Data Scientist Archetypes – Brandon Rohrer

Don't be alarmed – data science is impossibly huge

- Scala, Keras, R, TensorFlow, t-test, JSON, .csv, linear-or-logistic regression, anomaly detection, confidence interval, A/B testing,

Co-authors: Reps from Teradata, Endgame, DrivenData.org, Microsoft, Verily, Facebook, Pluralsight, LinkedIn

Data science is actually 3 different fields

- **Data Analysis** - taking raw information and turning it into knowledge that can be acted on
- **Data Modeling** - Using the data that we have and using it to estimate data we wish we had
- **Data Engineering** - Taking these analyses/modeling activities and making them work faster, more robustly, and on larger quantities of data.
- These are all distinct fields that an entire career could be based in

Data Analysis

- **Domain knowledge**
  - Translate a business need into a question
  - Involves making trade-offs in accuracy vs cost
  - Necessary for interpretation, how the numbers and labels represent in the real world
  - How much we can assume about the label's accuracy
- **Research**
  - Gather the data
    - Add logging to code
    - Sensors on production lines
    - Surveys
  - Design and conduct experiments
    - Plan what and how information will be gathered
    - Plan how it will be analysed after collection
- **Interpretation**
  - Given a large collection of data how can we accurately summarize, aggregate and visualize data.
  - How to turn a sea of bits into a nugget of knowledge
  - Visualization is especially important, as much as an art as a science
    - Requires an innate sense of human perception

Data Modeling (Machine Learning)

- Creating a simplified description of your data that can be used to estimate data that hasn't been measured. Either because there was no sensor for it, or because it hasn't happened yet.
- **Supervised Learning**
    - Use a large collection of labeled examples, distill out patterns
    - Labels are categorical => classification
    - Labels are numerical => regression
    - Can also be used for anomaly detection, by checking to see if new data is similar to what's been previously seen
- **Unsupervised Learning**
    - Data doesn't come with labels
    - Goal is to discover patterns in how the data is distributed
    - Clustering => Data points that are similar
    - Dimensionality reduction => Variables that behave similarly
- **Custom algorithm development**
    - Above learning methods are general and make no assumptions about their domain
    - For approaches with less data, learning isn't practical
    - Build in information we know – or can comfortably assume – about the domain

Data Engineering

- **Data Management**
    - Storing, moving and handling data
    - Trivial for a few thousands, or millions of data points, but might need new tools when the data is made of billions of data points (eg. Genome) and doesn't fit in RAM or even in a single hard-drive
- **Production**
    - Turning code that works well in prototype and making it ready for the world
        - Can build a jupyter notebook, but can't use that to handle all customer facing interfaces
        - Covers the gap of turning a prototype to smoothly running production code
    - Making code that is compatible with client systems
    - Needs to be made capable of ingesting data and publishing findings in the right format
    - Handle glitches and failures gracefully
- **Software Engineering**
    - Plays a role in both the above categories as well
    - Building things out of code that not only do things you want as fast as you want but can also be:
        - shared across a team
        - adapted to future changes
        - scaled up or down with evolving needs

Surprise 4th pillar: Data Mechanics

- Dirty work: everyone needs to do, but no-one wants to talk about
- **Data Formatting**
  - Making sure types are consistent
  - Strings in use are valid
  - Getting data out of strings
  - Type conversion
- **Value interpretation**
  - Dates and times
  - Units of measurements are consistent and documented
  - Missing values
- **Data handling**
  - Querying
  - Slicing
  - Joining

These 4 pillars can be used to define some archetypes of data scientists, using three tiers for each pillar: exposure, proficiency, mastery

---

Data Scientist**: Beginner** – has exposure in each of the 4 pillars

Data Scientist**: Generalist** – has proficiency in each of the 4 pillars

Data Scientist**: Diva** – Proficient in the 3 main pillars, but neglects Data Mechanics. (Anti-pattern)

Data Scientist**: Detective** – has mastery in Data Analysis, exposure in modelling and engineering, and proficient in mechanics

Data Scientist**: Oracle** – has mastery in Data Modeling, exposure in analysis and engineering, and proficient in mechanics

Data Scientist**: Maker** – has mastery in Data Engineering, exposure in analysis and modelling, and proficient in mechanics

Data Scientist**: Unicorn** – has mastery in all 4 pillars. Doesn't exist. Path to mastery in all fields requires an entire career in all of the other archetypes.