Cyber Data Analytics

Sicco Verwer

Andre Teixeira

Qin Lin

Guest today: Bert Wolters from Adyen!

Cyber Data Analytics

- Aim
  - identify and prevent cyber criminals using machine learning

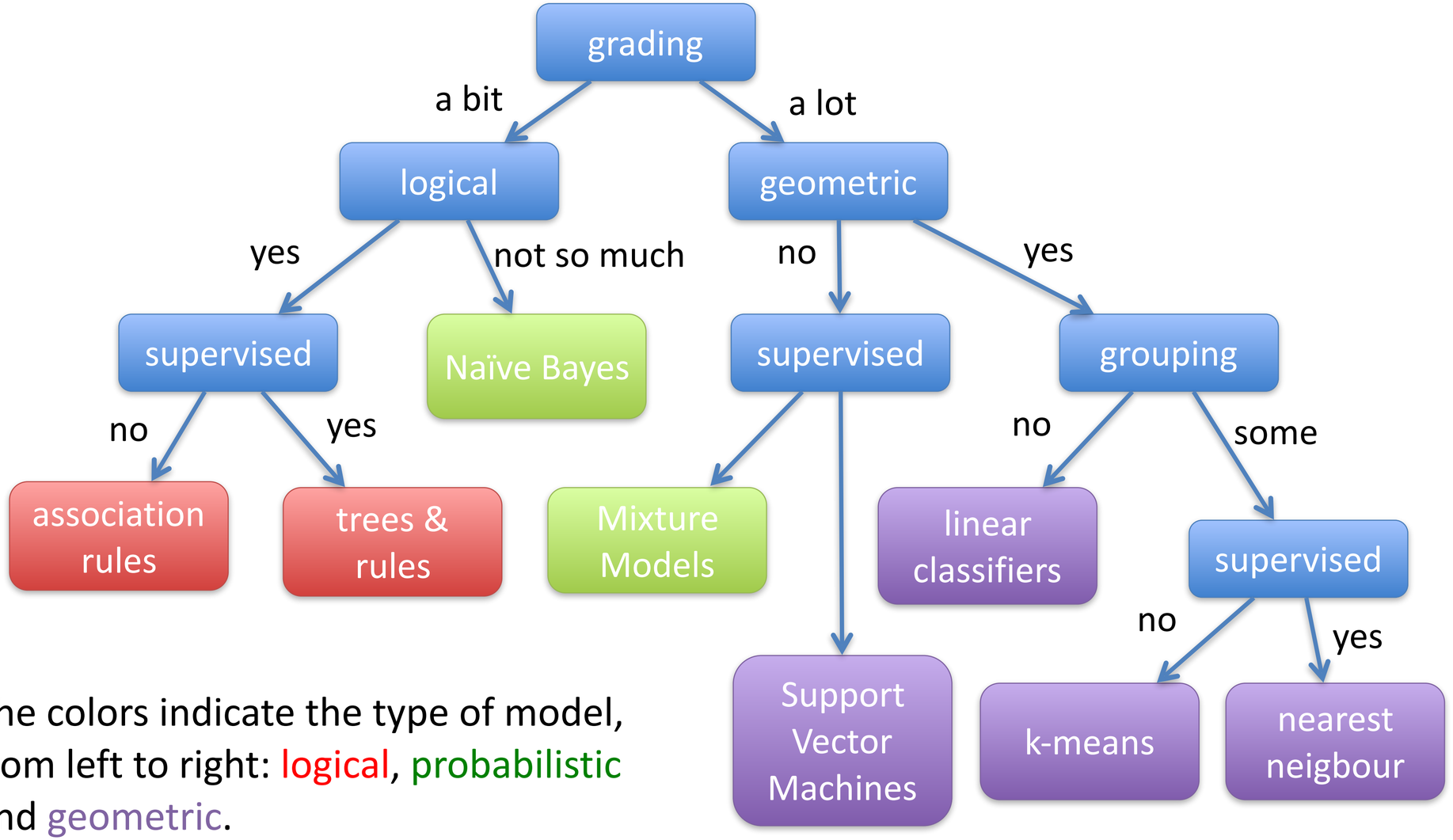
- How
  - use any method you can get your hands on
  - adapt it to cyber data:
    - huge structured data sets with few positives

Many kinds of models

- Geometric models use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- Probabilistic models view learning as a process of reducing uncertainty, modeled by means of probability distributions.
- Logical models are defined in terms of easily interpretable logical expressions.
- Characterized by their way of working:
  - Grouping models divide the instance space into segments; in each segment a very simple (e.g., constant) model is learned.
  - Grading models learning a single, global model over the instance space.

slides based on Peter Flach

ML Taxonomy



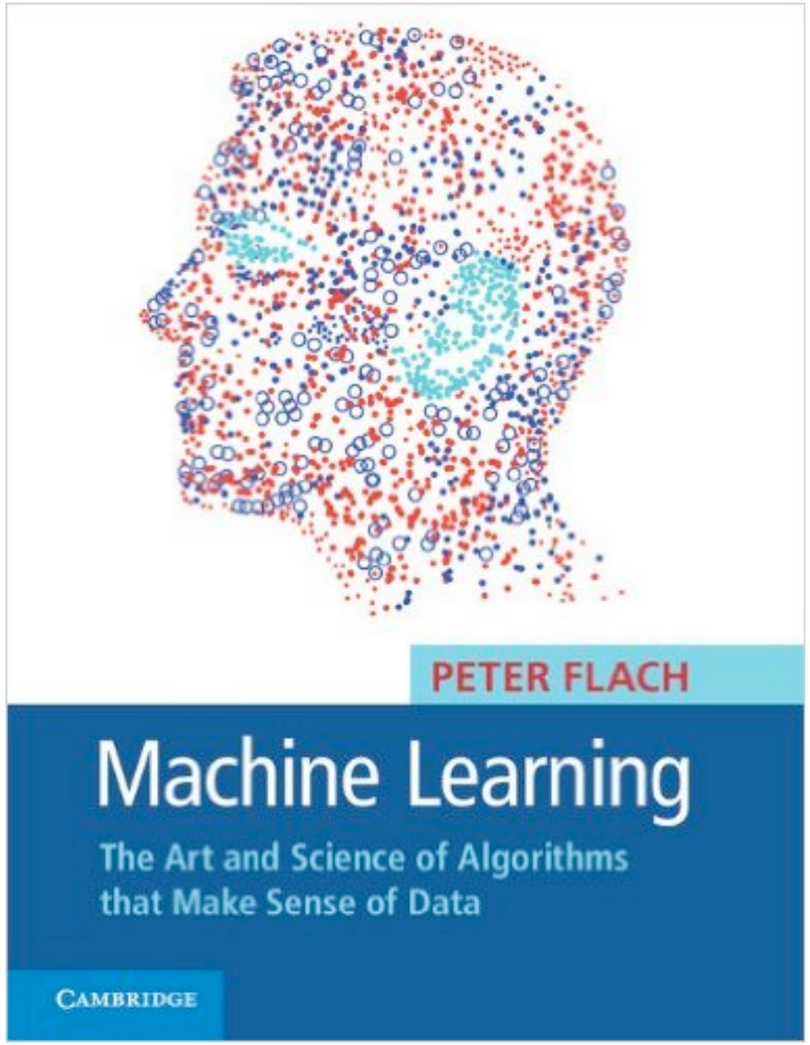
The colors indicate the type of model, from left to right: logical, probabilistic and geometric.

slides based on Peter Flach

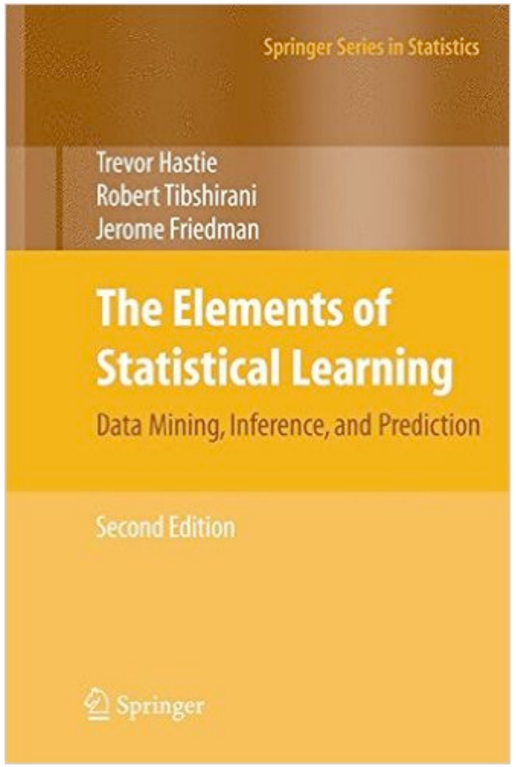
Need more background?

- Excellent text book, contains intuitive explanations of many of the basic machine learning methods and algorithms

- https://www.cs.bris.ac.uk/~flach/mlbook/



Free background (for all)



- Excellent textbook, quite some math, but also very good intuition

- http://statweb.stanford.edu/~tibs/ElemStatLearn/download.html

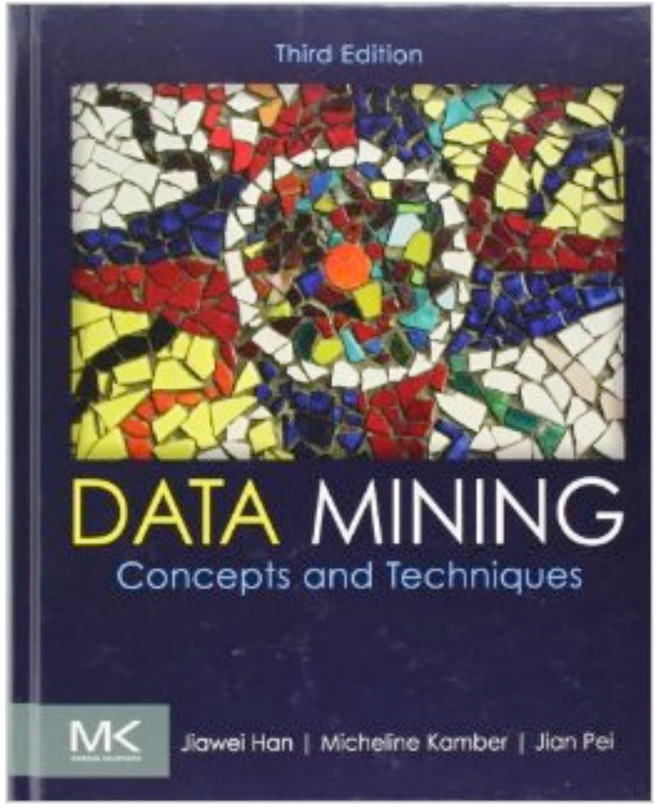
including some R code:

- http://statweb.stanford.edu/~tibs/ElemStatLearn/

Free background (from university)

- Good overview of basic data preprocessing, classification, pattern mining, and clustering

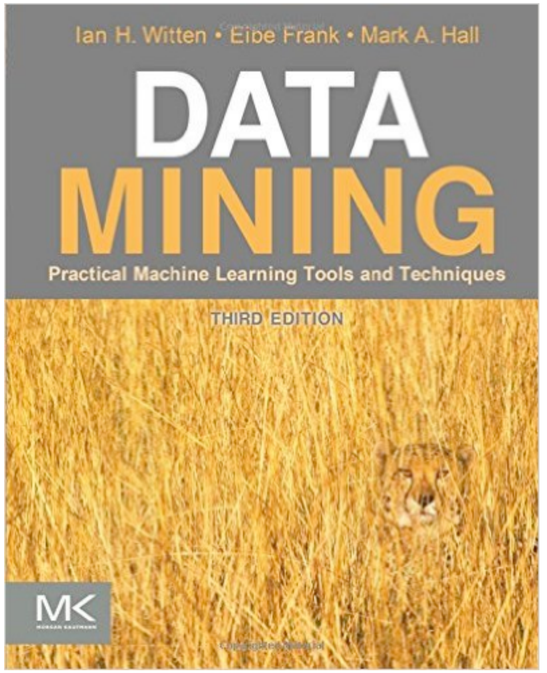
- little geometric, mostly logical



- http://www.sciencedirect.com/science/book/9780123814791

Free background (from university)

- Practical data mining in Weka



- http://www.cs.waikato.ac.nz/ml/weka/
- http://www.sciencedirect.com/science/book/9780123748560

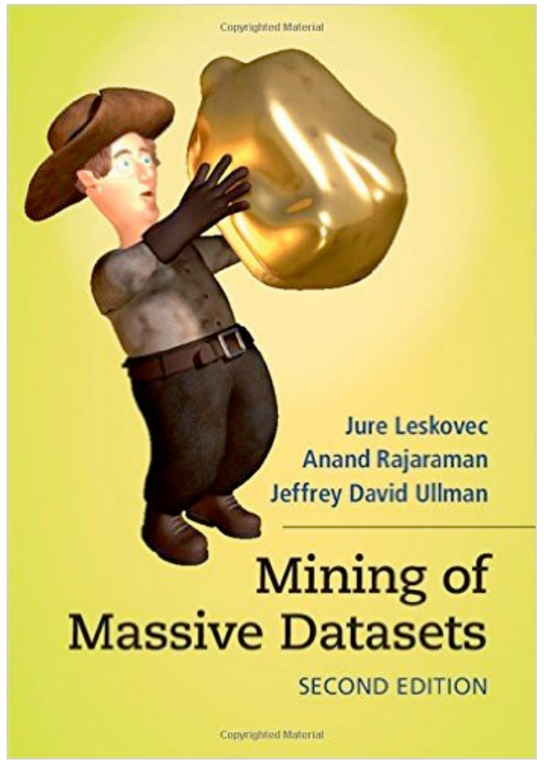
Free background (for all)

- Great text on learning

from big data

- Some content taught

in this course!



- http://infolab.stanford.edu/~ullman/mmds/book.pdf

Tools

- Scikitlearn – Python
  - http://scikit-learn.org/
- R – (e1071 package!, svmlib!)
  - https://www.r-project.org/
- Matlab (and prtools!)
  - http://prtools.org/
- Weka, see previous slide
- KNIME or RAPIDminer (GUIs)
  - https://www.knime.org/
  - https://rapidminer.com/

This course

- Understanding the inner workings and theory of algorithms that
  - week 1 – deal with unbalanced data
  - week 2 – detect stealthy attacks
  - week 3 – learn from data streams
  - week 4 – use distributed computation
  - week 5 – rely on multiple information sources
  - week 6 – operate on highly structured software data
  - week 7 – anonymize data and profile/track users

This course

- Understanding the inner workings and theory of algorithms that
  - week 1 – deal with unbalanced data
  - w
  -
  -
  -
  -
  -

Special focus:

Detecting anomalies
Sequential and structured data
Large data sets and many data sources
The detection game - discovery and ease of hiding

Rules

- Grade determined by lab-exam 50-50
- Lab average needs to be at least 6
  - 7 assignments
  - Write 2 A4-pages on the given exercises, deadline two weeks later, before lecture
  - Grade determined based more on quality of techniques used, problem solving capability, not so much on resulting classifier accuracy
- Closed-book exam on slide content, indicated scientific papers
  - Other references and books are background material, useful but do not need to be studied in detail

Lab/question hours

Reservering	Startdatum tijd	Einddatum tijd
36 (EWI) - Lipkenszaal	21/04/2016 13:30	21/04/2016 16:30
36 (EWI) - Lipkenszaal	28/04/2016 13:30	28/04/2016 16:30
36 (EWI) - Lipkenszaal	12/05/2016 13:45	12/05/2016 16:00
36 (EWI) - Lipkenszaal	19/05/2016 14:00	19/05/2016 16:00
36 (EWI) - Dijkstrazaal 9.150	26/05/2016 10:30	26/05/2016 12:30
36 (EWI) - Lipkenszaal	02/06/2016 13:30	02/06/2016 16:00
36 (EWI) - Timmanzaal	09/06/2016 9:30	09/06/2016 12:45
36 (EWI) - Vassiliadis zaal	09/06/2016 13:30	09/06/2016 16:00
36 (EWI) - Lipkenszaal	16/06/2016 13:30	16/06/2016 16:00

Credit Card Fraud Detection

adyen wherever people pay

Some facts
about Adyen

- $50 billion processed volume annually
- Global offices in 4 continents
- 590+ global employees
- 250+ payment methods, 150+ currencies
- All technology developed in-house
- One platform, all channels

wherever
people pay

Optimizing the value chain 3 benefits

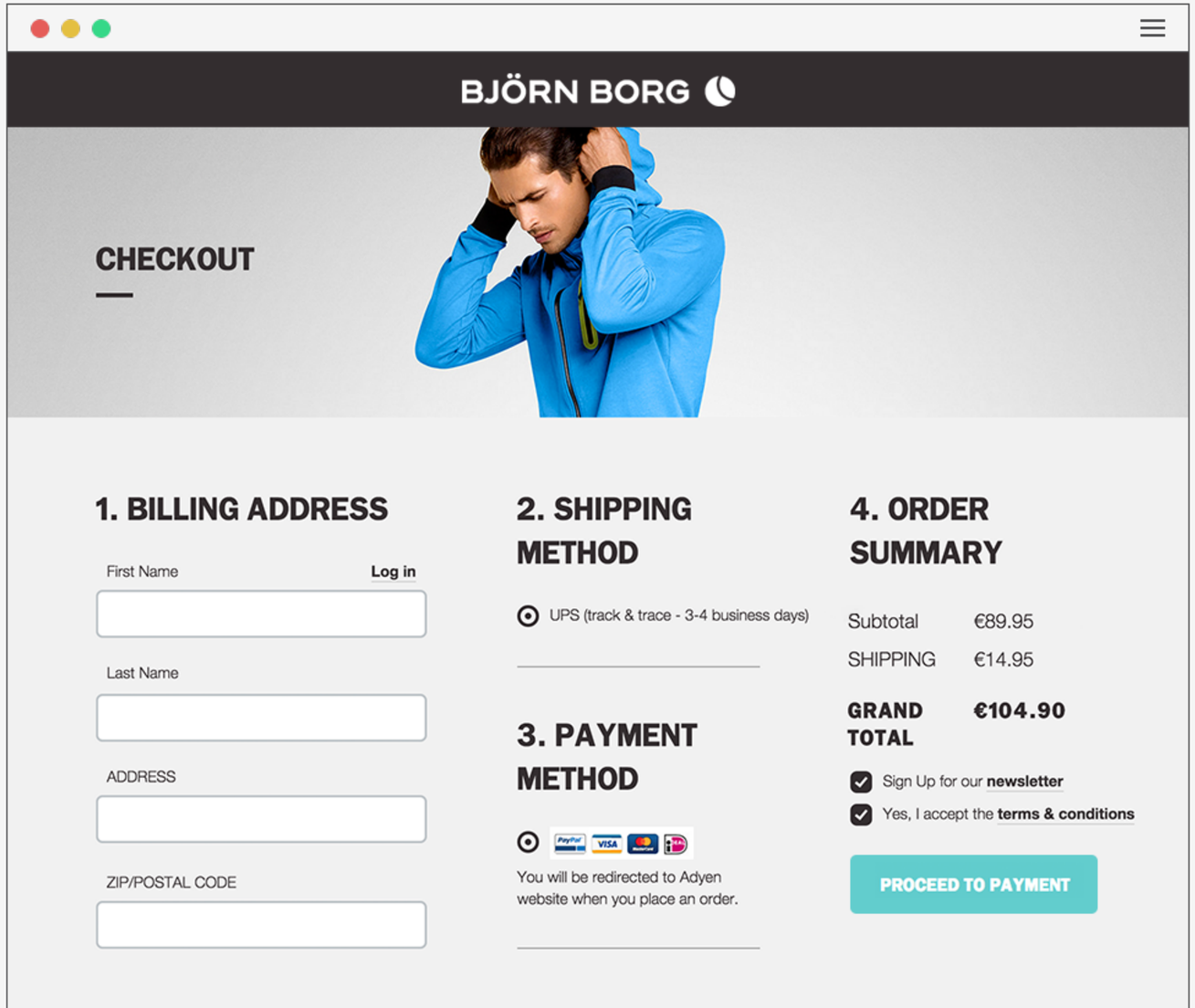
Merchants sell more via better data

A frictionless shopper experience

Global reach in one platform

The fraud game

How would you commit fraud?

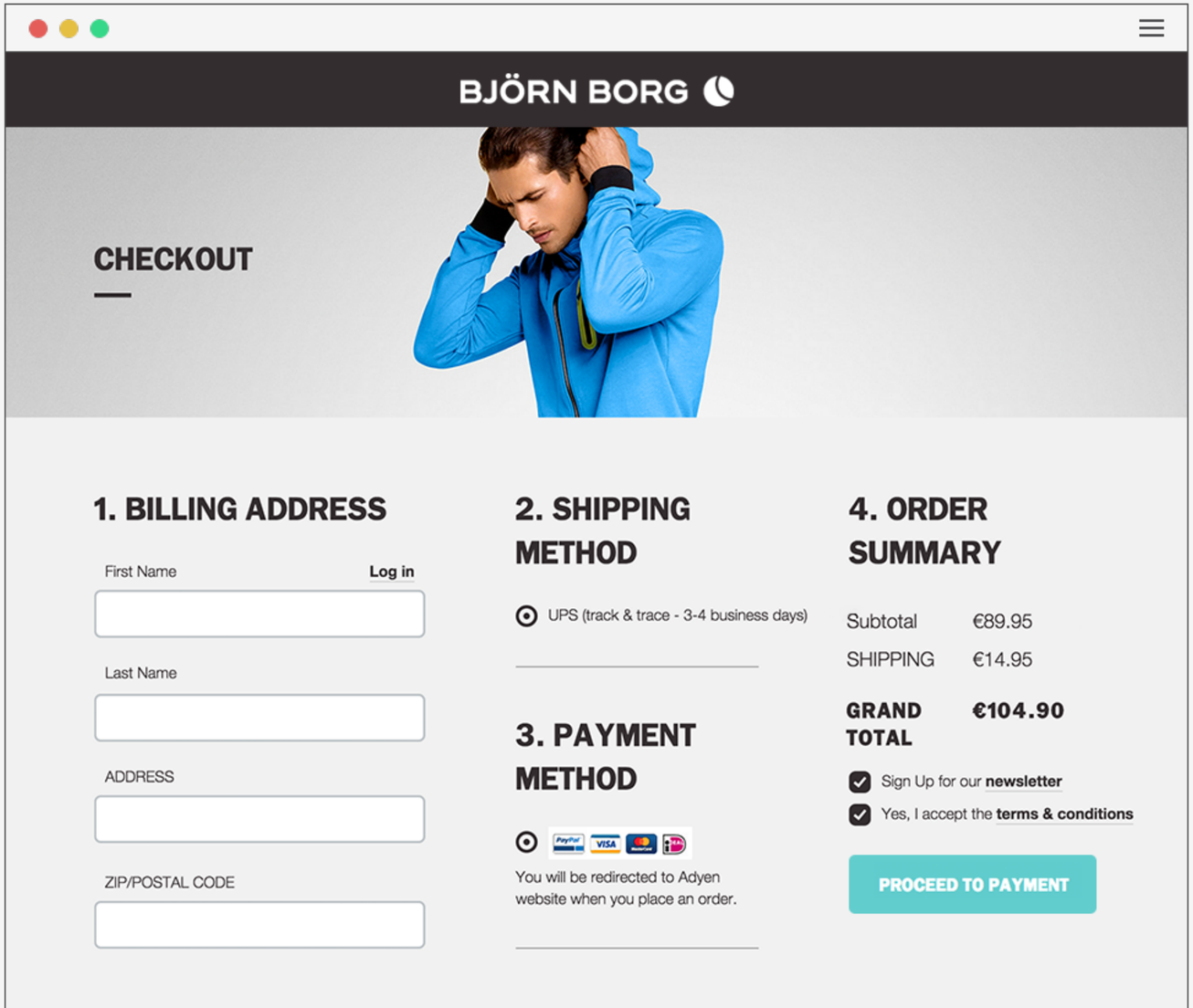


How would you commit fraud?

You will need:

- a lot of different stolen credit cards
- stay anonymous

How would you stop fraud?

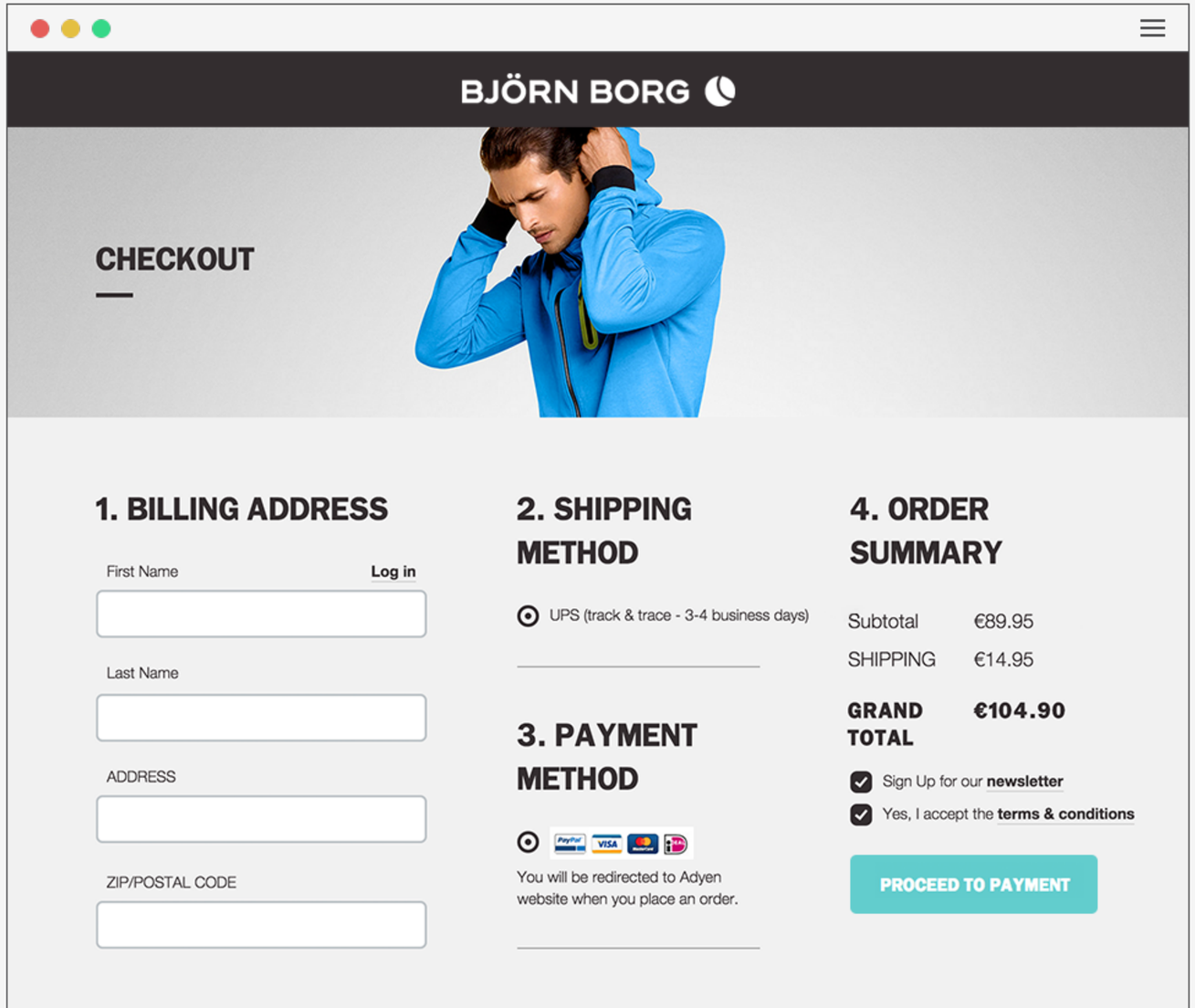


How would you stop fraud?

Ideas:

- detect the use of many different credit cards
- high velocity of details
- proxy IP detection

How would you commit fraud?

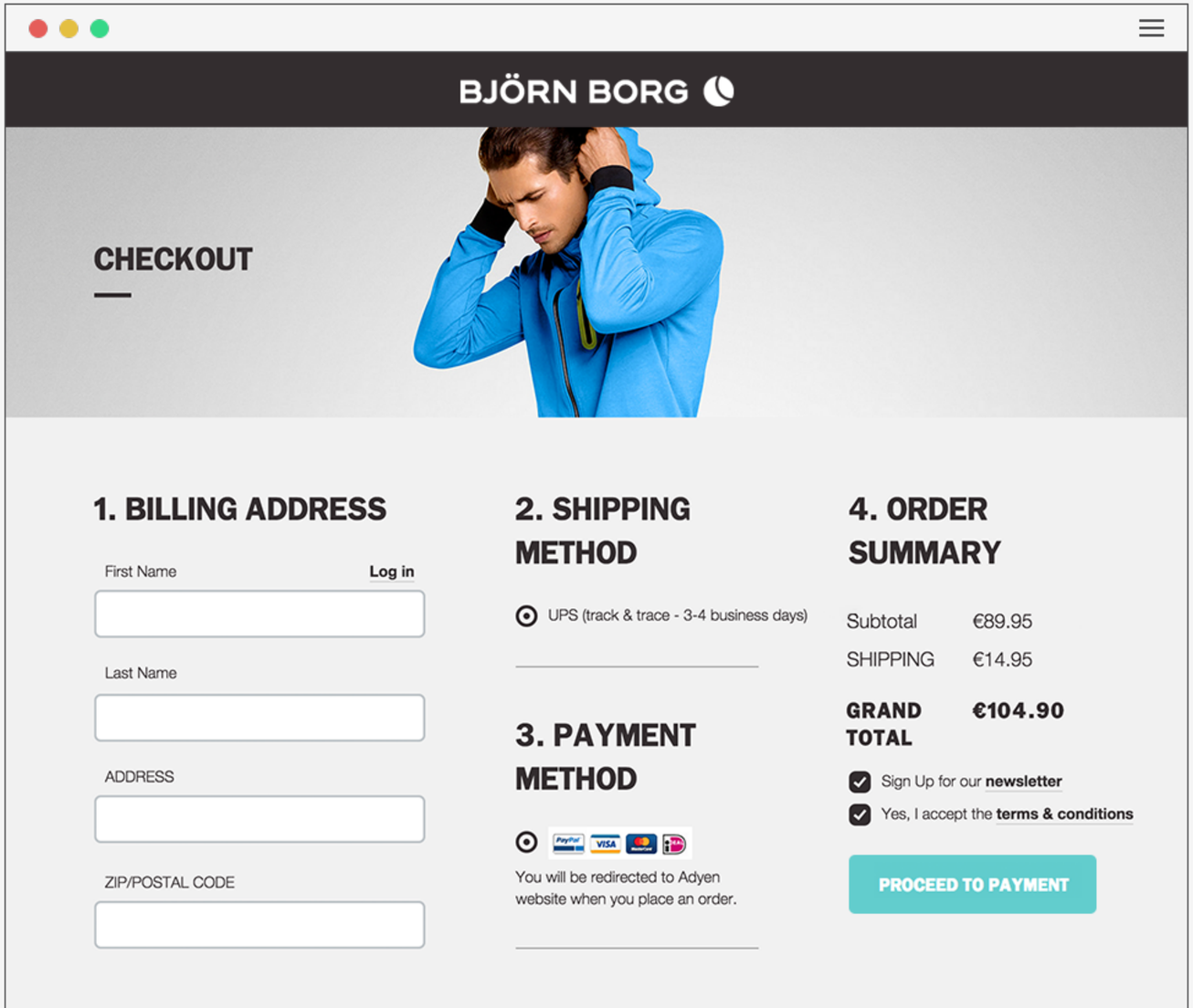


How would you commit fraud?

You will need:

- to change details all the time
- to behave as normal as possible

How would you stop fraud?



How would you stop fraud?

How would you stop fraud?

President Barack Obama's credit card was declined at a restaurant in New York City last month, he said.

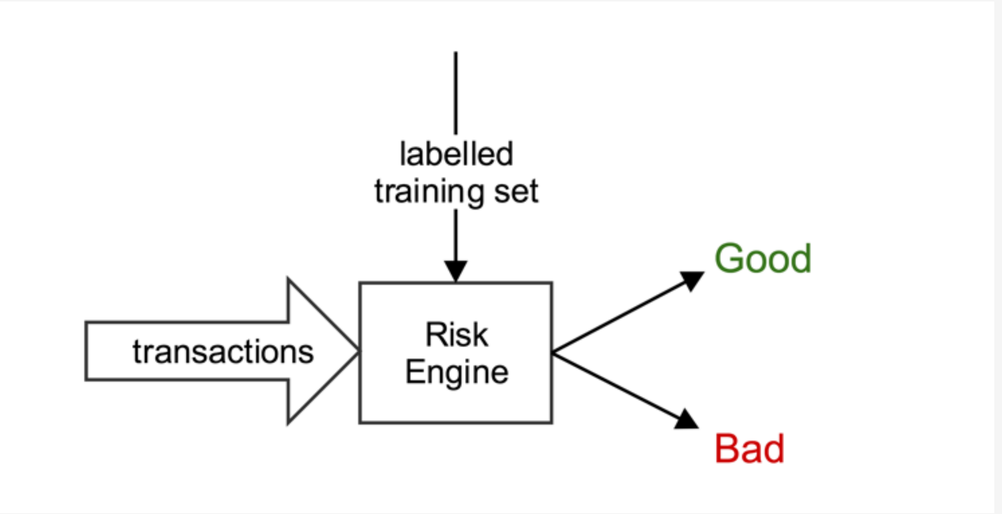
"It turned out I guess I don't use it enough," Mr Obama said. "They thought there was some fraud going on."

Machine Learning

Machine learning

Classic Approach:

- take a huge (labelled) data set
- train a classifier
- deploy the classifier in real time system



Machine learning - challenges

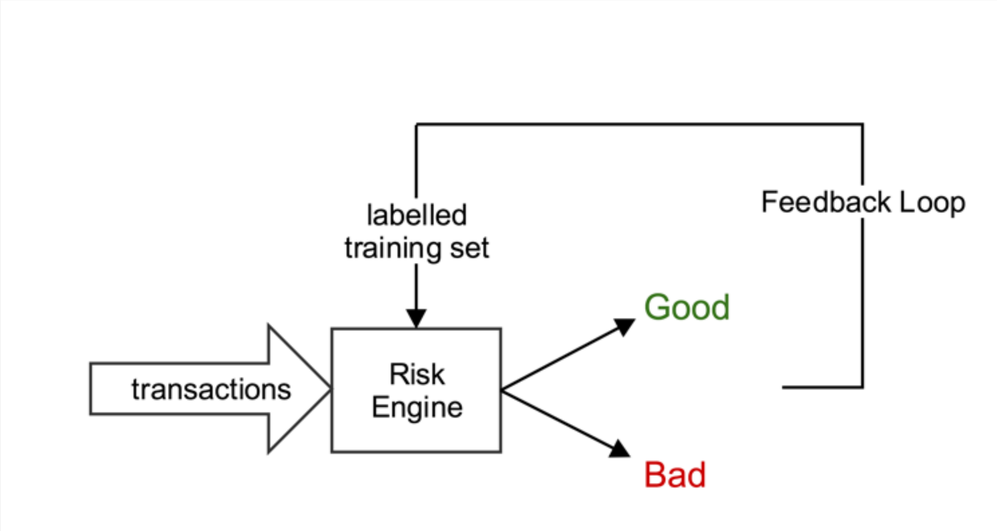
1. Skewed data sets (< 0.3% fraud). The most naive classifier has 99.7% accuracy!

Machine learning - challenges

2. Most data labelled, but with noise:
   - refusals
   - if you stop 100% fraud, you won’t get feedback
   - fraud not always reported

Machine learning - challenges

3. Patterns vary between merchants / markets and change over time

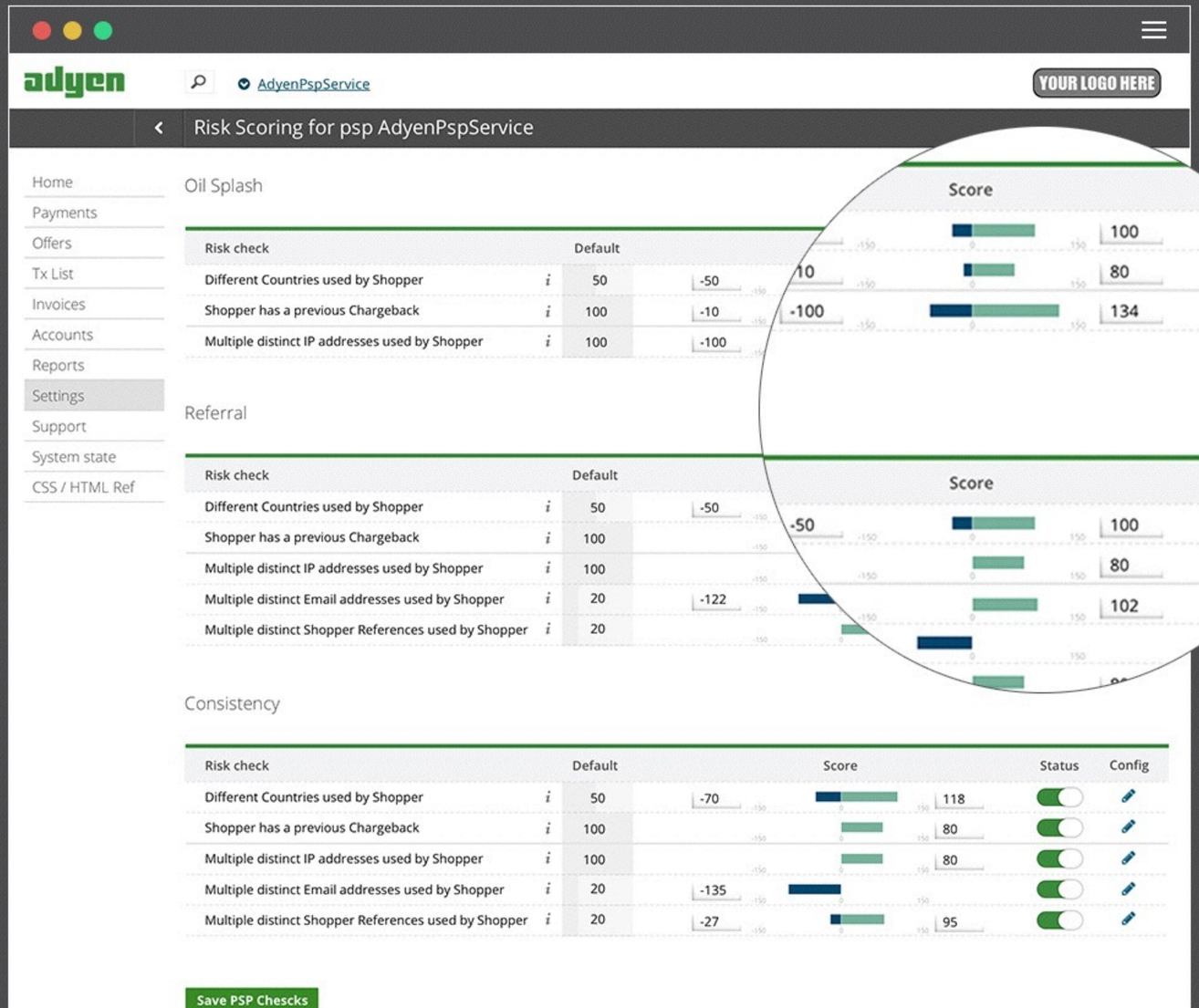


Machine learning - challenges

4. Merchants demand a white-box approach with full control (You need to explain why you stopped Obama!)

White box approach

Give control & Explain



White box approach

Give control & Explain

	Genuine	Refused By Risk	Chargeback
Current Configuration	29,698	3,723	434 (1.44%)
Simulated Configuration	27,827	5,711	317 (1.13%)

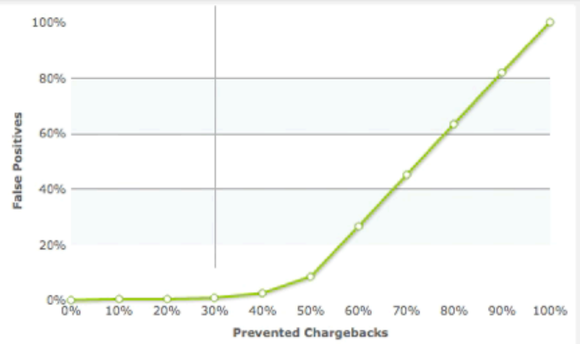
Expected Change in Results using Simulated Configuration

Simulated vs Current Configuration	Diff	Amount (EUR)
Chargebacks	-117	22,500.27
False Positives	+119	-7,837.34

Total		**14,662.93**
New Authorisation Requests	-1,752	-167,088.24

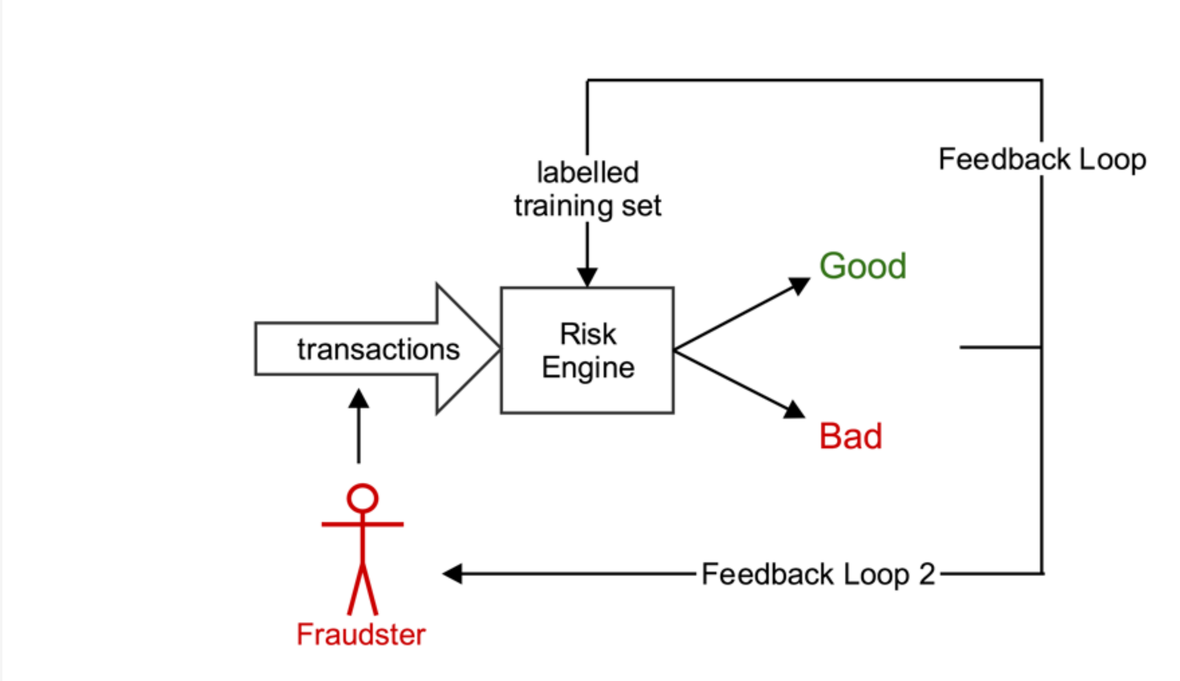
Proposed Risk Configurations

Few False Positives — *Low Chargeback Rates*



Machine learning - challenges

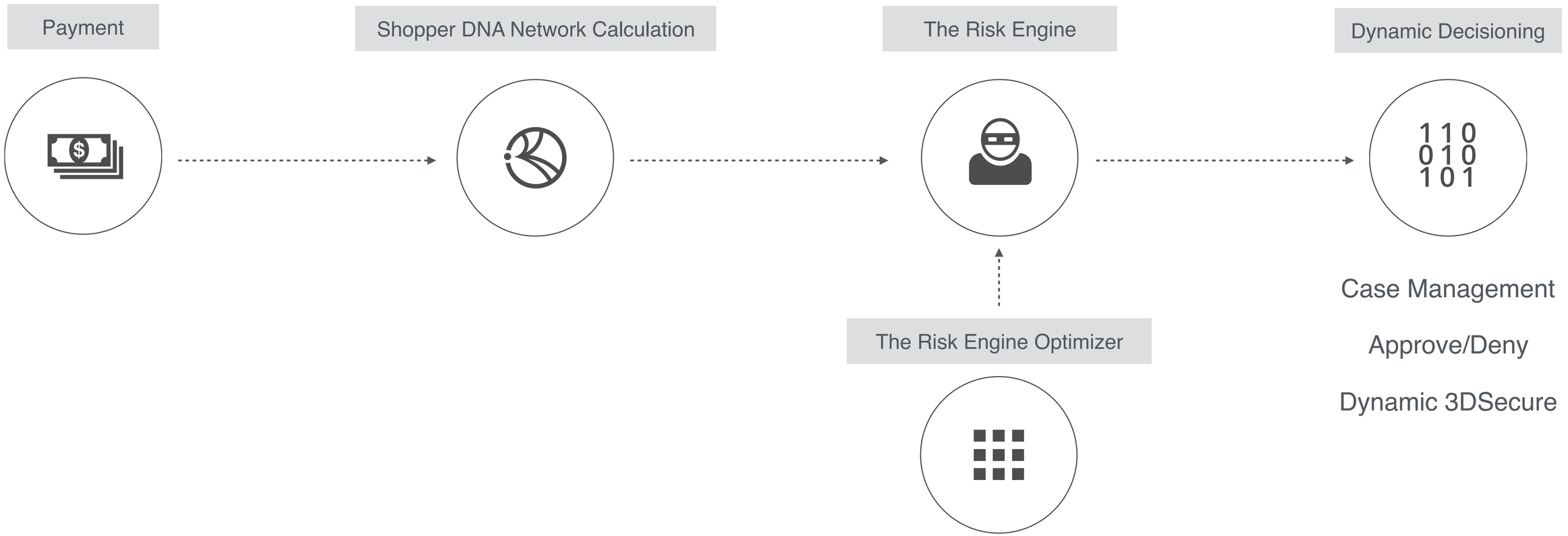
5. Don’t educate fraudsters!



Layered Machine Learning

Adyen RevenueProtect

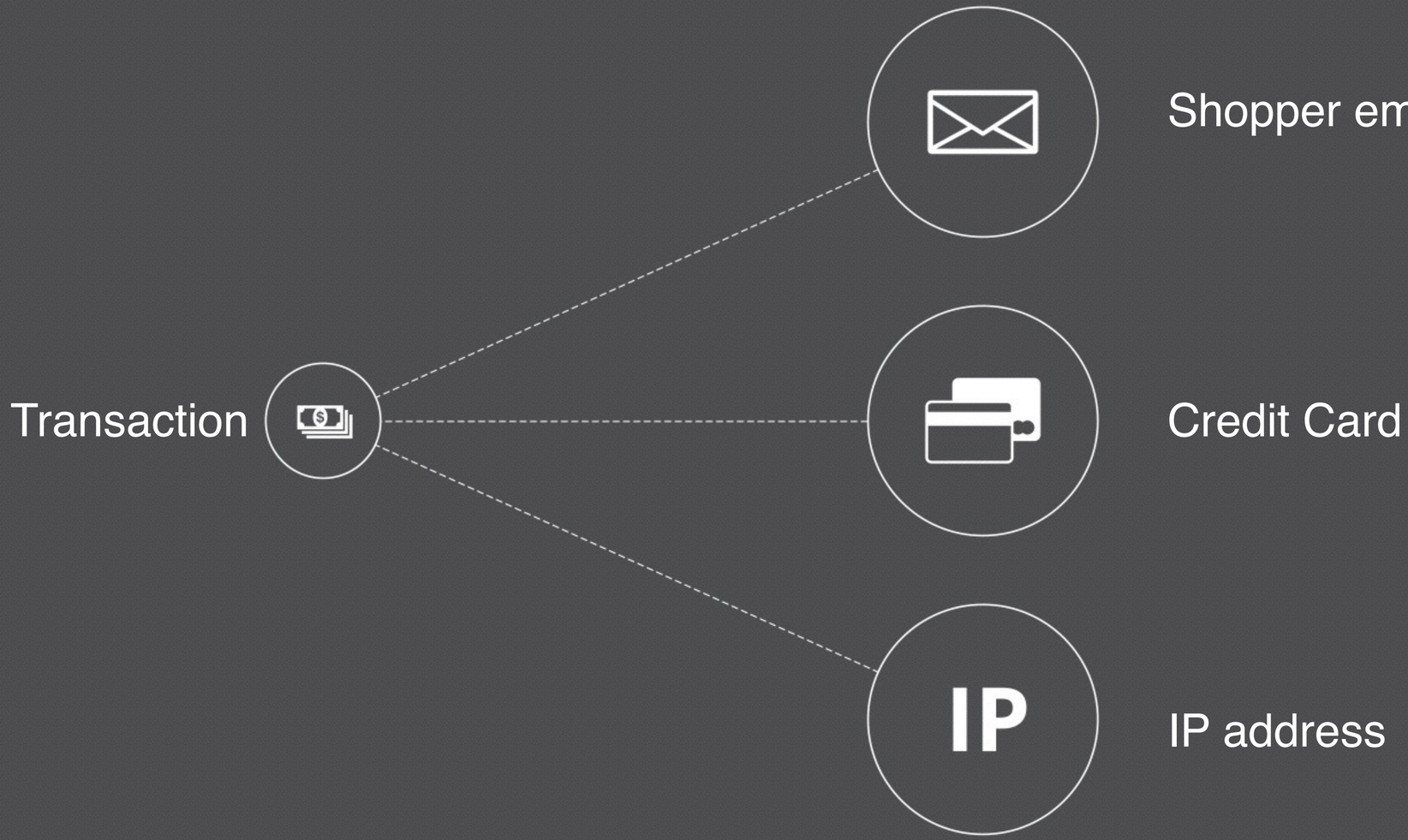
Architecture Overview

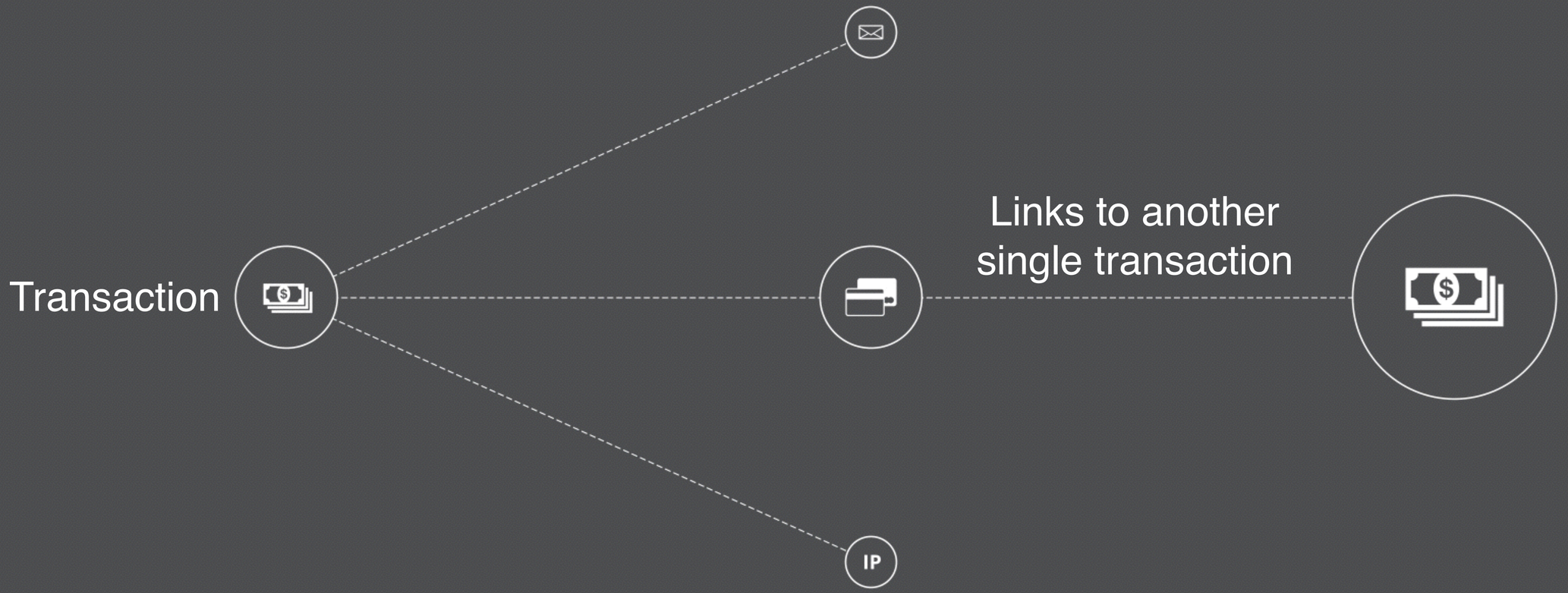


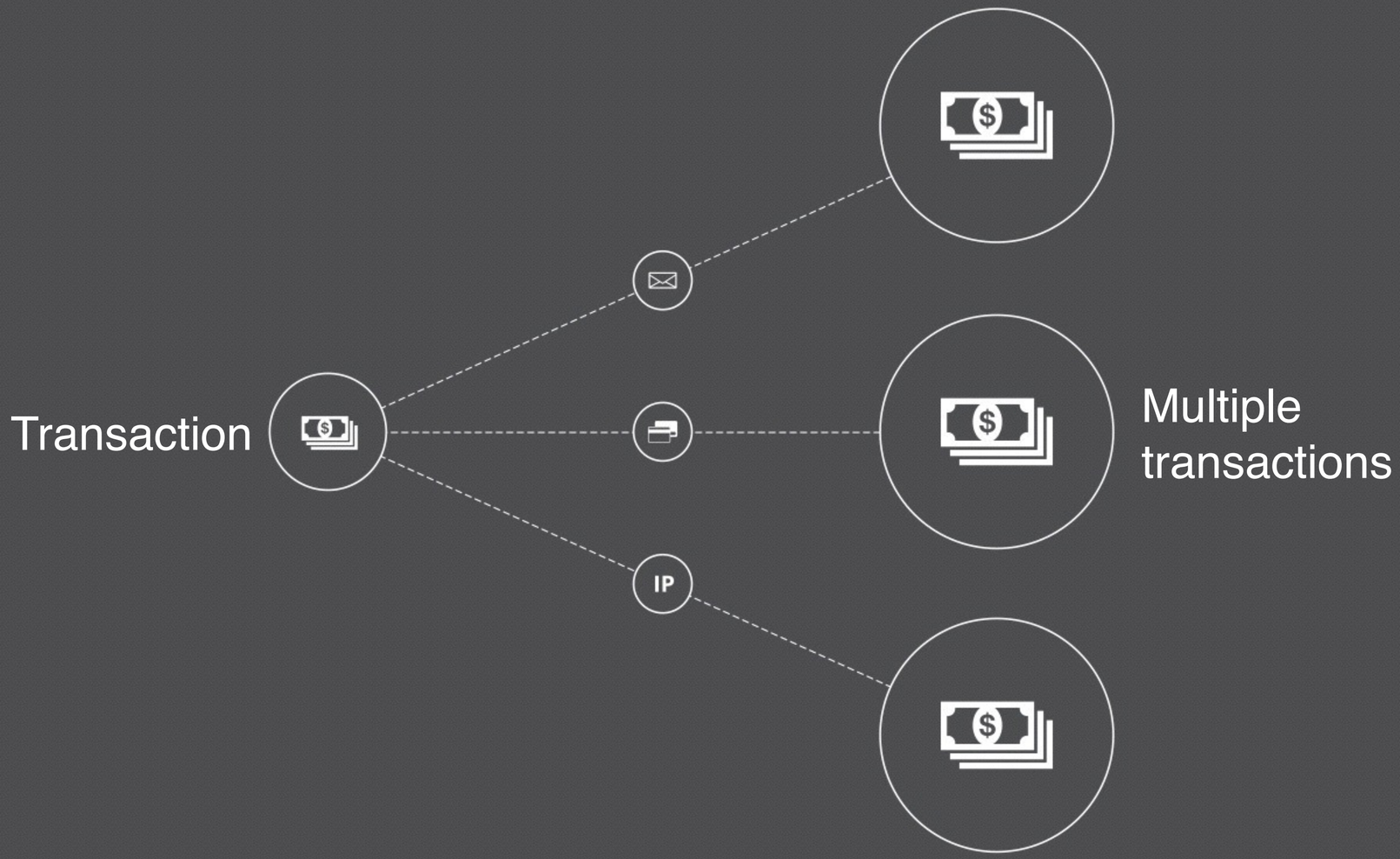
ShopperDNA

From Transaction to Network

A Transaction: The Old Way of Thinking




Transaction
IP
Links to another
single transaction


Transaction
IP
Multiple
transactions

Transaction

Multiple transactions

Oilsplash | Force layout

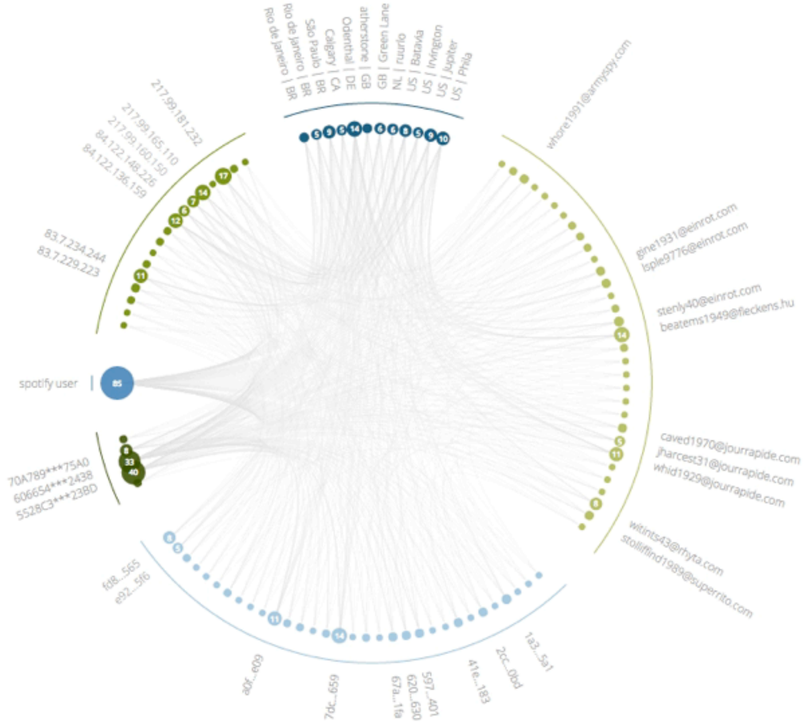
The Shopper

Number of Transactions

Total	**85**	**EUR**	**6479**
Authorised/Settled	11	EUR	897
Refused	10	EUR	597
Cancelled	9	EUR	697
Refunded	8	EUR	300
Chargeback	47	EUR	3988

Number of Identifiers

Card Number	5
Holder Name	1
IP Address	17
Shopper Address	12
Shopper Email	32
Shopper Reference	32



Type to search | Clear selection

Payment								Shopper			
Status	PSP reference	Creation time	Interaction	Method	Currency	Amount	Fraud score	Holder name	Card number	Reference	Address
Refused	4814203960608940	2015-01-04 19:27:40	ecommerce	VISA	EUR	99.00	-10	spotify user	606654***438	7d2920***5ae	DE | Odenthal

Oilsplash | Force layout

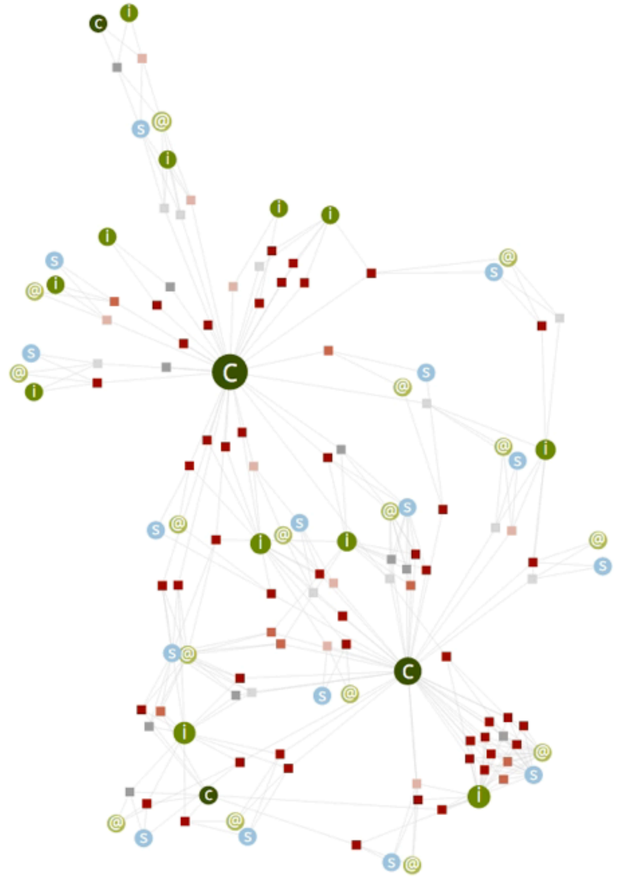
The Shopper

Number of Transactions

Total	**85**	**EUR**	**6479**
Authorised/Settled	11	EUR	897
Refused	10	EUR	597
Cancelled	9	EUR	697
Refunded	8	EUR	300
Chargeback	47	EUR	3988

Number of Identifiers

Card Number	4
Holder Name	0
IP Address	12
Shopper Address	0
Shopper Email	16
Shopper Reference	16

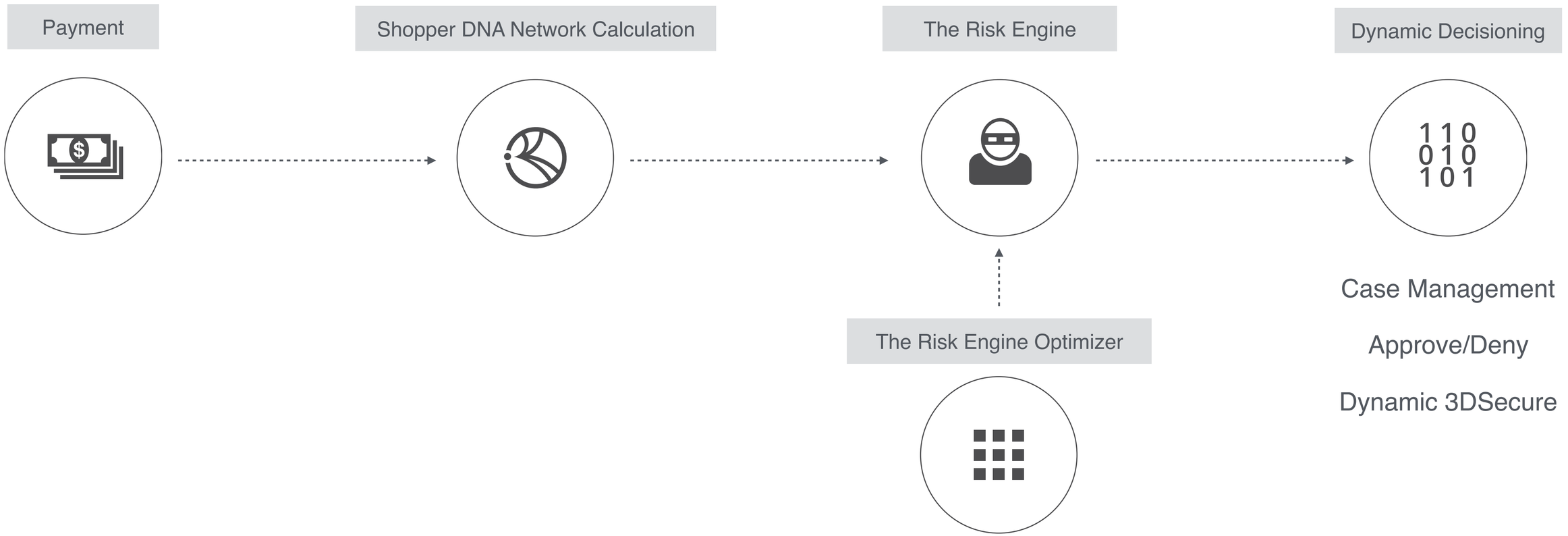


Type to search | Clear selection

	Payment							Shopper			
Status	PSP reference	Creation time	Interaction	Method	Currency	Amount	Fraud score	Holder name	Card number	Reference	Address
Refused	4814203960608940	2015-01-04 19:27:40	ecommerce	VISA	EUR	99.00	-10	spotify user	606654***438	7d2920***5ae	DE | Odenthal

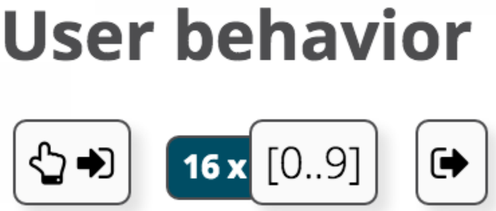
Adyen RevenueProtect

Architecture Overview

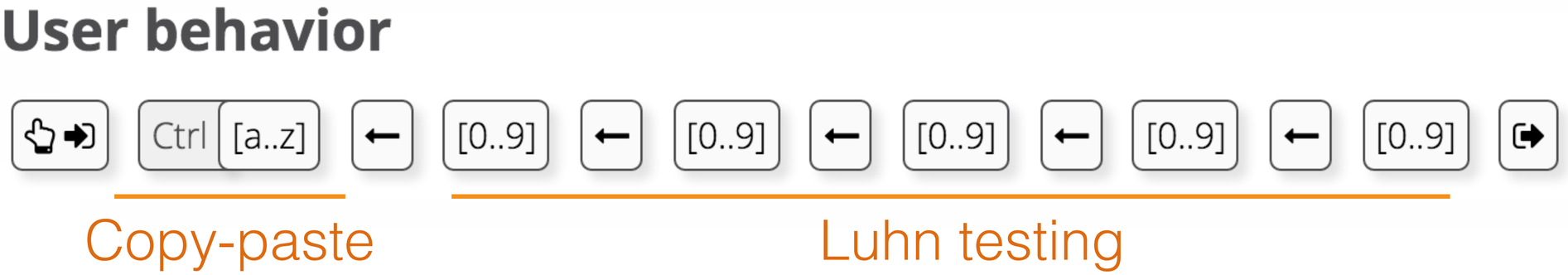


Interpreting Shopper Behavior

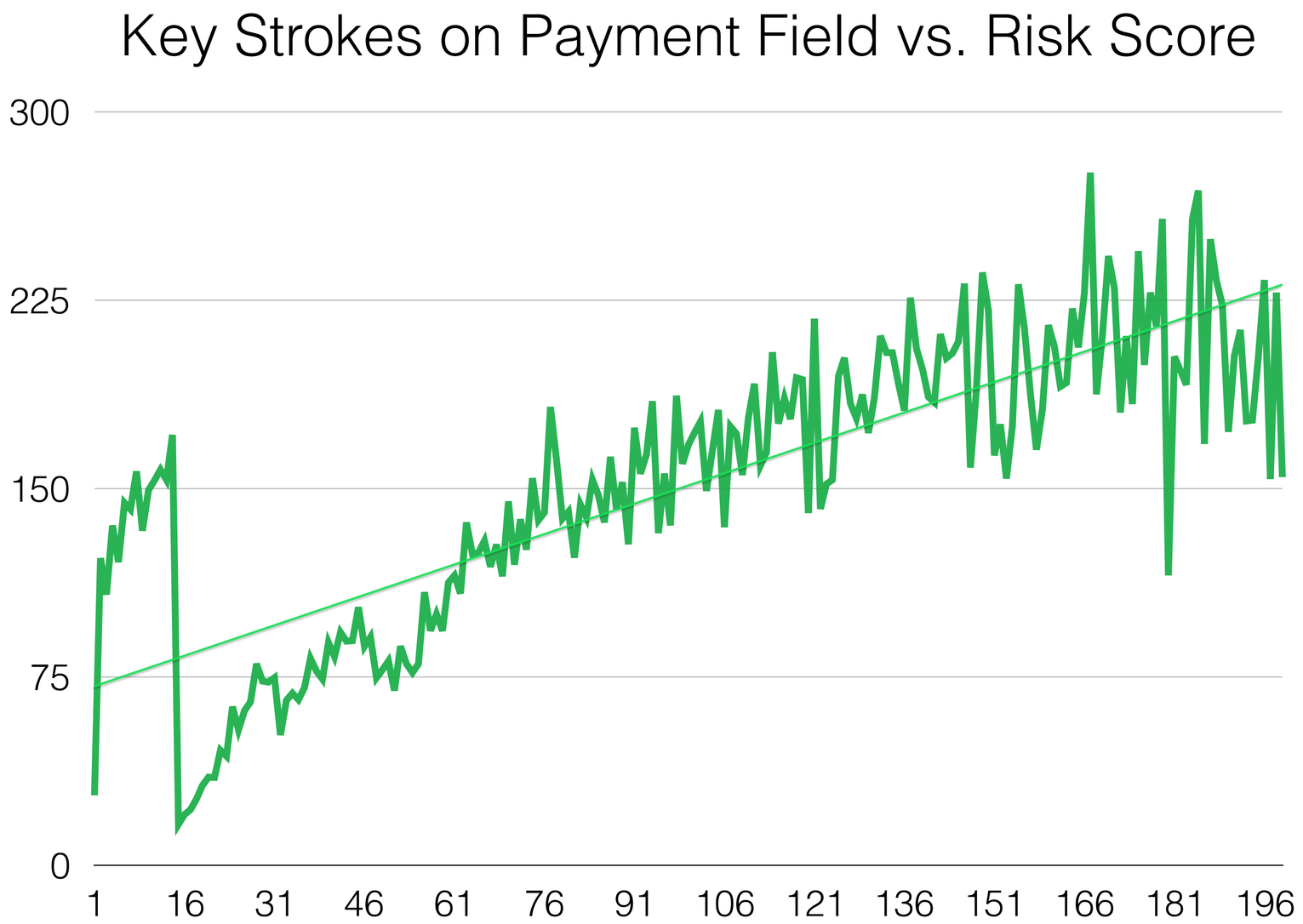
Expected:



Fraud:



Interpreting Shopper Behavior



Adyen enables companies to accept payments from their customers using any payment method over any sales channel.

Fraud Assignment

adyen wherever people pay

Your task this week

- Analyse the fraud data (about half million real records!)
- Apply your favourite ML tool
  - use at least one of the methods from today!
  - read and use the relevant cited papers from the survey
- Evaluate the obtained model(s)
- Write the performed steps on max. two A4s
  - figures in appendix
- The top 3 reports will get a price from Adyen!
- Deadline in two weeks, before lecture

Your task this week

- Analyse t
- Apply y
  - use at
  - read a
- Evaluat
- Write t
  - figures
- The top
- Deadlir

ey

3 key difficulties:

The detection game
Understandable (white-box) models
Imbalanced data

Reflect on these and use special evaluation metrics!

Credit card transaction data

- Every row is one transaction

09/11/15 14:26	MX	mccredit	530056	64800.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	01/07/15 23:03	Mexico Account	email68370	ip111778	card184798
09/11/15 14:27	MX	mccredit	547046	44900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	02/07/15 04:50	Mexico Account	email101299	ip78749	card151595
23/11/15 16:34	MX	mccredit	528843	149900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	02/07/15 14:30	Mexico Account	email278604	ip70594	card242142
23/11/15 16:34	MX	mccredit	547146	109900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	03/07/15 07:53	Mexico Account	email47409	ip113648	card181744
09/11/15 14:26	MX	visaclassic	477291	89900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	08/07/15 18:35	Mexico Account	email205501	ip83553	card97271

Credit card transaction data

- Every row is one transaction

09/11/15 14:26	MX	mccredit	530056	64800.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	01/07/15 23:03	Mexico Account	email68370	ip111778	card184798
09/11/15 14:27	MX	mccredit	547046	44900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	02/07/15 04:50	Mexico Account	email101299	ip78749	card151595
23/11/15 16:34	MX	mccredit	528843	149900.0	MXN	MX	Ecommerce	Chargeback	TRUE	0	02/07/15 14:30	Mexico Account	email278604	ip70594	card242142
23/11/15 16:34	MX	mccredit	547146	109900.0	MXN	MX						co unt	email47409	ip113648	card181744
09/11/15 14:26	MX	visaclassic	477291	89900.0	MXN	MX						co unt	email205501	ip83553	card97271

Which transactions are fraudulent?

Aggregated transaction data - daily

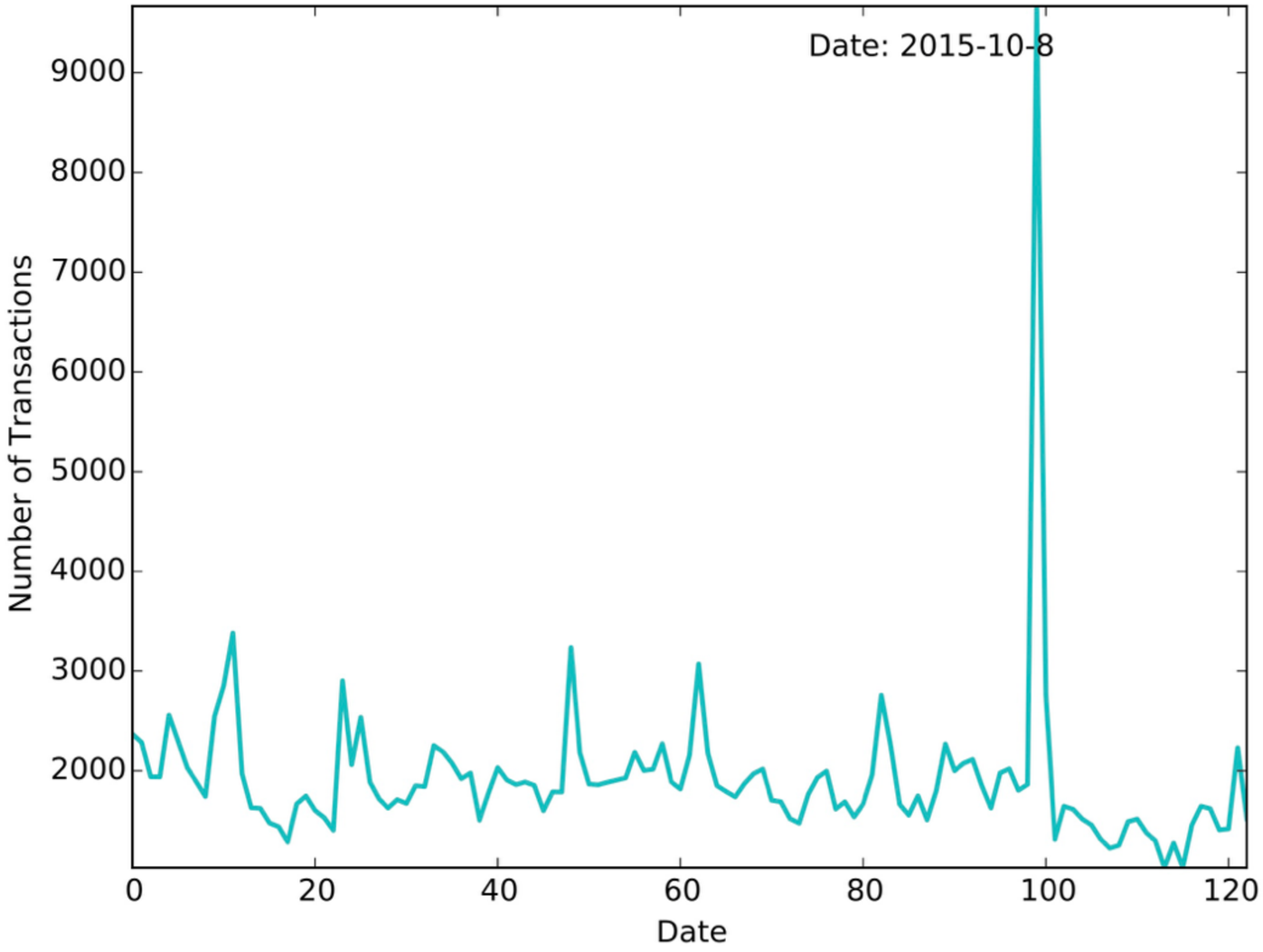
- Sum values per day

year	month	day	sum	nr fraud	sum ref	nr ref	sum OK	nr OK
2015	August	1	382900.0	3	2,79E+14	369	3,33E+13	1856
2015	August	2	316000.0	2	2,30E+14	341	3,47E+14	1843
2015	August	3	625440.0	5	3,09E+13	458	3,93E+13	2251
2015	August	4	248650.0	3	3,25E+14	400	3,66E+14	2192
2015	August	5	214600.0	3	2,39E+14	395	3,64E+14	2080
2015	August	6	262500.0	5	3,32E+14	457	3,43E+14	1924
2015	August	7	?	?	3,75E+14	402	3,59E+14	1985
2015	August	8	236700.0	5	1,76E+14	295	2,55E+12	1505
2015	August	9	16000.0	1	1,54E+14	291	2,66E+14	1782
2015	August	10	447280.0	5	1,40E+14	334	3,48E+14	2038
2015	August	11	52175.0	3	2,83E+14	411	3,34E+14	1911

Aggregated transaction data - daily

- Sum values per c

year	month	day	sum
2015	August	1	382900
2015	August	2	316000
2015	August	3	625440
2015	August	4	248650
2015	August	5	214600
2015	August	6	262500
2015	August	7	?
2015	August	8	236700
2015	August	9	16000
2015	August	10	447280
2015	August	11	52175

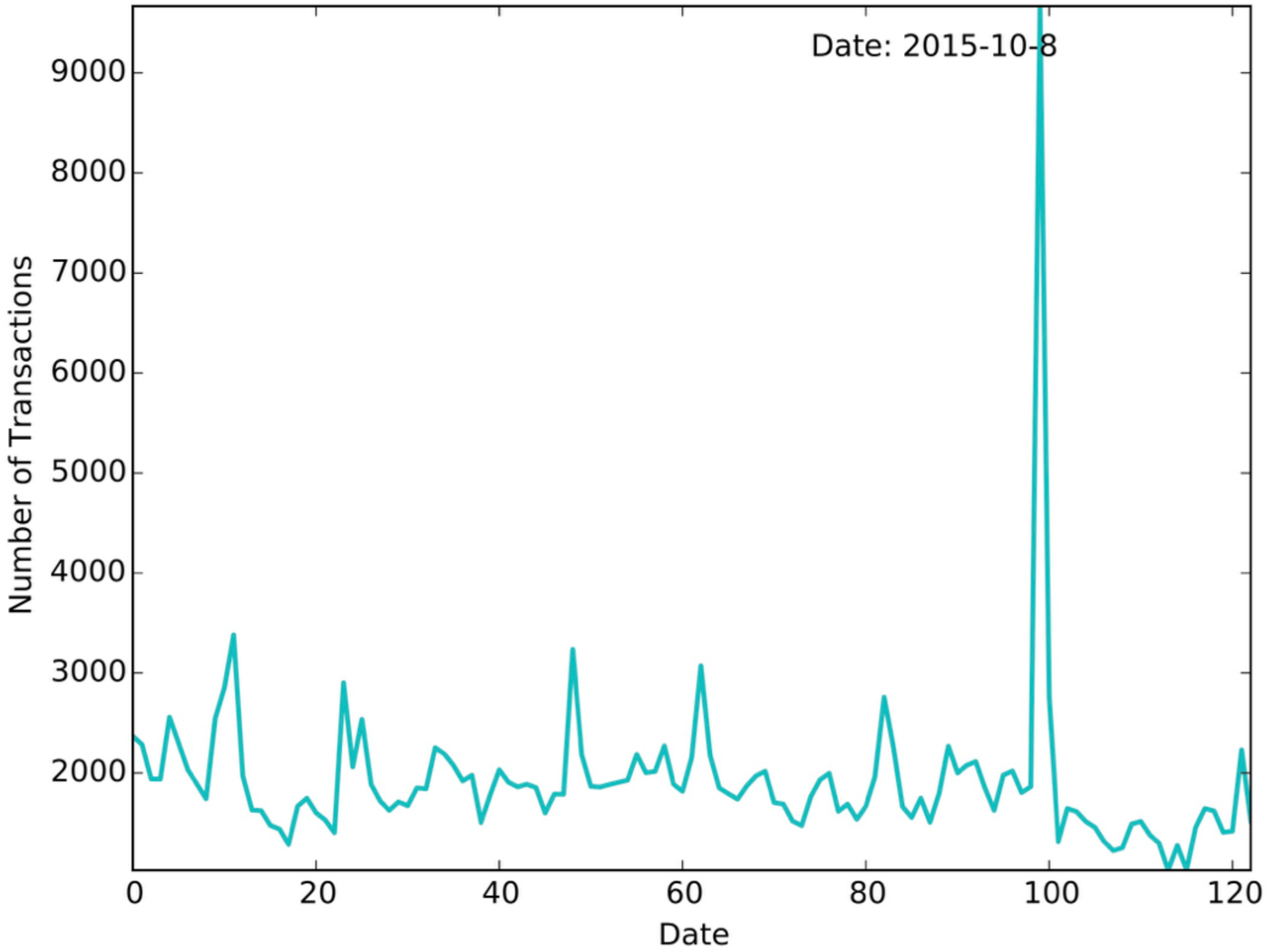


Aggregated transaction data - daily

- Sum values per c

year	month	day	sum
2015	August	1	382900
2015	August	2	316000
2015	August	3	625440
2015	August	4	248650
2015	August	5	214600
2015	August	6	262500
			?
			36700
			6000
			47280
2015	August	11	52175

Which days are high risk?

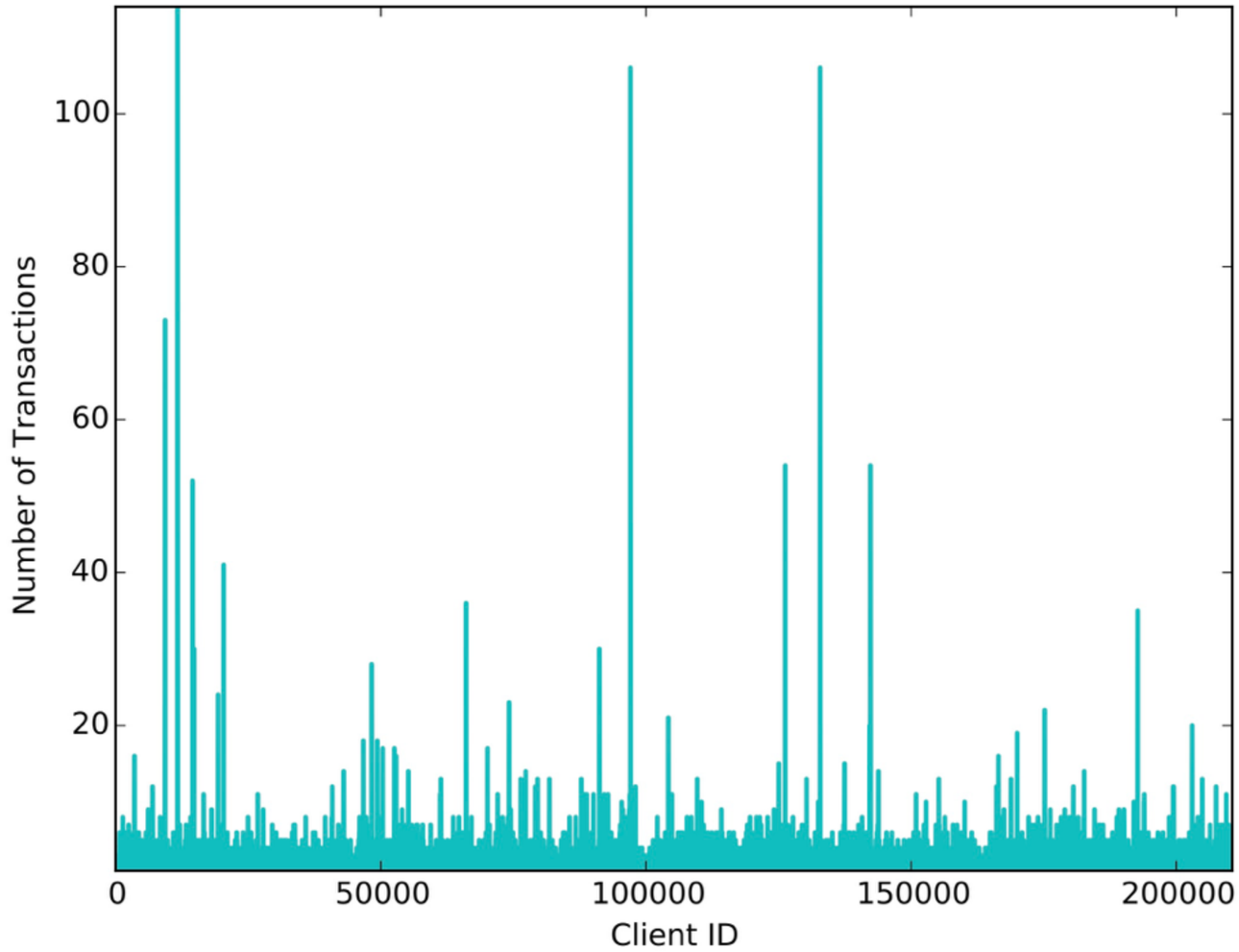


Aggregated transaction data - per card

	sum OK	nr OK	sum ref	nr ref	sum fraud	nr fraud
card13406	?	?	?	?	16000.0	1
card14719	?	?	?	?	16000.0	1
card15045	?	?	?	?	16000.0	1
card19731	?	?	?	?	16000.0	1
card73227	?	?	?	?	16000.0	1
card11088	?	?	?	?	15350.0	1
card25850	?	?	?	?	15300.0	1
card22996	?	?	?	?	15240.0	1
card14241	?	?	?	?	15000.0	1
card44926	?	?	51500.0	1	15000.0	1
card16444	?	?	?	?	14850.0	1
card21460	9300.0	2	32795.0	4	14000.0	2
card22317	?	?	?	?	13850.0	1

Aggregated transaction data - per card

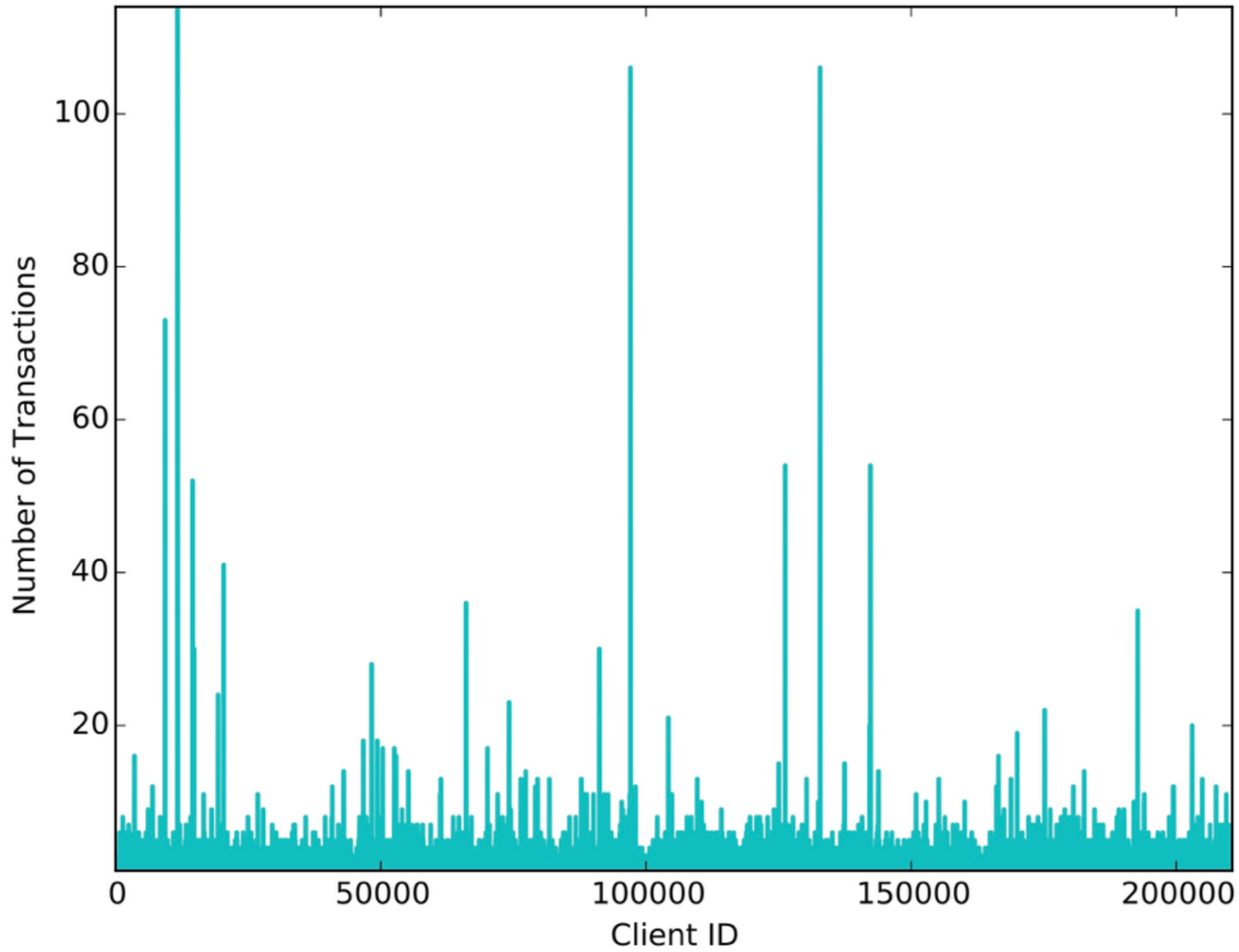
	sum OK	nr OK	sum ref	nr ref	sum fraud	nr fraud
card13406	?	?				
card14719	?	?				
card15045	?	?				
card19731	?	?				
card73227	?	?				
card11088	?	?				
card25850	?	?				
card22996	?	?				
card14241	?	?				
card44926	?	?				
card16444	?	?				
card21460	9300.0			2		
card22317	?	?				



Aggregated transaction data - per card

	sum OK	nr OK	sum ref	nr ref	sum fraud	nr fraud
card13406	?	?				
card14719	?	?				
card15045	?	?				
card19731	?	?				
card73227	?	?				
card11088	?	?				
card25850	?	?				
card22996	?	?				

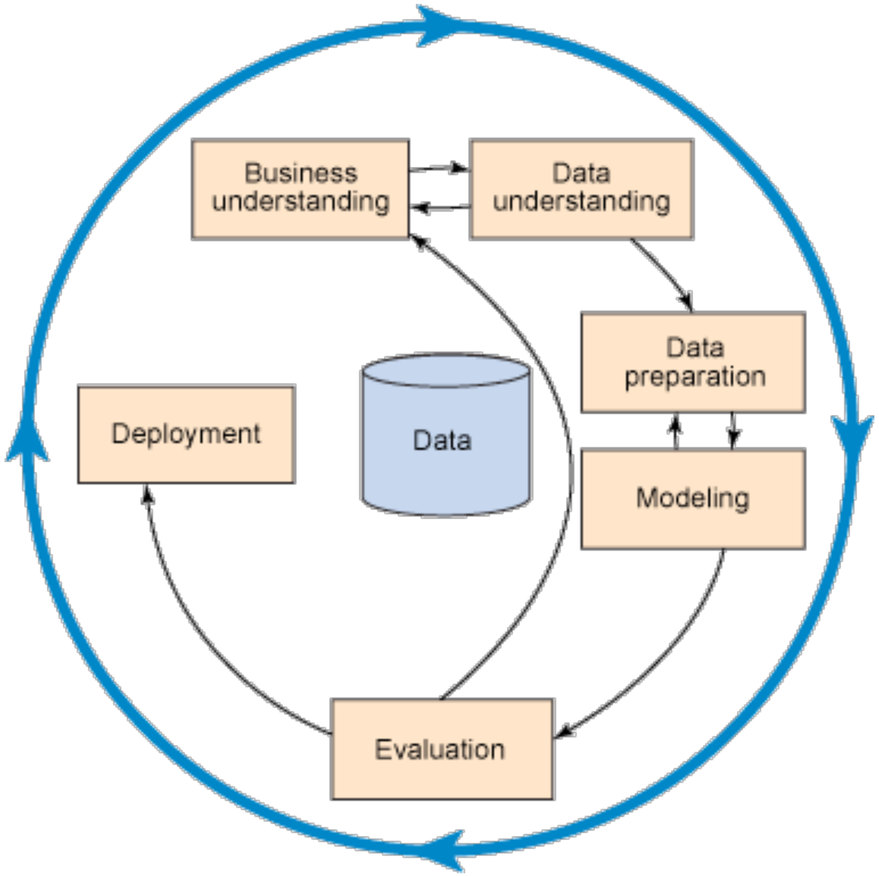
Which clients are fraudulent?



Many different views

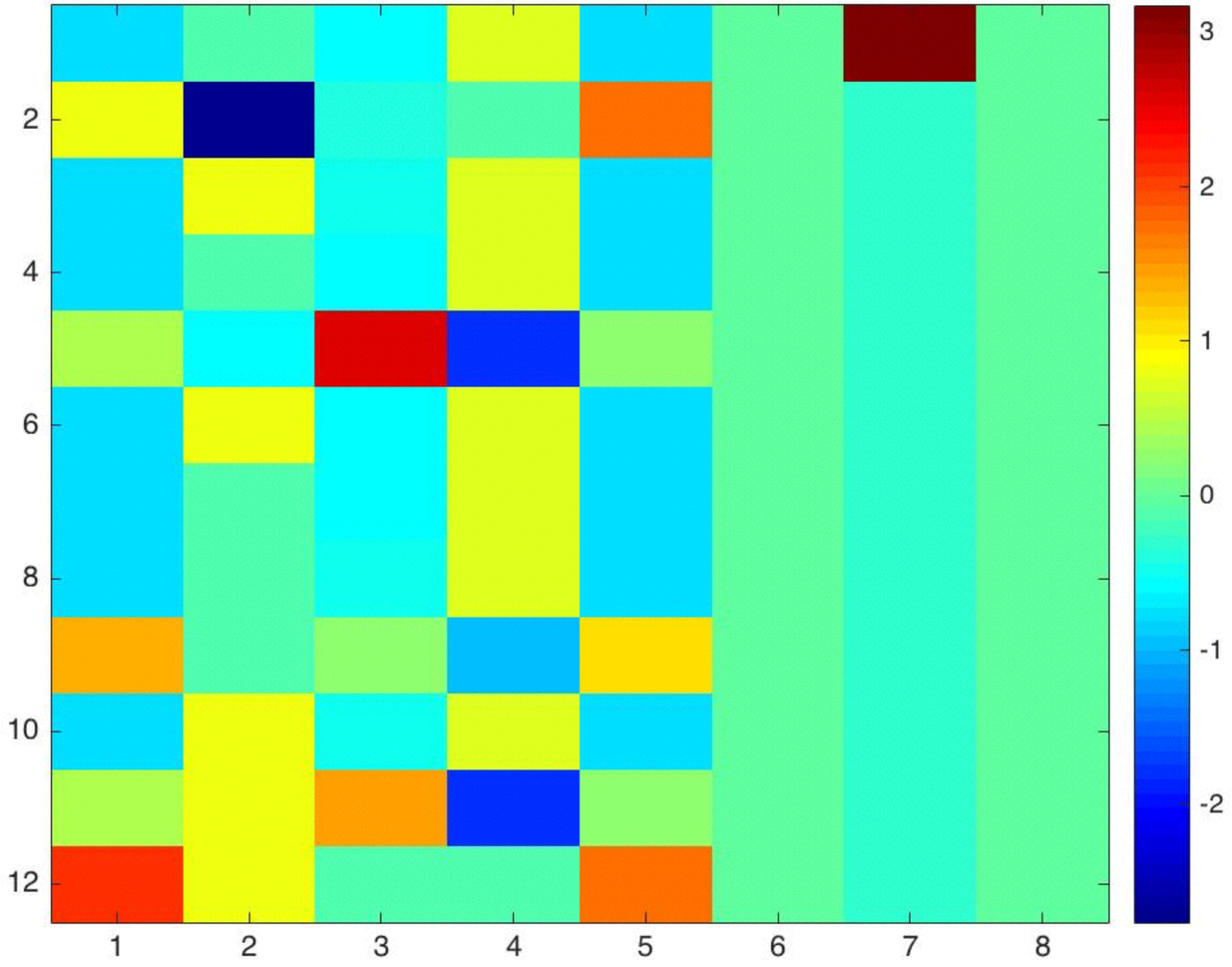
- Data can be aggregated in many different ways
  - understand the goal!
  - understand the data!
  - ***before modelling!***

- Key tool: visualisation
  - heat maps
  - cross tables
  - pair-wise scatterplots
  - parallel coordinates
  - …

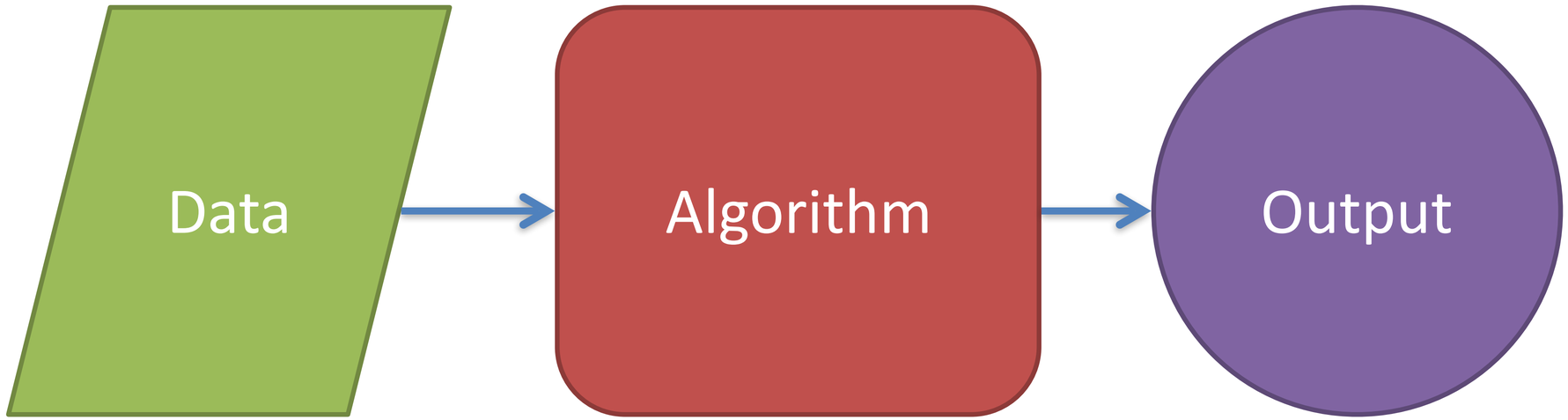


CRISP-DM

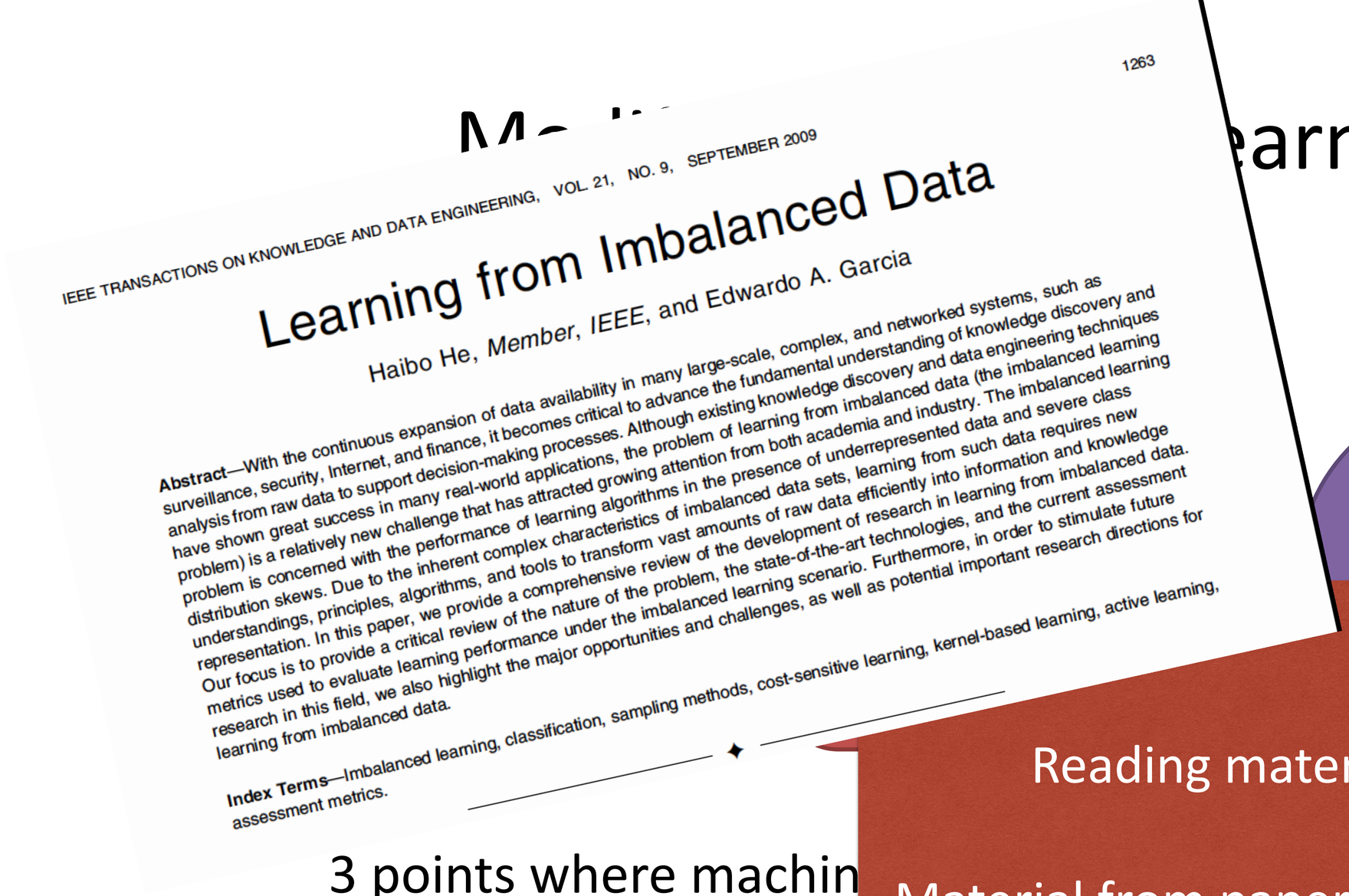
A heatmap of transactions



Modifying machine learning



3 points where machine learning can be modified



IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009

Learning from Imbalanced Data

Haibo He, *Member, IEEE*, and Edwardo A. Garcia

Abstract—With the continuous expansion of data availability in many large-scale, complex, and networked systems, such as surveillance, security, Internet, and finance, it becomes critical to advance the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation. In this paper, we provide a comprehensive review of the development of research in learning from imbalanced data. Our focus is to provide a critical review of the nature of the problem, the state-of-the-art technologies, and the current assessment metrics used to evaluate learning performance under the imbalanced learning scenario. Furthermore, in order to stimulate future research in this field, we also highlight the major opportunities and challenges, as well as potential important research directions for learning from imbalanced data.

Index Terms—Imbalanced learning, classification, sampling methods, cost-sensitive learning, kernel-based learning, active learning, assessment metrics.

Ma… earning

Output

3 points where machin…

Reading material for this week

Material from paper will be tested in exam

Cost-sensitive threshold modification

- Define a cost for false positives and false negatives

- When learning rankers, such as Naïve Bayes, set a decision threshold to
  - optimize cost
  - or obtain a predefined false positive rate

Modifying unbalanced input data

- Resampling
  - oversample minority class
  - undersample majority class
- Reweighting
  - assign large weights to minority class
  - assign small weights to majority class
- Adding
  - add artificial minority class instances

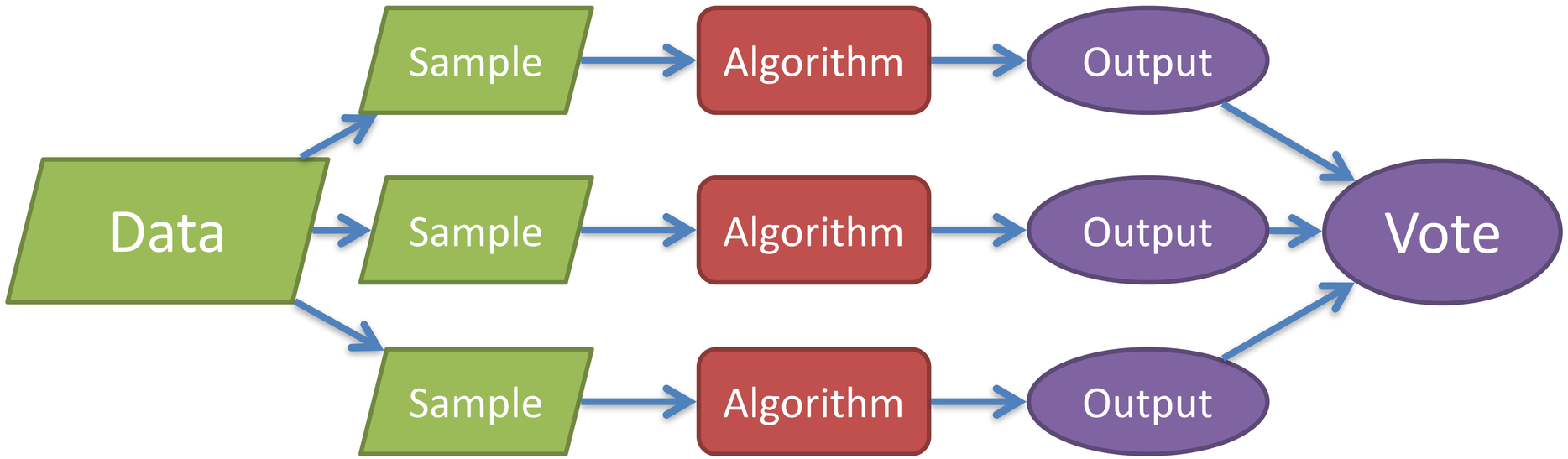
Oversampling

- Issues:
  - overfitting: the sampled examples provide false evidence of a positive label at a very specific point, possibly an anomaly
  - ignorance: some classifiers ignore multiple copies

Undersampling

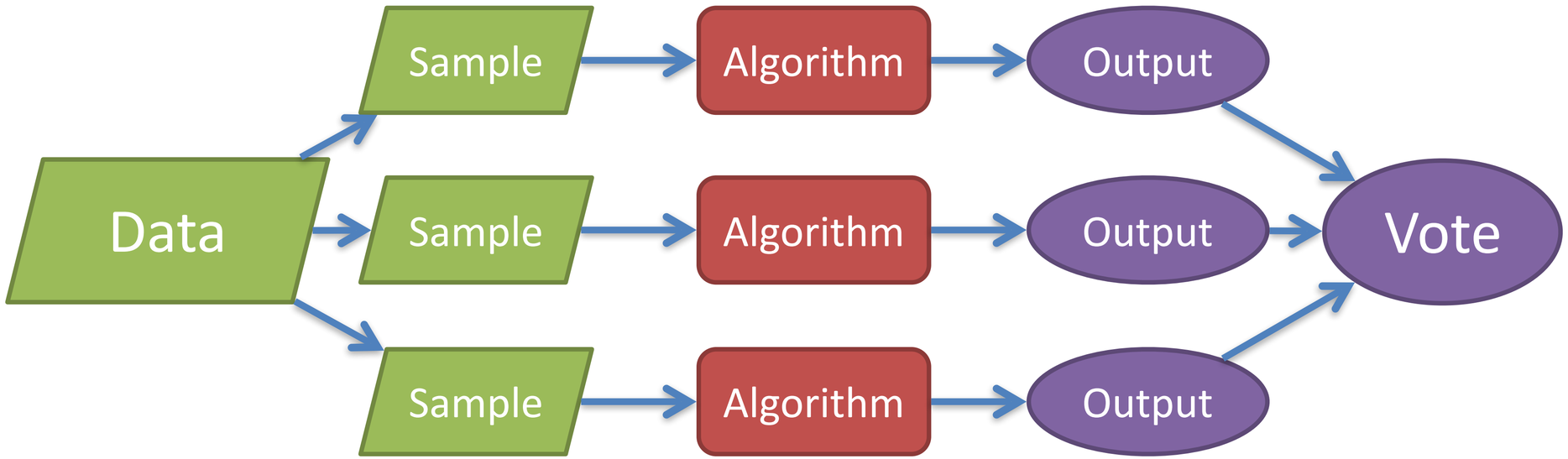
- Issues:
  - information loss: some examples can be essential for good performance!

Bagging



- Learn one model by sampling with replacement and combining results by majority voting

Imbalanced Bagging

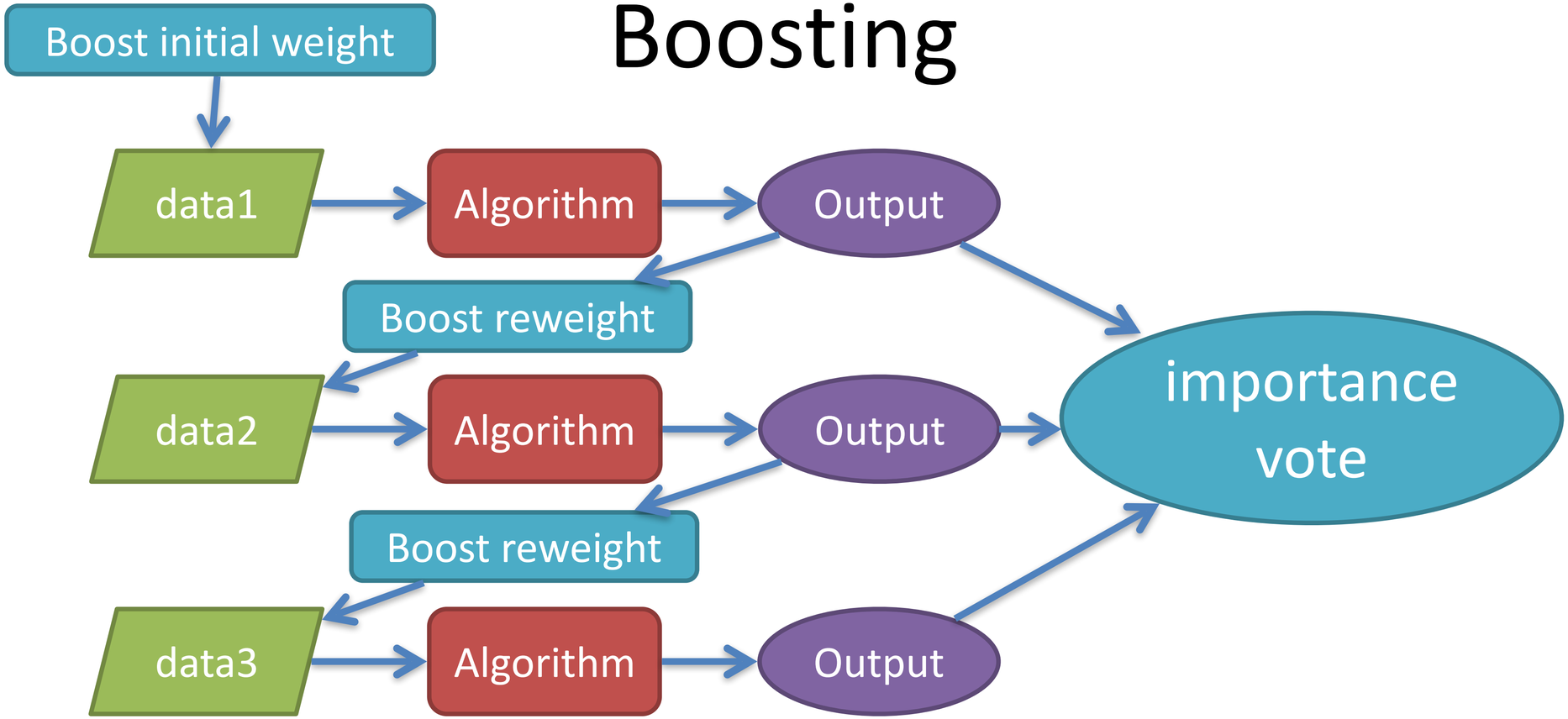


- Imbalanced sampling:
  - Sample all minority instances
  - Sample with replacement from majority instances

kNN-based sampling

- Sample majority instances that:
  1. are on average close to minority instances
  2. are closest to the furthest away monitory instances
  3. are close for each monitory instance
  4. …

Boosting



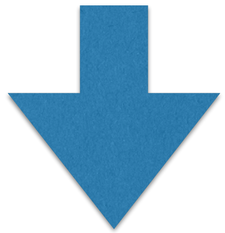
- Model importance determined by: $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
- Initialize weights: $D_1(i) = 1/m$
- Assign weights:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

Imbalanced Boosting

- Reweigh examples depending on cost:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$



$$D_{t+1}(i) = D_t(i) \exp(-\alpha_t C_i h_t(\boldsymbol{x_i}) y_i) / Z_t,$$

$$D_{t+1}(i) = C_i D_t(i) \exp(-\alpha_t h_t(\boldsymbol{x_i}) y_i) / Z_t,$$

$$D_{t+1}(i) = C_i D_t(i) \exp(-\alpha_t C_i h_t(\boldsymbol{x_i}) y_i) / Z_t.$$

Tips on bagging and boosting

- Bagging mainly reduced variance
  - use complex (overfitting) models
- Boosting mainly reduces bias
  - use simple “weak” (underfitting) models
- Beware of the black-box nature of ensembles
  - Can you still explain the outcome to a client?

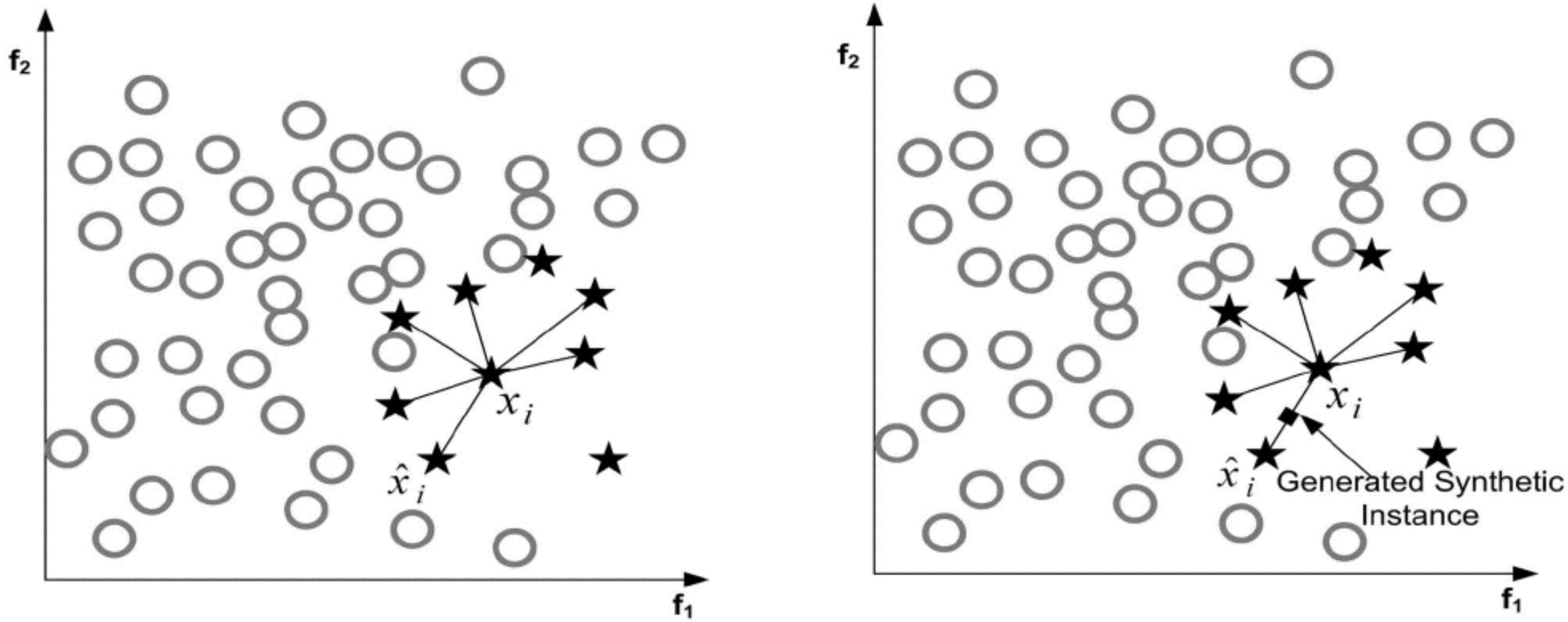
Adding synthetic minority instances

- Key technique: SMOTE
  - Synthetic Minority Oversampling Technique

1. For every minority instance i
2. Randomly select one of the the k nearest neighbors x
3. Compute the vector v(i,x) between i and x:
   - $i + v(i,x) = x$
4. Randomly select a point p along this vector
   - $p = i + rand(0,1) * v(i,x)$
5. Add p to the minority instances

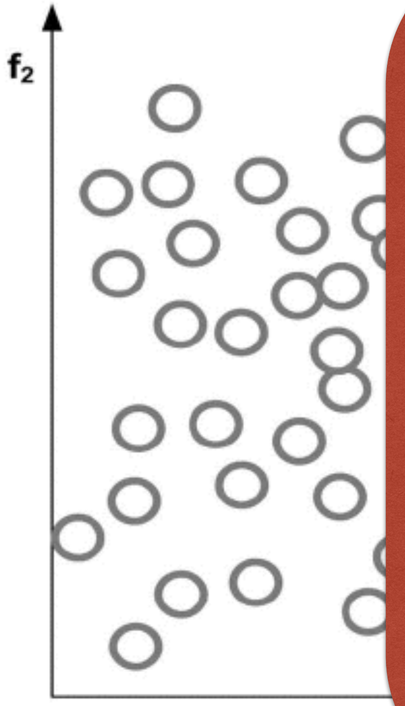
Adding synthetic minority instances

- Key technique: SMOTE
  - Synthetic Minority Oversampling Technique



Adding synthetic minority instances

- Key technique: SMOTE
  - Synthetic Minority Oversampling Technique

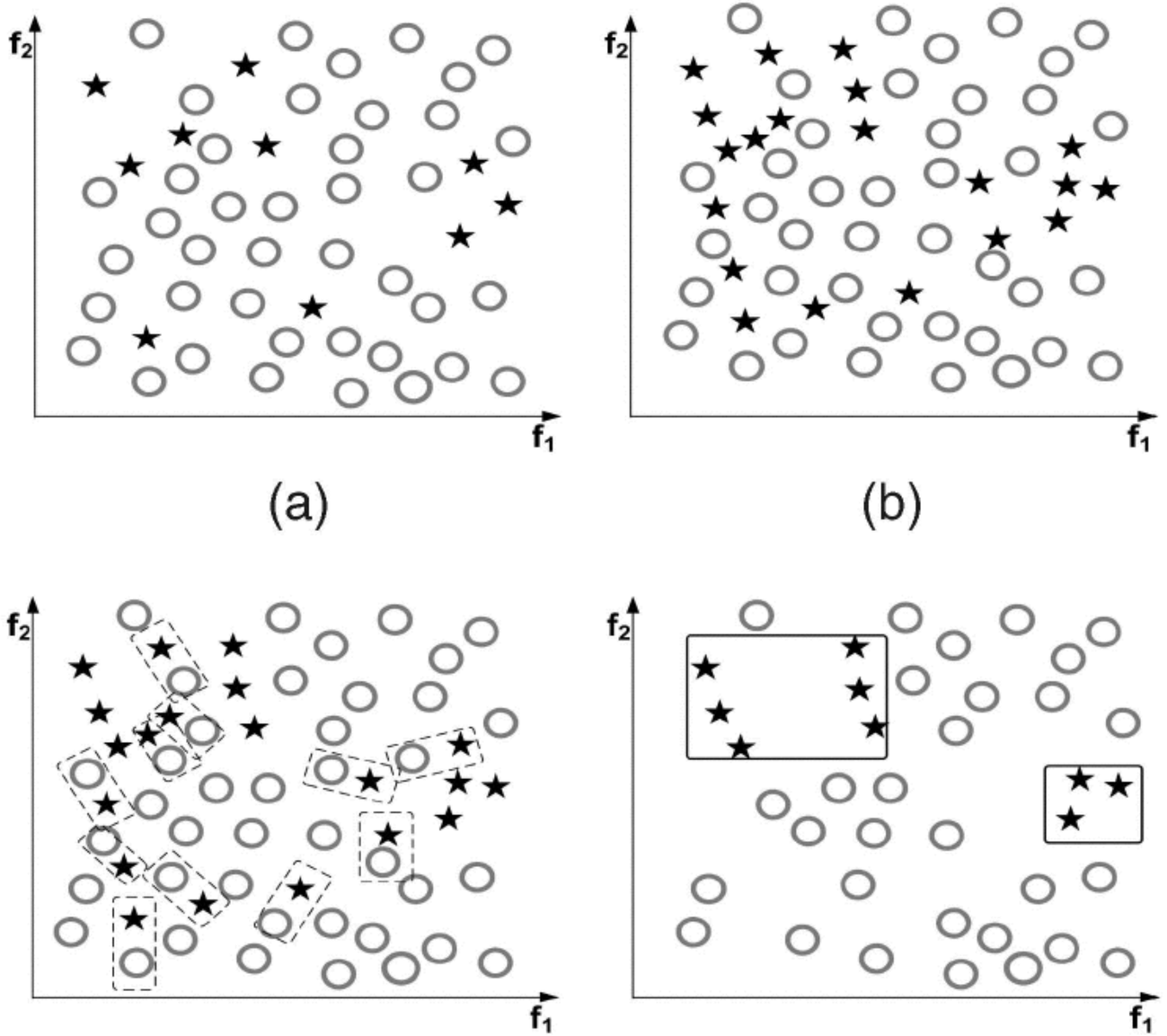


Demonstrated success on many problems

Other techniques possible:
clustering-based (k-Means)
imputation-based (mice R-package)
…

Reducing issues

- Some points to close to majority points
- Remove Tomek links
- …etc?…



Cost-sensitive decision trees

- Splitting using error rate
  - min(n+ / n, n- / n)
  - n+ = positives, n- = negatives, n = total
- Using impurity (Gini index):
  - 2 * (n+ / n) * (n- / n)
- Square Root Gini:
  - 2 * sqrt((n+ / n) * (n- / n))
  - is insensitive to changes in class distribution!

Cost-sensitive decision trees

- Splitting using error rate
  - min(n+ / n, n- / n)
  - n+ = positives, n- = negatives, n = total
- Using impurity (Gini index):
  - 2 * (n+ / n) * (n-
- Square Root G
  - 2 * sqrt((n+ / r
  - is insensitive to

Take care when pruning!

Suggested to use Laplace…

Weighted distances

- k-Nearest Neighbor class for instance i is:
  - sum[j=1..k](c(i,j)) / k
  - where c(i,j) is the class of the j-th nearest neighbor of i

- Weighted by distance:
  - sum[j=1..k](w(i,j) * c(i,j)) / k
  - where w(i,j) is the weight of the j-th neighbor, e.g.:
    - w(i,j) = (d(i,k) - d(i,j)) / (d(i,k) - d(i,1))

Weighted distances

- k-Nearest Neighb
  - sum[j=1..k](c(i,j)) /
  - where c(i,j) is the

- Weighted by dist
  - sum[j=1..k](w(i,j) * c(i,j)) / k
  - where w(i,j) is the weight of the j-th neighbor, e.g.:
    - w(i,j) = (d(i,k) - d(i,j)) / (d(i,k) - d(i,1))

Intuition:

fraud is anomalous
the closer to the anomaly
the stronger the vote of that anomaly

Measuring performance

- Standard measures:

$$\text{Recall} = \frac{TP}{TP+FN}$$

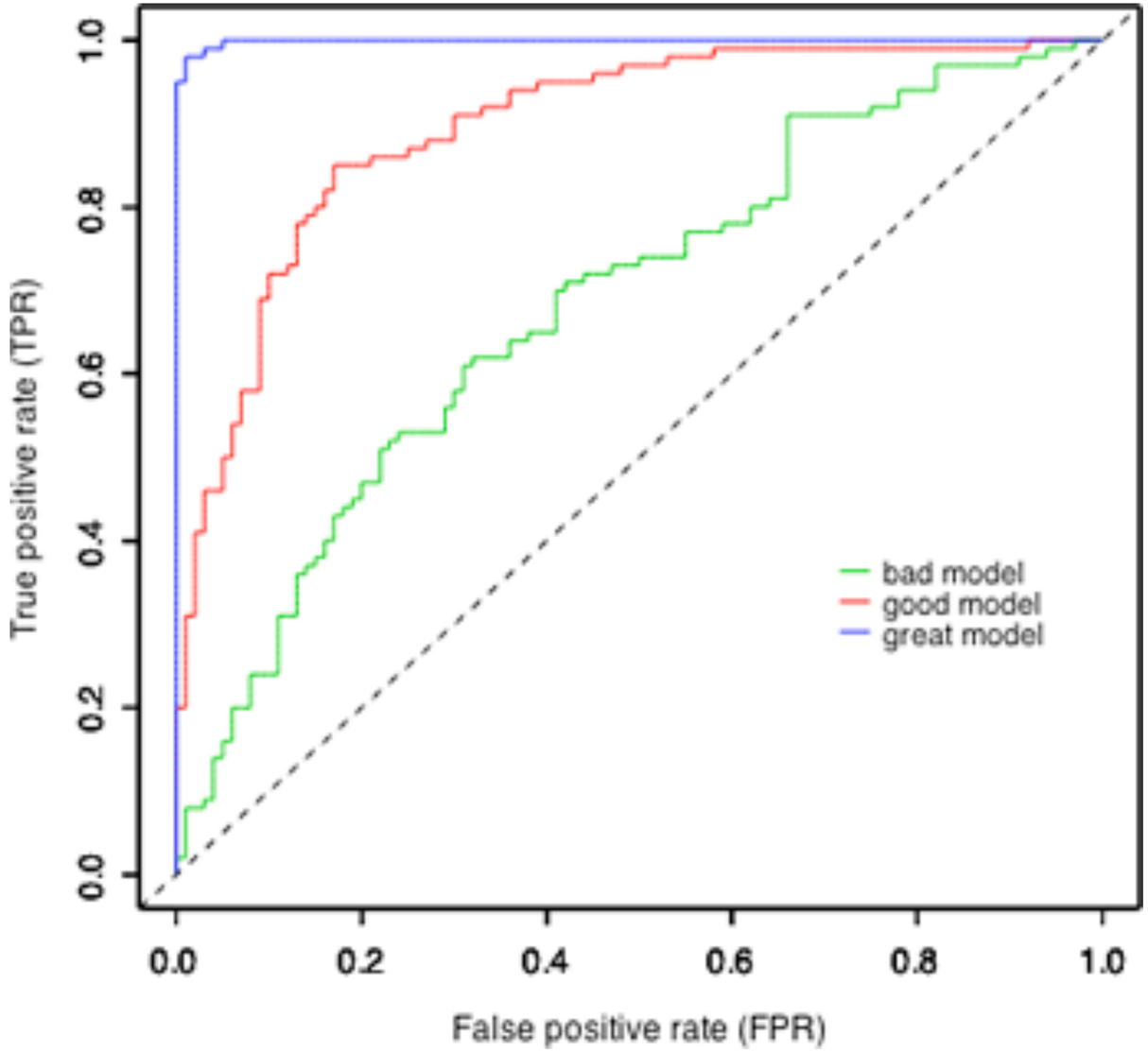
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

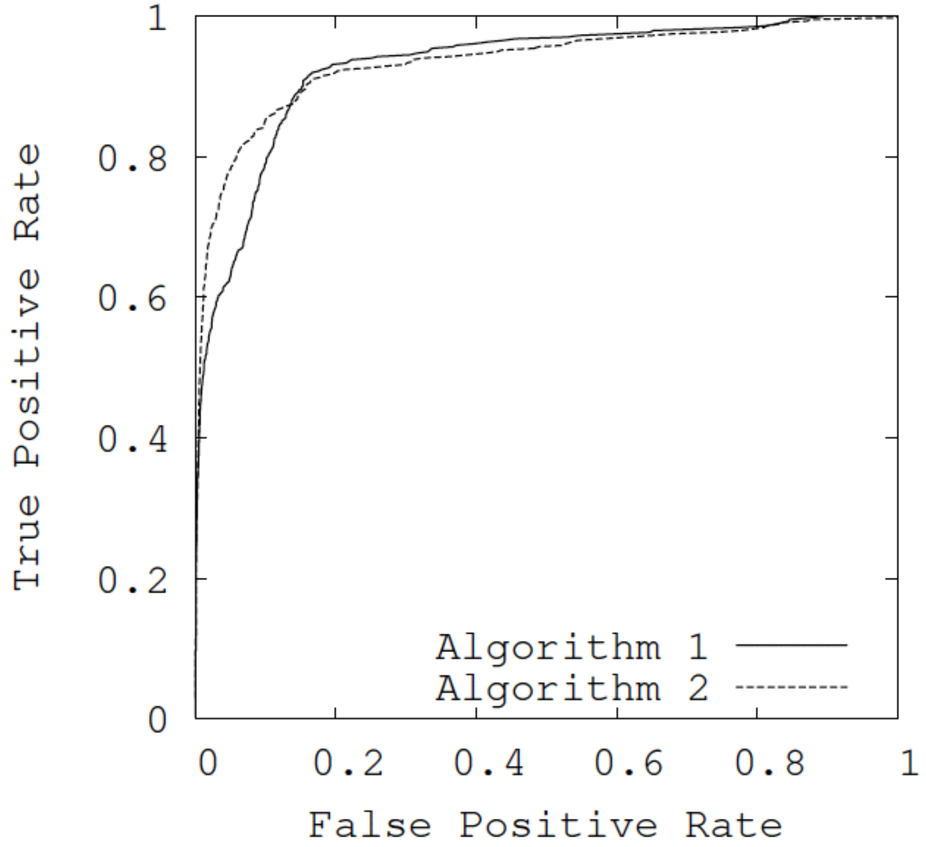
ROC-curves

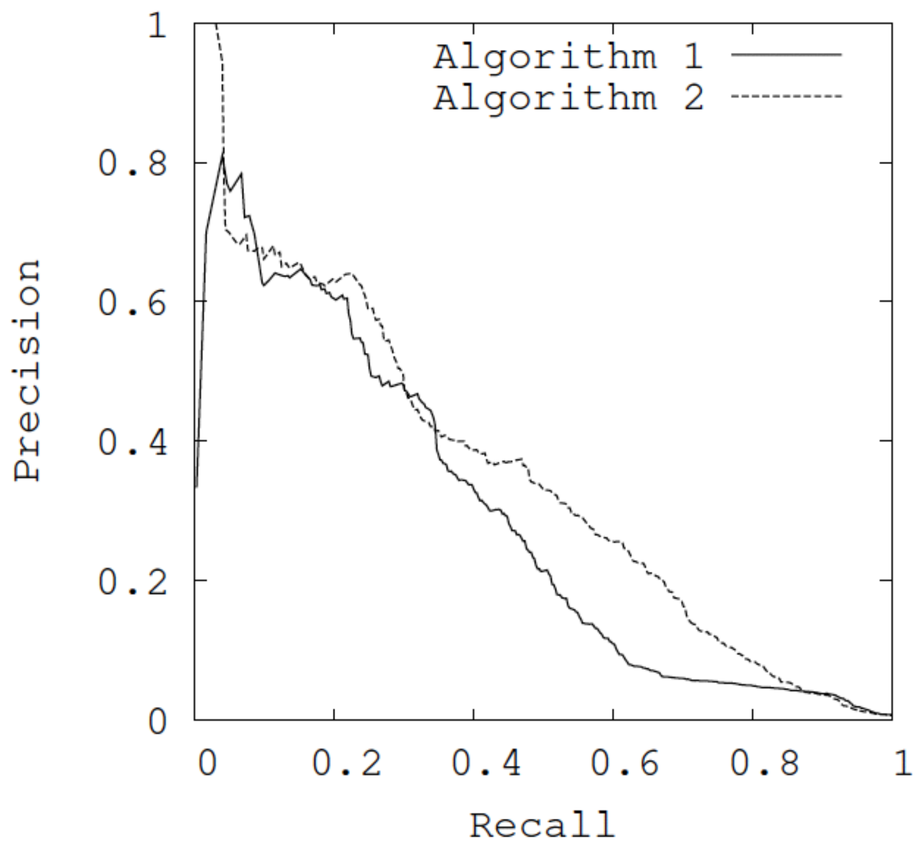
- Show TPR-FPR
- More to the top-left is better



PR-curves

- ROC curves often give an over-optimistic view
- Precision-Recall curves show differences more globally for imbalanced data, top-right is better





Cost-curves

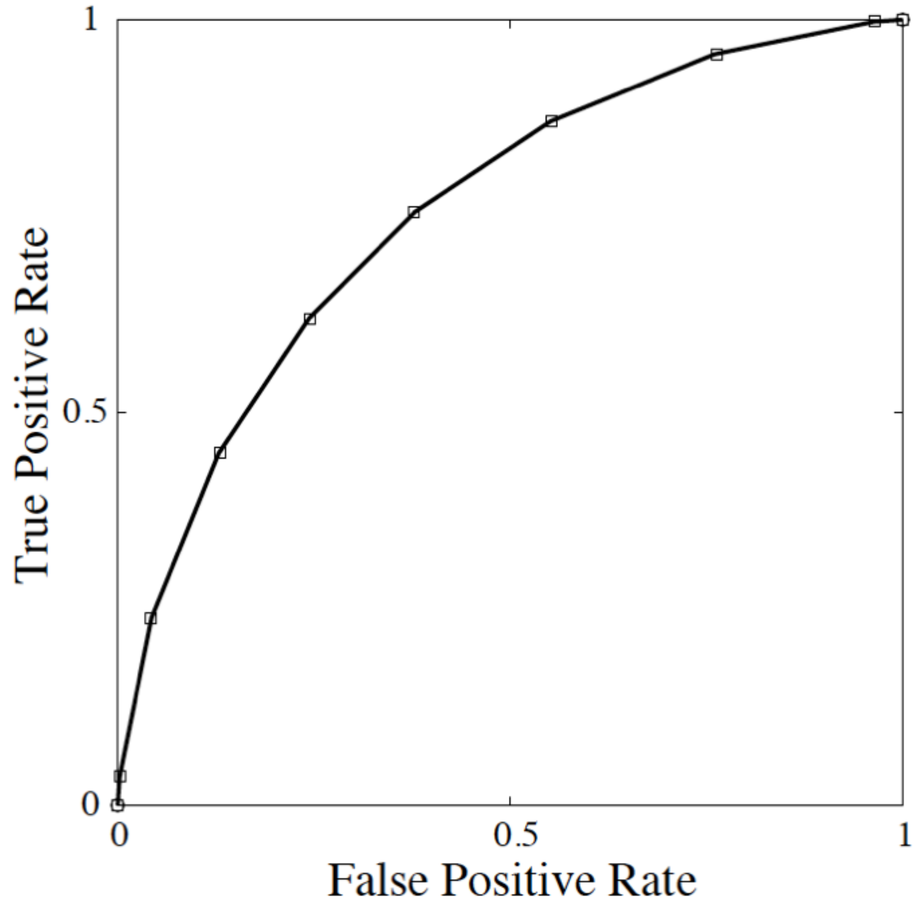
- Plot expected cost against probability of getting a positive example

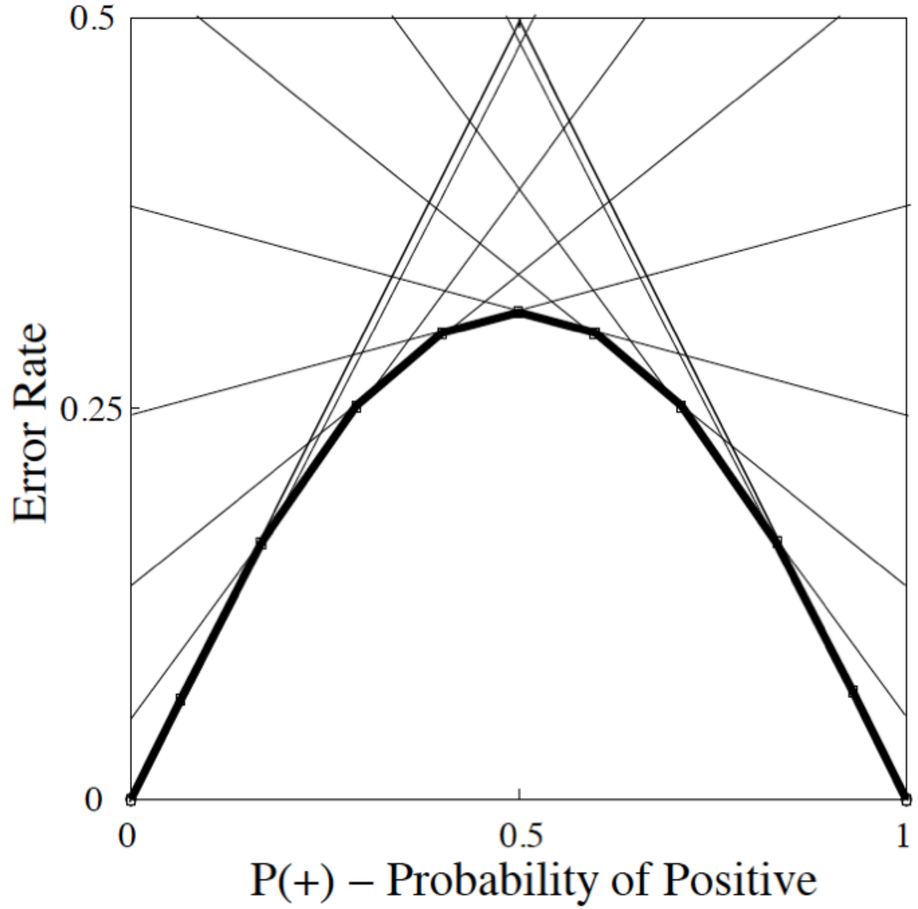
$$E[\texttt{Cost}] = FN * p(+) * \texttt{C}(-|+) + FP * p(-) * \texttt{C}(+|-)$$

- P(+) is simply the fraction of positive examples

Cost curves

- Every point in ROC-space is a curve in cost-space





Cost-curves

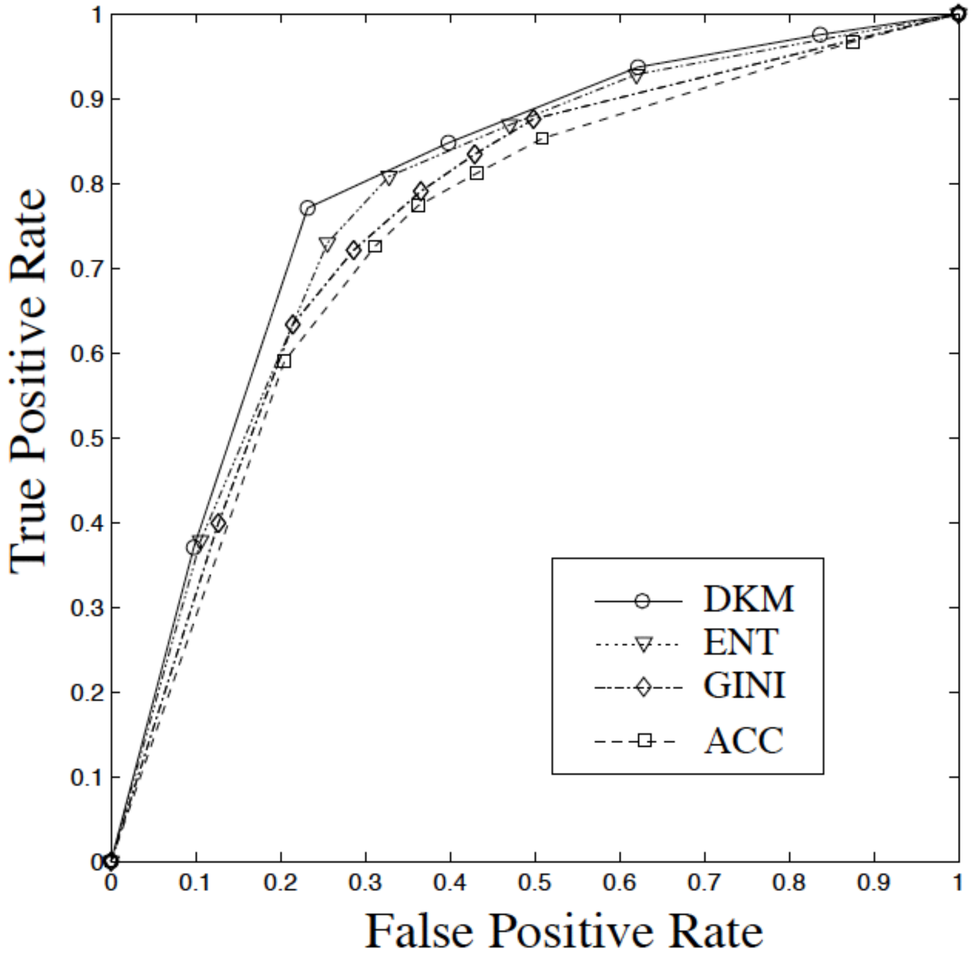
- Normalizing, and including cost in x-axis
  - “probability times cost”

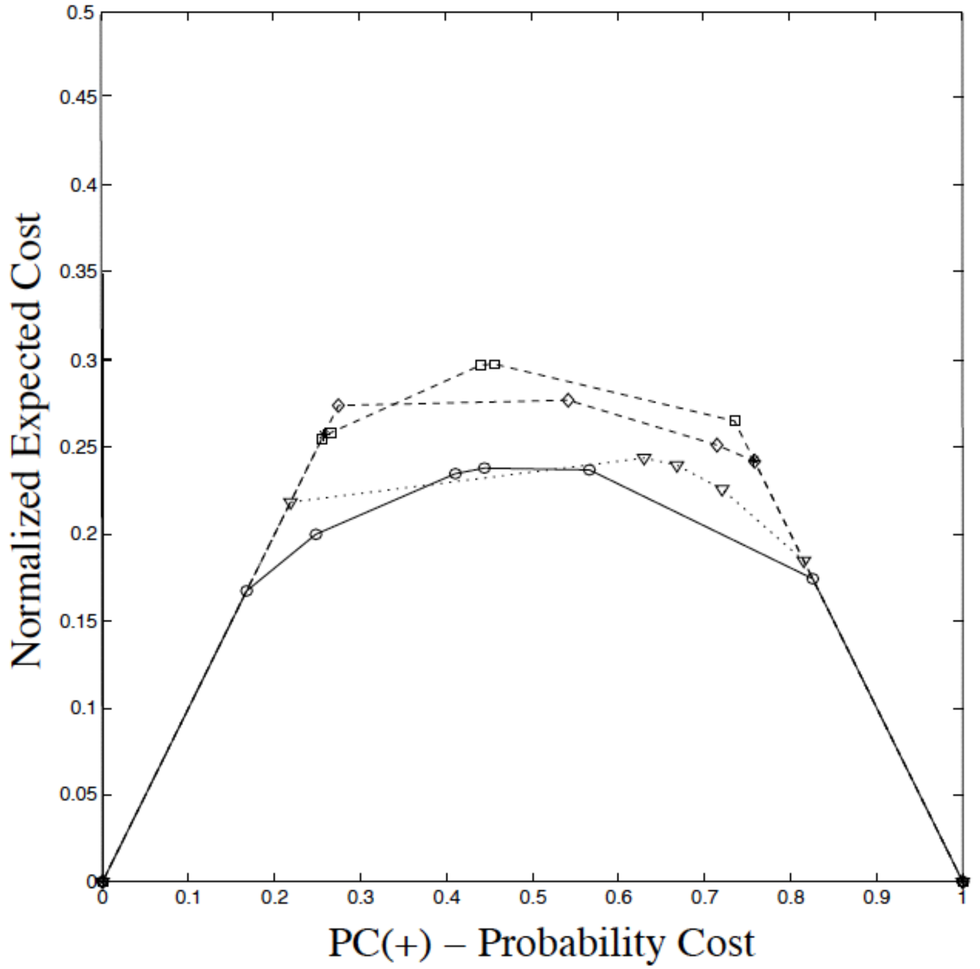
$$\texttt{Norm}(\texttt{E}[\texttt{Cost}]) = \frac{FN * p(+) * \texttt{C}(-|+) + FP * p(-) * \texttt{C}(+|-)}{p(+) * \texttt{C}(-|+) + p(-) * \texttt{C}(+|-)}$$

$$PC(a) = \frac{p(a) * \texttt{C}(\overline{\texttt{a}}|\texttt{a})}{p(+) * \texttt{C}(-|+) + p(-) * \texttt{C}(+|-)}$$

Cost curves

- Using different decision-tree splitting criteria

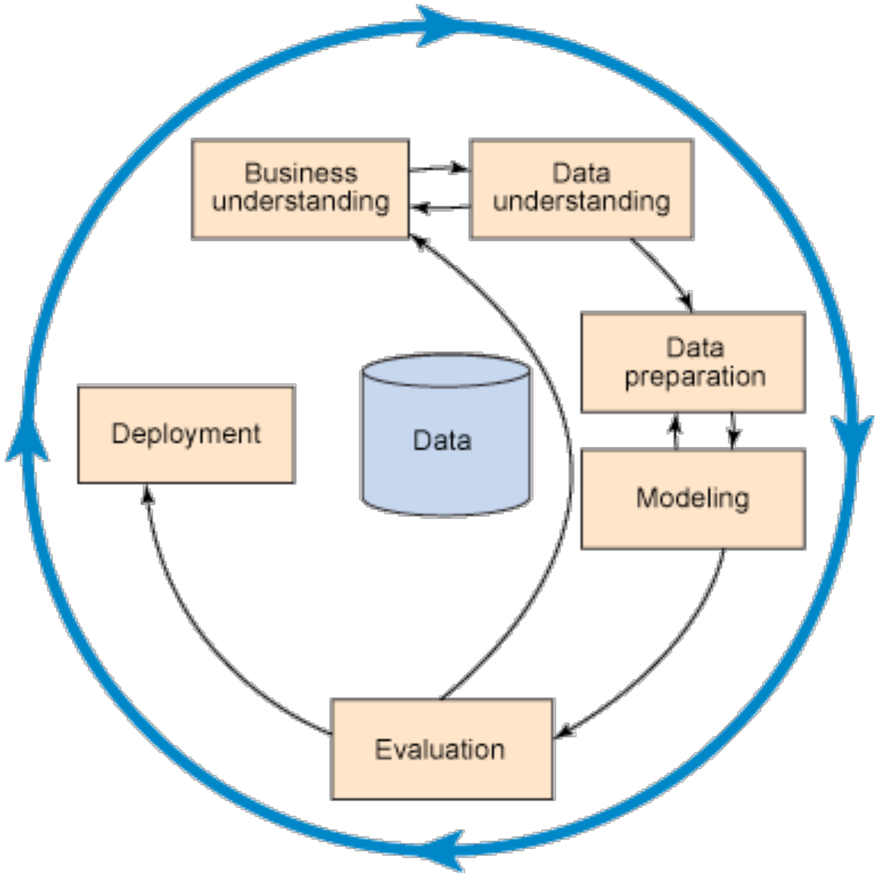




Please don’t forget:

- Data can be aggregated in many different ways
  - understand the goal!
  - understand the data!
  - ***before modelling!***

- Key tool: visualisation
  - heat maps
  - cross tables
  - pair-wise scatterplots
  - parallel coordinates
  - …



CRISP-DM