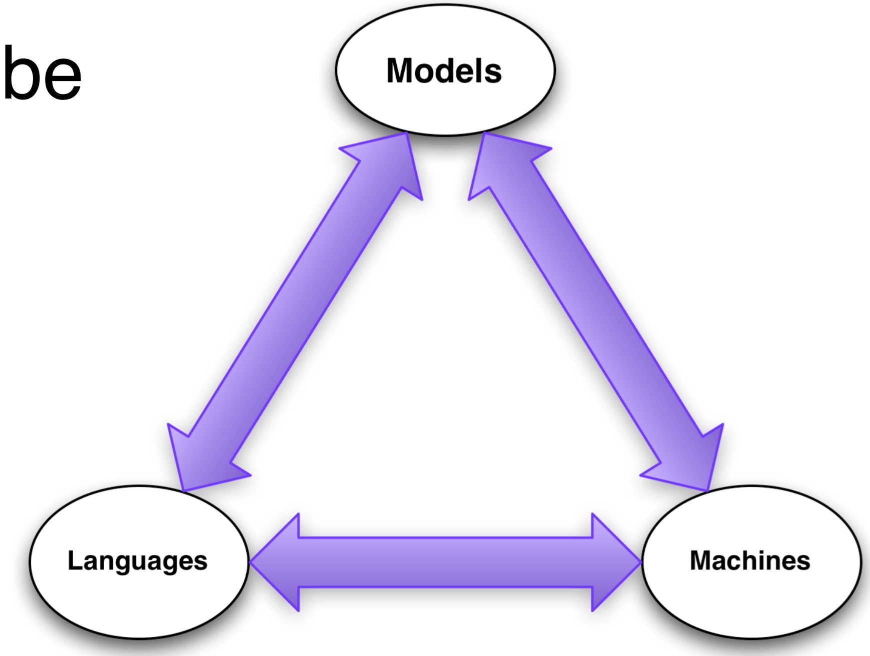
Computing in
Data Science Curricula

Some unspoken assumptions?

- Obviously DS is interdisciplinary but maybe some computationalists end up claiming data science by default by seeing it through a particular lens
- That's okay, and hardly unique to computing, but probably worth being upfront about it



The Pillars from Computing

- Basic Foundations: understanding data
  - Algorithms and algorithmic thinking, including data structures
  - Machine Learning (presumes statistics is happening somewhere)
  - Data curation (!= databases, but includes a lot of it)
  - Visualization and modeling
  - Computer systems

- Advanced Foundations: understanding large scale data
  - High Performance Computing, including algorithms for large data
  - Advanced Machine Learning, including advanced statistics

- Practicum: applying knowledge to real world problems, mastering tools of the trade

Algorithms and Algorithmic Thinking

- programming skills and tools, including data structures
- Big O, and algorithmic analysis
- Maybe some complexity (but maybe not the halting problem?!)
- sublinear algorithms
- random sampling
- randomized algorithms
- streaming algorithms
- ...and not too much more or it will end up a BS CS (interestingly, could still be accredited by ABET for computing)

Machine Learning

- Breadth of ML: Data finally trumps data structures
    - ...with focus on supervised and unsupervised
    - ...not so much on reinforcement and game theory
- Include emphasis on the empirical side of ML, resist just theory and derivations from first principles that we tend to do when left to our own devices
- Include emphasis on scalability, including direct ties to linear algebra scalability as well as dealing with problems of iterations over data
- KDD vs NIPS?

Data Curation, Visualization, and Modeling

- Storage, retrieval, and transmission of data... but not to the point of three database courses and two networking courses
- More emphasis on unstructured and error prone data (e.g. "real data from the web")
- Some security would be useful, policy and ethics would definitely be useful
- Data analysis for human understandability and interaction...as in HCI and DataViz

Systems and High Performance Computing

- Automation of analyzing big data
  - parallel algorithms
  - high performance software
- Programming for speed
- Special purpose programming (sometimes for special purpose hardware)
- Probably not organizations of hardware courses
- Systems includes software engineering, but perhaps more emphasis on prototyping

Tools of the Trade: Data Engineering?

- Facility with the major software tools that are now part of standard data science solution toolkit
- Pretend I had a slide full of thumb-sized logs of different software packages out there including: R, Spark, Hadoop, MapReduce, Pig, Hive, Floom, Hbase, Sqoop, Yarn, and probably some other random words.

- Applications to special kinds of data, perhaps, and may include appropriate specialization areas in computing like NLP
- Alternatively, some of this stuff may be MS level