Kavli IPMU, University of Tokyo

PREDICTING THE BEHAVIOR OF VIDEO GAME PLAYERS WITH MACHINE LEARNING

Pei Pei Chen
Machine Learning Engineer Lead
YOKOZUNA data

KAVLI IPMU
東京大学 国際高等研究所 カブリ数物連携宇宙研究機構
KAVLI INSTITUTE FOR THE PHYSICS AND MATHEMATICS OF THE UNIVERSE

YOKOZUNA data
A KEYWORDS STUDIO

Pei Pei Chen

MACHINE LEARNING ENGINEER LEAD

Specialized in deep learning techniques and machine learning applied to sequential analysis

Experienced on developing scalable and operational machine learning systems with big data infrastructure

+5 years of experience in game and music-related data science research

Co-author of 10 peer-reviewed articles in data science

What is Yokozuna Data?

Founded in 2015, joined Keywords Studios in 2018

to push back the frontiers of General Behavioral Machine Learning

and to revamp the video-game industry: Personalized games

YOKOZUNA data
A KEYWORDS STUDIO

PUSHING BACK THE FRONTIERS OF GAME DATA SCIENCE:

A state-of-the-art machine learning engine that predicts individual player behavior

YOKOZUNA data
A KEYWORDS STUDIO

Mission

To unlock the knowledge of big game databases

To convert unstructured data into actionable information in order to understand and predict individual player behavior

YOKOZUNA data
A KEYWORDS STUDIO

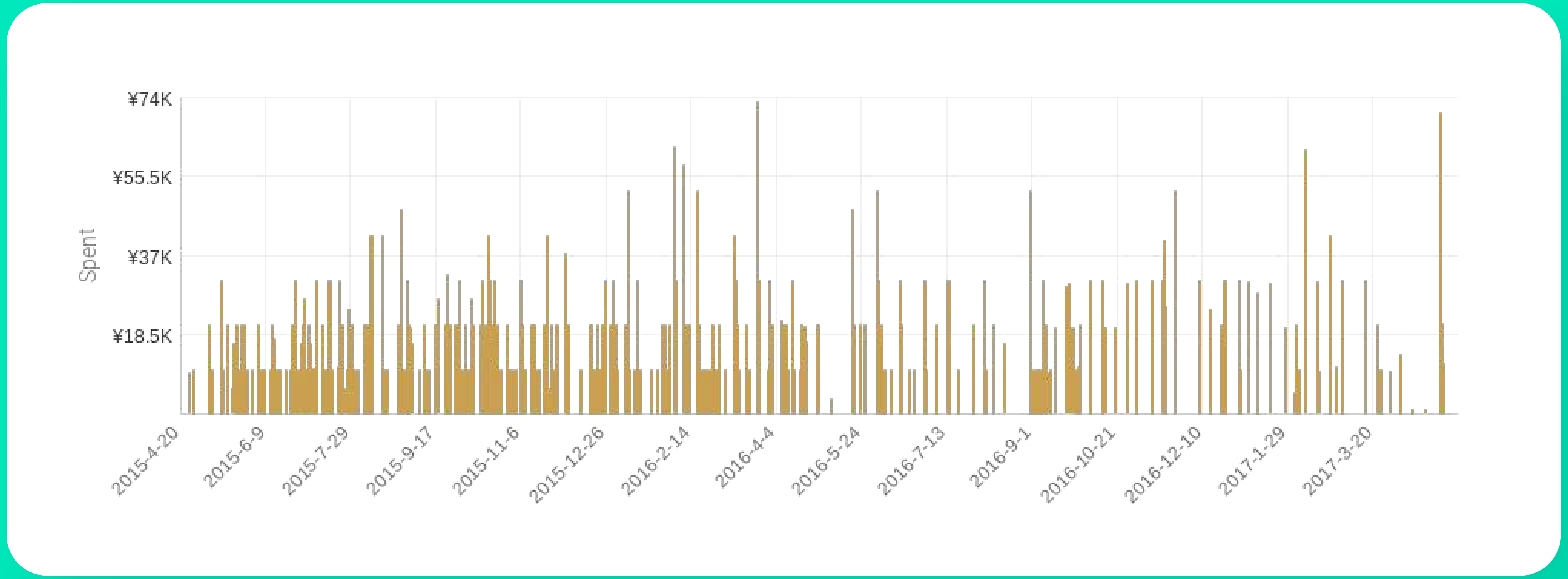
VIDEO

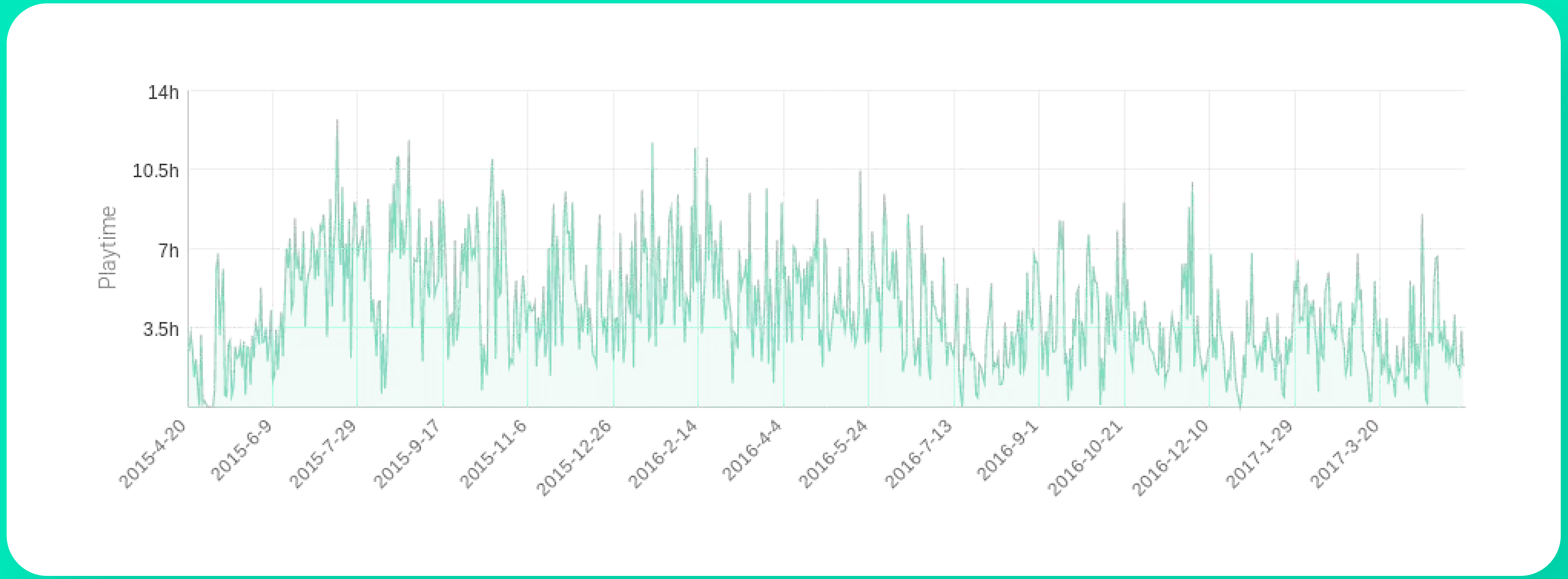
Highly sophisticated games allow players
to express nuanced emotions through their in-game actions

GAMES

VIDEO GAME DATA

Logins
Actions
In-app Purchases
Virtual Purchases
Items Selected
Playtime
Time Frame
Social Interactions
In-game Level-ups





THE TEAM


África Periáñez, PhD
Founder & CEO


Ana Fernández, MSc
SENIOR RESEARCH DATA SCIENTIST


Pei Pei Chen, MSc
MACHINE LEARNING ENGINEER LEAD


Shi Hui Tan, MSc
DATA SCIENTIST


Anna Guitart, MSc
DATA SCIENTIST


Jing Li, PhD
MACHINE LEARNING ENGINEER


Cristian Conteduca, MSc
BACKEND ENGINEER LEAD


Peng Xiao, BSc
BIG DATA ENGINEER


Nitin Kumar, MSc
FULL-STACK ENGINEER


Pooja Revanna, MSc
BACKEND ENGINEER


Dexian Tang, MSc
BIG DATA ENGINEER


Vitor Santos, MA
DESIGN & BUSINESS DIRECTOR


Álvaro de Benito, MA
PR & COMMUNICATION LEAD


Yu-Kai Hung, MSc
COMMUNITY MANAGER FOR ASIA


Omid Aladini, MSc
DATA INFRASTRUCTURE ADVISOR


Javier Grande, PhD
SCIENTIFIC EDITOR

YOKOZUNA DATA PEER-REVIEWED ARTICLES

CHURN PREDICTION IN MOBILE SOCIAL GAMES: TOWARDS A COMPLETE ASSESSMENT USING SURVIVAL ENSEMBLES

A. Periáñez, A. Saas,
A. Guitart and C. Magne
IEEE DSAA 2016 Montreal

DISCOVERING PLAYING PATTERNS: TIME SERIES CLUSTERING OF FREE-TO-PLAY GAME DATA

A. Guitart, A. Periáñez
and A. Saas,
IEEE CIG 2016 Santorini

GAMES AND BIG DATA: A SCALABLE MULTI-DIMENSIONAL CHURN PREDICTION MODEL

P. Bertens, A. Guitart
and A. Periáñez, A. Saas,
IEEE CIG 2017 New York

FORECASTING PLAYER BEHAVIORAL DATA AND SIMULATING IN-GAME EVENTS

A. Guitart, P. Chen,
P. Bertens and A. Periáñez
IEEE FICC 2018 Singapore

THE WINNING SOLUTION TO THE IEEE CIG 2017 GAME DATA MINING COMPETITION

A. Guitart, P. Chen,
A. Periáñez

A MACHINE-LEARNING ITEM RECOMMENDATION SYSTEM FOR VIDEO GAMES

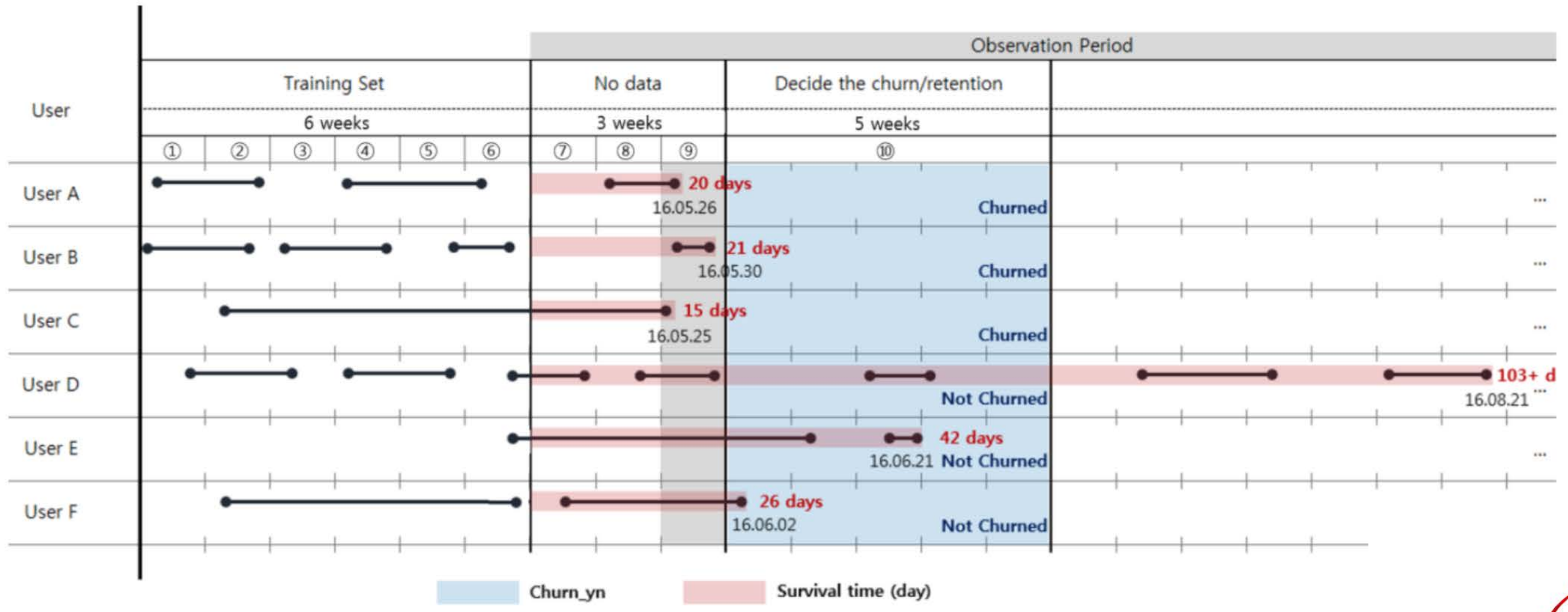
P. Chen, A. Guitart, P. Bertens,
A. Periáñez
IEEE CIG 2018 Maastricht

CUSTOMER LIFETIME VALUE IN VIDEO GAMES USING DEEP LEARNING AND PARAMETRIC MODELS

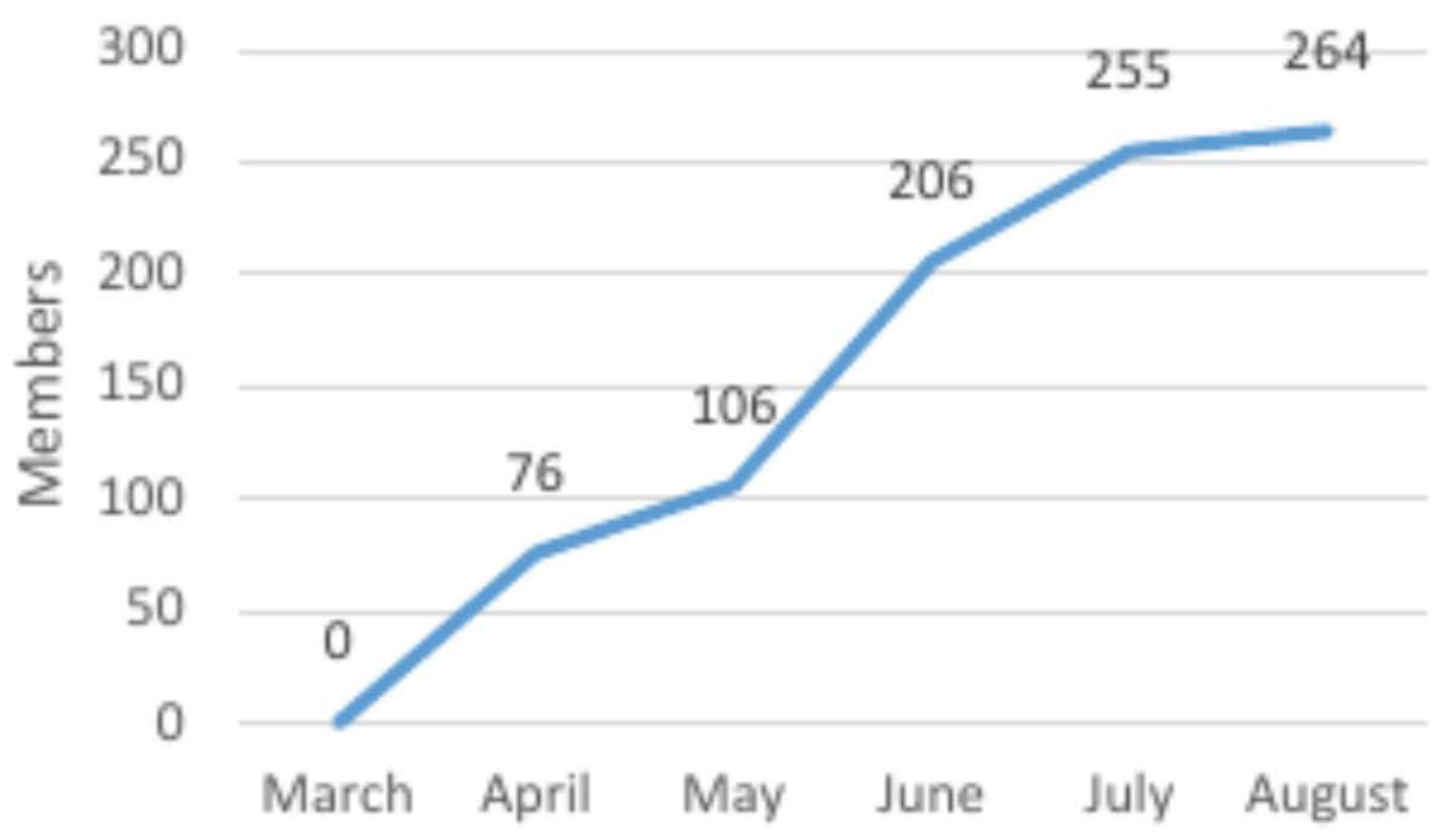
P. Chen, A. Guitart,
A. Fernández del Río and A. Periáñez
IEEE Big Data 2018 Seattle

THE WINNING SOLUTION TO THE IEEE CIG 2017 GAME DATA MINING COMPETITION

TWO TRACKS:
WHO IS GOING TO LEAVE THE GAME AND WHEN IS GOING TO DO IT



NUMBER OF REGISTRANTS IN COMPETITION



Results

Track 1 Which players will leave the game

Rank	Team	Test1 Score	Test2 Score	Total Score
	YokozunaData (Japan)	**0.610098**	**0.63326**	**0.62145**
	UTU (Finland)	0.60326	0.60370	0.60348
	TripleS (Korea)	0.57968	0.62459	0.60130
	TheCowKing	0.59370	0.60718	0.60036
	goedleio	0.57717	0.56205	0.58882

Track 2 When they will leave the game

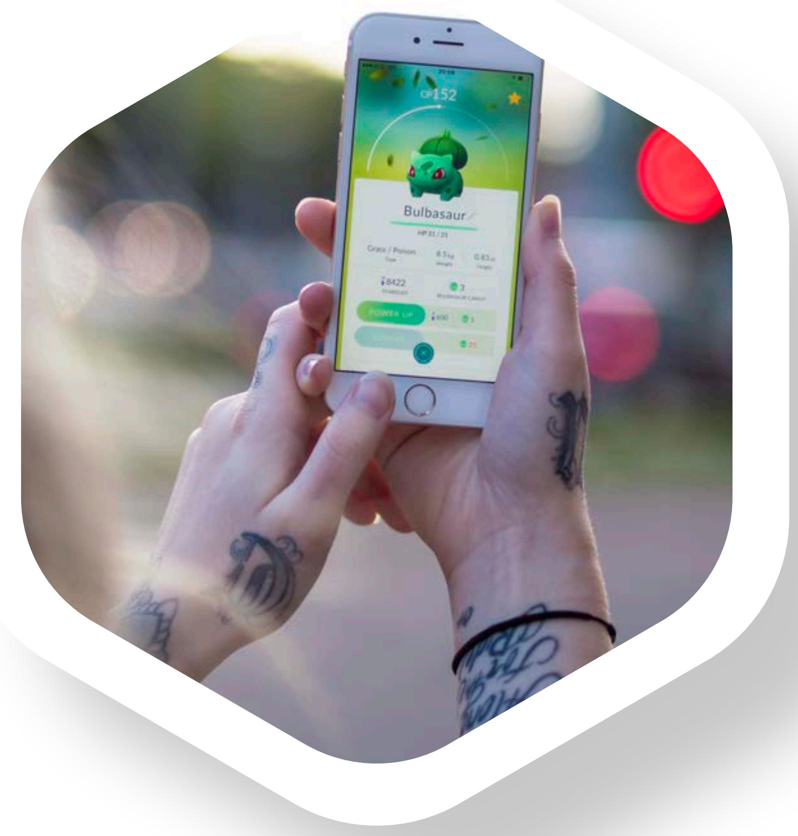
Rank	Team	Test1 Score	Test2 Score	Total Score
1	**YokozunaData (Japan)**	**0.883248**	**0.616499**	**0.726151**
2	IISLABSKKU	1.034321	0.679214	0.819972
3	UTU (Finland)	0.927712	0.898471	0.912857
4	TripleS (Korea)	0.958308	0.891106	0.923486
5	DTND	1.032688	0.930417	0.978888

WHAT CAN **MACHINE LEARNING** DO FOR VIDEOGAMES?

WHEN WILL PLAYERS LEAVE THE GAME?

DATE

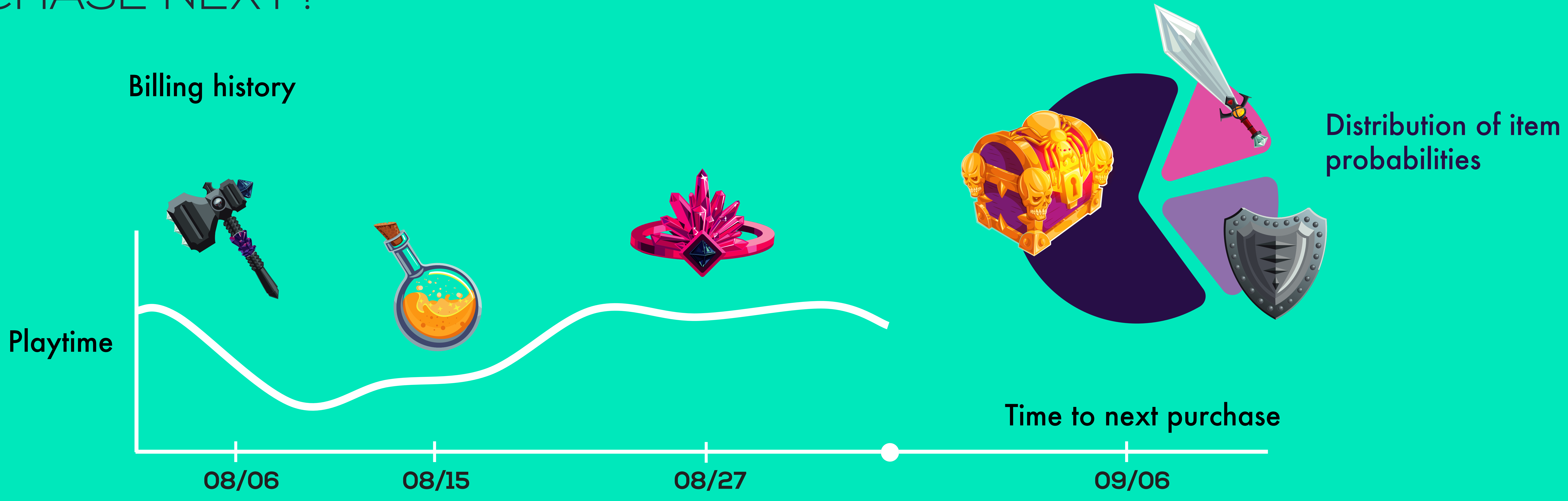
LEVEL



PLAYTIME

MONEY

WHICH ITEM WILL THEY PURCHASE NEXT?



GAME APPLICATIONS

- Players who may stop purchasing
- Upcoming churners

RECOMMENDS THE BEST
SEQUENCE OF **EVENTS** TO MAXIMIZE
PLAYER ENGAGEMENT CONSIDERING
EXTERNAL AND ENVIRONMENTAL FACTORS

PERSONALIZATION

Individual playlists

Product recommendations

Personalized film selection

Individual search results

PERSONALIZATION

Personalized matching

Which clan is your best opponent in Clash of Clans?

Who should you compete against in Mario Kart?

PERSONALIZATION

Engagement and retention-motivated actionable recommendations

Item recommendation system

Action recommendation

Rewards and discounts

CUSTOMER LIFETIME VALUE IN VIDEO GAMES

Deep Learning Approaches

Reference: Pei Pei Chen, Anna Guitart, Ana Fernández del Río and África Periáñez. "Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models." International Conference on Big Data (Big Data), 2018, IEEE, p. 2134-2140

DATA SET

THE GAME:
A RPG mobile game, freemium, social game with several millions of players worldwide

BEHAVIOUR LOGS:
logins, level-ups, purchases, playtime, actions, social, etc

TRAINING PERIOD:
2014-09-24 to 2016-04-30

EVALUATION PERIOD:
2016-05-01 to 2017-04-30

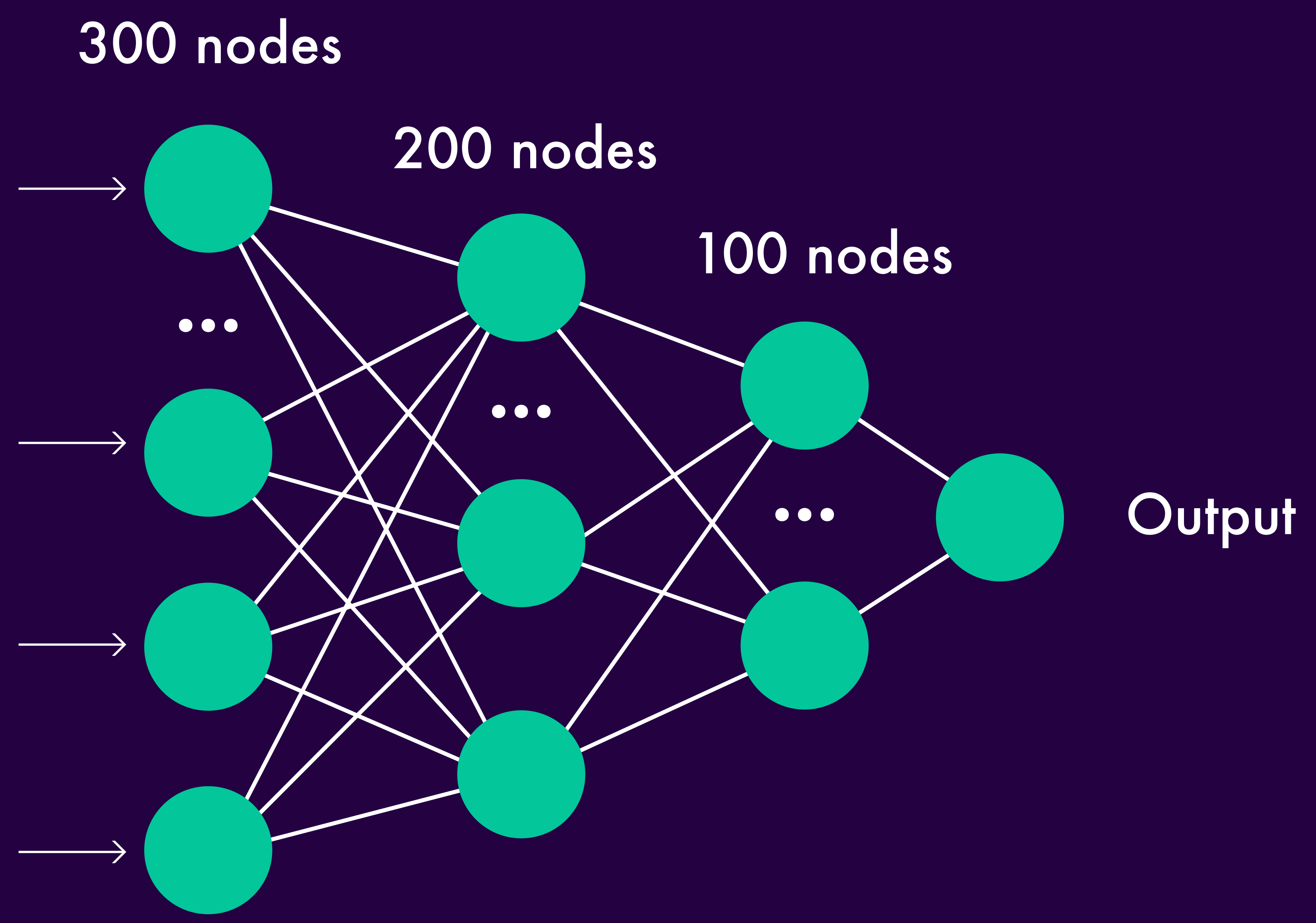
TARGET:
for every single paying user, predict LTV

DEEP MULTILAYER PERCEPTRON (MLP)

Features - statistical features (feature engineering beforehand required) e.g. average daily playtime, average days of one level-up, loyalty index (login days / lifetime)

TRAINING PROCEDURE

1. Initialize weights - Xavier initialization
2. Forward propagation
3. Calculate the total error - root-mean-square error
4. Back propagation
(Gradient descent optimization algorithm) - ADAM
5. Repeat 2 - 4 to minimize error



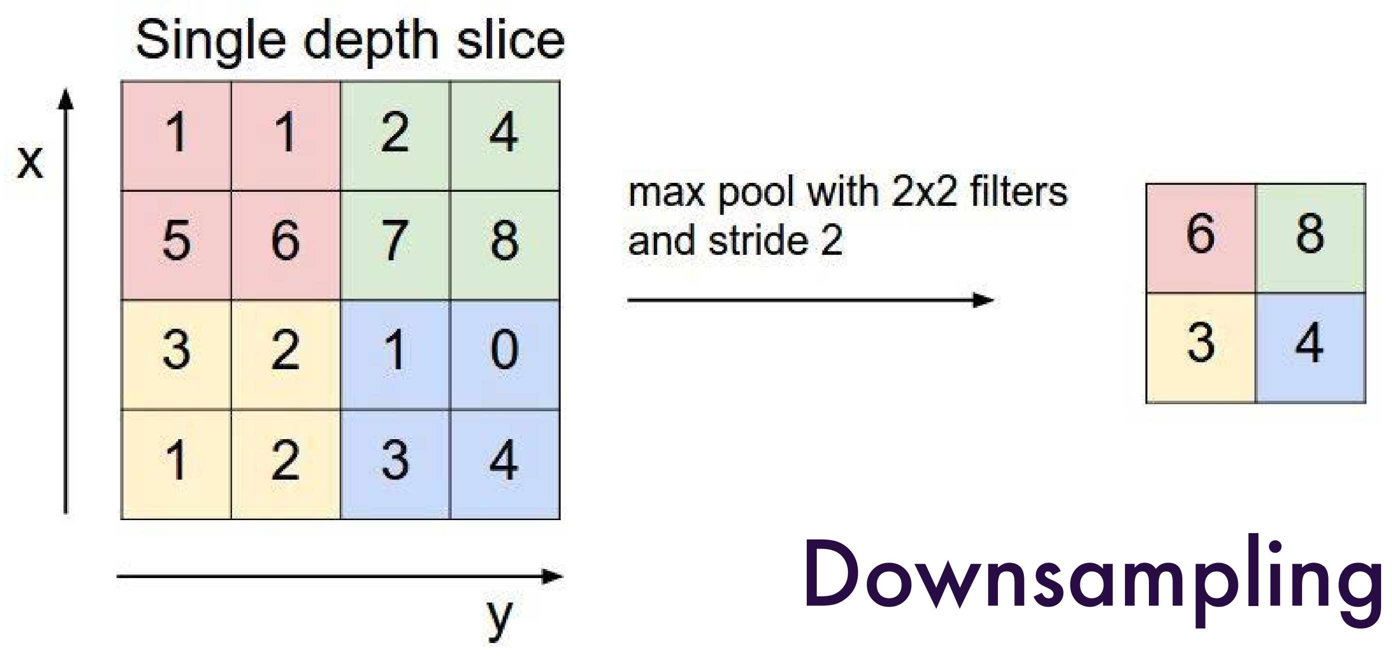
Reference: https://medium.com/deep-math-machine-learning-ai/chapter-7-artificial-neural-networks-with-math-bb711169481b

CONVOLUTIONAL NEURAL NETWORK (CNN, CONVNET)

Convolutional layers

Training:
Optimize weights and bias
to minimize error.

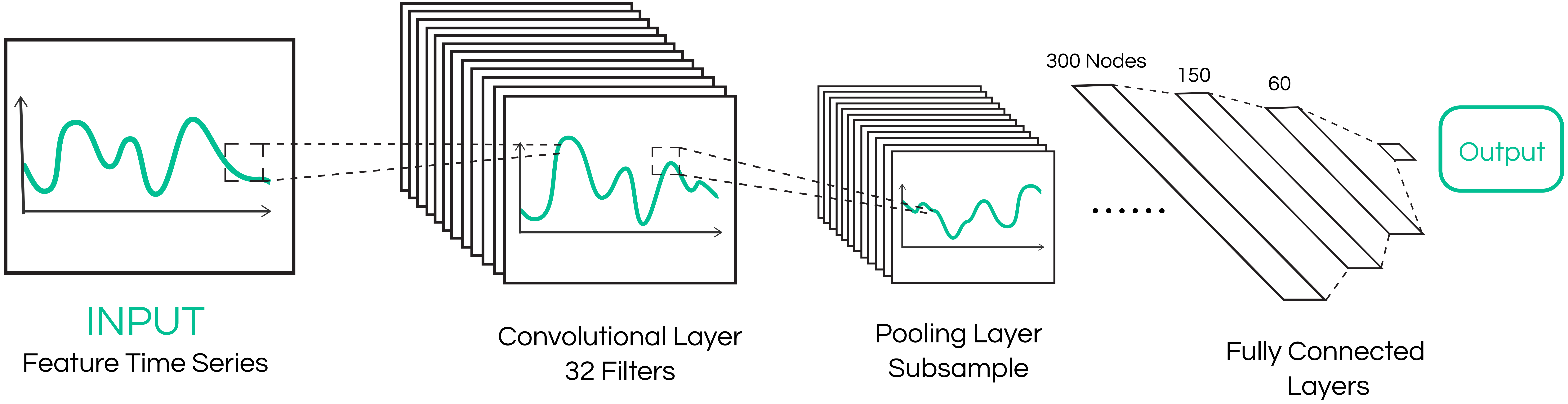
Max pooling layers



With filters that cover more than one input, CNNs can learn local connectivity between inputs.
Proposed by LeCun et al. in 1998 [1], CNNs have been widely applied to image processing, signal processing, and time series prediction.

[1] LeCun, Yann, et al. "Object recognition with gradient-based learning." Shape, contour and grouping in computer vision. Springer, Berlin, Heidelberg, 1999. 319-345.
Reference:
https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/
https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050

OUR CNN STRUCTURE (Simplified with only one time series)



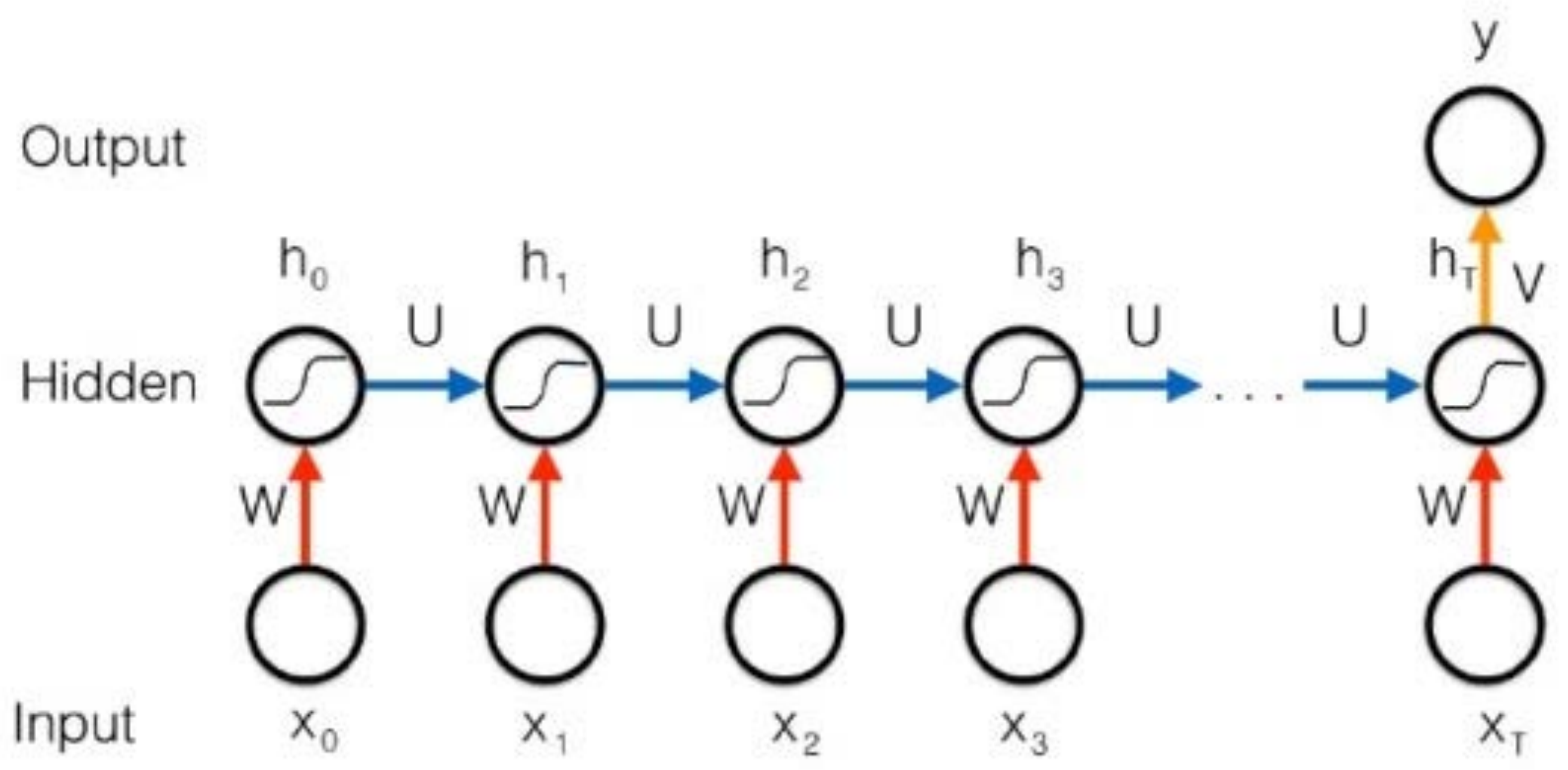
Feature daily time series input:
Playtime, level, sales, sessions, actions, number of purchases

* Values on days without login are filled up with zeros

Compared with MLP, CNN can help saving computation time on features calculation, and take relationship between time stamps into account.

RECURRENT NEURAL NETWORK (RNN)

RNNs are used to learn patterns in sequences of data, such as text, handwriting, the spoken word, and time series data.



$h_0 = \sigma(W \times x_0)$ $\quad h_t = \sigma(U \times h_{t-1} + W \times x_t)$ $\quad y = V \times h_t$

W: input to hidden weights
U: hidden to hidden weights
V: hidden to label weights

Reference: https://mlalgorithm.wordpress.com/2016/08/04/deep-learning-part-2-recurrent-neural-networks-rnn/

LONG SHORT-TERM MEMORY NETWORK (LSTM)

LSTM has to forget the gate to learn to memorize more important information in long time series.

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$ Input gate

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$ Forget gate

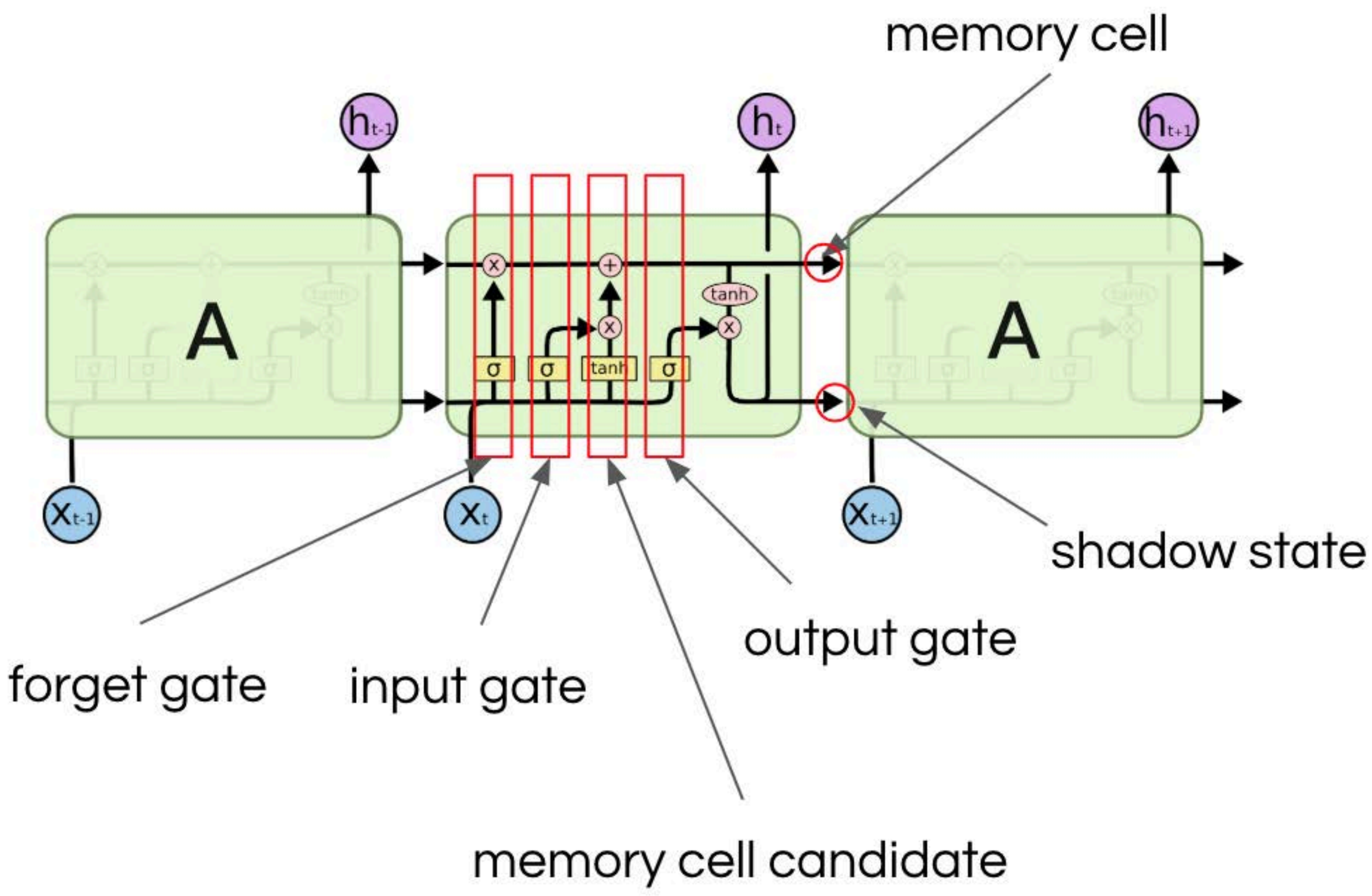
$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$ Output gate

$$\widetilde{c}_t = tanh(W h_{t-1} + U x_t + b)$$ Memory cell candidate

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c}_t$$ Memory cell

$$h_t = o_t \circ tanh(c_t)$$ Shadow state

$$y_t = h_t$$ Cell Output



Reference & image sources: http://colah.github.io/posts/2015-08-Understanding-LSTMs/
https://docs.google.com/presentation/d/1UHXrKL1oTdgMLoAHHPfMM_srDO0BCyJXPmhe4DNh_G8/

PARAMETRIC MODELS - PARETO / NBD (+ AVERAGE)

- Pareto distribution: obtain a binary classification
(indicating whether the customer is still active or not, the so-called dropout process)

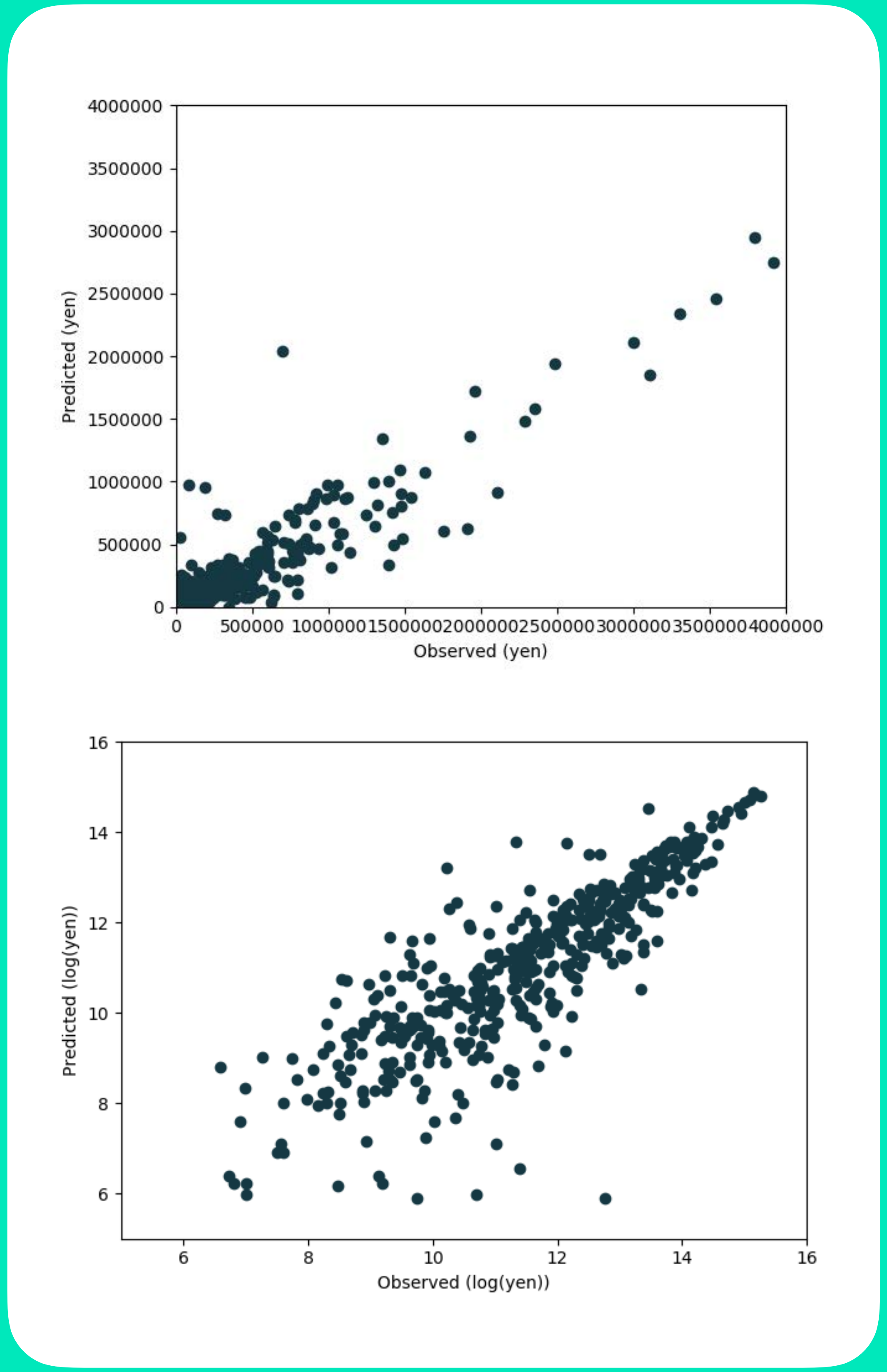
- Negative binomial distribution (NBD): estimate the purchase frequency

- Average: estimate monetary value

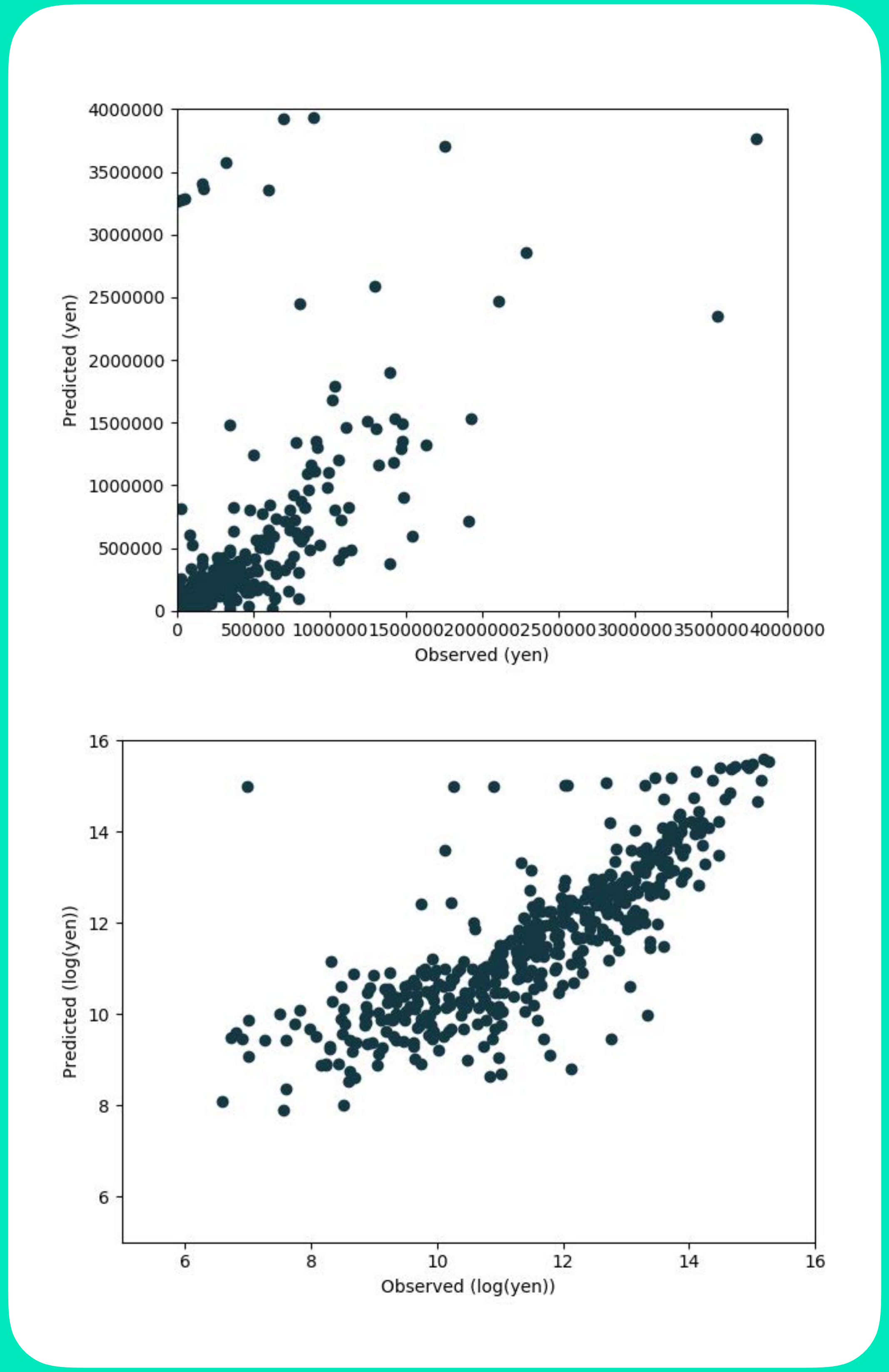
Pareto/NBD models and their extensions have been commonly used to estimate customer lifetime value (CLV) in many fields such as financial services, video games, and mobile prepaid subscribers.

Sifa, Rafet, et al. "Predicting purchase decisions in mobile free-to-play games." Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference. 2015.
Can, Zehra, and Erinç Albey. "Churn Prediction for Mobile Prepaid Subscribers." DATA. 2017.

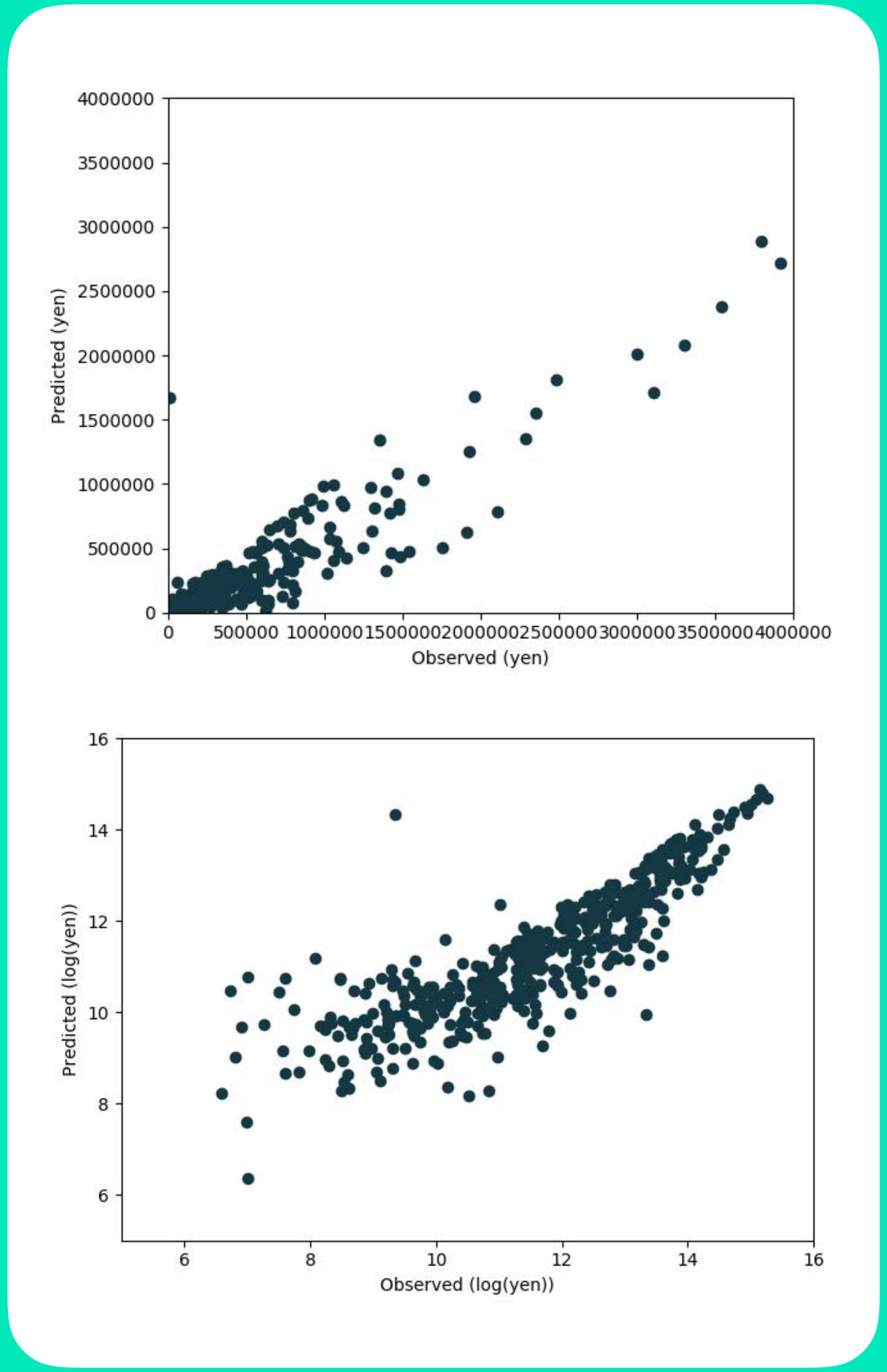
RESULTS VISUALIZATION



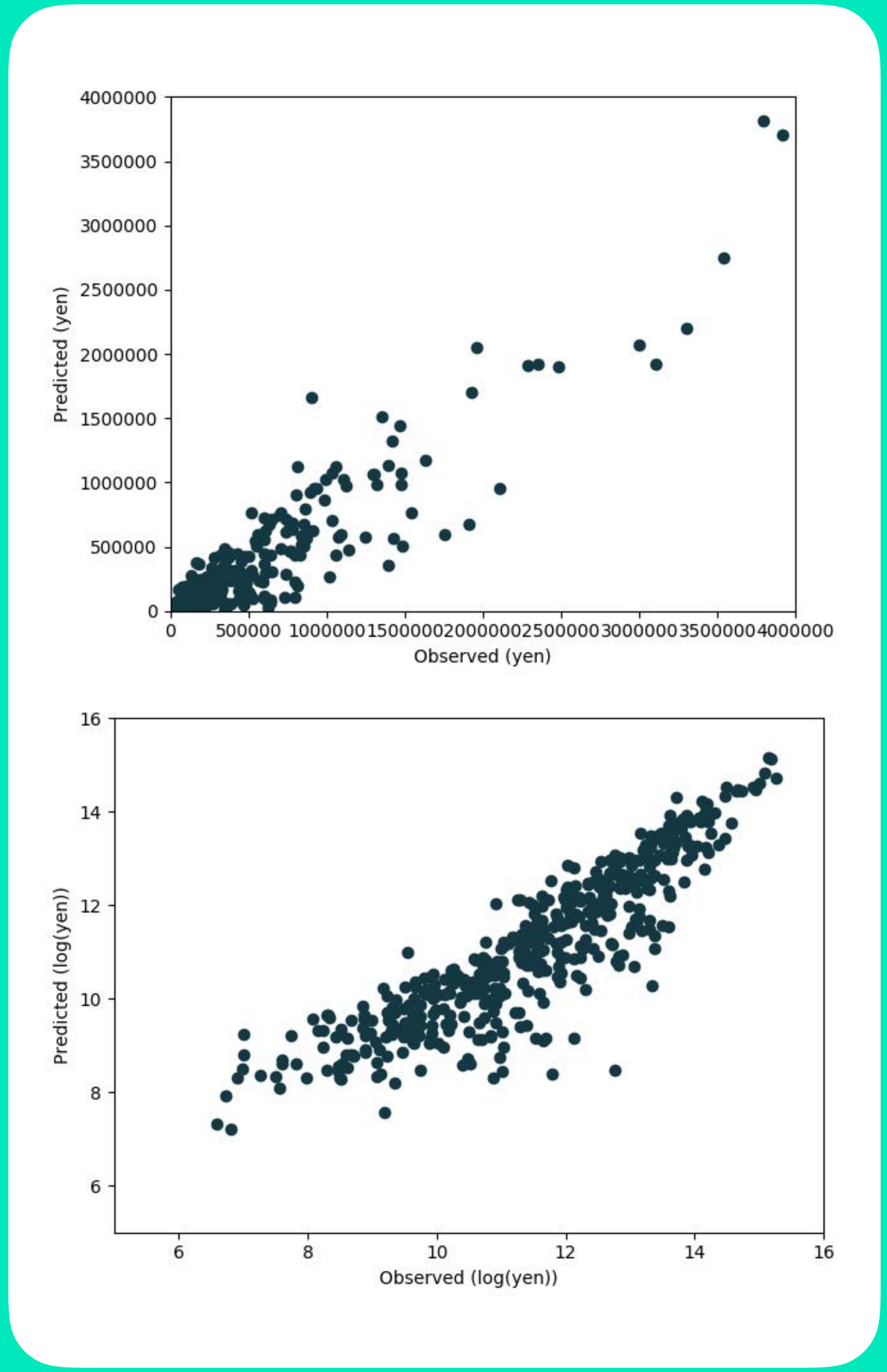
Pareto/NBD



MLP



CNN



LSTM

RESULTS - ERROR ESTIMATION

	RMSLE	SMAPE
Pareto/NBD	1.10	64.45
MLP	1.04	57.95
CNN	0.96	60.07
LSTM	0.87	54.23

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(o_i + 1) - log(p_i + 1))^2}$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|o_i - p_i|}{(|o_i| + |p_i|)/2}$$

* RMSLE: Root Mean Squared Logarithm Error
* SMAPE: Symmetric Mean Absolute Percent Error (the absolute differences between the observed and predicted values divided by their average)

CONCLUSIONS

- LSTM outperformed other models in terms of accuracy.
- Working with time series raw data, CNN and LSTM help saving features computation time.

COMPARISON BETWEEN LSTM AND CNN

- LSTM has flexibility on time series length. When adding new time stamps, we don't need to train again the model.
- CNN has more variables (filter size, stride length) to be decided before training. If the variable are selected well, the training can be more efficient and the results would be stable.
- LSTM learns all the relationship by itself. It helps saving time on parameter tuning, but is slower than CNN.

AN OPERATIONAL PREDICTION SYSTEM BIG DATA ENGINEERING INFRASTRUCTURE

SCALING TO INFINITY

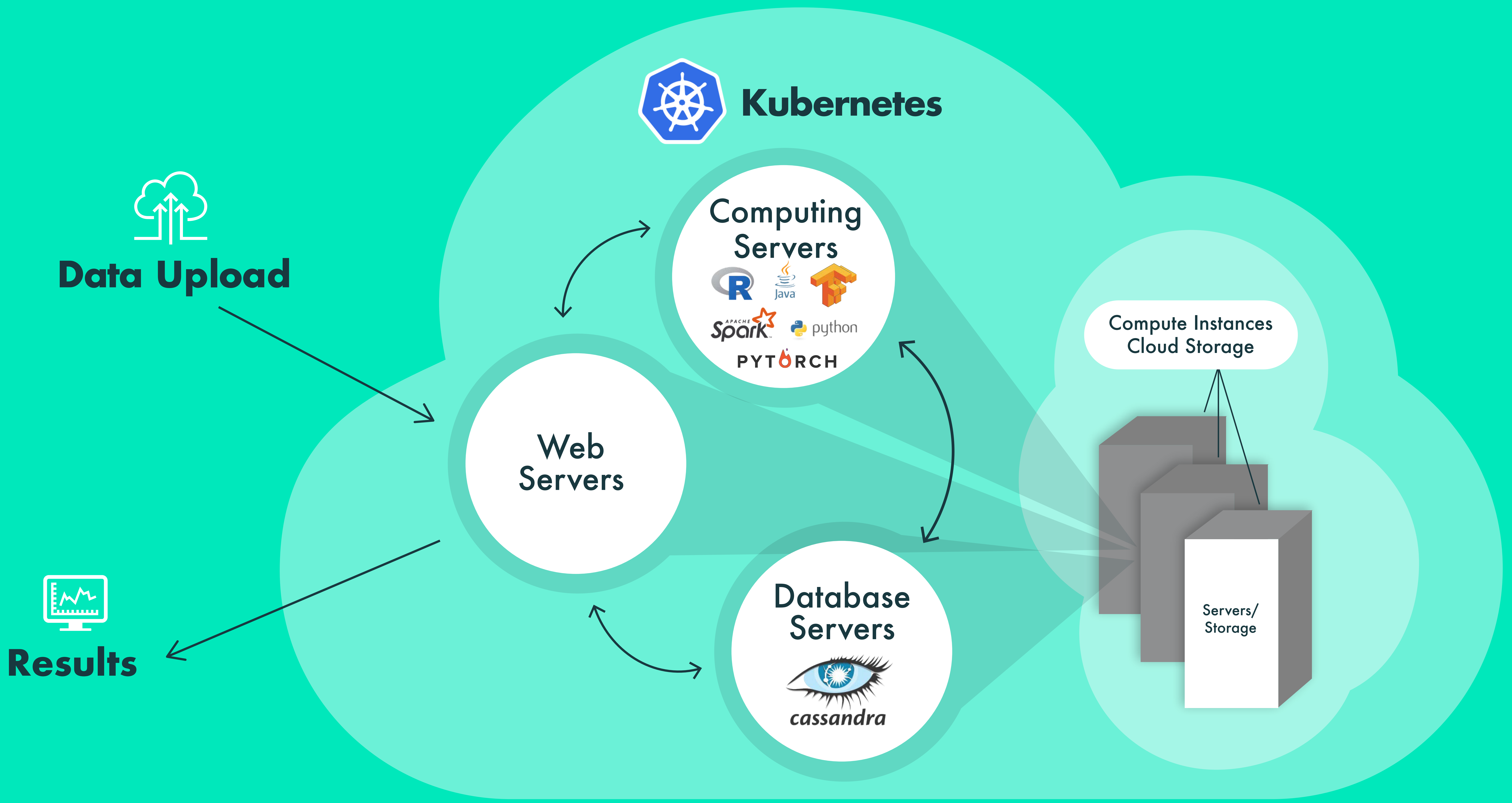
The Problem

Supporting data from thousands of games
and millions of Monthly Active Users (MAU)

The Solution

A CLOUD DISTRIBUTED-SYSTEM DESIGN FOR:
- Data upload
- Databases and storage
- Parallel computing for data processing
  and machine learning execution

UPLOAD → COMPUTE → RESULTS



CONTACT

ppchen@yokozunadata.com

@yokozunadata

linkedin.com/company/yokozunadata

www.yokozunadata.com

YOKOZUNA data

A KEYWORDS STUDIO

YOKOZUNA data
A KEYWORDS STUDIO

THANK YOU! :)