BUILDING FLEXIBLE DATA SCIENCE TEAMS

TERADATA. THE BEST DECISION POSSIBLE™

DATA SCIENCE VS BIG DATA

TERADATA. THE BEST DECISION POSSIBLE™

- There are many competing definitions of big data
  - 3/4/5 Vs
- For every definition there is an expert saying that it doesn't exist, or is irrelevant
- My definition is simple, and has two parts:
  - Big data is the raw material of **data science**
  - **Big** refers to the new importance of data

What is Data Science?

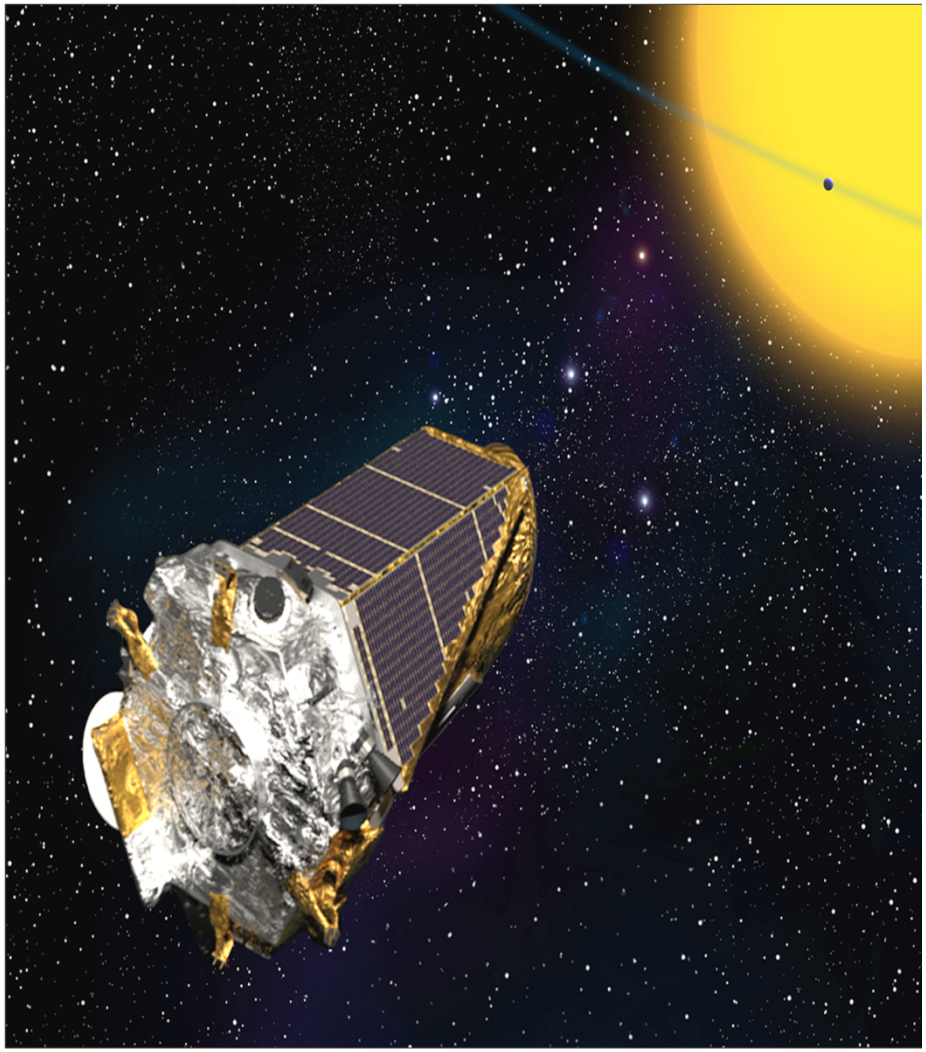
- Data Science is a discipline initially defined by a scientific approach to data
  - > In big science projects 60—70% of budgets were being spent dealing with data
    - – This was usually handled by the most junior researcher
  - > Dealing with data that was increasingly 'Big'
  - > Used to be cheaper to re-run an experiment than store and reuse date
    - – Large Hadron Collider
- Data scientists have certain features/behaviours that differentiate them from other disciplines

Data Science at work: The Kepler Mission

- What fraction of sun-like stars in our galaxy host potentially habitable Earth-size planets?

- Big Data:
  - >150,000 target stars
  - $6x10^6$ pixels collected and stored per ½ hour
  - ~40 GB downlinked each month
  - $>40\times10^9$ points in the time series over 3.5 years

- Big Processing Challenges
  - Instrument effects are large compared to signal of interest
  - Observational noise is non-white and non-stationary
  - ~100×106 tests per star for planetary signatures [O(N2)]
  - Stellar variations are higher than expected

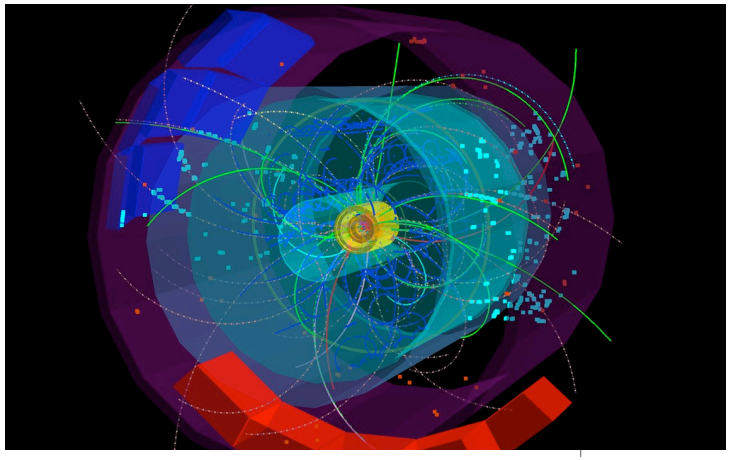


Learning 1 Data Science is an approach

- Data Science is about an attitude and approach to Big Data and data analysis

- It recognises an experimental approach to problem solving
  - Not dissimilar to data mining
  - Has similar issues around methodology

- Data scientists are more likely to take an end-to-end approach to problems

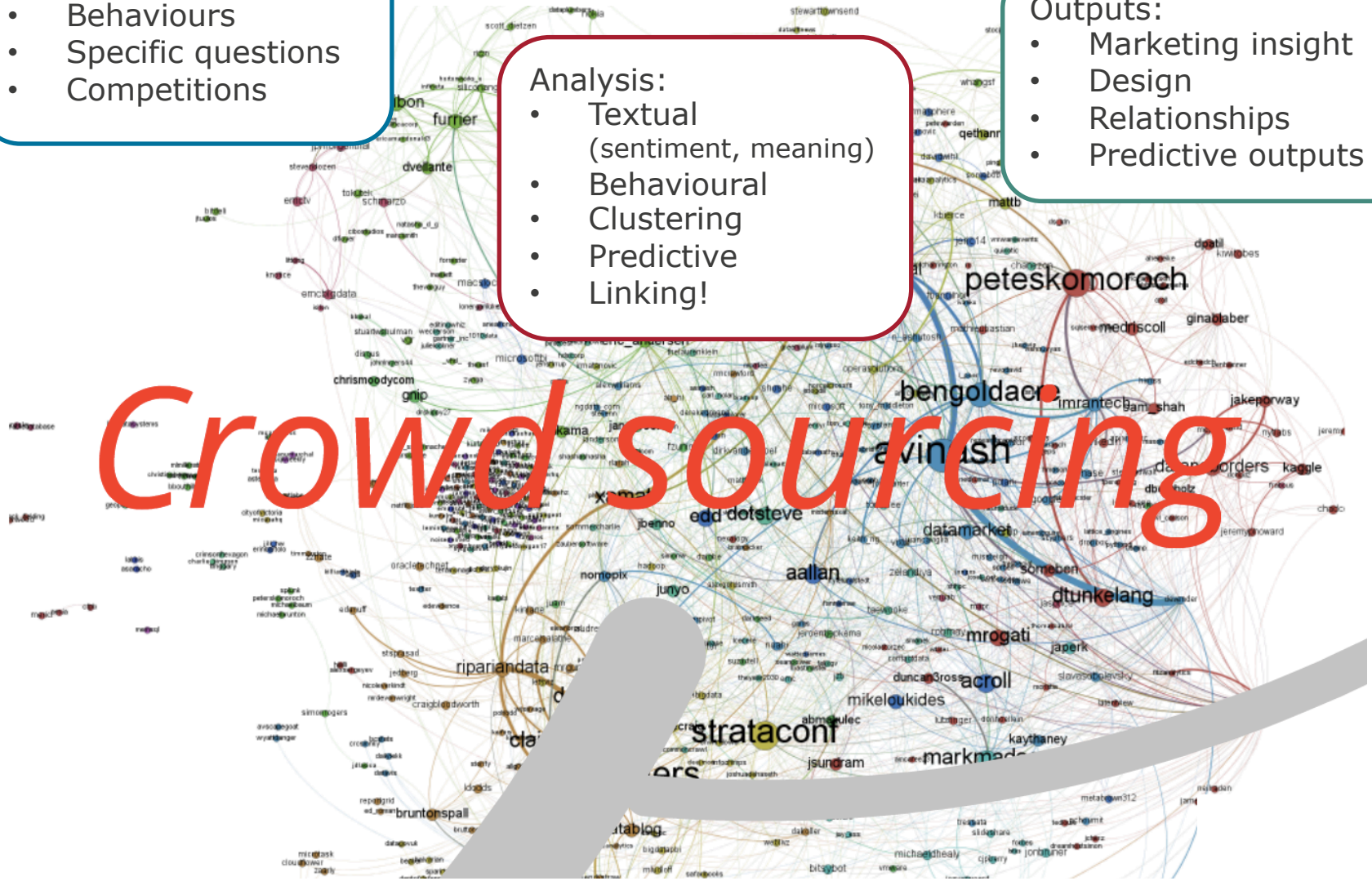


Examples of Data Science analytics

Data in:
- Social media
- Behaviours
- Specific questions
- Competitions

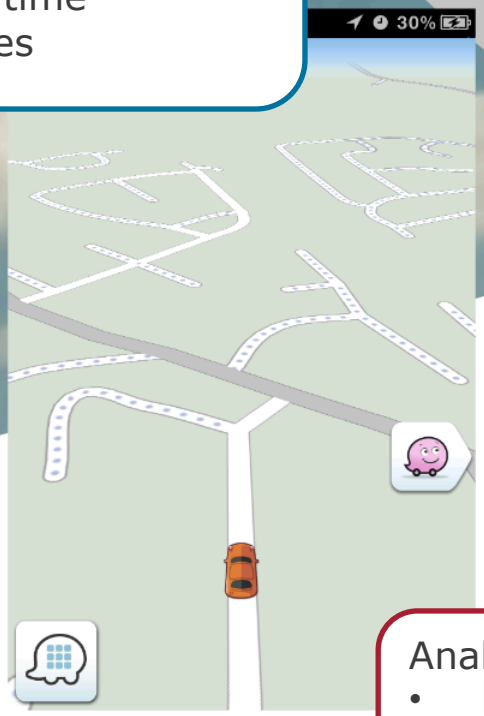
Analysis:
- Textual (sentiment, meaning)
- Behavioural
- Clustering
- Predictive
- Linking!

Outputs:
- Marketing insight
- Design
- Relationships
- Predictive outputs



Doing more with location

Data in:
- Location
- Passive/active
- Real-time
- Routes



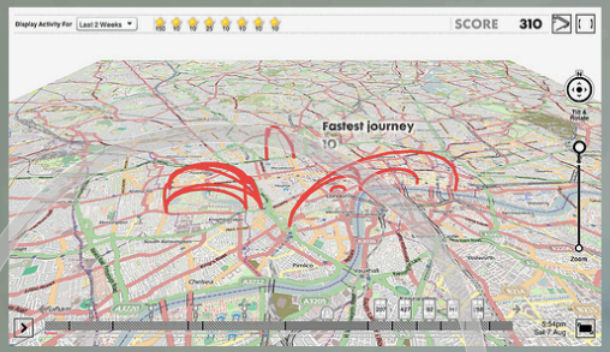
Location

Outputs:
- Mobile digital marketing
- Network analysis
- Partner offers
- Insight

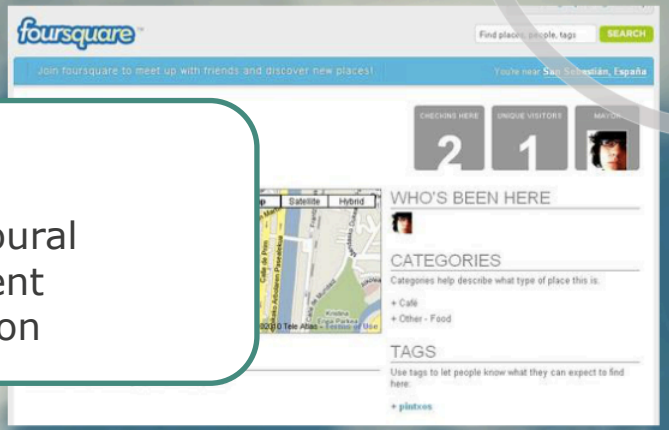
Analysis:
- Paths
- Behavioural
- Relational

Playing games with customers

Data in:
- Crowdsourced
- Apps
- Usage



Gamification



Outputs:
- Games!
- Behavioural alignment
- Education

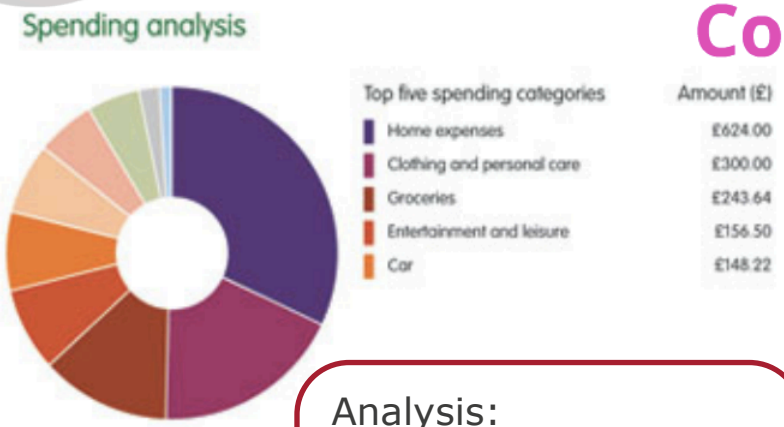
Analysis:
- More data!
- Behavioural
- Clustering
- Predictive

Giving customers control

Data in:
- Company provided
- Customer provided
- 3rd party

Quantified self

Outputs:
- Data and analysis to customer
- Solutions to third parties



Consumer data locker

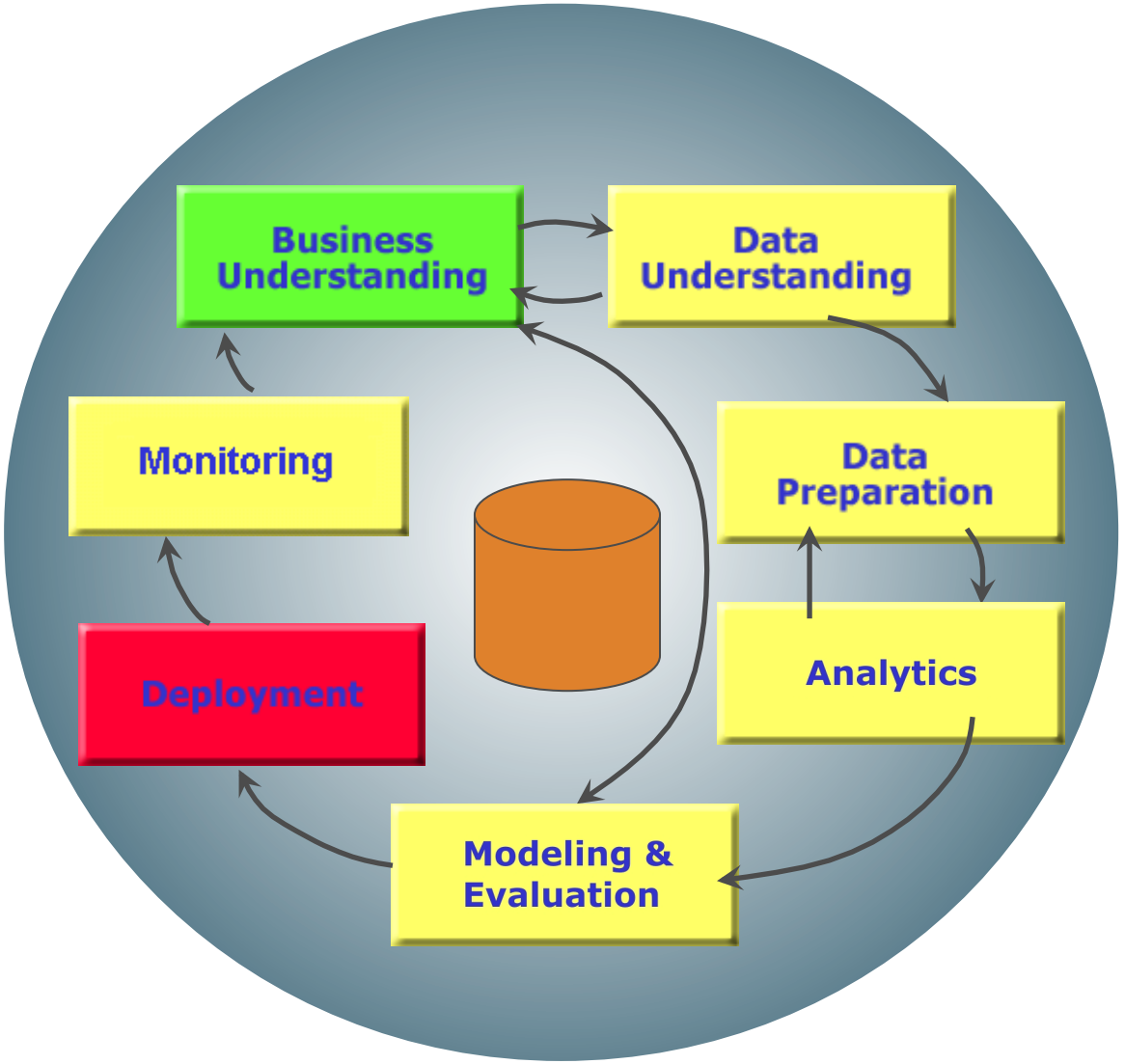
Analysis:
- Data quality
- Wider range of variables
- Supported user self-analysis

WHO ARE DATA SCIENTISTS?

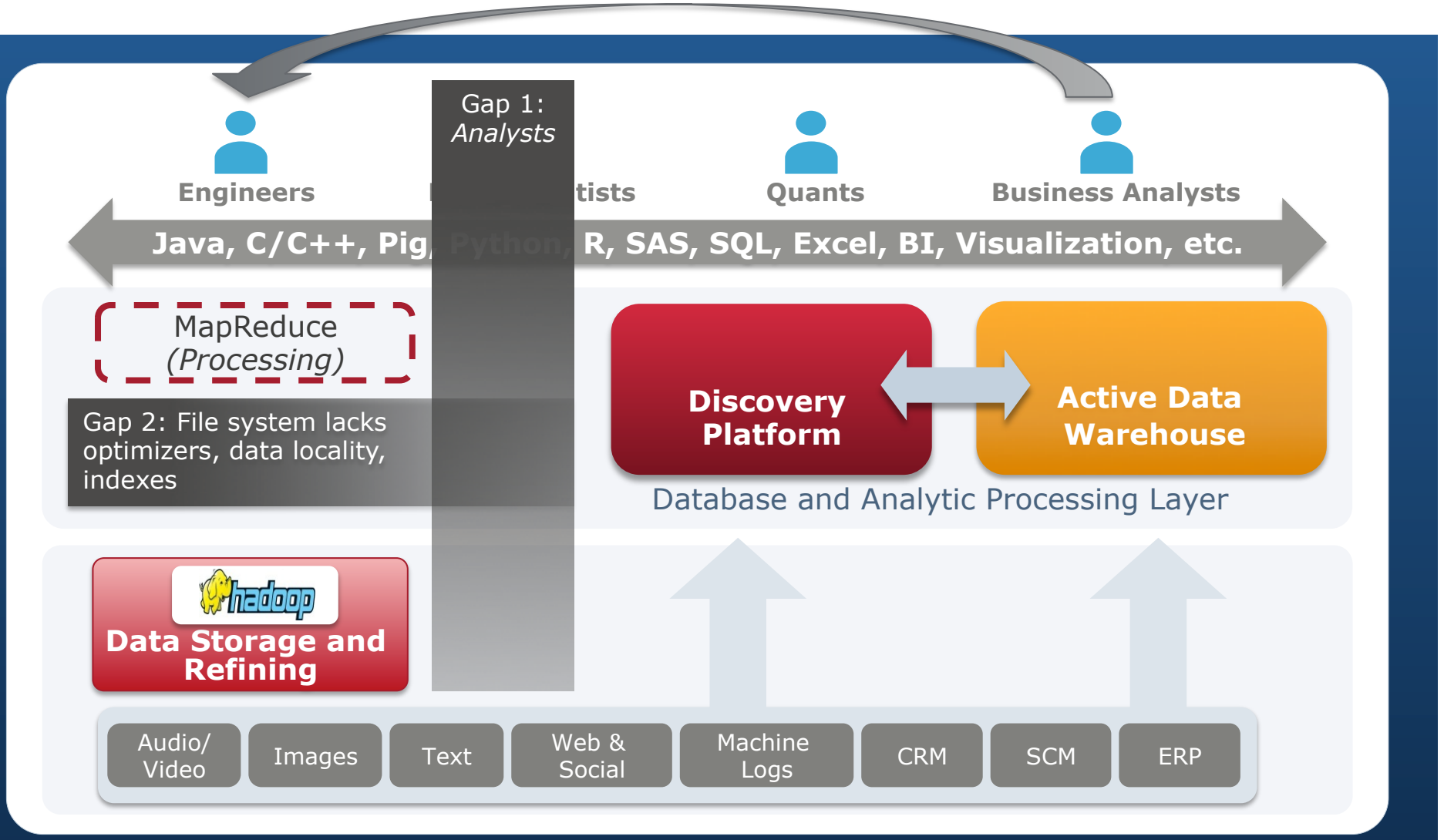
Learning 2 Data science types

- Followers of the yellow elephant
  - > Because Hadoop is the single biggest influencer on data science to date
  - > R is not far behind
- Hackers
  - > In that they are likely to want to write code
- Data miners
  - > Because they want to understand causality
- Communicators
  - > Need visualisation and descriptive skills

CRISP-DM still relevant: mapping to skills



The Big Data Architecture Today Has Gaps



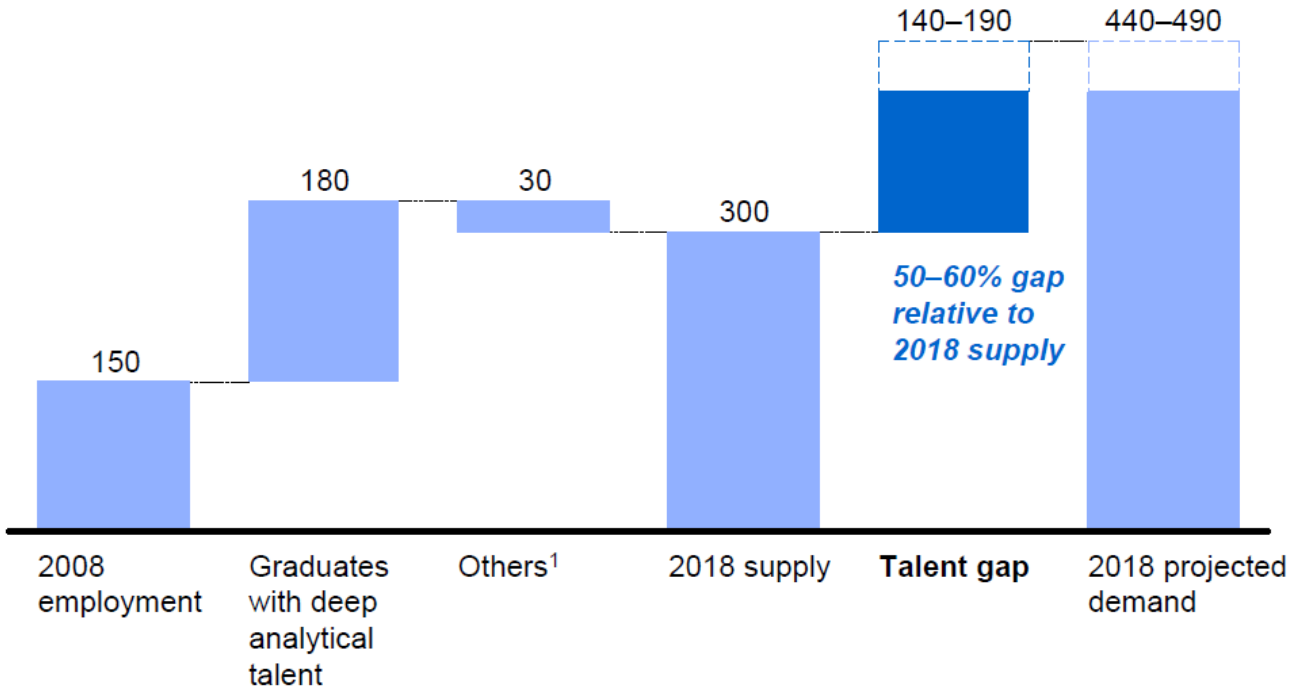
Where are the data scientists?

Exhibit 4

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



Source: McKinsey

Learning 3 Finding data scientists

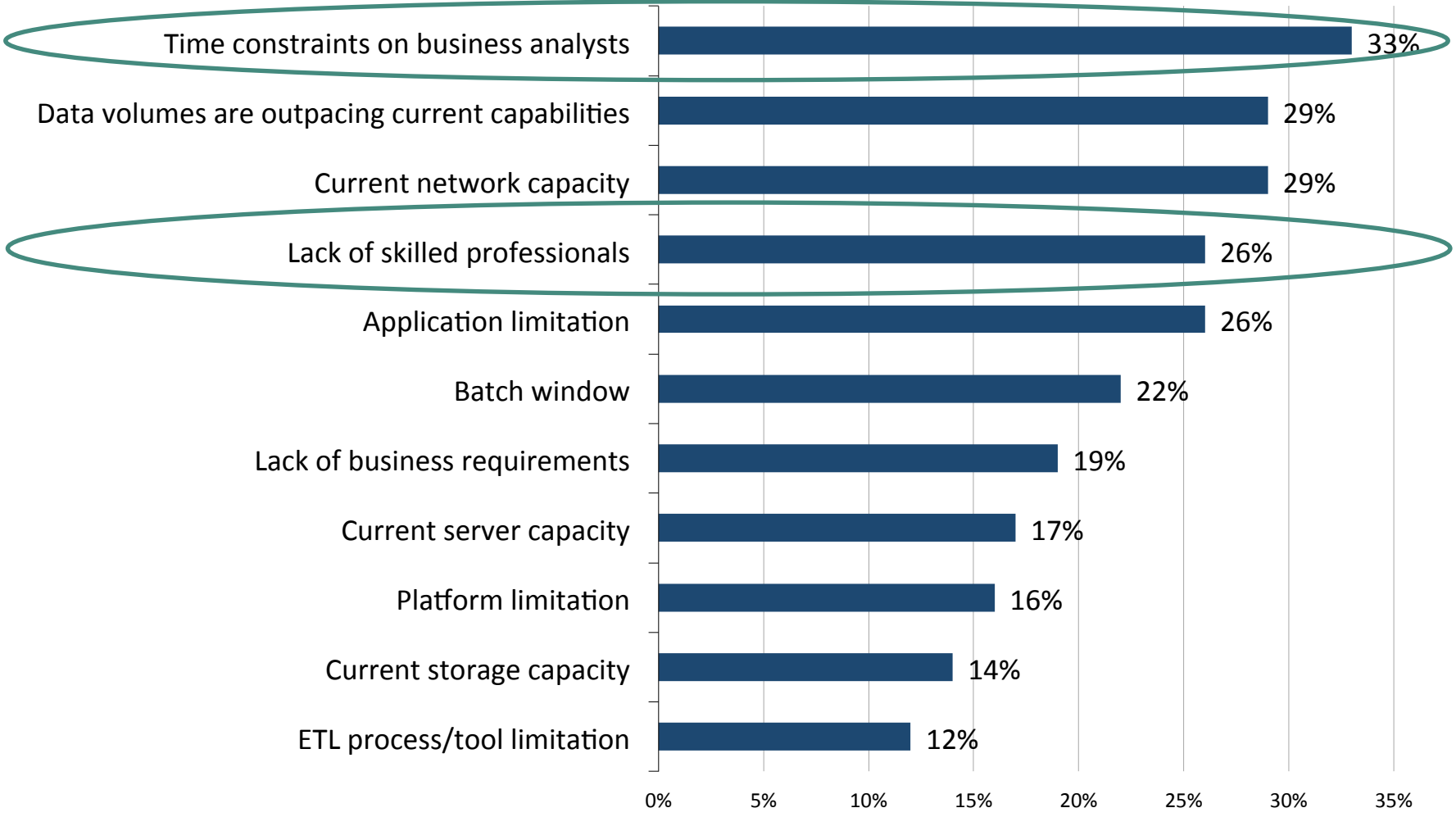
- Who exists in your organisation, and where do their skills map?
- How do you bridge the technical and business worlds?
- Where can you look for other resources?
  - > Universities?
  - > Internal?

BUILDING A DATA SCIENCE TEAM

TERADATA. THE BEST DECISION POSSIBLE™

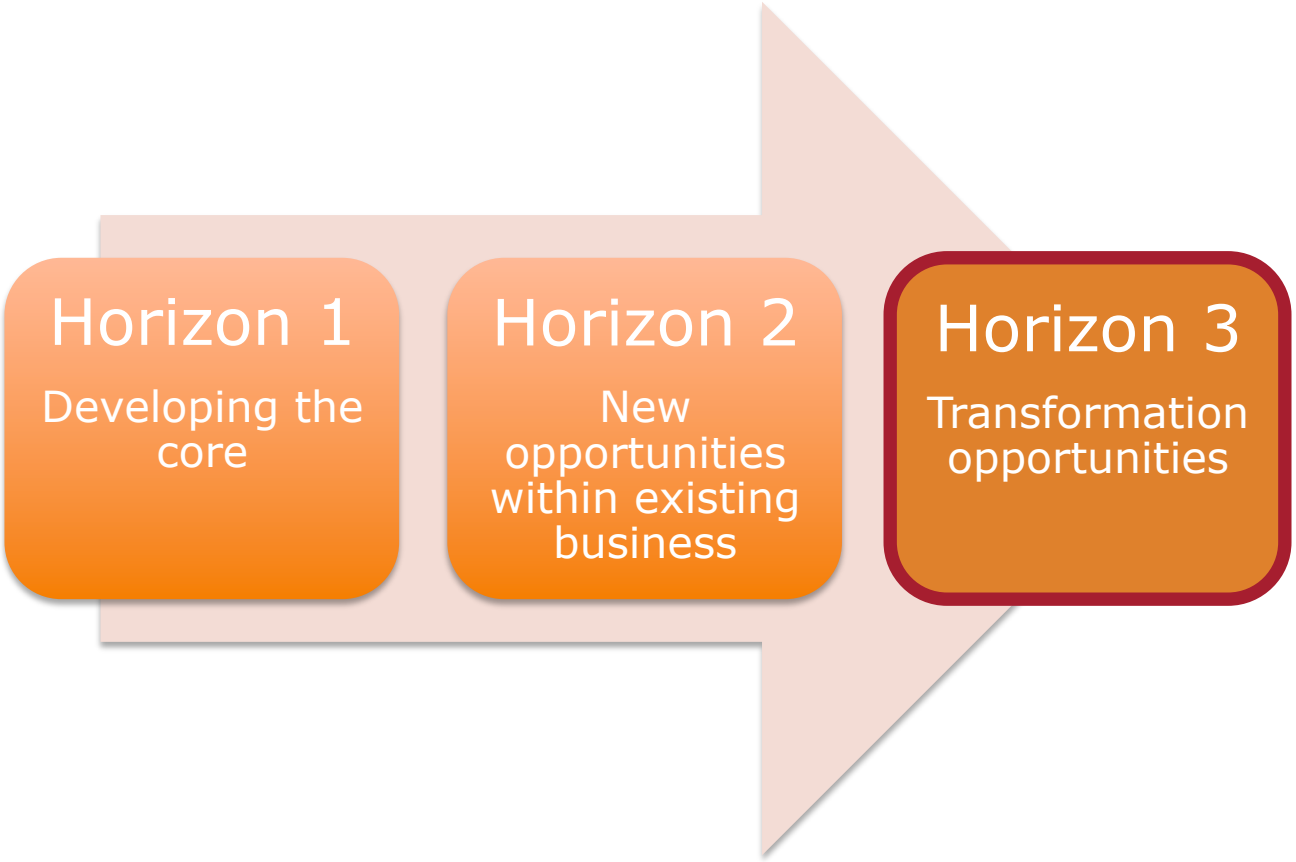
What is preventing your organization from conducting analytics on its largest data set more frequently?

(Percent of respondents, N=103, multiple responses accepted)



Source: Enterprise Strategy Group; April 5, 2012

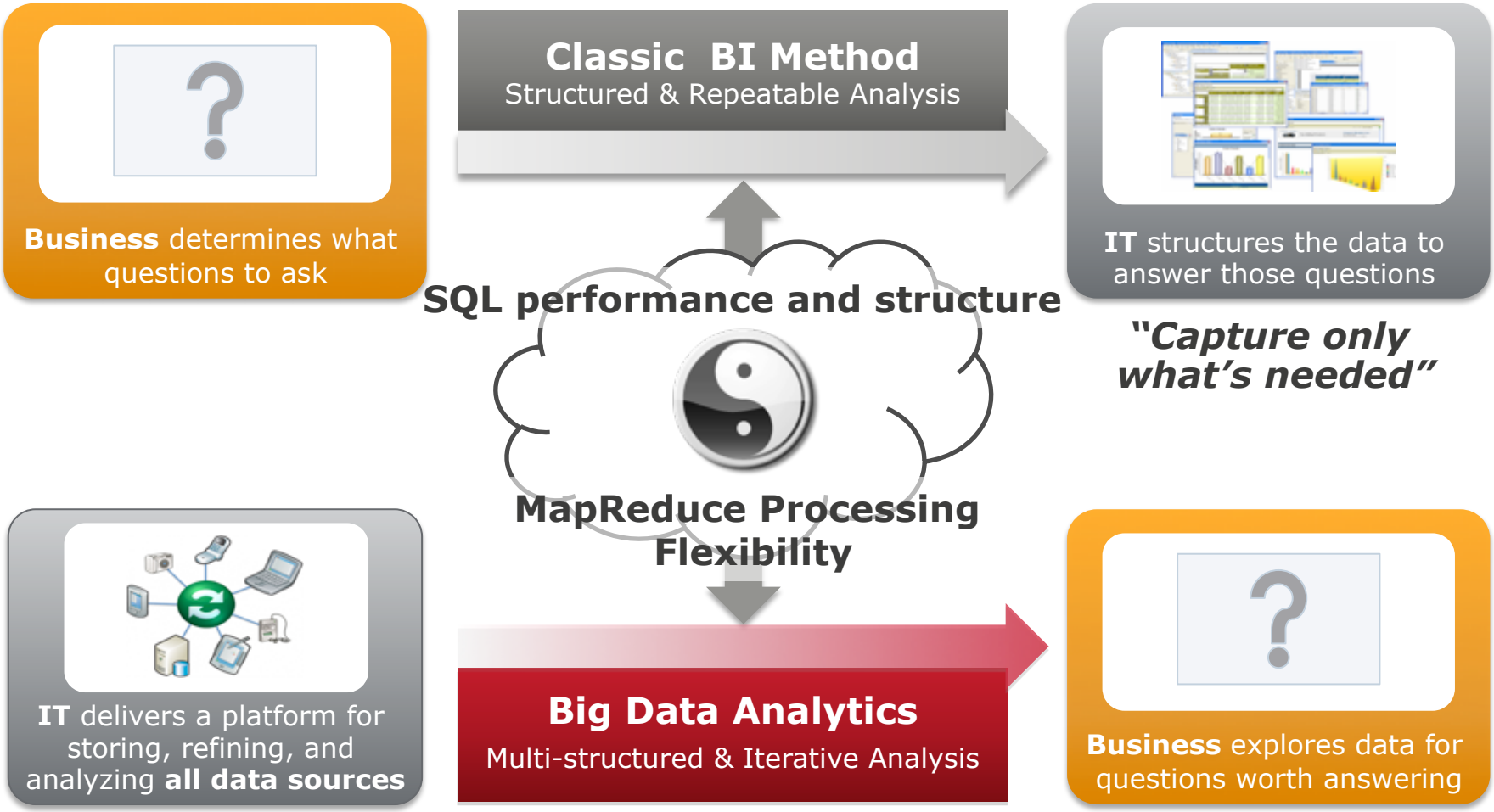
First goal should be innovation



Source: Steve Coley of McKinsey

The Data Science Process

Bridging Classic BI & Big Data Analytics Worlds



Fail slow: an example

- Data Innovation Group at company X
- Goal:
  > develop new, data focused products
- Initial success:
  > multi million € product designed and created in month 2
- Second success:
  > no second success
- Reason:
  > initial product still being 'hand cranked' by DIG
  > DIG became a DOG

Learning 4 Key elements for a team

- People
  - > Mix of skills
  - > Technical
  - > Business
  - > Data mining
- Infrastructure
  - > Platform and tools
  - > Flexibility
- Approach
  - > Flexible but methodology led
  - > Fail fast!
- Culture
  - > Senior support
  - > Head room
  - > Innovative

Learning 5 Decision points

- Can you give your Data Science team room for Horizon 3?
- Can you give access to experimentation platforms?
- Can you tolerate failure – the *fail fast* approach?
  - > Can you capture learning
- Do you have senior management support?
- Can you ensure that deployment doesn't become a hand-crafted solution?
- Can you move to more agile analytical approaches?

THANK YOU

TERADATA. THE BEST DECISION POSSIBLE™