UNCATEGORIZED

Why Big Data Science & Data Analytics Projects Fail

POSTED FEBRUARY 13, 2021 ADMIN

The Boston Dynamics humanoid robot tripping over a curtain and tumbling off stage? A prediction model that tips a teen's parents off to her pregnancy? A music recommendation engine that suggests Coldplay?

Yup. There are just too many big data, data science, and data analytics failure examples to cover in just one post. Indeed, the data science failure rates are sobering:

- 85% of big data projects fail (Gartner, 2017)
- 87% of data science projects never make it to production (VentureBeat, 2019)
- "Through 2022, only 20% of analytic insights will deliver business outcomes" (Gartner, 2019)

A one minute break for robot fails
This begs the question: Why do big data science and analytics projects fail?

You might think it has to do with data and processing. You're not wrong. These are certainly challenges. But there are bigger problems. So let's explore some big data failure examples and dive into what drives these failures.

8 Reasons Why Big Data Science and Analytics Projects Fail

1. Not having the Right Data

I'll start with the most obvious one.

Without data, you don't have a data science project. Yet, this data can be challenging to collect, create, or purchase. Even if you can get access to the data, you still have to overcome what seems like a mountain of issues such as:

- How do you secure the data?
- Is the underlying data biased?
- Can you ethically and legally use the data for your intended use case?
- Can you process the data in a timely and cost-appropriate manner?
- Is the data clean? (probably not in which case…) Can you clean the data?
- Do you know whether the data drifts over time?

With all these challenges (and a lot more), it's no surprise that a 2020 International Data Corporation survey finds that "Lack of adequate volumes and quality of training data remains a significant development challenge."

Data management deserves a post (or book) on its own but a few quick pointers:

- Have internal protocols including policies, checklists, and reviews to enforce proper data usage.
- Never assume data is clean. Assume it is dirty unless proven otherwise.
- Build production-grade cloud-based systems for data pipelines which include pro-active alerts and notifications to let you know if something looks off.
- Invest in data and cloud engineers to build these systems (which lead us to the next point…).

2. Not having the Right Talent

Finding, hiring, and retaining top tech talent is never easy. And the competition for qualified data talent is especially fierce. In fact, in 2020, QuantHub rated "data science/analytics" as the second most difficult skill set to find (after "cybersecurity").

And it's not just about finding qualified quant talent. You also have to find the right *combination* of qualified talent. Long gone are the days whereby the lone wolf data scientist executes a complete data science project on his own. Rather, full data science products span from idea incubation, deployment, and into machine learning operations. You'll need a wide range of roles to fully execute all these steps.

3. Solving the Wrong Problem

Ever started a project with a fuzzy idea of the goal? Or alternatively, a project with a clearly defined goal that is not realistic or that does not add any meaningful value?

You're not alone. We've all been there.

Indeed, Domino Data Labs reports that "We've seen large organizations hire 30+ PhDs without clear business alignment upfront. They then emerge from a six week research hole only to realize they had misunderstood the target variable, rendering the analysis irrelevant."

Oops!

To mitigate this risk, start your project on the right foot and ask the right questions before starting a project. Don't just take any request at face value. Rather, dive deeper to truly understand the underlying problem that needs to be solved.

4. Not Deploying Value

According to a 2020 Forbes article, only 15% of leading firms have deployed AI capabilities into production.

Why not more?

Enter the deployment gap. There are often technical, skillset, and motivational gaps between those who develop the models and those who maintain those models. This is often drawn out in comics as a data scientist blindly tossing a model over a high brick wall to some poor IT system engineer who is waiting to catch something he's never seen before with a fishing net.

You can do better though. Tear down that wall! Get rid of the fishing net. Have those responsible for the IT operations and those developing the models work together throughout the project life cycle so that deployment can be properly planned and executed. Or go even further and focus your team with a Machine Learning Operations mindset (learn more at https://ml-ops.org/).

5. Thinking Deployment is the Last Step

By the traditional definition of a project, a project ends whenever the scope is delivered. For data science, this is often the deployment phase.

However, your work doesn't end there. If you think it does, imagine…

- Not adjusting to the inevitable data and market changes. For example, how accurate would your flight demand prediction model during COVID times if you trained the model using data from 2019?
- Not knowing that your data feeds stop sending data or start sending corrupted data. Garbage in = Garbage out.
- Stakeholders losing interest or end-users not adopting your solution. That's frustrating (yet often preventable).
- Not making critical systems updates to maintain system uptime and proper security. Data hacks and data system failures crop up in the news nearly daily.

Some of these issues you can plan for. Others you cannot.

So plan for the unplanned. Ensure that you have the proper staffing and focus to allow your models to continue to add value beyond the initial deployment. Proactive planning and self-healing systems can mitigate these above scenarios.

6. Applying the Wrong (or No) Process

Without established and clear methodologies for data science project management, organizations often resort to ad hoc project management processes which can lead to inefficient information sharing, missed steps, and misinformed analyses.

Alternatively, other organizations try to apply approaches that other fields use. The most common mistake is to treat data science as another software function (they're both code, right?) which alienates data science from what it is.

Rather the best approaches combine the data science life cycle with an agile collaboration framework.

7. Forgetting Ethics

Models ruthlessly optimize what you tell them to. Done correctly this is a good thing.

However, it's a double-edged sword that can lead to serious ethical, market branding, and legal issues. Sure nefarious actors can intentionally create such issues. But often these issues rise by lack of oversight or by accident. Just consider these…

Data Science Failure Examples in Ethics

- **Racist Health Risk Scoring:** Healthcare providers used a health risk score to help determine whether they should offer proactive healthcare treatment to each patient. Good idea, right? However, the model used healthcare costs as a proxy for health risks. And because black patients tended to have lower health care costs, this health risk score inadvertently prioritized white patients for proactive treatment. (sciencemag.org, 2019)
- **Target Predicts Teen Pregnancy:** A Target advanced analytics team was tasked to predict whether a woman was pregnant so that they could offer targeted ads. They were successful in this prediction — very successful. But they ignored the wider privacy implementations which resulted in public backlash due to the "creepiness" factor. (forbes.com, 2012)
- **Cambridge Analytica Scandal:** London-based Cambridge Analytica goes down as one of the most notorious companies in violating privacy and data misuse. It took data from millions of Facebook profiles without user acknowledgment to create psychological profiles of voters. The result — public uproar, Facebook was fined $5 billion, and Cambridge Analytica went bankrupt. (nytimes.com, 2018)

As such it is critical to uncover potential ethical issues upfront and throughout the product life cycle. For starters ask yourself these 10 ethics questions.

8. Overlooking Culture

Culture still eats strategy for breakfast.

And recent surveys agree. In fact, for five straight years, the NewVantage Partner's 2021 survey finds that "executives report that cultural challenges – not technology challenges – represent the biggest impediment to successful adoption of data initiatives and biggest barrier to realizing business outcomes."

Remember that not everyone is on board with your data initiatives. Meet those who are uncomfortable where they are. Start by educating your staff, emphasize the importance of data-driven decision-making, and help everyone involved work through the change management process.

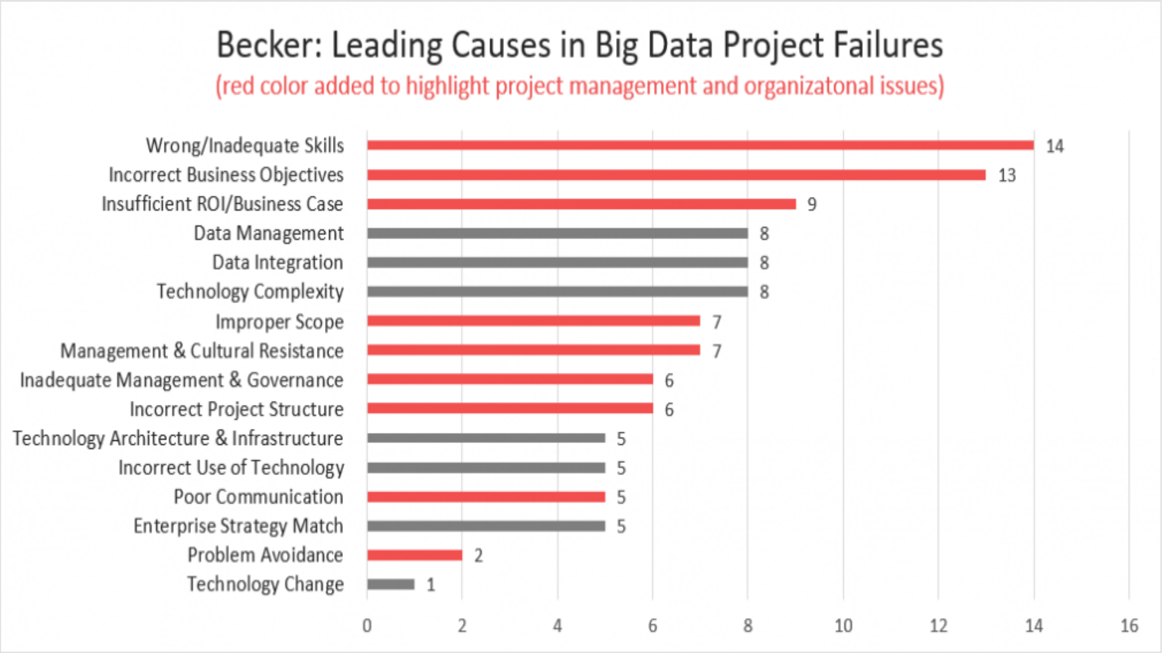
This change management process is often bigger than the technical changes.

Are these Issues New?

Not really. There's a wealth of documented analytics failures since before "data science" was coined as a term. We'll look at just two big data studies.

2017 Big Data Project Failure Study

David Becker clustered commentaries on big data project failures in a 2017 research paper. I further categorized these into technology-driven failures (in gray) and project management and organizational issue-driven failures (in red). 62% of the failures were due to these latter issues, not technical issues.



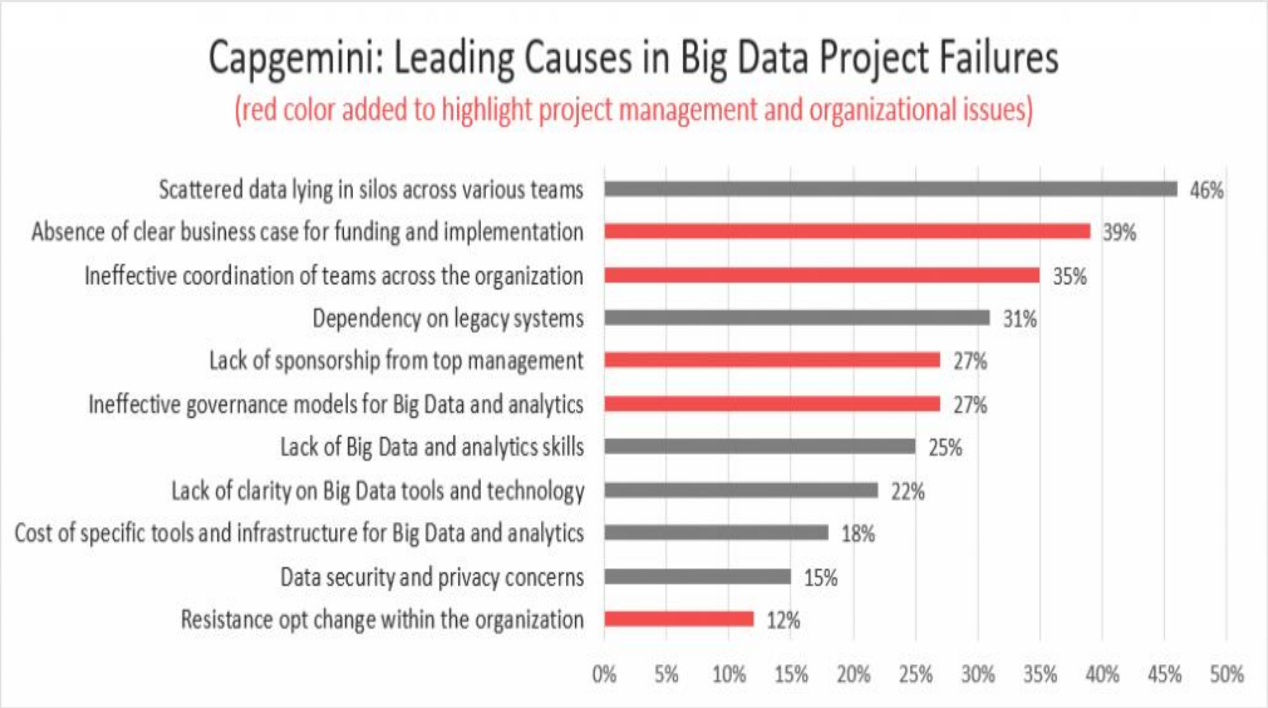
Becker (2017) based on clustering commentaries of 19 industry experts. I added the red/gray classification.

2014 Big Data Failure Study

Likewise, a 2014 Capgemini study found low success rates:

- "Only 27% of big data projects are regarded as successful"
- "Only 13% of organizations have achieved full-scale production for their Big Data implementations"
- "Only 8% of the big data projects are regarded as VERY successful"

Moreover, the main challenges were a mix of technical and project and organizational challenges that hinder success.



Capgemini (2014) based on survey of 226 respondents. I added the red/gray classification.

Concluding Advice

I can't personally verify whether data science projects indeed fail 85% of the time. However, given data science's experimental nature and unique challenges, I wouldn't be surprised.

Of course, you can't deliver projects successfully 100% of the time. And if somehow you do, you're probably very lucky, have a low-definition of success, or just have worked on one or two fortunate projects.

Regardless, identify and plan for the potential failure points, apply some of these tips, and perhaps engage in data science project management training. Then, you'd be more likely to succeed.

Learn More

Data Science Process Alliance: Given the requests we've had for training, Jeff and I have helped launch the DSPA which can help you learn how to better deliver projects.

- Individual training courses
- Enterprise consulting services

Other Posts: This is just one of several practical posts. Jump in and learn more through these topics:

- Data Science Teams
- Data Science Workflows
- Coordination Frameworks
- Agile Data Science
- Traditional Data Science Approaches
- Hybrid Data Science Approaches
- Emerging Data Science Approaches

Managing Machine Learning Projects

What is a Data Science Life Cycle?⟶