Preparation week 2 Data science

1. https://www.safaribooksonline.com/library/view/creating-a-data-driven/9781491916902/ch01.html - **What do we mean by data-driven?**

- Data-drivenness is about building tools, abilities, and, most crucially, a *culture* that acts on data
- Prereq: An organization must be collecting data.
  - Data undoubtedly is a key ingredient. Of course, it can't just be any data; it has to be the *right* data. The dataset has to be relevant to the question at hand. It also has to be timely, accurate, clean, unbiased; and perhaps most importantly, it has to be trustworthy.
  - This is not always easy. It is easy to bias conclusions and cleaning data is a long and tiresome process.
  - Even if you have quality data, it will only get you so far. It does not make you data driven. A small amount of usefull clean data is more worth that giants amounts of junk data.
- Data must be accessible and queryable
  - Having accurate, timely and relevant data is not sufficient to count yourself as being data driven. The data must be:
    - **Joinable -** The data must be in a form that can be joined to other enterprise data when necessary.
    - **Shareable –** There must be means to share the data within the organization
    - **Queryable -** there must be appropriate tools to query and search for the right data in a big database.
  - Now the data is accessible, but not sufficient enough yet. You need people with the right skills to use that data. That can mean the mechanics of filtering and aggregating data, such as through a query language or Excel macros, but it also means people who design and choose the appropriate metrics to extract and track.
  - **So, for an organization to be data-driven, there have to be humans in the loop, humans who *ask the right questions* of the data, humans who have the skills to extract the right data and metrics, and humans who use that data to inform next steps. In short, data alone is not going to save your organization.**
- **Reporting**
  - Data also needs to be presented in the right way. You could have a lot of valid and accurate data, but if you can't present it in a good way, the information won't mean anything.
- **Alerting**
  - Alerts are essentially reports about what is happening right now. They typically provide very specific data with well-designed metrics.

- **Hallmarks of data-drivenness**
  - There are a few characteristics that define a data-driven organization:
    - A data-driven organization may be continuously tested.
    - A data-driven organization may have a continuous improvement mindset.
    - A data-driven organization may have a continuous improvement mindset. It may be involved in repeated optimization of core processes.
    - A data-driven organization may be involved in predictive modeling, forecasting sales, stock prices, or company revenue, but importantly feeding the prediction errors and other learning back into the models to help improve them.
    - A data-driven organization will almost certainly be choosing among future options or actions using a suite of weighted variables.
  - At data driven organization will atleast do one of the following above.

- OK, now we have an organization that has high-quality data and skilled analysts who are engaged in these forward-looking activities. Surely, that makes it data-driven!
- Unfortunately, not necessarily. Like a tree falling in a forest with no one to hear it, if analysts are putting out analyses but no one takes notice, if they don't influence decision makers' decisions, which are still based on gut and opinion, it is not data-driven. Analytics has to inform and influence the influencers.

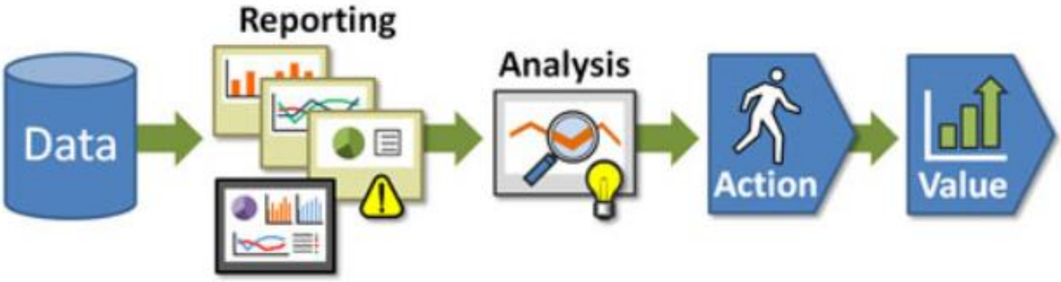

Figure 1: Dykes analytics value chain

- **Analytics**

  There are 8 levels of analytics:

  - *Standard reports*
    - What happened? When did it happen? *Example*: monthly financial reports.
  - *Ad hoc reports*
    - How many? How often? Where? *Example*: custom reports.
  - *Query drill down (or online analytical processing, OLAP)*

        - Where exactly is the problem? How do I find the answers? *Example*: data discovery about types of cell phone users and their calling behavior.
    - *Alerts*
        - When should I react? What actions are needed now? *Example*: CPU utilization mentioned earlier.
    - *Statistical analysis*
        - Why is this happening? What opportunities am I missing? *Example*: why are more bank customers refinancing their homes?
    - *Forecasting*
        - What if these trends continue? How much is needed? When will it be needed? *Example*: retailers can predict demand for products from store to store.
    - *Predictive modeling*
        - What will happen next? How will it affect my business? *Example*: casinos predict which VIP customers will be more interested in particular vacation packages.
    - *Optimization*
        - How do we do things better? What is the best decision for a complex problem? *Example*: what is best way to optimize IT infrastructure given multiple, conflicting business and resource constraints?

-
-