Fundamental Concepts in Data Insight:

Data Insight

Fundamentals for a General Audience

QA Ltd. owns the copyright and other intellectual property rights of this material and asserts its moral rights as the author. All rights reserved.

Data Insight

- Discuss: Why now?
  - Case Study: Predictive Policing
  - Where does the insight boom come from?
  - When did data insight start?
  - When did organizations start being data-driven?
- What is Data Analysis, Science and Insight?
  - What are traditional data practices?
  - What are the new insight practices?
- What is Data?
  - What is Insight?
  - What is inference?
  - How do you get Data Insight from Raw Data?
- What is a typical data insight workflow?
  - What techniques and technologies support Insight?
  - What tools does data insight require?
- What roles do data insight functions require?
  - How do you set up a data insight project?
  - What questions should I ask of a data project?
  - How do data practicioners solve problems?
- WB. What is Probability?
  - Review: What is probability?
  - WB. Case Study: Forensic Risk Analysis
- Group Discussion & Reflection

Overview

Discuss: Why now?

Case Study: Predictive Policing

> *The London MET developed the Gang Matrix to identify potential gang members and score them according to the risk they pose to society. Research by Amnesty International (2018) and Scott (2018) revealed the discriminatory nature of this predictive identification program, in which the majority of individuals were young black men. The MET was ordered to radically reform the matrix within a year by the Mayor of London, and are currently working on a new program called the 'Concern Hub' (Mayor or London, 2018a: Dodd, 2018; Crisp, 2019).*

- How could we approach automated risk profiling?
  - risk of reoffence, victimization, gang membership

Where does the insight boom come from?

- big data preceeds insight
  - computational power secondary
  - statistical techniques incredibly simple
- big data sources

When did data insight start?

- ancient?
    - babylon
    - tycho brahe
    - francis bacon
- "big data"
    - late 90s/early 2000s
- "data science"
    - 2010s

When did organizations start being data-driven?

- depends on market, organization type
    - governments tried mid 90s to 2010s
- many analytical problems require necessarily speculative methods
    - data relevant to problem needs to exist

What is Data Analysis, Science and Insight?

- data analysis = fact finding
- predictive analysis = predictive **associations**
- science = predictive **explanations**
- insight = actionable conclusions

What are traditional data practices?

- business intelligence
  - data analysis
  - relational
  - historical
  - factual
- What are traditional practices used for?
  - operational & retail data
  - analysis of historical trends
  - monitoring system performance
- analysis is passed to executive decision making for insight
  - data warehousing -> reporting -> business analysts
  - business analysts -> executives -> decisions

What are the new insight practices?

- data science & big data
    - predictive
    - explanatory
    - non-relational
    - future
    - probabilistic
- insight is the result of automated processes:
    - operational data -> predictive system -> realtime insight actions
    - data warehouse -> explanatory system -> generated insight reports
        - -> refined by scientists, analysts, ...
    - reports -> data scientists -> executives

What is Data?

- the measurement of an observable variable
  - age, sales, height
- a human categorization of bundles of such measurements
  - sex, product, city
- What is the value of Data?
  - evidence-based action
- How do you get data?
  - experiment
  - operational systems
  - external sources

What is Insight?

- levels of insight:
  - descriptive
  - descriptive & characterising
  - diagnostic, narrowly explanatory
  - predictive
    - associative (automation value)
    - **explanatory** (insight value)
  - **prescriptive**

What is inference?

- inference, def.
  - drawing conclusions
    - ... from data
- inference can be,
  - description which generalises
  - diagnosis
  - predictive associations
  - predictive explanations
- all inferential questions are probabalistic
  - (reliable,) reporting yields true facts
  - inference yields *claims* ("conclusions") which have a probability of being true

How do you get Data Insight from Raw Data?

- How does data become suitable for analysis?
  - raw data ->
  - clean data ->
  - problem-structured data ->
  - enriched data
  - ... -> analysis

What is a typical data insight workflow?

- obtain data (raw data)
- explore data (clean data)
- transform data
    - ETL transformation (problem-structured)
    - modelling transformation (enriched)
- model data
- evaluate
- deploy
- monitor

What techniques and technologies support Insight?

- Machine Learning (Low Quality, Automated)
  - Deep Learning (often what's meant by "Artifical Intelligence")
- Analytics (Medium Quality, Manual)
  - Fact-Finding
  - Diagnostics
- Statistical Modelling (High Quality, Automated/Manual Mix)
  - Experiment & Data Science
- Big Data
  - High complexity data sources

What tools does data insight require?

- data analysis
  - spreadsheets
  - notebooks
- data engineering
  - the cloud
  - databases & data systems
  - ETL tools
- automation, insight programming
  - the cloud
  - code editors

What roles do data insight functions require?

- technical
  - software engineer
  - data engineer
  - data analyst
- strategic-technical
  - data scientist
- what is a data scientist?
  - business analyst
  - project leader
  - software engineer
  - data analyst
  - data engineer

How do you set up a data insight project?

- identify problem
  - validate problem exists
  - illustrate with data
  - formulate hypothesis
- consider solution space
  - validate possible solution(s)
    - show they could work
    - formulate hypotheses
  - validate ROI on possible solution(s)
- define criteria for success
  - qualitative criteria
  - measurable metics

What questions should I ask of a data project?

- How do existing systems/approaches/conditions fail?
- What problem are we trying to solve?
- How will we know when it is solved?
- What existing systems/people does the probelm concern?
- Who does the problem impact (and how)?
- What possible solutions are there?
- How would we select between solutions?

How do data practicioners solve problems?

- investigation: what is the problem?
    - Q. What steps are relevant?
- ideation: what could we do to solve it?
    - Q. What steps are relevant?
- solution (design): what would a plausible solution look like?
    - Q. What steps are relevant?
- solution (build): how do we build it?
    - Q. What steps are relevant?

How do data practicioners solve problems?

- investigation: what is the problem?
    - problem understanding
        - domain
        - objective
        - impact
    - quantification & measurement
    - data exploration
- ideation: what could we do to solve it?
    - subproblems identified
    - strategies to solve subproblems / problems
    - hyopothesis generation
- solution (design): what would a plausible solution look like?
    - experimental design
    - product/implementation design
- solution (build): how do we build it?
    - ...

WB. What is Probability?

WB. What is Probability?

Review: What is probability?

- All probabilities are *conditional* (on what you know)
  - there is no such thing as **a** probability without qualification
  - nature has no probabilities

- $P(Claim|Evidence)$ aka. $P(Hypothesis|Evidence)$
  - note: mathematical notation *is just a shorthand for ordinary language!*

- This notation describes *a single number* between 0 and 1
  - $P(\ldots)$ means "probability of"
  - $P(X)$ means "probability of X" (here we *assume* some background information we're aware of)
  - $P(X|Y)$ means "probability of X given that we know that Y"
- eg.,
  - $P(RainTomorrow|RainingNow)$
  - $P(CrimeRates = High|LocalSurveyofResidents)$

Review: What is Probability?

- Common convention, $P(Y|X)$
  - Y is what we are trying to understand (we don't know it, it's **unobservable**)
  - X is what we know (its direclty **observable**)

- Something is *random* **relative to a claim** if knowing it makes no difference to the claim,
- $P(Coin = Heads|MyHappiness) = P(Coin = Heads)$
  - knowing how happy I am makes no difference to the outcome of the coin flip
  - almost nothing affects coin flips, this is why we regard them as random
  - $P(Coin = Heads|Heads - is - Magenet, IhaveaMagnet)$ is **not** $P(Coin = Heads)$!
- $P(CrimeisHigh) = P(CrimeisHigh|RecentNewspaperFrontpage)$
  - The probability *judgement* that crime is high *not knowing* the frontpage **is the same as** knowing it

Whiteboard: Case Study: Forensic Risk Analysis

- Can we use behaviour characteristics of suspects to automate a risk assement?
    - quick decision in the field based on local systems which can advise HIGH/LOW risk
- Solution
    - $Y \in$ HIGH, LOW
    - X observable features
        - age
        - location (ward boundary index, or postcode = latlong)
        - clothing colours (, gang identifiers, etc.)
        - ...
    - plot X_{age} vs *history* of Y
        - interpolate `S`-curve
        - read-off probability on the vertical
- Goal of case study:
    - illustrate the use of probability
    - illustrate the role of data and relationships in computing probabilties
- Refence Terms:
    - Risk Analysis
    - Binary Classification
    - Logistic Regression
    - Machine Learning

WB. Forensic Risk

WB. Forensic Risk

Group Discussion (25 min)

- Consider a data problem from your own work environment
  - What is the role of data here?
    - What are the data assets?
    - Where do they come from?
    - How are they used?
  - Brainstorm the ways data insight could solve the problem

- What skills does data insight require?
  - Consider:
    - mindset (eg., creative)
    - technical skills (eg., programming)
    - non-technical skills
    - technical knoweldge (eg., statistics)
    - non-technical knowldge
  - Consider:
    - How the roles mentioned help deliver insight?
    - What skills does each use?

WB. Review

WB. Review

Reflection

- The History of Data Insight
- Data Analysis vs. Data Science
  - The Hierachy of Interpretation
  - Traditional vs. Big Data
- Roles, Skills & Organizational Challenges