
Data
Analysis
Idea

Some Experience

Data analysis

ML with AWS

Deep learning

Computer vision

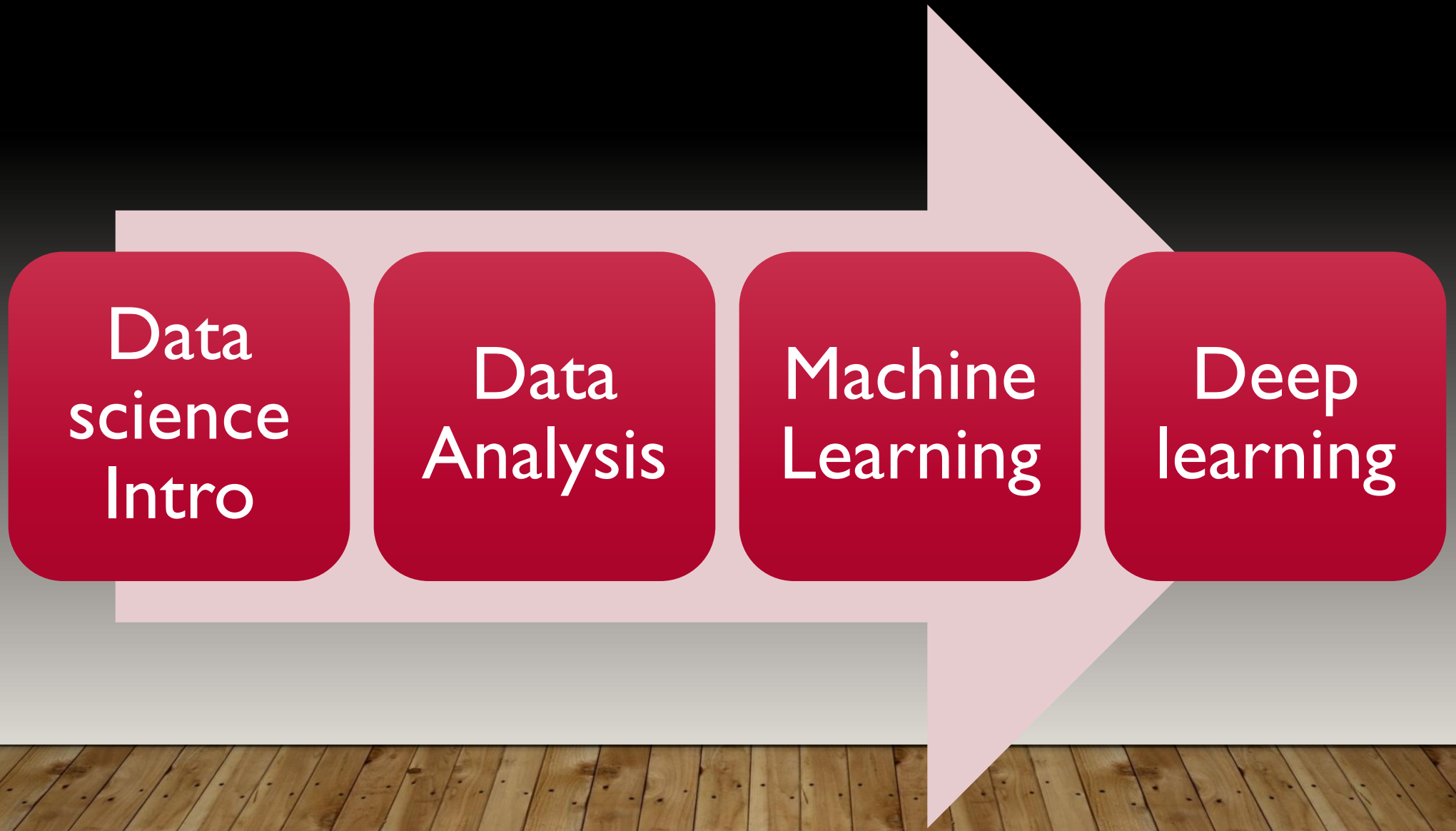
AI Analysis training

Machine learning engineer

Data Science
Squad mentor

Now it’s your turn

Agenda for all days



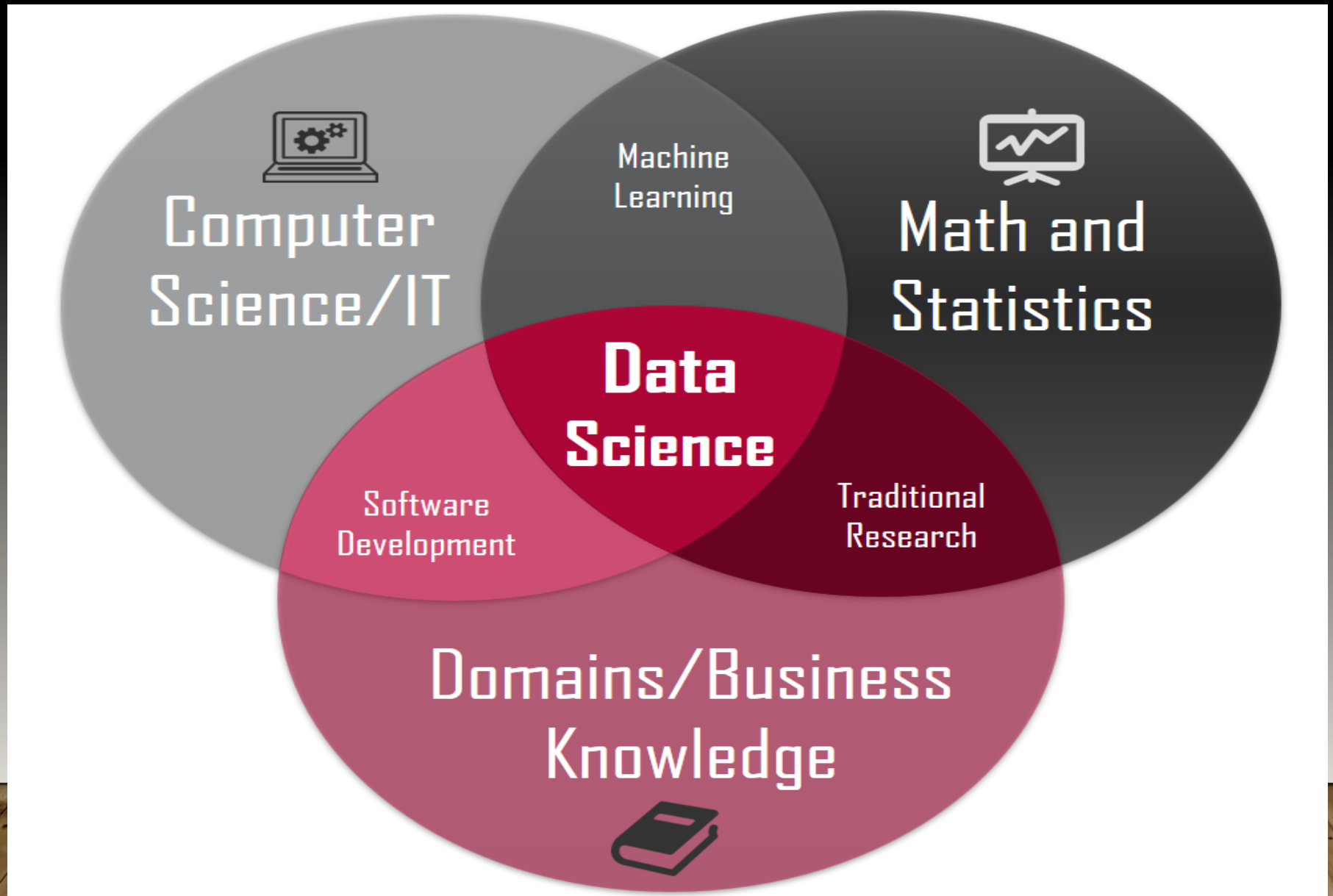
Agenda for Day one

- About Data science
- Projects Lifecyle
- Data in real world

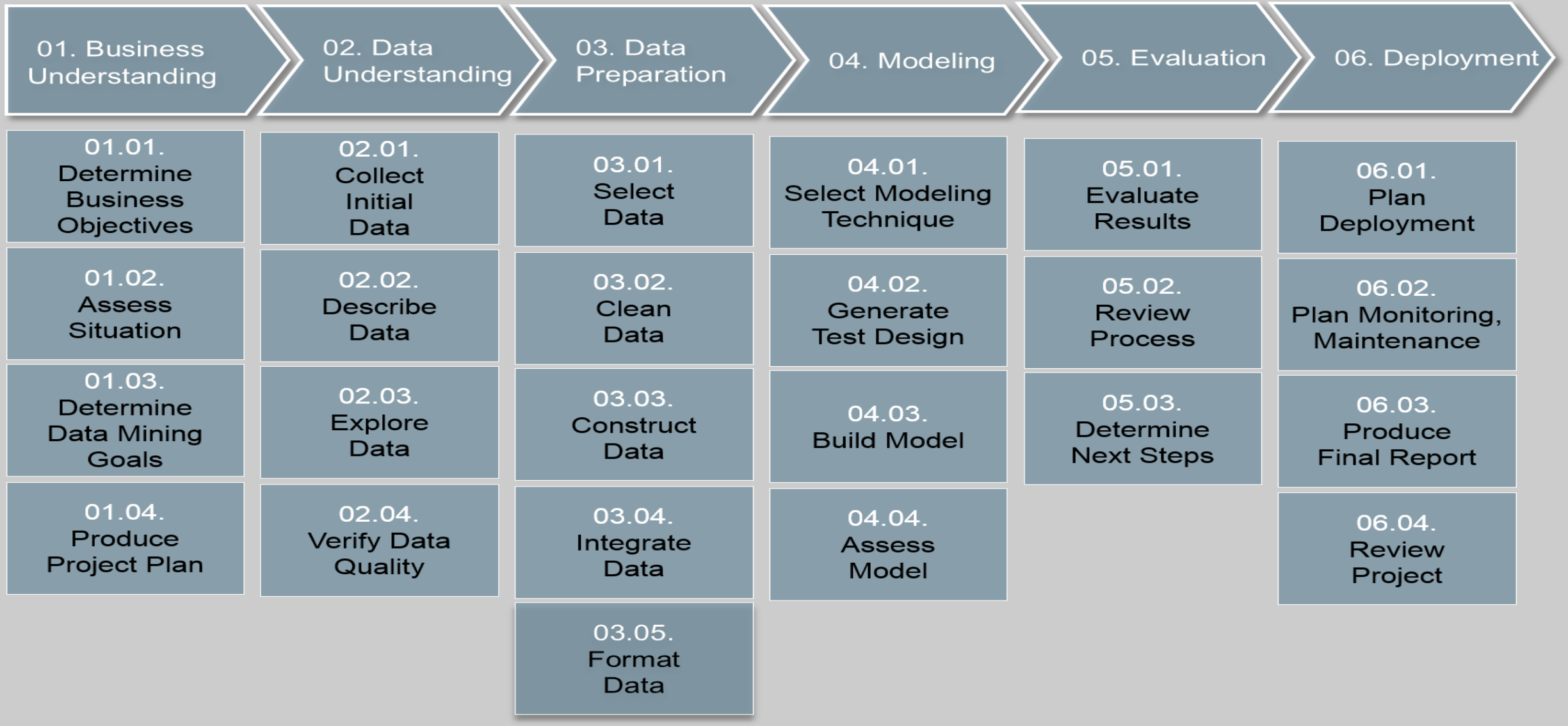
Why Data Science?

Amount of data	**A fuel of 21st Century**
High-Performance Computing	**An Answer to Complex Data**
Problem of Demand & Supply	**there aren't enough resources to convert this data into useful products.**

What is data science?



Data Science project life cycle




Data Collection
Methods
01 Forms and Questionnaires
02 Interview
03 Observation
04 Documents and Records
05 Focus Groups
06 Oral Histories
07 Combination Research
08 Online Tracking
09 Online Marketing Analytics
10 Social Media Monitoring
GlobalPatron

Data preparation

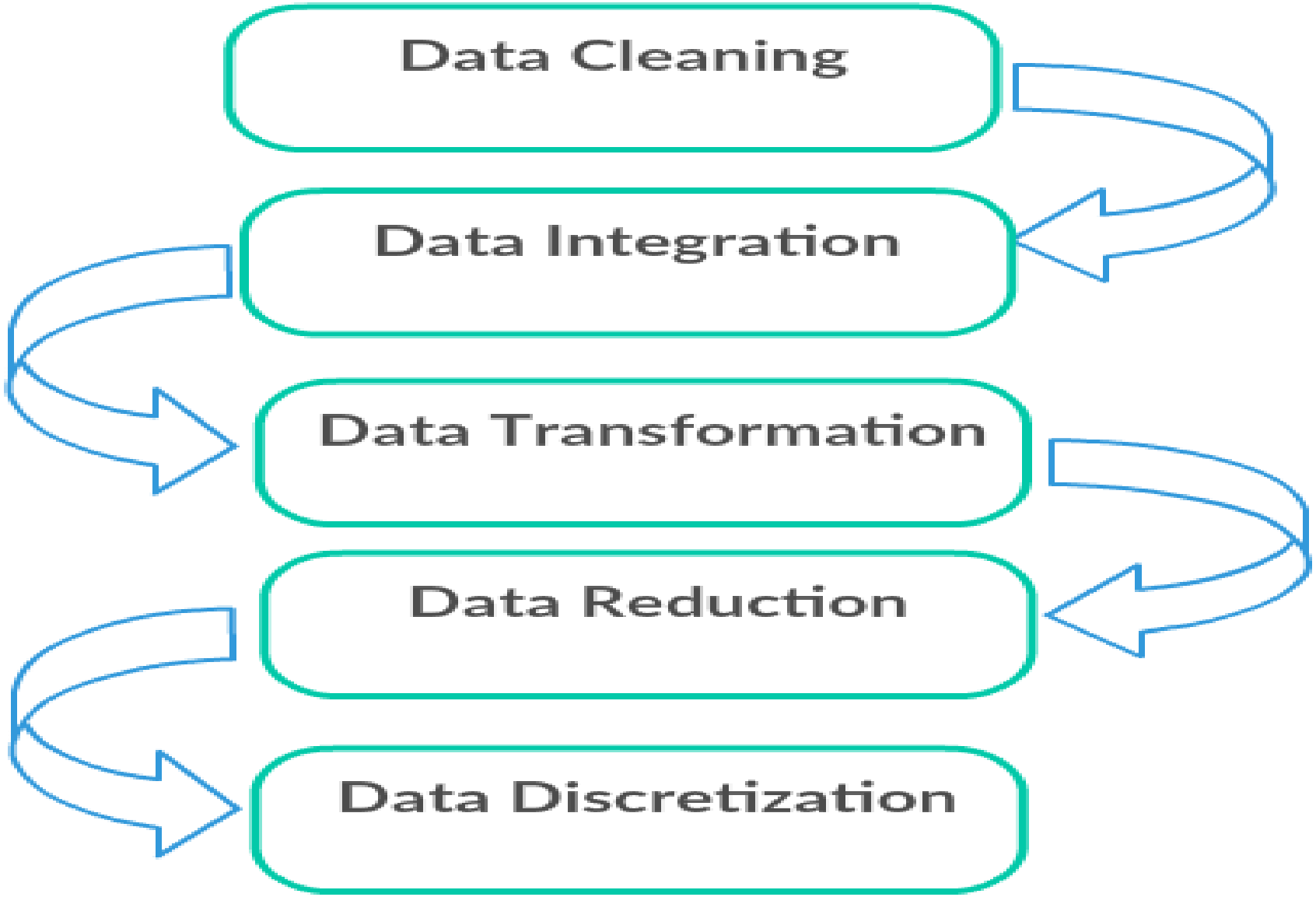
Data selection

Data preprocessing

Data transformation

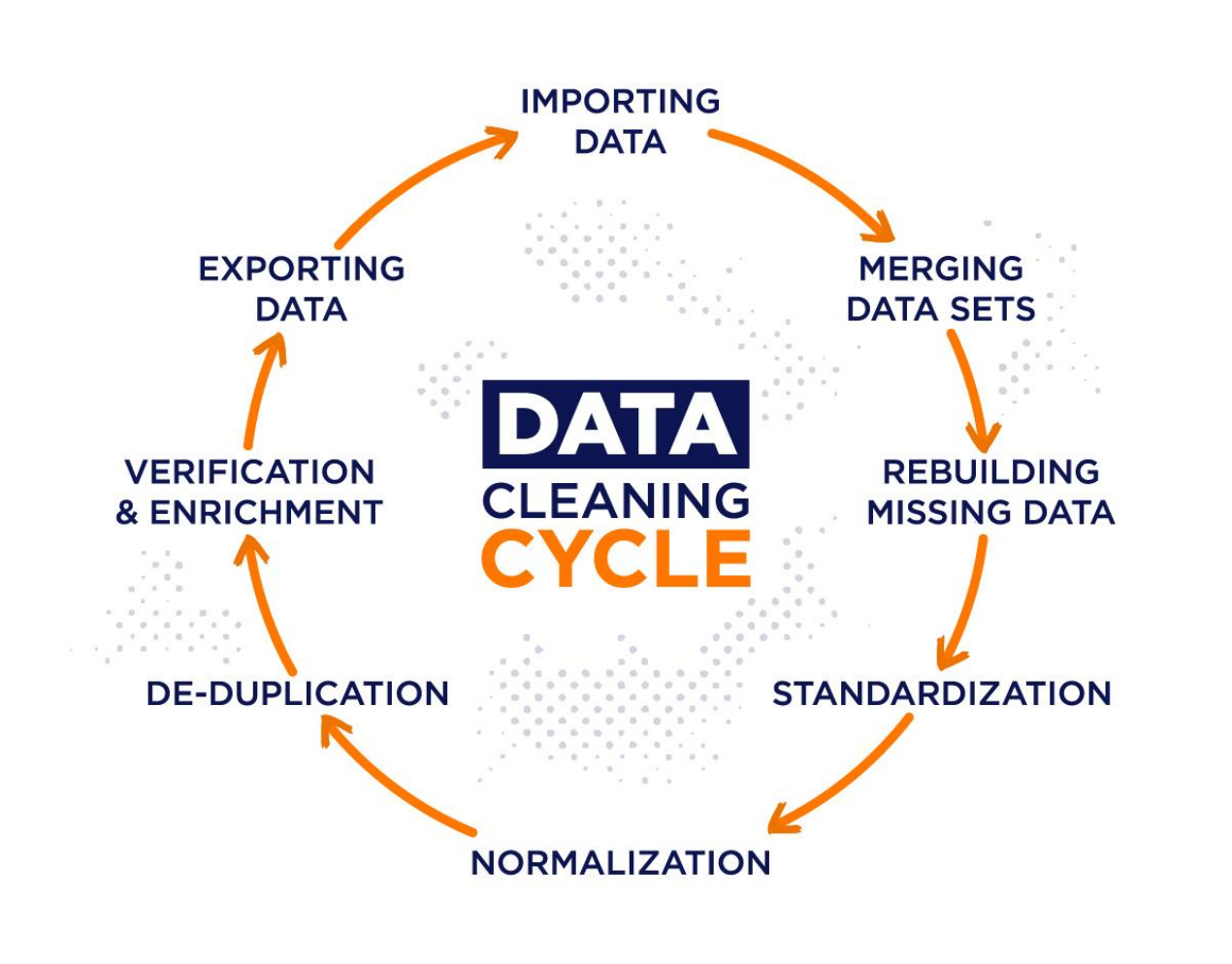
Machine learning algorithms depend highly on **the quality and quantity** of data. You must provide these algorithms with the **correct data**. Data preparation is a large subject that can involve many iterations, exploration, and analysis. **Becoming proficient at data preparation will make you a master at machine learning.**

Data preprocessing: Steps overview



Data cleaning

Complete missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies



Merging data

	group
employee	
Bob	Accounting
Jake	Engineering
Lisa	Engineering
Sue	HR

	hire_date
employee	
Lisa	2004
Bob	2008
Jake	2012
Sue	2014

	group	hire_date
employee		
Bob	Accounting	2008
Jake	Engineering	2012
Lisa	Engineering	2004
Sue	HR	2014

Missing values & un formatted data

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
20	Lois	NaN	4/22/1995	7:18 PM	64714	4.934	True	Legal
22	Joshua	NaN	3/8/2012	1:58 AM	90816	18.816	True	Client Services
27	Scott	NaN	7/11/1991	6:58 PM	122367	5.218	False	Legal
31	Joyce	NaN	2/20/2005	2:40 PM	88657	12.752	False	Product
41	Christine	NaN	6/28/2015	1:08 AM	66582	11.308	True	Business Development
49	Chris	NaN	1/24/1980	12:13 PM	113590	3.055	False	Sales
51	NaN	NaN	12/17/2011	8:29 AM	41126	14.009	NaN	Sales
53	Alan	NaN	3/3/2014	1:28 PM	40341	17.578	True	Finance
60	Paula	NaN	11/23/2005	2:01 PM	48866	4.271	False	Distribution
64	Kathleen	NaN	4/11/1990	6:46 PM	77834	18.771	False	Business Development
69	Irene	NaN	7/14/2015	4:31 PM	100863	4.382	True	Finance
70	Todd	NaN	6/10/2003	2:26 PM	84692	6.617	False	Client Services
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
939	Ralph	NaN	7/28/1995	6:53 PM	70635	2.147	False	Client Services
945	Gerald	NaN	4/15/1989	12:44 PM	93712	17.426	True	Distribution
961	Antonio	NaN	6/18/1989	9:37 PM	103050	3.050	False	Legal
972	Victor	NaN	7/28/2006	2:49 PM	76381	11.159	True	Sales
985	Stephen	NaN	7/10/1983	8:10 PM	85668	1.909	False	Legal
989	Justin	NaN	2/10/1991	4:58 PM	38344	3.794	False	Legal
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution

145 rows × 8 columns

	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
2	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

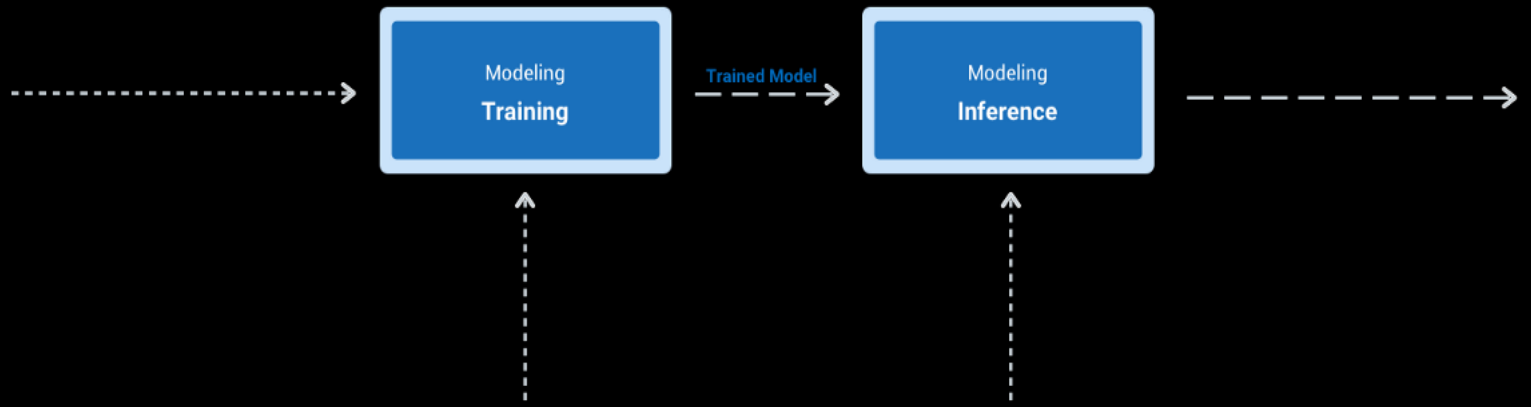
	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

Drop duplicates

Data modeling

Label Times

CUSTOMER	CUTOFF TIME	LABEL
Elisa	10/01/2018	1
Elisa	11/01/2018	0
Saul	09/01/2018	0
Saul	10/01/2018	0
Saul	11/01/2018	1
Sylvia	10/01/2018	1
Sylvia	11/01/2018	1



Predictions

CUSTOMER	PREDICTION DATE	PROBABILITY	PREDICTION
Elisa	12/01/2018	0.235	0
Saul	12/01/2018	0.785	1
Sylvia	12/01/2018	0.892	1

Features Matrix on Traning Data

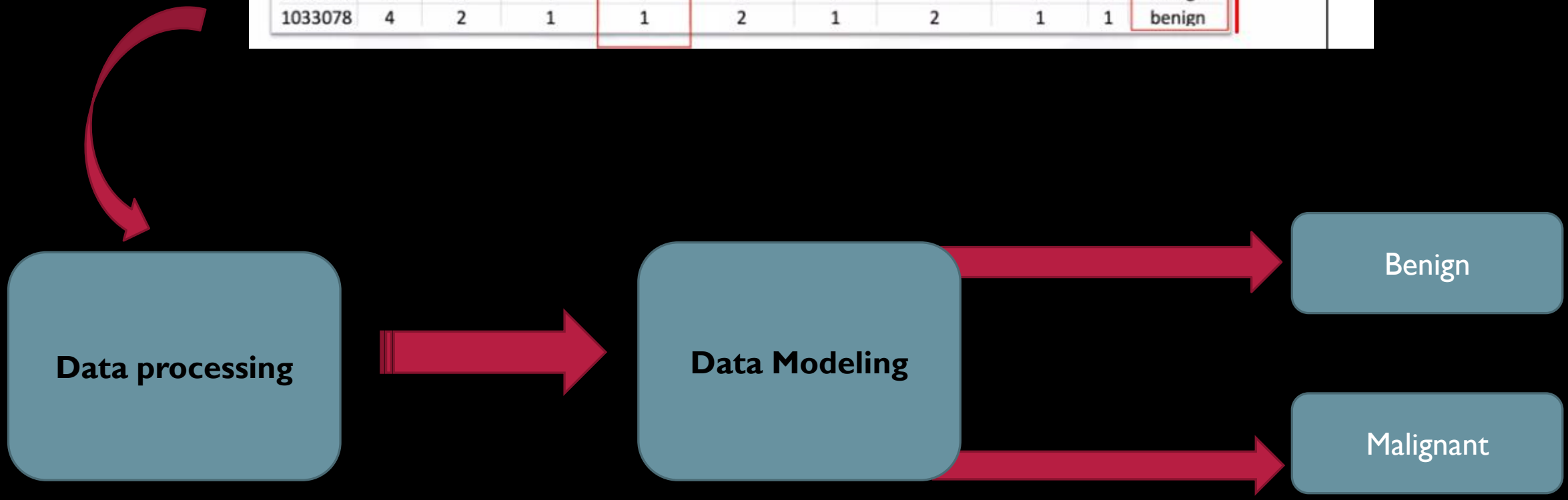
CUSTOMER	CUTOFF TIME	SUM (Transaction)	MEAN (Transaction)	COUNT (Transaction)	MODE (Day of Transaction)	MAX (days between Transaction)	NUMBER_UNIQUE (Products)
Elisa	10/01/2018	467	35.92	13	Thursday	12	115
Elisa	11/01/2018	814	203.5	4	Wednesday	25	2
Saul	09/01/2018	678	113	6	Friday	14	6
Saul	10/01/2018	693	69.3	10	Friday	56	14
Saul	11/01/2018	570	190	3	Wednesday	11	25
Sylvia	10/01/2018	768	64	12	Monday	50	32
Sylvia	11/01/2018	418	38	11	Tuesday	65	101

Features Matrix on New Data

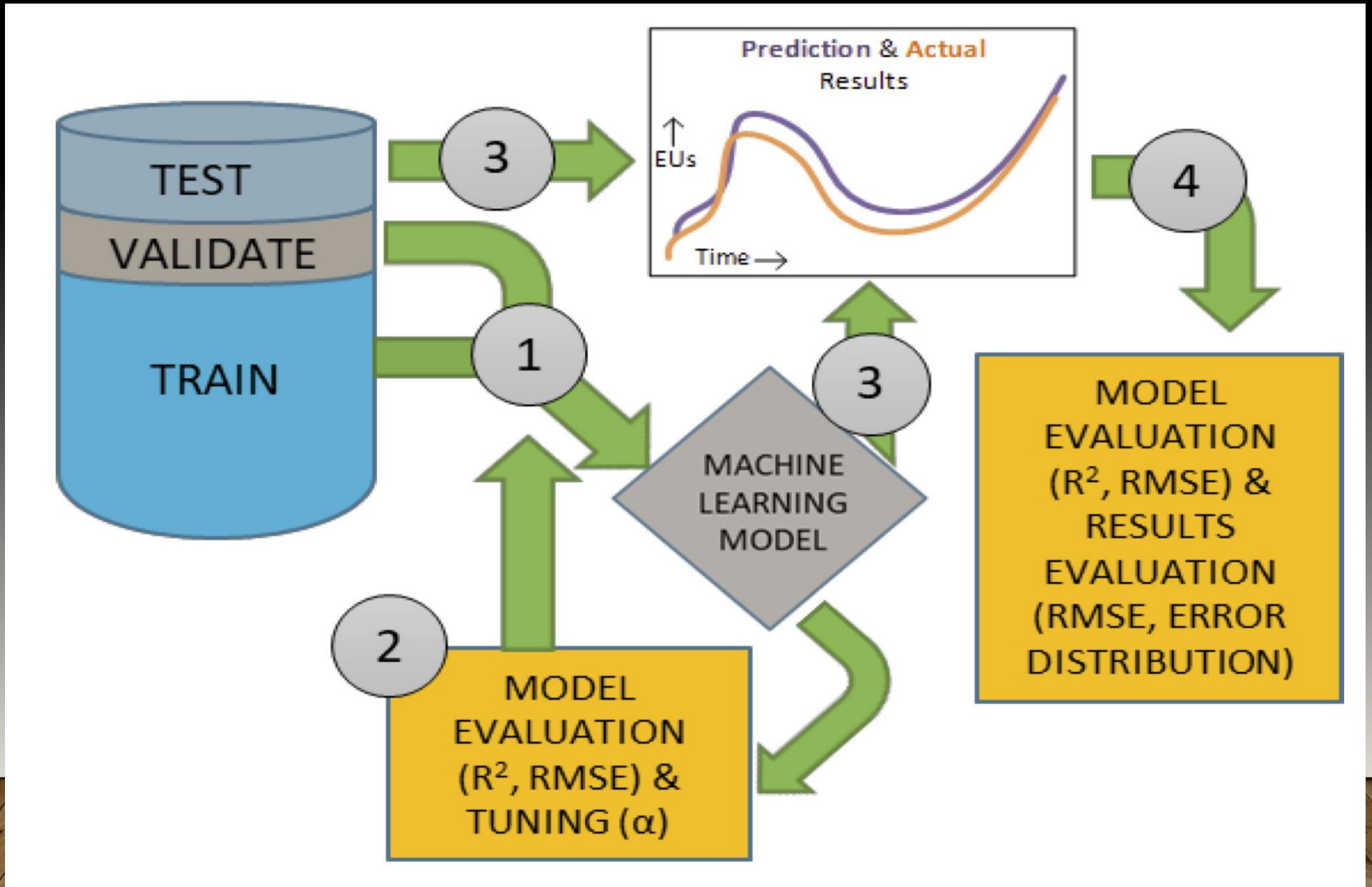
CUSTOMER	CUTOFF TIME	SUM (Transaction)	MEAN (Transaction)	COUNT (Transaction)	MODE (Day of Transaction)	MAX (days between Transaction)	NUMBER_UNIQUE (Products)
Elisa	12/01/2018	500	50	10	Thursday	12	120
Saul	12/01/2018	660	60	11	Friday	21	6
Sylvia	12/01/2018	450	50	9	Teusday	65	103

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

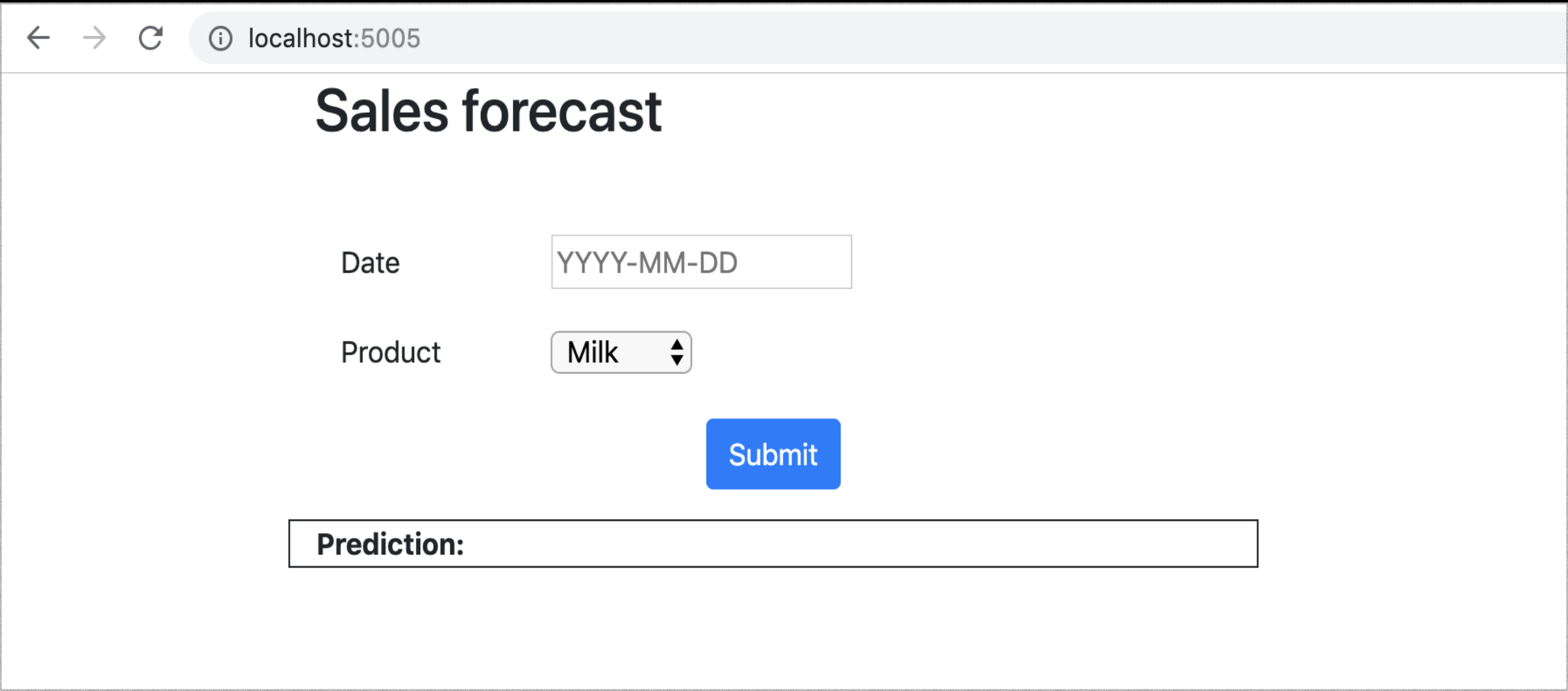
labels



Model evaluation



Model deployment



Data Science Applications

E-commerce

- Identifying Consumers
- Recommending Products
- Analyzing Reviews

Manufacturing

- Predicting Potential Problems
- Monitoring Systems
- Automating Manufacturing Units
- Maintenance Scheduling
- Anomaly Detection

Banking

- Fraud Detection
- Credit Risk Modeling
- Customer Lifetime Value

Healthcare

- Medical Image Analysis
- Drug Discovery
- Bioinformatics
- Virtual Assistants

Transport

- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

Finance

- Customer Segmentation
- Strategic Decision Making
- Algorithmic Trading
- Risk Analytics

Thank you!