The 4+1 Model of Data Science

Rafael C. Alvarado

11/8/22

Table of contents

1 Introduction **3**

2 The Image of the Pipeline **4**
2.1 An Arc with Four Zones 5
2.2 The Four Areas, Plus One 8
2.3 Two Principal Components 10
2.4 Final Representation 12
2.5 References 14

3 Concluding Remarks **15**

Appendices **15**

A Primary Sources **16**
A.1 Source List 16
A.1.1 Tukey on Data Analysis 16
A.1.2 Fayyad on KDD 17
A.1.3 Azevedo on SEMMA 18
A.1.4 Hayashi on Data Science 19
A.1.5 Wirth and Hipp on CRISP-DM 19
A.1.6 Mason and Wiggins on OSEMI 20
A.1.7 Ojeda, et al. on Data Science 21
A.1.8 Caffo, et al. on Data Science 21
A.1.9 Donaho on Data Science 22
A.1.10 Géron on Machine Learning 22
A.1.11 Das on Data Science 23
A.1.12 Dataman on Data Science 23
A.1.13 Porter on Data Science 23
A.2 Summary Table 24
A.3 References 30

B Relation to AI **31**

1 Introduction

Data Science is a complex and evolving field, but most agree that it can be defined as a combination of expertise drawn from three broad areas — computer science and technology, math and statistics, and domain knowledge — with the purpose of extracting knowledge and value from data. Many also associate it with a series of practical activities ranging from the cleaning and "wrangling" of data, to its analysis and use to infer models, to the visual and rhetorical representation of results to stakeholders and decision-makers.

This essay proposes a model of data science that is intended to go beyond the laundry-list definitions that dominate the discourse in the field today. Although these are not inaccurate, they do not get at the specific nature of data science or help distinguish it from adjacent fields such as computer science and statistics — fields whose members sometimes claim to already be doing data science. Without a clear understanding of its specific and unique nature, the field is subject to counterproductive turf battles in the academy as well as confusion in the workplace.[1]

We define data science in terms of a multi-part model that represents core areas of expertise in the field and how they are related to each other. These are the areas of **value**, **design**, **systems**, and **analytics**. A fifth area, **practice**, integrates the other four in specific contexts. Together, these areas belong to every data science project, even if they are often unconnected and siloed in the academy.

Unlike traditional academic disciplines, each area of the proposed model is inherently interdisciplinary, bringing together diverse and sometimes contrary perspectives under a common heading. The inherently interdisciplinary and pluralist nature of these areas is a distinctive feature of data science and a key differentiator between it and traditional disciplines.

The following describes how this model is derived and provides clues about how to interpret and apply the model to your own situation.

[1] It is increasingly the case that hiring managers in industry understand the role of data scientist differently than the academic programs that produce data scientists. One result of this is the proliferation of terms such as data engineer, machine learning engineer, and applied AI to define areas of work that have historically belonged to data science. Another result is the confusion of data scientist with the roles of data analyst and statistician.

2 The Image of the Pipeline

A review of the literature of data science definitions, from sources attempting to define the field explicitly, as well as from self-definitions from adjacent fields such as data analysis and data mining, reveals that most definitions invoke the image of a data processing pipeline — a sequence of actions through which data flows as it moves from the consumption of so-called raw data to production of actionable results. Consumed data may come from a variety of sources — databases or intentional experiments or sensors. Results may be equally various, from the communication of analytical results to stake-holders to the development of a data product for use on the web. For a detailed review of these sources, see the Appendix.

An analysis of a representative sample of essays that define a data processing pipeline shows that the various stories consist of elements drawn from a standard sequence of about twelve elements, give or take a few, depending on how one might expand or contract terms. These may signified by a core set of verbs or event types (which narratologists call functions), with the understanding that many synonyms are employed in the examples:

1. Understand
2. Plan
3. Collect
4. Store
5. Clean
6. Explore
7. Prepare
8. Model
9. Interpret
10. Communicate
11. Deploy
12. Reflect

No one definition includes them all, but some are more comprehensive than others, and different disciplines emphasize different parts.

For example, Hayashi's statistically-oriented definition of data science includes just three phases — design [1], collect [3], and analyze [8, 9] — with an emphasis on the experimental design phase in which data are actually produced through thoughtfully designed experiments [@hayashi1998].

Mason and Wiggins propose five — obtain [3, 4], scrub [5], explore [6], model [8], and interpret [9] — which highlights two conditions that define industrial data science, the simultaneous availability of data (one merely obtains it, say through web scraping) and their poor condition relative to analysis, i.e. the need to scrub and wrangle them into usable form [@mason2010].

The CRISP-DM model is the most comprehensive, with seven phases defined (if we include the unnamed but visually depicted function of storage), emphasizing the importance of understanding both the business proposition and data before anything is done with it [@wirth1999]. It also modifies the metaphor of the pipeline, representing it as a circular and iterative process. However, unlike Donoho's similarly comprehensive sequence (implied by the ordering of his six divisions of "greater data science"), it does not include a "meta" phase devoted to reflecting on the process as a whole [@donoho2017].

This twelve-part composite pipeline can be simplified by combining functions that naturally go together, by virtue of the expertise required to carry them out. This reduction yields about seven phases:

A understand and plan
B collect and store
C clean, prepare, and explore
D model and interpret
E communicate
F deploy
G reflect

Each of these may be considered a "chapter" in the story. Note that the number of verbs in each chapter title does not necessarily predict the length of its content. For example, the chapter on "model and interpret" covers a wide range of activities from a variety of perspectives, including classical statistics, machine learning, and computational simulation. It's a big and complicated chapter, but it is just one chapter among seven, even though many may consider it to be the most important chapter.

2.1 An Arc with Four Zones

To be sure, the middle chapter plays a central role in our story. If we think of the story as following a classical "there and back again" structure — a chiasmus pattern like X_1, Y_1, Z, Y_2, X_2 — then chapter D is the pivot, while chapters A, B, and C mirror E, F, and G. Thinking of the story in this way allows us to identify a parallel structure in the pipeline, connecting phases that are usually seen as separate. Specifically, we may visualize the pipeline as an arc, in which chapters in the first half of the pipeline mirror the those of the second half. We may then group chapters by the pairs formed in this way, yielding four zones — A and G belong to zone I, B and F to II, C and E to III, and D to IV — as in the following diagram:

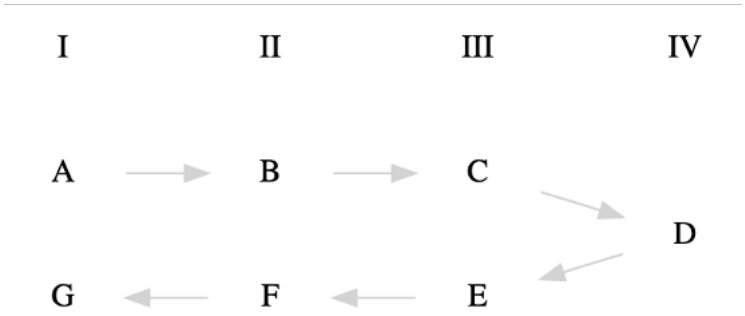

Figure 2.1: The Standard Sequence as a Narrative Arc

I	understand, plan	reflect
II	collect, store	deploy
III	clean, prepare, explore	communicate
IV	model, interpret	

Figure 2.2: The Arc Transposed

With this visualization, we can discern some interesting properties about the data science pipeline that are not obvious in the original sequential image. For one, the arc structure suggests that the two ends of the pipe are not separate; both make direct contact with the external world. The external world — natural or social — from which data are pulled is the same world into which data products are inserted. This insight echoes the CRISP-DM model, which connects A and G (actually F), except that the two ends of the arc model are not directly connected. Instead, they come into contact with — and are separated by — the world in all of its complexity and unpredictability. The relationship between the effects caused by our data products G and the data we pull from the world A is not given but a matter of discovery — and often surprise.

At this point, we can explore the unifying themes associated with the four zones in our arc model by transposing the preceding visualization, which draws attention to what is common to each pairing. This generates four candidate areas of data science expertise — activities that, although they appear on opposite ends of the pipeline, nevertheless share basic knowledge, know-how, and areas of concern.

Zone IV is the easiest to interpret in this way because as the pivot of the arc it is not paired. It represents the work of modeling a problem mathematically, as well as evaluating and interpreting the results of mathematical modeling. This work requires data to be available in a particular form — clean and organized, usually as "tidy" analytical tables.

Zone I is also relatively easy to interpret: the functions in this group each involve understanding the relationship between the pipeline and the external world, the messy interface between the enterprise of data science and the variety of real world situations in which it operates.

We may note in passing that I and IV can be contrasted in several ways — messy vs clean, exoteric vs esoteric, qualitative vs quantitative, existential vs essential, concrete vs abstract, etc.

When it comes to zones II and III, the interpretation of results is less straightforward. This is because the reality of the kind of work performed in these areas is not as clear-cut as it is for I and IV. Both II and III exhibit an internal complexity not found in the others, and the two are less clearly separable from each other than they are from the other two. One reason for this complexity is that here pure and applied forms of knowledge intermingle in ways that defy easy description from an academic perspective.

For example, the work of "data wrangling," often considered essential to data science, spans the two domains and involves a complex mixture of specific technological know-how and general scientific principles. It turns out that the relationship between these kinds of knowledge is highly contested, as evidenced by the reception of Donoho's "50 Years of Data Science," which has been criticized for separating science from engineering and demoting the importance of the latter [@donoho2017]. Regardless of the validity of this criticism, there is without doubt a long-standing conflict between data mining and data analysis over what counts as valid forms of knowledge, and this conflict emerges in the representation of zones II and III we find in our corpus.

We can take the conflict of interpretations over the status of technical knowledge in data science as a clue and use it to identify two broad dimensions that cross-cut the functions in zones II and III: technical know-how and abstract representation. Technical know-how $II´$ involves expertise in developing and deploying software and hardware designed to handle data at scale, including high-performance computing, big data architectures (such as Hadoop and its descendants), and data-oriented programming languages and libraries.

The topics associated with II' are highly specific and change rapidly relative to other forms of knowledge, and so are often omitted from, or under-represented in, academic curricula, even though to many they are the *sine qua non* of data science. Abstract representation $III´$, on the other hand, involves expertise in areas ranging from how data are to be modeled for capture and analysis to how the results of analyses are to be presented to non-expert decision-makers. These areas of knowledge strive for formal generality over the long run; they are often expressed as grammars or design languages, frequently with visual modes (such as entity-relationship models and unified modeling language UML). They also include other forms of visualization, such as the plots developed for exploratory data analysis, such as box plots, and those used to represent statistical facts and analytical results in dashboards and infographics.

2.2 The Four Areas, Plus One

We are now ready to define and name the areas of data science expertise that emerge from an analysis of the pipeline considered as an arc. In each case, we want to identify the common context shared by the paired activities in each zone as well as the tension that exists between them by virtue of their occupying opposite sides of the pipeline. In many cases, although we can identify a shared theme in each zone's work, the reality is that practitioners do not always interact or share disciplinary homes. One of the benefits of this model will be to identify these points of synergy and to identify new disciplinary boundaries.

Area I: Value

The area of value is defined by the relationship of data science to the world from which it draws data and into which it inserts data products. More broadly, it concerns the primary motivations of data science — why do we practice data science in the first place? It combines the traditional discipline of ethics with the professional activities of business planning, policy making, developing motivations for scientific research, and other activities that have a direct impact on people and the planet. This is the area where we determine what we do versus what we do not do, in order to maximize societal and environmental benefit and minimize harm. It is also the area that looks inward to the other data science areas and provides guidance on such issues as algorithmic bias or open science. Common activities include the forming of value propositions that initiate data science projects, research into how data is created and used "in the wild," understanding the ethics of data acquisition, manipulation, communication, and sharing, and the application of data products in the world.

Area II´: Design

The area of design is defined by the relationship between human and machine forms of representation. This relationship is bidirectional: human-generated data flowing into the pipeline must be represented for machine consumption (H2M, or $H \rightarrow M$), while analytically transformed data going out must be represented for human consumption (M2H, or $M \rightarrow H$). This area therefore includes expertise in human-machine interaction as it appears at the points of both consuming data and producing data products. Activities here include the representation and communication of captured data for the work of analytics, e.g. in database modeling, the curation of data, and of complex data and analytical results to humans to drive decision-making and influence behavior. It also includes the making of things, with purpose (i.e. to solve problems) and intent (meaning, concision, focus). A key part of the area is the broad practice of what is often called visualization, the translation of complex quantitative information into visual (and other sensory) forms that non-experts can understand. In slightly more technical terms, the area of design focuses on what Zuboff called "informating," the process by which the world is represented for computation and analytics, and also by which analytical models and results are represented to the world [@zuboff1995]. These two processes often produce competing representations — a private one *of* the world for the data scientist, and a public one *for* the world of the results of analytics. One task of this area is to reconcile these two representations.

Area III´: Systems

The area of systems is defined by the technological infrastructure that is common to the pipeline but concentrated in the activities of wrangling data, deploying data products, and building out systems to support these activities at scale. This area includes expertise in infrastructure systems and architectures to support working with big data — big in terms of volume, velocity, and variety — and building high performance systems in both development and production environments. It includes the broad areas of hardware and software as such — computer technology as opposed to computer science. Key activities include developing cloud resources, building performant pipelines to ingest and aggregate data, developing networks of resilient distributed data, and writing and using software to accomplish tasks. This area is often referred to as "data engineering" or "machine learning engineering," which, according to Owen, "is most of what Data Science is and Statistics is not" [@owen2015].

Area IV: Analytics

The area of analytics is defined by the practice of mathematical modeling based on data. This area includes what many consider to be the essence of data science, the combination of statistical methods with machine learning, along with information theory, optimization, network analysis, complexity theory, simulations, and other rigorous quantitative methods from a variety of fields. Although unified by a broad commitment to advanced mathematical models and computational algorithms, in reality this is a heterogeneous collection of competing schools and methods. Tensions include inference vs prediction, parametric vs non-parametric (kernel-based) methods, frequentist vs Bayesian statistics, analytic vs algorithmic solutions

(including simulations), etc. Key activities include clustering, pattern recognition, regression, rule mining, feature engineering, model selection, performance evaluation, and a host of other activities. Although currently dominated by statistical methods, this area also includes the rule-based methods that dominated the field of artificial intelligence before the more recent successes of statistical learning and deep learning.

Area V: Practice

The preceding four areas each represent areas of foundational knowledge, forms of expertise that can be taught as more or less separate subjects. In practice, however, these areas represent the interlocking parts of a division of labor that are integrated in the pipeline. This area consists of actual activities that brings people together to combine expertise from each of the four areas. It is characterized by data science teams working together and with external parties to develop solutions and projects that are responsible, authentic, efficient, and effective. Practice is also where the core areas of data science come into contact with a broad spectrum of domain knowledge and real world problems. The following diagram () shows the central, integrative role played by practice:

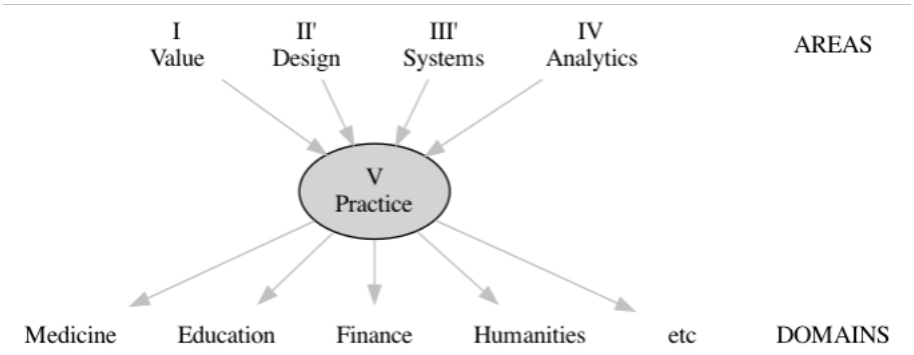

Figure 2.3: The Integrative Role of Practice

2.3 Two Principal Components

Is there a way to understand how the four primary areas are related to each other, beyond their being composed of functions from the same pipeline? Put another way, does the pipeline-as-arc model exhibit any structural features that will help us conceptualize the broader space of data science? Two such features stand out: (1) the opposition between concrete and abstract forms of representation, and (2) between human and machine processing.

Regarding the concrete and the abstract, it's clear that the arc model has a metric quality to it: as one moves toward the pivot point of analysis, one moves away from the concrete messiness of reality as experienced to the "tidy" and abstract world of mathematics; similarly, as one moves from the pivot back to the world, there is a requirement to convert esoteric results into more

humanly intelligible forms, often through a process of concretization; visualizations succeed by employing concrete metaphors that flesh out mathematical ideas that are notoriously detached from the imagination — no one can imagine, for example, n-dimensional spaces beyond a handful of dimensions. The arc describes a dialectic of abstraction and concretization that defines the ebb and flow and data science work.

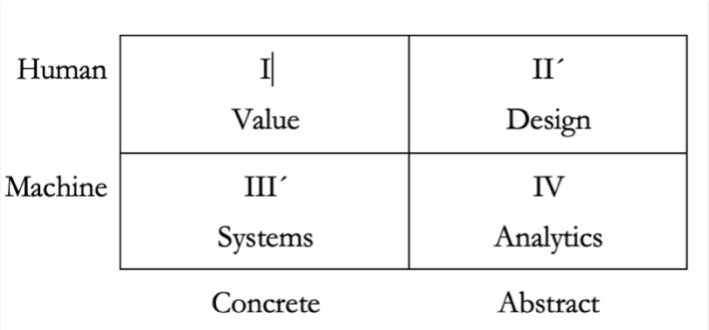

Figure 2.4: The Four Areas in Two Dimensions

The dimension of human and machine processing exhibits a similar duality, that between the conversion of information from humanly accessible forms, such as given by data acquired by instruments, into machine readable and processible forms, and the reverse. The process of moving from human to machine representations is a large part of what data capture, modeling, and wrangling is all about, while the process of converting the results of machine learning, broadly conceived, into humanly actionable form is what visualization and productization are all about. The reality of this dualism is captured by the concept of human-computer interaction (HCI), an established field that is applicable to both sides of the arc.

How do the four fundamental areas map onto these two dimensions? We can define each area as a combination of one pole from each duality; the four areas result from all possible permutations of the two dimensions. This produces the following high level characterizations of each area: (1) Value is concerned with concrete humanity, (2) Design with abstract humanity, (3) Analytics with abstract machinery, and (4) Systems is concerned with concrete machinery. All of these make intuitive sense, with the exception of Design. This is consistent, however, with the fact that the area of Design emerges from this analysis as an undervalued and not well understood area of expertise, even though Yau emphasized it early on (Yau 2009b). Indeed, one of the consequences of this analysis is to train our attention on this area of knowledge and to develop it further.

One exciting interpretation of the two dimensions defined here is that they correspond to two principal components that undergird the general field of data science. As components, these axes define two orthogonal dimensions within which all the specific topics of data science may, in principle, be plotted. The reality behind these axes may be that they represent cognitive styles associated with the division of labor implied by the data science pipeline.

PC1: Human versus Machine

The human-machine axis accounts for the most variance in the field. This seems evident from the fact that Conway's Venn diagram model of data science represents only the machine side of our model, with practice replaced by "substantive expertise" [@conway2010]. The human side — Value and Design — is left out, or short-changed by being lumped in with domain knowledge. The very fact that the human side has to be explained and added to the model suggests strongly that it defines a pole at some distance from the areas of knowledge described in Conway's model. The human pole refers to humanity understood as situated in their historical, social, and cultural milieu. It is synonymous with *human experience*. The machine pole refers to the technoscientific apparatus of formal, quantitative reasoning that operates on representations of the human and the world. In the context of data science, it is more or less synonymous with *machine intelligence*, broadly conceived to include machine learning but also other modes of analysis on the spectrum of prediction and inference. Given these poles, the human-machine axis represents the opposition between humanistic disciplines that seek to understand human experience as such, and the formal sciences that employ machine intelligence, broadly conceived, to interpret that experience as represented and aggregated in the form of data.

PC2: Concrete versus Abstract

The abstract-concrete axis accounts for the difference between two forms of knowledge, roughly between direct experience and the indirect representation of that experience enabled through data. Both the realm of Value and Systems involve immersion in the messy details of lived experience — and direct acquaintance with the devils in those details. This is the messy world of hacks and ironies. The realms of Design and Analysis, on the other hand, are founded on abstract representations that strive for clear and distinct purity, and which allow for deductive reasoning to succeed at the cost of simplifying assumptions and reduced representations. This is the orderly world of models. The concrete pole refers to situated knowledge, knowledge as understood by hackers and makers, but also ethnographers who seek to maximize thick description in their work. It represents *concrete materiality*. The abstract pole refers to formal knowledge, knowledge in the form of mathematical symbolism, deductive proofs, and algorithmic patterns. It is *abstract form*. Given these poles, the concrete-abstract axis is roughly the opposition between applied and pure forms of knowledge, between those that embrace materiality and those that seek purity of form.

2.4 Final Representation

The result of the preceding may be represented by the following graphic.

This visualization represents data science as composed of specific and complementary forms of knowledge. The vertical axis defines the dominant polarity between analysis — the *how* of data science, often identified entirely with it, contrasted with the *why* of data science, from which data science derives its meaning and value as a profession. The horizontal access defines

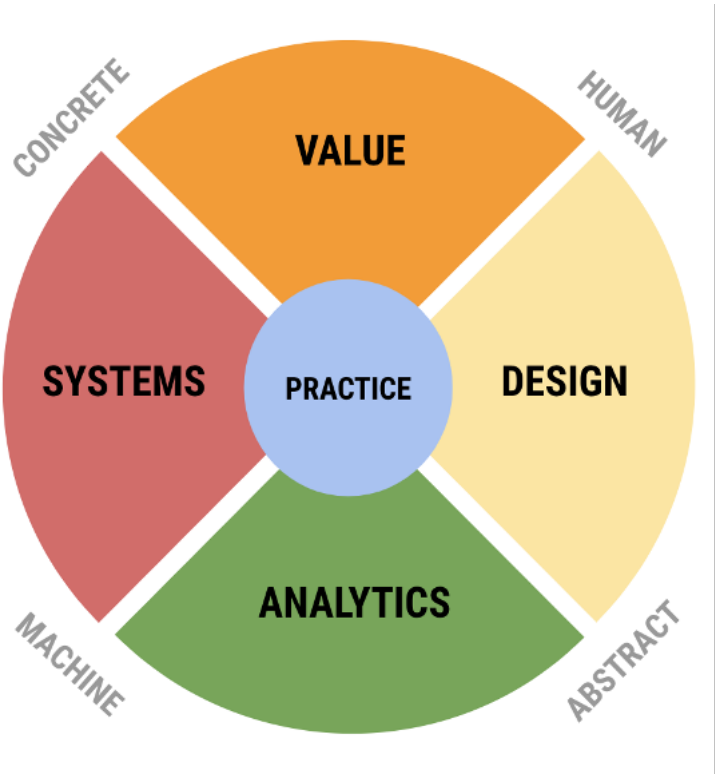

Figure 2.5: The 4+1 Model of Data Science

the polarity of methods that are often obscured in academic definitions of data science — the supporting practices that make the Analytics component work in the first place.

2.5 References

3 Concluding Remarks

The point of the $4 + 1$ model, abstract as it is, is to provide a practical template for strategically planning the various elements of a school of data science. To serve as an effective template, a model must be general. But generality if often purchased at the cost of intuitive understanding. The following caveats may help make sense of the model when considering its usefulness when applied to various concrete activities.

The model describes areas of academic expertise, not objective reality. It is a map of a division of labor writ large. Although each of the areas has clear connections to the others, the question to ask when deciding where an activity belongs is: *who would be an expert at doing it*? The realms help refine this question: the analytics area, for example, contains people who are good at working with abstract machinery. The four areas have the virtue of isolating intuitively correct communities of expertise. For example, people who are great at data product design may not know the esoteric depths of machine learning, and that adepts at machine learning are not usually experts in understanding human society and normative culture.

Each area in the model contains a collection of subfields that need to be teased out. Some areas will have more subfields than others. Although some areas may be smaller than others in terms of number of experts (faculty) and courses, each area has a major impact on the overall practice of data science and the quality of an academic program's activities. In addition, these subfields are in an important sense "more real" than the categories. We can imagine them forming a dense network in which the areas define communities with centroids, and which are more interconnected than the clean-cut image of the model implies.

The principal components abstract/concrete and human/machine are meant to help imagine the kinds of activities that belong in each area, through their connotations when combined to form the four bigrams — concrete human, abstract human, concrete machine, and abstract machine. For example, the area of value as the realm of the "concrete human" (or perhaps "concrete humanity") is meant to connote what the Spanish philosopher Unamuno called the world of "flesh and bone" within which we live and die, that is, where things matter. On the other hand, analytics as the realm of the "abstract machine" is meant to connote the platonic world of mathematical reasoning which, since Euclid, has been characterized by rigorous, abstract, deductive reasoning that has literally been described as an abstract machine (see Alan Turing).

At the center of this model and each area is people. Even in the area classified as "abstract machine," people and human thinking is at the center.

A Primary Sources

The materials on which the conclusions of this essay are drawn come from a variety of sources, from technical journals to blogs to internal documents. They also come from a range of viewpoints, from data analysis and statistics to data mining and data science *per se*. For the purposes of the essay, we select a more or less representative subset across these axes of variation.

For each source, we list the phases cited by the authors as fundamental to data processing, broadly conceived. These are aggregated in the table that follows, aligning the specific steps with the general categories used to produce a composite pipeline. Identifiers for each source are indicated by a short string in brackets that precedes the citation and is used in the first column of the table.

It should be noted that in most cases these phases are explicitly described as a process and often as a pipeline. When they are not, the implication is strong. In some cases, the process is likened to a cycle, emphasizing the connection between the endpoints of the pipeline, which is also emphasized by the 4+1 model.

The twelve documents listed below begin with Tukey's seminal essay on data analysis and end with examples of explainers of data science that have become quite frequent in recent years. Included also are the class definitions of the CRISP-DM and KDD processes which are the most developed pipeline models.

A.1 Source List

A.1.1 Tukey on Data Analysis

Key: `Tukey`
Source: @tukeyFutureDataAnalysis1962 URL
Field: Statistics, Data Analysis
Notes: * Distinguishes between mathematical statistics and data analysis. * Defines data analysis as: "… procedures for **analyzing** data, techniques for **interpreting** the results of such procedures, ways of **planning** the **gathering** of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (p. 2) * "Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation" * Defines data analysis as a science; contrasts

with mathematics, which is not *per se*. “Data analysis, and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics …” (p. 6) * “Data analysis, as used here, includes planning for the acquisition of data as well as working with data already obtained.” (p. 40) * Pipeline model implicit.

Model:

1. Planning: “ways of planning the gather of data to make its analysis easier”.
2. Gathering: the gathering of data.

3. Analyzing: the analysis of data with “all the machinery and results of (mathematical) statistics”.
4. Interpreting: “techniques for intepreting the results of” analysis.

A.1.2 Fayyad on KDD

Key: `KDD`
Source: @fayyadKnowledgeDiscoveryData1996 URL→
Field: Data Mining
Notes: * Summarization of prior article. * Described on p. 84. * Most thorough, but still not complete. * “The KDD process can involve significant iteration and may contain loops between any two steps.” * Exploration takes place in the data mining step. “Data Mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data ….” (p. 83) * KDD is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database.

Steps:

Short version

1. Selection: Creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. Pre-processing: Cleaning and pre processing the data in order to obtain consistent data.
3. Transformation: Transformation of the data using dimensionality reduction and other methods.
4. Data Mining: Searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction).
5. Interpretation/Evaluation: Interpretation and evaluation of the mined patterns.

Long version (derived from from Brachman & Anand 1996)

1. **Understanding**: "Developing an understanding of the application domain and the relevant prior knowledge, and identifying the goal of the KDD process from the customer's viewpoint."
2. **Creating**: "Creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed."
3. **Cleaning**: "Data cleaning and preprocessing: basic operations such as the removal of noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes."
4. **Reduction and projection**: "Data reduction and projection: finding usefull features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data."
5. **Method selction**: "Matching the goals of the KDD process (step 1) to particular data mining *method*: e.g., summarization, classification, regression, clustering, etc."
6. **Model selection**: "Choosing the data mining algorithm(s): selecting method(s) to be used for searching for patterns the data. This includes deciding which models and parameters maybe appropriate (e.g. models for categorical data are different than models on vectors over the reals) and matching a particular data mining methodwith the overall criteria of the KDDprocess (e.g., the end-user maybe more interested in understanding the model than its predictive capabilities …)."
7. **Data mining**: "Data mining: searching for patterns of interest in a particular representational form or a set of such representations: classification rules or trees, regression, clustering, and so forth."
8. **Interpreting**: "Interpreting mined patterns, possibly return to any of steps 1-7 for further iteration. This step can also involve visualization of the extracted patterns/models, or visualization of the data given the extracted models."
9. **Consolidating**: "Consolidating discovered knowledge: incorporating this knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This also includes checking for and resolving potential conflicts with prreviously believed (or extracted) knowledge."

A.1.3 Azevedo on SEMMA

Key: `SEMMA` Source: @azevedoKDDSEMMACRISPDM2008 Field: Statistics

Notes: * Developed by SAS institute in 1996 (source?). * Although developed for SAS Enterprise Miner, widely referenced as a model, especially in comparison to KDD and CRISP-DM. Bias towards statistis and to the SAS product is evident in the first step. * Descriptions are derived from Azevedo, et al. 2008.

Steps: 1. Sample: Sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. 2. Explore:

Exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas 3. Modify: Modification of the data by creating, selecting, and transforming the variables to focus the model selection process 4. Model: Modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome. 5. Assess: Assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

A.1.4 Hayashi on Data Science

Key: `Hayashi` Source: @hayashiDataScienceClassification1998 URL→
Field: Statistics

Notes:

- Hayashi adopted the term "data science" in the early 1990s to define a field that did not succumb to what he saw to be the errors of both statistics and data analysis. He felt that mathematical statistics had become too attached to problems of inference and removed from reality, while data analysis had lost interest in understanding the meaning of the data it deals with.
- "Data Science consists of three phases: design for data, collection of data and analysis on data. It is important that the three phases are treated with the concept of unification based on the fundamental philosophy of science …. In these phases the methods which are fitted for the object and are valid, must be studied with a good perspective." (p. 41)
- Critical of both statistics and data analysis.
- Viewed the pipeline as a spiral of diversification and simplification.
- Hard and soft results.

Steps:

1. Design: Surveys and experiments are developed to capture data from "multifarious phenomena".
2. Collection: Phenomena are expressed as multidimensional or time-series data; properties of the data are made clear. At this stage, data are too complicated to draw clear conclusions. (Representation)
3. Analysis: By methods of classification, multidimensional data analysis, and statistics, data structure is revealed. Simplification and conceptualization. Also yields understanding of deviations of the model, which begins the cycle anew. (Revelation)

A.1.5 Wirth and Hipp on CRISP-DM

Key: `CRISPDM` Source: @wirthCRISPDMStandardProcess1999 URL→
Field: Data Mining

Notes:

- "At the top level, the data mining process is organized into a small number of phases. Each phase consists of several second-level generic tasks."
- The process is cyclic and recursive.

Steps:

1. Business Understanding: Understanding project objectives and requirements from a business perspective. Development of a plan.
2. Data Understanding: Initial data collection and activities to get familiar with the data, e.g. to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. This is really two phases, which are combined because of their close relationship: *Collection* and *Exploration*.
3. Data Preparation: Construction of the final dataset. Tasks include table, record, an attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.
4. Modeling: Modeling techniques selected and applied, parameters calibrated.
5. Evaluation At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
6. Deployment: Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

A.1.6 Mason and Wiggins on OSEMI

Key: `OSEMI` Source: @masonTaxonomyDataScience2010 URL→
Field: Data Science

Notes: * Assumes context of web, available data. * Borrowed language from 5 Steps of a Data Science Project Lifecycle.

Steps:

1. Obtain: Gather data from relevant sources through APIs, web scraping, etc.

2. Scrub: Clean data and convert to machine readable formats. Clearning includes handling missing data, inconsistent labels, or awkward formatting; stripping extraneous characters; normalizing values, etc.
3. Explore: Find significan patterns and trends using statistical and data analytic methods, These include visualizing, clustering, performing dimensionality reduction.
4. Model: Construct methods to predict and forecast.
5. Interpret: Put the results to good use.

A.1.7 Ojeda, et al. on Data Science

Key: `Ojeda+` Source: @ojedaPracticalDataScience2014 URL→
Field: Data Science

1. Acquisition: The first step in the pipeline is to acquire the data from a variety of sources, including relational databases, NoSQL and document stores, web scraping, and distributed databases such as HDFS on a Hadoop platform, RESTful APIs, flat files, and hopefully this is not the case, PDFs.
2. Exploration and understanding: The second step is to come to an understanding of the data that you will use and how it was collected; this often requires significant exploration.
3. Munging, wrangling, and manipulation: This step is often the single most time-consuming and important step in the pipeline. Data is almost never in the needed form for the desired analysis.
4. Analysis and modeling: This is the fun part where the data scientist gets to explore the statistical relationships between the variables in the data and pulls out his or her bag of machine learning tricks to cluster, categorize, or classify the data and create predictive models to see into the future.
5. Communicating and operationalizing: At the end of the pipeline, we need to give the data back in a compelling form and structure, sometimes to ourselves to inform the next iteration, and sometimes to a completely different audience. The data products produced can be a simple one-off report or a scalable web product that will be used interactively by millions.

A.1.8 Caffo, et al. on Data Science

Key: `Caffo+` Source: @caffoExecutiveDataScience2015 URL→
Field: Data Science

1. Question

2. Get

3. Explore

4. Model

5. Interpret

6. Communicate

A.1.9 Donaho on Data Science

Key: `Donoho` Source: @donoho50YearsData2017 URL→

Field: Statistics

1. Gather
2. Prepare
3. Explore
4. Represent and transform
5. Compute

6. Model

7. Present

8. Meta

A.1.10 Géron on Machine Learning

Key: `Géron` Source: @geronHandsOnMachineLearning2017 URL→
Field: Data Science

1. Big picture

2. Get

3. Clean

4. Discover
5. Prepare Model
6. Fine tune

7. Launch

A.1.11 Das on Data Science

Key: `Das` Source: @dasDataScienceLife2019 URL→
Field: Data Science

1. Understand
2. Mine
3. Clean
4. Explore
5. Features
6. Model
7. Visualize

A.1.12 Dataman on Data Science

Key: `Dataman` Source: @datamanDataScienceModeling2020 URL→
Field: Data Science

1. Business
2. Data requirements
3. Collection
4. EDA
5. Modeling
6. Evaluation
7. Deployment
8. Monitoring

A.1.13 Porter on Data Science

Key: `Porter` Source: @porterFrameworkDataScience2020 Field: Statistics

1. Collect
2. Store and represent
3. Manipulation
4. Computing
5. Analytics

6. Communicate

7. Practice
8. Disciplinary

A.2 Summary Table

Understand

Plan

Collect

Store

Clean

Explore

Prepare

Model

Interpret

Communicate

Deploy

Reflect

Tukey

Plan

Gather

Analyze

Interpret

KDD

Selection

Preprocessing

Transformation

Data mining

Interpreting

SEMMA

Sample

Explore

Modify

Model

Assess

Hayashi

Design

Collect

Analyze

CRISPDM

Business

Collect

Explore

Preparation

Modeling

Evaluation

Deployment

OSEMI

Obtain

Scrub

Explore

Model

Interpret

Ojeda+

Acquire

Explore and understand

Wrangle

Analyze and model

Communicate

Operationalize

Caffo+

Question

Get

Explore

Model

Interpret

Communicate

Donaho

Gather

Prepare

Explore

Represent and transform; Compute

Model

Present

Meta

Géron

Big picture

Get

Clean

Discover

Prepare

Model; Fine tune

Launch

Das

Understand

Mine

Clean

Explore

Features

Model

Visualize

Dataman

Business

Data requirements

Collection

EDA

Modeling

Evaluation

Deployment; Monitoring

Porter

Collect

Store and represent

Manipulation

Computing

Analytics

Communicate

Practice

Disciplinary

A.3 References

B Relation to AI

The four areas of data science defined here are surprisingly analogous to the four approaches to artificial intelligence defined by Russel and Norvig in their classic textbook on the subject [@russell1995: 5]. Their four part model was generated by combining the axes thinking/acting with human/rational as follows:

(1) thinking humanly

(2) thinking rationally

(3) acting humanly

(4) acting rationally

Moreover, it is easy to see how the following analogies make sense:

$$abstract : concrete :: thinking : action$$

and

$$human : rational :: human : machine$$

In fact, it appears that the same space is shared by the 4+1 model of data science and Russell and Norvig's model of artificial intelligence. The difference that the former is focused on forms of labor carried out by the data scientist, whereas the latter concerns forms of intelligence built by AI specialists.

References