Kevin Chen

10/29/17

Data Science Foundations: Fundamentals

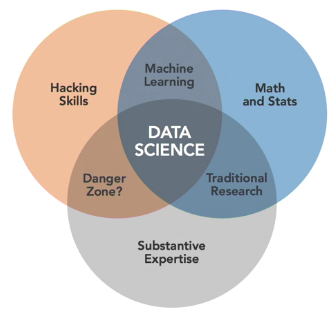
**What is Data Science?**

Demand

* Data Science is the field that joins statistics and programming in applied settings
* It is the analysis of diverse data (not fit into a common format). “Inclusive analysis”: includes all kinds of data.
* Data science shows rare qualities and high demand.
  + Rare qualities: It takes unstructured data and finds order, meaning, and value.
  + High demand because they provide insight and competitive advantage.
* There is a growing need for data science and may exist a shortage.
* High salaries.
* A need for technicians (who do the data science) and for generalists (contextually oriented manages who put those results into practice)

Venn Diagram

* Data Science Venn Diagram
  + Consist of three separate circles representing different areas. Taken together, they constitute Data Science



* Hacking skills (aka computer programming): gather and prepare data. Often in unusual data formats. Substantial creativity is required.
* Math and statistics: how to choose a useful procedure to answer questions and diagnose problems. Develop and improve procedures.
* Substantive Expertise: in your particular field, what is value, what are the goals, what constraints you have.
* Machine learning: “Black box” predictive models. We have variable going into it (what we’re trying to predict) and we have a model that puts it together
* Traditional Research: datasets and analyses are structured
* Danger zone: unlikely to happen.
* There are many roles and require diverse skills.

The Data Science Pipeline