PROBABILISTIC PROGRAMMING IN PYRO

SWAPNEEL MEHTA, NYU CDS

ABOUT ME

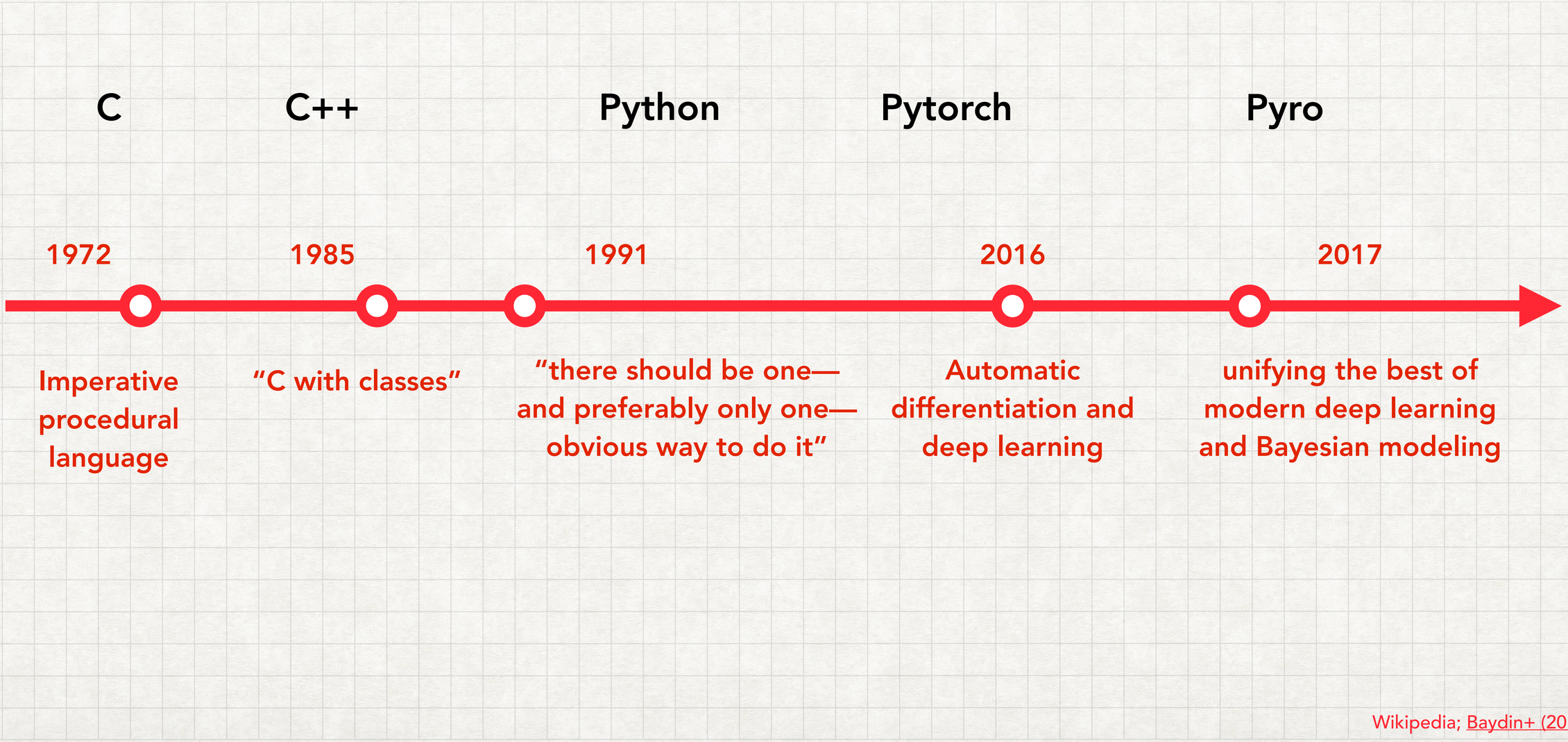
- Data Science Ph.D. student at NYU
- Probabilistic programming for simulation-based inference using social networks @ CSMAP
- High-energy physics and deep learning @ CERN
- Graph neural network-based video recommender systems @ Adobe

NYU Center for Data Science

CERN

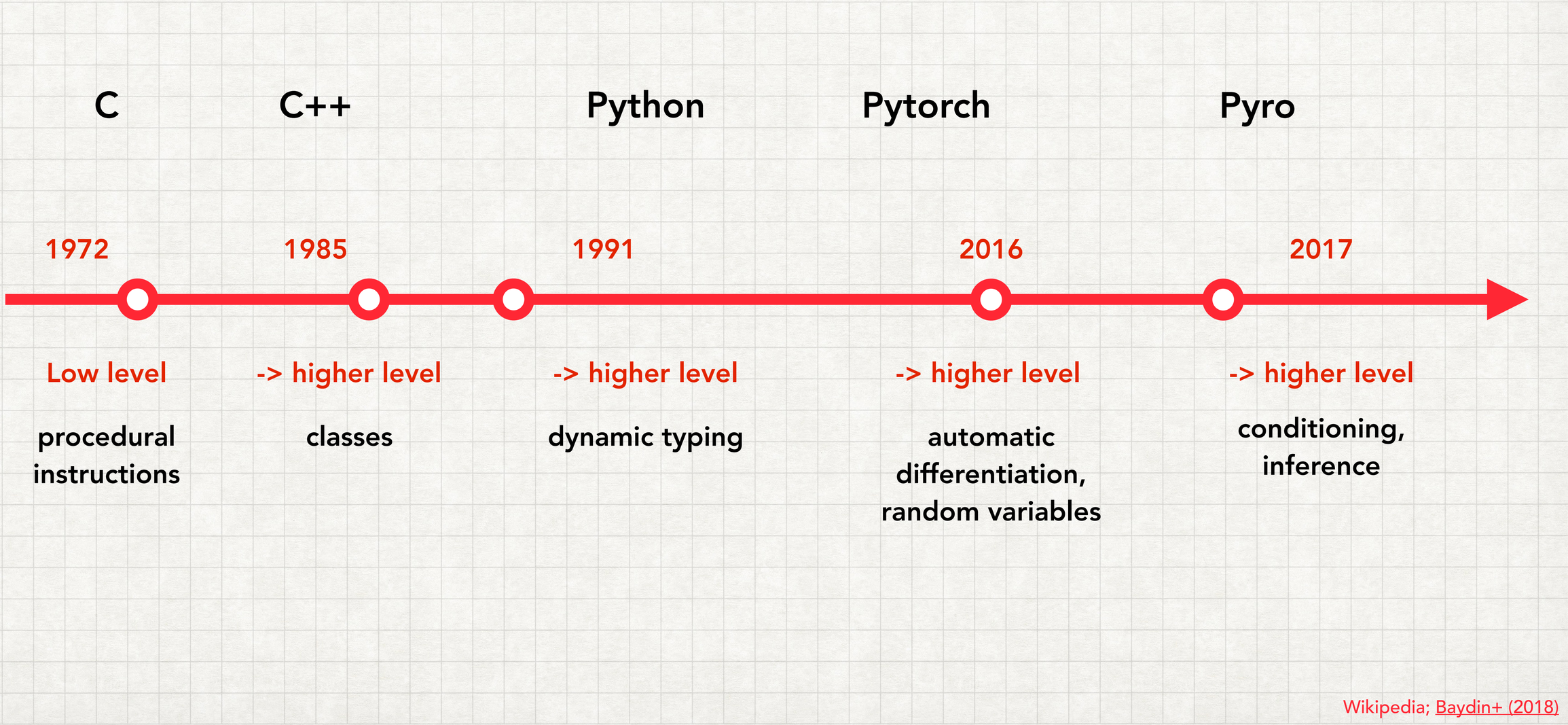
Adobe®

A (BIASED) HISTORY OF PL



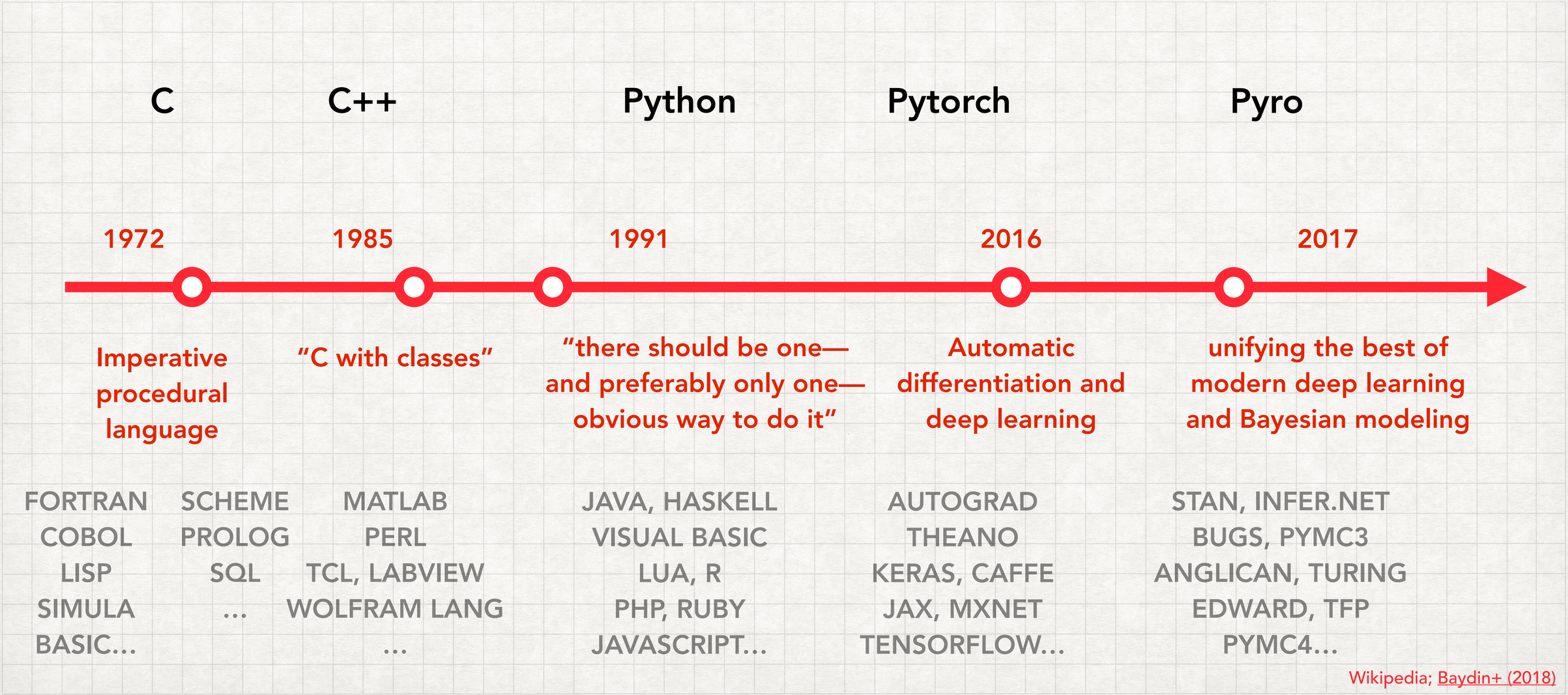
Wikipedia; Baydin+ (2018)

A (BIASED) HISTORY OF PL



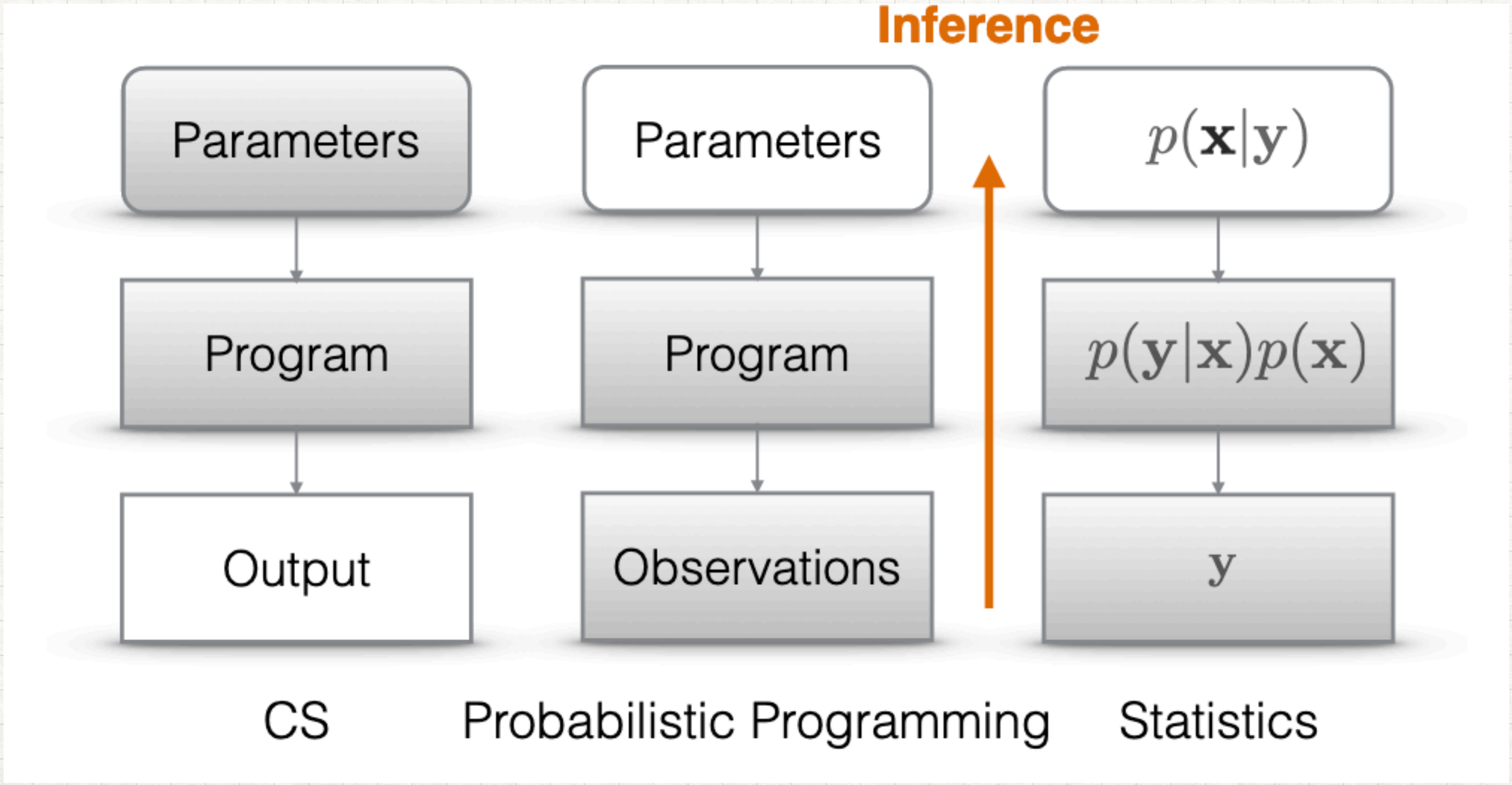
Wikipedia; Baydin+ (2018)

A (LESS-BIASED) HISTORY OF PL



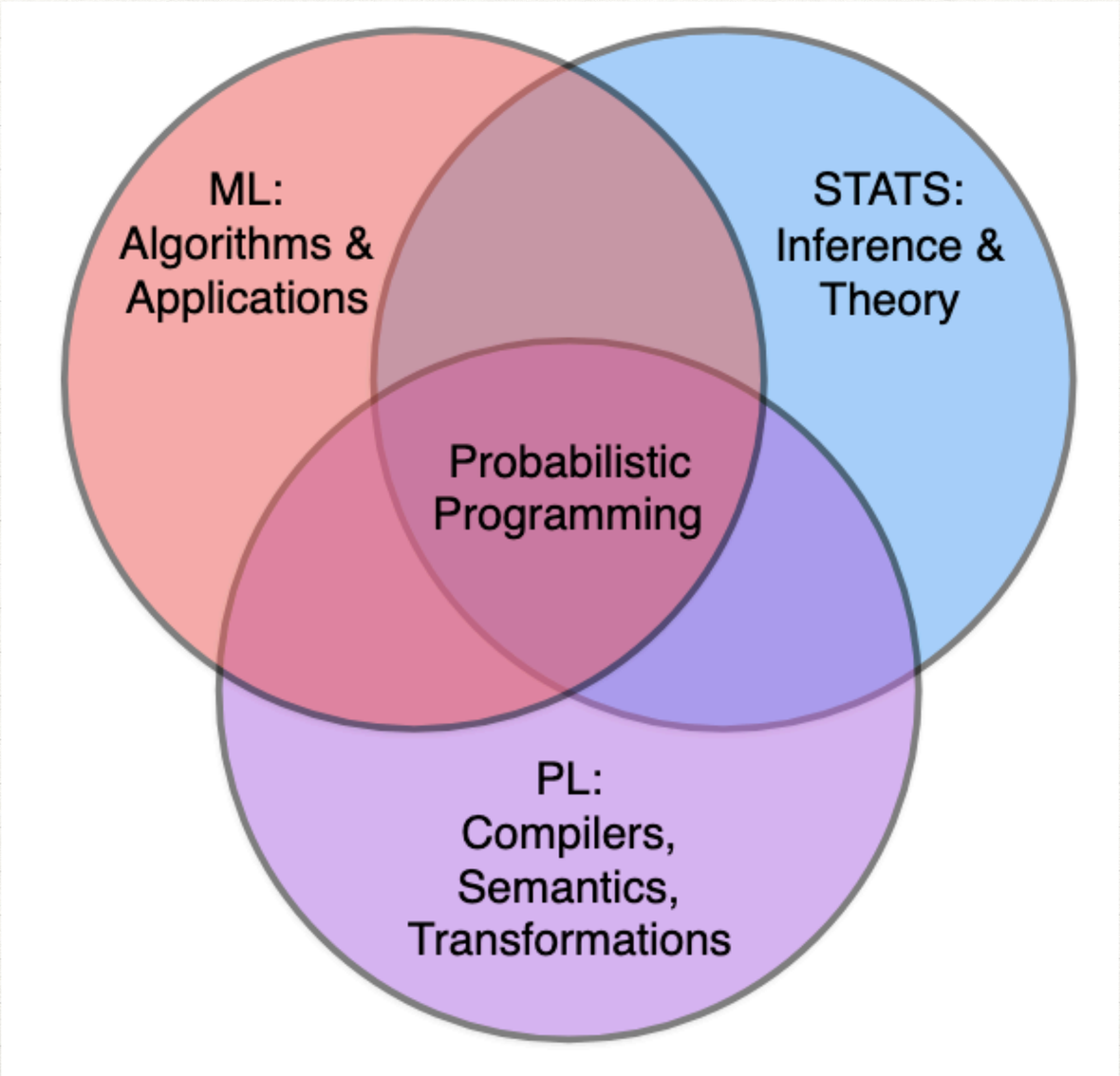
Wikipedia; Baydin+ (2018)

PROBABILISTIC PROGRAMMING



Frank Wood (2015)

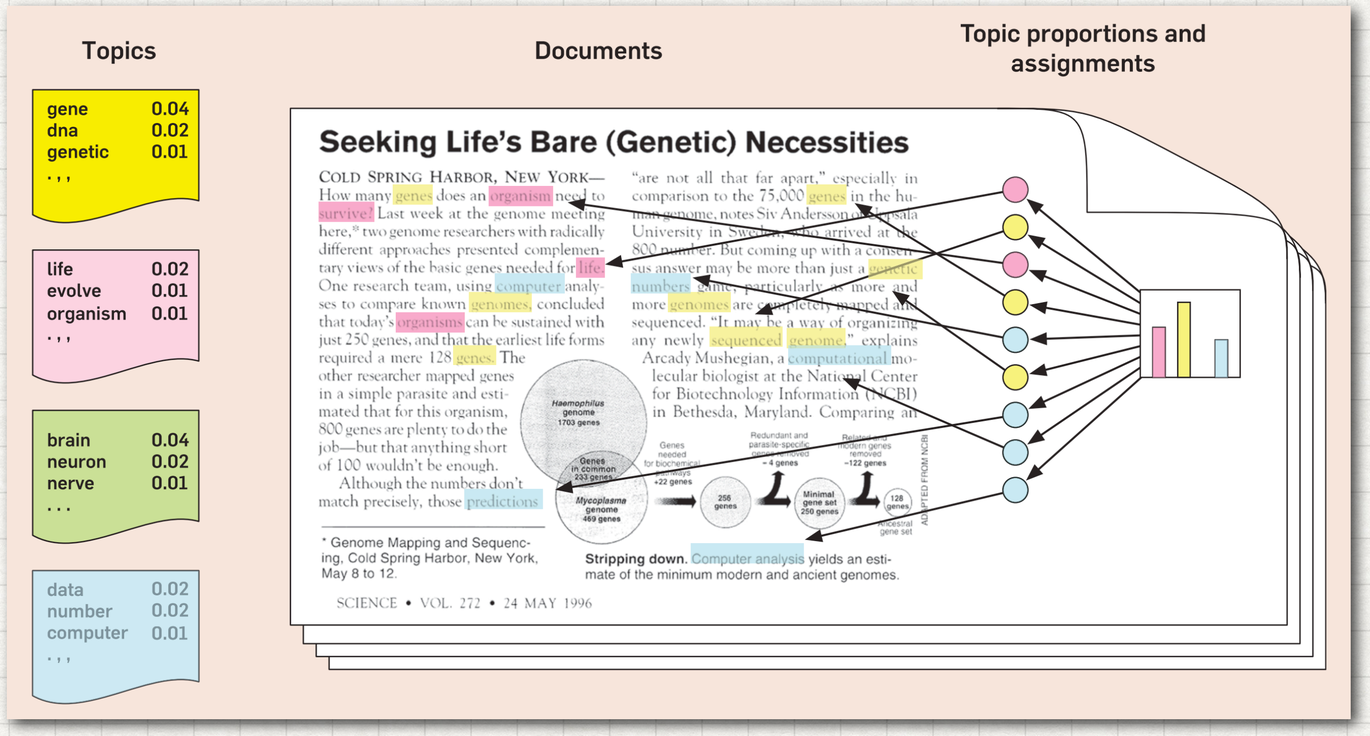
PROBABILISTIC PROGRAMMING



Frank Wood (2015)

PROBABILISTIC PROGRAMMING

- Scaling Bayesian inference to massive datasets has been hard!
- This has been an active area of research with encouraging results for specific problems
- A decade ago, custom inference algorithms were all the rage, but now we want general-purpose algorithms
- Eg. If someone asks you to write a custom file parser in Python, you're not gonna be happy :(

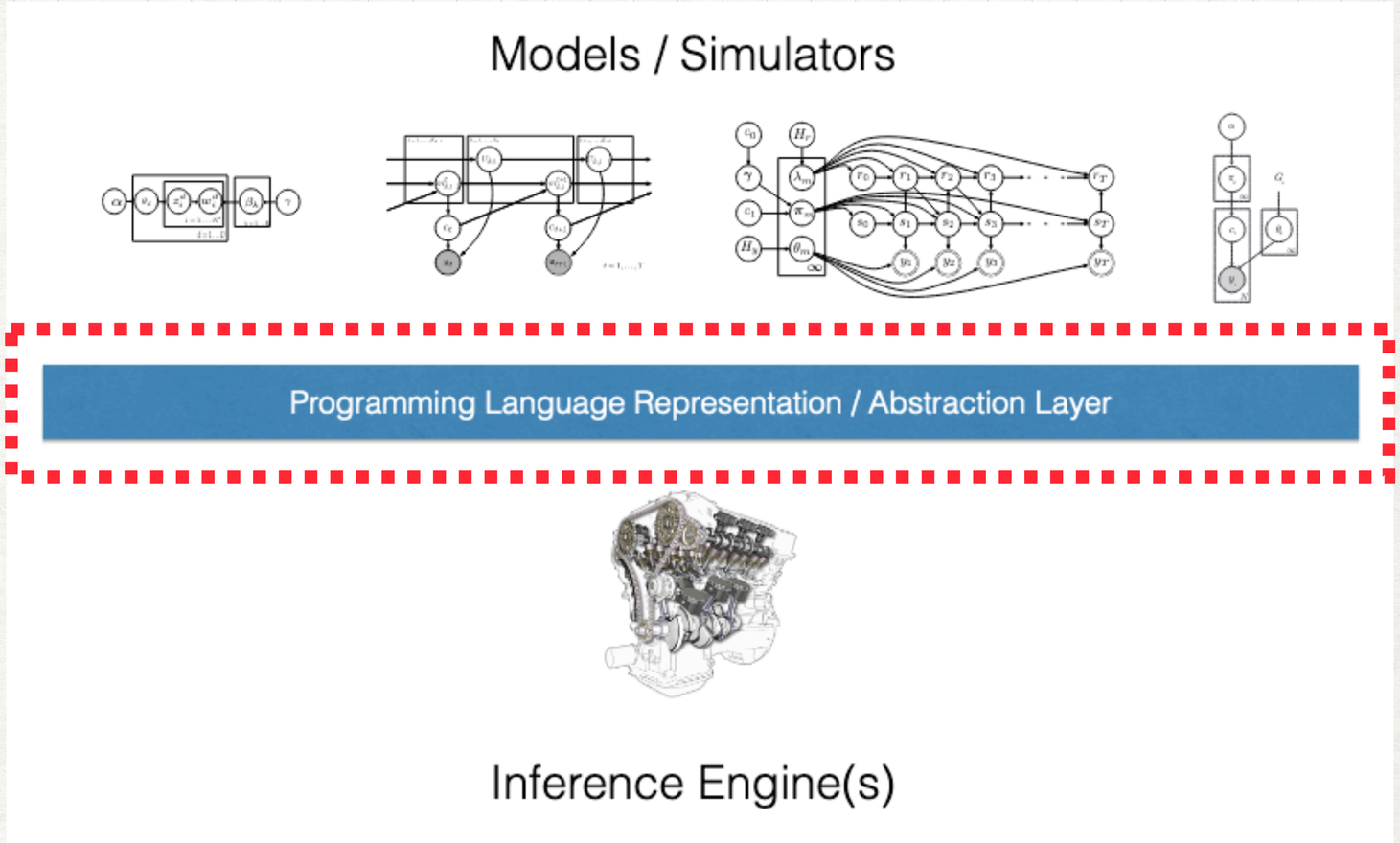


Topic Modeling

Srivastava+ (2017); Blei (2012)

PROBABILISTIC PROGRAMMING

- Modern PPLs ‘commodify’ inference
- Given any forward model, we can use the same inference algorithm!
- The emphasis is on deep ‘universal’ probabilistic programming
- This does for Bayesian inference what automatic differentiation did for deep learning



Ranganath+, 2013; Frank Wood (2015)

SIMULATION-BASED INFERENCE

- Step 1: Propose a forward generative model for a process
- Step 2: Collect real-world observations of the data-generating process being modeled
- Step 3: Use the observed dataset to infer the posterior distributions of the parameters of interest
- *Step 4: Use the fitted model for predictions and forecasting*

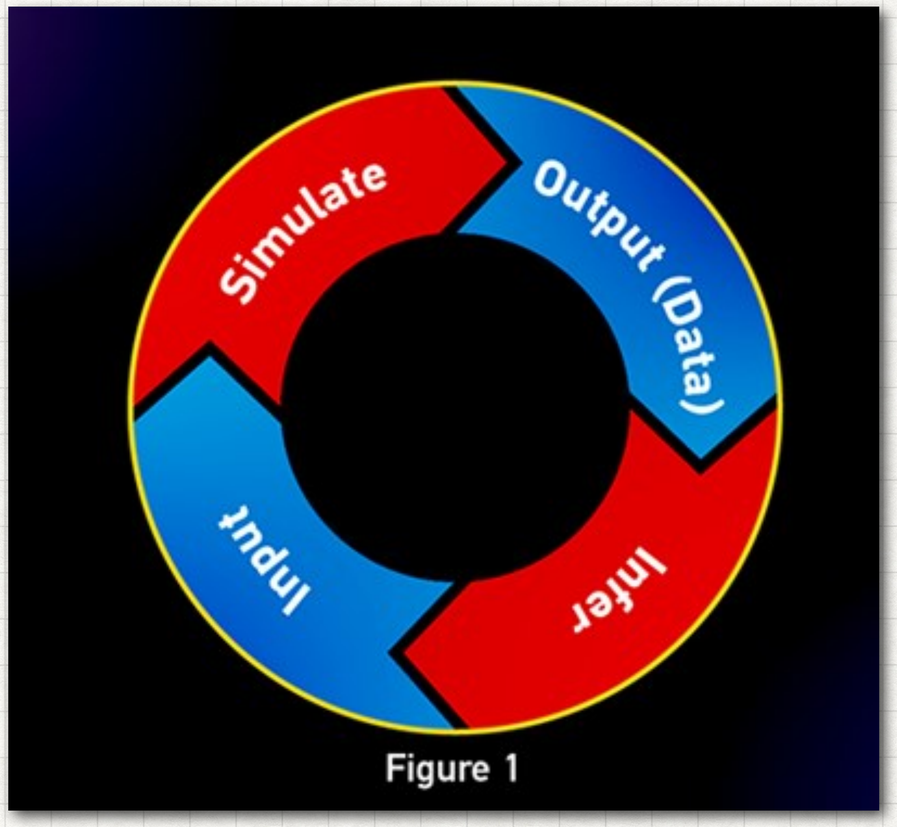


Figure 1

WorldQuant (2019)

LET’S DO SOME MATH.

BIO + PPL

- Bayesian Inference of Gene Expression (Jimenez-Jimenez+, 2021) uses Stan to infer params. of probabilistic models for gene expression at scale
- Inferring the parameters of arbitrary statistical phylogenetic models using PPLs (Ronquist+, 2021)
- (Dirmeier+, 2019) use PyMC3 for inferring causality from perturbation screening data via structured hierarchical models
- ProbRules (Grob+, 2019) uses PPLs for integrating multi-scale interaction data (temporal, spatial) with feedback
- Other applications, general directions (Shiffman, 2019)

Merrell+, 2020

RESOURCES

- ProbAI School 2021 (resources)
- Frank Wood's NeurIPS 2015 Presentation (slides)
- Pyro Slides by Taku Yoshioka
- My ProbProg Playlist (link)
- Coursework on ProbProg (UBC, links to others)
- Textbooks (Intro. To ProbProg, Foundations of ProbProg)