DATA SCIENCE LIFECYCLE FROM DATA TO IMPACT

Afshin Ashofteh
Link!

Nova Information Management School (Nova IMS)
Nova University of Lisbon

Email: aashofteh@novaims.unl.pt
Webpage: https://novaresearch.unl.pt/en/persons/afshin-ashofteh

25 February | 2.30 pm (GMT) | online

E-mail | Academic HP | ResearchGate | Discussion Group | YouTube

ABSTRACT:

This talk explores how data science and AI bring the impact of new data sources into our modern world through technologies like LLMs, AI agents, data spaces, web platforms, etc. Many professionals need new skills to automate, analyze, and optimize complex systems. Adopting these technologies requires updated knowledge, better methodologies, and improvements in quality, security, privacy, and legal frameworks. It also demands a diverse skill set for data scientists and AI specialists. This talk introduces a model integrating data science, data engineering, software development, Artificial Intelligence, and essential technical and soft skills for data professionals.

OUTLINE:

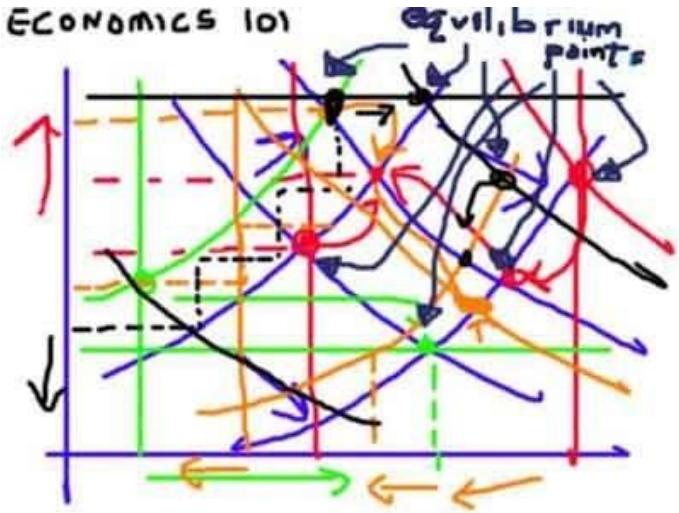
- Data Science Lifecycle
- New Technologies
- Examples

DATA ANALYTICS - HOUR 2:30!

TWO KINDS OF NOBELISTS IN FINANCIAL MODELING!

Academia

Specialists in the academy who have not practiced what they teach!



- Contain extremely elegant mathematics!
- Describes a world that does not exist!

(OVERFITING to PURE KNOWLEDGE)

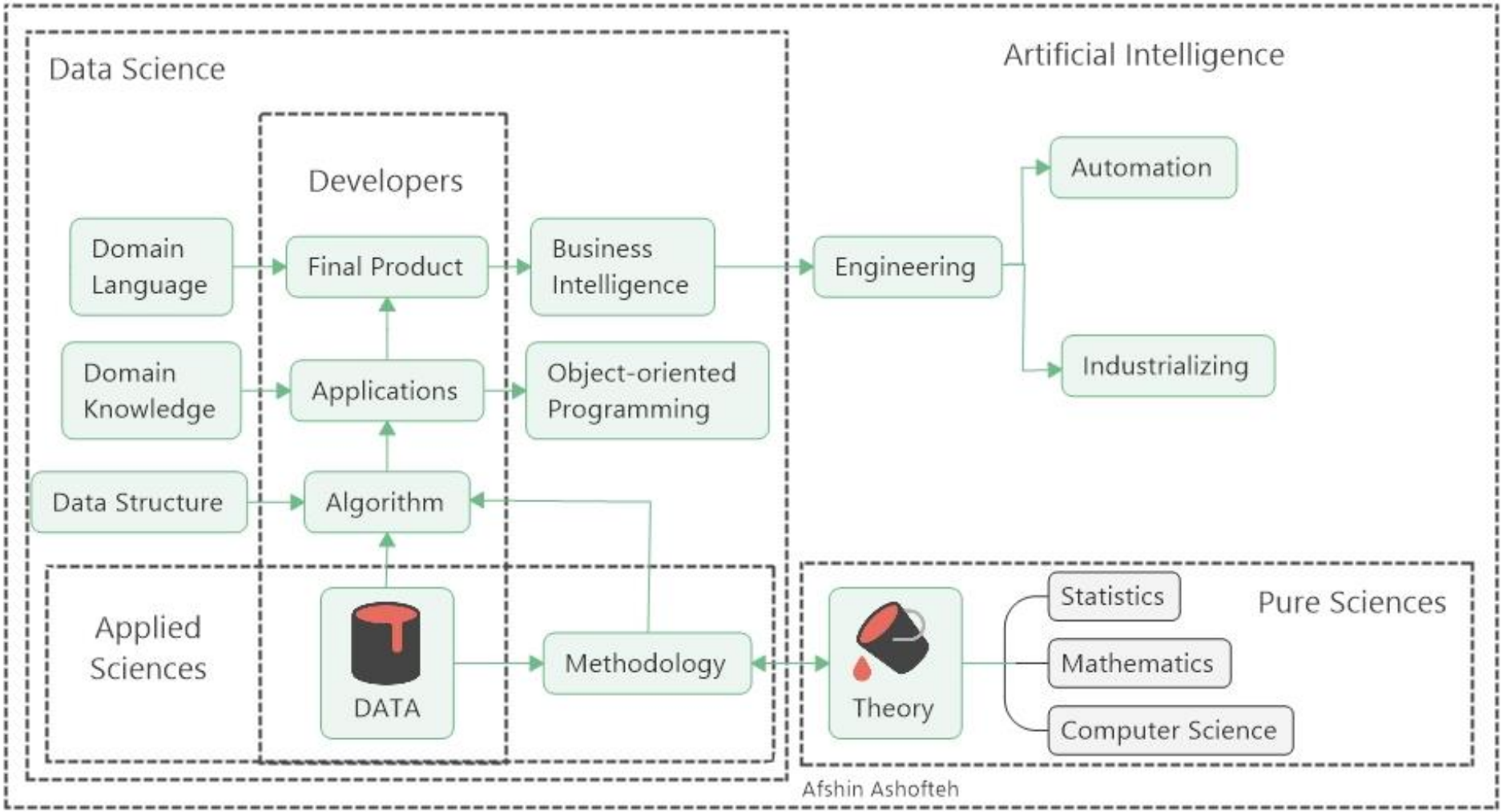
Industry

Specialists in the industry who offer explanations absent of any rigorous academic theory!

- Good fundamental analysts!
- Misuse mathematical tools to describe actual observations!

(OVERFITING to PRACTICAL KNOWLEDGE)

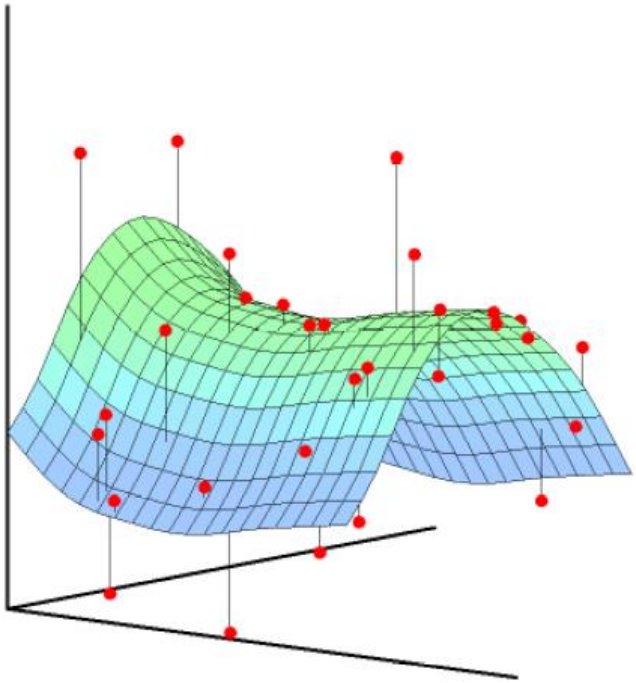
DATA SCIENCE SIMPLE LIFECYCLE



Link!

Ashofteh, A., & Bravo, J. M. (2021). Data Science Training for Official Statistics: a New Scientific Paradigm of Information and Knowledge Development in National Statistical Systems. Statistical Journal of the IAOS, 37(3), 771 – 789.

DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE

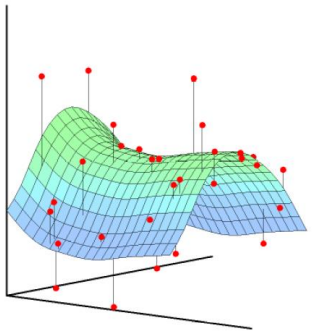


Observed Data

Real Target Model

$f_{\theta_0}(x)$

DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



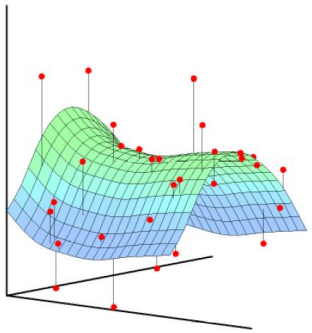
Candidate Models?

Real Target Model

$f_{\theta_0}(x)$

Observed Data

DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



Candidate Models?

?

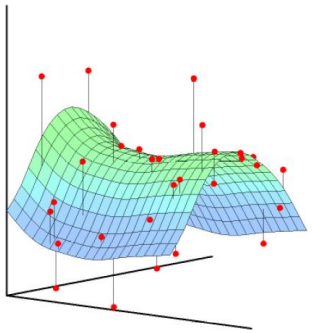
Distance?

Real Target Model

$f_{\theta_0}(x)$

Observed Data

DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



Candidate Models?

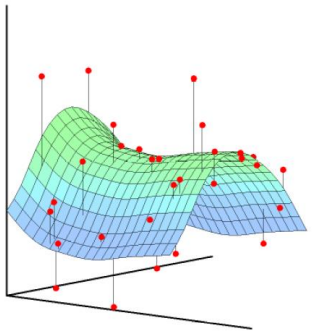
Standard Error

Real Target Model

$f_{\theta_0}(x)$

Observed Data

DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



A set of Algorithms with minimum S.E.

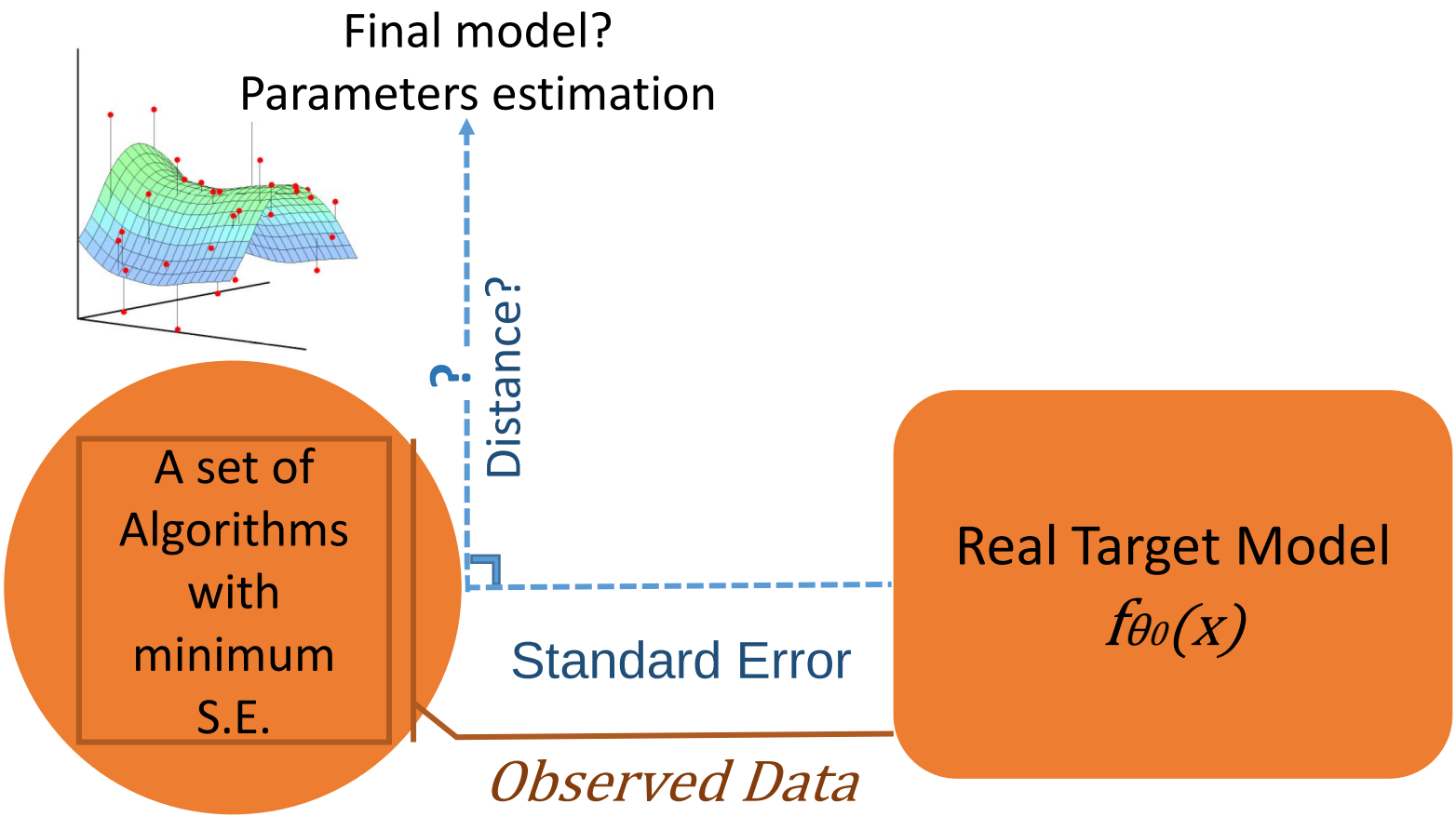
Standard Error

Observed Data

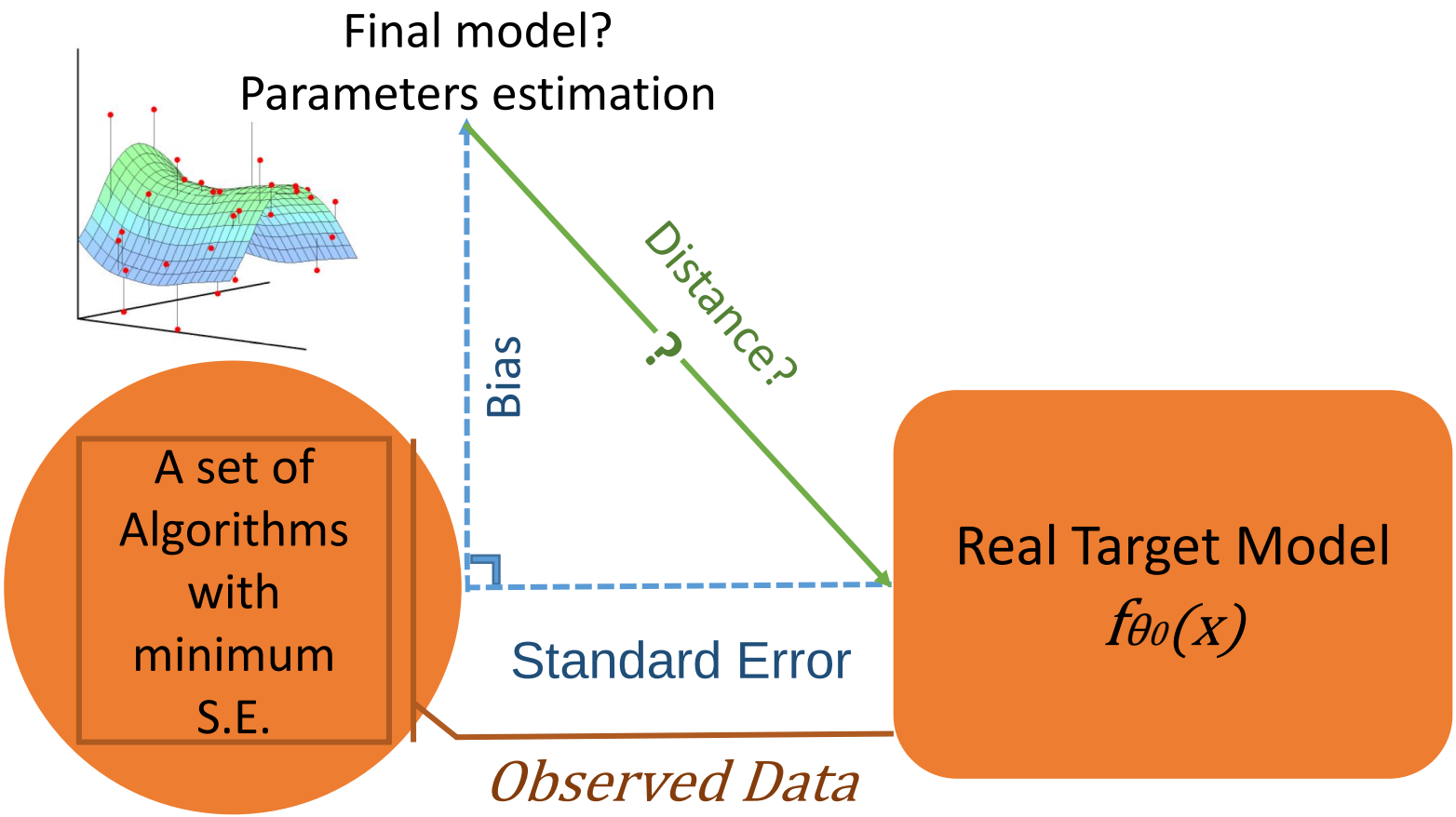
Real Target Model

$f_{\theta_0}(x)$

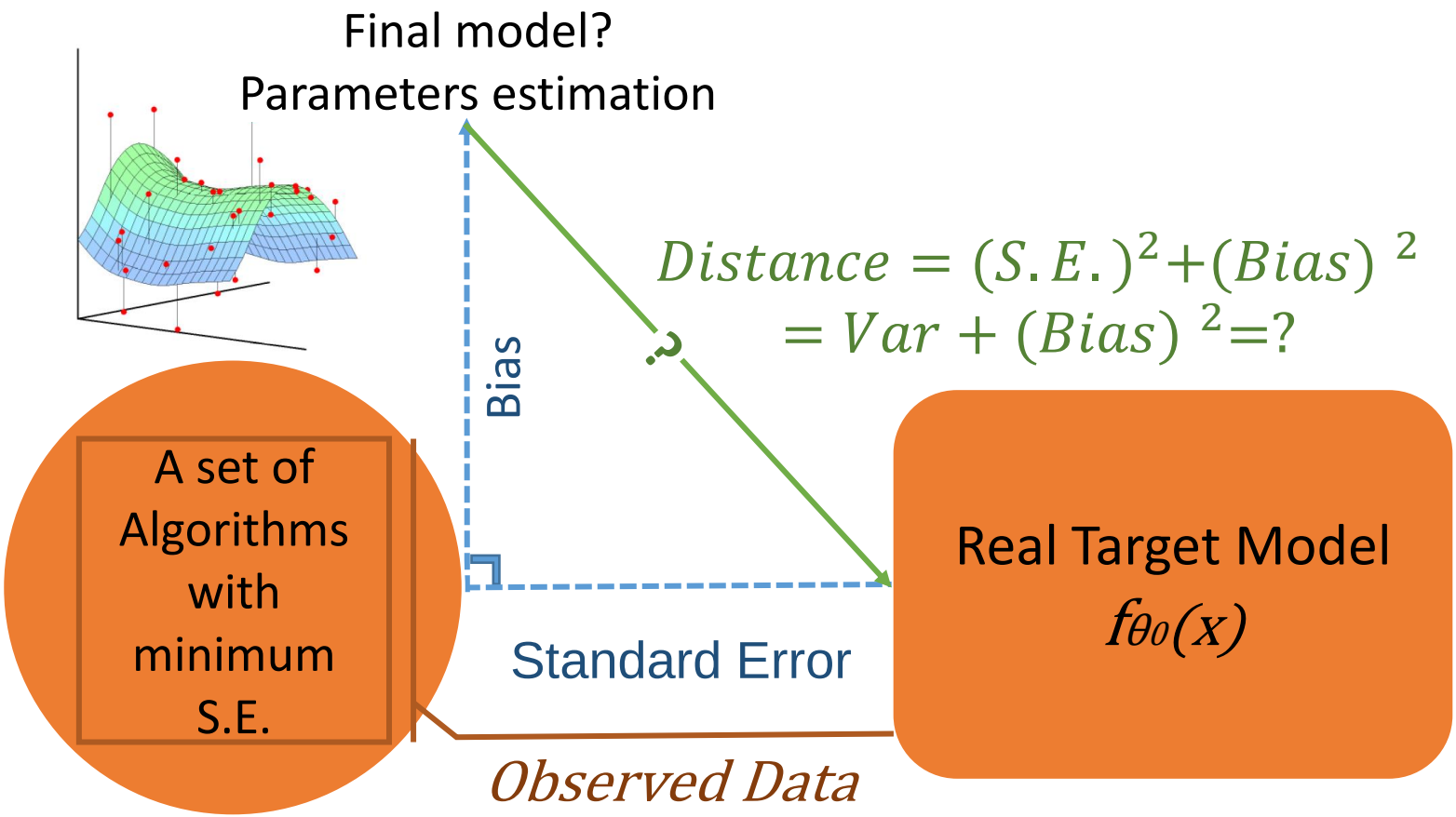
DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



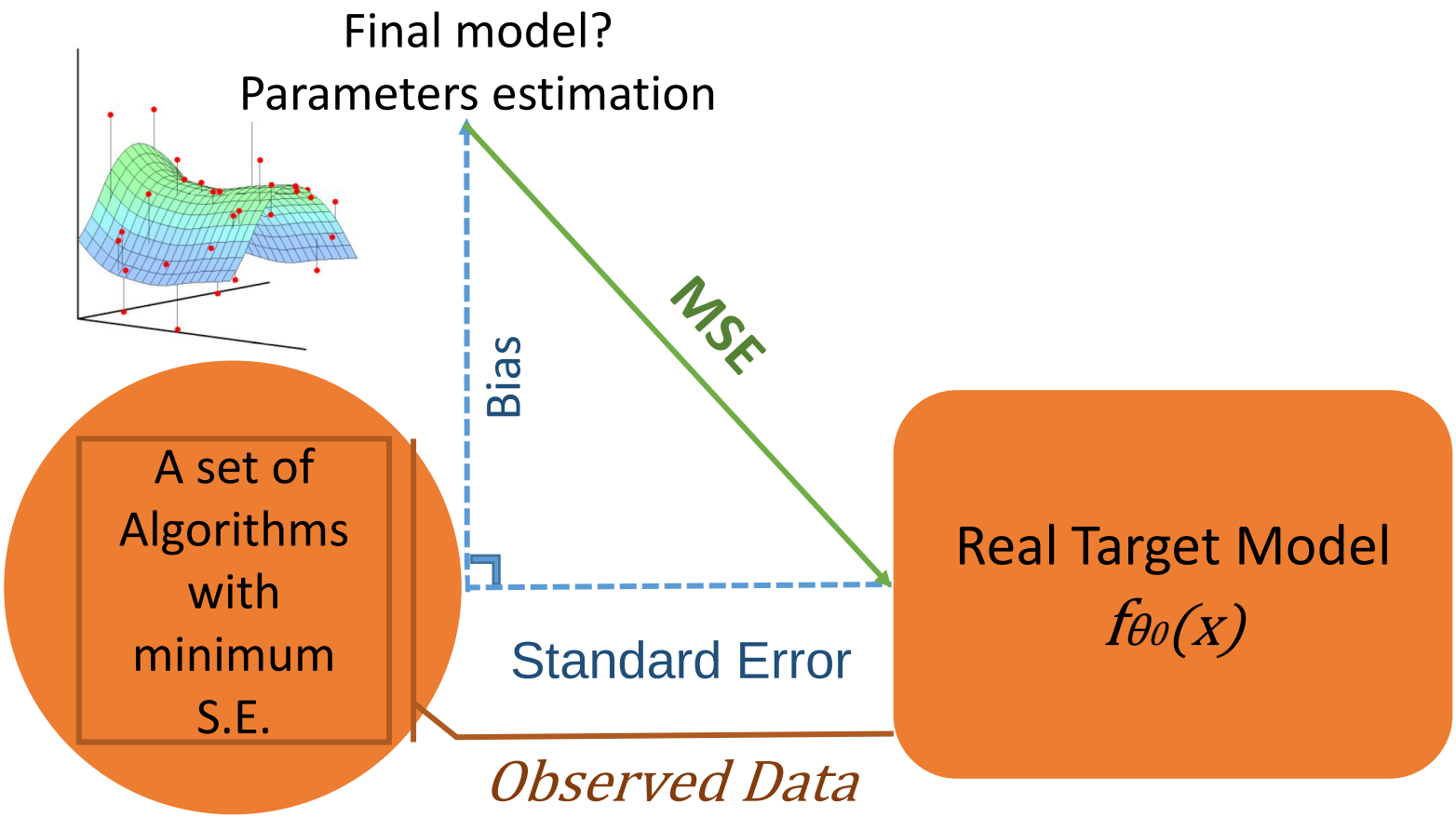
DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



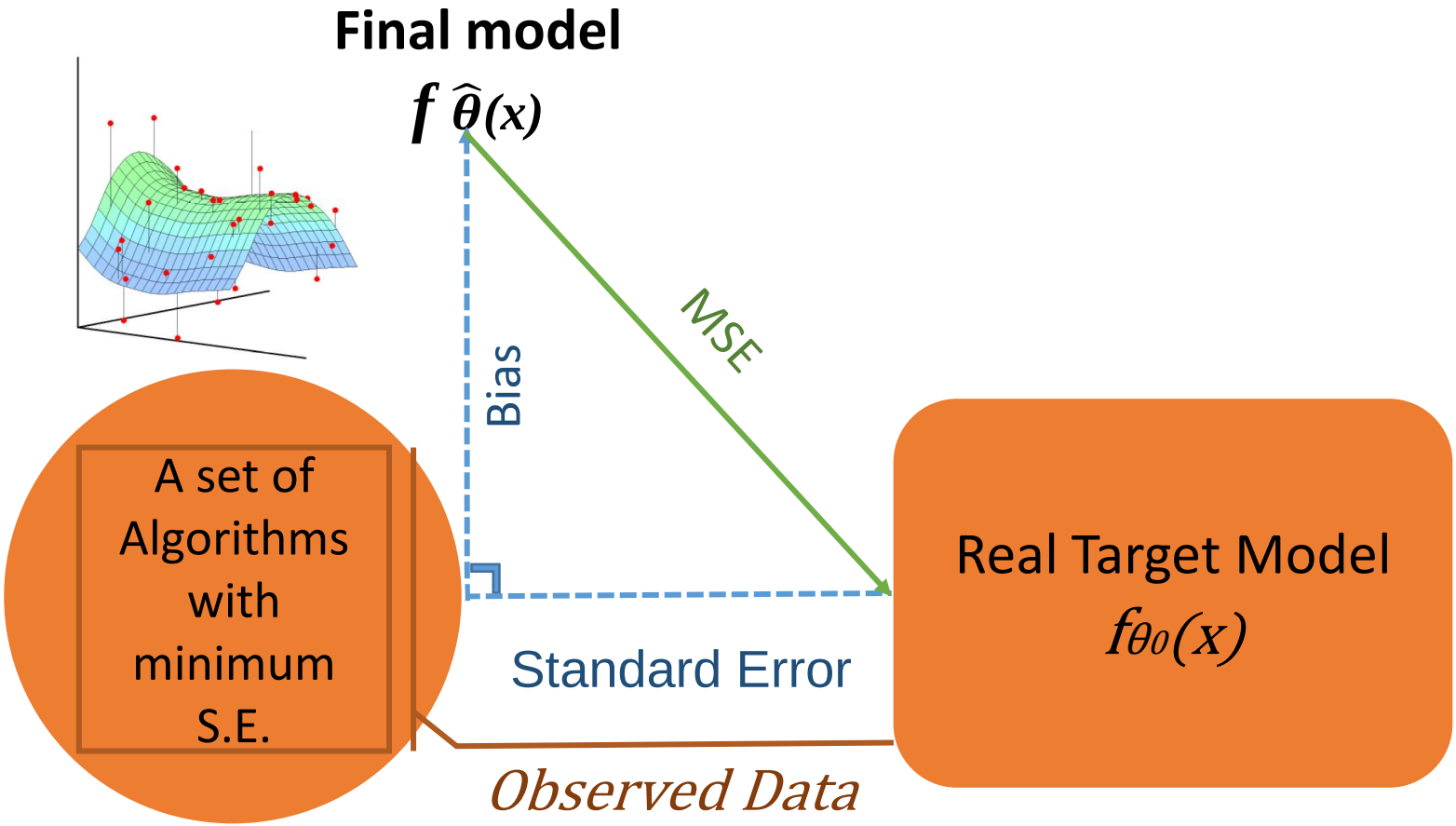
DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



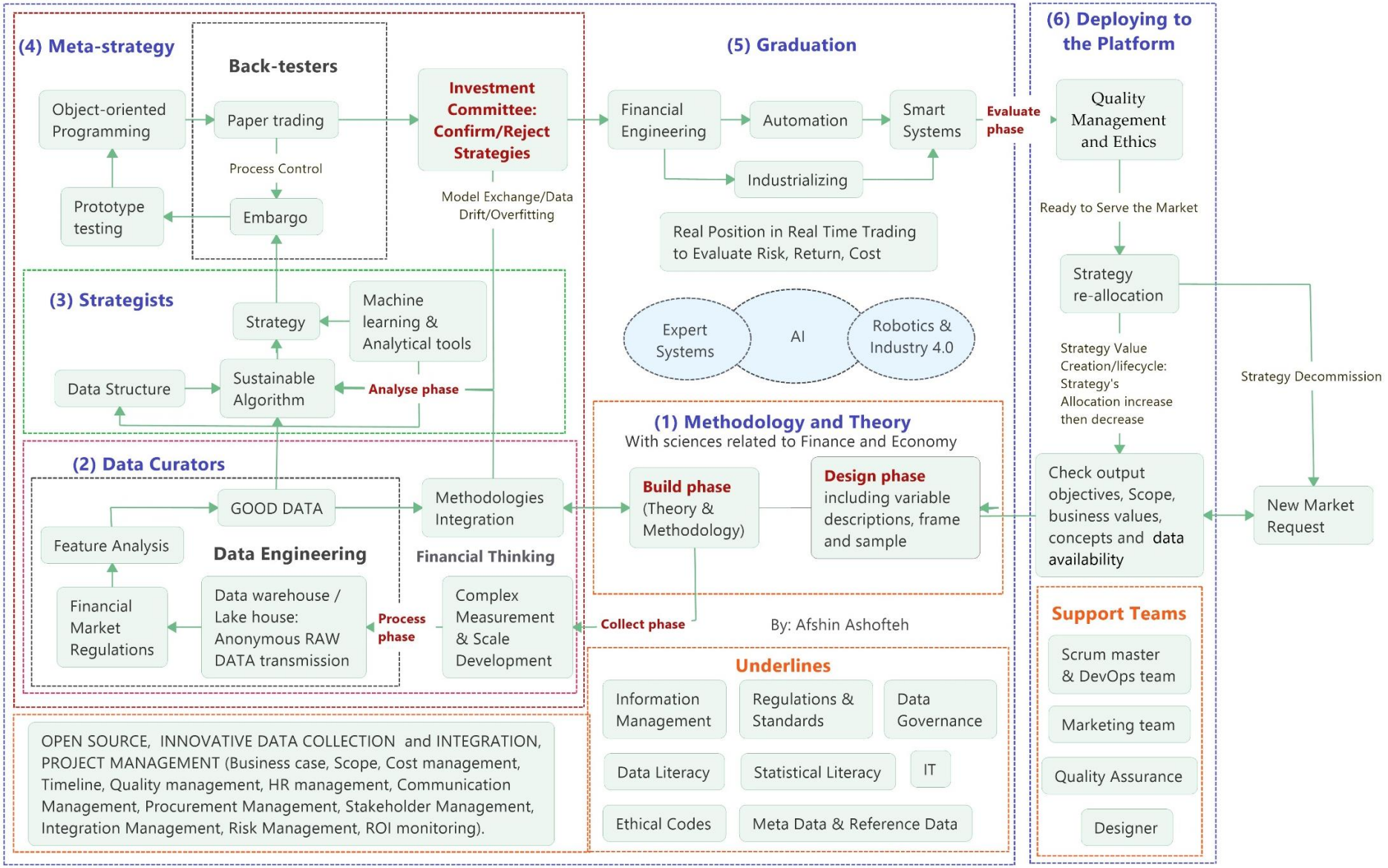
DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE



DATA SCIENCE AND MODELING IN SIMPLE LIFECYCLE

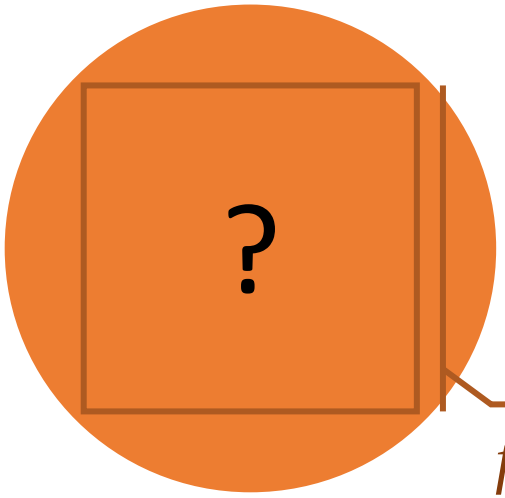


DATA SCIENCE COMPLETE LIFECYCLE



Paper: https://doi.org/10.1007/978-3-031-39864-3_2

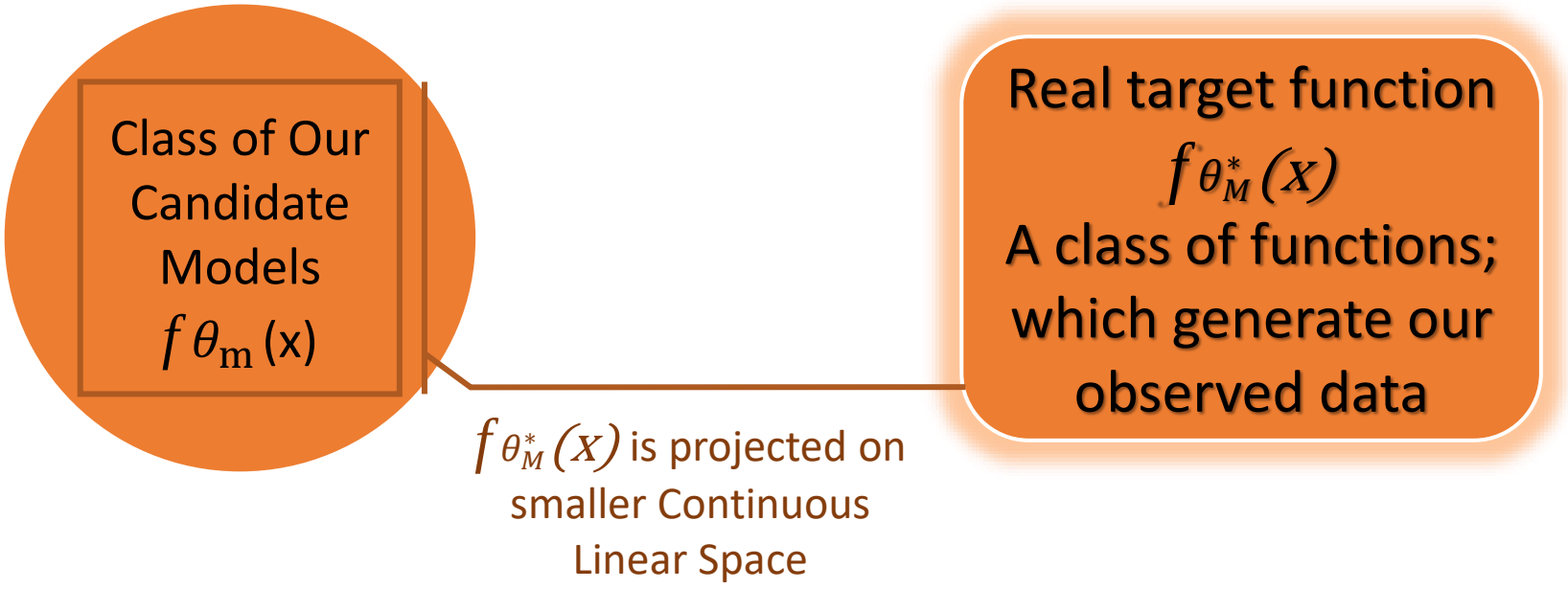
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



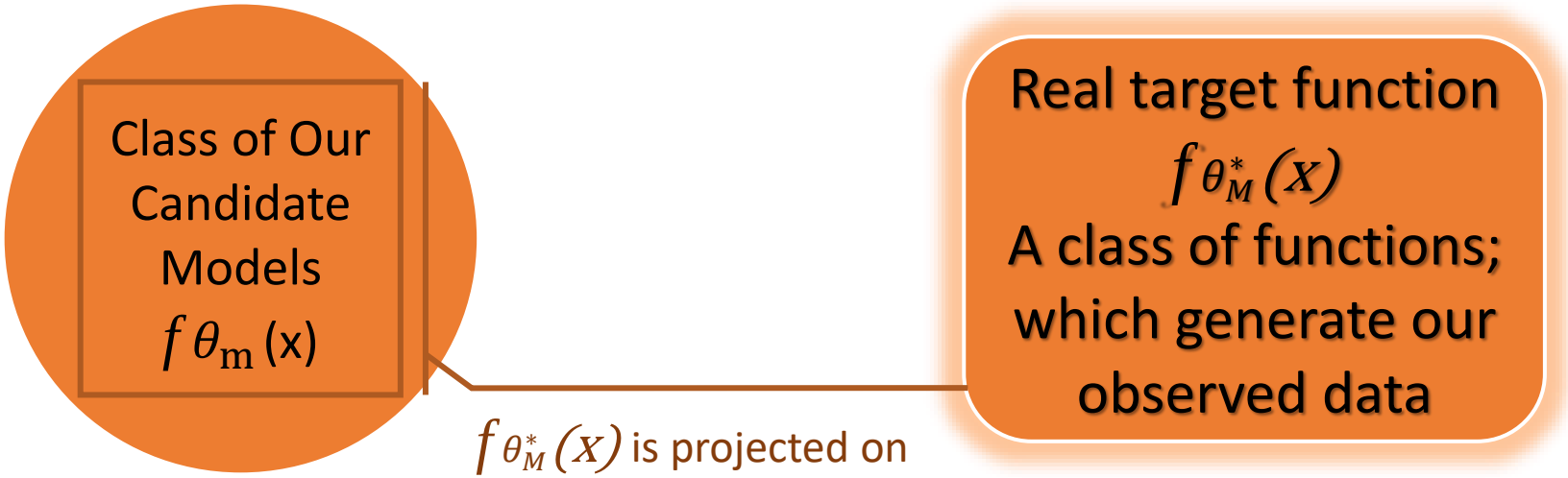
Real target function
$f\,\theta_M^*(X)$
A class of functions; which generate our observed data

$f\,\theta_M^*(X)$ is projected on smaller Continuous Linear Space

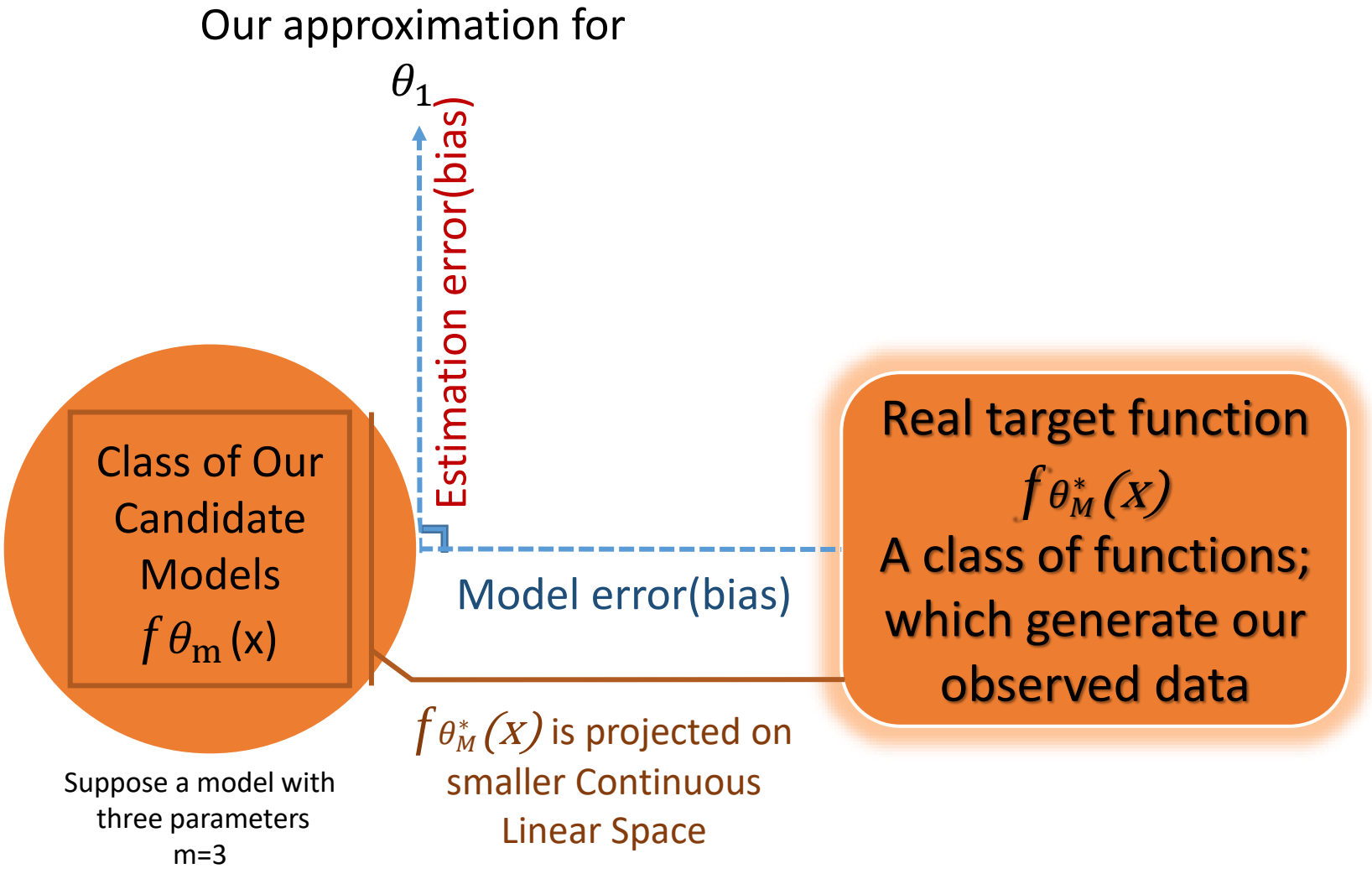
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



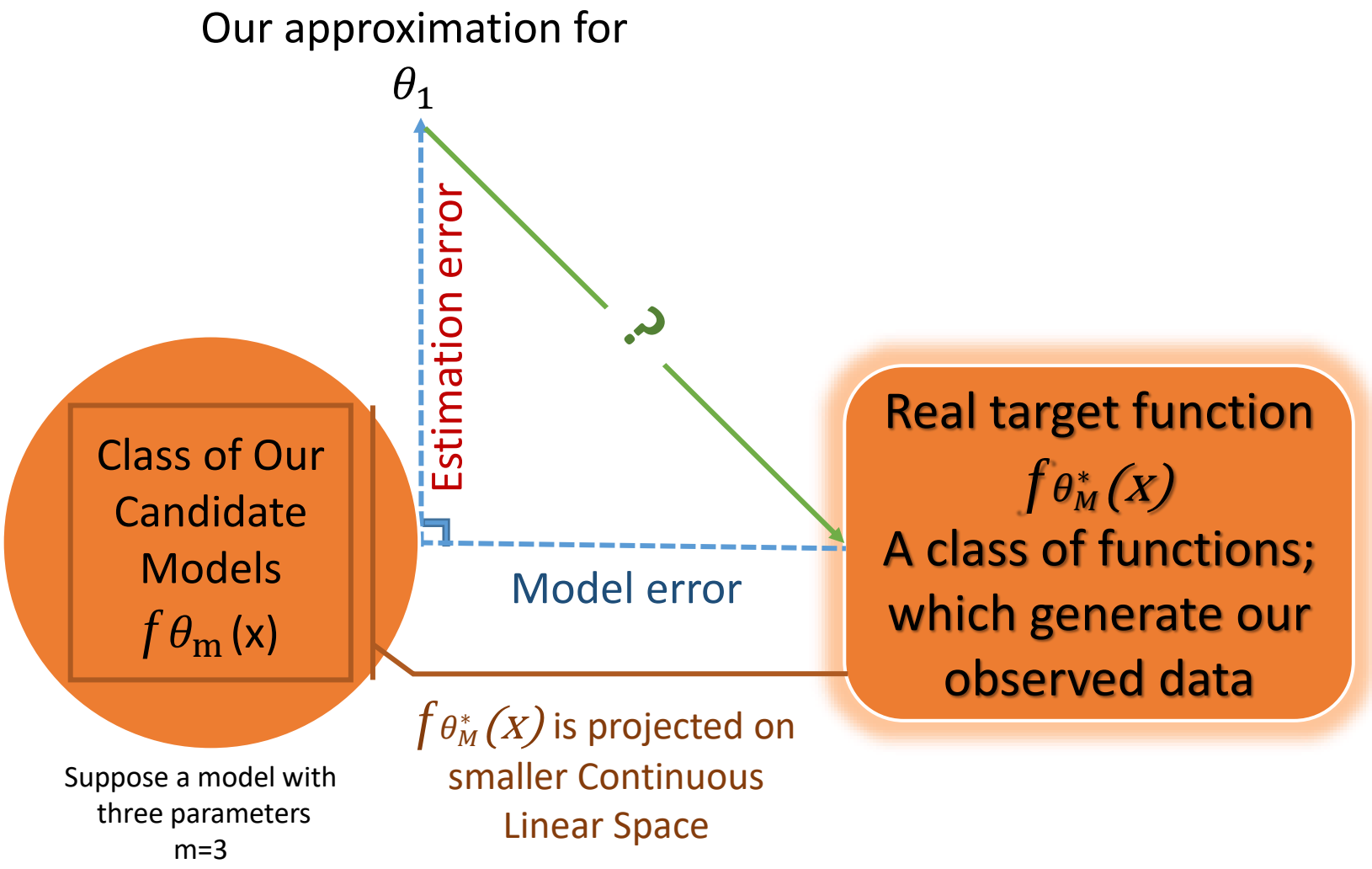
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



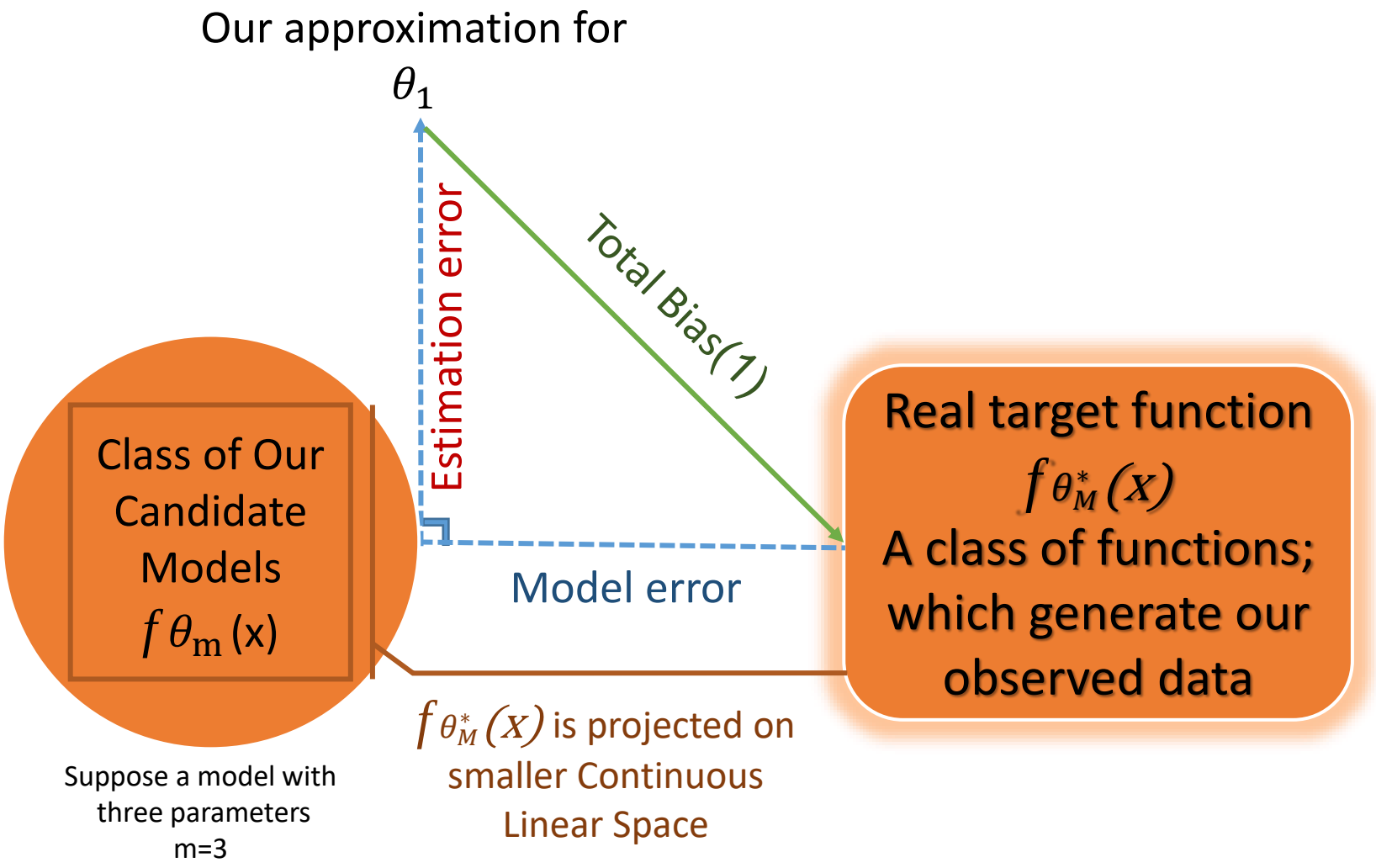
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



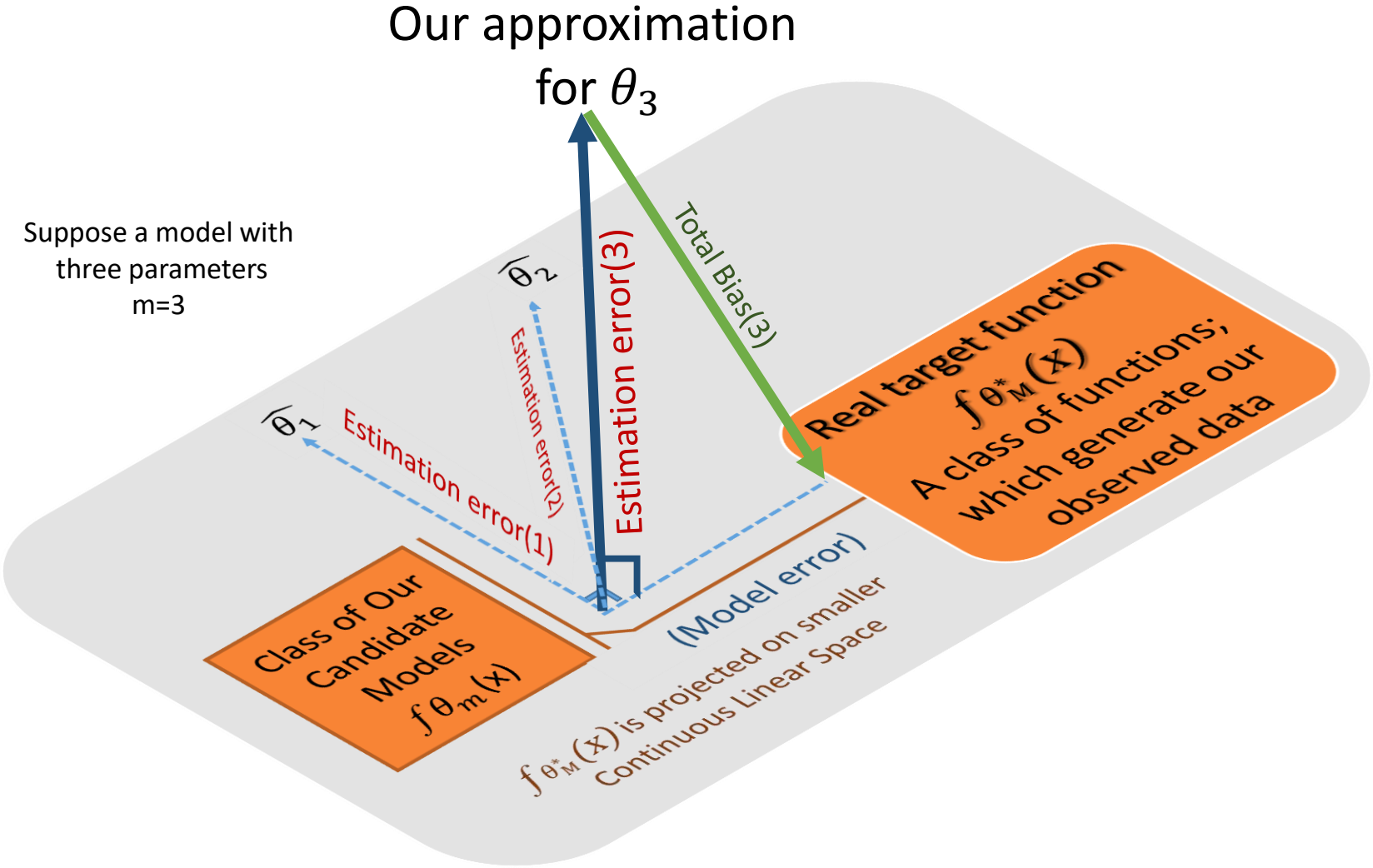
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

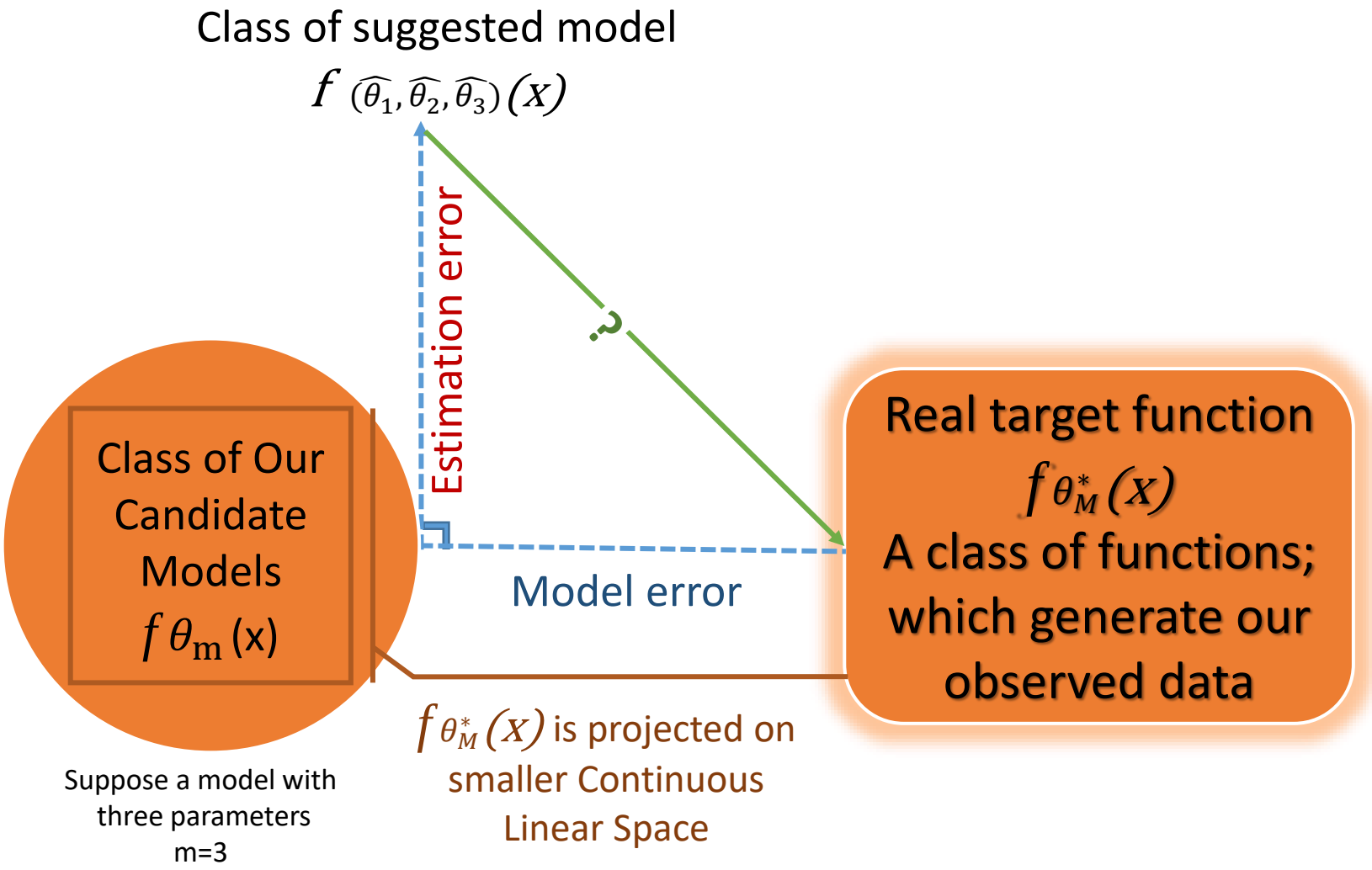


DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

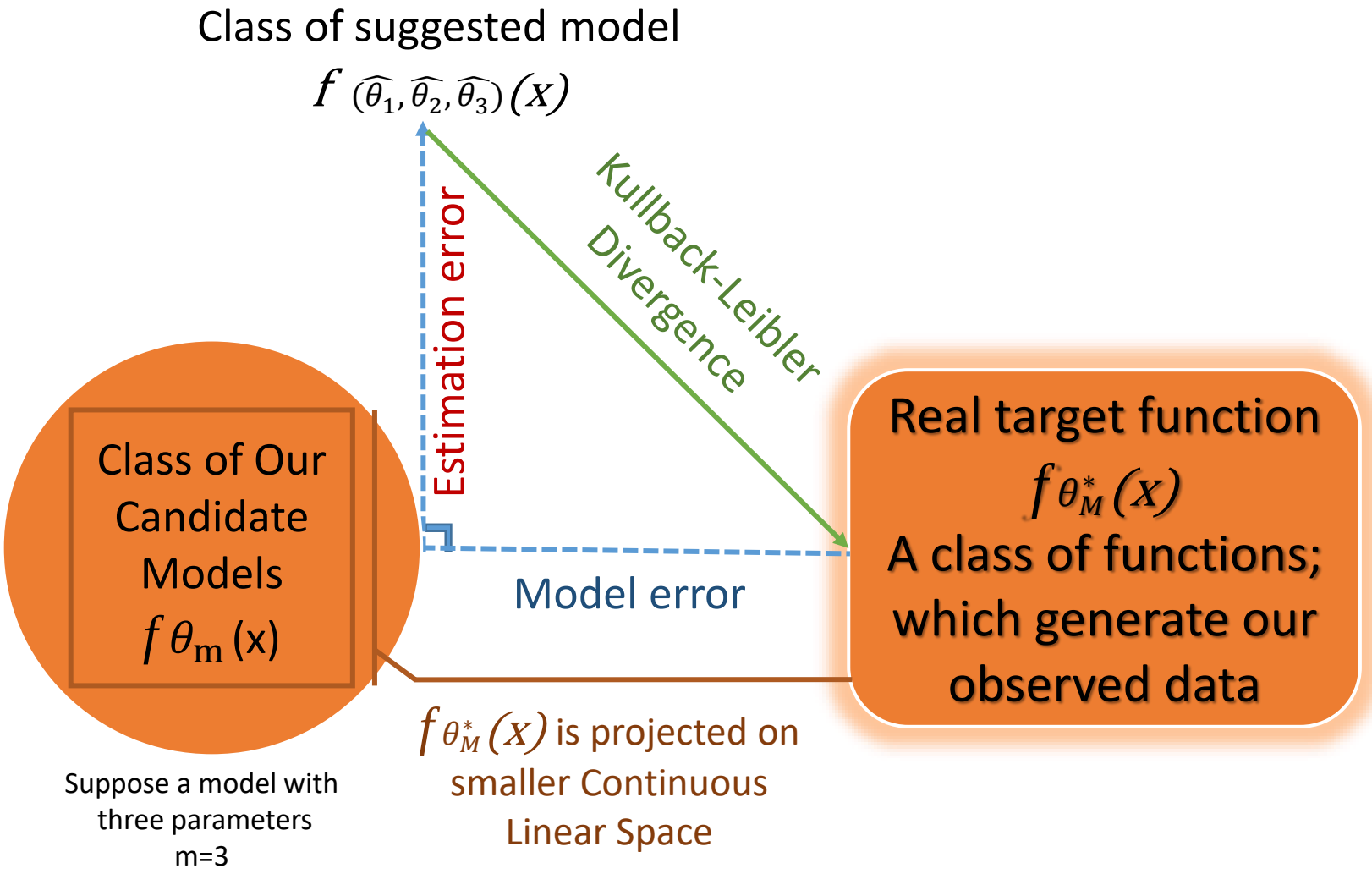


Check this video for more details: https://youtu.be/mVXqQCViEfo

DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

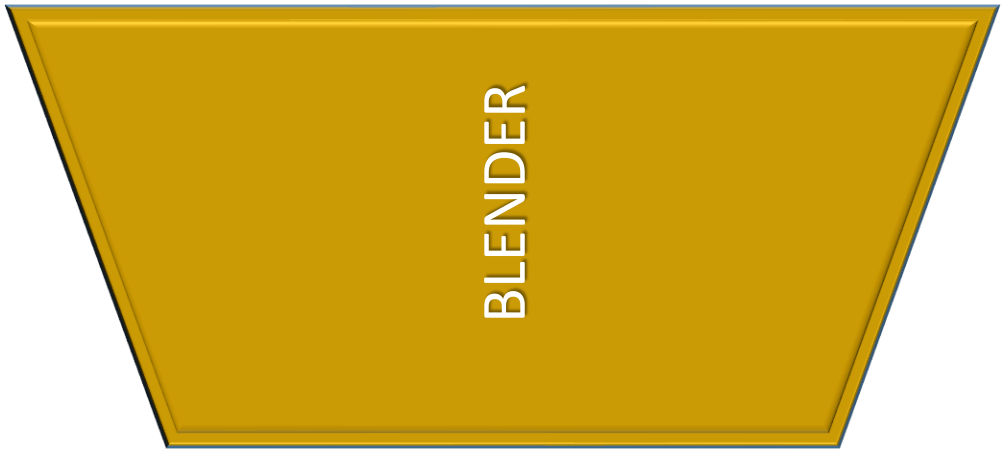


DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

Kullback-Leibler Divergence



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

$$KL(\theta^*, \theta) = \int \log\left\{\frac{f_M^*(x)}{f_m(x)}\right\} f_M^*(x)dx$$

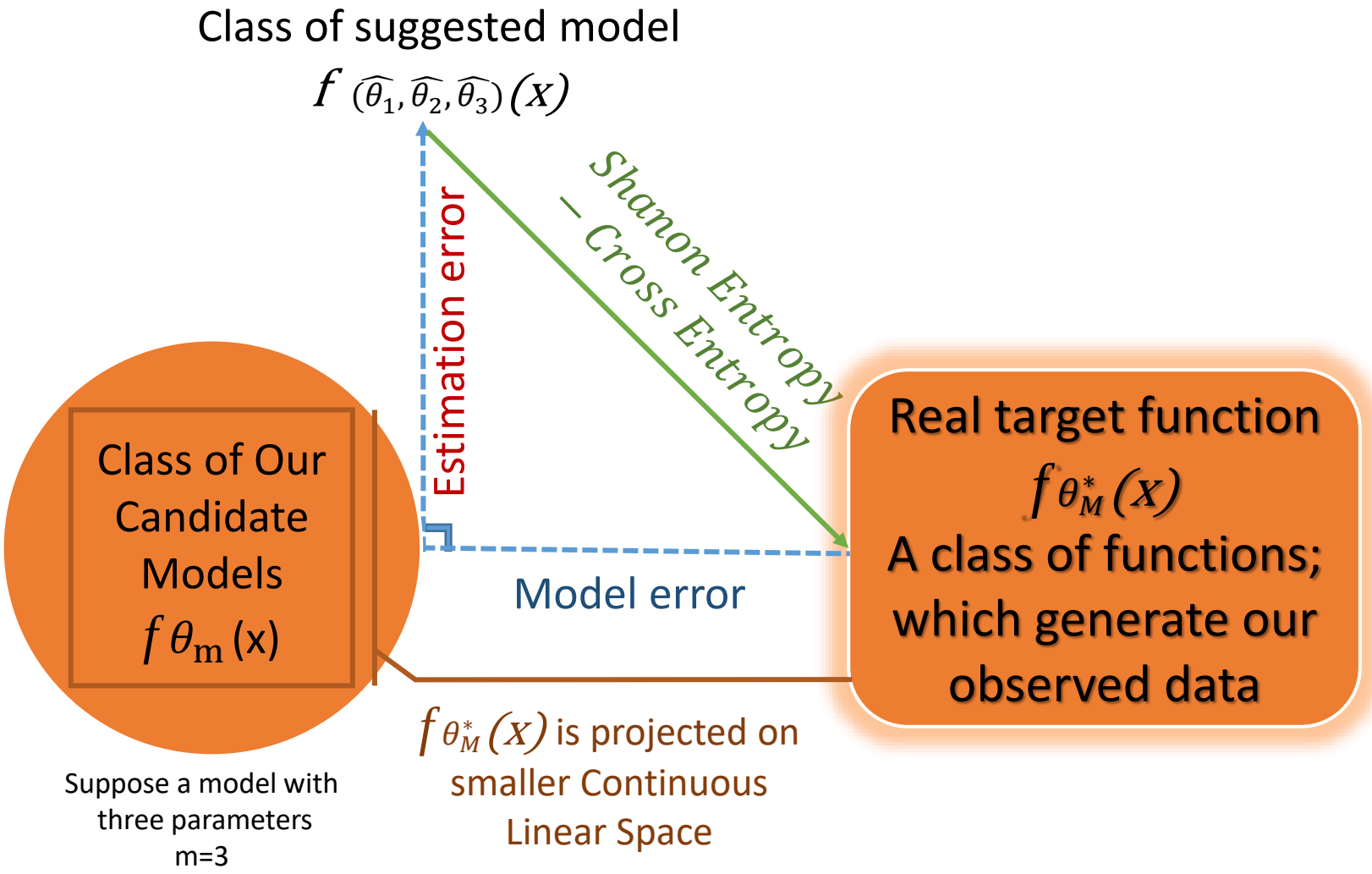
$$= \int \{\log f_M^*(x) - \log f_m(x)\} f_M^*(x)\, dx$$

$$= \int \log f_M^*(x)\, f_M^*(x)dx - \int \log f_m(x)\, f_M^*(x)dx$$

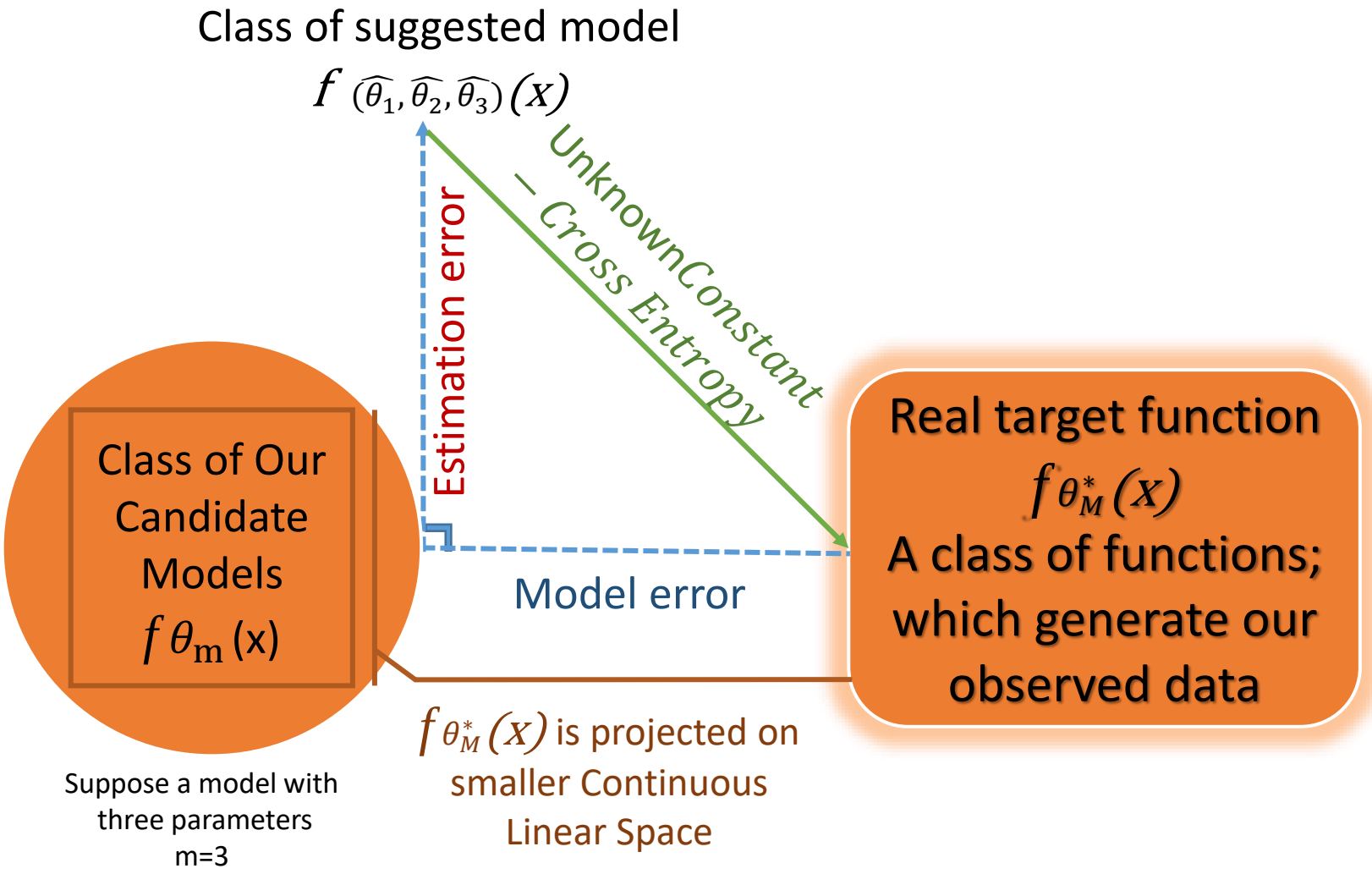
$$= H(\theta^*, \theta^*) - H(\theta, \theta^*)$$

$$KL(\theta^*, \theta) = Shanon\ Entropy - Cross\ Entropy$$

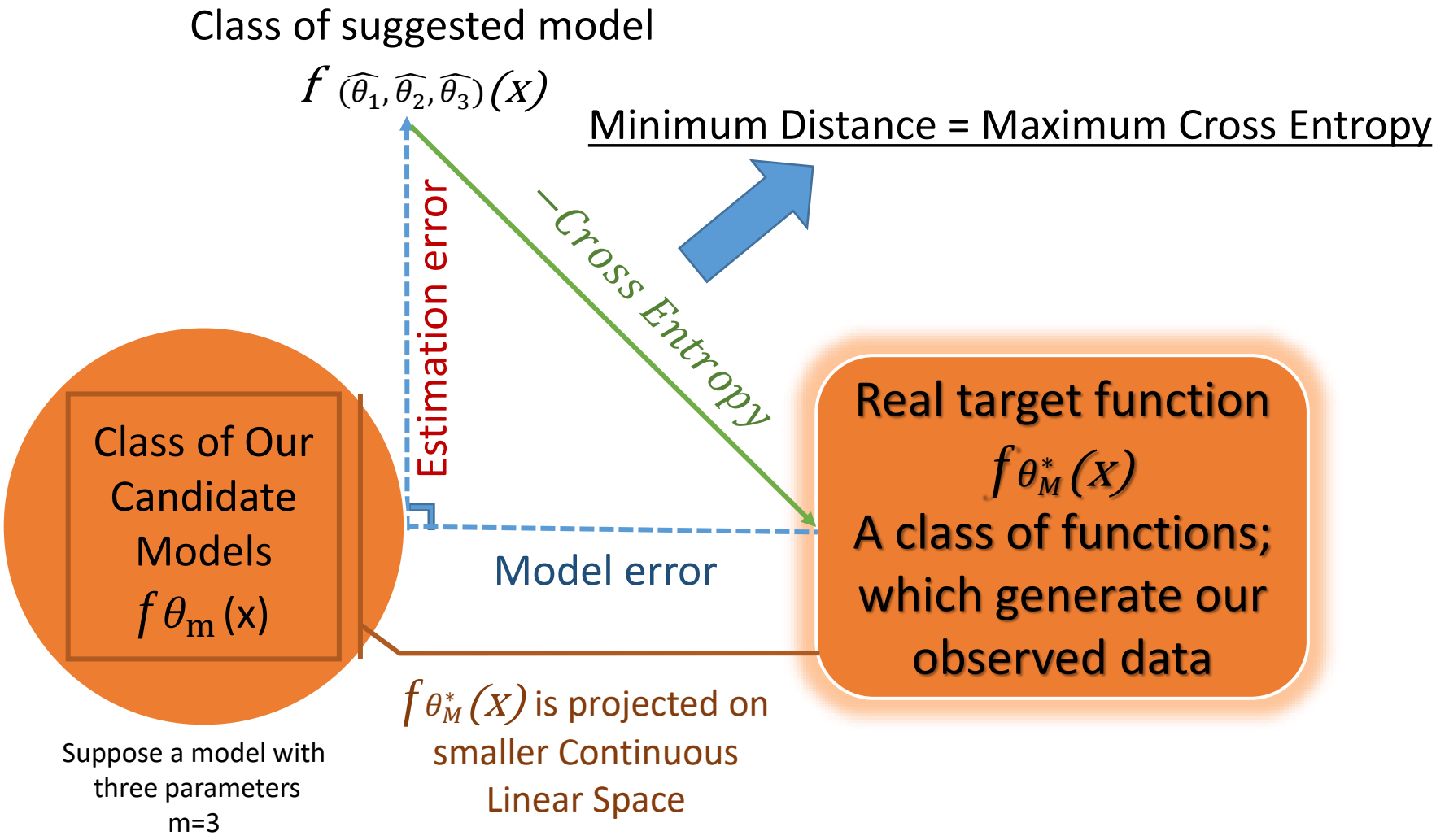
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



MAXIMIZING THE CROSS ENTROPY BY USING OBSERVED DATA

$$Cross\ Entropy = \int \log f_m(x) \,.\, f_M^*(x) dx$$

X is a Random Variable … We cannot use Riemann Integral … We should solve it Statistically … In Statistics we have:

$$\int xf(x)dx = E(X) = \bar{x} = Average = \sum_{i=1}^{n} x_i . p(x_i)$$

and **Strong law of Large numbers**

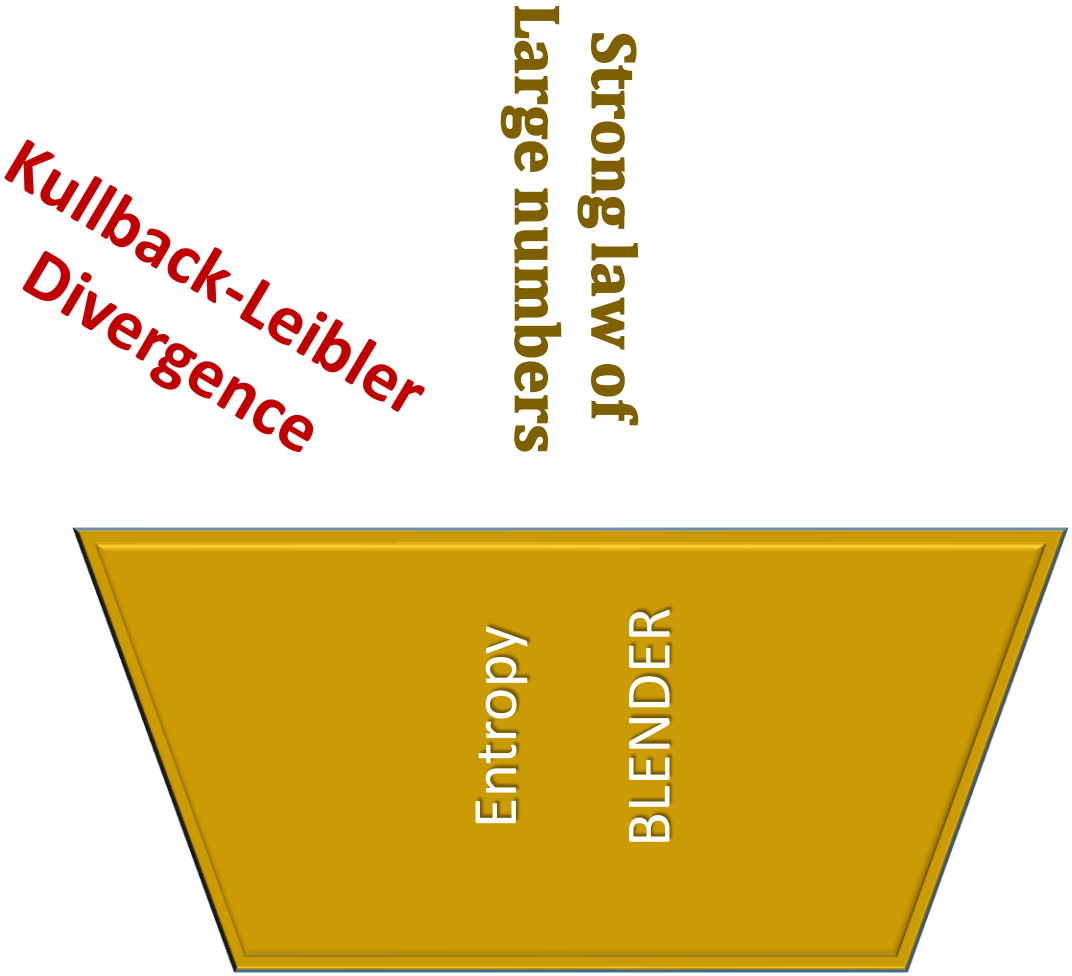
$$x_1, x_2, \ldots, x_n \; iid \quad \bar{x} \to \mu \quad as \;\; n \to \infty$$

DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

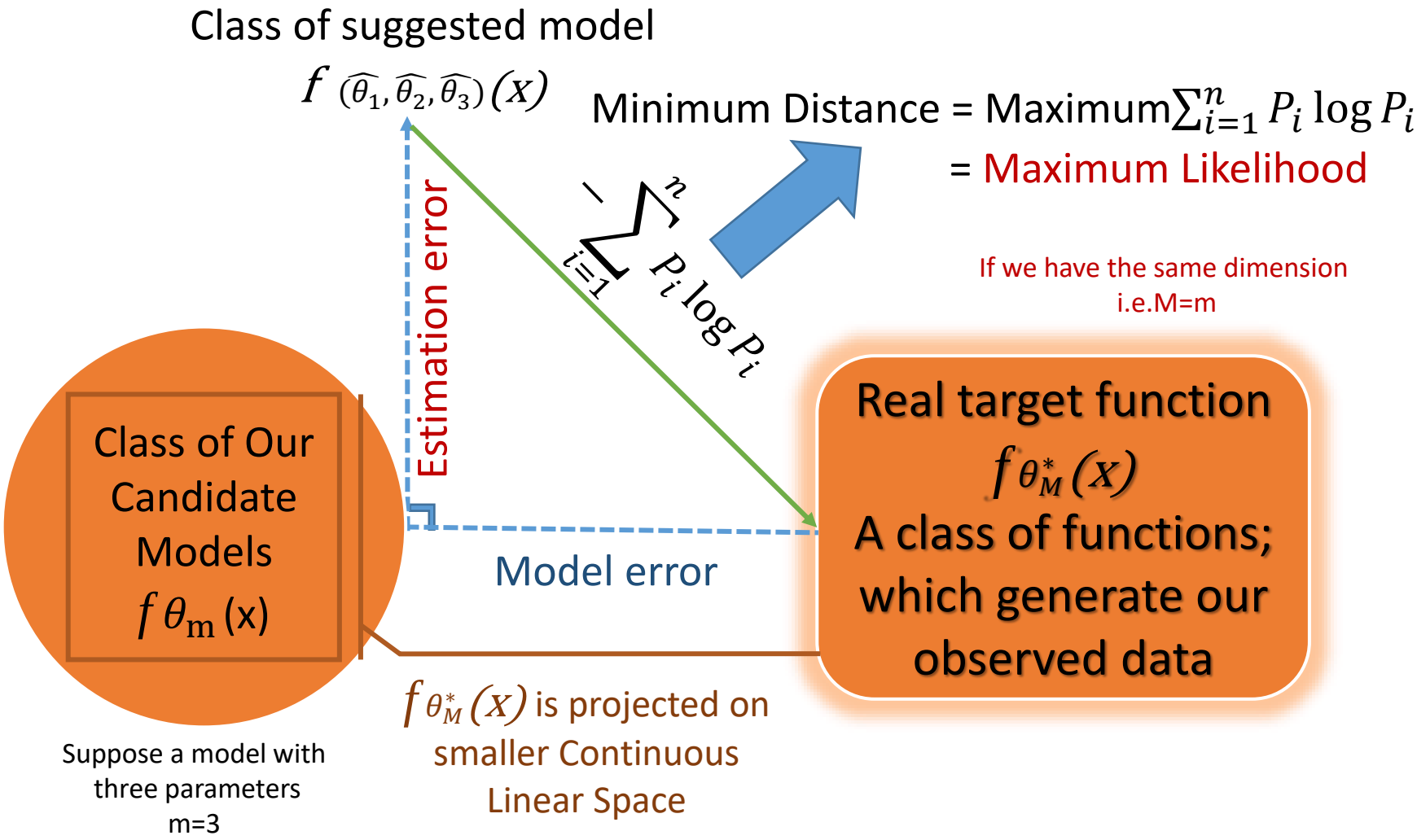
Solution: *calculate Average of observed data*

$$CrossEntropy \approx E_{\theta_0}(\log f_m(x)) = \sum_{i=1}^{n} f_m(x).\log f_m(x) \quad ; \; x_i \sim f_M^*(x)$$

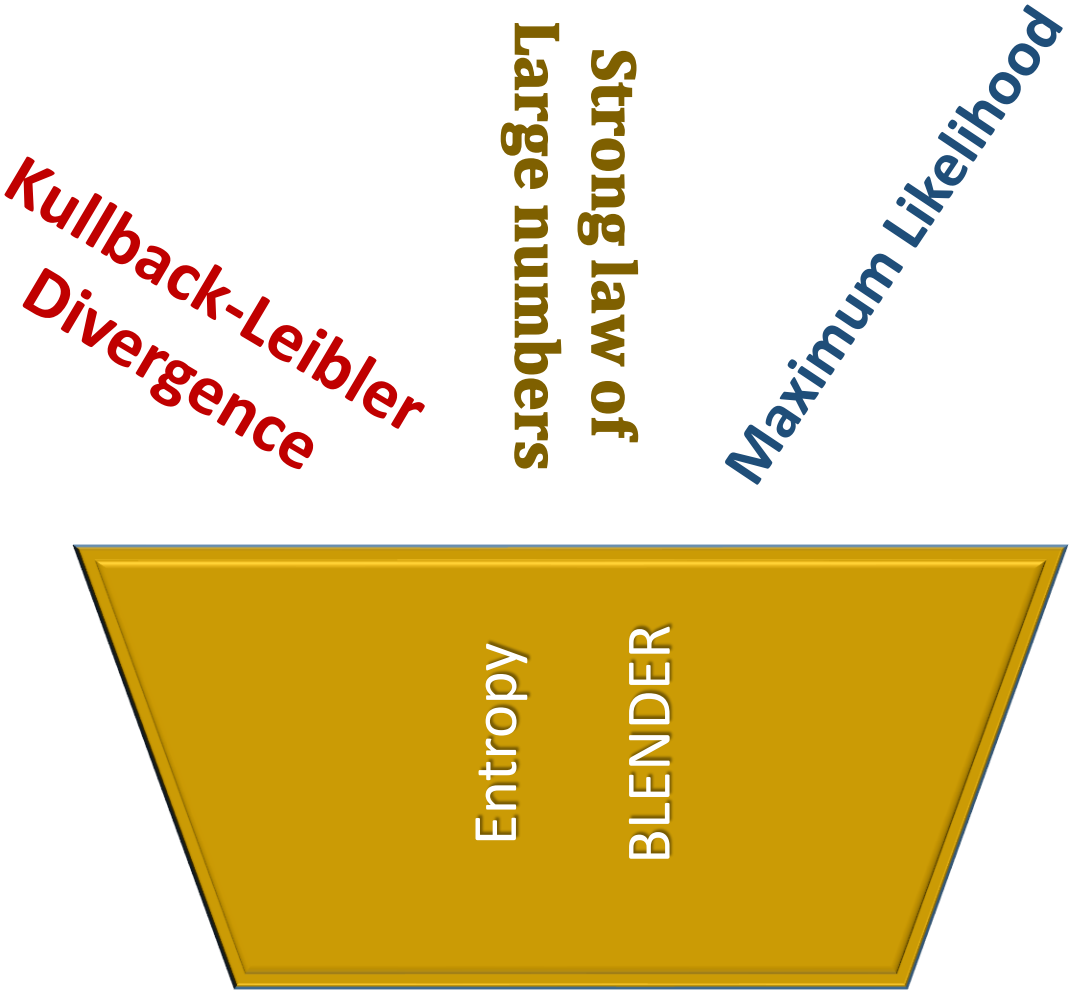
DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



$$Maximum\ Cross\ Entropy = \textbf{Maximum Likelihood}$$

But we don't have $f_M^*(x)$ and it is unknown!

Again Statistics helps:

Central Limit theorem

$$x_1, x_2, \ldots, x_n\ iid\ ; 0 < Var(x) < \infty\ ;\ \ \bar{X} \sim N(\mu_x, \frac{\sigma^2}{n})\ \ as\ \ n \to \infty$$

DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

Solution:

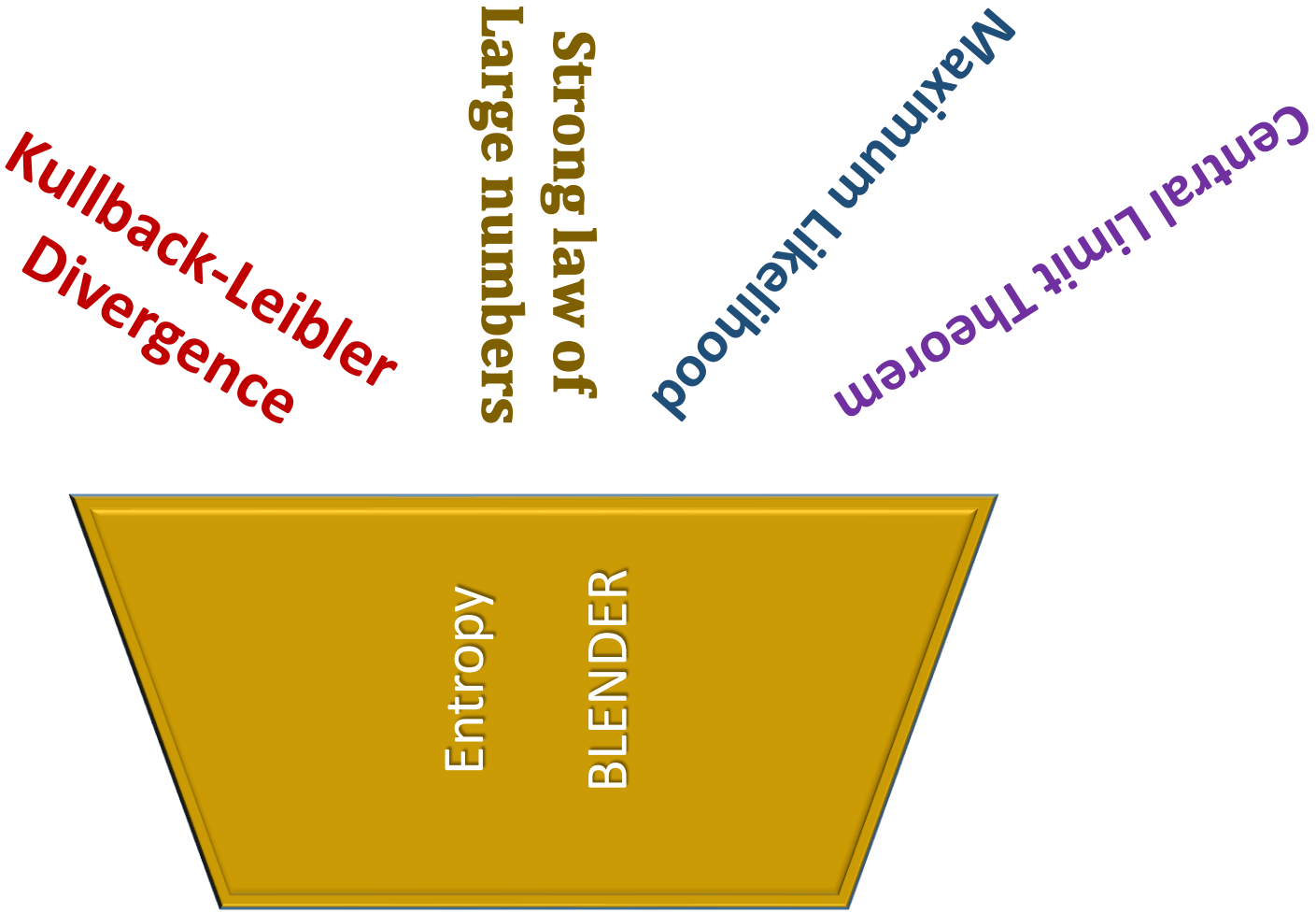
Use $C.L.T.$ and estimate $parameters\ of\ Normal\ Distribution$ from $observed\ data$ to calculate not only **Cross Entropy** by Maximum Likelihood but also **Confidence Interval** for Cross Entropy!

Even we can have a good **Outliers Detection**!

$$CrossEntropy$$

$$\approx E_{\theta_0}(\log f_m(x)) = \sum_{i=1}^{n} f_m(x).\log f_m(x) \qquad ;\ x_i \sim N(\hat{\mu}_x, \frac{\hat{\sigma}^2}{n})$$

DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE



DATA SCIENCE AND MODELING IN COMPLETE LIFECYCLE

We used these theorems:

- Kullback-Leibler Divergence
- Min K.L.D. = Max Cross Entropy = Max Likelihood
- Strong Law of Large Numbers
- Central Limit Theorem

K.L.D. = Kullback-Leibler Divergence

We need these assumptions:

- Class of candidate models should be Absolutely Continuous and Smooth to have second order derivative in all points
- For using average, we should remove outliers.
- To know about dimensions (m=M), we try to guess m by looking at Histogram.

TO SUM UP

It is a very brief review of a personal concept, which shows how I usually attack the curve fitting or modeling problems. First, we must know linear algebra. Why?

Because computer engines work only by **COMPUTATIONAL LINEAR ALGEBRA**. It is essential to make a function linear and put machines to work. Software such as R and Python are using linear algebra packages such as LAPACK and BLAS. They load these linear algebra packages for their computation. Even SAS does so. This is why we prefer **LINEAR SPACE** in comparison with the **NORM SPACE**. Additionally, in linear space, we prefer **ADDITIVE MODELS**:

$$y \approx f(x_1) + f(x_2) + \cdots + f(x_n),$$

why? Because simply they are easy to compute!

We can make multivariate analyses, Clustering, and Classification by using these additive univariate models without the need to reinvent the wheel.

Then, in linear space, we can use the tools of vector space, and we have angle θ as follows:

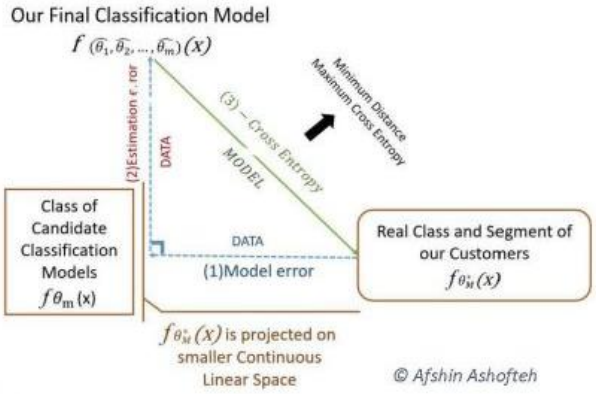
$$\hat{\theta} = argmax \log L(\theta) = argmax \log \prod_{i=1}^{n} f(x_i|\theta).$$

Why these tools and angles are important?

Because it gives us the ability to define the norm, vector, and direction.

For instance, we can define an inner product. ||x||*||Y||*cos(θ) and <X.Y>=$\sum x_i y_i$,which is important to measure and minimize the errors and defining the **DISTANCE MEASURES**. In this vector space, we are able to fit an approximate line, surface, or high dimension solution to our data! Isn't that amazing? Most of the time we need our outcome function to be smooth because simply we have many tools in mathematics for optimizing **SMOOTH FUNCTIONS**! We will be able to take derivatives for optimization and use averages in our theories. Therefore, smooth functions have good behavior in converging to an appropriate outcome in the modeling stage. If we want to have a smooth function, then it is necessary to solve the problem of **MISSING VALUES**, which make our function discrete in the middle of the distribution, and **OUTLIERS**, which make our function discrete in extreme values. With a class of smooth functions and a huge amount of good data, we are able to find the best model and the best estimation of parameters, only by using the data.

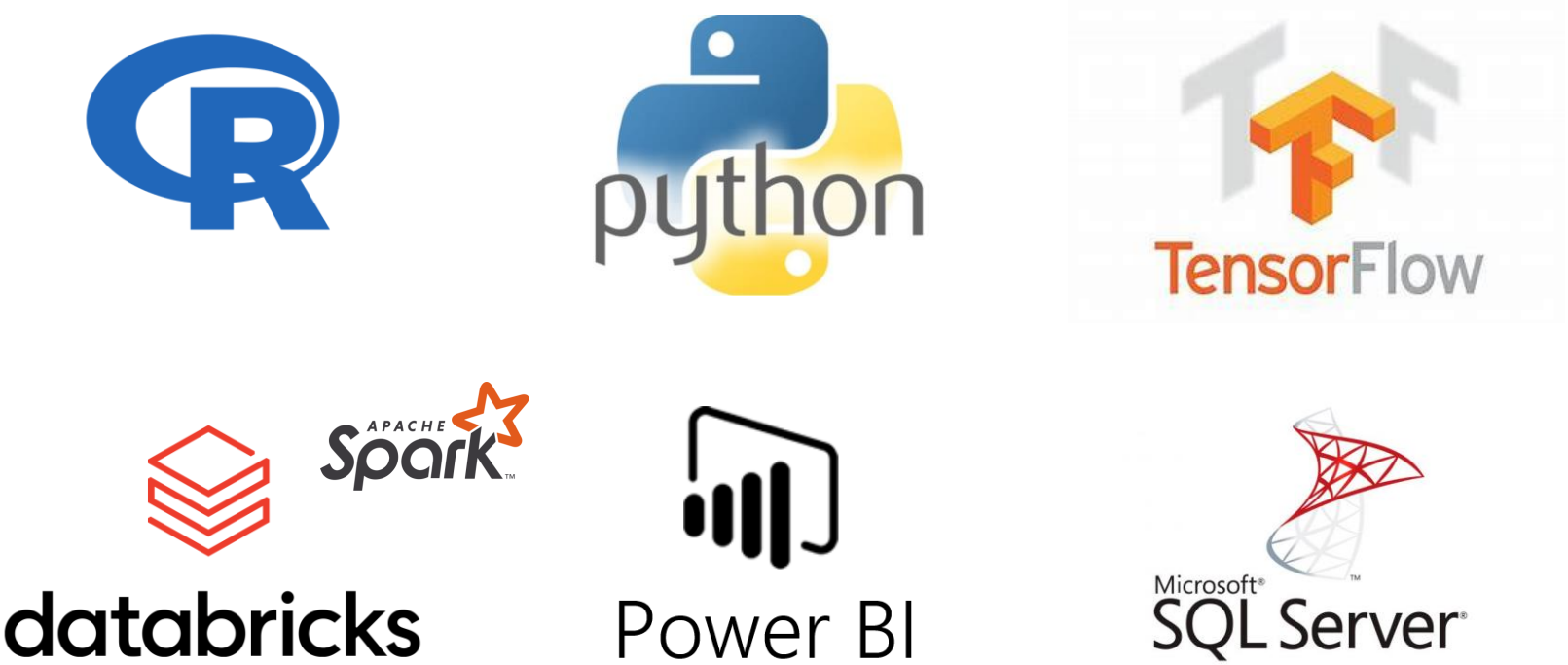
This is the methodology to solve the modeling problems! First, we will define the problem, clean the data set, use the data to make a suggested class of smooth functions and finally use the computational power to choose the best model with the best estimations of parameters. ■

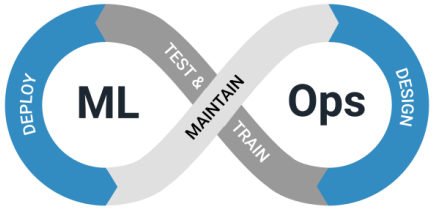


Link!

THE ERA OF TECHNOLOGY

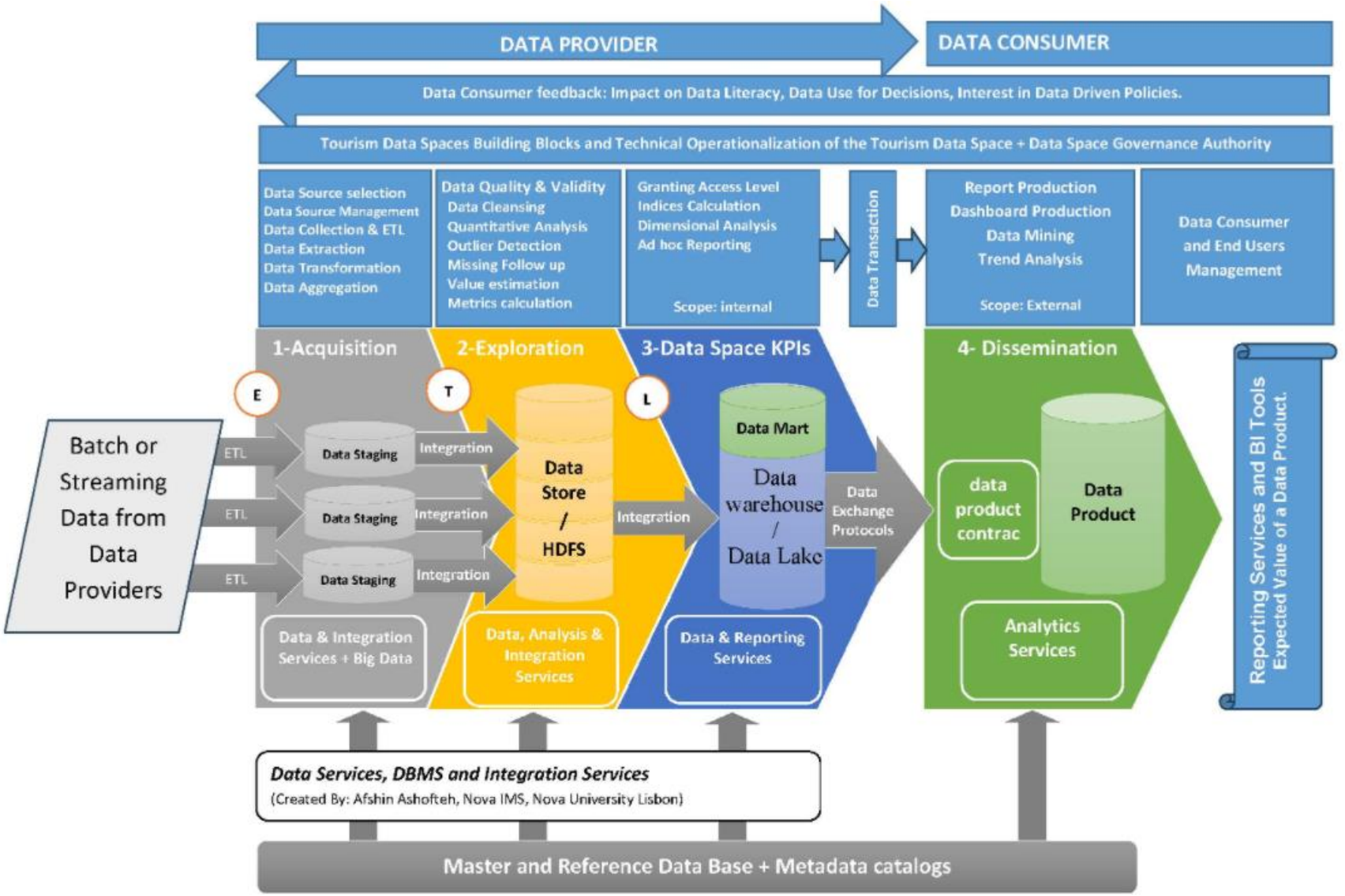
DATA SCIENCE TECHNOLOGIES





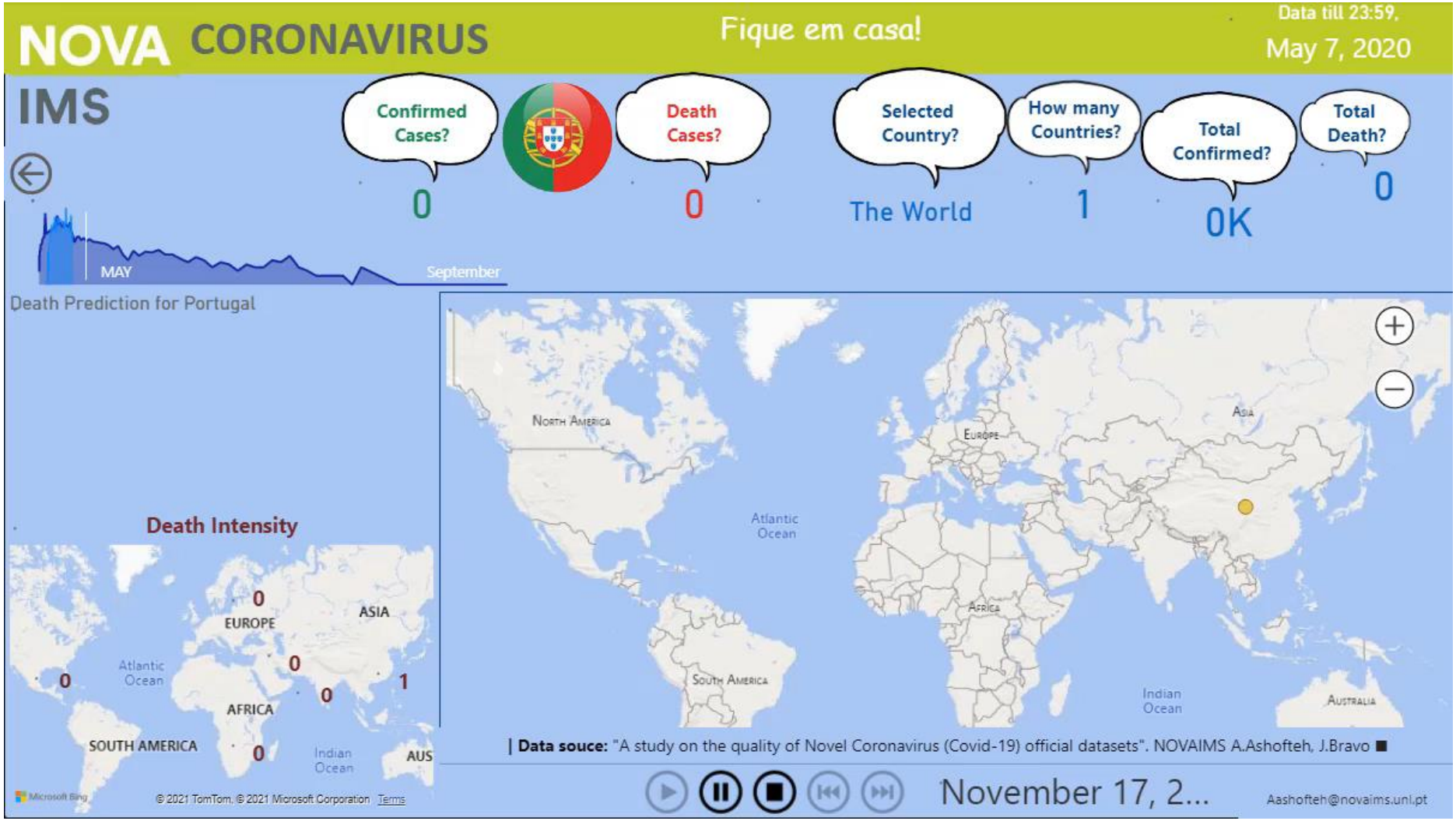
Paper1: https://doi.org/10.1109/COMPSAC61105.2024.00101
Paper2: https://doi.org/10.1109/MITP.2025.3532129

DATA SCIENCE TECHNOLOGIES



SAMPLE: POWER BI DASHBOARD

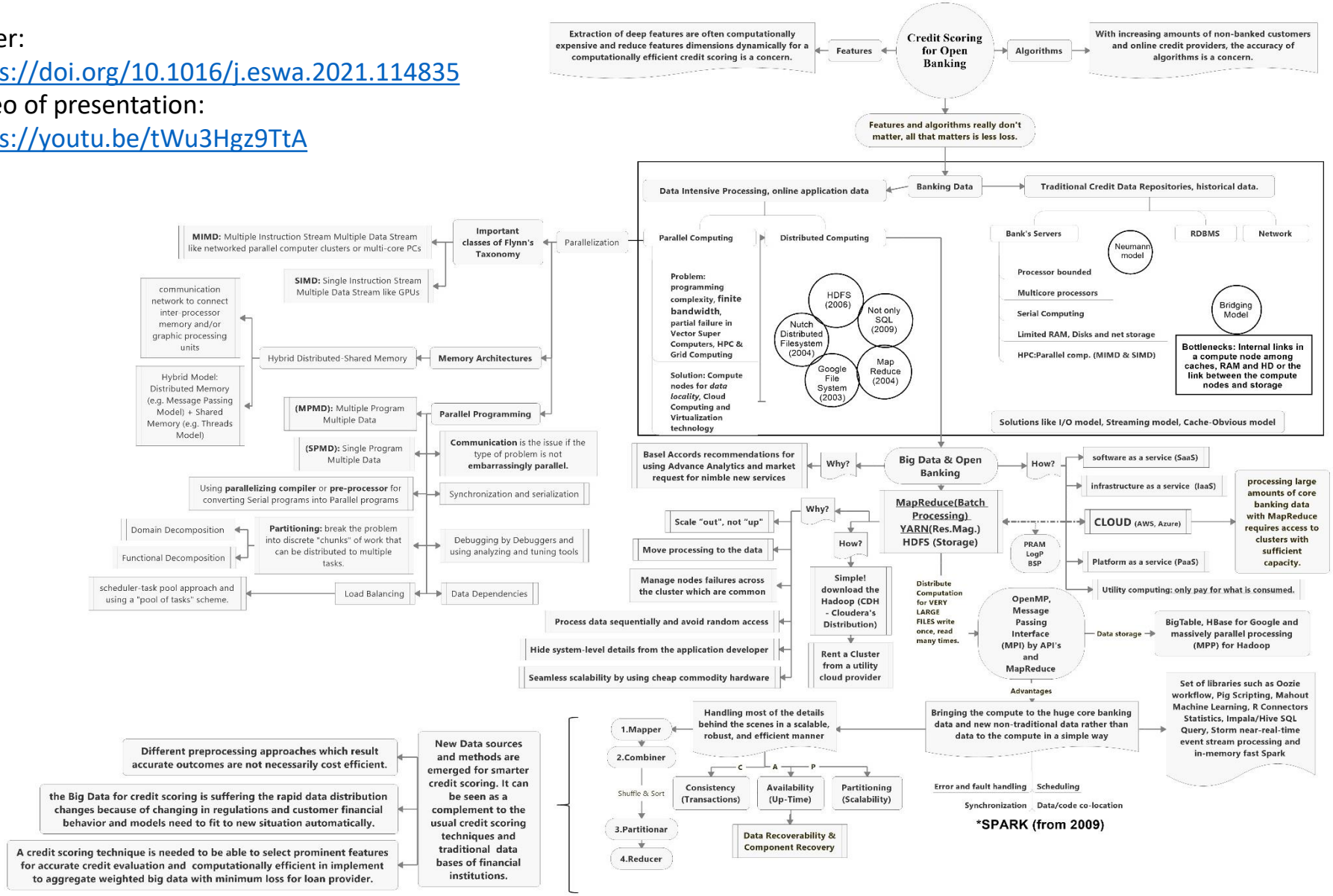
https://doi.org/10.3233/SJI-200674



THE ERA OF BIG DATA

BIG DATA STORY

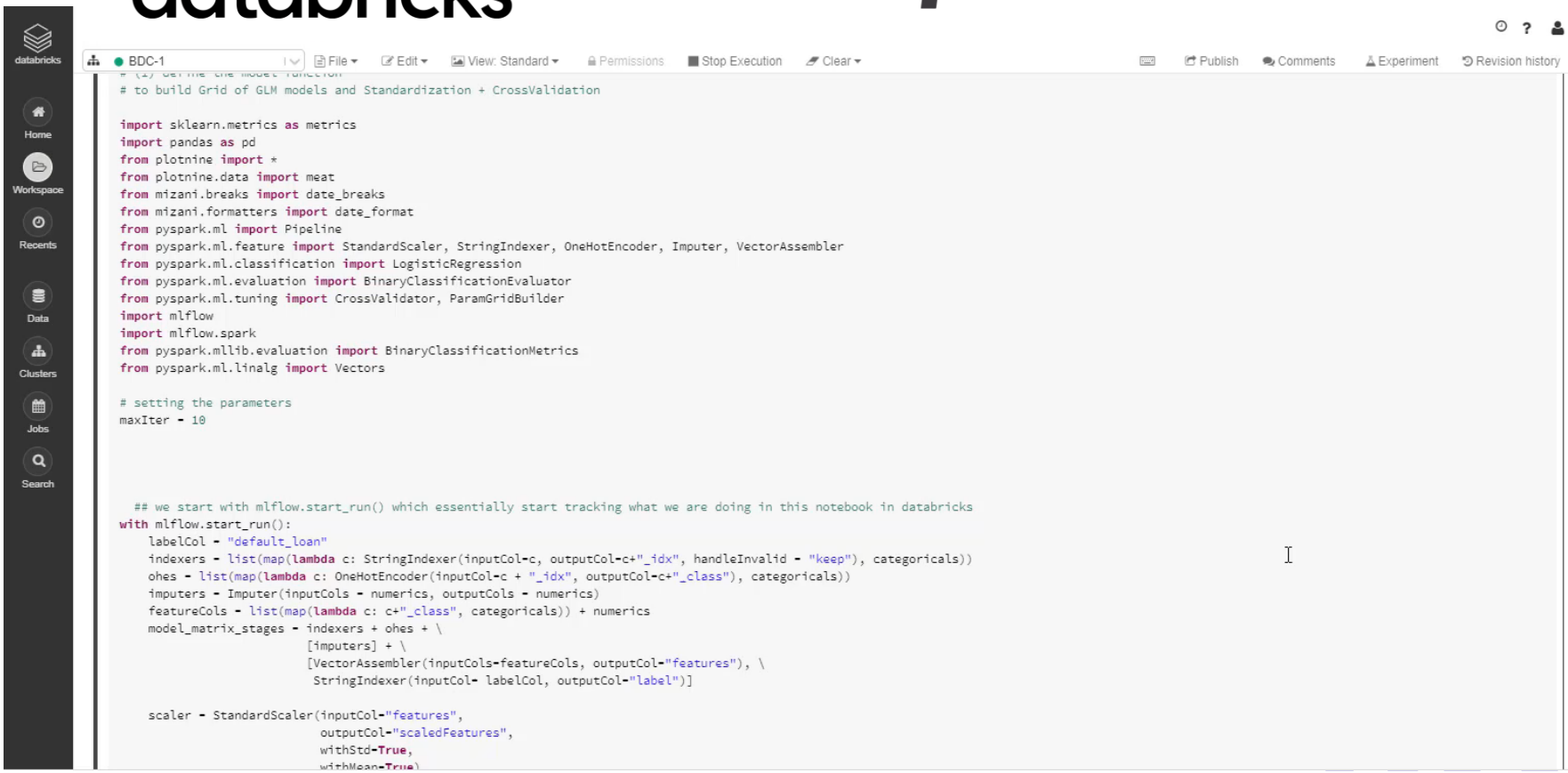
Paper:
https://doi.org/10.1016/j.eswa.2021.114835
Video of presentation:
https://youtu.be/tWu3Hgz9TtA



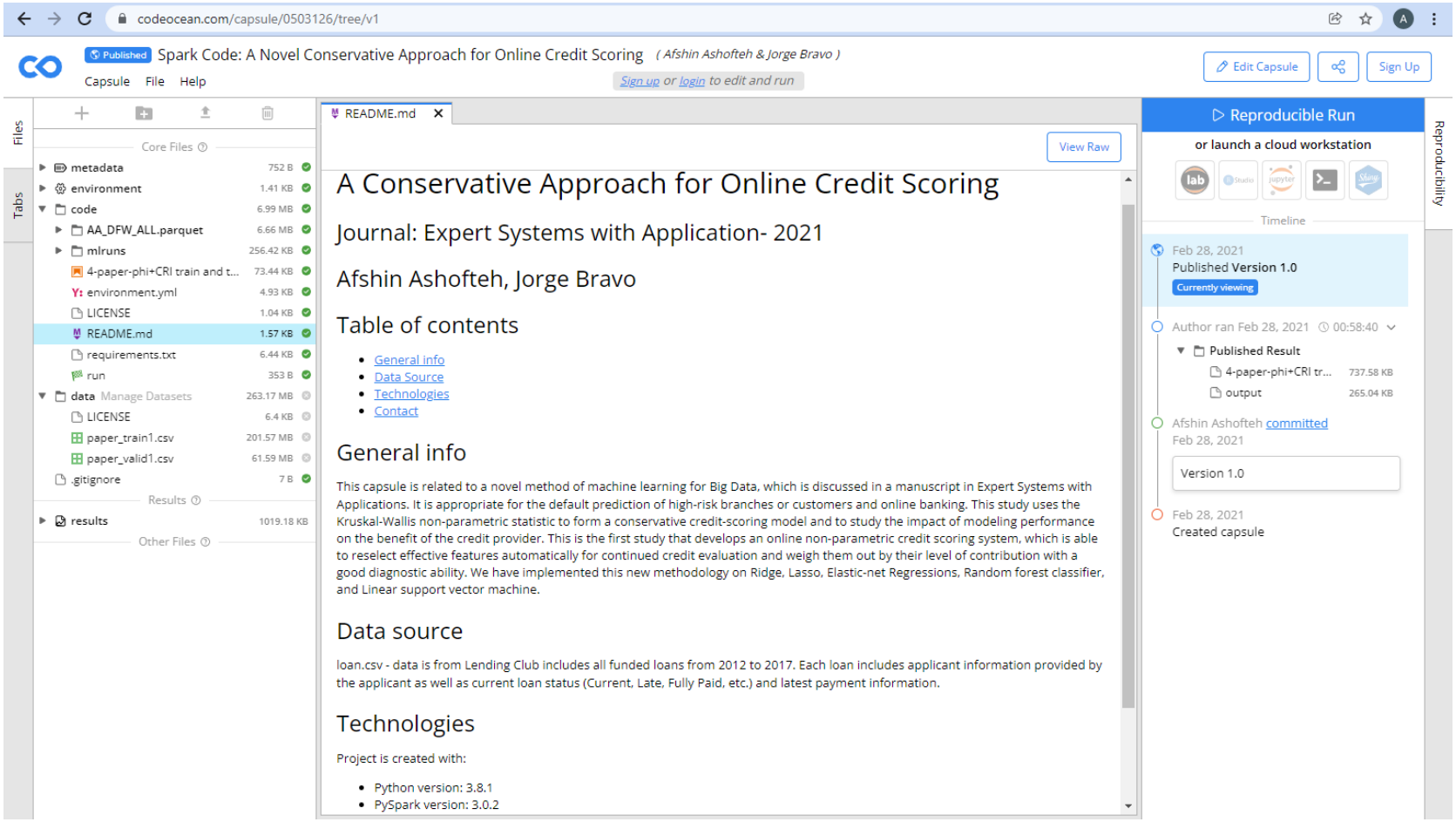
Chapter in Statistical Modeling and Simulation for Experimental Design and Machine Learning Applications (SimStat 2019): https://doi.org/10.1007/978-3-031-40055-1_14

databricks

APACHE Spark™



THE DATASET AND CODE IN PYSPARK ARE PUBLISHED BY CODEOCEAN



Afshin Ashofteh, Jorge Bravo (2021) Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. https://doi.org/10.24433/CO.1963899.v1

THE ERA OF COMPLEX MODELS

Ensemble models

LAYERED LEARNING AND ENSEMBLE LEARNING STRATEGY

DELMS for Time Series Modeling

Dynamic Ensemble Learning with an Intelligent Model Selection Strategy

- We use direct mapping from the time series of different countries, combining the intractable time series algorithms and predicting the ensemble model as the final output.

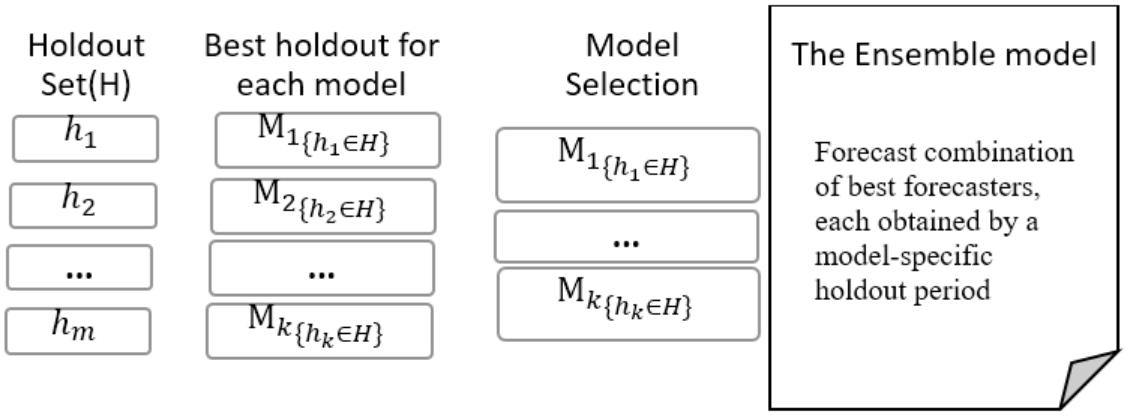


Figure 1 - Proposed strategy of ensemble learning.

The marginal posterior distribution across all models is

$$p(\Delta|y) = \sum_{k=1}^{K} p(\Delta|y, M_k) p(M_k|y)$$

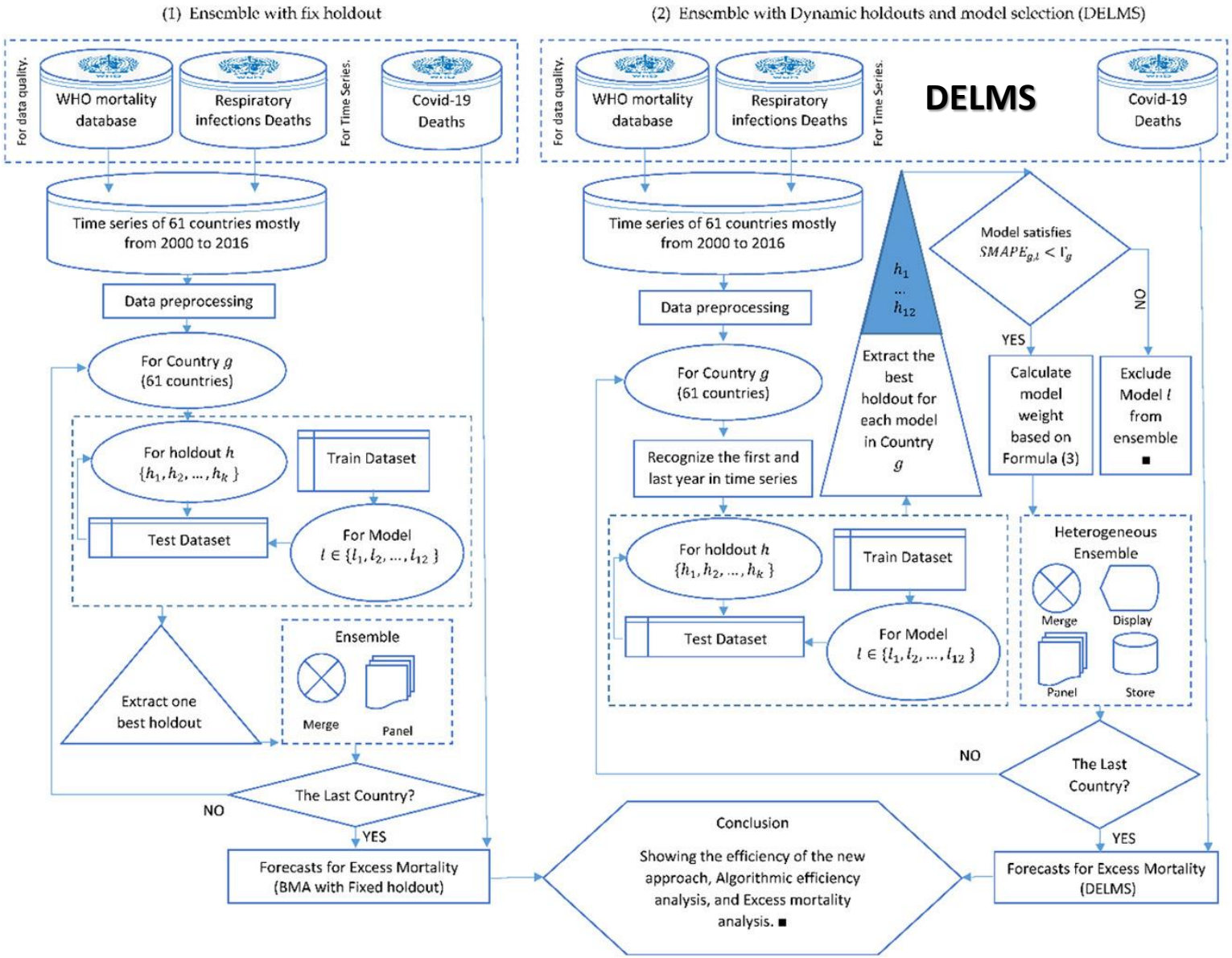
The weight assigned to each model M_k is given by its posterior probability

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}.$$

Publication: https://doi.org/10.1016/j.asoc.2022.109422

Video of presentation: https://youtu.be/7xabpWH4aFs

GRAPHICAL SUMMARY

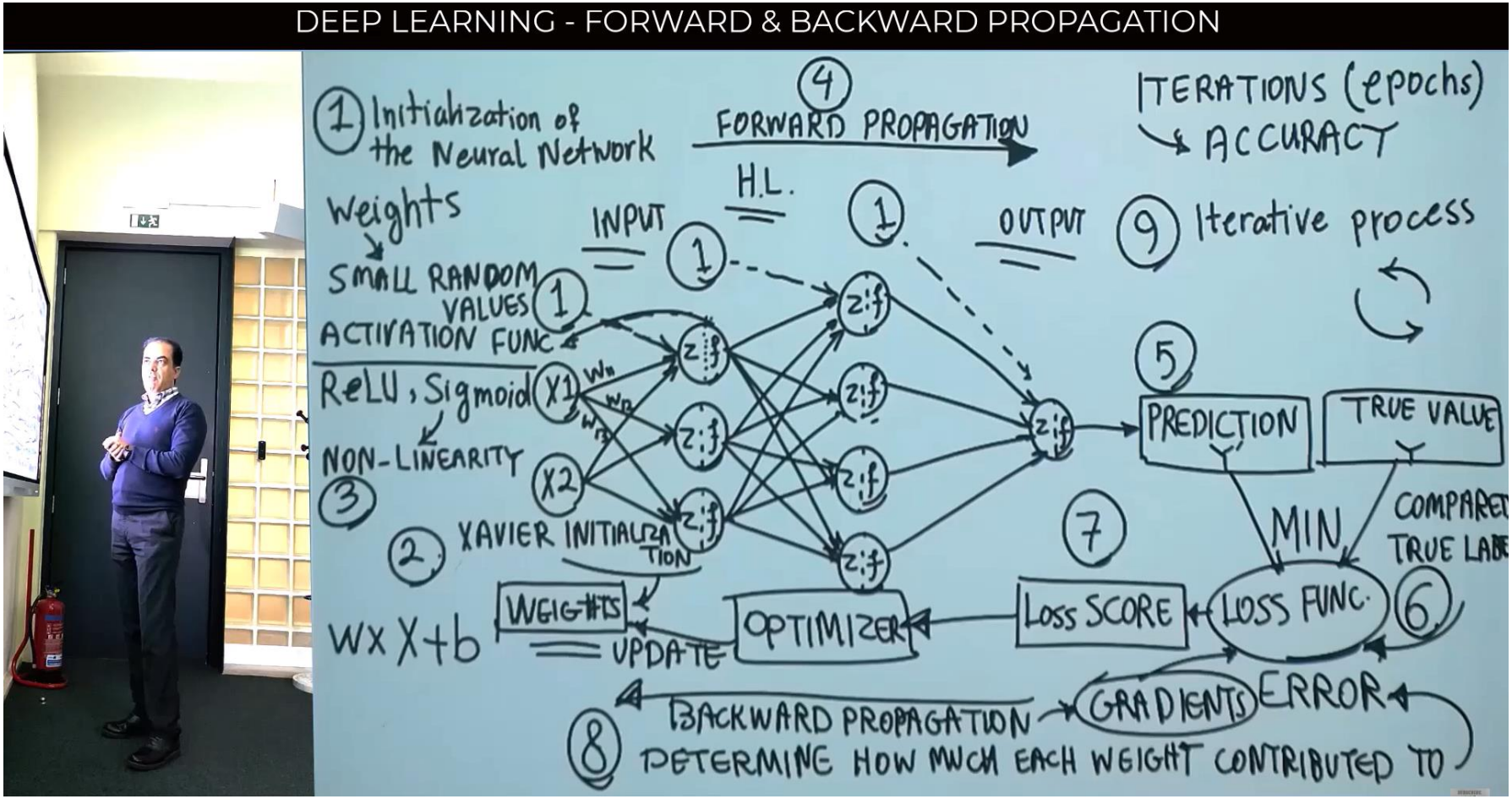


THE ERA OF DEEP LEARNING

Large Language Models (LLMs)

DEEP LEARNING

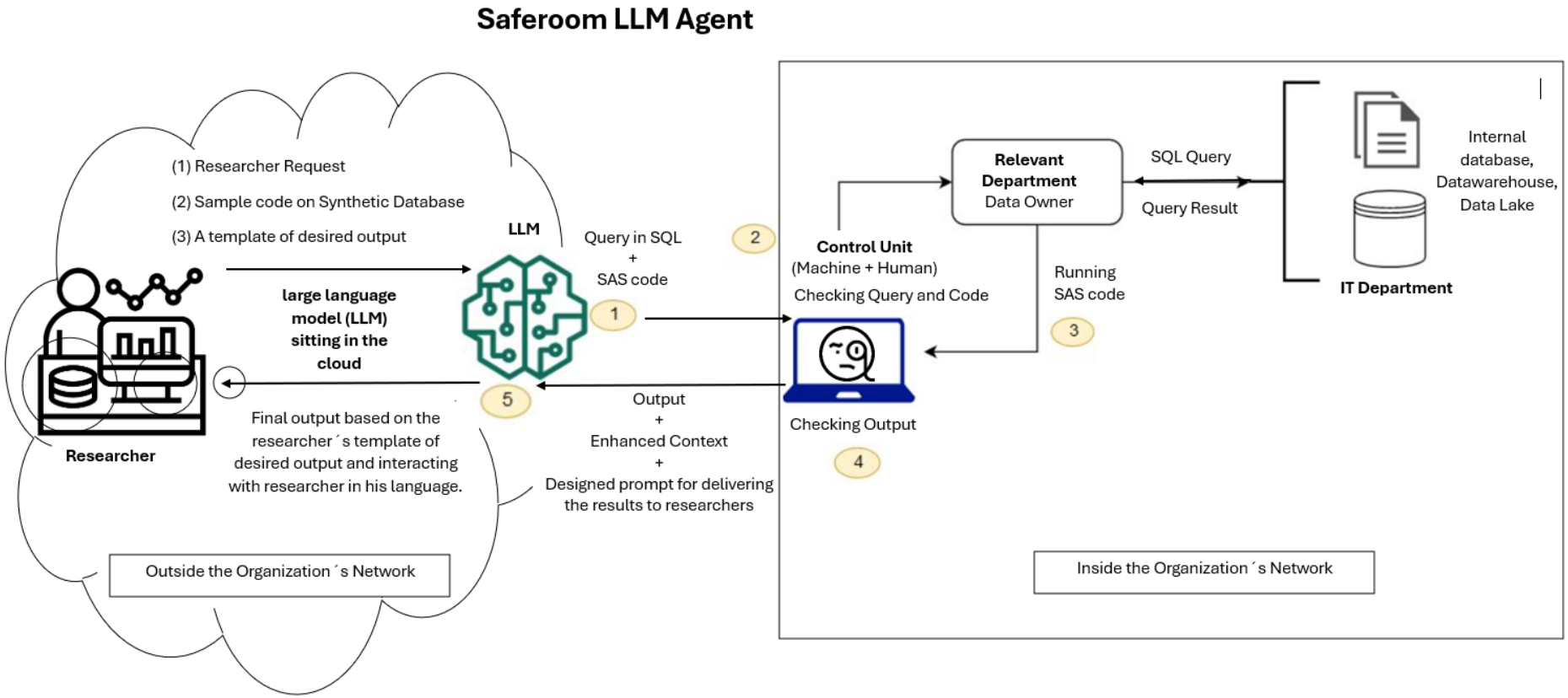
https://youtu.be/vLxgDQmM57Y



THE FOUNDATION OF LLMS

The

GRAPHICAL SUMMARY OF A PROJECT
LLMS FOR OFFICIAL STATISTICS



"SMARTEST AI"

Grok 3

deepseek

OpenAI

REFERENCES

Data Science Discussion Group
https://www.linkedin.com/groups/12420006/

Data Science YouTube
https://www.youtube.com/channel/UCTOuxIhJxcxNOntTpamJeAA

https://novaresearch.unl.pt/en/publications/a-conservative-approach-for-online-credit-scoring-2

https://novaresearch.unl.pt/en/publications/an-ensemble-learning-strategy-for-panel-time-series-forecasting-o

E-mail | Academic HP | ResearchGate

THANK YOU!

Afshin Ashofteh; PhD, PGDip, MBA, MSc

E-mail | Academic HP | Discussion Group | YouTube

Address: Campus de Campolide, 1070-312 Lisboa, Portugal

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa