Data Science Evaluation

What are the Data Science Challenges?

Bonnie Dorr

DSRP Aims to Address Data Science Challenges

Categories of Challenges: **The V's**

- **Volume:** data may be too large to store
- **Velocity:** data must be processed quickly
- **Variety:** data must be integrated from multiple sources of differently-formatted data
- **Veracity:** data may contain errors

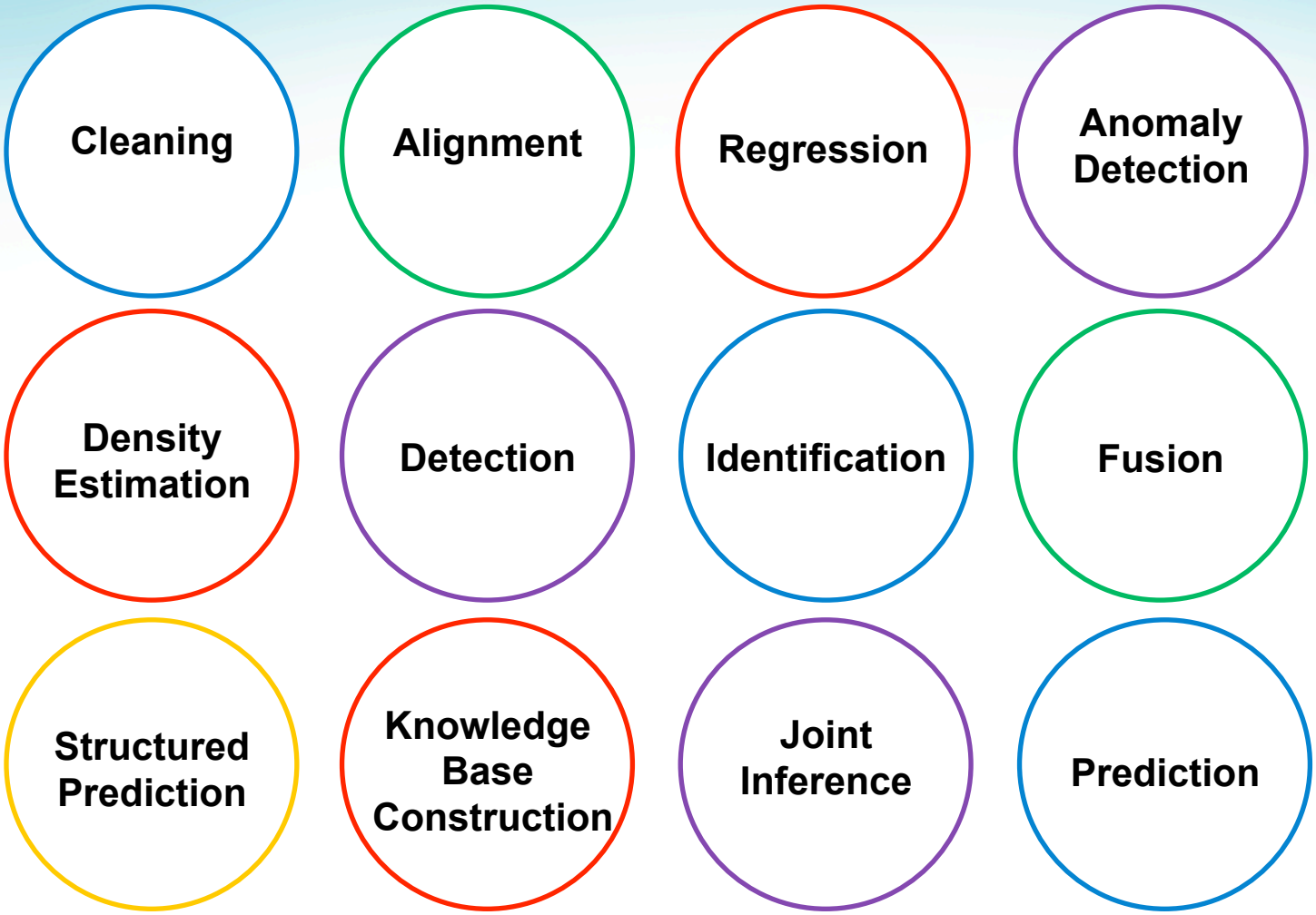
Examples of Data Science Challenges

Provenance	Origin of raw data? Is it current? What processes were applied to derive the data from its original sources?
Heterogeneity	How to use data from multiple large heterogeneous datasets?
Predictive Analytics	How to identify and distinguish trends from random fluctuation to provide a calibrated forecast of future value?
Knowledge Assimilation	How might algorithms understand data, e.g., infer causality?
Big Data Replicability	How to reproduce experimental findings given that truth may be hard to find, consistently?
Visualization of Knowledge	How to visually represent knowledge for decision making?
Data Uncertainty	How to handle gaps in knowledge due to potential for untrustworthy or inaccurate data?
Mitigating Error Propagation	How can algorithms mitigate cascading of error through data processing steps?
Data Privacy and Security	How do we manage data and develop algorithms in the face of privacy and security concerns/policies?

Examples of Data Science Challenges

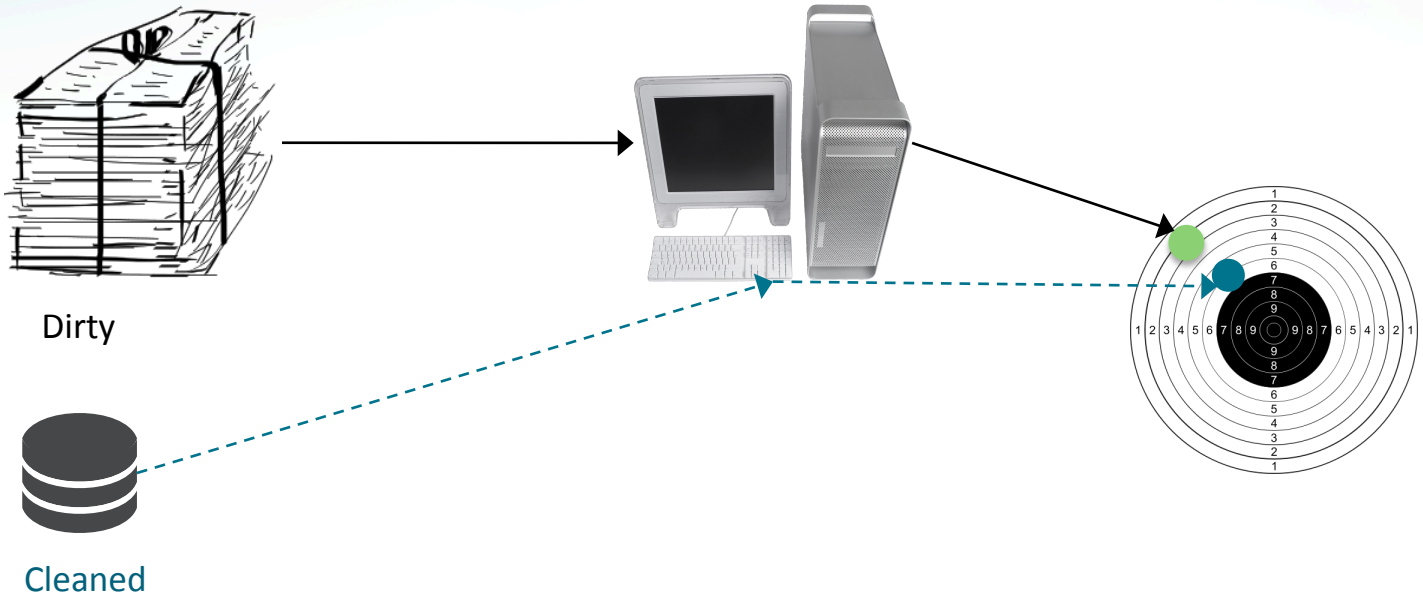
Provenance	Origin of raw data? Is it current? What processes were applied to derive the data from its original sources?
Heterogeneity	How to use data from multiple large heterogeneous datasets?
Predictive Analytics	How to identify and distinguish trends from random fluctuation to provide a calibrated forecast of future value?
Knowledge Assimilation	How might algorithms understand data, e.g., infer causality?
Big Data Replicability	How to reproduce experimental findings given that truth may be hard to find, consistently?
Visualization of Knowledge	How to visually represent knowledge for decision making?
Data Uncertainty	How to handle gaps in knowledge due to potential for untrustworthy or inaccurate data?
Mitigating Error Propagation	How can algorithms mitigate cascading of error through data processing steps?
Data Privacy and Security	How do we manage data and develop algorithms in the face of privacy and security concerns/policies?

Classes of Data Science Problems



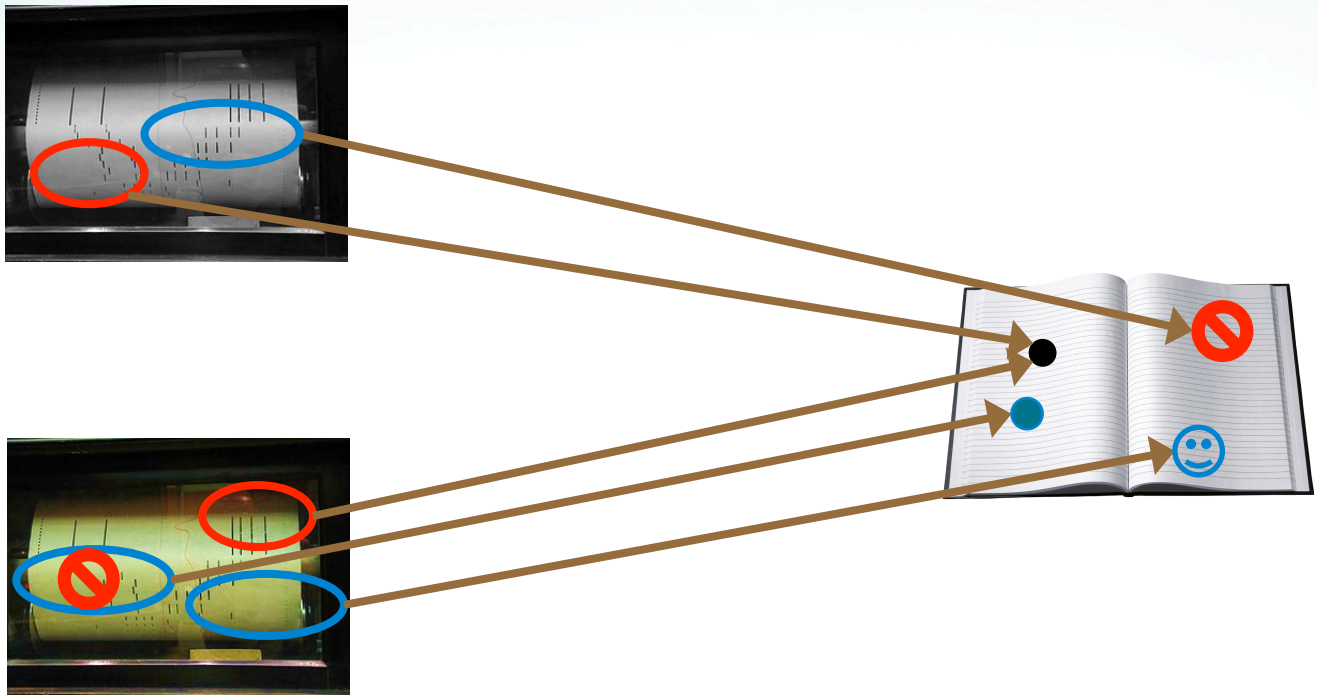
Data Science Challenge: Example #1

Dirty and Uncertain Data



Images used with permission from Wikimedia Commons; URLs given at the end of the presentation

Data Science Challenge: Example #2



Images used with permission from Wikimedia Commons; URLs given at the end of the presentation

Traffic Detection and Prediction is a Means, not our End Goal

- Traffic prediction is a **generalizable example** to serve as a **track archetype**
- Tracks established in **other domains** expected to have **similar challenges and problems**

Questions

- Alignment:
  - What is an alignment problem that you face in your domain?
  - How do you deal with it?
- Cleaning:
  - How do we tell when data is cleaner?
  - What errors in data are more problematic? What algorithms are more robust to errors?
  - What errors in data inhibit experiment reproduction, and how do we design experiments to mitigate the effects of these errors?
  - How do we track data to identify which points have been cleaned and how they have been changed?
- What Prediction/Forecasting challenges arise in your domain of expertise?

Image References

- Stack of Papers: https://commons.wikimedia.org/wiki/File:Stack_of_papers_tied.svg
- Target: https://commons.wikimedia.org/wiki/File:25_Meter_Precision_and_50_Meter_Pistol_Target.svg
- Pianola Paper: https://commons.wikimedia.org/wiki/File:Pianola_paper_tape.JPG