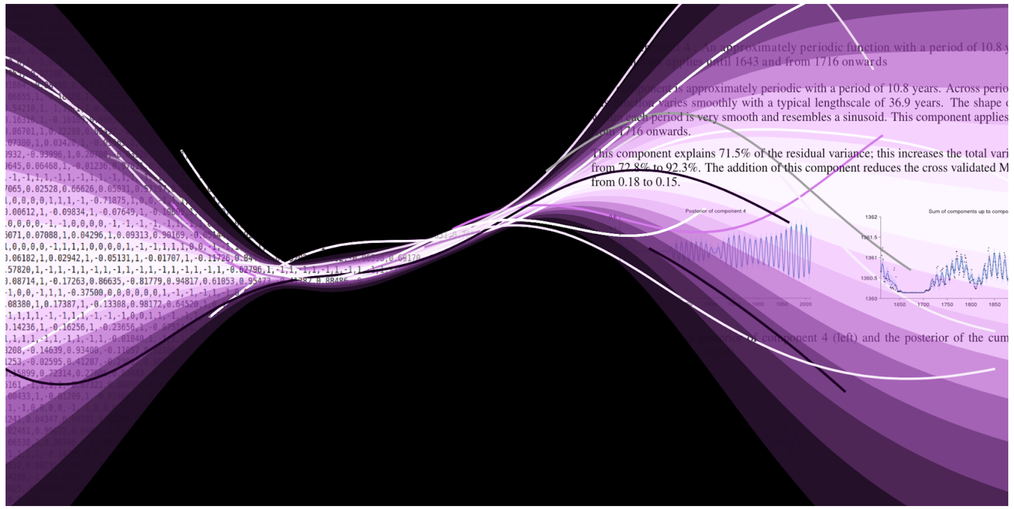
JSC370 and JSC470: Data Science II and III



David Duvenaud

January 2021

Intro Lecture

Goals of course

- Improve your conceptual understanding of reasoning, inference, statistics, and the capabilities of our current models.
- Set you up for your next steps professionally.
- Get to know each other
- Have fun and stay sane.

Focus of course

- Understanding what can and can't be inferred from data
- Use of collaborative tools
- Use of more advanced models (compared to JSC270 / JSC370)
- Communicating the results and possible next steps (decision support)
- Have fun!

Course Approach

- Guest lectures and projects designed around real data and problems
- Contrast to JSC270 (I think):
  - More focus on interfacing with larger team / organization
  - Fancier models
  - Setting you up with a portfolio / nice github profile
- One assignment on recent papers / developments
- https://padlet.com/duvenaud/c8nh6420vjawvoli

Evaluation

- Assignment 1: 18%
- Assignment 2: 18%
- Assignment 3: 18%
- Assignment 4: 18%
- Assignment 5: 18%
- Paper report: 10%

Schedule

- Tuesdays 3-5pm:
  - Alternating lectures and guest lectures.
- Thursdays 3-5pm:
  - Alternating tutorials or assignment presentations.

Guest speakers and projects

- Jan 19: Ben Allison, Principal Machine Learning Scientist at Amazon
  - Project on recommender systems
  - Amazon starting a machine learning group in Toronto

Guest speakers and projects

- Feb 2: Farah Bastien, Manager of the data science team at Maple Leaf Sports & Entertainment
  - (owns the Leafs & Raptors)
  - + team member
- Project analyzing game data

Guest speakers and projects

- Feb 23: Wanying Zhao, Impact Measurement and Evaluation Analyst at Ontario Trillium Foundation

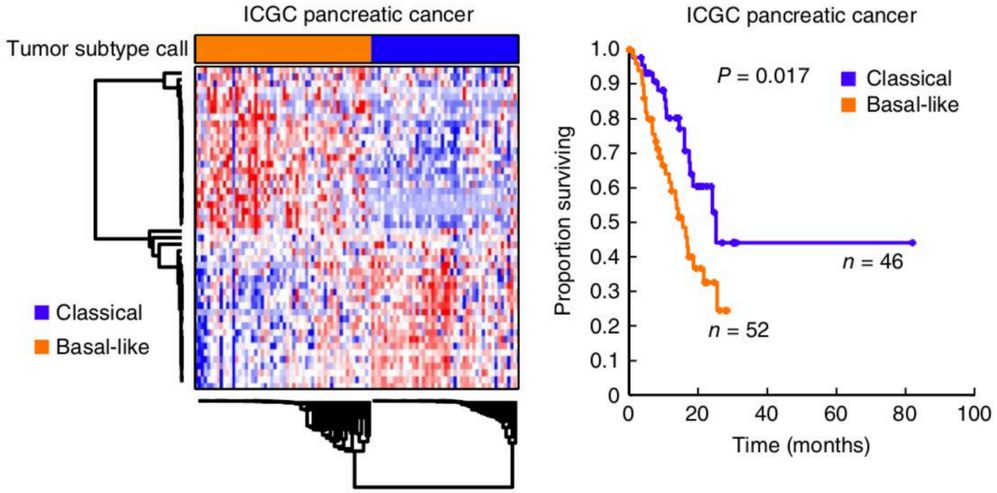
- Project related to measuring impact of charitable interventions
    - Confounding galore

Ontario
Trillium Foundation

Guest speakers and projects

- March 23: Dr. Robert Grant, Princess Margaret Cancer Centre, Oncology Division

- Project based on clustering / subtyping genomics data
    - Custom project for 470 folks



Last assignment: Paper presentations

- Just a graded journal club
- Need to keep up with new tools
- E.g. promising new model classes, case studies of model rollouts, unexpected impacts or problems, foundational issues
- I will provide a list, but you're free to propose one yourself

Tools of the trade: Git

- Version control is table stakes for industry, collaboration, your own sanity.
- Github demos add a lot to a resume.
- Assignments will be released, and due, through Github classroom.
- Don't make assignment repos public until after course is over!

GitHub

Tools of the trade: Python

- Suggested for predictive models: Jax, PyTorch
- Gotchas: need to learn both Python frameworks on top. Bad error messages.
- Will provide starter code / skeleton for at least most of the assignments
- Thursday: intro + tutorial
- Can use Jupyter / Colab for everything

Why not R?

- Less employer demand
- No reverse-mode autodiff!
  - Radford Neal is working on this.
- Other reasons:
  - Me and Harsh don't know much R.
  - Students write slow nested loops
  - Limited GPU support, limited composability.
- Can use R if you want, but we can't help you with it as well.

Dex: a typed array language built for speed

```
def map (f : a->b) (xs : n=>a) : n=>b =
  for i. f x.i
```

Flexibility

- Ragged and sparse arrays
- Algebraic data types (e.g. `Value|NaN|Missing`)

Correctness

- Dependent types for compile-time debugging (e.g. shape checking)
- Composable, zero-cost abstractions (e.g. run on any vector space)

Performance

- Fast nested loops + gradients (e.g. CTC loss)
- CPU, GPU, TPU backends, JAX interop

Ray tracer written in Dex
google-research.github.io/dex-lang/raytrace.html

Forum and Contact

- Discourse instance:
  - https://bb-2021-01.teach.cs.toronto.edu/c/jsc370
  - Should be able to login with your CS instances, or utoronto.ca ids (if they are the same as your CS ids). Please let me/TAs know if you can't log on.
- My email: duvenaud@cs.toronto.edu
- Harsh's email: harsh.panchal@mail.utoronto.ca
  - Please put "JSC370" or "JSC470" in subject line.

Collaboration

- Can collaborate / discuss with up to 2 others, even for individual assignments, but:
  - You must do the assignment on your own! Hand in a unique assignment.
  - You must list your collaborators on your assignment.

Questions regarding course setup?

Getting to know each other

- Will of course write letters of rec for jobs or grad school.
    - Participation gives me something to write about!
    - Forum participation leaves a paper trail.
- Lots of scope for custom projects, proofs of concept that could lead to research projects.

Getting to know each other

- TA: Harsh Panchal
  (pronounced "Hersh")
- Mech. Eng Master's Student
- Veteran of many data science internships + projects

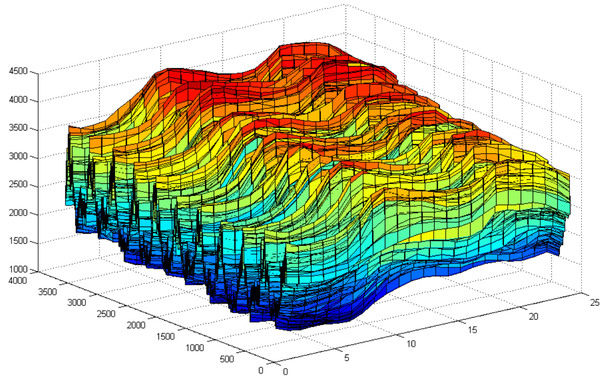
My teaching background

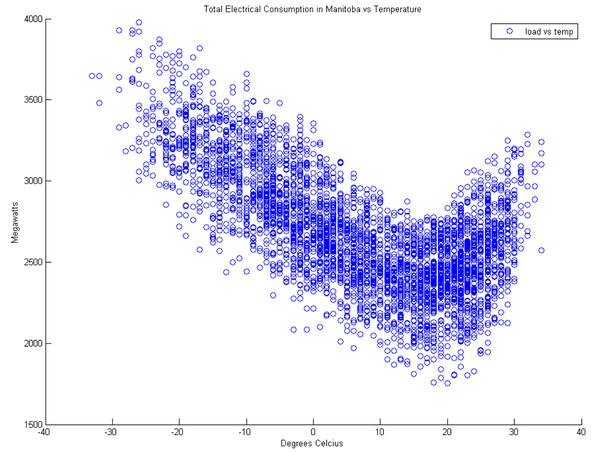
- I'm 51% CS and 49% stats
- Usually teach 4th year machine learning
  - My usual failure mode: going too fast / assuming too much background
  - The student is always right! Please speak up if anything is wrong.
- Building on Nathan Taback's course

My data science background

- Mostly from co-founding an energy prediction startup 14 years ago
  - Writing C# to scrape public electric utility data
  - Writing crappy models in MATLAB, making mistakes, running code I didn't understand at the time
  - Presenting results / selling consulting to power companies
  - Calling wind farm operators to ask what special values meant
- Used pandas + databases once or twice as a researcher

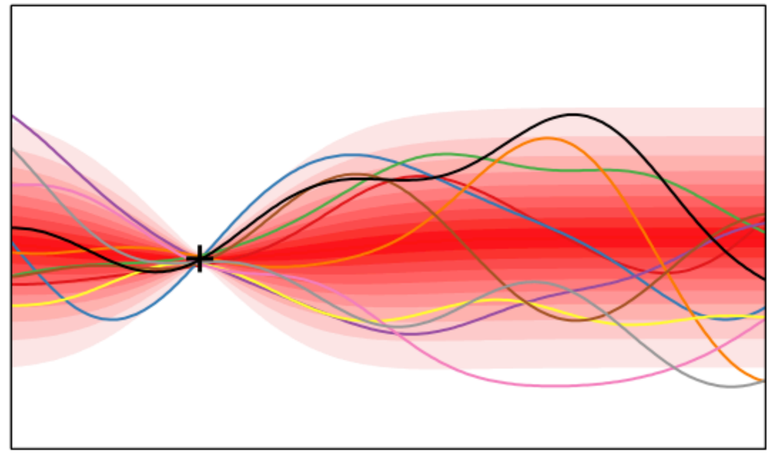
INVENIA

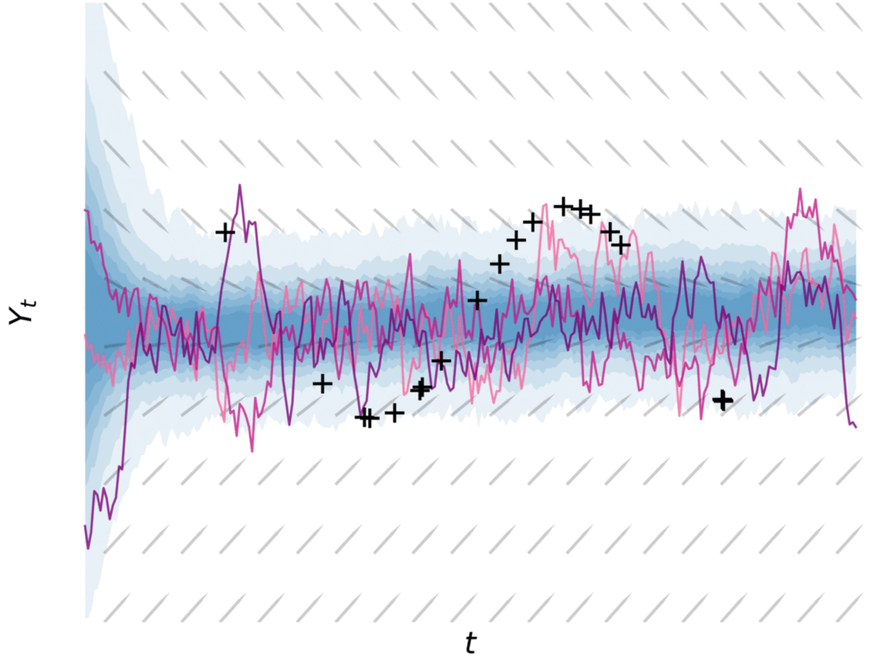




My research background

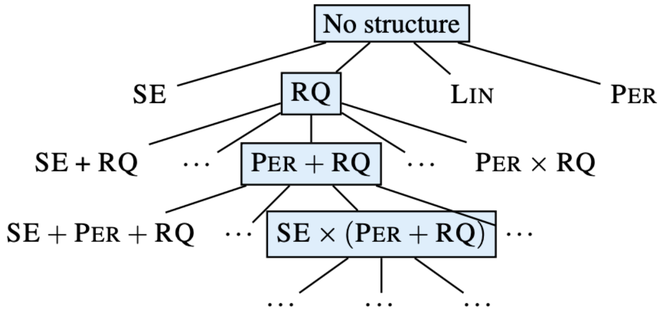
- Undergrad CS
- M.Sc. CS + Stats
- Ph.D. "Information Engineering", mostly Gaussian processes
- Current research: Probabilistic models, deep learning, continuous-time models.
    - Dealing with irregularly-sampled time series

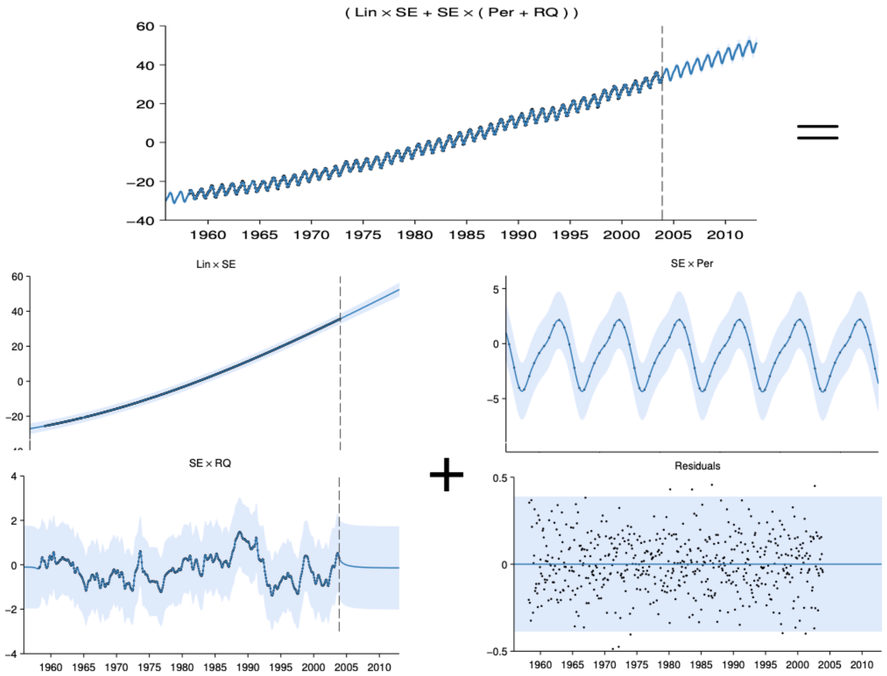




The Automatic Statistician

- Old project to search for a decomposable, interpretable model for time series data
- Roger Grosse did a similar project for vector / image data beforehand
- Idea is currently half-dead, neural architecture search is hot related area





Emphasis on Decision Support

- "Deterrence is the art of producing in the mind of the enemy... the fear to attack"

- Data science is the art of producing in the mind of the decision-maker... actionable understanding?

Proposed data gathering

Tukey: “The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data...”

- A report about limitations of existing data will usually be met with a question of what data *would* answer the question.
- Sometimes policies exist that avoid the need to identify the true answer.

Next steps

- Thursday: Tutorial on Python, Numpy, Pandas, Git, Jupyter / Colab
- Tuesday: Guest lecture by Ben Allison, and explanation of A1

- Any more requests?
  https://padlet.com/duvenaud/c8nh6420vjawvoli