CMPT 733 – Big Data Programming II

Automated Machine Learning (AutoML)

Instructor	Steven Bergner
Course website	https://sfu-db.github.io/bigdata-cmpt733/
Slides by:	Lydia Zheng and Jiannan Wang

Motivation

1. Machine learning is very **successful**
2. To build a traditional ML pipeline:
   - Domain experts with longstanding experience
   - Specialized data preprocessing
   - Domain-driven meaningful feature engineering
   - Picking right models
   - Hyper-parameter tuning
   - ……

H20 Driverless AI Demo

https://www.youtube.com/watch?v=ZqCoFp3-rGc

1. Will AutoML software replace Data Scientists?
2. How to approach AutoML as a data scientist?

AutoML Vision

For Non-Experts

AutoML allows non-experts to make use of machine learning models and techniques without requiring to become an expert in this field first

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

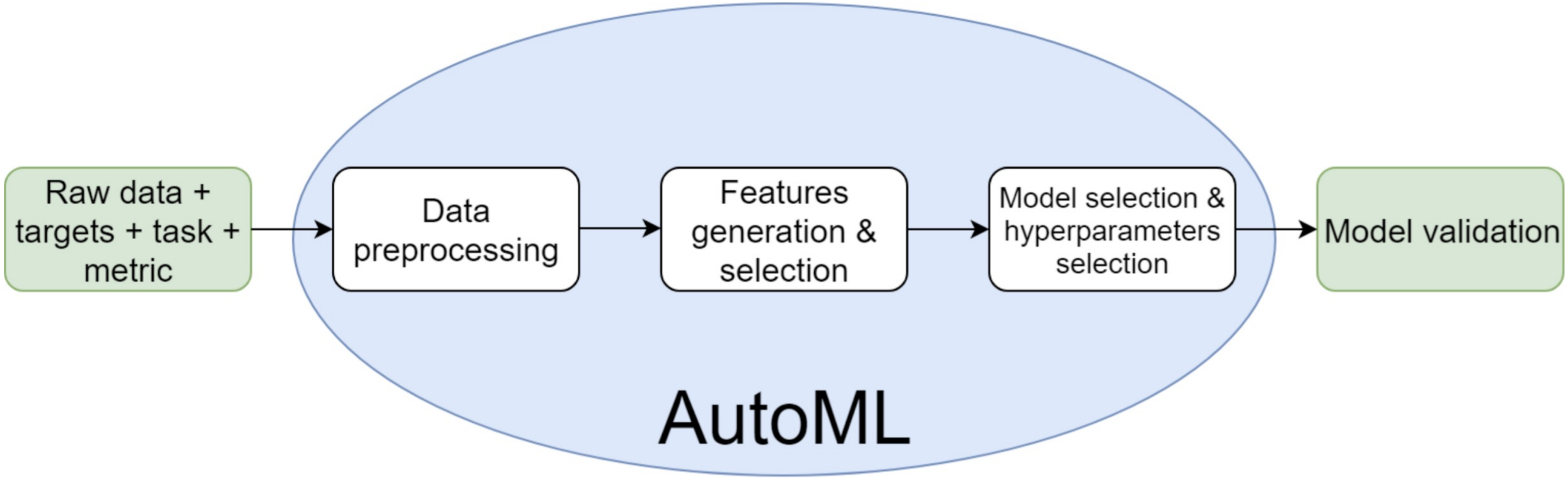
For Data Scientists

AutoML aims to augment, rather than automate, the work and work practices of heterogeneous teams that work in data science.

Wang, Dakuo, et al. "Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI." Proceedings of the ACM on Human-Computer Interaction 3.CSCW (2019): 1-24.

What is AutoML?

- Automate the process of applying machine learning to real-world problems


Raw data + targets + task + metric → Data preprocessing → Features generation & selection → Model selection & hyperparameters selection → Model validation

Outline

- Auto Feature Selection (Lecture 6)
- Auto Hyperparameter Tuning (Lecture 6)
- Auto Feature Generation (This Lecture) Neural Architecture Search (This Lecture)

Auto Feature Generation

Motivation

- The model performance is heavily dependent on quality of features in dataset
- It’s time-consuming for domain experts to generate enough useful features

Feature Generation

- Unary operators (applied on a single feature)
  - Discretize numerical features
  - Apply rule-based expansions of dates
  - Mathematical operators (e.g., Log Function)

- Higher-order operators (applied on 2+ features)
  - Basic arithmetic operations (e.g., +, -, ×, ÷)
  - Group-by Aggregation (e.g., GroupByThenAvg, GroupByThenMax)

Katz, Gilad, Eui Chul Richard Shin, and Dawn Song. "Explorekit: Automatic feature generation and selection." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.

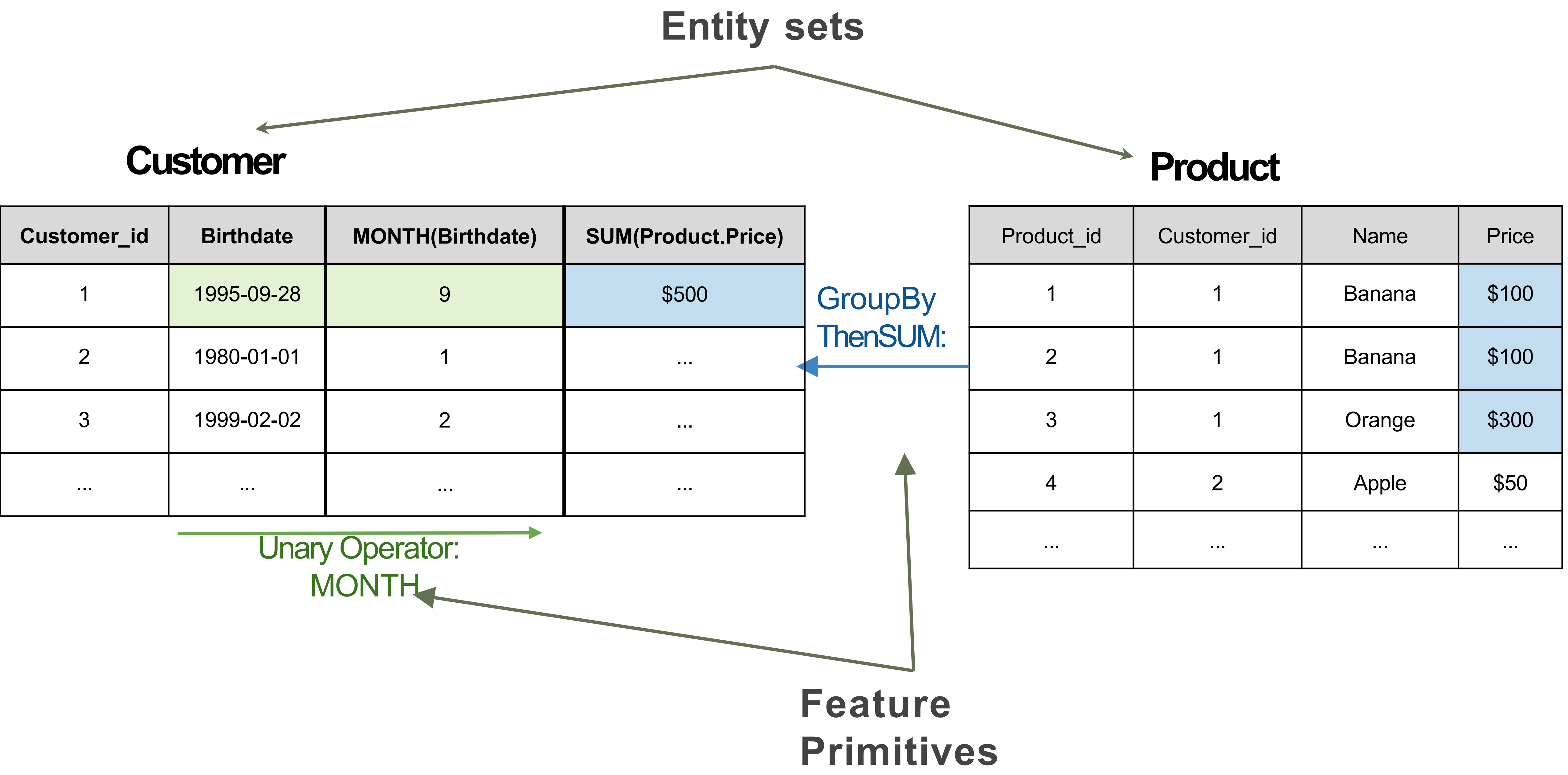
Featuretools

- An open source library for performing automated feature engineering
- Design to fast-forward feature generation across **multi-relational** tables

Concepts

- Entity is the relational tables
- An EntitySet is a collection of entities and the relationships between them
- Feature Primitives
  - Unary Operator: transformation (e.g., MONTH)
  - High-order Operator: Group-by Aggregation (e.g., GroupByThenSUM)

Entity sets



Customer

Customer_id	Birthdate	MONTH(Birthdate)	SUM(Product.Price)
1	1995-09-28	9	$500
2	1980-01-01	1	...
3	1999-02-02	2	...
...	...	...	...

Unary Operator:
MONTH

GroupBy
ThenSUM:

Product

Product_id	Customer_id	Name	Price
1	1	Banana	$100
2	1	Banana	$100
3	1	Orange	$300
4	2	Apple	$50
...	...	...	...

Feature
Primitives

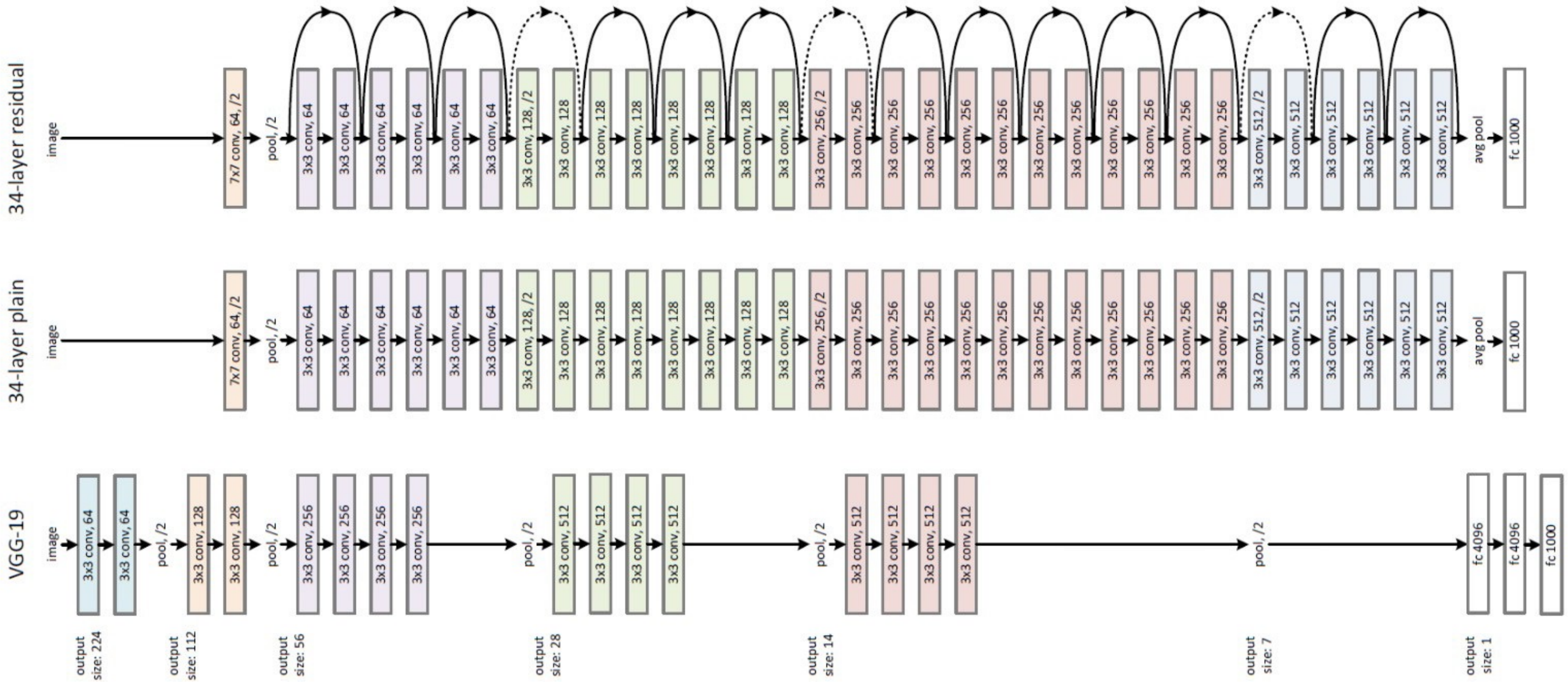
Outline

- Auto Feature Selection (Lecture 5)
- Auto Hyperparameter Tuning (Lecture 5)
- Auto Feature Generation (This Lecture) Neural Architecture Search (This Lecture)

Neural Architecture Search (NAS)

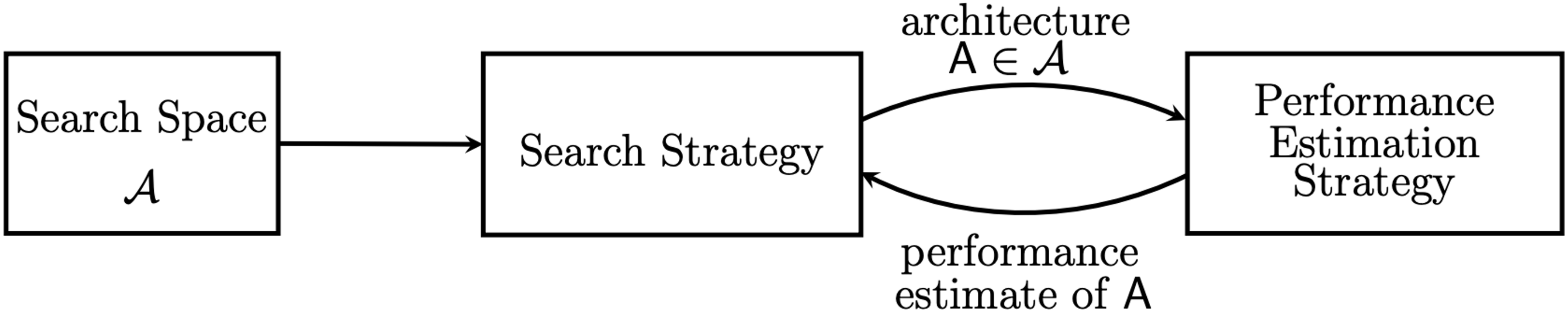
Motivation

How can someone come out with such an architecture?



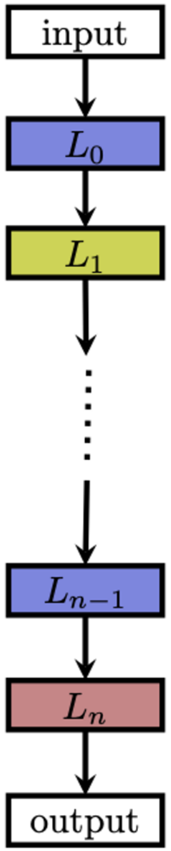
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." CVPR. 2016

Neural Architecture Search : Big Picture

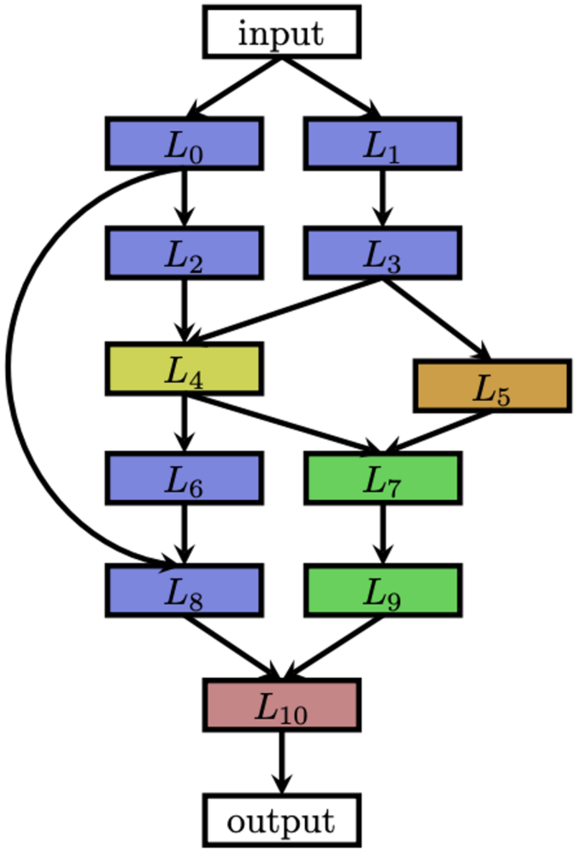


Search Space

- Define which neural architectures a NAS approach might discover in principle
- May have human bias → prevent finding novel architectural building blocks



Chain-structured



Multi-branch

Search Strategy

- **Basic Idea**
  - Explore search space (often exponentially large or even unbounded)
- **Methods**
  - Random Search
  - Bayesian Optimization [Bergstra et al., 2013]
  - Evolutionary Methods [Angeline et al., 1994]
  - Reinforcement Learning [Baker et al., 2017]
  - .....

Performance Estimation Strategy

- **Basic Idea**
  - The process of estimating predictive performance
- **Methods**
  - Simplest option: perform a training and validation of the architecture on data
  - Initialize weights of novel architecture based on weights of other architectures have been trained before
  - Using learning curve extrapolation [Swersky et al., 2014]
  - ……

Summary

What is AutoML and why we need it?

How AutoML works?

- Auto Feature Selection (Lecture 5)
- Auto Hyperparameter Tuning (Lecture 5)
- Auto Feature Generation (This Lecture)
- Neural Architecture Search (This Lecture)