

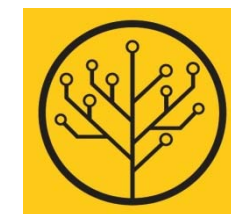


Картинка: <http://en.wikipedia.org/wiki/File:SpaghettiJunction0A.jpg>



# Эффективный анализ поведения пользователей с применением Software Process Mining

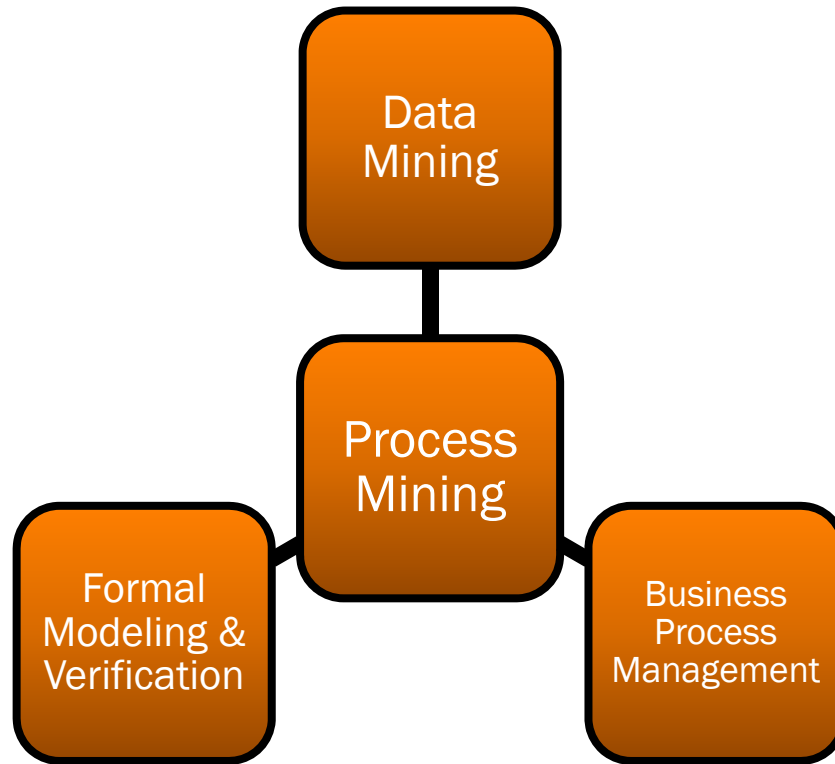
Алексей Мицюк, Лаборатория ПОИС ФКН НИУ ВШЭ



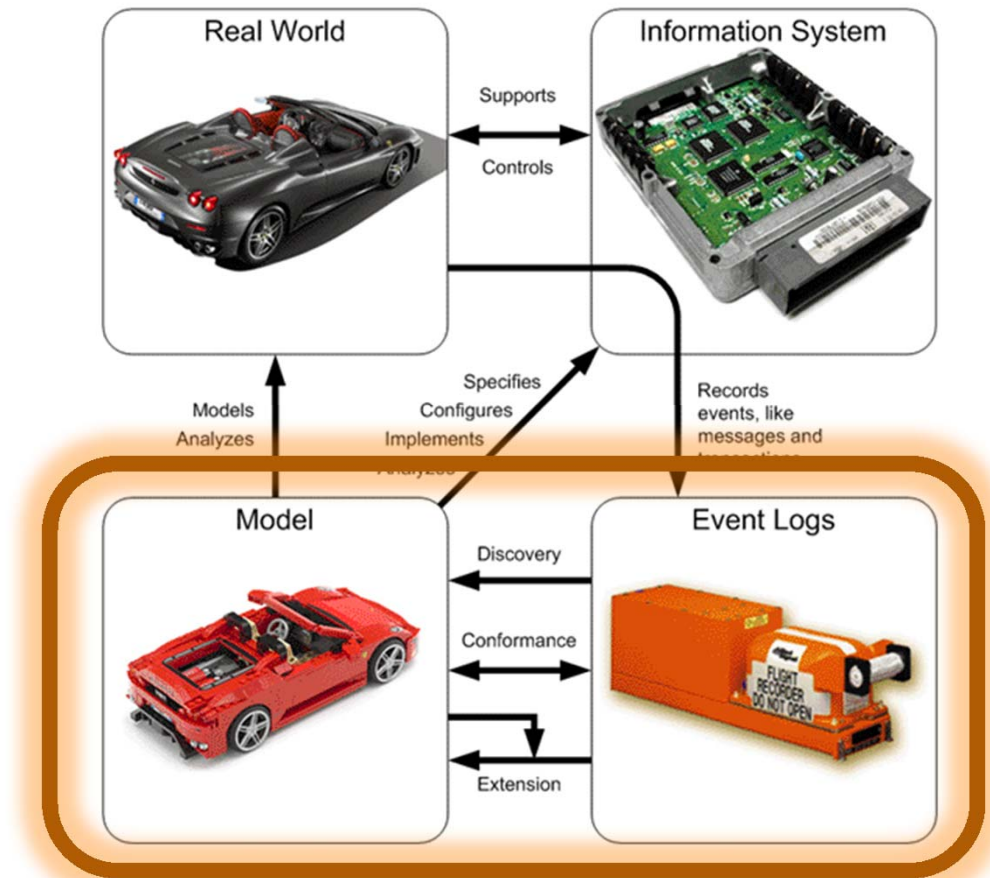
# Коротко

- В докладе будет рассмотрен вопрос применения Software Process Mining для анализа поведения пользователей информационной системы
- План доклада:
  - Что такое Process Mining?
  - Что означает Software в данном контексте?
  - Разбор двух примеров использования подхода
  - Что в итоге?

# Process Mining



# Process Mining



# Process Mining

## Process Discovery

- Лог событий ► Модель процесса

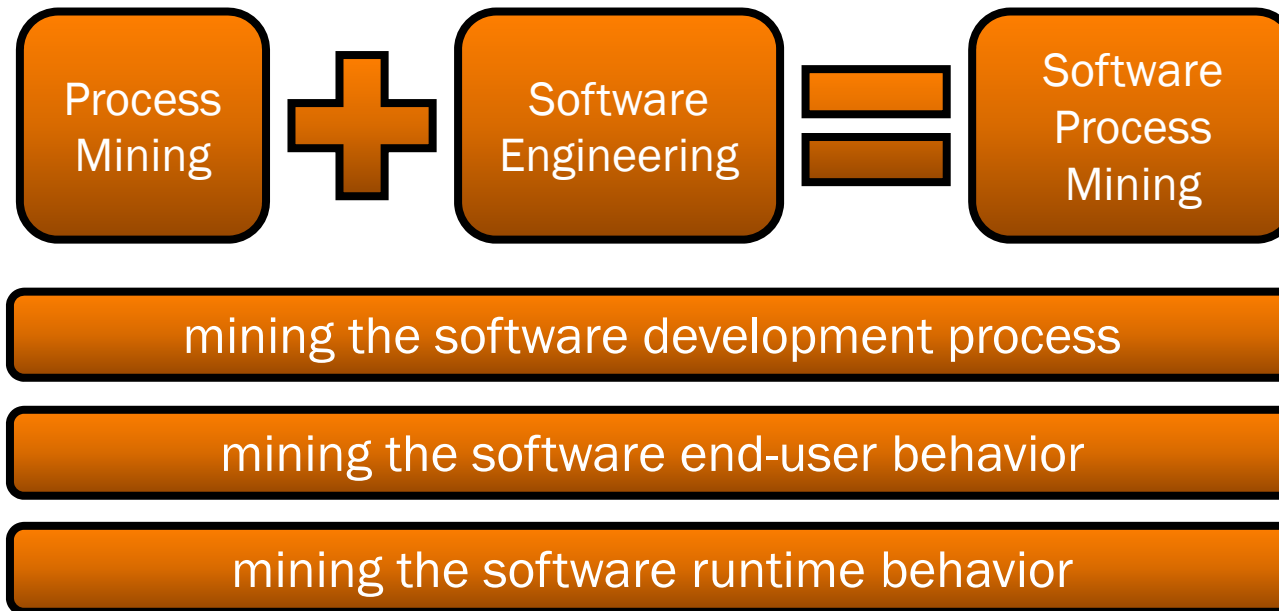
## Conformance Checking

- Лог событий + Модель процесса ► Мера соответствия

## Process Enhancement

- Лог + Модель + Мера соответствия + ... ► Улучшенная модель

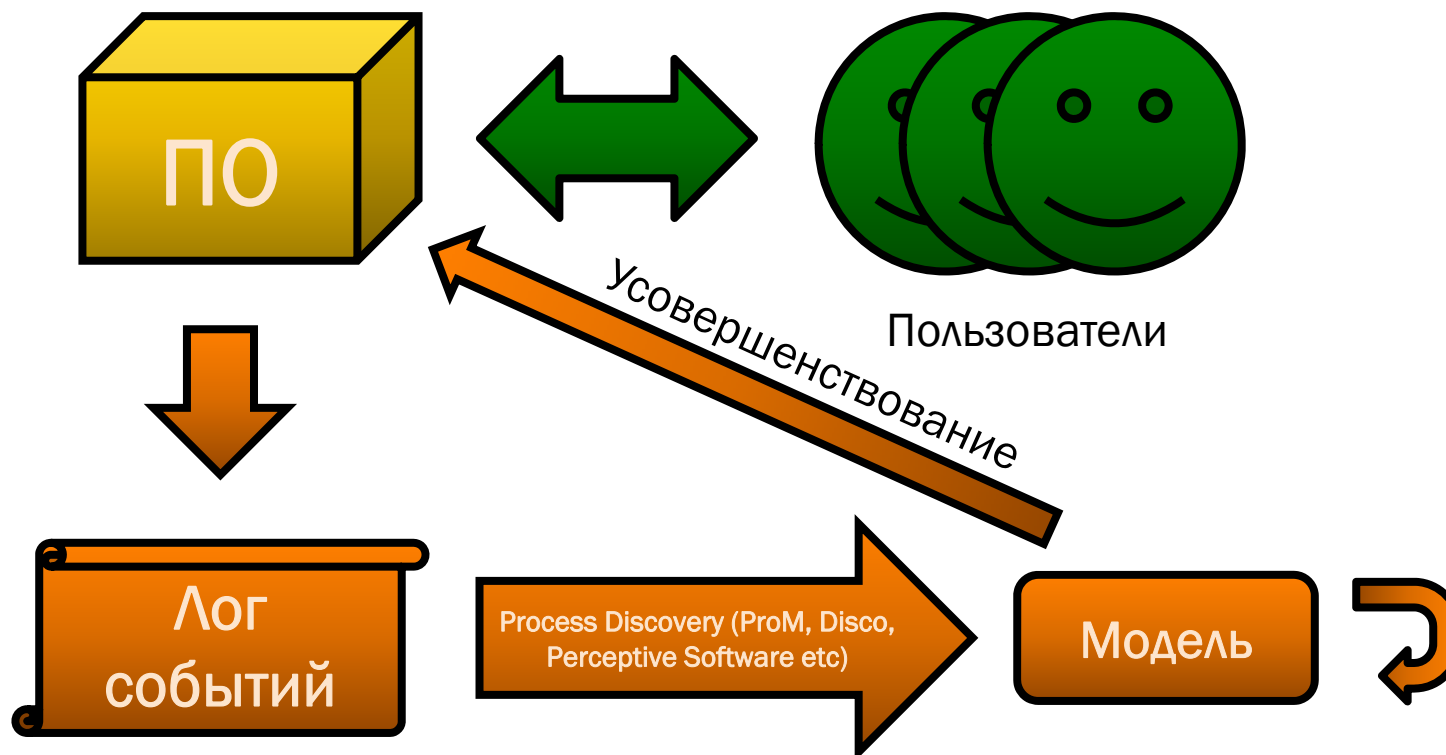
# Software Process Mining



Источник: V. Rubin et al (2014) Agile Development with Software Process Mining

# Software Process Mining

## Извлечение поведения пользователей



# Software Process Mining

## Извлечение поведения пользователей

1. Как ведут себя реальные пользователи?
2. Что НЕ получается у пользователей при работе с системой?
3. Какие функции нужны были бы, но отсутствуют?
4. Какие функции ПО НЕ используются, а какие более популярны?
5. Недостатки GUI – где пользователи попадают в тупик?
6. Особенности поведения для различных категорий пользователей, разных календарных периодов и т.д.



# Два примера применения Software Process Mining

Большая европейская система  
резервации туристических билетов

Российский портал по продаже  
электронных авиа-билетов

Один пример выполнялся в ЕС, другой в РФ

# Пример первый: исходные данные

Большая европейская система резервации туристических билетов

Более 10.000 фирм-клиентов в разных странах подключены к системе

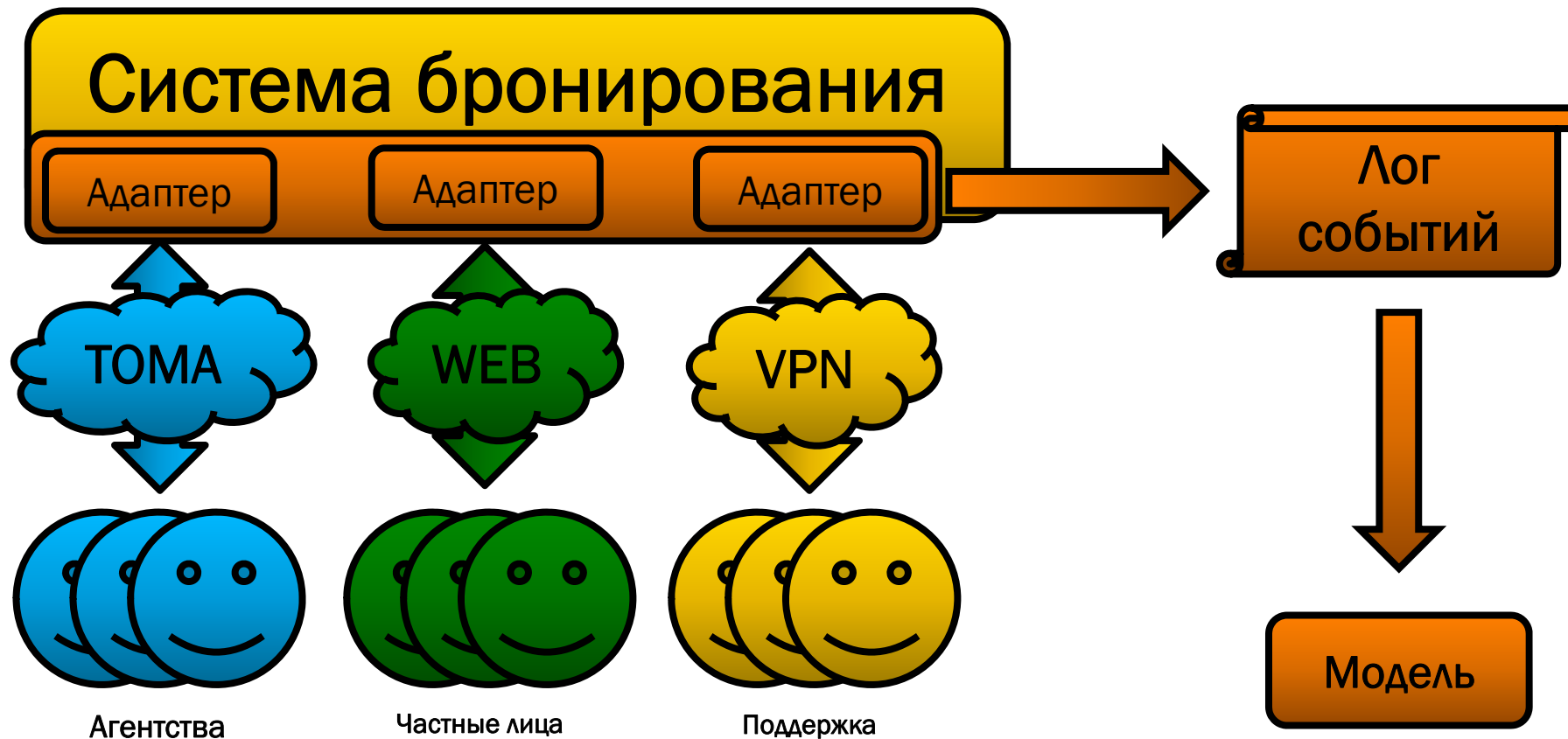
Система осуществляет бронирование разных типов: билеты, отели, страховки и т.д.

Агентства используют свои клиенты для доступа к системе по протоколу *ТОМА*\*

В качестве лога событий используется запись сообщений протокола, каждое из которых содержит идентификатор пользователя, время, операцию и введенные данные

\*Amadeus. Amadeus Toma Basic (Trainingshandbuch).  
<http://www.amadeus.com/at/documents/aco/at/de/Amadeus%20TOMA%20Basic.pdf>, 2014.

# Пример первый: исходные данные



# Пример первый: исходные данные

Case ID	Activity	Timestamp	User	Booking Code	Notification
Case 16	T1-HF-H:TES	2013-12-18 08:36:00:570	C05	BER	
Case 16	M5 Type 010Rsp 034	2013-12-18 08:36:04:717	C05		998
Case 16	T3-HF-H:HH00	2013-12-18 08:36:09:337	C05	BER	
Case 16	M5 Rsp	2013-12-18 08:36:09:337	C05		998

Код ответа

С помощью адаптеров собираются логи системы по дням

Размер логов большой (~30 Мб в день, много пользователей)

Формат логов хороший для process mining благодаря адаптерам (CSV)

Каждый Case содержит одно исполнение процесса: запрос предложения, оформление заказа, оплата, отправка квитанции и т.д.

## Пример первый: Задача

Выявить на этапе тестирования узкие места

Определить сценарий поведения типичного пользователя

Определить соответствие реального поведения системы и  
запроектированной архитектуры

# Пример первый: методика анализа

Использовался коммерческий инструмент от Fluxicon:



**Discover** your processes.

<http://www.fluxicon.com/>

Внутри Disco – алгоритм нечеткого извлечения (Fuzzy Miner)

C. Gunther and W. Aalst. Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics. In G. Alonso, P. Dadam, and M. Rosemann, editors, International Conference on Business Process Management (BPM 2007), volume 4714 of Lecture Notes in Computer Science, pages 328-343. Springer-Verlag, Berlin, 2007.

**Позволяет выделять типичное поведение в процессе  
и отфильтровывать редко встречающееся**

# Пример первый: методика анализа

Фрагмент нечеткой модели процесса по логу за 1 день



Прямоугольник: activity (действие)

Цвет прямоугольника: доля трасс, в которых действие выполняется или, например, среднее время выполнения действия (выбирается)

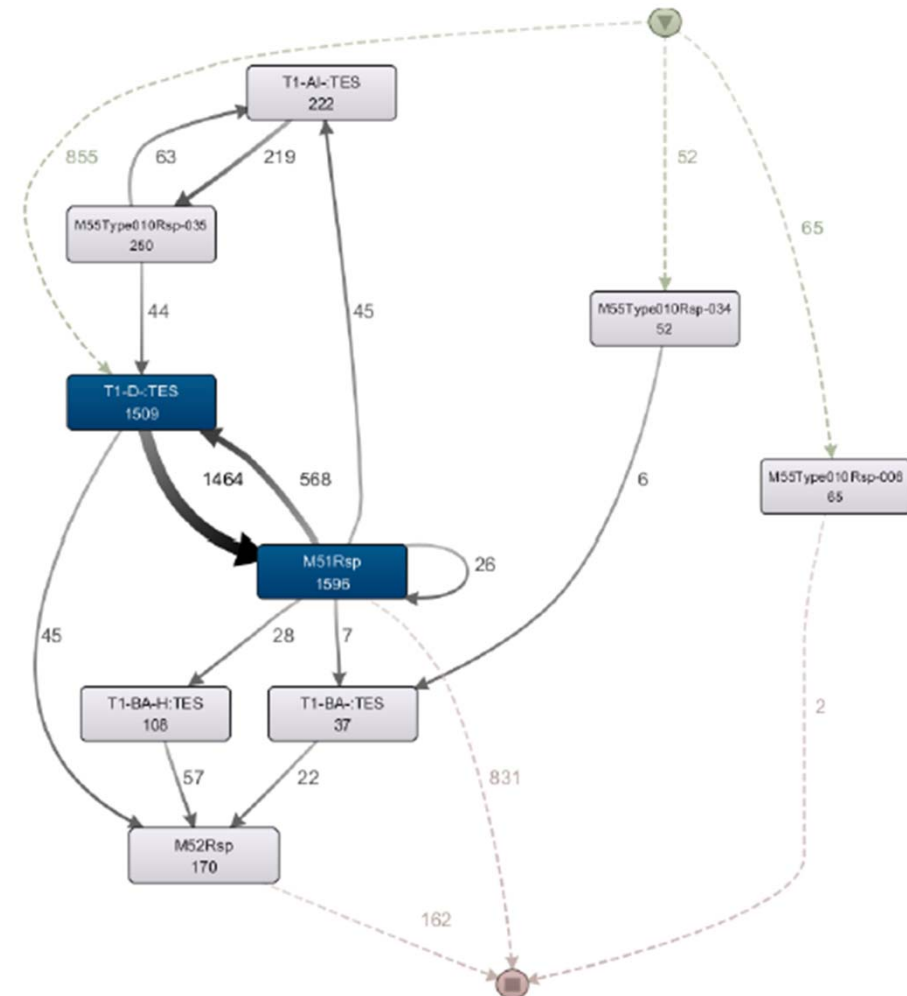
Дуги: последовательность действий в трассе

Толщина дуги: количество трасс, в которых действия идут друг за другом

# Пример первый

Модель наиболее часто-встречающегося поведения

- Часто исполняются действия запроса инфо об агентстве (T1-AI), бронирования разных типов (T1-D, T1-BA-H);
- M51 – код действия отправки сообщения об ошибке;
- Модель показывает, что очень часто в ответ на запрос бронирования отправляется сообщение об ошибке; оказалось, что имеет место ошибка в реализации модуля ПО.



Alexey Mitsyuk – HSE PAIS Lab / CEE SEC(R) 2014 / Moscow, Digital October, 23-24.10.2014



# Пример первый

Другая возможность: изучение всех вариантов исполнения процесса

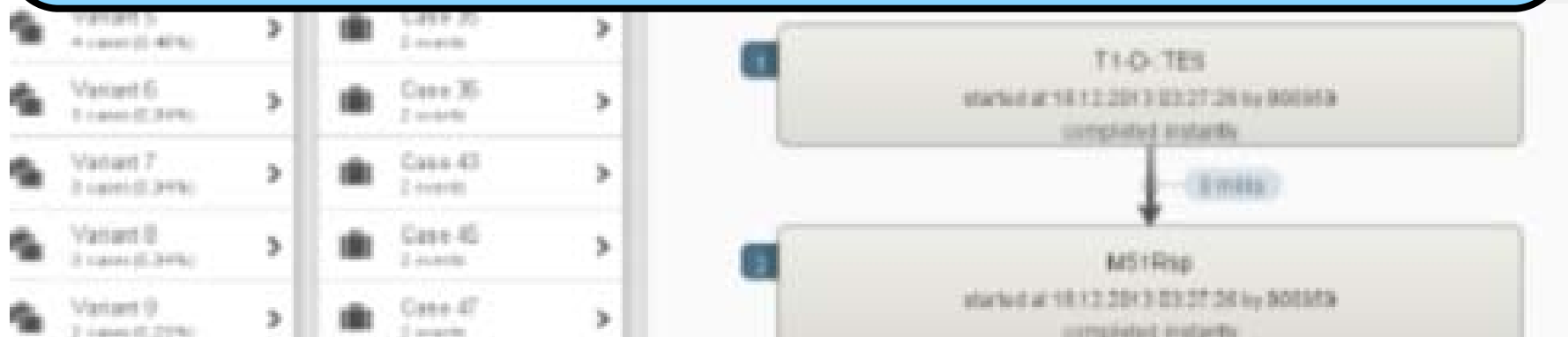
- Оказалось, что в ~65% негативных случаев (процесс не завершен успешно) пользователь после запроса бронирования получает сообщение об ошибке и уходит с портала, либо начинает новую сессию;
- Временные метки записей лога сверялись с исключениями, которые выдавались серверной частью системы и записывались в отдельный файл;
- Разработчики и поддержка до построения модели не могли понять, к какой части системы относятся исключения (содержимое стека вызовов Java-машины записывалось не полностью);
- Одного взгляда на модель разработчикам было достаточно, чтобы определить конкретные модули и исправить ошибку.

# Пример первый

Другая возможность: изучение всех вариантов исполнения процесса

Были выделены модели с позитивным поведением (процесс выполняется успешно для всех case'ов); таких оказалось большинство

В ходе дальнейшей разработки и усовершенствования разработчики стараются не затрагивать стабильные узлы, которые во всех case'ах исполняются верно



# Пример первый

## Модель длительностей исполнения (performance)

- Только запросы пользователя;
- Оказалось, что **большинство действий** выполняется в среднем за **десятки секунд**;
- **Некоторые** действия занимают **минуты**;
- Однако в **одном случае** средняя длительность между действиями оказалась **3.4 часа**;
- Это за туристические агенты, которые оставляют приложение запущенным, но не выполняют действий по 3-4 часа 😊

Среднее действие в GUI занимает 17 минут; а система отвечает на запрос не более, чем за 3 минуты.

Пользователи-агенты выполняют часто одни и те же операции поиска и бронирования, например, в случае тур-групп. Оказалось, что они каждый раз заново выполняют полный цикл: поиск, выбор позиции, бронирование. Разработчиками была добавлена функциональность для сохранения настроек поиска и бронирования, что существенно понизило нагрузку как вычислительную, так и сетевую.

Оказалось, что поведение агентов довольно **беспорядочно**: все выполняют операции в разных последовательностях. Что, в некоторой степени, определяется разным ПО.

## Пример второй: исходные данные

Российский портал по продаже электронных авиа-билетов

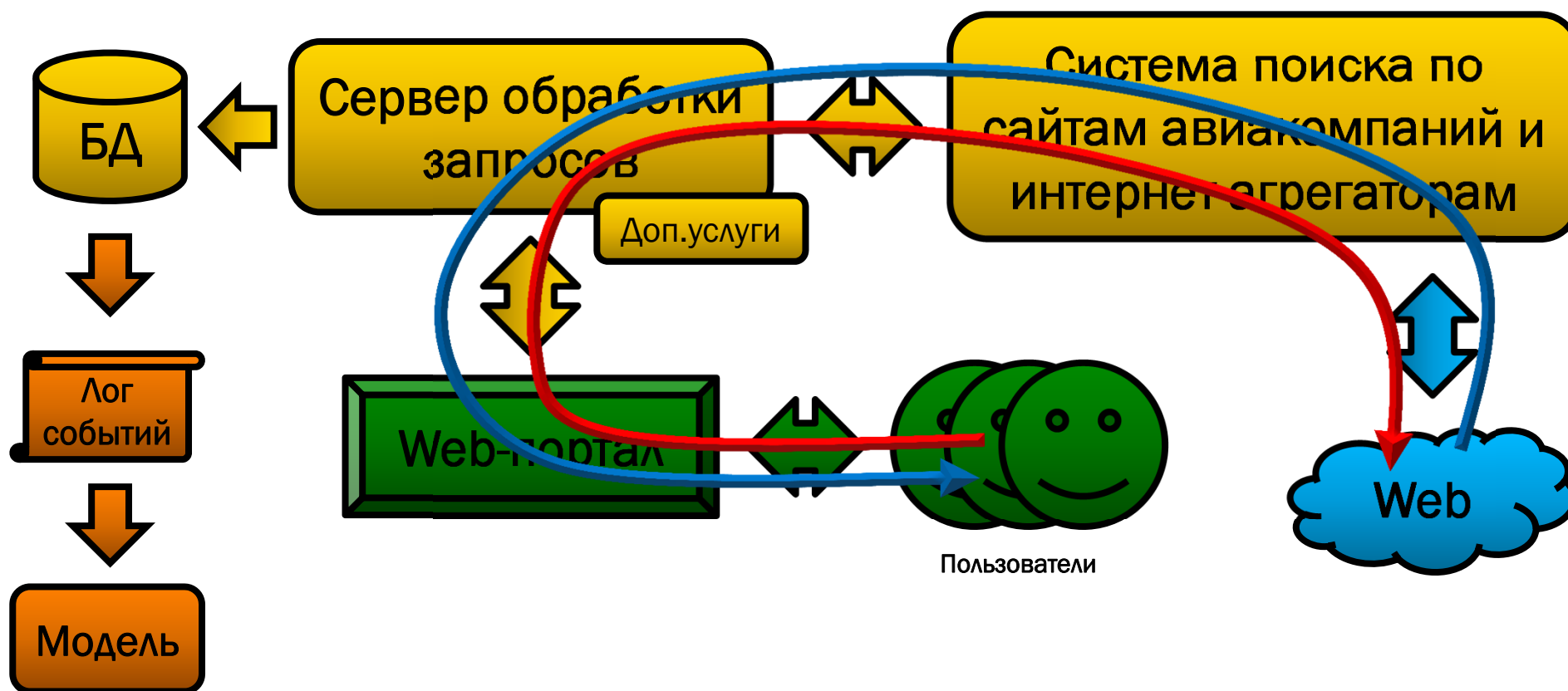
Web-портал помогает искать предложения от разных авиакомпаний

Позволяет осуществлять бронирование разных авиа-билетов, дополнительное страхование

Пользователи взаимодействуют с системой через сайт

Поведение пользователей на сайте записывается в базу данных, содержащую несколько таблиц, каждая из которых содержит определенную информацию

# Пример второй: исходные данные



## Пример второй: Задача

Согласно исследованию, средняя доля совершивших покупку из посетителей сайта, ниже, чем средняя по отрасли.

Почему?

Найти узкие места

## Пример второй: препроцессинг

БД

Лога событий в необходимом формате не было.

Для получения лога были использованы запросы SQL.

В качестве идентификатора экземпляра процесса был выбран идентификатор сессии.

По идентификатору сессии из таблиц выбирались соответствующие записи. В качестве действия были выбраны пары записей «действие пользователя» (click, scroll, select и т.д.) и «объект действия» (submit button, select booking button и т.д.)

Анализ проводился на данных за (примерно) месяц работы портала

Session ID	Activity
	OW-LO
	LE-CHE
	PT-CHE
	SUBMIT
1001014	WINDOW-UNL

Таким образом удалось получить журнал событий, содержащий все необходимое для Process Mining, а также кучу дополнительных данных (код заказа, направление, ОС заказчика и т.д.)

# Пример второй

Для анализа использовался ProM 6 Framework



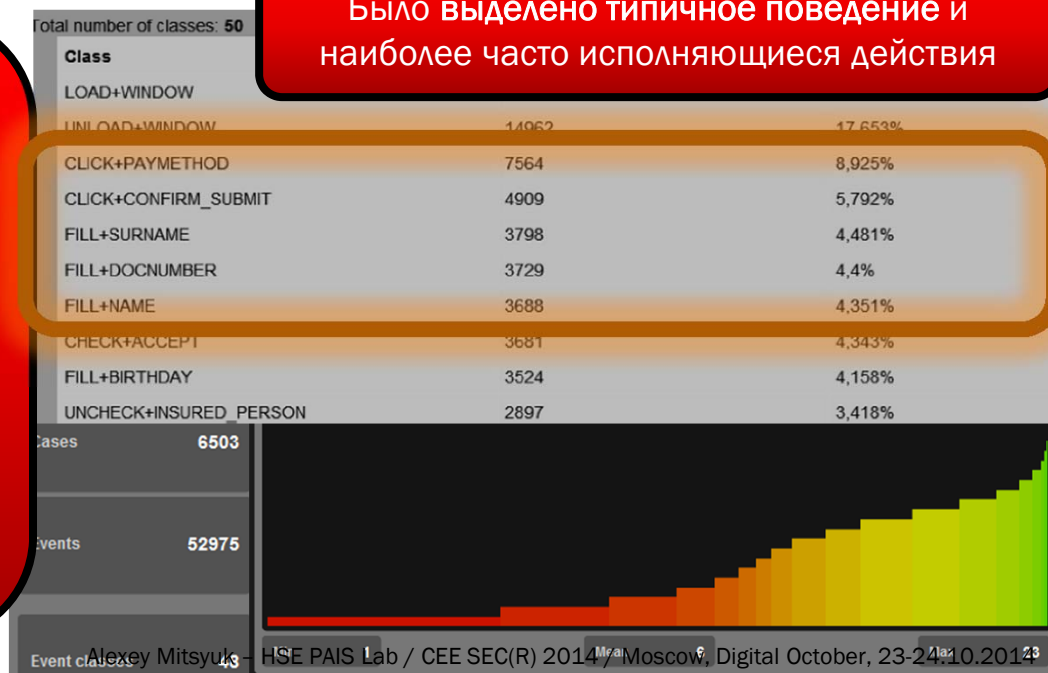
Прежде всего провели анализ лога

Очень удобно, что в нем содержались события загрузки и выгрузки страницы, отмечающие начало и окончание сессии.

Оказалось, что есть куча (около 40%) случаев, когда страница загрузилась, а потом выгрузилась через пол минуты.

Роботы были отфильтрованы!

Было выделено типичное поведение и наиболее часто исполняющиеся действия

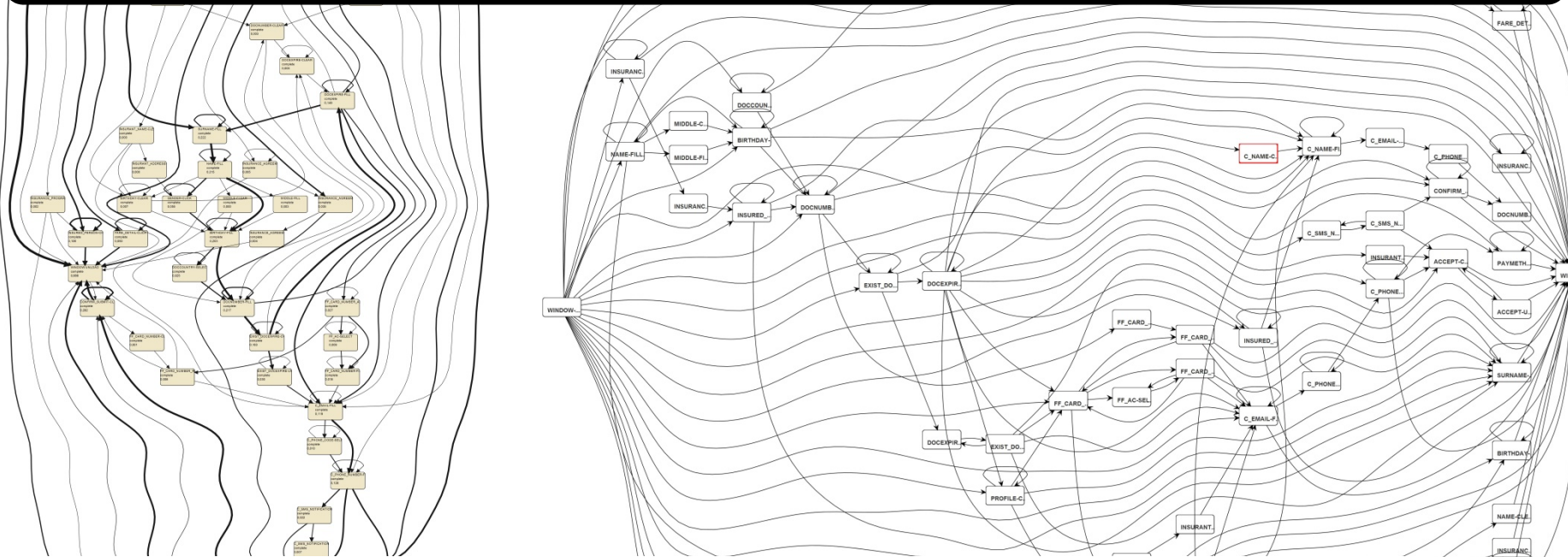






# Пример второй

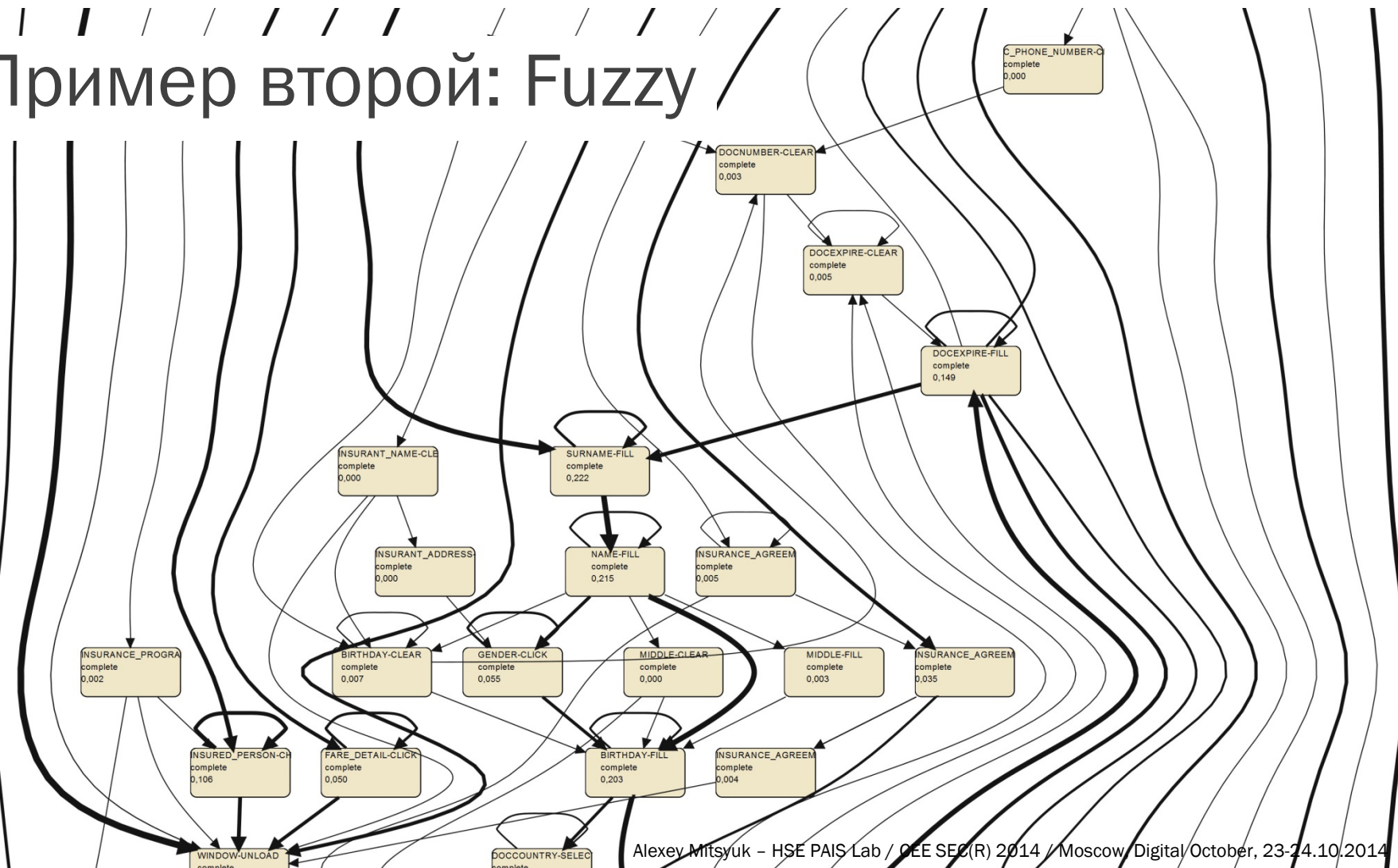
Модель поведения пользователей (без роботов) построена с помощью Fuzzy и Heuristics Miner



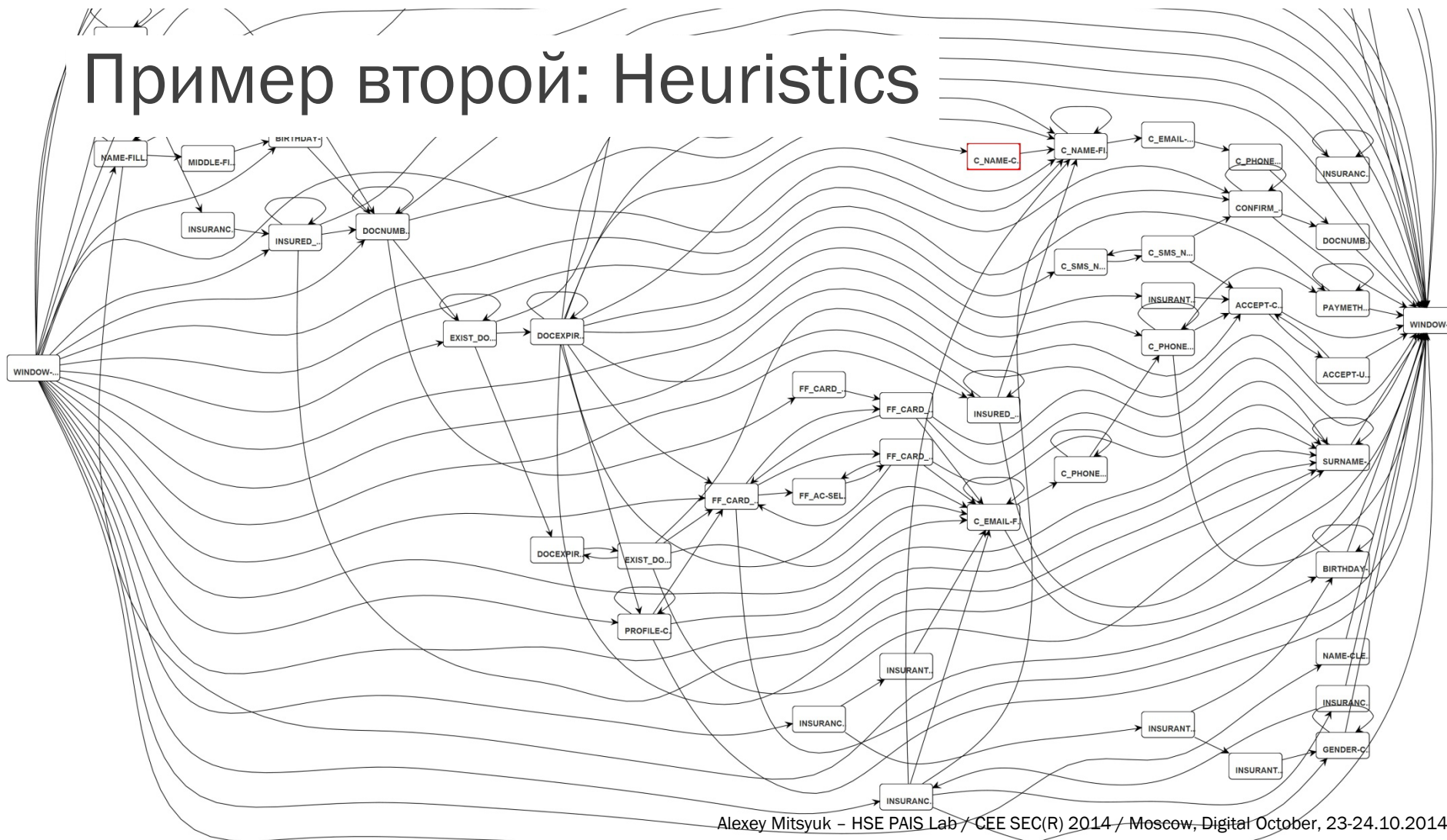
C. Gunther and W. Aalst. Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics. In G. Alonso, P. Dadam, and M. Rosemann, editors, International Conference on Business Process Management (BPM 2007), volume 4714 of Lecture Notes in Computer Science, pages 328-343. Springer-Verlag, Berlin, 2007.

A. Weijters, W. Aalst, and A. Medeiros. Process Mining with the Heuristics Miner-algorithm. BETA Working Paper Series, WP 166, Eindhoven University of Technology, Eindhoven, 2006.

# Пример второй: Fuzzy



# Пример второй: Heuristics



## Пример второй

### Что показывают эти жуткие модели?

- Последовательность действий пользователей (все варианты)
- Пользователи ведут себя по разному даже на линейно выстроенной странице
- Тем не менее, более половины ведут себя довольно единообразно (ввод даты и направления, поиск предложения, просмотр условий, ввод личных данных, выбор дополнительных опций, ввод платежных данных, покупка)
- Есть люди, которые платят без поиска предложения. Удивительно? Это тестировщики!
- Логи в день тестирования вообще отличаются от обычных: оказывается, тестировщики думают, что пользователи ведут себя не так, как они себя ведут 😊
- Есть около 20 случаев странного поведения (типа поиска десятков предложений или многократных попыток ввода различных платежных данных)

## Пример второй

Что пользователи делают чаще всего перед тем, как покинуть портал?

(1) Пользователи покидают портал, когда пытаются отправить свои данные. Почему? Оказывается, есть довольно много трасс в логе, которые завершаются многократным повторением события CONFIRM\_SUBMIT. Проблема решается расширением канала от web-портала к серверу, а также устранением ошибок в ПО.

Вообще, многие сценарии взаимодействия с сайтом были не предусмотрены разработчиками. Некоторые из элементов управления не работали в той последовательности, в которой их пытались использовать.

(2) Пользователи покидают портал, когда узнают условия оплаты. Почему? Оказалось, что недавно компания изменила условия продаж, а именно возврата средств при отмене заказа для некоторых типов билетов. Руководство не могло предположить, что усложнение порядка возврата так изменит картину покупок.

(3) Пользователи покидают портал, когда видят способы оплаты. Почему? Оказывается, портал испытывает проблемы с оплатой картами некоторых распространенных платежных систем, а также не принимает оплату через некоторые системы интернет-банкинга. Записи об этом уже даже есть в «жалобной книге» портала!

(4) Пользователи покидают портал, когда видят, что им предлагают дополнительно застраховаться. Почему? Здесь проблема неправильного проектирования GUI. Сначала пользователю показывался экран, где была отмечена доп. страховка, а снять её он мог только на следующем экране. Многие просто закрывали сайт.

## Пример второй

### Были предложены пути по исправлению ситуации

- Переработка GUI с изменением способа и порядка отображения предложений
- Изменение способа показа пользователю и выбора дополнительных услуг
- Исправление ошибок в ПО backend'a и повышение скорости отклика
- Усовершенствование схемы покупки
- Пересмотр тарифов и политики возврата билетов
- Расширение спектра поддерживаемых систем оплаты

## Что в итоге?

С помощью Process Mining можно синтезировать по логам информационных систем модели пользовательского взаимодействия с этими системами

Такие модели полезны администраторам/разработчикам и руководству компания для принятия технических или коммерческих решений

Модели позволяют выявить типичный способ взаимодействия пользователя с системой и сравнить его с запроектированным

Анализ моделей помогает выявить узкие места, ошибки, проблемы проектирования как back-, так и front- end составляющих ИС

## Что в итоге?

Разные системы, отличающиеся вопросы и условия,  
но схожая методика анализа

Анализ в обоих случаях принес реальную пользу

Поведение пользователей разное, но Process  
Mining дает усредненную модель

Можно *подсветить* исключительное или  
нежелательное поведение отдельно

Применение Process Mining для анализа поведения пользователей  
информационных систем зависит прежде всего от наличия  
**подходящих логов и от разумной постановки вопросов**



Спасибо за внимание!

[pais.hse.ru](http://pais.hse.ru)

Эффективный анализ поведения  
пользователей с применением  
Software Process Mining

Алексей Мицюк, Лаборатория ПОИС ФКН НИУ ВШЭ



Картинка: <http://en.wikipedia.org/wiki/File:Spaghetti Junction CA.jpg>