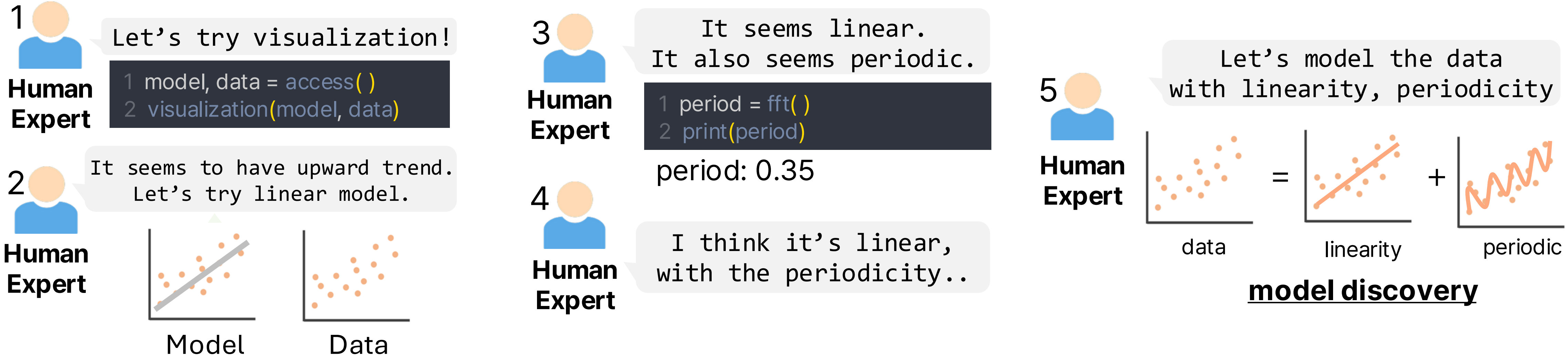
Automated Model Discovery via Multi-modal & Multi-step Pipeline

Lee Jung-Mok[1] Nam Hyeon-Woo[1] Moon Ye-Bin[1] Junhyun Nam[2] Tae-Hyun Oh[3]

[1]POSTECH [2]Samsung Electronics [3]KAIST

Task Definition

- Data Analysis by Model Discovery
  - Find the **interpretable model** that best express data for understanding data
  - Interpretable models: models composited with interpretable basis functions
  - Application: data scientists **manually finds model** through trial and error



Manual Model Discovery by Human Expert

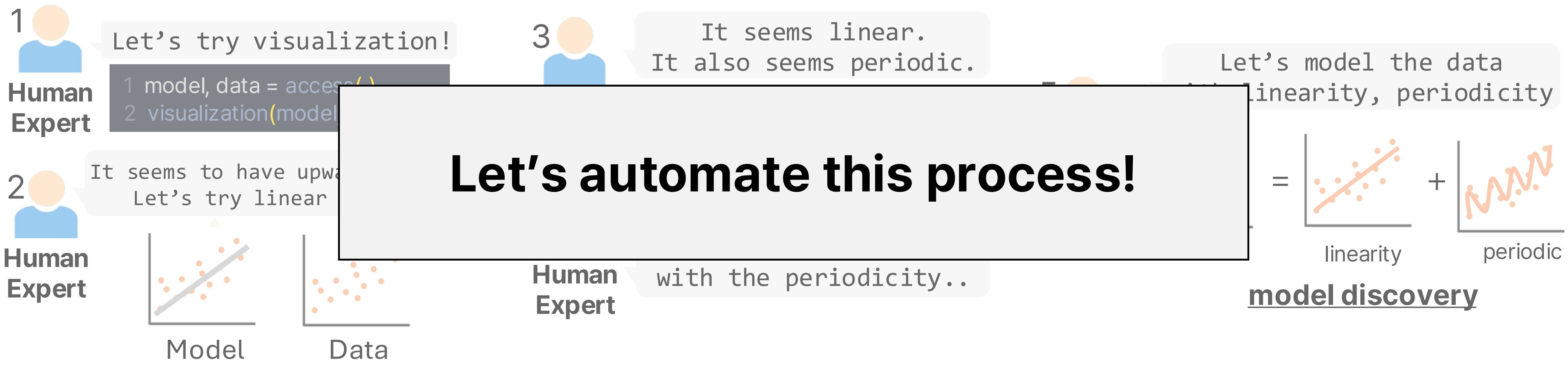
[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013

[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013

[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

Task Definition

- Data Analysis by Model Discovery
  - Find the **<u>interpretable model</u>** that best express data for understanding data
  - Interpretable models: models composited with interpretable basis functions
  - Application: data scientists **<u>manually finds model</u>** through trial and error



Manual Model Discovery by Human Expert

[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013
[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013
[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

Preliminary: Gaussian Process (GP)

- Gaussian Process Models
  - GP is specified by mean function $\mu(x)$, covariance function $\mathrm{k}(x, x')$
  - Assuming the mean function $\mu(x)$ to be zero, we **construct the covariance function (model)**

$$f \sim \mathcal{GP}(\mu(x), \boxed{k(x, x')})$$

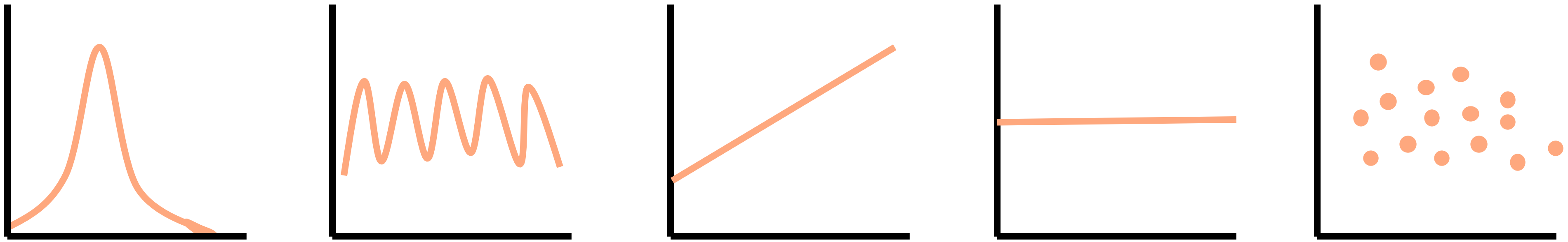
Mean Function

Covariance Function (**Model**)

[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013
[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013

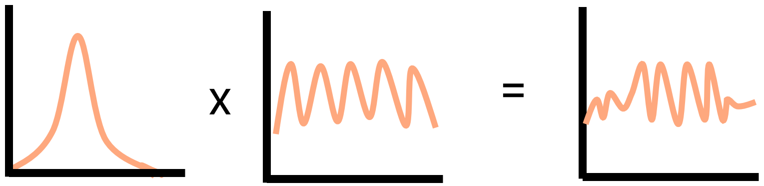
Preliminary: GP Model Discovery

- Basis Models of $k(x, x')$



Locality (SE) Periodicity Linearity Constant White Noise

- Constructing model composition
  - Adding (+) two basis models
  - Multiplying (x) two basis models
  - ChangePoint (CP) two basis models



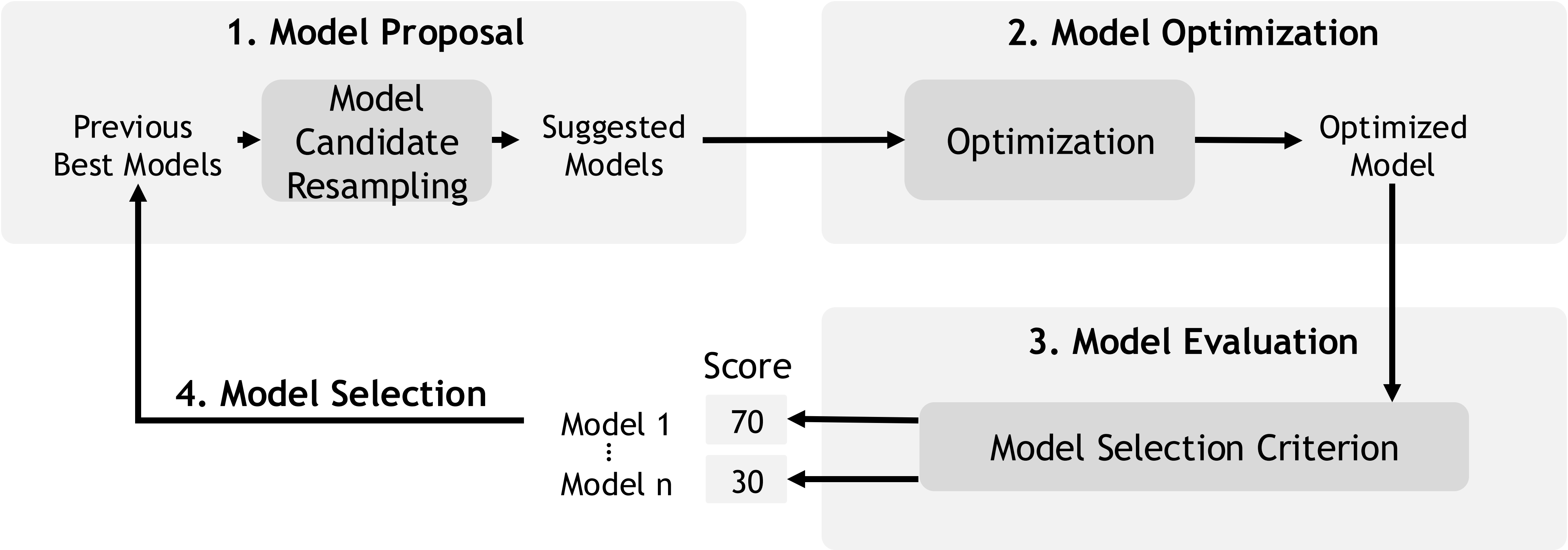
Locality (SE) Periodicity New Model Composition

Example model composition

[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013

Model Discovery Pipeline

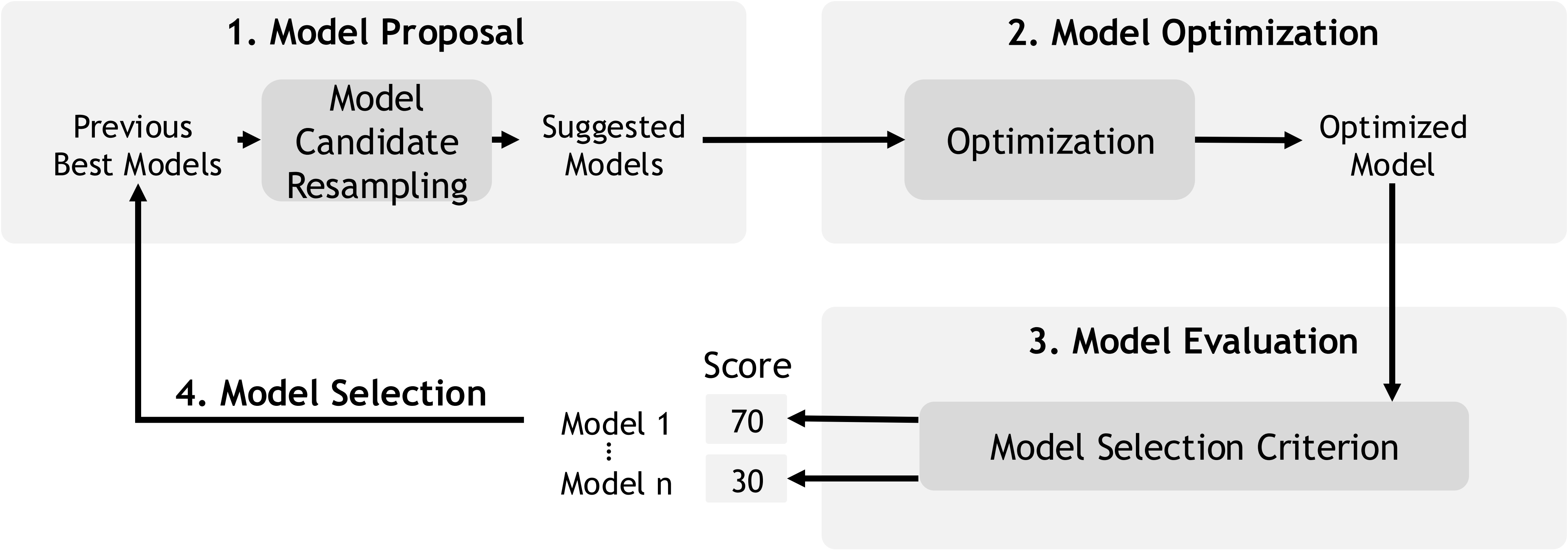
- Automated Model Discovery Pipeline [1, 2, 3]
  - **1) Model proposal**: proposing new model compositions
  - **2) Model optimization**: parameter fitting of proposed model composition



[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013
[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013
[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

Model Discovery Pipeline

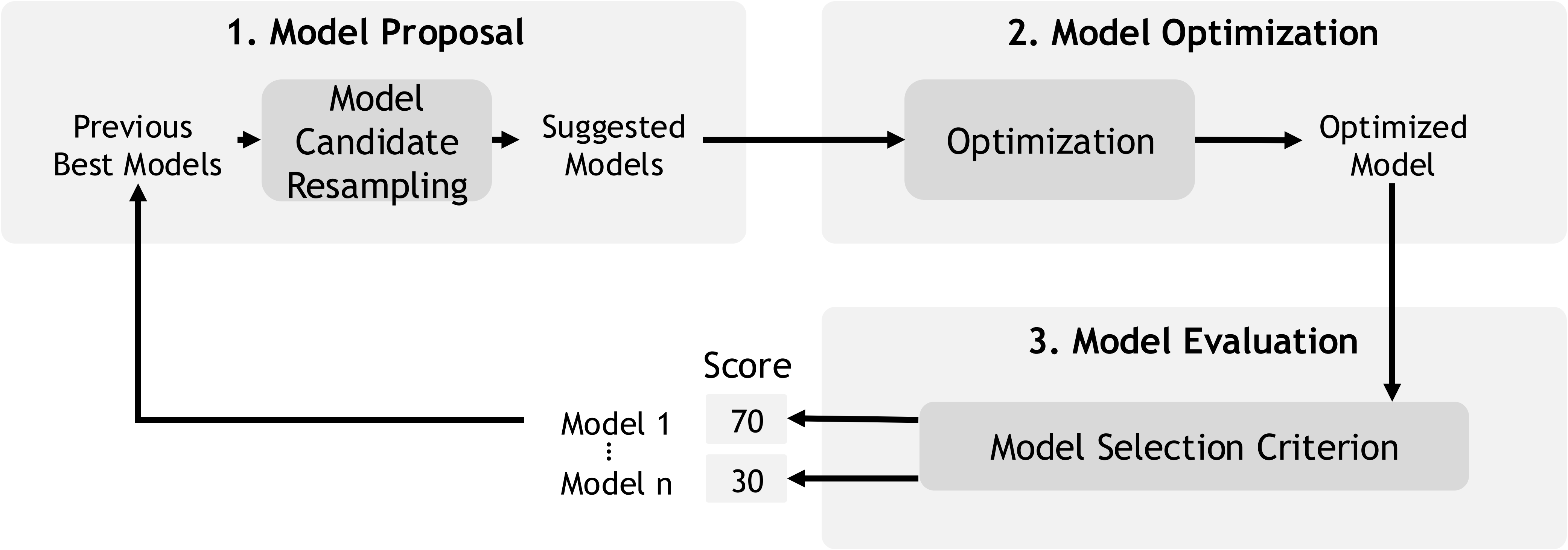
- Automated Model Discovery Pipeline [1, 2, 3]
  - **3) Model evaluation**: scoring each proposed models
  - **4) Model selection**: selecting best model among proposed models



[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013
[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013
[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

Model Discovery Pipeline

- Previous Model Discovery Pipeline [1, 2, 3]
  - [1, 2] adopts greedy search with **all model candidate resampling**
  - **[3] adopts LLM** for model proposal and evaluation



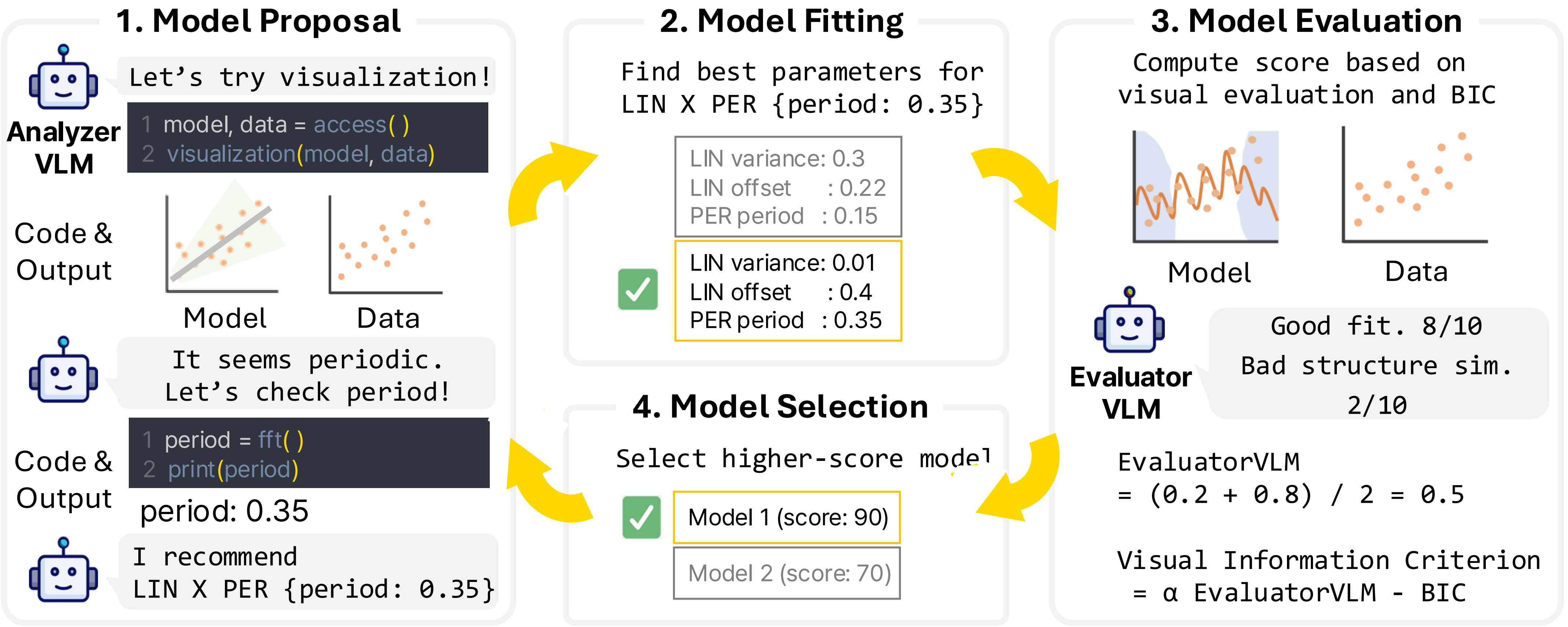
[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013

[2] Lloyd et al., Automatic Construction and Natural-Language Description of Nonparametric Regression Models, AAAI 2013

[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

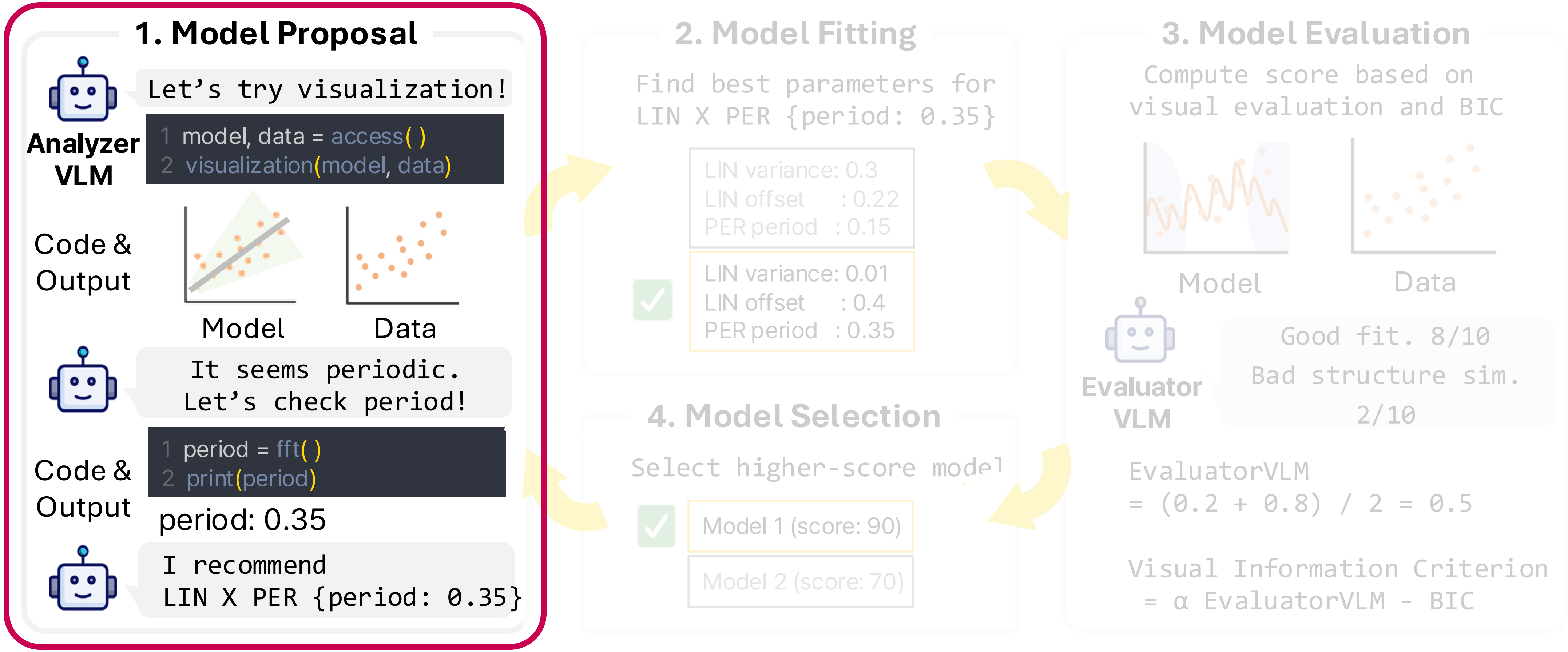
Proposed Pipeline

- Model discovery through **multi-step multi-modal agents**



1. Model Proposal

- **AnalyzerVLM** conducts multi-modal & multi-step analysis



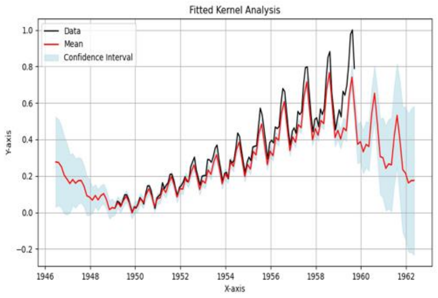
1. Model Proposal

- **AnalyzerVLM** conducts multi-modal & multi-step analysis
  - In each step, model chooses action: **Analyze** / **Execute** / **Propose**

step 1

Given the access code, visualize the data and model.

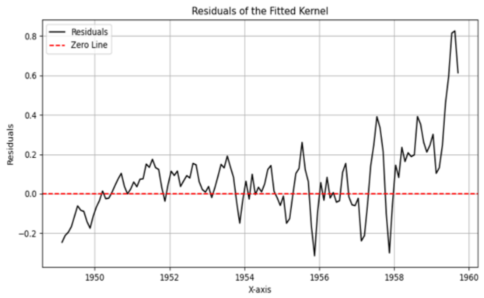
```
model, data = access( )
visualization(model, data)
```



step 2

Mostly follows the data, but some gap at the end. Check residual.

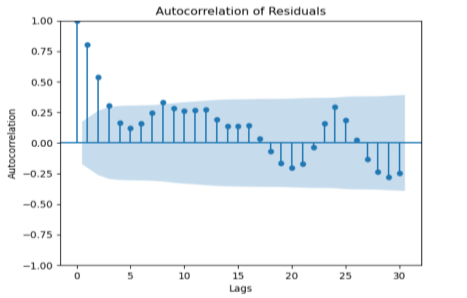
```
model, data = access( )
check_residual(model, data)
```



step 3

Residual at latter years. Check autocorrelation.

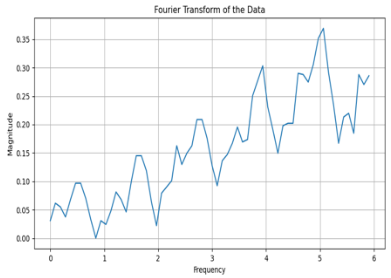
```
model, data = access( )
autocorrelation(model, data)
```



step 4

Big at first lags, shows feature. Visualize periodicity.

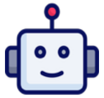
```
model, data = access( )
periodicity(model, data)
```



step 5

Enough analysis. Propose model with periodicity.

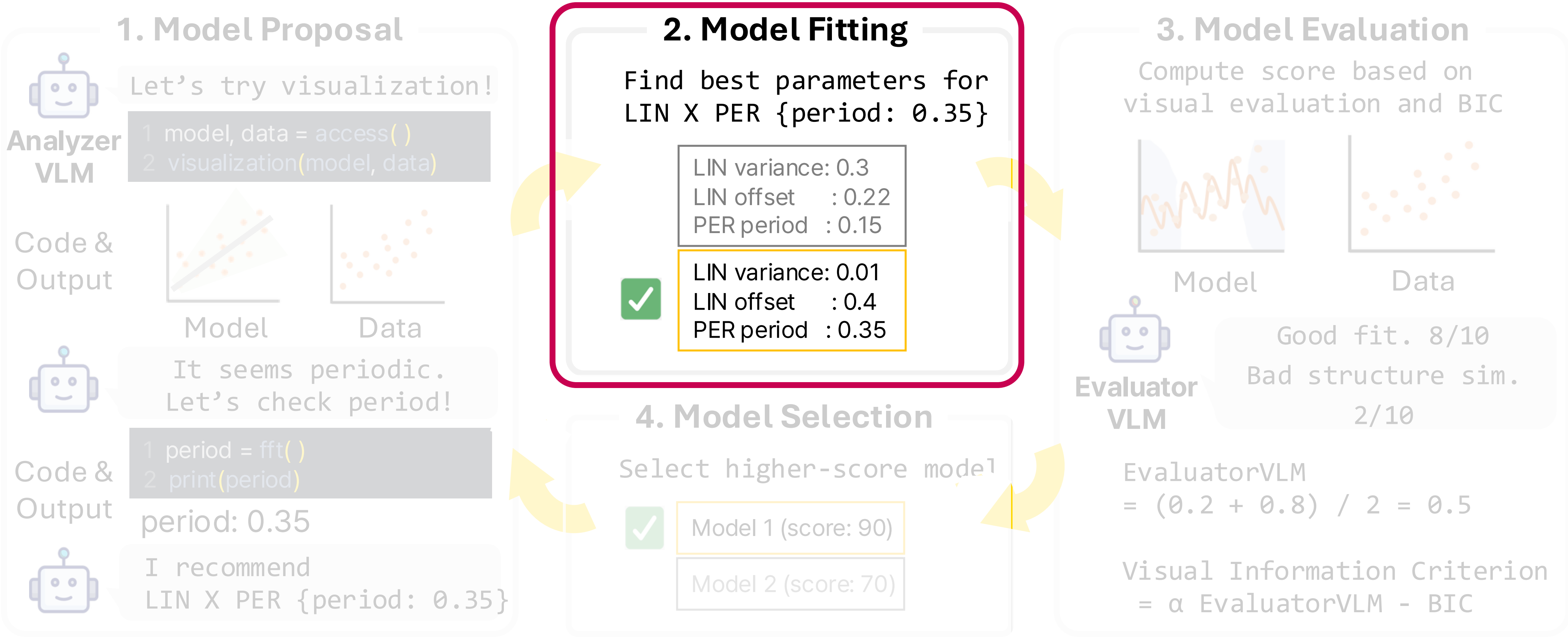
Model Proposal



LIN X PER

2. Model Fitting

- Model Parameter Fitting

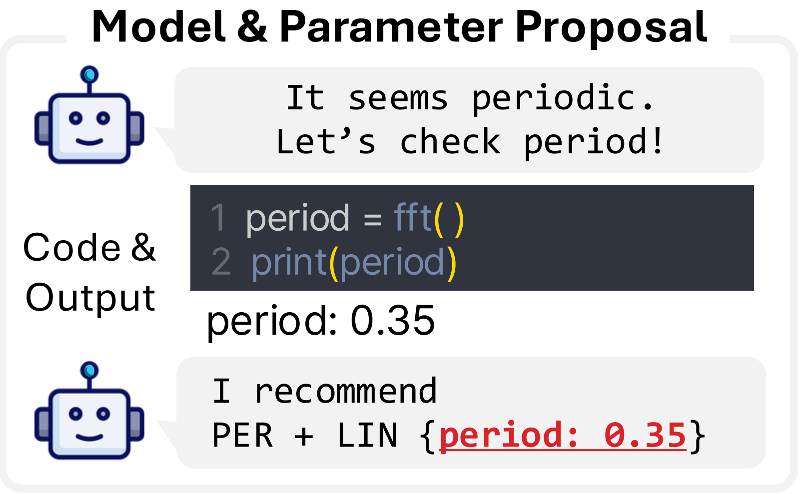


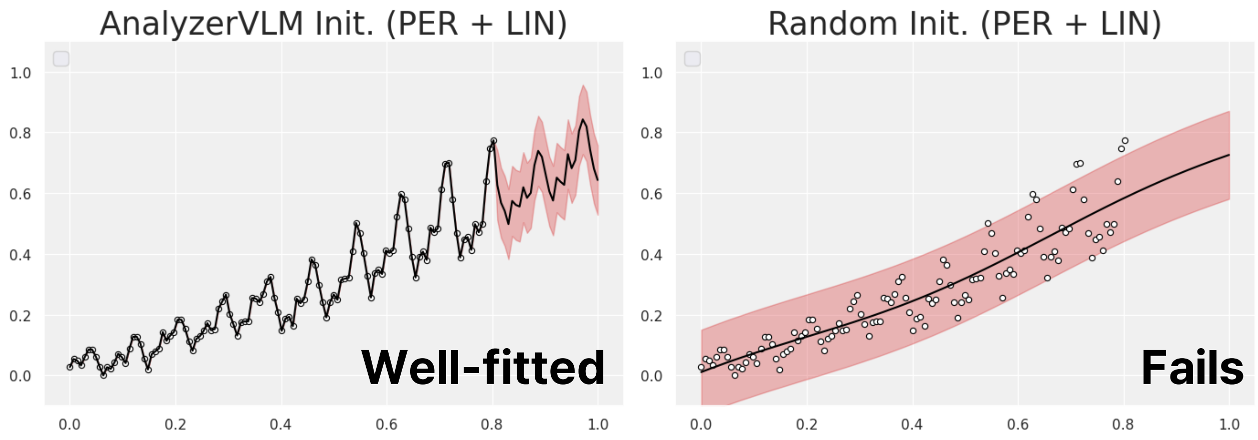
2. Model Fitting

- Model Parameter Fitting
  - Parameter fitting is done by **<u>maximizing log likelihood</u>**

$$\log p(\mathbf{y}|X,\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi$$

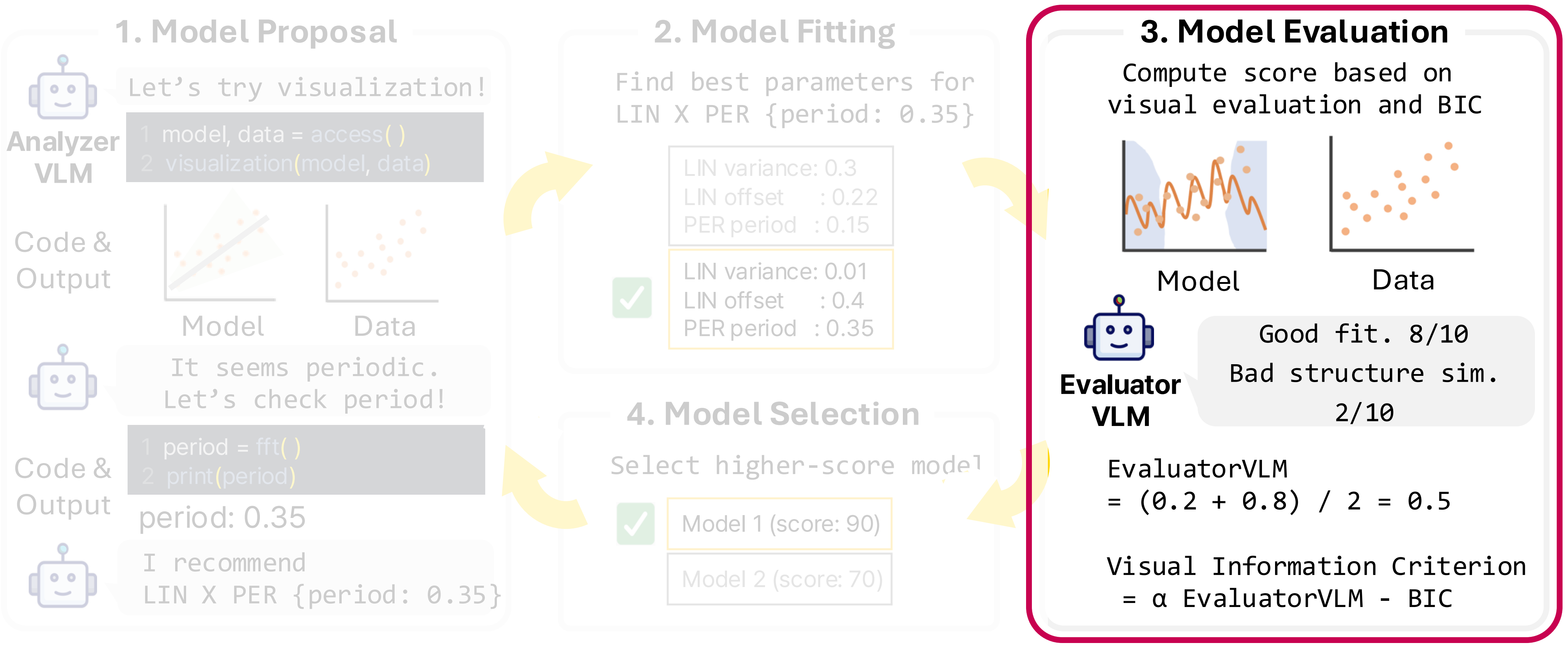
- Model parameter initialization





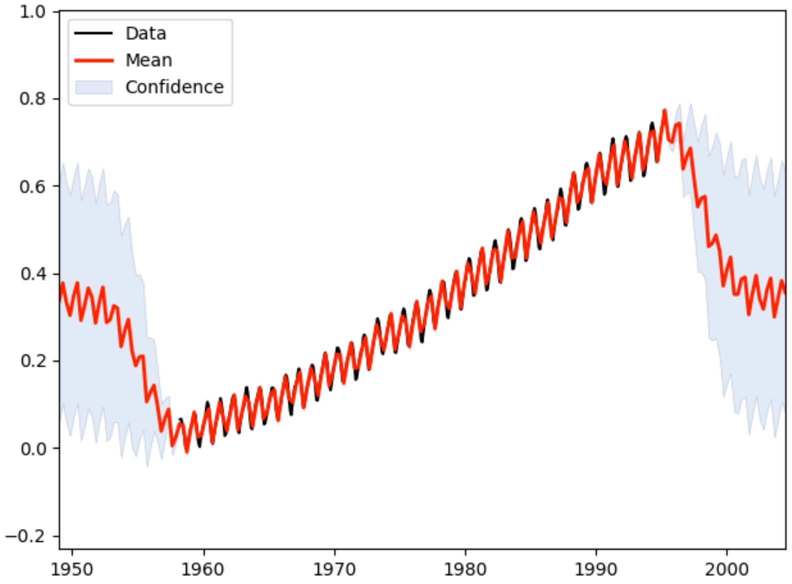
3. Model Evaluation

- **EvaluatorVLM**: Visiual Model Evaluation



3. Model Evaluation

- **EvaluatorVLM**: Visiual Model Evaluation



Visual Fitness

Evaluator VLM

Model's prediction is well-aligned with the data. **Good fitness.**

Visual Generalizability

Evaluator VLM

High uncertainty at both side, flattened predictions. **Bad generalizability.**

3. Model Evaluation

- Bayesian Information Criterion (BIC)

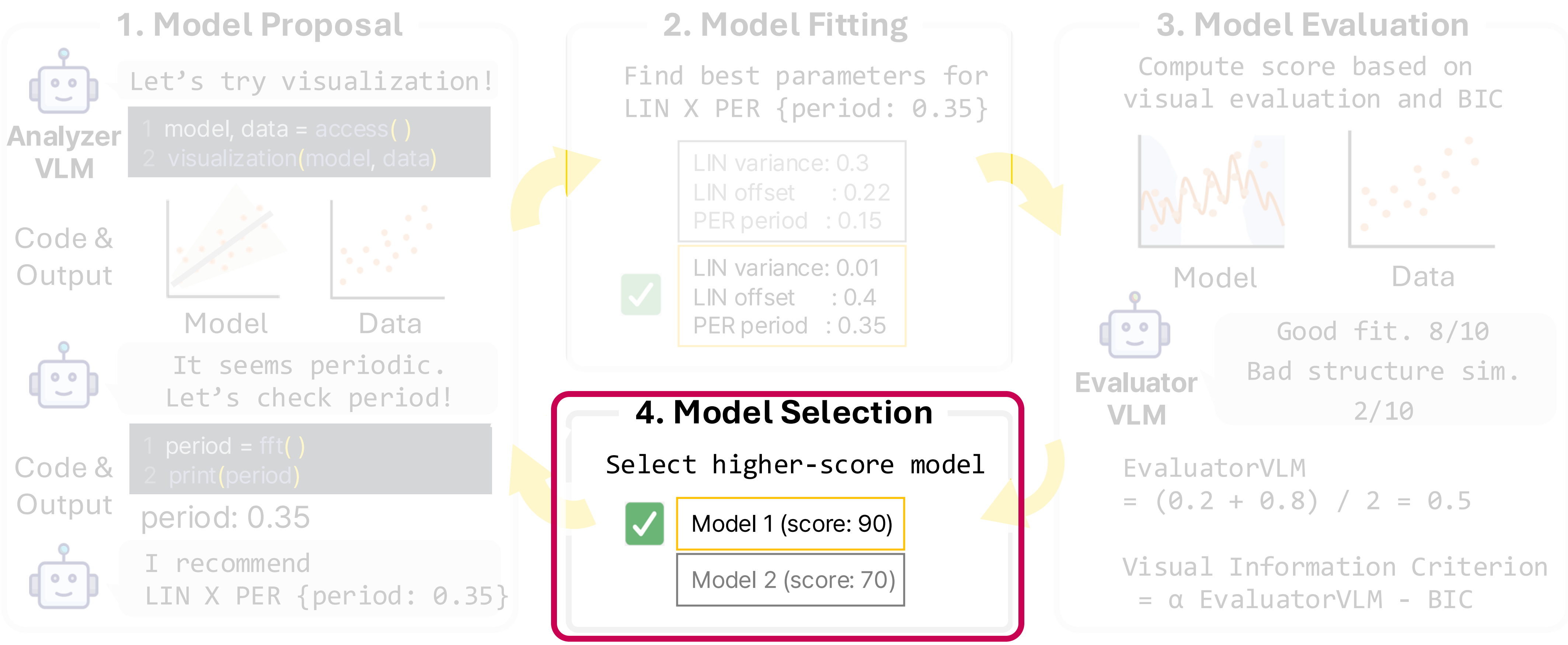
$$BIC = -2\log p(\underset{\text{data}}{D}|\underset{\text{model}}{M}) + \underset{\substack{\text{\# of model}\\ \text{parameters}}}{|M|}\log \underset{\text{data size}}{|D|}$$

- **<u>Visual Information Criterion (VIC)</u>**
  - We combine BIC with EvaluatorVLM's evaluation score

$$VIC = \alpha \cdot \underbrace{EvaluatorVLM(D, M, \theta^*)}_{\textbf{visual fitness + visual generalizability}} - BIC$$

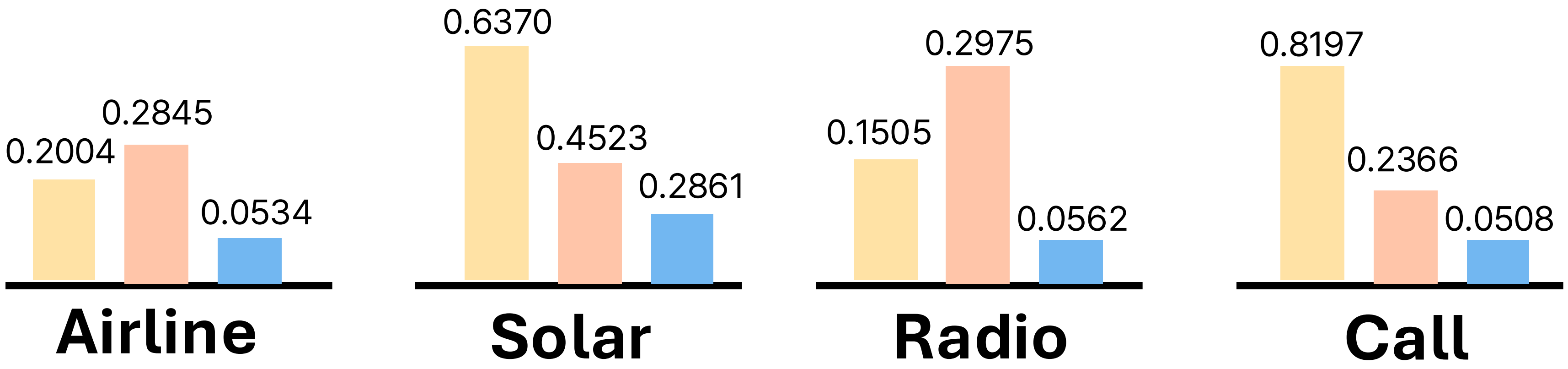
4. Model Selection

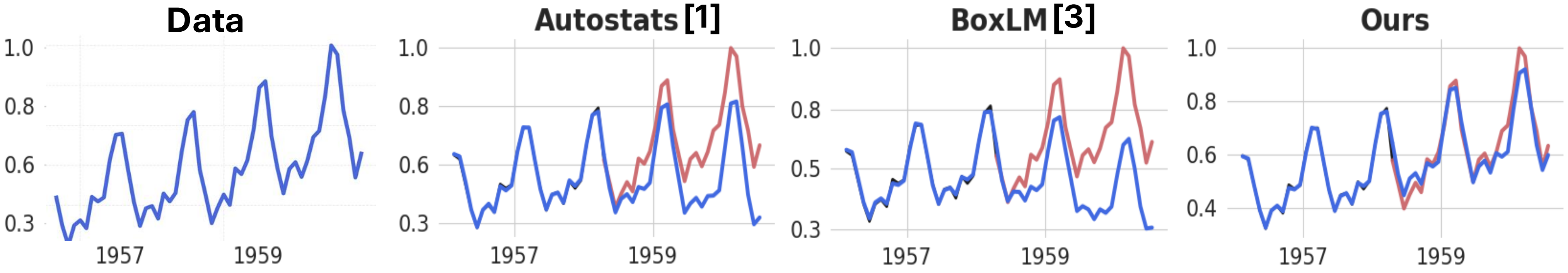
- Sort the model pool, and select best model



Experiment Results

- Our method's model prediction **<u>outperforms competing methods</u>**
  - Discovered model fits well to the data and has high generalizability



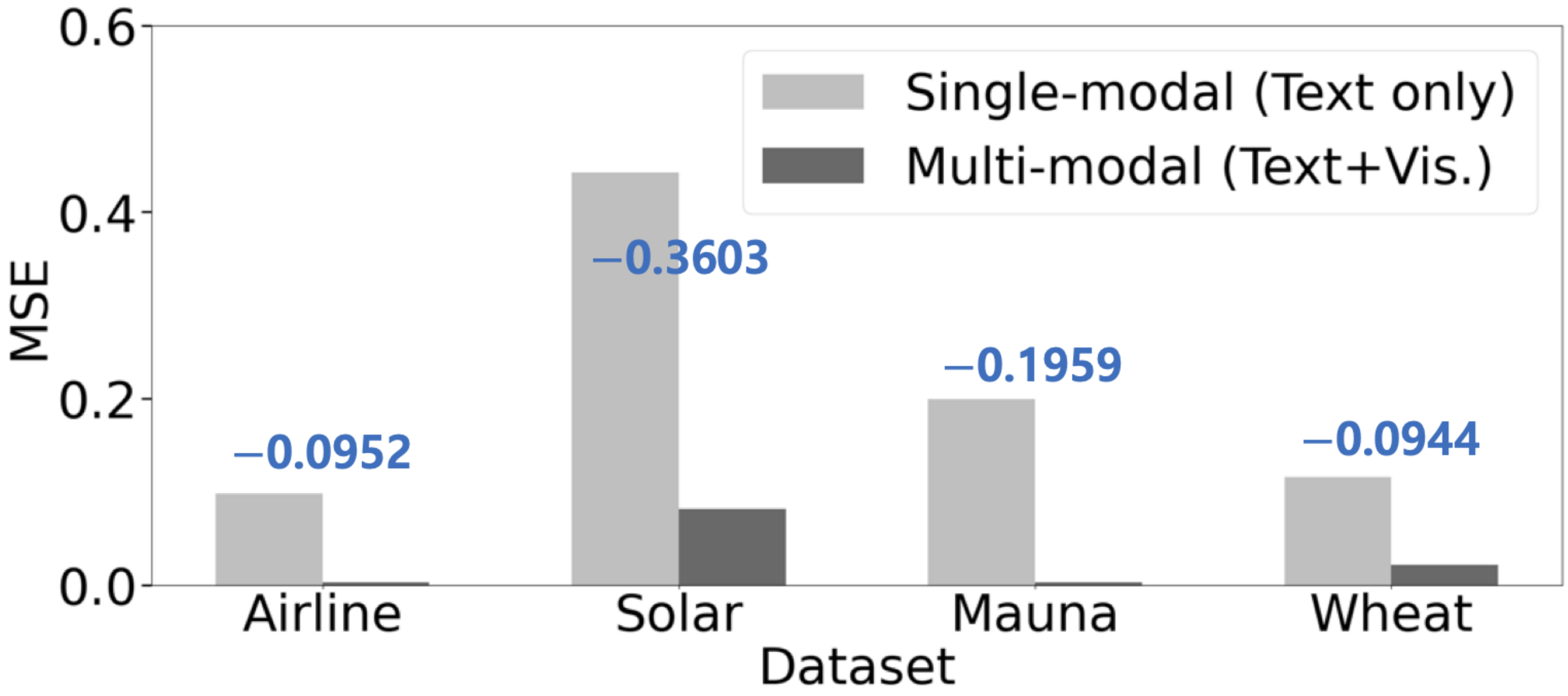


[1] Duvenaud et al., Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013

[3] Li et al., Automated Statistical Model Discovery with Language Models, ICML 2024

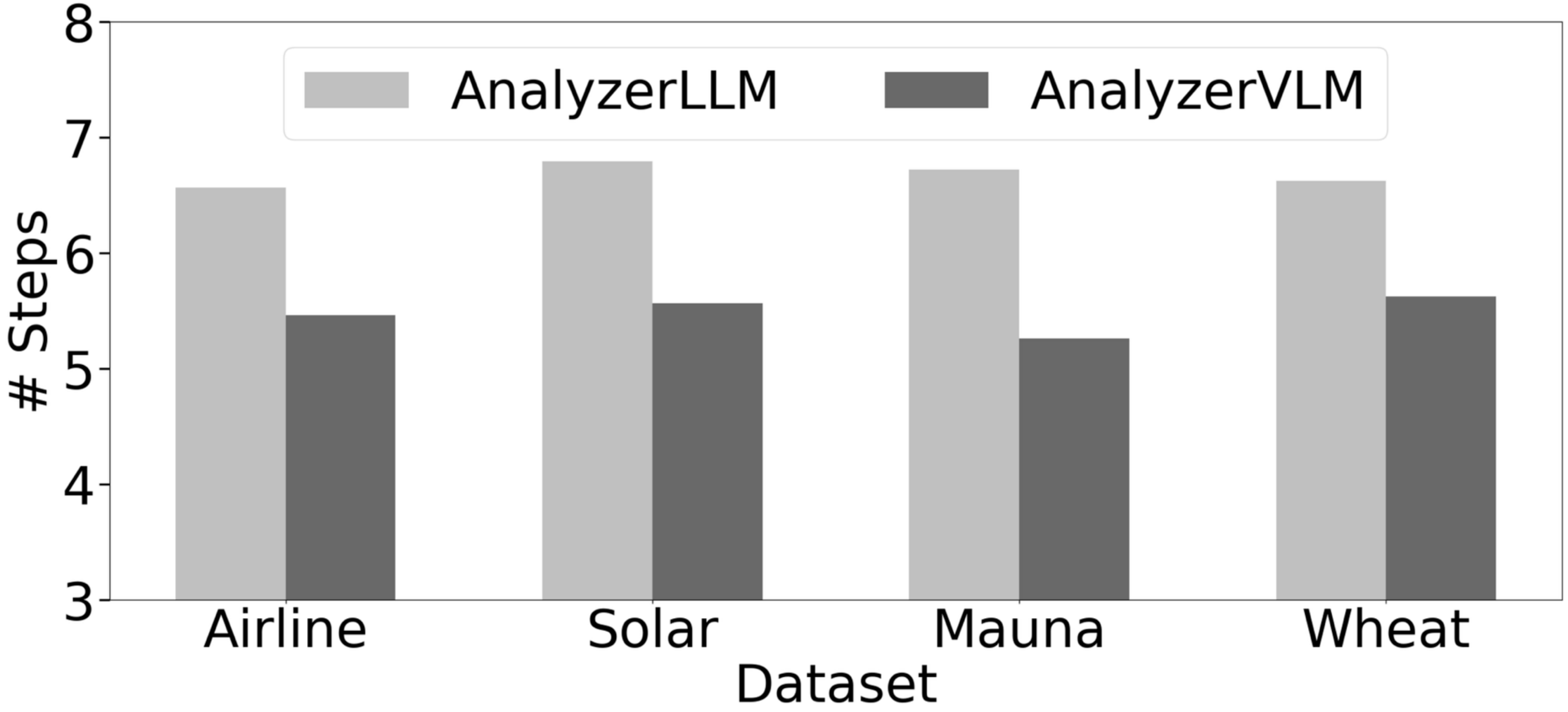
Ablation Study

- Without visual reasoning, it **shows downgraded performance**
- AnalyzerLLM is step-heavy due to inefficient textual reasoning



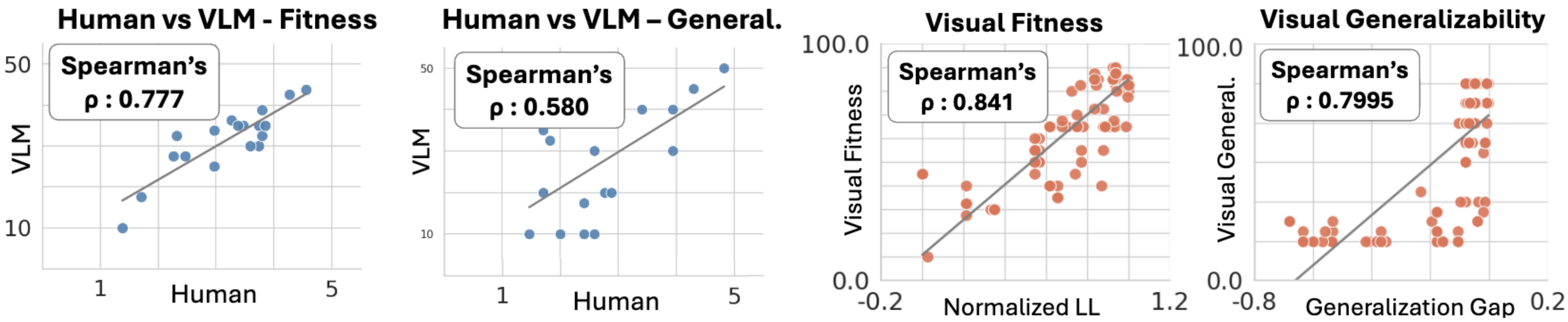
Ablation Study

- Without visual reasoning, it shows downgraded performance
- AnalyzerLLM is step-heavy due to **inefficient textual reasoning**



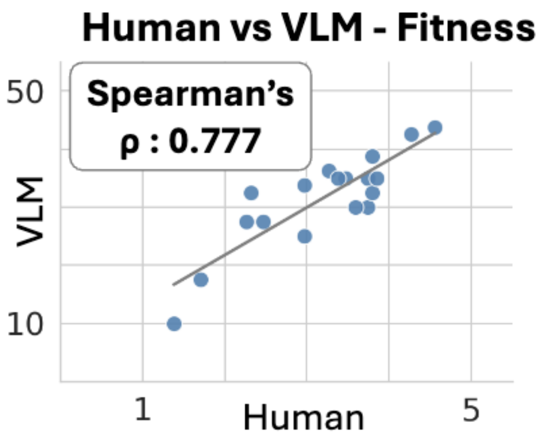
Correlation with VIC

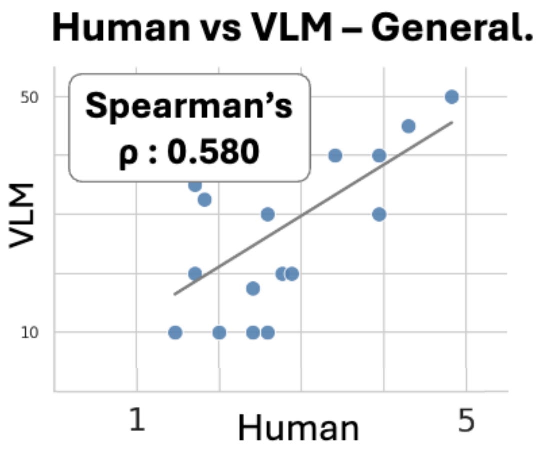
- Correlation of Visual Information Criterion (VIC)
  - Visual fitness and generalizability both **correlate highly with human scoring**

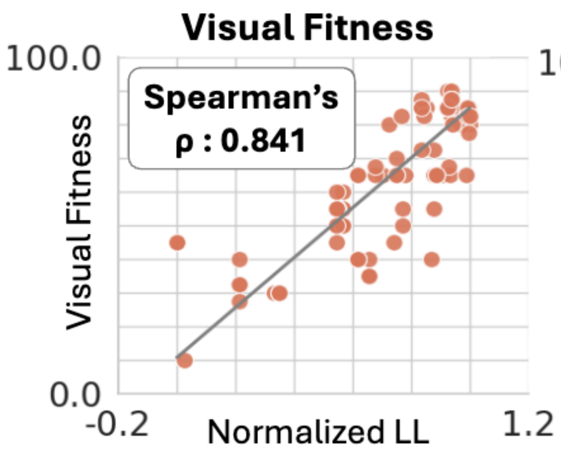


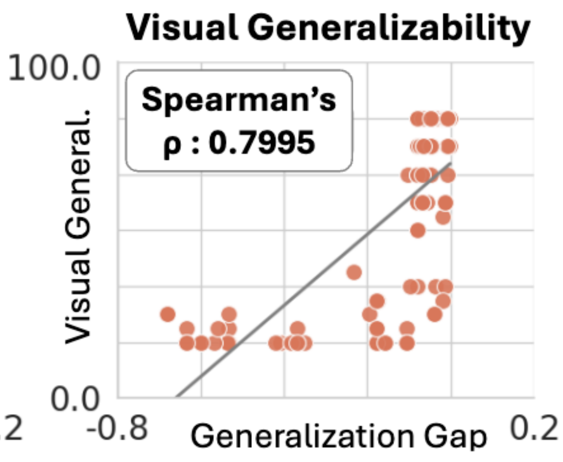
Correlation with VIC

- Correlation of Visual Information Criterion (VIC)
  - **Visual fitness**: shows high correlation with **normalized log-likelihood**
  - **Visual generalizability**: shows high correlation with **generalization gap**









Conclusion

- We propose **multi-modal, multi-step pipeline** for automated model discovery
  - **AnalyzerVLM** iteratively analyzes data to propose the most suitable model
  - **EvaluatorVLM** conducts visual evaluation: visual fitness, visual generalizability

- Through **visual reasoning**, AnalyzerVLM and EvaluatorVLM understands time-series patterns effectively
  - Which can lead to the **stronger generalization** compared to textual reasoning

Project page: https://mok0102.github.io/model-discovery
Email: jungmok@postech.ac.kr

Acknowledgement

This work was supported by Samsung Electronics Co., Ltd (Project Code: IO240508-09825-01).

T.-H. Oh was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00124, No.RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

T.-H. Oh was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

T.-H. Oh work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World).

T.-H. Oh was partially supported by the KAIST Cross-Generation Collaborative Lab Project.

Project page: https://mok0102.github.io/model-discovery

Email: jungmok@postech.ac.kr

Thank you

Any questions?