DATA SCIENCE CODE OF PROFESSIONAL CONDUCT

Terminology

Rule 1 - Terminology

(a) "Data" means a tangible or electronic record of raw (factual or non-factual) information (as measurements, statistics or information in numerical form that can be digitally transmitted or processed) used as a basis for reasoning, discussion, or calculation and must be processed or analyzed to be meaningful.

(b) "Data Science" means the scientific study of the creation, manipulation and transformation of data to create meaning.

(c) "Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

(d) "Data Quality" means rating the veracity of data.

(e) "Data Volume" means a measurement of the amount of data.

(f) "Data Variety" means the different types (written, numerical, sensor...etc) and structures (structured, unstructured, semi-structured) of data.

(g) "Data Velocity" means the measurable rate that data is collected, stored, analyzed and consumed.

(h) "Big Data" means large data sets that have different properties from small data sets and requires special data science methods to differentiate signal from noise to extract meaning and requires special compute systems and power.

(i) "Signal" means a meaningful interpretation of data based on science that may be transformed into scientific evidence and knowledge.

(j) "Noise" means a competing interpretation of data not grounded in science that may not be considered scientific evidence. Yet noise may be manipulated into a form of knowledge (what does not work).

(k) "Knowledge" means information backed by scientific evidence that creates meaning.

(l) "Machine Learning" means the field of study that gives computers the ability to learn without being explicitly programmed.

(m) "Algorithm" means a process or set of rules to be followed in calculations or other problem-solving operations to achieve a goal, especially a mathematical rule or procedure used to compute a desired result, produce the answer to a question or the solution to a problem in a finite number of steps.

(n) "Data Mining" means using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in data sets to discover new meaning in data.

(o) "Statistics" means the practice or science of collecting and analyzing numerical data in large quantities.

(p) "Statistically Significant" means a statistical assessment of whether observations reflect a pattern rather than just chance and may or may not be meaningful.

(q) "Correlation" means any of a broad class of statistical relationships involving dependence.

(r) "Spurious Correlation" means a correlation between two variables that does not result from any direct relation between them but from their relation to other variables.

(s) "Causation" means the relationship between cause and effect backed by scientific evidence (e.g. relationship between an event (the cause) and a second event (the effect), where the second event is understood as a consequence of the first).

(t) "Heuristics" means simple rules of thumb to assist in decision making or problem-solving by experimental and especially trial-and-error methods and the evaluation of feedback to improve performance. Simple and practical easy to apply rules of thumb that make life simple. These are necessary (we have no mental powers to absorb all information and tend to be confused by details) but they get us in trouble as we do not know we are using them when forming judgments.

(u) " Variable" means a value that may change within the scope of a given problem or set of operations and may be independent or dependent

(v) "Cherry picking" means pointing to individual cases or data that seem to confirm a particular position, while ignoring a significant portion of related cases or data that may contradict that position and may constitute scientific fraud, suppressing evidence, or the fallacy of incomplete evidence.

(w) "Correlation does not imply causation" is a phrase used in science and statistics to emphasize that a correlation between two variables does not necessarily imply that one causes

the other.

(x) "Substantial" when used in reference to degree or extent means a material matter of clear and weighty importance.

(y) "Predictive Analytics" means using techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts to help simulate scenario based decision making and make speculative, rationalistic and probabilistic predictions about future events (e.g. used in actuarial science, marketing, financial services, credit scoring, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields).

(z) "Nonpredictive approach" means designing and building stuff in a manner not based dependent on perturbations and thus durable and robust in changes in future outcomes.

(aa) "Ludic Fallacy" means mistaking the ecological complex real world to the well-posed problems of mathematics and laboratory experiments.

(bb) "Iatrogenics" means harm done by the healer, like the doctor doing more harm than good. Generalized Iatrogenics: By extension, applies to the side effects of data scientists, policy makers, researchers and academics.

(cc) "Naive Interventionism" means intervention with disregard to iatrogenics. The preference, even obligation, to “do something” over doing nothing. While this instinct can be beneficial in emergency rooms or ancestral environments, it hurts in others in which there is an “expert problem".

(dd) "Naive Rationalism" means thinking that the reasons for things are, by default, accessible to you.

(ee) “Confidential Information” means information that you create, develop, receive, use or learn in the course of employment as a data scientist for a client, either working directly in-house as an employee of an organization or as an independent professional. It includes information that is not generally known by the public about the client, including client affiliates, employees, customers or other parties with whom the client has a relationship and who have an expectation of confidentiality.

(ff) “Agency Problem” means moral hazard and conflict of interest may arise in any relationship where one party is expected to act in another's best interests. The problem is that the agent who is supposed to make the decisions that would best serve the principal is naturally motivated by self-interest, and the agent's own best interests may differ from the principal's best interests. The two parties have different interests and asymmetric information (the agent having more information), such that the principal cannot directly ensure that the agent is always acting in its (the principal's) best interests, particularly when activities that are useful to the principal are

costly to the agent, and where elements of what the agent does are costly for the principal to observe. Agents may hide risks and structure relationships so when he is right, he collects large benefits, when he is wrong, others pay the price. These also affect politicians and academics.

(gg) "Hammurabi Risk Management" means a builder has more knowledge than the inspector and can hide risks in the foundations.

(hh) "Ethical Inversion" means fitting one's ethics to actions (or profession) rather than the reverse.

(ii) "Protagoras Problem" means engaging in consequentially distorting assumptions "ifs" and calling it "science" and "evidence". The key is sincerity in assumptions.

(jj) "Narrative fallacy" means our need to fit a story, or pattern to series of connected or disconnected facts. The statistical application is data mining.

(kk) "Narrative discipline" is a discipline that consists in fitting a convincing and well-sounding story to the past. Opposed to experimental discipline. In medicine, epidemiological studies tend to be marred with the narrative fallacy, less so controlled experiments. Controlled experiments are more rigorous, free of cherry
picking.

(ll) "Rational Optionality" means not being locked into a given program, so one can change his mind as he goes.

(mm) "Subtractive knowledge" means you know what is wrong with more certainty than anything else. An application of via negativa.

(nn) "Via negativa" is the focus on what something is not, an indirect definition. In action, it is a recipe of what to avoid, what not to do - subtraction not addition, say, in medicine.
(oo) "Subtractive prophecy" means predicting the future by removing what is fragile from it, rather than naively adding to it. An application of via negativa.

(pp) "Thalesian thinking" focuses on exposure, payoff from decision.

(qq) "Aristotelian thinking" focuses on logic, the True-False distinction.

(rr) "Neomania" is a love of change for its own sake and forecasts the future by adding, not subtracting.

(ss) "Opacity " means the state or quality of being opaque (not transparent or hard to understand). Some things remain opaque to us, leading to illusions of understanding.

(tt) “Mediocristan” is a process dominated by the mediocre, with few extreme successes or failures (say income for a dentist). No single observation can meaningfully affect the aggregate. Also called “thin-tailed” or member of the Gaussian family of distributions.

(uu) “Extremistan” is a province where the total can be conceivably impacted by a single observation. Also called “fat-tailed”. Includes the fractal, or power-law family of distributions.

(vv) "Writing" or "written" denotes a tangible or electronic record of a communication or representation, including handwriting, typewriting, printing, photostating, photography, audio or videorecording, and electronic communications. A "signed" writing includes an electronic sound, symbol or process attached to or logically associated with a writing and executed or adopted by a person with the intent to sign the writing.

(ww) "Belief" or "believes" denotes that the person involved actually supposed the fact in question to be true. A person's belief may be inferred from circumstances.

(xx) "Fraud" or "fraudulent" denotes conduct that is fraudulent under the substantive or procedural law of the applicable jurisdiction and has a purpose to deceive.

(yy) "Informed consent" denotes the agreement by a person to a proposed course of conduct after the data scientist has communicated adequate information and explanation about the material risks of and reasonably available alternatives to the proposed course of conduct.

(zz) “Scientific method” means a method of research in which a problem is identified, relevant data are gathered, a hypothesis is formulated from these data, and the hypothesis is empirically tested. The data science method consists of the following steps: (1) Careful observations of data, data sets and relationships between data. (2) Deduction of meaning from the data and different data relationships. (3) Formation of hypotheses. (4) Experimental or observational testing of the validity of the hypotheses. To be termed scientific, a method of inquiry must be based on empirical and measurable evidence subject to specific principles of reasoning.

(aaa) "Knowingly," "known," or "knows" denotes actual knowledge of the fact in question. A person's knowledge may be inferred from circumstances.

(bbb) "Reasonable" or "reasonably" when used in relation to conduct by a data scientist denotes the conduct of a reasonably prudent and competent data scientist.

(ccc) "Reasonable belief" or "reasonably believes" when used in reference to a data scientist denotes that the data scientist believes the matter in question and that the circumstances are such that the belief is reasonable.

(ddd) "Reasonably should know" when used in reference to a data scientist denotes that a data scientist of reasonable prudence and competence would ascertain the matter in question.

Data Scientist - Client Relationship

Rule 2 - Competence

A data scientist shall provide competent data science professional services to a client. Competent data science professional services requires the knowledge, skill, thoroughness and preparation reasonably necessary for the services.

Rule 3 - Scope of Data Science Professional Services Between Client and Data Scientist

(a) Subject to paragraphs (b), a data scientist shall abide by a client's decisions concerning objectives of the services and shall consult with the client as to the means by which they are to be pursued. A data scientist may take such action on behalf of the client as is impliedly authorized to carry out data science professional services.

(b) A data scientist shall not counsel a client to engage, or assist a client, in conduct that the data scientist knows is criminal or fraudulent, but a data scientist may discuss the consequences of any proposed course of conduct with a client and may counsel or assist a client to make a good faith effort to determine the validity, scope, meaning or application of the data science provided.

Rule 4 - Communication with Clients

(a) A data scientist shall:

(1) reasonably consult with the client about the means by which the client's objectives are to be accomplished;

(2) act with reasonable diligence and promptness in providing services;

(3) keep the client reasonably informed about the status of the data science services;

(4) promptly comply with reasonable requests for information;

(5) consult with the client about any real, perceived and potentially hidden risks in relying on data science results; and

(6) consult with the client about any relevant limitation on the data scientist's conduct when the data scientist knows that the client expects assistance not permitted by the Code of Professional Conduct or other law.

(b) A data scientist shall explain data science results to the extent reasonably necessary to permit the client to make informed decisions regarding the data science.

Rule 5 - Confidential Information

(a) Confidential information is information that the data scientist creates, develops, receives, uses or learns in the course of employment as a data scientist for a client, either working directly in-house as an employee of an organization or as an independent professional. It includes information that is not generally known by the public about the client, including client affiliates, employees, customers or other parties with whom the client has a relationship and who have an expectation of confidentiality. The data scientist has a professional duty to protect all confidential information, regardless of its form or format, from the time of its creation or receipt until its authorized disposal.

(b) Confidential information is a valuable asset. Protecting this information is critical to a data scientists reputation for integrity and relationship with clients, and ensures compliance with laws and regulations governing the client's industry.

(c) A data scientist shall protect all confidential information, regardless of its form or format, from the time of its creation or receipt until its authorized disposal.

(d) A data scientist shall not reveal information relating to the representation of a client unless the client gives informed consent, the disclosure is impliedly authorized in order to carry out the representation or the disclosure is permitted by paragraph (e).

(e) A data scientist may reveal information relating to the representation of a client to the extent the data scientist reasonably believes necessary:

(1) to prevent reasonably certain death or substantial bodily harm;

(2) to prevent the client from committing a crime or fraud that is reasonably certain to result in substantial injury to the financial interests or property of another and in furtherance of which the client has used or is using the data scientist's services.

(f) A data scientist shall make reasonable efforts to prevent the inadvertent or unauthorized disclosure of, or unauthorized access to, information relating to the representation of a client, which means:

(1) Not displaying, reviewing or discussing confidential information in public places, in the presence of third parties or that may be overheard;

(2) Not e-mailing confidential information outside of the organization or professional practice to a personal e-mail account or otherwise removing confidential information from the client by

removing hard copies or copying it to any form of recordable digital media device; and

(3) Communicating confidential information only to client employees and authorized agents (such as attorneys or external auditors) who have a legitimate business reason to know the information.

(g) A data scientist shall comply with client policies that apply to the acceptance, proper use and handling of confidential information, as well as any written agreements between the data scientist and the client relating to confidential information.

(h) A data scientist shall protect client confidential information after termination of work for the client.

(i) A data scientist shall return any and all confidential information in possession or control upon termination of the data scientist - client relationship and, if requested, execute an affidavit affirming compliance with obligations relating to confidential information.

Rule 6 - Conflicts of Interest

(a) Except as provided in paragraph (b), a data scientist shall not provide professional data science services for a client if the services involves a concurrent conflict of interest. A concurrent conflict of interest exists if:

(1) providing services for one client will be directly adverse to another client; or

(2) there is a significant risk that providing professional data science services for one or more clients will be materially limited by the data scientist's responsibilities to another client, a former client or a third person or by a personal interest of the data scientist.

(b) Notwithstanding the existence of a concurrent conflict of interest under paragraph (a), a data scientist may represent a client if:

(1) the data scientist reasonably believes that the data scientist will be able to provide competent and diligent services to each affected client;

(2) the professional data science services is not prohibited by law; and

(3) each affected client gives informed consent, confirmed in writing.

Rule 7 - Duties to Prospective Client

(a) A person who consults with a data scientist about the possibility of forming a client-data scientist relationship with respect to a matter is a prospective client.

(b) Even when no client-data scientist relationship ensues, a data scientist who has learned information from a prospective client shall not use or reveal that information.

(c) A data scientist subject to paragraph (b) shall not provide professional data science services for a client with interests materially adverse to those of a prospective client in the same or a substantially related industry if the data scientist received information from the prospective client that could be significantly harmful to that person in the matter, except as provided in paragraph (d).

(d) When the data scientist has received disqualifying information as defined in paragraph (c), providing professional data science services is permissible if:

(1) both the affected client and the prospective client have given informed consent, confirmed in writing, or:

(2) the data scientist who received the information took reasonable measures to avoid exposure to more disqualifying information than was reasonably necessary to determine whether to provide professional data science services for the prospective client; and written notice is promptly given to the prospective client.

Rule 8 - Data Science Evidence, Quality of Data and Quality of Evidence

(a) A data scientist shall inform the client of all data science results and material facts known to the data scientist that will enable the client to make informed decisions, whether or not the data science evidence are adverse.

(b) A data scientist shall rate the quality of data and disclose such rating to client to enable client to make informed decisions. The data scientist understands that bad or uncertain data quality may compromise data science professional practice and may communicate a false reality or promote an illusion of understanding. The data scientist shall take reasonable measures to protect the client from relying and making decisions based on bad or uncertain data quality.

(c) A data scientist shall rate the quality of evidence and disclose such rating to client to enable client to make informed decisions. The data scientist understands that evidence may be weak or strong or uncertain and shall take reasonable measures to protect the client from relying and making decisions based on weak or uncertain evidence.

(d) If a data scientist reasonably believes a client is misusing data science to communicate a false reality or promote an illusion of understanding, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to

the proper authorities. The data scientist shall take reasonable measures to persuade the client to use data science appropriately.

(e) If a data scientist knows that a client intends to engage, is engaging or has engaged in criminal or fraudulent conduct related to the data science provided, the data scientist shall take reasonable remedial measures, including, if necessary, disclosure to the proper authorities.

(f) A data scientist shall not knowingly:

(1) fail to use scientific methods in performing data science;

(2) fail to rank the quality of evidence in a reasonable and understandable manner for the client;

(3) claim weak or uncertain evidence is strong evidence;

(4) misuse weak or uncertain evidence to communicate a false reality or promote an illusion of understanding;

(5) fail to rank the quality of data in a reasonable and understandable manner for the client;

(6) claim bad or uncertain data quality is good data quality;

(7) misuse bad or uncertain data quality to communicate a false reality or promote an illusion of understanding;

(8) fail to disclose any and all data science results or engage in cherry-picking;

(9) fail to attempt to replicate data science results;

(10) fail to disclose that data science results could not be replicated;

(11) misuse data science results to communicate a false reality or promote an illusion of understanding;

(12) fail to disclose failed experiments or disconfirming evidence known to the data scientist to be directly adverse to the position of the client;

(13) offer evidence that the data scientist knows to be false. If a data scientist questions the quality of data or evidence the data scientist must disclose this to the client. If a data scientist has offered material evidence and the data scientist comes to know of its falsity, the data scientist shall take reasonable remedial measures, including disclosure to the client. A data scientist may disclose and label evidence the data scientist reasonably believes is false;

(14) cherry-pick data and data science evidence.

(g) A data scientist shall use reasonable diligence when designing, creating and implementing algorithms to avoid harm. The data scientist shall disclose to the client any real, perceived or hidden risks from using the algorithm. After full disclosure, the client is responsible for making the decision to use or not use the algorithm. If a data scientist reasonably believes an algorithm will cause harm, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to the proper authorities. The data scientist shall take reasonable measures to persuade the client to use the algorithm appropriately.

(h) A data scientist shall use reasonable diligence when designing, creating and implementing machine learning systems to avoid harm. The data scientist shall disclose to the client any real, perceived or hidden risks from using a machine learning system. After full disclosure, the client is responsible for making the decision to use or not use the machine learning system. If a data scientist reasonably believes the machine learning system will cause harm, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to the proper authorities. The data scientist shall take reasonable measures to persuade the client to use the machine learning system appropriately.

(i) A data scientist shall use reasonable diligence when assigning value and meaning to the following concepts when conducting data science:

(1) "Statistically Significant"

(2) "Correlation"

(3) "Spurious Correlation"

(4) "Causation"

(j) A data scientist shall not engage in "Cherry picking" (pointing to individual cases or data that seem to confirm a particular position, while ignoring a significant portion of related cases or data that may contradict that position) when conducting data science. The data scientist understands that engaging in "Cherry picking" may constitute scientific fraud, suppressing evidence, or the fallacy of incomplete evidence.

(k) A data scientist shall not present incomplete evidence as real data science evidence. A data scientist may present a theory constituting incomplete evidence but shall label and clearly communicate the use of incomplete evidence.

(l) A data scientist shall use reasonable diligence to question assumptions and avoid engaging in consequentially distorting assumptions “if’s” and calling it “science” and “evidence” (also known

as the “Protagoras Problem”).

(m) A data scientist shall use reasonable diligence to recognize, disclose and factor “agency problems” when conducting data science. The prudent data scientist understands that agents may hide risks and structure relationships so when he is right, he collects large benefits, when he is wrong, others pay the price.

(n) A data scientist shall use reasonable diligence to detect, recognize, disclose and factor real, perceived and potentially hidden risks in using data science. The prudent data scientist understands that data creators and the designers and builders of data management systems have more knowledge than the data scientist and can hide risks in the foundations and interpretations / bias of the raw, created and manipulated data. The data scientist shall take reasonable remedial measures, including disclosure of risks to the client.

(o) A data scientist shall use the data science method which consists of the following steps:

(1) Careful observations of data, data sets and relationships between data;

(2) Deduction of meaning from the data and different data relationships;

(3) Formation of hypotheses;

(4) Experimental or observational testing of the validity of the hypotheses. To be termed scientific, a method of inquiry must be based on empirical and measurable evidence subject to specific principles of reasoning.

Maintaining the Integrity of the Data Science Profession

Rule 9 - Misconduct

It is professional misconduct for a data scientist to knowingly:

(a) violate or attempt to violate the Data Science Code of Professional Conduct, knowingly assist or induce another to do so, or do so through the acts of another;

(b) commit a criminal act related to the data scientist's professional services;

(c) engage in data science involving dishonesty, fraud, deceit or misrepresentation;

(d) engage in conduct that is prejudicial to methods of science;

(e) misuse data science results to communicate a false reality or promote an illusion of understanding.