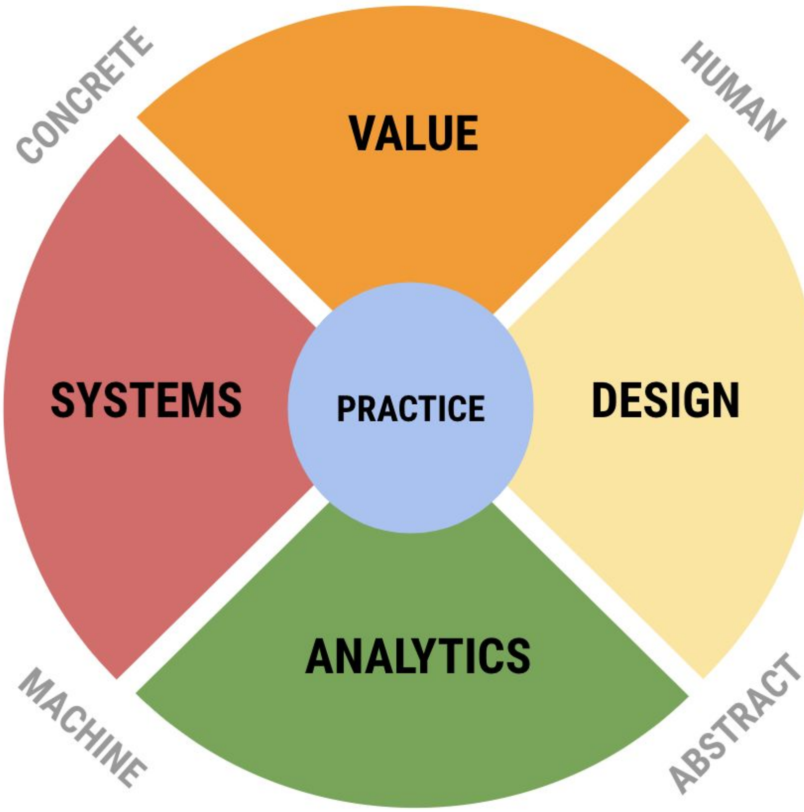
The 4 + 1 Model of Data Science

R.C. Alvarado | 19 October 2019 | The School of Data Science at UVa

DRAFT: DO NOT CITE WITHOUT PERMISSION OF AUTHOR



The 4+1 Model of Data Science (CC BY-SA 4.0)

Abstract

Data Science is a complex and evolving field, but most agree that it can be defined as the intersection of computer science and technology, math and statistics, and domain knowledge, with the purpose of extracting knowledge and value from data. We propose a new set of areas that builds on this definition but more accurately represents the field as it has grown and as it has come to be practiced. These are the areas of *value*, *design*, *systems*, and *analytics*. A fifth area, *practice*, connects the other four. Together, these areas define specific kinds of expertise that belong to every data science project but which are often unconnected and siloed in the academy. Unlike traditional disciplines, each new area is inherently interdisciplinary, bringing together diverse and sometimes contrary perspectives under a common theme. 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 Areas Defined

Value

The area of value 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 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.

- Key tensions: enterprise vs ethics, private interest vs public good (see Plato's *Republic*).
- Common theme: human value.
- Realm: concrete humanity.
- Keywords: ethics, justice, wealth, value, social good, motivation, meaning, *eudamonia*.
- Values: responsibility, diversity, inclusion, flourishing, excellence, *arete*.

Design

The area of design 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 human reality as data for the work of analytics, e.g. in database design, 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 humans 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. 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.

- Key tensions: discovery vs product, user-facing vs analyst-facing.
- Common theme: communication.
- Realm: abstract humanity.
- Keywords: communication, design thinking, representation, use interface, communication, visualization, visual and object languages, informatics, ontology, curation, HCI, informating, data product design, data discovery.
- Values: openness, authenticity, beauty, form and function.

Systems

The area of systems 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 pipelines 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.

- Key tensions: development vs production, volume vs speed.
- Common theme: building
- Realm: concrete machinery.
- Keywords: infrastructure, data systems, data engineering, the cloud, networks, hardware, software, programming languages, big data management, benchmarking, continuous integration, availability, cybersecurity.
- Values: speed, stability, robustness, resilience, uptime.

Analytics

The area of analytics includes what most consider to be the heart of data science, the combination of statistical methods with machine learning, along with optimization, signal processing, network analysis, and other rigorous quantitative methods from a variety of fields. Although unified by a broad commitment to advanced mathematical methods and algorithms, in reality this is a heterogeneous collection of competing methods and goals. Tensions include inference vs prediction, parametric vs non-parametric (kernel-based) methods, frequentist vs Bayesian statistics, analytic vs algorithmic solutions (including simulations), matrix vs graph-based methods, etc.

- Key tensions: inference vs prediction, analysis vs simulation.
- Common theme: Mathematical models and methods.
- Realm: abstract machinery.
- Keywords: prediction, inference, machine learning, statistics, operations research, AI, experimental design, causality, optimization, knowledge, models, feature engineering, data mining.
- Values: accuracy, precision, validity, truth, convergence, explainability.

Practice

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, effective, and efficient. Many activities that are considered an essential part of data science, such as data wrangling, actually exist only in practice, combining expertise in systems, design (representation), and analytics, and they are not usually taught in distinct classes. Practice is also where the core areas of data science come into contact with domain knowledge and real world problems. Although practice is depicted in the center of the diagram, it could also be viewed as the wider environment within which the field of data science operates. Or, practice may be envisioned as the water surrounding and connecting the four areas as islands:

Melting pingo wedge ice (*Wikimedia Commons*)

- Key tensions: goals vs resources; integration vs separation; planning vs play.
- Common theme: collaborative activity, praxis.
- Realm: umami (all four at once).
- Keywords: project management, stewardship, process, team sport, convergence, integration.
- Values: effectiveness, translation, impact, flow.

Interpreting the Model

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 may be purchased at the cost of intuitive understanding. The following caveats should help make sense of the model when considering its usefulness when applied to various concrete activities (discussed below).

The model emphasizes data science in relation to human and socially generated data. This is because data science, as opposed to the computational natural sciences, has historically been focused on the human, from human genomics and medicine to finance to education to social media. It should be understood that data science is not limited to working with human-related data, though. The field is likely to grow and merge with the natural sciences as the latter become more invested in working with data at scale.

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, we are pretty sure that 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.

The model describes academic areas for teaching and research, not service units for the university. For example, the area of systems is focused squarely on the internal needs of data science projects, not in providing infrastructure services to the university community. In fact, this area is likely to depend heavily on services that the university provides to the wider community, such as Research Computing.

Each area in the model is a big bucket containing 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 the school'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 areas of the model are like the components of a principal component analysis of the vector space of data science. They capture the variance that exists within the field, and, crucially, provide a framework for realigning (rebasing) the academy along a new set of axes. One effect of this is to both disperse and recombine older fields, such as computer science, statistics, and operations research, into new clusters. Thus we separate computer science subfields such as complexity analysis and database design. One possible salutary result of this will be the formation of new syntheses of fields that share concerns but differ in vocabularies and customs.

The dimensions 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 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). One reason these realms may be useful in grouping areas of expertise is that they may correlate with cognitive styles, although currently there is no research to support this claim (but see Appendix 2).

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.

Each area influences the others as a guiding principle. For example, although openness is part of design, it relates to how we build systems and perform analytics.

The entire model is conditioned on a concern with data. That is, when we say "design," we mean "data design," and so forth.

Uses of the Model

The usefulness of this model, and the proper context for its subsequent iteration, is in its application to the various areas of strategic decision-making that face a school of data science. Among these are:

Definition: The model can help us refine a definition of data science that will accurately encompass the variety of activities that fall within the scope of the field and remain consistent with prevailing definitions. We provide a definition below.

Curriculum design: The model can serve as categories for distribution requirements. See Appendix C.

Centers and collaboratories: Each area may be a center, or an umbrella within which centers are formed, with practice as a set of coordinated activities that involves capstones, graduate research, and collaboratories. This would also influence faculty hiring.

Architecture and spaces: Physical spaces might be organized on this model, perhaps as a quincunx, with areas associated with building sectors and practice as a literal commons.

Branding: The model may be used to symbolise the school, visually and verbally.

A Definition of Data Science

Given the validity of the five-part model of data science defined above, we may provide a succinct definition of data science as follows:

> Data science is a convergent field that integrates expertise from four broad areas of knowledge — value, design, systems, and analytics — with the purpose of extracting information, insight, and value from data in a responsible, authentic, and actionable manner.
>
> Data science concerns more than data analysis — it includes the broad and directly relevant contexts in which data analytic work takes place.

Derivation of the Model

[Insert earlier work on moving from divergent definitions to the pipeline.]

The model was originally derived through a structural analysis of narratives of data science. It is also corroborated by the results of text mining a corpus of data science documents from various sources.

Existing definitions in conflict

- As computational science (i.e. fourth paradigm).
- As Statistics 2.0 (i.e. data analysis; "50 years of Data Science").
- As Data mining and machine learning.

A pipeline in common

[Show example pipelines, including Cleveland's, Donoho's, Wiggins', etc., stacked to show common elements and sequential pattern.]

[Include the reconstituted pipeline.]

The pipeline is an arc with a structure

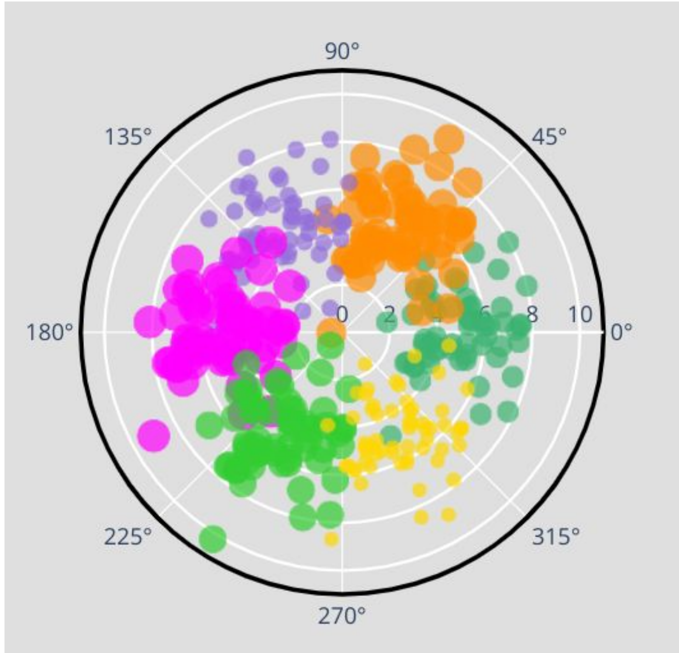
The pipeline defines a micronarrative with a four-part chiasmatic structure.

The onion model

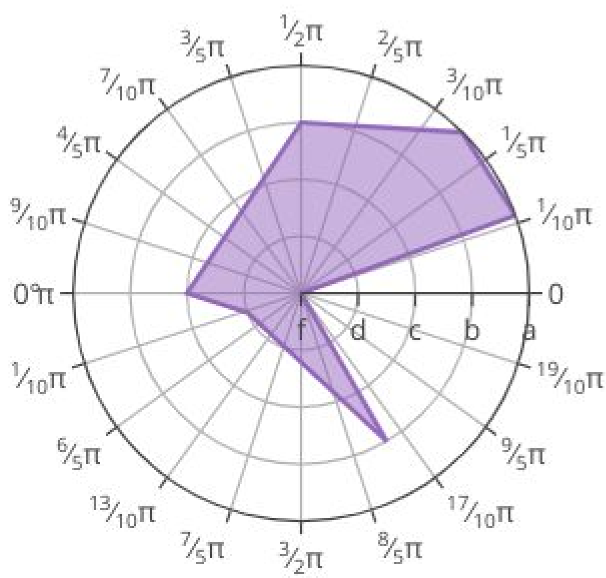
Realms

Applying the Model

[Describe various methods for using the model in a systematic way, e.g. a polar graph to plot topics and activities.]

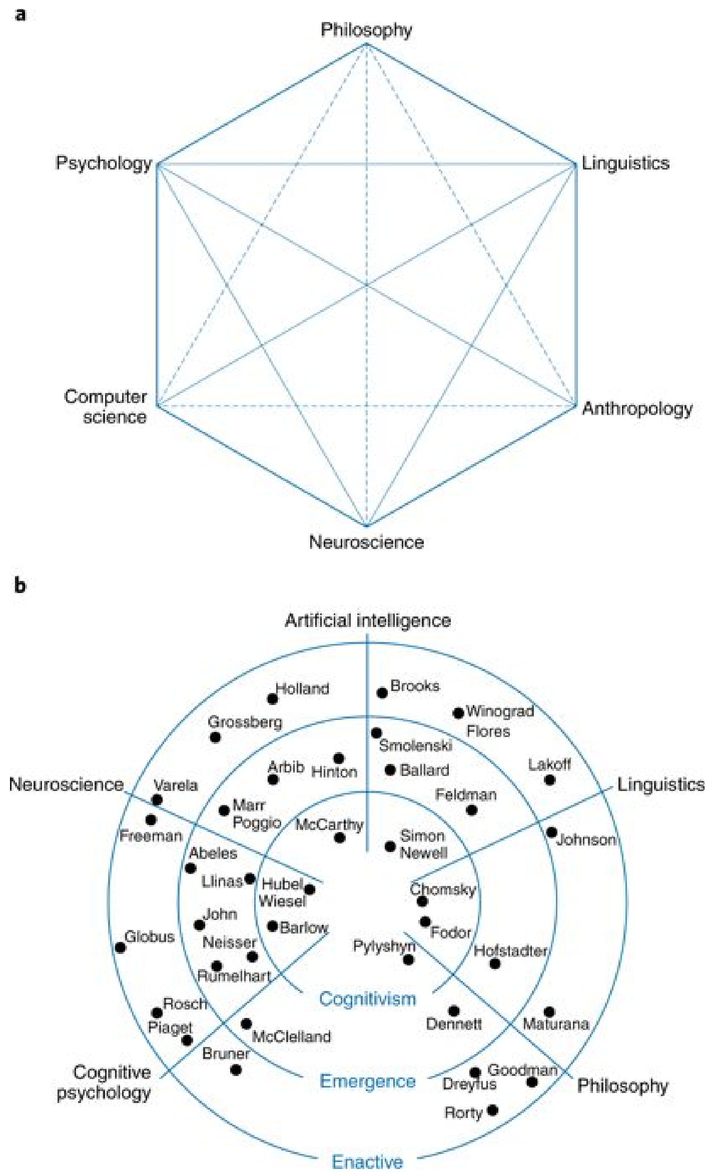


Example use of model as polar graph of categorical data. The spots represent specific topics, and their position on the graph shows that location within the areas of the model, now represented as quadrants in polar space.



An example of a plot of a topic in terms of how much it involves each area of expertise.

Here is another example, drawn from cognitive science, of how the model might be applied:

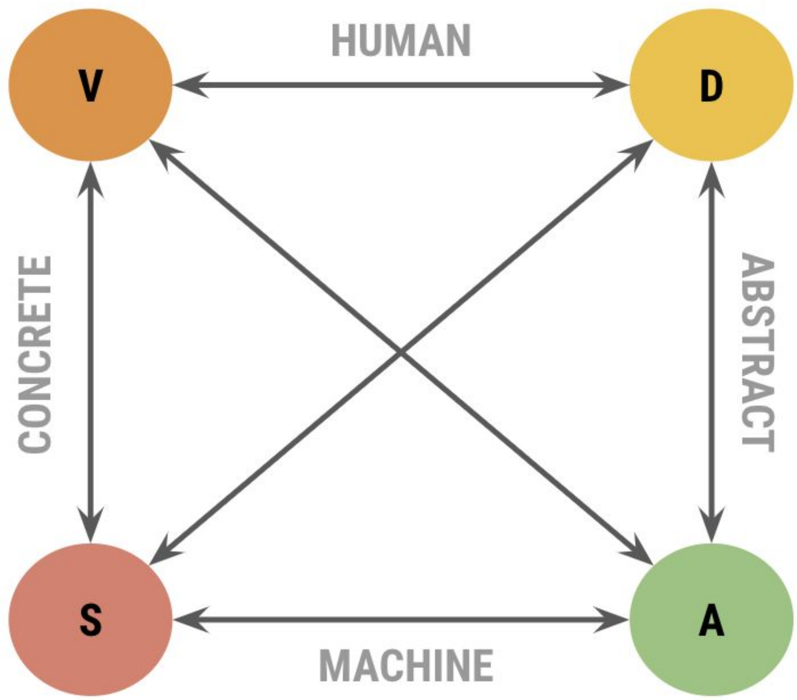


Example from Cognitive Science

Acknowledgements

Appendix 1: Relating the Areas

The central area of practice is the context within which each area is combined with the others. Within this area, it is helpful to consider how the four areas interact, given the idea that each area influences the others as a guiding principle. .



Value	+	**Design**	Openness, responsibility
Value	+	**Analytics**	Human-centered AI, algorithmic bias
Value	+	**Systems**	Sustainability, access, environmental impact
Design	+	**Analytics**	Elegance, literate programming, visualization
Design	+	**Systems**	Dashboards, engineering design
Analytics	+	**Systems**	Operational analytics, machine learning engineering

Appendix 2: Principal Components

The armature of the model consists of two major axes — the concrete and abstract axis, and the human and machine axis. These may be thought of as the two major 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. Here we give some explanation of the meaning and use of these axes within the model beyond what is described above. Principal components are understood to represent oppositions that define an implicit vector space onto which units if knowledge (to be specified) can be projected. As mentioned above, the reality behind these axes may be that they represent cognitive styles, a claim that, although unproven, is supported by the literary analogies described below.

PC1: Human versus Machine

The human-machine axis accounts for the most variance in the field. This seems evident from the fact that the traditional model of data science (Drew Conway's Venn diagram) describes only the machine side of our model (with practice replaced by "substantive expertise"). The human side — value and design — is left out, or at least 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 indirect representation of that experience. 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 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.

Literary analogies

The opposition between expertise in the concrete and the abstract is similar to that described by Herman Hesse in *Narcissus and Goldmund*, or perhaps that between Neuromancer and Wintermute in Gibson's *Neuromancer*.

The opposition between the human and the machine is similar to that described by Neal Stephenson between the Shaftoe and Waterhouse characters in *Cryptonomicon* and the *Baroque Cycle* or between Captain Aubrey and Stephen Maturin in Patrick O'Brian's *Master and Commander* series.

The opposition between apollonian and dionysian cultures described by Ruth Benedict also seems applicable. Apollo is clearly associated with the abstract and orderly, Dionysis with the concrete and messy.

Relation to Cognitive Architecture

The field of HCI has a concept of cognitive architecture, which has been described abstractly in terms very similar to the principal components described here.

Thinking / Acting

Humanly / Rationally

So:

Humanly acting

Humanly thinking

Rationally acting

Rationally thinking

Note that this is an analogy, not a possible replacement.

Compare to approaches to AI model (Russell and Norvig 2010)

Thinking : Acting :: Abstract : Concrete

Humanly / Rationally → Human / Machine

An analogy, not an identity!

Thinking Humanly "The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	**Thinking Rationally** "The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly "The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	**Acting Rationally** "Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Figure 1.1 Some definitions of artificial intelligence, organized into four categories.

Appendix C: Subareas and Courses

As stated above, one of the uses of the 4 + 1 model

Value

- Ethics of Big Data
- Social History of Data
- Value Propositions
- The Data-centric Organization
- Ethnography of Data
- Economics of Big Data

Design

- Forms of Data
- Data Product Design
- Data Discovery
- Data Curation
- Visualization and Communication
- Human-Centered AI
- Thinking through making / Making through thinking
- FAIR ecosystems
- Data sharing

Systems

- Database Systems
- Distributed High-Performance Architectures
- Cloud Computing
- Resilient Redundant Data
- Machine Learning Engineering
- Sensor Networks

Analytics

- Optimization Theory
- Information Theory
- Linear Models
- Graphical Models
- Data Mining
- Machine Learning
- Deep Learning
- Adversarial Models
- Bayesian Methods in Machine Learning
- Time-Series
- Text Analytics

Practice

- Project Management of Data Science

- Capstone Projects
- Ph.D. Research Projects
- Collaboratories
- Domain-specific courses

Liminal Topics

- Data preprocessing — between design (discovery) and analytics
- Explainable AI -- between design (communication) and analytics

Appendix D: Responses to Anticipated Criticisms

Where is data wrangling?

Short answer: Data wrangling is actually not thing, in the sense that it covers a broad variety of heterogeneous tasks. This is why it accounts for the proverbial 80% of a data scientist's time. It can include scraping data from websites and external sources, normalizing data formats (such as dates), migrating data from one platform for another, inferring data dictionaries, balancing observations to reduce bias, imputating missing values in a dataframe, feature engineering, ETL operations on source data, and so on. The fact that these tasks are often performed by a single person simply reflects the fact that many data scientists are "full stack" developers, a role that is changing as projects scale up and a division of labor becomes necessary. Given this, the set of things that may be described as data wrangling are distributed to different areas. For example, imputation is an activity within analytics, clarification of implicit data models belongs to design, and migrating data from one repository to another lies within systems. In practice, then, data wrangling refers to a more or less contiguous sequence of activities in the preprocessing stage of the data science pipeline. These activities may be accomplished by a single person or by members of a team of specialists.

Where does database administration go?

Short answer: The role of the DBA, which traditionally may include standing up and running databases as well as developing data models, is split between Systems and Design. This is a good thing, since it puts data modeling — a crucial but often hidden task — in the hands of people who are concerned to optimize models with respect to the human domain (value) and not simply in terms of systems efficiency.

This is a theory of everything — what use is it?

Short answer: We have to ascend the ladder of abstraction in order to reconstitute the categories we use to describe the field. Then we may descend the ladder with a new framework within which to situate concrete and detailed decisions. Otherwise we reproduce the old categories and data science continues to appear anomalous.

Remarks from Meeting 2019/11/06

Jon: How is this different from existing department? How will these topics flow? How do we avoid silos?

Don: How we implement this - we need to ensure tight linkage between course content. This is a challenge as we see with current MSDS. However, with this model we can design courses based off of this rather than what already exists.

Phil: Faculty will be part of centers rather than department silos. We need to have leadership in these areas.

Eric: These could be seen as departments, maybe we don't want to create centers based on these areas. Perhaps we require centers to have two areas (i.e. Design +systems).

Don: the devil is in the detail in how you create it.

Jon: Are we going to hire faculty in value, design, etc? How are we implementing across all academic areas?

Don: dependent on subcommittees development of curriculum, etc.

Tim: Reminiscent of a paper (https://en.wikipedia.org/wiki/4%2b1_architectural_view_model). Sets of attributes and dimensions.

Pete: this is not an org chart but trying to define a field.

Thoughts

- Differentiation through specification of method