TESTING A MACHINE LEARNING SOLUTION

May 24, 2023

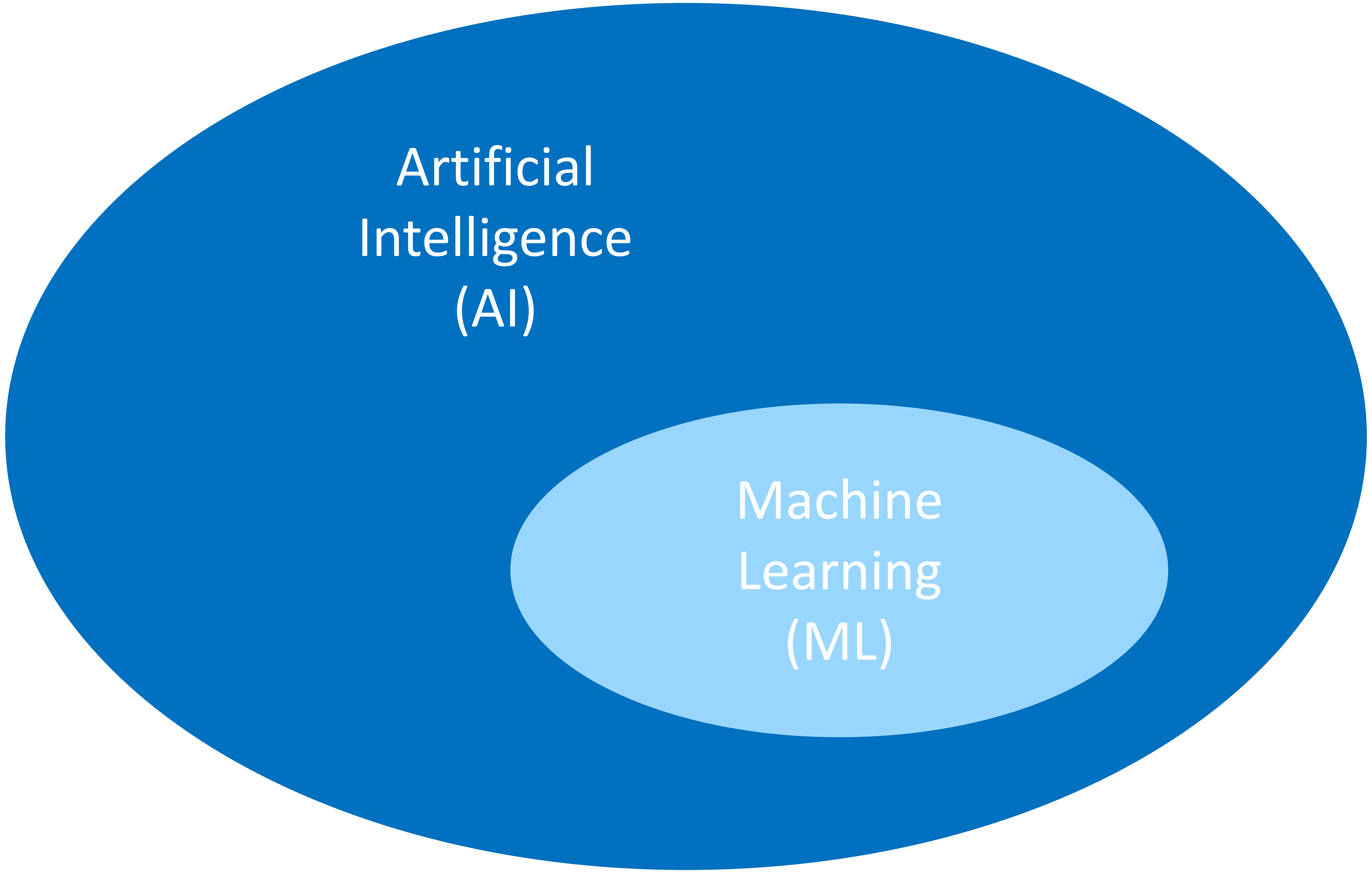
AGENDA

- Quick Overview of the AI/ML Landscape
- Testing Approaches to the Machine Learning Lifecycle

1

AI/ML Landscape

A Brief Clarification on Terminology



AI & ML: From Slow Growth to Soaring Heights

1950s
The idea that computers can think is conceptualized and the term Machine Learning is coined

1970s
People start to lose faith in AI/ML and investment drops

1980s – **1990s**
Resurgence – some common model types/algorithms are discovered

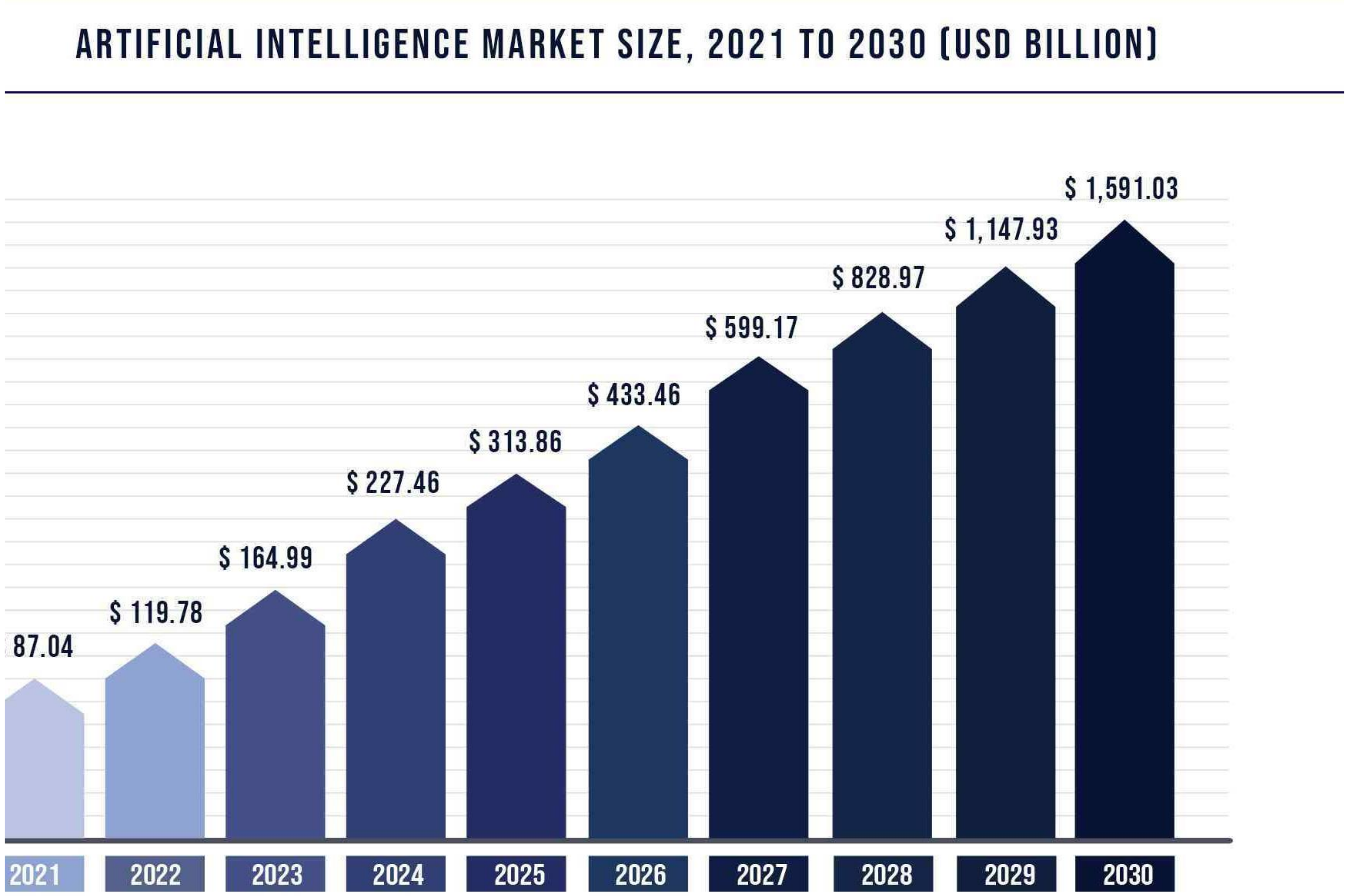
2000s – **2010s**
ML grows in popularity. Netflix prize is launched to create the best algorithm

Today
Massive models are being built to do almost anything conceivable

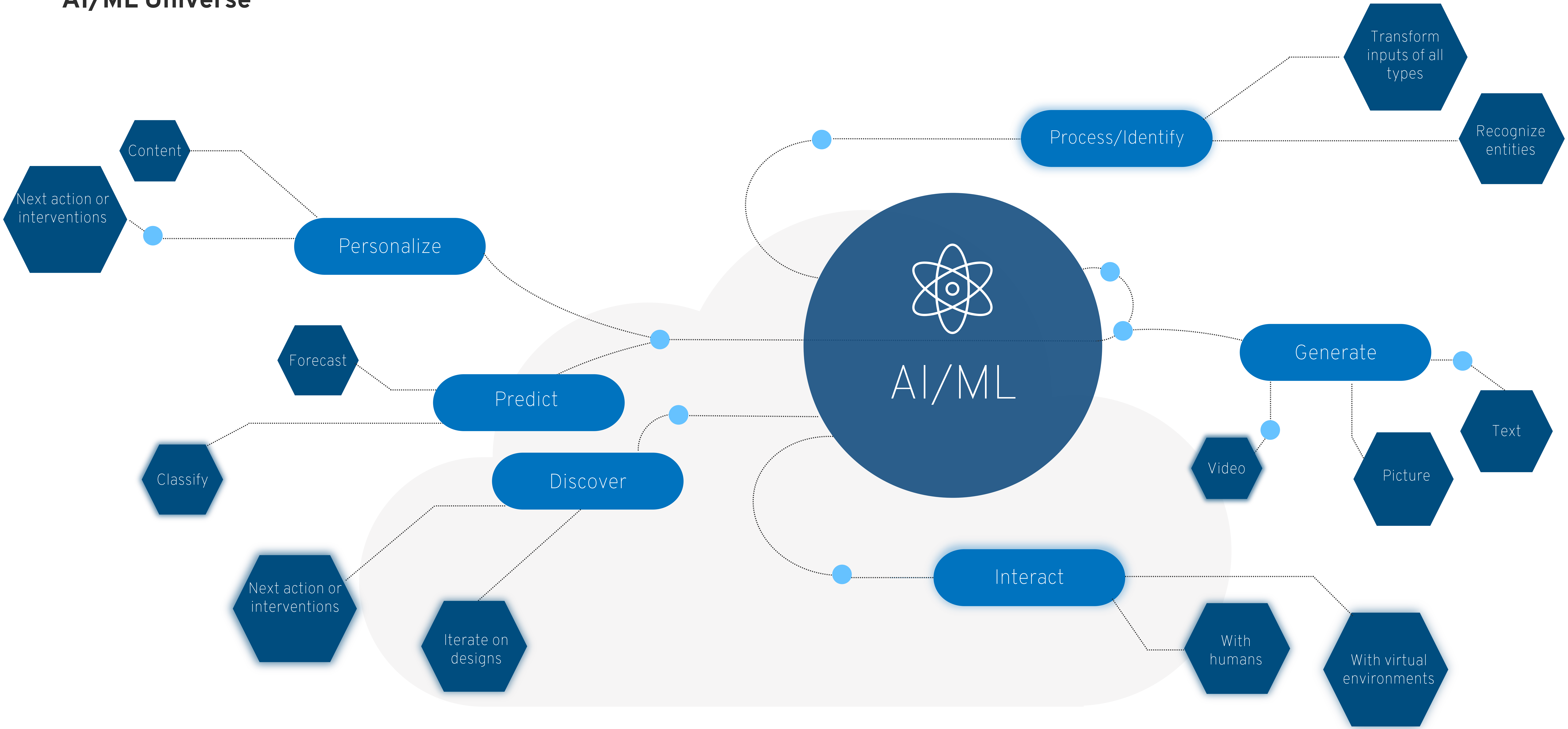
The future

- Everything we do will be augmented with AI/ML
- Modeling becomes easier, faster, and more accurate everyday – what data you have is the differentiator
- The rate at which things are researched/developed will increase very rapidly

Adoption & Market Growth



AI/ML Universe



2

Data Science Testing Challenges

Testing vs Evaluation

Testing

- Verifying that software product or application does what it is supposed to do
- Examples:
  - Unit tests
  - Regression tests
  - Integration tests

Evaluation

- Metrics and visualizations used to summarize to reliability and predictive performance of a model on validation or test data
- Examples:
  - Accuracy
  - F1 score
  - RMSE

Simple Example of Classic Testing Approach

```python
def addition(a, b):
    c = a + b
    return c

assert addition(1, 1) == 2
print("Success")
```

Most testing approaches require the ‘answers’ to be known and the solutions to be deterministic

Most machine learning models are stochastic, involve some element of randomness, and the answers are often unknown or entirely undefined

Advanced models are incredibly complex and difficult to understand, much less test

Deterministic vs Stochastic

Deterministic

- Produce the exact same results for a particular set of inputs
- Examples of deterministic concepts:
  - Accounting
  - Geometry
  - Converting units of measure
- Some simpler models are deterministic:
  - Linear Regression
  - Logistic Regression
  - Principal Component Analysis (PCA)

Stochastic

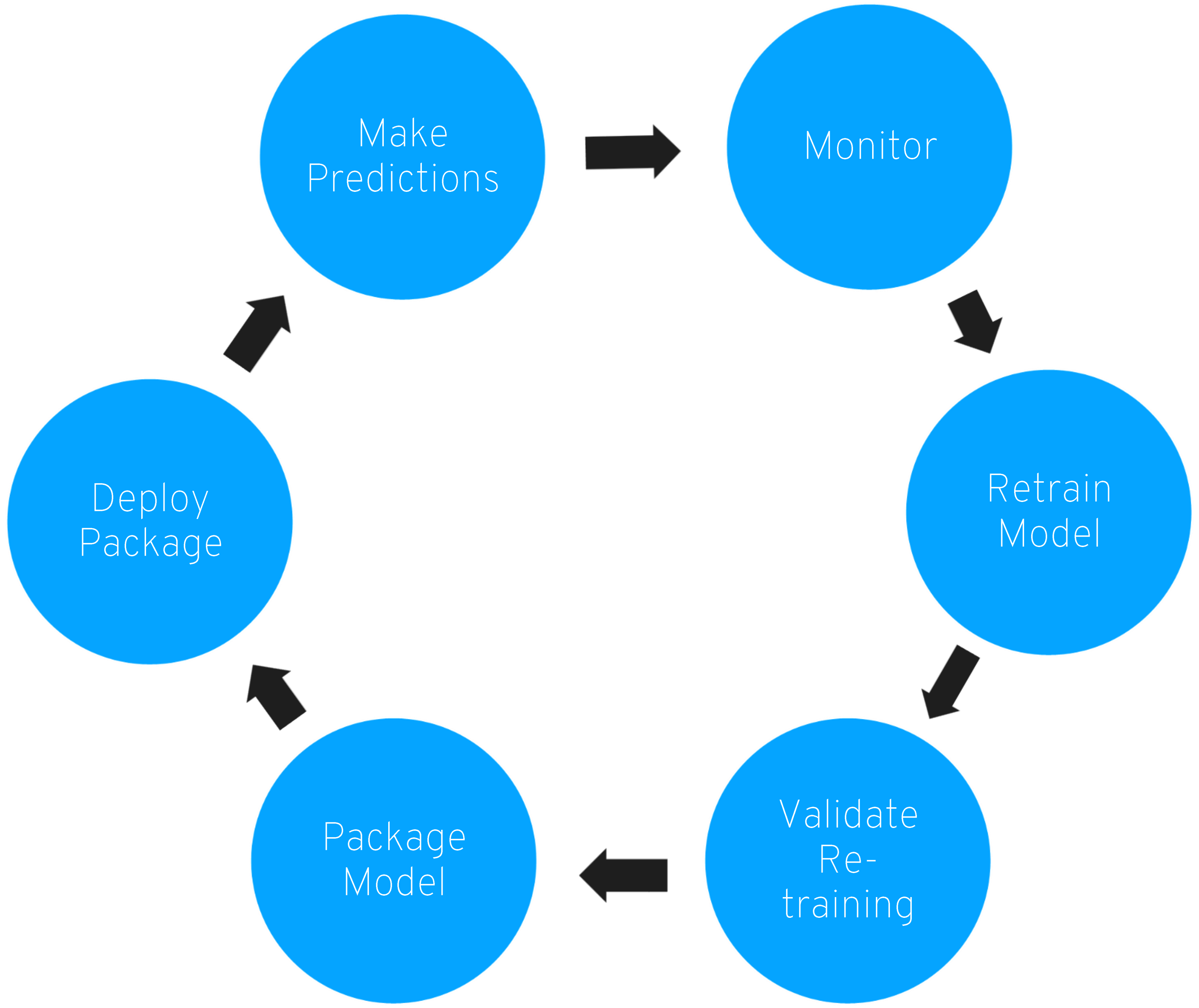
- Product differing results for a particular set of inputs
- Examples of stochastic concepts:
  - Monte Carlo simulation
  - Weather forecasts
  - Playing cards
- Most machine learning today is stochastic:
  - Any model involving a random seed hyperparameter
  - Any model with an element of probability or randomness

Why bother with stochastic models?

- 'Randomness' is a feature, not a bug
- Elements of randomness often produce better results
- Capable of deriving logic from complex data

Machine Learning Development Cycle

Model Development	Retraining Cycle



Model Development Common Testing Concerns

Model Development

Common Concerns

- Minor input data changes can sometimes have seemingly outsized impacts on predictions even with the same data set
  - Outlier removal
  - Adding/removing samples

Best Practices

- Stick to foundational data quality best practices
- Track the datasets that are being used to build models
- More data typically means less impact

Model Development Common Testing Concerns

Model Development

Problem Statement → Collect Data → **Develop Features** → Build & Evaluate Models → Package Model

Common Concerns

- Validation of derived features often doesn't happen

Best Practices

- Perform data unit tests before putting in production

Model Development Common Testing Concerns

Model Development

Problem Statement → Collect Data → Develop Features → **Build & Evaluate Models** → Package Model

Common Concerns

- Different iterations of the model produce different results
- Train-test splits are randomized

Best Practices

- Use a model registry tool to track models
- Fix the random number generator seed - only if absolutely necessary
- Perform appropriate number of k-folds validation

Model Development Common Testing Concerns

Model Development

<u>Common Concerns</u>

- ‘Black box’ logic

<u>Best Practices</u>

- Validate the output ranges
- Validate changes in inputs result in predictions that match your intuition
- Validate small perturbations affect the model how you would expect

Retraining Cycle Testing Concerns

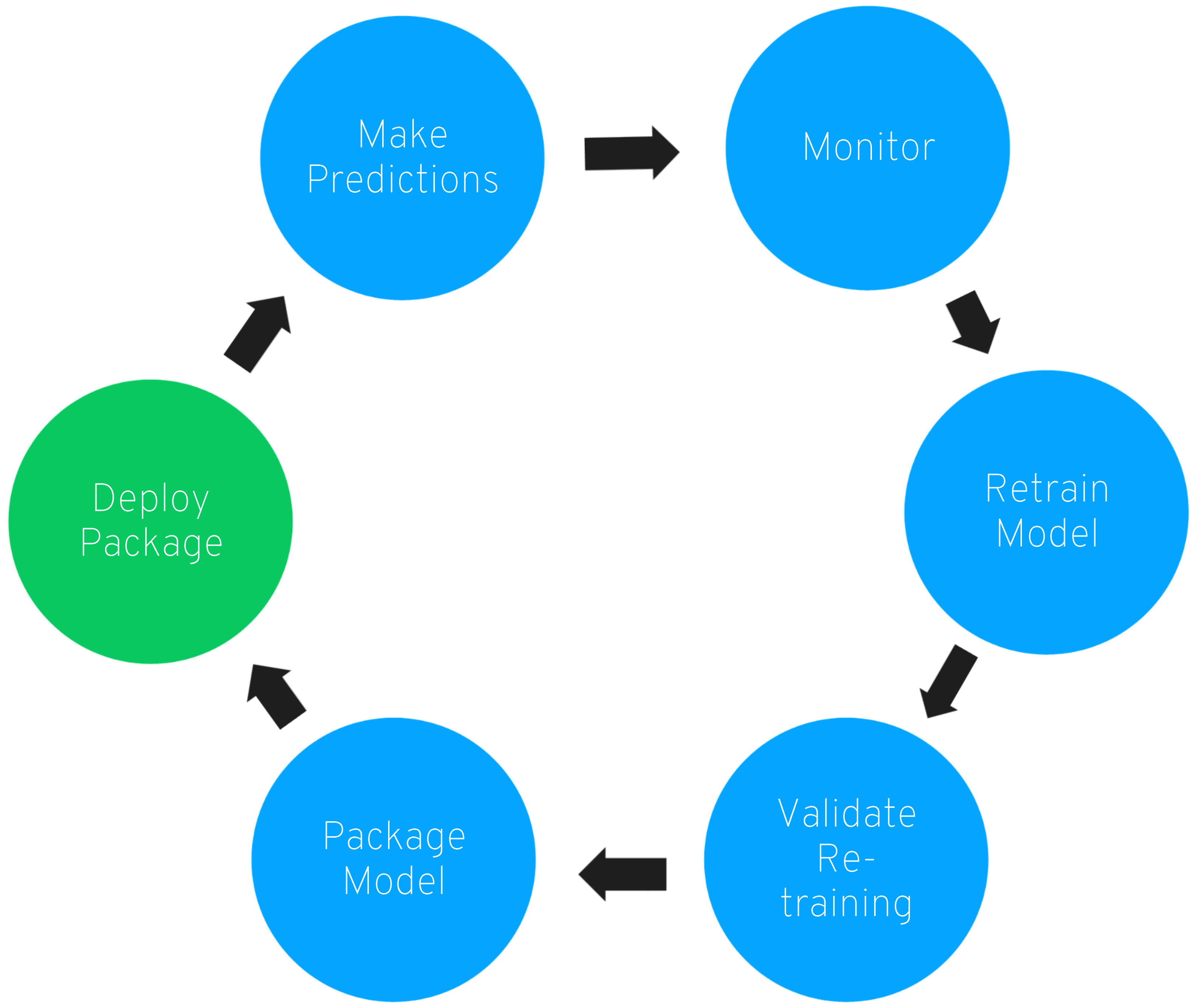
Common Concerns

- Differences in environment can cause different predictive values and subsequently model performance

Best Practices

- Keep the environment constant through development and production phase using containers or VMs

Retraining Cycle



Retraining Cycle Testing Concerns

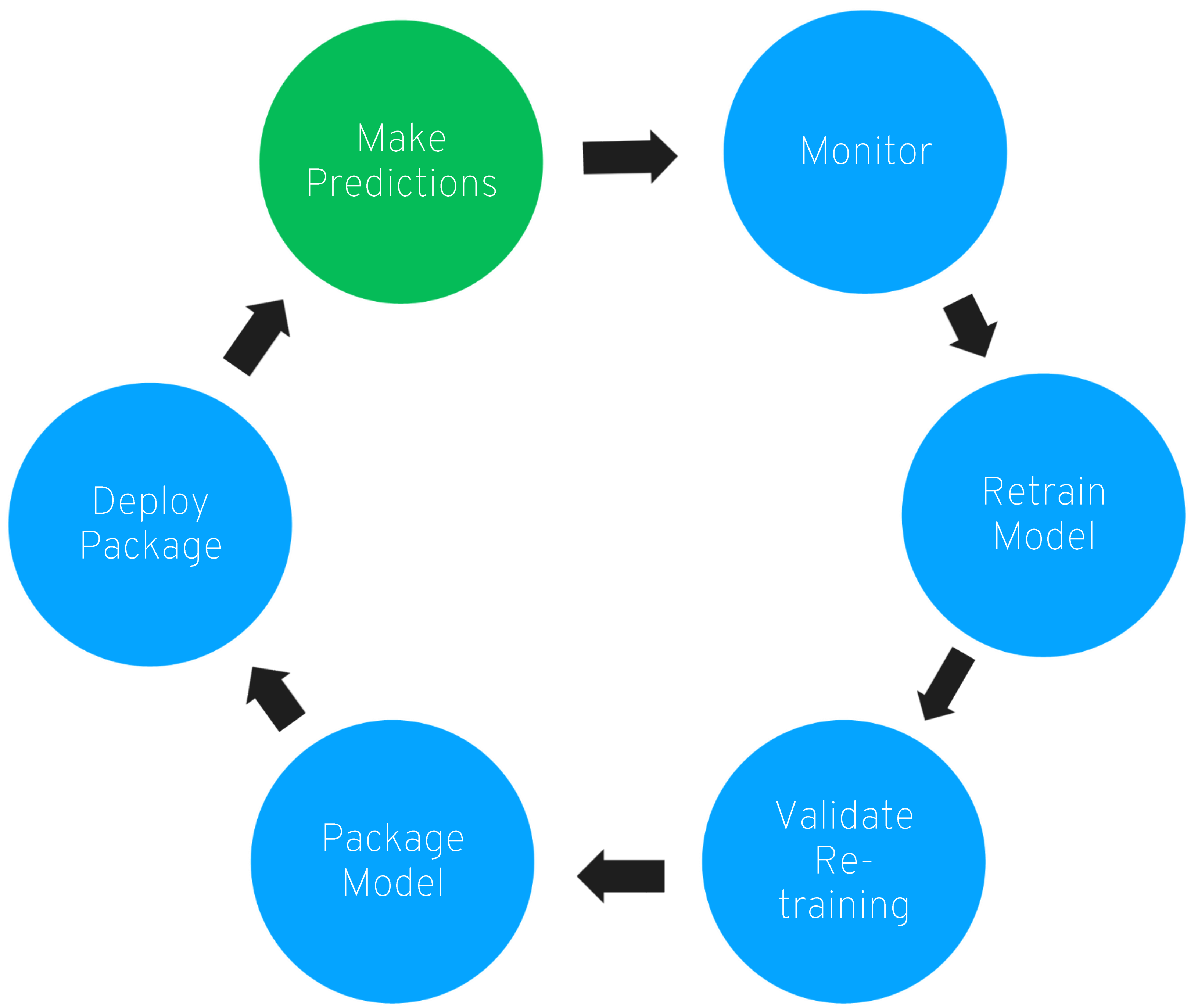
Common Concerns

- 'Correct' answer is undefined

Best Practices

- Validate your input data using traditional approaches
- Validate you are predicting reasonable values
- Validate predictions for high consequence examples

Retraining Cycle



Retraining Cycle Testing Concerns

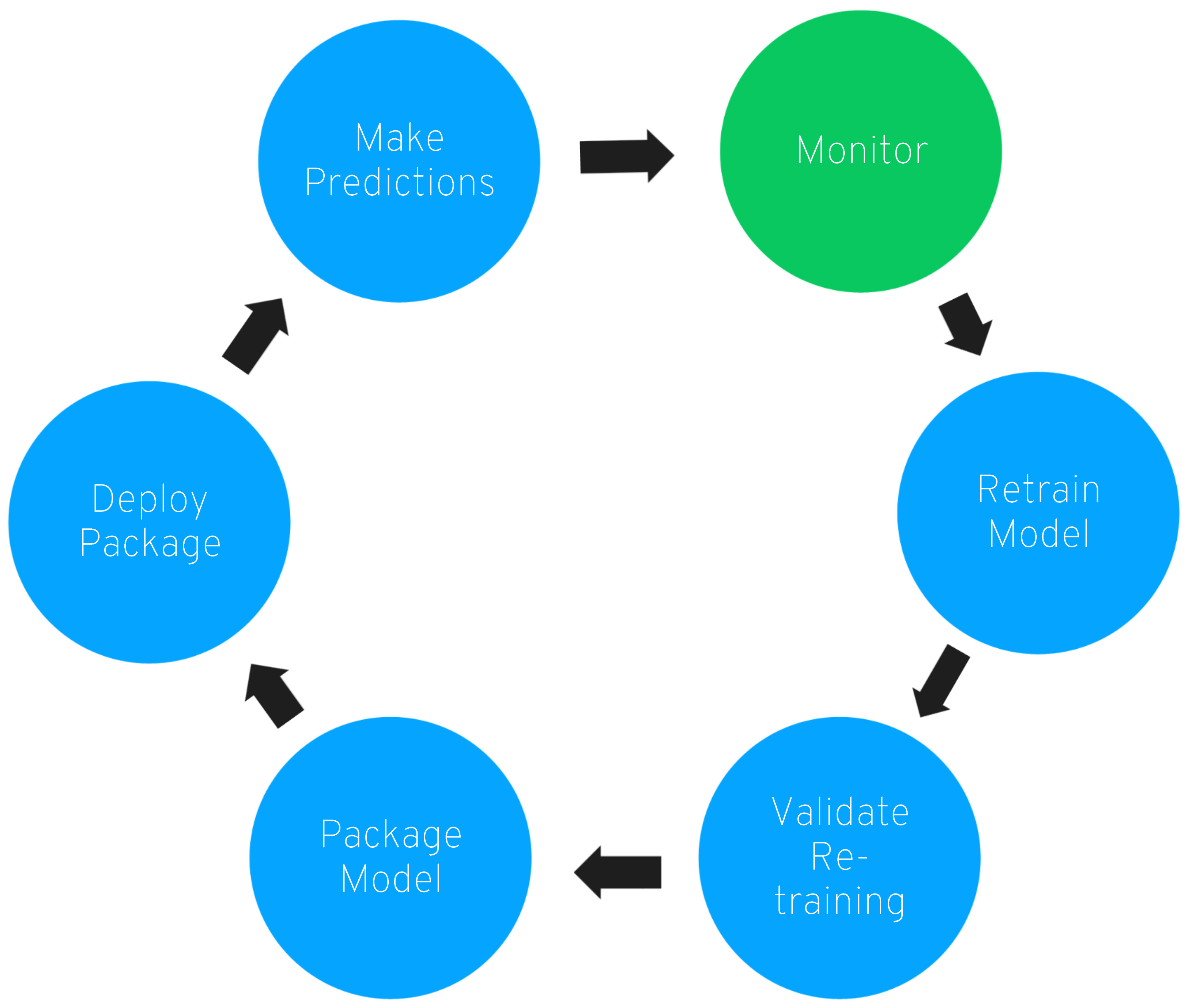
Common Concerns

- Model performance degrades over time, unlike typical software

Best Practices

- Monitor the performance of your model
- Monitor the distribution of your inputs
- Set thresholds for both to trigger re-training

Retraining Cycle



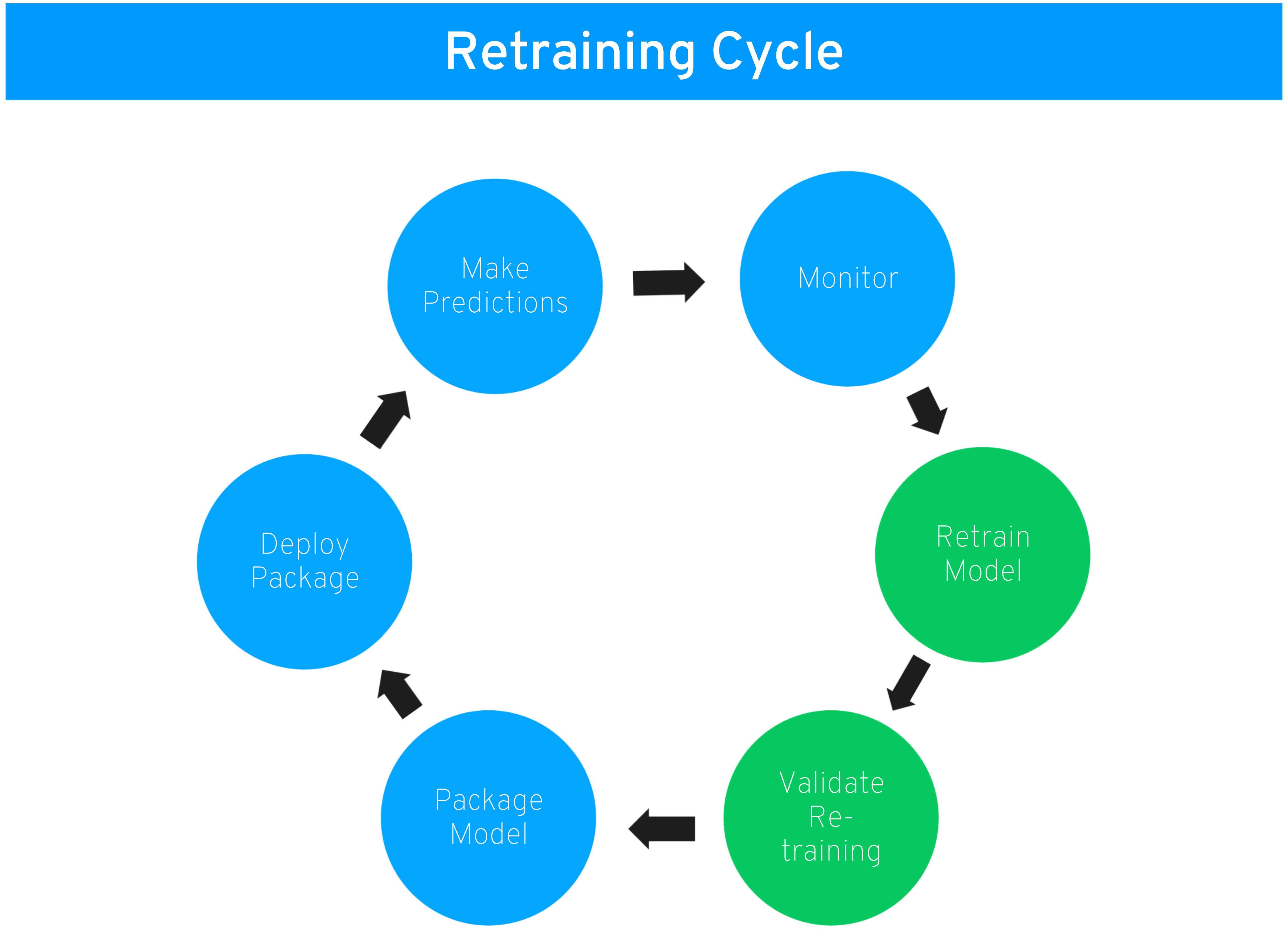
Retraining Cycle Testing Concerns

Common Concerns

- Same concerns as training in development process

Best Practices

- Go back to development process if model performance is sizably worse than previous iterations
- Track models in a model registry tool



Other machine learning concepts related to testing

BIAS

The idea that the model produces results that are systemically prejudiced due to erroneous assumptions in the build process

TOXICITY

The idea that the model can produce results that are unintentionally harmful when used in an uncontrolled environment

In Summary

- Stochastic nature is a feature not a defect
- Use the traditional testing methods when appropriate
- Work in partnership with data science counterparts on other aspects

Questions?

Steven Devoe
steven.devoe@spr.com
https://www.linkedin.com/in/stevendevoe/