Assignment 05

Reflection

Greg Sullivan

Washington University in St. Louis
Spring 2026.CAPS-CS.5035.01

Data Engineering Foundations
Professor Paul Boal
February 15, 2026

Reflections on the Importance of Data

As many organizations commence or continue a journey toward isolated experiments or enterprise-wide adoption of Artificial Intelligence (AI), we see some enjoy success while others seem to fail. While AI adoption is widespread, only a subset of organizations report measurable bottom-line impact (McKinsey & Company, 2023). Deloitte (2022) similarly reports that many organizations remain in pilot phases rather than scaling AI across core operations. As many rush to adoption, some seem to plan indefinitely and still others pay close attention to successes and failures of those who embarked on the journey before them. In this paper I reflect on some of the key reasons for successes and failures from personal experience, published research, surveys and expert opinions.

I'd like to start by exploring the success (and/or failure) rate of AI adoption inside various organizations. In the early stages of my career, I had the good fortune to support innovative individuals who fully understood the value of data in business. While there were many, I recall one person in particular, Mr. Grant Williams who had a mail order catalog company called Knights, Ltd. These were the early days of specialized mail order catalogs with Mr. Williams' catalog focusing almost entirely on the sale of women's shoes. I remember in one conversation he wanted to know if the red high-heel sales changed based on the position in each page where they appeared.

He taught me two important lessons a long, long time ago (think mid-1980's):

1. For his business to be the most successful, he needed a variety of data analytics performed to answer critical and insightful questions (for which there was no end to his creative thinking); and
2. No result was worth having if we couldn't trust the data upon which any analyses were conducted.

It is this second point that has resonated with me throughout my career and remains at the forefront of all my work to this day. In my current work I advise organizations on AI strategies, in which I always emphasize the importance of data integrity, data governance, data management, data provenance, metadata and data security (among others). In fact, in a recently published industry interview (Williams, 2026), I argued that most AI-related failures result from poorly governed data pipelines rather than the models themselves — a position that directly informs how I think about the importance of data engineering in AI initiatives.

The central problem I am exploring is this: If organizations hope to achieve meaningful success with AI adoption, what data must exist to support that effort, and how accessible is that data in practice? While many discussions focus on model performance and the data that supports the desired analytics, I argue that the more critical question is whether the underlying data — and the governance structures surrounding it — are sufficiently trustworthy and well managed to sustain AI at scale. Boston Consulting Group (2023) describes this disconnect between experimentation and value realization as the "AI impact gap," a gap I argue is often rooted in data governance immaturity.

The central question, then, is not simply whether AI adoption succeeds or fails, but what data (about the data) truly determines that outcome. Beyond model accuracy and computational power, what information must an organization possess about its own data — including governance, lineage, ownership, and integrity — to deploy AI responsibly and effectively? And how accessible is that information when organizations attempt to move from experimentation to enterprise scale?

What data exists to help us on our AI journey?

Organizations exist at all levels on the data maturity spectrum. I argue that there is little to no chance of discernible success in AI adoption without a high level of data maturity in the organization. Furthermore, data maturity is not something that can be added "later", after attempting to adopt AI. It is critical to get it right BEFORE beginning your AI adoption effort. That said, what data does exist to measurably support an AI implementation? Can we accurately track and trust this data? Can we set quantifiable goals and document policies? Do we have a formal exception governance and oversight process in place for when we must deviate from policy? These are just a few of the questions we should consider as we commence our enterprise AI deployment.

First, let's consider a list of data that might possibly serve our purpose: to ensure the best possible chance for success in our adoption of AI. To start with a broad list, I see 30+ different data sets that could be useful in tracking our success with AI adoption. I break these into five categories, listed in my suggested order of importance:

A. Data Governance & Integrity Data
These assess whether foundational data is reliable and controlled.

1. **Data lineage documentation** – Tracks where data originates, how it moves, and how it transforms; critical for explainability and error tracing.
2. **Data ownership registry** – Identifies accountable individuals for data domains; prevents ambiguity in stewardship.
3. **Data classification schema** – Defines sensitivity levels (public, confidential, restricted); essential for AI compliance.
4. **Data quality scorecards** – Quantifies completeness, accuracy, timeliness; indicates whether AI inputs are trustworthy.
5. **Metadata catalogs** – Documents meaning, structure, and usage of data assets; critical for scalability.
6. **Reference data dictionaries** – Standard definitions for shared data elements (e.g., "customer"); prevents semantic drift.
7. **Database access logs** – Records who accessed or modified data; supports accountability.
8. **Policy documentation (AI governance policy)** – Formal rules governing AI and data use.
9. **Audit trail logs** – Tracks changes to data over time; essential for forensic review.
10. **Data retention schedules** – Ensures data lifecycle management and regulatory compliance.
11. **Regulatory compliance reports** – Demonstrates adherence to legal frameworks.

B. Risk & Security Data
These assess vulnerability exposure and resilience.

12. **Security incident reports** – Past breaches or misuse involving data or AI.
13. **Vulnerability scan results** – Technical weaknesses in systems handling AI data.
14. **Third-party vendor risk assessments** – Evaluates external AI/data suppliers.
15. **Privacy impact assessments** – Assesses privacy risks from AI deployment.
16. **Third-party cybersecurity maturity assessment** – Evaluates external control effectiveness.
17. **Obligatory IT compliance assessments** – SOC, ISO, or similar compliance documentation.

C. Business Impact & Outcome Data
These measure whether AI delivers value.

18. **ROI measurements** – Financial return on AI investment.
19. **Productivity metrics pre/post AI** – Efficiency gains.
20. **Revenue impact attributed to AI** – Direct business growth.
21. **Customer satisfaction changes** – External perception impact.
22. **Quality improvements** – Error reduction or performance gains.
23. **Incident reports linked to AI errors** – Business consequences of AI failures.

D. Organizational & Cultural Data
These assess readiness and alignment. Ransbotham et al. (2020) found that organizations combining AI adoption with organizational learning are significantly more likely to generate value.

24. **AI strategy documentation** – Formal roadmap and intent.
25. **Change management plans** – Adoption support mechanisms.
26. **Employee AI training records** – Capability readiness.
27. **Survey results on AI trust** – Cultural acceptance.
28. **Executive steering committee minutes** – Oversight and governance activity.

E. AI Model & Technical Performance Data
These evaluate operational effectiveness.

29. **AI usage patterns** – Adoption levels.
30. **Model accuracy metrics** – Technical performance.
31. **Drift detection reports** – Performance degradation over time.
32. **Training dataset composition** – Bias and representativeness.
33. **Inference latency statistics** – System responsiveness.
34. **Model version history** – Governance of model evolution.
35. **Prompt library documentation** – Standardization of generative AI inputs.

Using this list of 35 potential data items, to each I'll assign the following:

- Priority (A/B/C), with A being the highest priority. Basically, this explains how important it is for this data item to exist with expectations of high quality and trustworthiness.
- Accessibility (1–4), with 1 being the least accessible. I assign this based on my personal experience in commercial organizations mostly subject to the compliance obligations of a US public company. As I venture further into the AI metadata, I do speculate a bit based on where most of the organizations I work with seem to exist these days relative to their AI maturity.

Following is that table along with brief explanations on why each data item important and why it is (or why it is not) accessible.

Data item	Importance (A/B/C)	Accessibility (1–4)	Why it's important	Why it's accessible (or not)
Data lineage documentation	A	2	We can't trust data from unknown sources	Something we should know, but often lack clarity on
Data Ownership	A	3	Should always be crystal clear	…and clearly documented
Data classification schema	A	2	Well-documented data classification categories are mandatory	Do we properly classify all data?
Data quality scorecards	A	1	Again, it's all about trust	Rarely recorded in most orgs
Metadata catalogs	A	2	Continuous monitoring is critical	If it does not exist, it can be created — but doing so requires organizational discipline.
Reference data dictionaries	B	3	We need a common "language"	If we don't have, easy to create
Database access logs	B	4	Various compliance requirements	Internal and external auditors
Policy documentation	A	2	A sign of data management maturity	Should exist; if not, they must be created and maintained
Audit trail logs	B	4	Data integrity and various compliance requirements	Easy to automate and should be spot-checked by humans
Data retention schedules	C	3	Cost of storage and compliance requirements	Should be well-maintained
Regulatory compliance reports	B	3	Internal testing and third-party validations are required	Auditor work product
Security incident reports	C	3	What went wrong and how we assure it doesn't happen again	Mandatory creation for each security incident
Vulnerability scan results	B	4	Knowing our risk level	Understanding our risk tolerance
Third-party vendor risk assessments	B	1	Sometimes the weakest link	Often the last frontier for many
Third-party cybersecurity maturity assessment	B	1	Are our vendors as secure as we are?	Often an afterthought
Privacy impact assessments	A	3	Compliance requirements	Internal or third-party developed
Obligatory IT compliance assessments	A	4	Compliance requirements	Internal or third-party developed
ROI measurements	A	1	Business goals	Financial metrics are normally easy to obtain
Productivity metrics pre/post AI	B	1	Is AI helping (or not)?	Challenging to measure
Revenue impact attributed to AI	B	1	Is our investment worthwhile?	Very challenging to measure
Customer satisfaction changes	B	3	Does it matter to our customers?	Very challenging to measure
Quality improvements	C	2	Are we getting better?	Metrics should exist
Incident reports linked to AI errors	B	1	How we know if something went wrong	We must take the time to document these incidents
AI strategy documentation	B	2	Provides baseline targets	Sometimes an afterthought
Change management plans	A	2	Sign of mature processes	Should exist, but don't always
Employee AI training records	C	4	Valuable, if tracked	Typically maintained by HR systems, though not always tied directly to AI effectiveness
Survey results on AI trust	C	2	Nice to have	Even when exists, often ignored
Executive steering committee minutes	C	4	Important for policy lineage	Should always exist
AI usage patterns	B	3	Track adoption and over time	Should be in our AI tools
Model accuracy metrics	B	2	Accuracy matters on trust	Sometimes arbitrary
Drift detection reports	B	1	More important over time	Sometimes arbitrary
Training dataset composition	C	2	Training alignment matters	Working with our training providers
Inference latency statistics	C	1	Not mission critical	AI tool logs
Model version history	B	3	At times, it could matter	AI tool tracking
Prompt library documentation	C	1	A luxury, but sign of maturity	Someone must do the work

Given I have a long list of potential data items – and not to mention a lot of work to gather and maintain those data items – it makes sense for us to find a convenient way to prioritize which of these data items will have the biggest impact and are most readily accessible. The chosen approach is to create a 2x2 grid with the rows providing separation by importance (those items rated A are considered of critical importance and those rated B or C are probably not needed). The columns are organized to convey how accessible the data item is in any given situation. In this case I say a rating of 1 or 2 are considered difficult to access and a rating of 3 or 4 are known to be readily accessible.

	Difficult to Access (1 & 2)	**Readily Accessible (3 & 4)**
Critical Importance (A)	Data Lineage Documentation Data Classification Schema Data Quality Scorecards Metadata Catalogs Policy Documentation Change Management Plans ROI Measurements	Privacy Impact Assessments Obligatory IT Compliance Assessments Data Ownership
Probably Not Needed (B & C)	Third-Party Vendor Risk Assessments Third-Party Cyber Maturity Assessments Productivity Metrics Pre/Post AI Revenue Impact Attributed to AI Incident Reports Linked to AI Errors AI Strategy Documentation Drift Detection Reports Training Dataset Composition Inference Latency Statistics Prompt Library Documentation	Reference Data Dictionaries Database Access Logs Audit Trail Logs Data Retention Schedules Regulatory Compliance Reports Security Incident Reports Vulnerability Scan Results Customer Satisfaction Changes Quality Improvements Employee AI Training Records Survey Results on AI Trust Executive Steering Committee Minutes AI Usage Patterns Model Accuracy Metrics Model Version History

Reflection and Conclusion

As Brynjolfsson and McAfee (2017) argue, digital transformation requires complementary organizational change. My review of AI-related data assets reinforces this principle: technology alone does not create value – disciplined data stewardship does. The data most critical to AI success is not model performance data, but governance and integrity data. In my view, nearly all foundational governance artifacts (lineage, ownership, quality, policy) fall into the **Critical + Difficult to Access** quadrant. Many reports (and personal observation) reveal organizations often rush toward AI tools while lacking visibility into the state of their own data ecosystem.

The 2x2 grid made visible a pattern I have observed professionally but never formally mapped. The exercise revealed a structural tension: the data required to justify and sustain AI adoption is often the hardest to assemble or measure. Many organizations may believe they are "AI ready" because technical metrics are available, while deeper governance gaps remain unexamined. Accessibility does not equal importance. Conversely, difficulty of access often signals organizational immaturity rather than irrelevance.

AI readiness is less about model sophistication and more about disciplined data stewardship. Data maturity cannot be retrofitted after AI deployment; it must precede it. Governance documentation, ownership clarity, and quality measurement represent institutional discipline – not technical complexity. Sustainable AI adoption is fundamentally an organizational problem before it is a technological one.

The exercise reinforced a lesson learned early in my career: results are meaningless without trustworthy data. My professional experience aligns with the findings of industry research – governance gaps, not model limitations, often explain failure.

AI does not fail because organizations lack algorithms; it fails because organizations lack disciplined data foundations. The real question is not whether we can build AI, but whether we have earned the right to trust its outputs. In that sense, AI adoption is less a technological leap and more a mirror reflecting the maturity of an organization's data culture.

References

Williams, K. (2026, January 8). *Cybersecurity in 2026: Experts predict what's next*. SmarterMSP. https://smartermsp.com/cybersecurity-in-2026-experts-predict-whats-next/

McKinsey & Company. (2023). *The state of AI in 2023: Generative AI's breakout year*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Deloitte. (2022). *State of AI in the enterprise (5th ed.)*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-in-the-enterprise.html

Boston Consulting Group. (2023). *From potential to profit: Closing the AI impact gap*. https://www.bcg.com/publications/2023/closing-the-ai-impact-gap

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2020). *Expanding AI's impact with organizational learning*. MIT Sloan Management Review and Boston Consulting Group. https://sloanreview.mit.edu/projects/expanding-ais-impact-with-organizational-learning/

Brynjolfsson, E., & McAfee, A. (2017). *Machine, platform, crowd: Harnessing our digital future*. W. W. Norton & Company.