

A woman with dark hair is looking upwards with an open mouth, as if in awe or excitement. Overlaid on her face and the background is a complex network of white lines and dots, resembling a data visualization or a neural network. The background has a warm, golden-yellow glow.

Data Science Life Cycle

Principles, Standards and Best Practices

Ivan Chen

April 2020

<https://github.com/chen115y/DataScienceTraining>

Agenda

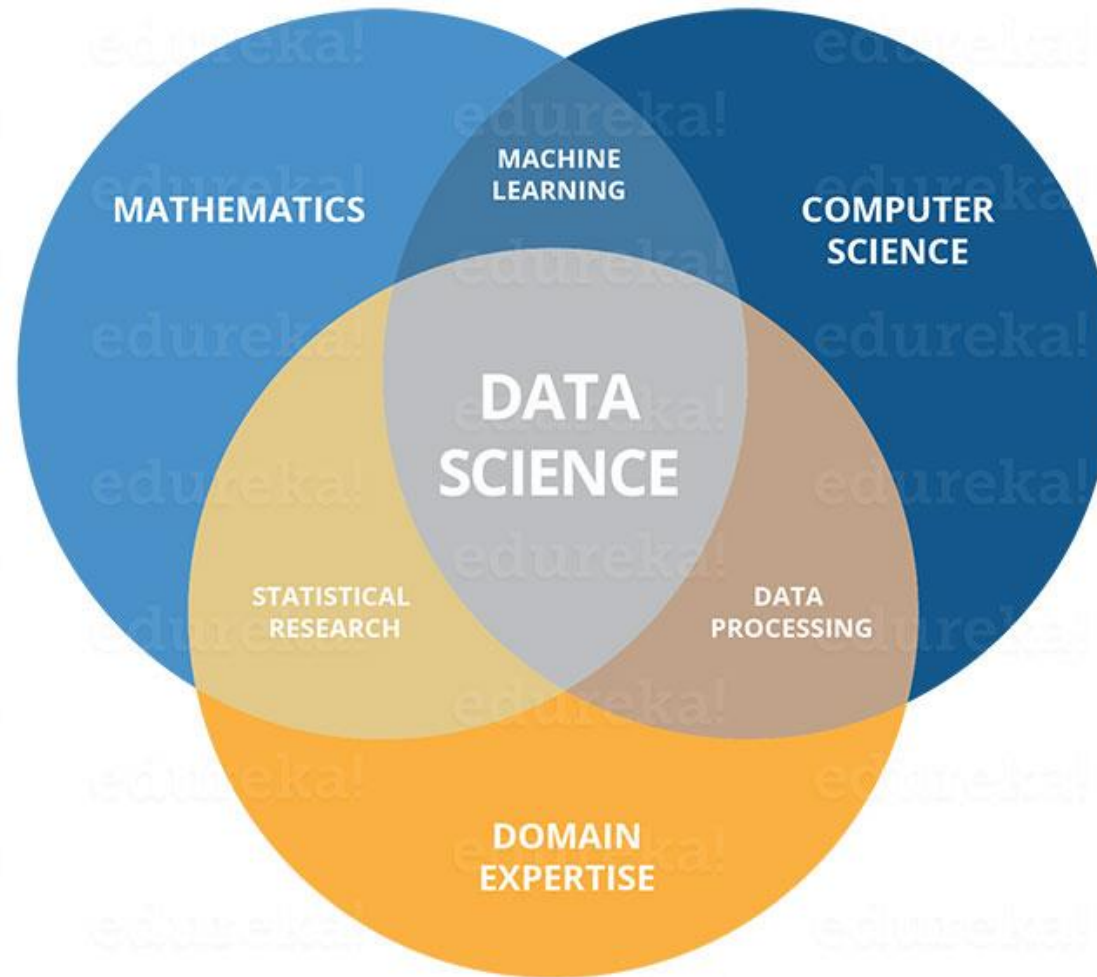
- Data Science – Basic Introduction
- Data Science Life Cycle & Architecture
- Data Science – Principles, Standards and Best Practices
 - Data Science Experiments
 - Data Science Modeling & Evaluation
 - Data Science Productionalization
- Q & A



What is Data Science?

Data science comprises three distinct and overlapping areas:

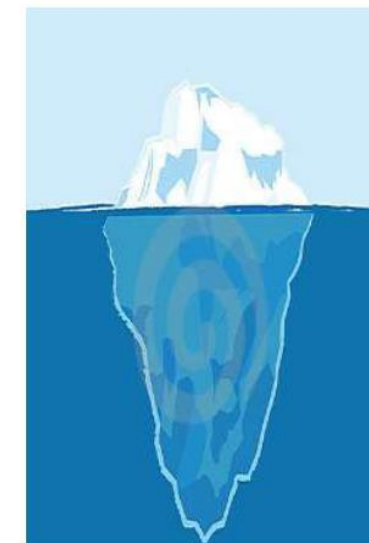
- The skills of a **statistician** who knows how to model and summarize datasets (which are growing ever larger);
- The skills of a **computer scientist** who can design and use algorithms to efficiently store, process, and visualize this data; and
- The **domain expertise**—what we might think of as “classical” training in a subject—necessary both to formulate the right questions and to put their answers in context.



Data Science is to use
Data as base,
Programming as legs,
Machine learning as
backbone and Business
logics as heart.

Key Components of Data Science

- Business Understanding ----- Data Science Requirement Analysis
 - Data Mining & Discovery
 - Data Exploration
 - Data Engineering
 - Feature Engineering
 - Visualization
 - Modeling & Evaluation
 - Software Implementation
 - Production Deployment & Monitoring ----- Data Science Productionalization
- Data Science Experiments
- Data Science Model Development



Machine Learning

Deployment

Application Development

Big Data Processing

Data Storage

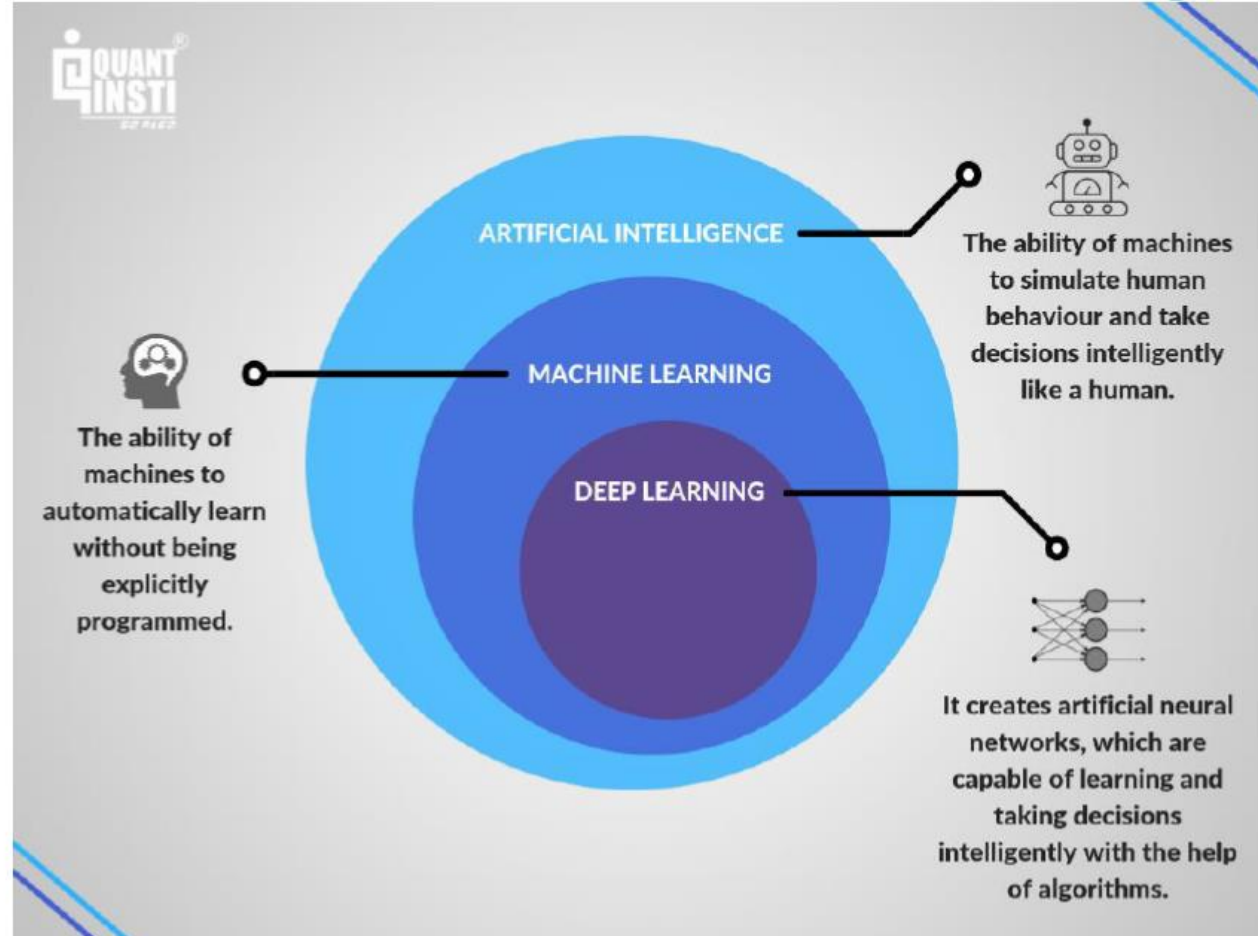
ETL

What is Machine Learning?

Traditional Programming



Machine Learning





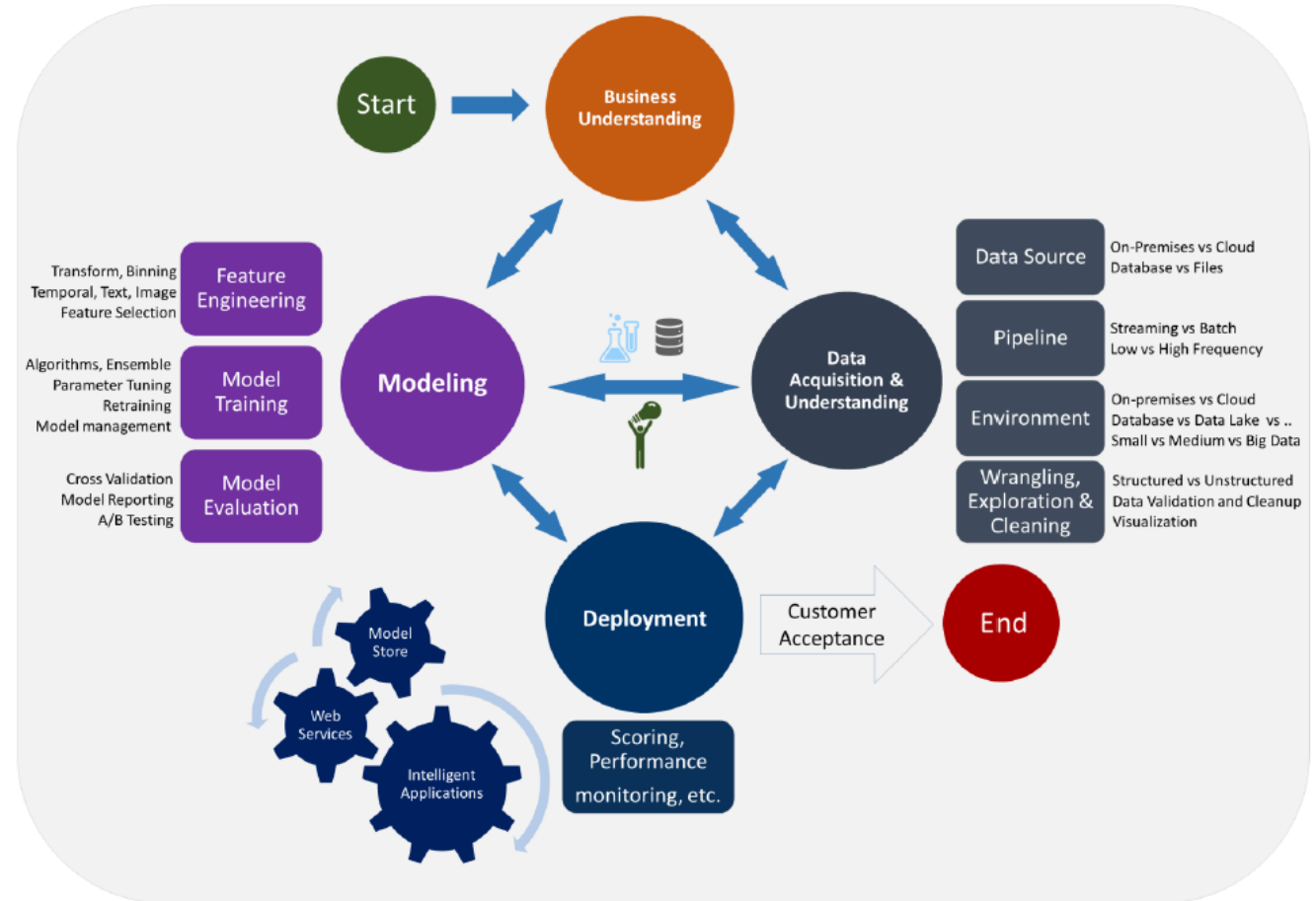
Why need Machine Learning or Use Cases for Machine Learning

- Problems for which existing solutions require **a lot of fine-tuning or long lists of rules**: one Machine Learning algorithm can often simplify code and perform better than the traditional approach.
 - For example, insurance under-writing processes
- Complex problems for which using **a traditional approach yields no good solution**: the best Machine Learning techniques can perhaps find a solution.
 - For example, image recognition
- Getting insights about complex problems and **large amounts of data**.
 - For example, text classification or sentiment analytics (NLP)

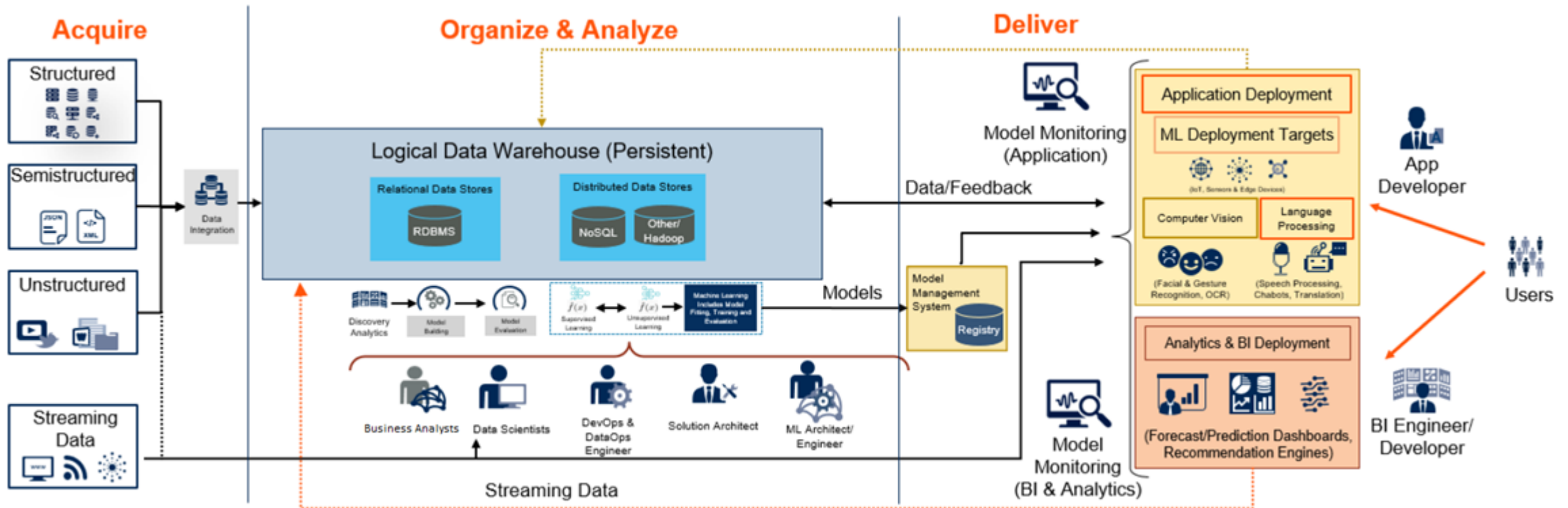


Data Science Life Cycle

Data Science Lifecycle




Data Science Architecture






Data Science Team Members

- **Product Sponsor or Owner**
 - **Data Scientist and Data Analyst**
 - **Machine Learning Architect and/or Engineer**
 - **Data Architect and/or Engineer**
 - **Business analyst**
 - **Data Visualization Engineer**
- 



Data Science – Principles, Standards and Best Practices

- Data Science Experiments
 - Data Science Modeling & Evaluation
 - Data Science Productionalization
- 

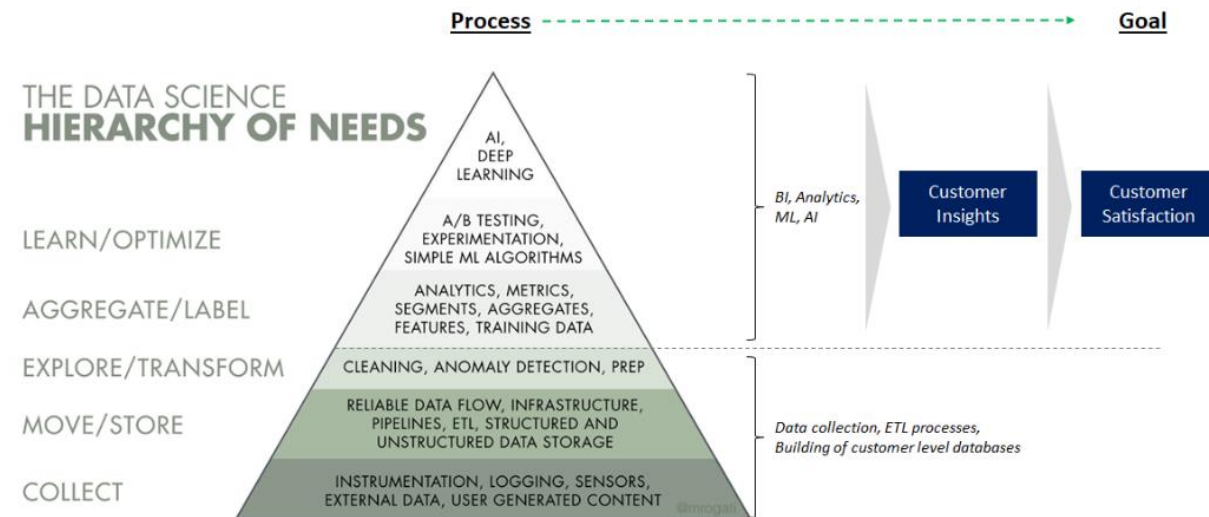
Data Science Experiments

■ Principles

- **Understand the Data Science Hierarchy of Needs**
- **Iterate fast:** Quickly setup a baseline approach and improve it with advanced technologies if needed and move on with this iteratively.
- **Data is no magic bullet:** Understanding the limitations of data and how machine learning algorithms work is important to know which models are worth building.
- **Models must be carefully evaluated by the business before moving to implementation stage.**

■ Best Practices

- **Use Jupyter Notebook with some templates**
- **Use Notebook extensions to help productivity**



Data Science Modeling & Evaluation

- **Principles**

- Build models that answer the right questions
- Analyze the best models and their errors

- **Best Practices**

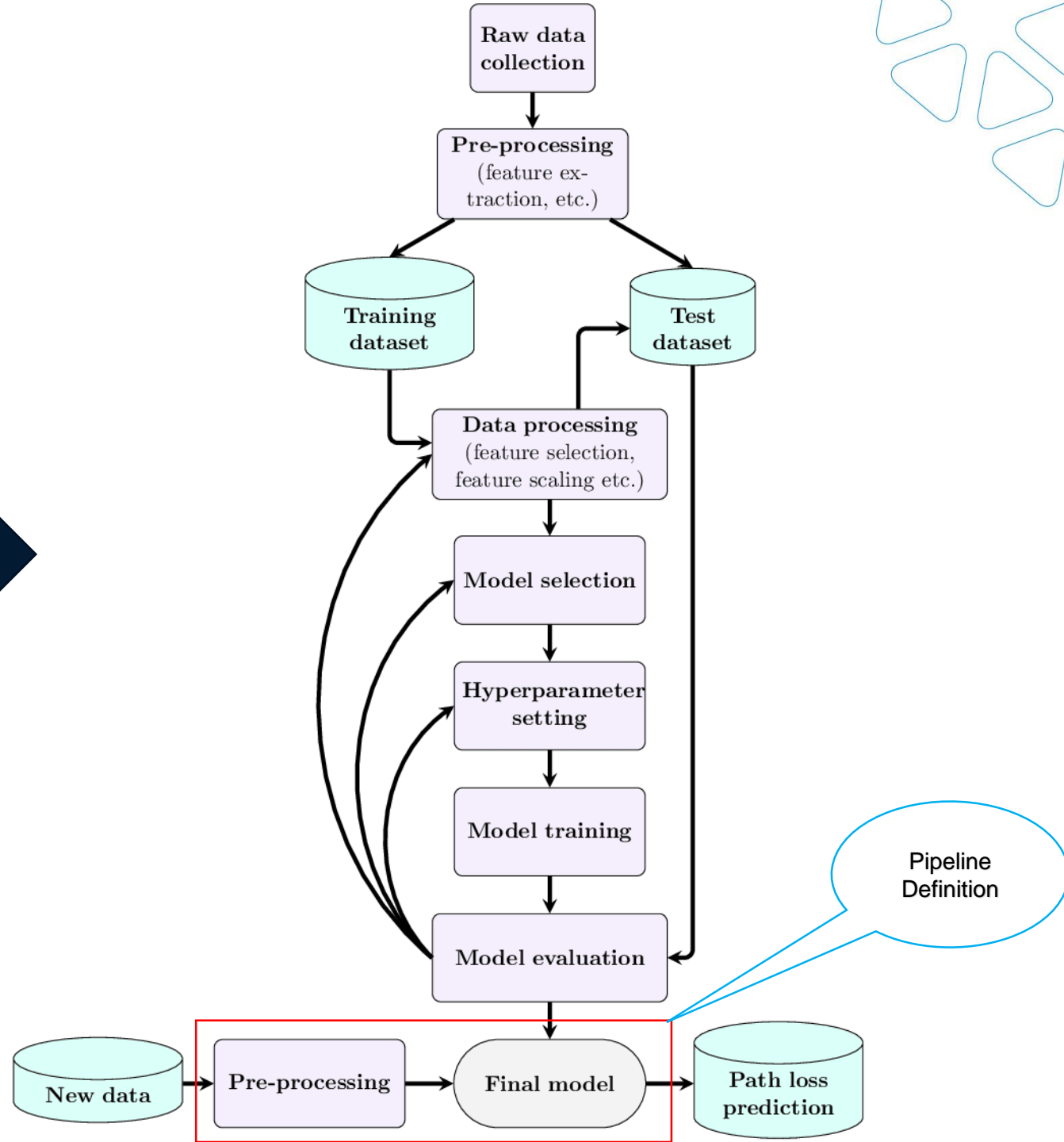
- Use machine learning model selection guidance or cheat sheets
- Be conservative when choosing modeling technology
- Better evaluation using cross-validation
- Use auto-search methods for fine-tune model hyperparameters.

- **Standards:**

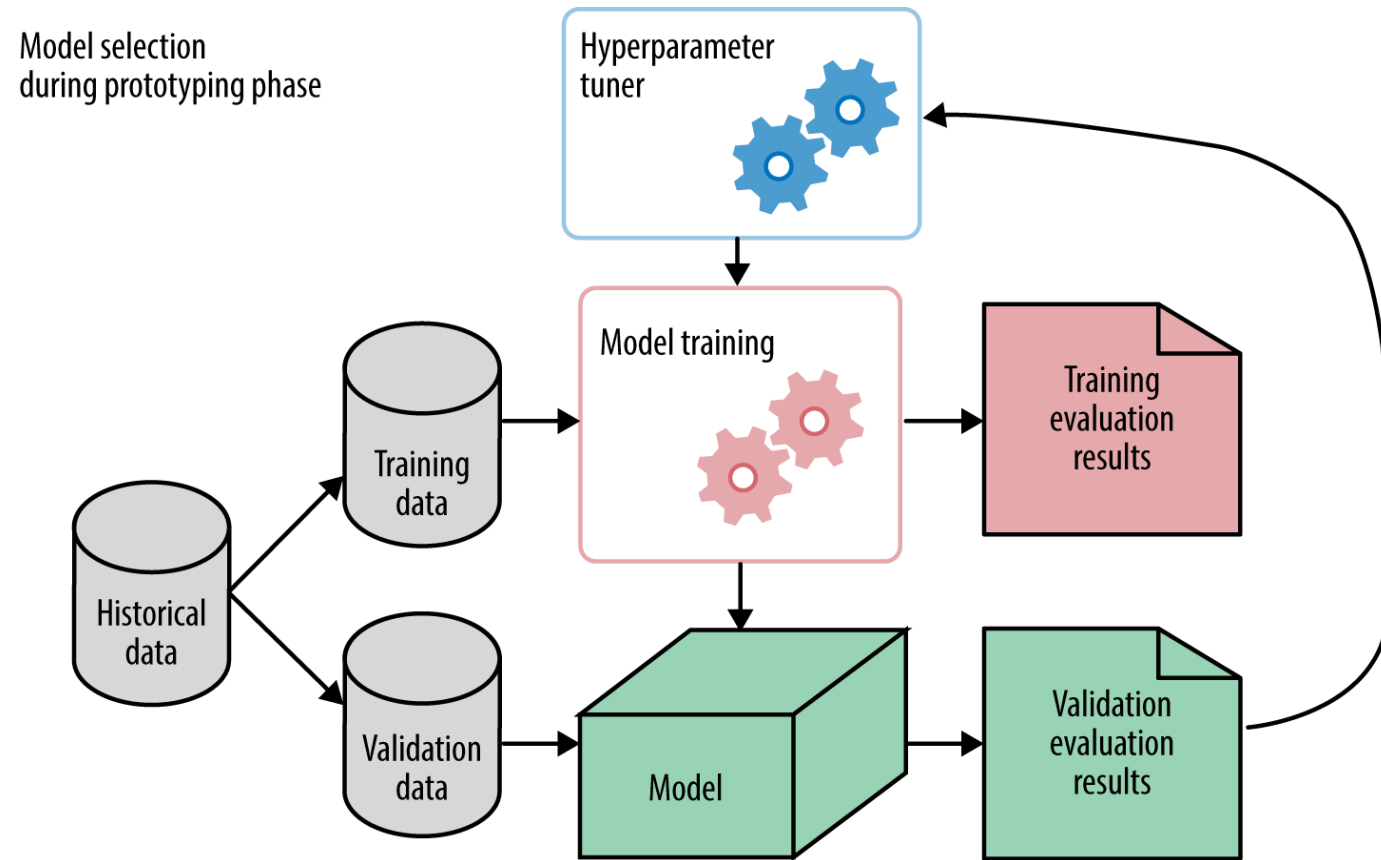
- Data science modeling work flow standards
- Data science model evaluation standards
- Python data science libraries and frameworks



Data Science Modeling Work Flow Standards



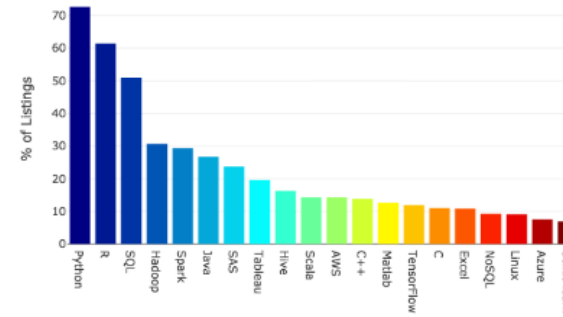
Model Evaluation Standards



Python for Data Science

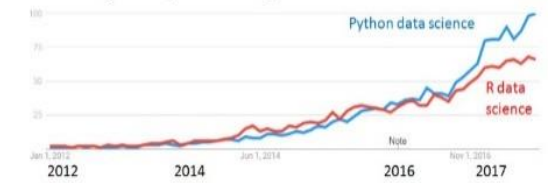
- Functional Scripting & Object-Oriented Programming
- General Purpose including Software Development, Data Science & Data Engineering
- Huge Open Source Libraries/Packages & Community
- Readability & Maintainability
- Less code base complexity
- Support by most all vendors

Top 20 Technology Skills in Data Scientist Job Listings



Medium

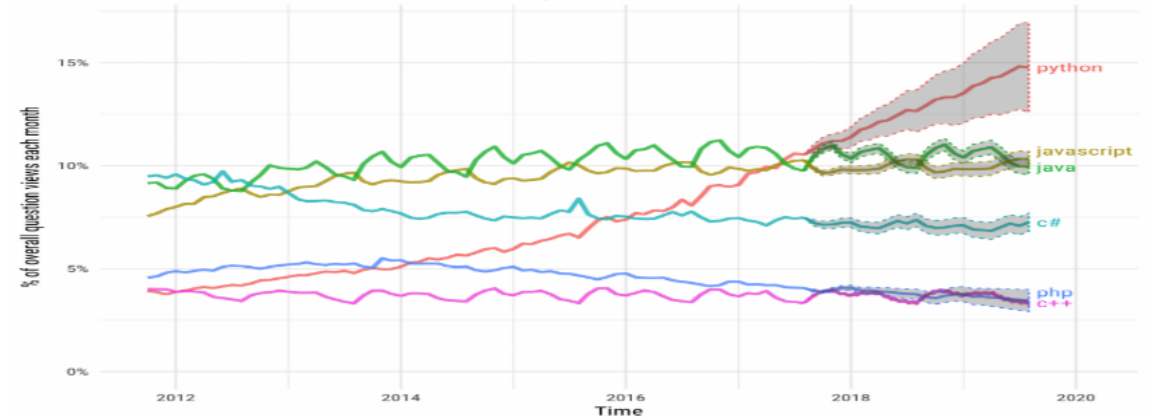
Google Trends, Jan 2012 – Aug 2017



KDNuggets

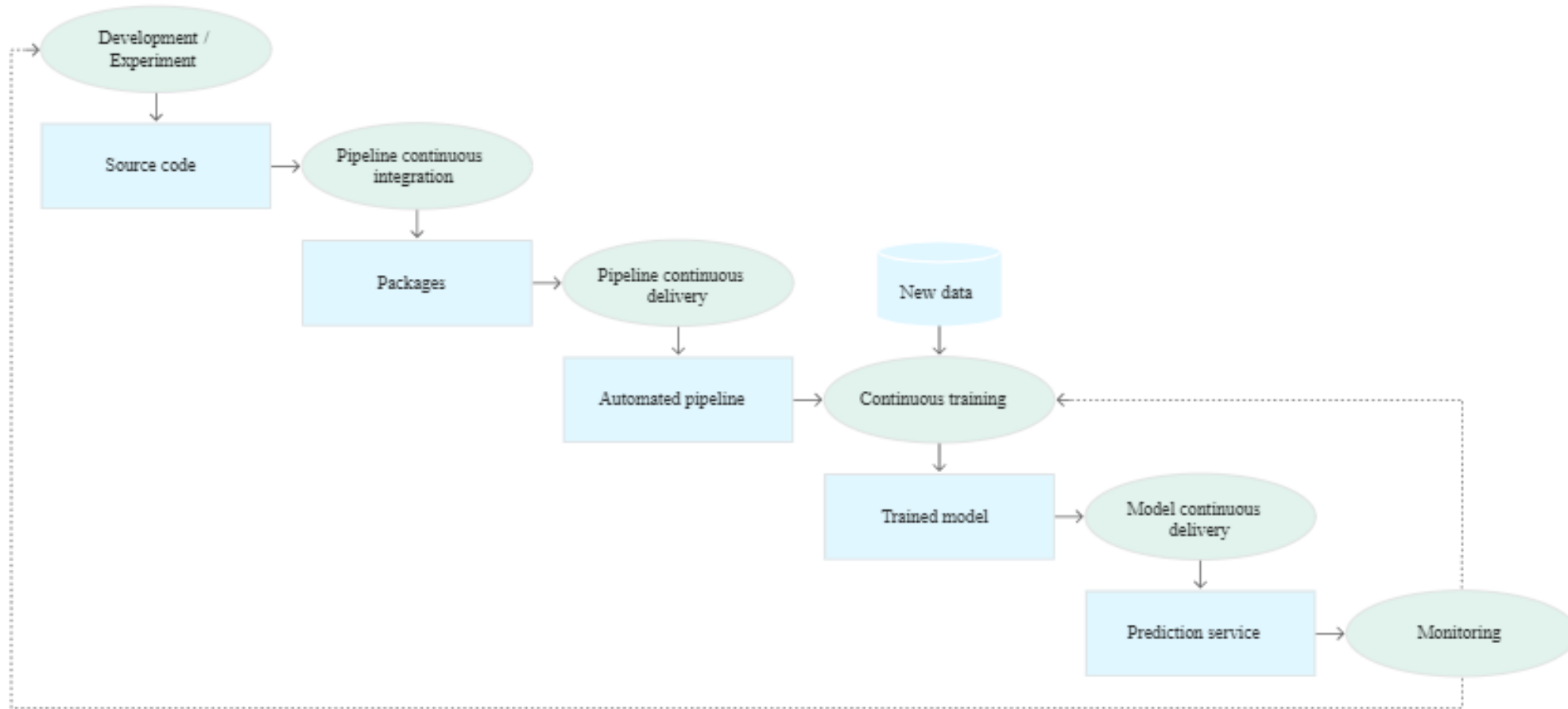
Projections of future traffic for major programming languages

Future traffic is predicted with an STL model, along with an 80% prediction interval.



Stack Overflow Blog

Data Science Productionalization – CI/CD Pipeline Best Practices



- <https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>







Q&A

<https://github.com/chen115y/DataScienceTraining>



Data Science & AI Services in Cloud

Business Use-case	AI — Solutions				
Insights	Machine Learning Platform	Amazon SageMaker	AI Platform and Cloud AutoML	Watson Studio	Azure Machine Learning Service
User Experience	Conversational Platforms	Amazon Lex	Dialogflow	Watson Assistant	Microsoft Bot Framework + Azure Bot Service
	Text Summarization/Analytics	Amazon Comprehend + Amazon Textract	Cloud Natural Language (NL) API + AutoML Natural Language + Document Understanding AI	Watson (NLU + Discovery + Knowledge Studio)	Azure Cognitive Services — Language
	Image Classification	Amazon Rekognition Image	Vision API and AutoML Vision	Watson Visual Recognition	Azure Cognitive Services — Computer Vision
	Streaming Video Processing	Amazon Rekognition Video and Amazon Kinesis Video Streams	Cloud Video Intelligence + AutoML Video	Watson Media	Azure Media Services — Video Indexer
Process Automation	IoT Platform	AWS IoT	Cloud IoT Core	Watson IoT	Azure IoT Central
	Contact Center	Amazon Connect	Contact Center AI	Customer Care Voice Agent	Dynamics 365 Virtual Agent for Customer Service