Towards Automating the Data Analytics Process

Chris Williams
with James Geddes, Zoubin Ghahramani, Ian Horrocks, Charles Sutton

April 2017

Figure credit: Aaron Bacall

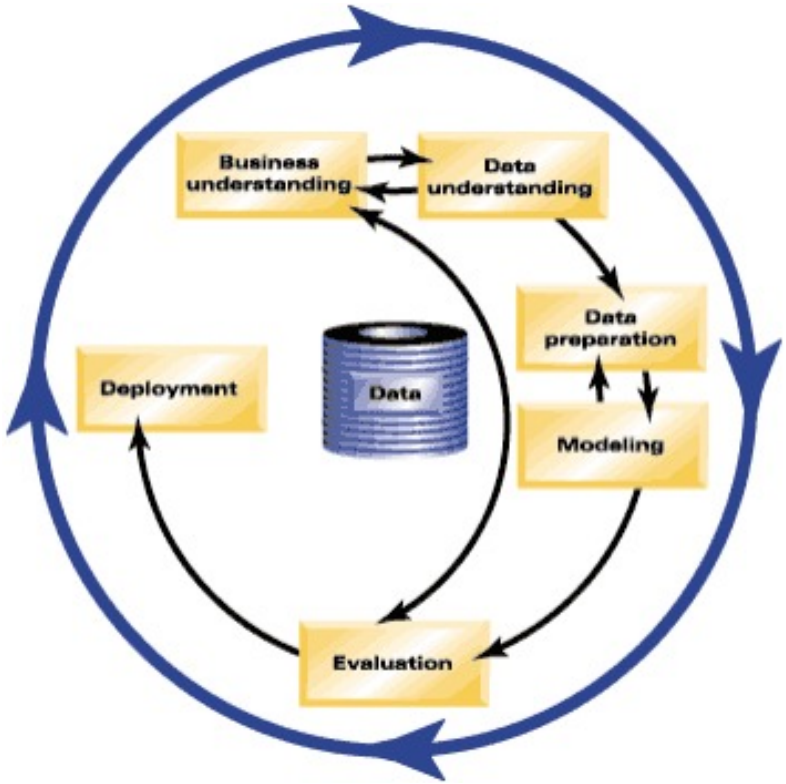
Towards Automating the Data Analytics Process

- 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?
- Our work complements the Automatic Statistician (Ghahramani et al), which is more concerned with the search for appropriate analysis models given clean data
- AI for Data Analytics project at the Alan Turing Institute. Initial funding from Lloyds Register Foundation, new funding from the ATI and the ATI Defence and Security Partnership: **hiring for 3 postdoc positions**

Outline

- CRISP-DM Methodology
- Five aspects of Data Wrangling
  1. Data Parsing
  2. Data Understanding
  3. Data Cleaning
  4. Data Integration
  5. Data Transformation
- The Automated Statistician
- We want your messy data sets!

CRISP-DM Methodology



Cross Industry Standard Process for Data Mining
Six Phases:

- Problem Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

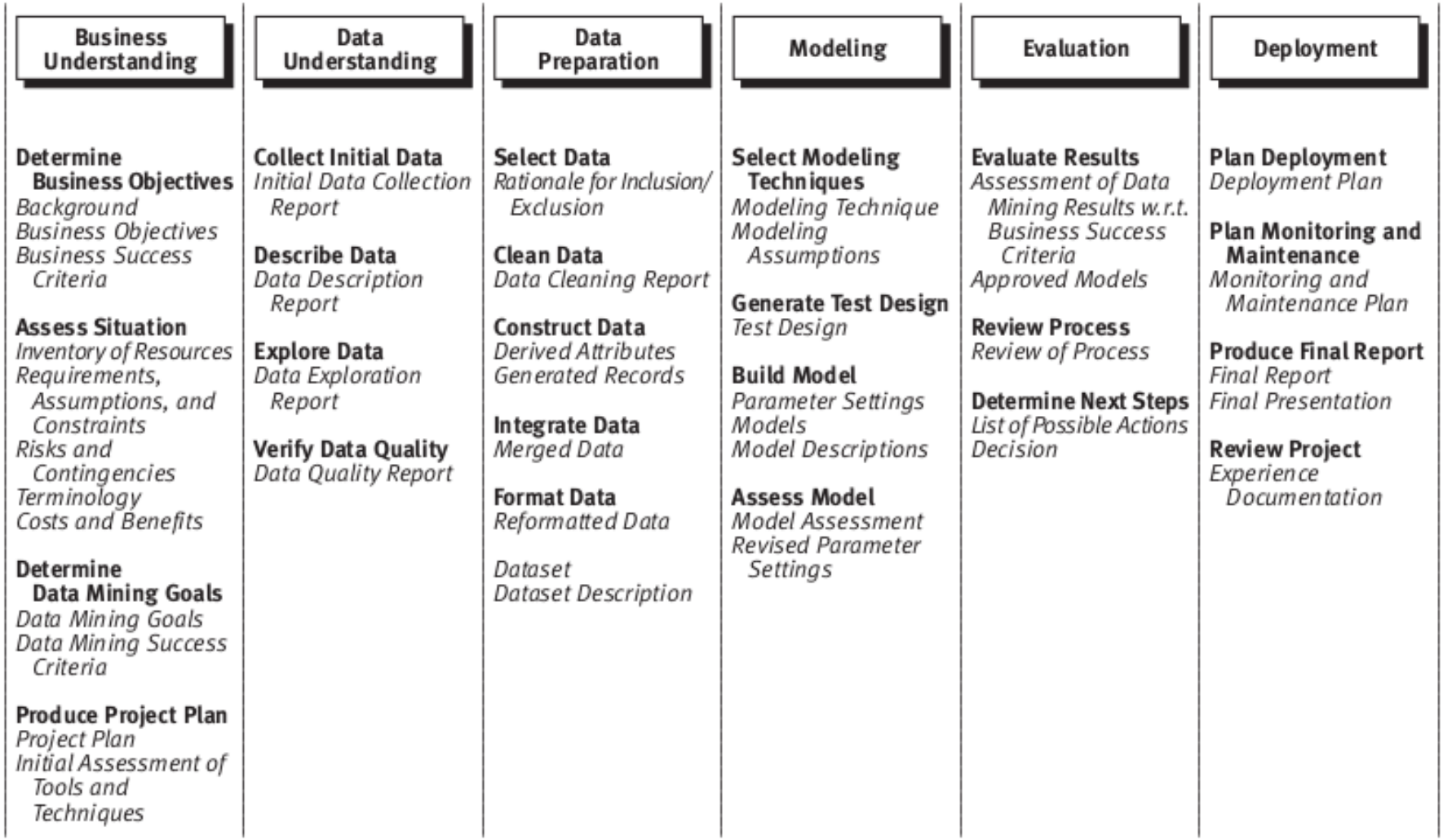


Figure credit: CRISP-DM 1.0 Step-by-step data mining guide, Chapman et al, 2000

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

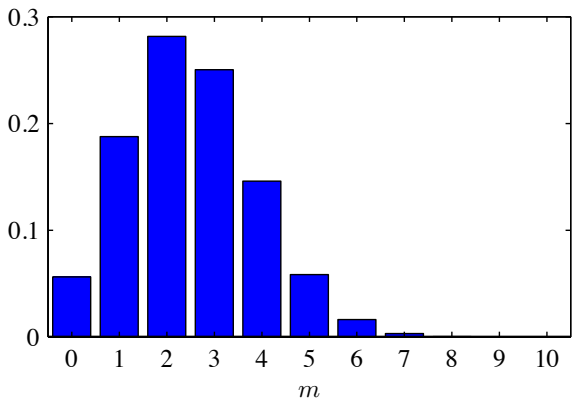
1. Data Parsing
2. Data Understanding
3. Data Cleaning
4. Data Integration
5. Data Transformation

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` 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



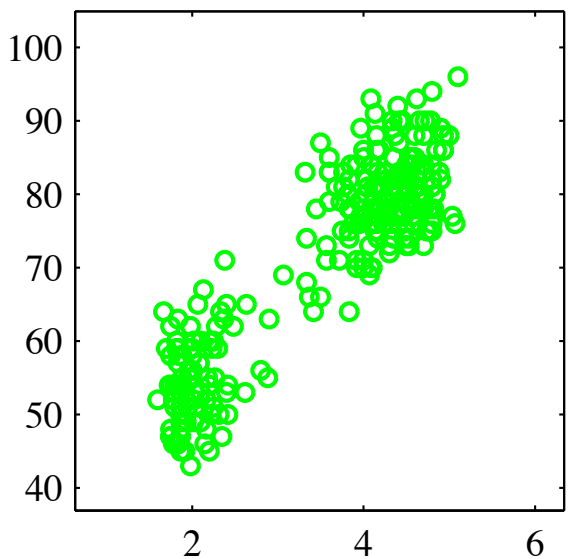


Figure credit: Chris Bishop PRML 2006

3. Data Cleaning

- Handling missing data
- Entity disambiguation
- Detecting anomalies

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 ... DANGER!!
- Why is data missing? Is it *missing at random* (MAR) or is there a systematic reason for its absence?

Let $\mathbf{x}_m$ denote those values missing, and $\mathbf{x}_p$ those values that are present.

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

- Model $p(\mathbf{x}_m|\mathbf{x}_p)$ 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)
- Reference: *Statistical Analysis with Missing Data* R. J. A. Little, D. B. Rubin, Wiley (1987)

Entity disambiguation

- Do “IBM”, “I.B.M.” and “IBM corp” refer to the same entity?
- “18/7/16” vs “18 July 2016”
- Aka record linkage, duplicate detection, ...
- Identify matching entries across data sources
- Needed when entities do not have a common identifier across sources (database key, e.g. national insurance number)
- Above examples need string matching, and special purpose handling of dates etc.
- Openrefine `openrefine.org` provides some good functionality for such tasks

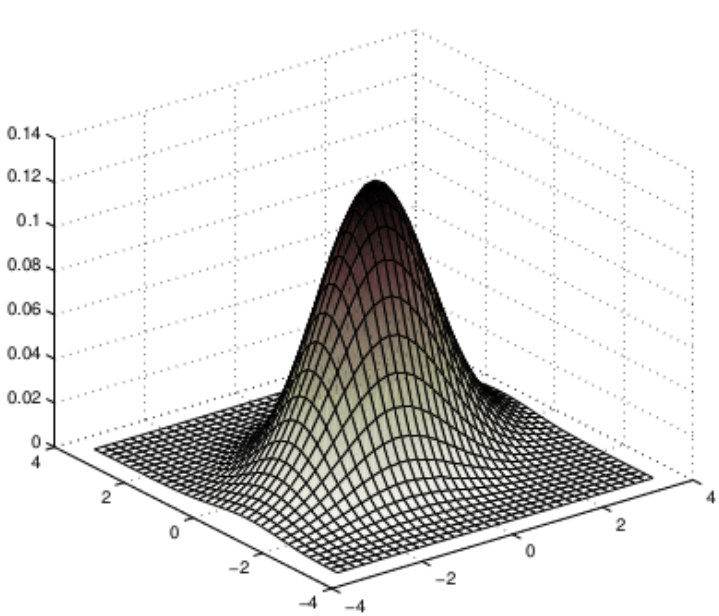
Record Linkage

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17

- Not only for people, e.g. astronomical objects in different wavebands
- Rule-based and probabilistic methods
- Fellegi IP & Sunter AB. A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210 (1969)

Anomaly Detection

- "An anomaly is defined as a pattern that does not conform to expected normal behavior" (Chandola, 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)



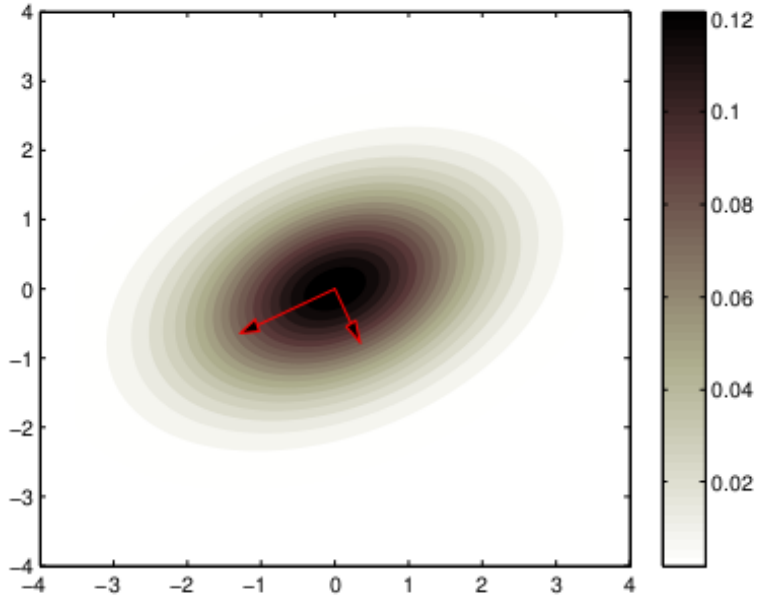


Figure credit: David Barber, BRML, 2012

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)

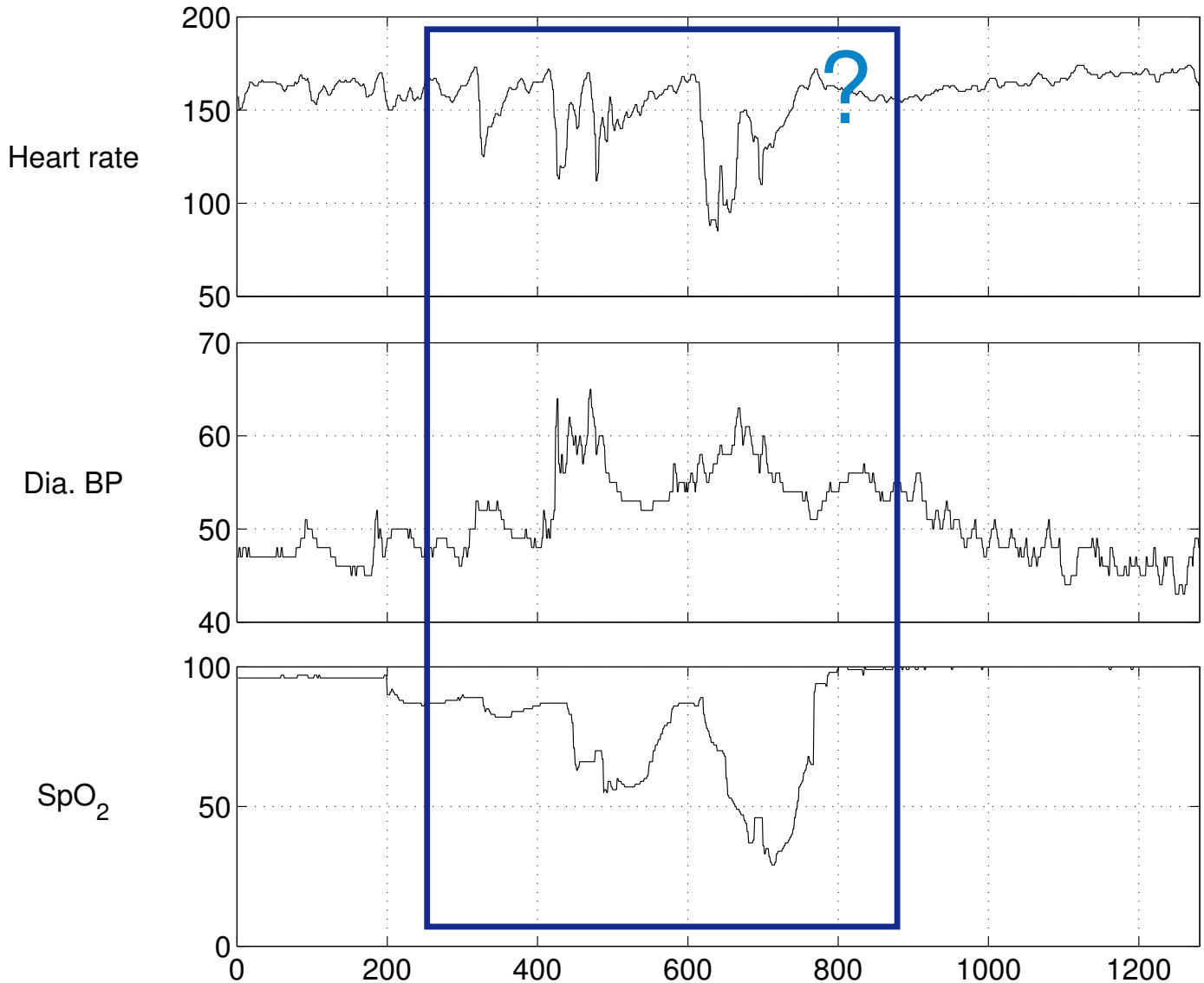
```
case 221568:  [0.000]
        payment_type = 2  (6744614 cases, 100.00% '1')
            tip_amount > 0 [1.95]

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]
```

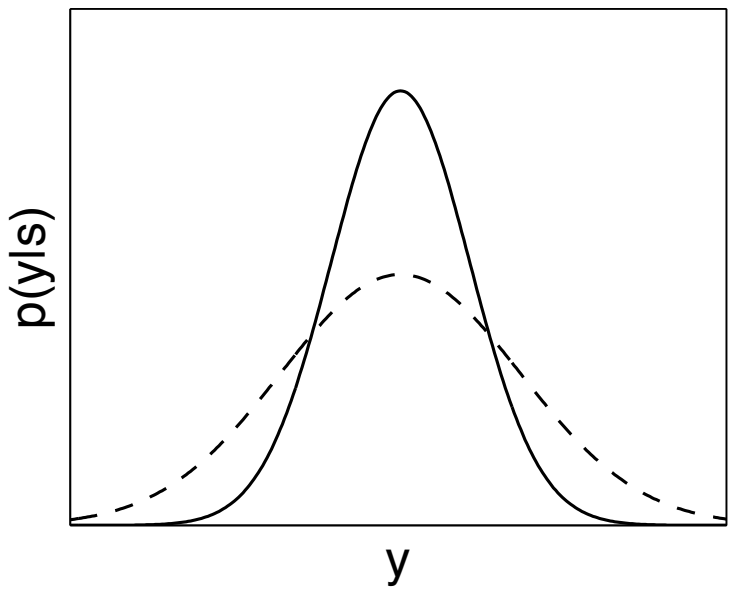
Only tips paid by credit card are supposed to be included.
`payment_type = 2` indicates a cash payment.

Anomaly Detection in Time Series



X-factor for static 1-D data

- For static data, we can use a model $\mathcal{M}_*$ representing 'abnormal' data points.



- The high-variance model wins when the data is not well explained by the original model

X-factor for dynamic data

Quinn, Williams and McIntosh (2009)

- Model 'normal' data with a linear dynamical system (LDS)

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \\ \mathbf{y}_t &\sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R}) \end{aligned}$$

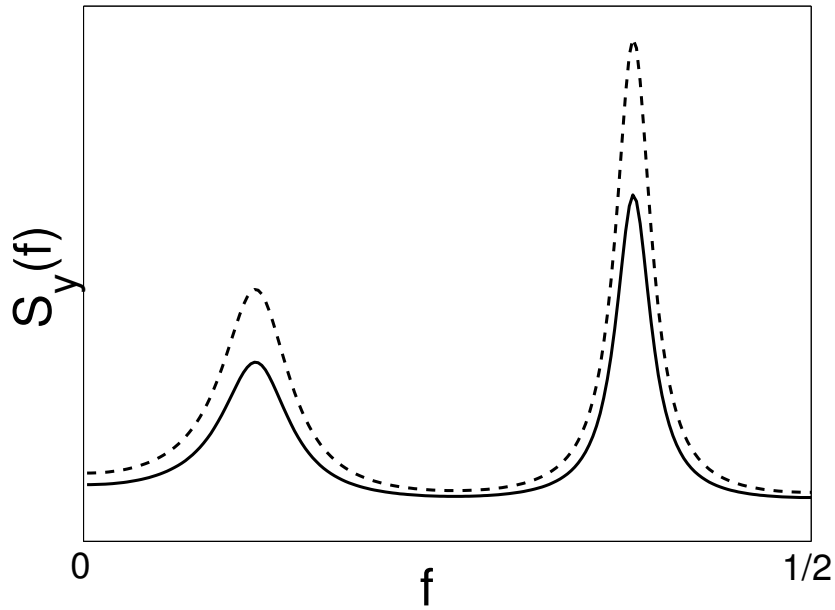
- Can construct an 'abnormal' dynamic regime analogously:

$$\text{Normal dynamics: } \{\mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}\}$$

$$\text{X-factor dynamics: } \{\mathbf{A}, \xi\mathbf{Q}, \mathbf{C}, \mathbf{R}\}, \quad \xi > 1.$$

- Build a switching LDS model that has a variable s_t at each time t that can be inferred from the time series data

Spectral view of the X-factor



- Plot shows the spectrum of a hidden AR(5) process, and accompanying X-factor
- More power at every frequency
- Dynamical analogue of the static 1-D case

4. Data Integration

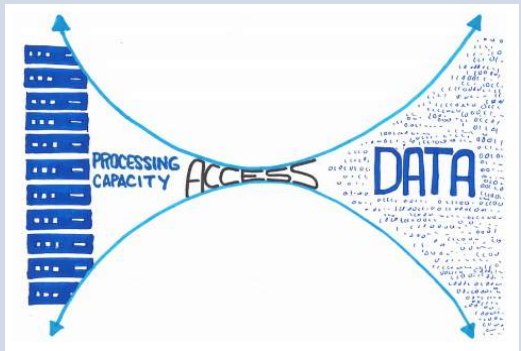
- Bringing together data from a number of different sources to form data matrix X
- Example: eBird data of bird species counts at a given location (lat/lon) and time
- Want to add climate and habitat information extracted from GIS databases
- In eBird these are provided as averages at small, medium and 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

Ontology Based Data Access (OBDA)

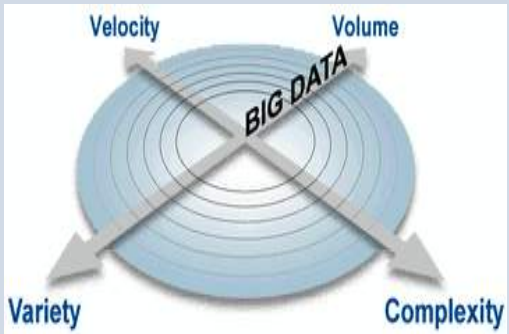
Next few slides from Ian Horrocks (Oxford)

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:Ian 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?

1 **Standardised language for exchanging data**

- W3C s
- RDF i triples
  - fo
  - all
- URIs
- Data h a
- Set of

Triple		
S	**P**	**O**
em1234	rdf:type	Person
em1234	name	“Eric Miller”
em1234	title	“Dr”
em1234	mailbox	mailto:em@w3.org
em1234	worksfor	w3c
w3c	rdf:type	organisation
w3c	hq	Boston
w3c	name	“W3C”
...	...	...

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**

$$\begin{aligned} \mathsf{Heart} &\sqsubseteq \mathsf{MuscularOrgan} \sqcap \\ &\quad \exists \mathsf{isPartOf}.\mathsf{CirculatorySystem} \\ \mathsf{HeartDisease} &\equiv \mathsf{Disease} \sqcap \\ &\quad \exists \mathsf{affects}.\mathsf{Heart} \\ \mathsf{VascularDisease} &\equiv \mathsf{Disease} \sqcap \\ &\quad \exists \mathsf{affects}.(\exists \mathsf{isPartOf}.\mathsf{CirculatorySystem}) \end{aligned}$$

DEPARTMENT OF COMPUTER SCIENCE

Information Systems Group

EPSRC
Engineering and Physical Sciences Research Council

Optique

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

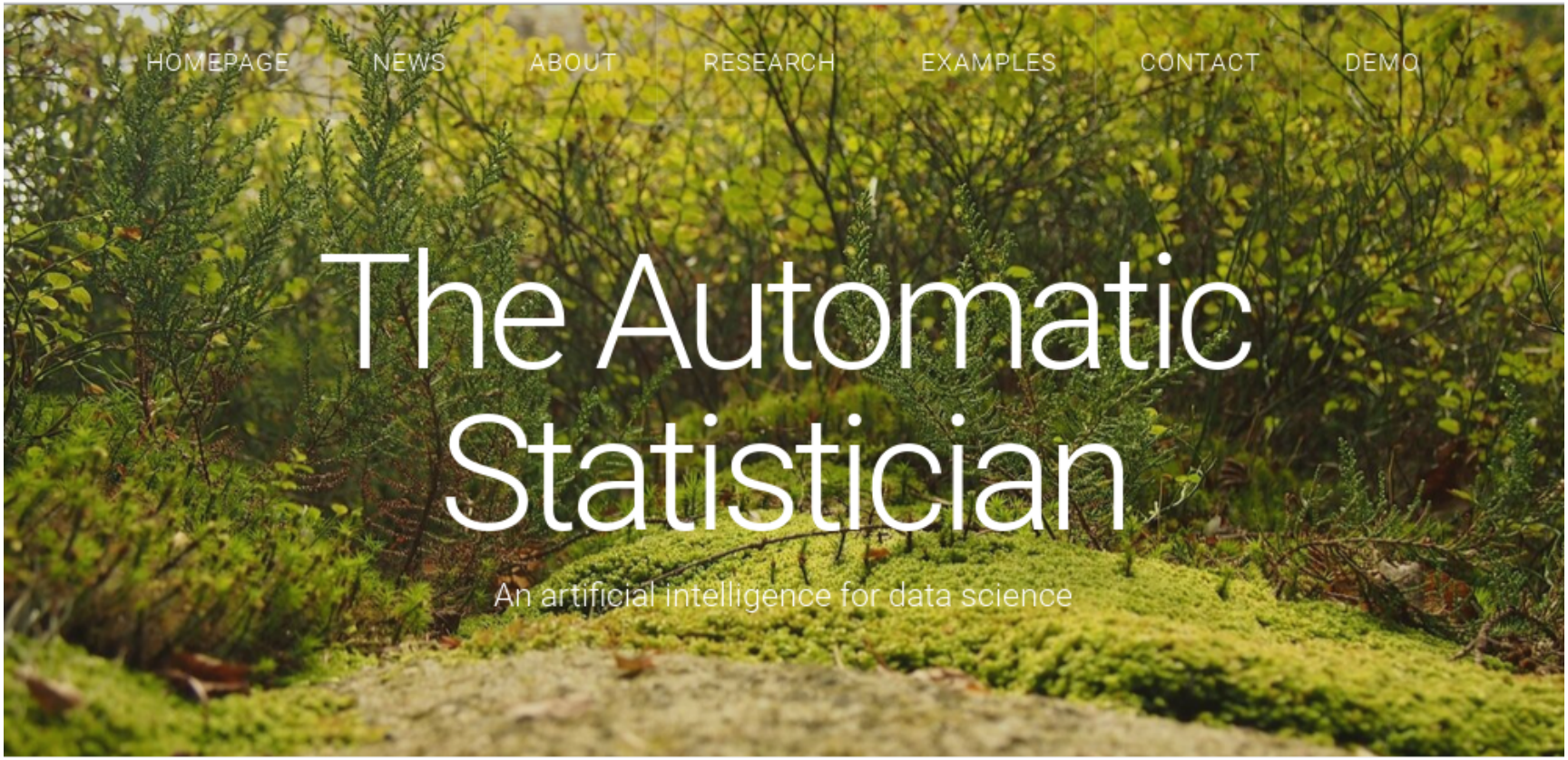
religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244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

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.

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

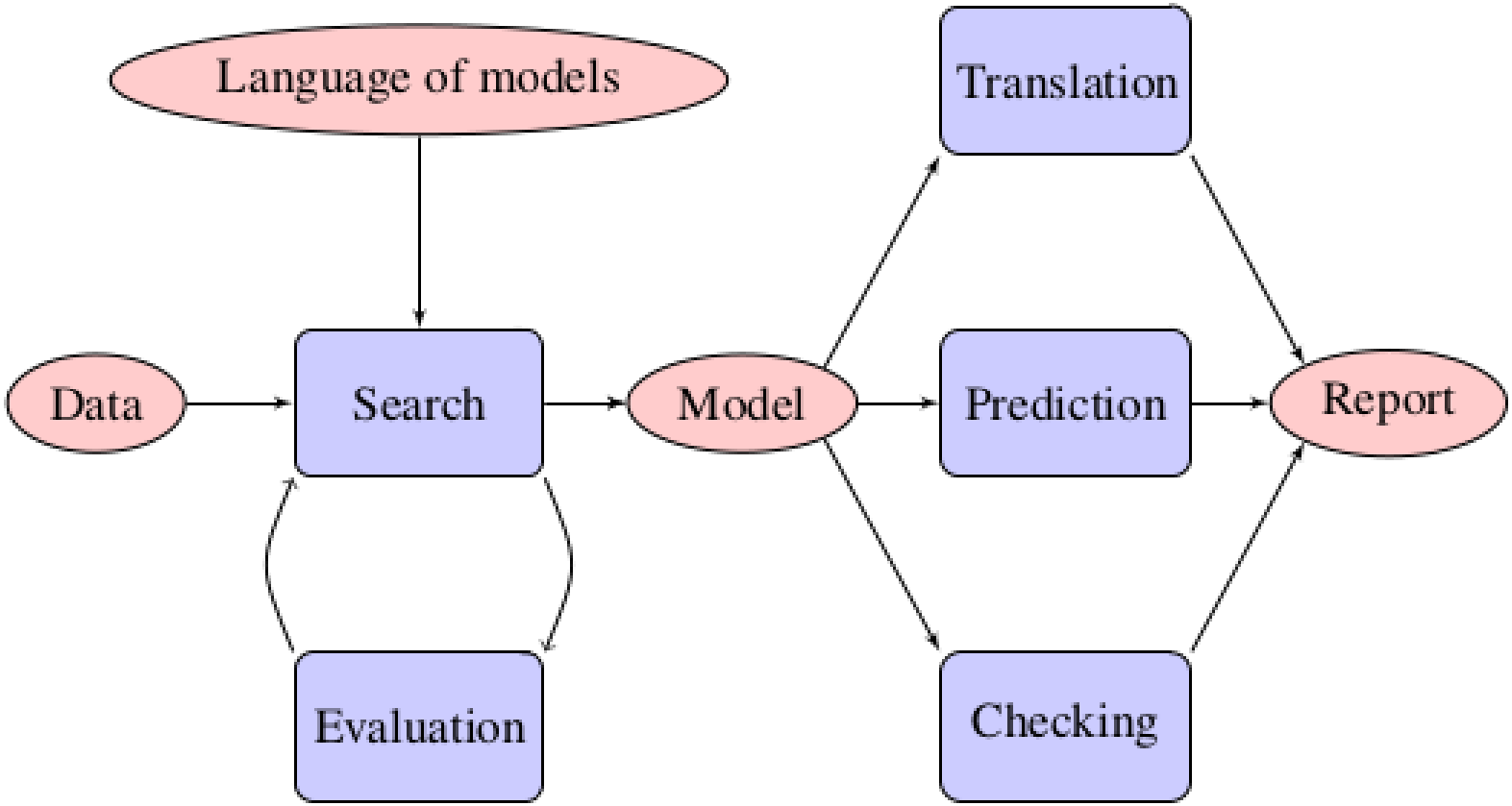
The Automated Statistician

`https://www.automaticstatistician.com/`



Next few slides from Zoubin Ghahramani

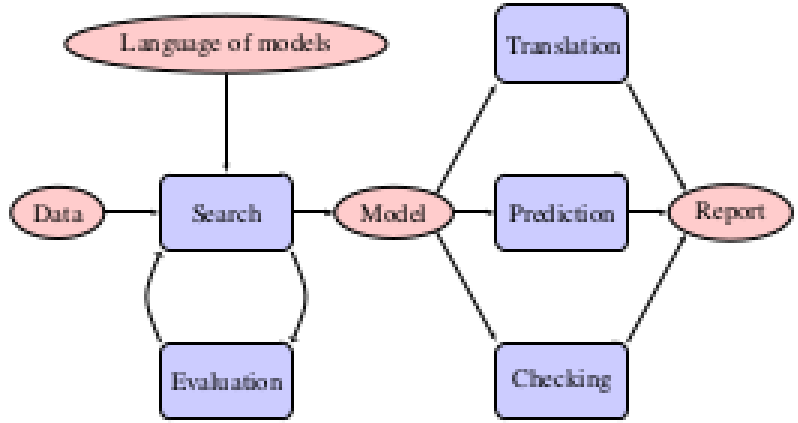
WHAT WOULD AN AUTOMATIC STATISTICIAN DO?



GOALS OF THE AUTOMATIC STATISTICIAN PROJECT

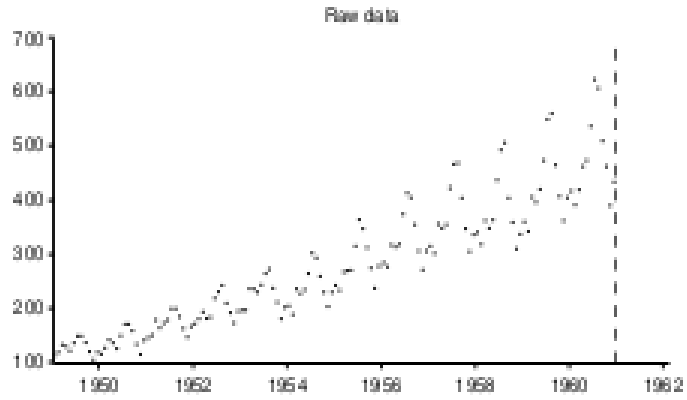
- Provide a set of tools for understanding data that require minimal expert input
- Uncover challenging research problems in e.g.
  - Automated inference
  - Model construction and comparison
  - Data visualisation and interpretation
- Advance the field of machine learning in general

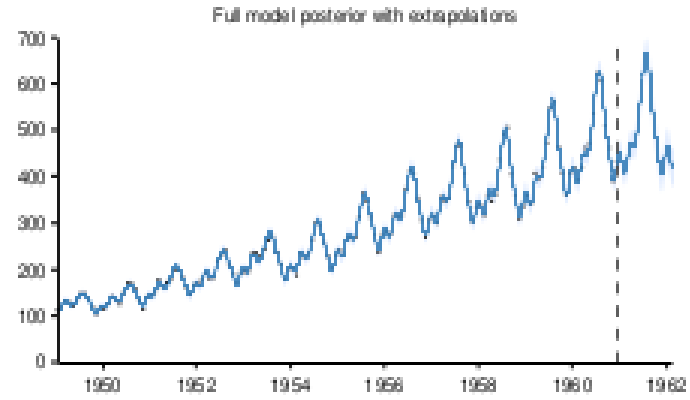
INGREDIENTS OF AN AUTOMATIC STATISTICIAN



- **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

PREVIEW: AN ENTIRELY AUTOMATIC ANALYSIS





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/wrangling-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 transformation 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)

Hiring Postdocs

- The AIDA team at Turing is hiring for **three** postdocs in the areas of
  - Data Acquisition and Transformation
  - Data Understanding
  - Data Quality and Cleaning
- If you are interested please contact Chris Williams `ckiw@inf.ed.ac.uk` or James Geddes `jgeddes@turing.ac.uk`