АНАЛИТИЧЕСКОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ SAS

СОДЕРЖАНИЕ (ЧАСТЬ 1)

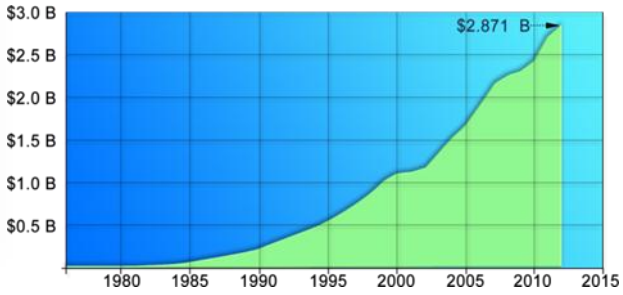
1. Интеллектуальный анализ данных: определения, задачи, методы. SAS Enterprise Miner: обзор возможностей и концепция SEMMA.
2. Задача прогнозирования (классификация, ранжирование, регрессия). Оценка качества моделей прогнозирования. Проблема переобучения. Методы формирования обучающих и контрольных выборок, sampling. Проклятие размерности, выбор значимых переменных.
3. Разведочный анализ данных: визуализация, поиск ассоциативных правил, анализ путей и последовательностей, кластеризация.
4. Модели прогнозирования и обработки данных на основе деревьев решений (SAS DT vs CHAID, CART, C4.5) и ансамблей моделей (bagging, boosting, blending, random forest).
5. Линейные модели прогнозирования. Линейная и логистическая регрессии. Подготовка данных (поиск и обработка выбросов, подстановка пропущенных значений). Выбор значимых переменных. Преобразование пространства признаков. Регуляризация. Kernel методы. (Stepwise repression, LARS/LASSO, PCR/PLS, SVM).

СОДЕРЖАНИЕ (ЧАСТЬ 2)

7. Нейронные сети (GLM, MLP, RBF, SOM) и глубинное обучение.
8. Анализ текстовых данных - Text mining.
9. Анализ временных рядов. Подготовка, моделирование и анализ временных рядов в SAS/ETS (EXPAND, TIMESERIES, MODEL, SEVERETY).
10. Регрессионные и автокорреляционные модели прогнозирования в SAS/ETS (ESM, AUTOREG, ARIMA). Методы выбора параметров моделей (наименьших квадратов и максимума правдоподобия).
11. Модели пространства состояний и векторные методы в SAS/ETS (SSM, UCM, VARMAX).
12. Автоматическая настройка моделей (FORECAST). Прогнозирование с использованием SAS Forecast Server.
13. Методы оптимизации (OPTMODEL).

КОМПАНИЯ SAS | ЛИДЕР В ОБЛАСТИ АНАЛИТИКИ

- 38 лет на рынке (с 1976 г.)
- Более 13 000 сотрудников в 400 офисах SAS в 56 странах
- Клиенты SAS - более 70 тысяч организаций в 140 странах мира.
- 91 компания из top 100 списка «2013 FORTUNE Global 500®».
- SAS занимает более 36% рынка аналитического ПО



Инвестиции в R&D
> 20 %

КОМПАНИЯ SAS В РОССИИ И СНГ

- **В России и странах СНГ компания SAS начала работу в 1996 году**
- **Полный спектр решений и услуг в области бизнес-аналитики:**
  - Консалтинг, внедрение, обучение, техническая поддержка
  - 120 сотрудников и стажеры
- **Крупнейшие клиенты SAS в России и СНГ:**
  - Все ведущие банки, включая топ 10 крупнейших российских банков (Альфа-банк, ЮниКредит банк, Райффайзенбанк, Ситибанк, GE Consumer Finance, Банк «Возрождение», Банк «Тинькофф Кредитные Системы», Райффайзен Банк Аваль, Приватбанк, Укрсиббанк, Банк Форум, Кредитпромбанк и др.)
  - Многие ведущие транспортные компании, включая РЖД и «Аэрофлот»
  - Крупнейшие компании из телекоммуникационного и топливно-энергетического сектора
  - Государственные организации: ЦБ РФ, Налоговый Комитет Республики Казахстан, ФТС и другие

КОМПАНИЯ SAS

ЛУЧШИЙ РАБОТОДАТЕЛЬ СРЕДИ ТРАНСНАЦИОНАЛЬНЫХ КОМПАНИЙ 2013

- SAS USA #2
- SAS Canada
- SAS Mexico
- SAS Argentina
- SAS China
- SAS Australia
- SAS India
- SAS Korea
- SAS Germany
- SAS Belgium
- SAS Norway
- SAS Portugal
- SAS Finland
- SAS France
- SAS Italy
- SAS RUSSIA #3

ЦЕЛЕВЫЕ ОТРАСЛИ

КОМПАНИЯ SAS

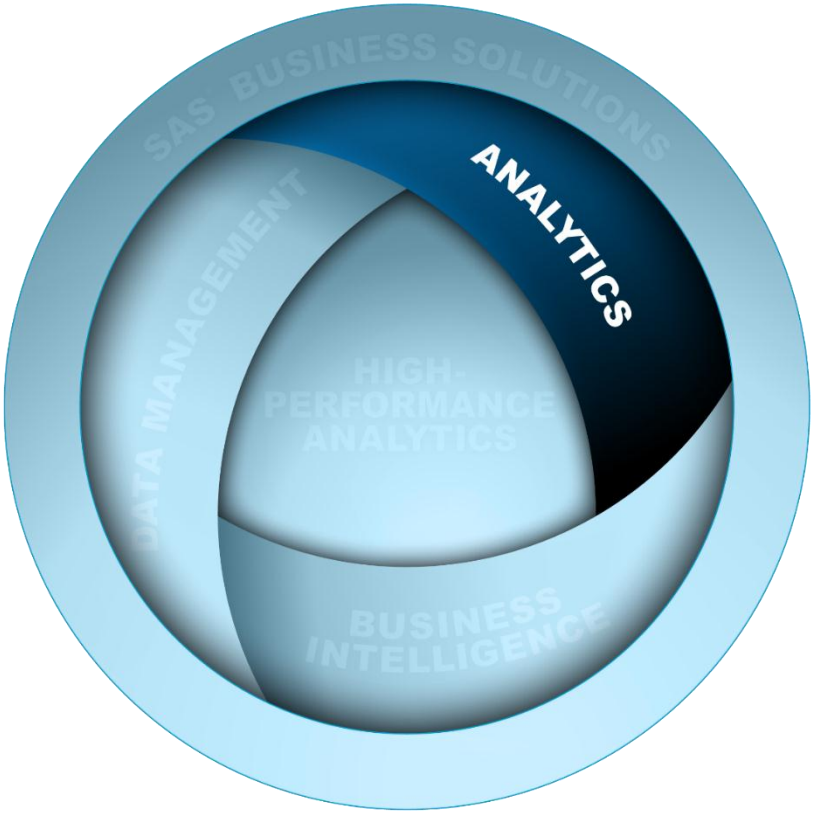
- Automotive
- **Banking**
- Capital Markets
- Casinos
- **Communications**
- Consumer Goods
- Defense & Security
- **Government**
- Health Care Providers
- Health Insurance
- High-Tech Manufacturing
- Higher Education
- Hotels
- Insurance
- K-12 Education
- Life Sciences
- Manufacturing
- Media
- Oil & Gas
- **Retail**
- Small & Midsize Business
- Sports
- **Travel & Transportation**
- Utilities

BUSINESS ANALYTICS FRAMEWORK

КОМПАНИЯ SAS

BUSINESS ANALYTICS FRAMEWORK

КОМПАНИЯ SAS



BUSINESS ANALYTICS FRAMEWORK

КОМПАНИЯ SAS

ПРОГРАММНЫЕ ПРОДУКТЫ

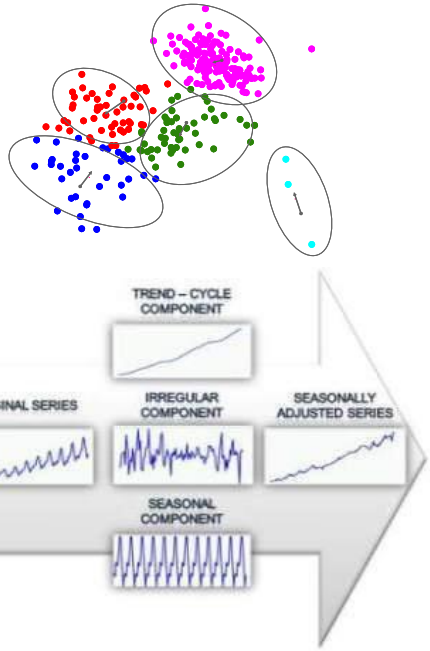
SAS ANALYTICS

SAS ANALYTICS | ДОСТОИНСТВА

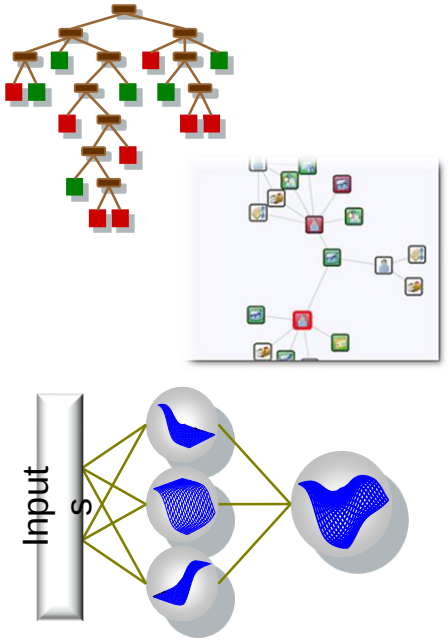
- **Исходные данные**. Обработка больших объемов данных сложной структуры из разных источников.
- **Глубина**. Реализованы самые современные методы анализа, которые постоянно совершенствуются, чтобы соответствовать самым последним достижениям.
- **Широта**. Совокупность методов:
  - Статистический анализ, визуализация и интеллектуальный анализа данных
  - Временные ряды, прогноз, эконометрика
  - Контроль и улучшение качества
  - Исследование операций
  - Имитационное моделирование
  - Анализ текстовых данных
- **Открытость.** Поддержка множества парадигм, основаных на многих дисциплинах, чтобы наилучшим образом формулировать и решать аналитические задачи.
- **Наглядность**. Поддерживается много графических методов визуального исследования данных, поиска взаимосвязей и неочевидных зависимостей для улучшения поддержки принятия решений.
- **Воспроизводилось**. Генерируемый код удовлетворяет большинству корпоративных и государственных требований к воспроизводимости и верифицируемости.
- **Независимость**. Работает на многих платформах.

SAS ANALYTICS | ПОДХОД

Выявление зависимостей



Построение моделей



Внедрение моделей

SAS/STAT® БАЗА ДЛЯ SAS® ANALYTICS

- дисперсионный анализ,
- байесовский анализ,
- категориальный анализ данных,
- кластерный анализ,
- описательная статистика,
- дискриминантный анализ,
- анализ распределения,
- подстановка пропущенных значениязначений,
- смешанные модели,
- многомерный анализ,
- непараметрическая статистика,
- регрессионный анализ,
- структурные уравнения,
- случайная выборка и анализ опросов,
- анализ выживаемости,
- … и многое другое.

SAS/IML® INTERACTIVE MATRIX PROGRAMMING

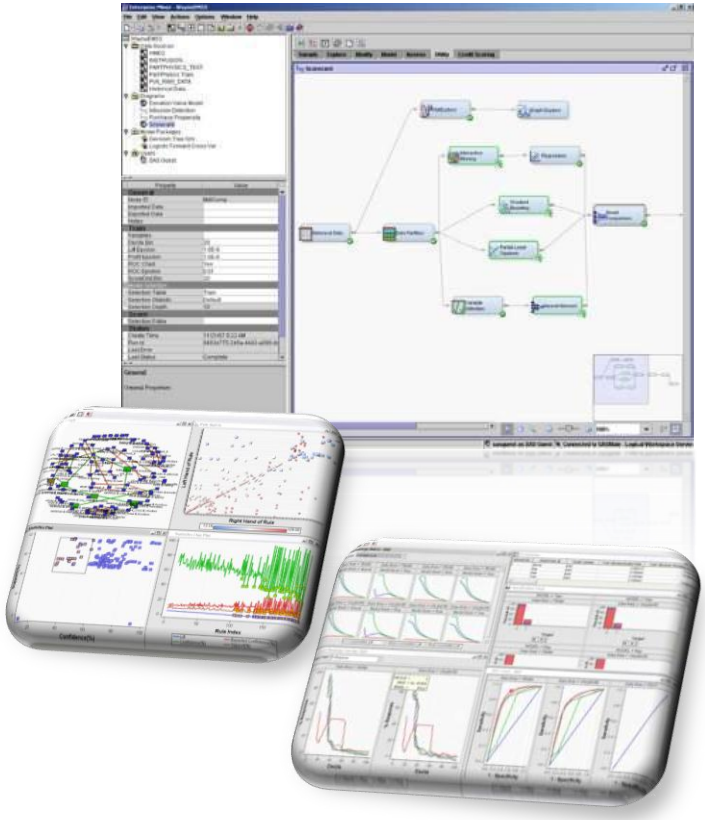
- Полный набор операций матричной алгебры
- Управляющие конструкции
- Линейная алгебра и статистические функции
- Временные ряды
- Численные методы
- Возможность совмещать разные языки в одной среде программирования

- Интерактивный анализ данных:
  - Разведочный анализ данных
  - Анализ распределений
  - Параметрическая и непараметрическая регрессия
  - Многомерный статистический анализ

SAS® ENTERPRISE MINER™

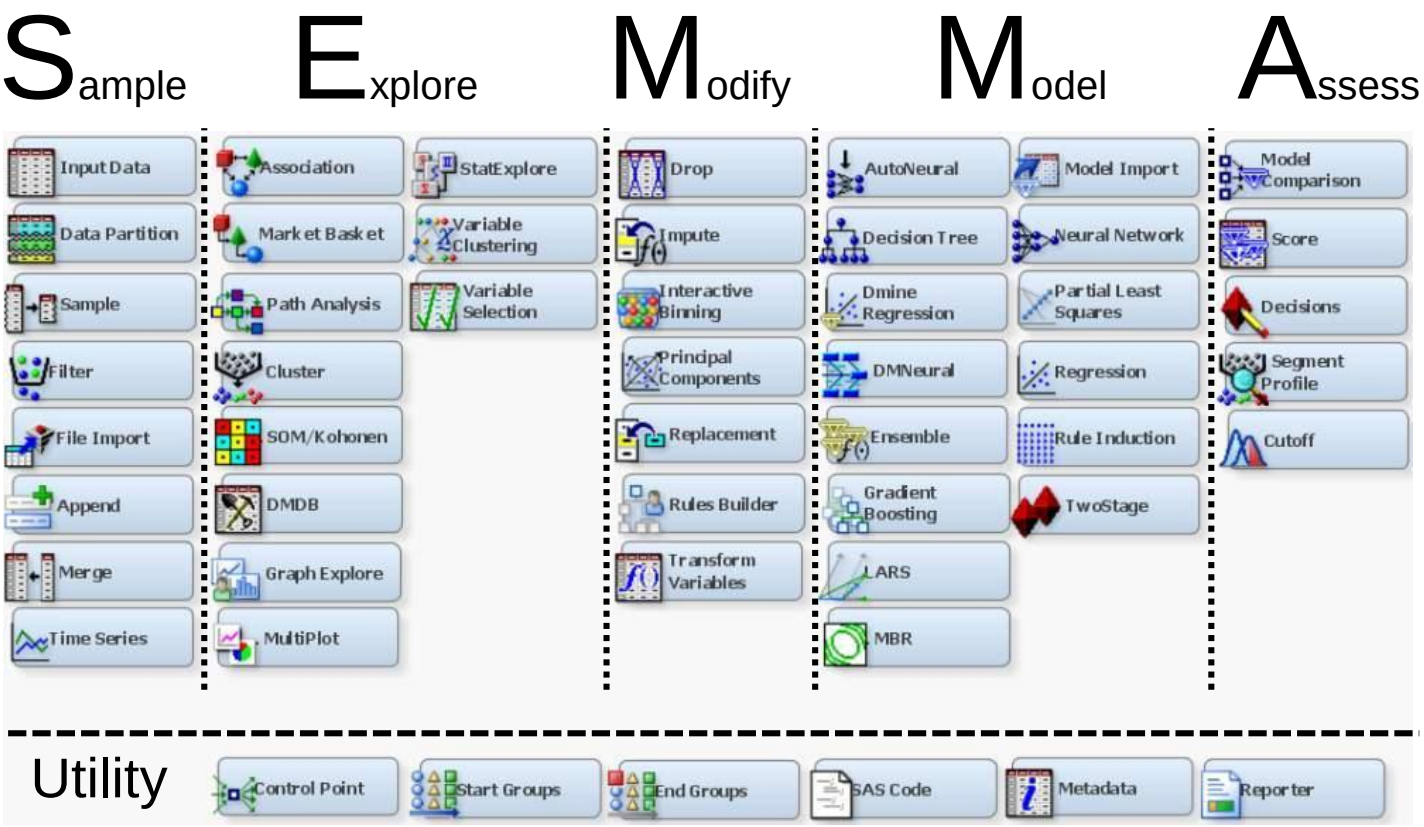
ИНСТРУМЕНТ DATA MINING

- Современное, корпоративное, легкое в использовании программное средство интеллектуального анализа данных
- Широкий набор средств подготовки и разведочного анализа дынных
- Набор эффективных алгоритмов построения моделей прогнозирования, в том числе параллельных
- Интерактивные средства сравнений, тестирования и валидации моделей
- Автоматизированные средства применения построенных моделей
- Открытая, расширяемая, гибкая архитектура



SAS® ENTERPRISE MINER™

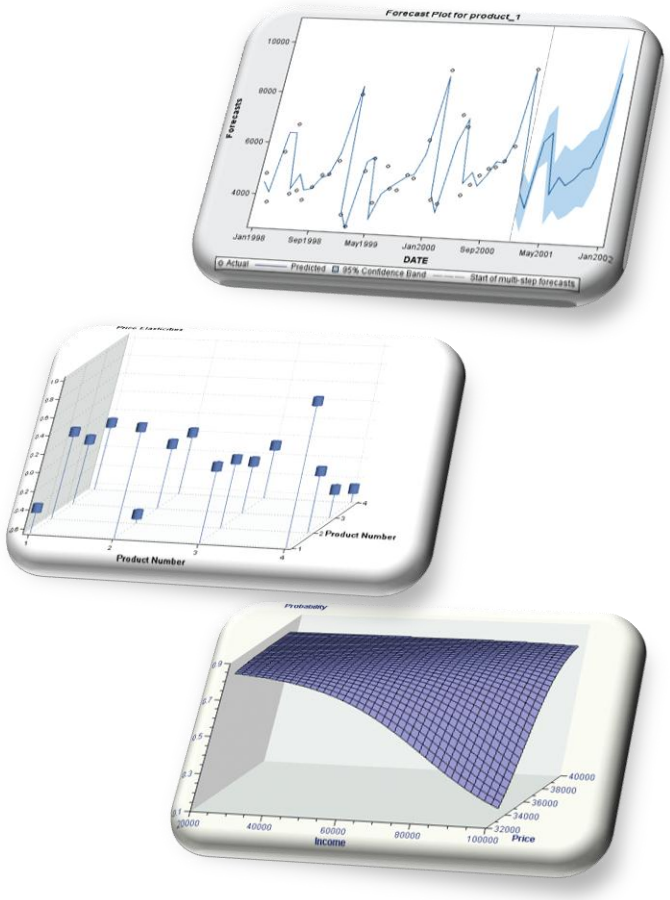
ПРОЦЕСС РАЗРАБОТКИ МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ



SAS/ETS®

ЭКОНОМЕТРИКА И ПРОГНОЗИРОВАНИЕ

- Интеграция методов анализа временных рядов, эконометрики, прогнозирования и имитационного моделирования бизнес процессов.
- Анализ влияния рекламных акций, оценка их эффективности для оптимизации маркетинговых компаний.
- Прогноз потребления и эффективное распределение ресурсов для производства, оборудования и управления персоналом.
- Моделирование поведения клиентов для максимизации эффективности маркетинговых компаний и понимания влияния характеристик клиентов на процесс выбора товаров и услуг.



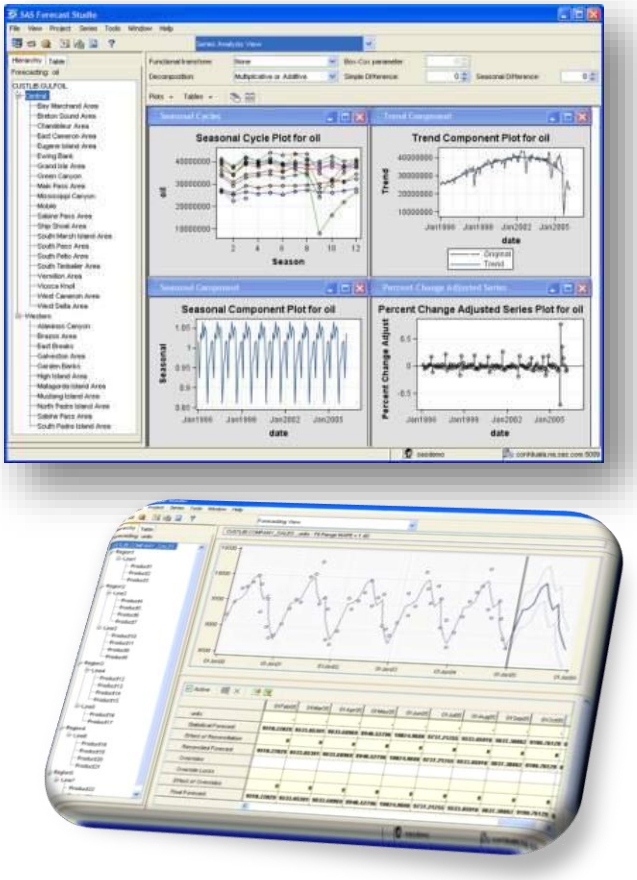
SAS/ETS®

МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ И ЭКОНОМЕТРИКИ

- Методы прогнозирования временных рядов
  - Выявление трендов, в том числе сезонных
  - Экспоненциальное сглаживание
  - Винтеровские методы (аддитивные и мультипликативные)
  - ARIMA
  - Методы на основе латентных состояний
  - Динамическая регрессия
  - Автоматическое выявление выбросов и исключений
  - Декомпозиция временных рядов
- Доступ к ряду коммерческих и правительственных информационных баз данных
  - FAME, DRI, Standard & Poor's (COMPUSTAT), Haver Analytics, и CRSP, Bureau of Economic Analysis, Bureau of Labor Statistics, International Monetary Fund …
- Эконометрический анализ
  - Построение и анализ совместных линейных и нелинейных регрессионных моделей
  - Анализ многомерного дискретного выбора
  - «Что-если» анализ
  - Методы Монте-Карло
  - Многомерные временные ряды
  - Регрессии с коррелированными и автокоррелированными ошибками
- Подготовка и управление данными
- Финансовый анализ

SAS® FORECAST SERVER АВТОМАТИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ

- Корпоративная среда для прогнозирования временных рядов с большим объемом информации
- Позволяет бизнес пользователям использовать все возможности аналитики SAS через пользовательский интерфейс SAS® Forecast Studio
- Автоматический выбор наиболее подходящих моделей и генерация прогноза
- Включает поддержку специальных календарей и событий, связанных с бизнес процессом



SAS® FORECAST SERVER

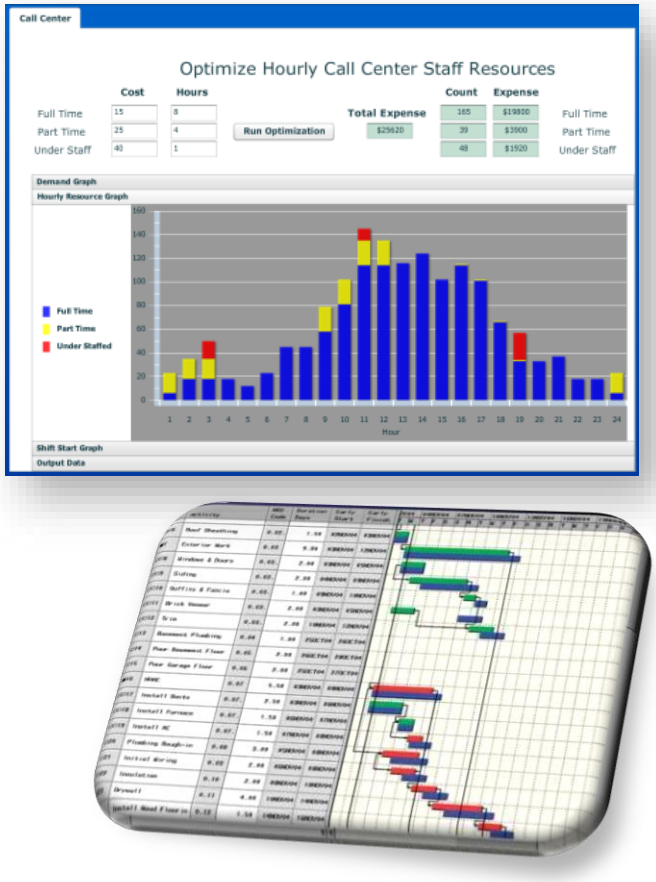
ВОЗМОЖНОСТИ И МЕТОДЫ

- SAS® Forecast Studio
  - Идентификация источников данных
  - Поддержка иерархий
  - История диагностики
  - Выбор и параметризация моделей
  - Генерация прогноза
- Автоматический прогноз с заданным уровнем автоматизации
  - Автоматический выбор событий и регрессоров для моделей
  - Оптимизация параметров моделей
- Управление данными временных рядов
  - Пользовательские иерархии
  - Иерархический прогноз
  - Правила поиска исключений и автоматический поиск выбросов
  - Консоль управления событиями для заданных типов прецедентных и будущих событий
  - Генерация кода для пакетной обработки
  - Анализатор сценариев для проверки влияния «что-если» факторов

ИССЛЕДОВАНИЕ ОПЕРАЦИЙ

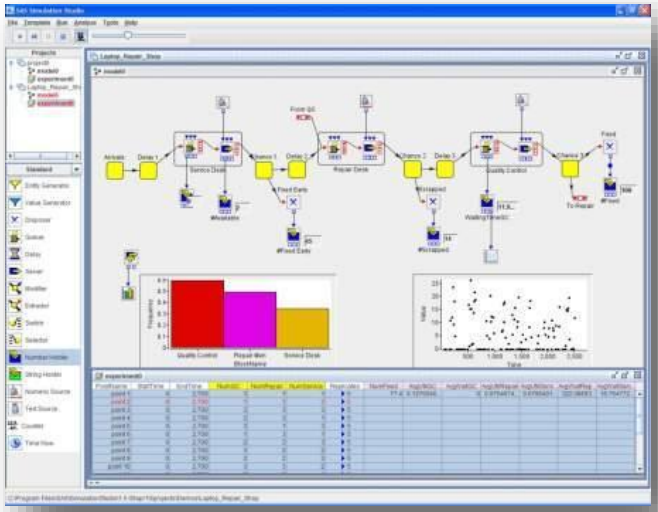
SAS/OR®

- Широкий спектр методов исследования операций и методов оптимизации.
- Интерактивное построение моделей, определение переменных и ограничений.
- Эксперименты с эффектами от изменения анализируемых данных.
- Простая индикация где и как используются в модели входные данные.
- Быстрые методы поиска решений для задач большой размерности.



SAS/OR® ОПТИМИЗАЦИЯ БИЗНЕС ПРОЦЕССОВ

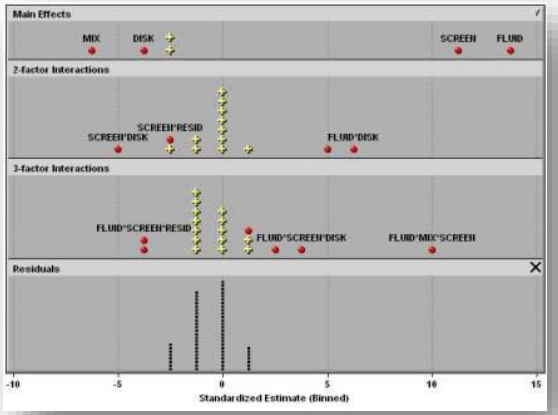
- Математическая оптимизация
  - Алгебраическая, символьная, дискретная оптимизация.
  - Единый язык представления линейных, целочисленных, смешенных, нелинейных, и квадратичных задач оптимизации.
  - Мощные алгоритмы решения задач условной и безусловной оптимизации.
- Дискретное событийное имитационное моделирование
- Математические методы управление проектами и ресурсами
- Методы освоенного объема
- Генетические алгоритмы
- Анализ решений



SAS/QC®

УЛУЧШЕНИЕ КАЧЕСТВА ДЛЯ БИЗНЕС ПРОЦЕССОВ ОРГАНИЗАЦИИ

- Позволяет улучшать качество продукции, оптимизировать процессы, и улучшать уровень удовлетворения клиентов
- Идентификация источников проблем с использованием методов статистического анализа
- Широкий набор средств идентификации и описания источников вариации процессов
- Средства планирования эксперимента и анализа экспериментальных результатов.
- Предоставление корпоративного представления процессов улучшения качества
- Обработка больших объемов данных от различных процессов



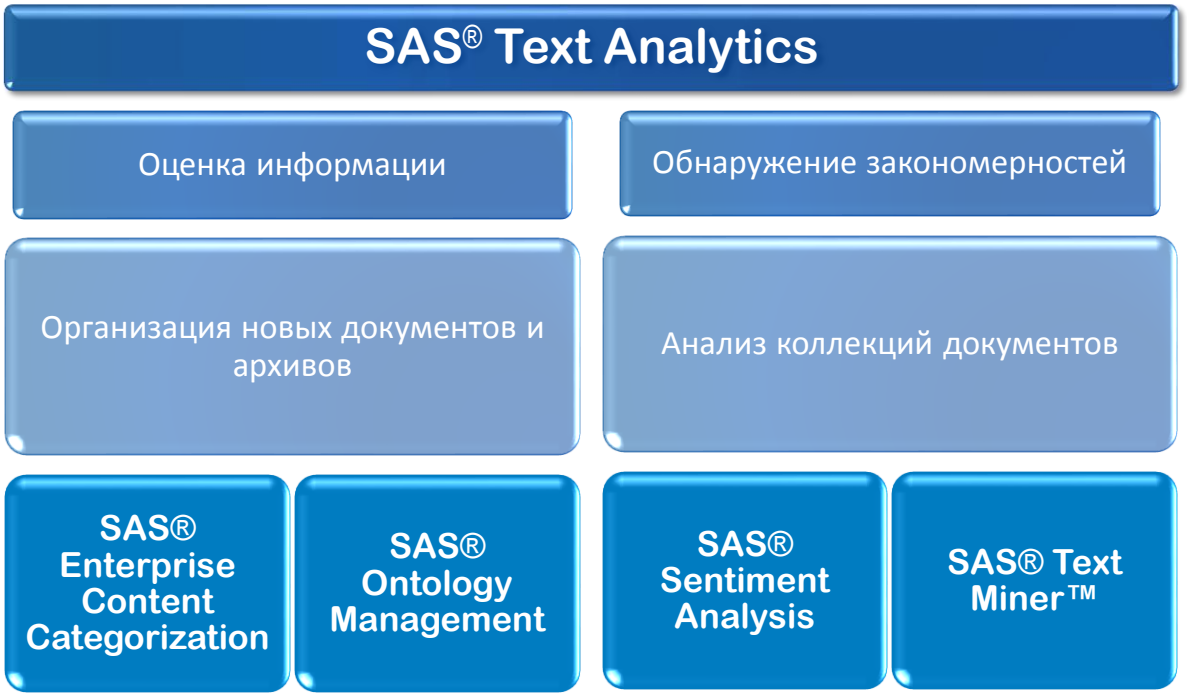
SAS/QC®

УЛУЧШЕНИЕ КАЧЕСТВА ДЛЯ БИЗНЕС ПРОЦЕССОВ ОРГАНИЗАЦИИ

- Базовые средства анализа качества на основе диаграмм Парето и Ишикава
- Статистическое управление процессами
  - Графики кумулятивных сумм, скользящих средних, нестандартные графики управления процессом, тренды и т.д.
- GAGE подход для оценки систем
- Анализ производительности процессов
  - Сравнительные гистограммы, CDF графики, вероятностные графики, Q-Q и P-P графики
  - Индексы производительности, достоверности, устойчивости, с доверительными интервалами и описательными статистиками
- Анализ надежности
  - Тесты с цензурированными данными, анализ моделей с множественными отказами, вероятностные графики и др.
- Анализ средств
  - Единичные и множественные переменные отклика, расчет границ решения, p графики, u графики, box графики
- Планирование экспериментов
  - Полный и частичный дизайн, D-оптимальный and A-оптимальный дизайн
  - ADX интерфейс для планирования эксперимента, анализа, генерации отчетов

КОНЦЕПЦИЯ

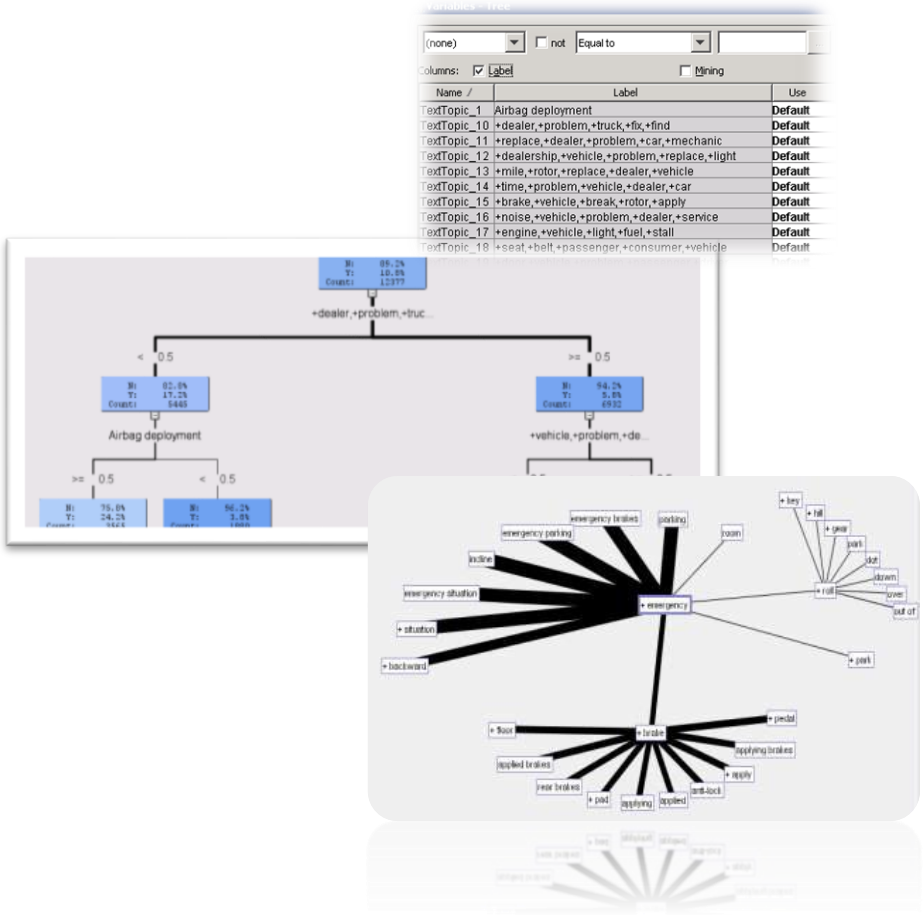
SAS ® TEXT ANALYTICS



SAS® TEXT MINER

ВЫЯВЛЕНИЕ СМЫСЛОВЫХ ЗАВИСИМОСТЕЙ В КОЛЛЕКЦИЯХ ТЕКСТОВ

- Богатый набор лингвистических моделей и средств
- Выявление знаний из текстовых коллекций
- Разведочный анализ и визуализация
- Автоматический поиск ассоциаций термов, групп тематик
- Включает пользовательские термины и настройки
- Структуризация текстов

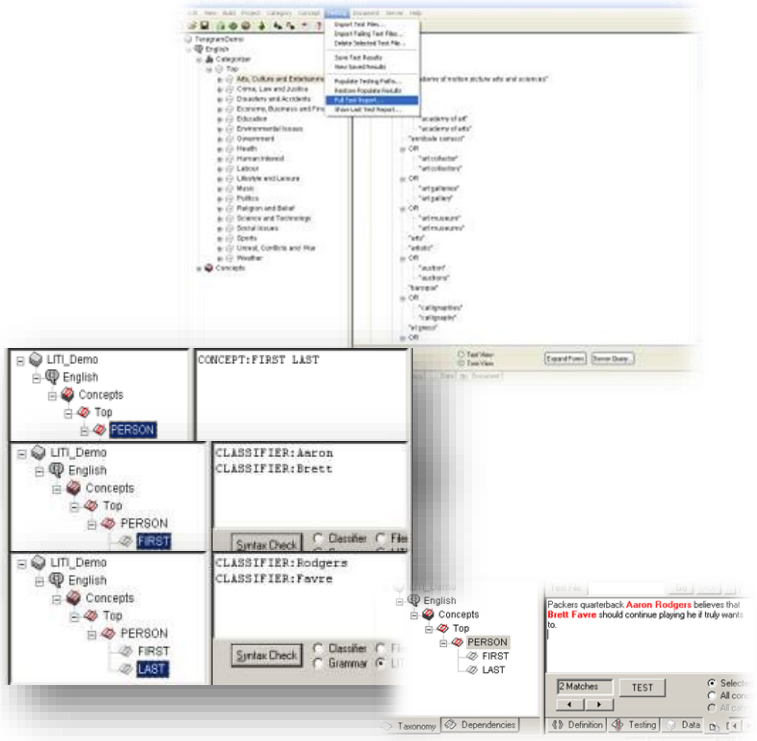


THE POWER TO KNOW.

SAS® ENTERPRISE CONTENT CATEGORIZATION

КАТЕГОРИЗАЦИЯ ДОКУМЕНТОВ ДЛЯ РЕЛЕВАНТНОГО ПОИСКА

- Natural Language Processing и углубленная лингвистика для автоматической категоризации больших объемов текстов на разных языках
- Разбор и анализ описаний событий, создание метаданных для автоматизации процесса
- Совместное создание и управление таксономиями для фактов и событий
- Контролируемые правила для включения/исключения документов для анализа
- Значительное улучшение точности информационного поиска



SAS® SENTIMENT ANALYSIS

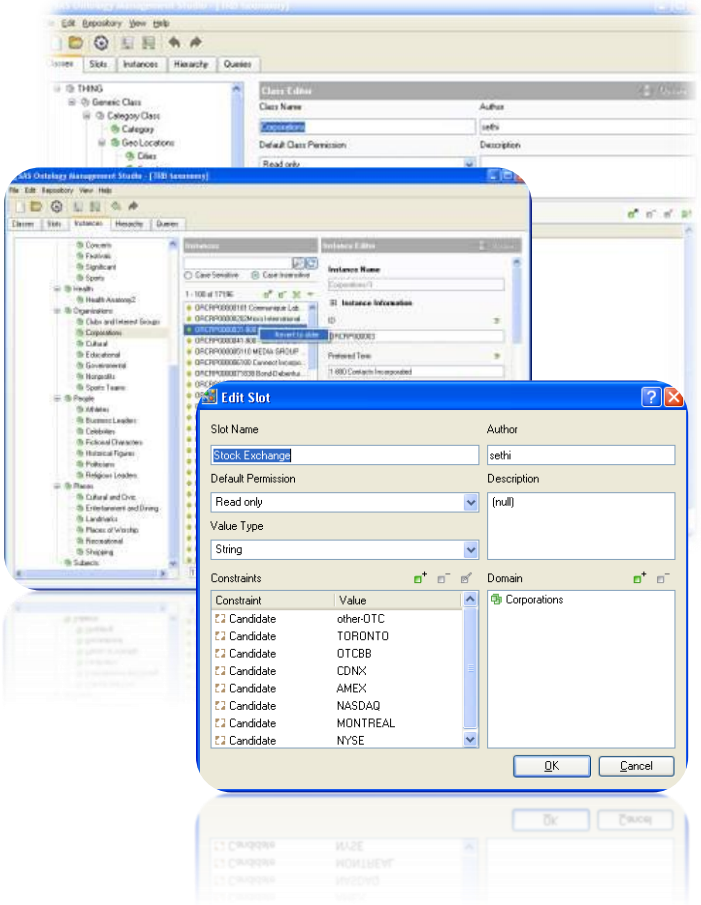
ОБНАРУЖЕНИЕ, АНАЛИЗ И ИНТЕРПРЕТАЦИЯ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ВЫСКАЗЫВАНИЙ

- Автоматически обнаруживает и интерпретирует эмоциональную окраску высказываний
- Отслеживает высказывания о целевом концепте (событии, организации, бренде, продукте, услуге)
- Уникальный гибридный подход на основе комбинации статистических и лингвистических методов
- Редактирование, тестирование и улучшение моделей
- Реализация множества моделей для анализа в реальном времени

SAS® ONTOLOGY MANAGEMENT

УПРАВЛЕНИЕ И АНАЛИЗ СЕМАНТИКИ

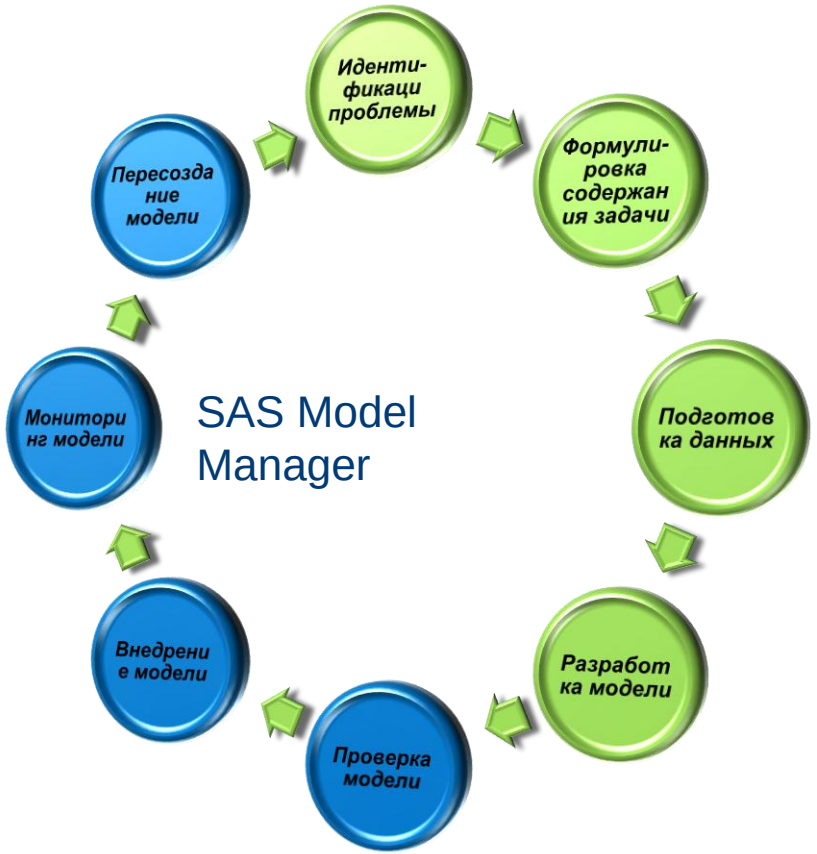
- Анализ информации в различных источниках в корпорации и поиск ассоциации между ними
- Определение взаимосвязей для поиска ответов на сложные вопросы
- Предоставление явной, консистентной, определенной процедуры управления информацией
- Сохранение данных о предметной области в архивах компании



ЖИЗНЕННЫЙ ЦИКЛ АНАЛИТИЧЕСКОГО ПРОЦЕССА

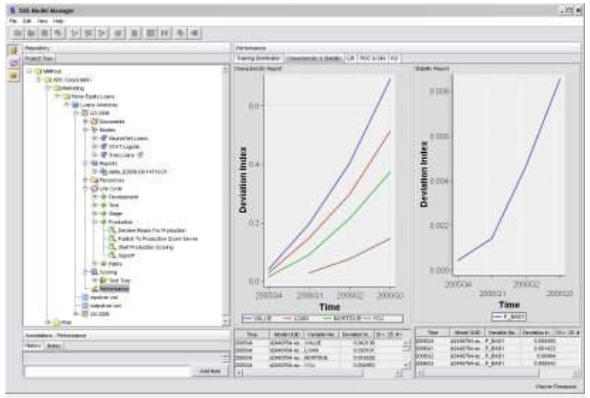
SAS ® MODEL MANAGER

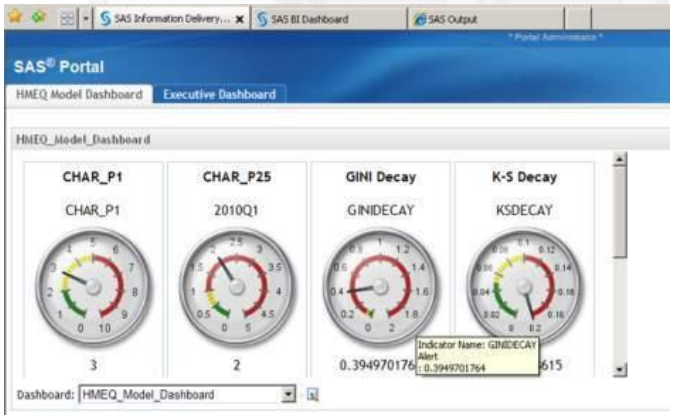
- Автоматизированный
- Контролируемый
- Воспроизводимый
- Управляемый
- Проверяемый
- Внедряемый
- Совместный и разделяемый
- Применение «внутри» СУБД
- Робастный



SAS ® MODEL MANAGER

- Центральный безопасный репозитарий для организации моделей
- Шаблоны для регистрации моделей
- Проверка бизнес-логики применения моделей перед внедрением
- Интеграция с SAS® Scoring Accelerator для применения «внутри» СУБД
- Отчеты с оценками производительности и сравнения моделей в течении всего жизненного цикла

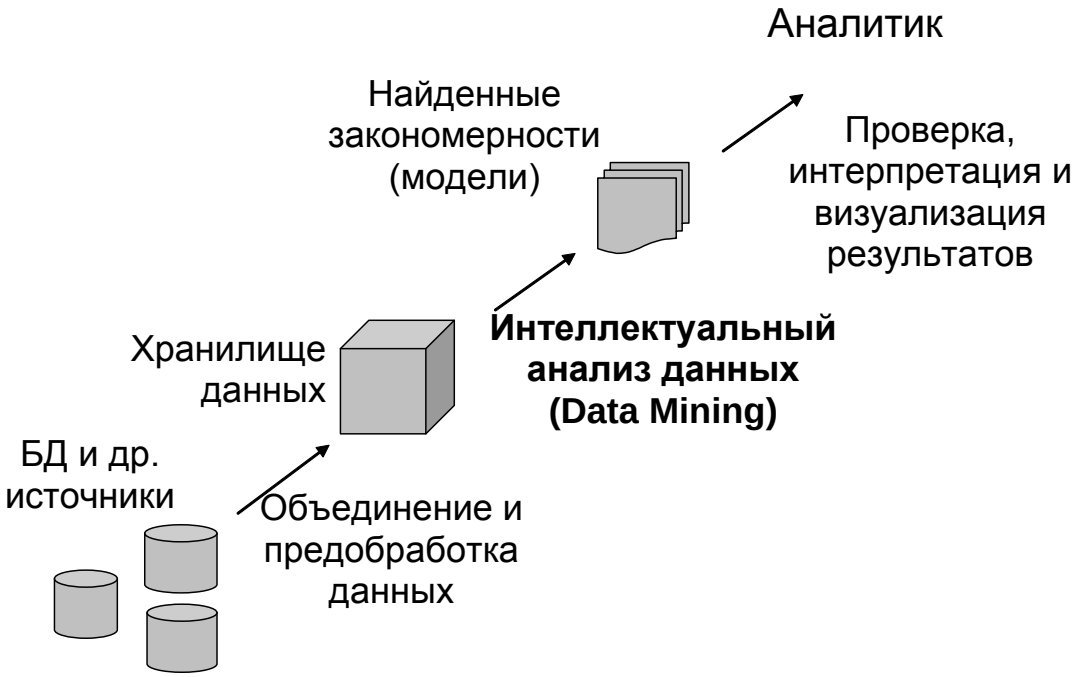




SAS ENTERPRISE MINER

ВВЕДЕНИЕ И ОБЗОР ВОЗМОЖНОСТЕЙ

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ (DATA MINING)



Системы *интеллектуального анализа данных* (ИАД) – класс программных систем поддержки принятия решений, задачей которых является поиск *скрытых, ранее неизвестных, содержательных и потенциально полезных* закономерностей в *больших объемах разнородных, сложно структурированных* данных.

Han J., Kamber M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2000

ЭВОЛЮЦИЯ ТЕХНОЛОГИЙ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ

- … — 1960-е:
  - Файлы и файловые архивы
- 1960-е:
  - Первые СУБД, иерархические, сетевые и т.д.
- 1970-е:
  - Реляционная модель данных, реляционные СУБД
- 1980-е:
  - «Продвинутые» СУБД (объектно-реляционные и объектные, «расширенные» реляционные, дедуктивные и др.)
  - «Специализированные» СУБД (гео-,научные, инженерные и др.)
- 1990-е — …:
  - Мультимедийные БД, WWW, хранилища, витрины данных,OLAP, Data Mining

АКТУАЛЬНОСТЬ И НЕОБХОДИМОСТЬ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ (ИАД)

- Проблема больших объемов («Data explosion»):
  - Средства автоматического сбора данных, повсеместное внедрение СУБД, электронный документооборот, WWW, мультимедийные архивы и т.д. приводят к росту объемов и усложнению структуры хранимой информации.
- Традиционные средства не справляются:
  - Информационный поиск и стат. анализ не везде помогают – много данных, сложная структура и нужно знать точно, что искать.
  - Вывод: много данных, но мало информации для аналитика.
- Необходимо:
  - Наличие программных средств автоматизированного анализа данных большого объема и сложной структуры.

ПРОЦЕСС ИАД (1)

- Анализ предметной области:
  - выявление и формулировка необходимых априорных знаний о предметной области, целей анализа, задач приложения, сценариев использования
- Формирование и подготовка данных для анализа:
  - поиск (или выбор) «сырых» данных
  - предобработка данных (нормализация, дискретизация, обработка пропущенных значений, удаление артефактов, проверка консистентности)
  - уменьшение размерности, выбор значимых характеристик, расчет интегральных показателей и инвариантов
- Определение типа решаемой задачи анализа и формализация:
  - классификация, прогнозирование, кластеризация, поиск исключений, ассоциативный анализ и т.д.

ПРОЦЕСС ИАД (2)

- Выбор или разработка алгоритма анализа:
  - определение ограничений и требований к алгоритму по точности, размеру, интерпретируемости, скорости построения и применения получаемых моделей, по типу исходных данных
- Непосредственно «Data mining»:
  - применение выбранного алгоритма анализа для поиска закономерностей выбранного типа и построение моделей
  - визуализация, преобразование, удаление избыточности, оценка точности, достоверности моделей и т.д.
- Применение построенных моделей:
  - Descriptive data mining - информирование аналитика, «описательные» модели, основная цель – визуализация
  - Predictive data mining – прогнозирование неизвестных значений или характеристик в «новых» данных с помощью построенных моделей, основная цель – прогноз

МЕСТО ИАД В ПРОЦЕССЕ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

ОСНОВНЫЕ ТИПЫ ИСХОДНЫХ ДАННЫХ

- Транзакционные
  - Объекты анализа – «события» различной структуры с числовыми и категориальными атрибутами и с временной меткой
- Табличные
  - Объекты анализа представлены в виде реляционных таблиц, возможно взаимосвязанных (заданно ER-схемой), имеют разнотипные атрибуты
- Временные ряды и числовые данные большого объема
  - Обработка результатов наблюдений, научных экспериментов, характеристик технологических процессов
- Электронные тексты на естественном языке
  - анализ содержимого документов

ЗАДАЧИ ИАД = ТИПЫ ВЫЯВЛЯЕМЫХ ЗАКОНОМЕРНОСТЕЙ

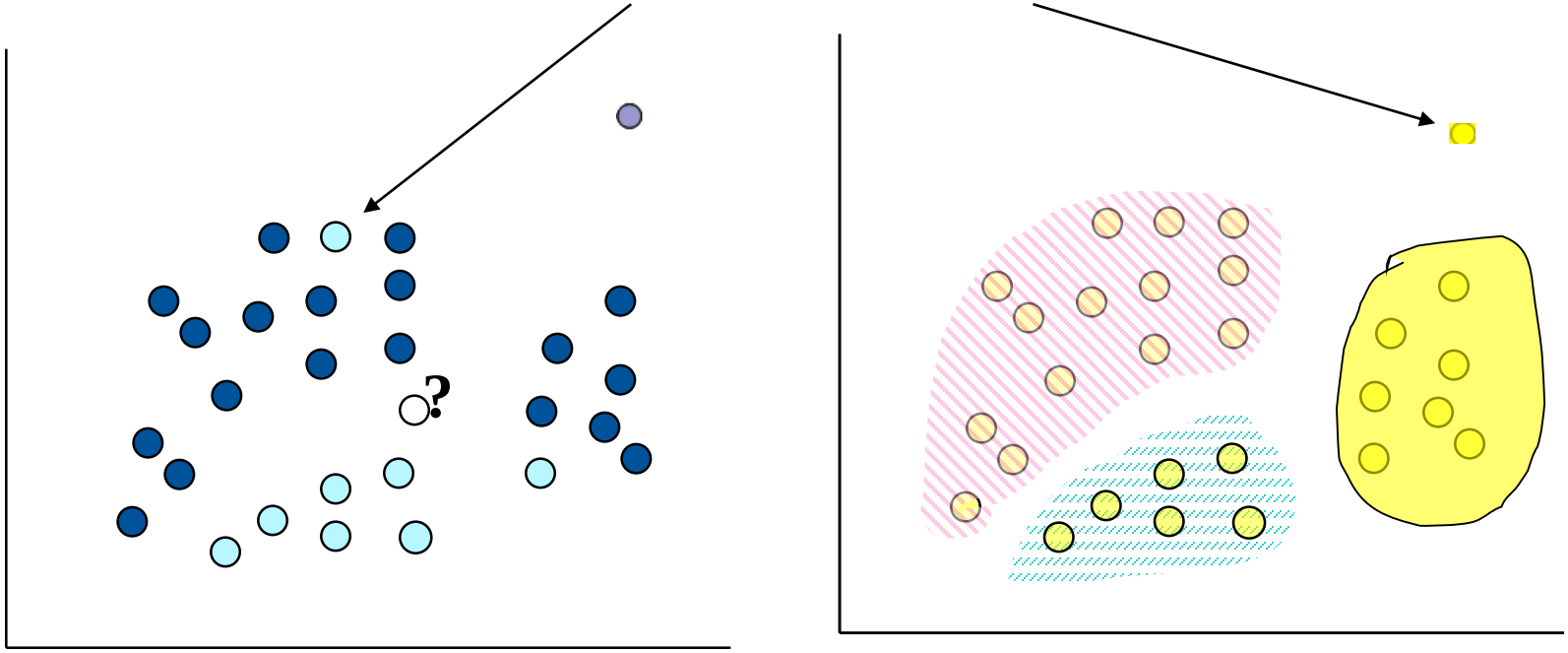
- Классификация («Обучение с учителем»)
  - Отнесение объектов к заранее определенным категориям
- Ранжирование («Обучение с учителем»)
  - Оценка степени соответствия объектов одной или более заранее определенным категориям
- Прогнозирование («Обучение с учителем»)
  - На основании известных значений атрибутов анализируемого объекта определяются значения неизвестных атрибутов
- Ассоциации («Обучение без учителя»)
  - Выявление зависимостей между атрибутами в виде правил или аналитических зависимостей
- Кластеризация («Обучение без учителя»)
  - Выделение компактных подгрупп «похожих» объектов
- Выявление исключений («Обучение с учителем и без»)
  - Поиск объектов, которые своими характеристиками значительно отличаются от остальных

ДАННЫЕ ДЛЯ АНАЛИЗА

- Объект анализа (или прецедент, или кейс, или ситуация, …) задается набором признаков (или атрибутов, или свойств, …)
- Признаки бывают:
  - Категориальные - нет расстояний, не задан порядок
  - Ординальные (порядковые) – нет расстояний
  - Числовые – есть расстояние
- «Размеченный» набор данных – для каждого объекта выделен один или более признаков, которые могут быть неизвестны и которые нужно предсказывать, тогда задача обучения «с учителем», иначе «без учителя» («неразмеченный» набор данных):
  - «Выходные» признаки - нужно предсказывать (они же отклики, или «зависимые переменные», или …)
  - «Входные» признаки, которые считаются всегда известными (они же входы, или «независимые переменные», или регрессоры, …)

ОБУЧЕНИЕ «С УЧИТЕЛЕМ» И «БЕЗ»

аномалии тоже разные

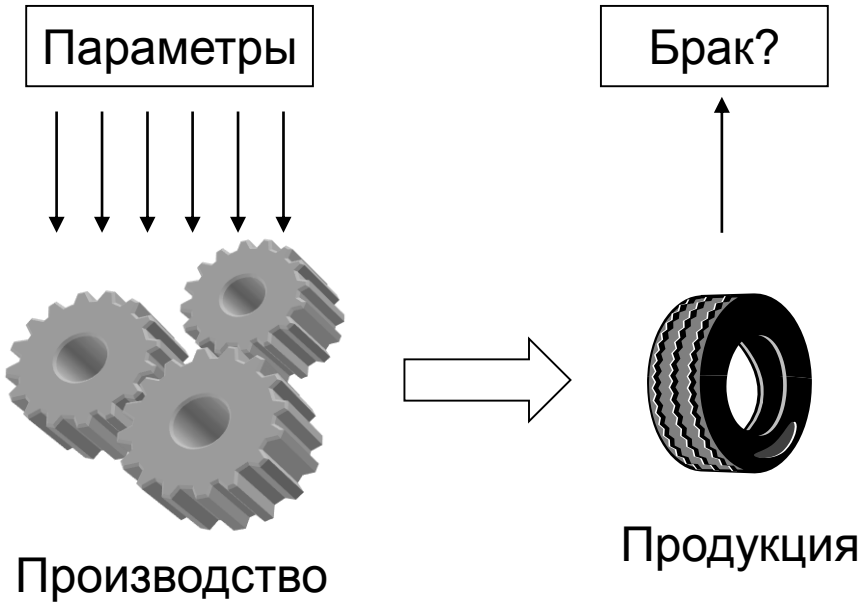


КЛАССИФИКАЦИЯ

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известен его класс
- Цель:
  - Построить классификатор – функцию или алгоритм, который в зависимости от свойств объекта предсказывает его класс
- Приложения:
  - Компьютерная безопасность
  - Производство- прогнозирование качества изделий
  - Распознавание образов

ПРИМЕР: АНАЛИЗ И ПРОГНОЗИРОВАНИЕ БРАКА В ТЕХНОЛОГИЧЕСКОМ ПРОЦЕССЕ

Какие параметры производственного процесса влияют на качество продукции?



$$Quality = F(X_1, \ldots X_n),$$

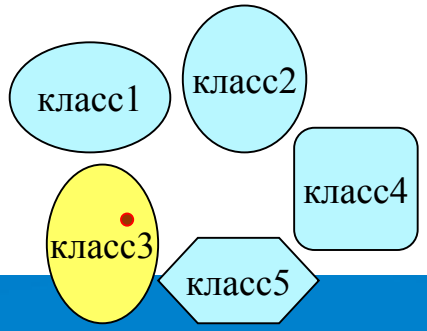
где X_i — i-ая характеристика производственного процесса,

РАНЖИРОВАНИЕ

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известен его класс или несколько не взаимоисключающих классов
- Цель:
  - Построить функцию или алгоритм ранжирования, который в зависимости от свойств объекта вычисляет степень его соответствия классам
  - Результат ранжирования: в рамках каждого класса можно упорядочить объекты по степени соответствия данному классу, и наоборот, в рамках каждого объекта можно упорядочить классы по степени соответствия данному объекту
- Приложения:
  - Документооборот - рубрикация документов
  - Кредитование - оценка заемщика

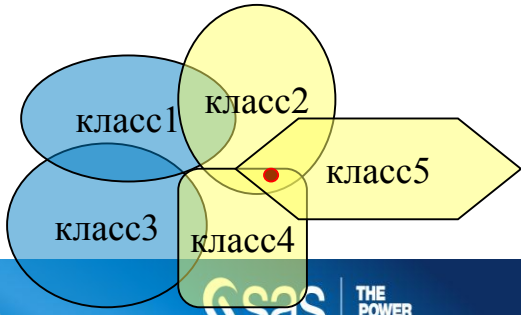
ПРИМЕР: РАНЖИРОВАНИЕ МНОГОТЕМНЫХ (MULTI-LABEL) ДОКУМЕНТОВ

Задача ранжирования документов – определение степени принадлежности документа к одному или нескольким классам (из предопределённого набора классов) на основании анализа совокупности признаков, характеризующих документ



Традиционная задача классификации (multi-class classification)

Задача классификации с пересекающимися классами (multi-label classification)

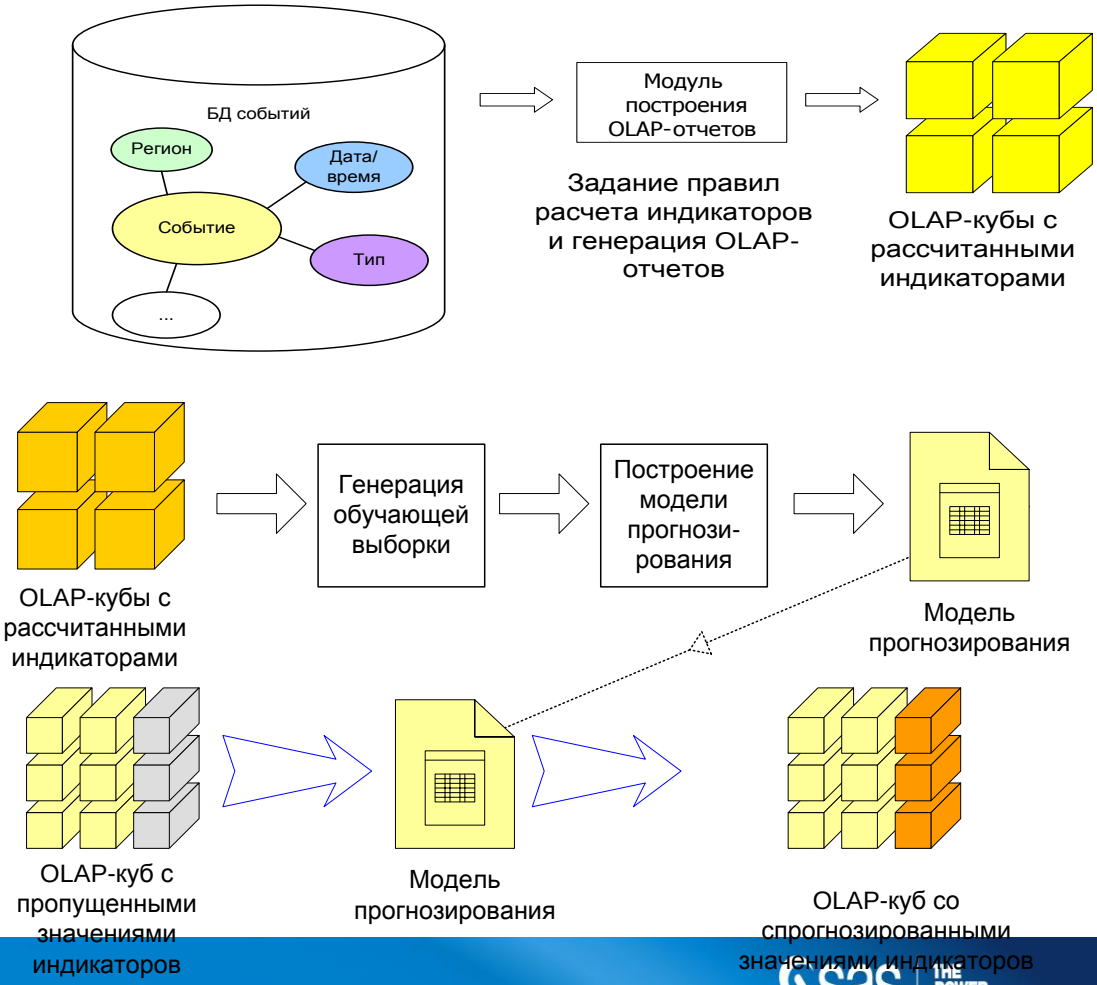


ПРОГНОЗИРОВАНИЕ

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известно значение некой числовой величины, которое необходимо спрогнозировать
- Цель:
  - Построить функцию, которая в зависимости от свойств объекта предсказывает значение данной величины
- Приложения:
  - Финансы - прогноз курсов валют, цен на нефть и др., оценка ожидаемых доходов или убытков предприятия
  - Маркетинг – прогнозирование числа новых клиентов или убыли старых
  - Прогноз электропотребления

ПРИМЕР: ПРОГНОЗИРОВАНИЕ РАЗВИТИЯ ОБСТАНОВКИ

- Проведение статистического анализа и вычисление индикаторов, описывающих ситуацию
- Определение тенденций и прогнозирование значений индикаторов



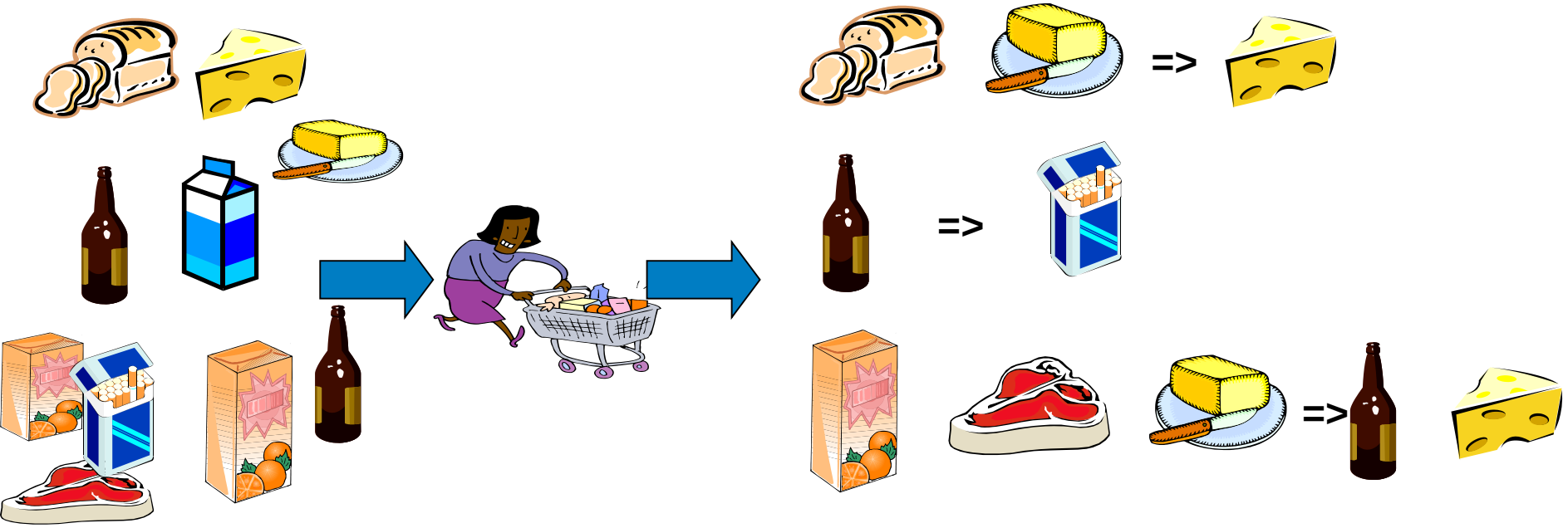
ПОИСК АССОЦИАЦИЙ

- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти зависимости между значениями атрибутов, например, в виде правил «если … то …»
- Приложения:
  - Маркетинг и рекомендательные системы - анализ зависимостей между покупаемыми товарами или услугами
  - Финансовый анализ – поиск зависимостей между значениями индексов и другими финансовыми параметрами
  - Медицина – анализ результатов исследований

ПРИМЕР: АНАЛИЗ «КОРЗИНЫ ПОКУПАТЕЛЯ»

Ассортимент супермаркета

Интересные правила



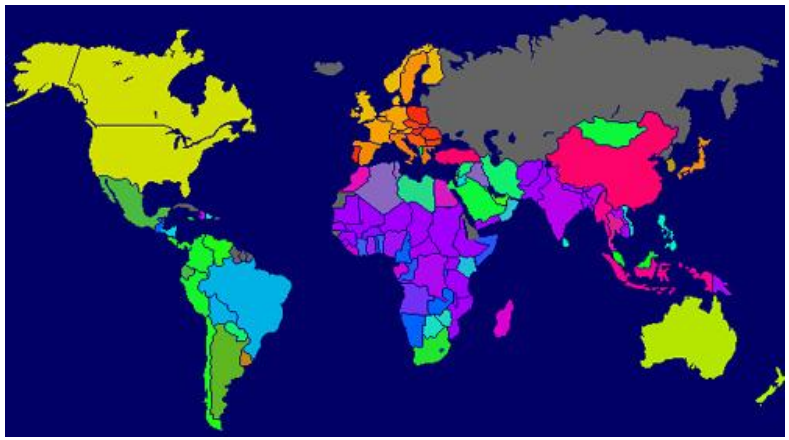
Задача Определить интересные правила в предпочтениях покупателей при выборе товара

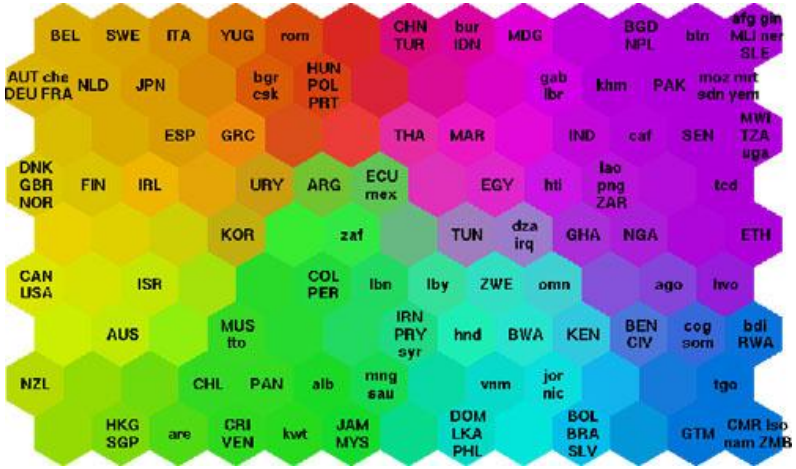
КЛАСТЕРИЗАЦИЯ

- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти «непохожие» группы «похожих» объектов
- Приложения:
  - Маркетинг – сегментация клиентов, рынков, товаров и т.д.
  - Производство – выявление типовых состояний и ситуаций
  - Индексирование документов

ПРИМЕР: КЛАСТЕРИЗАЦИЯ И ВИЗУАЛИЗАЦИЯ С ПОМОЩЬЮ SOM

- Когерентные области:
  - Близкие группы стран (по заданным стат. показателям) в исходном пространстве – рядом на решетке (свойство SOM) и одним (или спектрально близким) цветом
  - Группы стран (категории, области) - кластеры





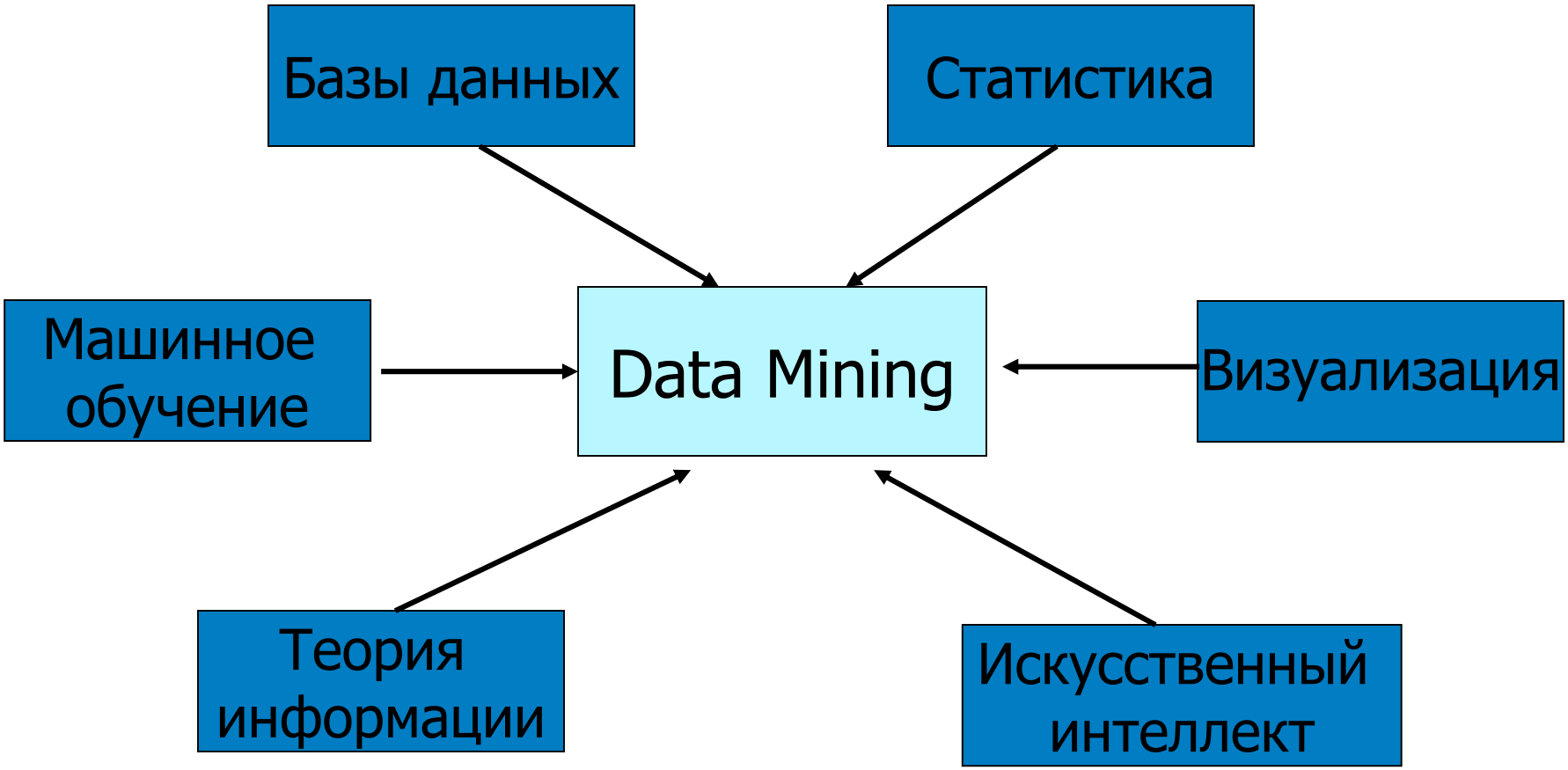
ВЫЯВЛЕНИЕ ИСКЛЮЧЕНИЙ

- Дано:
  - тренировочный набор («размеченный» или нет) – для каждого объекта известны значения его свойств
- Цель:
  - Найти наиболее «непохожие» объекты
- Приложения:
  - Безопасность – подозрительные финансовые транзакции, звонки, люди, организации
  - Производство – выявление нештатных ситуаций
  - Медицина – диагностика

ПРИМЕР: ВЫЯВЛЕНИЕ МОШЕННИЧЕСТВ

- Проблема:
  - мошенничать могут легальные пользователи
  - правилами (сигнатурами) тяжело выявить «новые» или «замаскированные» сценарии мошенничеств
- Примеры мошенничеств
  - Кредитные карты
  - Страховые случаи
  - Мобильные звонки
  - Инсайдеры
- Проблемы
  - Реальное время
  - Велика цена ошибок и первого, и второго рода
  - Аномалия (необычное действие пользователя) еще не значит мошенничество

МЕТОДЫ АНАЛИЗА



ОТЛИЧИЯ ИАД СИСТЕМ (1)

- Наличие «обучения»
  - модели формируются на основе анализируемых данных, а не экспертных знаний (в отличие от традиционных экспертных систем и систем информационного поиска)
  - структура модели и искомые зависимости заранее не известны (в отличие от стандартных статистических пакетов, ориентированных на расчет статистик, проверку гипотез и оценку параметров распределений)

ОТЛИЧИЯ ИАД СИСТЕМ (2)

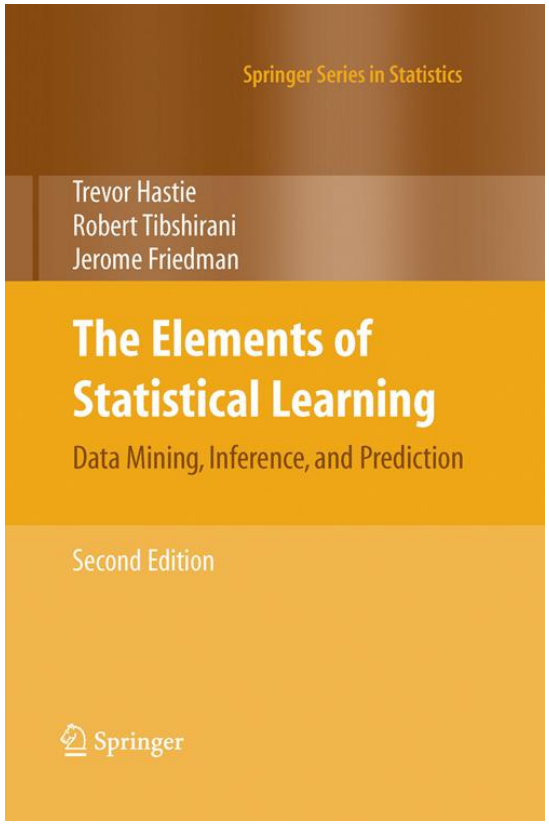
- Наличие большого объема данных сложной структуры
  - зачастую скорость работы алгоритмов в ИАД важнее отклонений по точности (“quick and dirty solution”)
  - большинство алгоритмов работают с исходными данными в виде числовой матрицы признаков, сложная структура реальных объектов в ИАД приводит к необходимости решать задачу построения пространства характеристик и отображения в него свойств исходных объектов
  - перечисленные особенности отличают ИАД системы от традиционных систем машинного обучения, в которых, как правило, решается обратная задача – построение достоверной модели в условиях малой обучающей выборки

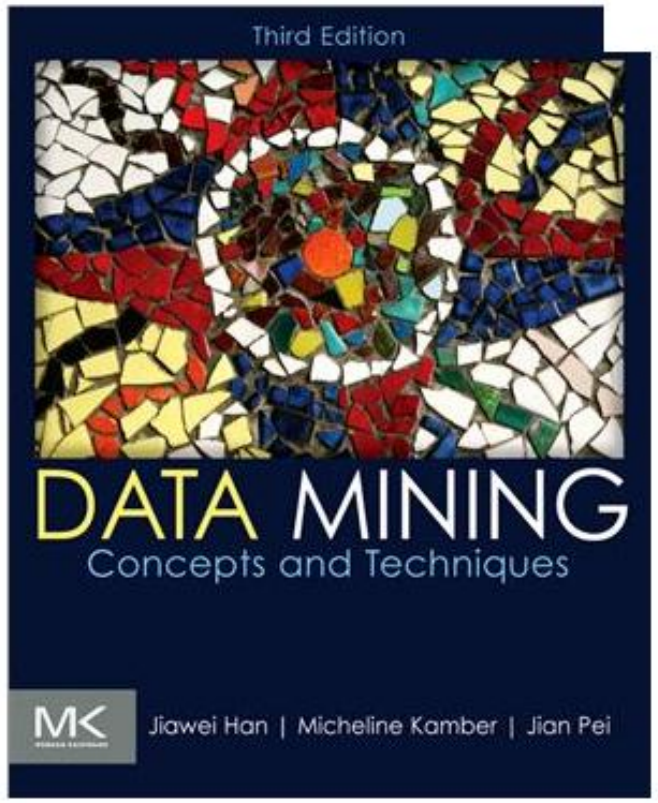
ОТЛИЧИЯ ИАД СИСТЕМ (3)

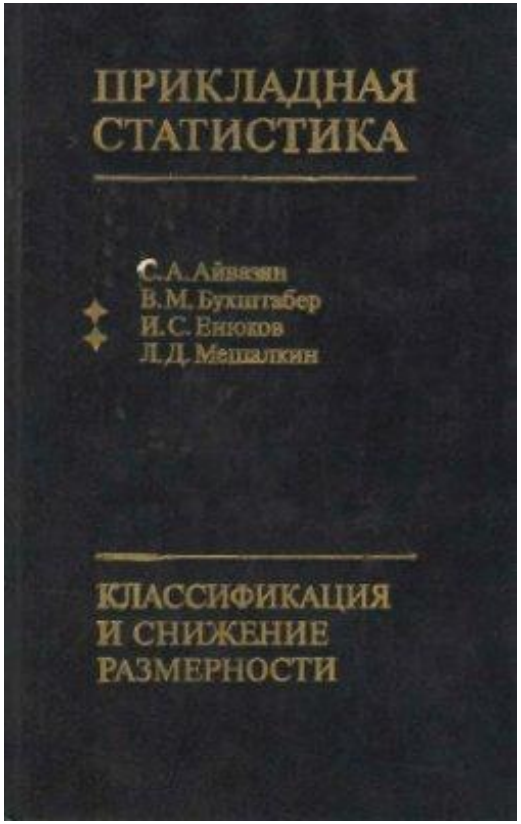
- Наличие аналитика
  - в сценарии работы любой системы ИАД всегда присутствует аналитик, даже если полученная в результате модель далее используется для автоматической классификации
  - аналитик формирует тренировочные наборы, производит настройку алгоритмов, обучение, анализирует полученные модели и принимает решения об их дальнейшем использовании
  - таким образом, системы автоматической классификации, кластеризации и распознавания образов, даже использующие возможность обучения, не являются системами ИАД

ЛИТЕРАТУРА

http://www-stat.stanford.edu/~tibs/ElemStatLearn







SAS ENTERPRISE MINER

- Программный продукт компании SAS Institute Inc., Cary, NC, USA,
- Реализует ИАД процесс в соответствии с концепцией SEMMA и обладает следующими характеристиками:
  - Удобный GUI, позволяющий начать работать с «0», в том числе «бизнес-пользователю»
  - Возможность создавать и обрабатывать в фоновом режиме пакеты задач
  - Мощные средства предобработки, агрегации и «разведочного анализа» данных
  - Современные алгоритмы прогнозного и описательного интеллектуального анализа данных (многие из них - запатентованные разработки SAS)

SAS ENTERPRISE MINER

- характеристики:
  - Развитые бизнесс-ориентированные средства сравнения и выбора моделей, построения отчетов, управления моделями, встроенные возможности поддержки принятия решений
  - Автоматизированный процесс применения моделей «внутри» продукта и «вне» («генерация» кода, реализующего ИАД процесс)
  - «Открытая» расширяемая архитектура (возможно встраивание своего кода)
  - Масштабируемые вычисления (пока для части методов)
  - Богатый набор встроенных прикладных решений (не входит в стандартный пакет)

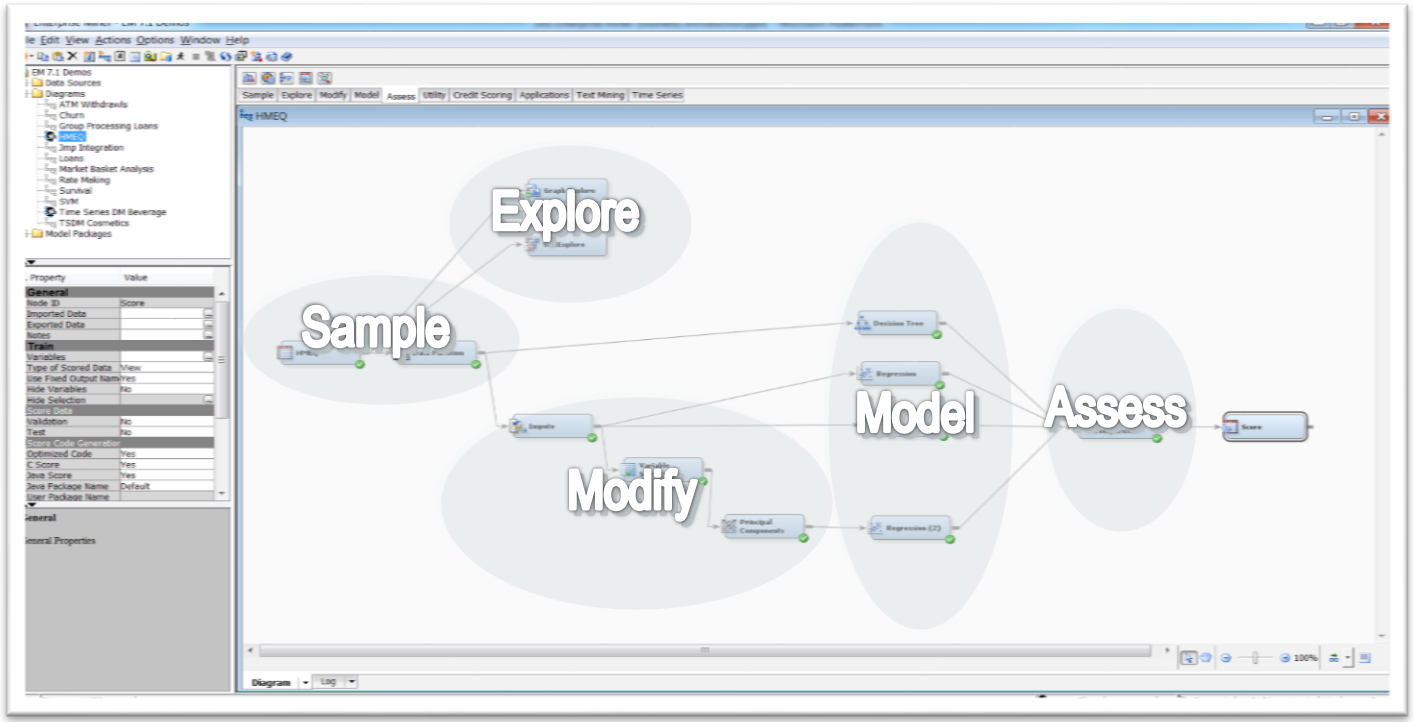
ONDEMAND SOLUTION

- «Облачная» Saas модель:
  - Хостится, управляется и конфигурируется SAS
  - Пользователь ставит только Java клиент
  - Полные функциональные возможности по сравнению со стандартной версией
  - Доступен «все время» «отовсюду»
  - Есть возможность загружать свои данные для анализа и «разделять» результаты работы

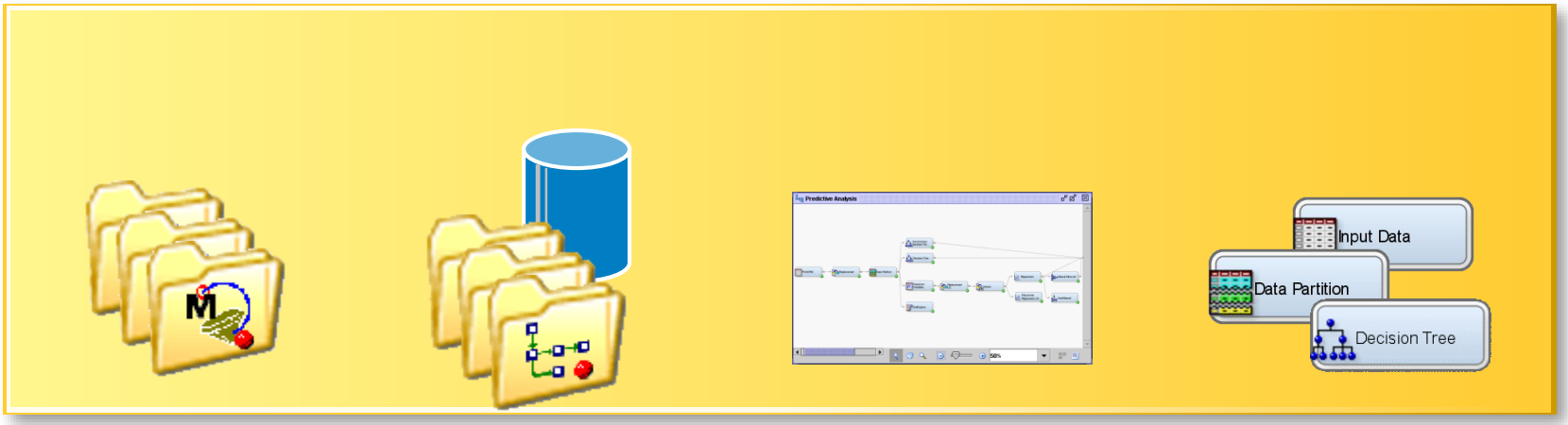
ОРГАНИЗАЦИЯ РАБОЧЕГО ПРОСТРАНСТВА В SAS ENTERPRISE MINER

- Проекты
- Источники данных
- Диаграммы
- Процессы
- Задачи

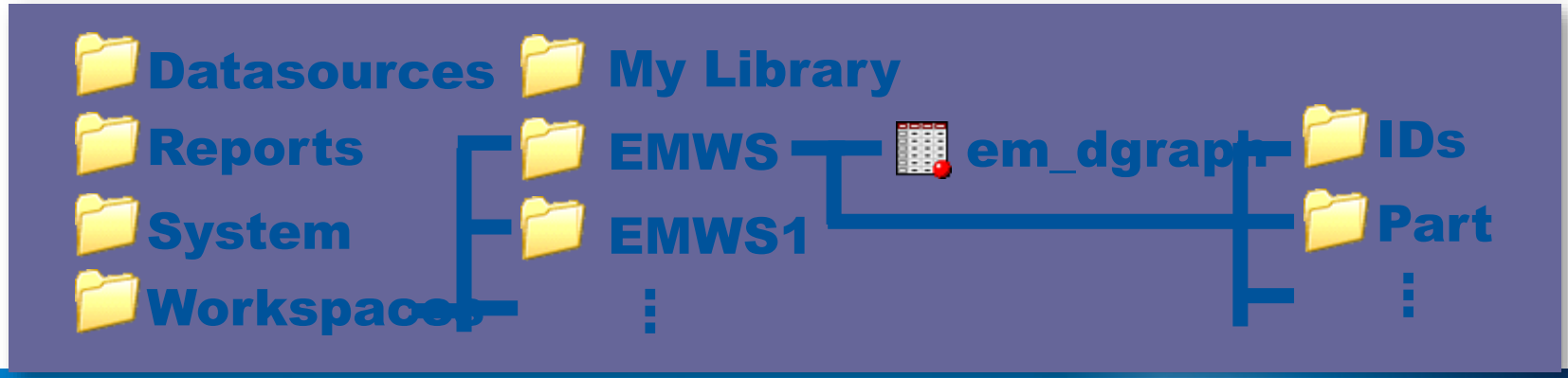
- Основная структура данных - табличная



ОСНОВНЫЕ СУЩНОСТИ DATA MINING ПРОЕКТА



Projects | Libraries and Diagrams | Process Flows | Nodes



КОНЦЕПЦИЯ SEMMA

- Sample (Выборка данных)
  - Создание наборов данных для анализа из источников «сырых» данных (только выбирает, не создает новых значений и не видоизменяет данные)
- Explore (Исследование данных)
  - Разведочный анализ данных, включает ряд алгоритмов «обучения без учителя» и богатые средства визуализации
- Modify (Преобразование данных)
  - Алгоритмы преобразования данных, включая алгоритмы уменьшения размерности, выбора значимых признаков и т.д.
- Model (Построение моделей)
  - Построение моделей прогнозирования
- Assess (Оценка моделей)
  - Выбор и сравнение моделей

КОНЦЕПЦИЯ SEMMA

- Sample (Выборка данных):
  - Подключение источников данных «внешних» и «внутренних»
  - Определение ролей источников, выделение структуры прецедентов, задание ролей и типов характеристик прецедентов
  - Разбиение на тренировочный, тестовый и валидационный наборы (несколько стратегий)
  - Очистка данных (удаление ненужных прецедентов)
  - Случайная выборка (несколько стратегий)
  - «Вертикальная» и «горизонтальная» склейка данных
  - Агрегирование транзакций во временной ряд

КОНЦЕПЦИЯ SEMMA

- Explore (Исследование данных)
  - Богатые графические средства визуализации
  - Различные методы кластеризации (включая SOM, LVQ, k-means) и средства визуализации
  - Методы ассоциативного анализа (включая иерархические правила, а также анализ последовательностей и связей)
  - Методы уменьшения размерности (выбор ключевых характеристик)
  - Методы «кластеризации» переменных

КОНЦЕПЦИЯ SEMMA

- Modify (Преобразование данных)
  - Методы импутации пропущенных значений (точечные оценки и на основе прогнозных моделей)
  - Поиск главных компонент
  - Интерактивная дискретизация (зависящая от отклика)
  - Богатые средства определения пользовательских процедур преобразования данных
  - Стандартные преобразования числовых и дискретных характеристик с возможностью автоматического выбора оптимального преобразования

КОНЦЕПЦИЯ SEMMA

- Model (Построение моделей)
  - Модели на основе деревьев решений для задач классификации и регрессии с различными критериями построения деревьев
  - Регрессионные модели, включая линейную, полиномиальную, логистическую, LASSO, PLS, собственные разработки SAS
  - Нейро-сетевые модели, включая многослойные персептроны, радиально-базисные сети, GLM, а также методы «оптимального» выбора архитектуры сети и собственные разработки SAS
  - MBR kNN
  - Комбинированные модели для прогнозирования редких событий
  - Ансамбли (голосущие, усредняющие, бустинг, баггинг, …)

КОНЦЕПЦИЯ SEMMA

- Assess (Оценка моделей)
  - Вычисление оценок качества моделей
  - Графические средства сравнения качества и визуализации найденных закономерностей
  - Средства выбора оптимального порога для задач принятия решений
  - Средства интеграции в процесс поддержки принятия решений
  - Средства применения моделей

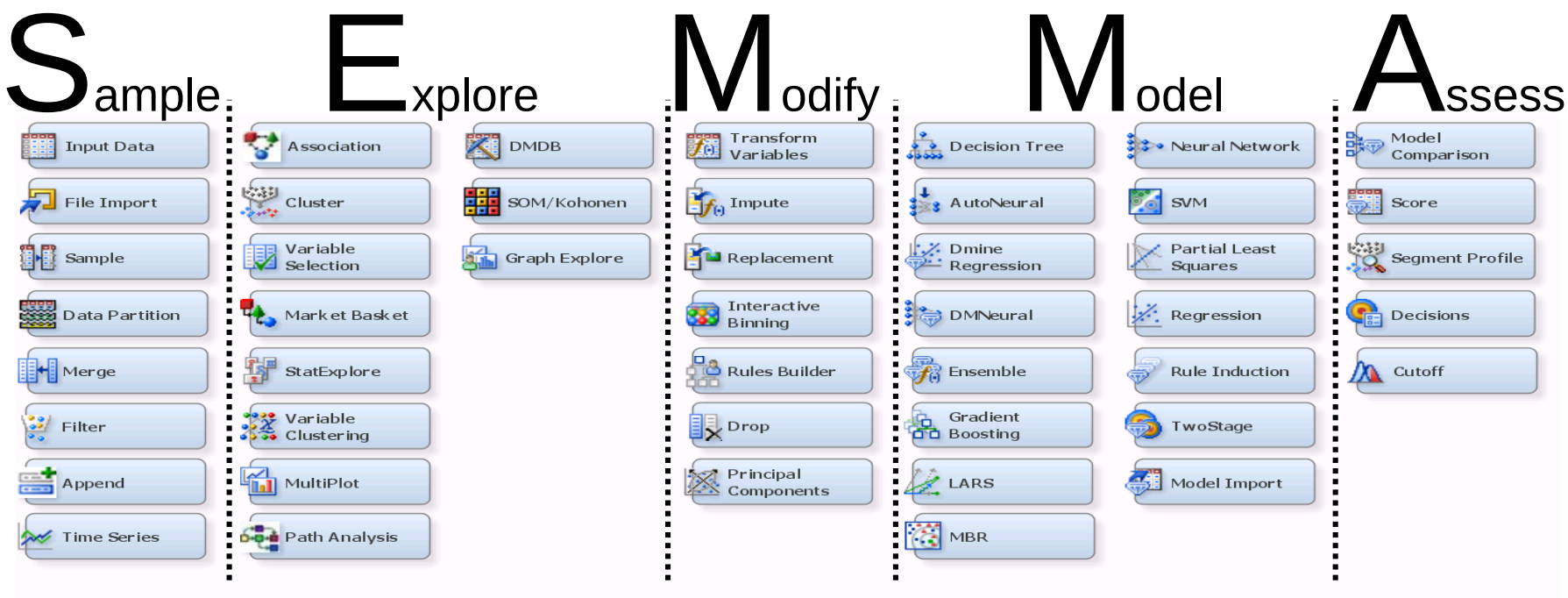
КОНЦЕПЦИЯ SEMMA

- Служебные компоненты
  - Экспорт результатов и создание отчетов
  - Редактирование метаданных
  - Создание своих компонент, интеграция с SAS Code, R
  - «Циклы» для кросс валидации, бустрепинга, бэггинга и др.
- Высокопроизводительные компоненты
  - «Параллеьльные» версии стандартных алгоритмов и методов обработки данных (регрессии, GLM, деревья решений, PCA, выбор значимых переменных, partition и др.)
  - Специальные «параллельные» алгоритмы (SVM, Random Forest, deep learning NN и др.)

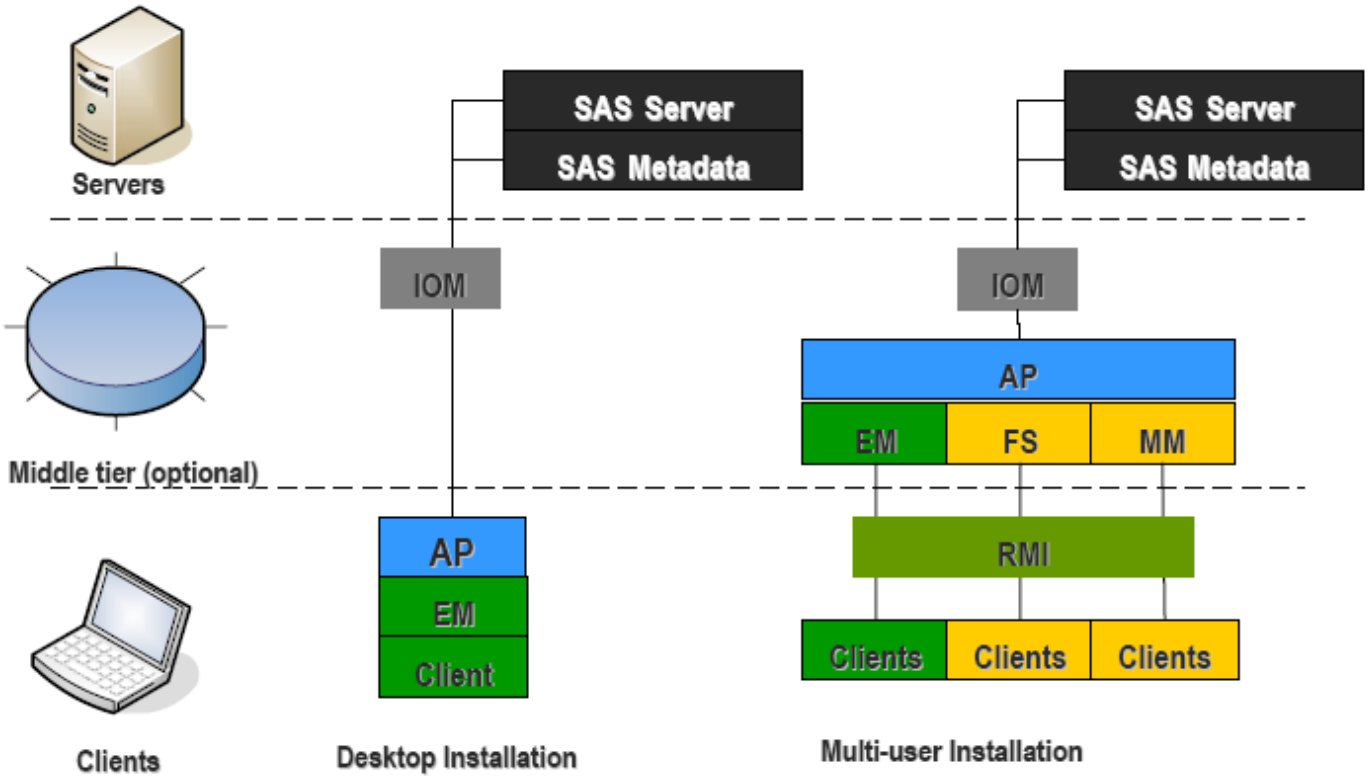
КОНЦЕПЦИЯ SEMMA

- Text mining
  - Импорт корпксов текста
  - Фильтрация и парсинг
  - Кластеризация и рубрикация
  - Выявление ключевых слов и тематик
- Анализ временных рядов
  - Импорт данных
  - Сглаживание, кластеризация
  - Декомпозиция и уменьшение размерности

КОНЦЕПЦИЯ SEMMA



3-Х УРОВНЕВАЯ АРХИТЕКТУРА



EM = SAS Enterprise Miner
AP = SAS Analytics Platform
FS = SAS Forecast Server
MM = SAS Model Manager

ONDEMAND SOLUTION

- «Облачная» Saas модель:
  - Хостится, управляется и конфигурируется SAS
  - Пользователь ставит только Java клиент
  - Полные функциональные возможности по сравнению со стандартной версией
  - Доступен «все время» «отовсюду»
  - Есть возможность загружать свои данные для анализа и «разделять» результаты работы