Анализ рисков в мире больших данных

Профессор Барт Бэсенс

Департамент принятия решений и информационного менеджмента,
Лёвенский католический университет (Бельгия)

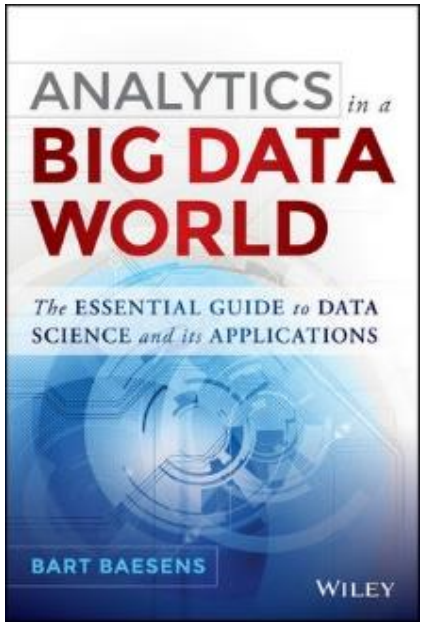
Школа менеджмента, Университет Саутгемптона (Великобритаия)

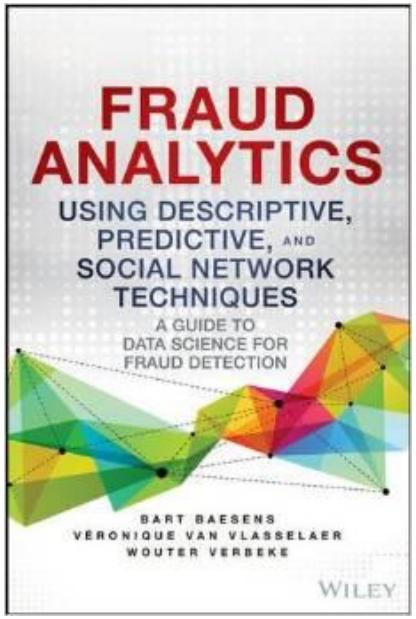
Bart.Baesens@kuleuven.be
Twitter/Facebook/Youtube: DataMiningApps
www.dataminingapps.com

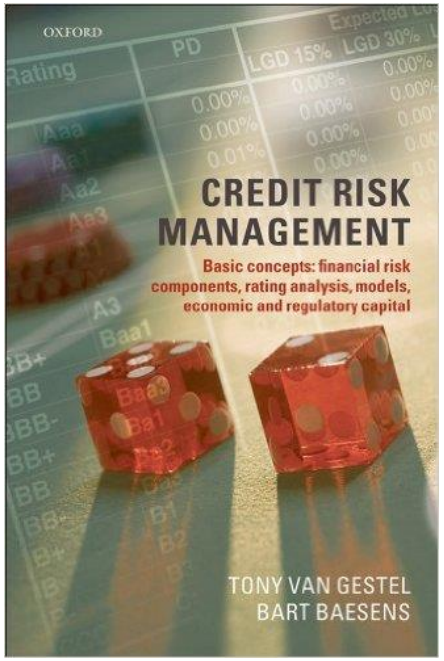
Докладчик: Барт Бэсенс (Bart Baesens)

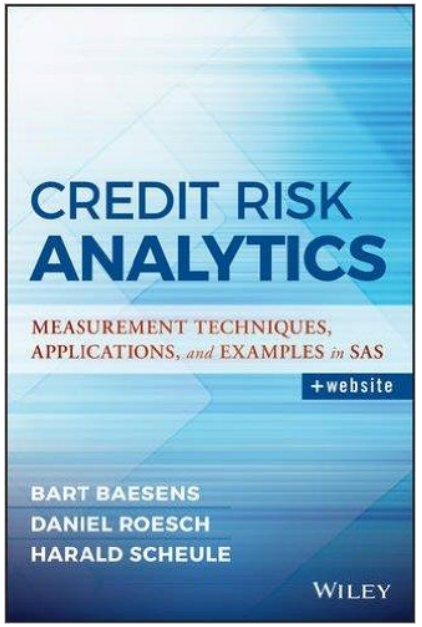
- Закончил Лёвенский КУ (Бельгия)
  - Бизнес-инженер по направлению подготовки «Бизнес-информатика», 1998
  - PhD в области прикладной экономики, 2003
- Тема диссертации: Разработка интеллектуальных систем кредитного скоринга с применением технологий машинного обучения
- Профессор больших данных и аналитики в Лёвенском католическом университете (Бельгия)
- Преподаватель Университета Саутгемптона (Великобритаия)
- Научные интересы: большие данные и аналитика; кредитный риск; обнаружение мошенничества; маркетинг, …
- Youtube/Facebook/Twitter: DataMiningApps
- www.dataminingapps.com
- Bart.Baesens@kuleuven.be

Книги


ANALYTICS in a
BIG DATA
WORLD
The ESSENTIAL GUIDE to DATA SCIENCE and its APPLICATIONS
BART BAESENS
WILEY

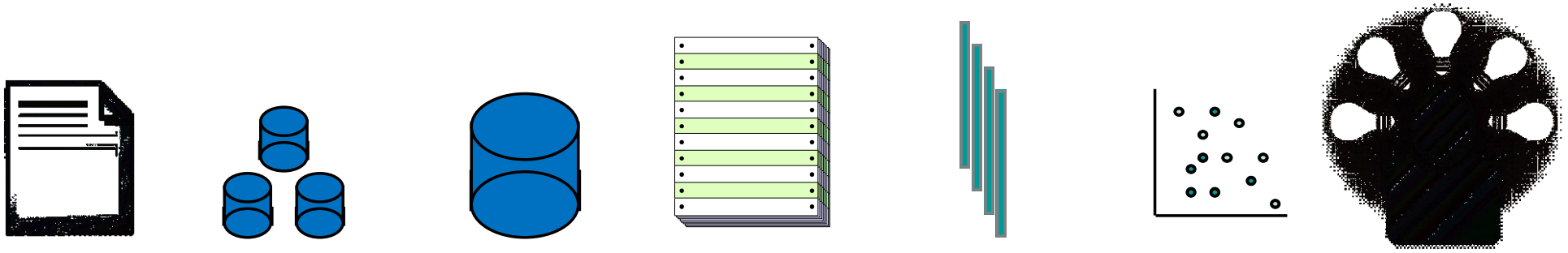

FRAUD
ANALYTICS
USING DESCRIPTIVE, PREDICTIVE, AND SOCIAL NETWORK TECHNIQUES
A GUIDE TO DATA SCIENCE FOR FRAUD DETECTION
BART BAESENS
VÉRONIQUE VAN VLASSELAER
WOUTER VERBEKE
WILEY


OXFORD
CREDIT RISK
MANAGEMENT
Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital
TONY VAN GESTEL
BART BAESENS


CREDIT RISK
ANALYTICS
MEASUREMENT TECHNIQUES, APPLICATIONS, and EXAMPLES in SAS
+website
BART BAESENS
DANIEL ROESCH
HARALD SCHEULE
WILEY

Мы живем в мире потоков данных!

Модель аналитического процесса



Постановка бизнес-задачи

Определение источников данных

Выбор данных

Очистка данных

Трансформация данных

Анализ данных

Интерпретация, оценка и развертывание модели

Предобработка

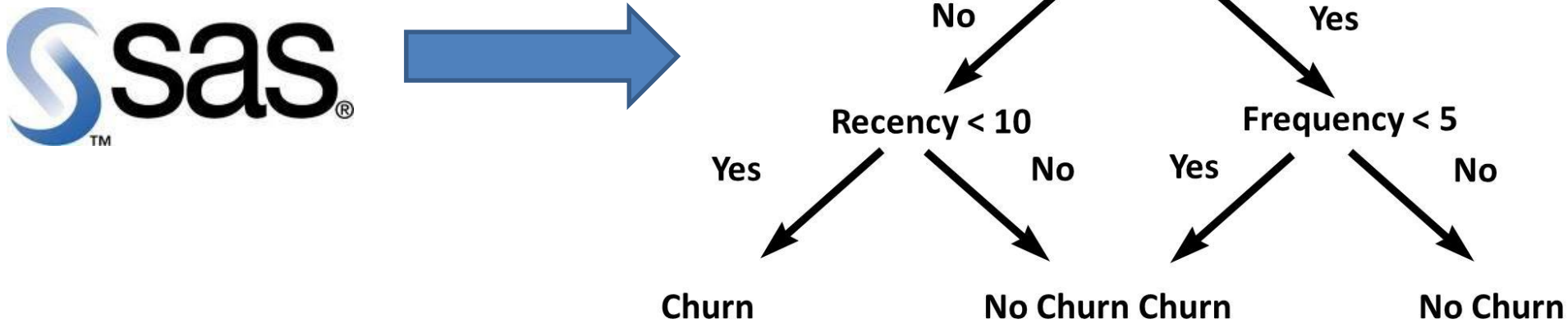
Аналитика

Постобработка

Почувствуйте вибрации!

Пример из области маркетинга

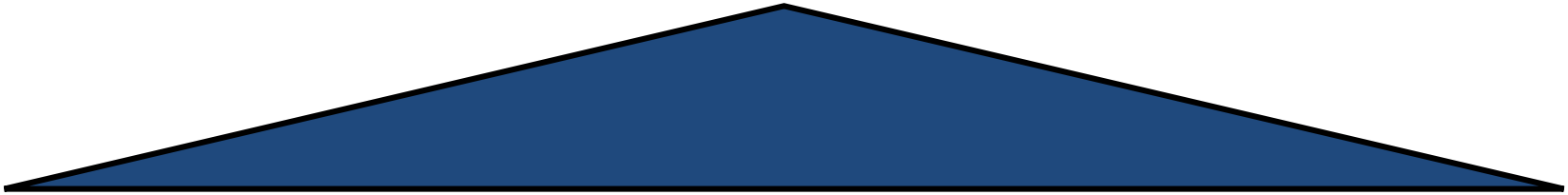
Customer (имя клиента)	Age (воз-раст)	Recency (срок с последней покупки)	Frequency (частота покупок)	Monetary (сумма)	Churn (отток)
John	35	5	6	100	**Yes**
Sophie	18	10	2	150	**No**
Victor	38	28	8	20	**No**
Laura	44	12	4	280	**Yes**



Пример из области риск-менеджмента

- Как никогда раньше аналитические модели определяют стратегические решения по управлению рисками в финансовых учреждениях!
- Минимальный собственный капитал (защитный капитал) и состояние финансового учреждения полностью определяются, в том числе:
  - аналитикой кредитного риска
  - аналитикой рыночного риска
  - аналитикой операционного риска
  - …
- Обычно для построения всех этих моделей используется бизнес-аналитика!
- Часто является предметом регулирования (например, Basel II, Basel III, Solvency II, …)!
- Ошибки в моделях напрямую влияют на рентабельность, платеже-способность, стоимость акций, макроэкономику, …, общество в целом!

Базельское соглашение



Колонна 1: Минимальные требования к капиталу

- Кредитный риск
- Операционный риск
- Рыночный риск

Колонна 2: Процедура пересмотра

- Надежные внутренние процессы оценки риска (ICAAP)
- Надзор регулятора

Колонна 3: Рыночная дисциплина и раскрытие информации

- Раскрытие:
  - Профиля риска банка
  - Качественной и количественной информации
  - Процессов управления рисками
  - Стратегии управления рисками

Компоненты кредитного риска

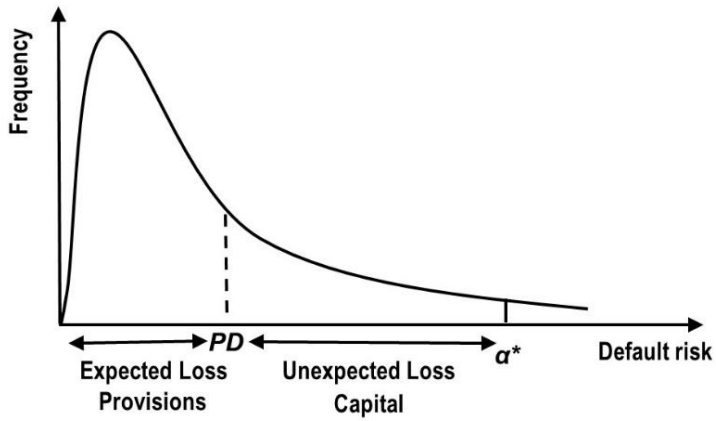
- **Вероятность дефолта (PD - Probability of default, десятичн.):** вероятность дефолта контрагента в течение года (Ст. 4, ЕС)

- **Требования под риском дефолта (EAD - Exposure at default, ден.ед.):** сумма задолженности

- **Доля потерь при дефолте (LGD - Loss given default, десятичн.):** отношение потерь при дефолте к общей сумме задолженности (Ст. 4, ЕС)

- **Ожидаемые потери (EL - Expected loss)**
  EL = PD x LGD x EAD

- **Неожиданные потери (UL - Unexpteced loss)**
  UL = f(PD, LGD, EAD)



Пример таблицы балльной оценки

Барьер = 500

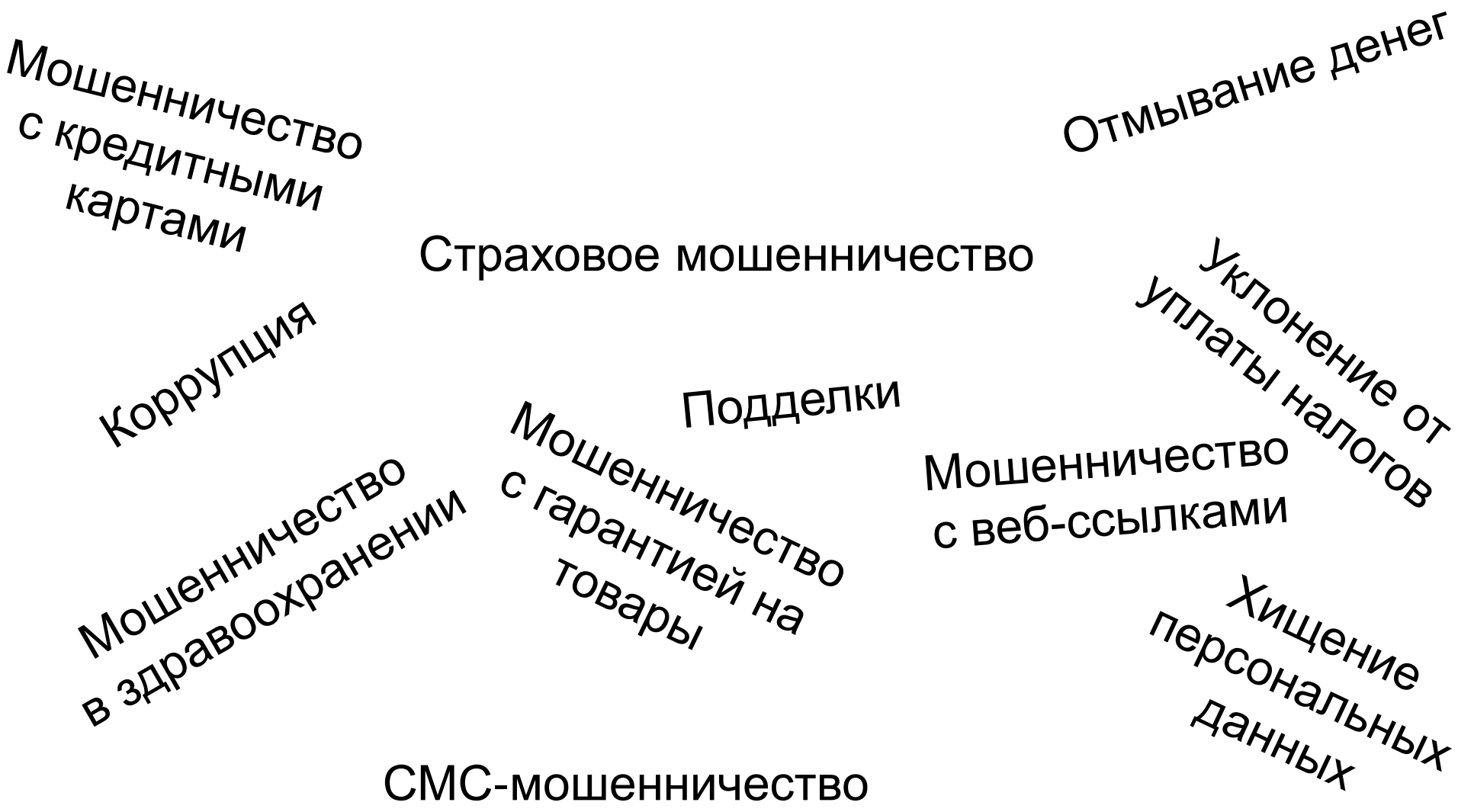
Новый заемщик обращается за кредитом:

Возраст	**32**	**120 баллов**
Пол	**Ж**	**180 баллов**
Зарплата	**$1,150**	**160 балов**
Итого		**460 баллов**

Не выдавать кредит

Показатель	Значение	Баллы
ВОЗРАСТ 1	до 26	100
ВОЗРАСТ 2	26 – 35	120
ВОЗРАСТ 3	35 – 37	185
ВОЗРАСТ 4	более 37	225
ПОЛ 1	М	90
ПОЛ 2	Ж	180
ЗАРПЛАТА 1	до 500	120
ЗАРПЛАТА 2	501-1000	140
ЗАРПЛАТА 3	1001-1500	160
ЗАРПЛАТА 4	1501-2000	200
ЗАРПЛАТА 5	более 2000	240

Обнаружение мошенничества



Определение мошенничества (Baesens, Van Vlasselaer, Verbeke, 2015)

“Мошенничество – это необычные,
хорошо продуманные,
старательно скрываемые,
развивающиеся во времени
и, как правило, тщательно организованные
преступные действия, которые проявляются
в различных видах и формах.”

Основные проблемы выявления мошенничества

- Мошенники стараются постоянно быть умнее, чем аналитические модели
- Цена невыявленного мошенничества крайне высока
- Ни в коем случае нельзя беспокоить хороших клиентов и блокировать их счета
- Мошенничество - иголка в стоге сена
- Очень асимметричные наборы данных
  - Напр., доля мошенничеств с кредитными картами обычно < 0.5%
- Ключевой показатель – операционная эффективность
  - Напр., время принятия решений при мошенничествах с кредитными картами обычно < 8 секунд
- Необходимо обрабатывать большие объемы данных!

Подходы к выявлению мошенничества

- Экспертные оценки
- Описательная аналитика
- Предсказательная аналитика
- Аналитика социальных сетей

Описательная аналитика

- Обнаружение выбросов/аномалий
- Выявление фактов мошенничеств

Дата	Время	День недели	Продол-жительность	Вызыва-ющий	Вызыва-емый	Мошенничество
18/11/2016	11:15	Wed	18 mins	Paris	Rome	
19/11/2016	9:44	Thu	15 mins	Paris	Prague	
19/11/2016	19:20	Thu	22 mins	Helsinki	Madrid	
20/11/2016	12:05	Fri	34 mins	Paris	Paris	
20/11/2016	14:34	Fri	12 mins	Dublin	Rome	
20/11/2016	15:58	Fri	53 sec	Paris	Paris	
21/11/2016	01:45	Sat	30 sec	Amsterdam	Moscow	Подозрительно
22/11/2016	01:50	Sun	15 sec	Amsterdam	Moscow	Подозрительно
22/11/2016	10:45	Sun	10 min	Dublin	Paris	
22/11/2016	23:20	Sun	24 sec	Madrid	Frankfurt	Подозрительно
23/11/2016	15:18	Mon	14 min	Amsterdam	Amsterdam	
23/11/2016	23:54	Mon	18 sec	Amsterdam	Berlin	Подозрительно

Предсказательная аналитика

Исторические данные

Клиент	Возр.	Станд. сумма	Личность	…	Мошенничество
John	30	-1,22	Идент.		Нет мошенн..
Sarah	25	3,44	Не идент.		Мошенн..
Sophie	52	0,78	Идент.		Нет мошенн..
David	48	0,22	Идент		Нет мошенн..
Peter	34	-2,54	Не идент.		Мошенн.

```
proc logistic data=frauddata;
   class Identity;
   model Fraud=StdAmount Identity …;
run;
```

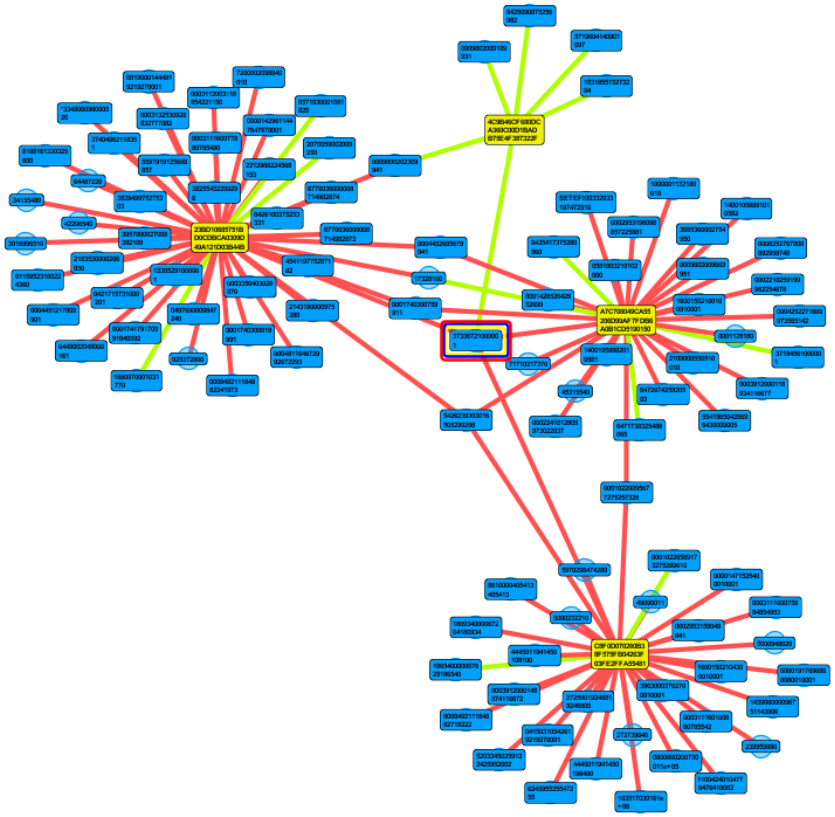
$$P(\text{fraud} = \text{yes} \mid StdAmount, \text{Identity}, \text{Income}, \ldots) = \frac{1}{1 + e^{-(0.10 + 0.78\,StdAmount + 0.50\,Identity - 0.80\,Income \ldots)}}$$

Новые данные

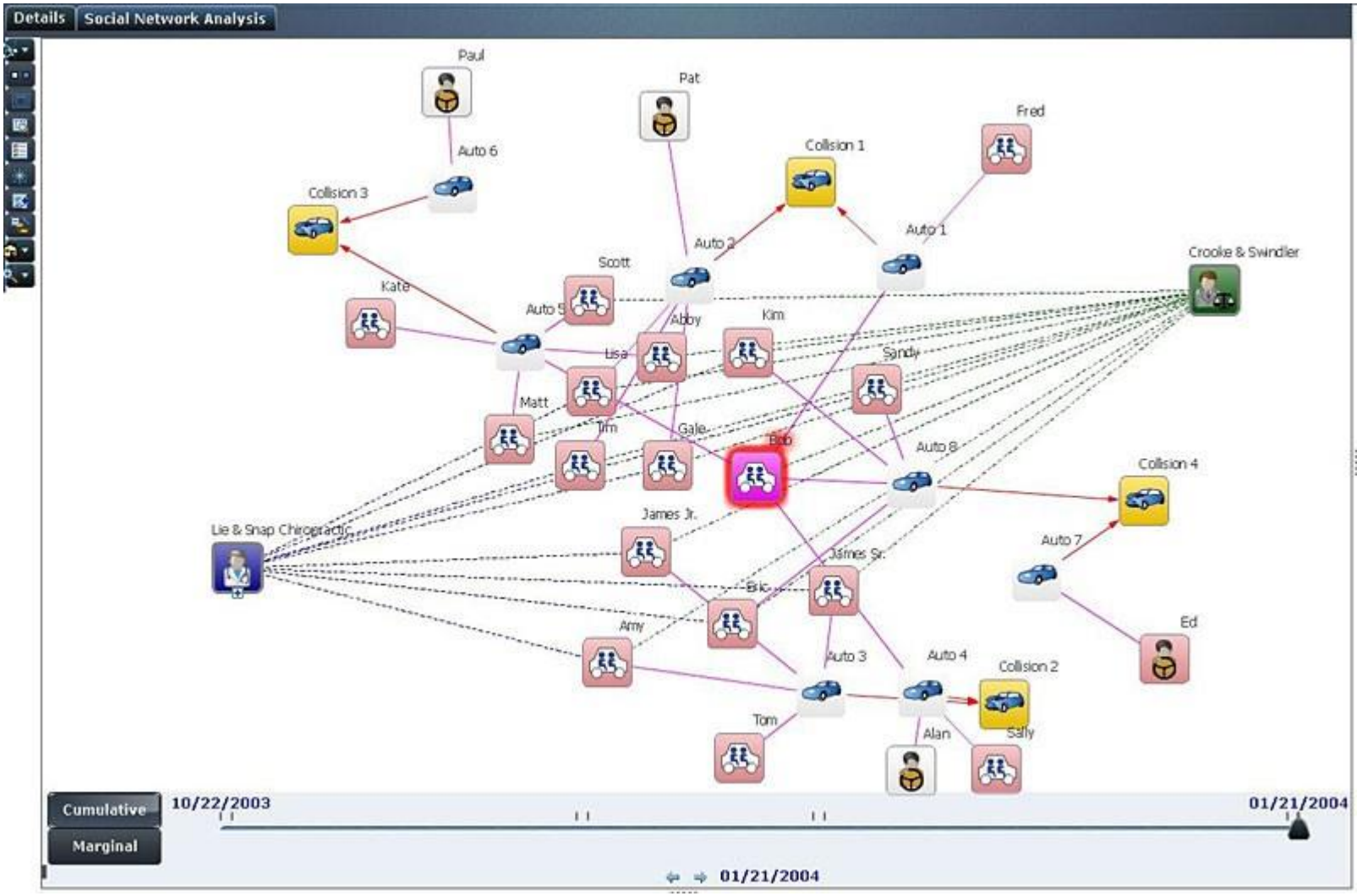
Клиент	Возр.	Станд. сумма	Личность	…	Мошенничество
Emma	28	0,18	Идент.		0,14
Will	44	-1,78	Не идент.		0,72
Dan	30	0,06	Идент.		0,06
Bob	58	4,15	Не идент.		0,88

Аналитика социальных сетей

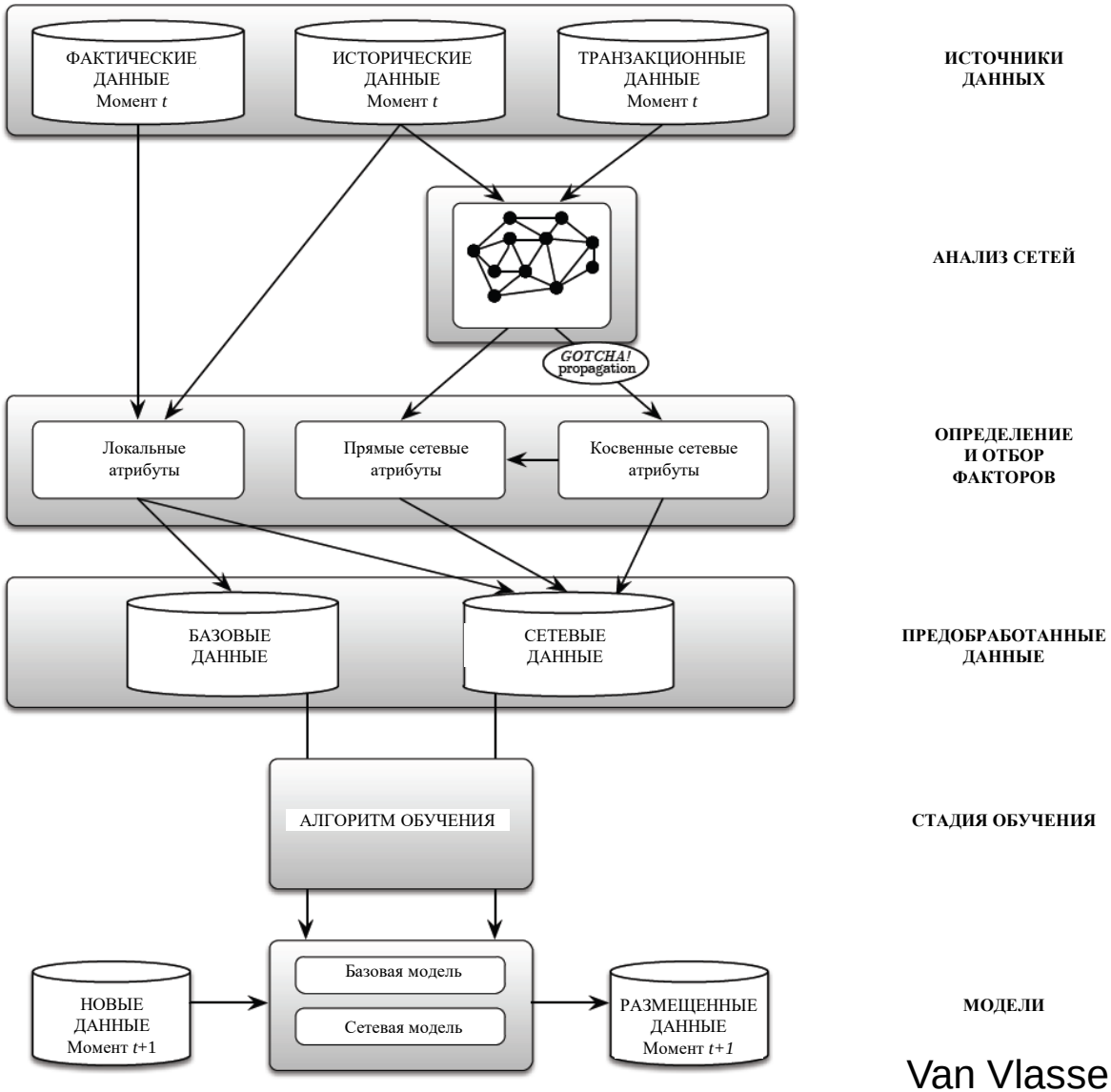
- Мошеннические транзакции по кредитным картам:
- Украденными кредитными картами (желтые узлы) часто платят в одних и тех же магазинах (синие узлы)
- Чтобы покрыть мошенничество, магазины проводят и *законные* сделки



Аналитика социальных сетей



Gotcha! (англ. - Попался!)



Van Vlasselaer, Baesens, 2016.

Заключительные замечания

- Интеграция данных
- Качество данных
- Правильные специалисты
  (специалисты по наукам о данных!)
- Соответствие нормативным требованиям
- Экономический эффект
- Конфиденциальность
- Проверка на исторических данных

Дополнительная информация

Электронный курс
«Продвинутая аналитика в мире больших данных»

https://support.sas.com/edu/schedules.html?id=2169&ctry=US

Курс начинается с обзора основных понятий модели аналитического процесса: предобработка данных, аналитика, постобработка. Затем обсуждаются деревья решений и модели ансамблей (случайные леса), нейронные сети, машины опорных векторов, байесовские сети, анализ выживаемости, социальные сети, мониторинг и тестирование аналитических моделей на статистических данных. На протяжении всего курса мы активно используем наш практический и научный опыт. Для лучшего понимания в курс включены различные реальные примеры (e.g. кредитный скоринг, предсказание оттока клиентов, выявление мошенничества, сегментирование потребителей и т.д.) и мини-кейсы. Курс состоит из пятиминутных видеороликов общей продолжительностью более 20 ч. Для облегчения понимания материала приводятся тестовые вопросы. После регистрации Вы получаете код, который дает неограниченный доступ в течение 1 года ко всем материалам курса (видео, тестам, скриптам, …). Курс фокусируется на концепции и методологии моделирования, а не на программных продуктах SAS. Для доступа к курсу достаточно иметь компьютер, планшет, смартфон с доступом к интернету. Программное обеспечение SAS не требуется.

Дополнительная информация

Электронный курс «Моделирование кредитного риска»

https://support.sas.com/edu/schedules.html?ctry=us&id=2455

Курс охватывает как базовые, так и более продвинутые способы моделирования, проверки и стресс-тестирования вероятности дефолта (PD), доли потерь при (LGD) и требований под риском дефолта (EAD). The E-learning course covers both the basic as well some more advanced ways of modeling, validating and stress testing Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD) models. На протяжении всего курса мы активно используем наш практический и научный опыт. Для лучшего понимания в курс включены различные реальные примеры и мини-кейсы из области розничного и корпоративного кредитования. Курс состоит из пятиминутных видеороликов общей продолжительностью более 20 ч. Для облегчения понимания материала приводятся тестовые вопросы. После регистрации Вы получаете код, который дает неограниченный доступ в течение 1 года ко всем материалам курса (видео, тестам, скриптам, …). Курс фокусируется на концепции и методологии моделирования, а не на программных продуктах SAS. Для доступа к курсу достаточно иметь компьютер, планшет, смартфон с доступом к интернету. Программное обеспечение SAS не требуется.

Дополнительная информация

Электронный курс «Выявление мошеничества с помощью описательной аналитики, предсказательной аналитики и аналитики социальных сетей»

https://support.sas.com/edu/schedules.html?ctry=us&id=1912

Потери типичной организации от мошенничества составляют около 5% годового дохода. В данном курсе мы показываем, как для борьбы с мошенничеством можно использовать модели машинного обучения. Обсуждается использование моделей обучения с учителем (с размеченными данными), обучения без учителя (с неразмеченными данными) и анализ социальных сетей (с сетевыми данными). Эти технологии можно применять к широкому кругу прикладных задач: страховое мошенничество, мошенничество с кредитными картами, отмывание денег, мошенничество в здравоохранении, смс-мошенничество, мошенничество с веб-ссылками, уклонение от уплаты налогов, подделка и т.д. В курсе рассматриваются как теоретические, так и технические идеи, а также детали практической реализации. Кроме того, преподаватель активно использует свои научные результаты. Для лучшего понимания в используются различные реальные примеры и мини-кейсы.