Recap of Part I: Data Science

Data science has advanced in large part by pooling techniques and goals from statistics, operations research, and computing. In turn, these fields have changed due to data's impact. Continual learnings from these fields and others are further improving data science. Data scientists increasingly must consult other disciplines to craft solutions that meet the needs of a broad coalition.

While data science does not solve all problems, it has already had extraordinary impact, with its full potential yet to be realized. With improved processing of ever larger amounts of data, data science will continue to grow in importance and performance. A field of study is in part defined by its most important terms, so we conclude this section by reviewing 16 terms we think are critical to understanding data science.

Table I.1 reiterates our introductory section's five key terms.

Table I.2 reiterates three key statistics terms that data science uses. Inference is absolutely central to almost everything in this book, and understanding correlation and causation and their differences is essential for effectively applying and understanding data science techniques and results.

Operations research contributes to the major data science focus of optimization as illustrated by the terms in Table I.3.

From computing, Table I.4 includes three key terms.

From our ethics discussion, Table I.5 describes three key terms motivating ethical considerations when applying data science.

Table I.1 *Key terms from the definition of data science.*

Data science	The study of extracting value from data, as insights or conclusions.
Insights	Understanding what may arise from a new hypothesis that can be tested against data, from an apt visual chart, or from interactively exploring a complex model of the data, or trying out different scenarios and seeing the implications.
Conclusions	Learnings from data science of the form of prediction, recommendation, clustering, classification, transformation, or optimization.
Model	A representation of a subject system – an abstraction that emphasizes key ideas about the system and ignores extraneous details.
Big data	A body of techniques for conceiving, designing, developing, and running systems that gather, store, and process vast amounts of information.

Table I.2 *Key terms from statistics.*

Inference	The process of drawing conclusions about the properties of a population or a system. Inference is often used to test a falsifiable hypothesis, which is one which can be disproven.
Correlation	The relationship between two variables. They could be positively correlated (e.g., if one goes up, the other tends to go up), or negatively correlated (e.g., if one goes up, the other tends to go down), or perhaps uncorrelated (e.g., like the outcomes of rolling one fair die and then rolling another).
Causation	The relationship where an intervention in one variable (“the cause”) contributes to a change in the value of another variable (“the effect”). In searching for causal relationships, we are aided by a search for a mechanistic relationship between the cause and the effect, as in smoking’s causal relationship with cancer, or the causal relationship observed by John Snow between drinking-water contamination and cholera. Often, an interpretable relationship requires knowledge of one or more intermediaries and a pathway; for example, water contamination causing water-borne pathogens, and these pathogens causing cholera. Excellent data science sometimes leads to understanding causation. However, often, experimentation outside the realm of an established body of data is needed to definitely determine causation. It bears repeating that correlation does not imply causation, but correlation is correlated with causation.

Table I.3 *Key terms from operations research.*

Optimization	The selection of actions or values needed to generate a most desired outcome, usually subject to constraints mirroring those in the real world. For example, finding the shortest travel distance path for visiting specified cities. Optimization may be subject to constraints, such as travel time between any two cities not taking longer than a specified value, or vaccination availability being subject to certain fairness constraints.
Objective function	Represents a precise statement of a metric on which different outcomes can be compared, where a better outcome has a better value (which can be higher or lower). Sometimes, this is very simple to state. For example, in manufacturing, one wants to maximize how many parts can be made with a certain quantity of raw material. Objective functions can also be limited by constraints, such as requiring equity across subgroups.

Table I.4 *Key terms from computing.*

Algorithm	A clearly specified procedure that a computer can follow to perform a task.
Artificial intelligence	The study and construction of programs that act intelligently. They achieve their goals by examining their inputs and then taking appropriate actions.
Machine learning	A process that uses data to create a model, which a program can use to reach conclusions.

Table I.5 *Key terms from ethics.*

Respect for persons	Ensuring the freedom of individuals to act autonomously based on their own considered deliberation and judgments. Often summarized as *informed consent*, this principle also includes having sufficient transparency to make judgments.
Beneficence	This emphasizes *not* merely “do no harm,” but instead seeks to maximize the benefit from using data science both directly and for society at large. Doing so requires careful consideration of the immediate risks and benefits, as well as a commitment to monitor and mitigate new harms as results occur.
Justice	The consideration of how risks and benefits from using data science are distributed. This includes the notion of a *fair* distribution. Fair may not mean “equal,” but instead that benefits accrue according to factors such as one’s effort, contribution, merit, or need.