数据挖掘导论

Introduction to Data Mining

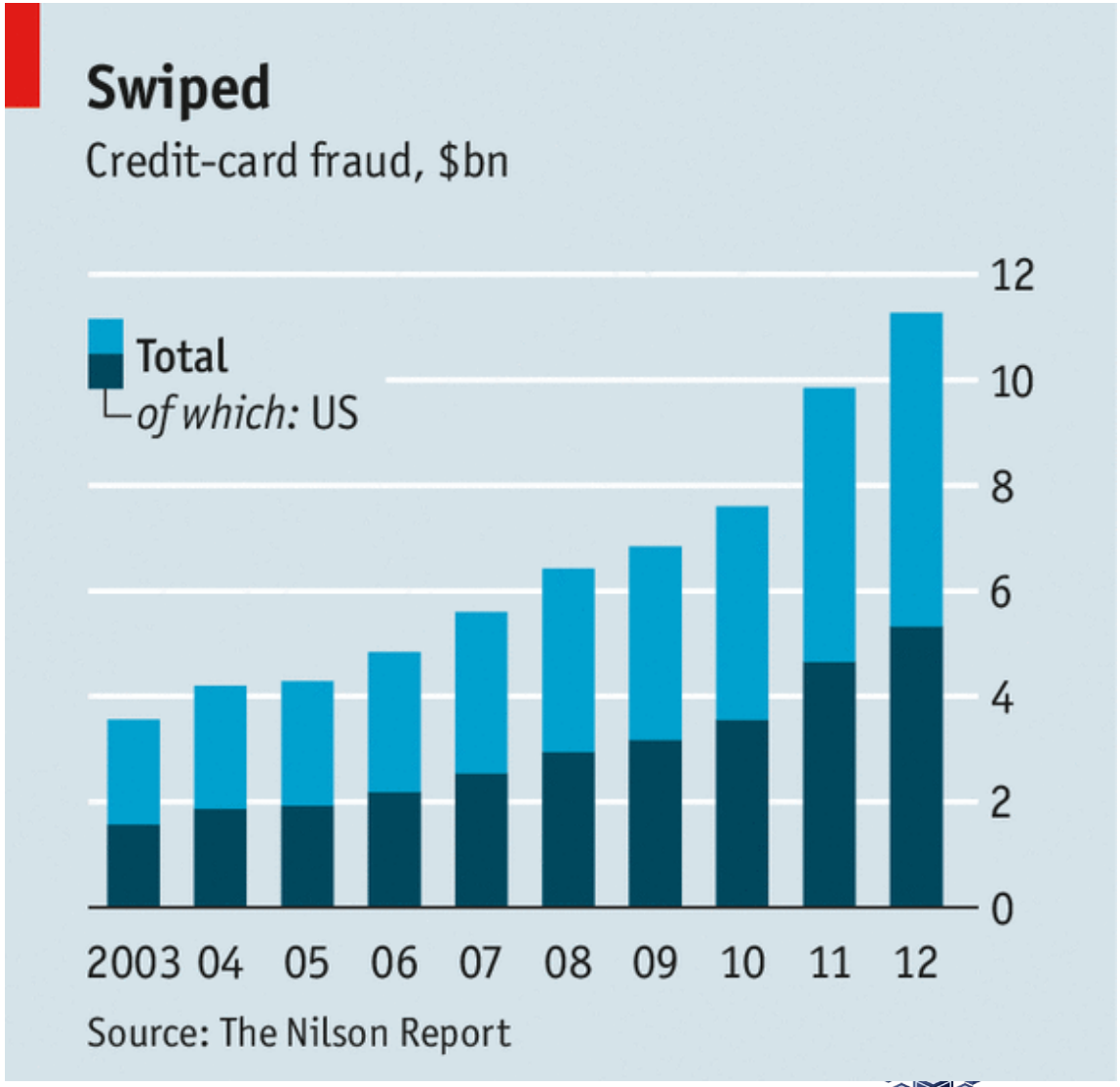
Outlier Detection Applications and Solutions

D I-LAB 数据智能实验室 DATA INTELLIGENCE LABORATORY
浙江大学 Zhejiang University

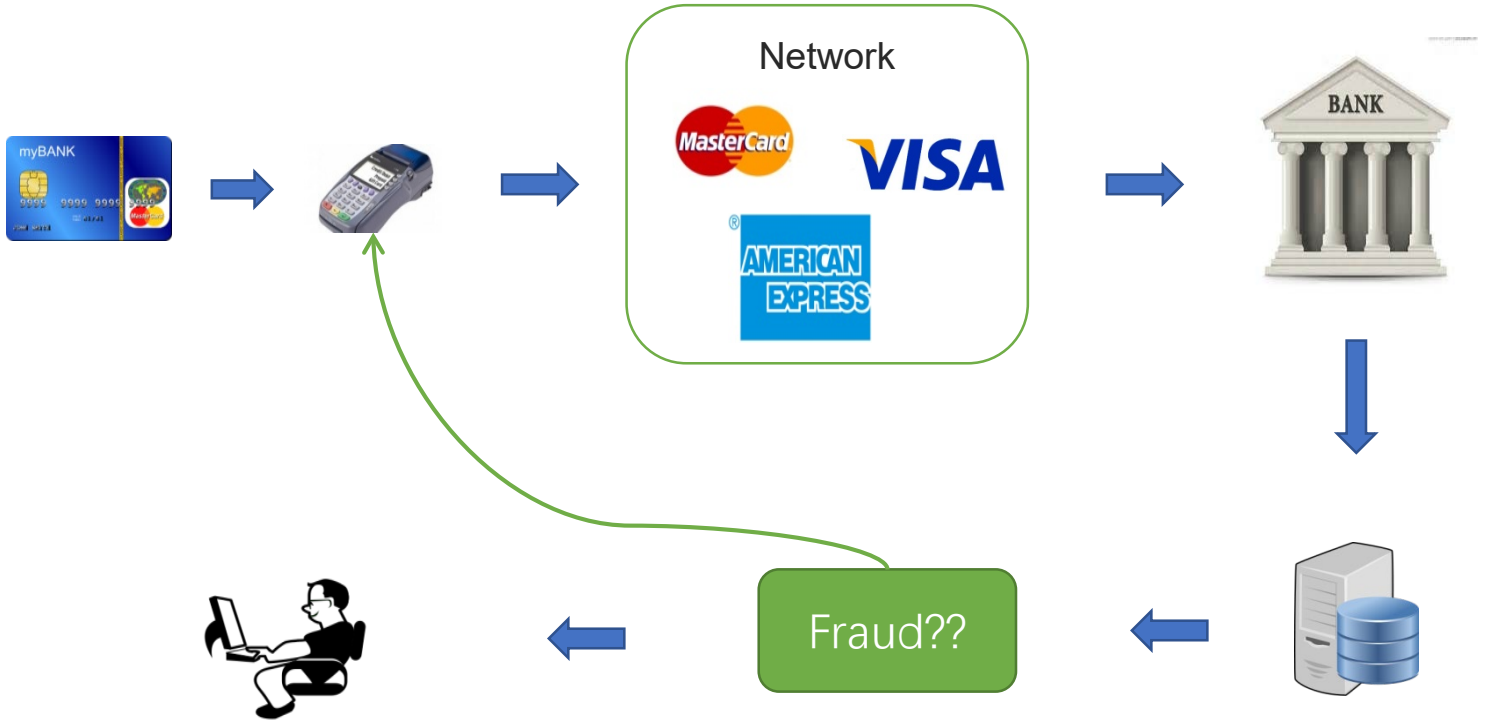
Agenda

- Credit Card Fraud Detection
- Spam Email Detection
- Trajectory Detour Detection
- Summary

Credit Card Fraud

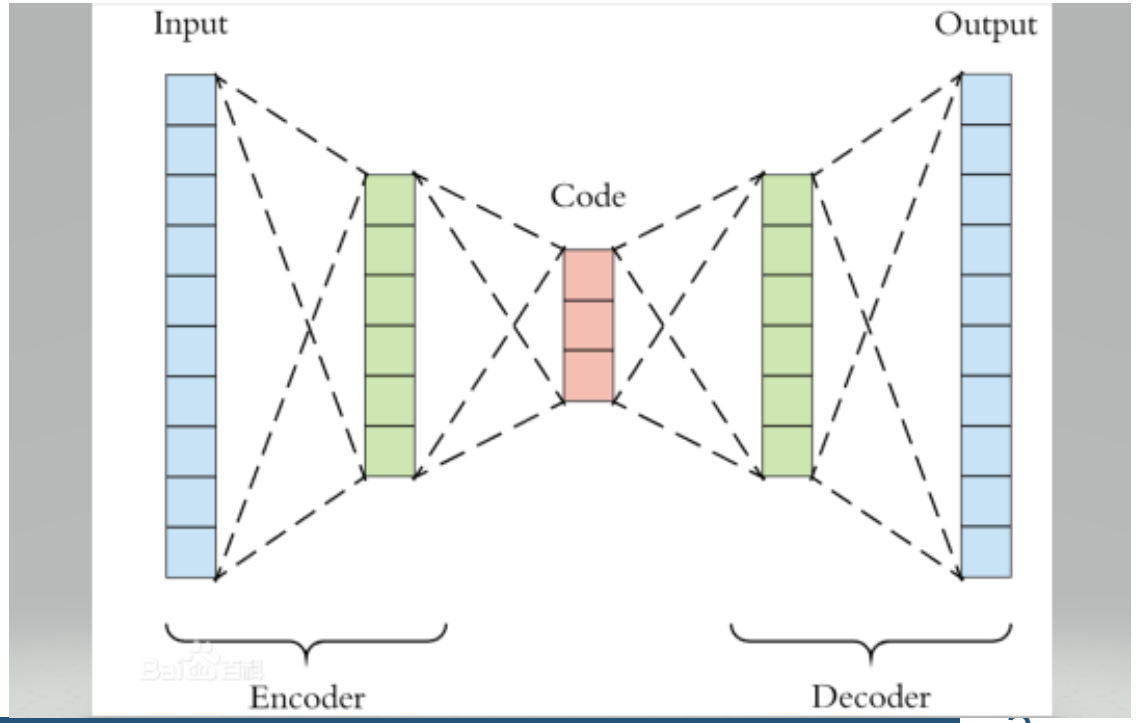


Credit Card Fraud



AutoEncoder

- An autoencoder consists of 3 components: encoder, code and decoder
- Compress the input into a lower-dimensional code and then reconstruct the output from this representation
- The goal is to get an output identical with the input.



AutoEncoder

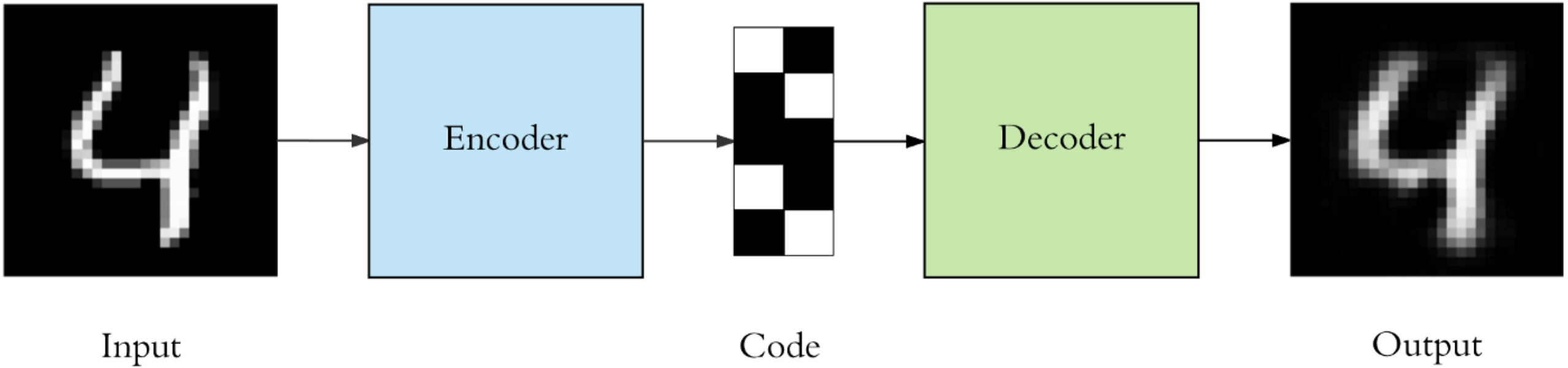


Image source: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

AutoEncoder for Fraud Detection

- AutoEncoder is forced to learn a condensed representation from which to reproduce the original input
- We feed it only normal transactions, which it will learn to reproduce with high fidelity
- If a fraud transaction is sufficiently distinct from normal transactions, the auto-encoder will have trouble reproducing it with its learned weights, and the subsequent reconstruction loss will be high
- Anything above a specific loss (treshold) will be flagged as anomalous and thus labeled as fraud
- Example available at
  - ✓ https://www.kaggle.com/code/robinteuwens/anomaly-detection-with-auto-encoders/notebook

Database

Raw attributes

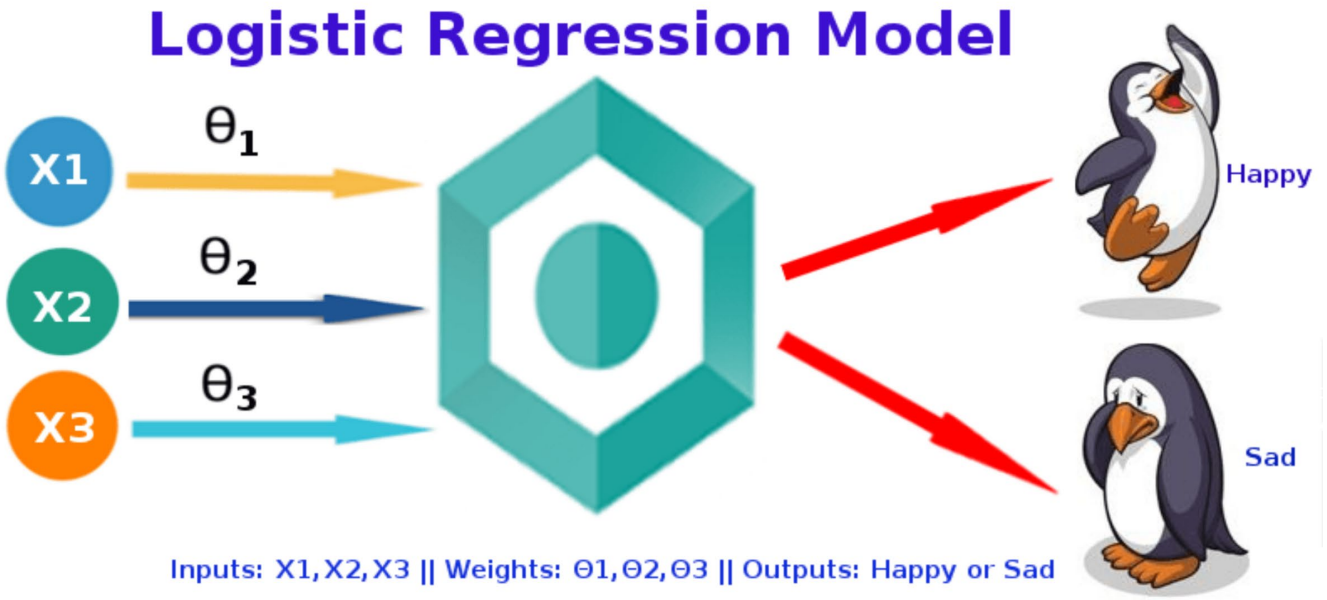
TRXID	Client ID	Date	Amount	Location	Type	Merchant Group	Fraud
1	1	2/1/12 6:00	580	Lux	Internet	Airlines	No
2	1	2/1/12 6:15	120	Lux	Present	Car Renting	No
3	2	2/1/12 8:20	12	Bel	Present	Hotel	Yes
4	1	3/1/12 4:15	60	Lux	ATM	ATM	No
5	2	3/1/12 9:18	8	Fra	Present	Retail	No
6	1	3/1/12 9:55	1210	Lux	Internet	Airlines	Yes

Other attributes

- ✓ Age, country of residence, postal code, type of card

The following slides of credit card fraud are copied from Alejandro Correa Bahnsen

Logistic Regression

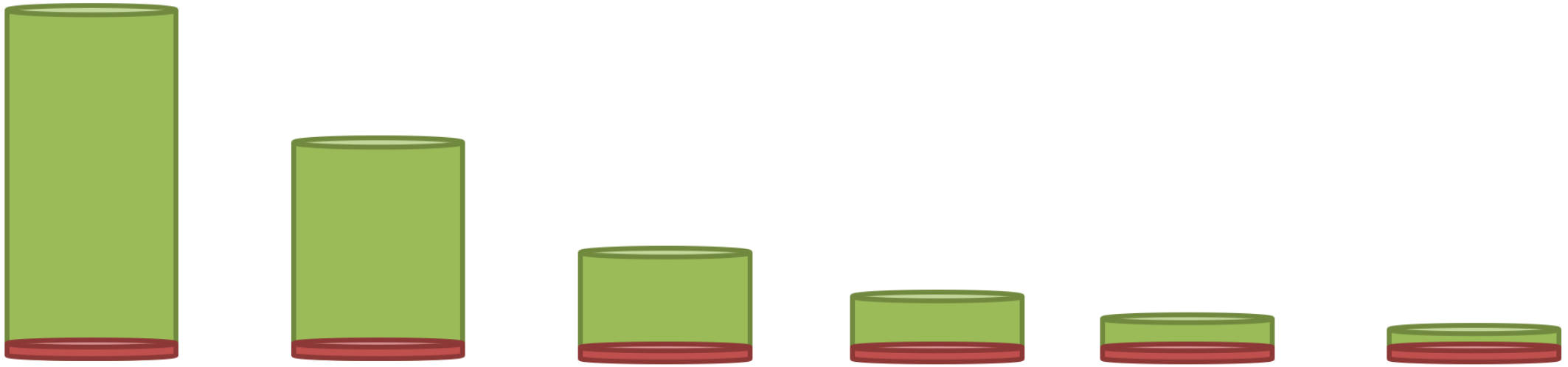


		True Class (y_i)	
		Fraud (y_i=1)	Legitimate (y_i=0)
Predicted class (p_i)	Fraud (p_i=1)	0	1
	Legitimate (p_i=0)	1	0

Logistic Regression

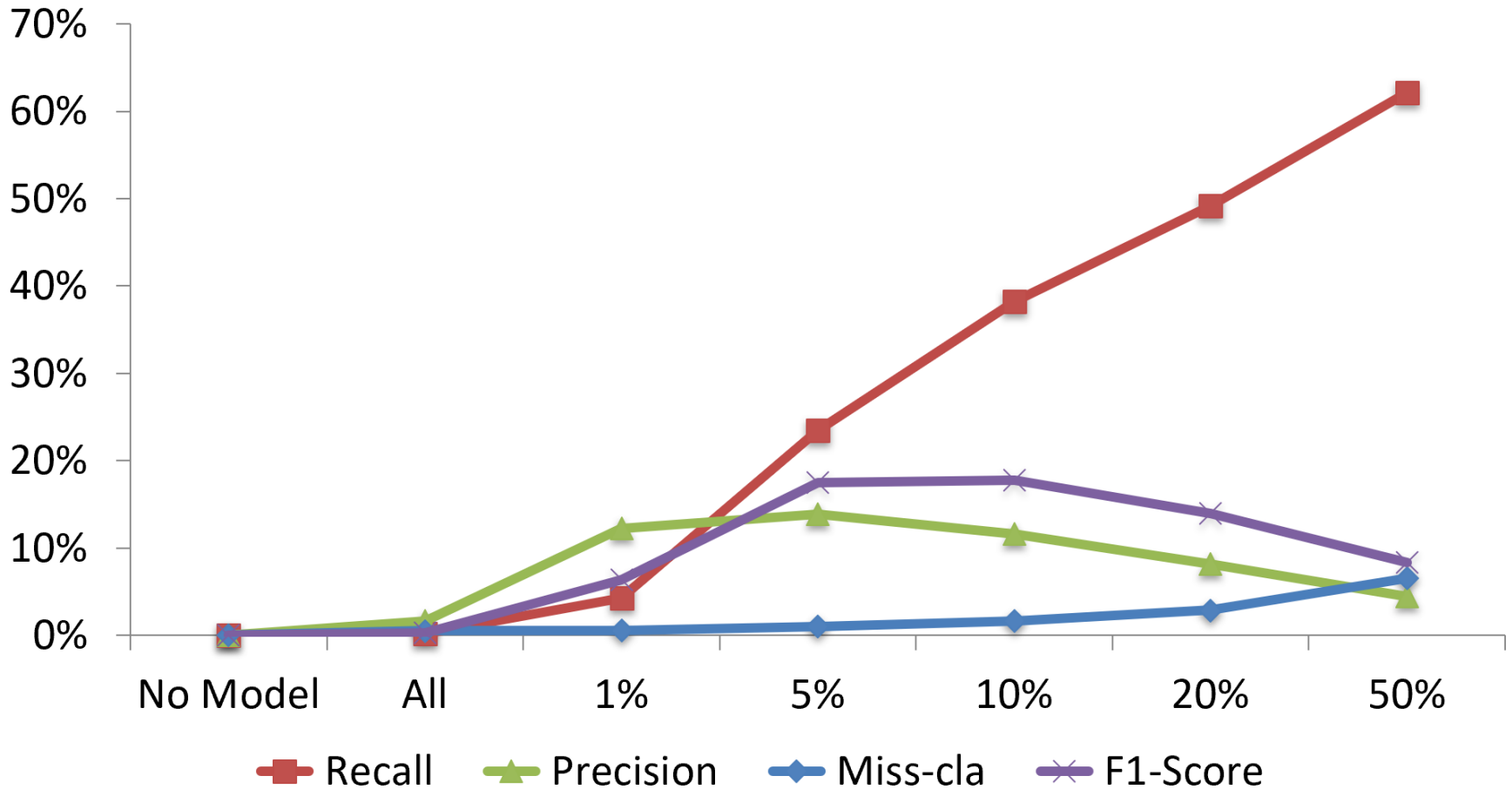
Different sampling strategies

- ✓ Select all the frauds and a random sample of the legitimate transactions.



Logistic Regression

Results



Cost Analysis

- **True Positive and False Positive**
  - ✓ Customer service center need to contact the customer for the issue
- **False Negative**
  - ✓ When a fraud is not detected, the cost is much much larger

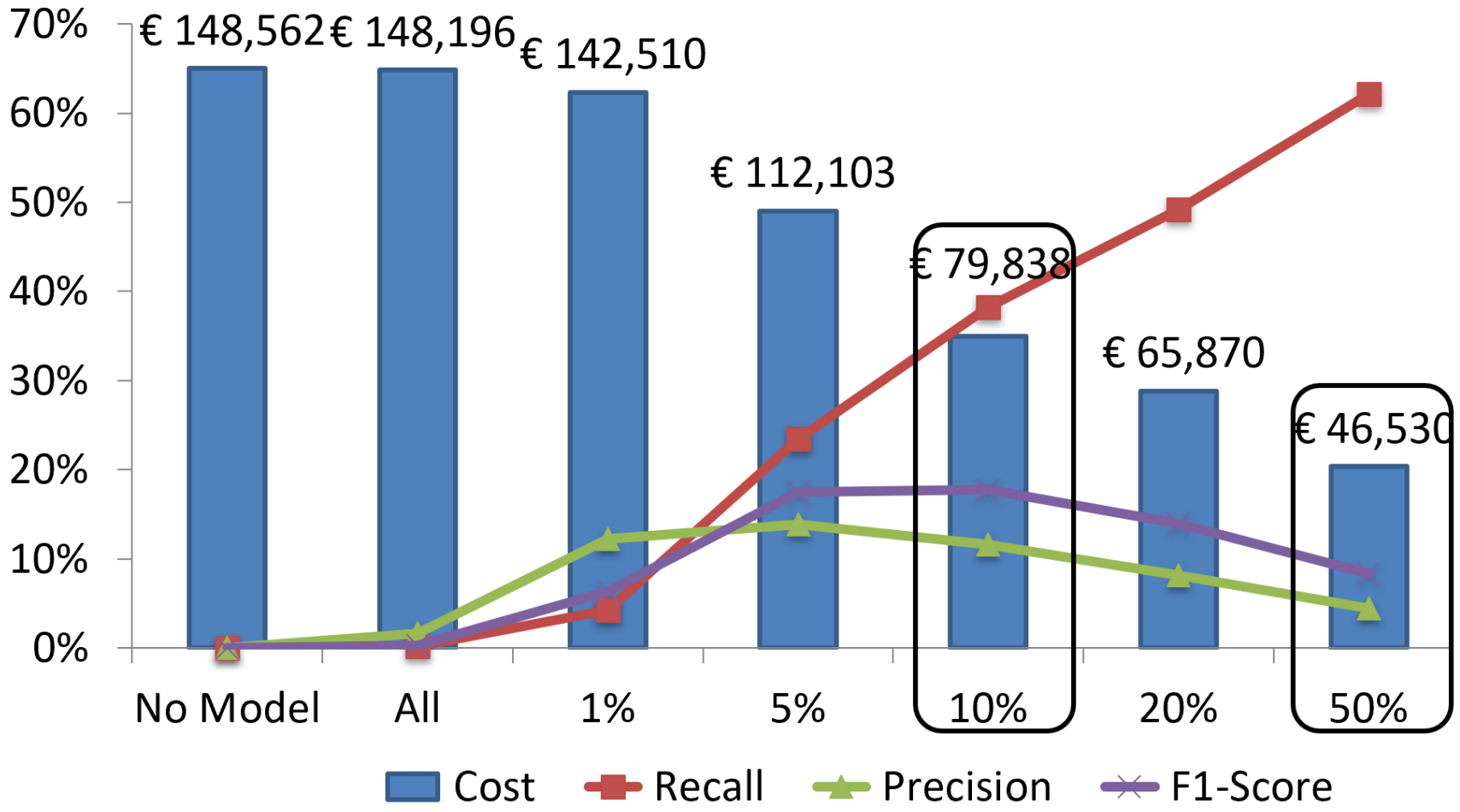
http://www.xinhuanet.com › politics · Translate this page

最高法：银行卡被盗刷银行应赔偿损失 - 新华网

May 25, 2021 — 发生伪卡盗刷交易或者网络盗刷交易，信用卡持卡人基于信用卡合同法律关系请求发卡行返还扣划的透支款本息、违约金并赔偿损失的，人民法院依法予以支持； ...

Logistic Regression

Results



Logistic Regression

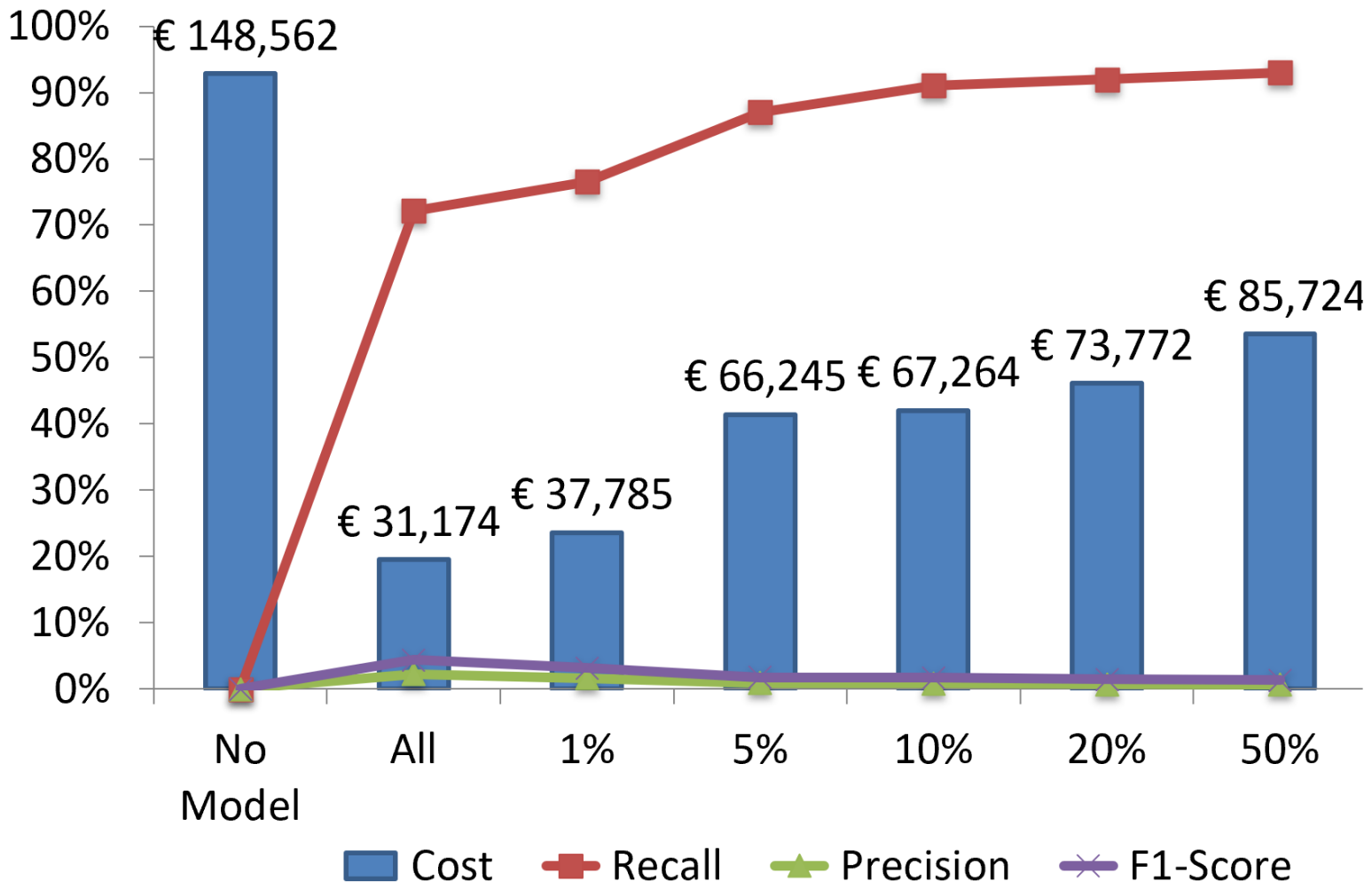
Improvement

- ✓ Select all the frauds and a random sample of the legitimate transactions.
- ✓ Model selected by cost, is trained using less than 1% of the database, meaning there is a lot of information excluded
- ✓ The algorithm is trained to minimize the miss-classification (approx.) but then is evaluated based on cost
- ✓ Why not train the algorithm to minimize the cost instead?

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[y_i\left(p_\theta^*(x_i)Ca + (1 - p_\theta^*(x_i))Amt_i\right) + (1 - y_i)p_\theta^*(x_i)Ca\right]$$

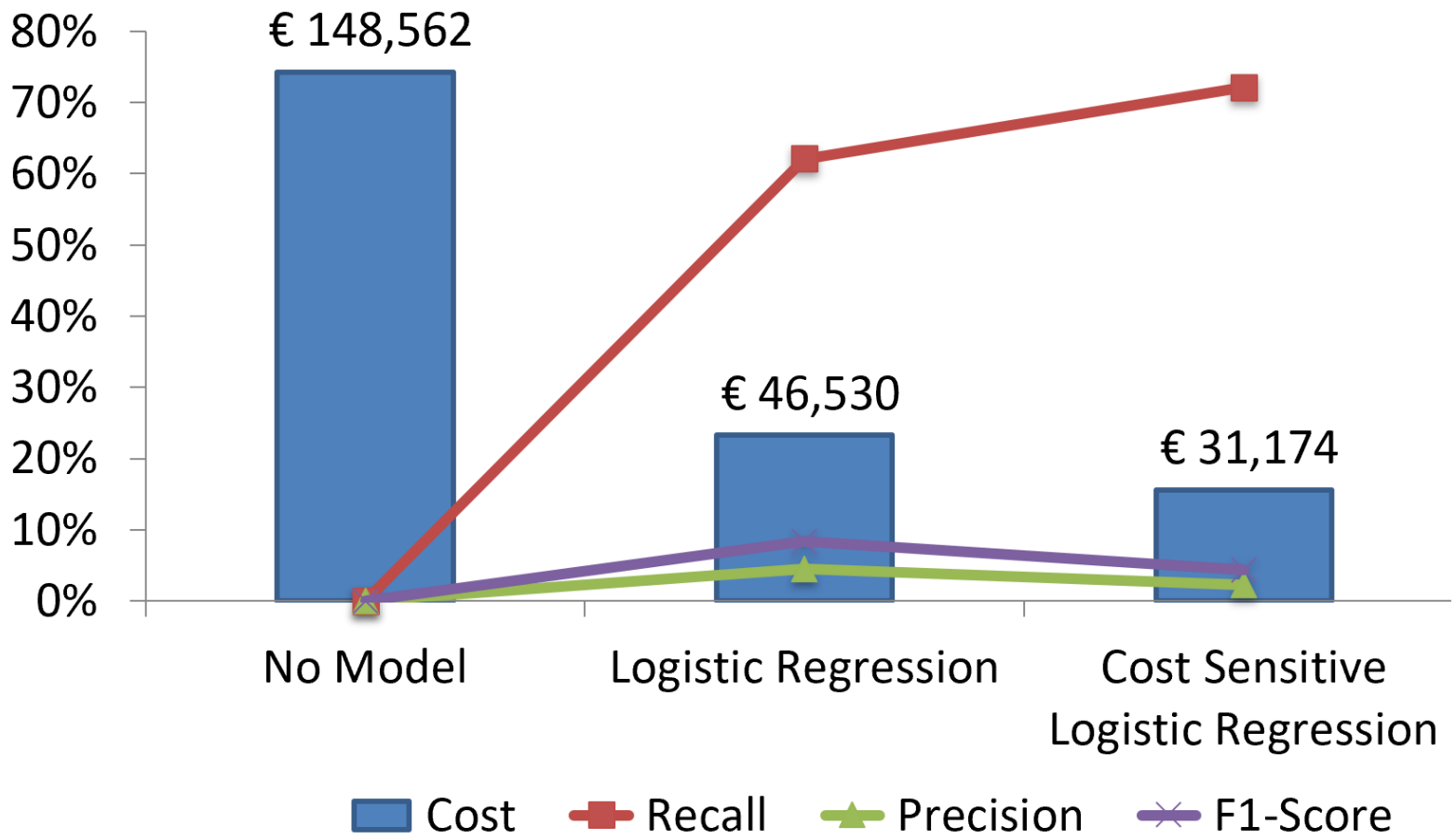
Logistic Regression

Results



Logistic Regression

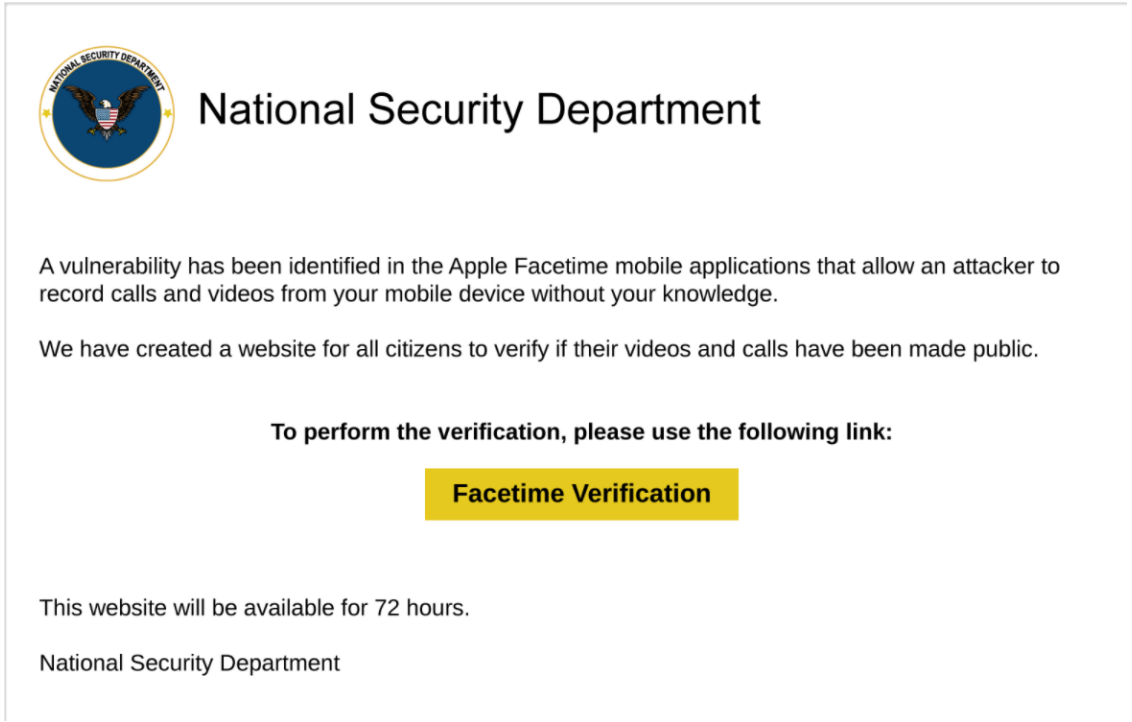
Results



Agenda

- Credit Card Fraud Detection
- Spam Email Detection
- Trajectory Detour Detection
- Summary

Spam Emails



National Security Department

A vulnerability has been identified in the Apple Facetime mobile applications that allow an attacker to record calls and videos from your mobile device without your knowledge.

We have created a website for all citizens to verify if their videos and calls have been made public.

To perform the verification, please use the following link:

Facetime Verification

This website will be available for 72 hours.

National Security Department

b)

How to Make Money Online in 2019

Our members are making around $500-2000 per day without effort? and their bank account is growing fabulously.

All this is possible due to the new patented system of increasing profits, which allows you to earn at least $1000 per day with minimum investment.

New passive income method 2019! Register now if you don?t want to regret later!

Make 1000 Euro daily!

http://mymoneyhere.info/go145

Spam Emails

Re:李春瑶Rfg5

会务组-Dale： 京公网安备11010802032783Copyright@2007-中国科学报社AllRightsReserved

前天 (2)

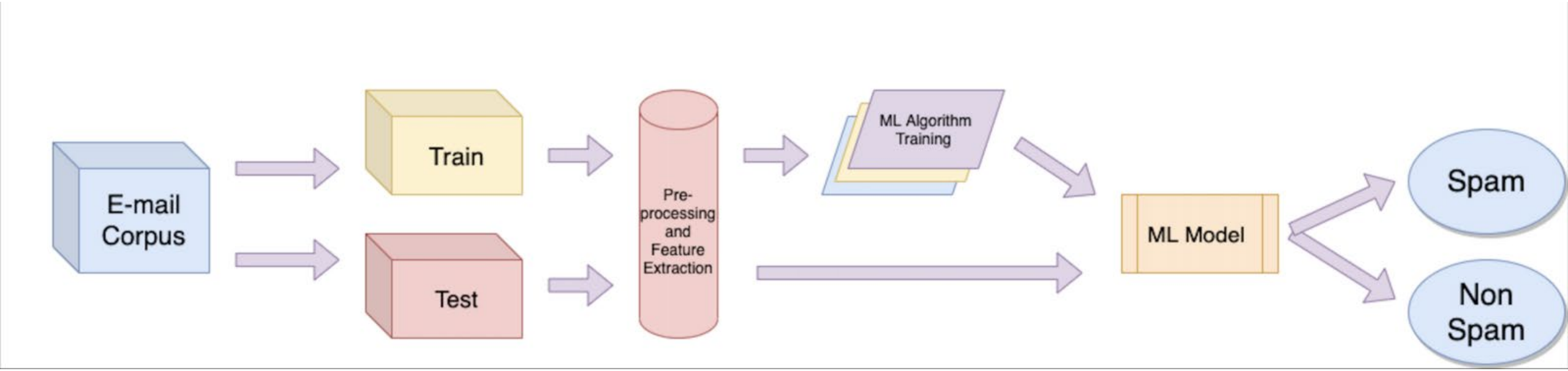
Submit Papers for Publication

netjourncalls： Dear Colleague, African Educational Research Journal (ISSN 2354-2160) is a peer-reviewed open access

未读：顶级SSCI期|刊源，正|刊发|表YRryR

iasrs编辑部小曾875： 尊敬的作者： 您好！ 重点推荐，顶级期刊，代表作期刊，SSCI期刊源； CiteScore排名，分区JCR:Q1 中科|

Binary Classification with Machine Learning



https://www.enjoyalgorithms.com/blog/email-spam-and-non-spam-filtering-using-machine-learning

Binary Classification with Machine Learning

- **Important attributes:** Frequency of repeated words, Number of semantic discrepancies, an Adult content bag of words, etc.
- **Additional Attributes:** Sender account features like Sender country, IP address, email, age of sender, Number of replies, number of recipients, website address.
  Note: These web addresses are converted in the word format only. For example, https://www.google.com/ can be converted to "HTTP google."
  Sometimes these processes are called Normalization.
- **Less important attributes:** Geographical distance between sender and receiver, Sender's date of birth, Account lifespan, Sex of sender, and Age of the recipient.

https://www.enjoyalgorithms.com/blog/email-spam-and-non-spam-filtering-using-machine-learning

Multi-Class Spam Email Detection

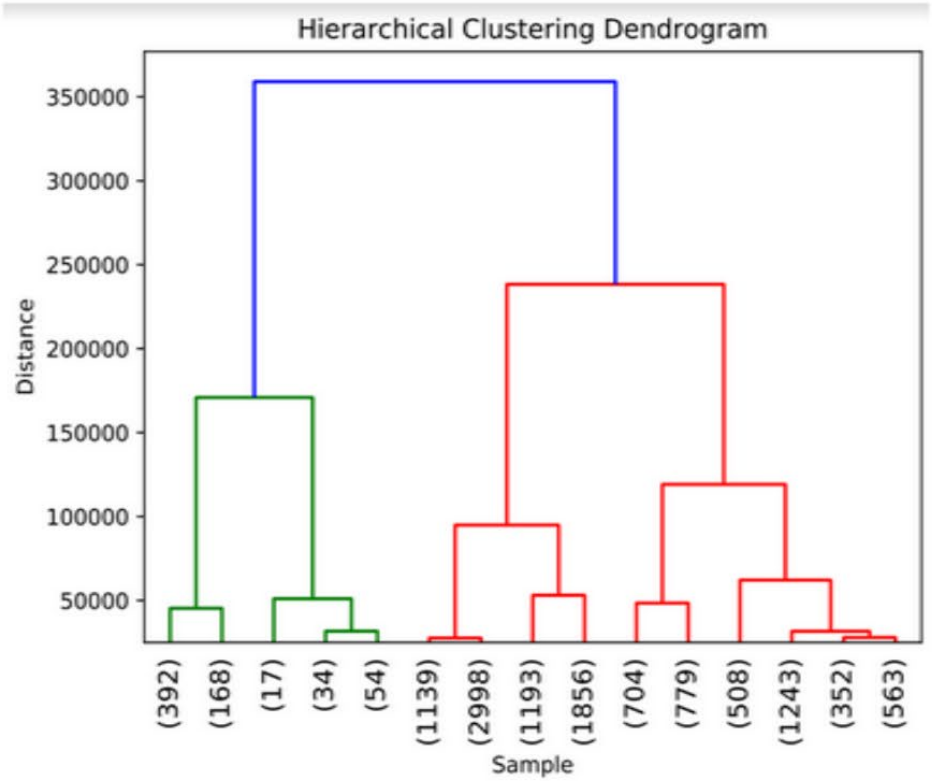
a)

Spam Emails Resources

b)

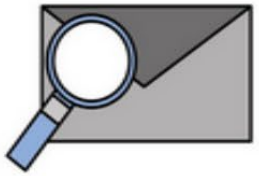
Creation of SPEMC-11K

c)



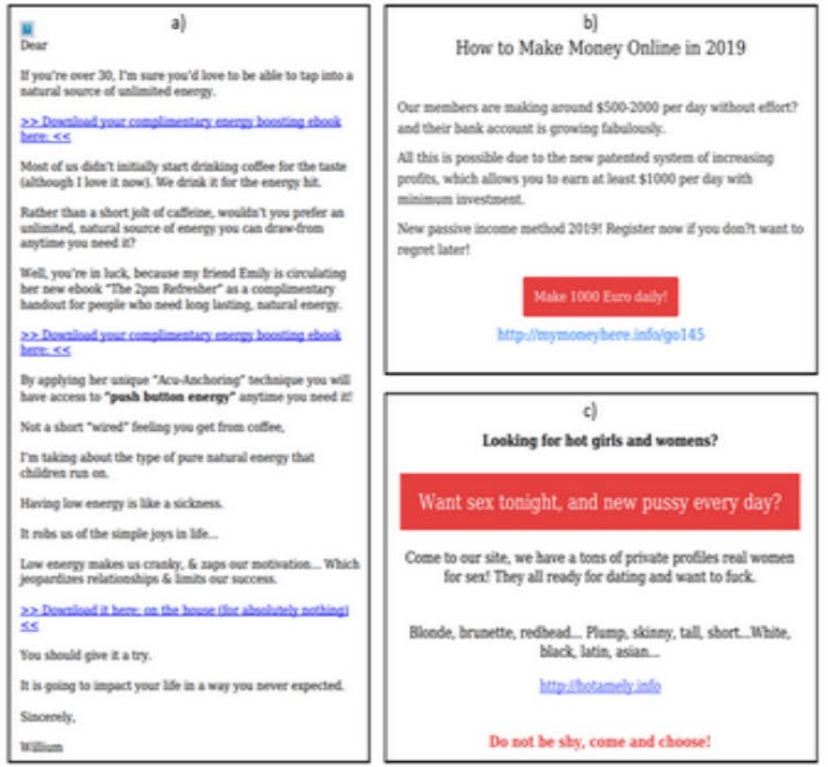
Hierarchical Clustering

d)



Visual Inspection

e)



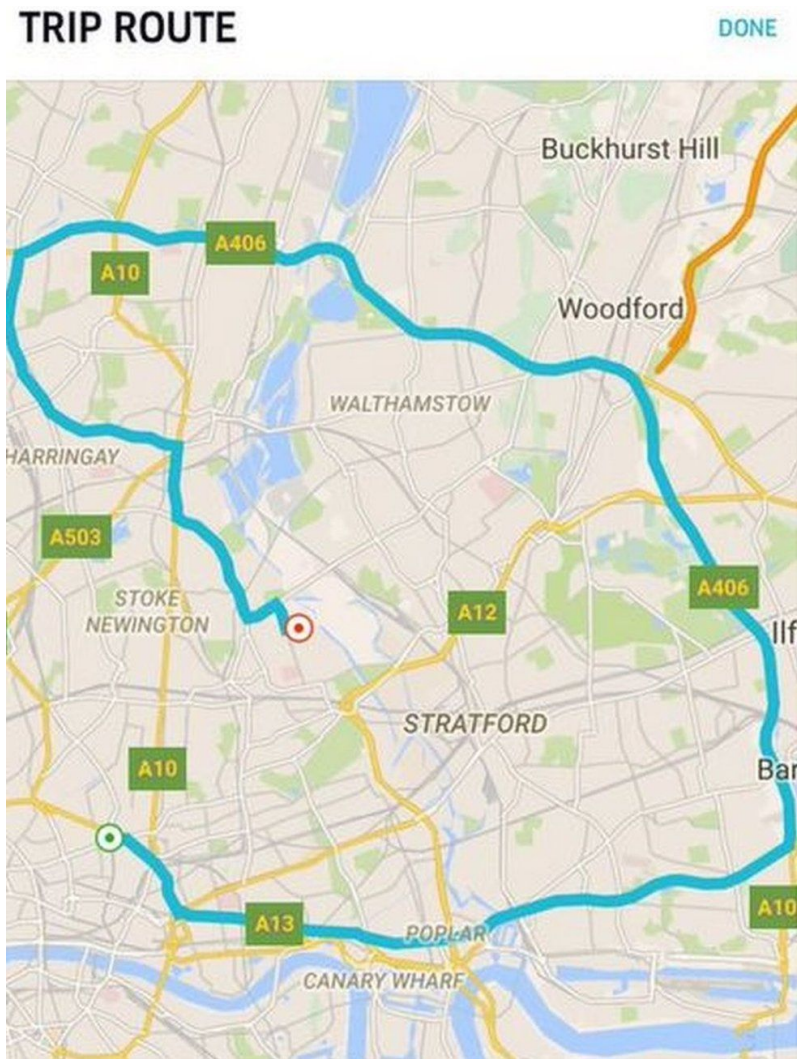
Classes Definition

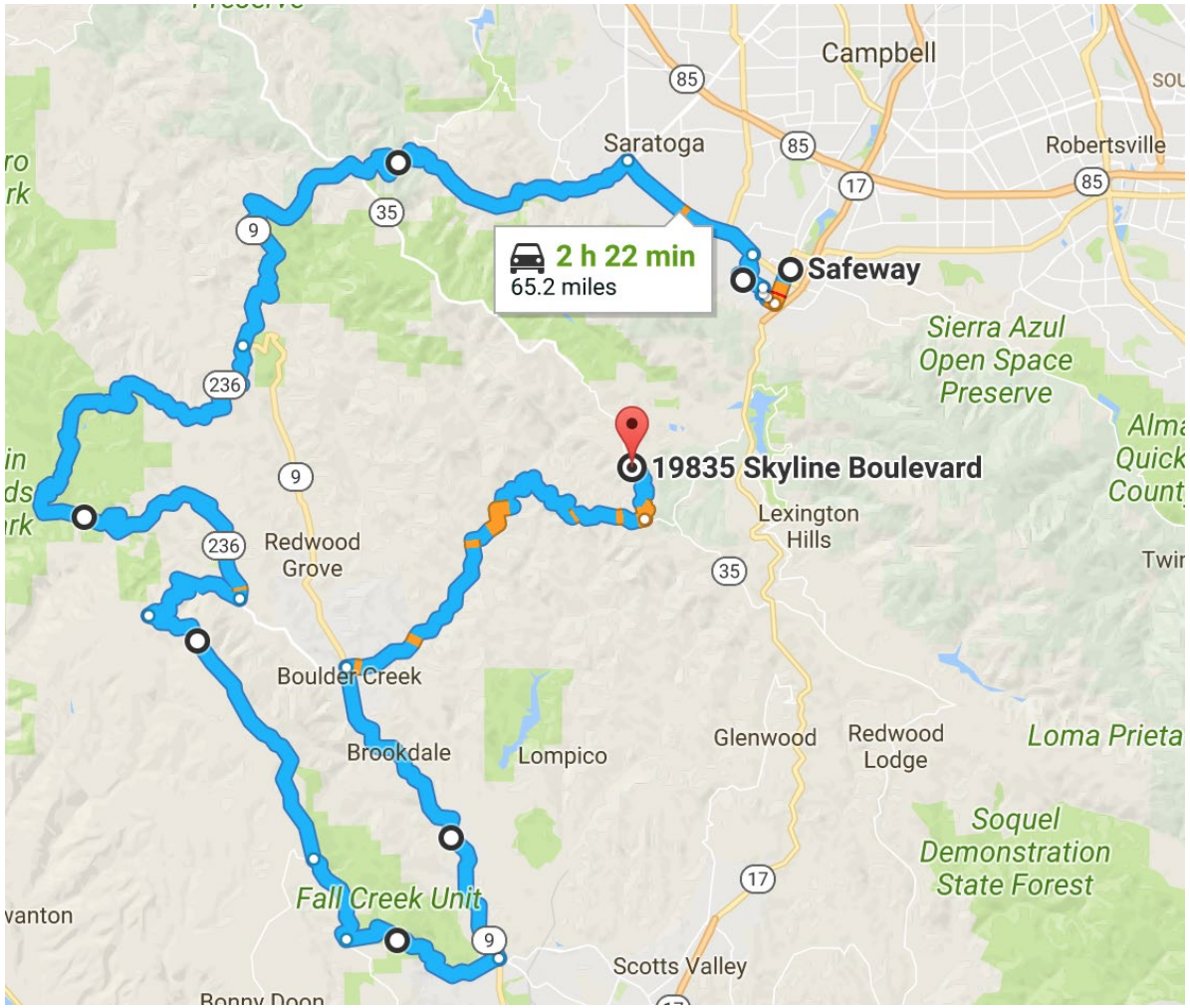
https://arxiv.org/pdf/2005.08773v1.pdf

Agenda

- Credit Card Fraud Detection
- Spam Email Detection
- Trajectory Detour Detection
- Summary

Trajectory Detour Detection





Trajectory Detour Detection

Offline Classification Model

- ✓ Distance-based feature: the ratio between the actual route and the reference route
- ✓ Time-based feature: the ratio between the actual travel time and the reference travel time
- ✓ The model is trained with logistic regression

$$X^{(1)} = \frac{Dis(atr)}{Dis(r(s_1, s_n, t_1))} - 1$$

$$X^{(2)} = \frac{At(atr)}{Et(r(s_1, s_n, t_1), t_1)} - 1$$

https://arxiv.org/pdf/1910.06949.pdf

Trajectory Detour Detection (TKDE 2021)

候鸟迁徙

沙漠马拉松

欧式空间

路网空间

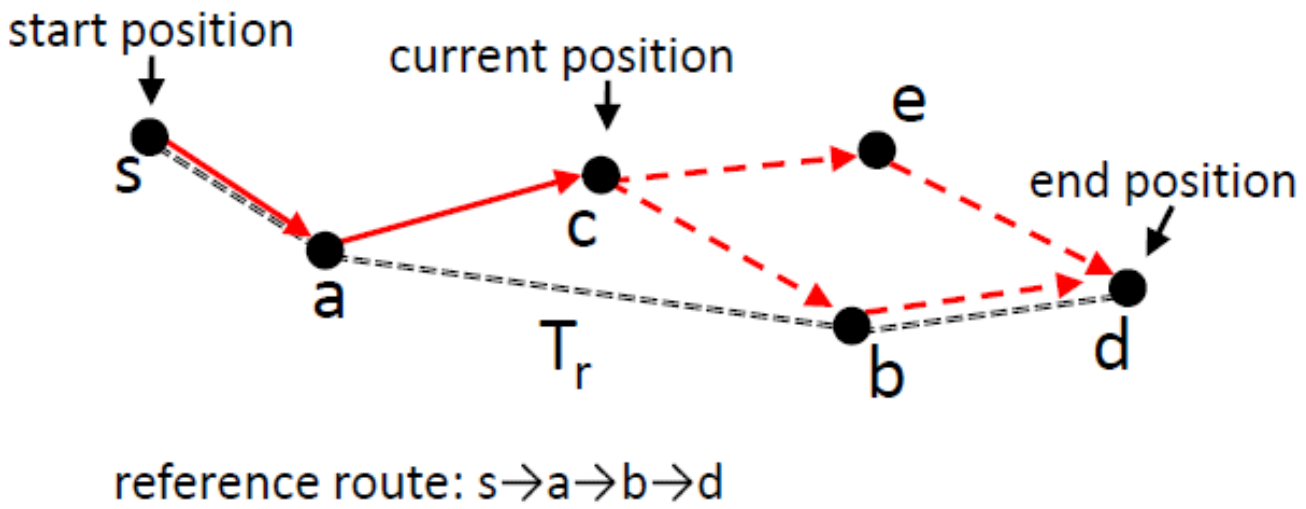
问题定义

偏移距离

- 从当前位置到终点d的所有可能路径中与参考路径T_r的轨迹距离最小值

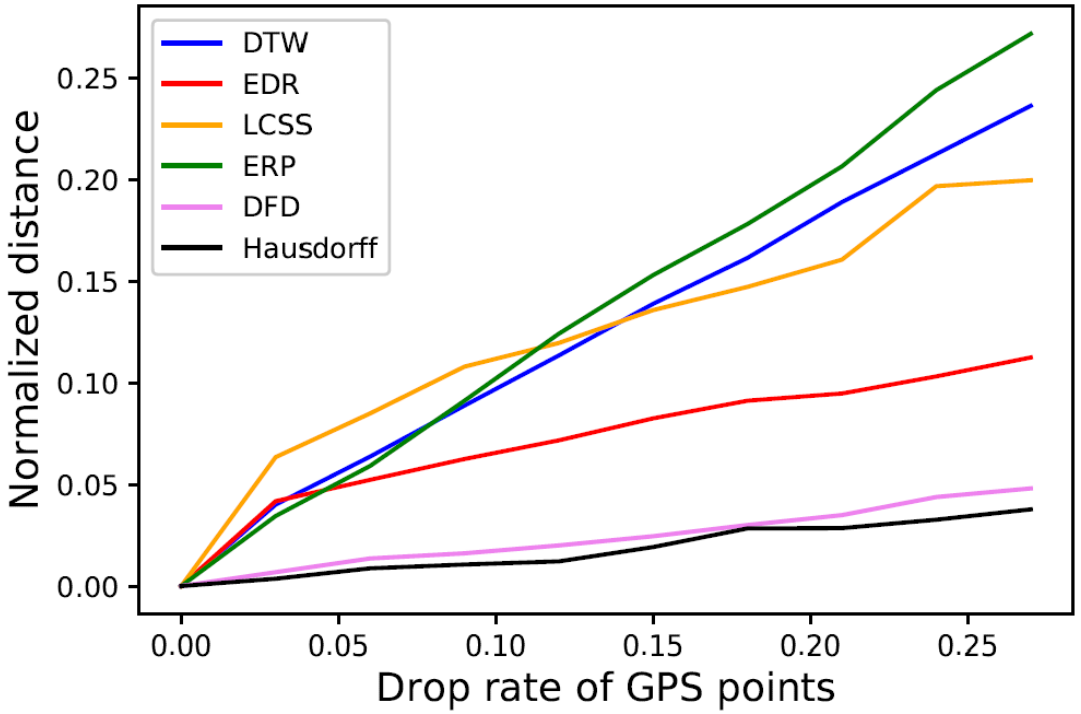
$$dist_p(T_{s\rightsquigarrow p_c}, T_r) = \min_{p_c \rightsquigarrow d} dist(T_{s\rightsquigarrow p_c \rightsquigarrow d}, T_r)$$

- 建立在已有的轨迹相似性函数基础之上

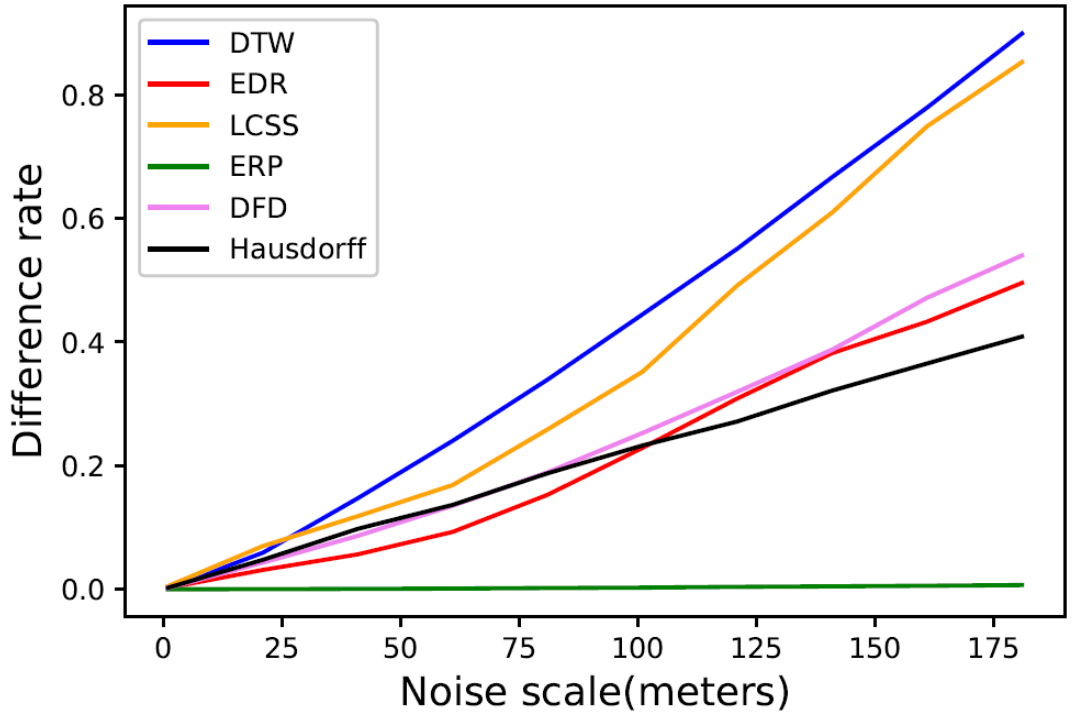


选择哪个轨迹相似性函数?

对采样率的敏感程度

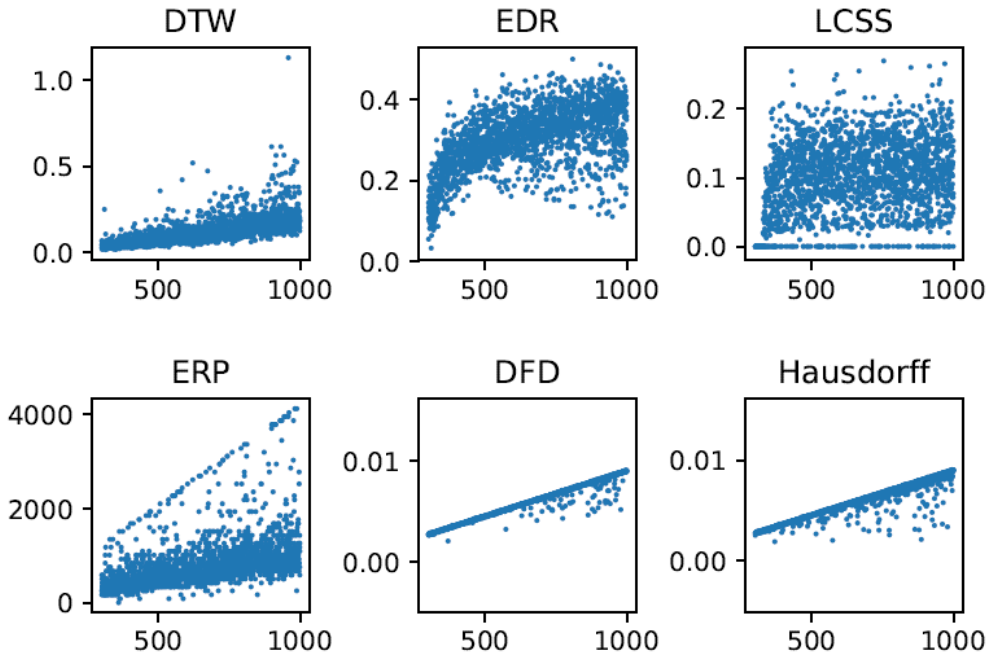


对采样误差的敏感程度



选择哪个轨迹相似性函数?

- 对位置偏离的敏感程度

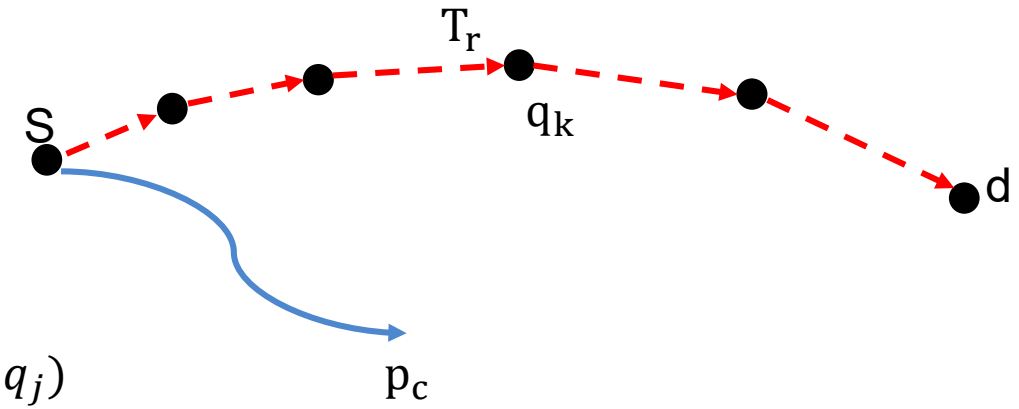


综合3方面的考量，选择DFD为CTSS的轨迹距离函数

欧式空间的偏移距离

- 偏移距离

$$\mathrm{dist_p}(\mathrm{T_{s\rightsquigarrow p_c}}, \mathrm{T_r}) = \min_{\mathrm{p_c \rightsquigarrow d}} \mathrm{dist}(\mathrm{T_{s\rightsquigarrow p_c \rightsquigarrow d}}, \mathrm{T_r})$$



- 如何寻找距离最小的路径T_*?
  - $DFD(c,j) = DFD(p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_c, q_1 \rightarrow q_2 \rightarrow \cdots \rightarrow q_j)$
  - $k = argmin_{1 \le j \le m} DFD(c,j)$
  - Construct $T_* = (p_1 \rightarrow p_2 \rightarrow \cdots p_c \rightarrow q_k \rightarrow q_{k+1} \rightarrow \cdots \rightarrow q_m)$
- 我们可以证明出如下两个性质
  - $DFD(T_*, T_r) = DFD(c,k)$
  - $dist_p(T_r, T_{s \rightsquigarrow p_c}) = DFD(T_*, T_r)$

如何快速找到最小的 $DFD(c,k)$?

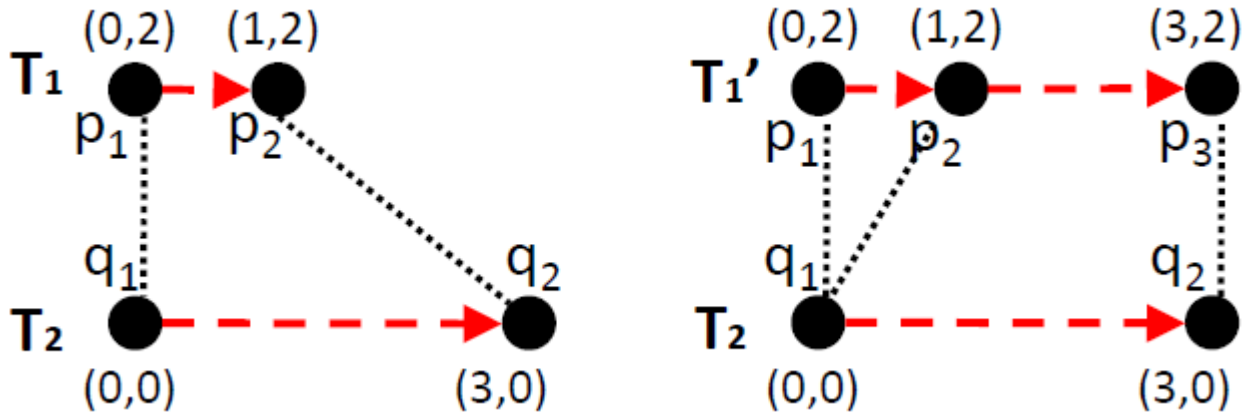
Input: Deviation threshold δ; Reference route T_r; Partial route $T_{s \rightsquigarrow p_c}$;
Output: $\min_{q_j \in T_r} DFD(c,j)$

```
1  d_min ← ∞;
2  for q_j ∈ T_r do
3  |   darray[j] ← dist(p_c, q_j);
4  end
5  Split darray into segments;
6  for each segment S do
7  |   if The state is NON-INC then
8  |   |   q_b ← the end point in S;
9  |   |   d_min ← min(d_min, DFD(c,b));
10 |   end
11 |   if The state is INC then
12 |   |   for each point q_j in S do
13 |   |   |   d_min ← min(d_min, DFD(c,j));
14 |   |   |   if DFD(c,j) ≤ dist(p_c, q_{j+1}) then
15 |   |   |   |   break;
16 |   |   |   end
17 |   |   end
18 |   end
19 end
20 return d_min;
```

路网情况下的偏移距离计算

挑战

- 无法像欧式空间那样直接构建最优路径
- DFD不具备monotonic property，无法利用Apriori算法在枚举空间进行剪枝



$$2\sqrt{2} = DFD(T_1, T_2) > DFD(T_1', T_2) = \sqrt{5}$$

路网情况下的偏移距离计算

加速策略

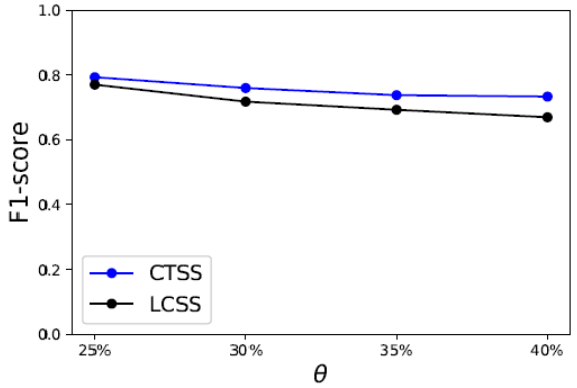
- 基于贪心算法找一条路径，如果该路径偏移距离小于δ，则此位置无异常
- 如果当前路网顶点p'到参考路径任意点距离都大于δ，则可以排除包含p'的路径

Lemma 3 *Given a trajectory* T' *with a point* p' *such that* $\min_{q_j \in T_r} \mathrm{dist}(p', q_j) \geq \delta$, *we have* $\mathrm{dist}_p(T', T_r) \geq \delta$.

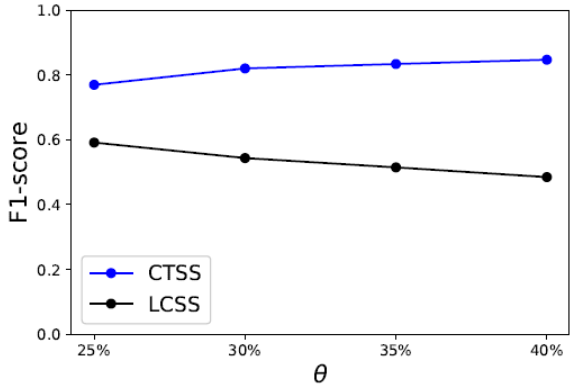
- 利用欧式空间的最短偏移距离作为lower bound

Lemma 4 *If* $\min_{q_j \in T_r} \mathrm{DFD}(c+t, j) \geq \delta$, *then for any trajectory* T' *extended from* T_{cand}, *we have* $\mathrm{DFD}(T', T_r) \geq \delta$.

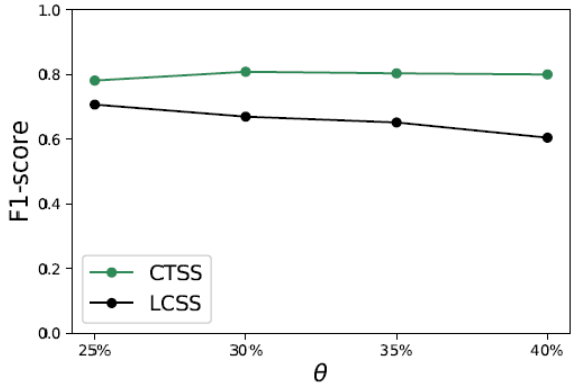
CTSS用于检测司机绕路



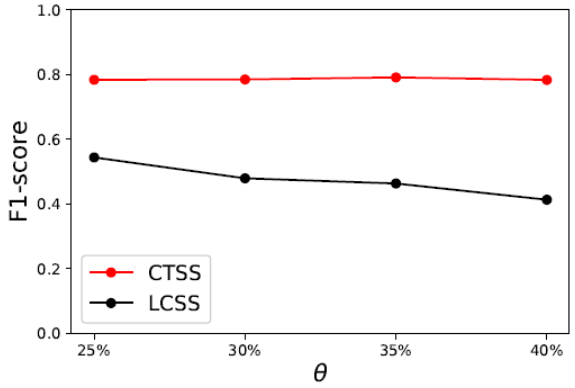
(a) Porto (6 ~ 8km)



(b) Porto (14 ~ 16km)



(c) Singapore (6 ~ 8km)



(d) Singapore (14 ~ 16km)

Agenda

- Credit Card Fraud Detection
- Spam Email Detection
- Trajectory Detour Detection
- Summary