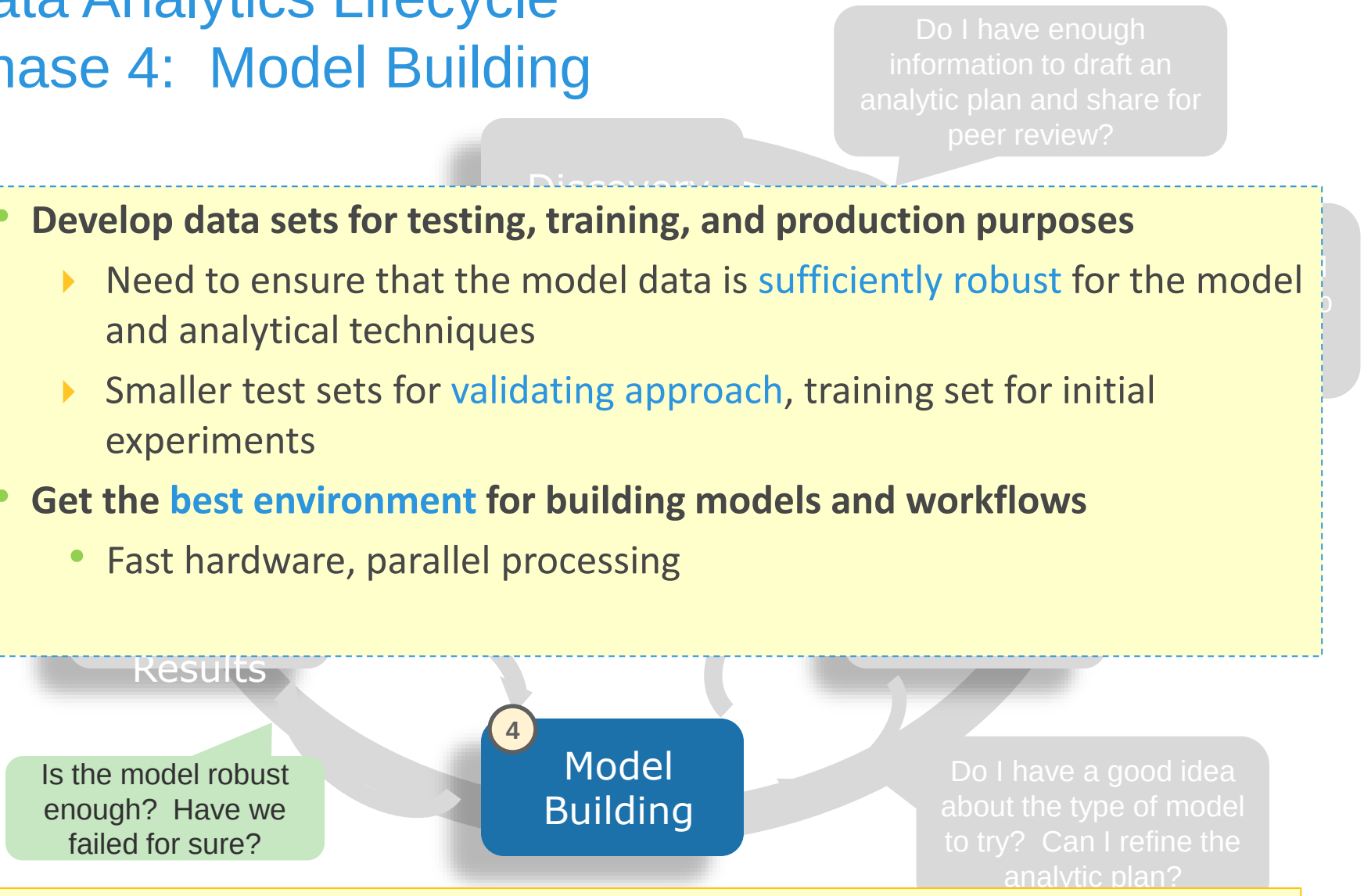
Module 2
Data Analytics Lifecycle

Sample Research: Churn Prediction in Other Verticals

- After conducting research on churn prediction, you have identified many methods for analyzing customer churn across multiple verticals.
- At this point, a Data Scientist would assess the methods and select the best model for the situation

Market Sector	Analytic Techniques/Methods Used
Wireless Telecom	DMEL method (data mining by evolutionary learning), Neural network, **decision tree**, hierarchical neurofuzzy systems, rule evolver, L**ogistic regression**.
Retail Business	**Logistic regression**, ARD (automatic relevance determination), **decision tree**
Daily Grocery	MLR (**multiple linear regression**), ARD, and **decision tree**
Retail Banking	**Multiple regression**

Data Analytics Lifecycle
Phase 4: Model Building



- **Develop data sets for testing, training, and production purposes**
  - Need to ensure that the model data is sufficiently robust for the model and analytical techniques
  - Smaller test sets for validating approach, training set for initial experiments
- **Get the best environment for building models and workflows**
  - Fast hardware, parallel processing

- **Useful Tools for this phase**: R, PL/R, SQL, Alpine Miner, SAS Enterprise Miner

Data Analytics Lifecycle
Phase 4: Model Building

Do I have enough information to draft an

- In this phase, the model is fit on the training data and evaluated (scored) against the test data.
- Generally this work takes place in the sandbox, not in the live production environment.
- The phases of Model Planning and Model Building overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model.
- Some methods require the use of a training data set, depending on whether it is a supervised or unsupervised algorithm for machine learning.
- Although the modeling techniques and logic required to build this step can be highly complex, the actual duration of this phase can be quite short, compared with all of the preparation required on the data and defining the approaches.
- In general, plan to spend more time preparing and learning the data (Phases 1-2) and crafting a presentation of the findings (Phase 5), where phases 3 and 4 tend to move more quickly, although more complex from a conceptual standpoint.

Data Analytics Lifecycle
Phase 4: Model Building

Do I have enough information to draft an

As part of this phase, you'll need to conduct these steps:

1) Execute the models defined in Phase 3
2) Where possible, convert the models to SQL or similar, appropriate database language and execute as in-database functions, since the runtime will be significantly faster (execute R models on large data sets as PL/R or SQL (PL/R is a PostgreSQL language extension that allows you to write PostgreSQL functions and aggregate functions in R).
3) Use R (or SAS) models on file extracts for testing and small data sets
4) Assess the validity of the model and its results (for instance, does it account for most of the data, and does it have robust predictive power?)
5) Fine tune the models to optimize the results (for example modify variable inputs)
6) Record the results, and logic of the model

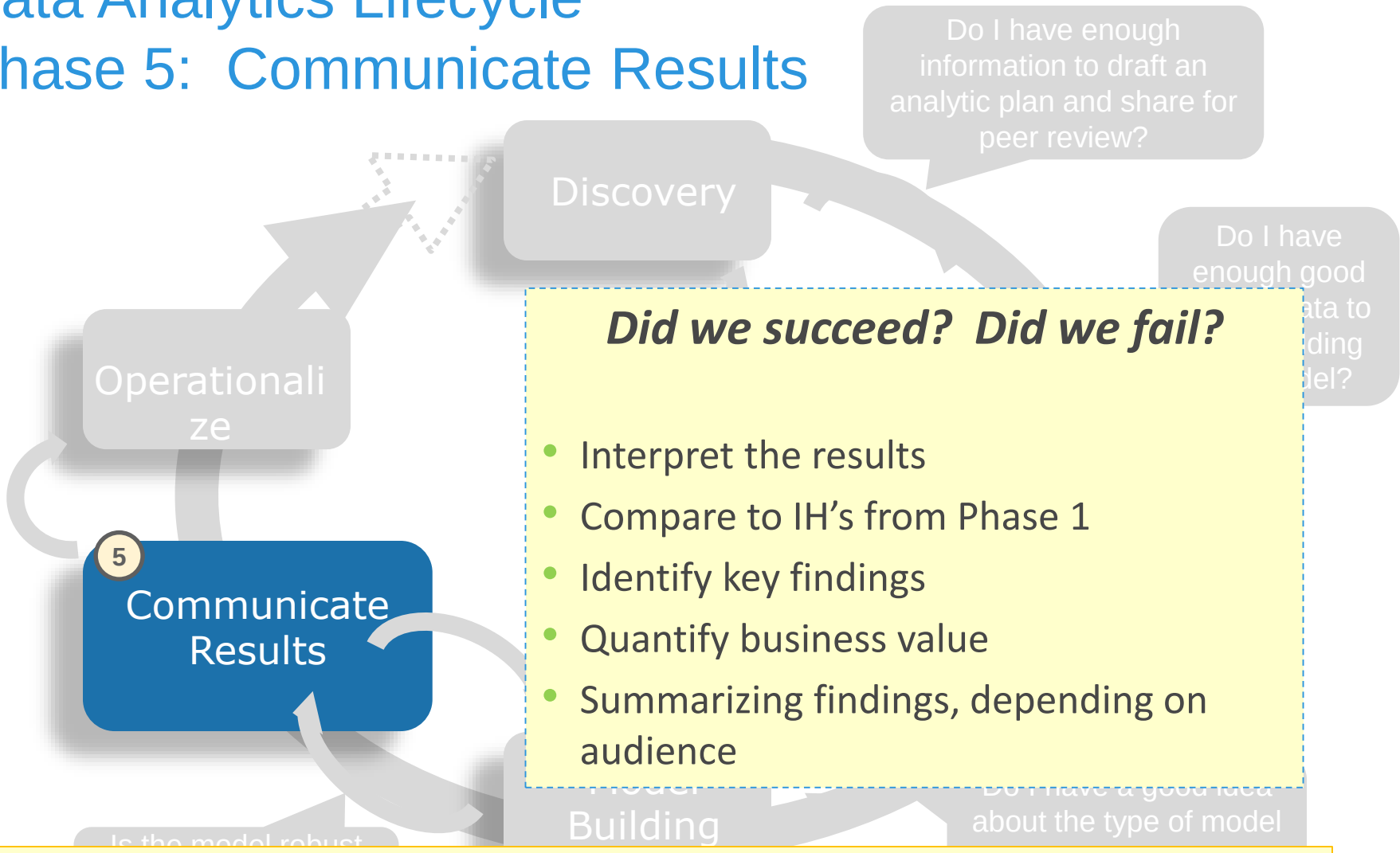
Data Analytics Lifecycle
Phase 4: Model Building

Do I have enough information to draft an

While doing these iterations and refinement of the model, consider the following:

- Does the model look valid and accurate on the test data?
- Does the model output/behavior makes sense to the domain experts? That is, does it look like the model is giving "the right answers", or answers that make sense in this context?
- Is the model accurate enough to meet the goal?
- Is it avoiding the kind of mistakes it needs to avoid? Depending on context, false positives may be more serious or less serious than false negatives, for instance.
- Do the parameter values of the fitted model make sense in the context of the domain?
- Do you need more data or more inputs? Do you need to transform or eliminate any of the inputs?
- Do you need a different form of model? If so, you'll need to go back to the Model Planning phase and revise your modeling approach.

Data Analytics Lifecycle
Phase 5: Communicate Results

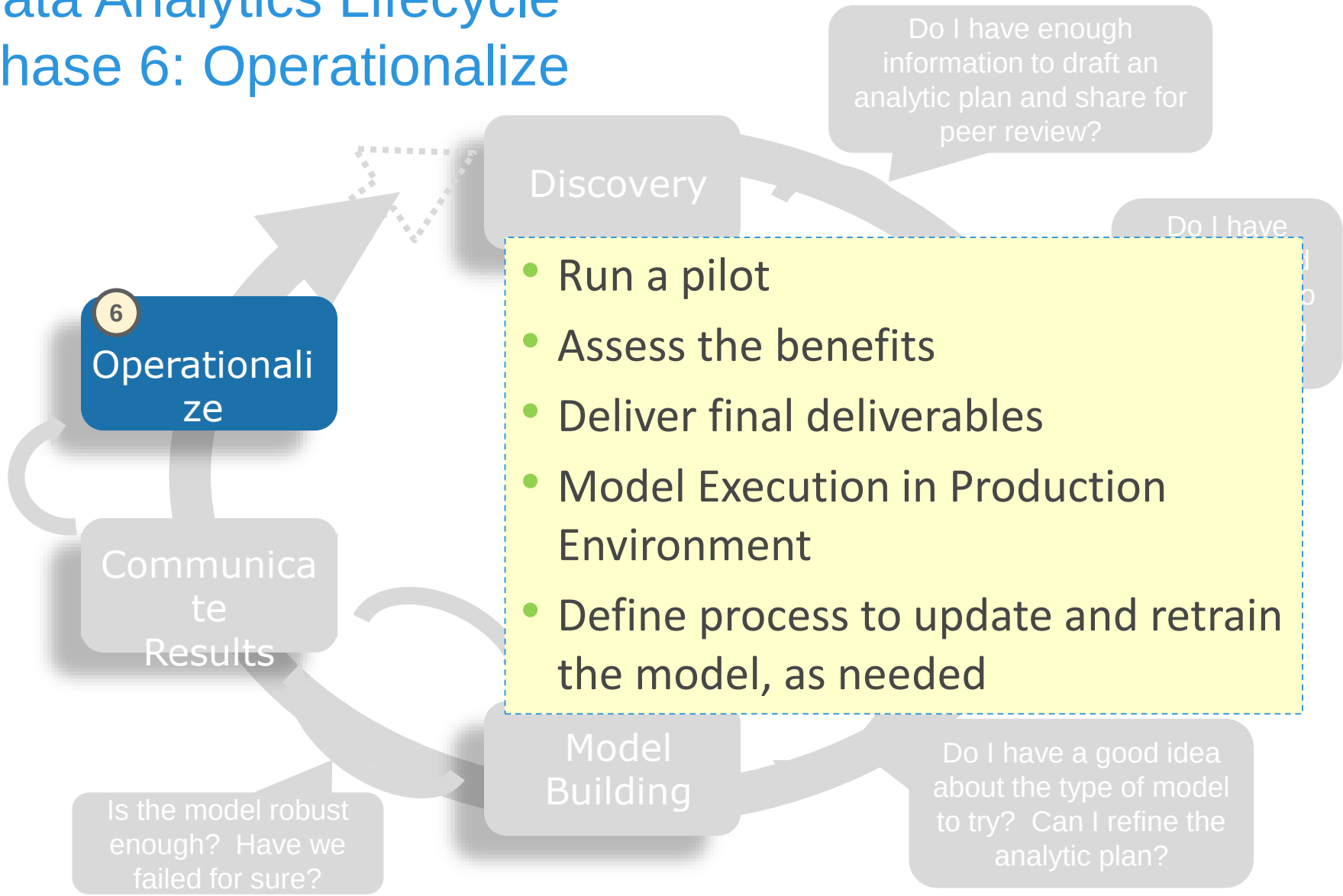


Did we succeed? Did we fail?

- Interpret the results
- Compare to IH's from Phase 1
- Identify key findings
- Quantify business value
- Summarizing findings, depending on audience

For the ABC Bank Case Study,
what would be some possible results and key findings?

Data Analytics Lifecycle
Phase 6: Operationalize



Data Analytics Lifecycle
Phase 6: Operationalize

Do I have enough information to draft an

- Run a pilot
- In phase 4, you scored the model in the sandbox, and
- In phase 6 represents the first time that most analytics approach deploying the new analytical methods or models in a production environment.
- Rather than deploying this on a wide scale basis, we recommend that you do a small scope, pilot deployment first.
- Taking this approach will allow you to limit the amount of risk relative to a full, enterprise deployment and learn about the performance and related constraints on a small scale and make fine tune adjustments before a full deployment.

Data Analytics Lifecycle
Phase 6: Operationalize

Do I have enough information to draft an

Consider running the model in a product environment for a discrete set of single products, or a single line of business, which will test your model in a live setting.

This will allow you learn from the deployment, and make any needed adjustments before launching across the enterprise.

Keep in mind that this phase can bring in a new set of team members – namely those engineers who are responsible for the production environment, who have a new set of issues and concerns.

They want to ensure that running the model fits smoothly into the production environment and the model can be integrated into downstream processes.

Data Analytics Lifecycle
Phase 6: Operationalize

Do I have enough information to draft an

After deploying the model, conduct follow up to reevaluate the model after it has been in production for a period of time.

Assess whether the model is meeting goals and expectations, and if desired changes (such as increase in revenue, reduction in churn) are actually occurring.

If these outcomes are not occurring, determine if this is due to a model inaccuracy, or if its predictions are not being acted on appropriately.

If needed, automate the retraining/updating of the model. In any case, you will need ongoing monitoring of model accuracy, and if accuracy degrades, you will need to retrain the model. If feasible, design alerts for when model is operating "out-of-bounds".

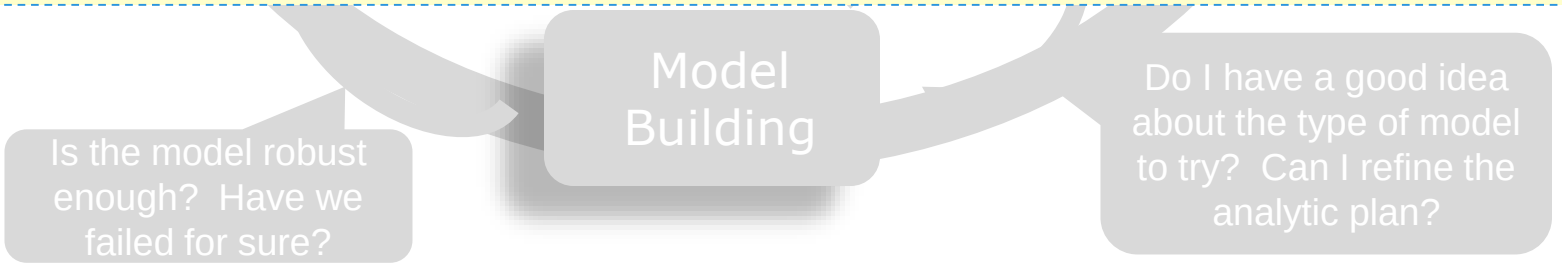
Data Analytics Lifecycle
Phase 6: Operationalize

If feasible, design alerts for when model is operating "out-of-bounds".

This includes situations when the inputs are far beyond the range that the model was trained on, which will cause the outputs of the model to be inaccurate.

If this begins to happen regularly, retraining is called for.

Many times analytical projects yield new insights about a business, a problem, or an idea that people may have taken at face value or thought was impossible to big into.



Analytic Plan

Components of Analytic Plan	Retail Banking: ABC Bank
Phase 1: Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates.
Data	5 months of customer account history.
Phase 3: Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn.
Phase 5: Result & Key Findings	Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save $3 million in lost of customer revenue and avoid $1.5 million in new customer acquisition costs each year.

Key Outputs from a Successful Analytic Project, by Role

Role	Description	What the Role Needs in the Final Deliverables

Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	• **Sponsor Presentation** addressing:
		  • Are the results good for me?
		  • What are the benefits of the findings?
		  • What are the implications of this for me?

Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	• **Sponsor Presentation** addressing:
		  • What's the business impact of doing this?
		  • What are the risks? ROI?
		  • How can this be evangelized within the organization (and beyond)?

Project Manager	Ensure key milestones and objectives are met on time and at expected quality.	
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective	• Show the **analyst presentation** • Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met	• Show the **analyst presentation** • Share the **code**

4 Core Deliverables to Meet Most Stakeholder Needs

1. **Presentation for Project Sponsors**
   - "Big picture" takeaways for executive level stakeholders
   - Determine key messages to aid their decision-making process
   - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp
2. **Presentation for Analysts**
   - Business process changes
   - Reporting changes
   - Fellow Data Scientists will want the details and are comfortable with technical graphs (such as ROC curves, density plots, histograms)
3. **Code** for technical people
4. **Technical specs** of implementing the code

Analyst Wish List for a Successful Analytics Project

Data & Workspaces

- Access to all the data, including aggregated OLAP data, BI tools, raw data, structured and various states of unstructured data as needed
- Up-to-date data dictionary to describe the data
- Area for staging and production data sets
- Ability to move data back and forth between workspaces and staging areas
- Analytic sandbox with strong compute power to experiment and play with the data

Tools

- Statistical/mathematical/visual software of choice for a given situation and problem set, such as SAS, Matlab, R, java tools, Tableau, Spotfire
- Collaboration: an online platform or environment for collaboration and communicating with team members
- Tool or place to log errors with systems, environments or data sets

Concepts in Practice
Greenplum's Approach to Analytics

Magnetic

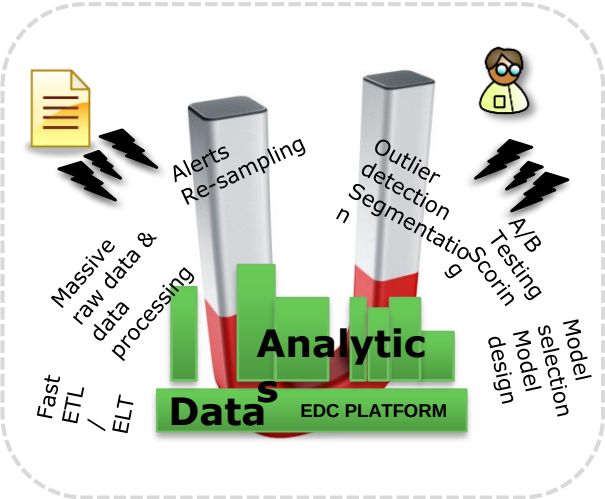
Attract all kinds of data

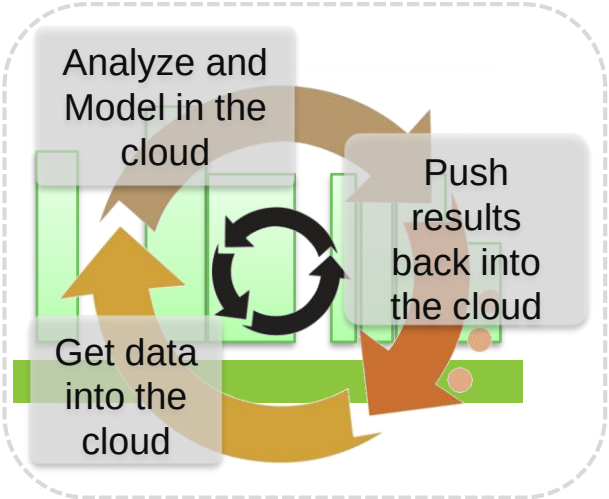
Agile

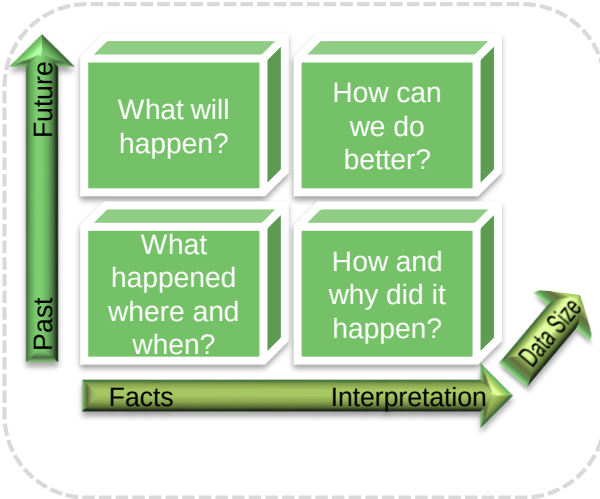
Flexible and elastic data structures

Deep

Rich data repository and algorithmic engine







Source: MAD Skills: New Analysis Practices for Big Data, March 2009