- Big Data
  - Roles (from V2 taxonomy)
    - Data Consumer
    - System Management
    - Big Data Security and Privacy
    - Data Provider
    - Big Data Framework Provider
    - Big Data Platform Specialists
  - Characteristic
    - defining
      - Volume
        - by reason
          - Accumulation over time — *web history scenario*
          - Collection from large number of data sources — *IoT scenario*
          - Redundancy of data record
            - high degree of de-normalization
              - polyglot persistence
              - materialized views
            - large numbers of copies with slight or no differences — *often observed in log data*
          - Very large records — *DNA sequence*
        - by lifecycle position
          - Big at moment of creation (Big in the Beginning) — *web scan*
          - Big at the moment of analysis (Big in the Middle) — *explosion of temporary data during analysis*
          - Big derivative (Big at the End) — *result of data enrichment*
      - Velocity
        - by reason
          - rate of acquisition of new data records — *click stream log*
          - rate of update for existing data records — *state changes for monitored objects*
          - rate of churn among data sources — *keeping track of "smart ammunitions"*; *this case is like the other two combined*
        - by time scale
          - off-line
          - near-line
          - on-line
      - Variety
        - of data format
          - unstructured
          - semistructured
            - arbitrary nested structure of documents
            - "anything goes" fields
          - structured
            - multiple ways for schema definition
              - xml
              - json
              - SQL
              - custom
                - proprietary
                - public
            - multiple options for schema availability
              - on-line — identified by URI
              - in-line — located in the same system/file as data
              - implied — located "in the head" of expert
            - unstructured data inside of structured documents — *XML CDATA*
        - of storage media
          - *tape*
          - *SSD*
          - *Memory*
        - of data access latency
          - distance from CPU
            - CPU cache
            - memory
            - disk
            - NAS
          - variability of network latency
        - of syntax
          - natural language
            - dialects
          - storage encoding
            - versions
            - character sets
          - wire encoding
            - versions
            - compression
        - of errors
          - measurement error
          - processing error
          - human error
        - of certainity
          - in consistency
            - eventually consistent
            - always consistent
          - in integrity
            - uncertain completeness of some documents
            - possible tempering with some documents
          - in veracity
            - uncertain source
              - record about source is lost
              - source is not credible
            - uncertain time
              - time zone
              - clock drift
              - time encoding
          - in validity
            - out of range
            - wrong format
        - of missing data
          - data present but unaccessible
            - due to security restrictions
            - due to privacy restrictions
            - due to lost metadata
            - due to ambiguity of results — *different values reported from different replicas*
          - data absent in storage
            - phenomena did not exist
            - not collected
            - lost in transmission
            - access failure
              - temporary — *loss of network connection to storage*
              - permanent — *loss of media*
              - at required time scale — *on-line request for data on tape*
        - of transparency
          - transparent
            - plain text
              - ASCII
              - Unicode
              - EBCDIC
            - binary
              - real number representation
                - floating point
                - fixed point
              - endianness — *http://en.wikipedia.org/wiki/Endianness*
          - partially opaque data
            - obfuscated
            - encrypted
        - ? of meaning
          - multiple interpretations
            - of data
            - of observed phenomena
          - multiple units and scales
          - multiple inferred relationships
    - describing
      - Location
        - of storage
          - Attached — *Mainframe*
          - Local network — *Hadoop cluster*
          - Geographically distributed private — *Cloud storage*
          - Geographically distributed public — *Open Data*
        - of source
          - Attached — *MRI machine*
          - Local Network — *enterprise data warehouse*
          - Remote — *Cloud managed IT on premises data*; *IoT sensor data*
      - Dynamics
        - as static
        - as accumulating
          - appending to existing records
          - creating new records
        - as changing
          - by update
            - Transactional — *Teradata DB*
            - Eventually consistent
              - without update history — *Hadoop*
              - with update history — *Cassandra*
          - by streaming — with significant temporary retention — *Kafka*
      - Sensitivity
        - Open
        - Private
        - Secret
      - Content
        - Machine generated — *cpu utilization*
        - Human generated — *click stream*
        - Data indirectly revealing information about a human — *geo-location of a device carried by human*
      - Time span
        - perpetual — *wayback machine*
        - current — *search engine data*
        - limited time window — *rotating logs*
  - Stakeholder
    - An entity revealing data about itself, or about own property.
    - Originator
      - definition
        - It is possible to define originator, also, as owner of infrastructure that collects data (e.g Facebook, wireless provider)
        - *Legal definition needs to be defined, perhaps, by the Regulator*
      - interested in
        - Monetization
        - Privacy
        - Protection of personal or trade secret(s)
      - risk
        - unintended exposure
        - identity theft
        - excessive monetary or in-kind cost of revealing data
        - inconvenience/annoyance
    - Owner
      - definition
        - entity with valid legal claim to data
      - interested in
        - Governance
        - Monetization
        - Preventing unauthorized access
      - risk
        - loss of exclusivity for owned data — need for digital rights mgmt?
        - high cost of data retention
    - Operator
      - definition
        - human or corporation operating data storage, movement or processing
      - interested in
      - risk
    - Curator
      - definition
        - entity in custody of choosing, maintaining, and increasing quality of a collection of data
      - interested in
        - Monetization
          - refined data
        - Data Integration
        - Data Quality
          - normalization
          - disambiguation
          - refinment
          - views
            - via polyglot persistence
            - via schema on read
        - Creation/collection of metadata
      - risk
        - effort wasted on not useful data
        - data bloat
        - misinterpretation of raw data
    - Analyst
      - interested in
        - Ease of access to data
          - in standard query languages
          - in common data interchange formats
        - Speed of access to data
        - Data Quality
          - fidelity of data
          - granularity of data
          - availability of data suitable for processing by analytic tools
        - Monetization — analytics
        - Creation of analytic
          - reports
          - data
          - predictions
      - risk
        - oversteppping privacy expectations of data originator
        - producing misleading results — garbage in garbage out
    - Customer
      - definition
      - interested in
        - Analytics
        - Data driven features
        - New knowledge/insights
        - Minimization of cost
      - risk
        - decisions based on erroneous insights or predictions
    - Regulator
      - definition
        - a government or private entity establishing laws, guidelines, and regulations for collection, processing, and use of data
        - *while data was "small" it had no significant impact on society as a whole, hence this stakeholder is specific to Big Data*
      - interested in
      - risk