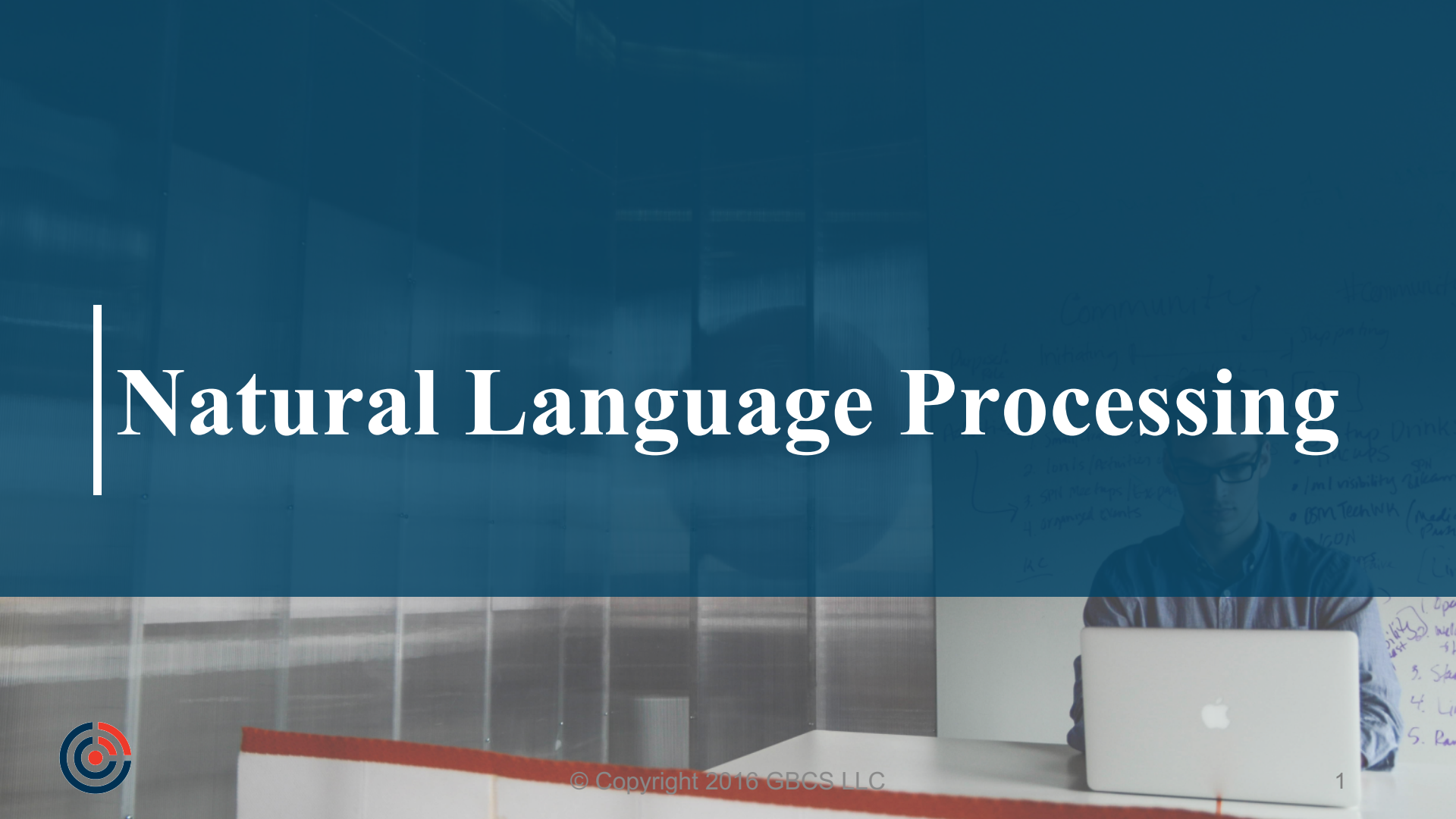
Natural Language Processing



Hello!

Zia Khan

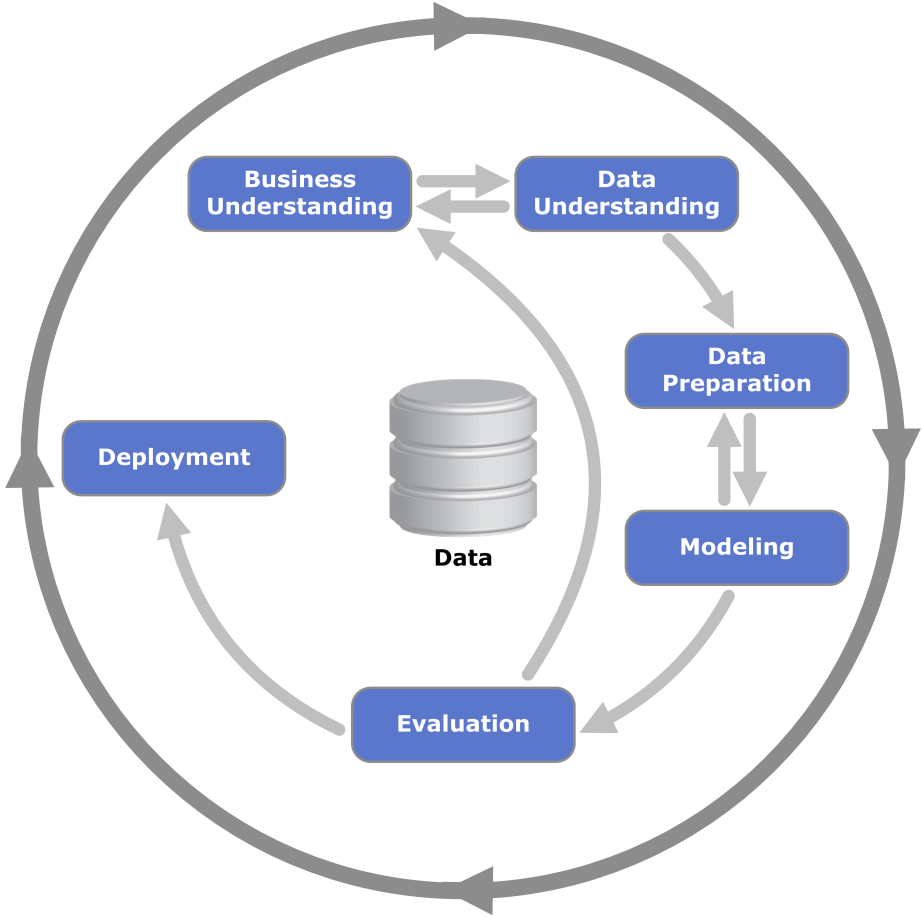
zia@thedevmasters.com

CRISP-DM	1
Business Objective	2
Data Understanding	3
Data Preparation	4
Modeling	5

CRISP-DM

Overview

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



CRISP-DM 1

Business Objective 2

Data Understanding 3

Data Preparation 4

Modeling 5

2. Business Understanding

This case study is designed to evaluate your ability to turn a vague problem into an interesting story with data insights. From your analysis we are looking to understand your approach, programming skill, resourcefulness, and comfort with text data.

2. Text data for analytics

The goal is to build an algorithm that can predict if a campaign will be successful in receiving funding. We have a hypothesis that the information in the free text title may allow us to go beyond the predictive powers of the other variable columns.

CRISP-DM 1

Business Objective 2

Data Understanding 3

Data Preparation 4

Modeling 5

Data Understanding

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110815 entries, 0 to 110814
Data columns (total 15 columns):
ID                  110815 non-null int64
name                110814 non-null object
category            110815 non-null object
main_category       110815 non-null object
currency            110815 non-null object
deadline            110815 non-null object
goal                110815 non-null int64
launched            110815 non-null object
pledged             110815 non-null float64
state               110815 non-null object
backers             110815 non-null int64
country             110815 non-null object
usd pledged         109999 non-null float64
usd_pledged_real    110815 non-null float64
usd_goal_real       110815 non-null float64
dtypes: float64(4), int64(3), object(8)
memory usage: 12.7+ MB
```

CRISP-DM	1
Business Objective	2
Data Understanding	3
Data Preparation	**4**
Modeling	5

Data Understanding

```
1 df.head()
```

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_re
0	1015685046	Organic Tattoo remains 10 days (Self use & cus...	Accessories	Fashion	CAD	1/1/2016	6000	12/7/2015 18:21	100.0	failed	2	CA	74.84	70.
1	1019043170	Handcrafted leather wallet | key holder | card...	Accessories	Fashion	EUR	1/1/2016	2000	12/11/2015 10:52	2102.0	successful	48	BE	2314.01	2282.
2	1036288991	The Liaisons: a new twist on the old standards	Jazz	Music	USD	1/1/2016	5000	11/9/2015 20:12	5630.0	successful	116	US	5630.00	5630.
3	1045749249	The Many Encounters of Bosley Bear	Children's Books	Publishing	USD	1/1/2016	20000	11/24/2015 2:08	101.0	failed	2	US	101.00	101.
4	1048577059	Swift & Co | Innovative Men's Footwear	Footwear	Fashion	GBP	1/1/2016	40000	11/17/2015 12:30	2246.0	failed	19	GB	3416.59	3273.

Data Prep and Modelling

The campaign titles (“name” variable in the data) that people use when creating their campaign have many formats. To extract the maximum information from these text titles you need to carry out a feature engineering step:

Create 3 features from the titles. For each feature explain your thought process for why you chose to create this feature, summary data supporting the creation of this feature, and your analysis around the feature validation. Make sure to discuss the value of these new features within the context of the other variables in the data set.

Modeling

- Clean and fill missing values
- Feature engineering
- Split the dataset into train and test
- Train model (use cross validation)
- Define evaluation metric
- Evaluate the model
- Determine feature importance
- Presentation

Deliverables:

1. Code from your analysis
2. PowerPoint presentation of your case study findings
3. 20 minute presentation of your insights and analysis

Thanks!

Thanks!!

Any questions?

zia@thedevmasters.com