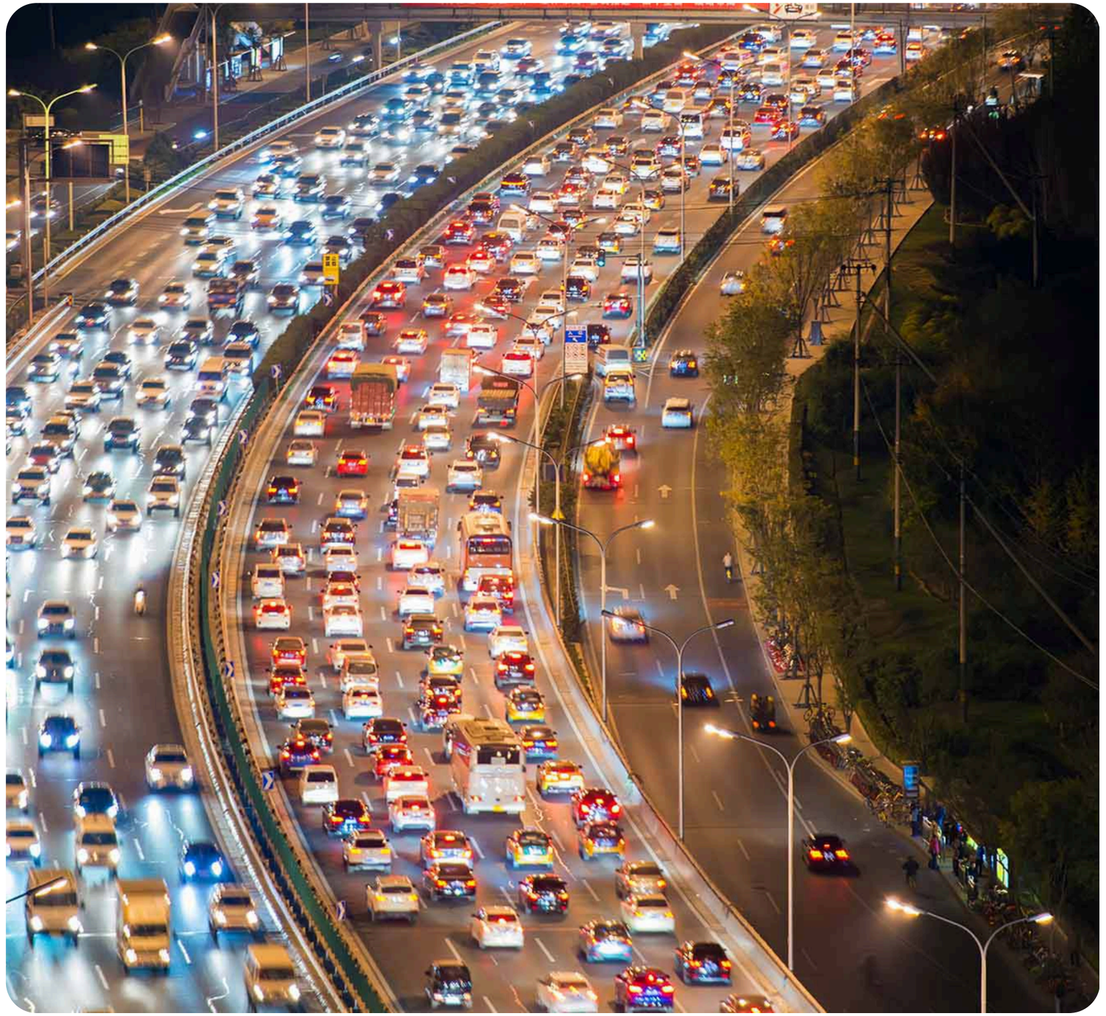
CATCH-UP

WEB TRAFFIC TIME SERIES FORECASTING

Presented by: Johan S. Beltrán Merchán, Edison D. Álvarez Varela,
Yader I. Quiroga Torres, Julián D. Celis Giraldo

INTRODUCTION AND OBJECTIVE

- The traffic to 145,000 Wikipedia pages needs to be predicted based on a historical data set.
- The goal is to fill in the missing information based on current information and then make predictions.



THE DATASET

01/07/2015-01/03/2017

train.csv

Page	########	########	########	###
2NE1_zh.wikipedia.org_all-access_spider	18	11	5	
2PM_zh.wikipedia.org_all-access_spider	11	14	15	
3C_zh.wikipedia.org_all-access_spider	1	0	1	
4minute_zh.wikipedia.org_all-access_spider	35	13	10	
52_Hz_I_Love_You_zh.wikipedia.org_all-access_spider				
5566_zh.wikipedia.org_all-access_spider	12	7	4	
91Days_zh.wikipedia.org_all-access_spider				
A'N'D_zh.wikipedia.org_all-access_spider	118	26	30	
AKB48_zh.wikipedia.org_all-access_spider	5	23	14	
ASCII_zh.wikipedia.org_all-access_spider	6	3	5	
ASTRO_zh.wikipedia.org_all-access_spider				
Ahq_e-Sports_Club_zh.wikipedia.org_all-access_spider	2	1	4	
All_your_base_are_belong_to_us_zh.wikipedia.org_all-a	2	5	5	
AlphaGo_zh.wikipedia.org_all-access_spider				
Android_zh.wikipedia.org_all-access_spider	8	27	9	

key.csv

Page	Id
!vote_en.wikipedia.org_all-access_all-agents_2017-01-01	bf4edcf969af
!vote_en.wikipedia.org_all-access_all-agents_2017-01-02	929ed2bf52b9
!vote_en.wikipedia.org_all-access_all-agents_2017-01-03	ff29d0f51d5c
!vote_en.wikipedia.org_all-access_all-agents_2017-01-04	e98873359be6
!vote_en.wikipedia.org_all-access_all-agents_2017-01-05	fa012434263a
!vote_en.wikipedia.org_all-access_all-agents_2017-01-06	48f1e93517a2
!vote_en.wikipedia.org_all-access_all-agents_2017-01-07	5def418fcb36
!vote_en.wikipedia.org_all-access_all-agents_2017-01-08	77bd08134351
!vote_en.wikipedia.org_all-access_all-agents_2017-01-09	5889e6dbb16f
!vote_en.wikipedia.org_all-access_all-agents_2017-01-10	5f21fef1d764
!vote_en.wikipedia.org_all-access_all-agents_2017-01-11	6f07e1b8815a
!vote_en.wikipedia.org_all-access_all-agents_2017-01-12	228e54b5dea0
!vote_en.wikipedia.org_all-access_all-agents_2017-01-13	da1b34963ed7

sample submission.csv

Id	Visits
bf4edcf969af	0
929ed2bf52b9	0
ff29d0f51d5c	0
e98873359be6	0
fa012434263a	0
48f1e93517a2	0
5def418fcb36	0
77bd08134351	0

WHAT HAVE WE IDENTIFIED?

Input components

- Input
- relationships
- Data characteristics

Process

- Feature engineering characteristics
- Model training and forecasting
- Validation

Output components

- Forecast in CSV
- SMAPE evaluation results

WHAT HAVE WE IDENTIFIED?

Complexity and sensitivity

- Structural complexity
- Intensive processing
- High sensitivity

Chaos and randomness

- Dynamic Chaos
- Inherent Randomness
- Unstable Prediction

SYSTEM REQUIREMENTS

01 FR1: DATA MANAGEMENT

02 FR2: PREDICTION ENGINE

03 FR3: FEATURE ENGINEERING

04 FR4: VALIDATION AND TESTING

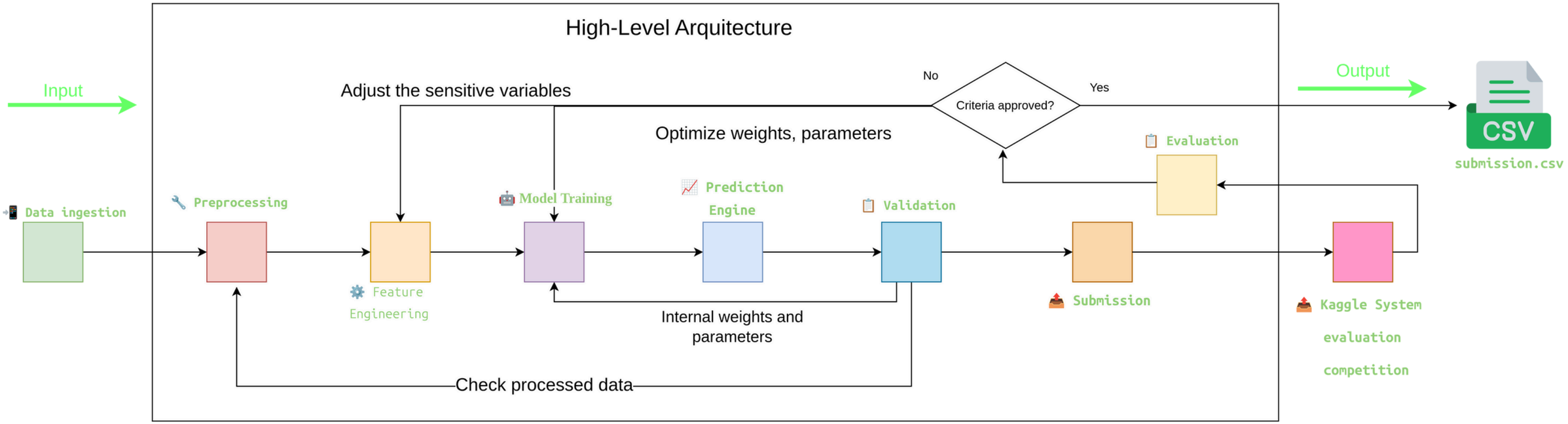
05 NFR1: SCALABILITY FOR LARGE DATA INGESTION

06 NFR2: SCALABILITY REQUIREMENTS

07 NFR3: RELIABILITY REQUIREMENTS

08 NFR4: ACCURACY REQUIREMENTS

HIGH-LEVEL ARCHITECTURE



- HOLISTIC APPROACH
- MODULARITY AND FUNCTIONAL DECOMPOSITION
- FEEDBACK

- OPEN SYSTEM
- HOMEOSTASIS
- EQUIFINALITY

- ADAPTATION AND EVOLUTION
- ENTROPY
- GOAL ORIENTATION

ADDRESSING SENSITIVITY AND CHAOS

Strategies for sensitive variables

- viral pikes and unpredictable events
- Manage bot and spider traffic
- Missing data and zeros

Handle caotic factors

- continuous adaptation and drift monitoring
- Complex seasons and non-linear patterns
- Handle low traffic series

Monitoring and error handling routines

- validation checkpoints
- cascading fallback
- post-prediction analysis

WHAT'S NEXT?


Design a robust system
Apply software engineering
Data and event-based simulations
Documentation and report

THANK YOU FOR YOUR ATTENTION

Thank you for joining us on this data project

WIKIPEDIA

JOHAN SEBASTIÁN
BELTRÁN MERCHÁN
jsbeltranm@udistrital.edu.co

EDISON DAVID
ÁLVAREZ VARELA
edalvarezv@udistrital.edu.co

YADER IBRALDO
QUIROGA TORRES
yiquirogat@udistrital.edu.co

JULIÁN DAVID
CELIS GIRALDO
jdcelisg@udistrital.edu.co