85% OF DATA SCIENCE PROJECTS FAIL[1]

… and here is why

- Today, it is evident that to capture the full value from data and analytics, Companies have to deeply transform their organizations, but a complete transformation can take years
- In the meanwhile Analytics transformation take place, Corporates can get immediate benefits by managing Data Science as a discipline, or in other words, by standardize their approach of the entire Data Science project lifecycle

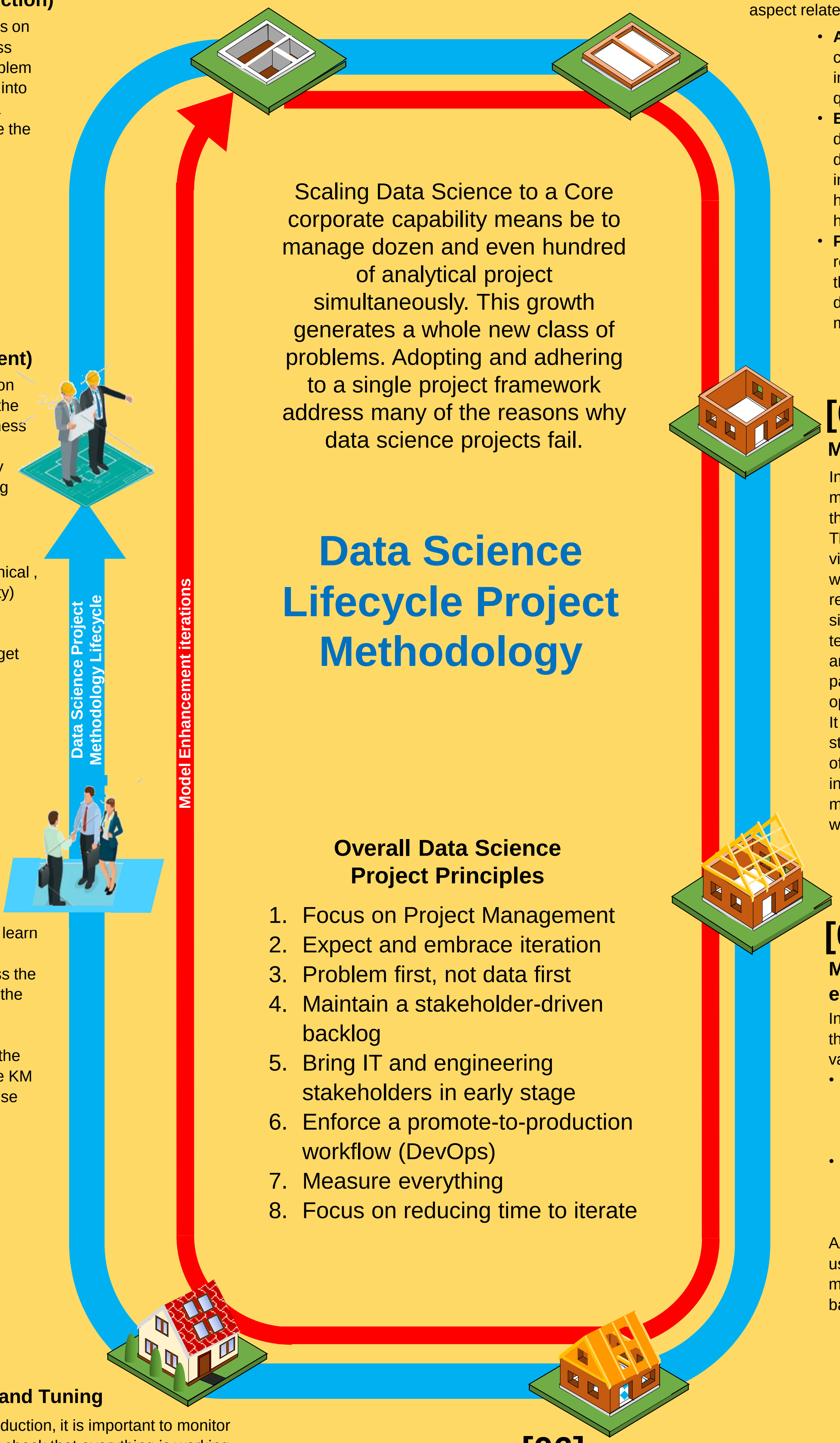
Source : [1] https://twitter.com/nheudecker/status/928720268662530048

STANDARDIZE DATA SCIENCE PROJECTS

is key to the success of an Analytics organization

- One of the key aspects of a success full analytics transformation is the ability to select, manage, scale, and accelerate data science projects and manage them in a coherent portfolio.
- This means that any analytics organization must to develop and adopt a data science lifecycle project methodology.
- Only organization that are able to develop a disciplined data science practice can scale data science to a core Corporate capability

Project framework for data science projects


Data Science Project Methodology Lifecycle
Model Enhancement iterations

Scaling Data Science to a Core corporate capability means be to manage dozen and even hundred of analytical project simultaneously. This growth generates a whole new class of problems. Adopting and adhering to a single project framework address many of the reasons why data science projects fail.

Data Science Lifecycle Project Methodology

Overall Data Science Project Principles

1. Focus on Project Management
2. Expect and embrace iteration
3. Problem first, not data first
4. Maintain a stakeholder-driven backlog
5. Bring IT and engineering stakeholders in early stage
6. Enforce a promote-to-production workflow (DevOps)
7. Measure everything
8. Focus on reducing time to iterate

[01]

Project Selection (Portfolio Management)

This initial phase focuses on understanding if and how the project fits within the business strategy. If done well, the ideation stage dramatically de-risks a project by driving alignment across stakeholders.

Main outputs:

- Feasibility studies (technical , economics and capability)
- Business case review
- Project prioritization
- Initial resource and budget allocation

[02]

Problem Definition (Business Requirements Collection)

This second phase focuses on understanding the Business objectives and needs (problem definition), translate these into requirements and create a preliminary plan to achieve the objectives.

[03]

Data Acquisition, Exploration and Preparation

This phase focuses covers all the aspect related to data understanding.

- **Acquisition**: data collection, data investigation, identify data quality problems
- **Exploration**: data analysis, discover insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information
- **Preparation**: all activities related to the construction the final dataset, that is the data that will be fed into the modeling tool.

[04]

Model development

In this phase, one or more model are built to address the business problem.
The model creation is done via several iterations in which the model(s) is refined and enhanced (start simple). Various modeling techniques can be selected and applied, and their parameters are calibrated to optimal values.
It is important to bring stakeholder in this phase, as often they have solid intuition about what features matter and in what direction work.

[05]

Model testing and evaluation

In this phase, is tested both the correctness and the value of the model

- **Correctness**, tests if the model is corrected from technical and analytical point of view
- **Evaluation** verifies with the business users, if the model properly achieves the business objectives.

A/B tests or similar are used to check if the new model performs better that baseline.

[06]

Model deployment in production

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It means applying the models within an organization's decision-making processes.

[07]

Model Monitoring and Tuning

After the model is in production, it is important to monitor the model to proactively check that everything is working properly.
In this phase, the model params are tuned using real data to optimize the model outputs value

[08]

Knowledge management

In this phase, all lessons learn (technical, statistical and project) are shared across the team and are inserted in the knowledge management systems.
Datasets created during the project are inserted in the KM system tool for future reuse

www.andreafabrizi.org