



HOW TO BUILD AND DEPLOY MACHINE LEARNING PROJECTS

Litan Ilany, Advanced Analytics

litan.ilany@intel.com

AGENDA

- Introduction
- Machine Learning: Exploration vs Solution
- CRISP-DM
- Data Flow considerations
- Other key considerations
- Q&A

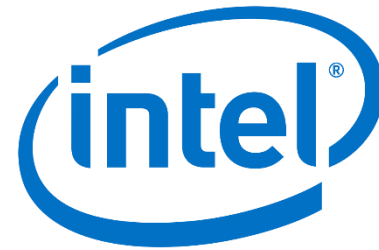
INTRODUCTION – LITAN ILANY

litan.ilany@intel.com



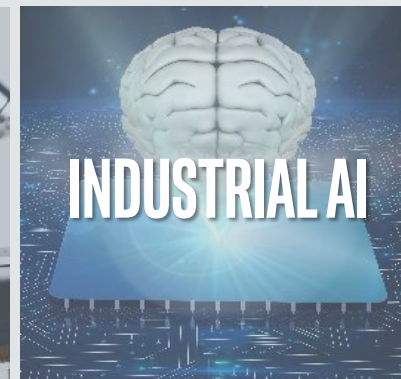
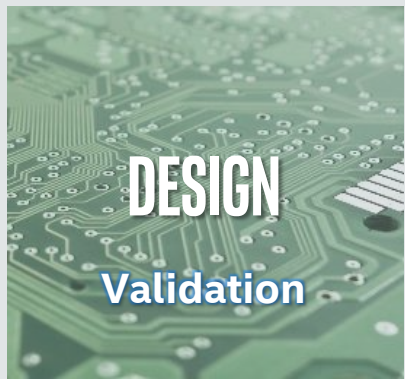
- Data Scientist Leader at Intel's Advanced Analytics team.
- Owns a M.Sc. degree in Information-Systems Engineering at BGU (focused on Machine-Learning and Reinforcement-Learning)
- Married + 2, Live in Kiryat Motzkin

ADVANCED ANALYTICS TEAM



RADICAL IMPROVEMENT OF CRITICAL PROCESSES

HELP BUILDING AI COMPETITIVE PRODUCTS



BREAKTHROUGH TECHNOLOGY THAT SCALES

MACHINE LEARNING

- Statistics
- Pattern recognition
- Generalization / Inductive Inference
- Types of learning:
 - Supervised vs Unsupervised Learning
 - Passive vs Active & Reinforcement Learning
 - Batch vs Online Learning



ML – ALGORITHM VS SOLUTION

- “Given a data matrix...” – does not exist in real life
- Pareto Principle (80/20 rule)
 - Technical aspects
 - Business needs
 - Extreme cases

ML PROJECT - GO / NO-GO DECISION

BUSINESS FEASIBILITY

Problem definition
is clear

Partner willing to
invest / change

Enough ROI / impact

DATA FEASIBILITY

Data measures what
they care about
("signal")


Enough accessible &
connected data

Data is accurate 

EXECUTION FEASIBILITY

Technology is
accessible

Model can be
executed in a timely
manner and size

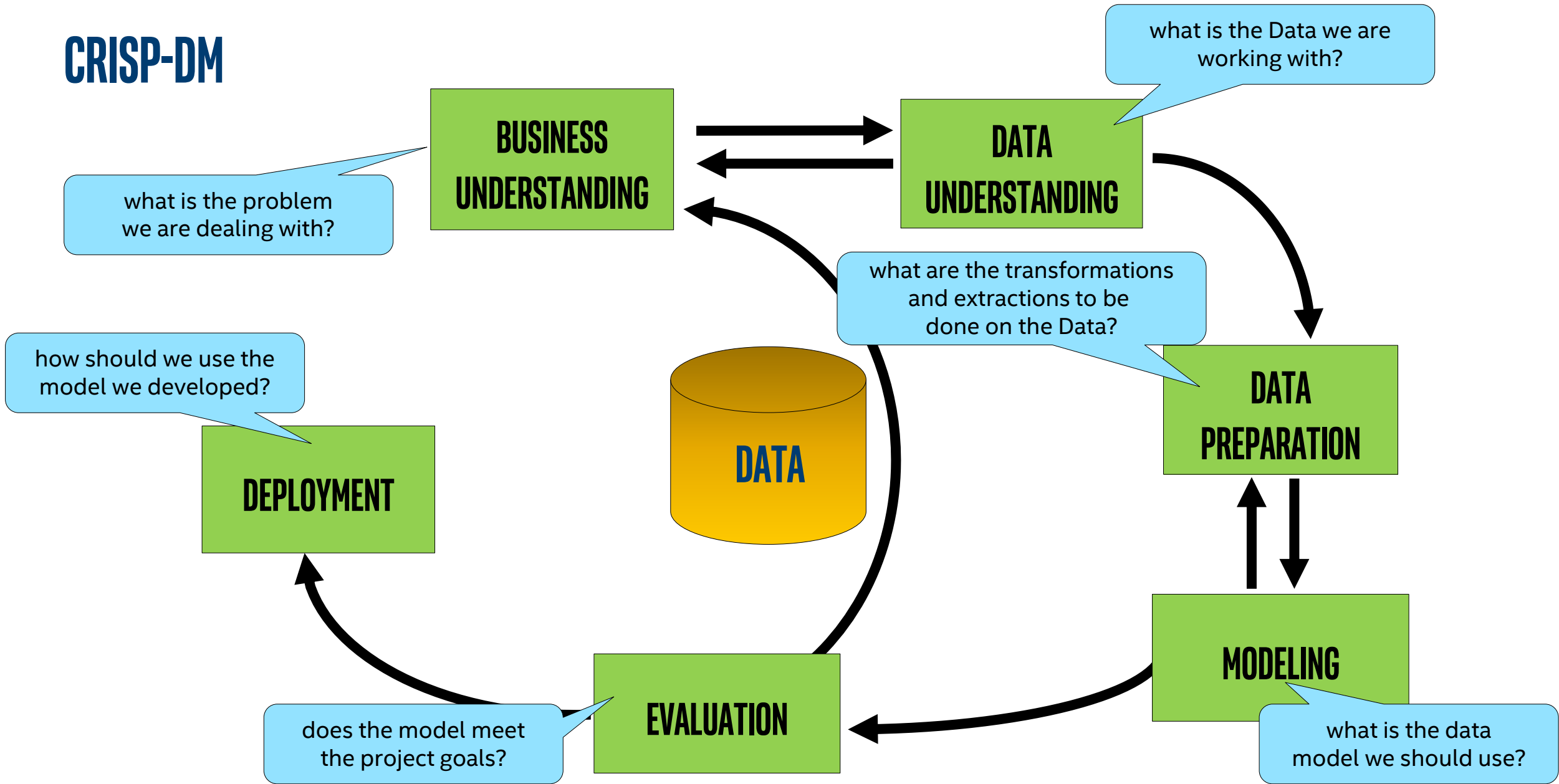
Model's I/O flow is
reasonable 

CRISP-DM

Cross-Industry Standard Process for Data Mining

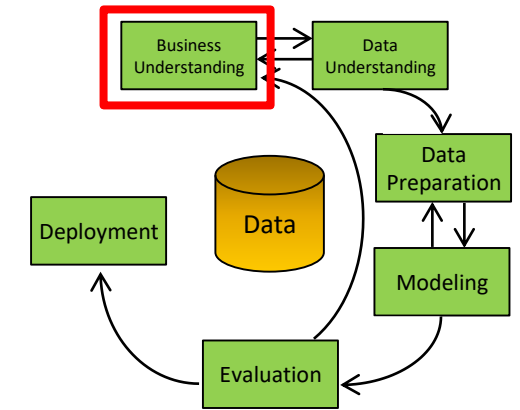
- A structured methodology for DM projects
- Based on practical, real-world experience
- Conceived in 1996-7

CRISP-DM



CRISP-DM: BUSINESS UNDERSTANDING

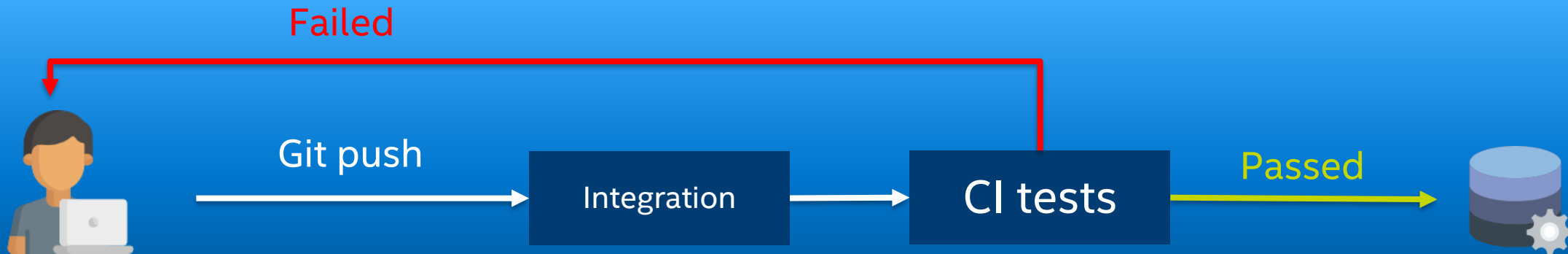
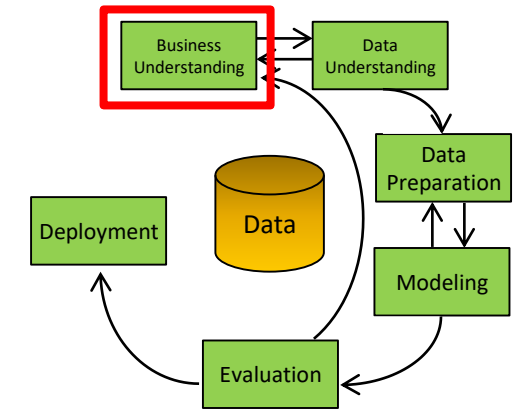
- Determine business objective
- Assess situation
- Determine data mining goals and success criteria
- Determine project plan



CRISP-DM: BUSINESS UNDERSTANDING - EXAMPLE

Example: Smart CI

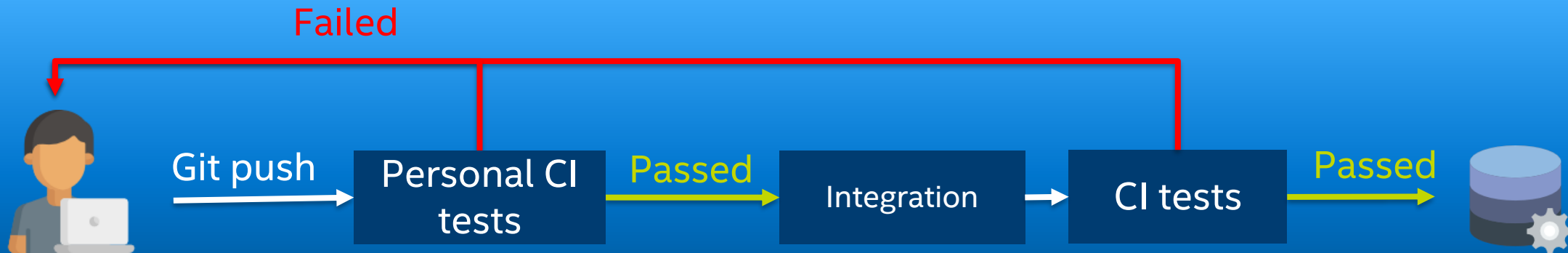
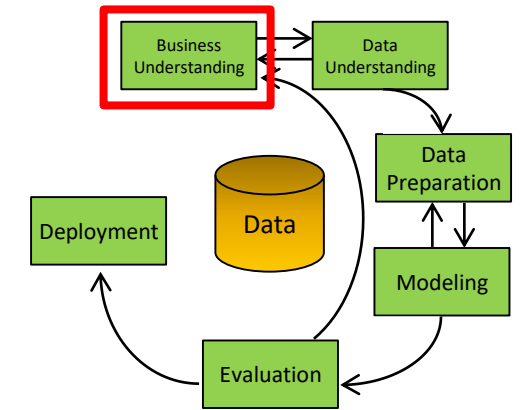
- Each git-push is integrated with the main repository – after tests series passes
- Multi git-push (can't check one-by-one)
- Bug in code causes entire integration to fail



CRISP-DM: BUSINESS UNDERSTANDING - EXAMPLE

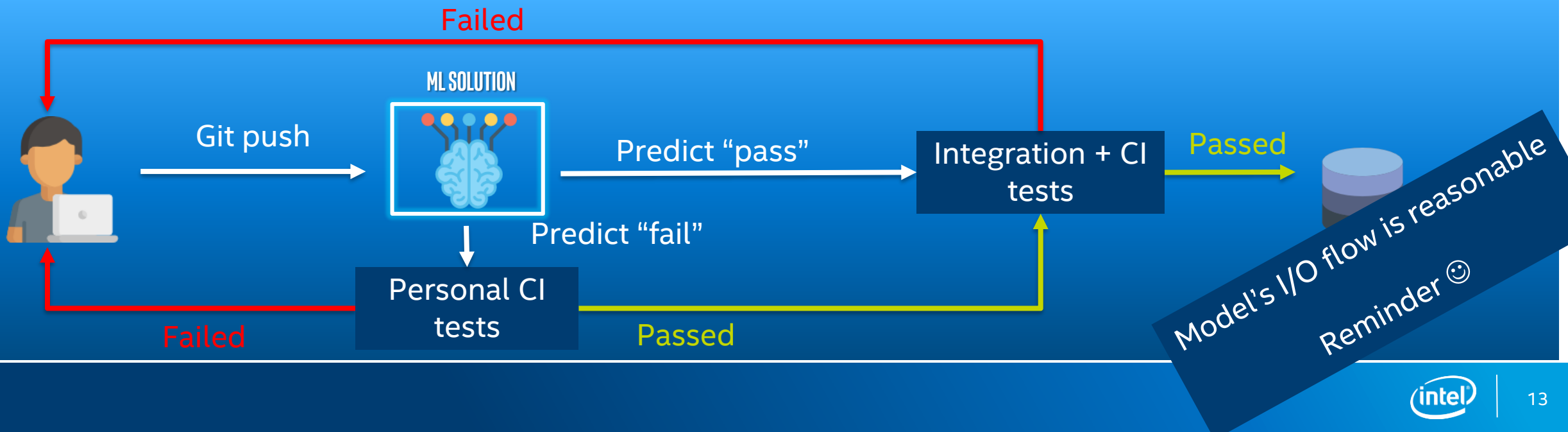
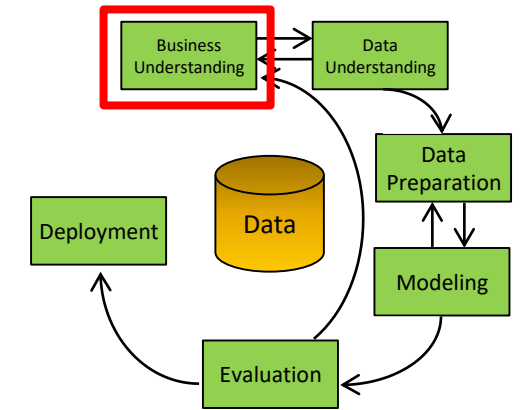
Example: Smart CI

- Each git-push is integrated with the main repository – after tests series passes
- Multi git-push (can't check one-by-one)
- Bug in code causes entire integration to fail



CRISP-DM: BUSINESS UNDERSTANDING - EXAMPLE

- Goals and success criteria:
 - Reduce Turnaround Time (TAT)
 - At least 20% time reduction
- Project plan

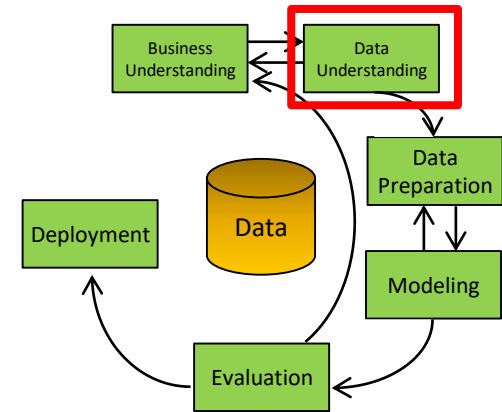


CRISP-DM: DATA UNDERSTANDING

- Collect initial data
- Describe data
- Explore data
- Verify data quality

Example:

- Git-log files (unstructured data):
 - Commits – numerical / binary
 - Files, Folders – numerical / binary
 - Lines – numerical
- Git DB (structured data):
 - Users – categorical
 - Timestamps, etc.
- Historical tests results (labels)

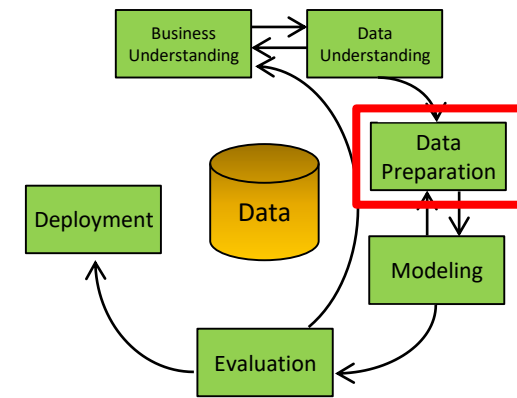


CRISP-DM: DATA PREPARATION

- Integrate data from multi sources
- Format data
- Feature extraction
- Clean data
- Construct data
 - Derive attributes – transformation
 - Reduce imbalance data
 - Fill in missing values
- Feature selection

Example:

- Generate features from log
- Generate and clean user-features
- Normalize counters
- Thousands of features – remove unnecessary ones
- Data balancing (if needed)

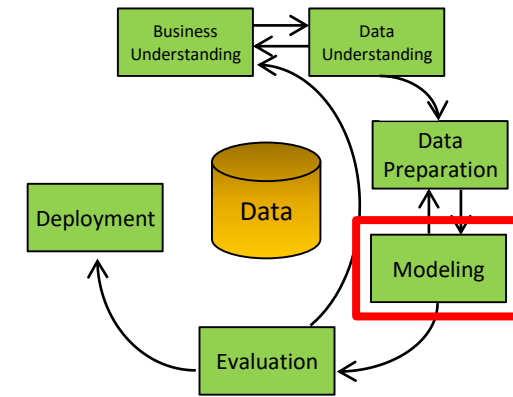


CRISP-DM: MODELING

- Select modeling technique
 - Consider computer resources, computation time, number of features, business needs
- Generate test design
 - Train/Test split, Cross validation
 - Simulation (chronological order)
- Build model
- Assess model

Example:

- We'll check various ML models with various hyperparameters
- Simulation, weekly training phase



CRISP-DM: MODELING – EXAMPLE (SMART CI)

- Model assessment:
 - Which model to choose?
 - How can we measure it?

Model A

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	55	18	73
Predicted fail	15	12	27
Total	70	30	100

Model B

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	35	5	40
Predicted fail	35	25	60
Total	70	30	100

CRISP-DM: MODELING – EXAMPLE (SMART CI)

- Model assessment:
 - Which model to choose?
 - How can we measure it?

Measure	A	B
Accuracy	$(55+12)/100 = 67\%$	$(35+25)/100 = 60\%$

Model A

Model B

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	55	18	73
Predicted fail	15	12	27
Total	70	30	100

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	35	5	40
Predicted fail	35	25	60
Total	70	30	100

CRISP-DM: MODELING – EXAMPLE (SMART CI)

- Model assessment:
 - Which model to choose?
 - How can we measure it?

Measure	A	B
Accuracy	$(55+12)/100 = 67\%$	$(35+25)/100 = 60\%$
Precision	$55/73 = 75\%$	$35/40 = 87\%$

Model A

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	55	18	73
Predicted fail	15	12	27
Total	70	30	100

Model B

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	35	5	40
Predicted fail	35	25	60
Total	70	30	100

CRISP-DM: MODELING – EXAMPLE (SMART CI)

- Model assessment:
 - Which model to choose?
 - How can we measure it?

Measure	A	B
Accuracy	$(55+12)/100 = 67\%$	$(35+25)/100 = 60\%$
Precision	$55/73 = 75\%$	$35/40 = 87\%$
Recall	$55/70 = 76\%$	$35/70 = 50\%$

Model A

Model B

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	55	18	73
Predicted fail	15	12	27
Total	70	30	100

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	35	5	40
Predicted fail	35	25	60
Total	70	30	100

CRISP-DM: MODELING – EXAMPLE (SMART CI)

- Model assessment:
 - Which model to choose?
 - How can we measure it?

Measure	A	B
Accuracy	$(55+12)/100 = 67\%$	$(35+25)/100 = 60\%$
Precision	$55/73 = 75\%$	$35/40 = 87\%$
Recall	$55/70 = 76\%$	$35/70 = 50\%$
FPR*	$18/30 = 60\%$	$5/30 = 17\%$

*Lower is better

Model A

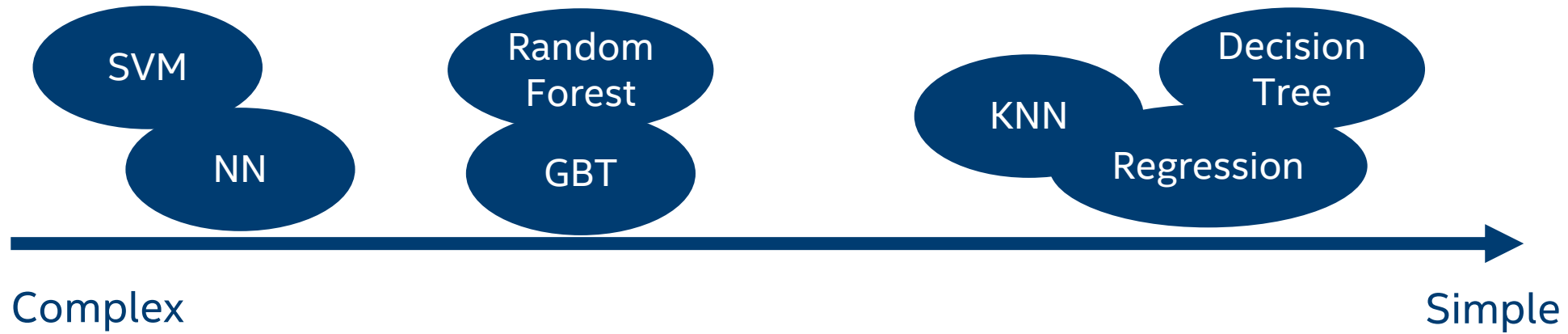
Model B

Predicted \ Actual	push Passed	push failed	Total
Predicted pass	55	18	73
Predicted fail	15	12	27
Total	70	30	100

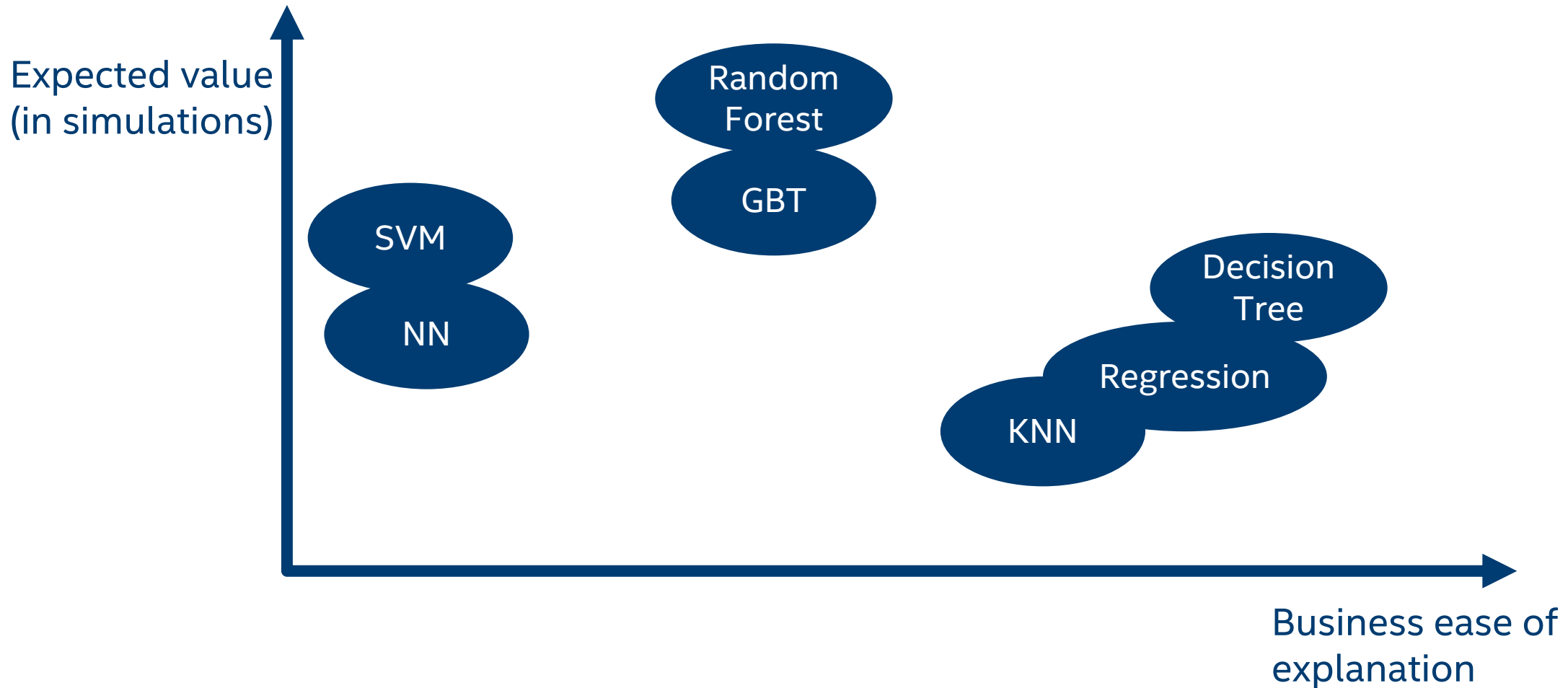
Predicted \ Actual	push Passed	push failed	Total
Predicted pass	35	5	40
Predicted fail	35	25	60
Total	70	30	100

CRISP-DM: MODELING – EXAMPLE (SMART CI)

Business ease of explanation



CRISP-DM: MODELING – EXAMPLE (SMART CI)



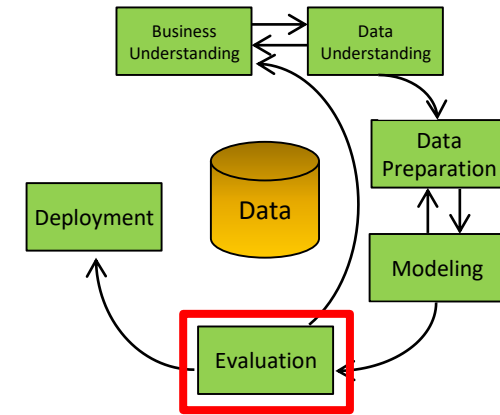
CRISP-DM: EVALUATION

- Evaluate results
 - In terms of business needs
- Review Process
- Determine next steps

Example:

Predicted \ Actual	push Passed	push failed
Predicted pass	TP	FP
Predicted fail	FN	TN

- TAT reduction:
 - TP = 50% reduction (X2 faster)
 - FN = 0% reduction
 - FP = -500-5000% reduction (X5-50 slower)

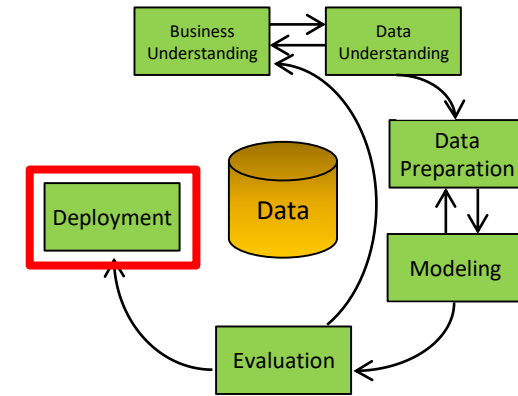


CRISP-DM: DEPLOYMENT

- Plan and deploy the model
- Plan monitoring and maintenance process

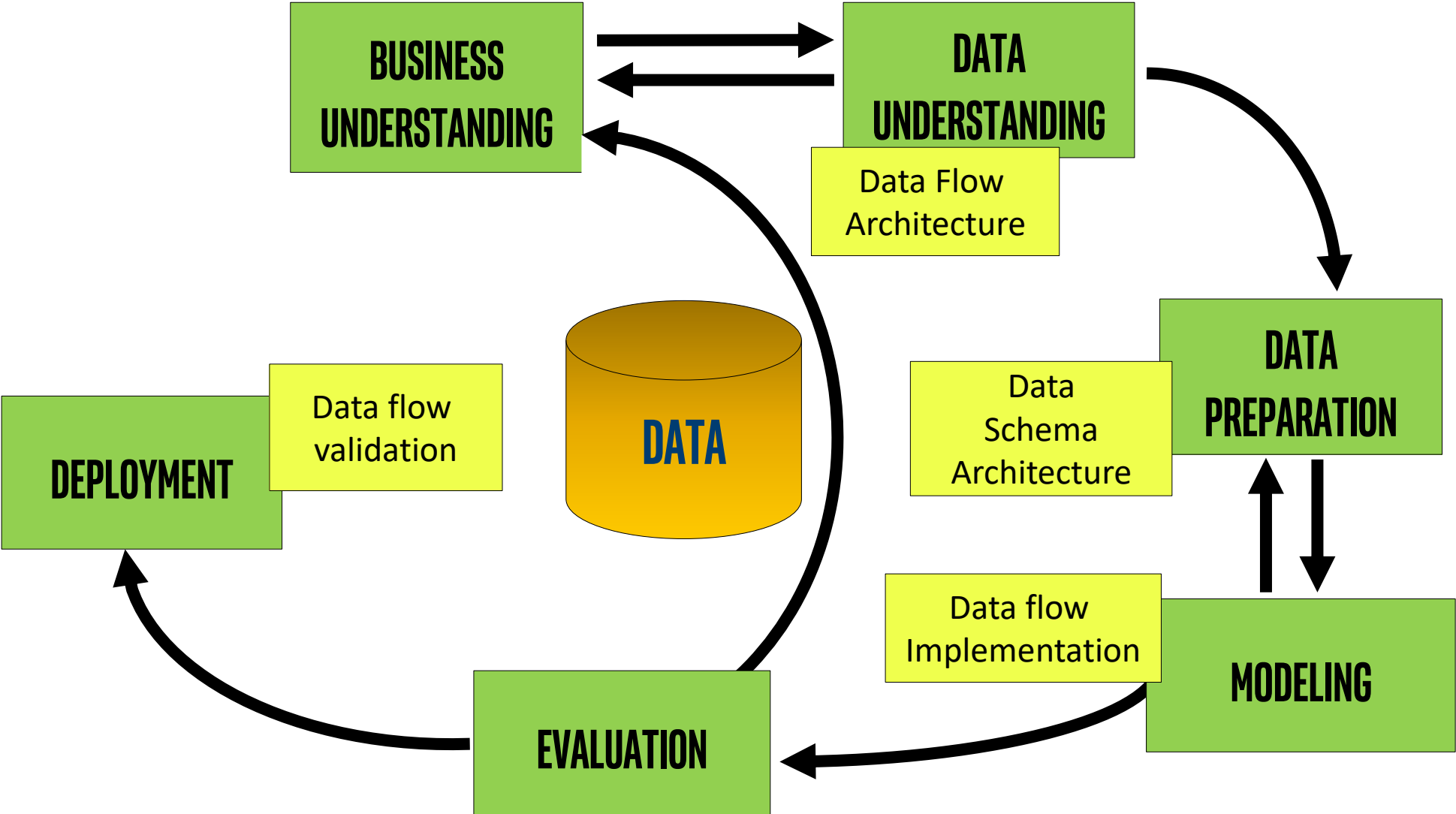
Example:

- Integrate with existing CI system
- Weekly automatic process that will train the model
- Weekly automatic process that will monitor the model's performance and suggest better hyper parameters (if needed)



Data is Accurate
Reminder 😊

CRISP-DM: DATA FLOW



OTHER KEY CONSIDERATIONS

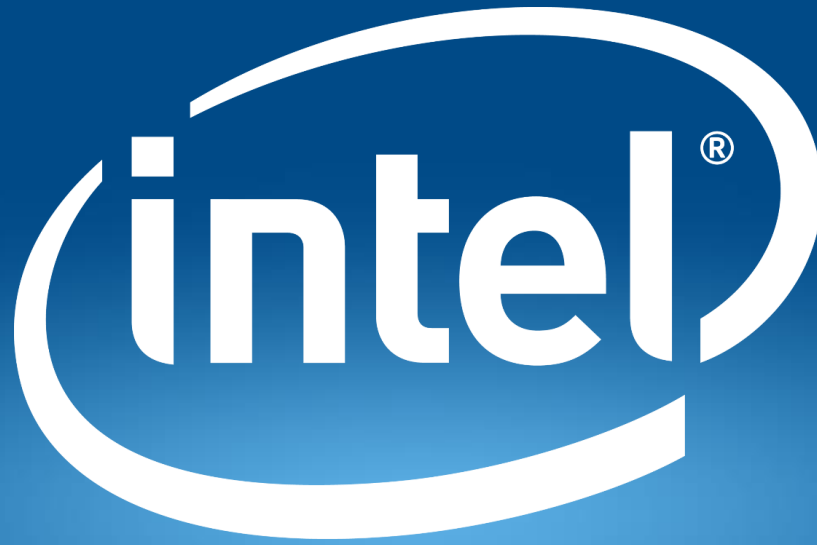
- Use Git (or other version control platform)
- Automate the research process (trial-and-error)
- Use Docker containers
- TEST YOUR CODE (don't think of it as black box)
- ML Technical Debt – code and data

REFERENCES

[CRISP-DM \(Wikipedia\)](#)

[4 things DSs should learn from software engineers](#)

[Machine Learning: The High Interest Credit Card of Technical Debt](#)



experience
what's inside™