MCKNIGHT
CONSULTING GROUP

Strategies for Machine Learning Success

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444

Partial Client List

CONSUMER PRODUCTS/RETAIL

Brinker International, Caterpillar, Superior Communications, Coach, Demdaco design for the home, Nabisco, The Clorox Company, Domino's, Life Fitness, Dunkin' Donuts, Manheim Auctions, Inc., Nordson, Hasbro, Just for Feet, Time Warner Cable, TimeWarner, Nutrisystem, RAC Rent-A-Center, Viking, Safeway, Weiler, Red Robin, Smart&Final, PLI Practising Law Institute

FINANCIAL

SchoolsFirst Federal Credit Union, Fidelity Investments, Bank One, American Credit Acceptance, Litton Loan Servicing, CoBank, Standard & Poor's, MFS Investment Management, State Street, سامبا samba, VISA, Commerzbank, MoneyGram, Scotiabank, Waddell & Reed, TD Ameritrade

INSURANCE/HEALTHCARE

AIG, amedisys, Anthem, Argo Group, Delta Dental, Fresenius Medical Care, SSMHealth, LIO, USAA, U.S. Renal Care, Pekin Insurance, Validus Re, United Healthcare, Molina Healthcare, DentaQuest, Erie Insurance, Navitus, AAP American Associated Pharmacies

PUBLISHING

A.H. Belo Corporation, The San Diego Union-Tribune, Forbes, The Star-Ledger

OTHER

BASF We create chemistry, Zodiac Aerospace, claritas, PointLeader, Nielsen Media Research, Intermedia Cloud Communications

GOVERNMENT AND UTILITIES

Bonneville Power Administration, DONG energy, FPL Energy Services, LCOG, Ministry of Immigration and Integration, Enable Midstream Partners, British Columbia Ministry of Finance

EDUCATION

National Heritage Academies, UF University of Florida, Cal Poly

PHARMACEUTICAL

AmerisourceBergen Specialty Group, teva, Pfizer

TELECOMMUNICATIONS

france telecom, Telia, pinger, verizon

Big/Analytic/Vector/Mixed Data Management

databricks
Pinecone
Google BigQuery
snowflake
Azure Synapse
MySQL HeatWave
teradata.
amazon REDSHIFT
CLOUDERA
DIAMANTI
Exasol
VERTICA
Starburst
ACTIAN™ activate your data
Azure HDInsight
SingleStore
Apache Impala
amazon EMR
Yellowbrick
Cloud Dataproc
ORACLE AUTONOMOUS DATABASE
ORACLE

Data Movement and APIs

Airbyte
Informatica
Qlik Replicate®
Fivetran
SYNADIA
CONFLUENT
talend
apigee
MATILLION
Stitch
Kong
Astera

Data Management

SNOWPLOW
SAP Datasphere
alteryx DESIGNER CLOUD
Profisee
splunk>
NGINX Part of F5
Collibra
IMMUTA™
elasticsearch
Azure MLOps

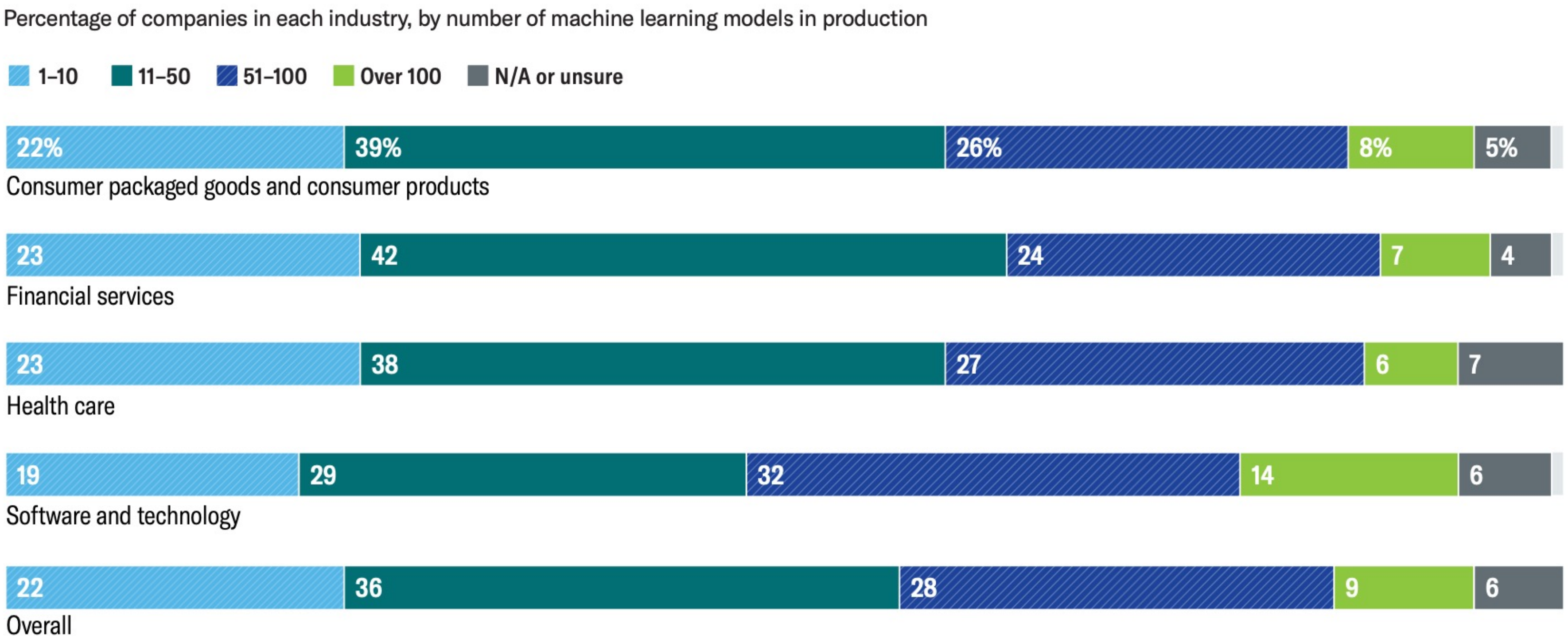
Operational/Transactional Data Management

Couchbase
DataStax
Cockroach
MongoDB.
PostgreSQL
yugabyteDB
Google Cloud Spanner
neo4j
ORACLE
redis
Azure Cosmos DB
SQLite
Amazon Aurora
Microsoft® SQL Server®

Machine Learning

- Supervised Learning
  - Data = features and a label
- Unsupervised Learning
  - Data is unlabeled
- Reinforcement Learning
  - Giving feedback to machine

ML Uptake is Strong



Harvard Business Review

MCKNIGHT CONSULTING GROUP

Use Cases for ML

	Flow optimization	Modeling and analytics	Predictive insights	Threat and risk analysis
Public Sector	Traffic flow management	Smart city planning	Autonomous routing	Situational Awareness
Oil and Gas	Pipeline modelling	Drilling patterns and asset utilization	Intelligent planning	Safety assurance
Manufacturing	Supply chain optimization	Production optimization	Predictive maintenance	Fault identification
Retail	Supply chain optimization	Customer experience	Segmentation analysis and forecasting	Fraud and theft identification
Healthcare	Patient care pathway optimization	Disease research and drug creation	Early diagnosis of conditions	Patient safety
Technology	Operational efficiency	Log analysis	Capacity planning	Cybersecurity and zero-day detection

1. When you think ML, think BIG

Where to Look for ML Opportunities

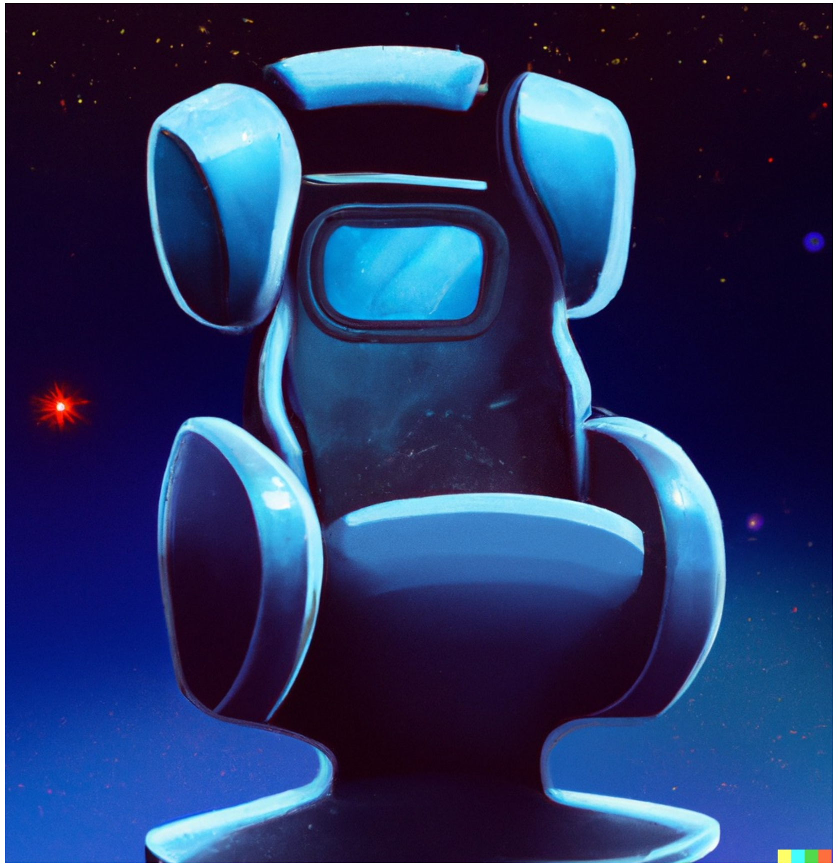
- The products you make and the services you offer
- The supply chain for those products and services
- Business operations (hiring, procurement, after-sale service, etc.)
- The intelligence used in determining and designing your product and service set
- The intelligence used in the marketing/approval funnel for your products and services

2. Align ML with business goals

- You must have a firm grasp of the business issue you are attempting to address and the value you are providing
- You should specify your machine learning solution's success criteria, anticipated results, and key performance indicators (KPIs)
- In addition to engaging in communication with stakeholders and users, it is advisable to ascertain their requirements, anticipations, and input
- You can guarantee that your ML solution is pertinent, practical, and influential by ensuring that it is in line with the objectives of the organization

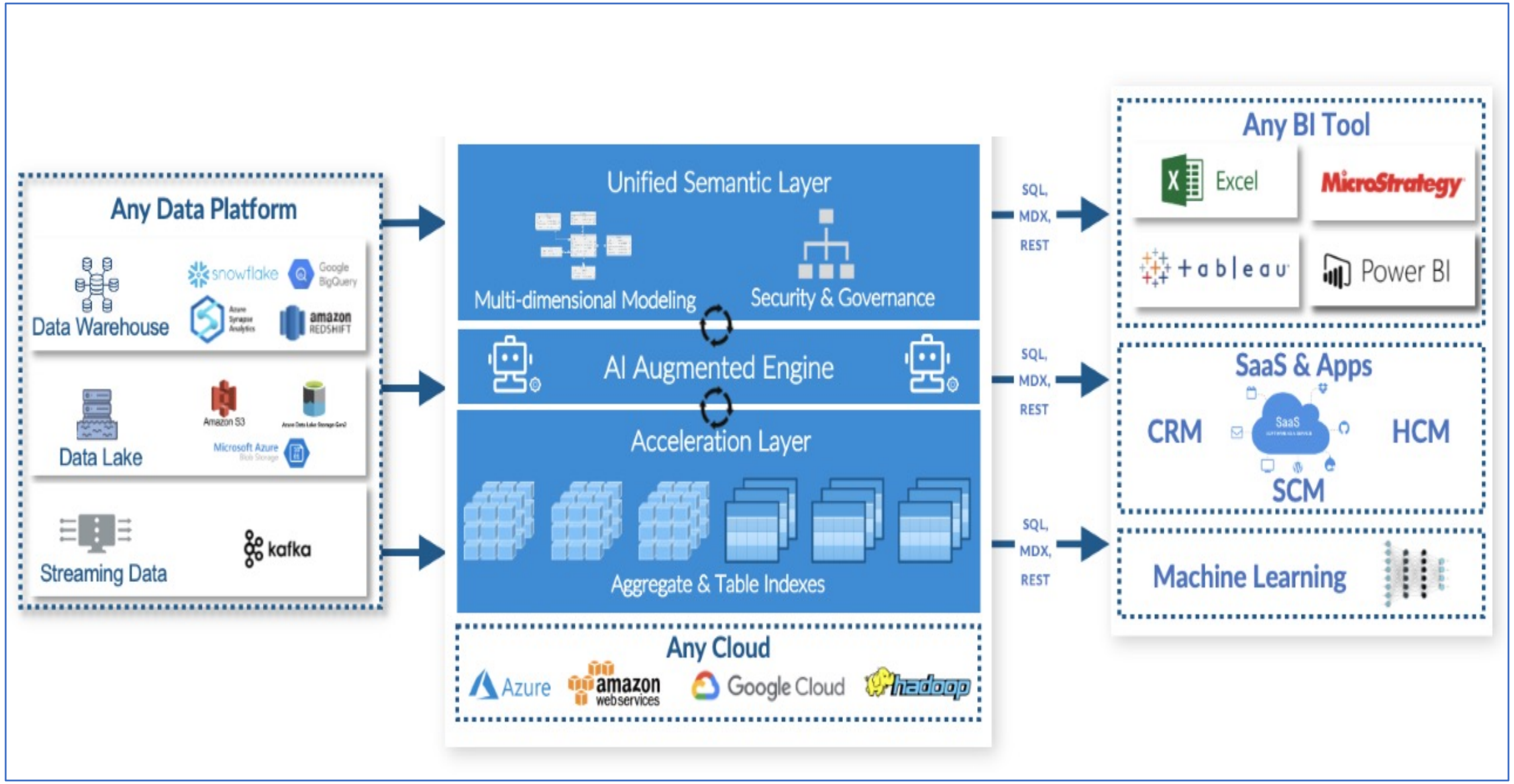
3. Boost productivity and effectiveness

- Automate and streamline repetitive tasks
- Improve decision making
- Personalize experiences
- Optimize resource allocation
- It is imperative to weigh the compromises that accuracy, efficiency, scalability, and cost entail in accordance with your specific needs
- By optimizing efficiency and performance, one can guarantee the robustness, agility, and sustainability of their machine learning solution



Machine Learning Stack

4. Machine Learning Stack



Major ML Stacks

Category				
01-Dedicated Compute	Azure Synapse	Amazon Redshift ra3.4xlarge	Google BigQuery Annual Slots	Snowflake
02-Storage	Azure Synapse SQL Pool	Amazon Redshift Managed Storage	Google BigQuery Active Storage	Snowflake
03-Data Integration	Azure Data Factory	AWS Glue	Google Dataflow Batch	Talend Cloud Data Integration
04-Streaming	Azure Stream Analytics	Amazon Kinesis	Google Dataflow Streaming	Kafka Confluent Cloud
05-Spark Analytics	Azure Databricks Premium Tier	Amazon EMR + Kinesis	Google Dataproc	Azure Databricks Premium Tier
06-Data Exploration	Azure Synapse	Amazon Redshift Spectrum	Google BigQuery On-Demand	Snowflake
07-Data Lake	Azure HDInsight	Amazon EMR	Google Dataproc	S3
08-Business Intelligence	Power BI Professional	Amazon Quicksight	Google BigQuery BI Engine	Tableau
09-Machine Learning	Azure Machine Learning	Amazon SageMaker	Google BigQuery ML	Amazon SageMaker
10-Identity Management	Azure Active Directory P1	Amazon IAM	Google Cloud IAM	Amazon IAM
11-Data Catalog	Azure Purview	AWS Glue Data Catalog	Google Data Catalog	Alation Data Catalog

5. Know What You're Building

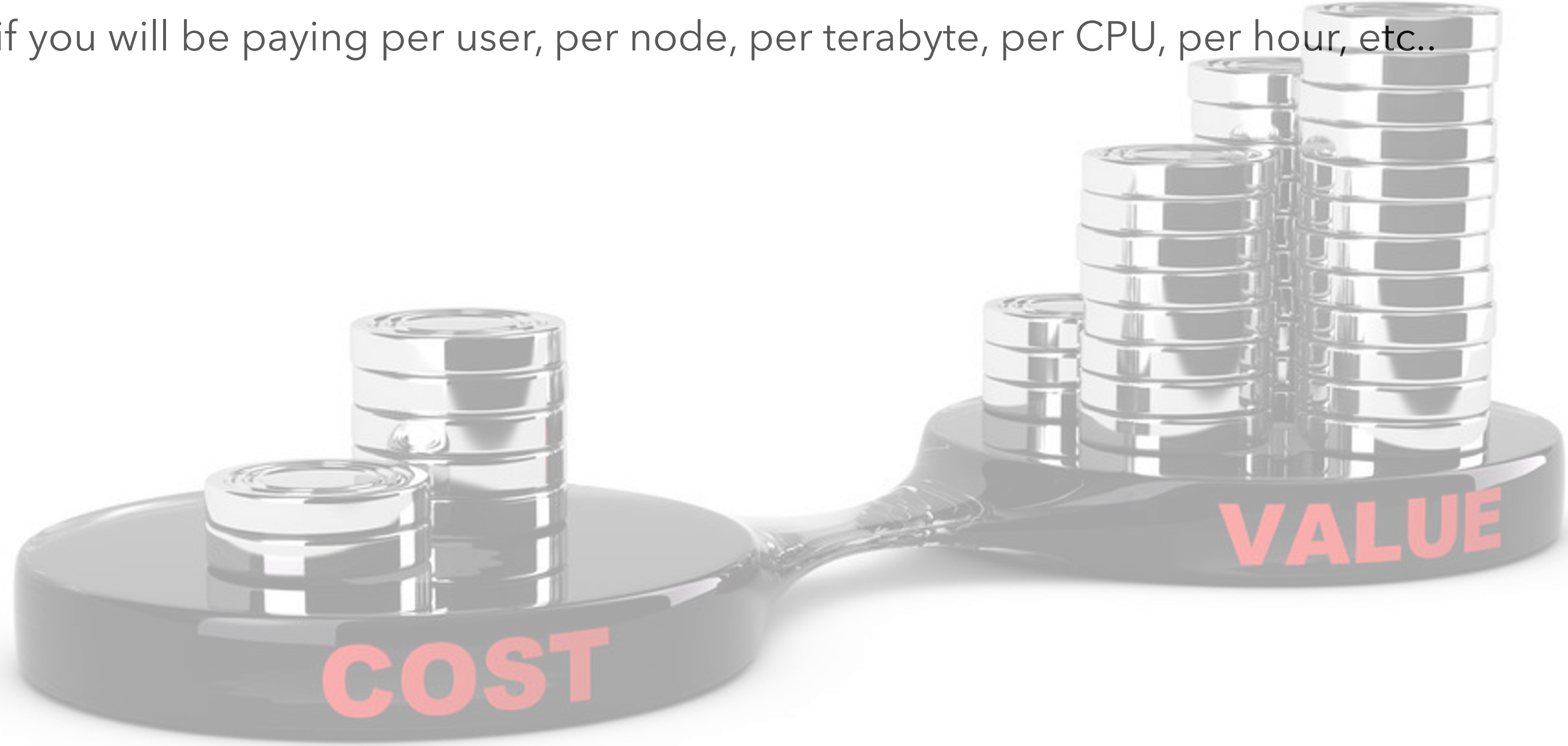
- An ML **program** which will provide analytics, automation, personalization, etc. for several projects
- A **project** which will use ML
- ML insights for a **project**
- The inclusion of new projects into an existing ML **program**

ML Costs

- Large Project Stack costs between $7M-$23M (to get full ML-based project to production) and $19M-$43M over 2 years for the enterprise.
- Buyer Beware
  - The total cost of ownership of cloud analytics platforms scales up too. Demand for ML at your company will only increase in the coming years.
- Hardware (CPU, memory, and input/output) is often the biggest performance bottleneck of a database management system.
  - Most cloud analytical products scale hardware in powers of 2
  - In many systems, you can add more memory here or more CPU there at a more fractional cost.
- The true gauge of value is price-performance. Thus, we recommend that you demand reliable performance at a predictable price from your ML platform.
- The true gauge of project efficacy is ROI.

6. Be Cost Conscious About the ML Stack

- Be on the lookout for cost optimizations like not paying when the system is idle, compression to save storage costs, and moving or isolating workloads to avoid contention.
- Look for the ability to directly operate on compact open file formats Parquet, ORC, Iceberg, Hudi
- Also, costs can spin out of control if you have to pay a separate license for each deployment option or each machine learning algorithm.
- Finally, also consider if you will be paying per user, per node, per terabyte, per CPU, per hour, etc..



6a. Watch for Pricing Gotchas

- Concurrency Scaling
- Serverless RPU Usage
- Extra Costs i.e., SageMaker costs for Redshift

	vCPU	Memory	Addressable storage capacity	I/O	Price
Dense Compute DC2					
dc2.large	2	15 GiB	0.16TB SSD	0.60 GB/s	$0.25 per Hour
dc2.8xlarge	32	244 GiB	2.56TB SSD	7.50 GB/s	$4.80 per Hour
RA3 with Redshift Managed Storage*					
ra3.xlplus	4	32 GiB	32TB RMS	0.65 GB/s	$1.086 per Hour
ra3.4xlarge	12	96 GiB	128TB RMS	2.00 GB/s	$3.26 per Hour
ra3.16xlarge	48	384 GiB	128TB RMS	8.00 GB/s	$13.04 per Hour

Instance name	RI upfront fee	RI monthly fees*	RI effective hourly rate**	Effective price per TB per year***	Savings over On-Demand	On-Demand rate
dc2.8xlarge	$0	$2,774.00	$3.800	$13,003.13	21%	$4.8000
dc2.large	$0	$146.00	$0.200	$10,950.00	20%	$0.2500
ra3.16xlarge	$0	$6,663.44	$9.128	$1,249.40	30%	$13.0400
ra3.4xlarge	$0	$1,665.86	$2.282	$312.35	30%	$3.2600
ra3.xlplus	$0	$554.95	$0.760	$208.10	30%	$1.0860

7. Benchmark Your ML Stack

- What are you benchmarking?
  - Training performance
  - Loading performance
  - Inference performance
  - With concurrency?
  - Ease of use
- Competition
- Models, Data, Efficacy
- Scale
- Cost
- Number of runs/cache
- Number of nodes
- Tuning allowed
- Vendor Involvement
- Any free third party, SaaS, or on-demand software
- Any not-free third party, SaaS, or on-demand software
- Instance type of nodes
- Measure Price/Performance!

8. Don’t Ignore the UX

- User-centric design
- Clear and intuitive interfaces, explanations
- Seek and integrate user feedback, as well as assess and enhance user retention and satisfaction
- Effective Communication of Results
- Enhanced Trust and Confidence
- Positive Brand Perception and Reputation

9. Prepare: Corporate Requirements > Data

- The split of the necessary AI/ML between the 'edge' of corporate users and the software itself is still to be determined
- **Math**
  - floating point arithmetic, deep statistics, and linear algebra
- **GPUs**
- **Python**
  - easy to program and it good enough
  - NumPy and pandas libraries are available
- **TensorFlow**
  - adds a computational/symbolic graph to Python
- R and MATLAB
  - optimized for math with features such as direct slice and dice of matrices and rich libraries to draw from
- Java and **Scala**
  - work well with Hadoop and **Spark** respectively

ML Data

10. Focus More on Data than the Algorithms

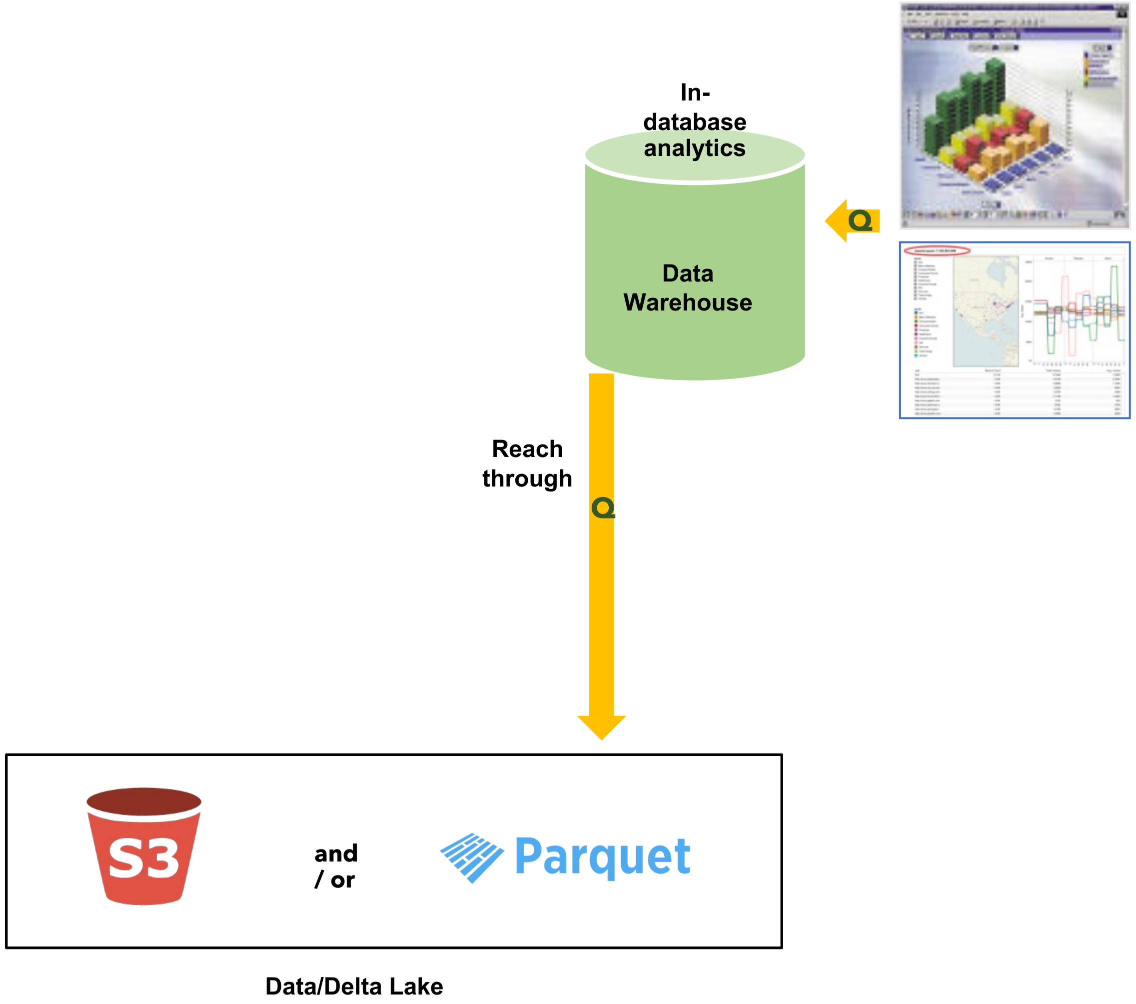
You'll need data for ML but without a discrete focus on it, you will not get it well

- Do it with data specialists
  - Data modeling, integration, quality
- It's Operational and Real-Time
- Let the data infrastructure create analytical/empowering elements
- Make the data stores discrete projects
  - With high touchpoints with ML applications
- Focus on Total Cost of Ownership first for Justification of data stores
- Build to Scale
  - It doesn't take much longer to consider all known requirements

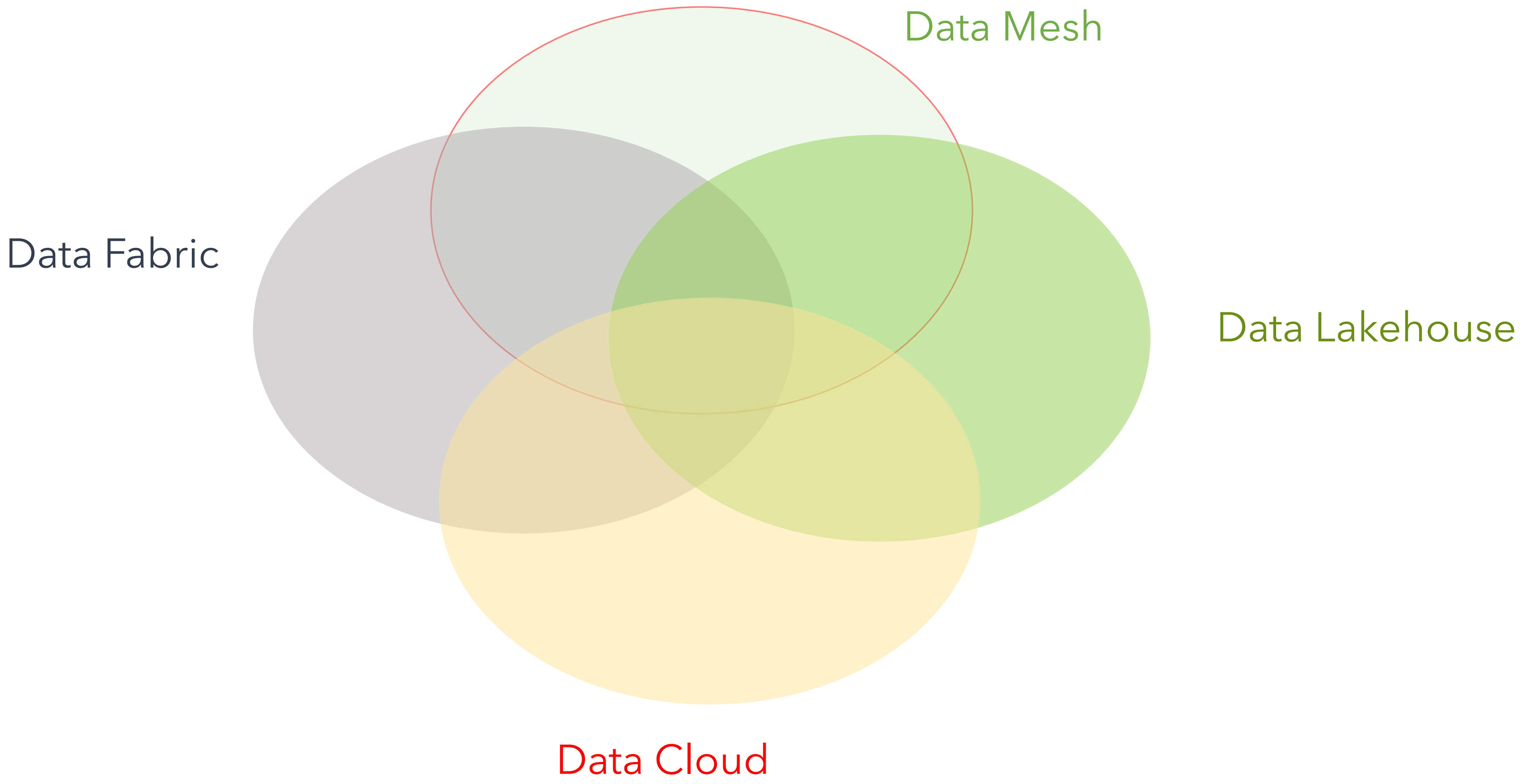
ML Data

- Call center recordings and chat logs
  - Content and data relationships as well as answers to questions
- Streaming sensor data, historical maintenance records and search logs
  - Use cases and user problems
- Customer account data and purchase history
  - Similarities in buyers and predict responses to offers
- Email response metrics
  - Processed with text content of offers to surface buyer segments
- Product catalogs and data sheets
  - Sources of attributes and attribute values
- Public references
  - Procedures, tool lists, and product associations
- YouTube video content audio tracks
  - Converted to text and mined for product associations
- User website behaviors
  - Correlated with offers and dynamic content
- Sentiment analysis, user-generated content, social graph data, and other external data sources
  - Mined and recombined to yield knowledge and user-intent signals

11. Get ML Data from a Data Lakehouse+



12. Use Architectural Pattern(s)



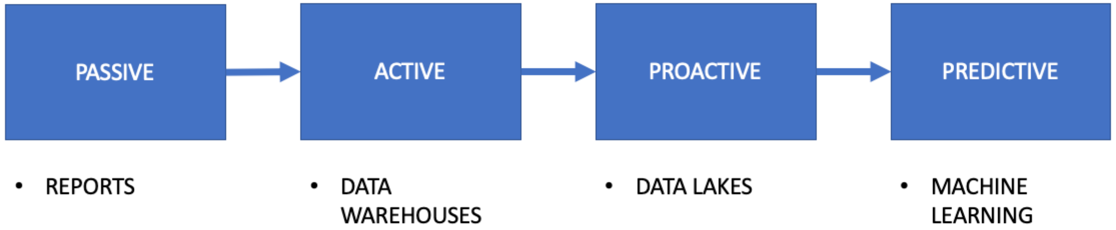
13. Make Sure MDM is in the Environment

Industry	Subject Areas	ML Applications	Objectives with MDM
Retail	Customers, locations, Menu items, ingredients, store locations	Improve customer list management and make real-time AI-based recommendations, Speed up new menu item introduction and ensure consistency across stores	Improve average order size, Foundation for loyalty program, CCPA compliance, enable online ordering, Fraud detection, real-time recommendations
Healthcare	Patients, providers, locations, supplies, donors, reference data	Enable best clinical practices	Enable data sharing for research and operational efficiency
Manufacturing	Ingredients and recipes, Financial hierarchies	Manage multiple large rollup hierarchies, Manage ingredients and variations to control the manufacturing process	Understand margin and pricing across regions; enable and expand ecommerce, Improve business processes and service while eliminating clerical errors (order fulfilment, billing, etc.), Enable nimbler FP&A (what-if scenarios, tax optimization, etc.), supply chain management
Financial	Customer, product, channel	Customer management, risk management, audit support, regulatory compliance	Anti money laundering

ML Processes

14. Adopt MLOps Early

- ML initiatives can work in isolation from each other, resulting in difficulties aligning workflows between ML and other teams.
- To be effective, ML training requires large quantities of high-quality data, which creates significant overheads across data access, preparation, and ongoing management.
- ML/data science work requires a large amount of trial and error, making it hard to plan the time required to complete a project.



From ML to MLOps

- Many companies have built strong ML capabilities
- Few businesses have been successful in putting the majority of their ML models into production, leaving a sizable amount of value untapped
- Machine learning operations, also known as MLOps, are a set of standards, tools, and frameworks that are used to scale ML to reach its full potential
- Three main objectives of MLOps, which concentrates on the entire life cycle of ML model design, implementation, testing, monitoring, and management, are as follows:
  - To create a highly <u>repeatable</u> procedure for the entire life cycle of a model, from feature exploration to model deployment in production.
  - Data scientists and analysts should be <u>shielded</u> from the complexity of the infrastructure so they can concentrate on their models and plans.
  - Develop MLOps so that it <u>scales</u> without a horde of engineers, along with the number of models and modeling complexity.

15. Strive for Iterative Pipelines

- **Reproducibility** – As with software configuration management and continuous integration, ML pipelines and steps, together with their data sources and models, libraries and SDKs, need to be stored and maintained such that they can be repeated exactly as previously.
- **Reusability**– To fit with principles of continuous delivery, the pipeline needs to be able to package and deliver models and code into production, both to training and target environments.
- **Manageability** – The ability to apply governance, linking changes to models and code to development activities (for example through sprints) and enabling managers to measure and oversee both progress and value delivery.
- **Automation** – As with DevOps, continuous integration and delivery require automation to assure rapid and repeatable pipelines, particularly when these are augmented by governance and testing (which can otherwise create a bottleneck).

16. Use a stepwise progression for model development and evaluation

1. Determine which machine learning algorithm is most suitable
2. Data preprocessing and feature engineering
3. Divide data into validation, testing, and training sets
4. Assess the efficacy of the model
5. Tune hyperparameters of the model
6. Check for bias
7. Document the process of developing the model

17. Package the Model and the Deployment

- Package the model along with its dependencies
- Enable version management and revert functions
- Constantly monitor the efficacy of the model
- Retrain the model at regular intervals
- Consider both concept drift and data drift
- Consider and revise the governance model

18. Consider More Than Accuracy for the Models

- Robustness
- Interpretability
- Performance
- Managability
- Scabality

19. Employ Explainable AI (XAI) Techniques

- Improve Decision-Making: Gain deeper insights into the rationale behind AI decisions, leading to better-informed and defensible choices.
- Identify Bias and Fairness: Unmask potential biases within AI models, enabling proactive mitigation and fairer outcomes.
- Reduce Risk and Compliance Concerns: Address regulatory requirements and mitigate risks by offering transparent and auditable AI decision-making processes.
- Foster Collaboration and User Acceptance: Enhance human-AI collaboration by providing stakeholders with a clear understanding of AI models and their limitations.
- Enable Error Detection and Debugging: Quickly identify and address errors within AI models through more transparent and interpretable decision-making processes.
- Demystify the "Black Box": Demystify the inner workings of complex AI models, making them more accessible and understandable for wider audiences.

20. Address Ethical Considerations

- Responsible Data Collection: Ensuring that data is gathered responsibly, ethically, and in a way that respects and protects individual rights
- Responsible Development: Taking into account how AI is developed and used and how ethical considerations, including fairness and privacy, are addressed
- Trustworthiness: Ensuring safety, security, and accuracy in AI systems
- Explainability: Making sure AI systems are transparent and explainable as possible so that decisions made by them are comprehendible and understandable
- Discrimination: Reducing the possibility of algorithms making biased decisions based on factors like race, gender, ethnicity, or religion
- Privacy: Protecting user data and ensuring user autonomy with regard to their own data

21. Plan for Data Drift

- Model decay/drift is inevitable
- Changes to the environment affect model input
- Implement statistical tests and anomaly detection algorithms for monitoring the performance of ML models over time
- Implement version control for both data and models to track changes and facilitate rollbacks to previous versions if model performance deteriorates due to data drift
- Regularly retrain ML models with new data to adapt to changes in the underlying data distribution

22. Support ML with Data Governance

- Ensuring Data Quality and Reliability
- Promoting Data Transparency and Explainability
- Enhancing Data Security and Privacy
- Facilitating Data Collaboration and Sharing
- Managing Data Lifecycle and Disposal
- Supporting Regulatory Compliance
- Improving Model Fairness
- Enhancing Trust and Adoptability

23. Plan to Deal with Big Change

- Organizations implementing ML have recognized the need to make significant changes
- People instinctually don't like change to begin with
- When you add artificial intelligence coming into the workplace that's going to even make the issue worse if you don't get ahead of it
- Demonstrate how it can help the company instead of having the fear grow with people thinking it's going to hinder them or even worse replace them

24. Stay Abreast of Industry Developments

Summary

- ML Opportunities are Everywhere
- Think Big and Produce Business Results
- Choose the ML Stack with Careful Consideration
- Focus More on the Data than the Algorithms
- Adopt MLOps Early
- Prepare the Organization for the Change

MCKNIGHT
CONSULTING GROUP

Strategies for Machine Learning Success

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444