The Data Life Cycle

generation	collection	processing	storage	management	analysis	visualization	interpretation
"People generate data: every search query we perform, link we click, movie we watch, book we read, picture we take, message we send, and place we go contribute to the massive digital footprint we each generate" *[Think also of historical source documents]*	*"Not all data generated is collected, perhaps out of choice because we do not need or want to, or for practical reasons…. Deciding what to collect defines a filter on the data we generate"*	*"everything from data cleaning, data wrangling, and data formatting to data compression, for efficient storage, and data encryption, for secure storage"*	*"the bits are laid down in memory"*	*"We are careful to store our data in ways both to optimize expected access patterns and to provide as much generality as possible. We need to create and use different kinds of meta data for these dimensions of heterogeneity to maximize our ability to access and modify the data for subsequent analysis"*	*"all the computational and statistical techniques for analyzing data for some purpose: the algorithms and methods that underlie artificial intelligence (AI), data mining, machine learning, and statistical inference, be they to gain knowledge or insights, build classifiers and predictors, or infer causality"*	*"helps present results in a clear and simple way that a human can readily understand and visualize"*	*"we provide the human reader an explanation of what the picture means. We tell a story explaining the picture's context, point, implications, and possible ramifications"*

Jeannette M. Wing (2019) "The Data Life Cycle" *Harvard Data Science Review*

Lessons for big data

generation	collection	processing	storage	management	analysis	visualization	interpretation
Dataset bias (Deon C.2)	*Informed consent (Deon A.1; DEDA 28–29)*	*PII exposure & anonymity (Deon A.3; DEDA 5, 19–21; DHR)*	*Data security & breaches (Deon B.1, DHR)*	*Access & reuse (DEDA 10– 11)*	*Algorithm/ machine bias (DHR)*	*Honest representation (Deon C.3; DEDA 9)*	*Explainability (Deon D.4; DEDA 1)*
Data quality (Deda 3–4)	*Collection bias (Deon A.2)*	*Discrimination through proxy variables (Deon D.1)*	*Right to be forgotten (Deon B.2)*	*Data retention plan (Deon B.3)*	*Honest representation (Deon C.3)*	*Embodiment & affect (FDV 3.5)*	*Communication bias (Deon D.5)*
Documenting sources (DEDA 2)	*Metric selection (Deon D.3; DEDA 26)*	*Algorithm/ machine bias (DHR)*	*Encryption (DEDA 6)*	*Compliance (DEDA 13, 22)*	*Privacy in analysis (Deon C.4)*		*Consider contexts (FDV 3.4)*
	Binaries (FDV 3.1)			*Responsibility (DEDA 14, 17)*	*Auditability (Deon C.5)*		*Future implications (DEDA 27)*
				Rollback possible? (Deon E.2)			
				Labor & power (FDV 3.3 & 3.6)			

use/deployment

Fairness across groups / Discrimination (Deon D.2; DEDA 15; DHR)

Targeting, manipulation, injury (DHR)

Unintended uses (Deon E.4; DEDA 12)

throughout

Missing perspectives (Deon C.1; DEDA 25; FDV 3.2 & 3.3)

Concept drift (Deon E.3)

Awareness of bias (DEDA 23–24)

Transparency (DEDA 18)

Redress (Deon E.1; DHR)

Choose one area and read the links there. What's missing? Add it to the map.

generation	collection	processing	storage	management	analysis	visualization	interpretation
Dataset bias (Deon C.2)	*Informed consent (Deon A.1; DEDA 28–29)*	*PII exposure & anonymity (Deon A.3; DEDA 5, 19–21; DHR)*	*Data security & breaches (Deon B.1, DHR)*	*Access & reuse (DEDA 10– 11)*	*Algorithm/ machine bias (DHR)*	*Honest representation (Deon C.3; DEDA 9)*	*Explainability (Deon D.4; DEDA1)*
Data quality (Deda 3–4)	*Collection bias (Deon A.2)*	*Discrimination through proxy variables (Deon D.1)*	*Right to be forgotten (Deon B.2)*	*Data retention plan (Deon B.3)*	*Honest representation (Deon C.3)*	*Embodiment & affect (FDV 3.5)*	*Communication bias (Deon D.5)*
Documenting sources (DEDA 2)	*Metric selection (Deon D.3; DEDA26)*	*Algorithm/ machine bias (DHR)*	*Encryption (DEDA 6)*	*Compliance (DEDA 13, 22)*	*Privacy in analysis (Deon C.4)*		*Consider contexts (FDV 3.4)*
	Binaries (FDV 3.1)			*Responsibility (DEDA 14, 17)*	*Auditability (Deon C.5)*		*Future implications (DEDA 27)*
				Rollback possible? (Deon E.2)			
				Labor & power (FDV 3.3 & 3.6)			

use/deployment

Fairness across groups / Discrimination (Deon D.2; DEDA 15; DHR)

Targeting, manipulation, injury (DHR)

Unintended uses (Deon E.4; DEDA 12)

throughout

Missing perspectives (Deon C.1; DEDA 25; FDV 3.2 & 3.3)

Concept drift (Deon E.3)

Awareness of bias (DEDA 23–24)

Transparency (DEDA 18)

Redress (Deon E.1; DHR)