From Data Mining Processes to Data Science Trajectories

Peter Flach (EuADS President)

What is Data Science? (1/5)

- The Science *of* Data?
  - machine learning & statistics research
- Doing Science *with* Data?
  - data-intensive research
- Applying Science *to* Data?
  - data-intensive X

What is Data?

Data can broadly be described as factual information about events, situations, circumstances etc. in the world around us.

Data is often hailed as the "new oil" or the "new electricity", to highlight its transformative potential as a driver of prosperity in the twenty-first century.

Data ≠ Oil

However, the absence of a fixed physical form as well as data's often transient nature imply that

the potential for and challenges of collecting, processing, and exploiting data are incomparable with physical resources.

What is Knowledge?

While data is a vehicle for describing the world around us, *knowledge is the carrier of understanding*:

- (K in) Domain knowledge is indispensable for understanding the meaning of data and for processing and exploiting it in a productive way.
- (K out) Further knowledge is produced by applying analytics to data.

Data + Knowledge = Value

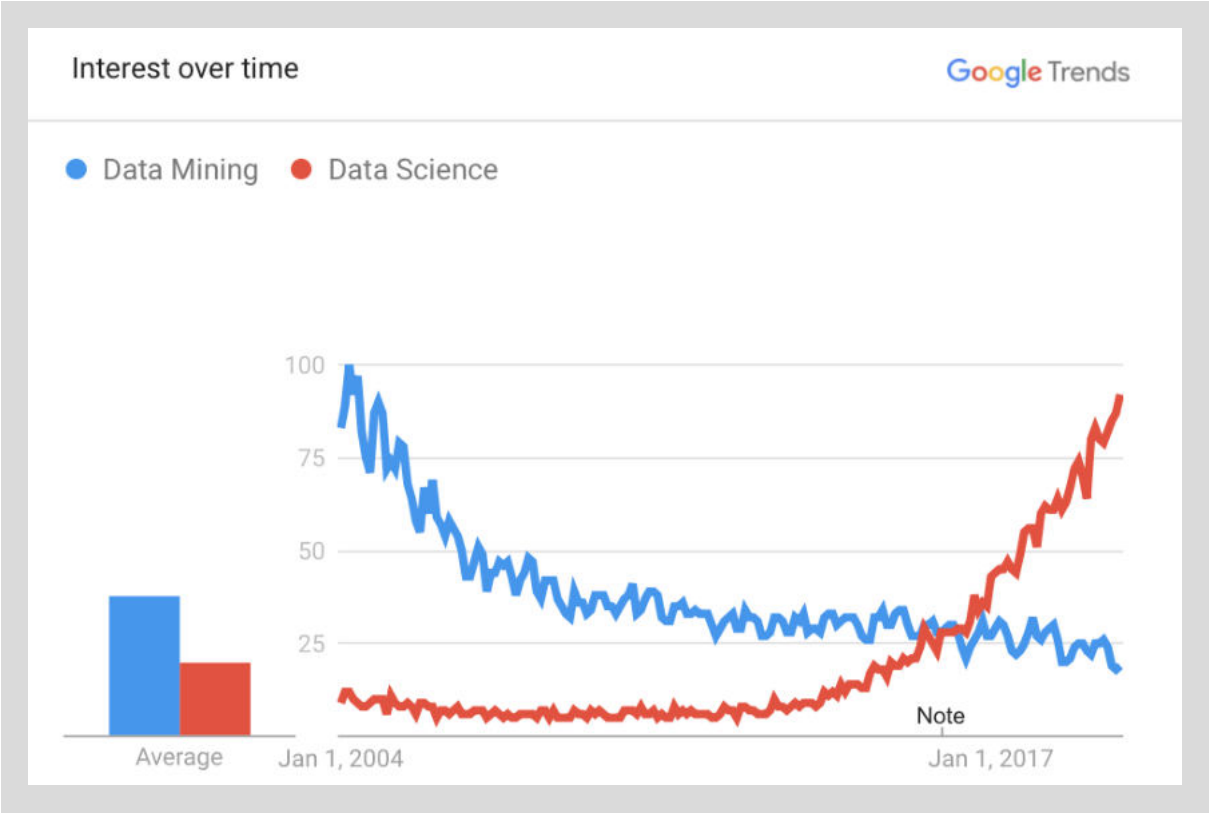
Data Science is a family of *disciplines operating at the junction of data and knowledge*, building on rich data and domain knowledge to produce **value** in a variety of forms.

These disciplines range from methodological subjects (statistics, machine learning, data mining) to data-driven applications in other domains (psychology, social science, economics, history, among many others).

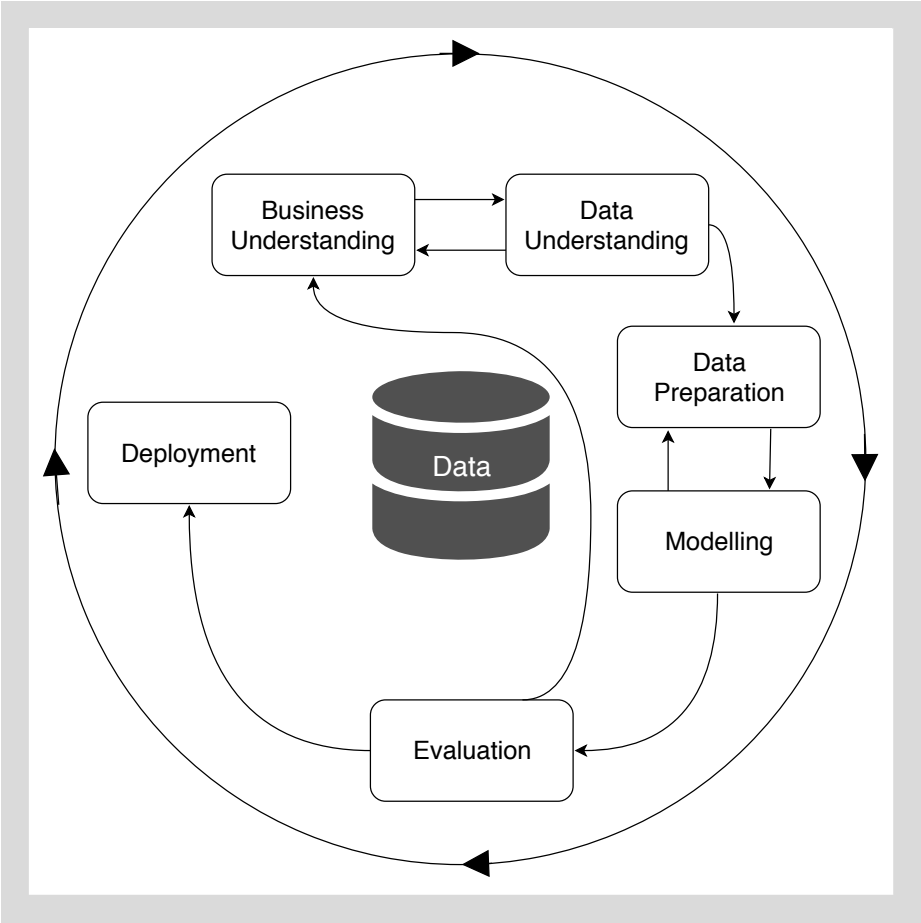
The Value of Data Science

- scientific knowledge and models
- societal value
- economic value
- personal value
- ...

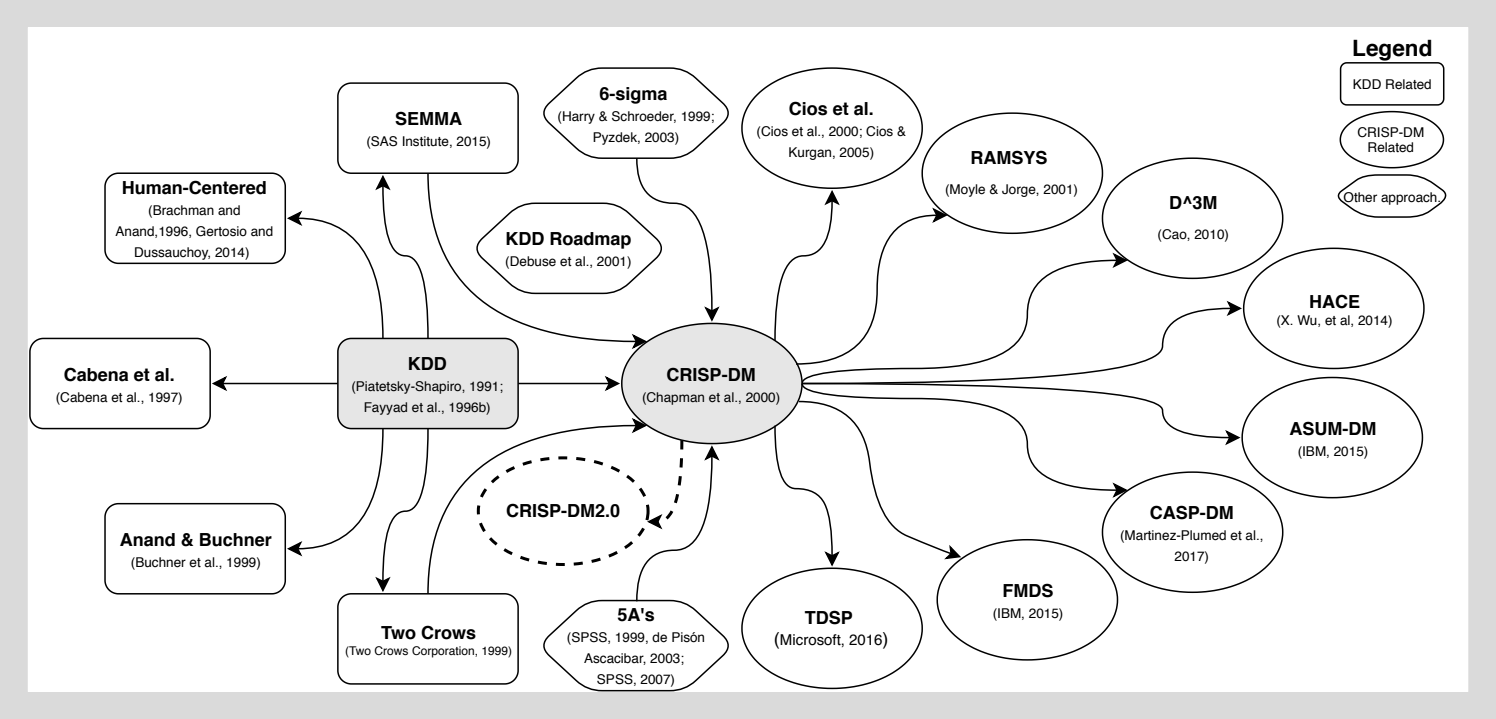
From Data Mining to Data Science (2/5)



CRISP-DM (1999)



CRISP-DM evolution



Adapted from G. Mariscal, O. Marban, and C. Fernandez: A survey of datamining and knowledge discovery process models and methodologies, Knowledge Engineering Review 25(2):137-166, 2010.

Data takes centre stage

Contemporary Data Science *starts from the data*:

- We know or suspect there is value in these data, how do we unlock it?
- What are the possible operations we can apply to the data to unlock and utilise their value?

While moving away from the process the methodology becomes less prescriptive and more inquisitive: things you **can** do to data rather than things you **should** do to data.

From mining to prospecting

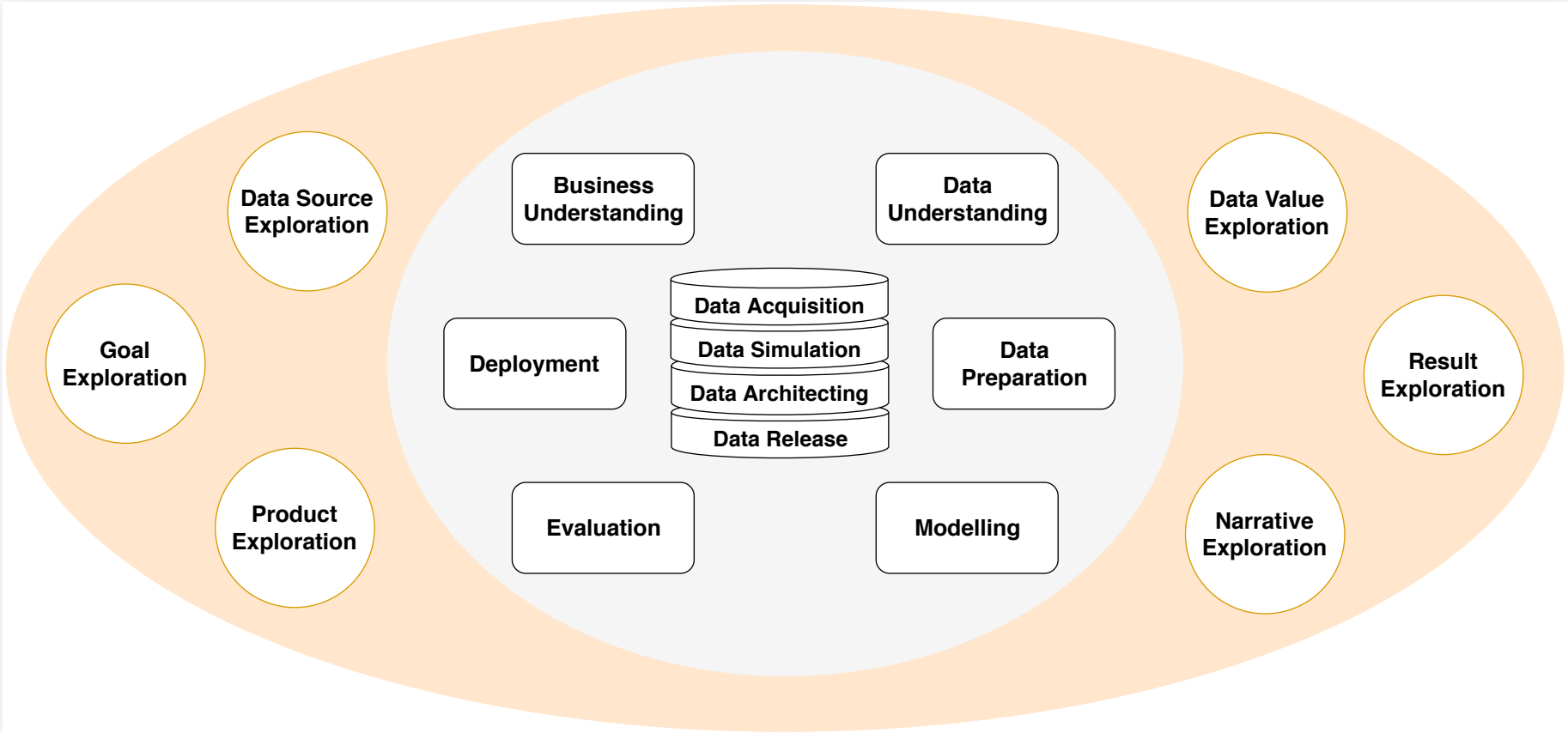
If data mining is like mining for precious metals, Data Science is like *prospecting*: searching for deposits of precious metals where profitable mines can be located.

Such a prospecting process is fundamentally **exploratory** and can include some of the following activities:

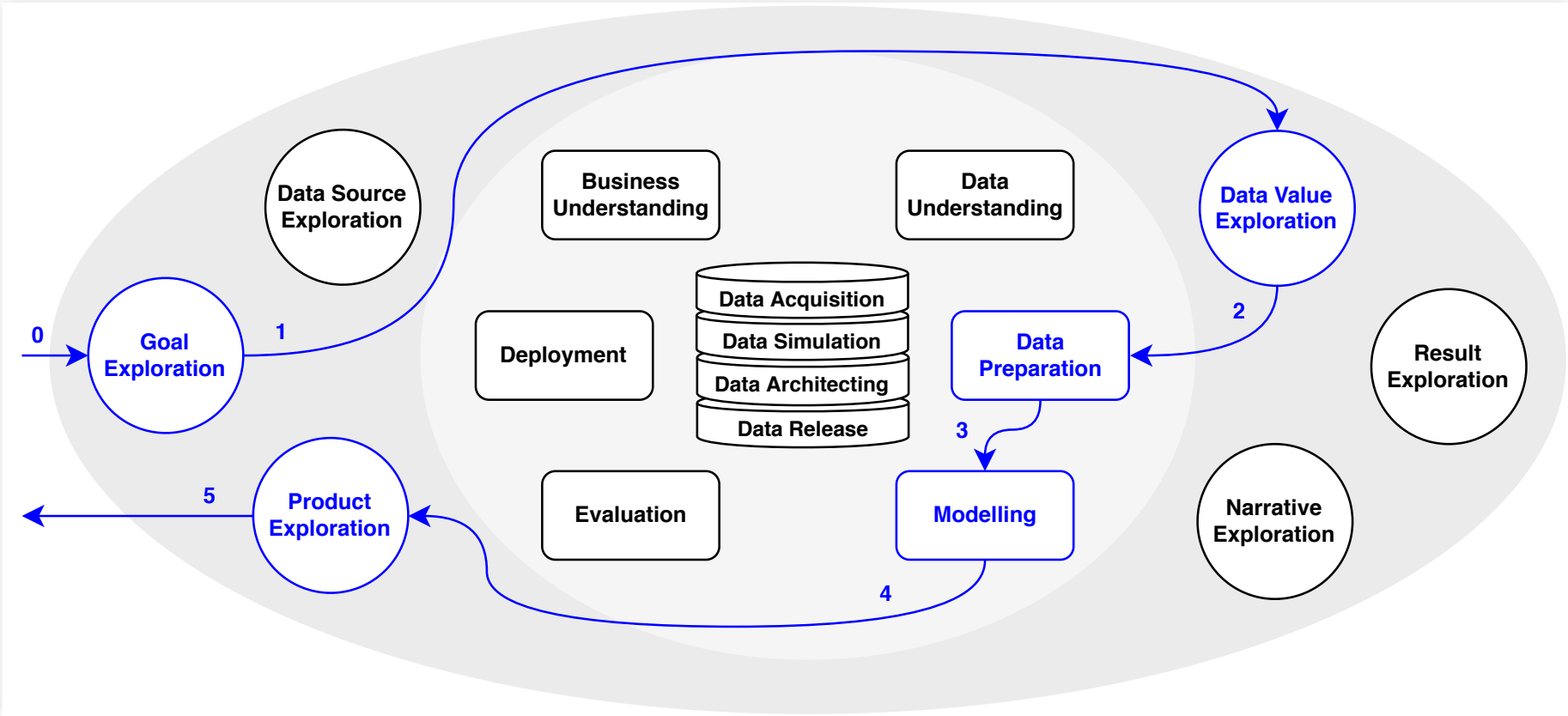
Exploratory activities in Data Science

- *Goal exploration*: finding business goals which can be achieved in a data-driven way;
- *Data source exploration*: discovering new and valuable sources of data;
- *Data value exploration*: finding out what value might be extracted from given data;
- *Result exploration*: relating Data Science results back to the business goals;
- *Narrative exploration*: extracting valuable stories (e.g., visual or textual) from the data;
- *Product exploration*: finding ways to turn the value extracted from the data into a service or app.

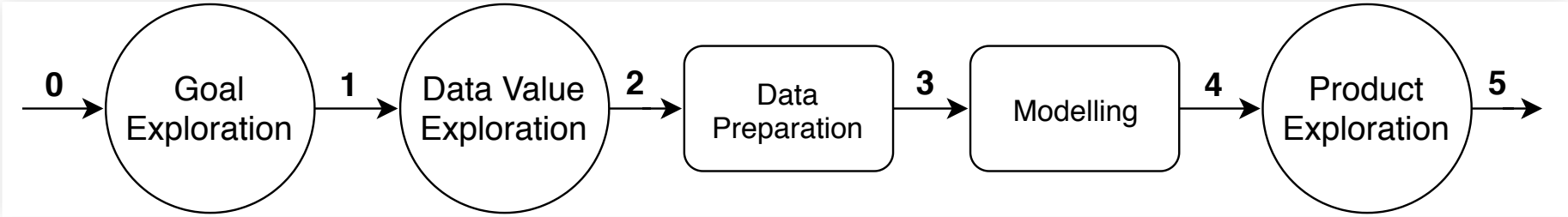
DST space


Data Source Exploration
Goal Exploration
Product Exploration
Business Understanding
Data Understanding
Deployment
Data Acquisition
Data Simulation
Data Architecting
Data Release
Data Preparation
Evaluation
Modelling
Data Value Exploration
Result Exploration
Narrative Exploration

Travelling through DST space

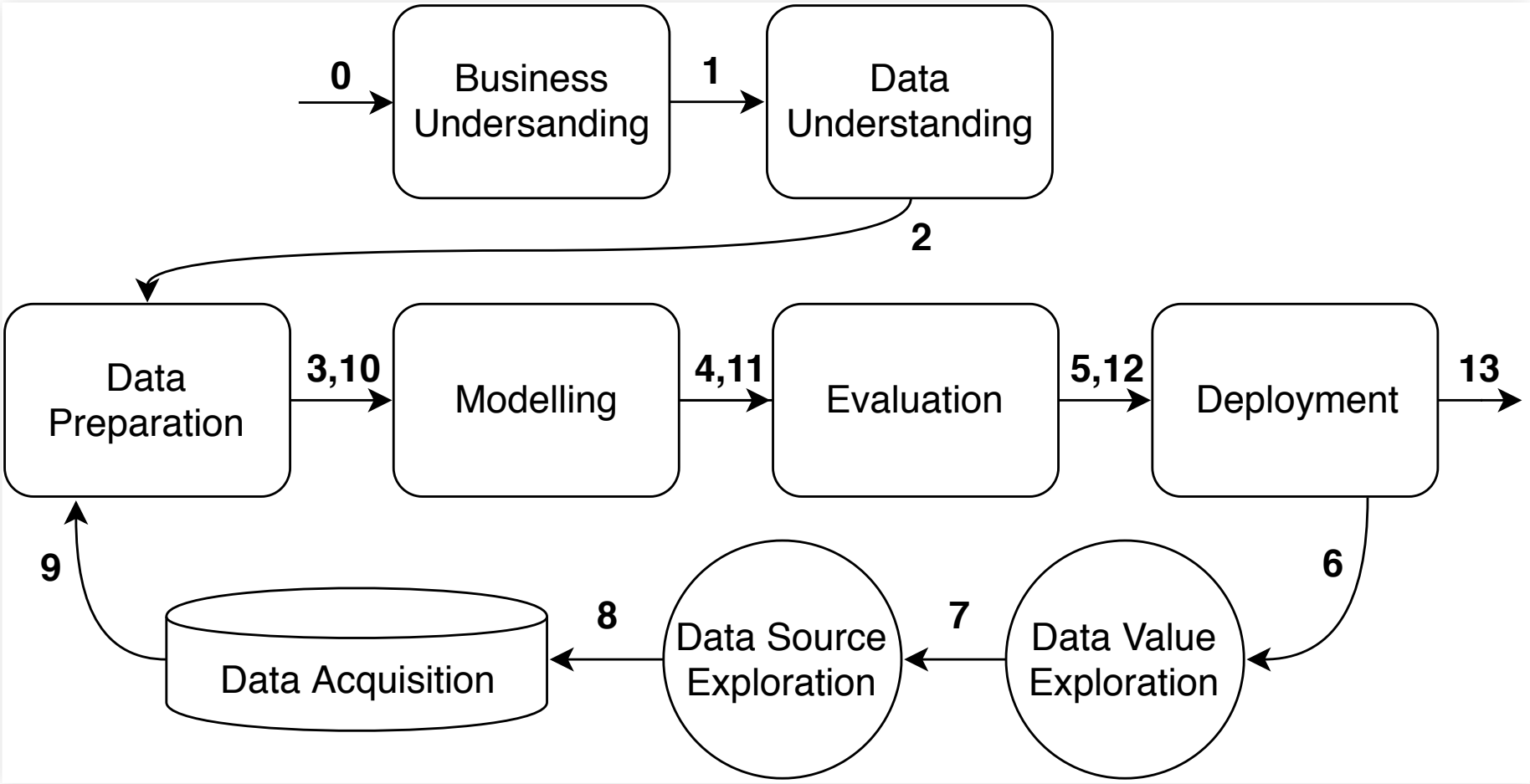


A Data Science Trajectory

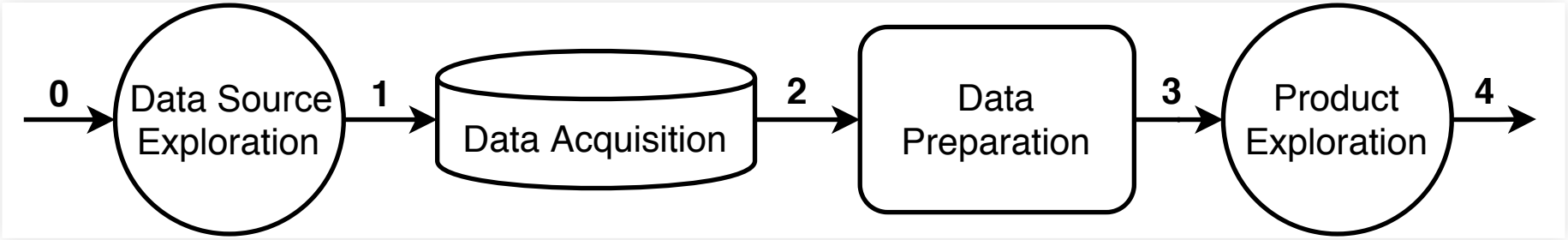


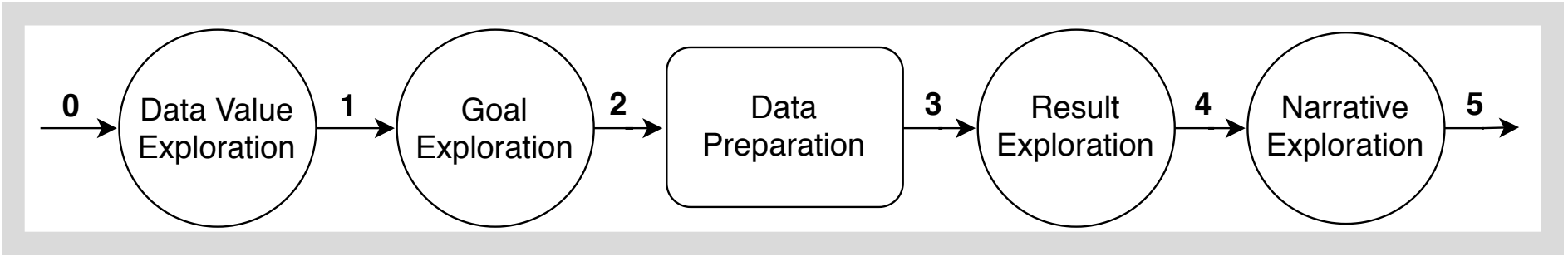
- *Goal Exploration*: activity recommender for tourists
- *Data Value Exploration*: get third-party location data
- *Data Preparation*: create user-location-activity ratings
- *Modelling*: train a recommender system
- *Product Exploration*: explore most appropriate end-user product/presentation

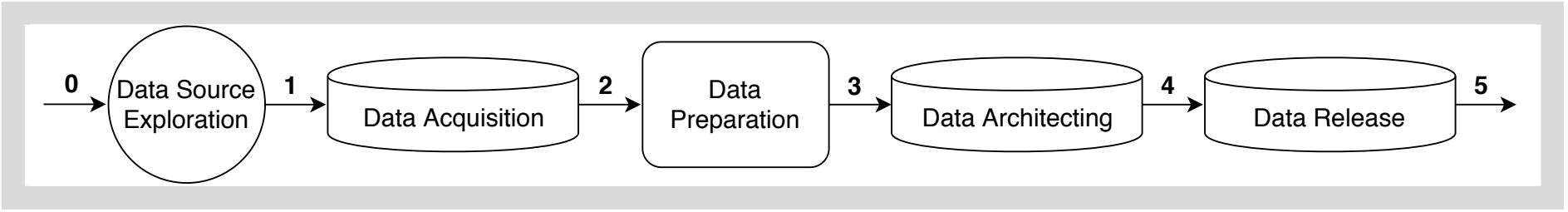
Looping


0
Business Undersanding
1
Data Understanding
2
Data Preparation
3,10
Modelling
4,11
Evaluation
5,12
Deployment
13
9
Data Acquisition
8
Data Source Exploration
7
Data Value Exploration
6

Not all trajectories require Modelling







Watch this space!

CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. Fernando Martinez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, Jose Hernandez-Orallo, Meelis Kull, Nicolas Lachiche, Maria Jose Ramirez-Quintana and Peter Flach. (Under review, 2019)

The Human Factor (3/5)

Data Science is for, about, by and with humans.

Responsible Data Science means taking the human factor into account at all stages.

Here I offer a few personal thoughts, also inspired by parallel discussions and developments in Human-Centred AI.

Fairness, Accountability and Transparency

These are important aspects contributing to Responsible Data Science.

But aspirations are not enough, and much more work is needed to work out

- *how to define them?*
- *how to measure them?*
- *how to achieve them?*

Fair or not?

	Men applied	Men admitted	%	Women applied	Women admitted	%
	2691	1198	45%	1835	557	30%
A	825	512	62%	108	89	82%
B	560	353	63%	25	17	68%
C	325	120	37%	593	202	34%
D	417	138	33%	375	131	35%
E	191	53	28%	393	94	24%
F	373	22	6%	341	24	7%

Humans in the Loop

There is a clear need to model the entire Data Science *ecosystem*, including human actors in their various roles.

This will allow keeping track of artefacts as they are created and travel through the ecosystem (*provenance*).

Ultimately such a model needs to be *causal* so that we can reason about the consequences of (not) doing something.

Caveats

Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."

Data for All?

If Data has intrinsic value, we need to consider

- data ownership;
- who profits from the additional value created.

Currently, users often give their data away freely in exchange for a service. Is this sustainable?

GDPR

More control for individuals over their personal data:

- the need for the individual's clear *consent* to the processing of personal data;
- easier *access* by the subject to his or her personal data;
- the rights to *rectification*, to erasure and 'to be forgotten';
- the right to *object*, including to the use of personal data for the purposes of 'profiling';
- the right to data *portability* from one service provider to another.

Consent

For consent to be valid it must be informed consent. For this to be the case it must be:

- given voluntarily (with no coercion or deceit);
- given by an individual who has capacity;
- given by an individual who has been fully informed about the issue.

Is consent informed if it is hidden in many pages of dense legalese?

Ethics to the rescue?

It is important to discuss and reach consensus on ethical issues raised by Data Science.

It is equally important to realise that

- *ethics existed long before Data Science & AI came to the fore;*
- *ethics is not prescriptive and may need to be followed by regulation and legislation to become effective.*

About EuADS (4/5)

In such a multi-disciplinary and varied landscape it is important to provide bridges between disciplines and balance the interests of stakeholders.

To support this, EuADS aims to bring together information about resources, best practice etc.:

- training programmes
- conferences, events & organisations
- funding opportunities & job openings
- technology infrastructure
- legal & ethical frameworks

EuADS activities so far

- Co-organised the *European Conferences on Data Analysis* (ECDA)
  - Luxembourg 2013, Bremen 2014, Colchester 2015, Wroclaw 2017, Paderborn 2018, Bayreuth 2019
- Organised the *European Data Science Conference* (Luxembourg, Nov. 2016).
- Guest-edited a special issue on *Data Science in Europe* of the Int. J. of Data Science and Analytics (2018).
- Organised the first European *Summer School on Explainable Data Science* (Luxembourg, Sept. 2019).

EDSC 2016

The EDSC programme consisted of invited plenary talks, symposia, workshops and panel discussions on:

- The question of trust, transparency and provenance.
- Legal aspects of Data Science such as data protection, data privacy and data access.
- Support for navigating the complex chain from raw data to actionable outputs .
- The role of Data Science in medicine and health care .
- How to define the field's methodological substance.

See Int.J. Data Science and Analytics 6(3), 2018.

Explainable Data Science Summer School

- Human-centric data exploration: *Tijl De Bie* (U Gent, B)
- Explanations from a psychological perspective: *Christos Bechlivanidis* (University College London, UK)
- XAI - Science and technology for the explanation of AI decision making: *Fosca Giannotti* (Information Science and Technology Institute Pisa, I)
- Making Data Science Intelligible: *Simone Stumpf* (City, U of London, UK)
- Causal inference from empirical data: *Jilles Vreeken* (CISPA Helmholtz Center for Information Security, D)

Two-way platforms

- *Data Science jobs*
- Data Science experts
- Data Science projects
- Data Science education
- Data marketplace
- Data Science for Europe

The European dimension

Data can be an integrative force to handle, and benefit from, European diversity.

Data can guide major European projects (e.g., environmental, socio-economic, equality & diversity).

Efforts towards a more data-driven European development need to be pooled and coordinated on different levels (e.g., science & technology, politics, public engagement).

How can you get involved?

- Become a EuADS member.
- Get your organisation to become an institutional member.
- Become a national representative.
- Help organise events such as this one.
- Help build our two-way platforms.
- ...

Navigate to EuADS.org.

Sabine Krolak-Schwerdt, 1958-2017

Recap (5/5)

- Whether seen as the science of data, doing science with data, or applying science to data: *data science is inherently multi-disciplinary* and requires sustained interaction between disciplines.
- Tracing its evolution from data mining exposes the *essentially exploratory nature of data science*.
- The *human factor raises important questions*, e.g. regarding ownership of data and results, and where regulation/legislation is needed.
- The **EuADS mission** is to help articulate, support and advance these challenges and opportunities.