ML Developers for Real Estate Developers

Ekapol Chuangsuwanich

Joint work with Parichat Chonwiharnphan, Pipop Thienprapasith, Proadpran Punyabukkana, Atiwong Suchato, Naruemon Pratanwanich, Ekkalak Leelasornchai, Nattapat Boonprakong, Panthon Imemkamon

About me

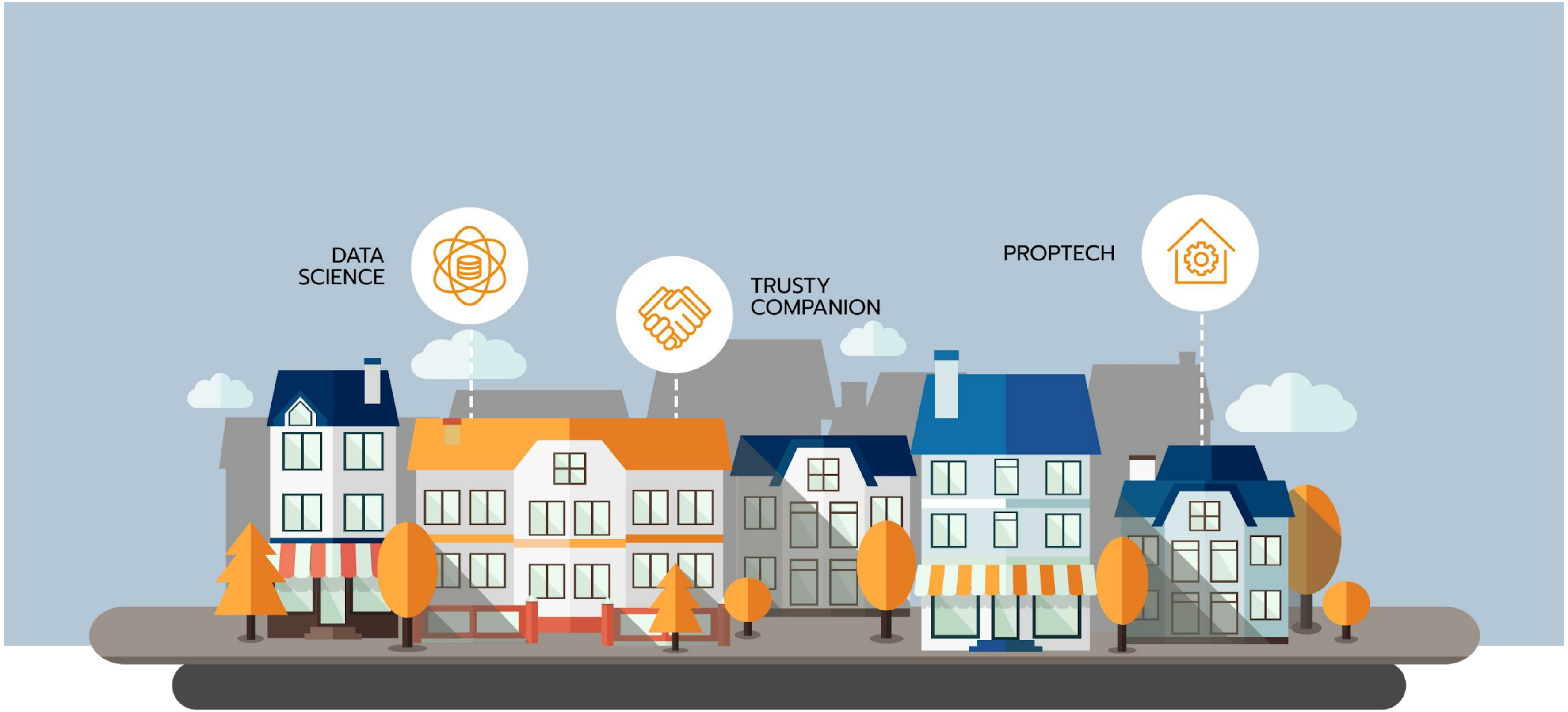
Lecturer at Chulalongkorn University

Research focus: ASR, NLP, Bioinformatics, or anything interesting

Various industry collaborations

Ex-intern Google Speech team, a tensorflow fanboy

About HomeDotTech



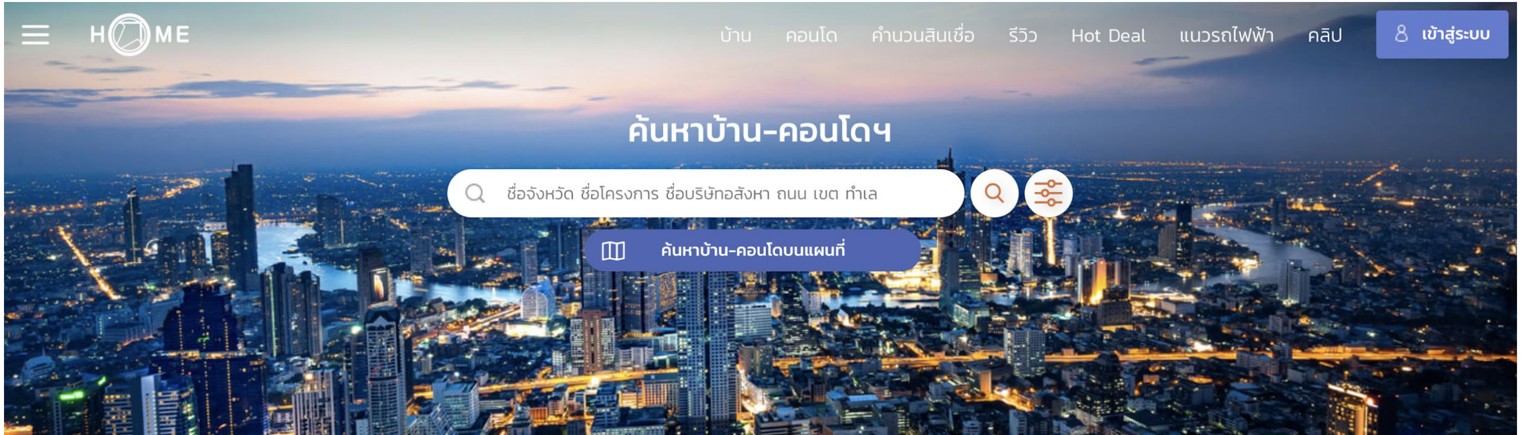
About HomeDotTech

Part of Home Buyer's Group

http://home.co.th

One of the most visited Real Estate Listings website in Thailand

~2,000,000 page views per month



Real Estate

The most expensive purchase for most people

Little prior experience

Top complaints to the Office of the Consumer Protection Board (สคบ.)

Homedottech’s mission is to help with the home buying process.

Data science for Real Estate

Consumer

Matching

Social listening

(Real Estate) Developers

Lead generation and smart marketing

Social listening

Project development
- Customer segmentation
- Trend prediction
- Pricing

Data science for Real Estate

Consumer

Matching

Social listening

(Real Estate) Developers

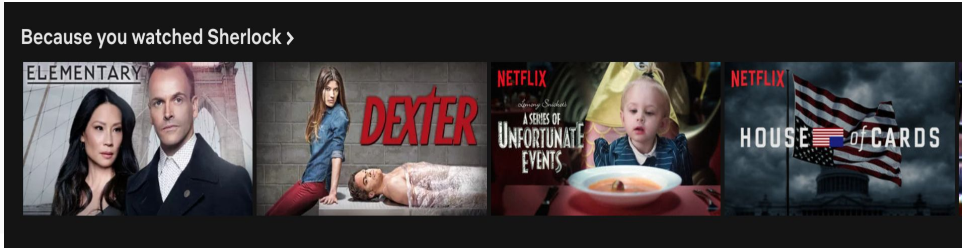
Lead generation and smart marketing

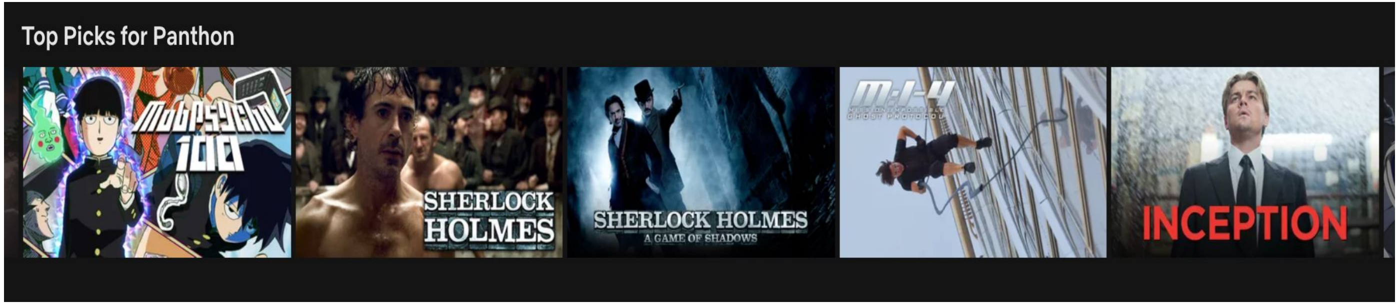
Social listening

Project development
- Customer segmentation
- Trend prediction
- Pricing

Recommendation systems

Goal: predict user's preference toward an item



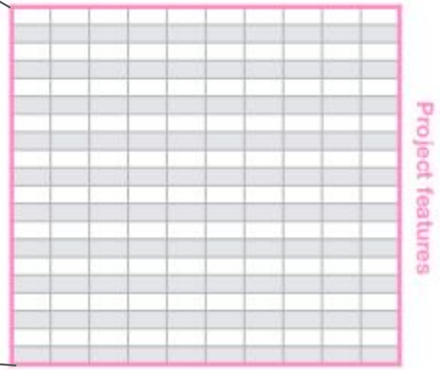


Information for recommendation systems

There’s many information available

Product and info

	1	1	1	1	0
	0	1	0	1	0
	1	1	0	1	0
	0	0	1	0	1
	1	1	1	0	1

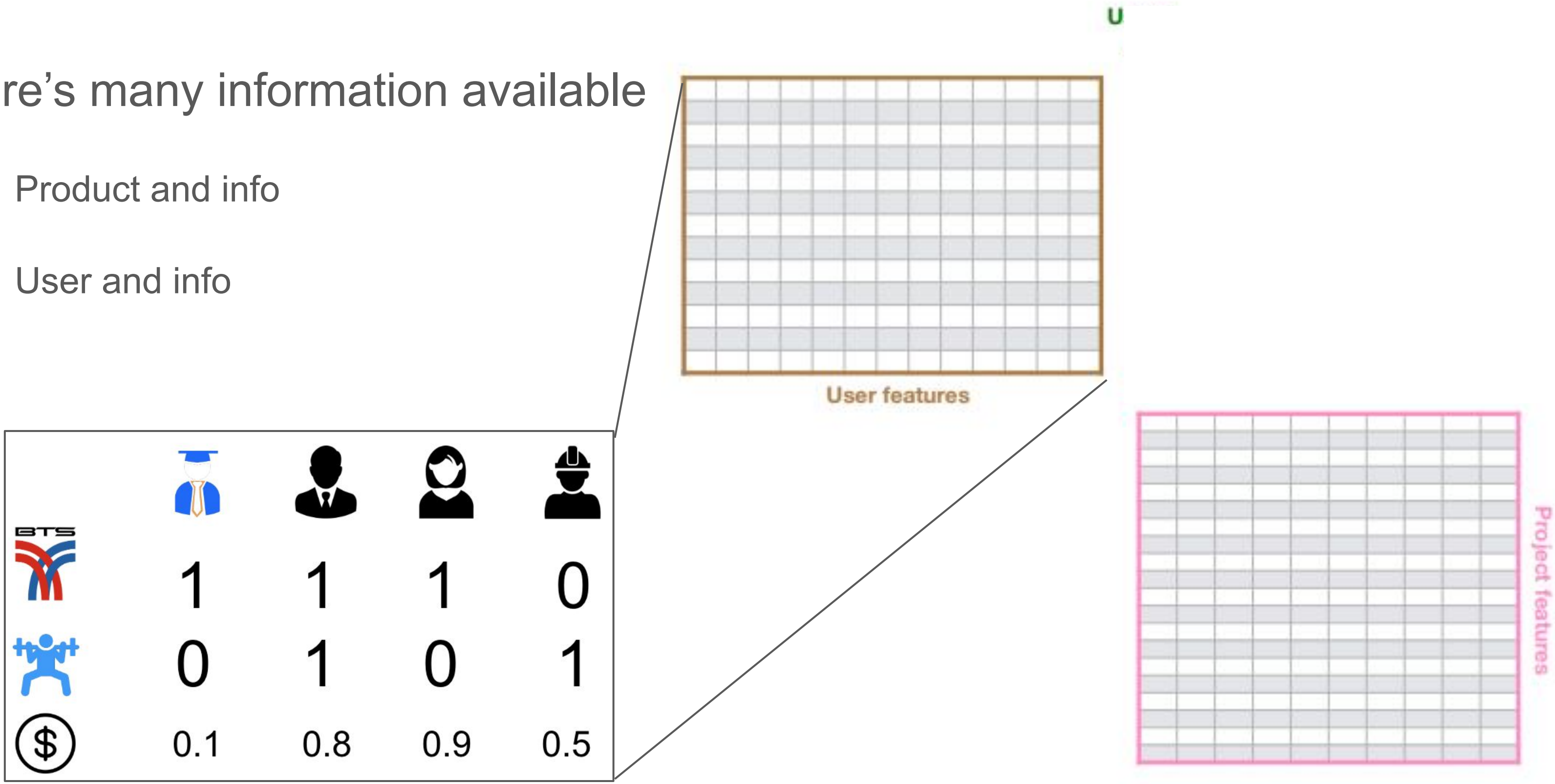


Icons by Freepik

Information for recommendation systems

There's many information available

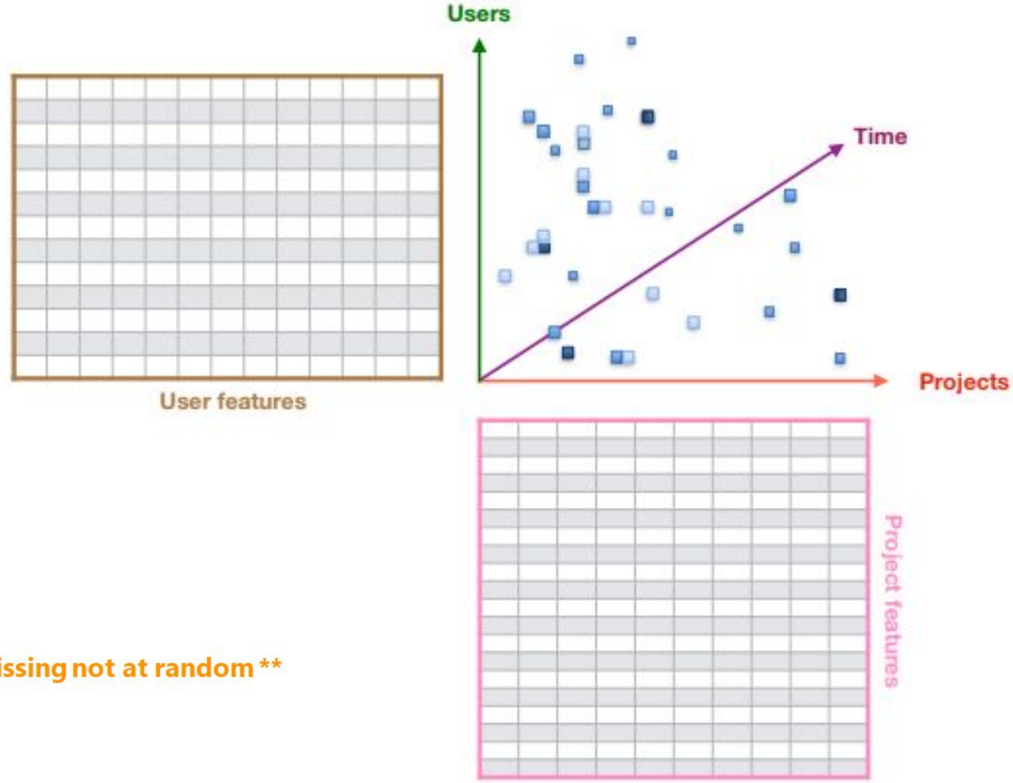
- Product and info
- User and info



Information for recommendation systems

There's many information available

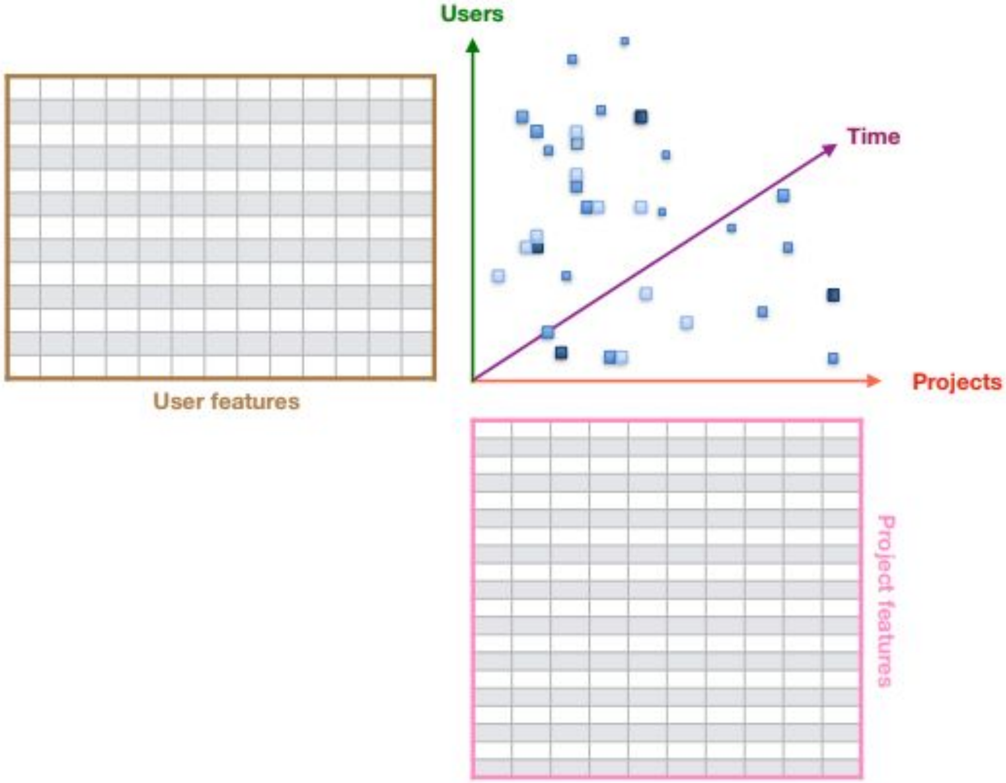
- Product and info
- User and info
- Interactions between product and user
  - Rating
  - Time
  - Missing interactions



** Missing not at random **

Information in the clicks

User interactions (views of the projects) can provide interesting insights

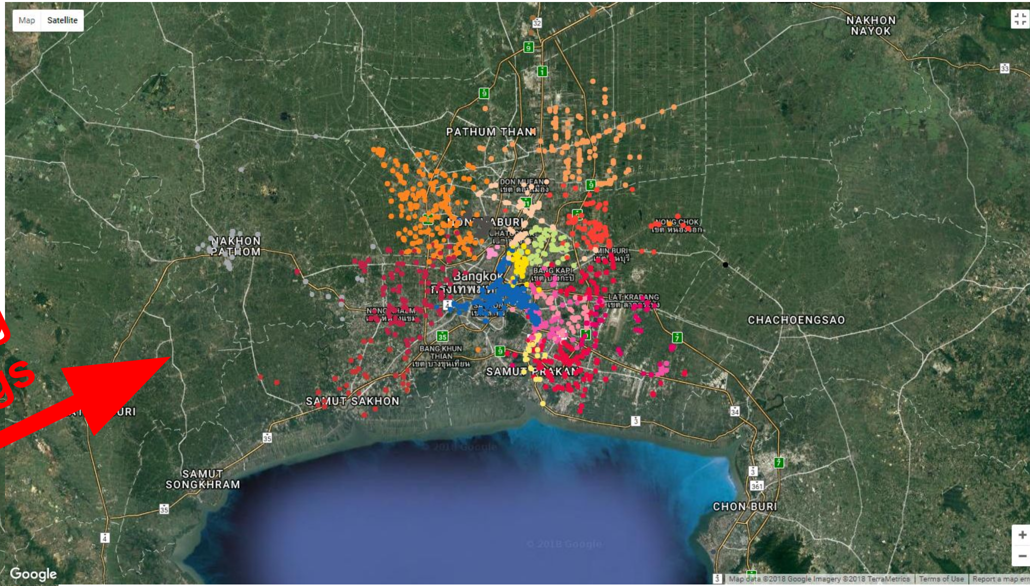


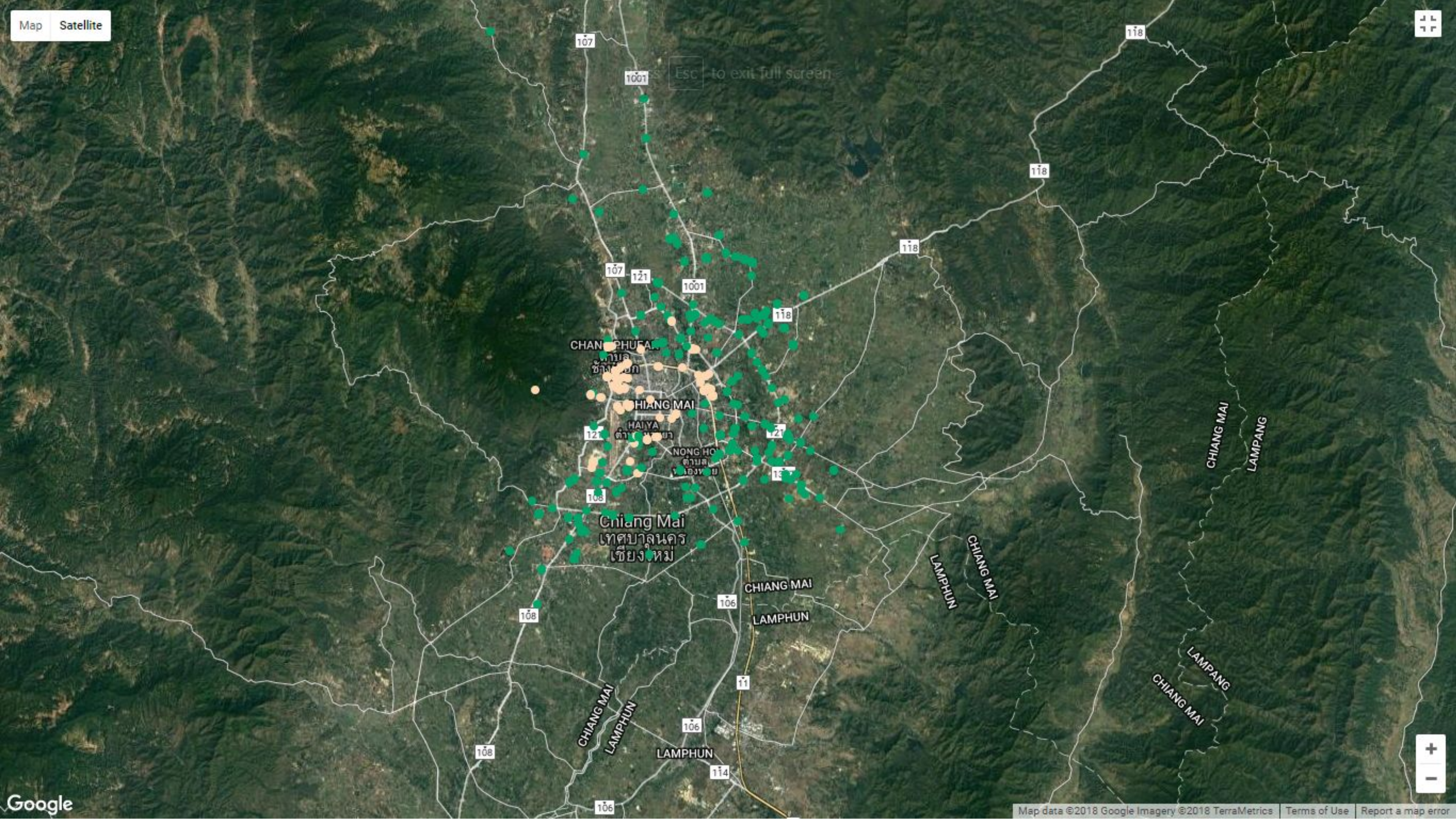
Product segmentation from user interactions

Run k-means clustering on view logs to cluster the real estates in Thailand.

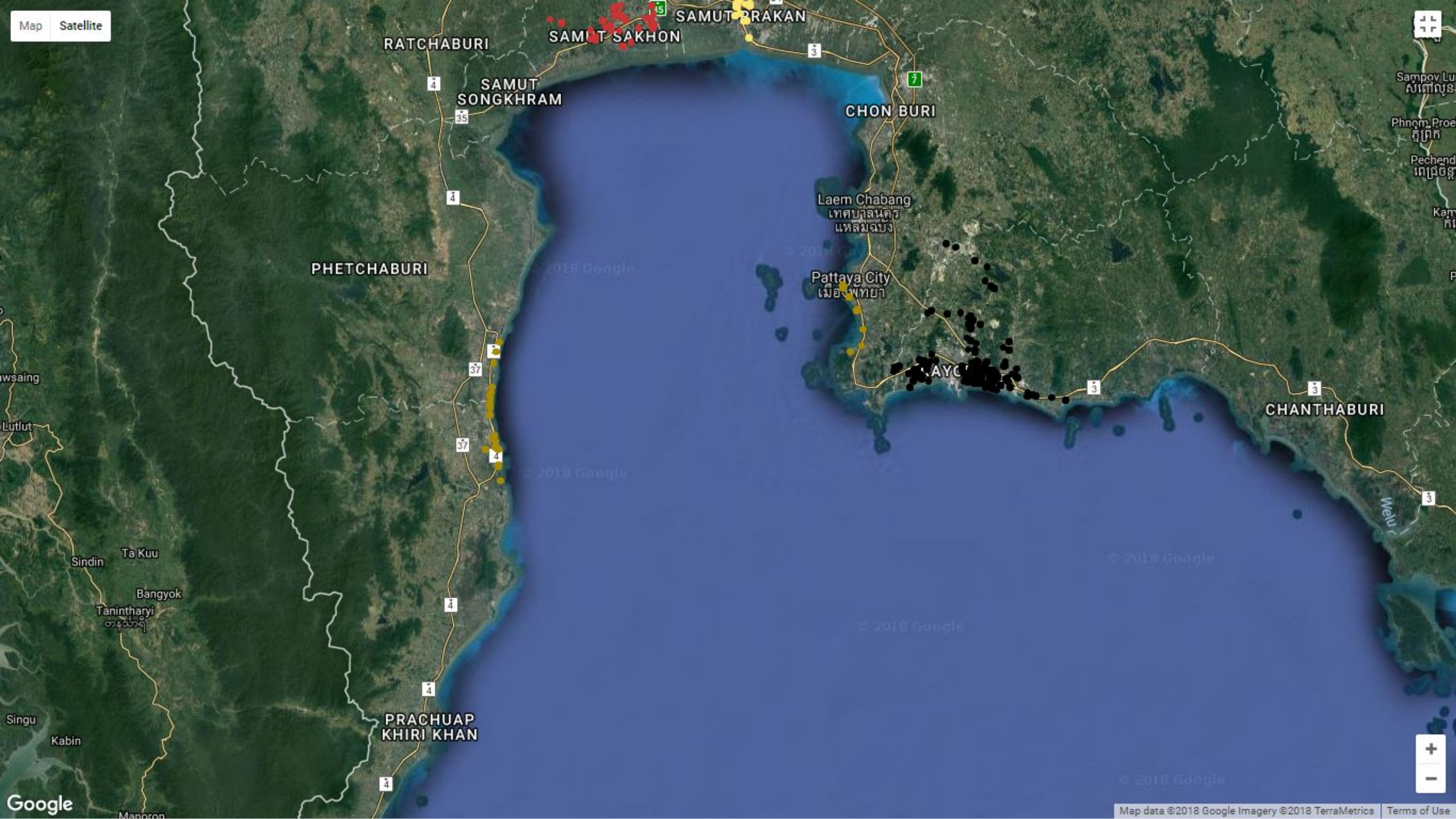
We can also cluster viewers.

Clustering using logs

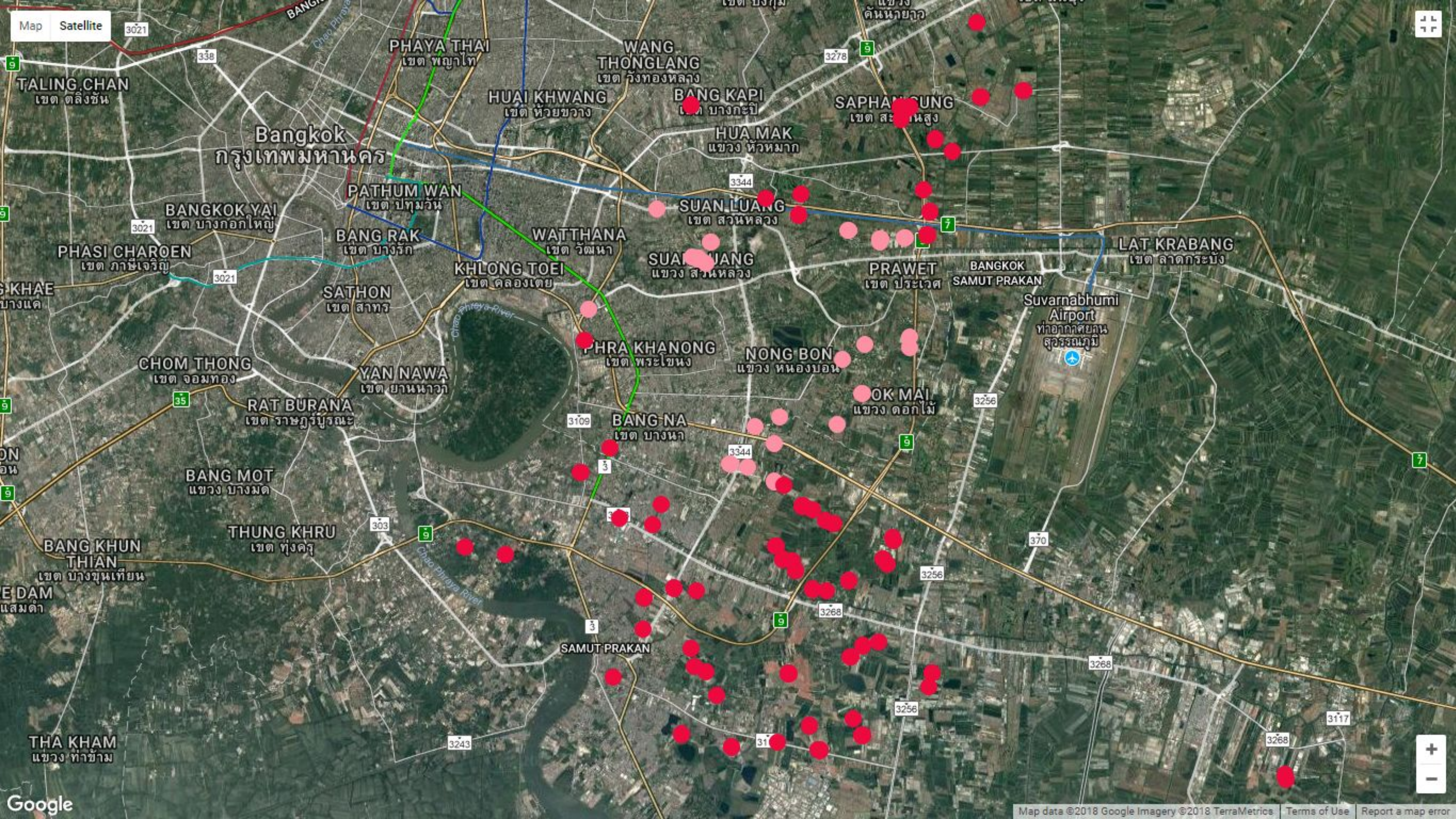




Map
Satellite
107
1001
118
121
108
106
11
114
CHIANG MAI
HAI YA
Chiang Mai
เทศบาลนคร
เชียงใหม่
CHIANG MAI
LAMPHUN
LAMPHUN
CHIANG MAI
LAMPANG
Google
Map data ©2018 Google Imagery ©2018 TerraMetrics
Terms of Use
Report a map error



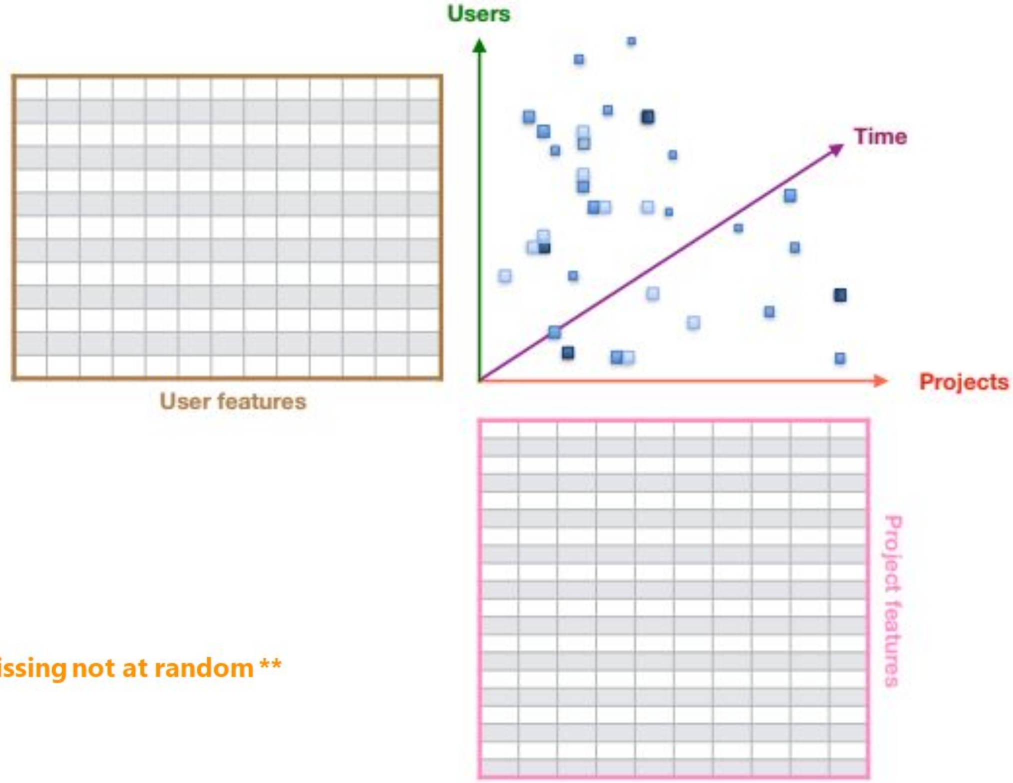
Map
Satellite
RATCHABURI
SAMUT SAKHON
SAMUT PRAKAN
SAMUT SONGKHRAM
CHON BURI
Laem Chabang
Pattaya City
PHETCHABURI
CHANTHABURI
PRACHUAP KHIRI KHAN
Lutlut
Sindin
Ta Kuu
Bangyok
Tanintharyi
Singu
Kabin
Google
Map data ©2018 Google Imagery ©2018 TerraMetrics
Terms of Use


Map
Satellite
PHAYA THAI
WANG THONGLANG
HUAI KHWANG
BANG KAPI
SAPHAN SUNG
HUA MAK
TALING CHAN
Bangkok
กรุงเทพมหานคร
PATHUM WAN
SUAN LUANG
BANGKOK YAI
BANG RAK
WATTHANA
PHASI CHAROEN
KHLONG TOEI
LAT KRABANG
PRAWET
BANGKOK SAMUT PRAKAN
SATHON
Suvarnabhumi Airport
PHRA KHANONG
NONG BON
CHOM THONG
YAN NAWA
RAT BURANA
BANG NA
BANG MOT
THUNG KHRU
BANG KHUN THIAN
SAMUT PRAKAN
THA KHAM
Chao Phraya River
Google
Map data ©2018 Google Imagery ©2018 TerraMetrics
Terms of Use
Report a map error

Context information

There's many information available

- Product and info
- User and info
- Interactions between product and user
  - Rating
  - Time
  - Missing interactions

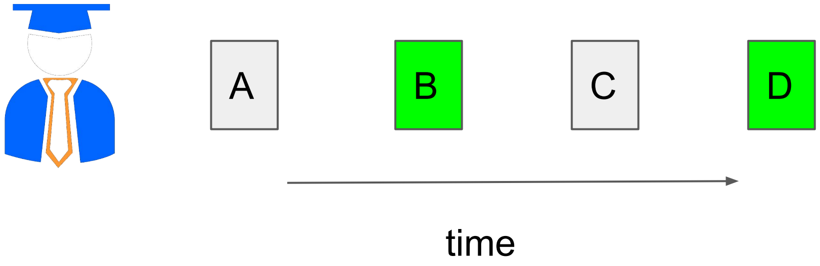


** Missing not at random **

Autoregressive recommendation model

Modeling time information (sequence)

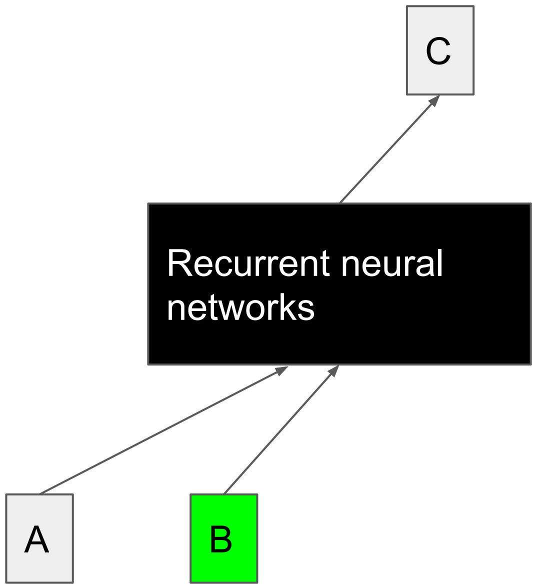
Recurrent Neural Networks



Autoregressive model

Modeling time information (sequence)

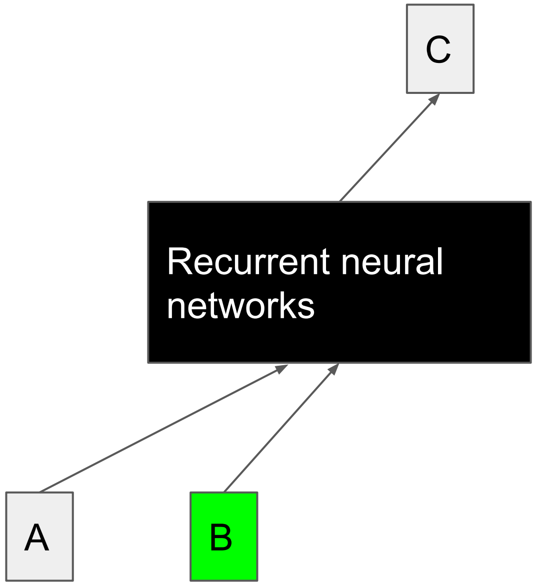
Recurrent Neural Networks



Autoregressive model

Modeling time information (sequence)

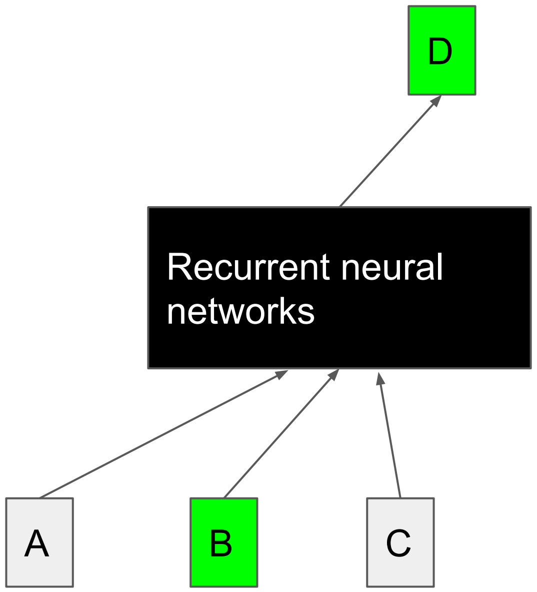
Recurrent Neural Networks



Autoregressive model

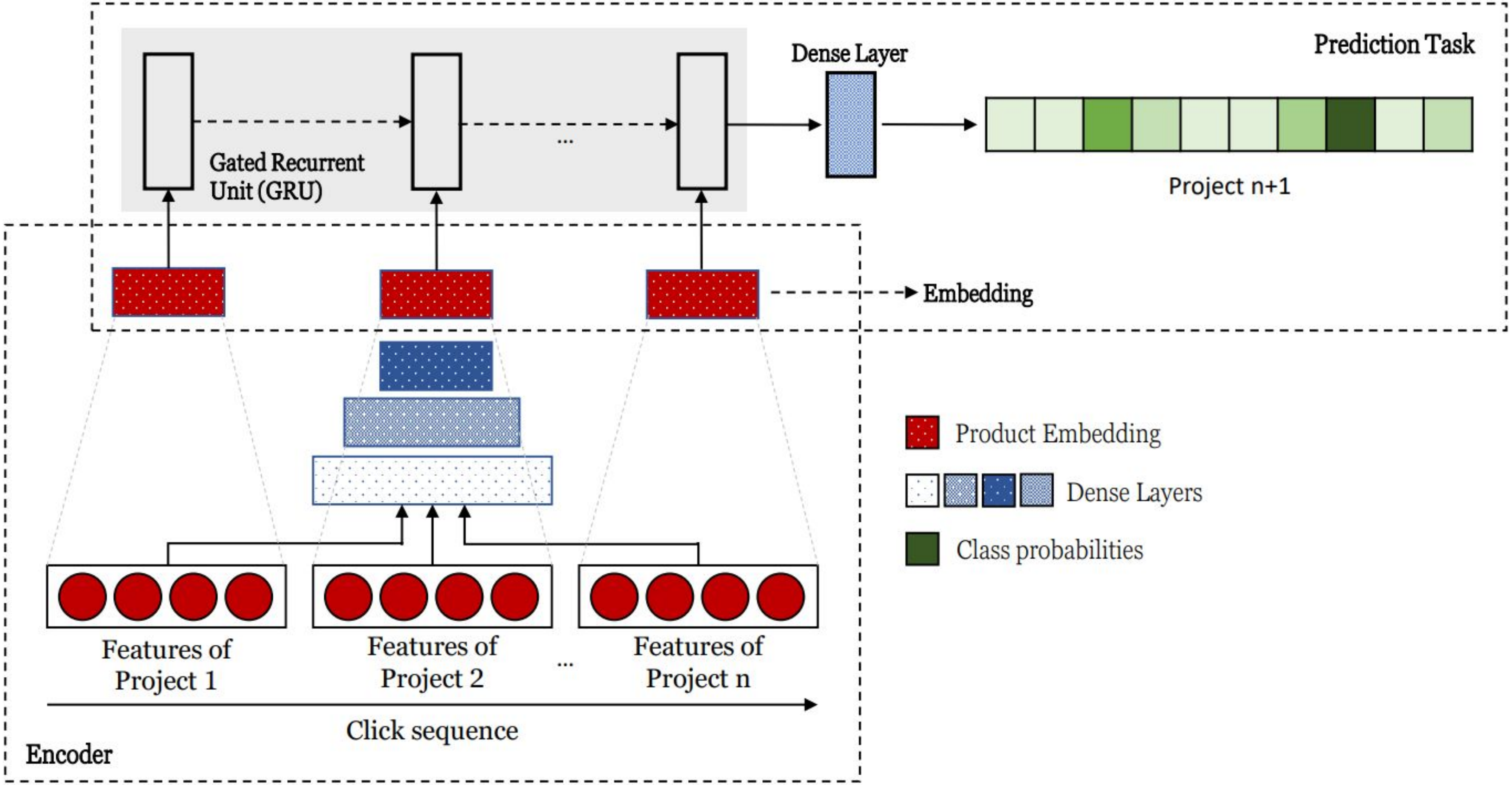
Modeling time information (sequence)

Recurrent Neural Networks



P Covington, Deep Neural Networks for YouTube Recommendations. 2016

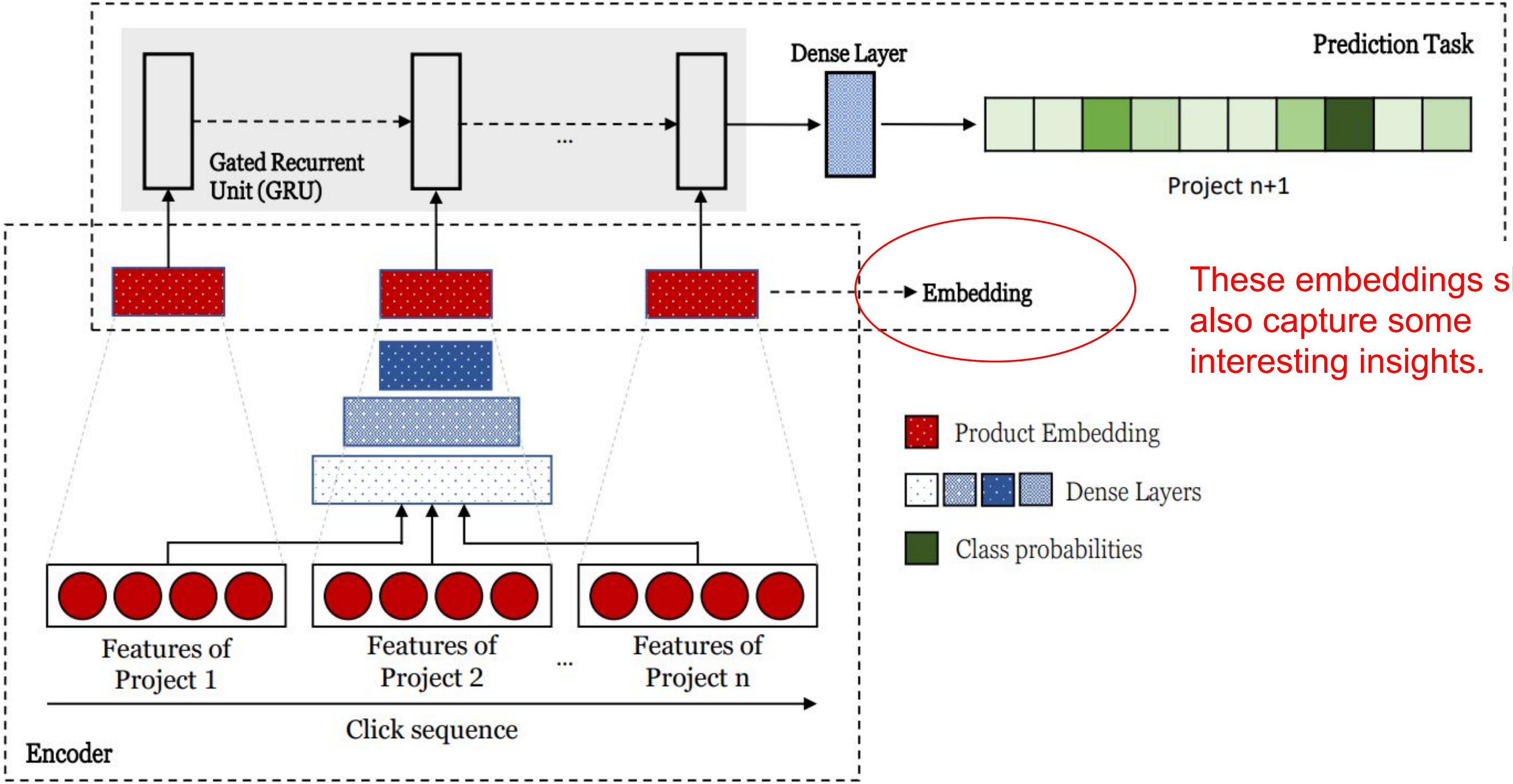
A Beutel, Latent Cross: Making Use of Context in Recurrent Recommender Systems, 2018



Price: 2,900,000
Type: Home
District: Nonthaburi
Facility: [Security, Park]

Price: 3,900,000
Type: Home
District: Bangkok
Facility: [Fitness, Security, Park]

Price: 5,900,000
Type: Home
District: Bangkok
Facility: [Fitness, Security, Park, Pool]



These embeddings should also capture some interesting insights.

Price: 2,900,000
Type: Home
District: Nonthaburi
Facility: [Security, Park]

Price: 3,900,000
Type: Home
District: Bangkok
Facility: [Fitness, Security, Park]

Price: 5,900,000
Type: Home
District: Bangkok
Facility: [Fitness, Security, Park, Pool]

HOMEHOP

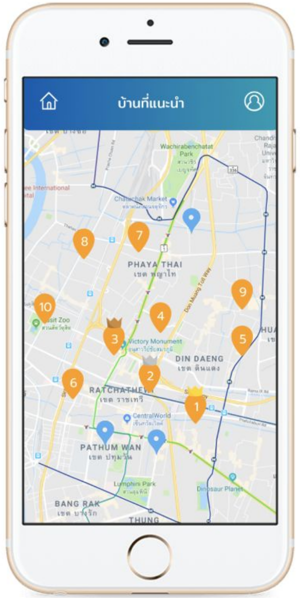
Home recommender app based on user’s lifestyle and commute.

Persona

ไลฟ์สไตล์ (persona) ในการเลือกซื้อบ้าน ซึ่งแบ่งประเภท โดย AI จากข้อมูลผู้ใช้กว่า 10 ล้านคน

Daily life travel

วิธีและเวลาปกติที่คุณเดินทางจากบ้านไปยังที่ทำงาน หรือสถานที่ต่างๆ ในชีวิตประจำวันของคุณ



Affordable Price

ช่วงราคาบ้านที่คุณต้องการ หรือสามารถจ่ายได้

Traffic data from iTIC

ข้อมูลการจราจร จากมูลนิธิศูนย์ข้อมูลจราจรอัจฉริยะไทย เพื่อแนะนำโครงการที่จะใช้เวลาเดินทางน้อยที่สุด

1. เลือกแผนการเดินทาง

ระบุสถานที่ต่างๆ ที่คุณมักจะเดินทางไปในแต่ละวัน เช่น บ้าน โรงเรียนของลูก สถานที่ทำงาน ห้างสรรพสินค้าที่มักเดินทางไปบ่อยๆ เป็นต้น พร้อมทั้งระบุเวลาตั้งแต่ออกจากบ้าน จนถึงเวลากลับถึงบ้าน

2. เลือกวิธีการเดินทาง

เลือกวิธีการเดินทาง เช่น เดินทางโดยรถยนต์ส่วนตัว รถประจำทาง เรือ รถไฟฟ้า โดยระบบจะคำนวณเวลาการเดินทางจากข้อมูลจราจร ของมูลนิธิศูนย์ข้อมูลจราจรอัจฉริยะไทย

3. เลือกช่วงราคาบ้านที่คุณต้องการ

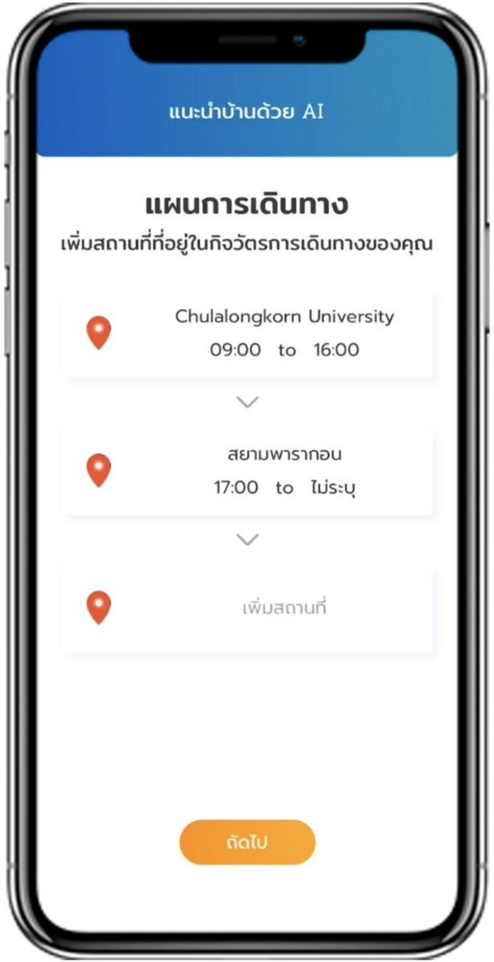
เลือกช่วงราคาบ้านที่คุณต้องการจะซื้อ

4. เลือกเพอร์โซนา

ระบุเพอร์โซนา (persona) หรือ ไลฟ์สไตล์ของคุณ เช่น เน้นประโยชน์ใช้สอย หรือเน้นความหรูหรา

5. ประมวลผล

ยืนยันข้อมูล แล้วสนุกไปกับการเลือกบ้านที่โปรแกรมแนะนำ ได้ทันที!



Data science for Real Estate

Consumer

~~Matching~~

Social listening

(Real Estate) Developers

Lead generation and smart marketing

Social listening

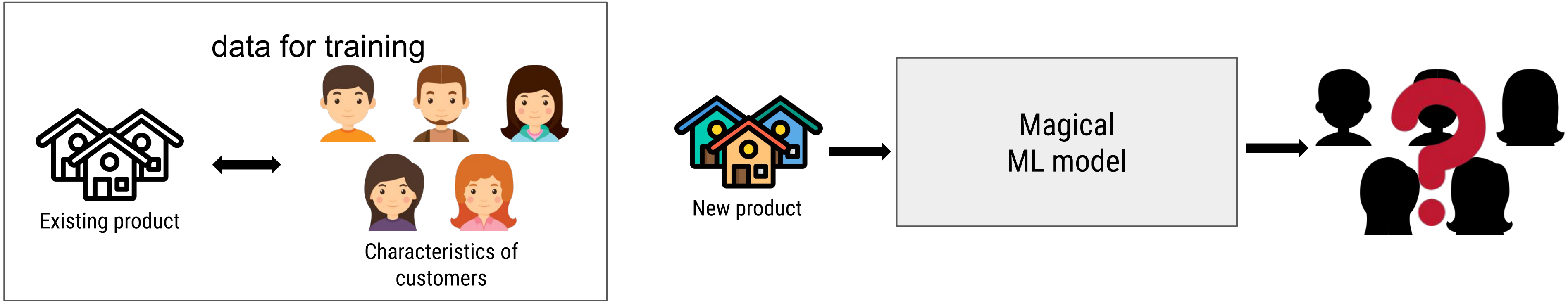
Project development
- Customer segmentation
- Trend prediction
- Pricing

ML for product development

For real estates, no two products are the same. Development based on gut feeling.

Make some informative guess about a new product

- popularity
- the type of potential buyers
- whether to add or remove some features
- the best marketing channel



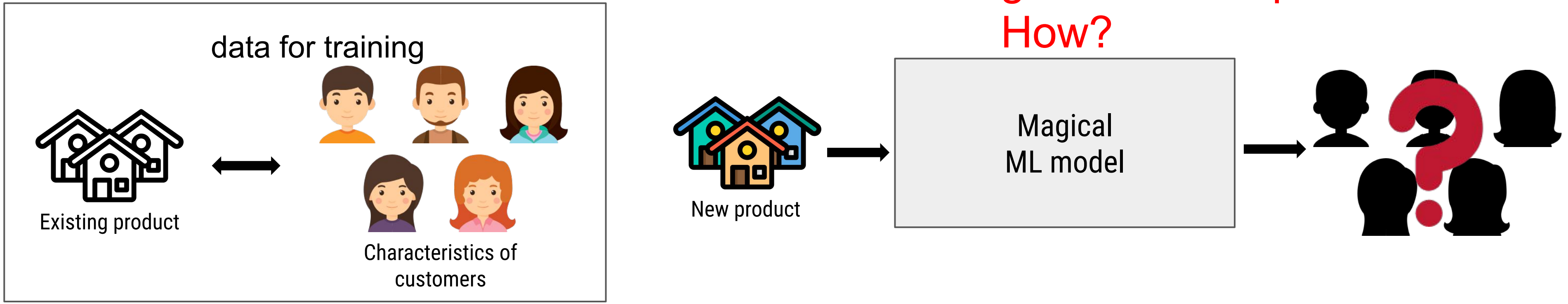
ML for product development

For real estates, no two products are the same. Development based on gut feeling.

Make some informative guess about a new product

- popularity
- the type of potential buyers
- whether to add or remove some features
- the best marketing channel

We want to learn the distribution of the user given some input. How?

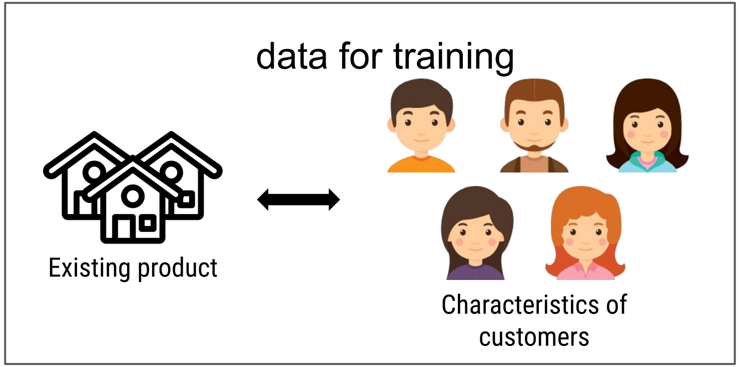


ML for product development

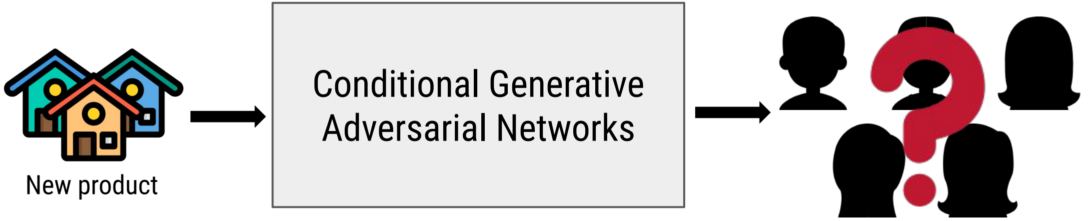
For real estates, no two products are the same. Development based on gut feeling.

Make some informative guess about a new product

- popularity
- the type of potential buyers
- whether to add or remove some features
- the best marketing channel



GAN!



Generative Adversarial Networks (GANs)

0.1, -0.3, .. → Generator →  → Discriminator → Real or Fake

z   x $Y=f(x)$

Consider a money counterfeiter

- He wants to make fake money that looks real
- There's a police that tries to differentiate fake and real money.

The counterfeiter is the adversary and is generating fake inputs. – Generator network

The police is try to discriminate between fake and real inputs. – Discriminator network

Generative Adversarial Networks (GANs)

0.1, -0.3, .. → Generator →  → Discriminator → Real

Z

X

Generative Adversarial Networks (GANs)

0.1, -0.3, .. → Generator → 

→ Discriminator → Fake

Z

X

Generative Adversarial Networks (GANs)

real x

Discriminator

Real

gradient

fake x

The discriminator learns to differentiate real and fake data

Learns by backpropagation

Generative Adversarial Networks (GANs)

0.1, -0.3, .. → Generator →  → Discriminator → Fake

z

x

gradient

The generator learns to be better by the gradient given by the discriminator

Conditional GAN (CGAN)

0.1, -0.3, ..

z

1.0

c

Generator

x

1.0

c

Discriminator

Real or Fake

Y=f(x,c)

GAN can be conditioned (controlled) to generate things you want by concatenating additional information

Conditional GAN (CGAN)

0.1, -0.3, ..

Z

-1.0

C

Generator

X

-1.0

C

Discriminator

Real or Fake

Y=f(x,c)

GAN can be conditioned (controlled) to generate things you want by concatenating additional information

Conditional GAN (CGAN)

0.1, -0.3, ..

z

-1.0

c

Generator

x

1.0

c

Discriminator

Real or Fake

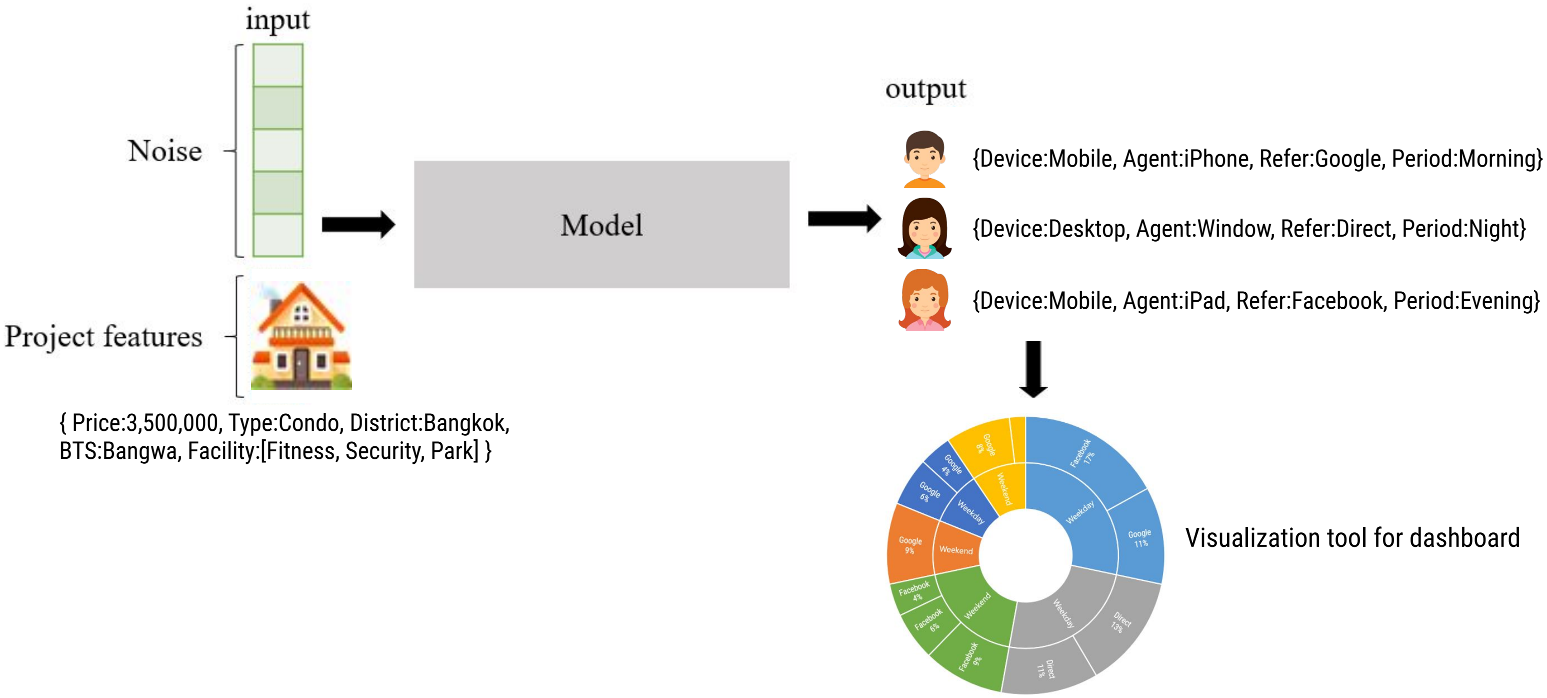
$Y=f(x,c)$

GAN can be conditioned (controlled) to generate things you want by concatenating additional information

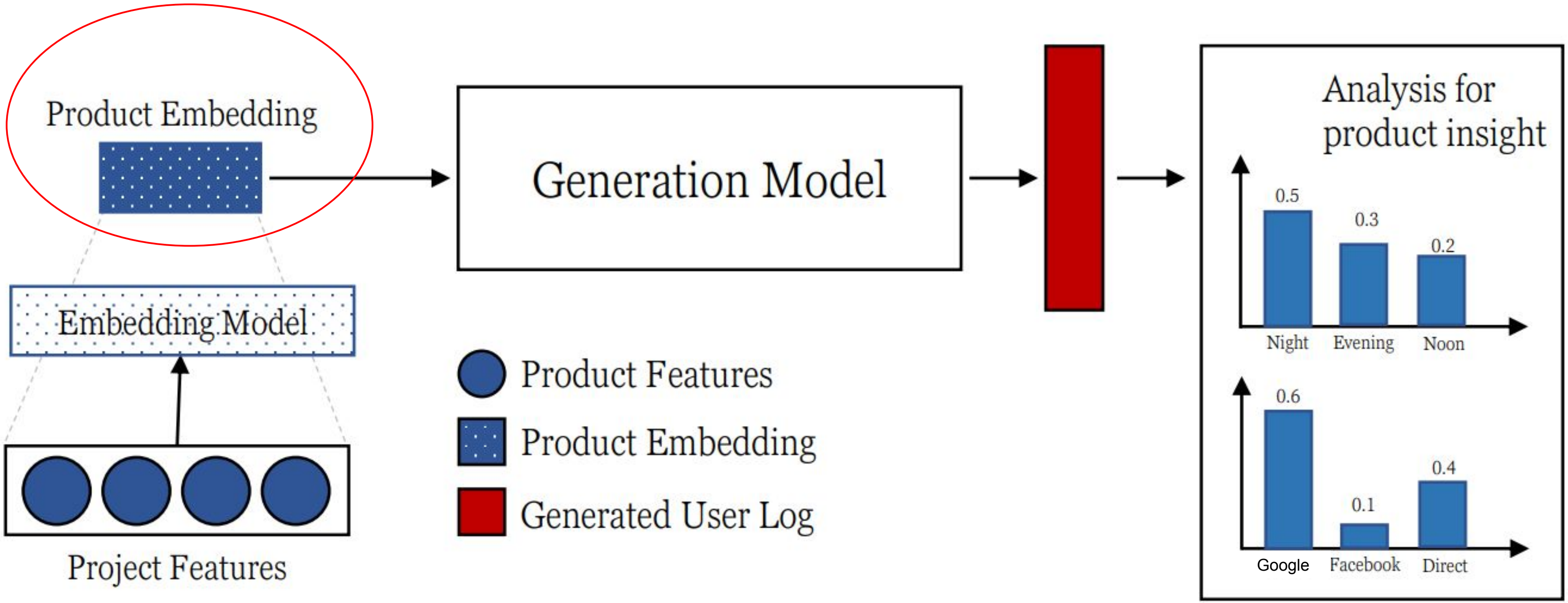
Example of CGAN applications

Globally and Locally Consistent Image Completion [Iizuka et al., 2017]
StackGAN: Text to Photo-realistic Image Synthesis with Stacked GANs [Zhang et al. 2017]

Overview of our system



Embedding learned from our recommender system



Why GAN?

vs supervised learning

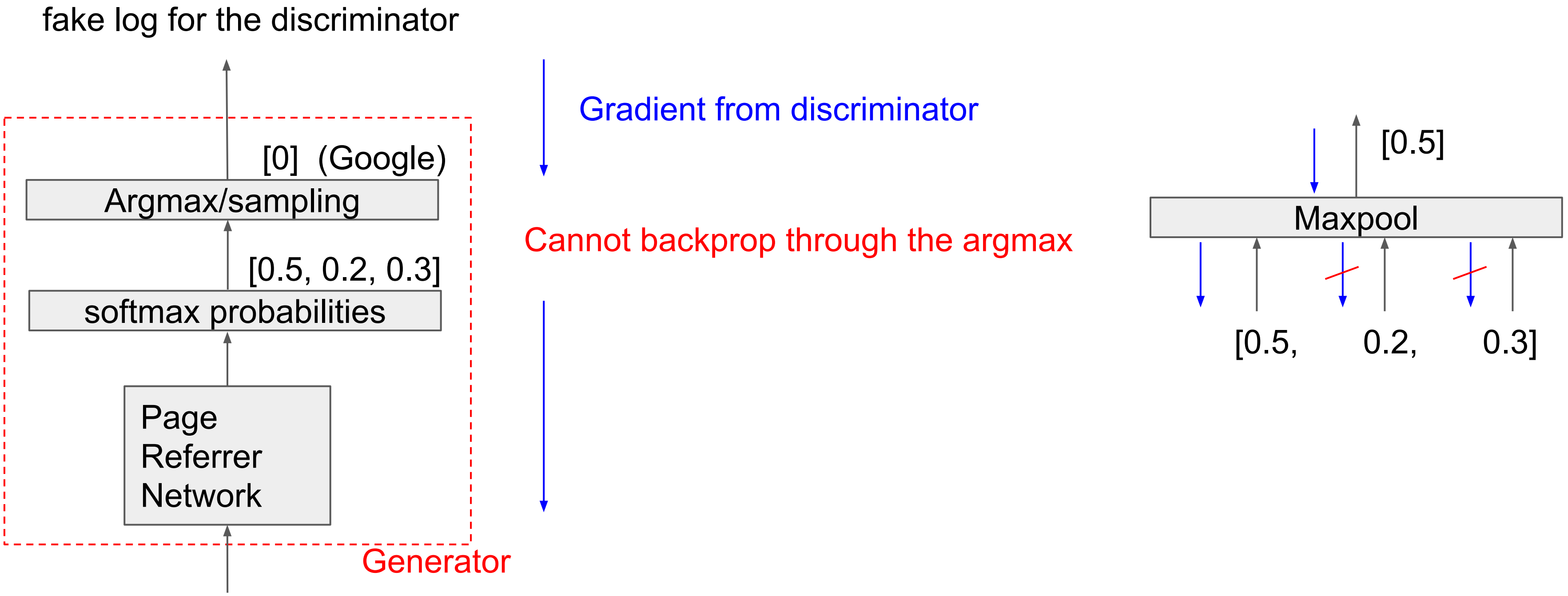
- supervised learning yields one correct answer (not learning the distribution)
- cannot be used to generate examples

vs other distribution learning methods

- non-parametric
- better than other methods for multi-modal distributions
- generate things that differ from the training data but still “realistic”

GAN for discrete output

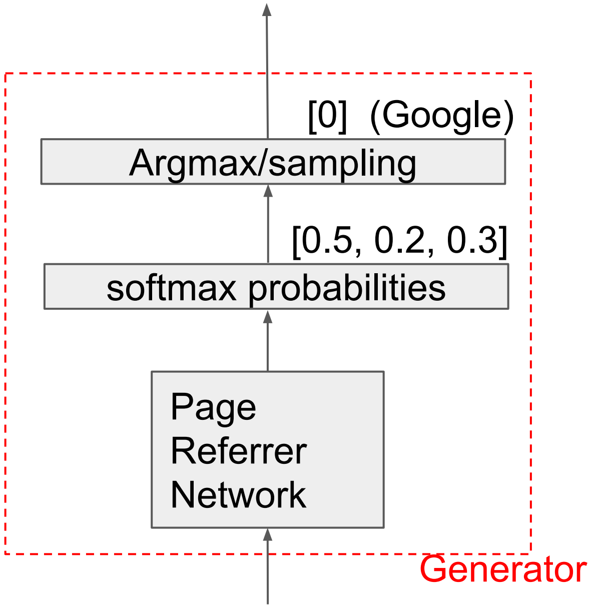
Unlike images, generating discrete output includes a sampling process



GAN for discrete output

Unlike images, generating discrete output includes a sampling process

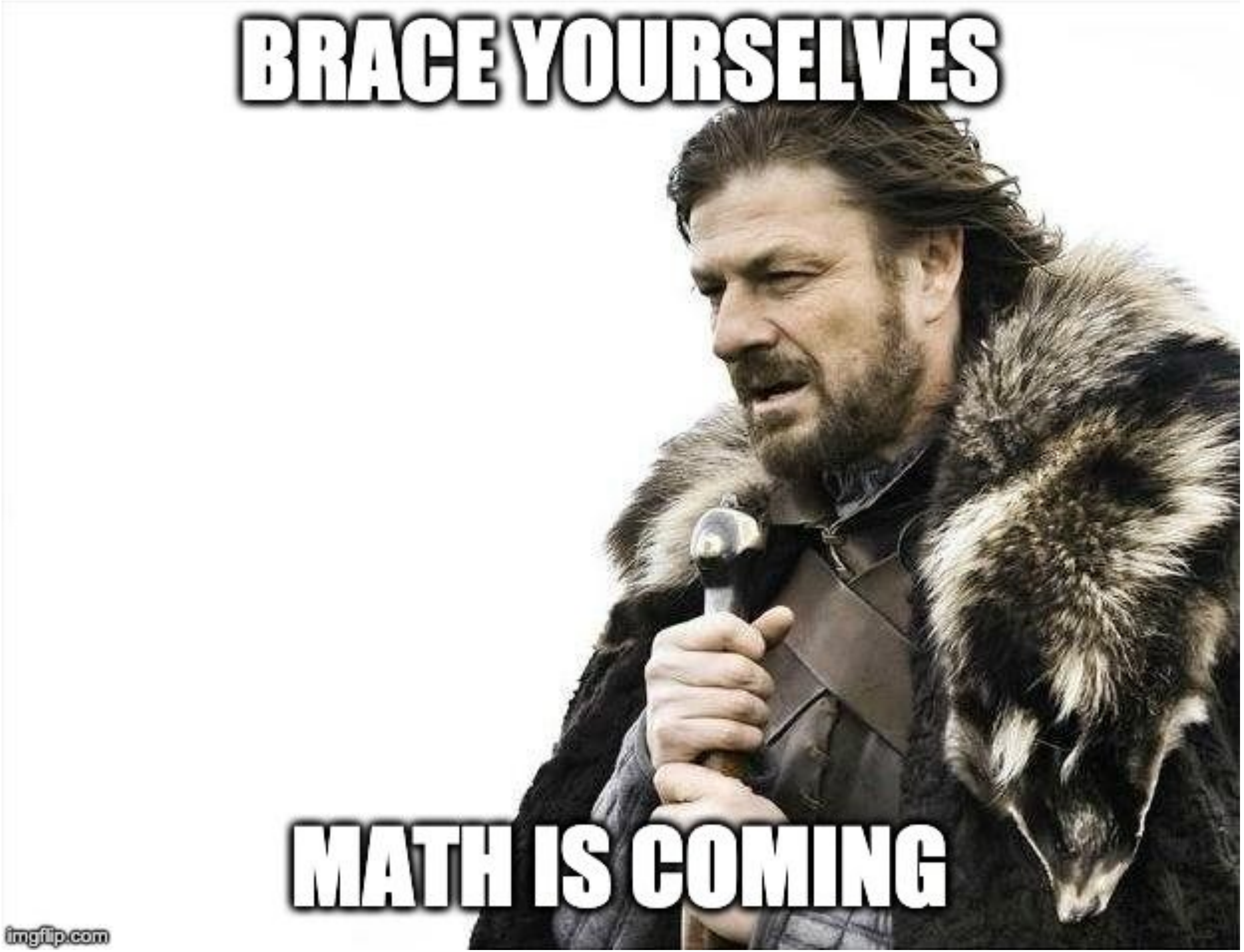
fake log for the discriminator

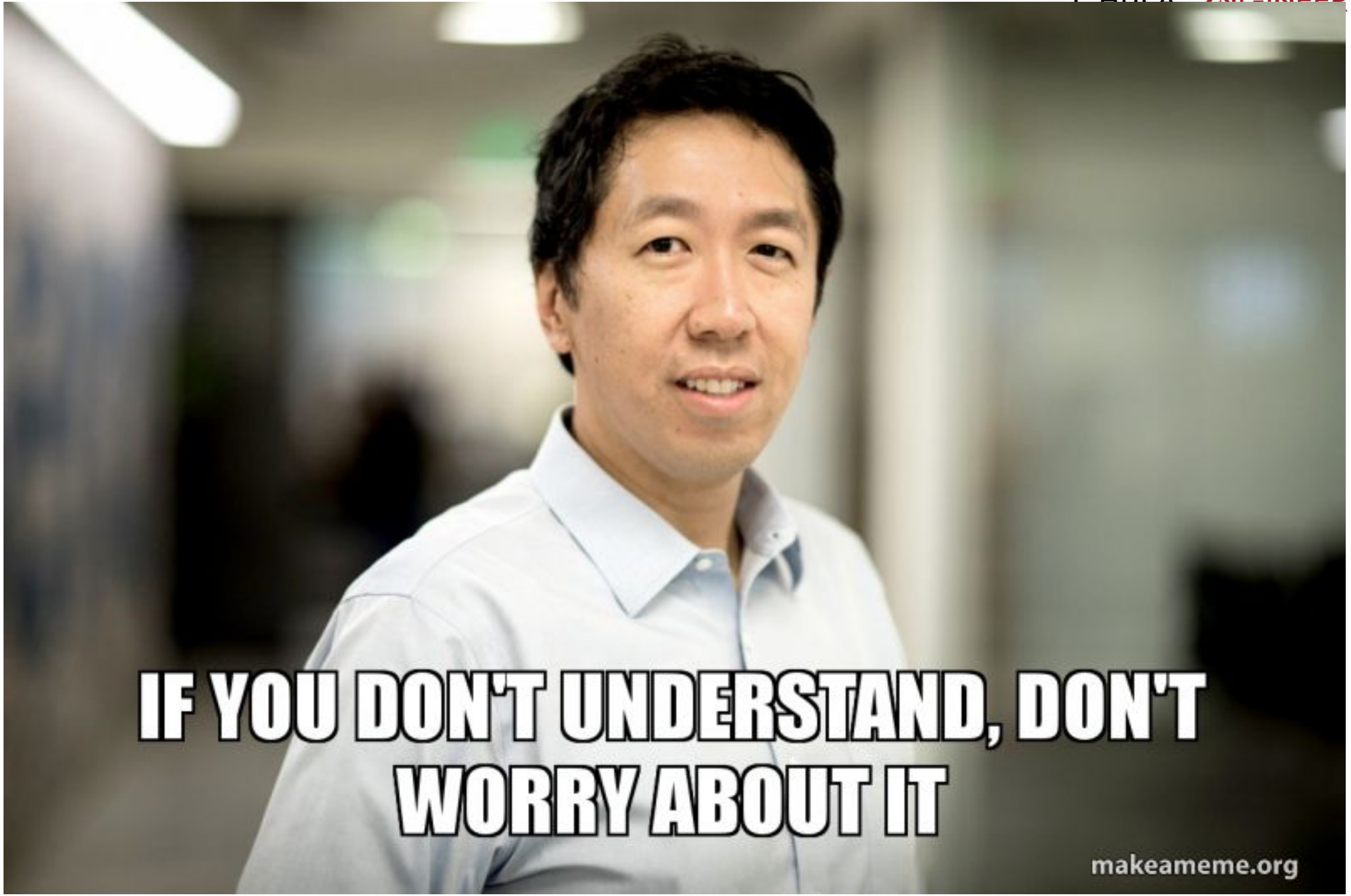


Gradient from discriminator

Cannot backprop through the argmax

Two popular methods: REINFORCE, Gumbel-Softmax approximation (https://arxiv.org/abs/1611.01144)


BRACE YOURSELVES
MATH IS COMING
imgflip.com


IF YOU DON'T UNDERSTAND, DON'T WORRY ABOUT IT
makeameme.org

Gumbel Softmax

Sampling from a softmax can be done via the Gumbel-max trick

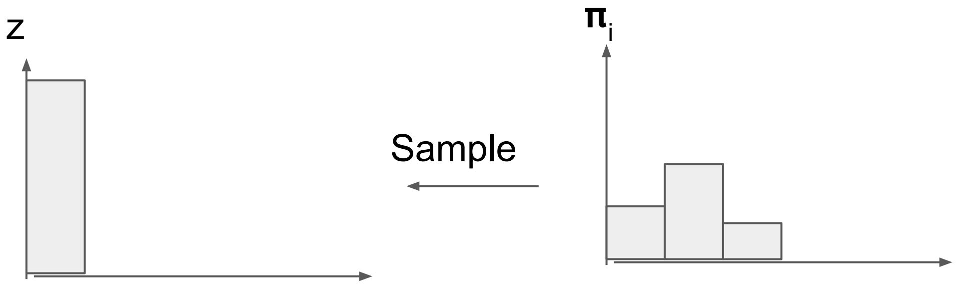
random value generated from Gumbel dist.

$$z = \mathrm{one\_hot}\left(\arg\max_i \left[g_i + \log \pi_i\right]\right)$$

Final output
Ex. [1, 0, 0]

index for discrete output

prob values from softmax
Ex. [0.5, 0.2, 0.3]



https://arxiv.org/pdf/1611.01144.pdf

Gumbel Softmax

Approximate the argmax term with **y** (continuous)

random value generated from Gumbel dist.

$$z = \text{one\_hot}\left(\boxed{\arg\max_i \left[g_i + \log \pi_i\right]}\right)$$

prob values from softmax

index for discrete output

$$\boxed{y_i} = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)} \qquad \text{for } i = 1, \ldots, k.$$

Gumbel Softmax

Approximate the argmax term with **y** (continuous)

random value generated from Gumbel dist.

$$z = \texttt{one\_hot}\left(\boxed{\arg\max_i \left[g_i + \log \pi_i\right]}\right)$$

prob values from softmax

index for discrete output

Temperature parameter

$$\boxed{y_i} = \boxed{\frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)}} \quad \text{for } i = 1, \ldots, k.$$

This rescales the distribution

Gumbel Softmax

y at small T is similar to an argmax but can be backpropagated through

Temperature parameter

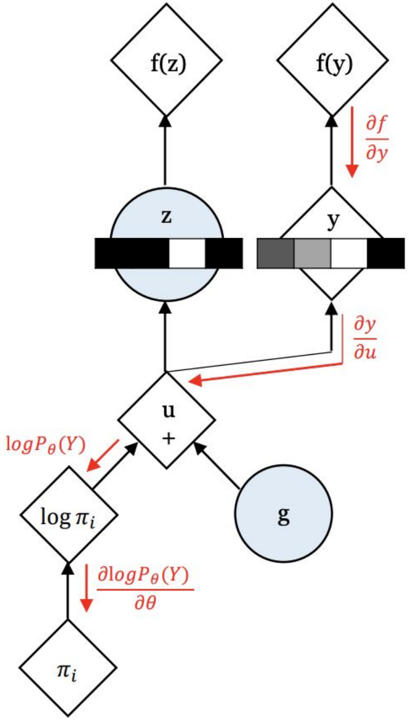
$$\boxed{y_i} = \boxed{\frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)}} \qquad \text{for } i = 1, \ldots, k.$$

This rescales the distribution

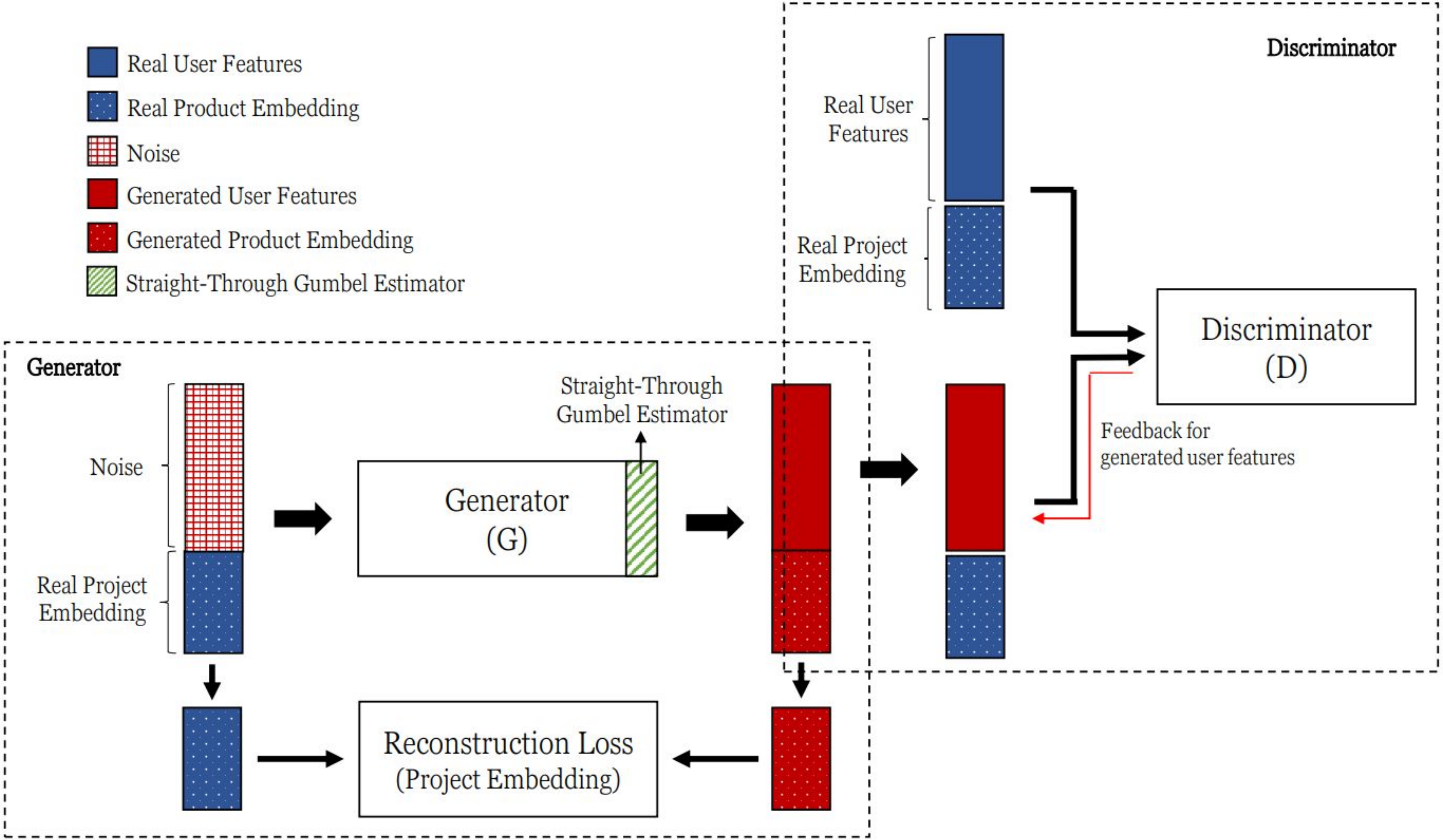
Straight-through Gumbel estimator

Forward to the discrim.

Backward from the discrim.



The generator generates both the argmax and the Gumbel version. The discriminator uses the argmax version as input. However, the gradient is passed through the Gumbel version.

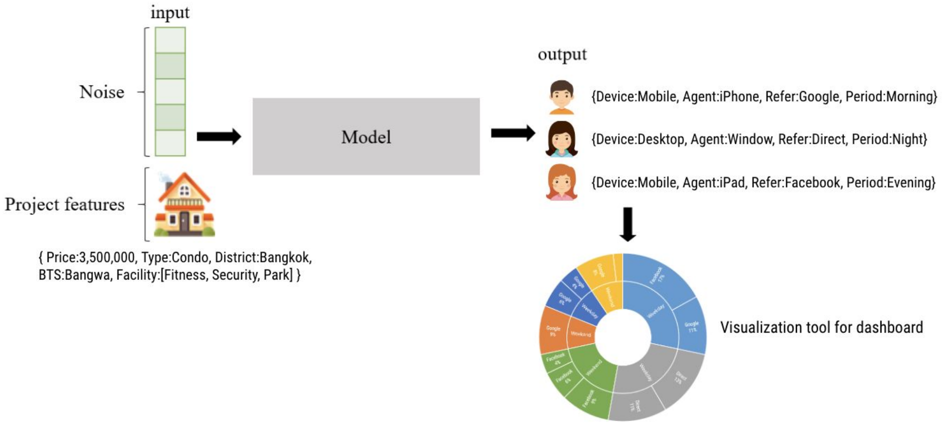

Real User Features
Real Product Embedding
Noise
Generated User Features
Generated Product Embedding
Straight-Through Gumbel Estimator
Discriminator
Real User Features
Real Project Embedding
Discriminator (D)
Feedback for generated user features
Generator
Noise
Real Project Embedding
Generator (G)
Straight-Through Gumbel Estimator
Reconstruction Loss (Project Embedding)

Experimental setup

~5000 projects, ~2 million log entries

- Held out 50 random projects as novel projects to generate
- Measure the distribution of generated logs vs real data

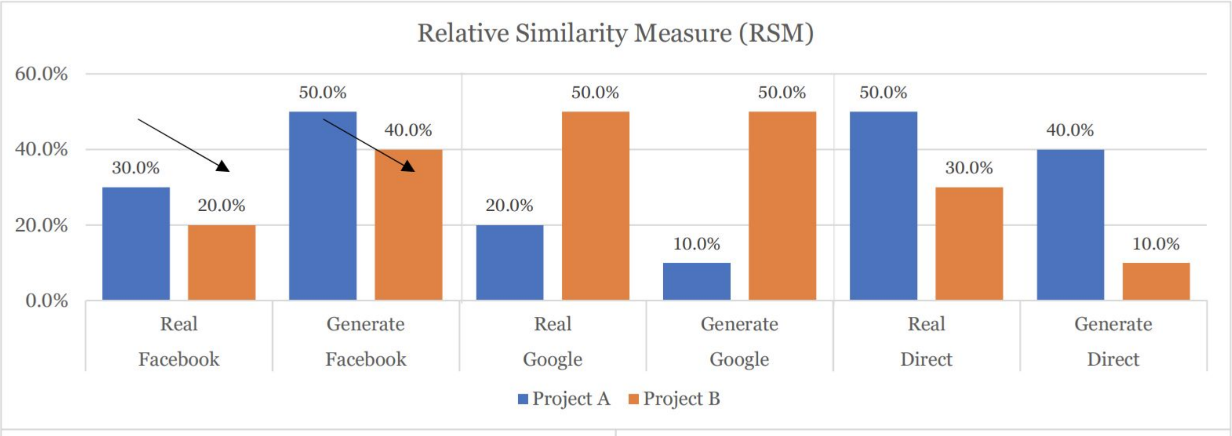
Average the performance over 10 runs



Metrics

RSM

- Relative measure
- Across project pairs



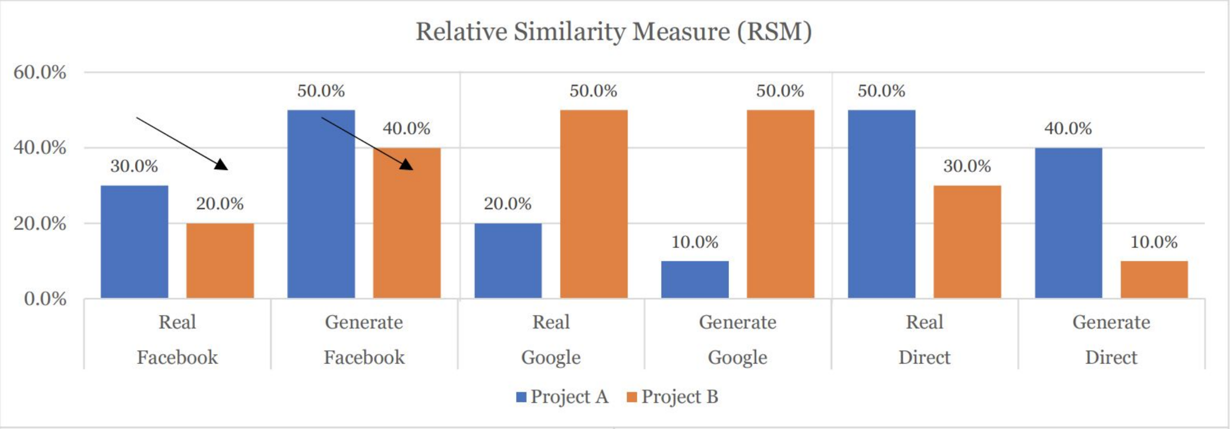
Metrics

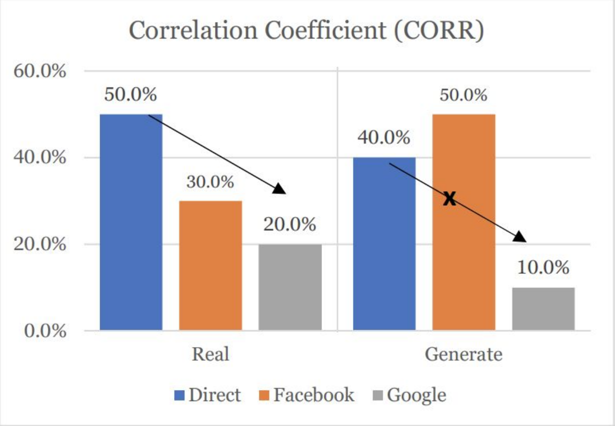
RSM

- Relative measure
- Across project pairs

Correlation

- Relative measure
- Within a project





Metrics

RSM

Relative measure

Across project pairs

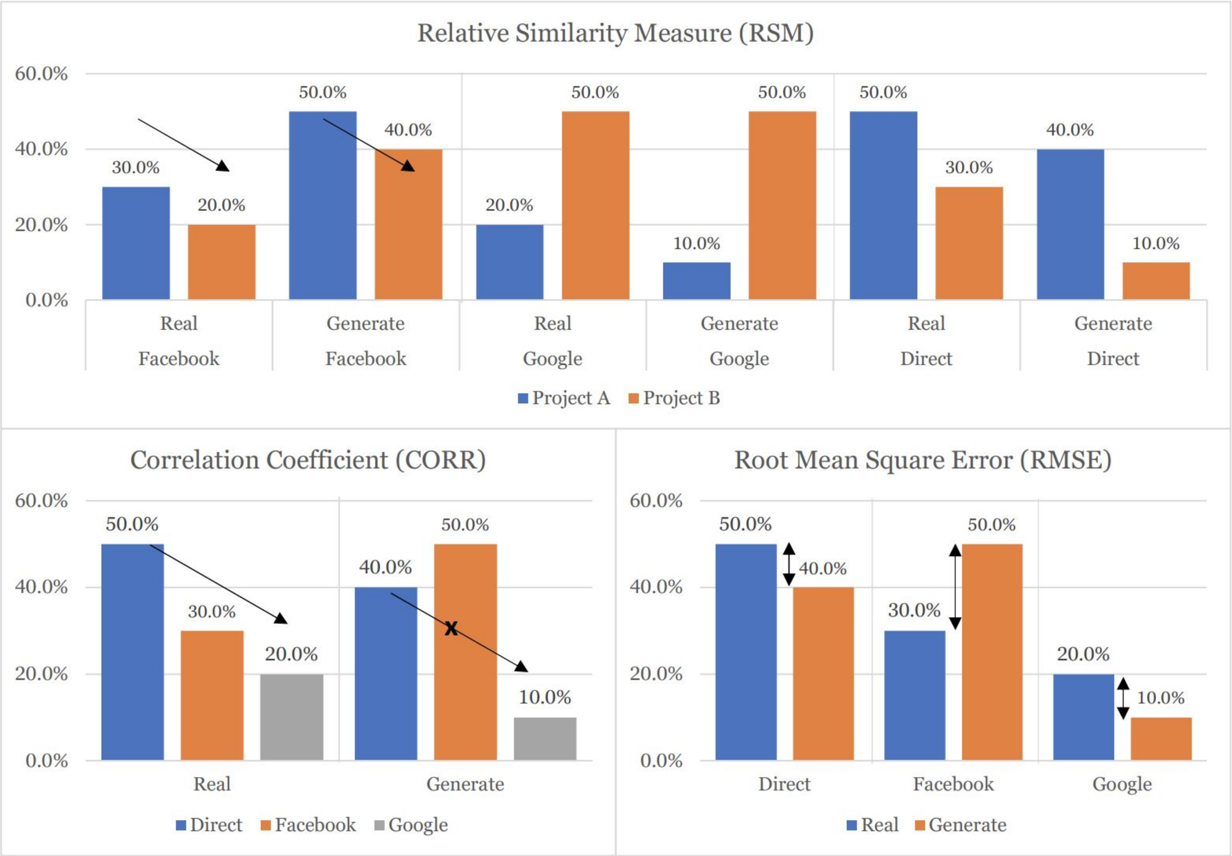
Correlation

Relative measure

Within a project

RMSE

Absolute measure



Results

Model	RSM	CORR	RMSE
NN with Rec. Emb	54.7%	71.6%	28.0%

Use the most similar project in the training data based on recommendation embeddings

Results

Model	RSM	CORR	RMSE
GAN with Rec. Emb	72.5%	88.9%	16.2%
NN with Rec. Emb	54.7%	71.6%	28.0%

Our model with recommender embedding

Use the most similar project in the training data based on recommendation embeddings

Results

Model	RSM	CORR	RMSE
GAN with Rec. Emb	72.5%	88.9%	16.2%
GAN with AutoEncoder Emb	69.7%	87.8%	18.1%
GAN with product features	67.9%	86.6%	18.2%
NN with Rec. Emb	54.7%	71.6%	28.0%

Our model with recommender embedding

Our model with embeddings learned from Autoencoder
Our model with product features instead of embedding

No knowledge about relationships between different products

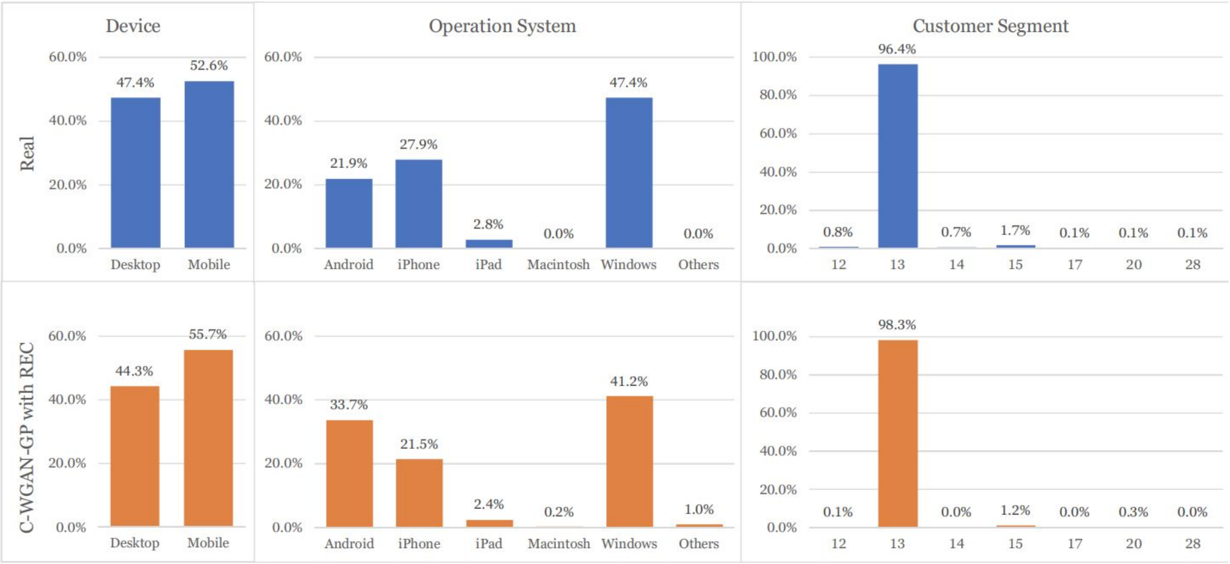
Results

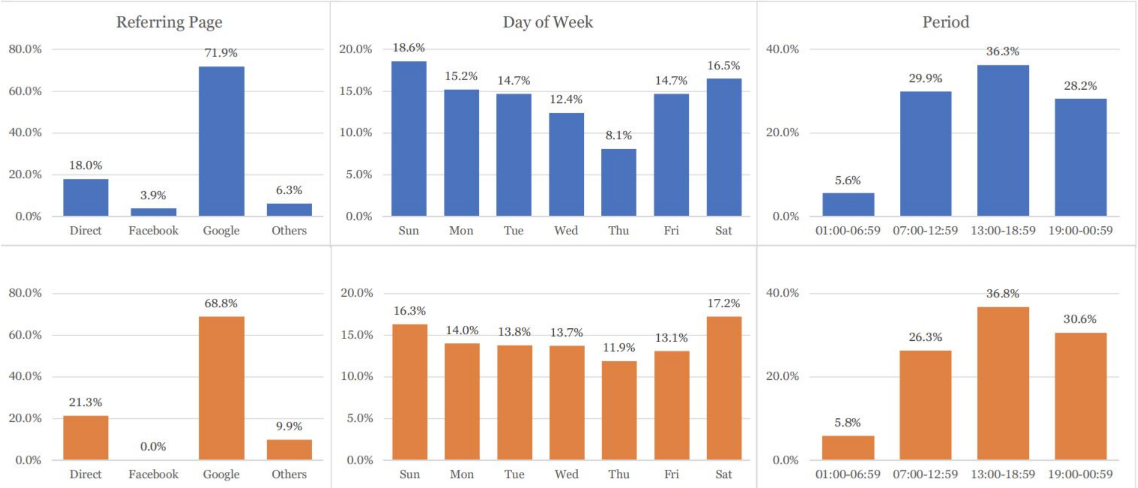
Model	RSM	CORR	RMSE
GAN with Rec. Emb	72.5%	88.9%	16.2%
GAN with AutoEncoder Emb	69.7%	87.8%	18.1%
GAN with product features	67.9%	86.6%	18.2%
VAE with Rec. Emb	65.3%	85.6%	20.3%
NN with Rec. Emb	54.7%	71.6%	28.0%

Our model with recommender embedding

Instead of GAN use VAE

CHULA ENGINEERING
Foundation toward Innovation
HOME DOT TECH





Data science for Real Estate

Consumer

Matching

- Autoregressive Recommender system

(Real Estate) Developers

Project development

- GAN-based distribution learning

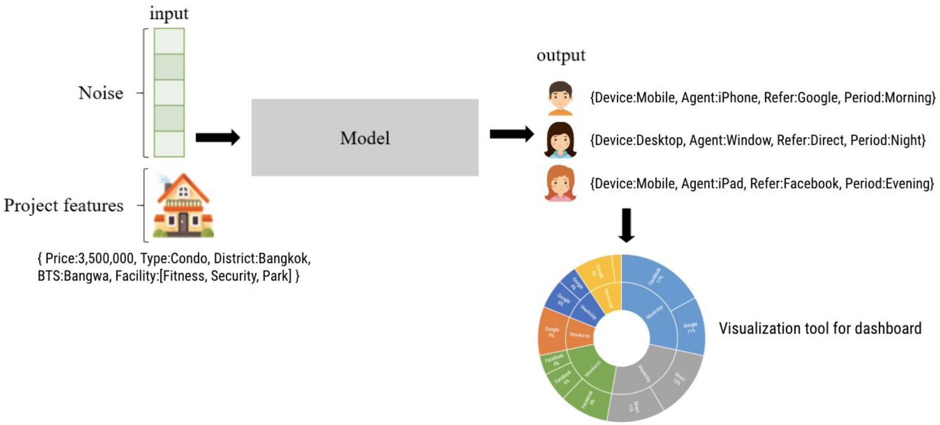
Team

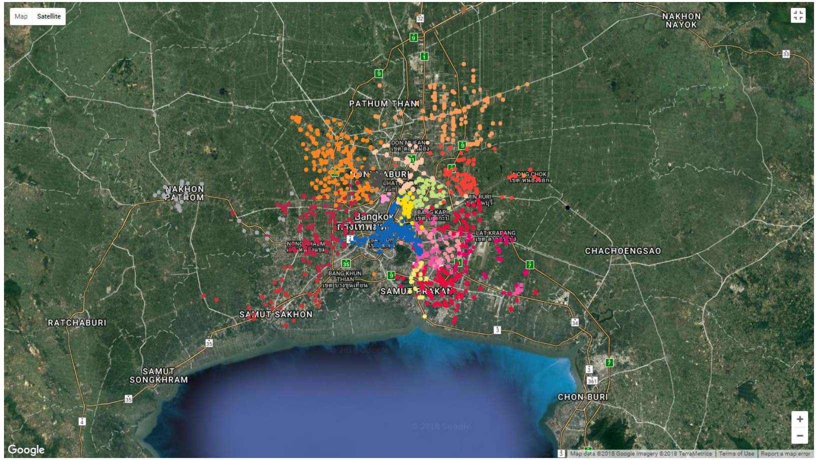


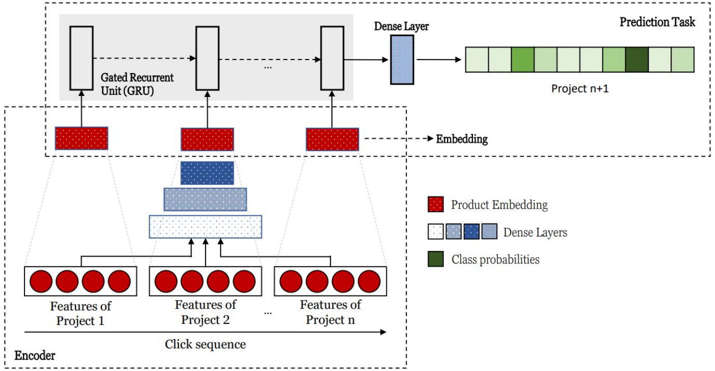


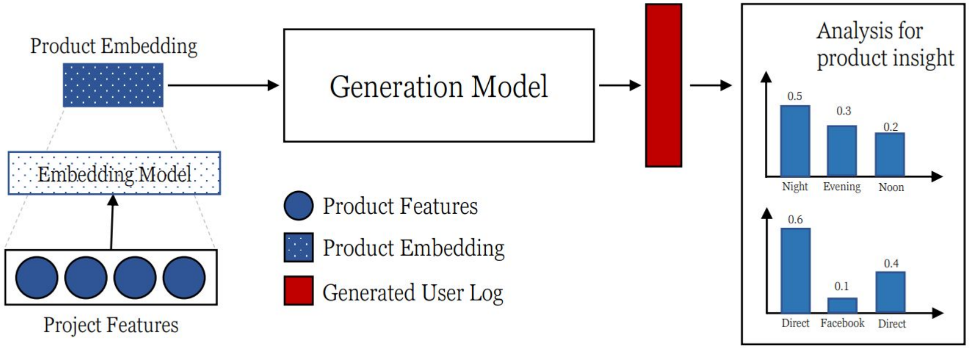


Questions?


input
Noise
Project features
{ Price:3,500,000, Type:Condo, District:Bangkok, BTS:Bangwa, Facility:[Fitness, Security, Park] }
Model
output
{Device:Mobile, Agent:iPhone, Refer:Google, Period:Morning}
{Device:Desktop, Agent:Window, Refer:Direct, Period:Night}
{Device:Mobile, Agent:iPad, Refer:Facebook, Period:Evening}
Visualization tool for dashboard


Map
Satellite
NAKHON NAYOK
PATHUM THANI
NAKHON PATHOM
CHACHOENGSAO
RATCHABURI
SAMUT SAKHON
SAMUT SONGKHRAM
CHON BURI
Google


Dense Layer
Prediction Task
Gated Recurrent Unit (GRU)
...
Project n+1
Embedding
Product Embedding
Dense Layers
Class probabilities
Features of Project 1
Features of Project 2
Features of Project n
Click sequence
Encoder


Product Embedding
Generation Model
Analysis for product insight
Embedding Model
Project Features
Product Features
Product Embedding
Generated User Log
0.5
0.3
0.2
Night
Evening
Noon
0.6
0.1
0.4
Direct
Facebook
Direct