- Workshop
  - Data science crash course
    - Preliminaries: learning from data
    - Classic techniques
      - Classification
        - Example: spam filtering
        - Example: patient triage
      - Regression
      - Clustering
        - Example: where to put retail stores
        - Example: identifying related products
        - Example: market segmentation
      - Recommenders
    - Feature engineering
      - Data cleaning
        - Imposing structure
        - Imposing constraints
        - Correcting corrupt or spurious records where possible
        - Rejecting corrupt or incomplete records when necessary
        - Example: address correction
        - Example: infrastructure log timestamps
      - Normalization
        - Simple example: temperature scale ranges
        - Application-focused example: fixed ranges to prepare for linear regression, PCA, or similar techniques
      - Encoding
        - one-hot encoding
        - hashing
      - Natural-language features
        - Classical vs statistical techniques
        - Stemming
        - Stopwords
        - TF-IDF
        - Topic modeling
        - Word2Vec
      - Dimensionality reduction
        - Motivation: most high-dimensional data has relatively few meaningful dimensions
        - PCA
        - random projection
        - tree-based approaches
    - Introducing Spark
      - Fundamental ideas
        - Data-parallel computation
        - The RDD: laziness, immutability, and resilience
        - Example program
      - Structured data
      - Machine learning
    - Deploying an app on OpenShift