IRDS: Data Mining Process

Charles Sutton
University of Edinburgh

“Data Science”

- Our working definition
  - Data science is the study of the computational principles, methods, and systems for extracting knowledge from data.
- A relatively new term. A lot of current hype…
  - “If you have to put ‘science’ in the name…”
- Component areas have a long history
  - machine learning
  - databases
  - statistics
  - optimization
  - natural language processing
  - computer vision
  - speech processing
  - applications to science, business, health….
- Difficult to find another term for this intersection

The term “data mining”

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.
>
> — Hand, Mannila, Smyth, 2001

The term “data mining”

Data mining is the analysis of (often large) **observational data sets** to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner. — Hand, Mannila, Smyth, 2001

not collected for the purpose of your analysis

The term “data mining”

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data **in novel ways** that are both understandable and useful to the data owner.
>
> — Hand, Mannila, Smyth, 2001

Many “easy” patterns already known

e.g., pregnant example from association rule mining

The term “data mining”

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both **understandable and useful** to the data owner.
>
> — Hand, Mannila, Smyth, 2001

Tradeoff between

- predictive performance
- human interpretability
  Ex: neural networks vs decision trees

Before I get too far ahead of myself…

What problem am I trying to solve?

Problem Types

- Visualization
- Prediction: Learn a map **x** —> y
  - Classification: Predict categorical value
  - Regression: Predict a real value
  - Others
    - Collaborative filtering
    - Learning to rank
    - Structured prediction

supervised learning

- Description
  - Clustering
  - Dimensionality reduction
  - Density estimation
  - Finding patterns
    - Association rule mining
    - Detecting anomalies / outliers

unsupervised learning

Prediction Examples

- Classification
  - Advertising
    - Ex: Given the text of an online advertisement and a search engine query, predict whether a user will click on the ad
  - Document classification
    - Ex: Spam filtering
  - Object detection
    - Ex: Given an image patch, dose it contain a face?
- Regression
  - Predict the final vote in an election (or referendum) from polls
  - Predict the temperature tomorrow given the previous few days
- Sometimes augmented with other structure / information
  - Structured prediction
    - Spatial data, Time series data
    - Ex: Predicting coding regions in DNA
  - Collaborative filtering (Amazon, Netflix)
  - Semi-supervised learning

Description Examples

- Clustering
  - Assign data into groups with high intra-group similarity
    - (like classification, except without examples of “correct” group assignments)
  - Ex: Cluster users into groups, based on behaviour
    - Social network analysis
  - Autoclass system (Cheeseman et al. 1988) discovered a new type of star,
- Dimensionality reduction
  - Eigenfaces
  - Topic modelling
- Discovering graph structure
  - Ex: Transcription networks
  - Ex: JamBayes for Seattle traffic jams
- Association rule mining
  - Market basket data
  - Computer security

Data Science Process

Preparation

- Understanding the problem
  - What are your goals?
  - What is the “signal” in the data?
  - What would success be? Is it likely?
- Collect data
  - Think about / mitigate biases: selection, non-response, etc.
  - Think about causes of noise or missing data
- Data preparation
  - Data wrangling
  - Scraping
  - Data integration
  - Data cleaning
  - Data verification, scrubbing, auditing
- Data exploration
  - Exploratory visualisation

Data Science Process

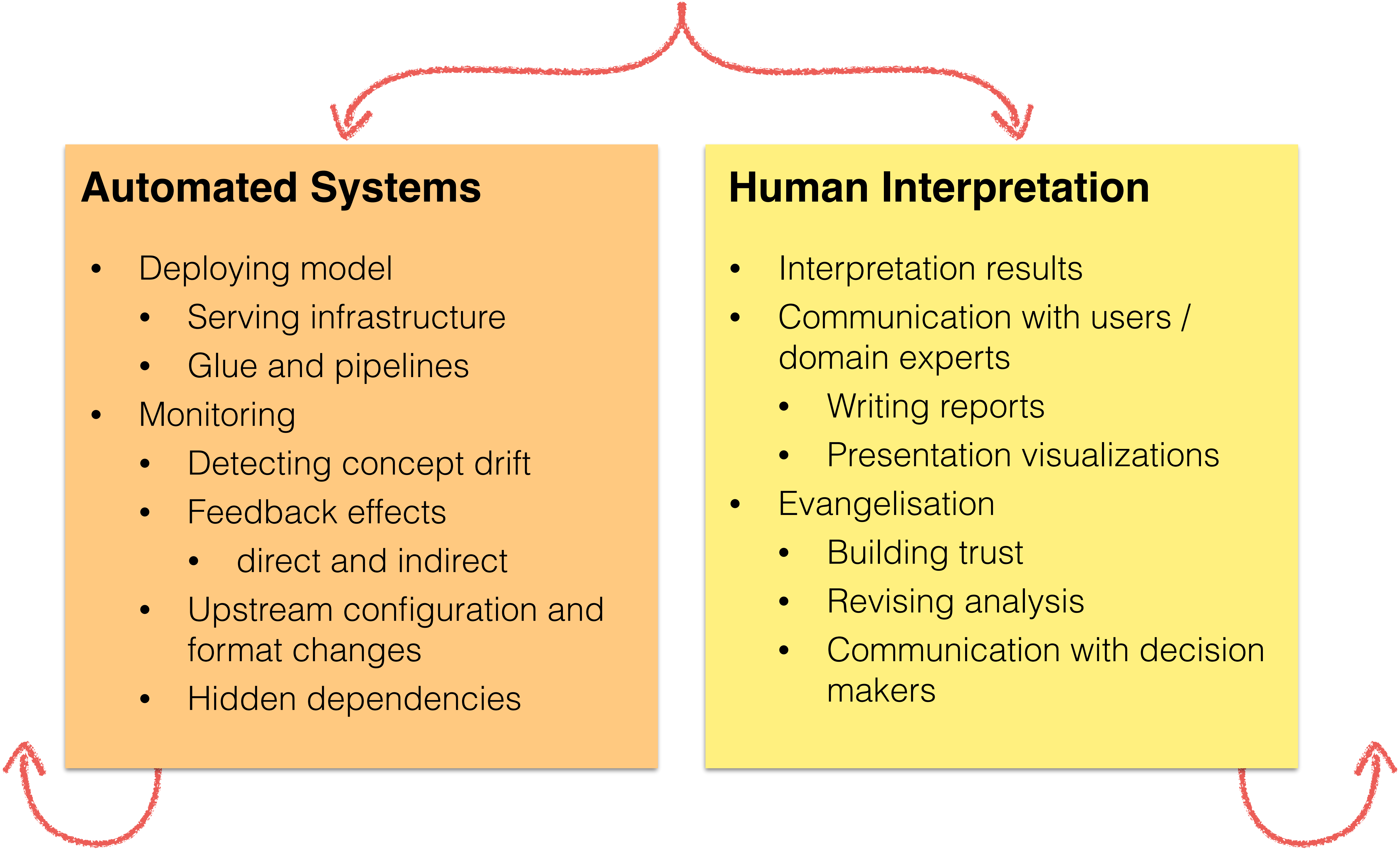
(part 2)

Development and Debugging

- Feature engineering
- Model building
    - Choose algorithm
    - Train / fit to data
        - (possibly parallel, distributed, GPU)
    - Hyperparameter exploration
- Evaluation and debugging
    - Evaluation metrics and procedures
        - (e.g., cross-validation, temporal or stratified sampling of development sets)
    - Developing diagnostics
    - Revising previous steps in process
    - Sensitivity to model parameters
    - Evaluative visualization

Data Science Process

(parts 3a and 3b)



Automated Systems

- Deploying model
  - Serving infrastructure
  - Glue and pipelines
- Monitoring
  - Detecting concept drift
  - Feedback effects
    - direct and indirect
  - Upstream configuration and format changes
  - Hidden dependencies

Human Interpretation

- Interpretation results
- Communication with users / domain experts
  - Writing reports
  - Presentation visualizations
- Evangelisation
  - Building trust
  - Revising analysis
  - Communication with decision makers

Roadmap

In the next few weeks, we'll talk about

- Visualization
- Feature extraction
- Evaluation and debugging

Further Reading

The “process” flowchart combines ideas from:

- D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, 2015.
- K. Wagstaff. Machine learning that matters. In *International Conference on Machine Learning (ICML)*, 2012.
- P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0 Step-by-step data mining guides*. 2000.

These readings are not examinable.

(there is no exam for this course)