Visual Analytics: Illuminating the Employees' Paths

Евгения Новикова

С.н.с. Лаборатории проблем компьютерной безопасности СПИИРАН

План доклада

- Визуальная аналитика
  - Ключевые элементы
  - Процесс визуального анализа данных
  - Задачи и вызовы визуальной аналитики для решения задач информационной безопасности
- Исследование журналов систем контроля доступа методами визуального анализа

Мотивация

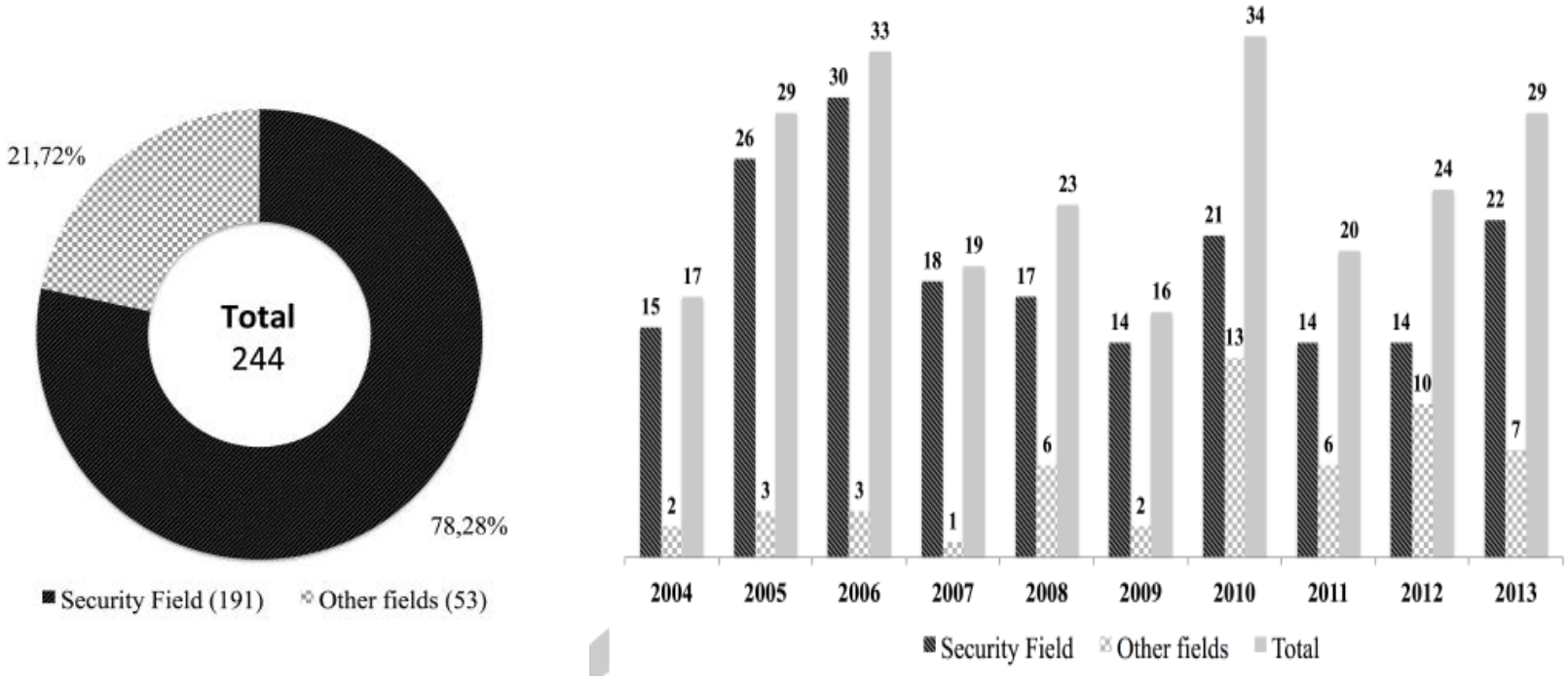
…Data visualization is the only approach that scales to the ever changing threat landscape and infrastructure configurations…

Р. Марти (2016)

Конференции

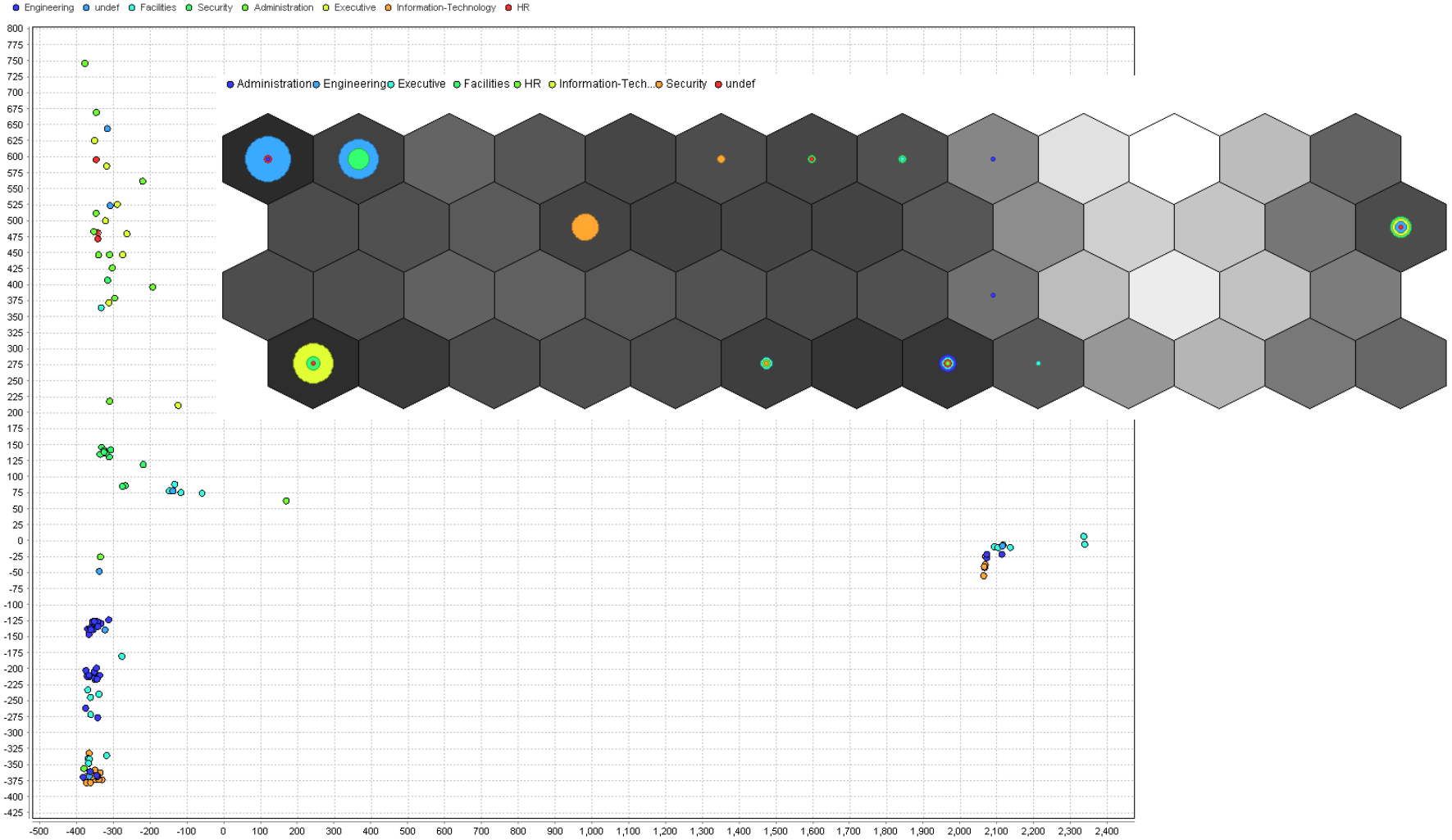
- IEEE Conference on Visual Analytics Science and Technology (IEEE VAST);
- IEEE Symposium on Visualization for Cyber Security (VisSec);
- Eurovis workshop on Visual Analytics

Актуальность



Guimarães V. T., Freitas C. M. D. S., Sadre R., Tarouco L. M. R. and Granville L. Z. A Survey on Information Visualization for Network and Service Management // *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 285-323, Firstquarter 2016.

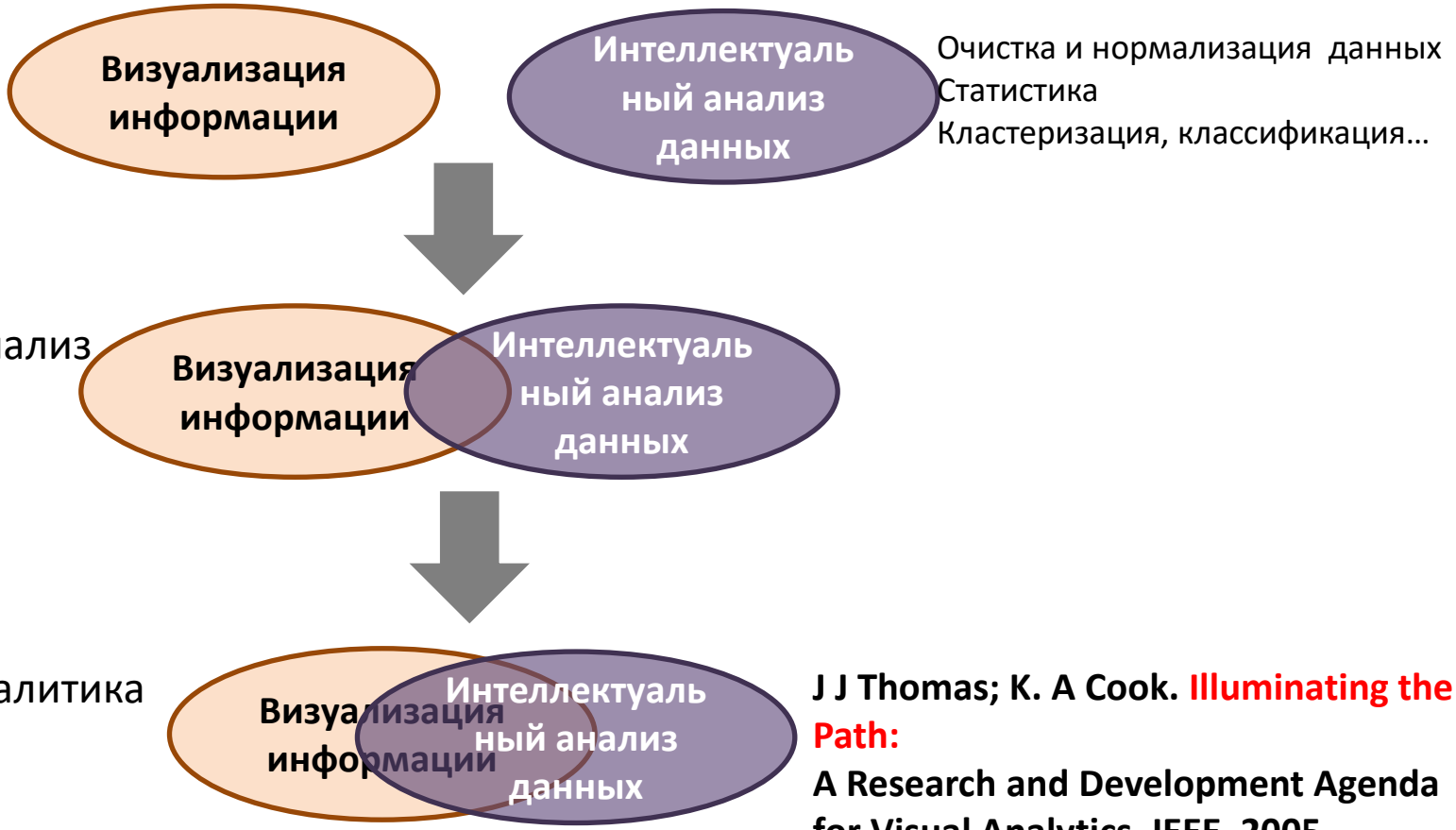
Визуализация данных vs визуальная аналитика



Визуальная аналитика (ВА)

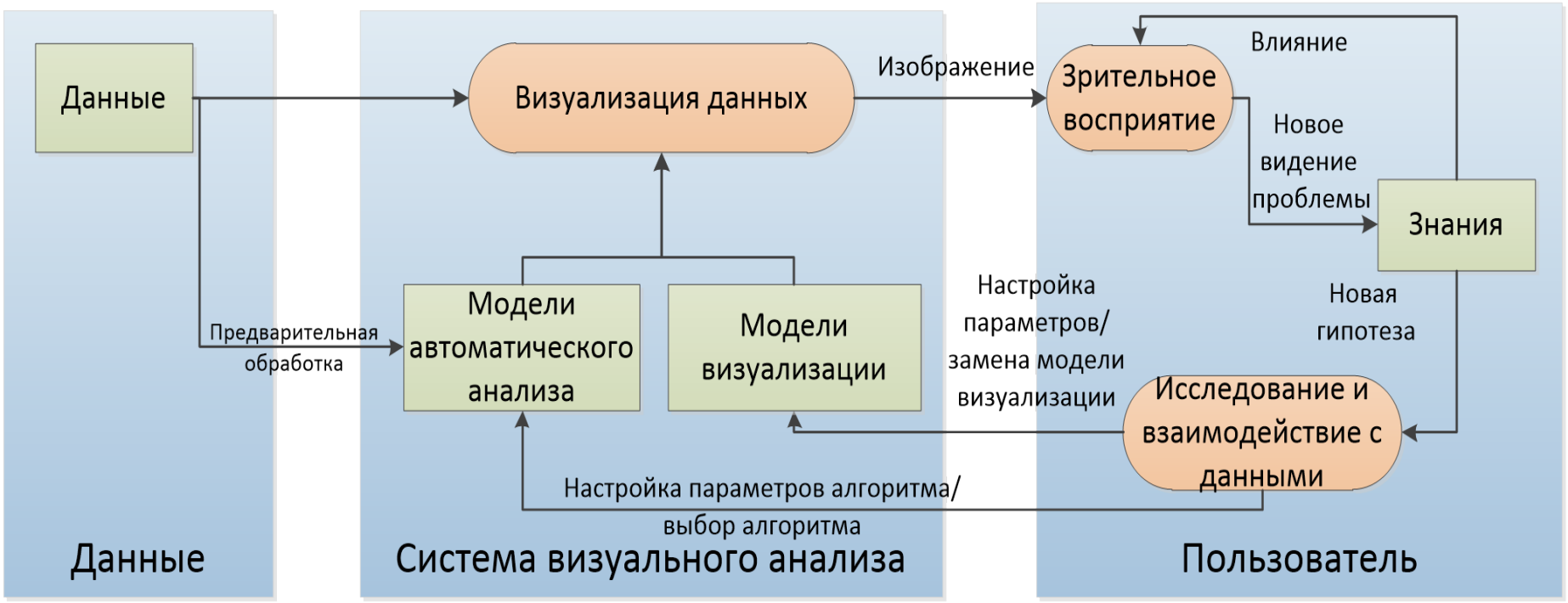
Визуальная аналитика – это междисциплинарное научное направление, формируемое на стыке когнитивной графики и автоматического анализа данных.

Этапы развития BA



J J Thomas; K. A Cook. Illuminating the Path:
A Research and Development Agenda for Visual Analytics, IEEE, 2005

Процесс анализа данных методами ВА



Задачи ВА

Зависит от **роли пользователя** информационной системы

- представление информации публике, обоснование полученных результатов;
- мониторинг текущей ситуации (оперативный контроль);
- исследование данных (исторический анализ данных);
- **верификация корректности работы моделей автоматического анализа.**
  - визуализация конечного и промежуточных результатов автоматических моделей, алгоритмов и методов анализа
  - 2015 - оценка корректности функционирования моделей автоматического анализа (S. Walton, E. Maguire, M. Chen. A visual analytics loop for supporting model development // Proceeding of *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Chicago, IL, 2015, pp. 1-8.)

Классификация методик визуального анализа

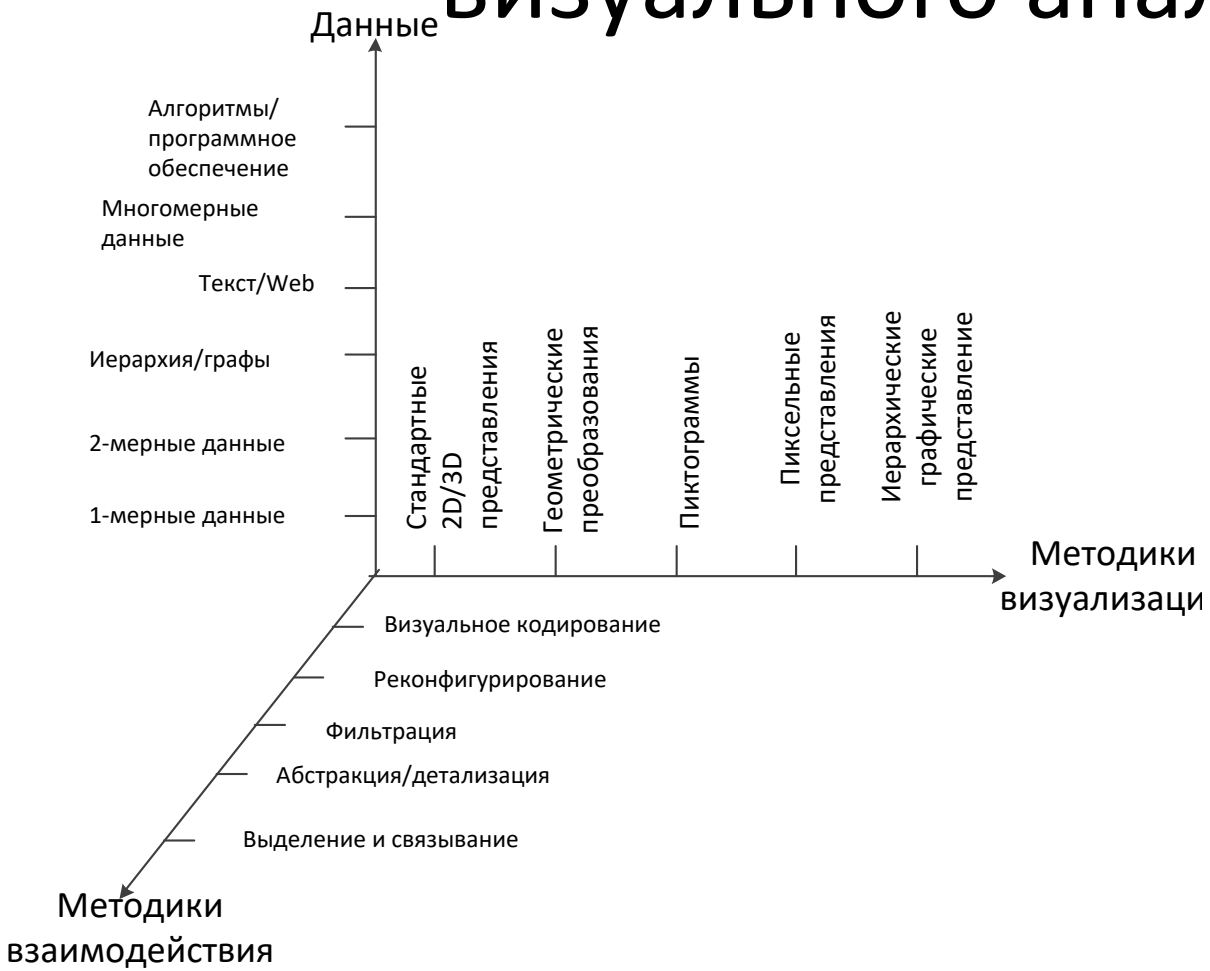
Определение 1. Модель визуализация VM − это множество, состоящее из трех сущностей: $VM = \{D, vt, IT\}$, где

- D − данные,
- $vt : D \to V$ − методика визуализации
- $IT = \{it_i : (V, D, vt) \longrightarrow V\}_{i=0}^{n}$ − множество методик взаимодействия

IT, определенными для заданной методики визуализации vt в визуальном пространстве V.

Keim D., Ward M.. Visual Data Mining Techniques // Intelligent Data Analysis. Chapter 11. Springer Verlag, 2 edition, 2002. Pp. 403-427

Классификация методик визуального анализа



Keim D., Ward M.. Visual Data Mining Techniques // Intelligent Data Analysis. Chapter 11. Springer Verlag, 2 edition, 2002. Pp. 403-427

Модель анализа ВА

Определение 2. Модель визуального анализа VAM – это множество, состоящее из четырех взаимосвязанных сущностей:

$VAM = \{D, VT, DM, IT\}$, где

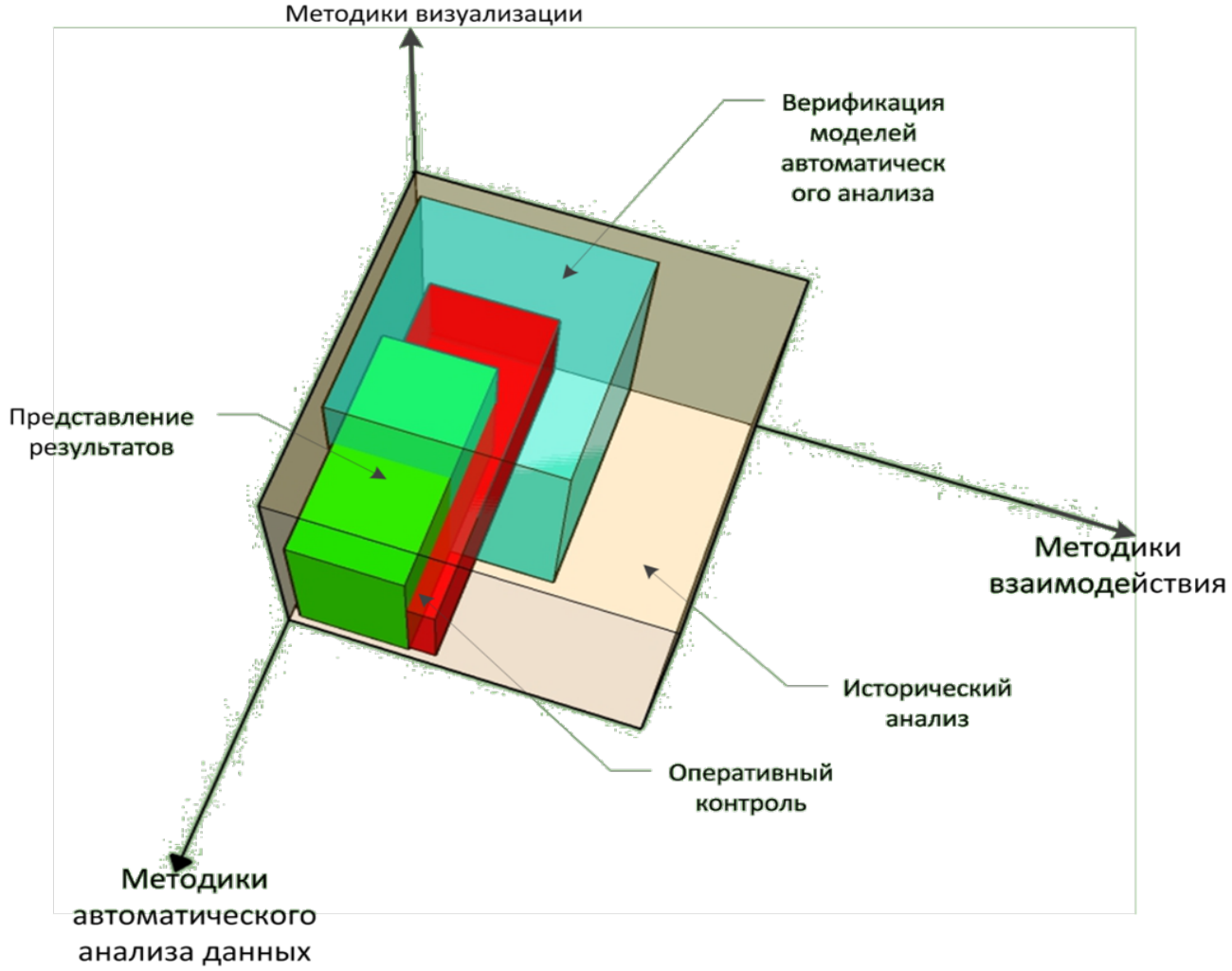
- D – данные,
- $VT = \{vt_i; vt_i : D \longrightarrow V\}_{i=1}^{n}$ – множество методик визуализации данных, преобразующих пространство данных D в визуальное пространство V;
- $DM = \{dm_j; dm_j : D \longrightarrow D'\}_{j=1}^{l}$ – множество моделей автоматического анализа данных, позволяющих преобразовать исходное множество данных D в пространство D′, дополнив его данными об исходном множестве (статистические данные, информация о его структуре и т.д.);
- $IT = \{it_i; i = 1 \div n; it_k : (V, D, VT, DM) \longrightarrow V'\}_{k=0}^{m}$ – множеством методик взаимодействия IT, определенными над заданными множеством данных D, множеством методик визуализации VT , множеством моделей автоматического анализа DM, в визуальном пространстве V.

Задачи ВА

Зависит от **роли пользователя** информационной системы

- представление информации публике, обоснование полученных результатов;
- мониторинг текущей ситуации (оперативный контроль);
- исследование данных (исторический анализ данных);
- **верификация корректности работы моделей автоматического анализа.**
  - визуализация конечного и промежуточных результатов автоматических моделей, алгоритмов и методов анализа
  - 2015 - оценка корректности функционирования моделей автоматического анализа (S. Walton, E. Maguire, M. Chen. A visual analytics loop for supporting model development // Proceeding of *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Chicago, IL, 2015, pp. 1-8.)

Классификация моделей ВА



Применение методик ВА для решения задач ИБ

- оперативный контроль периметра компьютерной сети;
- **обнаружение внутренних нарушителей в режиме реального времени и на основе анализа исторических данных;**
- оценка уровня защищенности каждого хоста в отдельности и всей сети в целом;
- исследование инцидентов безопасности, включая киберпреступления;
- формирование отчетов различного уровня и типа;
- изучение вредоносного программного обеспечения;
- определение признаков атак для формирования правил их обнаружения;
- проведение аудита безопасности, включая оценку и верификацию политик безопасности;
- верификация автоматических моделей анализа, используемых для управления информационной безопасностью.

Вызовы ВА для ИБ

- большой объем данных;
- многообразие источников данных;
- отсутствие связи (синхронизации) между источниками данных;
- качество данных;
- **формирование паттерна нормального поведения ИС;**
- **отслеживание развития информационных угроз (реагирование)**

Обнаружение внутреннего нарушителя

- Основные источники данных
  - 1) логи HTTP протокола;
  - 2) логи сетевых потоков;
  - 3) логи файловой системы,
  - 4) журналы систем управления базами данными,
  - 5) логи почтовых серверов;
  - 6) данные операторов сотовой связи ;
  - 7) журналы различных приложений, таких как MS Word, MS Power Point, MS Excel, JPG, and TXT
- Специализированное программное обеспечение, осуществляющее мониторинг активности сотрудников внутри организации, движения транспортных средств в основном предназначено
  - расчет проводимого времени на рабочем месте,
  - оценка динамики опозданий,
  - продуктивность работы

Iffat A. Gheyas and Ali E. Abdallah. Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis // Big Data Analytics, vol 1 (6). 2016

Цели и задачи исследования

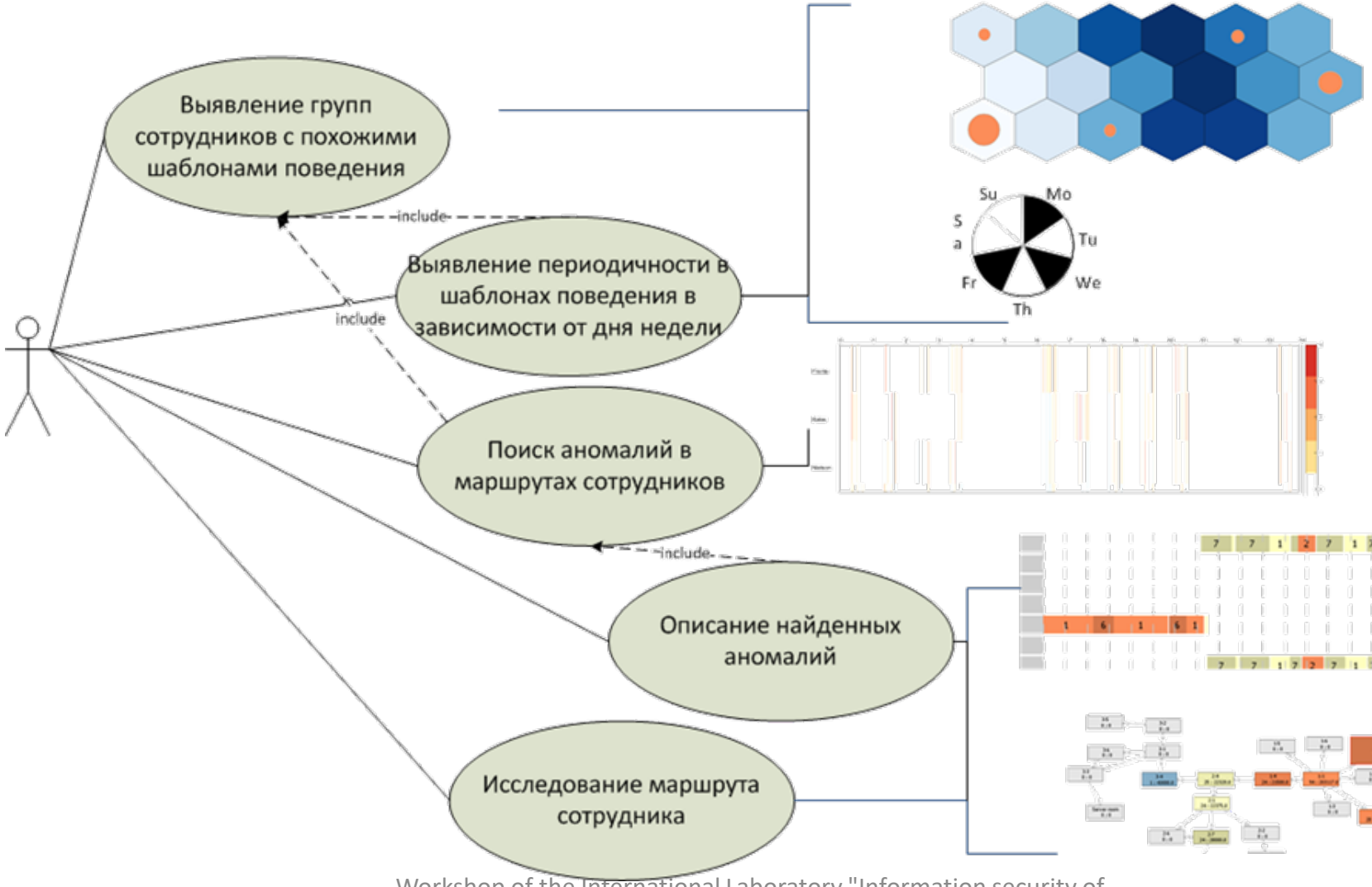
- Цель

  поддержка и соблюдение

  политик безопасности и регламента

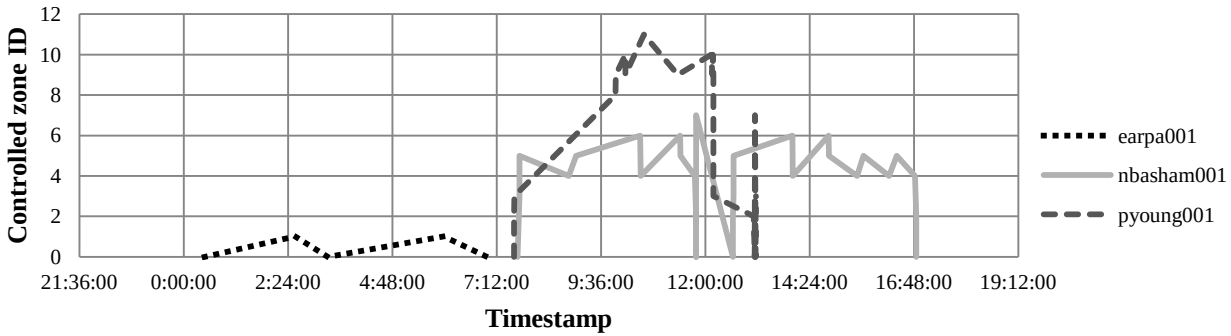
  технологических работ
- Задачи
  - Разработать систему визуального анализа
    - Определять шаблоны поведения сотрудников в условиях, когда их должностные обязанности неизвестны
    - Выявлять аномальные действия в перемещениях сотрудников
    - Исследовать журналы системы контроля доступа

Задачи системы визуального анализа журналов систем контроля доступа



Этап предварительной обработки данных

- Общий формат исходных данных
  - **Журнал <timestamp, employee ID, controlled zone ID>**
  - Данные о должностях и рабочих зонах сотрудников
  - Планы этажей
- Появление записей с proximity-датчиков носит нерегулярный характер
- Интервал между записями для одного сотрудника может длиться от нескольких секунд до нескольких часов
- Дискретизация траекторий в последовательности фиксированной длины



Этап предварительной обработки данных

$E = \{e_i\}_{i=1}^{n}$ - множество сотрудников

$Z = \{z_j\}_{j=1}^{m}$ - множество зон

$T = \{t_k : t_i < t_j; i < j\}_{k=1}^{p}$ - множество упорядоченных врем. инт.

$LOGS = \{(e_i, z_j, t_k)\}, i = 1 \div n,\ j = 1 \div m, k = 1 \div p$

- множество записей журналов датчиков контроля доступа

Процесс дискретизации: T_0 – наблюдаемый период времени

$T_0 = \{\Delta t_l : \Delta t_i = \Delta t_j; \Delta t_i = [t_i; t_{i+1}); \Delta t_{i+1} = [t_{i+1}; t_{i+2}); i \neq j; i, j \leq l;\}_{l=1}^{r}$

Для каждого сотрудника, зоны и интервала времени z_j , Δt_l , e_i

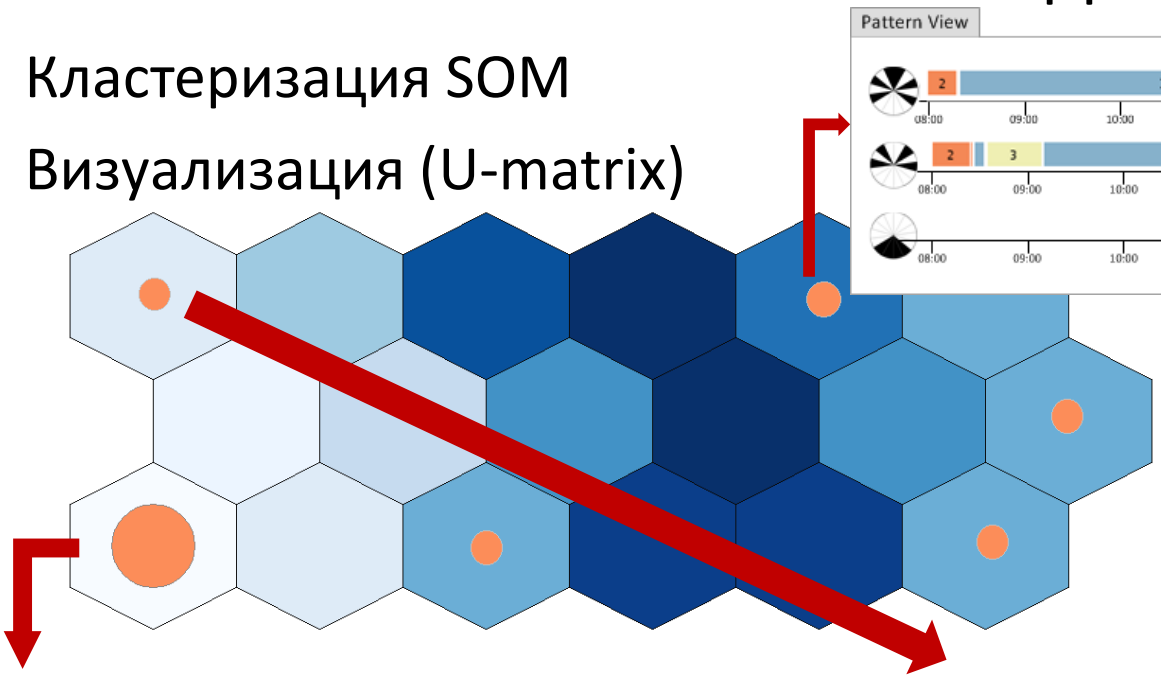
$n_{z_j}^{\Delta t_l}$ - Количество визитов в зону z_j в течение интервала Δt_l

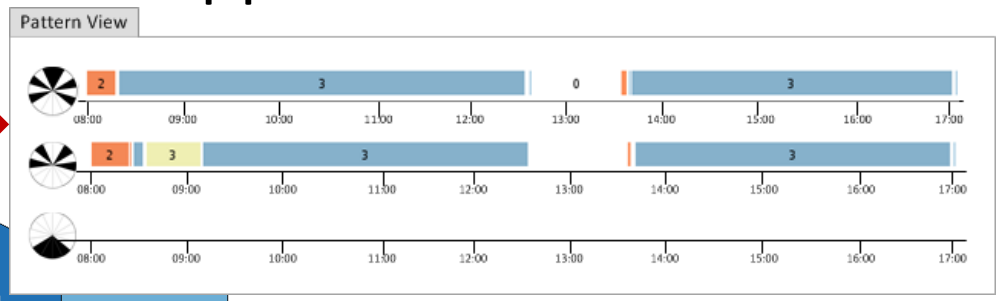
$\Delta t_{z_j}^{\Delta t_l}$ - Продолжительность нахождения в зоне z_j в течение Δt_l

$LOGS = \{(n_{z_j}^{\Delta t_l}; \Delta t_{z_j}^{\Delta t_l})\}_{e_i}, i = 1 \div n,\ j = 1 \div m, l = 1 \div r$

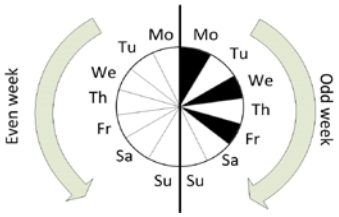
Определение групп сотрудников с похожим поведением и выявление периодичности в шаблонах поведения

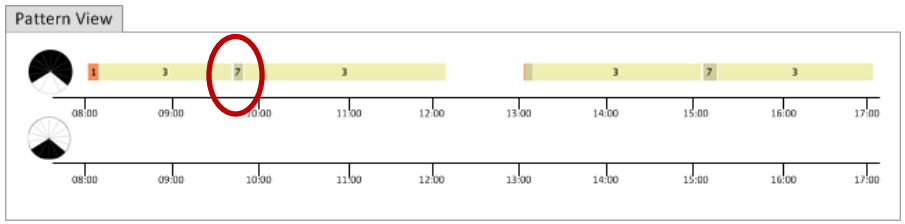
- Кластеризация SOM
- Визуализация (U-matrix)

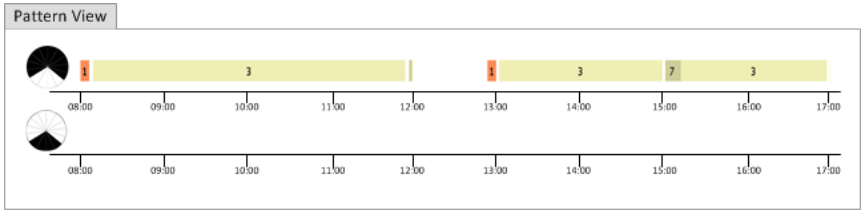




Глиф WeekCircle

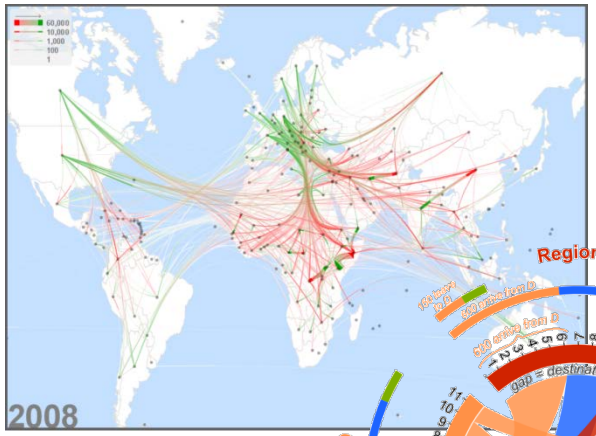




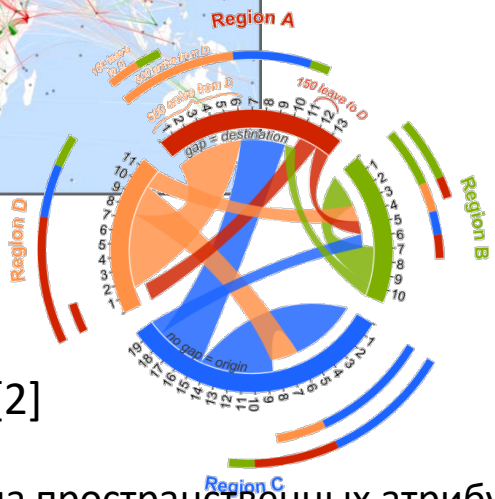


Визуализация траекторий

I Географические карты



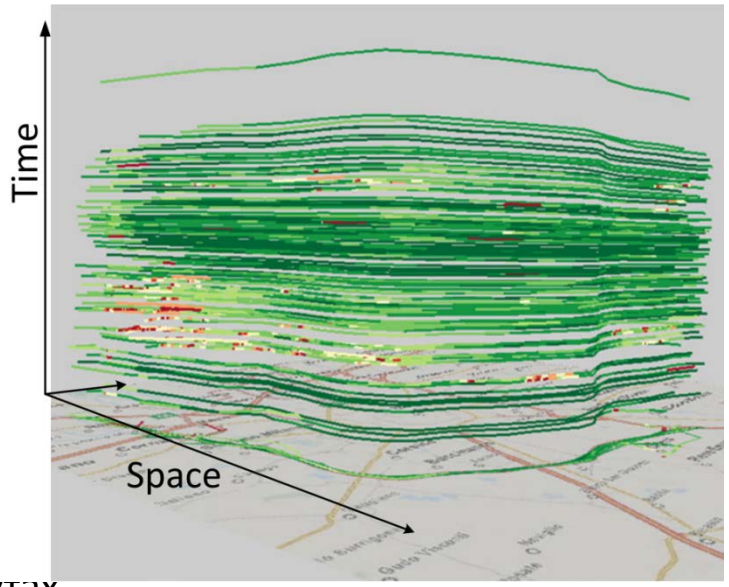
[1]



[2]

- Сосредоточен на пространственных атрибутах маршрутов
- Не может отображать пространственные и временные атрибуты маршрутов одновременно

II Визуализация траекторий с помощью пространственно-временного куба [3]

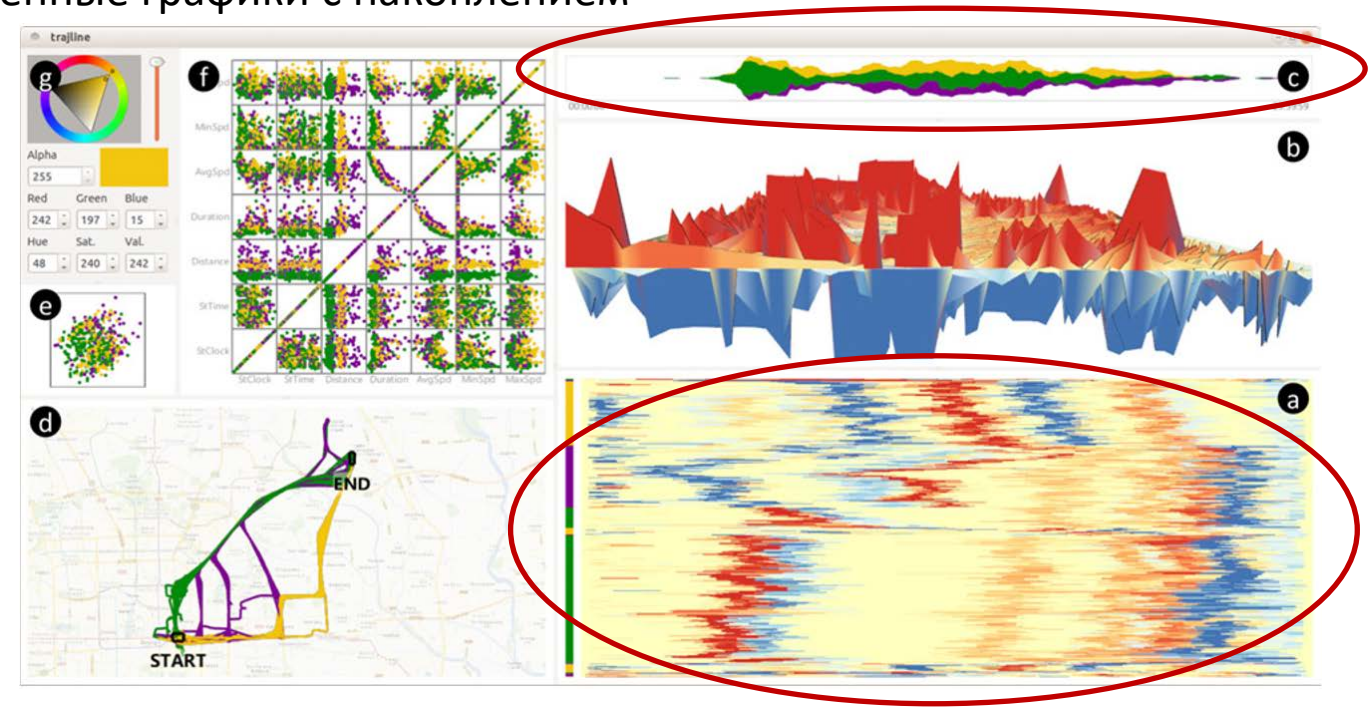


- Отображает пространственные и временные атрибуты маршрутов одновременно s
- Страдает от загромождения отображаемых траекторий

[1] I. Boyandin, E. Bertini, D. Lalanne, “Using Flow Maps to Explore Migrations Over Time. In: Proceedings of Geospatial Visual Analytics Workshop in conjunction”, the 13th AGILE International Conference on Geographic Information Science (GeoVA), 2010.
[2] N. Sander, J. Abel, R. Bauer, J. Schmidt, “Visualizing Migration Flow Data with Circular Plots”, 2014.
[3] G. Andrienko , N. Andrienko, H. Schumann, C. Tominski, “Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall”, 2014, pp. 157-163.

Визуализация траекторий

III Временные графики с накоплением



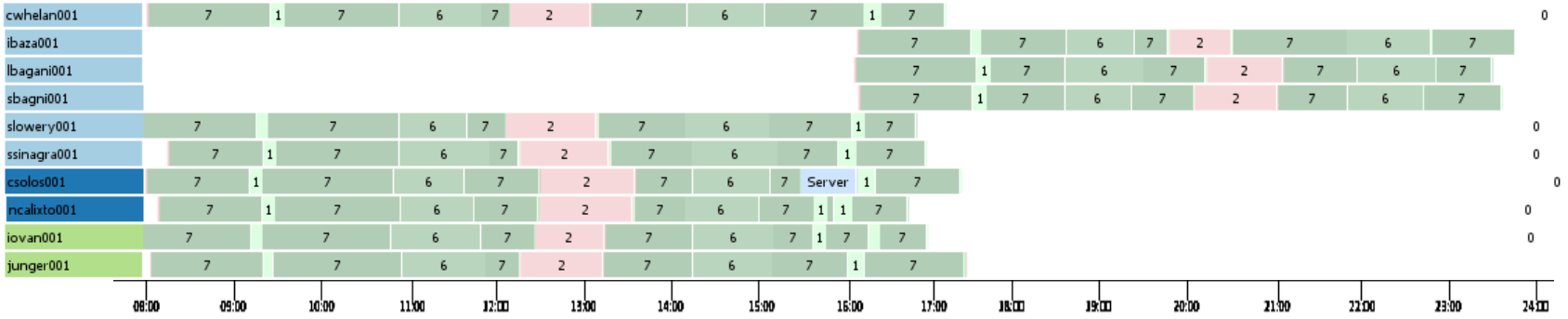
Z. Wang, X. Yuan, “Urban Trajectory Timeline Visualization”, International Conference on Big Data and Smart Computing (BIGCOMP), 2014.

Анализ маршрутов сотрудников

Модель визуализации BandView

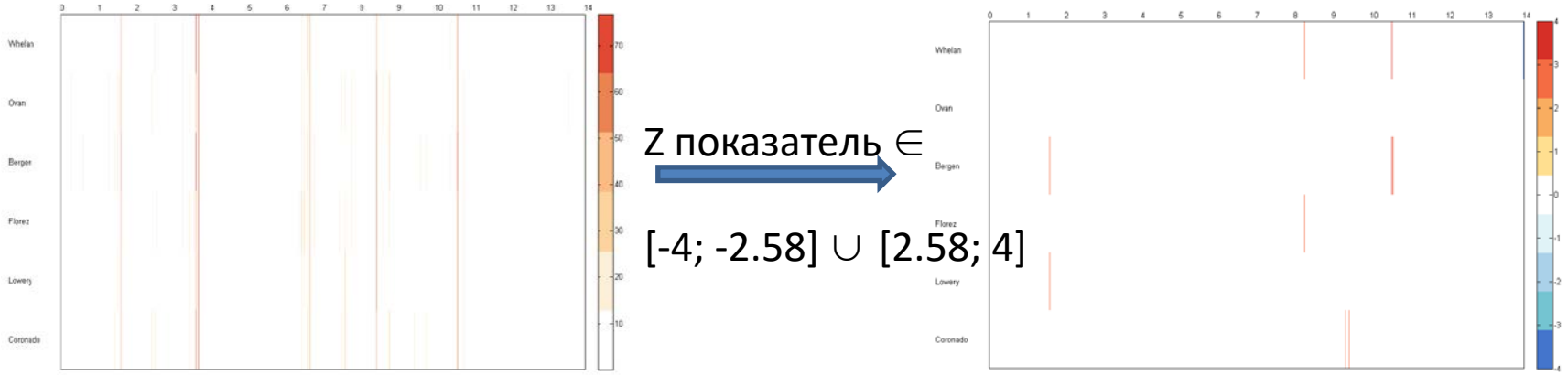
- Модель визуализации на основе графика с накоплением
- Используется для представления
  - Сырых данных
  - Шаблонов поведения
  - Аномалий
- Отображает пространственные и временные атрибуты перемещений одновременно

  - Окрашенные полоски соответствуют частям траектории
  - Цвет соответствует пространственным атрибутам зоны
  - Длина окрашенного сегмента определяется продолжительностью пребывания в контролируемой зоне



Поиск аномалий

- Вычисление расстояний атрибутов траектории от центроида:
- Механизм оценки отклонений на основе z-показателя
  - [-4; -2.58] ∪ [2.58; 4] - потенциальные аномалии
  - [-1.65; 1.65] ожидаемые значения
- Отображение с помощью тепловых карт
  - Механизм фильтрации по z – показателю, времени, контролируемой зоне



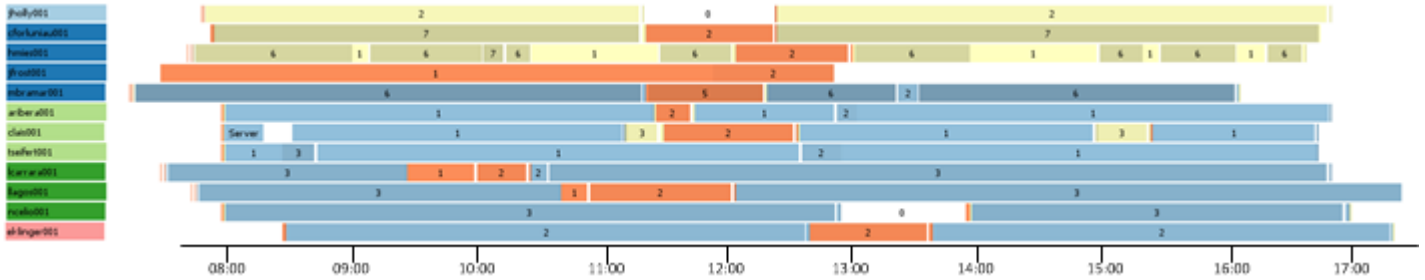
Эксперименты и оценка эффективности

- Набор данных Mini-Challenge 2 of the VAST Challenge 2016
  - <*timestamp, type, prox-id, floor, zone*>
  - План этажей с контролируемыми зонами
  - 8 отделов
- Логи системы контроля доступа небольшой компании по разработке ПО в СПб
  - Система отключается с 9.00 до 18.00
  - Работают только турникеты на входе/выходе

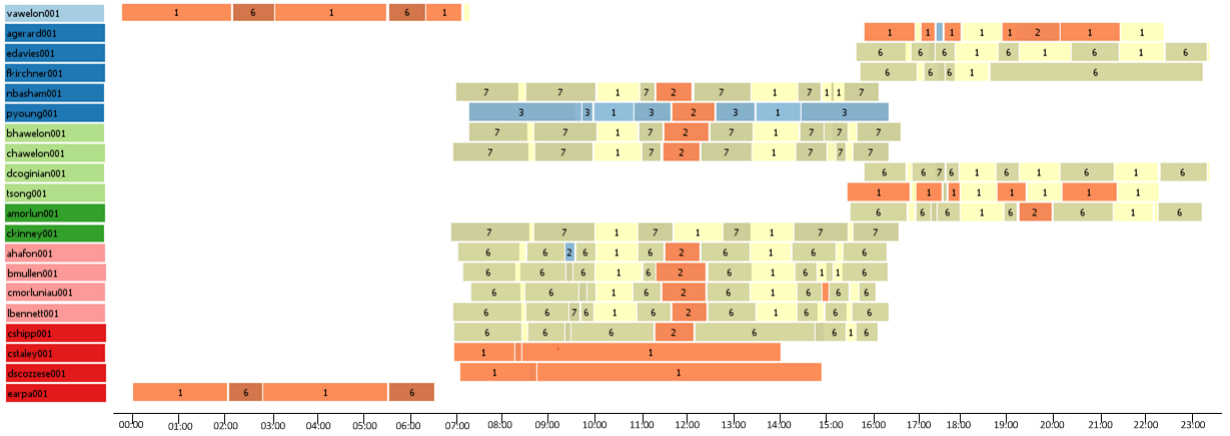
Эксперименты и оценка эффективности

- Примеры маршрутов

Отдел администрации



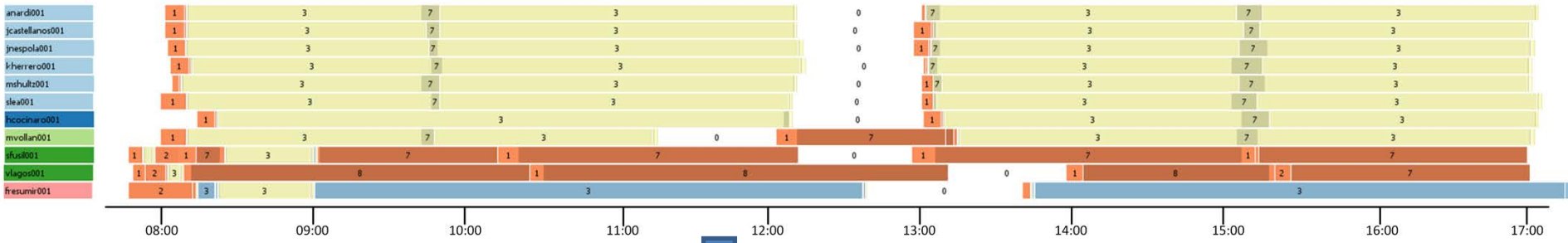
Отдел Facility

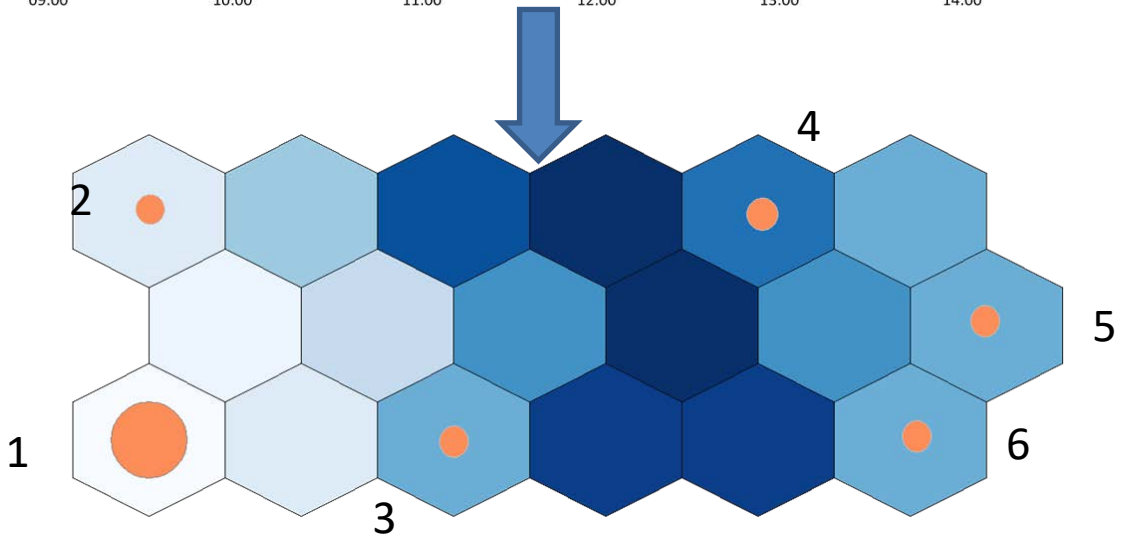


Эксперименты и оценка эффективности

- Отдел безопасности

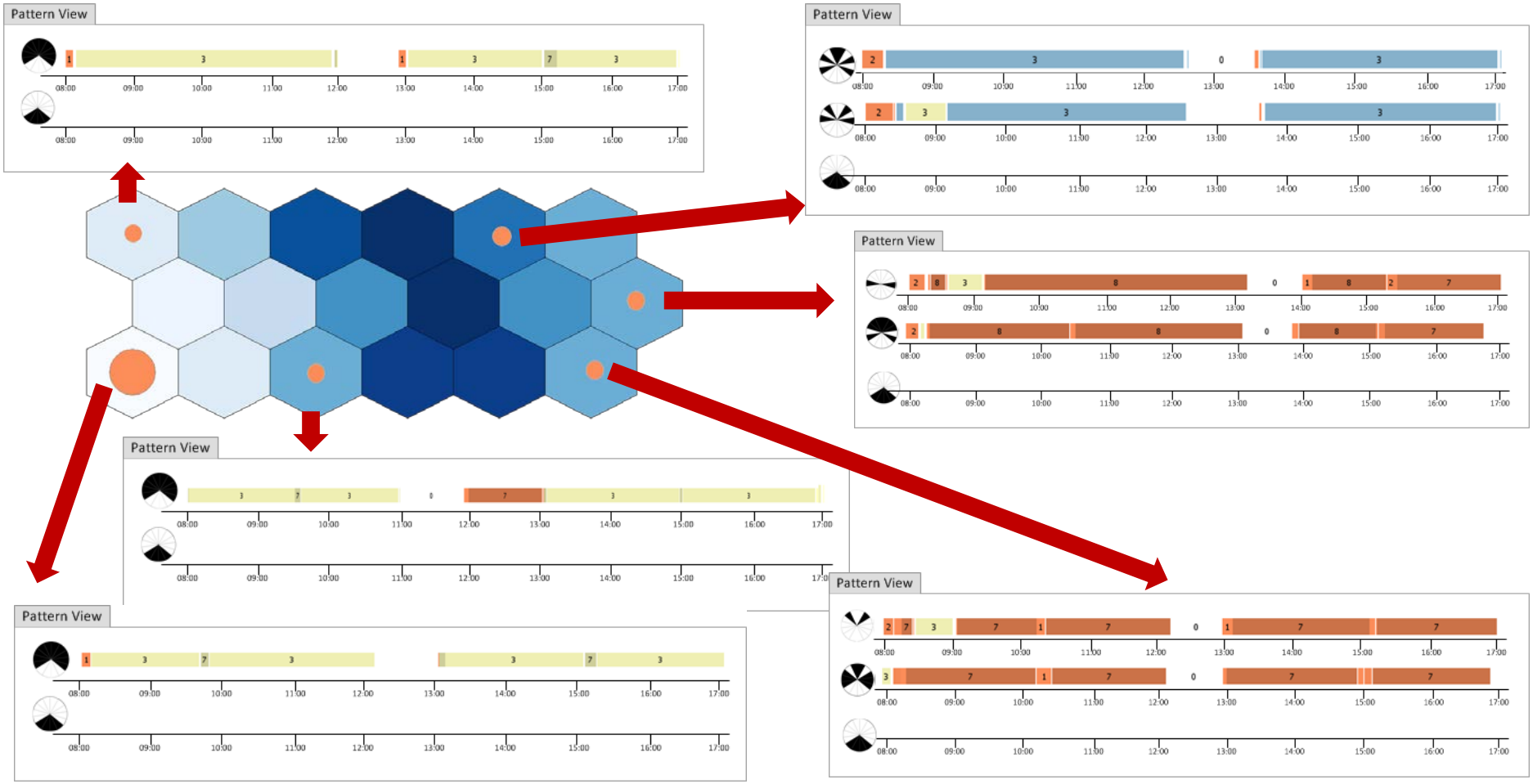
BandView – визуализация журнала за 31.05





Эксперимент

- Шаблоны маршрутов сотрудников безопасности

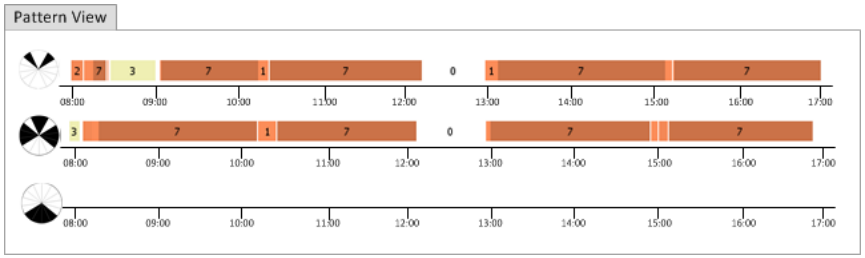


Типы обнаруженных аномалий

- работник теряет свой электронный пропуск (он может восстановить ее в тот же в тот же день, на следующий день);
- сотрудник забывает использовать свой электронный пропуск;
- работник проводит нестандартное время в типичной для него зоне
  - нестандартная длительность нахождения в контролируемой зоне может быть как слишком длинной, так и слишком короткой;
  - сотрудник может выйти на работу в выходные дни
  - нестандартная длительность может быть связана с отсутствием посещения контролируемой зоны в определенное время интервал;
- сотрудник посещает нестандартную контролируемую зону;
- два сотрудника используют одну бесконтактную карту.

Поиск аномалий:посещение нетипичной зоны

Шаблон поведения для сотрудника снизу-справа

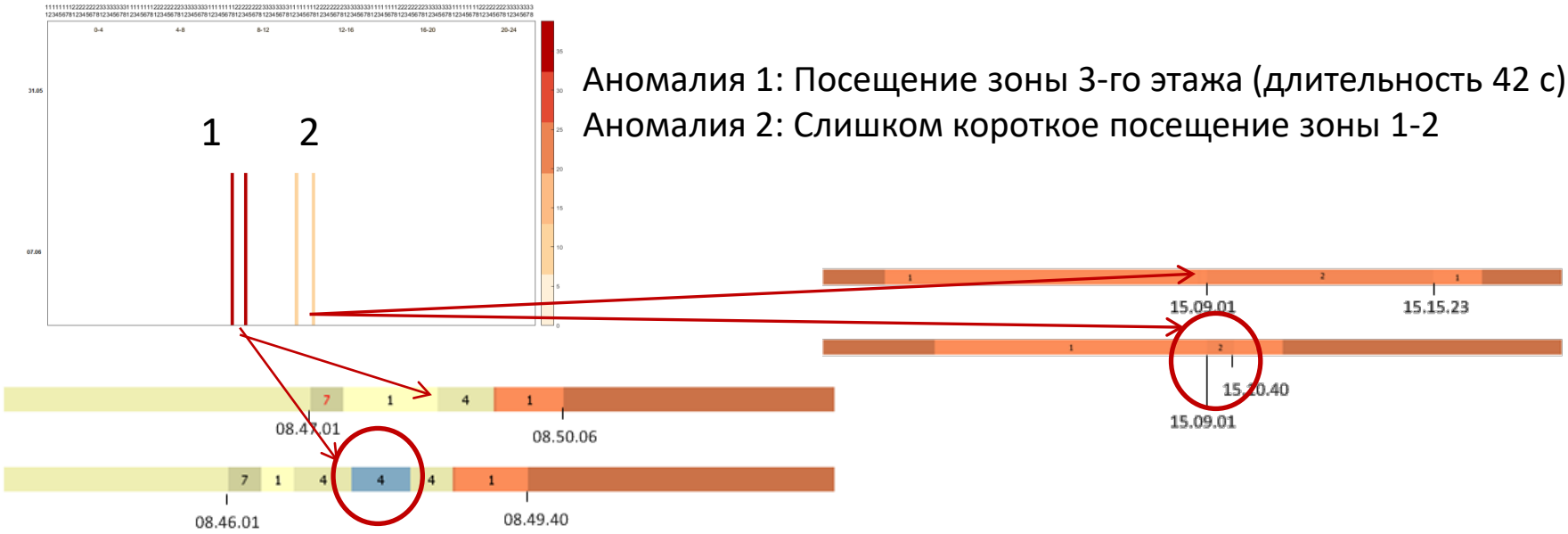


Вторник

Остальные рабочие дня

Выходные

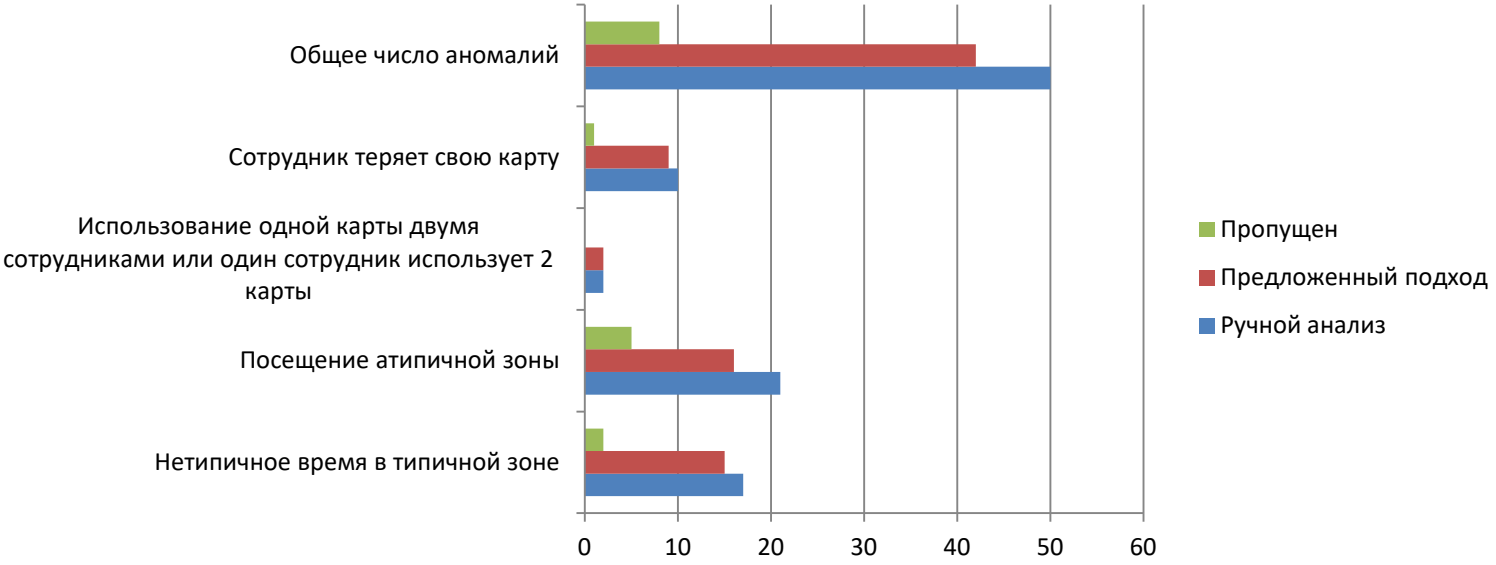
Обнаруженные аномалии



Аномалия 1: Посещение зоны 3-го этажа (длительность 42 с)

Аномалия 2: Слишком короткое посещение зоны 1-2

Эксперименты и оценка эффективности



$$\Pr ecision = \frac{tp}{tp + fp} = 0.79$$

$$\mathrm{Re}\, call = \frac{tp}{tp + (N - tp)} = 0.84$$

$$F - measure = 2\frac{\Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} = 0.81$$

Результаты

- Разработана методика визуального анализа, которая позволяет
  - Выявлять группы сотрудников, имеющих одинаковые паттерны передвижения
  - Выявлять периодичность в движении сотрудников
  - Выявлять возможные аномалии и исследовать их характер

Дальнейшая работа

- Разработка методики анализа существующих механизмов взаимодействия между сотрудниками
- Разработка методик корреляции данных, полученных от различных источников данных (видео наблюдение, системы жизнеобеспечения здания)
- Разработка моделей анализа для мониторинга передвижений сотрудников в режиме реального времени
- доработка программного прототипа системы визуального анализа, выполнение оценки эффективности

Спасибо за внимание!

- Вопросы?

- Контакты

Новикова Евгения Сергеевна

novikova@comsec.spb.ru