Three Models for Learning Data Science

Soubhik (sho-bik) Barari

Research Methodologist / Data Scientist
NORC at the University of Chicago

Adjuct Assistant Professor
Columbia University

IC²S²

July 18th, 2024

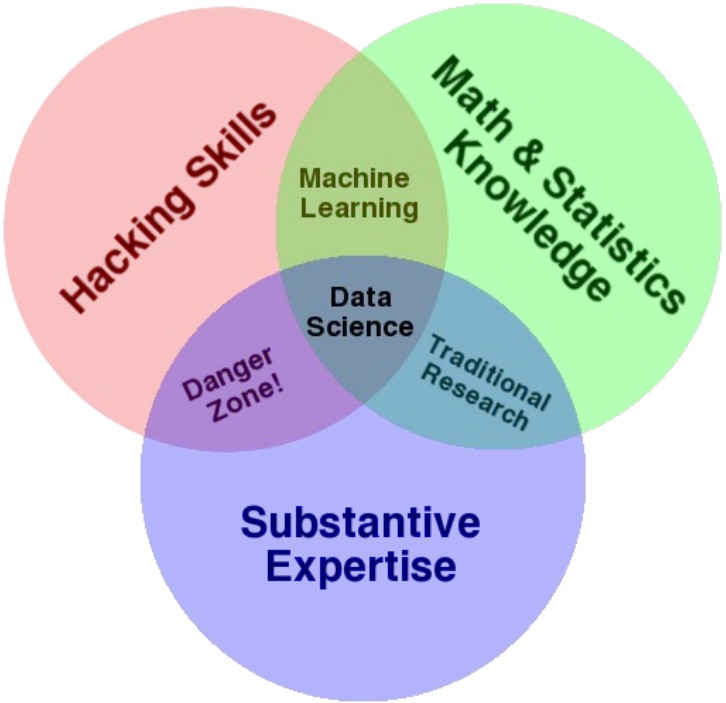
About me

I’m a computational social scientist.

- I work on data science, applied survey methodology, and public opinion research at **NORC at the University of Chicago**.
- Previously, I’ve worked in data science / applied research roles at **SurveyMonkey** and **Microsoft Research**.
- I’m a political scientist by training (**Ph.D. Harvard 2024**).

So you want to be a data scientist?

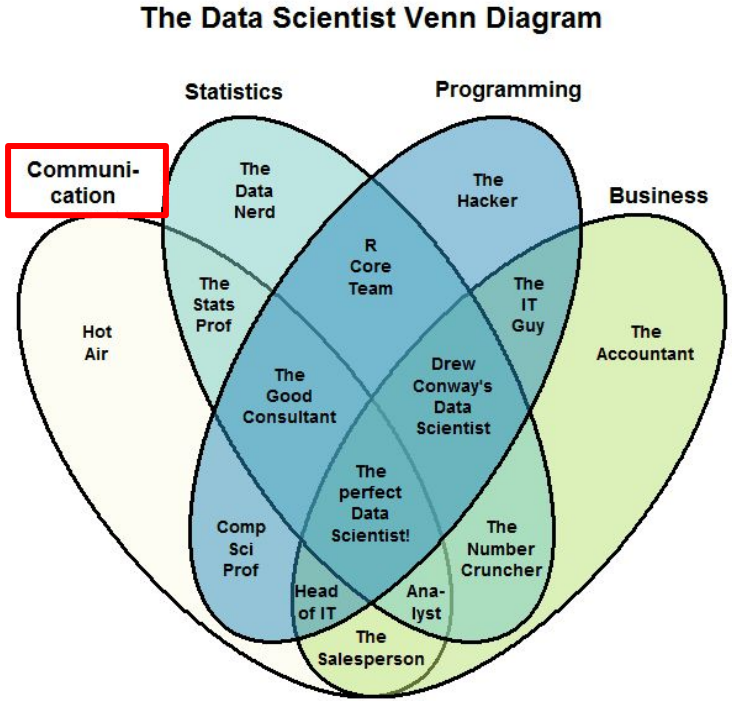
You'll need to get good at these three things:



Drew Conway (2010)

So you want to be a data scientist?

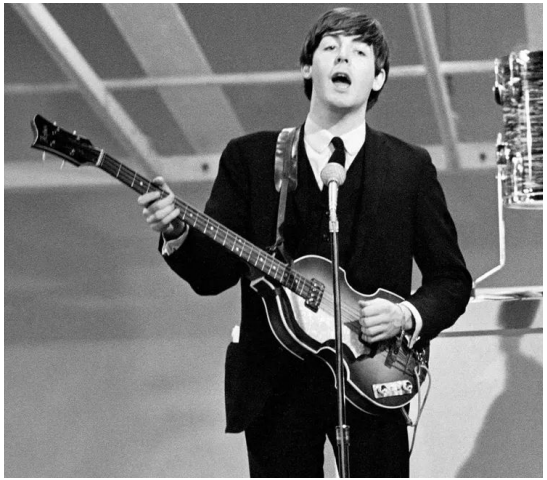
You'll need to get good at these three things (plus a **fourth**):



Stephen Kolassa (2014)

Three different models for learning from three different 20th century rock maestros

Jimi Hendrix



Paul McCartney (of The Beatles)

Brian May (of Queen)

Replication-Based Learning: Jimi Hendrix

Model:

Observation → Replication → "Riffing" → Adaptation

- Completely self-taught guitar by "reverse-engineering" songs from the radio.
- No formal music theory training / knowledge.
- Adapted right-handed guitar / playing style for his left-handedness.

Translation to data science:

- Reproduce your favorite social science study using public replication code (e.g. `dataverse.harvard.edu`).
- Fork an open-source package (e.g. `tidymodels`) and build an extension.
- Intern as a data scientist in "industry", learn what you need on the job.

Output-Based Learning: Paul McCartney

- Well-versed, but never played covers very well.
- More of an artist (substantive) than a musician (methodologist), i.e. driven by *lyrics/vocals*, not guitar!
- Notably, well-trained in piano from a young age.

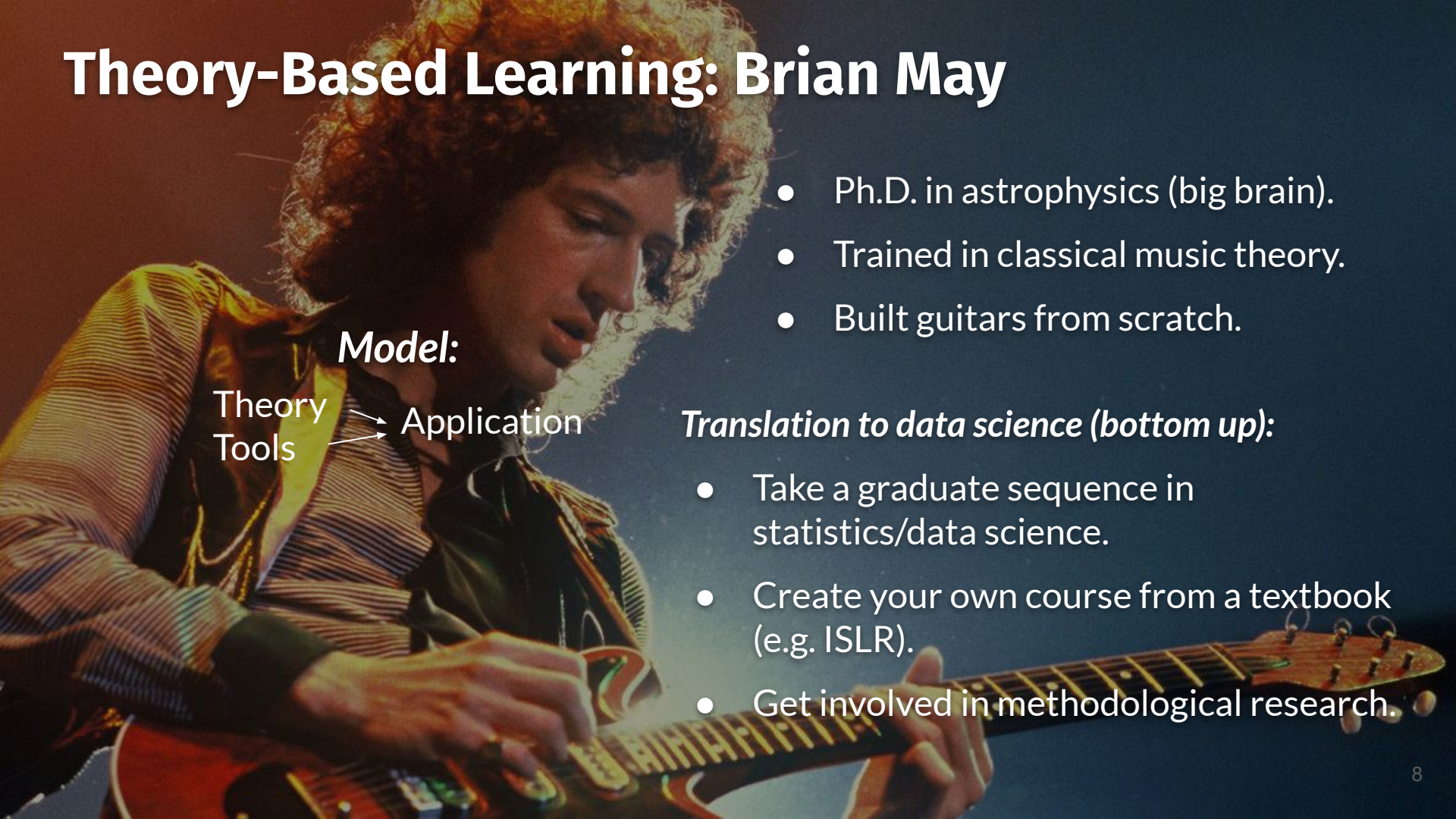
Model:

Inspiration → Output ↔ Learning

Translation to data science (top-down):

Have a "product" (e.g. a dashboard, graphic, blog post, paper, dissertation) in mind and figure out (e.g. stackexchange, textbooks) what you need at every step.

Theory-Based Learning: Brian May



- Ph.D. in astrophysics (big brain).
- Trained in classical music theory.
- Built guitars from scratch.

Model:

Theory
Tools
→ Application

Translation to data science (bottom up):

- Take a graduate sequence in statistics/data science.
- Create your own course from a textbook (e.g. ISLR).
- Get involved in methodological research.

Which model is right for me?

If you “think” like a…

- **Engineer / developer** (who is new to data science) → Be like Hendrix

- **Social scientist / domain expert** (with some coding/stats skills) → Be like Mccartney

- **Statistician / methodologist** (approaching a new domain) → Be like May

Are these models mutually exclusive? ***No!***

What these models (and maestros) have in common

- Frequent and intense **communication**:
  - Intense collaboration (i.e. being in a band!)
  - Intense interactions with other maestros (the “scene”).
  - Intense feedback mechanisms (both helpful and unhelpful).
- Baseline of **foundational knowledge** (i.e. the Brian May model).
- Development of **intuition**, rather than *just* knowledge or skills.
- Eventual **limitations**.
  - But see all of the above for how to overcome!

A few last reminders

- **Data science has many rules, but *there are no rules for learning data science*.**
  - But, how you learn will shape the kind of data scientist you are.

- **Know yourself and what "default" approach works for you.**
  - Recognize the limitations of singularly adopting one approach.

- **Find your community.**


THANK YOU
Let's keep the conversation going:
barari-soubhik@norc.org
More on specific skills to learn:
github.com/jaeyk/ic2s2-training-css-tutorial
IC2S2