Towards Automating the Data Analytics Process

Chris Williams

February 2017
He expects so much from me but his data is so flawed.
Alan Turing
– heritage and inspiration

© Alan Turing Institute
The Alan Turing Institute is the “UK’s national institute for data science”
ATI Mission

(from https://turing.ac.uk/)

- To undertake data science research at the intersection of computer science, mathematics, statistics and systems engineering;
- To provide technically informed advice to policy makers on the wider implications of algorithms (including developing a strong ethical foundation for data science);
- To enable researchers from industry and academia to work together to undertake research with practical applications;
- To act as a magnet for leaders in academia and industry from around the world to engage with the UK in data science and its applications.
ATI Headquarters

- at the British Library
- in the heart of the new “Knowledge Quarter”
ATI Strategic Priorities

as of June 2016

[Diagram showing the following areas:
- Engineering
- Technology
- Defence & security
- Smart cities
- Financial services
- Health & wellbeing

Each area is connected to:
- Mathematical representations
- Inference and learning
- Systems and platforms
- Understanding human behaviour]
Strategic Partners

- **Lloyds Register Foundation**
  Building the emerging field of data science and the role of data-centric engineering within it

- **Intel**
  A research programme focused on high-performance computing and data analytics. In addition, Intel will dedicate a hardware architecture team at the Institute’s facilities

- **GCHQ**
  GCHQ and the ATI have agreed in principle to work together with the wider national security community for the benefit of data science and analytics research in the UK

- **HSBC**
  HSBC and the Turing are partnering to advance research into economic data science
Turing came on stream in October 2016

- ~ 90 Faculty Fellows (part time)
- ~ 20 Research Fellows (full time)
- ~ 15 ATI PhD students plus 25 one-year enrichment students
HE EXPECTS SO MUCH FROM ME BUT HIS DATA IS SO FLAWED
Towards Automating the Data Analytics Process

▶ Common view that up to 80% of work on a data mining project is involved in **data understanding** and **data preparation**

▶ What can we do to try to improve that?

▶ ATI team on *AI for Data Analytics* (started work recently)
  ▶ James Geddes
  ▶ Zoubin Ghahramani
  ▶ Ian Horrocks
  ▶ Charles Sutton
  ▶ Chris Williams

▶ Our work complements the Automatic Statistician (Ghahramani et al), which is more concerned with the search for appropriate analysis models given clean data
Outline

- What is Machine Learning?
- CRISP-DM Methodology
- Five aspects of Data Wrangling
  1. Data Parsing
  2. Data Understanding
  3. Data Cleaning
  4. Data Integration
  5. Data Restructuring
What is Machine Learning?

- It’s about finding patterns in data, and using the patterns to make predictions.
- Focus here on **Supervised Learning**: Given dataset $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$ or $\mathcal{D} = (X, y)$ learn a predictor that given a new $x^*$ makes a useful statement about the associated $y^*$.
- Classification and regression problems.
- Data is usually formed into a rectangular table, with rows as *training examples or cases*, and columns being *features, attributes or covariates*.

**Figure credit: Kevin Murphy (2012)**
Example: Categorizing Documents

Input: Text of Document

Tory fury as Lib Dem peers join Labour to delay boundary review
Review delayed for five years, seriously damaging David Cameron's chances of winning overall majority in 2015

Patrick Wintour, political editor
The Guardian, Monday 14 January 2013 19:45 GMT
Jump to comments (120)

David Cameron's chances of winning an overall majority in 2015 have been seriously damaged by the delay to the boundary review. Photograph: Kerim Okten/EPA

Coalition relations plummeted on Monday when the Liberal Democrats were accused by Conservatives of double crossing, cynicism, cheating and opportunism as Nick Clegg's peers joined Labour to delay a constituency boundary review that had been likely to gift the Tories 20 extra seats.

The review will now be delayed for five years, leaving the next election to be fought on the existing constituency boundaries, so seriously

NFL playoffs: Colin Kaepernick and Ray Lewis keep Super Bowl dreams alive
An exhilarating weekend of NFL playoffs saw Peyton Manning, Russell Wilson, Aaron Rodgers and Rob Gronkowski depart; while Ray Lewis, Colin Kaepernick, and Matt Ryan go through

Golden Tate celebrates a Seattle touchdown as the Seahawks contributed to a helter skelter weekend of NFL action. Photograph: Mike Ehrmann/Getty Images

John Fox the real villain in Denver's defeat
In the heady moments that followed Baltimore's double overtime victory over the Broncos on Saturday, Ray Lewis cut to the chase: "When all the emotions calm down," he said, "this will probably go down as one of the greatest victories in Ravens history."

If that is not such a big claim to make for a team which is still just 16 years old, then the truth is that this game will be remembered far outside Baltimore's city limits. It might just merit consideration in among the league's best-ever playoff games. (On what certainly felt like one of the best-ever weekends.) At 4hrs 11mins, it was the longest NFL game since 1987 and there were not a lot of dull moments in that time.

Label: Politics Sports
Slide from Charles Sutton
Example: Categorizing Documents

- Make a list of sport terms?
- What about this:

Supreme court ruling: Medicaid expansion becomes political football
States may opt out of a programme offering health coverage to 16 million of America's poorest after supreme court ruling

Chris McGreal and Richard Adams
guardian.co.uk, Friday 29 June 2012 14:23 BST

Even this simple task is a bit more complicated.
Example: Learning Robot Inverse Dynamics

- Calculate the torques that must be applied to make the robot move in the way prescribed by its current task
- A regression problem
Unsupervised Learning

- Clustering
- Principal components analysis
- More sophisticated models (e.g. topic models, sparse coding)
- Also based on data matrix $X$
CRISP-DM Methodology

Cross Industry Standard Process for Data Mining
Six Phases:

▶ Business Understanding
▶ Data Understanding
▶ Data Preparation
▶ Modelling
▶ Evaluation
▶ Deployment
<table>
<thead>
<tr>
<th>Business Understanding</th>
<th>Data Understanding</th>
<th>Data Preparation</th>
<th>Modeling</th>
<th>Evaluation</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Determine Business Objectives</strong></td>
<td><strong>Collect Initial Data</strong></td>
<td><strong>Select Data</strong></td>
<td><strong>Select Modeling Techniques</strong></td>
<td><strong>Evaluate Results</strong></td>
<td><strong>Plan Deployment</strong></td>
</tr>
<tr>
<td>Background</td>
<td>Initial Data Collection Report</td>
<td>Rationale for Inclusion/Exclusion</td>
<td></td>
<td>Assessment of Data Mining Results w.r.t. Business Success Criteria</td>
<td>Deployment Plan</td>
</tr>
<tr>
<td>Business Objectives</td>
<td><strong>Describe Data</strong></td>
<td><strong>Clean Data</strong></td>
<td>Modeling Technique</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business Success Criteria</td>
<td>Data Description Report</td>
<td>Data Cleaning Report</td>
<td>Assumptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Explore Data</strong></td>
<td><strong>Construct Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inventory of Resources</td>
<td>Data Exploration Report</td>
<td>Derived Attributes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requirements, Assumptions, and Constraints</td>
<td></td>
<td>Generated Records</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risks and Contingencies</td>
<td><strong>Integrate Data</strong></td>
<td><strong>Generate Test Design</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminology</td>
<td>Merged Data</td>
<td>Test Design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs and Benefits</td>
<td><strong>Format Data</strong></td>
<td><strong>Build Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reformatted Data</td>
<td>Parameter Settings Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model Descriptions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Determine Data Mining Goals</strong></td>
<td><strong>Verify Data Quality</strong></td>
<td><strong>Assess Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Mining Goals</td>
<td>Data Quality Report</td>
<td>Model Assessment Revised Parameter Settings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Mining Success Criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Produce Project Plan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Assessment of Tools and Techniques</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure credit: CRISP-DM 1.0 Step-by-step data mining guide, Chapman et al, 2000
Towards (Semi-)Automation of Data Understanding and Data Preparation

- Study what people do over a wide variety of tasks
- Identify what issues they detect and how they fix them
- Provide advanced tools that:
  - Propose fixes
  - Implement them if approved
  - Otherwise avoid repeated pestering
- Build an interactive assistant that will step the analyst through all the issues in the current dataset
Five aspects of Data Wrangling

1. Data Parsing
2. Data Understanding
3. Data Cleaning
4. Data Integration
5. Data Restructuring
1. Data Parsing

- Understanding the data format: names and types of each field in a file
- The *Data Dictionary* should provide this information, but in reality that this information is often out-of-date or incomplete
- If the information exists it can provide:
  - What are the entities and attributes?
  - What is the meaning of a table?
  - What are the constraints on particular fields?
  - What physical units is some value in?
  - Which fields comprise a candidate key?
- Some work e.g. by the PADS project [http://pads.cs.tufts.edu](http://pads.cs.tufts.edu) on the inference of the structure and properties of an ad hoc data source
- Important to learn (and carry over knowledge) from previous datasets
2. Data Understanding

- aka Exploratory Data Analysis (Tukey, 1977)
- Displaying single variables (outliers, multimodality etc)
- Displaying two or more variables (scatterplots etc)
- Projection methods (e.g. PCA, projection pursuit, t-SNE)
- Displaying *local* patterns in datasets, e.g. frequent itemsets

Figure credit: Chris Bishop PRML 2006
3. Data Cleaning

- Noisy data (e.g. mis-spellings)
- Handling missing data
- Detecting anomalies
Noisy data

- Do “IBM”, “I.B.M.” and “IBM corp” refer to the same entity?
- “18/7/16” vs “18 July 2016”
- Need string matching, and special purpose handling of dates etc.
- Openrefine openrefine.org provides some good functionality for such tasks
Missing data

- Lack of a value for a variable within an observation.
- Might be coded as “NA” or a value that is inconsistent with the type of the attribute (e.g. “NaN”)
- But can be coded as e.g. 0 when it is not clear if this value is inconsistent ...
- Why is data missing? Is it missing at random (MAR) or is there a systematic reason for its absence?
Let $x_m$ denote those values missing, and $x_p$ those values that are present.

If MAR, some “solutions” are to *impute* the missing data

- Model $p(x_m|x_p)$ and average (correct, but hard)
- Replace missing data with its global mean value (?)
- Look for similar (close) input patterns and use them to infer missing values (crude version of density model)
Anomaly Detection

- The “identification of items, events or observations which do not conform to an expected pattern or other items in a dataset” (Banerjee and Kumar, 2009)
- Can be at the level of the whole record, or an attribute in a record
- May arise because an error has occurred in data measurement or transmission; but may also arise from correct measurement of an unusual situation
- Usually handled by building a model of normality, and detecting low probability events/records/observations (outliers)
Figure credit: David Barber, BRML, 2012
Conditional anomaly detection

- Example tool: GritBot (Quinlan, 2015)
  https://www.rulequest.com/gritbot-info.html
- Example: application to the NY taxi dataset (11 million trips)

```python
case 221568: [0.000]
    payment_type = 2 (6744614 cases, 100.00% '1')
    tip_amount > 0 [1.95]

case 447324: [0.000]
    payment_type = 2 (6744614 cases, 100.00% '1')
    tip_amount > 0 [2.66]

case 494846: [0.000]
    payment_type = 2 (6744614 cases, 100.00% '1')
    tip_amount > 0 [3.86]
```

Only tips paid by credit card are supposed to be included. payment_type = 2 indicates a cash payment.
4. Data Integration

- Bringing together data from a number of different sources to form data matrix $X$
- Example: eBird data of bird species counts at a given location (lat/lon) and time
- Want to add climate and habitat information extracted from GIS databases
- In eBird these are provided as averages at small, medium and large scales. Will need to do interpolation in order to achieve this
- When fusing multiple datasets which have partial overlap, we may have two or more values for a particular attribute that ought to be the same but are not
Record Linkage

- Aka entity disambiguation/linking, duplicate detection, ...
- Identify matching entries across different data sources
- Needed when entities do not have a common identifier across sources (database key, e.g. national insurance number)
- Not only for people, e.g. astronomical objects in different wavebands
- Rule-based and probabilistic methods
Ontology Based Data Access (OBDA)

Next few slides from Ian Horrocks (Oxford)
Motivation

- Huge **quantity** of data increasing at an exponential rate
- Identifying & accessing **relevant** data is of critical importance
- Handling data **variety & complexity** often turns out to be main challenge
- **Semantic Technology** can seamlessly integrate heterogeneous data sources
How Does it Work?

1. **Standardised language for exchanging data**
   - W3C standard for data exchange is **RDF**
   - RDF is a simple language consisting of `<S P O>` **triples**
     - for example `<eg:lan eg:worksAt eg:Oxford>`
     - all S,P,O are URIs or literals (data values)
   - **URIs** provides a flexible **naming scheme**
   - Data has a flexible structure, with no fixed **schema**
   - Set of triples can be viewed as a **graph**
How Does it Work?

Standardised language for exchanging data

- W3C standard for data exchange is RDF
- RDF is a simple language consisting of triples
- URIs provide a flexible naming scheme
- Data has a flexible structure, with no fixed schema
- Set of triples can be viewed as a graph

<table>
<thead>
<tr>
<th>Triple</th>
<th>S</th>
<th>P</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>em1234</td>
<td>rdf:type</td>
<td>Person</td>
<td></td>
</tr>
<tr>
<td>em1234</td>
<td>name</td>
<td>“Eric Miller”</td>
<td></td>
</tr>
<tr>
<td>em1234</td>
<td>title</td>
<td>“Dr”</td>
<td></td>
</tr>
<tr>
<td>em1234</td>
<td>mailbox</td>
<td><a href="mailto:em@w3.org">mailto:em@w3.org</a></td>
<td></td>
</tr>
<tr>
<td>em1234</td>
<td>worksfor</td>
<td>w3c</td>
<td></td>
</tr>
<tr>
<td>w3c</td>
<td>rdf:type</td>
<td>organisation</td>
<td></td>
</tr>
<tr>
<td>w3c</td>
<td>hq</td>
<td>Boston</td>
<td></td>
</tr>
<tr>
<td>w3c</td>
<td>name</td>
<td>“W3C”</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
How Does it Work?

2 Standardised language for exchanging vocabularies/schemas
- W3C standard for vocabulary/schema exchange is OWL
- OWL provides for rich conceptual schemas, aka ONTOLOGIES

```
Heart ⊆ MuscularOrgan
∃isPartOf.CirculatorySystem
HeartDisease ≡ Disease
∃affects.Heart
VascularDisease ≡ Disease
∃affects.(∃isPartOf.CirculatorySystem)
```
How Does it Work?

3 Standardised language for querying **ontologies+data**

- W3C standard for querying is **SPARQL**
- SPARQL provides a rich query language comparable to SQL
  
  - ?x worksfor ?y .
  - ?y rdf:type organisation .
  - ?y hq Boston .
  
  - Select ?x
    where { ?x worksfor ?y .
    - ?y rdf:type organisation .
    - ?y hq Boston . }
  
  - Q(?x) ← worksfor(?x,?y) ∧ organisation(?y) ∧ hq(?y,Boston)
Semantic Technology

Rich **conceptual schemas** used to integrate heterogeneous sources

- **User Centric**
  - Schema modelled according to user intuitions
  - Independent of physical structure/storage of data

- **Declarative**
  - Improved understandability
  - Easier design, maintenance and evolution

- **Logic-based semantics**
  - Precise and formally specified meaning
  - Machine processable

- **Used at both design and query time**
  - Check validity and consequences of design
  - Easier query formulation and enriched query answers
5. Data Restructuring

- Sample rows of a table
- Project (delete columns)
- Feature construction (adding new cols as functions of existing ones)
- The tables we have may not be what we want, need to re-format
- Tidy data (Hadley Wickham, 2014)
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table
- Tidy Data, Hadley Wickham, J. Statistical Software 59(10), 2014
  https://www.jstatsoft.org/article/view/v059i10
### Tidy data: example

<table>
<thead>
<tr>
<th>religion</th>
<th>&lt;$10k</th>
<th>$10-20k</th>
<th>$20-30k</th>
<th>$30-40k</th>
<th>$40-50k</th>
<th>$50-75k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agnostic</td>
<td>27</td>
<td>34</td>
<td>60</td>
<td>81</td>
<td>76</td>
<td>137</td>
</tr>
<tr>
<td>Atheist</td>
<td>12</td>
<td>27</td>
<td>37</td>
<td>52</td>
<td>35</td>
<td>70</td>
</tr>
<tr>
<td>Buddhist</td>
<td>27</td>
<td>21</td>
<td>30</td>
<td>34</td>
<td>33</td>
<td>58</td>
</tr>
<tr>
<td>Catholic</td>
<td>418</td>
<td>617</td>
<td>732</td>
<td>670</td>
<td>638</td>
<td>1116</td>
</tr>
<tr>
<td>Don’t know/refused</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>11</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>Evangelical Prot</td>
<td>575</td>
<td>869</td>
<td>1064</td>
<td>982</td>
<td>881</td>
<td>1486</td>
</tr>
<tr>
<td>Hindu</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>34</td>
</tr>
<tr>
<td>Historically Black Prot</td>
<td>228</td>
<td>244</td>
<td>236</td>
<td>238</td>
<td>197</td>
<td>223</td>
</tr>
<tr>
<td>Jehovah’s Witness</td>
<td>20</td>
<td>27</td>
<td>24</td>
<td>24</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>Jewish</td>
<td>19</td>
<td>19</td>
<td>25</td>
<td>25</td>
<td>30</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.
WHAT WOULD AN AUTOMATIC STATISTICIAN DO?
GOALS OF THE AUTOMATIC STATISTICIAN PROJECT

- Provide a set of tools for understanding data that require minimal expert input

- Uncover challenging research problems in e.g.
  - Automated inference
  - Model construction and comparison
  - Data visualisation and interpretation

- Advance the field of machine learning in general
Ingredients of an Automatic Statistician

- An open-ended language of models
  - Expressive enough to capture real-world phenomena...
  - ...and the techniques used by human statisticians
- A search procedure
  - To efficiently explore the language of models
- A principled method of evaluating models
  - Trading off complexity and fit to data
- A procedure to automatically explain the models
  - Making the assumptions of the models explicit...
  - ...in a way that is intelligible to non-experts
Four additive components have been identified in the data:

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.
We want your messy data sets!

- We are hoping to create a set of challenges which include all the day-to-day problems that beset the typical data scientist.
- Examples collected so far at https://alan-turing-institute.github.io/turing-tests/
- We welcome new examples—we need:
  - Data that is *publically available*
  - A clear *analysis task* associated with the data
  - If possible the scripts *etc* used to carry out the data wrangling
  - If available a paper giving details of the task and analysis
- **Contact:** James Geddes jgeddes@turing.ac.uk
Summary

- Common view that up to 80% of work on a data mining project is involved in **data understanding** and **data preparation**
- The AI for Data Analytics team seeks to reduce the time and effort needed for these phases by developing advanced tools to propose fixes
- There are a diverse set of issues covering data parsing, data understanding, data cleaning, data integration, data restructuring and more
- Aim to build an interactive assistant that will step the analyst through all the issues in the current dataset
- Need to learn from past experience, i.e. across datasets
- We welcome your input (as per the previous slide)