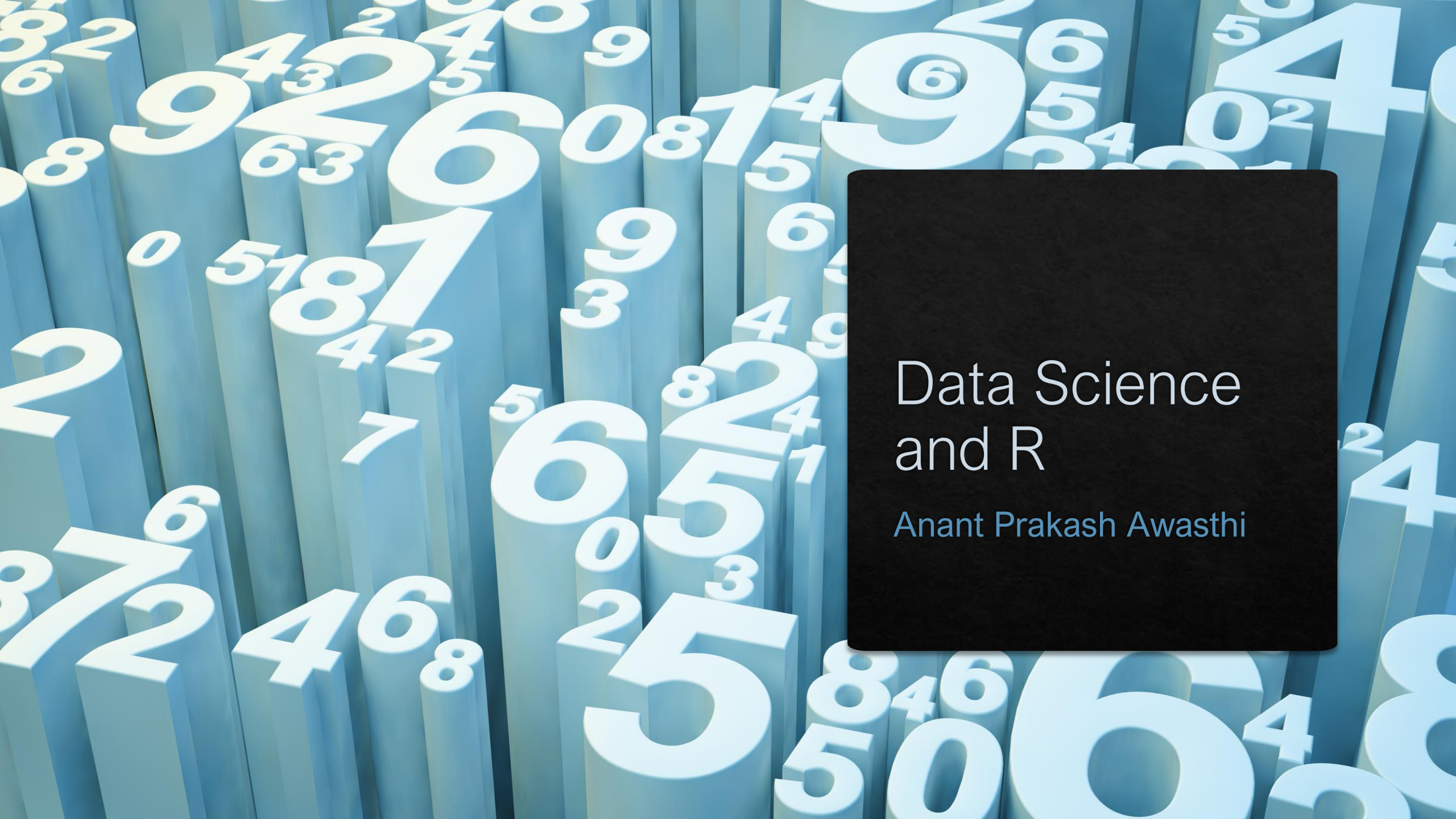


Data Science and R

Anant Prakash Awasthi

Before We Start

- The content delivered by me in this session is totally based on my experience.
- We might have different opinions, I always welcome different healthy conversation. Feel free to shoot a question.
- R is based on an open source culture and there are various ways to do a task. Your way might be more efficient than mine. I will be happy to learn.
- If any of my statement doesn’t go inline with you, it doesn’t mean that I am not respecting you. I respect each one of us.

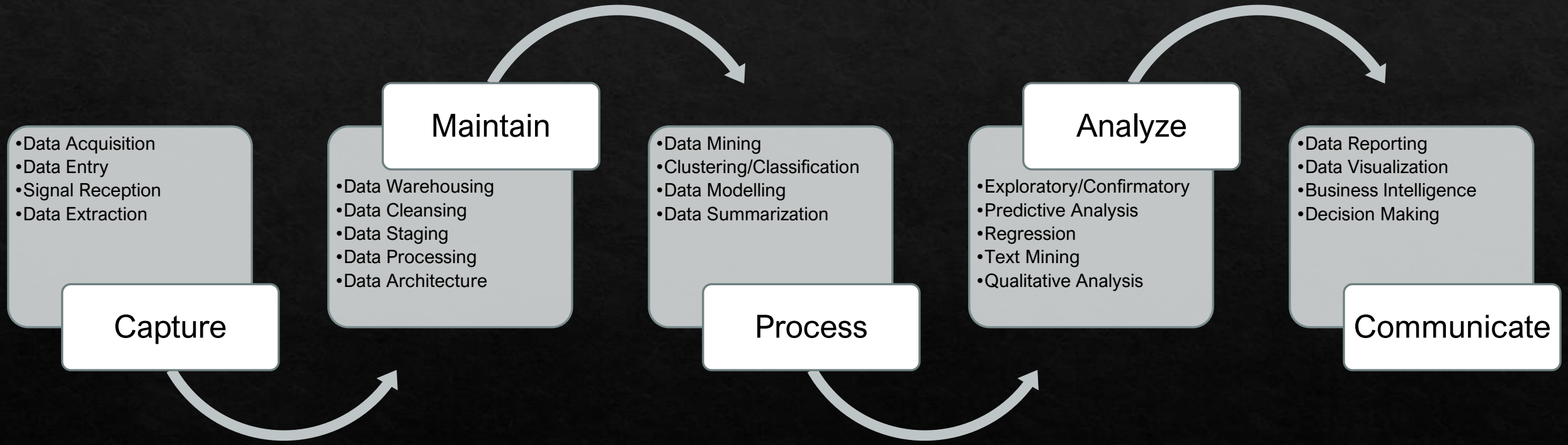
Data Science always starts with Data

What is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

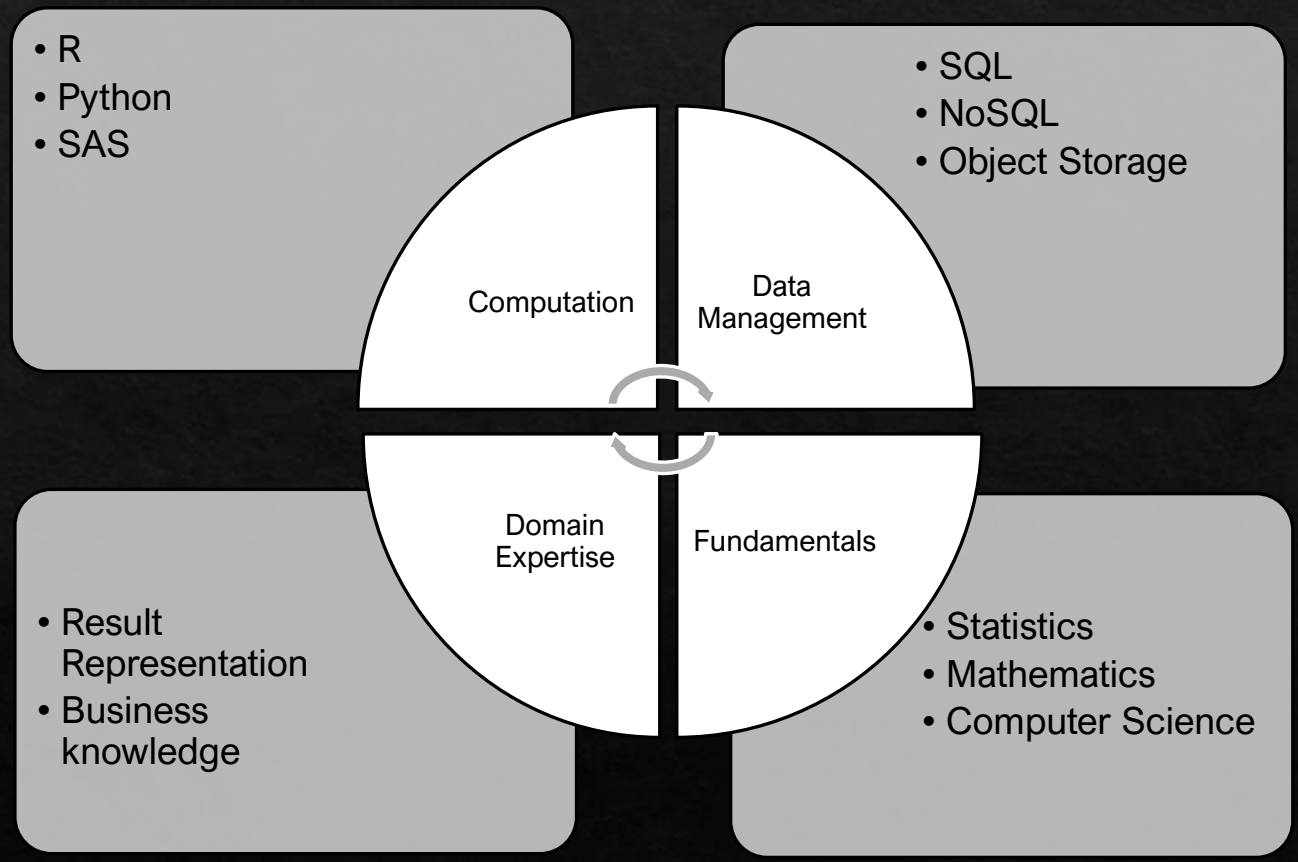
https://en.wikipedia.org/wiki/Data_science

The Data Science Life Cycle



https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

Data Science in a nutshell



Global Trends and Demand as of 2020
Kaggle DS Survey Results

Page Link

Data

Results

Our Dear R!!

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

The Gentlemen

Robert Gentleman

George Ross Ihaka

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

- Open Source
- Exemplary Support for Data Wrangling
- The Array of Packages
- Quality Plotting and Graphing
- Highly Compatible
- Platform Independent
- Eye-Catching Reports
- Machine Learning Operations
- Statistics
- Continuously Growing

- Weak Origin
- Data Handling
- Basic Security
- Complicated Language
- Lesser Speed
- Spread Across various Packages

https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

R and its different flavours

Our Loved Classic R

Microsoft R Open

Anaconda R Environment

Oracle Enterprise R

Who uses R and How R is used?

Facebook uses R to update status and its social network graph. It is also used for predicting colleague interactions with R.

Ford relies on Hadoop. It also relies on R for statistical analysis as well as carrying out data-driven support for decision making.

Thomas Cook uses R for prediction and also Fuzzy Logic Systems to automate price settings of their last-minute offers.

Google uses R to calculate ROI on advertising campaigns and to predict economic activity and also to improve the efficiency of online advertising.

Microsoft uses R for the Xbox matchmaking service and also as a statistical engine within the Azure ML framework.

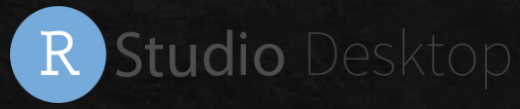
Trulia, the real-estate analysis website uses R for predicting house prices and local crime rates.

https://data-flair.training/blogs/r-applications/

Adoption of R as Software Components

Getting and Setting R

Engine  https://mran.microsoft.com/download

IDE 

https://www.rstudio.com/products/rstudio/

Rtools https://cran.r-project.org/bin/windows/Rtools/

https://www.dunebook.com/best-r-programming-ide/


tidyverse

The tidyverse is a collection of open source R packages introduced by Hadley Wickham and his team that "share an underlying design philosophy, grammar, and data structures" of tidy data.

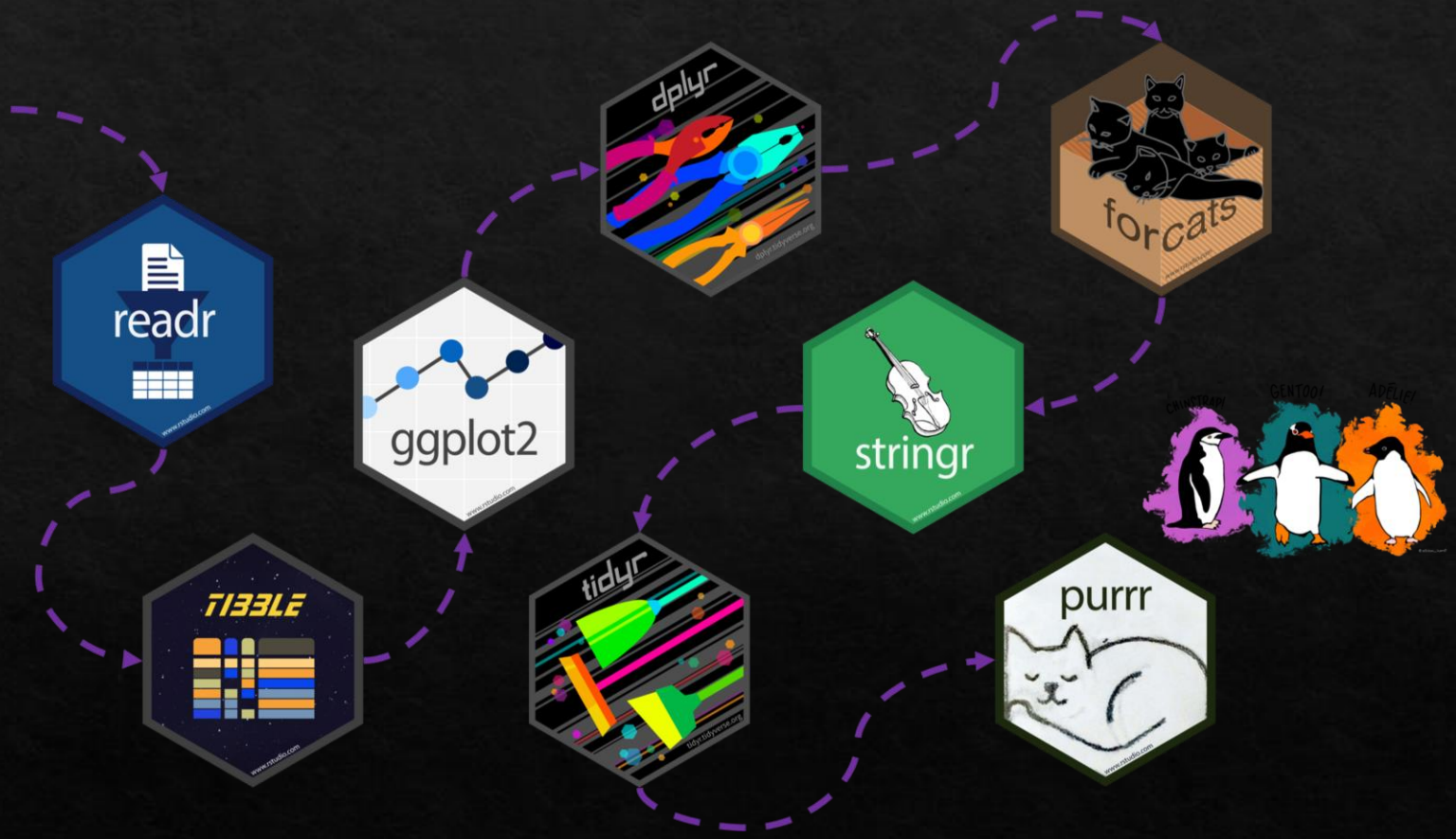
- https://en.wikipedia.org/wiki/Tidyverse

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

- https://www.tidyverse.org/

- Resources
  - Website : https://www.tidyverse.org/
  - Book : R for Data Science
  - Community Resources : https://www.tidyverse.org/learn/
  - Github Repositories : https://github.com/tidyverse

Tidyverse Family



Before we move to hands-on

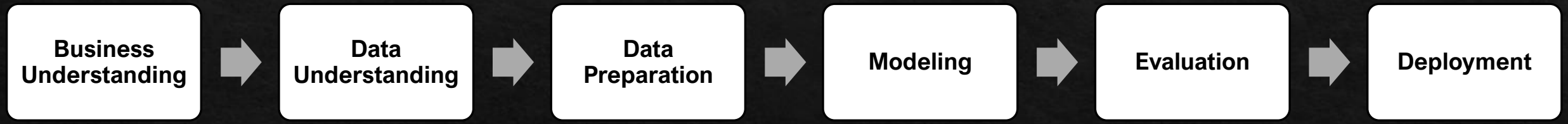
One more point to discuss

 R Only R 

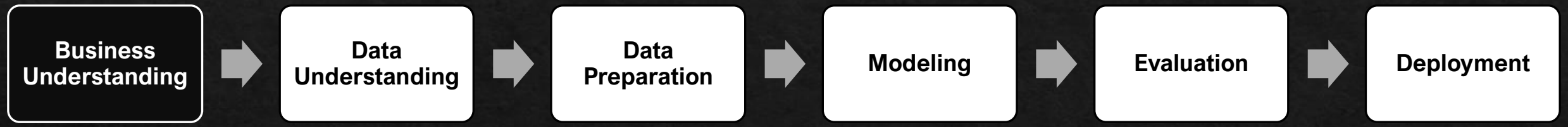
CRISP-DM

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model with six phases that naturally describes the data science life cycle. It's like a set of guardrails to help you plan, organize, and implement your data science (or machine learning) project.

Solution Building Process

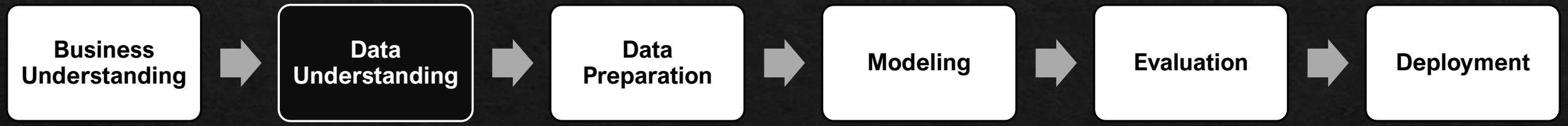


Solution Building Process



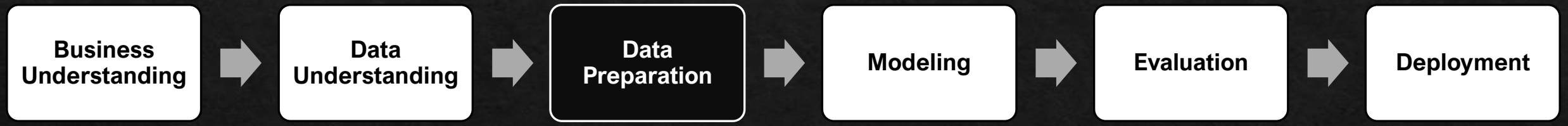
1. Determine business objectives
2. Assess situation
3. Determine data mining goals
4. Produce project plan

Solution Building Process



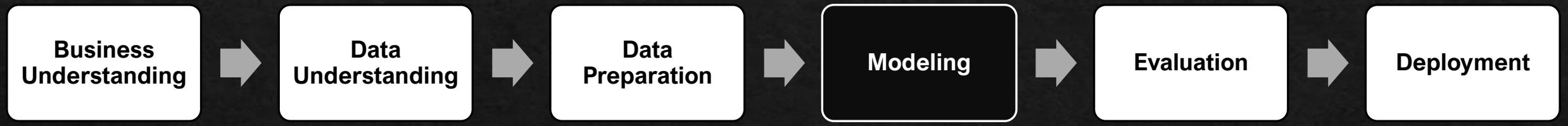
1. Collect initial data
2. Describe data
3. Explore data
4. Verify data quality

Solution Building Process



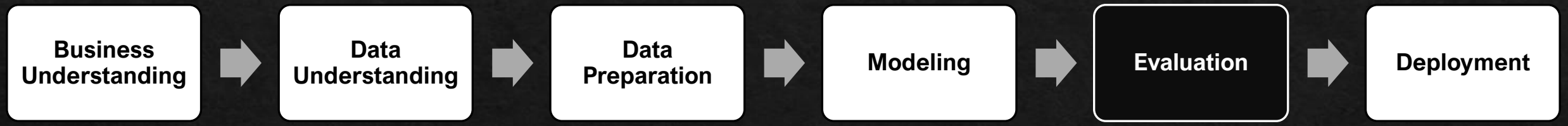
1. Select data
2. Clean data
3. Construct data
4. Integrate data
5. Format data

Solution Building Process



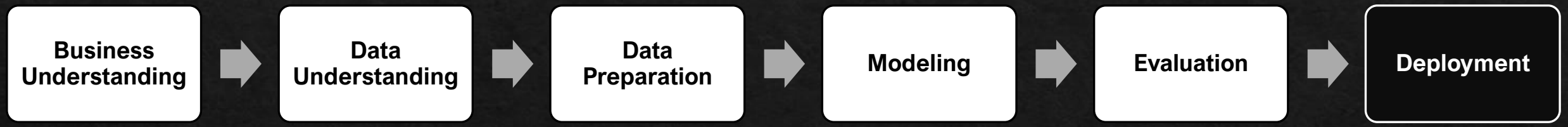
1. Select modeling techniques
2. Generate test design
3. Build model
4. Assess model

Solution Building Process



1. Evaluate results
2. Review process
3. Determine next steps

Solution Building Process

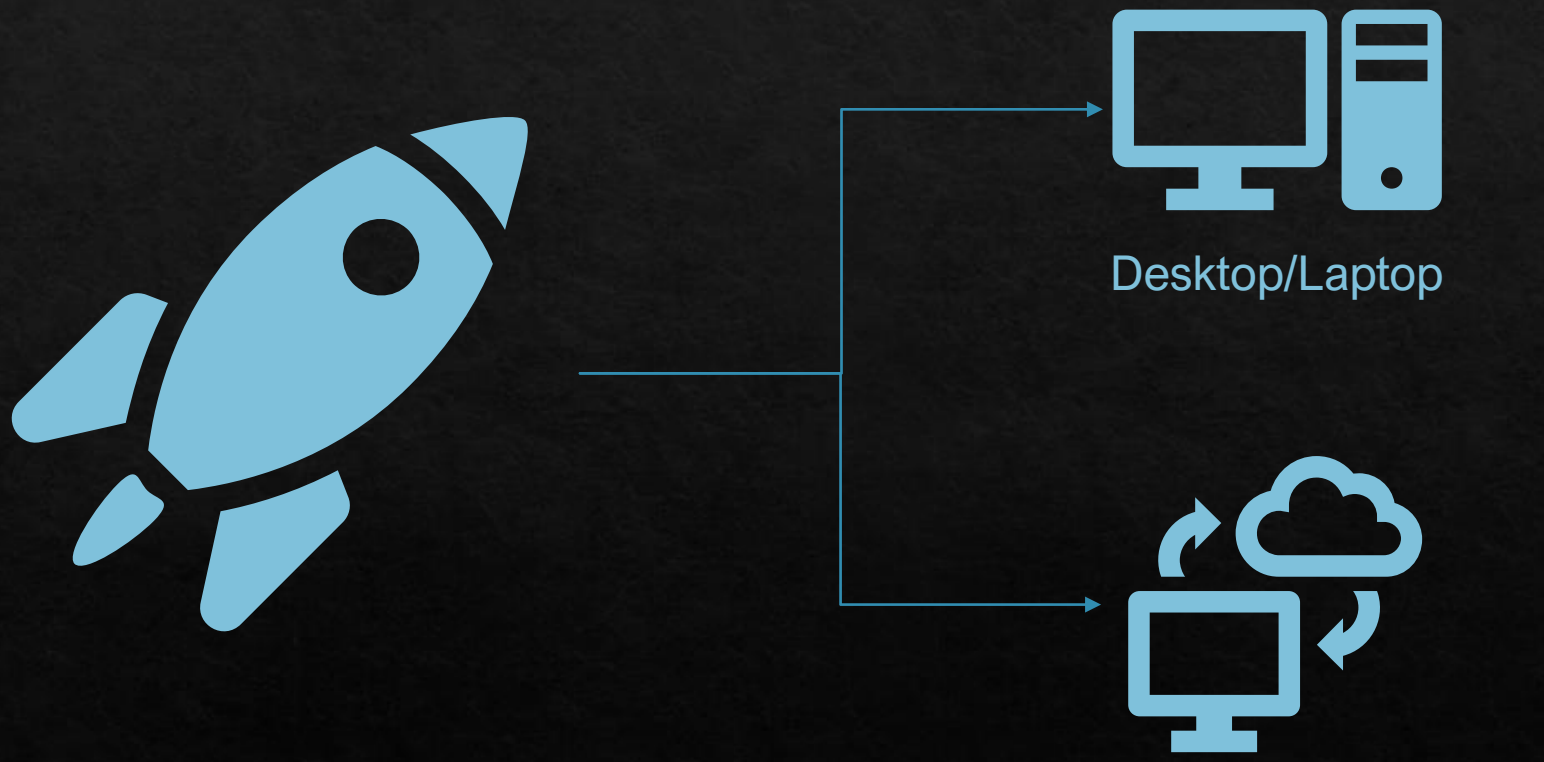


1. Plan deployment
2. Plan monitoring and maintenance
3. Produce final report
4. Review project

Case Studies

- Attendance Percentage in January
- Best and Worst Performing class
- Days with Maximum and Minimum Attendance

Lets go to Play Ground!!



Lets Connect!

anant.awasthi@outlook.com

+91 888.469.2929