



**DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI
DEL TERRITORIO E DELLE COSTRUZIONI**

**RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE**

Metodi di Progettazione per la Data Science

SINTESI

RELATORI

Prof. Ing. Gualtiero Fantoni
*Dipartimento di Ingegneria Civile e Industriale,
Università di Pisa*

Ing. Filippo Chiarello
*Dipartimento di Ingegneria Civile e Industriale,
Università di Pisa*

IL CANDIDATO

Gianluca Ruggi
g.ruggi@studenti.unipi.it

Sessione di Laurea Magistrale del 29/04/2020
Anno accademico 2018/2019
Consultazione non consentita

Sommario

La presente Tesi è incentrata sul rapporto tra un campo di studi nato nell'ultimo decennio, quello della Data Science, ed i temi legati alla gestione dei processi e al Design Thinking.

Il lavoro è diviso in due parti: la prima ha avuto come obiettivo quello di inquadrare il processo da un punto di vista Gestionale, dapprima analizzando le fonti disponibili in letteratura, ed in seguito da un questionario diretto ai professionisti del settore, in modo da reperire informazioni di interesse su parametri del processo ed in particolar modo sui rischi e le criticità che lo contraddistinguono.

La seconda parte del lavoro, chiariti gli aspetti fondamentali del processo, ha avuto come oggetto l'applicazione del Design Thinking che ha portato a dare una visione gestionale al processo tipicamente eseguito da chi si occupa della Scienza dei Dati.

Le analisi svolte hanno avuto come principali risultati una nuova mappatura del processo ed un primo metodo per la progettazione del processo, aprendo una nuova strada improntata al design, per integrare al massimo delle possibilità la Data Science nei processi aziendali.

Abstract

This Thesis focuses on the relation between Data Science, Project Management and Design Thinking.

The work is divided into two parts: the first has the goal of framing the process from an Engineering Management point of view, first of all by analyzing the available sources in the literature, and subsequently by developing a questionnaire to be answered by Data Science professionals. This allowed to map and collect information about relevant parameters of the Data Science process and especially about risks and critical aspects that distinguish it.

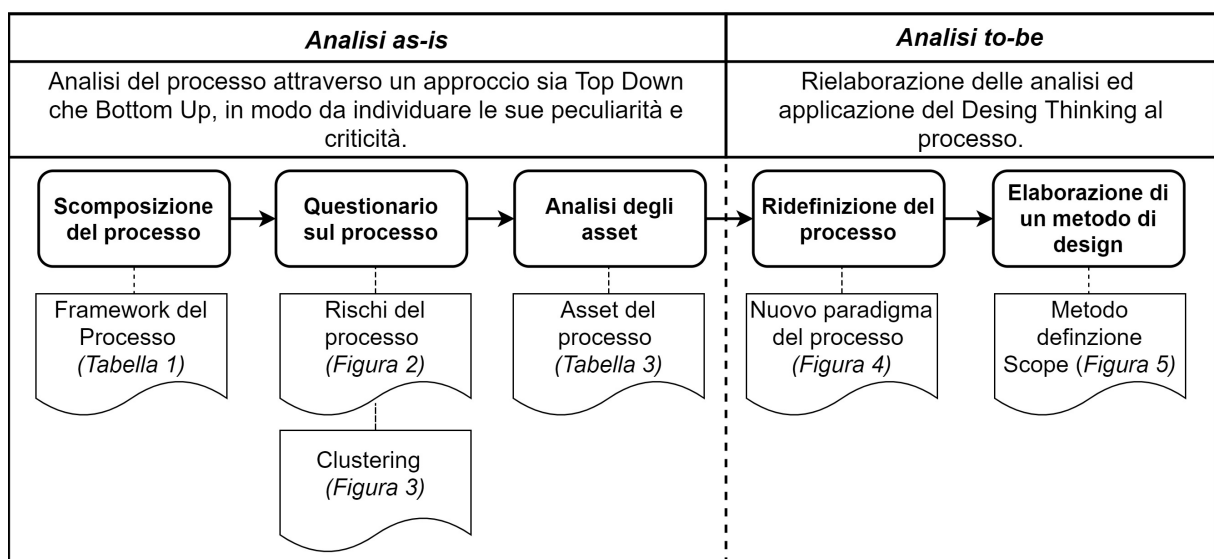
The second part of the work, clarified the fundamental features of the process, focuses on the study and application of Design Thinking, reviewing the paradigm of Data Science process itself.

The most important findings coming from the analysis are a new map of the Data Science process and a method for activities design, opening a new design focused path, that helps to better integrate Data Science into the business processes.

1. Introduzione

Il lavoro svolto ha l'intento di affrontare con la lente dell'ingegneria gestionale i processi di Data Science. Questo campo di studi è solo parzialmente nuovo, dato che combina conoscenze da ambiti diversi e già largamente affrontati come Computer Science, Ingegneria del Software, Statistica applicata, Intelligenza Artificiale e Data Management. Ciò che è nuovo, ed è la motivazione per la quale è sensato rivisitare tali processi da un punto di vista gestionale, è il tema dell'integrazione di queste diverse discipline (e quindi competenze). *"L'obiettivo dei progetti di Data Science è quello di estrarre conoscenza e conclusioni dai dati raccolti"*¹. Questa visione traccia un sentiero condiviso, ma che nel corso della trattazione verrà messo in discussione. Il focus della tesi è infatti posto sul fornire una prospettiva gestionale, volta al Design Thinking, al tema della Data Science, gettando le basi per colmare alcune delle lacune che dovranno essere affrontate dalla ricerca, per integrare al massimo delle possibilità la Data Science nei processi aziendali. Infatti, *"mentre molto è stato scritto in termini di utilizzo di algoritmi, molto meno è stato scritto su metodologie, strumenti e framework che potrebbero consentire ai team di essere più efficaci"*². La gran parte della letteratura a riguardo prende in esame solamente metodologie per la gestione dell'intero processo, fornendo spesso strumenti dallo scarso contenuto operativo. È necessario fare un passo avanti e scendere ad un livello di dettaglio più fine, sviluppando strumenti relativi a specifiche fasi e attività. Il lavoro svolto è stato organizzato come definito in Figura 1:

Figura 1 Metodologia utilizzata per lo sviluppo della presente Tesi.



¹ Rybicki, J. (2018, September). Best Practices in Structuring Data Science Projects. In International Conference on Information Systems Architecture and Technology (pp. 348-357). Springer, Cham.

² Saltz, J. S. (2015, October). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2066-2071). IEEE.

L'obiettivo primario è stato quello di aprire la strada verso questo tipo di approccio, spostando il punto di vista sul processo e sulla Data Science.

I risultati proposti mostrano la necessità di sottrarsi ad una visione strettamente tecnica del processo della Scienza dei Dati, dando spazio ad una nuova prospettiva, guidata dalla progettazione, ossia, da un processo sistematico e intelligente, oltre che creativo, atto a indirizzare il processo verso le giuste domande e decisioni, usando tecniche non necessariamente composte da formule o algoritmi³

2. Scomposizione del processo

Il primo passo compiuto per lo studio del processo è stata una sua scomposizione in parti omogenee più facili da comprendere e da analizzare, definendo le fasi e le attività che lo compongono. Le fonti aventi gli stessi propositi già presenti in letteratura⁴, mostravano il principale difetto di non presentare in maniera chiara e definita le attività collegate alle fasi del processo, rimanendo ad un livello di dettaglio troppo aggregato. L'obiettivo qui posto invece, facendo un passo avanti rispetto ai precedenti lavori, è stato quello di collegare le fasi del processo ad un numero ben definito di attività, che fossero espresse in una forma chiara, sintetica e coerente.

Mentre le fasi erano già definite a livello macro, ed è stato quindi possibile prenderle come riferimento⁵, per le attività è stato necessario uno sforzo di rielaborazione maggiore, approfondendo le descrizioni delle singole fasi presenti nei testi ed individuando i punti in comune. Il risultato è riportato in Tabella 1:

Tabella 1 Nuovo framework del processo di Data Science

<i>n</i>	<i>Fase</i>	<i>n</i>	<i>Attività</i>
1	setting the research goal	1.1	understand business problem
		1.2	define the research goal
		1.3	frame the problem analytically
2	data retrieving	2.1	define the data requirements
		2.2	collect data
		2.3	check data quality
3	data preparation	3.1	combine multiple data sources
		3.2	cleanse data
		3.3	transform data

³ Dym, C. L., & Little, P. (1999). Engineering design: A project-based introduction. John Wiley and sons.

⁴ Cielen, D., Meysman, A., & Ali, M. (2016). Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co.; Rollins, J. B. (2015). Foundational Methodology for Data Science, IBM.

⁵ Cielen, D., Meysman, A., & Ali, M. (2016). Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co..

4	data exploration	4.1	analyze data
		4.2	understand the data
5	data modeling	5.1	define the modeling technique
		5.2	build the model
		5.3	check model's quality
6	presentation and automation	6.1	develop research report
		6.2	communicate model features
		6.3	communicate research results

Tale suddivisione del processo in fasi ed attività, è stata poi utilizzata per costruire le domande del questionario e come riferimento per lo studio del processo.

3. Questionario

Al fine di reperire informazioni sul processo attraverso un approccio Bottom-Up, è stato elaborato un questionario diretto ai professionisti che lavorano nel campo della Data Science, in modo da sfruttare la loro esperienza diretta. Il questionario aveva l'obiettivo di reperire informazioni riguardo:

- le caratteristiche del processo e delle sue fasi;
- la suddivisione delle attività del processo tra i Data Scientist;
- l'approccio dei Data Scientist verso i metodi e il design;
- le criticità ed i rischi del processo e delle sue fasi.

Il lavoro è stato organizzato seguendo quattro passi sequenziali:

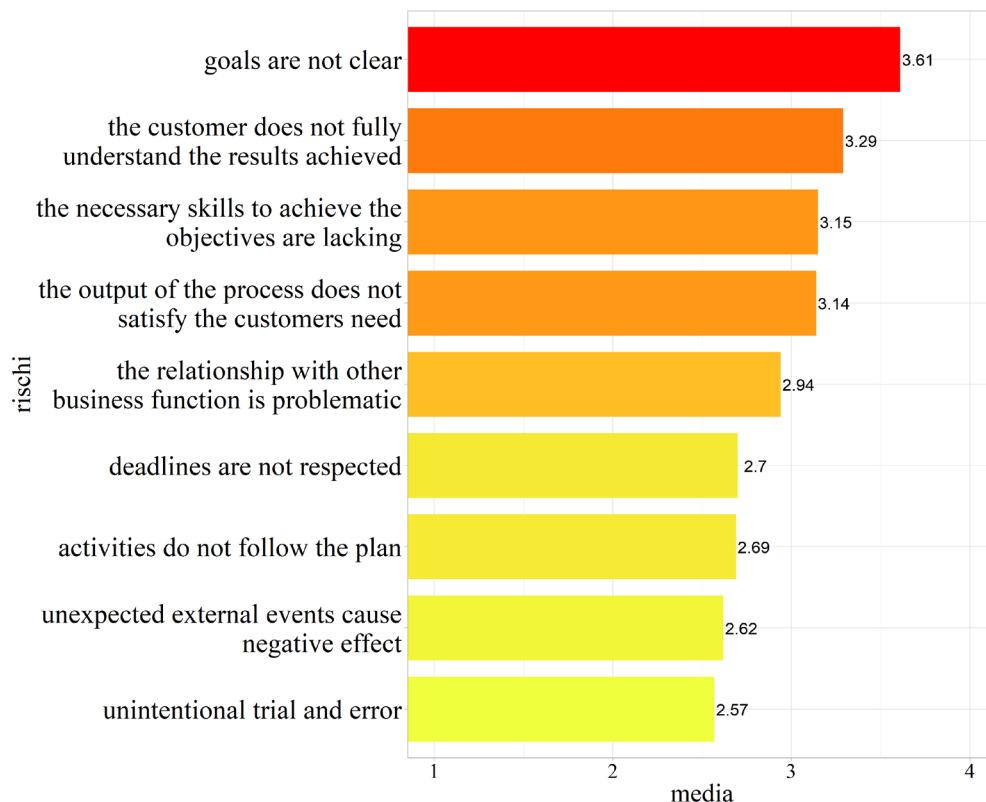
- Progettazione del questionario: si sono individuati gli obiettivi della raccolta dati e sulla base di questi si è elaborata una prima versione del questionario;
- Testing del questionario: la prima versione è stata sottoposta all'attenzione dei Data Scientist operanti presso aziende quali, Texty s.r.l., ErreQuadro s.r.l., TOI s.r.l. e GATE s.p.a., rielaborando il questionario sulla base dei commenti ricevuti, è stata approntata una versione finale;
- Pubblicazione del questionario: per la pubblicazione, i canali scelti sono stati i Social Network, in particolare LinkedIn, Facebook e Reddit, postando unicamente in pagine o gruppi riguardanti l'argomento Data Science. La pubblicazione ha seguito criteri rigidi, in modo da ricevere risposte unicamente dal target desiderato, ed è iniziata in data 20/12/2019, e terminata in data 14/01/2020, raccogliendo 75 risposte (tempo di compilazione medio di 12 minuti, per un totale di circa 14 ore totali).
- Analisi dei dati: i dati raccolti, dopo essere state sottoposti ad un controllo, che ha portato all'eliminazione di 4 risposte anomale, sono stati analizzati attraverso il

Software di programmazione R Studio. I risultati più rilevanti sono riassunti nella Tabella 2.

Tabella 2 Descrizione delle analisi maggiormente rilevanti tra quelle effettuate sui dati ricavati dal questionario.

Argomento	Descrizione	Analisi	Scopo
Rischi del processo (Figura 2).	I rispondenti dovevano valutare, in una scala da 1 a 4, l'impatto di 9 rischi sullo svolgimento dell'intero processo.	Media del totale dei punteggi assegnati dai rispondenti ad ogni rischio.	Individuare i rischi più impattanti sul processo.
Attività del processo (Figura 3).	I rispondenti dovevano valutare il tempo da loro personalmente allocato su ognuna delle 17 attività del processo.	Clustering dei rispondenti con algoritmo PAM.	Valutare la suddivisione in ruoli e la divisione delle competenze nella Data Science.

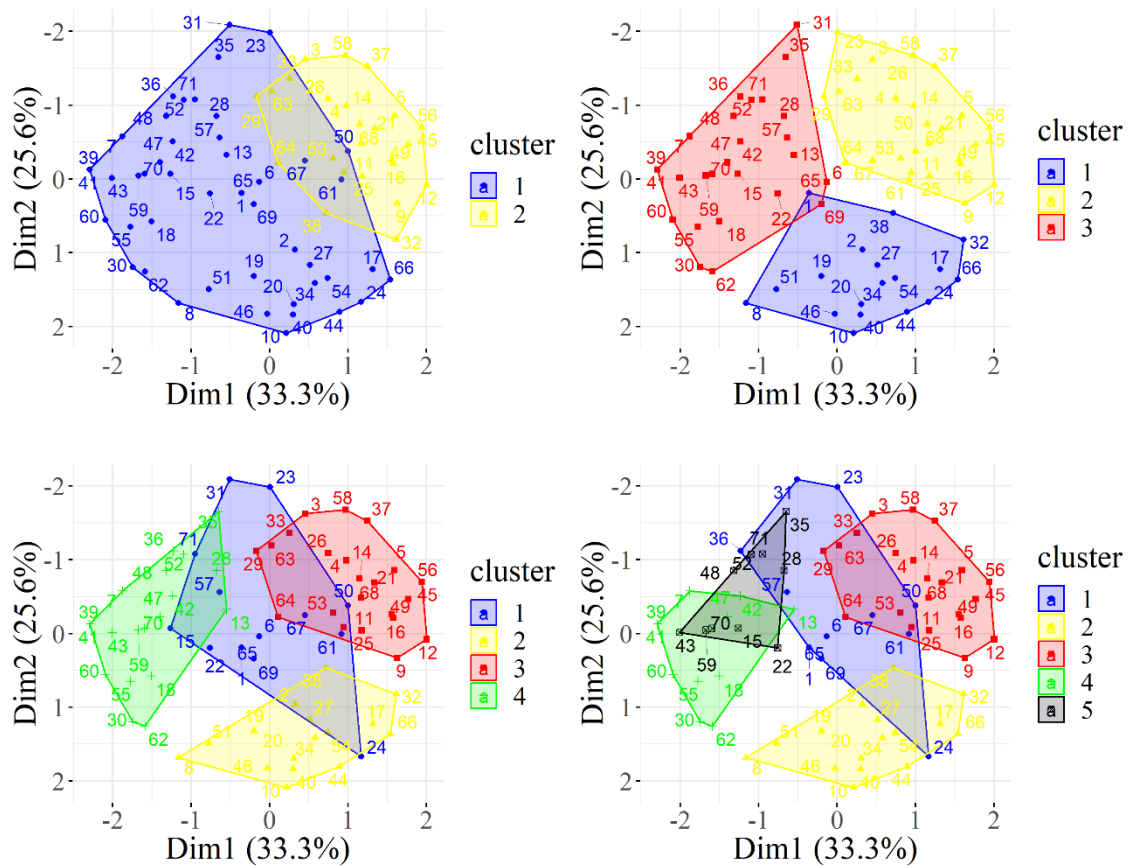
Figura 2 Grafico a colonne raffigurante le medie dei punteggi ottenuti dai rischi proposti.



Si noti come due dei rischi più rilevanti rappresentati (Figura 2), il primo “goals are not clear” ed il quarto “the output of the process does not satisfy the customer needs”, possano essere direttamente ricondotti alla prima fase, nella quale dovrebbe aver luogo la parte più

cospicua di progettazione e pianificazione del processo. Ciò dimostra come vada posta grande attenzione a questa fase, dato che dalla corretta conduzione della stessa potrebbe dipendere fortemente la buona riuscita di tutte le altre fasi.

Figura 3 Clustering dei rispondenti in base alle attività svolte all'interno del processo.



I punteggi assegnati alle attività sono stati accorpati secondo le fasi del modello (Tab. 1), per ridurre la complessità dell'analisi. I dati si sono mostrati poco strutturati avendo *Silhouette media* massima pari a 0.25, con $k = 3$. Questo (Figura 3, in alto a destra), che è stato ritenuto il più significativo, mostra tre cluster che si differenziano principalmente per, Cluster 1: *data modeling*; Cluster 2: *data retrieving* e *data preparation*; Cluster 3: *presentation and automation*. La principale considerazione riguarda come le fasi del processo ed i compiti delle figure professionali coinvolte siano elementi scarsamente definiti. Non di rado, infatti, la figura professionale del Data Scientist viene intesa come un esperto di tutti i domini coinvolti nel processo ed in grado di svolgere tutte le attività previste. In definitiva, i risultati mostrati evidenziano la necessità di una maggiore formalizzazione del processo, così da ridefinire ruoli e competenze in modo che vi sia una divisione chiara dei compiti all'interno del team.

4. Analisi degli asset

Avendo definito con maggiore chiarezza fasi e caratteristiche, sono stati posti come oggetto di analisi gli asset operativi. Il processo è stato messo in analogia con i processi di produzione industriale, individuando similitudini e differenze nella funzione e nel valore degli asset impiegati. Adoperando come chiave di lettura del confronto il passaggio da risorse fisiche a capitale intellettuale, dalla materia all'informazione, ed utilizzando come riferimento il noto strumento delle 4M, si è costruito un alias per il processo di Data Science, individuando sette categorie, delle quali le più significative sono riportate in Tabella 3:

Tabella 3 Asset operativi maggiormente rilevanti nel processo di Data Science.

Categoria	Descrizione
Dati	Costituiscono la materia prima del processo. Vengono trasformati per estrarre valore. Possono essere espressi in varie forme, quali immagini, video, tabelle, testo, eccetera.
Software	Sono sistemi, integrati sui dispositivi Hardware, che costituiscono l'interfaccia tra le macchine e le persone ed effettuano le trasformazioni necessarie sui dati. Vi sono vari livelli di Software (e.g. sistemi operativi, linguaggi di programmazione, singoli script di analisi).
Modelli	Costituiscono l'anello di congiunzione tra i metodi ed i Software ed attraverso essi vengono formalizzati gli algoritmi che definiranno come eseguire l'elaborazione delle informazioni contenute nei dati. Dalla scelta di questi deriva anche la definizione delle fasi e delle attività specifiche da svolgere nel processo.
Metodi	Sono impiegati dalla componente umana del processo per la progettazione concettuale e per la gestione del processo, (e.g. procedure e frame work), mantenendo la stessa funzione ricoperta nei processi di produzione industriali.

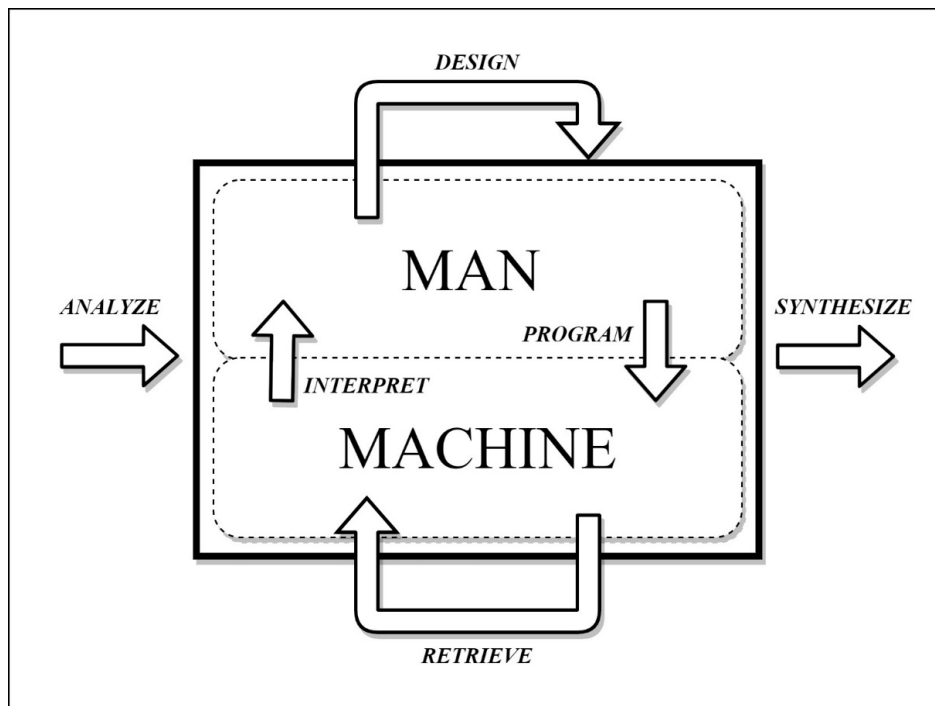
5. Un nuovo paradigma del processo di Data Science

In relazione alla definizione di Data Science fornita nel capitolo 1., viene proposto un passo avanti, passando da un'ottica Data-to-Knowledge, come quella precedentemente espressa, ad un'ottica Knowledge-to-Knowledge. Questa visione ha la caratteristica di essere di più ampio respiro rispetto alla precedente: i dati non vengono più considerati come il punto di

partenza, ponendo invece in input al processo, grazie alla lente gestionale, un'analisi della realtà esterna e aziendale dalla quale tutti gli altri passaggi saranno trainati.

Il processo è stato dunque rappresentato avendo come focus la sua caratteristica peculiare: un'intensa interazione uomo-macchina che, di passaggio in passaggio trasforma il dato, aggiungendo valore. Mappando questi passaggi, si è costruito lo schema rappresentato in Figura 4. In esso sono rappresentate sei interazioni principali, non necessariamente ordinate cronologicamente, descritte da frecce (accompagnate da verbi), e quattro elementi distinti, rappresentati da quadrilateri: persone, macchine, il processo stesso e l'ambiente esterno.

Figura 4 Rappresentazione del complesso di interazioni costituente il processo di Data Science, secondo la lettura proposta.



Vengono di seguito commentate le interazioni rappresentate dalle frecce:

- **Analyze**: Il team formula un'analisi della realtà ed individua e formalizza il problema che verrà affrontato nel processo;
- **Design**: Sulla base dell'analisi in input, le persone progettano il processo tenendo in considerazione anche fattori esterni (freccia esce dal quadrilatero del processo), definendo il complesso di interazioni che lo comporranno (punta della freccia cade sul quadrilatero del processo), in modo da ottenere l'output desiderato utilizzando al meglio le risorse disponibili;
- **Program**: le persone interagiscono con le macchine attraverso il codice, determinando i criteri secondo i quali esse cercano ed elaborano i dati. Il codice è da considerarsi una forma di dato computer-friendly;

- Retrieve: I dati, dai quali verrà estratto valore attraverso le dovute trasformazioni, devono essere reperiti dalle macchine per essere immagazzinati ed elaborati. Essi derivano dall'ambiente esterno (freccia esce dal quadrilatero del processo);
- Interpret: l'interazione macchina-uomo si configura i come l'interpretazione da parte delle persone dei dati di ritorno dalle macchine. Questi hanno già subito un primo processo di sintesi e sono quindi in formato human-friendly (e.g. grafici, immagini, video, tabelle);
- Synthesize: l'output del processo è un'interpretazione dei dati elaborati nelle analisi. Questi, in questa fase, sono ulteriormente sintetizzati per essere comunicati agli stakeholders esterni al processo.

In definitiva, l'obiettivo al quale un'organizzazione dovrebbe mirare sarebbe quello di definire un set di metodi e di procedure per regolare ognuno dei passaggi descritti.

6. Metodo di definizione del Project Scope

In questa sezione si presenta un esempio di metodologia che andrà a comporre il set relativo ad ognuna delle interazioni descritte precedentemente, incentrato sulla freccia relativa alla progettazione del processo, "Design". Il metodo è stato scelto seguendo un criterio data-driven, seguendo i risultati del questionario (Figura 2), in modo da agire sul rischio a maggior impatto sul processo: "goals are not clear", "gli obiettivi non sono chiari".

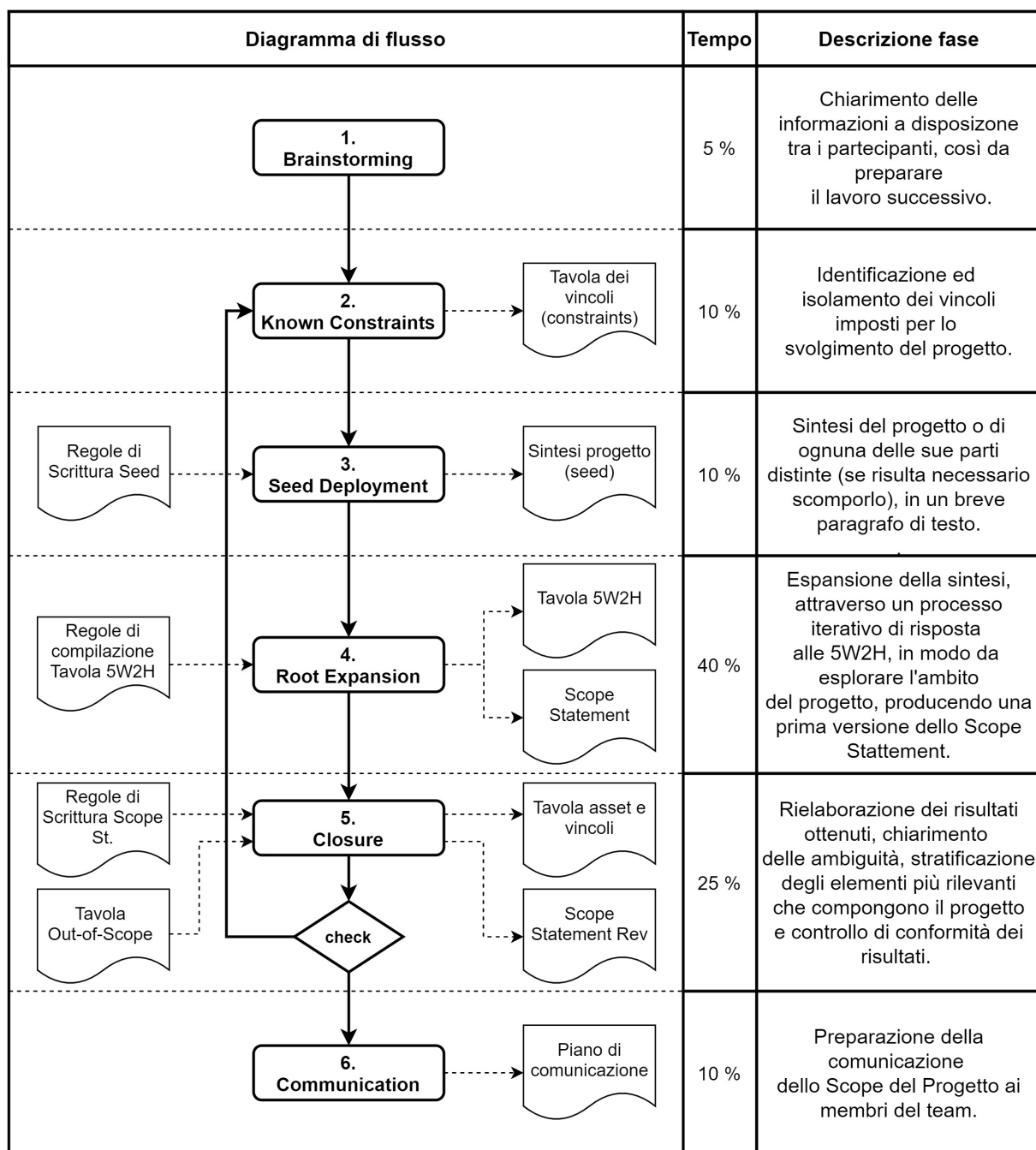
Per intervenire con efficacia su questo rischio è necessario agire prima che esso si verifichi, spendendo del tempo prima dell'inizio delle attività operative, in modo che:

- non vengano sprecate risorse nello svolgimento di attività non necessarie;
- il processo venga svolto in un ambiente sereno e favorevole alla cooperazione, avendo chiari gli obiettivi.

Si è deciso quindi di agire sul processo di definizione del Project Scope, con l'obiettivo di fornire una metodologia strutturata per la definizione dello Scope Statement nell'ambito della Data Science. È interessante sottolineare come la letteratura esistente posta sotto analisi (anche al di fuori del campo della Data Science) non presenti fonti che spieghino in maniera soddisfacente come questo processo debba essere eseguito.

Il metodo elaborato risulta suddiviso in sei fasi distinte e comprende un insieme di regole ed un insieme di documenti in output definiti. Tali fasi vengono descritte in Figura 5:

Figura 5 Metodo di elaborazione dello Scope Statement nell'ambito della Data Science.

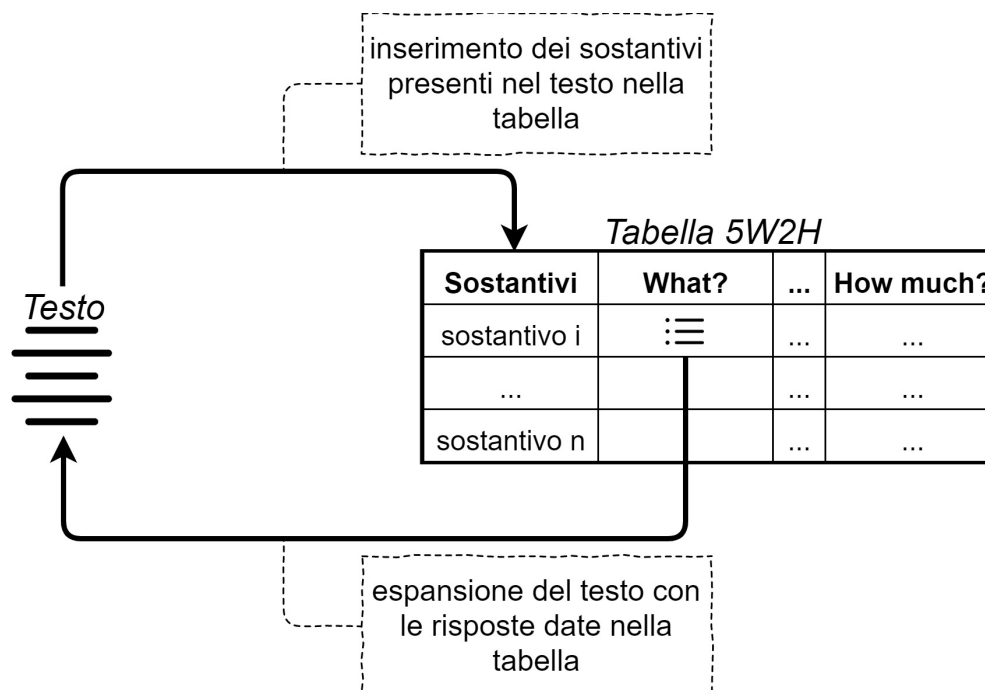


Il cuore del procedimento mostrato in Figura 5 risiede nella fase 4. che è stata pensata con lo scopo di favorire un'esplorazione sistematica dell'ambito del progetto ed ha come output una prima versione dello Scope Statement. Consiste, infatti, nell'espansione progressiva di un breve testo che sintetizza l'ambito del progetto (preparato nella fase 3.), attraverso un processo iterativo di analisi per mezzo delle 5W2H.

Il procedimento seguito è il seguente: vengono estratti i sostantivi presenti nel testo per essere poi inseriti in un'apposita tabella ed analizzati singolarmente rispondendo alle domande relative alle 5W2H. Si genera così, eventualmente, altro testo da aggiungere

all'elaborato. Ciò permette di alimentare il ciclo che viene ripetuto fino a quando necessario. Questo procedimento è schematizzato nella Figura 6:

Figura 6 Ciclo di espansione del testo dello Scope Statement attraverso l'analisi dei sostantivi per mezzo delle 5W2H.



In tal modo, dopo un certo numero di iterazioni, partendo dal breve testo preparato in fase 3., si arriva ad ottenere un testo più ampio costituente lo Scope Statement.

Gli altri output rilevanti del metodo hanno lo scopo di evidenziare gli elementi più importanti di un progetto di Data Science, anche in riferimento a quanto espresso nel paragrafo 4.

7. Considerazioni finali

Il lavoro svolto ha avuto come principale risultato quello di spostare la prospettiva da una visione strettamente tecnico-informativa del processo ad una visione ingegneristica, volta al Management ed al Design. Ciò è stato ottenuto sia attraverso un approccio Bottom-Up con il questionario, sia con un approccio concettuale, Top-Down. La fase di studio del processo era indispensabile ed è stata condotta in maniera approfondita, occupando così gran parte del tempo disponibile. Un corretto ampliamento del lavoro sarebbe stato quello di portare avanti in parallelo all'approccio concettuale, delle prove pratiche per i metodi, come dei Case Study, unico modo per stabilirne la reale validità.

Gli sviluppi futuri sul tema, dovranno proseguire nella direzione proposta, incentrandosi sulla formalizzazione del processo e sull'elaborazione di metodologie, procedure e prassi precise per ognuna delle fasi e le attività svolte. Queste dovranno inoltre essere testate in contesti reali, unico modo per misurarne gli effetti ed eventualmente ridisegnarle.