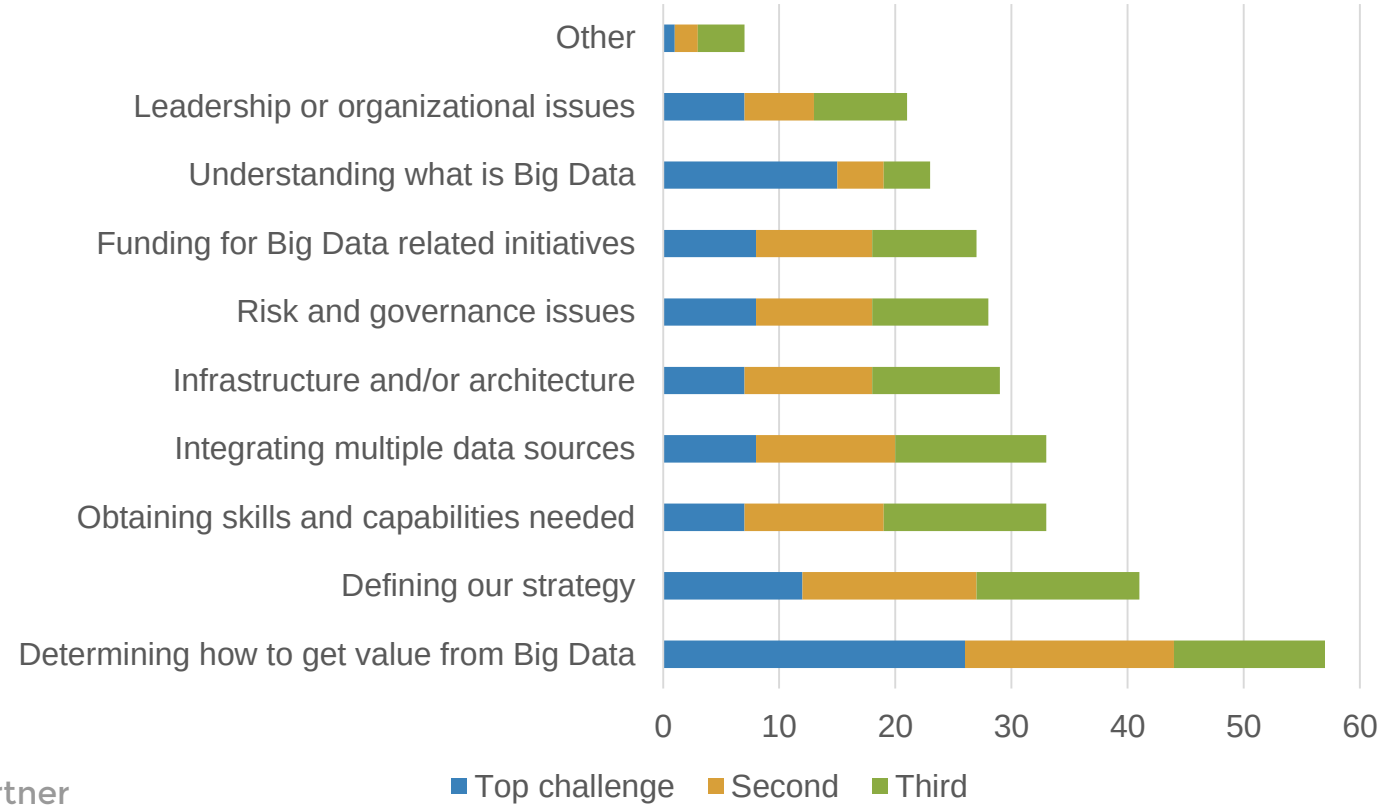
DATA SCIENCE PROJECT METHODOLOGY

Sergey Shelpuk
sergey@shelpuk.com

DATA SCIENCE IS ALL ABOUT **BUSINESS**

TOP BIG DATA CHALLENGES



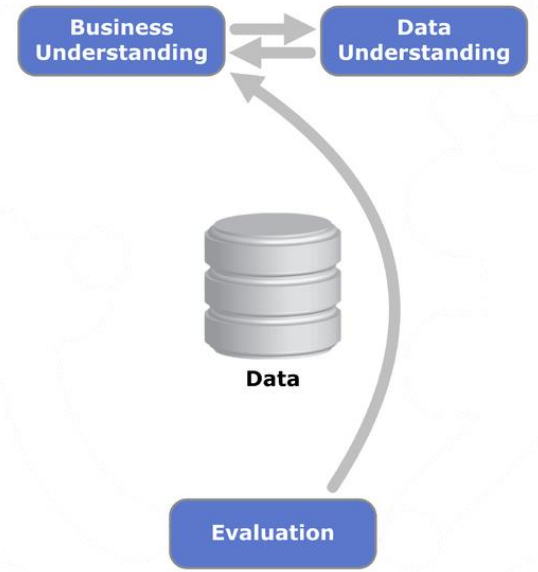
METHODOLOGY IS A KEY TO SUCCESS

Cross-Industry Standard Process for Data Mining (CRISP-DM)

BUSINESS UNDERSTANDING

Determining Business Objectives

1. Gather background information
   - Compiling the business background
   - Defining business objectives
   - Business success criteria
2. Assessing the situation
   - Resource Inventory
   - Requirements, Assumptions, and Constraints
   - Risks and Contingencies
   - Cost/Benefit Analysis
4. Determining data science goals
   - Data science goals
   - Data science success criteria
4. Producing a Project Plan



EXAMPLE OF THE PROJECT PLAN

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data scientists, DB engineers	Data problems, technology problems
Modeling	2 weeks	Data scientists	Technology problems, inability to build adequate model
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data scientist, DB engineers, implementation team	Economic change, inability to implement results

READY FOR THE DATA UNDERSTANDING?

From a business perspective:

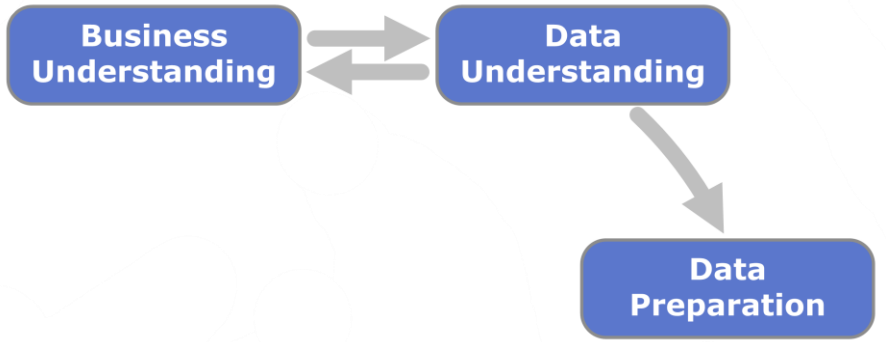
- ✓ What does your business hope to gain from this project?
- ✓ How will you define the successful completion of our efforts?
- ✓ Do you have the budget and resources needed to reach our goals?
- ✓ Do you have access to all the data needed for this project?
- ✓ Have you and your team discussed the risks and contingencies associated with this project?
- ✓ Do the results of your cost/benefit analysis make this project worthwhile?

From a data science perspective:

- ✓ How specifically can data mining help you meet your business goals?
- ✓ Do you have an idea about which data mining techniques might produce the best results?
- ✓ How will you know when your results are accurate or effective enough? (Have we set a measurement of data mining success?)
- ✓ How will the modeling results be deployed? Have you considered deployment in your project plan?
- ✓ Does the project plan include all phases of CRISP-DM?
- ✓ Are risks and dependencies called out in the plan?

DATA UNDERSTANDING

1. Collect initial data
   - Existing data
   - Purchased data
   - Additional data
2. Describe data
   - Amount of data
   - Value types
   - Coding schemes
3. Explore data
4. Verify data quality
   - Missing data
   - Data errors
   - Coding inconsistencies
   - Bad metadata

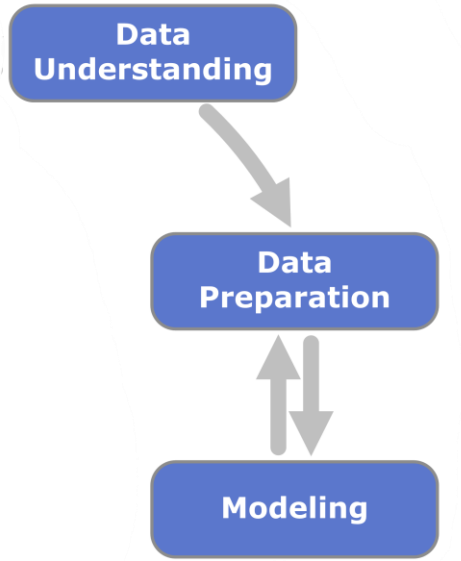


READY FOR THE DATA PREPARATION?

- ✓ Are all data sources clearly identified and accessed? Are you aware of any problems or restrictions?
- ✓ Have you identified key attributes from the available data?
- ✓ Did these attributes help you to formulate hypotheses?
- ✓ Have you noted the size of all data sources?
- ✓ Are you able to use a subset of data where appropriate?
- ✓ Have you computed basic statistics for each attribute of interest? Did meaningful information emerge?
- ✓ Did you use exploratory graphics to gain further insight into key attributes? Did this insight reshape any of your hypotheses?
- ✓ What are the data quality issues for this project? Do you have a plan to address these issues?
- ✓ Are the data preparation steps clear? For instance, do you know which data sources to merge and which attributes to filter or select?

DATA PREPARATION

1. Select right data
   - Select training examples
   - Select features
2. Clean data
   - Fill in missed data
   - Correct data errors
   - Make coding consistent
2. Extend data
   - Extend training examples
   - Extend features
2. Format data
   - Put data in a format for training the model

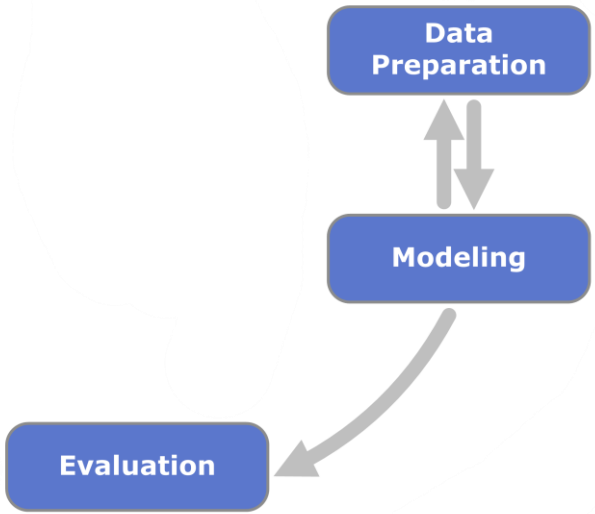


READY FOR THE MODELING?

- ✓ Based upon your initial exploration and understanding, were you able to select relevant subsets of data?
- ✓ Have you cleaned the data effectively or removed unsalvageable items? Document any decisions in the final report.
- ✓ Are multiple data sets integrated properly? Were there any merging problems that should be documented?
- ✓ Have you researched the requirements of the modeling tools that you plan to use?
- ✓ Are there any formatting issues you can address before modeling? This includes both required formatting concerns as well as tasks that may reduce modeling time.

MODELING

1. Select modeling techniques
   - Select data types available for analysis
   - Select an algorithm or a model
     Define modeling goals
   - State specific modeling requirements
2. Build the model
   - Set up hyperparameters
   - Train the model
   - Describe the result
3. Assess the model

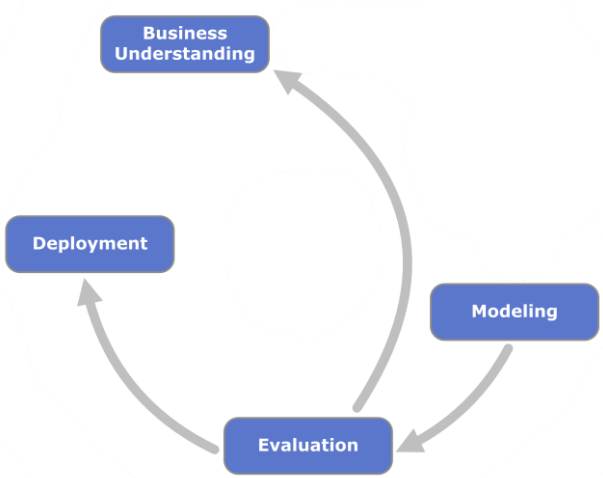


READY FOR THE EVALUATION?

- ✓ Are you able to understand the results of the models?
- ✓ Do the model results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration?
- ✓ From your initial glance, do the results seem to address your organization's business question?
- ✓ Have you used analysis nodes and lift or gains charts to compare and evaluate model accuracy?
- ✓ Have you explored more than one type of model and compared the results?
- ✓ Are the results of your model deployable?

EVALUATION

1. Evaluate the **results**
   - Are results presented clearly?
   - Are there any novel findings?
   - Can models and findings be applicable to business goals?
   - How well do the models and findings answer business goals?
   - What additional questions the modeling results have risen?
2. Review the **process**
   - Did the stage contribute to the value of the results?
   - What went wrong and how it can be fixed?
   - Are there alternative decisions which could have been executed?
2. Determine the **next steps**



DEPLOYMENT

1. Planning for deployment
   - Summarize models and findings
   - For each model create a deployment plan
   - Identify any deployment problems and plan for contingencies
2. Plan monitoring and maintenance
   - Identify models and findings which require support
   - How can the accuracy and validity be evaluated?
   - How will you determine that a model has expired?
   - What to do with the expired models?
2. Conduct a final project review

THANK YOU


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data