**Agenda:**

Begin with discussion of deliverable.

Consider Constraints

Review Concept Organization

Begin discussion of conceptual hierarchy

**Constraints:**

Restrict to what is different now that we have “big data”

Not trying to create a taxonomy of the entire data lifecycle and all data types

Keep terms independent of a specific tool

**Plan**:

1. Begin with concepts
2. Assign terms
3. Wordsmith definitions

Plan for check-point with other subgroups on the Wednesdays we meet as a whole

Save the definition of big data till after we’ve clarified component terms

Begin with a simple hierarchy of concepts

**Big Data Scope**:

**(1) Dataset characteristics**

**(2) Data Storage and Access Paradigms – Big Data Engineering**

**(3) Ubiquitous Sensors – Internet of things**

**(4) Changing Analytics Paradigm – Data Science**

**(5) Changing Analytics Engineering**

**(6) Data Lifecycle?**

**(7) ?**

**(1) Dataset characteristics**:

Volume – amount

Velocity – rate

Variability – changing rate

Variety/Complexity – data types (structured, unstructured, etc)

Variety/Complexity – numbers of datasets

Variety/Complexity – different refresh rates or timescales

**(2) Data Storage Paradigms:**

Flat files

Relational database

Big table

Name-value pairs

Graph

**(3) Ubiquitous Sensors – Internet of things**

**(4) Changing Analytics Paradigm – Data Science**

Statistics – rigorous causal analysis of carefully sampled data

Data Mining – approximate causal analysis of repurposed data carefully sampled

Data Science – probabilistic analysis/trending of large selection or even entire dataset

Data Science – correlation not necessarily causation

Data Science – determine the questions and not the answers

Data Science – getting an answer by solving a simpler problem

Veracity – precision/accuracy/timeliness of the data

Provenance – data history

Meta-data – data about data

Cleanliness

?

**(5) Changing Analytics Engineering**

Cluster storage

Mapreduce

**(6) Data Lifecycle**

A simple lifecycle is:

Collect

Curate

Analyze

Act

Or the military OODA

Observe

Orient

Decide

Act

**From CRISP-DM**

Business Understanding

Objectives goals

Data Mining goals

Plan

Data Understanding

Collect initial data

Describe data

Explore data

Verify Data Quality

Data Preparation

Select data

Clean data

Construct data

Integrate data

Format data

Modeling

Select modeling technique

Generate test design

Evaluation

Evaluate results

Review process

Determine next steps

Deployment

Plan deployment

Plan monitoring and maintenance

Produce final report

Review project

Do we have any changes here because of big data?