**DATA**

* Data are observations of real-world phenomena
* For instance,
  + stock market data might involve 🡪 “Which stocks I should invest in?”
    - observations of daily stock prices
    - announcements of earnings by individual companies
    - even opinion articles from pundits .
  + Personal biometric data can include 🡪 “How can I live a healthier lifestyle?”
    - measurements of our minute-by-minute heart rate, blood sugar level, blood pressure, etc.
  + Customer intelligence data includes 🡪 “How can I understand my customers’ changing tastes, so that my business can serve them better?
    - Observations such as “Alice bought two books on Sunday,”
* there’s always measurementnoise and missing pieces.

**Machine Learning**

Machine learning fits mathematical models to data in order to derive insights or make predictions

**Models**

ML models take features as input

**Features**

* Feature is a numeric representation of an aspect of raw data. Features sit between data and models in the machine learning pipeline
* Right features can ease the difficulty of modelling and enable the pipeline to output results of higher quality
* Right features can only be defined in the context of both the model and the data; since data and models are so diverse, it’s difficult to generalize the practice of feature engineering across projects.

**Feature engineering**

* Feature engineering is the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model.

**Workflows**

* Workflows with data are frequently multistage(involves a lot of stages), iterative(done again and again) processes.
  + For instance,
    - stock prices are observed at the exchange
    - Then aggregated by an intermediary like Thomson Reuters
    - Then stored in a database
    - Then bought by a company, converted into a Hive store on a Hadoop cluster
    - Then pulled out of the Hive store by a script,
      * subsampled(divided into samples according to the requirements)
      * massaged(to help get more info regarding the data)
      * cleaned by another script
      * dumped to a file, and converted to a format that you can try out in your favorite modellinglibrary in R, Python, or Scala.
    - The predictions are then dumped back out to a CSV file and parsed by an evaluator, and
    - the model is iterated multiple times, rewritten in C++ or Java by your production team, and run on all of the data before the final predictions
    - are pumped out to another database