

The Growing Symbiosis between Data Governance and AI



Hello everyone! Thank you for inviting me to present at the New York City chapter of DAMA. I'm excited to be here with you today.

This slide shows two approaches to portraying the Fearless Girl statue with the Wall Street Bull. The first one is a real picture taken by a human being, and the second is AI-generated.

This illustrates some of the creativity that AI can unleash; I certainly could never create the second image here on my own.

But it also shows the potential for "deepfakes" and reality distortion. If someone had never seen the actual statues, they might come away convinced that the AI-generated image is the reality.

Your Speaker Today



Dan Power, Senior Data Leader

Data Strategy | Data Governance | Data Quality | Data Management | Analytics

Some background on your speaker today:

I'm Dan Power. Formerly, I was a Managing Director & Head of Data Governance at State Street Global Markets. I've just started a new position this week as Head of Data Governance at Raymond James.

By the way, please feel free to ask questions as we go; no need to hold your questions to the end. I'm fine with you just putting your hand up, and I'll answer your questions to the best of my ability.

To introduce today's topics: in Part 1, we're going to discuss how Data Governance is an essential foundation of – and driver for – successful & responsible AI, and then in Part 2, we'll discuss what AI can do to enhance and optimize Data Governance.

I've spent my whole career in the "data world" – 24 years as a consultant, and 12 years in corporate America. I've worked in various industries, but mainly financial services, high tech, and health care.

I helped launch the Data Governance program at State Street in 2017. Before that, I ran a boutique consulting firm called Hub Designs from 2007 to 2017, focusing on data strategy, data governance, data quality, master data management, and analytics.

I've been an employee at companies like Dun & Bradstreet, CSC Consulting, and Deloitte, and have had consulting clients such as Apple, AIG, LPL Financial, Red Hat, and EMC.

I'll be the first to admit that I'm not an expert on AI, but as a longstanding data governance person, who has sat in both the business and in IT, I do have thoughts on where AI is headed.

Part 1: Data Governance as a Foundation for AI

- Introduction to Part 1
- About Data Governance
- The Crucial Connection: Data Governance and AI
- Enhancing Collaboration and Communication
- Timeframe and Geographic Impact
- Examples / Use Cases
- Future Trends and Emerging Technologies
- Summary of Part 1



So, how can Data Governance help with the challenges of doing AI successfully in a corporate setting?

We're not going to talk much today about public AI offerings like OpenAI's ChatGPT and Google's Bard, as interesting as they are.

They're trained on broad & deep datasets from a wide variety of sources on the Internet. This has led to some copyright, privacy and intellectual property questions. The New York Times is suing OpenAI and Microsoft for copyright infringement, for using millions of their articles to train ChatGPT.

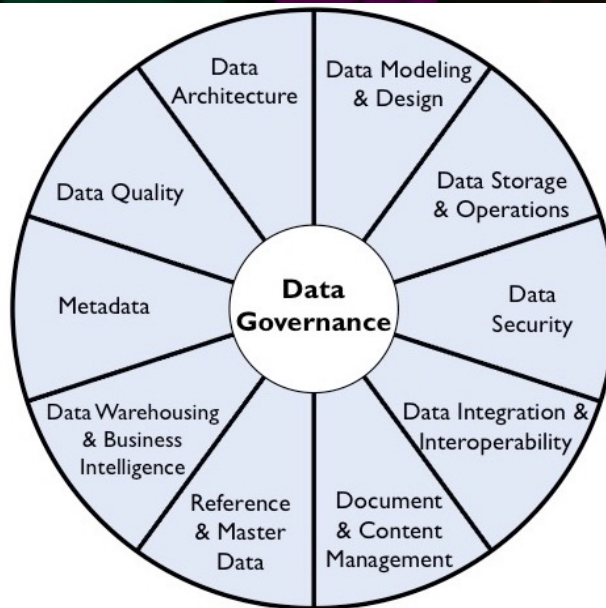
There's also a class-action lawsuit by a California law firm alleging that OpenAI violated the copyrights and privacy of millions of users by scraping data from the Internet to train its models. And there's another class action lawsuit from the Authors Guild on behalf of dozens of best-selling writers.

In a corporate setting, of course, the copyright question shouldn't be too much of an issue. But there are other challenges to doing AI successfully and responsibly, that the discipline of Data Governance can help with, and a consensus seems to be emerging that doing AI Governance well requires doing Data Governance well as a starting point.

So, the first part of today's discussion will be about Data Governance as the underlying foundation and accelerator for successful and responsible AI. The second part will then be about how AI can improve, automate, and optimize Data Governance.

Think of it as the "yin & yang" of Data Governance and AI. Why does AI need Data Governance? And why does Data Governance need AI?

What Is Data Governance?



Copyright© 2017 DAMA International

What Data Governance Is, What It Isn't

Let's make sure we all understand what I mean by "Data Governance".

I'm sure you're all familiar with the "DAMA wheel". According to DAMA International,

"Data Governance is the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets. Data governance is the discipline of cataloging and defining important data, assigning ownership of data, and incorporating governance of data into everyday business processes."

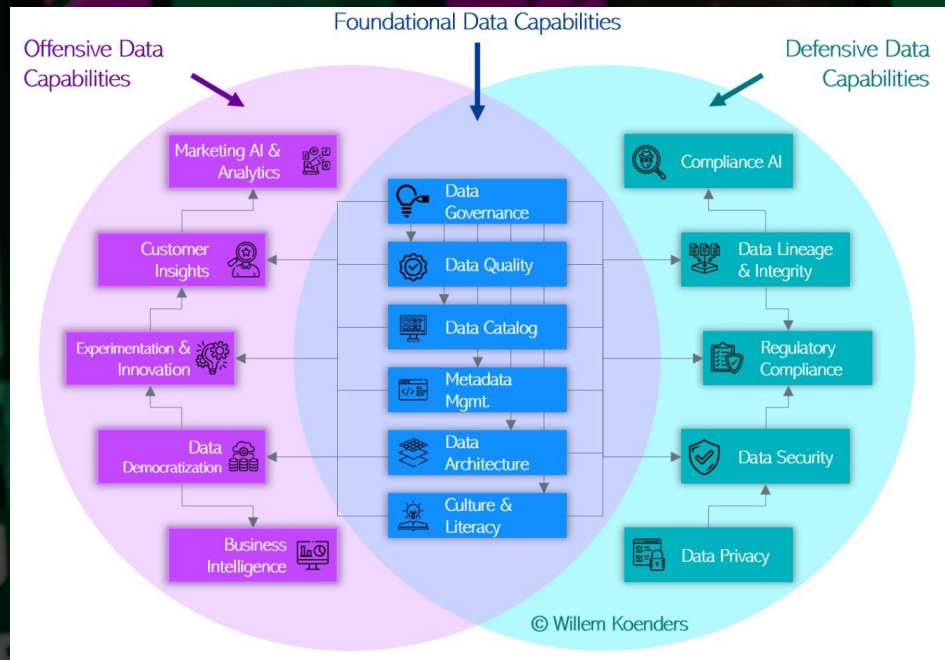
So, Data Governance is not the same thing as Data Management, Data Quality, Master Data Management, Data Stewardship, etc. Data Governance is the broad umbrella, which is related to and incorporates (but is distinct from) other concepts like metadata management, data quality, data architecture, data modeling, data storage, data privacy, data retention, data security, reference & master data, etc.

I'm a big fan of the work of DAMA and the Enterprise Data Management Council.

I could go on about what Data Governance is for a while, but we're here to talk about the relationship between Data Governance and AI. Data Governance is essential for successful AI, as it increases the quality, and improves the reliability and security of the data, and hence the AI models' output.

Before we move on, there's another view of Data Governance that I'd like to talk about for a few minutes. It shows a lot of the same concepts, but it arranges them into categories such as foundational, offensive, and defensive, which turns out to be very helpful.

What Is Data Governance?



What Data Governance Is, What It Isn't

This image is from Willem Koenders (<https://medium.com/@willemkoenders/offensive-vs-defensive-data-strategy-do-you-really-need-to-choose-c04f0387dbc3>). He breaks data governance down into these three categories.

Before we start on that though, let's talk about what I call the "Five Business Drivers":

- Increasing Revenue [offensive]
- Decreasing Costs and Improving Efficiency [defensive]
- Improving Regulatory Compliance [defensive]
- Enhancing the Quality of Decision Making [offensive]
- Increasing Agility and Decreasing Time to Value [offensive]

It's critical that the company's Data Strategy be reflective of and directly support the Business Strategy. And the Data Governance approach selected must be a good fit with that Data Strategy.

Koenders describes the offensive strategy as focusing on driving positive outcomes, and the defensive strategy as focusing on preventing negative outcomes. But don't feel like you need to choose only one or the other.

The right answer, most of the time, is a mixture of offensive and defensive approaches. With the reality of limited time and resources, you need to carefully select which offensive, foundational, and defensive elements of Data Governance to prioritize, but you should be prepared to use and prioritize some combination of all three.

Significance of AI in the Modern Landscape



What AI Is, What It Isn't

Now let's talk about AI. There's no universally accepted definition of AI. The field is changing so fast that things that seem like science fiction today, end up being widely available in a few months.

The Department of Defense (in the 2021 National Defense Authorization Act) defined AI as "the ability of machines to perform tasks that normally require human intelligence." I like that – specific enough to be useful, but broad enough to leave plenty of room for the field to develop and change.

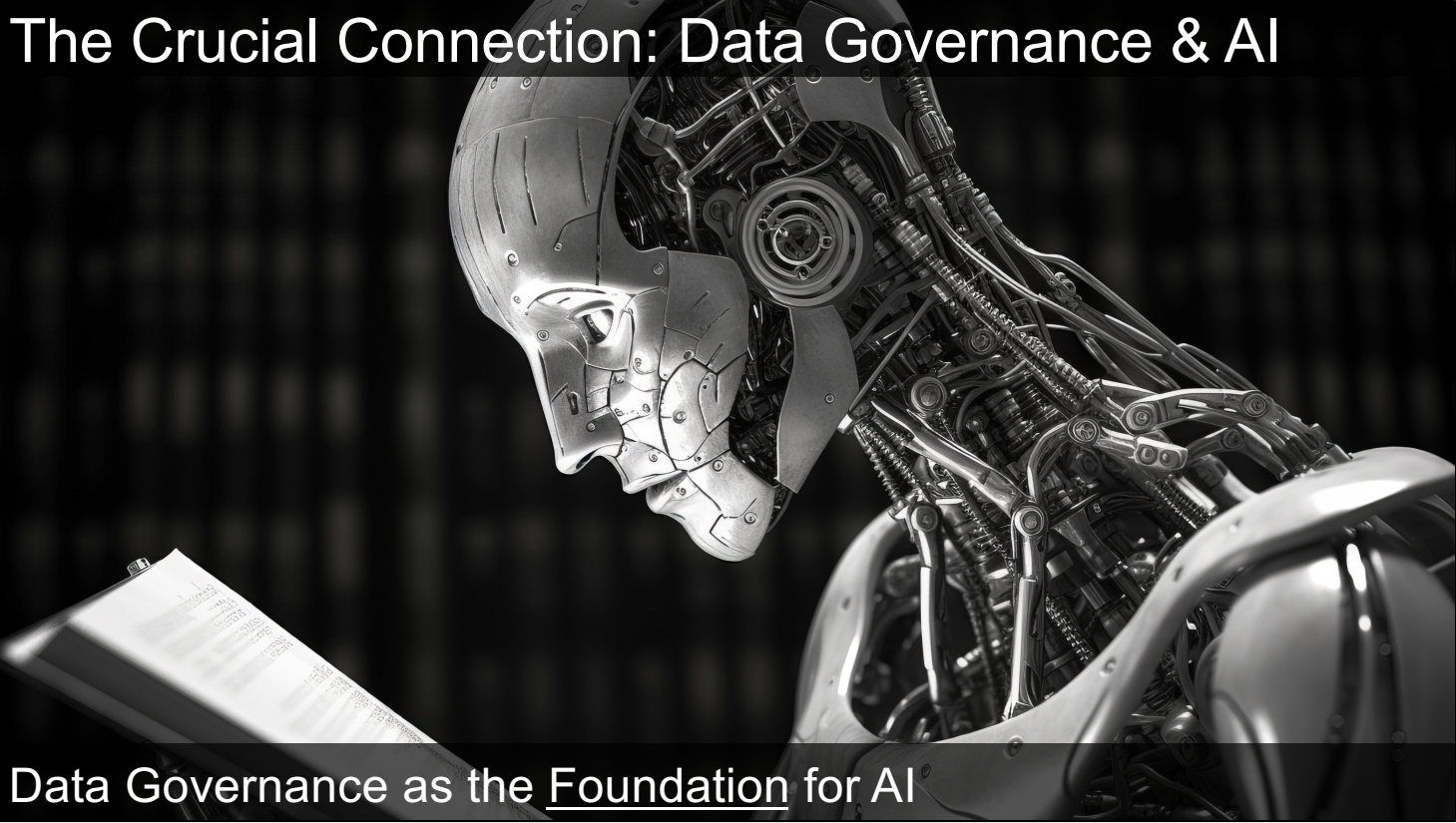
At the simplest level, machine learning uses algorithms trained on large datasets to create models that allow systems to perform tasks like making song recommendations, identifying the fastest way to travel to a destination, or translating text from one language to another. We've all used it by now.

Generative AI like ChatGPT has received seemingly endless press coverage and hype. This type of AI can generate text, images, video, music, etc. And there are other types such as artificial narrow intelligence, reactive AI, predictive AI, and limited memory AI.

Within finance and banking, AI can enhance real-time decision-making and risk management. Traders are using algorithms to identify market trends and make quicker and more informed trades. Financial institutions also use AI today in credit scoring for better lending decisions. And AI excels in fraud detection and anti-money laundering by analyzing vast data sets to spot anomalies.

But it's not a panacea. AI is trained on huge quantities of human data and fine-tuned by an army of human workers. And it's not really intelligent per se. It simply takes data inputs and manipulates them probabilistically. We have to work out, as humans in society, the best way to collaborate with AI and to have it embody the human values that are critical to us.

The Crucial Connection: Data Governance & AI



Data Governance as the Foundation for AI

Why do we care about this in the first place? Let's talk about some of the challenges in developing AI.

Data exists in silos; it's highly fragmented. How do we connect the AI models we're developing to all of the relevant source systems across business units, in both on-premise and cloud environments?

And to be able to trust an AI model's outputs, we need to be able to trust its inputs. That's common sense, of course – the inverse of “Garbage In, Garbage Out”. We need to ensure that the data used for training and deploying AI models is accurate, complete, timely and consistent (ACT+C).

I've heard the saying “data is the new oil” – but I prefer “data is the new water”, because I use a lot of analogies with water: sourcing, filtration, distribution, and consumption.

AI requires large volumes of data – training data to train the model, and then test data to validate that the model generates the desired outputs, performs as expected, and is free of bias.

Once the model is “live”, it then operates on actual production data, and starts generating outputs. Changes in the underlying business or in the input data over time can cause models to drift, so monitoring and managing that production models is important.

It helps AI teams tremendously not to have to start from a “clean sheet of paper” when it comes to data policies, standards and procedures. If there's a mature data governance practice, that means there's a data catalog, and standards for data quality, retention, privacy, and security already in place. This provides a starting point for new standards around responsible & ethical use of AI, model risk management, transparency, and compliance with new AI executive orders and emerging laws.

The Crucial Connection: Data Governance & AI



Data Governance as an Accelerator for AI

I believe that Data Governance is a critical factor in doing AI projects more quickly and more successfully.

As far as Data Governance is concerned, AI is just another downstream consumer of data.

All technology projects have (or should have) certain resources available to them, like the enterprise data catalog, data stewardship teams, data quality & observability tools, etc.

And all technology projects have (or should have) certain obligations in return, in terms of providing artifacts & deliverables, and following enterprise data governance policies and standards for data retention, data privacy, and data security.

So, to answer the “why should I care” question: if you don’t care, your AI project will take longer, be more expensive, have a lower chance of succeeding, and may ultimately run afoul of the Internal Audit, Compliance, Cybersecurity, and Privacy groups in your firm, as well as outside regulators.

If you take the time to get to know what your enterprise Data Governance program offers and requires, your project will go faster & more smoothly, your data scientists will complain less about the amount of time they spend finding and fixing the data, and result will be more transparent, reliable, trusted, and compliant with the enterprise’s goals and objectives.

You may already have an AI team and capability in place, but it’s definitely to your advantage to reach out to and foster closer collaboration with your enterprise and business unit data governance teams.

Enhancing Collaboration and Communication



Bridging the Gap Between Data and AI Teams

Why do I say this? The Data Governance people are heavily networked with various parts of your company. They know the data owners and SMEs in the business, and a lot of the technology people.

They understand the data and have usually worked with it for years. They know the data quality rules and the data lineage. They have access to the right people in the business, operations and IT, and are used to being facilitators in the collaboration and communication process between business and IT.

So, by bringing the enterprise and business unit Data Governance teams into the AI process, they're going to save the AI team a lot of time in navigating the business, as well as the technology, political, compliance, privacy, legal, and security processes.

Think of them as expert guides to the enterprise and to the data. And they're typically good diplomats and facilitators in the process of getting access to the data, understanding the data, and documenting the end results, so the AI projects don't end up increasing the technical debt of the company. You could do the AI development process without them, but why would you want to?

The Enterprise Data Governance program can usually provide helpful collaboration tools and platforms around source systems and data.

This includes things like an enterprise data catalog, data quality and observability tools, data lineage tools, as well as showing the AI teams how to leverage the catalog for data security and data privacy purposes. Give your AI teams a head start when it comes to data sourcing, getting access to, understanding, and assuring the quality of the data they're using to train and feed their models.

Enhancing Collaboration and Communication



Establishing a Common Vocabulary

One of the first things to make sure of is that there's a common vocabulary around information assets.

There are applications, databases, schemas, tables, and columns. And don't forget views, indices, flat files, integration pipelines, ETL tools, data virtualization, business intelligence, advanced analytics, predictive tools, metadata management platforms like enterprise data catalogs, data modeling tools, master & reference data management platforms, etc. It can get pretty confusing pretty fast.

In addition to creating a glossary of business terms and abbreviations, you'll need a similar glossary of the technical terms, because people love to use two different terms for the same thing, or two terms that sound the same but are actually quite different. It will help new people get up to speed faster if there's an accepted glossary of the business and technical terms. It will ensure that everyone in the business who works with data & technical concepts has the same understanding as the technology and AI people in the enterprise.

Another thing the Data Governance people can help with is plugging the AI teams into cross-functional groups across the various business units, and with people in groups like Risk Management, Compliance, Internal Audit, Legal, Privacy, Data Retention, and Cybersecurity.

Regulators such as the SEC, the Federal Reserve, the Consumer Financial Protection Bureau, the Federal Deposit Insurance Corporation and Federal Trade Commission, the National Credit Union Administration, and the Office of the Comptroller of the Currency definitely want to see that the AI teams are connected with these internal groups, and that they're following all applicable regulations around data privacy, retention, security, as well as things like the Fair Credit Reporting Act, Equal Credit Opportunity Act, and the updated Gramm-Leach-Bliley Act, just as an example.

Core Principles - Understanding Your Data



What's Available Where? Who Owns The Data? Who Are The Experts?

There is some basic “blocking and tackling” that the Data Governance teams can help with:

- What's available where (in other words, finding data sources)?
- Who owns the data and who are the experts (i.e. finding data owners and SMEs)?
- At the source system level, what is the underlying data model?
- Are there useful intermediary sources (operational data stores, data warehouses, data lakes)?
- What are the data quality rules that affect data as it flows through the ecosystem?
- What is the lineage for data that has gone through several hops before it gets to the AI team?
- How trustworthy is the underlying data we're going to use to train our model?
- Are there historical or other types of bias “baked into” the datasets we'll be feeding to AI?
- Are there data privacy laws & regulations we need to consider?
- Is there personal & private information included in the datasets we're planning to use?
- What's the impact of the existing General Data Protection Regulation and the new European Union AI Act?
- What US and state legislation will impact us? What about the recent AI executive order?

There's lots to think about here, and many aspects where the AI team would benefit by being able to tap into the underlying data community in the firm.

It might seem like more “bureaucracy” at first, but down the road, it will end up saving a lot of time & effort.

My experience with Data Governance is that after an initial “investment period”, it ends up as a big productivity booster for technology teams.

Increase Productivity Through a Data Catalog



Covering Databases, Schemas, Tables, Columns, Data Products, Other and Information Assets

Let's talk about increasing productivity specifically through the enterprise data catalog.

In an ideal world, every application, every database, every information asset in the enterprise would be self-describing and cataloged in the enterprise data catalog. Data elements (columns) in databases, flat files and other datasets would be defined with clear, understandable definitions.

We would be employing Data Governance best practices for AI, and our data catalog would have built-in data quality and data lineage capabilities that would allow us to create and maintain a comprehensive inventory of the data assets and their origins, transformations, and usage.

However, we don't live in that ideal world. There's a little problem called the "scale issue". At my last company, we had over 3,000 applications and databases in the enterprise. Just in my own business unit, we had more than 260 applications and databases. And we had a broad historical mixture of the underlying technology: IBM DB2, Oracle DBMS, Mongo DB, Hadoop, a blend of off the shelf packages and heavily modified vendor products, and a lot of completely "home grown" systems. Just in my business unit, we had about 60 IT projects underway at any given time.

The approach we took to the data catalog question was, we defined five criteria that would qualify an application or database as a "key application". This ended up being about 25% of the total inventory, but it meant we could focus at first on those critical applications, and still cover about 80% of the meaningful landscape.

We did that for the first three years, and when we were ready, we pivoted to the "everything else" category, and started adding them to the catalog too, although to a slightly lower level of detail.

Data Quality & Data Observability



Increasing Levels of Trust and Confidence in the Data

So where do data quality and data observability fit into this picture?

A good high-level definition is “accurate, complete, timely, and consistent” (ACT+C), which I learned at D&B. *Accurate*: Is the underlying data accurate? *Complete*: Does the dataset include everything needed? *Timely*: How recently has the data been updated? *Consistent*: Does the data comply with the company’s formatting and other standards?

There are many other definitions of data quality, including things like “fitness for purpose”. But remember, once you’ve lost a data consumer’s trust in the data, it’s very hard to get it back again.

Gartner defines the expanded category of Data Observability as:

1. **Data content (legacy data quality)**: monitors data based on metrics such as completeness, uniqueness and accuracy, and identifies anomalies, patterns and violations of business rules.
2. **Data flow and pipeline**: observation of status and changes as data is moving through its pipeline. It monitors drifts in schema, codes or configurations and identifies bottlenecks.
3. **Infrastructure and compute**: collects and analyzes signals from the data infrastructure layer around resource consumption like compute, performance, storage and network.
4. **Users, usage and utilization**: provides critical information about data users and usage. Alerts are sent out if unusual usage is detected.
5. **Financial allocation**: By adopting the concept of FinOps and cloud governance, this reveals the underlying cost associated with each dataset and data pipeline. It provides critical information in resource consumption (e.g., the most expensive datasets). This helps with the fact that everyone says that 30% of their cloud spend is wasted – but which 30%?

Data Architecture & Data Lineage



How Does Data Flow Across the Ecosystem?

One thing my Data Governance team and I spent a lot of time on is data lineage. Sometimes, it was required by a regulator, and in other cases, we did it to evaluate the trustworthiness of the data on a critical management report or dashboard.

We looked at data lineage at an app-to-app level by default, and at a data element level where needed. In one case, we traced about 100 data elements for a single regulatory report, and it took about four person-months (two people for two months, 100% of their time).

There's no arguing that being able to trace data back to its original source system improves the trust and confidence level of the users of that data.

One of the issues we'd run into was the complexity this introduced. Most of our systems were tied to at least five other systems. One database we looked at had almost fifty different downstream applications that it fed. It usually meant manually reading through tens or hundreds of thousands of lines of source code to document the origins, transformations, and destinations of the data elements.

And visualizing that resulting complexity had its challenges – we used to call the data lineage diagrams “spaghetti diagrams” because, unless you only looked at selected data elements, it quickly exploded out into an unreadable mess.

But tracking and document the data flows, and the impact analysis this enables for downstream applications, reports, and models, increases everyone's trust in the flows, and increases confidence that upstream system changes won't inadvertently upset downstream consumers' use of the data. We'll talk about how AI can help with this in Part 2.

Data Security and Data Privacy



Protecting Data and AI Models

Like “mom and apple pie”, data privacy and security are good things that take an awful lot of work.

We need to strike the right balance between data that is locked down so tightly that no one can use it at all (which, of course, drives everyone crazy) and the other extreme where anyone can get any kind of data they want (we called it the “data black market”).

We have to protect the data and the AI models against unauthorized access, and from manipulation and distribution.

The data security approach – in terms of defining & enforcing the data access roles, policies, and permissions – also cannot be so burdensome that people are constantly trying to game the system and work around it.

Data privacy has many components – protecting personal & private data, complying with GDPR and other data privacy regulations, working with Cybersecurity and Computer Information Security Officers (CISOs) to protect critical data. And the data security professionals are constantly working to protect the enterprise from external and internal threats, without imposing too heavy of a footprint on employees trying to get their everyday work done.

AI has the potential to introduce more and stronger tools for the “bad guys” to use. And it adds complexity for the professionals in the Privacy Office and Cybersecurity groups. But ultimately, the AI teams, the Data teams, and the Privacy and Security folks will all have to work together, because these tasks are too big and too different from each other for one large group to do it all.

Data Lifecycle Management



Data Lifecycle: Design ➡ Creation ➡ Storage ➡ Processing ➡ Analysis ➡ Deletion

Data goes through a pretty well-defined life cycle.

There's a requirements gathering and design phase. Data is then captured or created, stored, processed, analyzed, and then (hopefully) deleted in accordance with the firm's retention policies.

Data scientists can spend up to 80% of their time with the lifecycle stages leading up to data science, what people call "data munging" or "data wrangling". This consists of finding good data sources, profiling them for usability, getting access, harmonizing multiple sources into a single format, developing and applying data quality and transformation rules, and loading the data into the desired platform.

Then the data science, analysis, predictive models, and AI model training can take place. But if you're a data science or AI team member, and you spend 80% of your time just getting the data to the point where you can do your actual work, it's pretty disheartening.

Not naming any names, but I've seen data science groups experiencing much higher-than-expected turnover because the team members had to do so much unexpected data engineering, and they felt like they had been sold a bill of goods when they joined the company.

So, the Data Governance team can support a Data Engineering practice, to take at least some of this workload off the AI and Data Science teams' shoulders and help them get back to what they were hired for (and what they really enjoy doing).

The burden for the entire data lifecycle shouldn't fall only on the AI teams.

Data Governance Framework: Policies & Procedures



Focus on Increasing Value for the Business and Technology Teams

One of the first components of a robust Data Governance Framework is to develop policies & procedures, preferably at the enterprise level, but also at the Line of Business level where needed.

From *Compliance Bridge* (<https://compliancebridge.com/policy-vs-procedures>):

- Policies are the overarching tenets of your organization. They set direction, guide and influence decision-making. They provide a framework of management philosophies, aims and objectives.
- Procedures have a narrow focus; they describe, step by step, what actions to take in specific instances. Procedures have a beginning and an end and should be strictly followed to achieve the desired outcomes.

And from *FR Secure* (<https://frsecure.com/blog/differentiating-between-policies-standards-procedures-and-guidelines>):

- Standards are mandatory courses of action or rules that support formal policies.
- They are compulsory and must be enforced to be effective (this also applies to policies).

This, of course, will or has taken some time to develop. I'd start with the policy, add the related standards where necessary, and then add the procedures needed as "how to" or training documents.

I'm not a fan of any of these for their own sake, or as what we call a "check box exercise". Focus on the ones that really add value – usually one per major area. Some examples include the Enterprise Data Governance Policy, or a Regulatory Reporting Accountability Policy. Once you have them drafted, you can beta test or review them with the various teams, including the AI teams.

Data Governance Framework: Data Stewardship



A Critical Part of Data Governance, and Subject Matter Expertise for AI

Data Stewardship is the second component, and I could talk about that for quite a while. It's a central part of a Data Governance program, which itself can take several different forms.

Think first about your Operating Model. Is the Data Governance program going to be centralized, which is hard to do at large firms, decentralized, which is hard to coordinate when you have many different moving parts, or federated, which usually strikes the right balance of central coordination and autonomy for the various lines of business.

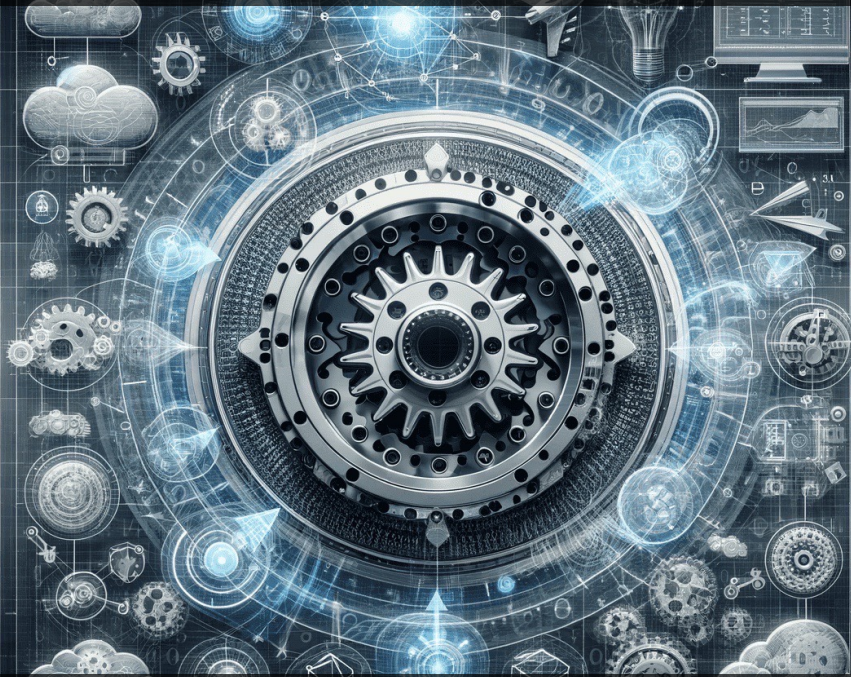
Once you've selected the "best fit" Operating Model, then think about where everyone will be sitting, and whether they're dedicated to Data Governance and Data Stewardship, or not.

My last firm used a federated operating model and had a mixture of dedicated and non-dedicated Data Governance Leads, who headed up the LOB Data Governance teams, with both dedicated and non-dedicated Data Stewards reporting to them (either directly or on a dotted line basis). In other words, we used the "all of the above" option.

Think about whether these people are situated in the business, which is recommended for several reasons, or in IT, which is sometimes necessary, especially when first launching a DG program. Again, we had both – LOB data governance people were usually in the business, as part of Operations or Risk Management, and the centralized Chief Data Office was located in IT and reported directly to the CIO.

I was my business unit's dedicated Head of Data Governance, and I had a team of 3-5 dedicated Data Stewards working for me, with each steward handling one large and one small area of the business unit. Regardless of the structure you use, those DG teams become a resource for your AI efforts.

Data Governance Framework: Metadata Management



Bringing Together Business and Technical Metadata

The third component of a Data Governance Framework is metadata management. Metadata, of course, is literally “data about data”.

Usually, metadata management is done in an enterprise data catalog. If you think about the typical database-driven application, whether it’s vendor-provided or built “in house”, regardless of the user interface on top of it, there’s going to be a lot of information somewhere about the design and behavior of that application.

That design information will either be embedded in the database itself, if you’re lucky, or distributed all over the place in other platforms or on your network, in the form of source code, Word & PDF documents, SharePoint sites, Excel spreadsheets, Confluence, Jira, ServiceNow, etc.

The catalog gives you one central place to pull all of that information together – a “single source of truth” for your enterprise metadata.

This can be a time-consuming & painful process of course, and you need to stay on top of it and keep it refreshed. Still, it’s better than the alternative of not having the catalog, because once you find the original developers have all moved on or retired, you’ll have to start every project you do with reverse engineering the application and the database, and you will have to do that over and over again.

Once your catalog is reasonably up to date and complete, it becomes a valuable resource when it comes to data sourcing, ownership, and finding subject matter experts on the data and applications. Built properly and used correctly, it will save your technology and AI teams a lot of time and provide them with a unified “business and technical” view of the enterprise’s data.

Dependence of AI on High Quality Data



High Quality Data is Essential for Development of Effective AI models



This part seems obvious to me. You can't just slap "quality" on at the end of a business process.

Data flows from the original source or transactional systems, across the enterprise to intermediary repositories like operational data stores, data warehouses, and data lakes, and now increasingly into cloud-based analytical environments and lake houses such as Snowflake and Databricks.

As a "data person", the dependency of AI on high quality data, free of bias, seems unmistakable. But I had an interesting panel discussion with a smart AI practitioner, who seemed to think his approach and his systems could compensate even for biased or low-quality data. The AI could "just deal with it".

If that's true, great. But usually the process of sourcing and getting access to the data, preparing and harmonizing it, applying transformation rules and labeling it, and then using it to train an AI model relies on data quality steps being performed somewhere along that journey.

After discovery and connection to the data sources, typically you profile the data to better understand potential quality issues. Then you create business & technical rules that the data needs to meet. Once you turn on those rules, you monitor on a batch basis or periodic intervals to see where the data might be violating those rules. Those exceptions are then researched to diagnose and remediate the root causes, and after remediation, the cases or incidents are closed out, allowing you to report on them to look for patterns and trends.

Impact of Data Quality on AI Model Performance



Critical to Building User Trust and Acceptance

Let's take a typical use case. The AI team assembles a dataset on sales or financial transactions, and employees can use AI to do natural language queries or get predictions for the next 12 months.

If the AI team wants to (or has to) “roll their own” approach to data quality, there's a very good chance the result will be different from the answers that the more traditional systems, reports, and dashboards would give for those same questions. This inconsistency, of course, would absolutely crater the users' trust in the AI system, once it was discovered.

If someone gets a quarterly income statement straight from the financial system, and someone else does a natural language query via AI, and they get two different answers for the same question (such as the firm's net profit for the most recent quarter), that's going to be a problem.

But if people get the same answers to the question, because terms are defined and used consistently and transparently throughout the enterprise, data is flowing in pipelines with well understood lineage, datasets are certified as reliable and managed as actual products, using product management disciplines, then data consumers, whether they're looking at a management or regulatory report, or a dashboard or data visualization, or the output from an AI model, will get the same answers consistently, because everything is built off the same foundation. If the data foundation for AI is built properly, the AI models and their outputs will be trusted and adopted by the business.

What we have to be mindful of, and find a way to overcome, is what it says in an article by Kevin Brewer in the Jan. 2024 edition of the *Journal of Accountancy*, that said “about 50% of senior finance and accounting professionals said they do not fully trust their data.”

Ensuring Reliable and Trustworthy AI Outputs



AI Sometimes Produces Incorrect or Nonsensical Outputs (“Hallucinations”)

I’ve been reading more articles and hearing more conference sessions about the need to enhance Data Quality to better support AI. To me, Data Quality has always been the “red headed stepchild at the family picnic” – something everyone agrees is important but somehow, the company never gets around to funding properly.

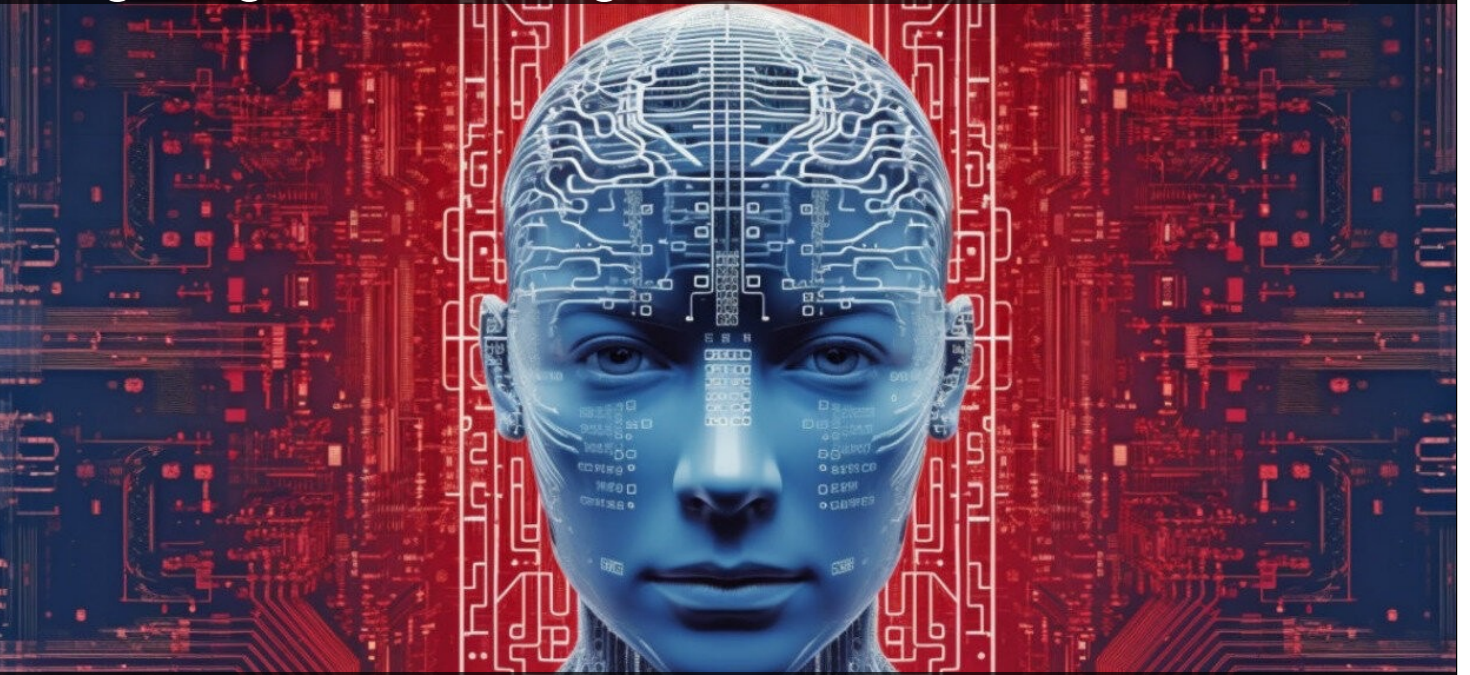
But the old saying “Garbage In, Garbage Out” is still true in the early AI age that we’re in. You can’t build reliable, accurate AI models on bad data. If your training data is flawed, or even radically different from your production data, your model will make bad decisions or predictions in the end. So, the data needs to be high quality in terms of its inherent accuracy, completeness, and timeliness, and consistency with your larger dataset.

To improve your models’ robustness, you want to apply rigorous data quality checks on the input side. You don’t want to leave this up to your AI development teams. Make sure the checks you’re applying on the “data prep” side are consistent with your overall approach to data for analytics and reporting.

And of course, none of this is a “one and done” process. You’ll need continuous monitoring and testing not only of your data, but of your models’ performance as well. Otherwise, you can have “model drift” that jeopardizes trust and adoption of your AI efforts.

For me, hallucinations are one of the biggest problems with AI. Imagine the effect on trust if senior management or regulators are questioning the output of your models, and the answer turns out to be that the AI was just having a bad day – hallucinating. We probably wouldn’t tolerate that in output from a human or a traditional system, and I don’t think we’ll be tolerating it in output from AIs.

Mitigating Risks through Data Governance



Know What You Know, and What You Don't Know

There are many different types of risks associated with AI. It's a fast moving, emerging technology with few regulatory guardrails so far. In many ways, it's a new "Wild West" out there, and the risks to our society, our democracy, our businesses, and our industries are hard to predict.

For me, not violating regulations in a business is a given. But the brand and reputational risk of proceeding with AI initiatives without careful internal oversight is substantial.

For example, algorithms are not always transparent, and can sometimes produce unfair or biased results. Trying to reduce everything about human behavior and the real world to numbers, vectors, and statistical probability brings a real danger that things which aren't easily quantifiable will fall out of the equation.

AI can vastly improve the success rates in financial services areas like Anti-Money Laundering and fraud detection, but we need to ensure the idea of a "human in the loop" becomes the norm. The human capacity for spotting anomalies, making good decisions based on judgement and experience, and using compassion and empathy to know when to make exceptions, cannot be overstated.

AI can bring with it risks based on the data, the models, how it's deployed, the underlying ethics of the use case, as well as security and privacy issues, legal and regulatory risks, and business risks of both financial and reputational damage.

Data Governance can help with most, if not all, of these risks. If the business and IT people in the organization really understand the data – what they have and what they don't have – that will help the company to know what they're trying to do with AI, and what the risk factors are.

Addressing Bias and Fairness in AI



Take a Hard Look at the Data from Under-Represented Groups

When the Apple & Goldman Sachs credit card came out, people (including Steve Wozniak, a co-founder of Apple) pointed out that the male member of a married couple was receiving a credit limit ten to twenty times higher than the credit limit of the female member of the couple.

This highlights the “creativity” of AI models. Even if we’re careful not to feed it data about gender, the machine can still decode the patterns and conclude that women are historically not as creditworthy.

AI depends completely on the collected and tagged data used to train the models. If that underlying data or tagging contains encoded bias, the model outputs will perpetuate those biases. And how the statistical weights are assigned during development can have a big impact on the model’s fairness.

Some ideas on how to address this:

- It goes without saying that training data must be highly accurate and free of historical bias
- Make sure under-represented groups are included proportionally in the training data
- Go the extra mile to build diverse development teams that include under-represented groups such as women, people of color, neurodiverse individuals, people of different ages, etc.
- Experiment with different datasets from both internal and external sources
- Explicitly test, test, and test again for bias. Compare the model outputs and business outcomes for men vs. women, white vs. non-white, young vs. old, and other demographic categories
- Make sure to include and empower your stakeholders and your community – customers, investors, regulators, employees, suppliers, etc. – so they can help you spot unforeseen outcomes
- Have a way to identify and report harmful outcomes directly back to your AI teams

Broader Ethical Considerations in AI Development



Increase Accountability & Transparency, Minimize Bias & Other Harms

Often, we're trying to capture or automate the decision process a human being would use. That's hard obviously, and it's even harder when we try to remove opportunities for bias to creep in.

I was at a conference recently when someone used a scenario of consumer and mortgage lending in a discussion of making sure AI development was ethical. Yes, that scenario is heavily quantified and relatively easy to develop. But don't forget the practice of redlining – institutionalized race-based housing discrimination – began in the 1930s and didn't officially end until the Sixties and Seventies.

Lending algorithms at the time also didn't prevent the subprime mortgage crisis and the Global Financial Collapse of 2007-2008. Granted there were lots of different causes of this crisis, but let's not overestimate the power of our AI systems to foresee the future.

Lots of AI systems would fail in the face of "black swan" events like 9/11, the COVID-19 pandemic, the wars in Ukraine and Israel, huge weather catastrophes, spiking inflation rates and interest rates, etc.

We need to remember that, just like in investing, when it comes to AI, **past performance does not guarantee future results.**

Everything in our modern world is vulnerable to these shocks, but it highlights the need to monitor our AI models, and to watch for model drift or changes in performance over time. And don't expect models to automatically or magically take into account major changes in the world, our industry, our companies and businesses, or our stakeholder groups. Like any sophisticated piece of technology, AI requires more, not less, accountability and transparency to minimize bias and other harms and maintain acceptable levels of performance.

Legal and Regulatory Compliance



If You're Not In, Then You Will Be Out!

Compliance is becoming increasingly complex and expensive, so companies are focusing on shifting from spending money on people to spending money on technology.

This represents a great opportunity for data and AI, of course. People spend up to 90% of their time manually collecting and organizing information, and only 10% actually analyzing and using the data.

Financial services firms are focusing on complying with existing regulations, even as there are new regulations on AI coming from European regulators, and being discussed in the U.S. at both the federal and state level. As companies move from manual processes with a high potential for human error, to automated and AI-driven processes, it's critical to be able to provide data with the granularity the regulators expect, so firms can provide the reports and forecasts the regulators require.

Every regulatory report provides both an obligation, obviously, and also an opportunity. Imagine being able to not only report recent transactions and results as required, but also to forecast them for the next 3, 6 or 12 months. The reports and the datasets behind them can be used to drive the training of AI models that project the actuals out into useful forecasts of the future.

Firms are looking into or starting to build AI-based approaches to compliance, risk management, transaction monitoring, anti-money laundering, financial crimes, fraud detection, insider trading, as well as regulatory reporting.

If we can improve the regulatory compliance process through automation of manual tasks using AI, we can drive better efficiency, leading to significant cost savings, and we can improve decision-making and reduce time to value vs. traditional methods.

GDPR, CCPA, and Other Data Protection Laws



Make Sure You're Following Relevant Data Privacy Laws

There are a number of existing laws relating to data privacy that impact AI:

- **GDPR** – the General Data Protection Regulation (significant aspects of which apply outside of Europe)
- **CCPA** – California Consumer Privacy Act (applies to California citizens, but corporations are trying to comply across the U.S., especially since other states are proposing or adopting similar regulations)
- **ADA** – Americans with Disabilities Act – employers' use of algorithmic decision-making may violate Title I of the ADA
- The **Fair Credit Reporting Act** and the **Equal Credit Opportunity Act** – controls how companies can use data in making decisions to grant credit, employment, insurance, housing or similar decisions
- **Illinois Biometric Information Privacy Act** – protections on collecting and processing biometric identifiers such as fingerprints, facial geometry, retinal scans, etc. with significant penalties
- The FTC is proposing **new protections** to combat AI impersonation of individuals

This list is by no means exhaustive, and it's changing rapidly over time. Within the last few months, the European Parliament issued new AI regulations, and there was a new U.S. Executive Order. Ten states included AI regulations in larger consumer privacy laws that were passed or went into effect in 2023. So your AI teams will be working closely with internal Legal, Compliance, Privacy, and Audit teams.

Aligning AI with Evolving Regulatory Requirements



AI Brings Both Risks and Opportunities for Regulatory Compliance

So, the regulatory landscape continues to evolve, and inevitably AI continues to evolve at an even faster pace.

Data Governance can assist with this, of course. Regulators now routinely request to see the data lineage, business definitions, transformations, data quality rules, and business controls for the data elements involved in a report. While AI can assist with some of these requests, which we'll talk about in Part 2 of today's presentation, most of these tasks are considered to be classic Data Governance functions.

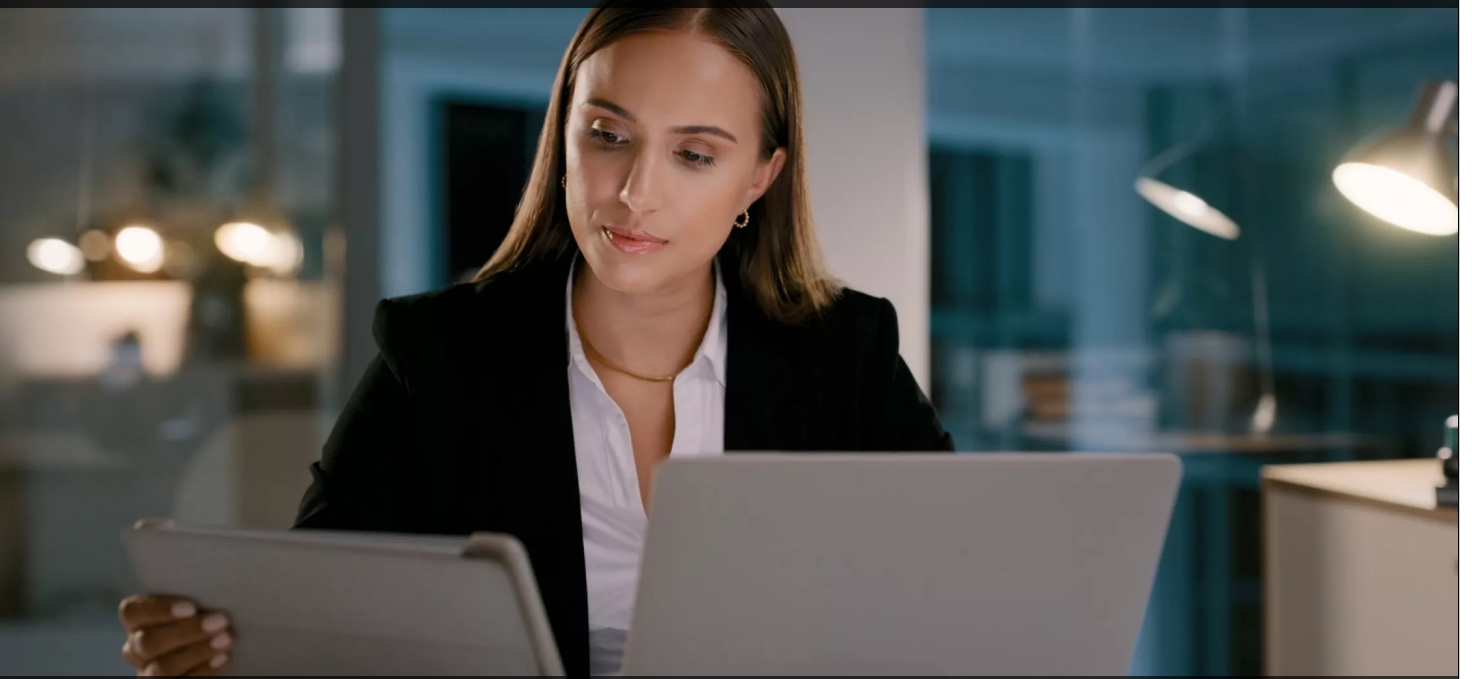
But certainly, AI can assist in internal compliance functions "keeping up with the Joneses" at the regulators. The pace of new regulations shows no signs of slowing down, and manual, spreadsheet-driven approaches are time-consuming, error-prone, and require ever larger head counts to maintain.

AI is now being used to monitor global regulatory developments and matching those up with vast internal datasets to find new requirements. This is a productivity boost for internal compliance teams.

However, the traditional three lines of defense are not going away, just getting better, more efficient, and allowing for better collaboration with the regulators.

Every new set of regulations brings both risks and opportunities, and as the complexity of today's financial services business continues to rise, augmenting human intelligence with machine intelligence may be the only way to stay ahead of the curve, and to balance innovations on the business and technology fronts with risk mitigation, both in using AI to augment people, and in the risk of developing AI itself.

Best Practices for Integrating Data Governance into AI Initiatives



To Better Support AI, We Need to Integrate Data Governance with AI

Let's talk about some best practices for integrating Data Governance capabilities into AI Initiatives.

If you're in Data Governance, reach out to the AI teams to make them aware of what you've already done that's available to them, so the AI teams don't inadvertently reinvent the wheel.

Make sure you involve the company's leadership team and all the different stakeholder groups, internally and externally.

Stress the importance of continuous monitoring and evaluation of the data going into your AI models, the performance of the models themselves, the outcomes they produce, and the importance of ethics and adherence to responsible AI guidelines you've hopefully developed by this point.

Plan for a process of iterative improvement, adoption, and adaptation. AI will ultimately be embedded into everything you do, but that won't happen overnight. Your Data and AI teams will have time to work together and to figure out the best "division of labor" that adds the most business value.

I believe that AI won't succeed over time without a strong, symbiotic relationship with Data Governance. I think we're already starting to see that. AI is a huge structural change coming to our economies, our industries, and our businesses, and we have a responsibility to help plan for that and make it as smooth a transition as possible.

Engaging Data Governance to help with the data-related tasks for the models your companies build will help drive success for your AI initiatives.

Summarizing Part 1

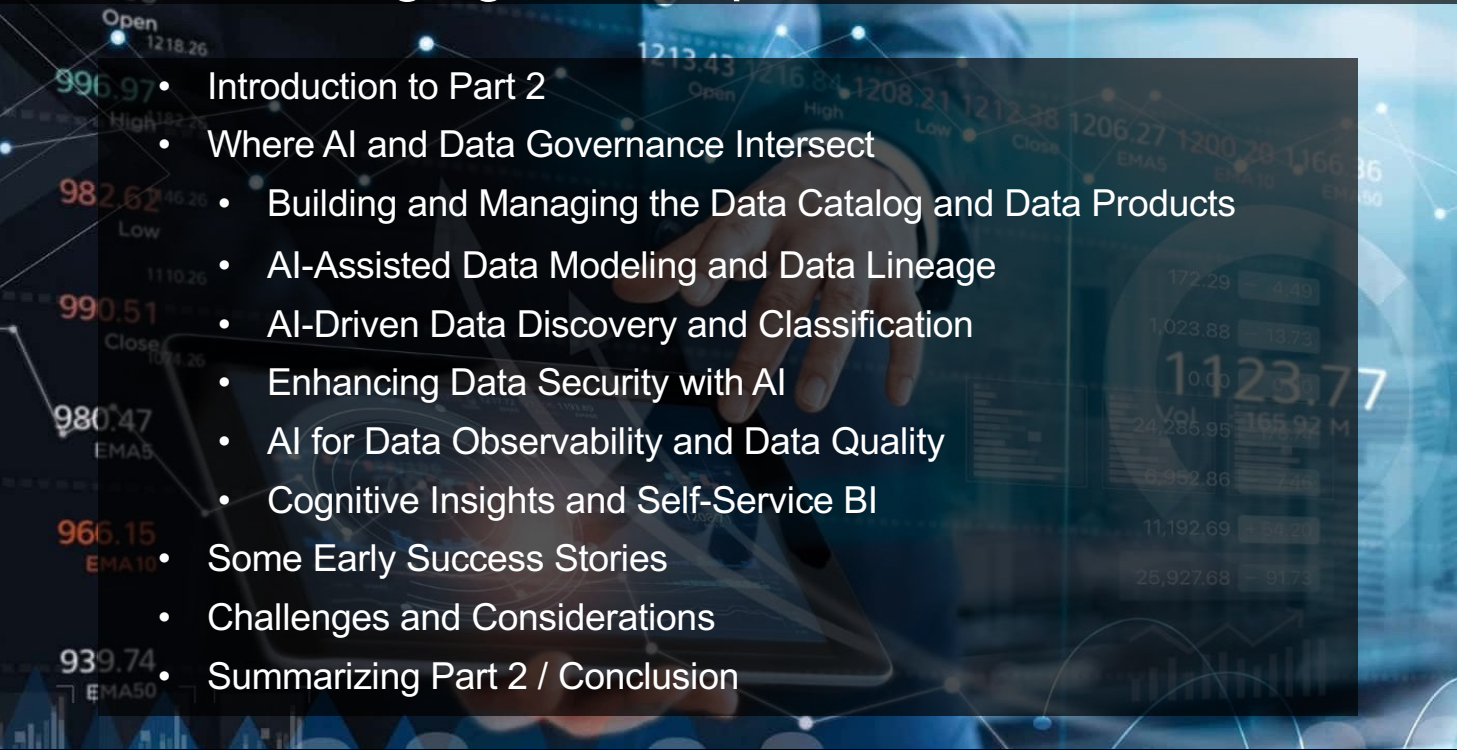
Wrapping Up, Then A Short Break

What we've talked about so far in Part 1 today:

- The importance of Data Governance as a foundation and accelerator of successful & responsible AI
- Understanding what Data Governance and AI are, and what they are not
- The importance of connection between the Data and AI teams
- Helping the AI teams understand the data, through collaboration, a data catalog, and data lineage
- Increasing levels of trust & confidence in the data through data quality & data observability
- How to protect your enterprise data and AI models – the importance of Privacy and Security
- Data lifecycle management, and supporting the AI teams so that burden doesn't fall on them alone
- Data Governance framework – the right policies, procedures, and standards to support AI
- How data stewardship brings critical subject matter expertise on the source systems and the data
- Metadata management – providing a unified business & technical view of the enterprise's data
- How AI depends on high quality, high integrity data that's free of historical bias
- About 50% of senior finance and accounting professionals don't fully trust their data
- Producing reliable and trustworthy AI outputs, and dealing with hallucinations
- Leveraging Data Governance to help with risk management for AI initiatives
- Ensuring AI systems are fair and unbiased to under-represented groups
- Explainability, trust, accountability, and transparency – now and when required by new regulations
- Data and AI can help financial companies with compliance, risk management, financial crimes, etc.
- Working closely with Legal, Privacy, and Security to comply with privacy and data protection laws
- Keeping pace with the development of AI itself, and the development of new regulations on AI

Let's take a quick break, and we'll move on to Part 2: what AI can bring to improving, optimizing, and automating Data Governance.

Part 2. Leveraging AI to Improve Data Governance

- 
- Introduction to Part 2
 - Where AI and Data Governance Intersect
 - Building and Managing the Data Catalog and Data Products
 - AI-Assisted Data Modeling and Data Lineage
 - AI-Driven Data Discovery and Classification
 - Enhancing Data Security with AI
 - AI for Data Observability and Data Quality
 - Cognitive Insights and Self-Service BI
 - Some Early Success Stories
 - Challenges and Considerations
 - Summarizing Part 2 / Conclusion

So, how we can leverage AI to improve Data Governance? I'm highlighting several use cases where I see an advantage to incorporating AI into our Data Governance and Data Management tools & platforms. Some of these are more futuristic than others.

A disclaimer – most of the Data Governance products I'm familiar with make somewhat limited use of AI so far. I'm familiar with Data Governance platforms from Collibra, IBM, Informatica, and BigID, and with Data Quality / Observability tools from Acceldata, Siffler, and Anomalo, and I'm sure there are many others. I'll mostly refrain from mentioning specific vendors.

These tools are just starting to incorporate AI-driven functionality. Think of it as a spectrum, with completely human-driven Data Governance at one end, and completely "lights out", AI-driven Data Governance at the other. We're looking for the Goldilocks spot, with a lot of AI-driven optimization and automation, but still with the "human in the loop" in the Data Governance process.

I'm not talking about vendors talking about using their tools to support AI, which we covered in Part 1. Here I'm talking about automated, intelligent functionality that helps advance and support Data Governance itself. And I'm trying to separate the current reality from the incessant hype we hear every day.

Where AI and Data Governance Intersect



What's the Opportunity Here?

Data Governance must deal directly with the “scale issue” at large enterprises.

My last firm had over 3,000 applications and databases to govern, and of course, only a finite number of people in the enterprise and business unit data governance groups. You can't solve this problem by throwing bodies at it.

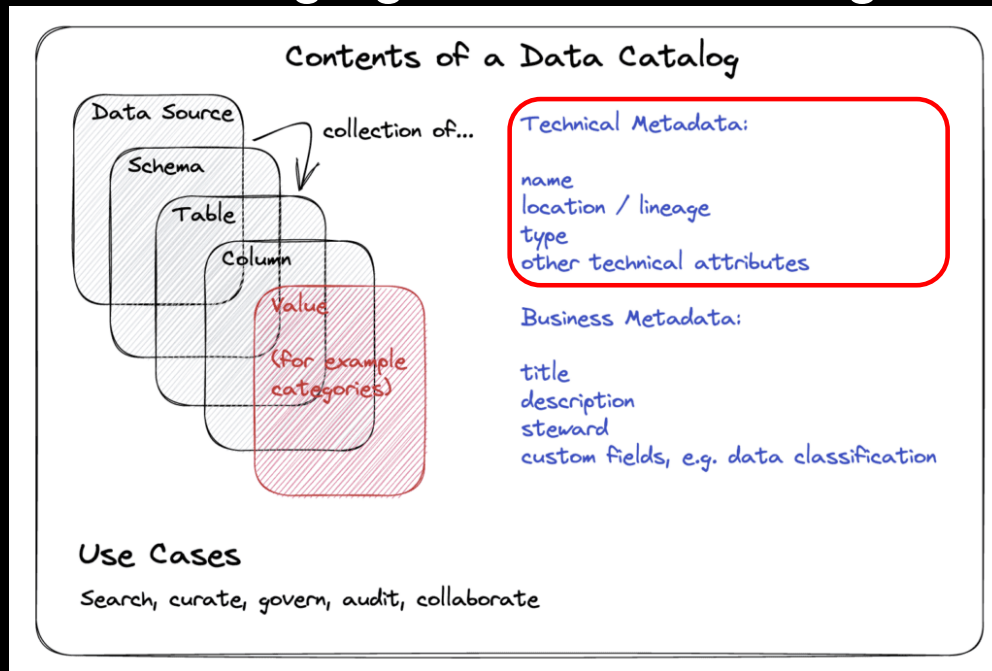
Anything AI can do will make a big difference. Of course, Data Governance can support AI. But there's a lot that AI can do to support Data Governance.

For example, think of the data and project lifecycle. Metadata is created about the applications, databases, and projects going on at any given time. A lot of that metadata is captured informally, in Word and Excel documents, in systems like Confluence, Jira and ServiceNow, and in SharePoint sites and shared network drives. It tends to get lost when a project team turns over, which happens a lot when one release “goes live” and another project phase starts up.

Building a data catalog and keeping it up to date as the underlying systems change and evolve is a big help to everyone. But it's an awful lot of work.

A modern data catalog is an essential tool. But what can we do to make the metadata management process more automated, and less manual and labor intensive?

Building and Managing the Data Catalog



Driving the Initial Population of the Metadata

Say your team is working on the initial population step of building & managing your catalog. How do you use a catalog that can at least partially automate that process?

You've got to know that the application or data source exists – either through finding it manually or through an automated scan. Then, you need access to its structure.

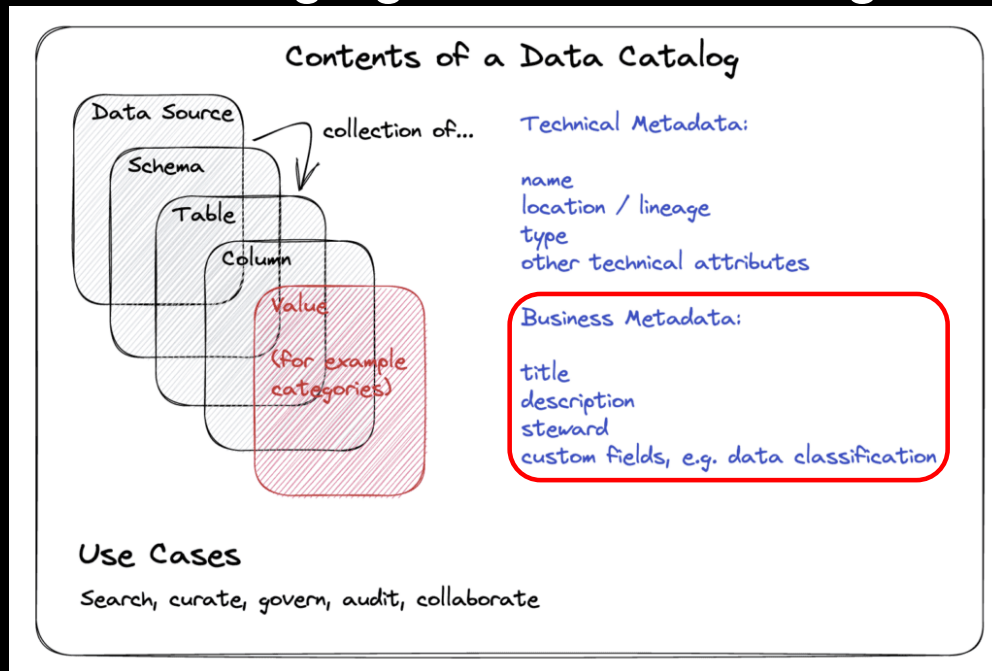
Once you know it exists and you have permission, you “read it into” your catalog. Hopefully, no manual steps needed here. At this stage, you've got the entire structure in the catalog, but not much more than the names of the database, schemas, tables, and columns. What people call the “technical metadata”. And there can be thousands of columns in a large database application, and hundreds or thousands of databases.

This is the first place where AI can be helpful, in terms of recognizing patterns, relationships, repetition, and connecting columns with other database columns to show potential lineage either to other databases or internally in the current database.

The curation and approval of that suggested metadata most likely has to be done by a human, at least to correct errors and to fill in things the machine ingestion step missed.

You'll still have the scale problem to some extent, but hopefully applying AI here saves the data stewards a lot of time.

Building and Managing the Data Catalog



Staying On Top of Changes

The next phase though is maintenance. Your data catalog should have an automated way to schedule refreshes of the catalog, usually tied to the frequency of new releases by the development team.

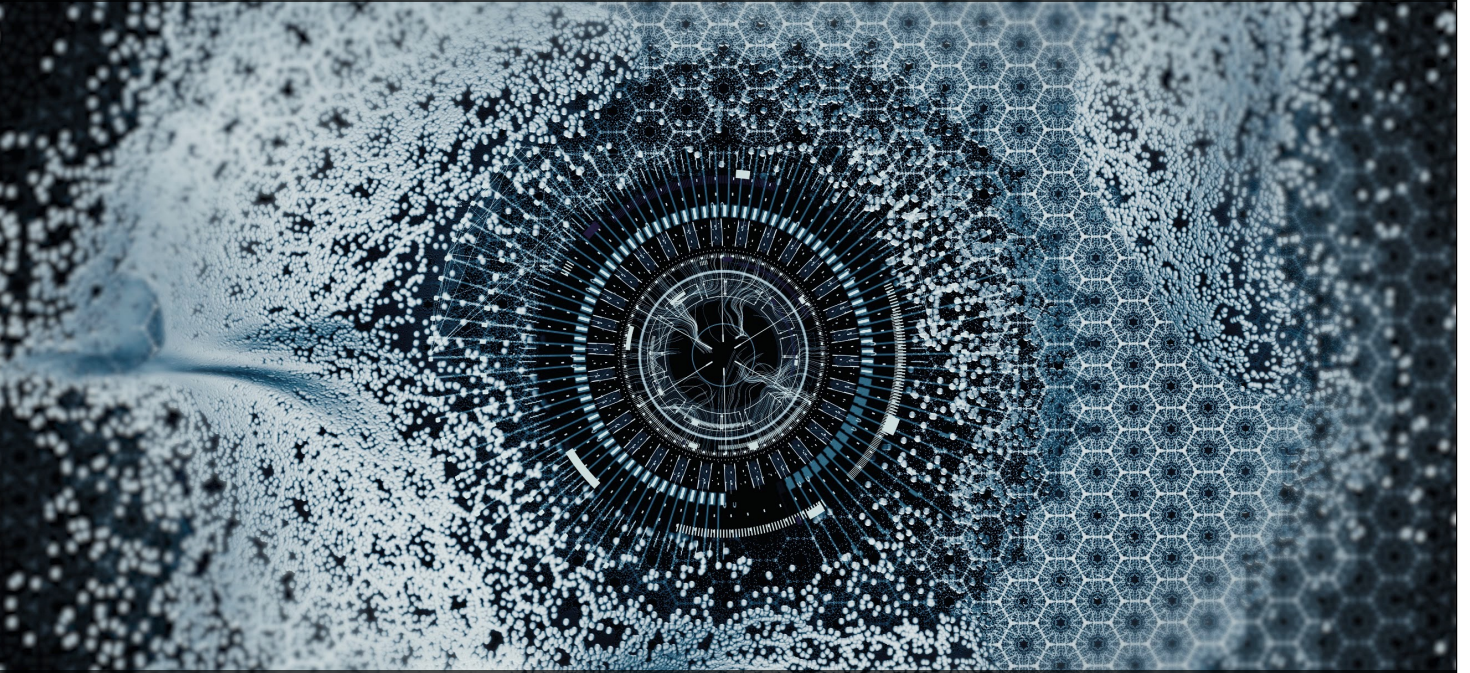
An AI-enabled catalog would automatically pick up on newly added objects (schemas, tables, and columns), as well as changed objects (different data type, size, etc.), and deprecated objects.

The data stewards responsible for that application should only have to deal with those changed objects, and all of their previous work should be protected from being overwritten by the refresh process.

In addition, the AI enrichment process that was applied in the first intake should be available on subsequent refreshes, so new assets come in with suggested definitions, privacy classifications, relationships and lineage, and user-defined categories.

The data catalog should be the single source of truth for the information assets of the whole firm, and it should be – as much as possible – self-updating, in order to maximize the productivity of the data stewards and development teams working to build and enhance the application.

Building and Managing the Data Catalog



Mapping Technical Metadata to the Business Glossary

Another area where AI can help is in building and maintaining the mapping between the technical metadata at the physical level (schemas / tables / columns), and the business glossary at the logical level. Every asset should eventually be tied to a business definition. But it should be a one-to-many relationship – between the business definitions in the glossary and the technical metadata.

For example, the business glossary will probably have several thousand entries and definitions, vs. potentially millions of technical metadata assets.

By mapping the logical to the physical, the business can manage the definitions and changes at the logical level, and they automatically apply to and update the associated physical assets, i.e. Customer ID / Number to all of the instances of that column.

This is much more practical and sustainable, and it's a good place for AI to assist. Once a human steward connects a business glossary entry to a physical data asset like a column, the AI can propose similar linkages to all of the similar physical data assets.

When there are new incoming data sources, those schemas, tables, and columns can be compared for similarity to existing assets, and then mapped back to the business glossary, at least in a proposed fashion that can then be reviewed by a human steward. Another area for AI to deliver huge time savings and productivity for data governance.

Building and Managing Data Products



AI Can Assist with Initial Creation and Ongoing Updates

This is a relatively new approach – to treat underlying datasets, and the data consumption “endpoints” that use them, as products in their own right, and to use basic product management disciplines and techniques in managing them.

Managing data as a product makes realizing value easier in the short term, and makes it more straightforward to build additional “consumption use cases” over time, which provide additional ROI from the initial investment to build that particular data product.

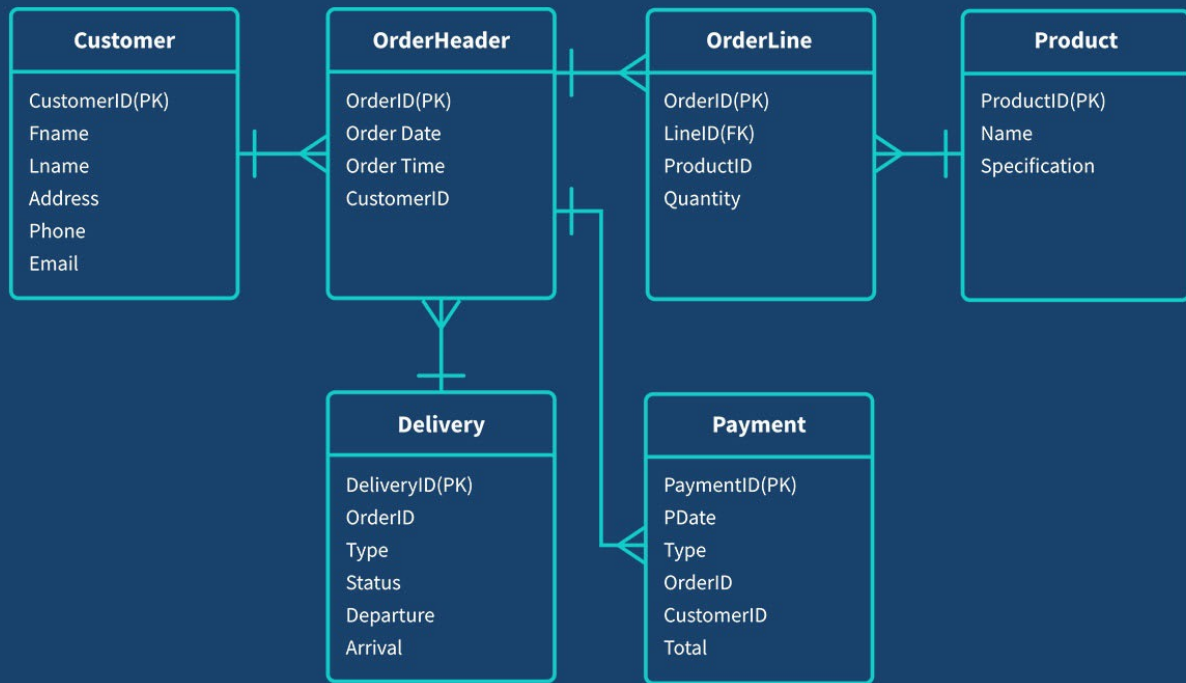
AI can help with the initial creation of data products, with the impact analysis to keep them updated over time as the upstream systems evolve, with the analysis of who’s using them and how much they’re costing, and with managing potential sprawl.

Typically, data products are described in the data catalog, and made available in a self-serve “data marketplace” for consumption by the right people at the right time.

There are dedicated “data product builders”, or you can develop and test them another way, as long as the result is available in the catalog for access through the marketplace.

Data products are usually accompanied by a provider / consumer contract, agreement, or service level agreement, and may be associated with a “data mesh” or “data fabric” approach.

AI-Assisted Data Modeling



A Little More Futuristic

Data modeling at the conceptual, logical and physical levels helps us understand data. But the practice of human beings modeling data seems to have become outmoded, and certainly financial pressures and the scale issue we discussed earlier are factors.

I'm hopeful that AI can generate diagrams like this (either now or soon), at least to a first draft level, by analyzing data sources and relationships in the data and metadata.

This would be a big time saver for data engineers and data scientists, who otherwise have to spend hours manually generating or reverse engineering those data models.

Large companies, in financial services particularly, are dealing with potentially thousands of complex, legacy data sources, and usually don't have complete, up to date data models for every source system.

But for people in the "data community" at these firms, that's an issue that gets in the way of their understanding those source systems and creating robust data products that support all of the different use cases they need to support, such as AI and self-service business intelligence.

AI can help push data modeling to the next level with automatically generated data models that can then be reviewed and tweaked as needed by human developers.

AI-Assisted Data Lineage



Not Perfect Yet But Improving

Data lineage involves being able to record and visualize data as it flows through your ecosystem.

Some benefits include improvements to data governance and data ops, more robust regulatory compliance, increased trust in the quality of the data, impact analysis before making system changes, and data migration (including migrating data to the cloud).

Data lineage may be built into your data governance platform, or it may require a separate product.

There can also be multiple levels of data lineage. For example, some cases only require application to application lineage (database A feeds data to database B). Other uses require getting down to the individual column level, including tracking all the transformations data goes through along its travels.

Fully automated, 100% accurate data lineage in a large heterogenous technology environment is probably still not achievable, even with AI assistance. But tracking data lineage manually is so labor intensive that any technology that improves the process is worth researching. One high priority lineage request from a regulator took two people from my team, dedicated full-time for two months just to trace the data lineage for that one report. So, applying AI here is a big benefit.

AI-Driven Data Discovery and Classification



Having managed a team of 3-5 data stewards in a capital markets business for six years, there were lots of areas in our data catalog where applying AI could improve things.

One example: AI can help data stewards to organize & manage the catalog by enhancing its **Data Classification** capabilities. Reviewing millions of data elements for sensitivity levels really isn't feasible.

AI can facilitate the classification of sensitive and private data, ensuring that data is only available to the people & systems that have a permissible purpose to access it. Other users and systems either can't see it at all or can only see an obfuscated version.

This application of AI is primarily privacy and security related, protecting data from external threats and internal people whose jobs don't require access. It's particularly important to classify data that is spread across many locations or hosted in the cloud, to help prevent exposure of multiple copies of sensitive data.

Usually classification involves three or four levels of sensitivity: high, medium, or low, although they may be called something else in your organization. AI can provide suggested labels and classification, but it should still be a human being that approves whether the classification levels make sense.

AI-Driven Data Discovery and Classification



Data Categorization for Grouping

Categorization is very similar to classification, and they're often used interchangeably. Categorization is all about grouping things that are related or alike in some way. For example, you can categorize data elements like Street_Address, City, State_Province, ZIP_PostalCode, and Country under a category named "Address".

AI-enhanced categorization functionality in the data catalog is helpful from both a regulatory compliance and a self-service perspective. As teams and users seek to better understand and use the enterprise's data, categorization gives them a chance to facet or narrow down their searches to assets with similar category assignments.

Creating an ontology for categorization purposes is complex and usually incomplete. AI can look at all of the technical & business metadata, as well as the data values, to find things that are similar to each other, and should be assigned to the same category.

Sometimes this will use human-provided groupings, such as a category called Contact Info to be applied to phone numbers, email addresses, and social media handles. And sometimes the machine itself will suggest relationships, such as identifiers that are common across multiple datasets, suggesting those rows are related somehow.

This provides a much-needed way to group related assets for compliance and self-service business intelligence purposes.

AI-Driven Data Discovery and Classification



Data Tagging for Use by AI

Tagging can add even more useful intelligence about a data asset. Data tagging involves labeling and/or annotating data to make it more identifiable and usable by algorithms.

For example, is this asset machine-generated or human-entered? Is it derived data (calculated from other columns) and if so which ones?

In theory, the business and technology teams should be able to create any number of tags, and AI would assign them when a new asset is ingested into a catalog. This tagging will be helpful to AI if that data eventually helps train a model.

This assumes that your AI model can not only read the data you're feeding it, but the metadata on that data, including tags, which is available from your data catalog.

Of course, this concept applies to both unstructured data like e-mails or media assets (photos and videos), and the structured data we've been focusing on today.

Mislabeling and relevance, as well as potential bias, can be big concerns though, plus the ability to use multiple language and cultural contexts. The pattern recognition and outlier detection capabilities of AI make it a good candidate for automatically applying tags to enterprise data, but there's lots of room for improvement here.

Enhancing Data Security with AI



AI-Powered Threat Detection

It's important for the Governance, Privacy, Security and AI teams to collaborate and have a common inventory of the information assets of the company. When you don't, each group ends up with its own perspective, and "never the twain shall meet".

With a single repository of enterprise information assets and metadata, you can let the Privacy folks look at what they need, and the Security team, and so on. But it's the same underlying inventory, and you can extend it and add to it as needed.

Cyberattacks are increasingly sophisticated and ambitious. And nation states have joined the fray – attacking financial institutions because "that's where the money is".

By using AI to help with data discovery and classification, it's easier to protect assets because now at least you know where everything is. And you can employ AI driven threat detection as well, to pick up on subtle attacks you otherwise might miss.

Every information asset needs to be protected, and the Security team can't work from a spreadsheet – it's too easy for assets to fall through the cracks.

The only way to stay on top of all the vulnerable assets is to scan the environment frequently and incorporate AI into the threat detection and real-time monitoring platforms. If the "bad guys" are going to be using AI, the "good guys" need to as well.

Enhancing Data Security with AI



Continuous Monitoring and Adaptive Security Measures

One of the biggest challenges is the sheer amount, variety and growth rate of data that companies have to protect. The “scale issue” affects you here in a different way. It’s not just the large number of applications, databases, repositories, data lakes, etc. It’s also the speed at which those data stores are growing over time, and the complexity of the ecosystem – the different types of data and the ways those datastores are connected.

It’s a constant “arms race” between the cybercriminals & state actors, against the security professionals and CISOs.

Automated detection of phishing can identify patterns, content, and user behavior.

Helping companies move from passwords to more secure biometrics and multi-factor authentication, and monitoring access patterns for unauthorized attempts to access restricted data, both enhance the robustness and sophistication of security measures.

And there are companies that can encrypt or tokenize all data leaving your corporate network, whether it’s headed to a cloud solution like Salesforce.com or a trading partner or outside processor.

AI has a large and growing role in adapting rapidly to the security environment, as new threats appear and known threats increase in frequency and sophistication.

Enhancing Data Security with AI



Synthetic Data for Model Training, Development, and Testing

Another challenge for data and AI teams is getting usable data. It can be difficult to get access to highly protected production data for training and developing AI models, and for the subsequent testing that has to be done.

One way to address this is by removing sensitive data altogether, things like Social Security Number, Gender, Date of Birth, contact information, etc.

Another way is to obfuscate data – to tokenize it or encrypt it. That allows AI to detect patterns in the data, without letting anything truly sensitive enter the training datasets.

The stakes are raised even further when the highly protected production data is transferred to “the cloud”, as it usually needs to be for use in AI.

But AI can solve this problem differently. AI can generate synthetic data from production data, by feeding it into a model that generates realistic synthetic data.

Synthetic data generation tools exist today to help data scientists and AI teams to create useful copies of personal and sensitive data. It can actually reduce bias by up-sampling under-represented groups and introducing fairness constraints to the generation process.

AI for Data Observability and Data Quality



Automated data cleansing and standardization is another area where AI can help.

There's a common principle called "Shift Left" – in a data flow diagram where data is generally flowing from left to right, the closer to the source you can address an issue, the better. It's less difficult and expensive to prevent problems up front, or to catch them right at the source, before they spread further into the company and its systems.

There's a two-fold savings here: first, you can save developers a lot of time by giving them a faster way to build data quality rules into their applications and data flows vs. hand coding.

Second, those rules and interventions are more likely to be consistent across the enterprise, instead of varying from application to application. And you can promote re-use of the rules, so people don't have to start from a clean sheet of paper every time.

Another way to apply AI to data observability and quality is through anomaly detection. AI is very good at detecting patterns and outliers and can notify a data steward when data deviates from the norm, even if there isn't an explicit rule in place. This is an area where all of the data observability vendors are adding AI-driven functionality fast and furious and introducing flexible ChatGPT-type capabilities to their products.

AI for Data Observability and Data Quality



Predictive Analytics and Data Quality Results Monitoring

Predictive analytics in the area of data quality and observability are another helpful AI capability.

You can use an AI model to look at all of the data quality results from a particular system or dataset, and project those results forward to see if particular types of issues are becoming more or less common. And you can automatically collapse all of the individual rule violations of a given kind down to one incident, so you don't treat them every incident in the same period separately.

Really, data quality and profiling is all about asking the business and IT folks, "did you know the data was doing this?" Sometimes they don't and want to fix it, so you can help them develop a business rule to automatically correct that error or notify the appropriate person whenever it happens. An example would be if a security price jumps more than a certain percentage from one day to the next.

Sometimes they say, "yes, we know about that and it's not a problem", such a field that's blank 100% of the time that the business just doesn't use.

Monitoring the profiling results and rule violations is a great way to find out proactively when something goes wrong, rather than waiting for downstream people to notify you.

Cognitive Insights and Self-Service BI



Improving Data Governance and Business Decision-Making

Using AI to generate data insights is another potential use case. When you can look across all the databases and datasets, you can uncover patterns and trends more easily.

Your AI models will be able to compare and analyze data and metadata across the whole firm.

One place this is helpful is in uncovering redundant data – copies of the same data that have proliferated into multiple locations. Maybe at the time that was necessary, but now organizations may have hundreds of copies of the same data scattered around – in databases, data warehouses, data lakes, and migrating to the cloud.

That's not good from a regulatory and security perspective, and it's not good from a cost perspective. Finding that redundant data can be a potentially big savings.

Using AI to create predictive analytics for better decision-making can benefit even something as straightforward as regulatory reporting.

It's great to report current actual results as required by the regulators. But imagine being able to project those figures out into the future to understand what lies around the corner for the business. That would give us an alert to make changes while there's still time to do something about it.

Cognitive Insights and Self-Service BI



Monitoring Data Governance Objectives and Performance

The first place to try enabling data-driven strategies is in Data Governance itself.

- What are the corporate objectives, and what's the corresponding business strategy?
- How does the data strategy align with the business strategy?
- What are the short and long term milestones & deliverables of the data strategy?
- How can AI help to deliver those data strategy and Data Governance milestones & deliverables?
- How can we measure progress towards those data strategy checkpoints?

AI should either help us deliver on our objectives – to do things better, faster, or cheaper – or it should help us measure our progress by automating our insights into the business and IT performance against those our objectives.

If AI can help us improve organizational performance as a whole, and data governance and Chief Data Office performance in particular, that's a big win. We already know that simply throwing bodies at the problem won't be enough, because of the various aspects of the "scale problem". There's too much data, of too many different types, in too many different places, growing too quickly; managing the governance, privacy, and security aspects of all that data can feel like "boiling the ocean".

Bringing the insights, optimization, and automation that AI provides to the table may mean the difference between the failure and success of large scale data governance.

Some Early Success Stories: American Express



Applying AI to Data Governance and Data Management

It's hard to find success stories that aren't tied to a software vendor. And some companies have realized that their advanced data governance & management techniques, and their leading edge application of AI, are a key competitive advantage, so they don't talk about it much in the press or at conferences.

However, there are some companies incorporating AI into their product offerings and into their data governance and data management infrastructures.

American Express has made progress in incorporating AI and business intelligence in their transformation initiatives.

Through intensive customer data analysis, the company identified customer segments that were driving revenue growth, which helped them better understand those groups' behaviors, preferences, and spending habits. Leveraging the latest analytics and AI technologies, Amex has advanced towards its goal of providing a better customer experience, improving profitability, and being seen as an industry leader.

Amex also employs AI in its tenth generation fraud detection model, which helps it process billions of credit card transactions, and in its application of natural language processing to customer service. Amex prioritizes data quality, ethical and transparent data practices, regular data cleansing, robust data governance, and ongoing learning.

Some Early Success Stories: Mastercard



Applying AI to Data Governance and Data Management

Another card company worth mentioning here is Mastercard.

I've heard JoAnn Stonier speak at several conferences. She's Mastercard's Fellow for Responsible AI and Data (and their former Chief Data Officer and Chief Privacy Officer). She's extremely knowledgeable about data strategy, data governance, and ethical & responsible use of AI and machine learning.

She describes generative AI as the new strategic battleground, with unique risks & challenges. Most companies are experimenting with Gen AI on a limited scale, a small percentage are starting to implement it, and the rest are looking into it or just waiting.

She emphasizes the importance of understanding your privacy and security obligations in using AI, how to manage the risks in a rapidly changing landscape, balancing safety and innovation, keeping bias out of the data, and the importance of committing to ethical and responsible use of AI from the top down.

Mastercard warns "industry leaders must approach the technology with great caution and care", and highlights its early successes in complex, multi-party payments, improving security and preventing fraud, and developing cultural & explicit process guardrails for combating biased datasets, insufficient privacy protections, and AI hallucinations. But the company says, "the biggest risk of AI is not using it."

Some Early Success Stories: JPMorgan Chase

The image shows the J.P. Morgan logo in a large, metallic, three-dimensional font. The logo is mounted on a glass building facade, with the glass reflecting a blue sky and some clouds. The logo is centered and takes up a significant portion of the upper half of the image.

Applying AI to Data Governance and Data Management

JPMorgan Chase is another example from the financial services industry. The company has been investing in data, AI, business intelligence, and dashboards for several years.

Teresa Heitsenrether is JPMorgan Chase's chief data and analytics officer and is heavily involved in overseeing adoption of AI. She's responsible for data use, governance and controls, and aims to harness AI technologies to develop new products, improve productivity and enhance risk management.

Lori Beer, JPMC's Global CIO, explained JPMorgan Chase's approach to responsible AI:

"We take the responsible use of AI very seriously, and we have an interdisciplinary team, including ethicists, data scientists, engineers, AI researchers and risk and control professionals helping us assess the risk and build appropriate controls to prevent unintended misuse, comply with regulation, and promote trust with our customers and communities. We know the industry is making remarkably fast progress, but we have a strong view that successful AI is responsible AI."

We've made tremendous progress building a competitive advantage for JPMorgan Chase. We have over 900 data scientists, 600 machine learning engineers and about 1,000 people involved in data management. We also have a 200-person top notch AI research team looking at the hardest problems in the new frontiers of finance."

Challenges and Considerations



Hallucinations and the Impact on Trust

With any new technology, there's always resistance to adoption and lack of trust. AI seems particularly vulnerable here, with people tending to be either AI optimists or predictors of AI doom.

AI hallucinations emerge in certain conditions, when AI models generate results that are factually incorrect, misleading, or just plain inappropriate for the question.

The recently decided Air Canada case – where an AI-powered chatbot gave inaccurate guidance on bereavement fares – illustrates the potential for reputational and financial damage. Companies are being held accountable for their AI systems.

Two New York lawyers relied on ChatGPT for legal research. The judge in the case noted there were internal citations and quotes that didn't exist, and sanctioned & fined them.

The AI hallucination issue not only erodes trust, it is attracting regulatory attention. The European Union's AI Act and the US Executive Order on AI include requirements to meet data transparency obligations. Results must be explainable and traceable, and high-risk AI systems must include appropriate human oversight.

While AI can deliver powerful new capabilities, it does bring some new risks as well.

Challenges and Considerations



Overcoming Organizational and Data Silos

Another challenge for successful AI (and Data Governance) are organizational silos, and fragmented or siloed data.

A certain amount of autonomy and self direction on behalf of large business units and corporate functions is to be expected at large financial institutions. The resources, headcount, and data belong to the business owners. They are usually the ones closest to the individual businesses and know best what will work for them.

But not being able to access or use data from across the enterprise can be deadly to data and AI teams' initiatives. The end result is incomplete, inconsistent datasets, lots of duplicate platforms and processes, increased cost, and reduced collaboration.

Having the right approach to data integration across silos, building virtual or physical repositories that are enterprise-wide, and having a robust Data Governance & Management program can all help to reduce the number of existing silos, and prevent new ones from being created.

Ultimately though, overcoming these organizational and data silos requires cultural change. I highly encourage you to think of the “people” aspects first – to recognize that your data and AI efforts are first and foremost exercises in organizational change management.

Challenges and Considerations



Prioritizing Use of AI for Data Governance

Traditional Data Governance has been around for several years now, and even practitioners like me wouldn't describe it as "sexy". AI, however, very much is sexy, at least since November 2022.

While your Data Governance efforts would undoubtedly benefit from applying AI, to help with scale issues, to apply policies consistently, to assist with privacy and security challenges, etc. – you'll have to make your case for the AI budget dollars and resources. Don't be surprised if there are revenue or customer related use cases ahead of you.

There are two reasons to be optimistic here: as your Data Governance team finds ways to assist and accelerate the AI model development process, as we discussed in Part 1 today, hopefully that's going to build some political and social capital that you can use to get your requests – at least some of them – considered.

Secondly, over time we're going to see more AI-based functionality being built into the popular underlying Data Governance and Data Management platforms. Most of them will tell you they're doing that now but be sure to push them on this. If it's there at all, it's at a pretty superficial level. But the potential to apply AI in order to give themselves a marketing and sales advantage is powerful, so over the next few years, we'll see them add more authentic areas of innovation, automation and intelligence.

Future Trends and Emerging Technologies



The Growing Symbiosis between Data Governance and AI

The first trend I want to highlight today is a gradual strategic shift in how organizations manage and deploy their data, in order to make their AI development process more efficient and repeatable.

There's a growing symbiosis between Data Governance and AI, a developing consensus that a robust Data Governance program can act as a foundation and accelerator for successful and responsible AI, and that in turn, AI can help automate & optimize Data Governance.

What may have started off as two separate-but-equal programs, with Data Governance reporting to one part of the corporate hierarchy and AI reporting somewhere else, increasingly start to come together. The top of the house is realizing the opportunities for overlap, collaboration, and cooperation between the data and AI teams.

Society as a whole is starting to wrestle with how to handle the disruption that AI can bring. We saw the Hollywood writers' and actors' strikes in 2023. Forrester's "2024 State of AI Report" suggests that a wave of disruption is coming.

JPMorgan Chase sees the potential for a massive workforce productivity boost over the next few years. But there could be mass-scale white-collar job realignment 4-8 years from now.

Future Trends and Emerging Technologies



Navigating Challenges and Disruption

Right now, widespread adoption of AI is still hampered by a general shortage of AI skills in the workforce, difficulty in integrating existing infrastructure with new AI capabilities, and concerns about data security, privacy and intellectual property rights.

There are a number of interesting court battles going on, and a lot of companies are waiting for the regulatory framework to emerge and mature.

Many are calling the emergence of AI so far “the Fourth Industrial Revolution”. We are still in the early days, but we are already seeing increasing uncertainty about whether it will bring positive or negative changes, or a little bit of both.

One of my favorite authors, William Gibson, said “The future is already here — it's just not very evenly distributed.” That seems to be a good description of AI today.

In some businesses and industries, AI's progress is unmistakable. In others, it's made some improvements but it's still too early to tell. And in other areas, there seems to be a real potential for economic disruption.

Many firms choose not to put themselves out there as innovators or accelerators, choosing instead to be “fast followers”. But almost everyone wants to avoid being seen as a laggard by their employees, customers, investors, and the market at large.

Summarizing Part 2



Wrapping Up

What we've talked about so far in Part 2:

- How AI can help Data Governance deal with the “scale issue” in large enterprises
- Building and managing the enterprise data catalog: initial population and staying on top of changes
- Mapping between technical and business metadata – more practical and sustainable
- Using AI and product management disciplines to help build and manage data products
- How AI can assist with data modeling for large numbers of legacy data sources
- Where AI can help record and visualize data lineage as data flows through the ecosystem
- AI for data classification for sensitivity, data categorization for grouping, and data tagging for AI
- AI-powered threat detection, and using AI in continuous monitoring and adaptive security
- Using synthetic data for training and testing models
- Putting AI to work in Data Observability and Quality - “Shift Left” and Data Quality Results Analysis
- AI facilitating insights & predictions, self-service BI, and monitoring data governance performance
- Some early success stories in the financial services industry
- Challenges: hallucinations & their impact on trust, overcoming silos, getting access to AI for DG
- Trends: symbiosis between Data Governance and AI, and navigating challenges and disruptions

Just one more slide, putting together both Part 1 and Part 2, and looking ahead over the next few years.

Conclusion



Putting It All Together

I'm inherently an optimist when it comes to human nature, technology, business, everything. There's a great scene at the end of Season 1 of HBO's *True Detective*.

It's just one story. The oldest. Light versus dark.

Well ... it appears to me that the dark has a lot more territory.

Yeah, you're right about that. But you're looking at it wrong ...

How's that?

Well, once there was only dark. You ask me, the light's winning.

Take a look at the really big picture. Most people want to do the right thing. Most companies want to improve the lives of their community, their investors, their customers, and their employees (and of course, make some money at the same time). Most legislators and regulators want to improve the lot of their constituents and ensure a fair playing field in the economy and the markets.

I think we'll find a way to overcome the challenges, improve our companies, bring together the best capabilities of Data Governance and AI, and navigate the disruptions brought on by AI and the Fourth Industrial Revolution.

There will be a role for innovation and acceleration, for fast followers, and even for laggards. It won't happen all at once, but my experiences working with a broad swath of young people, experienced people, men & women, people of color, neurodiverse people, and tons of other groups gives me tremendous hope for the future, for our children and grandchildren. It may be a bumpy road, but I think it trends up and to the right.

Questions?



I'm happy to answer any questions you may have at this point.

Contact Information



Dan Power
(617) 306-3091

powerd@hubdesigns.com
<https://linkedin.com/in/powerd>

My contact information is

Dan Power
(617) 306-3091
powerd@hubdesigns.com
<https://linkedin.com/in/powerd>

The best way to reach me is usually via e-mail or LinkedIn.

Have a great day! I hope you enjoyed this presentation and thank you again for having me here today.