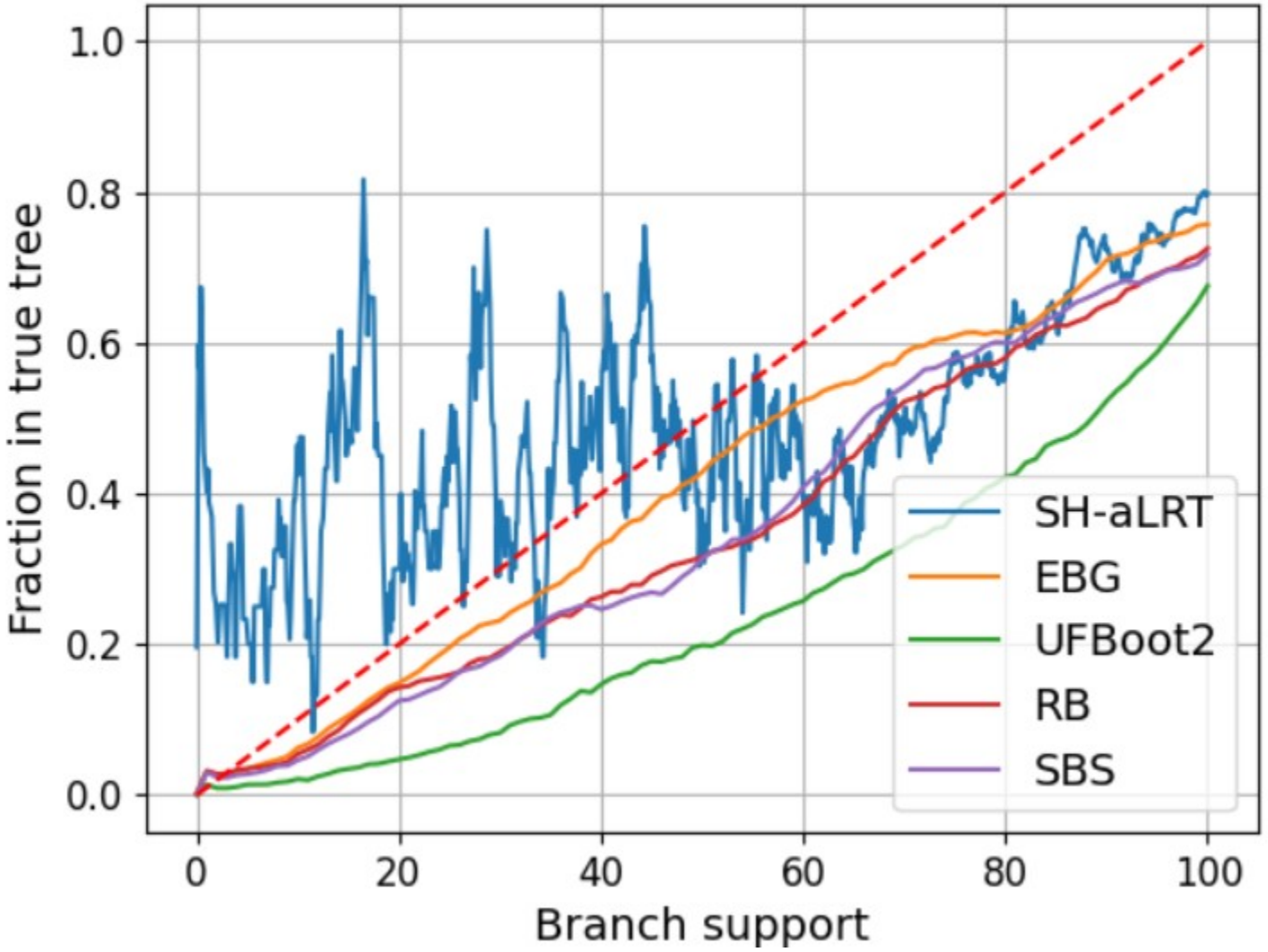
Light Gradient-Boosting Model from tree-based boosting ensemble framework.

Parsimony: 85%

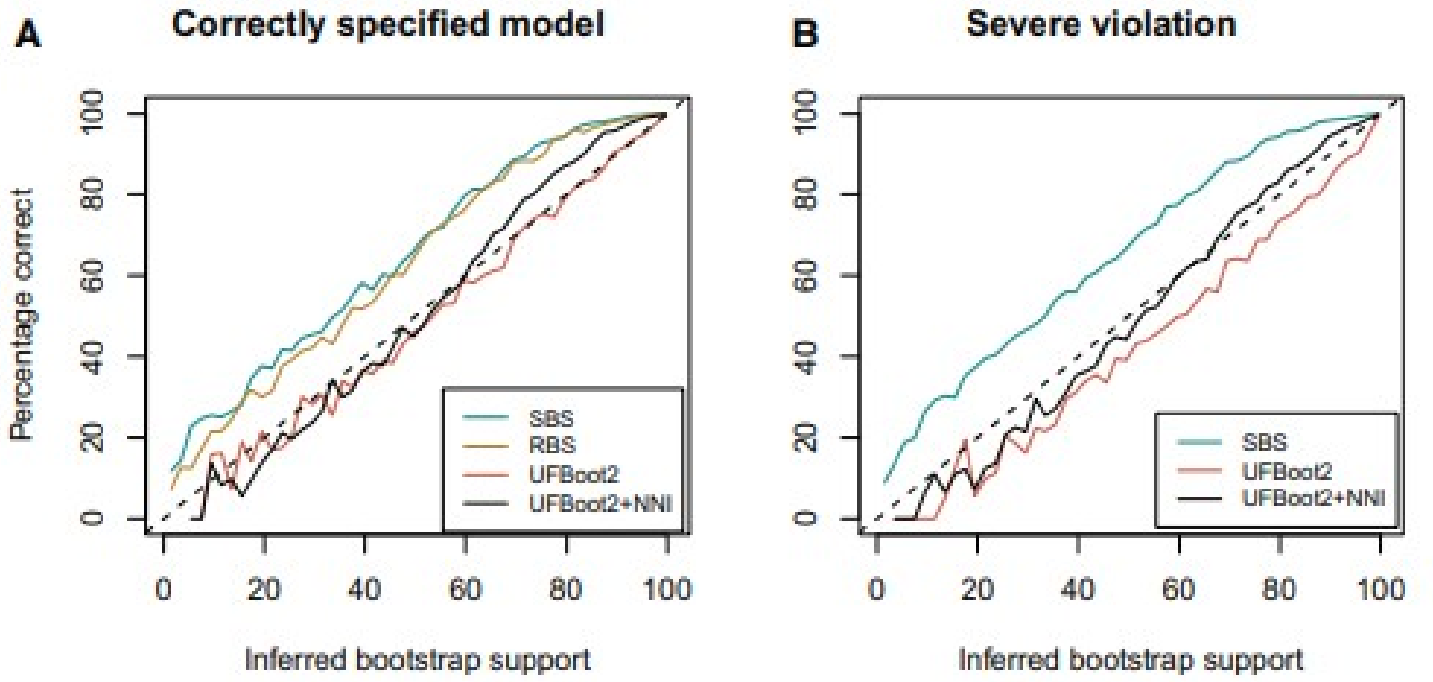
Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

PBS = **P**arsimony **B**ootstrap **S**upport from *200* parsimony bootstraps
PS = **P**arsimony **S**upport from *1000* parsimony starting trees

Accuracy – Simulated Data



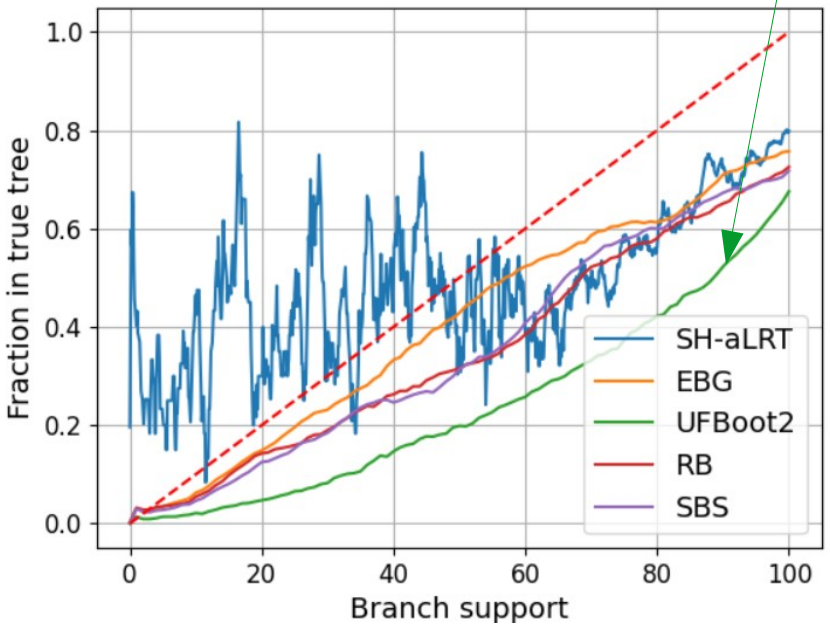
But ...

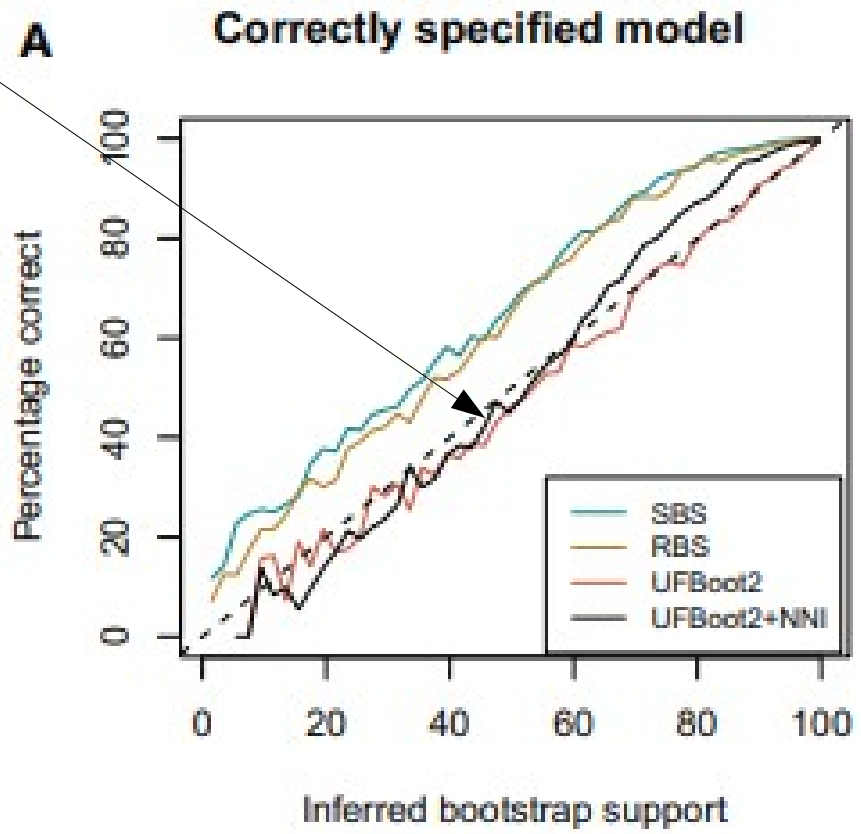


Accuracy on simulated data from UFBoot2 paper

But ...

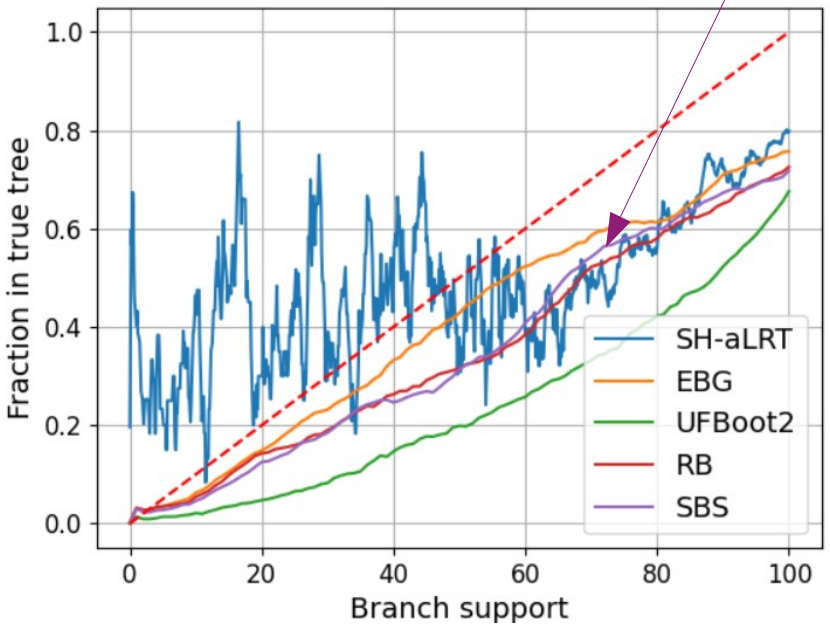
????

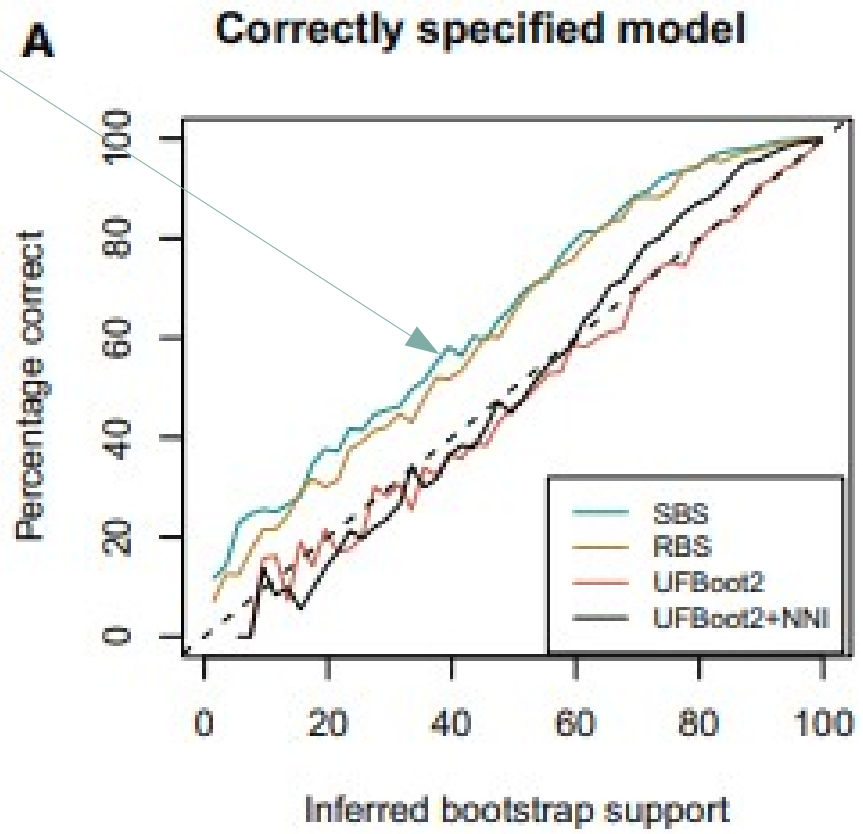

Fraction in true tree
Branch support
SH-aLRT
EBG
UFBoot2
RB
SBS


A
Correctly specified model
Percentage correct
Inferred bootstrap support
SBS
RBS
UFBoot2
UFBoot2+NNI

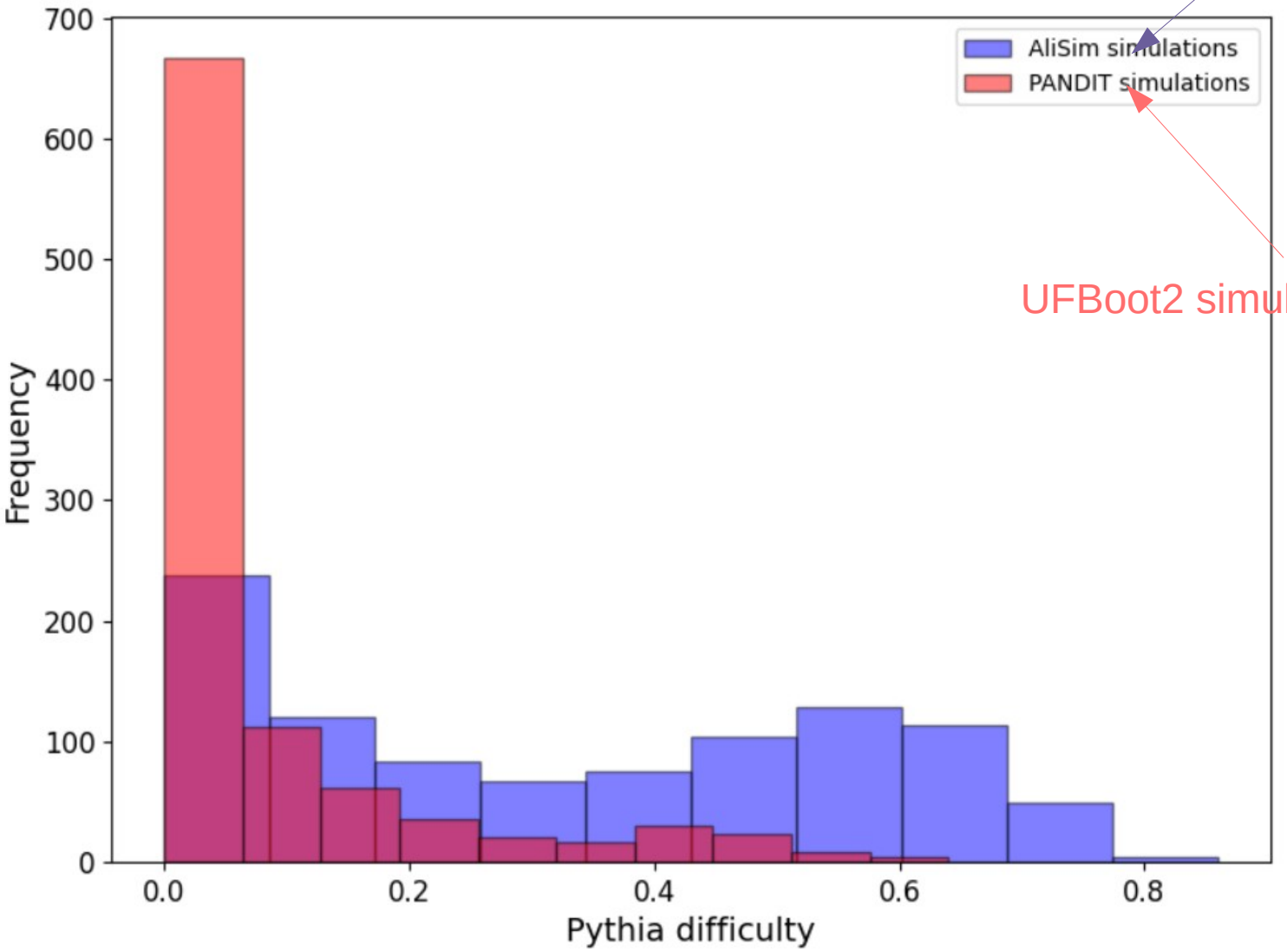
But ...

????


Fraction in true tree
Branch support
SH-aLRT
EBG
UFBoot2
RB
SBS


A
Correctly specified model
Percentage correct
Inferred bootstrap support
SBS
RBS
UFBoot2
UFBoot2+NNI

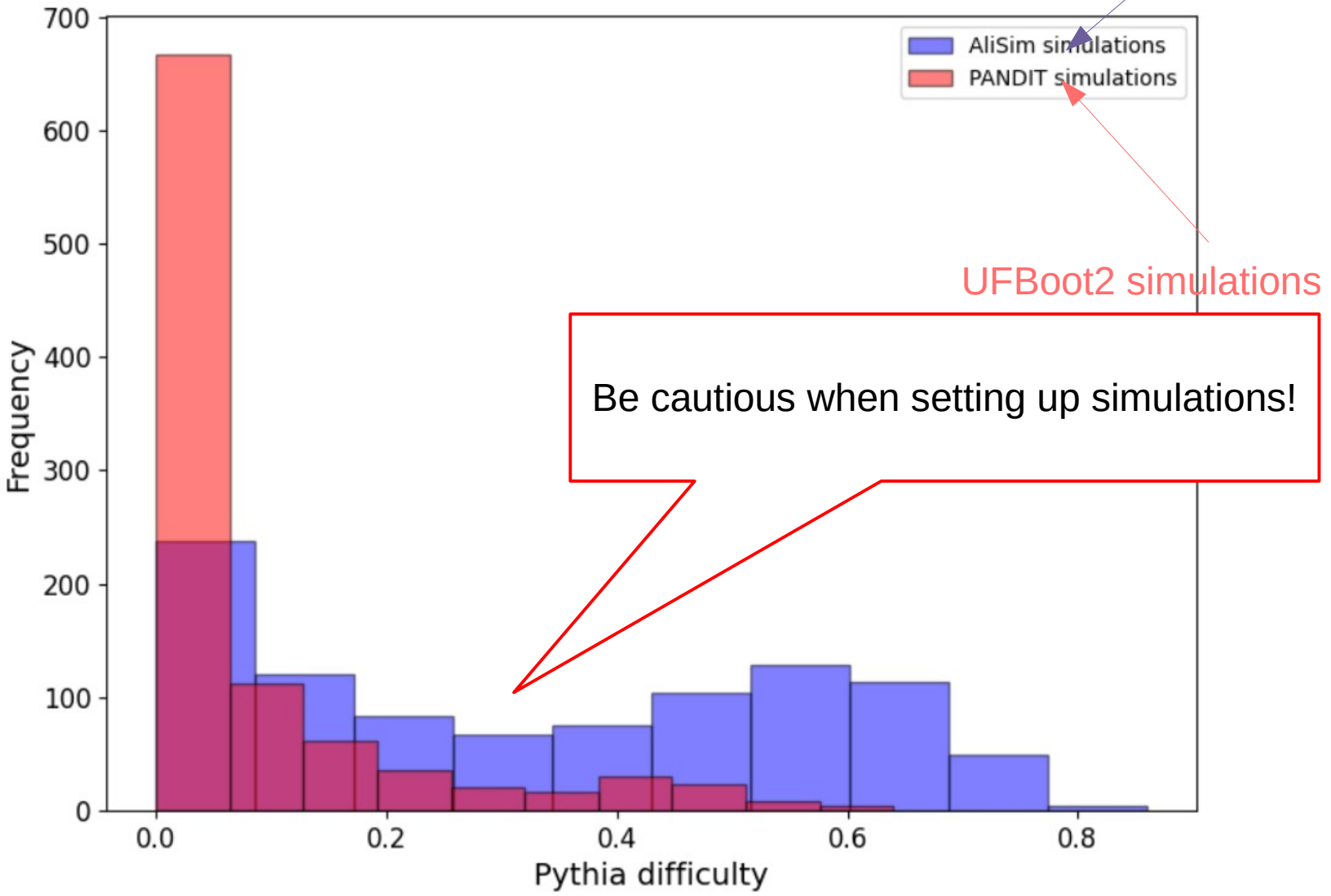
Difficulty Distribution



Our simulations

UFBoot2 simulations

Difficulty Distribution



Simulated Data Suck!

JOURNAL ARTICLE

Simulations of Sequence Evolution: How (Un)realistic They Are and Why

Johanna Trost, Julia Haag ✉, Dimitri Höhler, Laurent Jacob, Alexandros Stamatakis, Bastien Boussau Author Notes

Molecular Biology and Evolution, Volume 41, Issue 1, January 2024, msad277, https://doi.org/10.1093/molbev/msad277

Published: 20 December 2023 **Article history** ▾

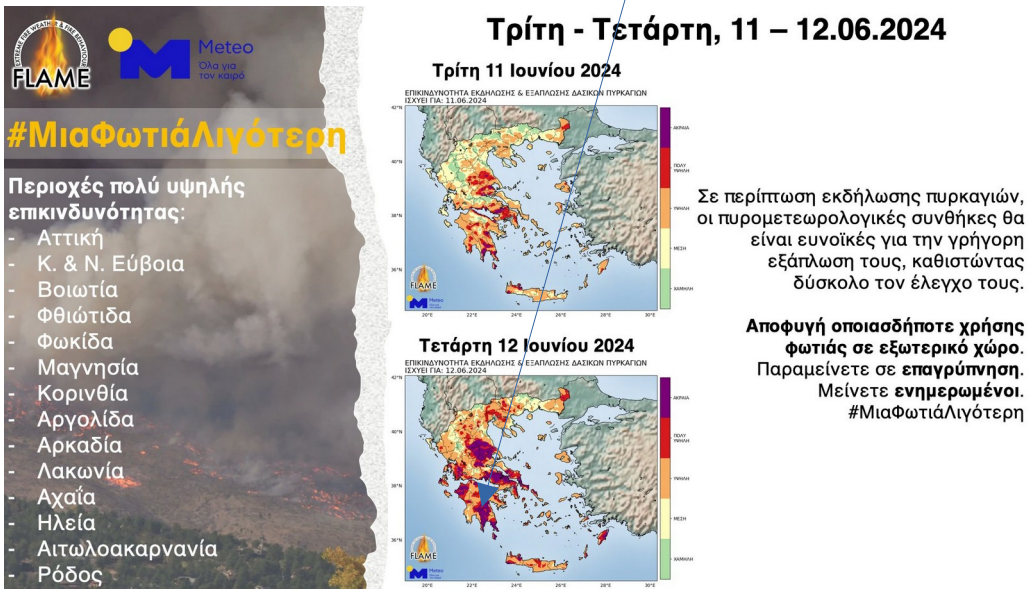
We can distinguish between empirical and simulated molecular data with high accuracy using two distinct and independently developed machine learning based classification approaches!

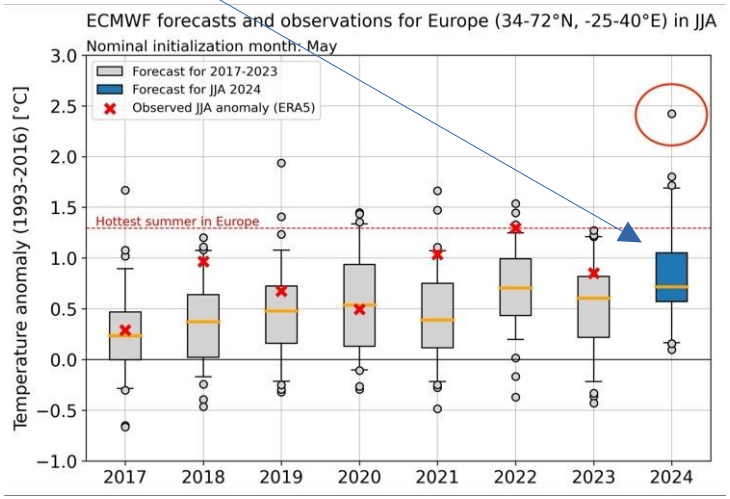
Outline

- Recent Advances in Phylogenetic Inference
- **Energy-efficient Computing**

Motivation
Rural Community Southern Crete

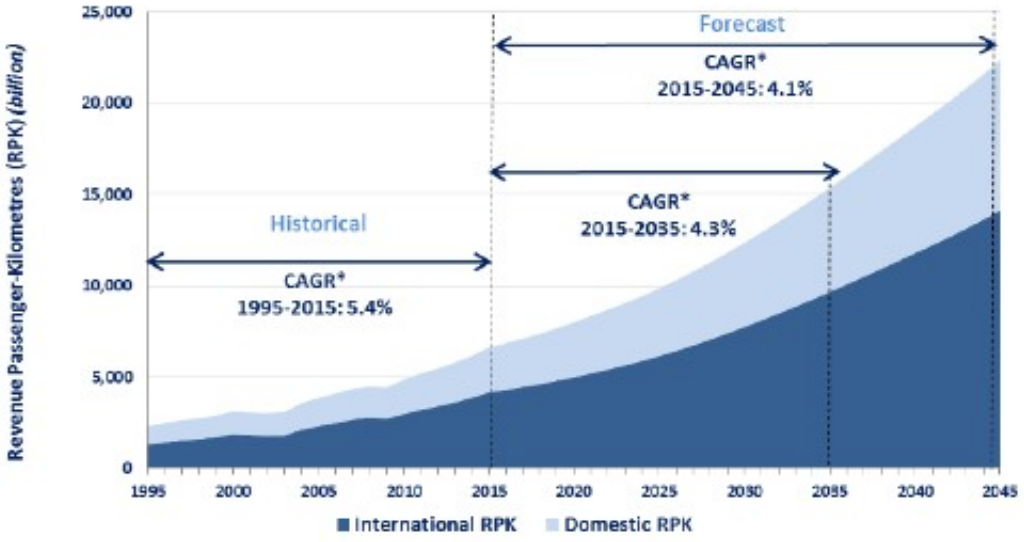
- Crete
  - Water temperature around Crete was 3 degrees above average in April
  - Local fire chief was extremely nervous beginning of April
- Medium term summer temperature forecast is worrisome (European Centre for Medium-Range Weather Forecasts)
- Fire danger is **extreme** in many regions of Greece



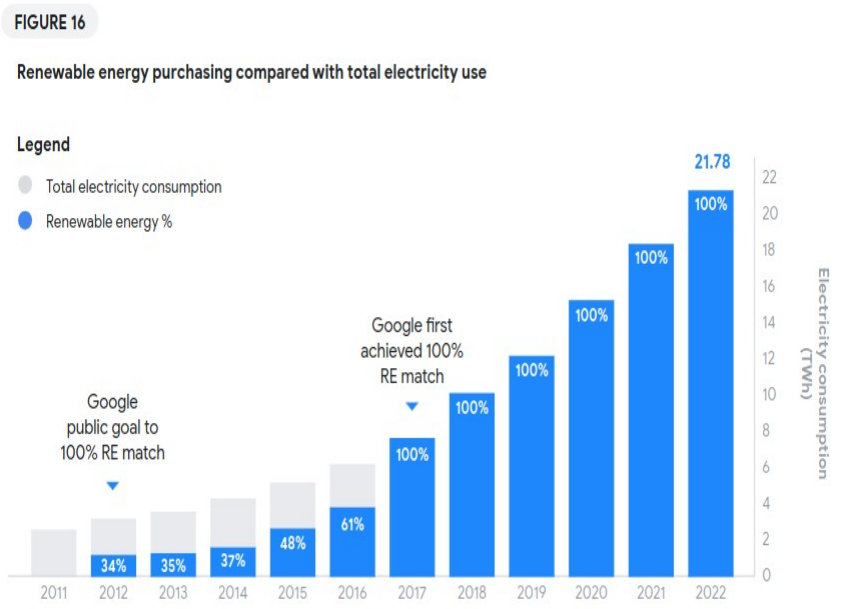


Efficiency is insufficient

- Efficiency gains != savings (Jevons 1865)
  - "But we optimize jet turbines!" → see above & below



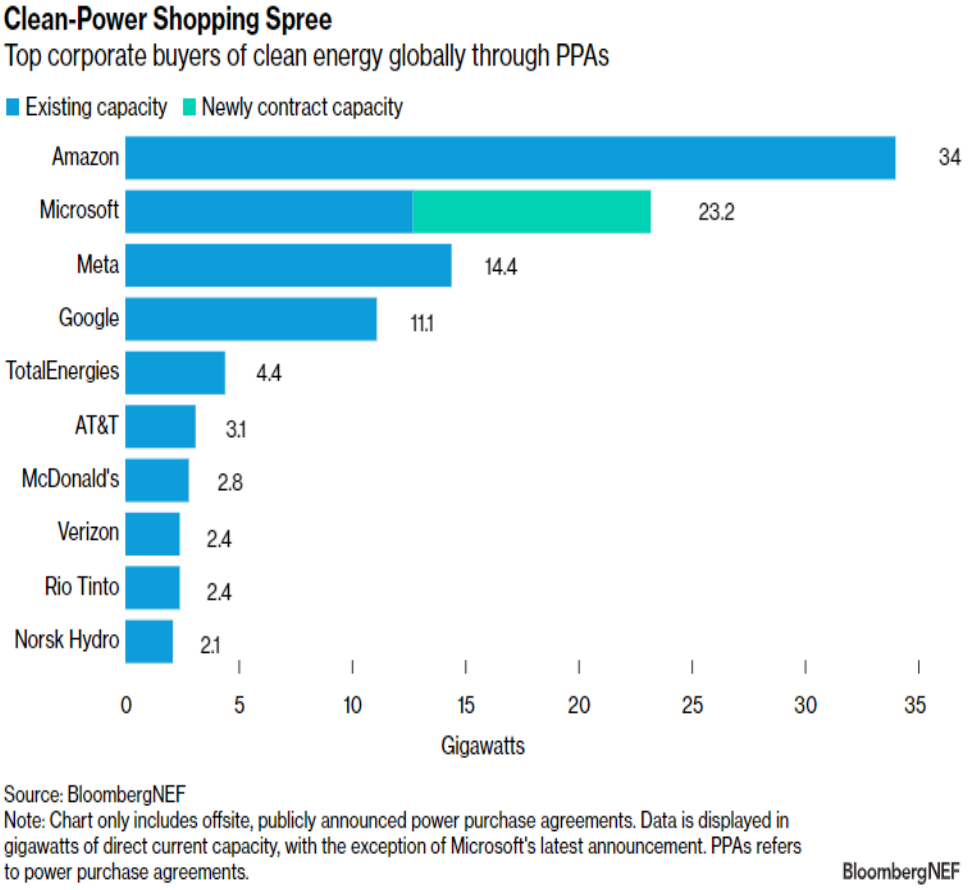
Data: ICAO



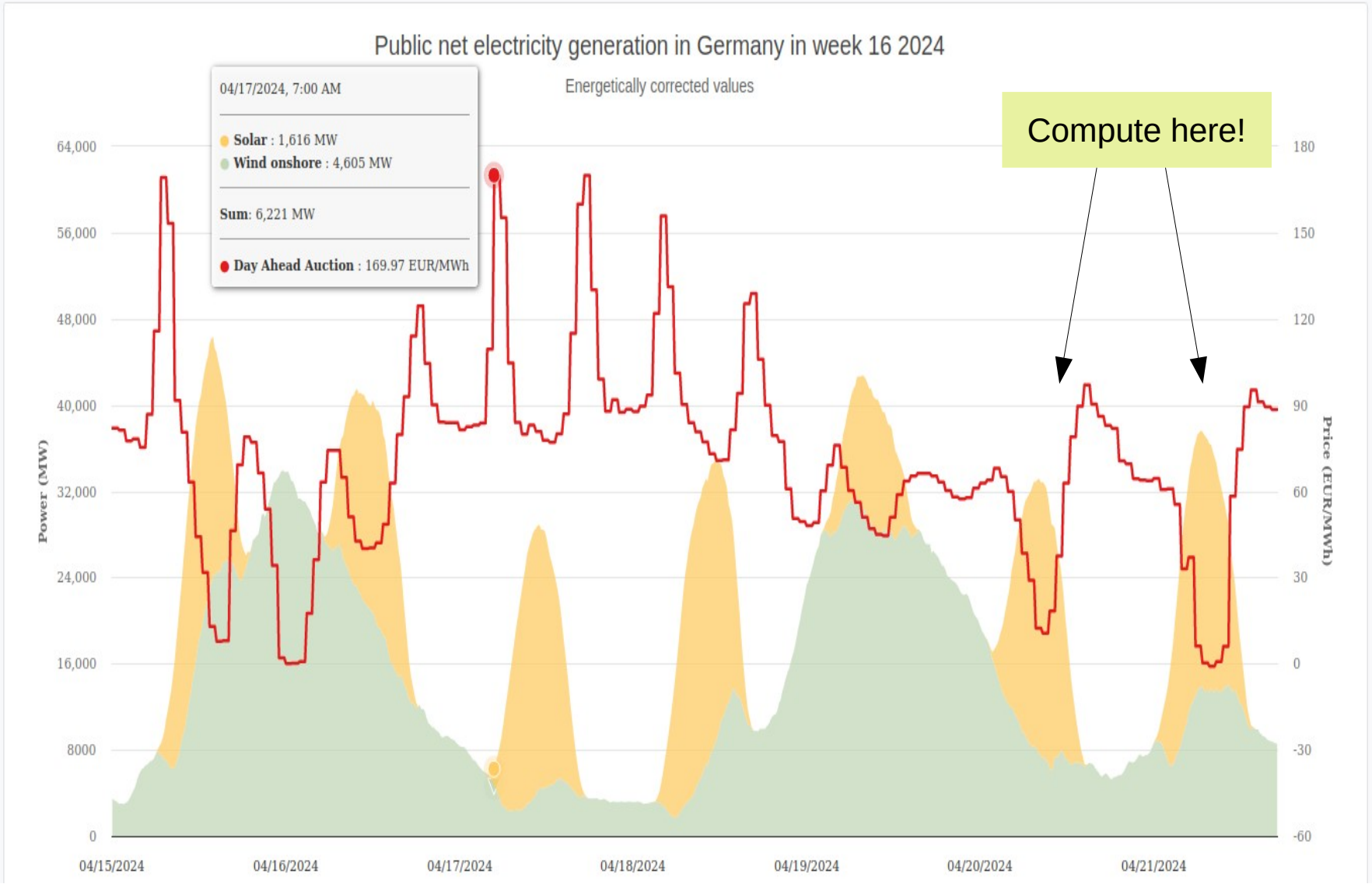
Google Environment report (2023)

Decarbonizing the energy system

- Data centers have a role to play
  - Renewable **P**ower **P**urchase **A**greements (PPAs)
  - Heat reuse (partially)
  - Flexible demand

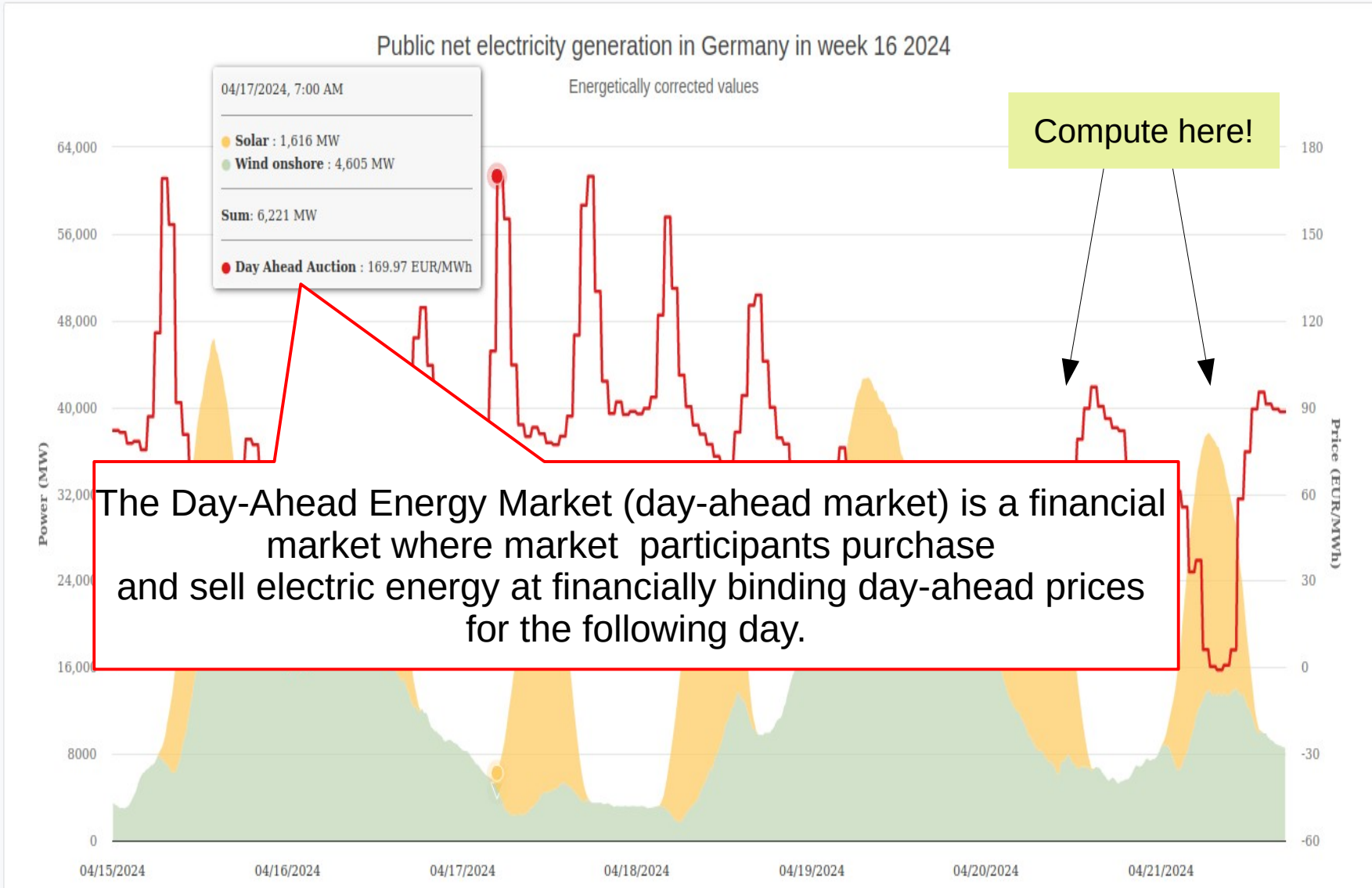


Timing the electricity market



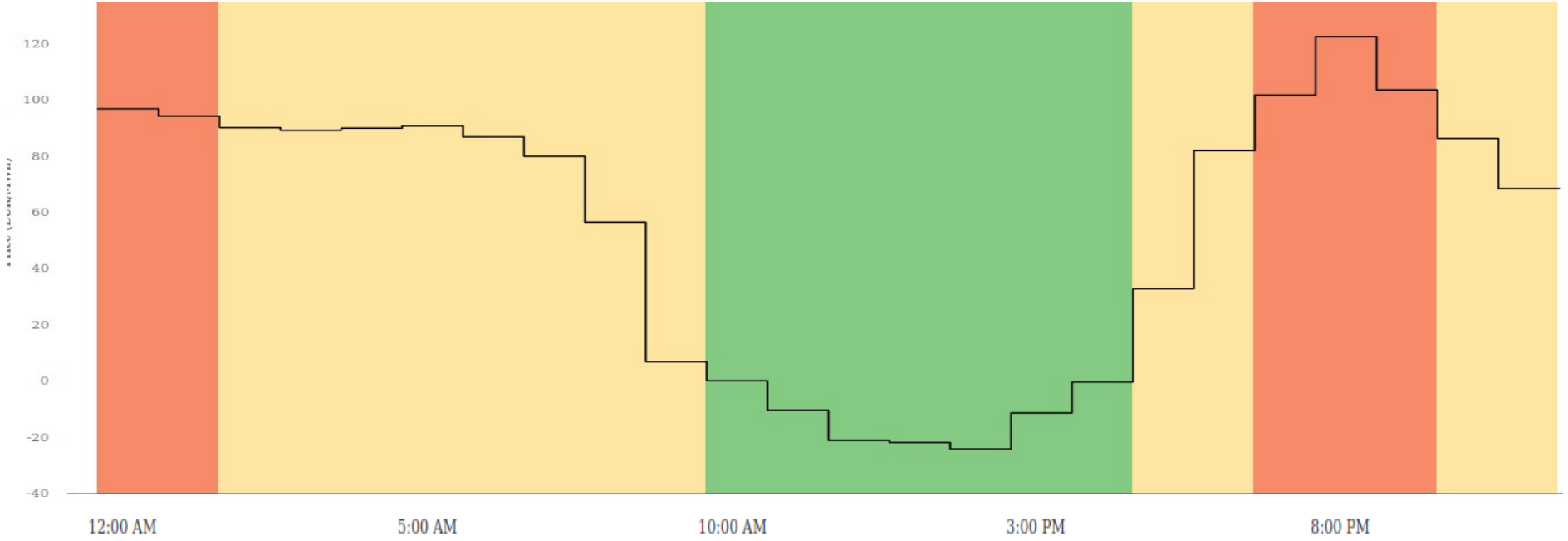
energy-charts.info

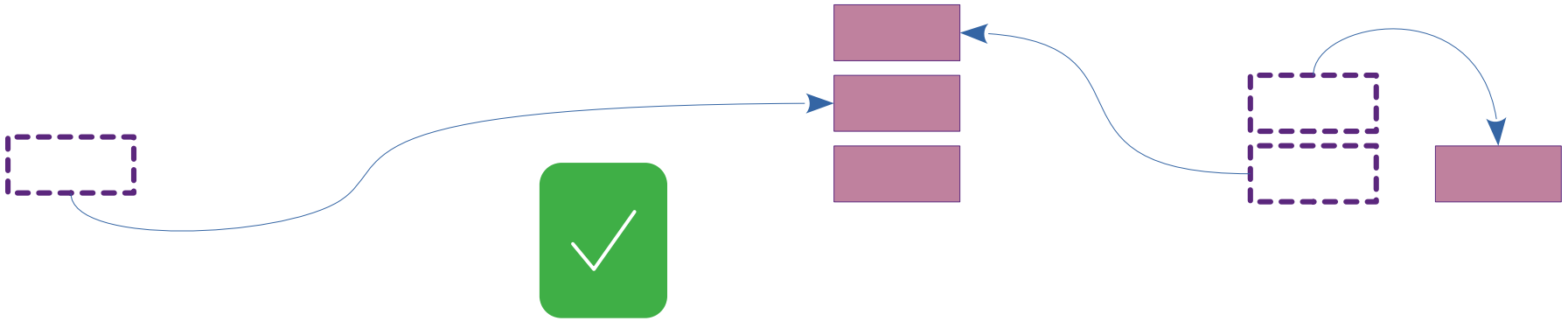
Timing the electricity market



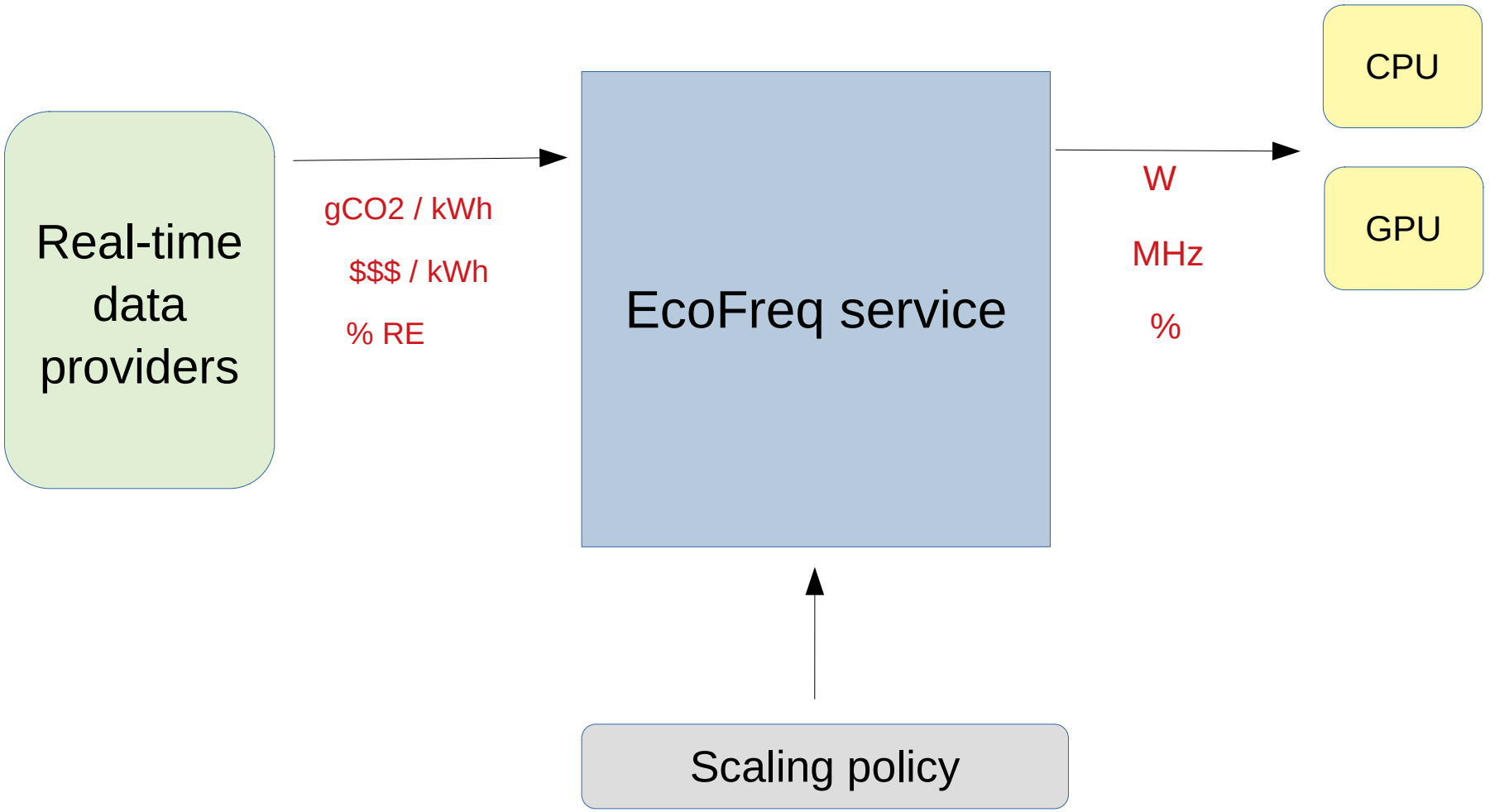
energy-charts.info

Carbon-aware scheduling



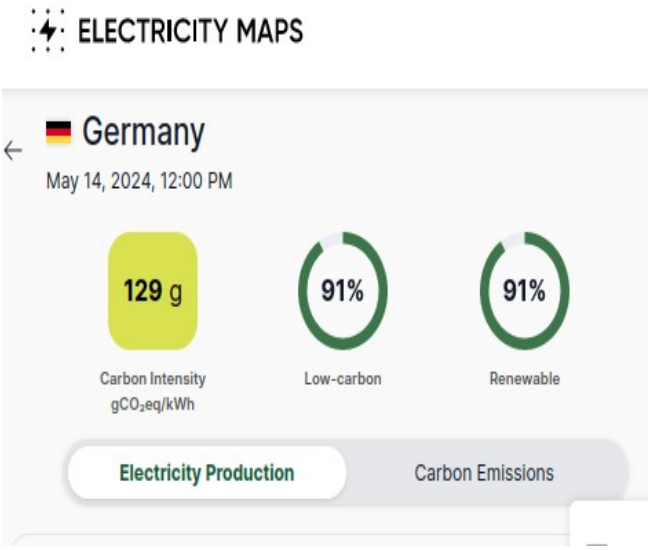


Dynamic carbon-aware power scaling

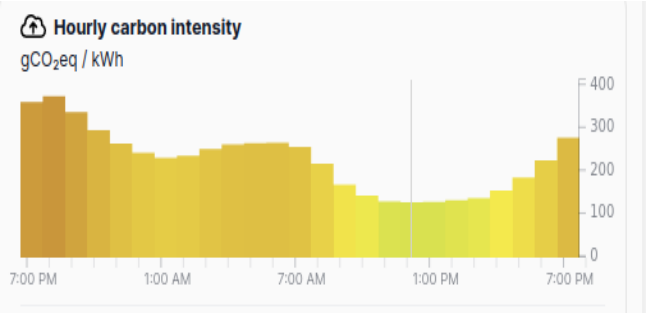


Carbon signal(s)

- Carbon intensity: CI [gCO_2/kWh]
- Share of renewable



Real-time data providers



Energy-Charts

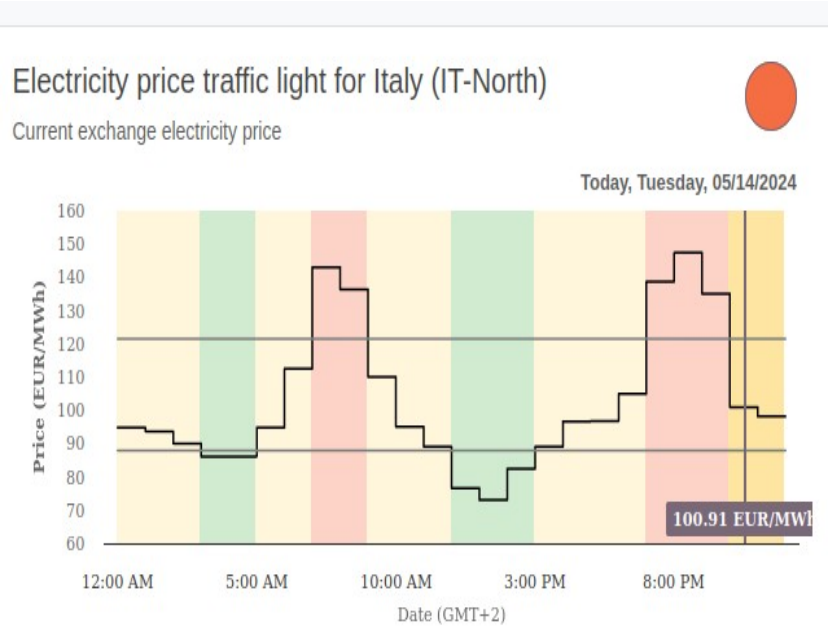
WattTime

nationalgridESO

Alternatively: Carbon “traffic light”

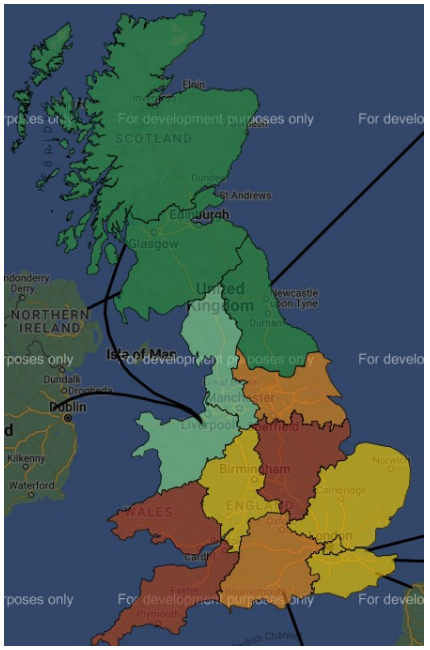
- Discrete signal, e.g. green / yellow / red
- Ideally, reflects *regional* CI (Carbon Intensity)

EU + CH + NO



https://energy-charts.info

UK



https://carbonintensity.org.uk/

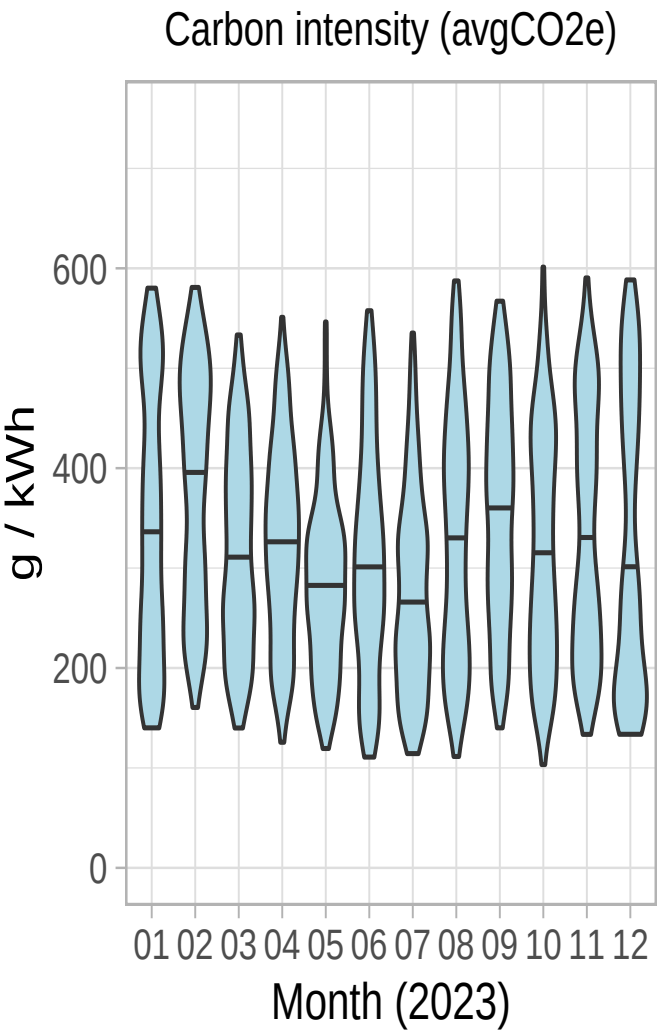
Germany, BW

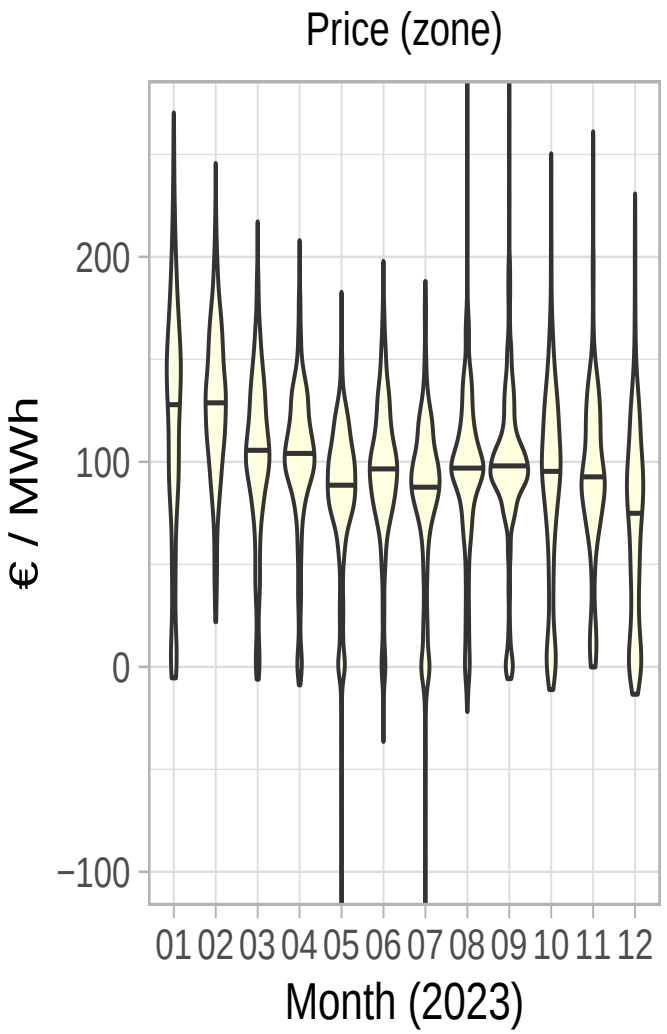
https://www.stromgedacht.de/

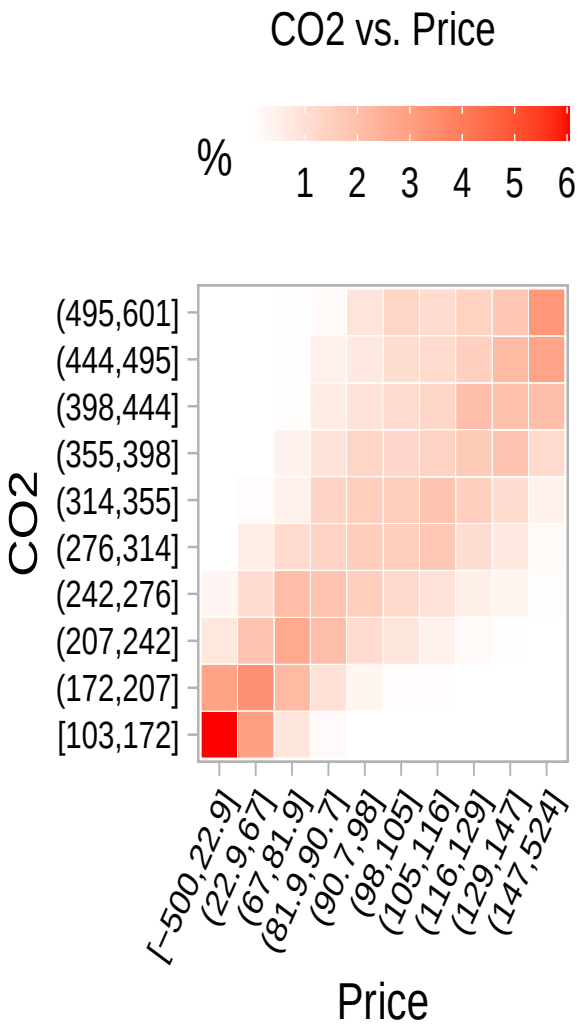
Test Carbon Intensities 2023 Historical Data

	Carbon intensity		Price	
Region	Type	Signal	Market	Type
Germany	average	continuous	wholesale	national

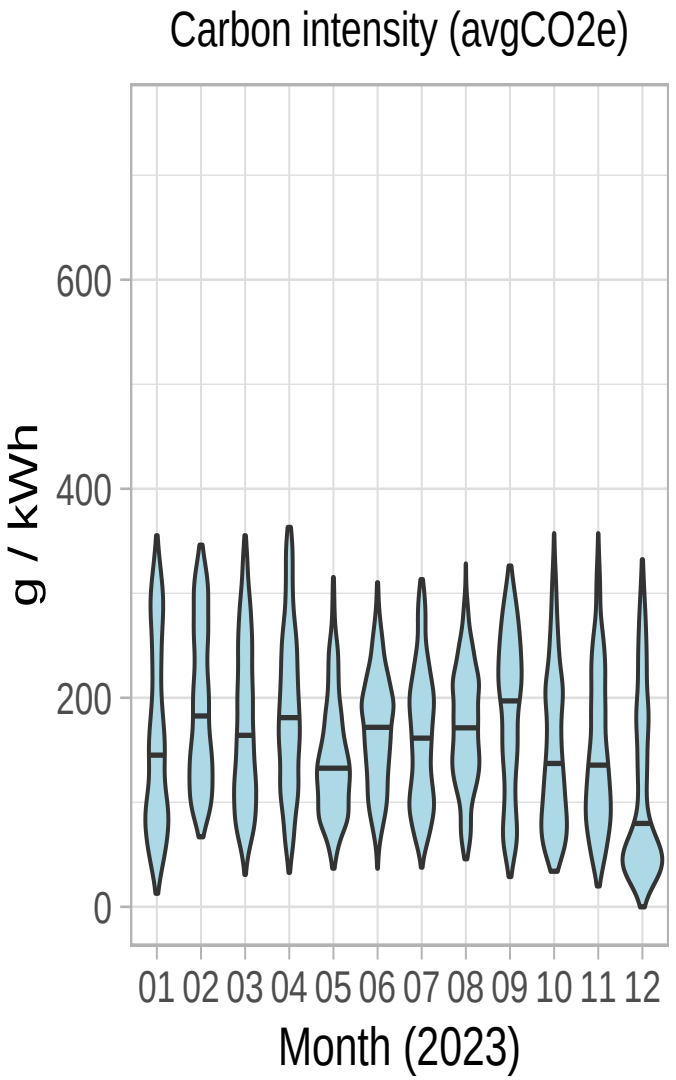
Carbon and price profiles: Germany (2023)

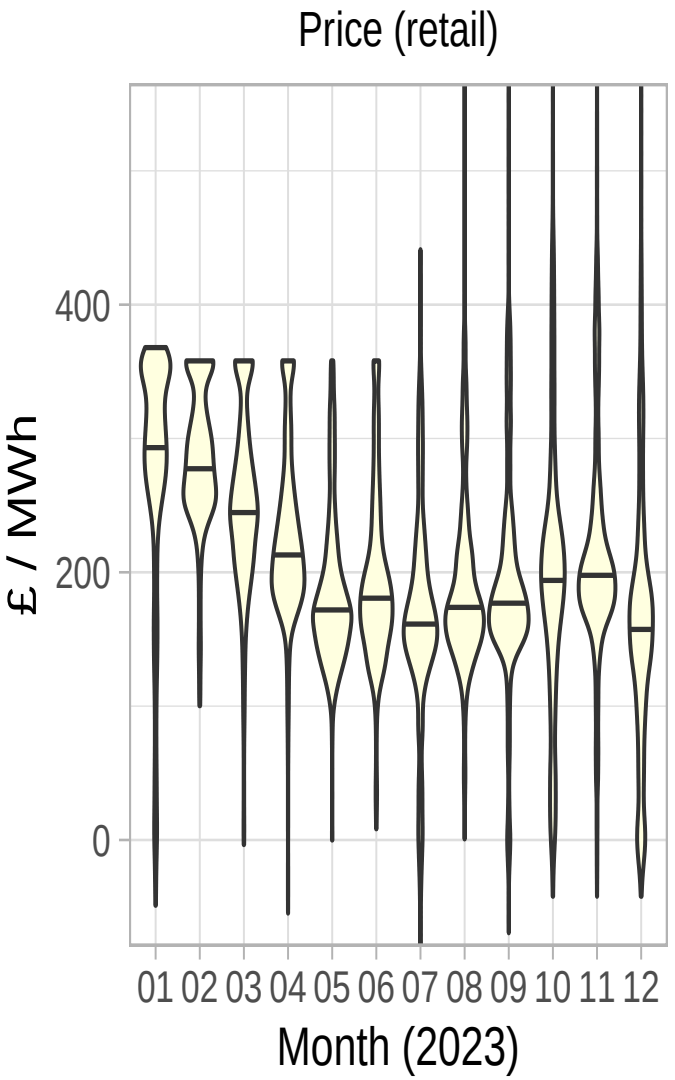


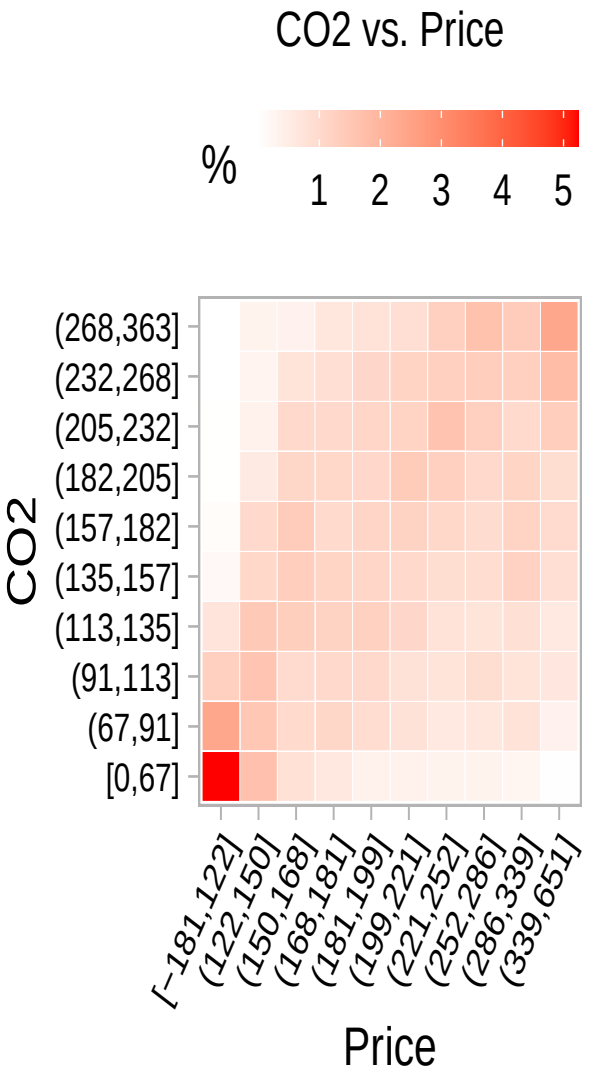




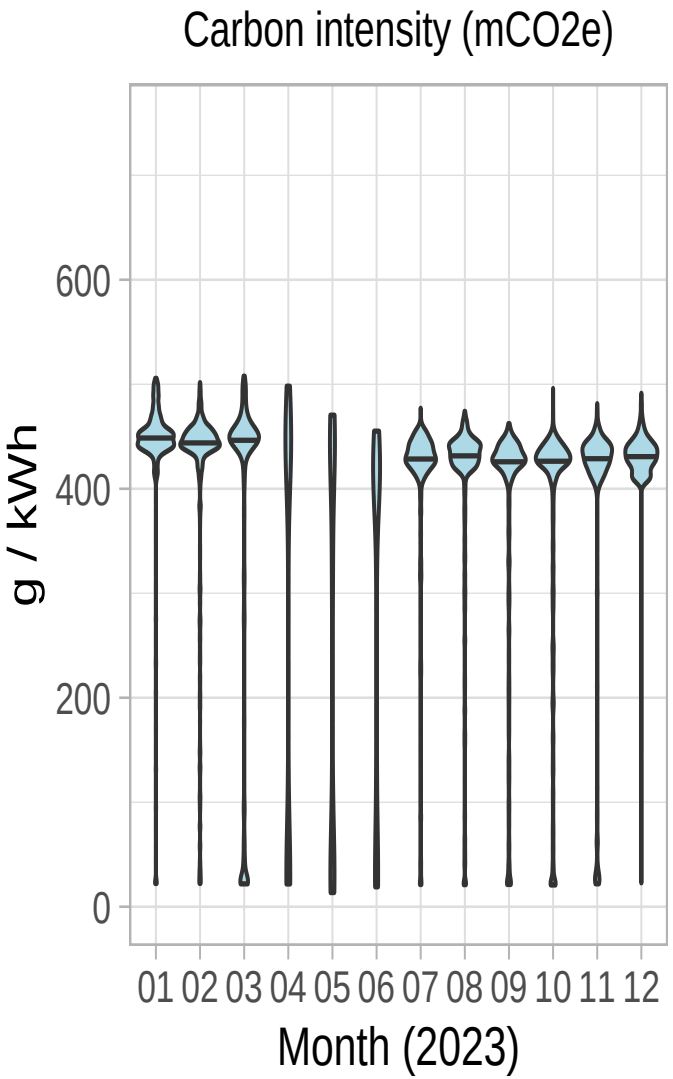
Carbon and price profiles: London (2023)

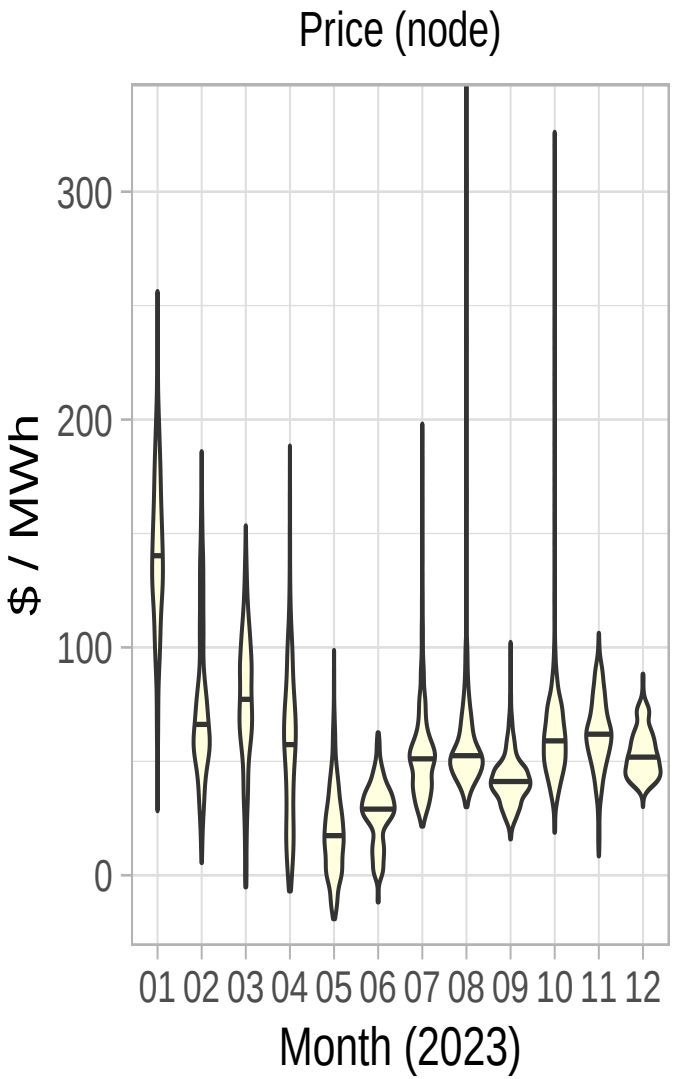


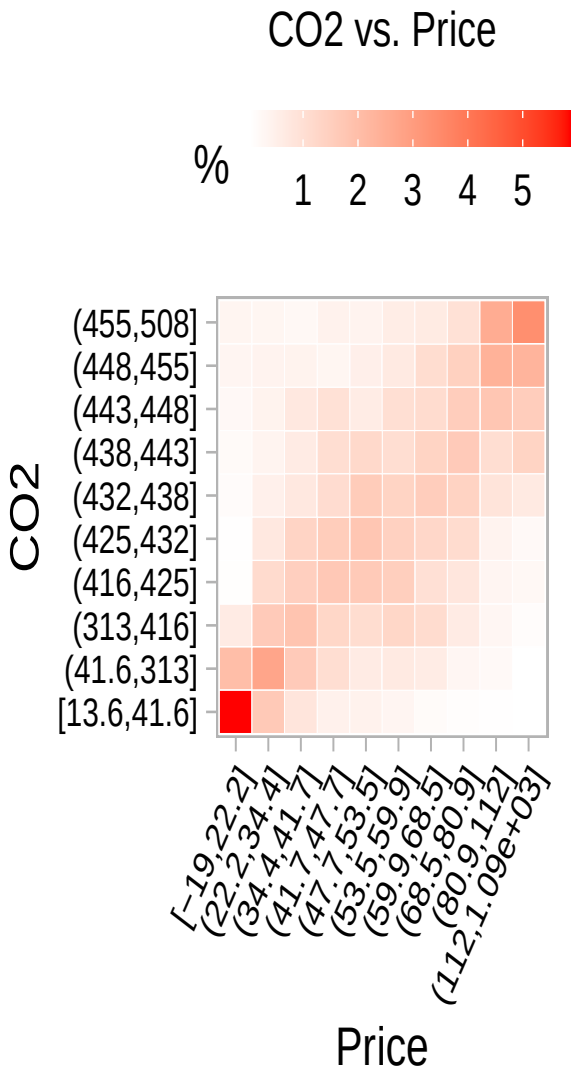




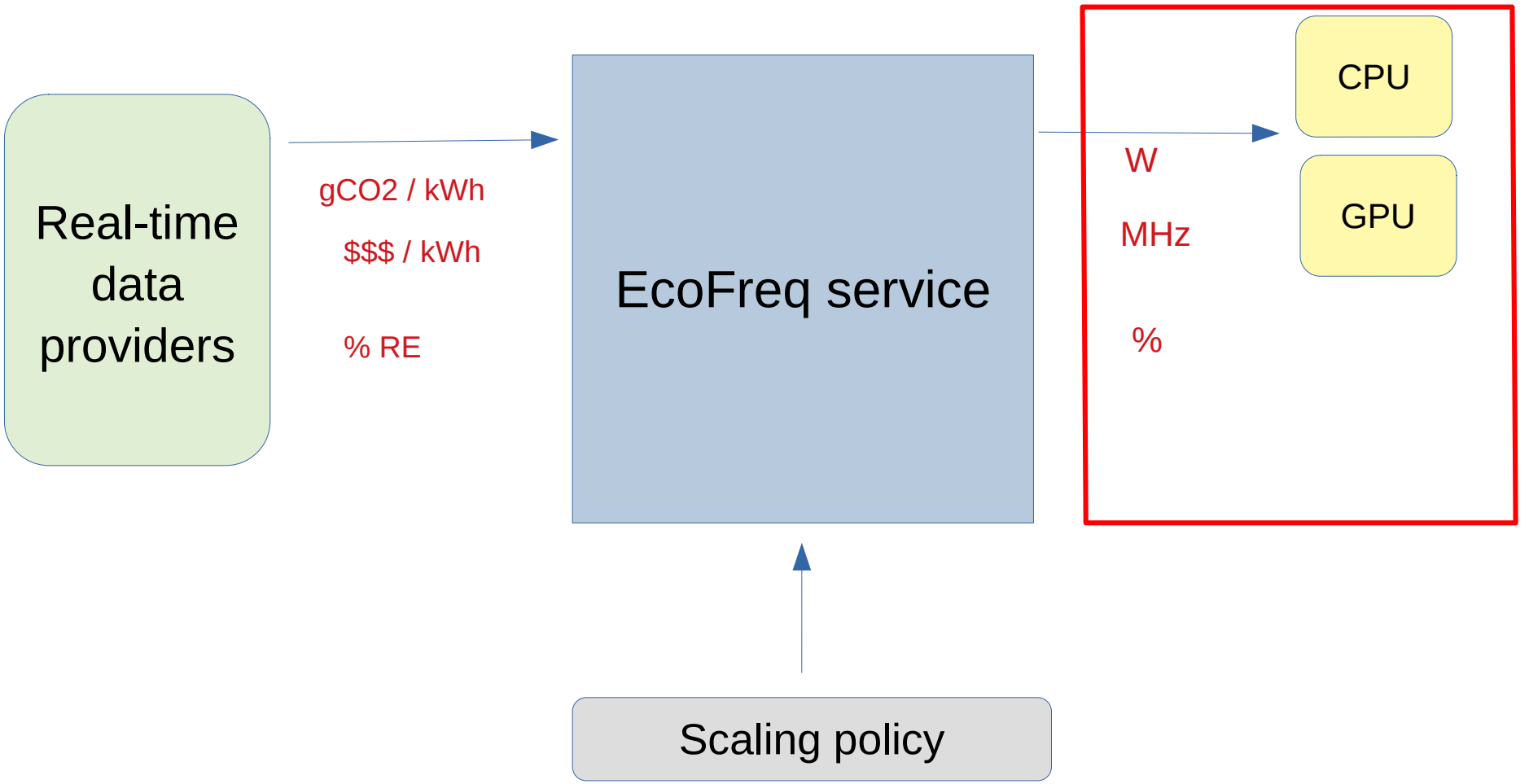
Carbon and price profiles: N. California (2023)







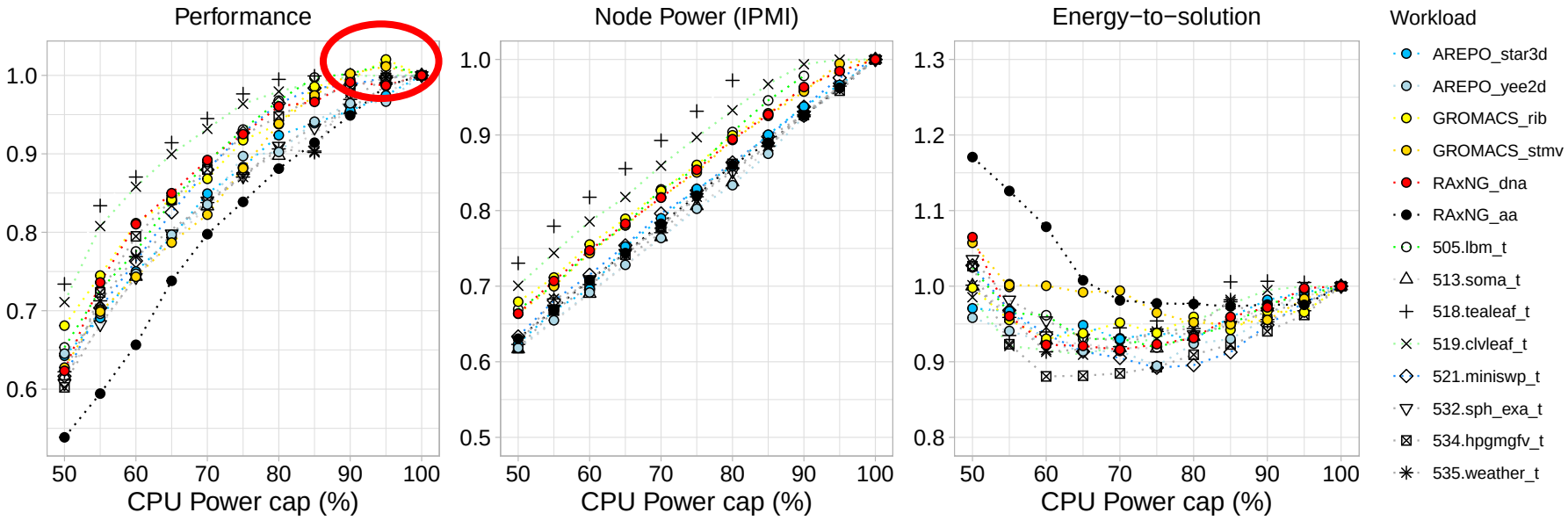
Dynamic carbon-aware power scaling



Experimental Setup

- Energy data
  - Historical data from 2023
- Hardware
  - 2x `Intel Xeon CascadeLake`
  - 2x `Intel Xeon IceLake` + 4x `NVIDIA A40`
- Software
  - `SpecHPC 2021`
  - + real-world tools: `Arepo`, `Gromacs`, `RAxML-NG`
  - 14 workloads (total)

Power profiles: RAPL powercap - CPU



System: `Xeon Platinum 8260, Cascade Lake`, 48C, 768 GB RAM

Scaling policy

- Germany & North California

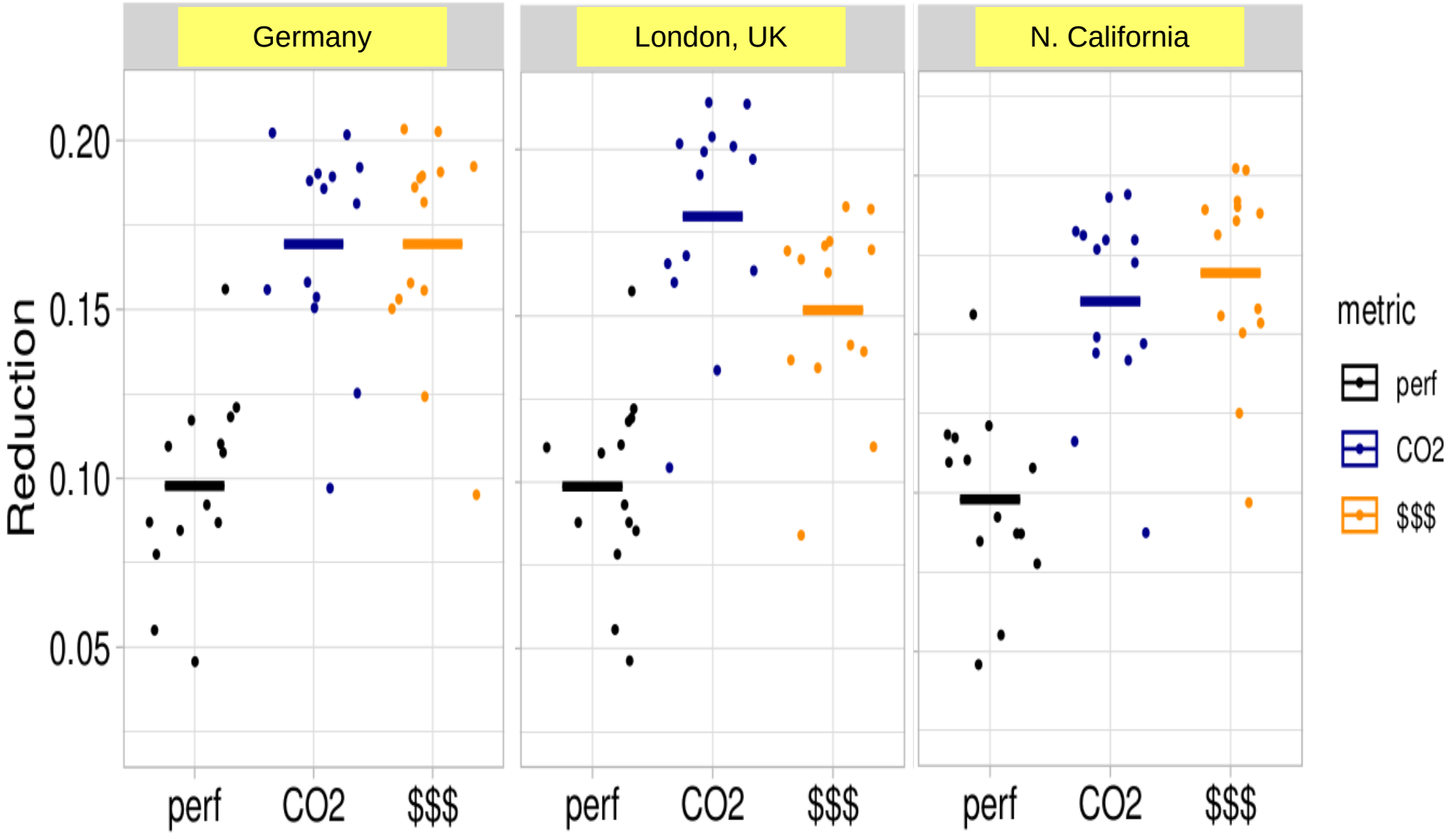
$$P_{lim}(CI) = \begin{cases} 100\%\ \text{TDP} & \text{if } 0\% < CI <= 33\% \\ 80\%\ \text{TDP} & \text{if } 33\% < CI <= 66\% \\ 60\%\ \text{TDP} & \text{if } 66\% < CI <= 100\% \end{cases}$$

- London

$$P_{lim}(CI) = \begin{cases} 100\%\ \text{TDP} & \text{if } CI \text{ in } ['low', 'very\ low'] \\ 80\%\ \text{TDP} & \text{if } CI \text{ in } ['moderate'] \\ 60\%\ \text{TDP} & \text{if } CI \text{ in } ['high', 'very\ high'] \end{cases}$$

TDP = Thermal Design Power
CI = Carbon Intensity

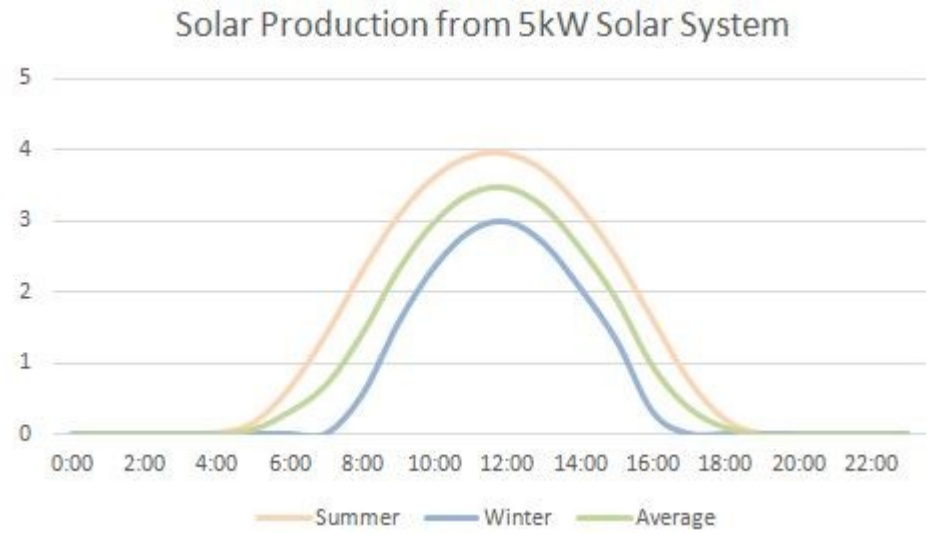
Evaluation



2023 electricity market data; 3-step scaling policy (100%/80%/60%); 14 HPC workloads, Xeon Cascade Lake

Features & Future Work

- Features & Setup
  - Specify Power Policy
  - Specify Electricity Data Provider
  - Put hardware to sleep on idle
- Future Work
  - Adapt to PV systems that are using net-metering



Contact

Oleksiy Kozlov
Staff scientist, HITS gGmbH
alexey.kozlov@h-its.org
https://cme.h-its.org/exelixis

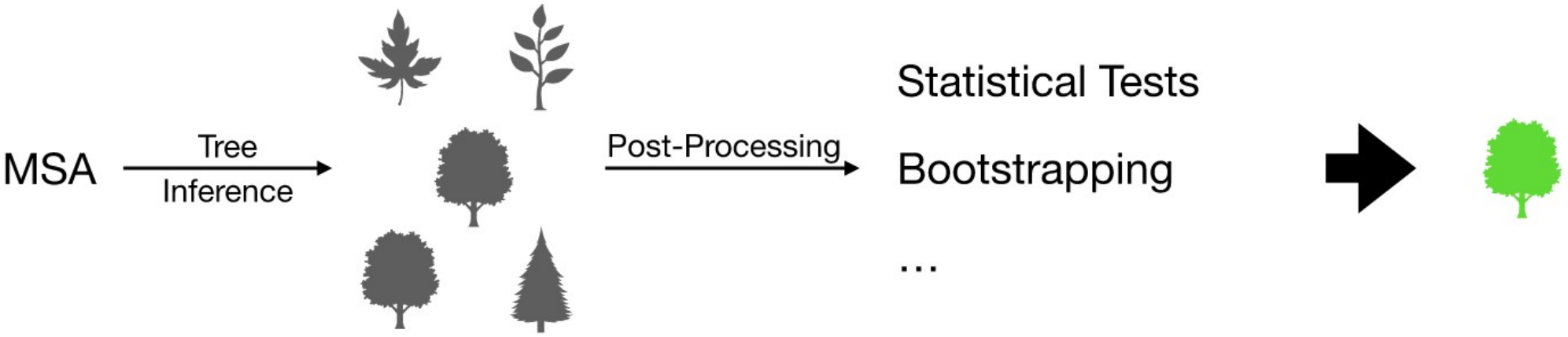
HITS

- Code: https://github.com/amkozlov/eco-freq
- Paper: https://ieeexplore.ieee.org/document/10528928

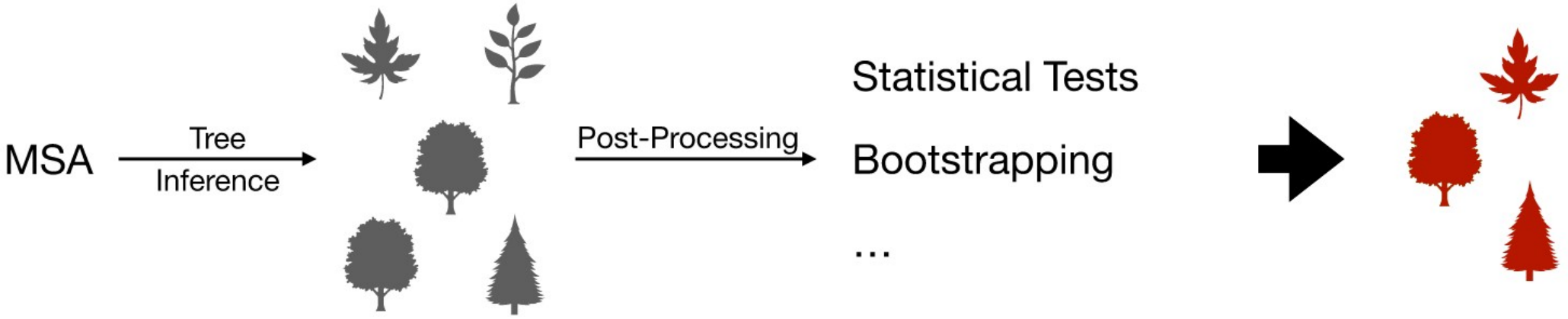
Thank you for your attention

Listaros village, Crete

Easy


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...

Difficult


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...

What does Difficulty mean?

Difficulty = ruggedness of the tree space

- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

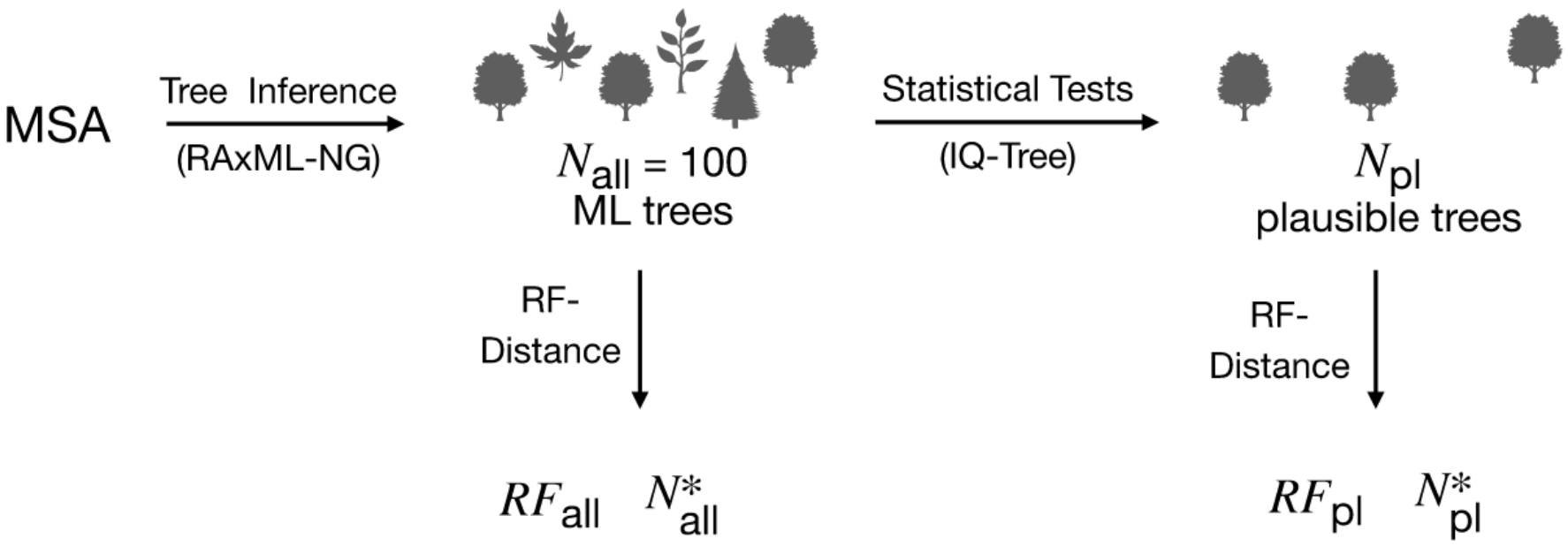
Predicting Difficulty with `Pythia`

- `Pythia` = Boosted Tree Regressor
- Supervised Regression Task
  - Predict difficulty between **0** (easy) and **1** (difficult)
  - Ground truth difficulty as training target based on 100 distinct Maximum Likelihood tree inferences
- Initially trained on 4K empirical MSAs
  - Mean absolute error: 2.5%

`Pythia` developments

- New release (May 19, 2023)
  - Trained on 12K datasets
    - 11,108 DNA MSAs
    - 979 Protein MSAs
    - 460 Morphological MSAs
  - Two new features
  - Improved accuracy
    - Mean absolute error: 0.07 (previously 0.09)
    - Mean absolute percentage error: 1.7% (previously 2.5%)
- Using `Pyhtia`
  - See next slides

Definition of Difficulty



$$\text{difficulty(MSA)} = \frac{1}{5} \cdot \left[RF_{all} + \frac{N^*_{all}}{N_{all}} + RF_{pl} + \frac{N^*_{pl}}{N_{pl}} + \left(1 - \frac{N_{pl}}{N_{all}} \right) \right]$$

Prediction Features

- Eight Features
  - 4 MSA attributes
    - Sites-over-taxa
    - patterns-over-taxa
    - **%** gaps
    - **%** invariant sites
  - 2 MSA information metrics
    - Shannon entropy
    - Bollback multinomial test statistic
  - 2 Parsimony-tree-based features
    - Infer 100 parsimony trees
      - → average RF-Distance
      - → **%** unique topologies

Using `Pythia`

- **Prior** to tree inference
  - → determine analysis & post-analysis setup
  - → adjust/modify MSA
  - → explore data filtering & assembly strategies
  - → adjust user expectations about data

Pythia