Введение в методы интеллектуального анализа данных (Data Mining)

к.ф.-м.н. М.И. Петровский (michael@cs.msu.su), SAS Certified Data Scientist

лаборатория «Технологий программирования»
ВМиК МГУ им. М.В. Ломоносова

Задачи курса

- Познакомить с предметной областью:
  - дать основные определения и терминологию, обсудить прикладные задачи
- Рассмотреть основные задачи Data Mining:
  - и популярные алгоритмы на основе методов машинного обучения для их решения
  - меньше теорем, больше алгоритмов и понимания как их настраивать и использовать
- Дать практический опыт работы с промышленными системами Data mining:
  - Практические задания в SAS Enterprise Miner (не все рассматриваемые в курсе алгоритмы есть в Enterprise Miner)

Содержание курса (1/3)

1. Введение
2. Выявление структур в данных (обучения без учителя)
   - Поиск ассоциативных правил (алгоритмы apriori и fp-tree) и тематическое моделирование (методы главных компонент, неотрицательная матричная факторизация)
   - Кластеризация (иерархическая, метрическая, вероятностная)

Содержание курса (2/3)

3. Задача прогнозирования (обучение с учителем)
   - Виды задач прогнозирования, проблема переобучения, оценка и сравнение моделей, простейшие методы прогнозирования (kNN и Naïve Bayes)
   - Методы на основе деревьев решений и их ансамблей
   - Регрессионные модели (отбор и преобразование пространства признаков, регуляризация, обобщенные линейные модели)

Содержание курса (3/3)

3. Задача прогнозирования (обучение с учителем)
   - Нейронные сети (MLP, RBF, борьба с переобучением, SOM, задачи глубинного обучения)
   - Методы опорных векторов для задач классификации и регрессии
4. Выявление аномалий

ВСЕГО 9 ЛЕКЦИЙ и ПРАКТИЧЕСКИЕ ЗАДАНИЯ!!!
Итог = практические задания + посещаемость + экзамен

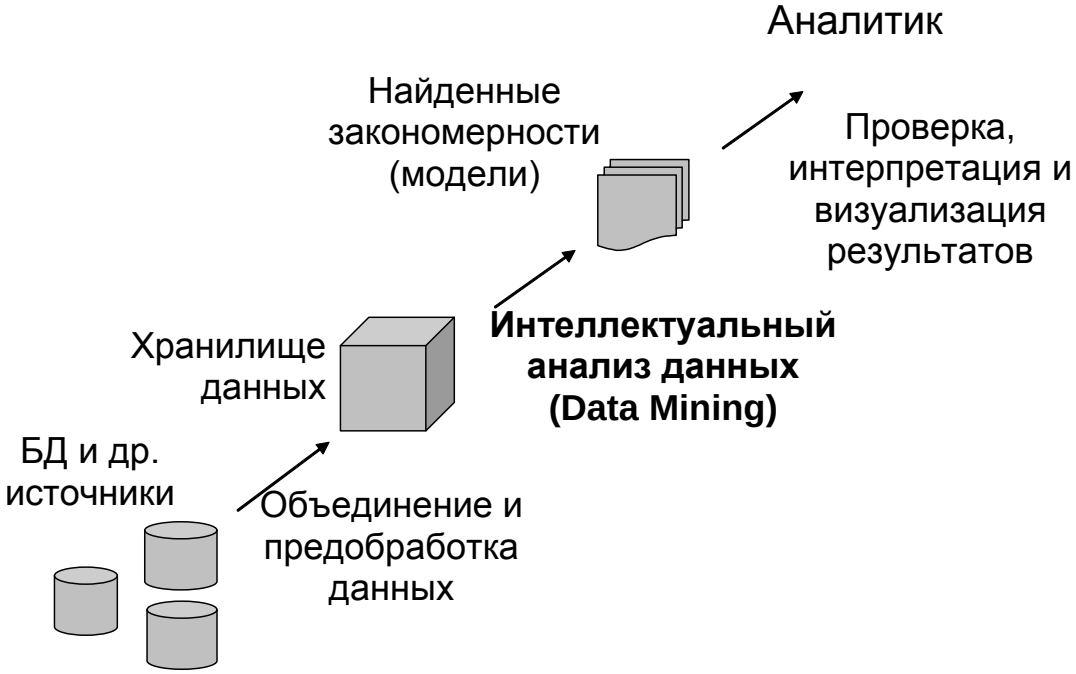
Эволюция технологий хранения и обработки данных

- … — 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.

Интеллектуальный анализ данных (Data Mining)



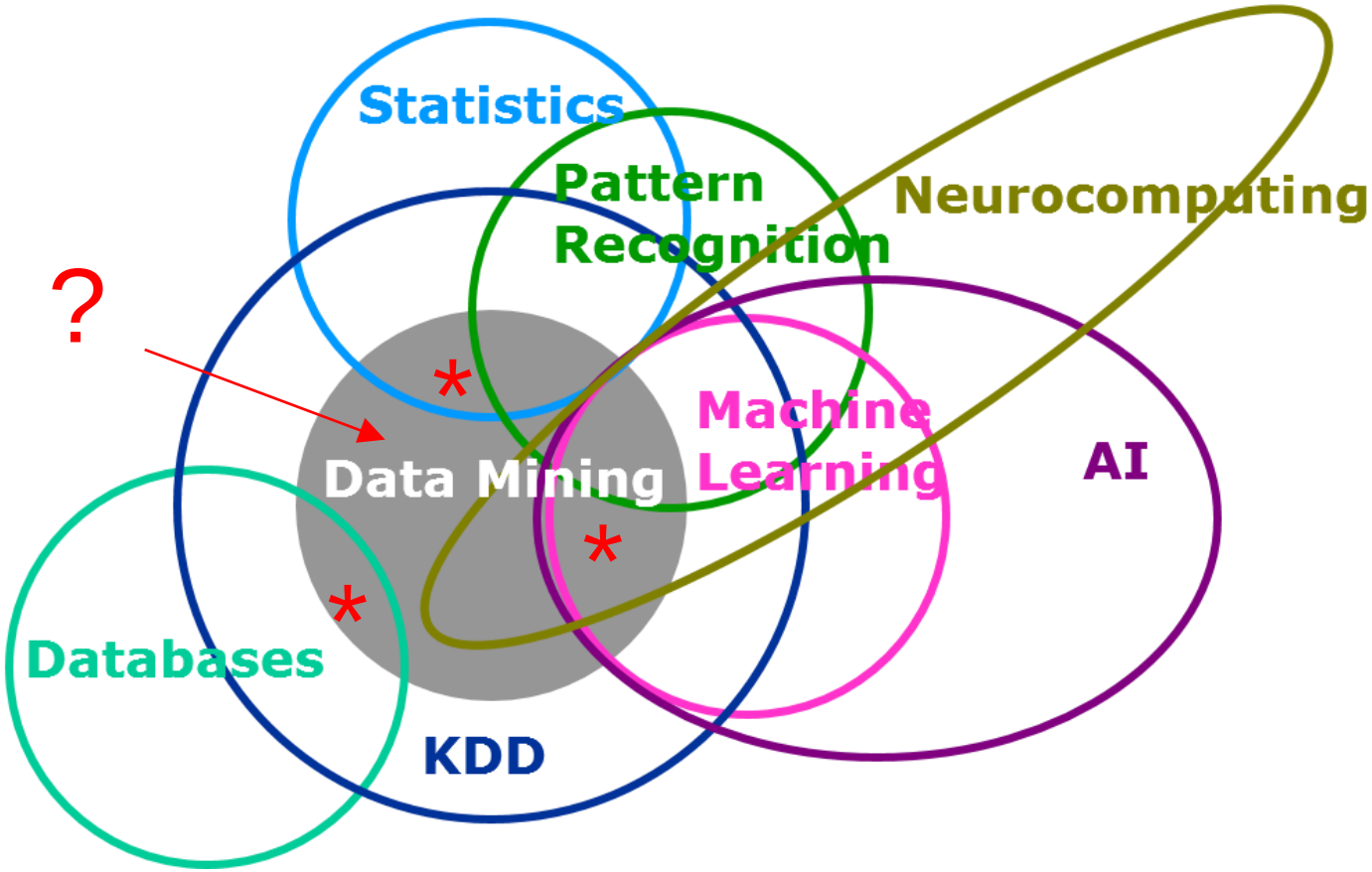
Системы *интеллектуального анализа данных* (ИАД) – класс программных систем поддержки принятия решений, задачей которых является поиск *скрытых, ранее неизвестных, содержательных и потенциально полезных* закономерностей в *больших объемах разнородных, сложно структурированных* данных.

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

Краткая история ИАД

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- Другие конференции по data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

Место Data mining среди современных подходов анализа данных



Обратите внимание на пересечения областей!

Кто такой Data Scientist?

- «три в одном»:
  1. Понимает прикладную предметную область, в которой строит модель, а в идеале является экспертом в этой области
  2. Владеет математическим аппаратом прикладной статистики и машинного обучения, знает много алгоритмов и методов анализа данных, умеет их настраивать, а в идеале может разрабатывать собственные
  3. Является программистом и специалистом по системам хранения данных (в том числе Big Data), может писать код для эффективной выборки и обработки больших объемов сложно структурированных данных, а в идеале может сам проектировать хранилища и программировать методы анализа данных

Раньше это было три разных человека, с разным образованием, но сейчас поняли что это не эффективно и надо объединять!!!

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, возможно, реализация подсистемы сбора (консолидации)
  - 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:
  - 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
- Проверка моделей и представление результатов анализа:
  - 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

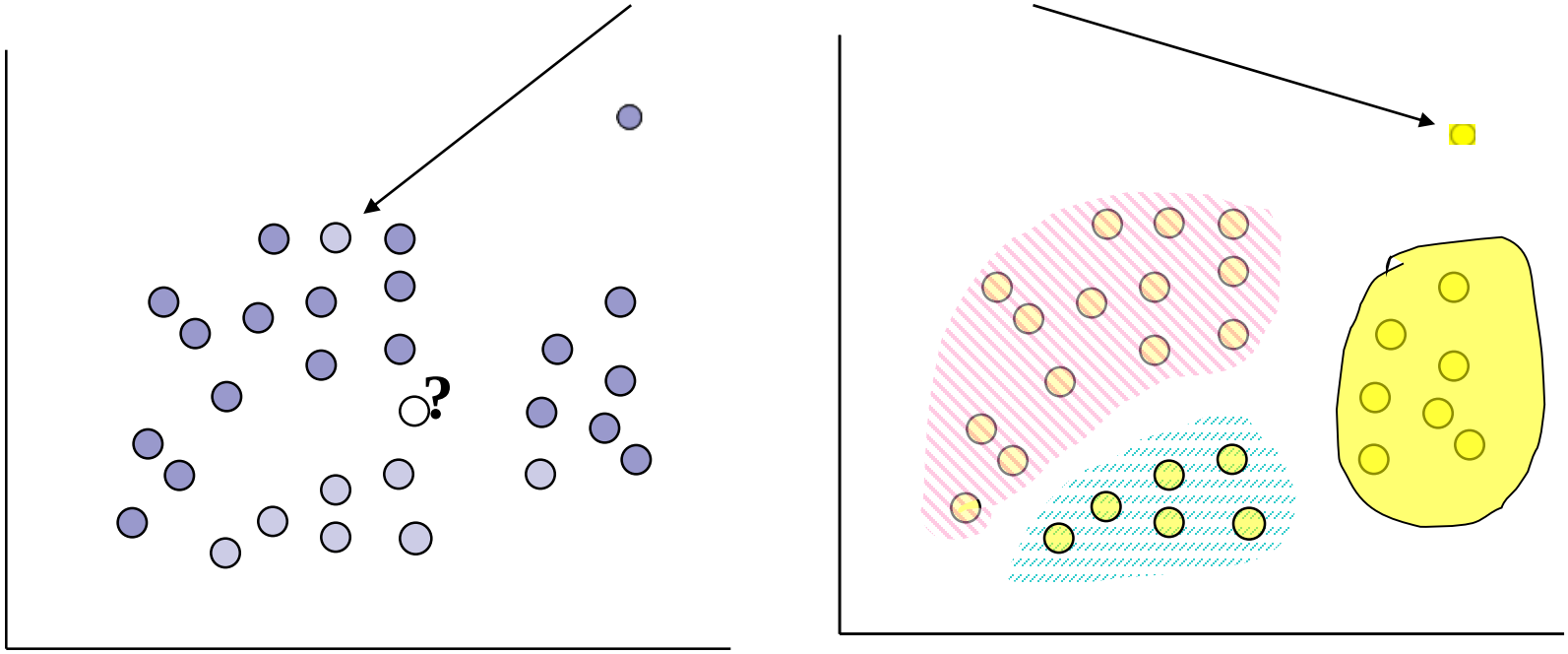
- Транзакционные
  - Объекты анализа – «события» различной структуры с числовыми и категориальными атрибутами и с временной меткой
- Табличные
  - Объекты анализа представлены в виде реляционных таблиц, возможно взаимосвязанных (заданно ER-схемой), имеют разнотипные атрибуты
- Временные ряды и числовые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

Данные для анализа

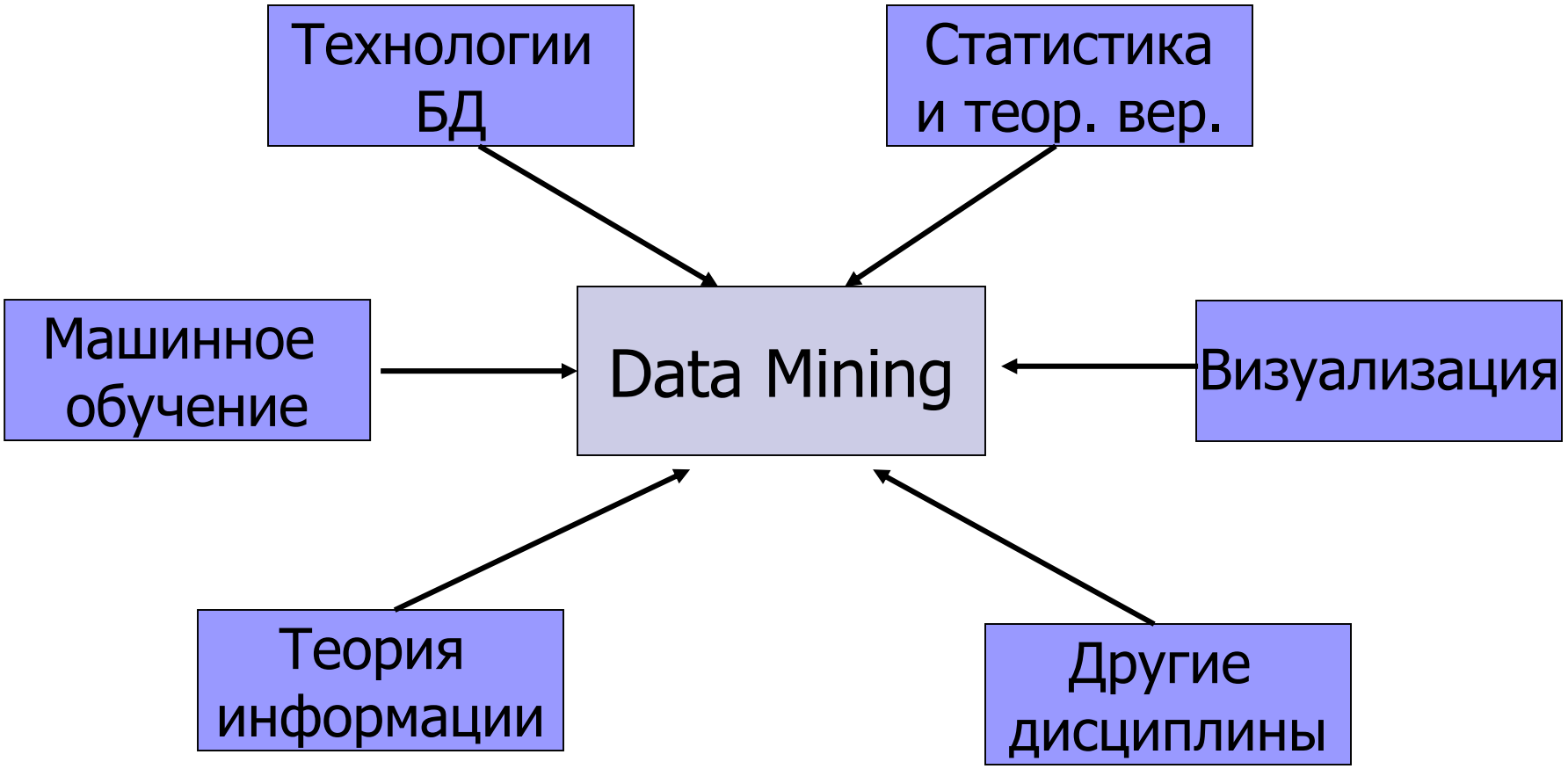
- Объект анализа (или прецедент, или кейс, или наблюдение, …) задается набором признаков (или атрибутов, или свойств, …)
- Признаки по типам бывают:
  - Категориальные - нет расстояний, не задан порядок
  - Ординальные (порядковые) – нет расстояний
  - Числовые – есть расстояние
- «Размеченный» набор 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



Методы анализа



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, выявление скрытых свойств объектов
- 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

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

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 и информационный поиск - рубрикация документов
  - Кредитование - оценка заемщика
  - Рекомендательные системы

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

Поиск ассоциаций

- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти зависимости между значениями атрибутов в виде правил «если … то …»
  - Найти аналитические зависимости между атрибутами и выявить скрытые признаки и характеристики
- Приложения:
  - Маркетинг и рекомендательные системы - анализ зависимостей между покупаемыми товарами или услугами
  - Финансовый анализ – поиск зависимостей между значениями индексов и другими финансовыми параметрами
  - Латентно-семантический анализ текстов

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, товаров и т.д.
  - Производство – выявление типовых состояний и ситуаций
  - Индексирование документов

Выявление исключений

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

Отличия ИАД систем (1)

- Наличие «обучения»
  - база знаний 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 статистических пакетов, ориентированных на расчет статистик, проверку гипотез и оценку параметров распределений)

Отличия ИАД систем (2)

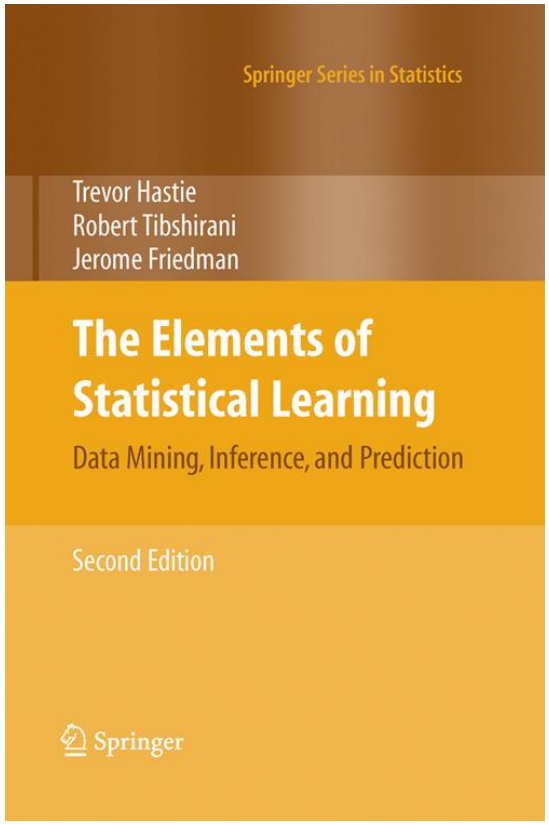
- Наличие большого объема данных сложной структуры
  - зачастую скорость работы алгоритмов в ИАД важнее отклонений по точности (“quick and dirty solution”)
  - большинство алгоритмов работают с исходными данными в виде числовой матрицы признаков, сложная структура реальных объектов в ИАД приводит к необходимости 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

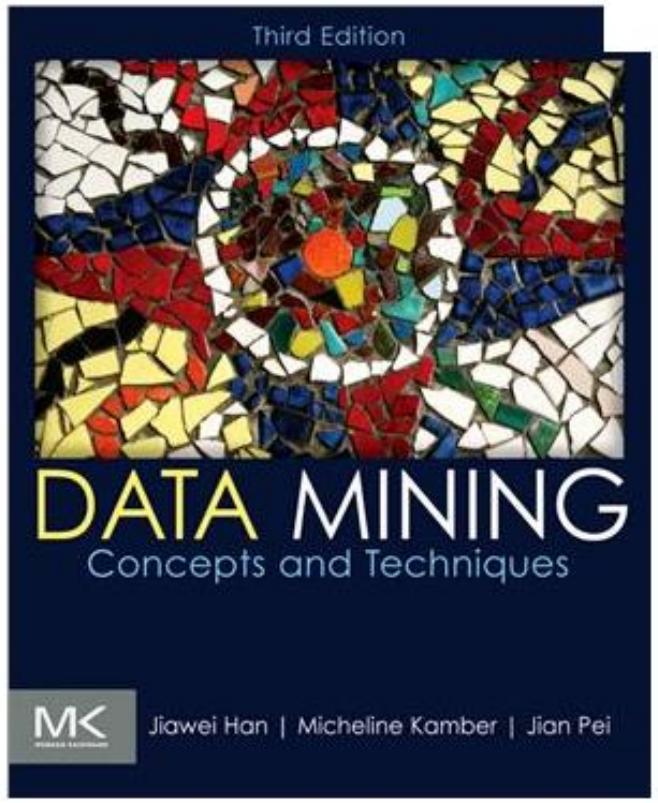
Отличия ИАД систем (3)

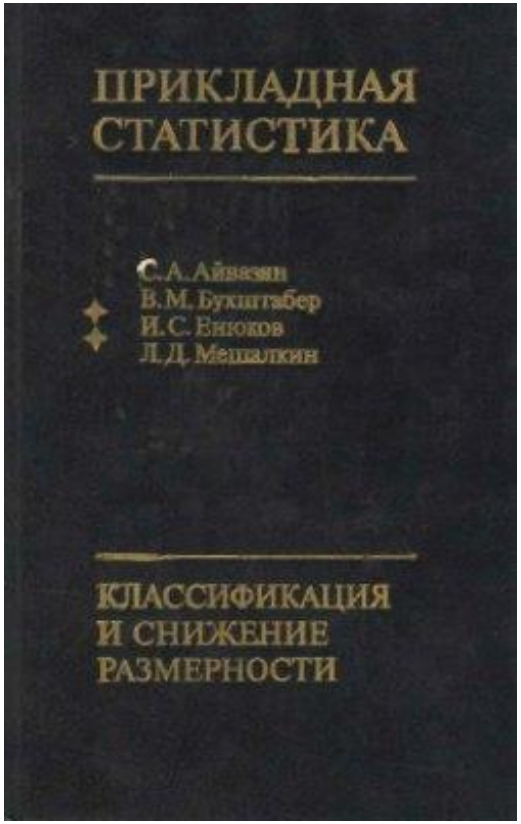
- Наличие человека - аналитика как оконечного потребителя результатов работы ИАД системы
  - в сценарии работы любой системы ИАД всегда присутствует аналитик, даже если полученная в результате модель далее используется для автоматической классификации
  - аналитик формирует тренировочные наборы, производит настройку алгоритмов, обучение и дообучение, анализирует полученные модели и принимает решения об их дальнейшем использовании
  - таким образом, системы автоматической классификации, кластеризации и распознавания образов, даже использующие возможность дообучения, 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)

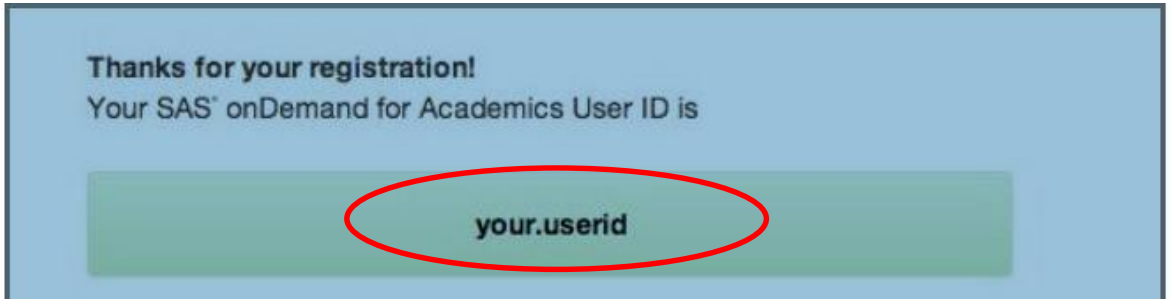
OnDemand Solution

- «Облачная» 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

SAS OnDemand for Academics

РЕГИСТРИАЦИЯ И УСТАНОВКА ПО

- Step 1: Зарегистрируйте SAS профайл и получите **Academic User id** https://odamid.oda.sas.com/SASODARegistration/

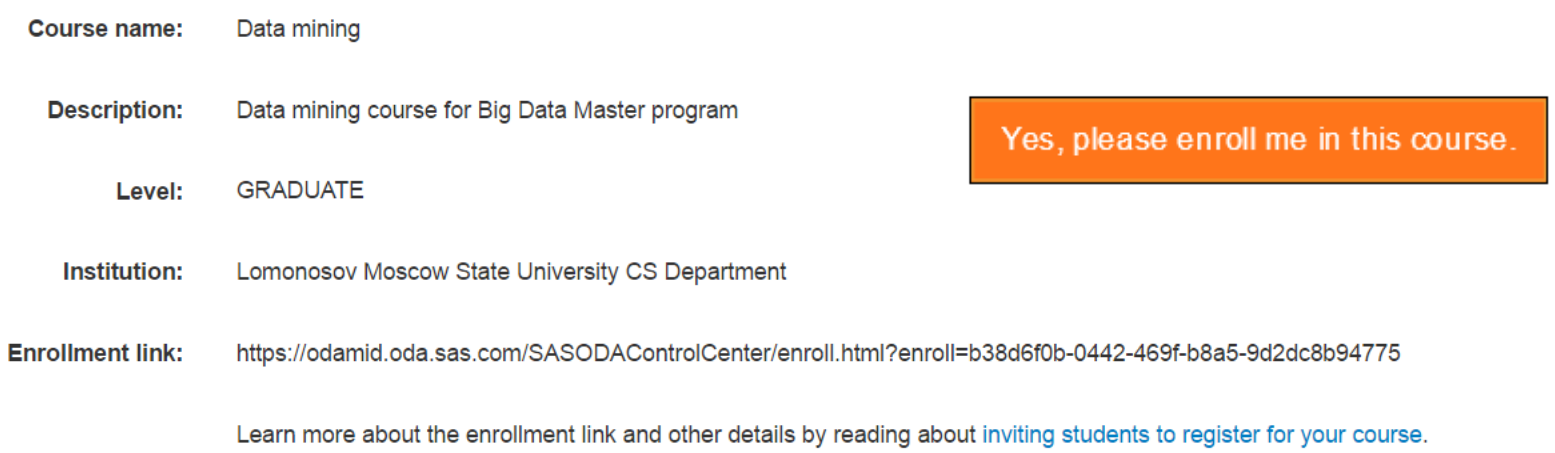


- Детальные инструкции: http://support.sas.com/software/products/ondemand-academics/manuals/EnterpriseMinerStudent.pdf

SAS OnDemand for Academics

РЕГИСТРАЦИЯ И УСТАНОВКА ПО

- Step 2: Подпишитесь на курс по ссылке: https://odamid.oda.sas.com/SASODAControlCenter/enroll.html?enroll=b38d6f0b-0442-469f-b8a5-9d2dc8b94775

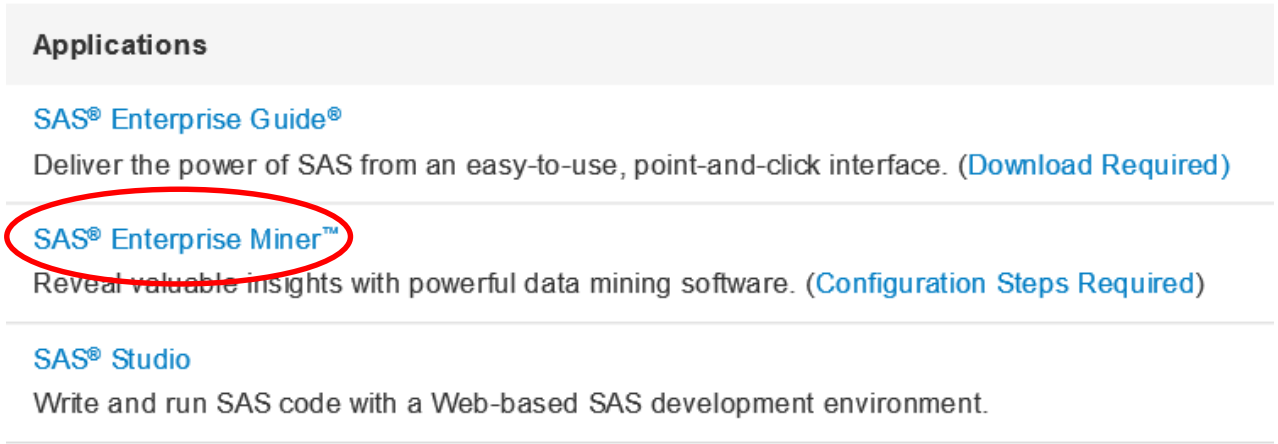


- Инструкции: http://support.sas.com/software/products/ondemand-academics/manuals/EnterpriseMinerStudent.pdf

SAS OnDemand for Academics

РЕГИСТРАЦИЯ И УСТАНОВКА ПО

- Step 3: Скачайте SAS Enterprise Miner

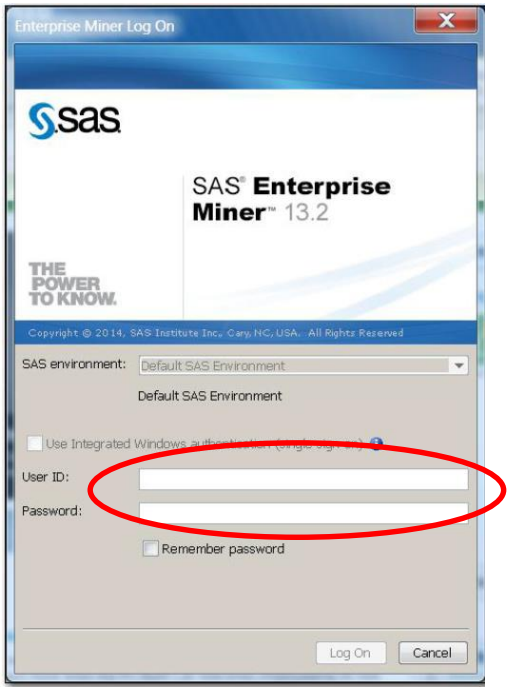


- Инструкции: http://support.sas.com/software/products/ondemand-academics/manuals/EnterpriseMinerStudent.pdf

SAS OnDemand for Academics

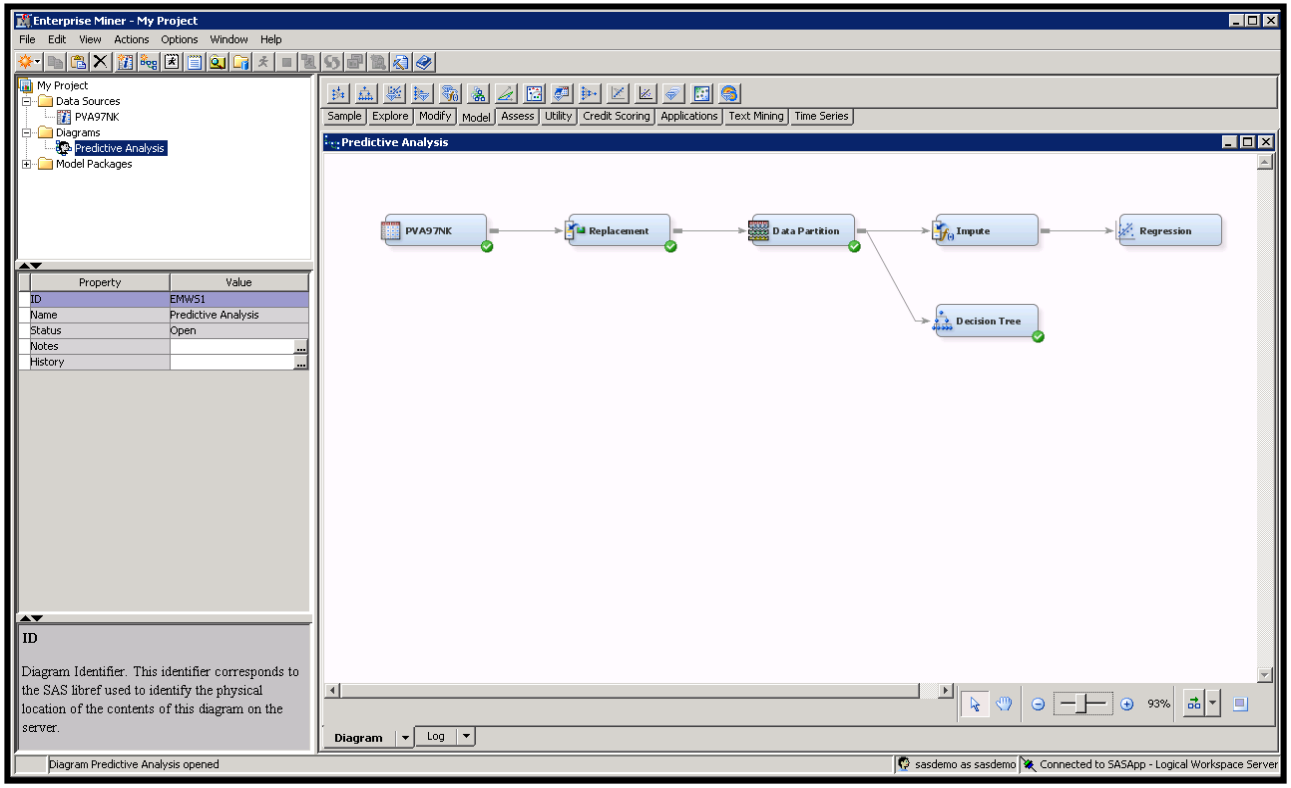
РЕГИСТРАЦИЯ И УСТАНОВКА

- Step 4: Вход в SAS Enterprise Miner



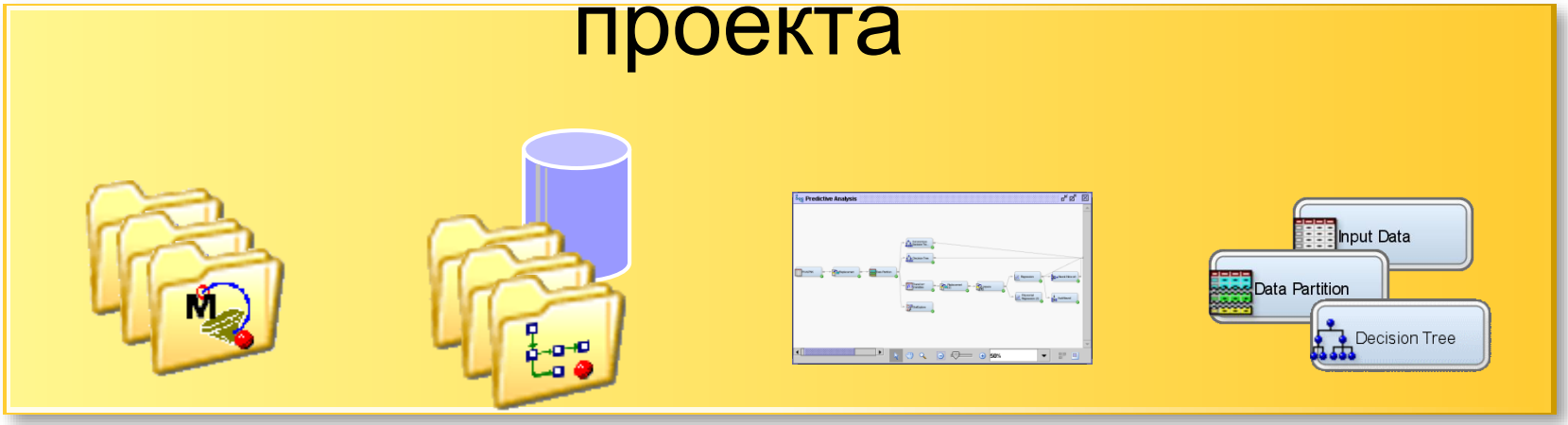
- Инструкции:
- http://support.sas.com/software/products/ondemand-academics/manuals/EnterpriseMinerStudent.pdf

SAS Enterprise Miner

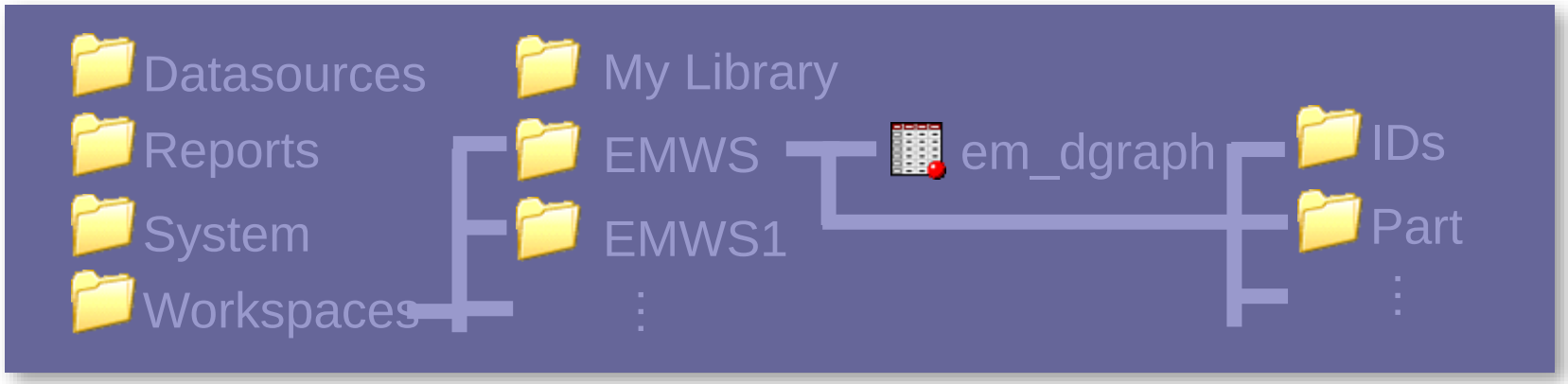


Одна сессия – один рабочий проект

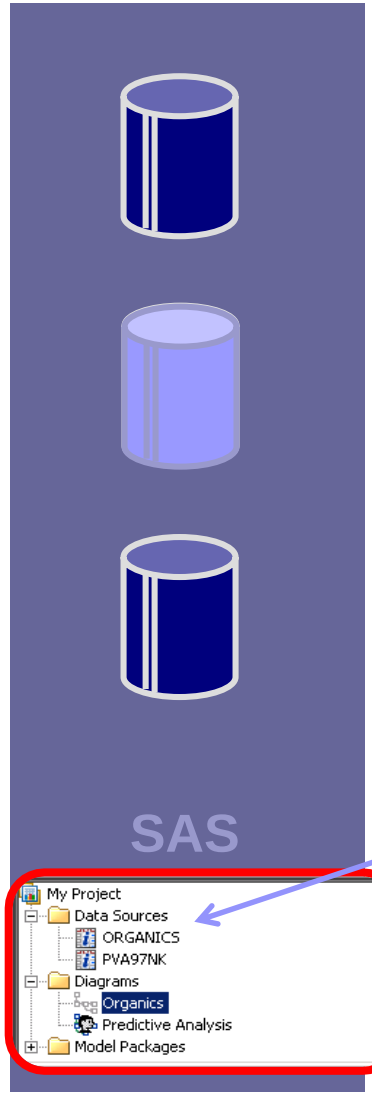
Основные сущности data mining проекта



Projects | **Libraries and Diagrams** | **Process Flows** | **Nodes**

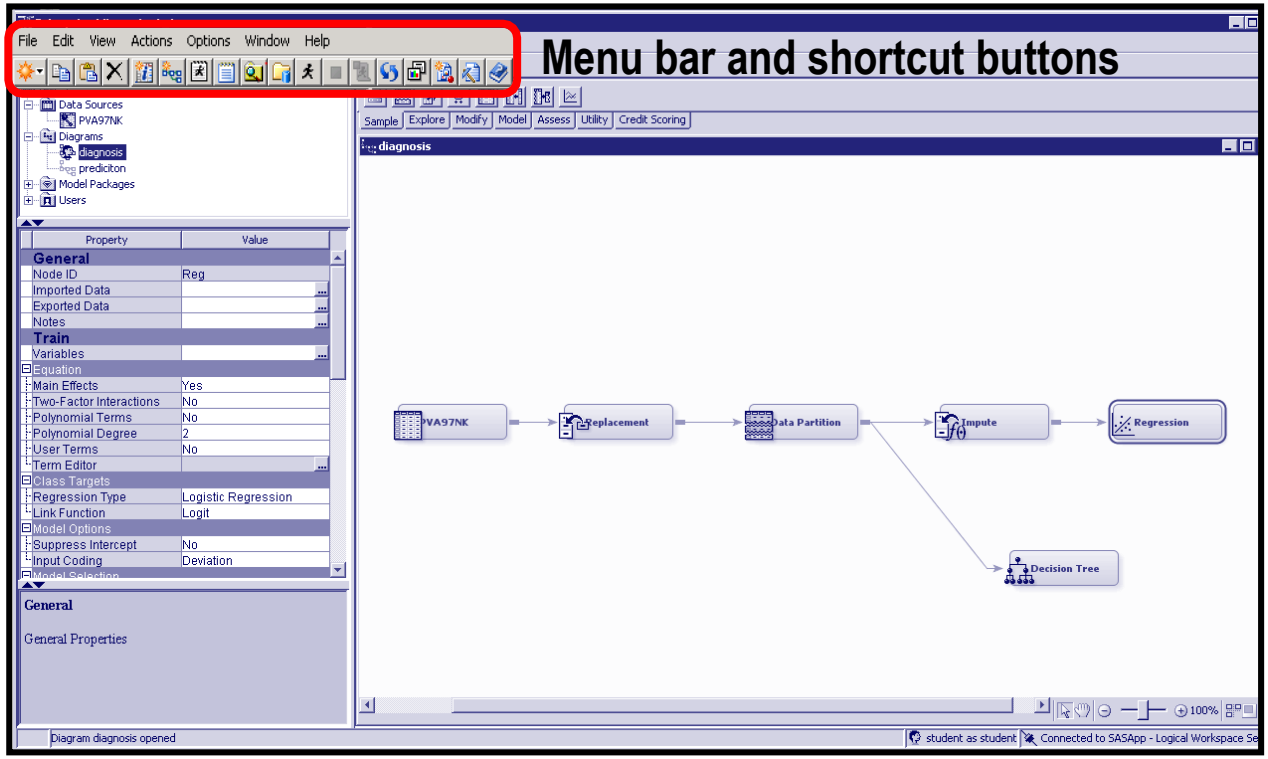


Подключение источника данных

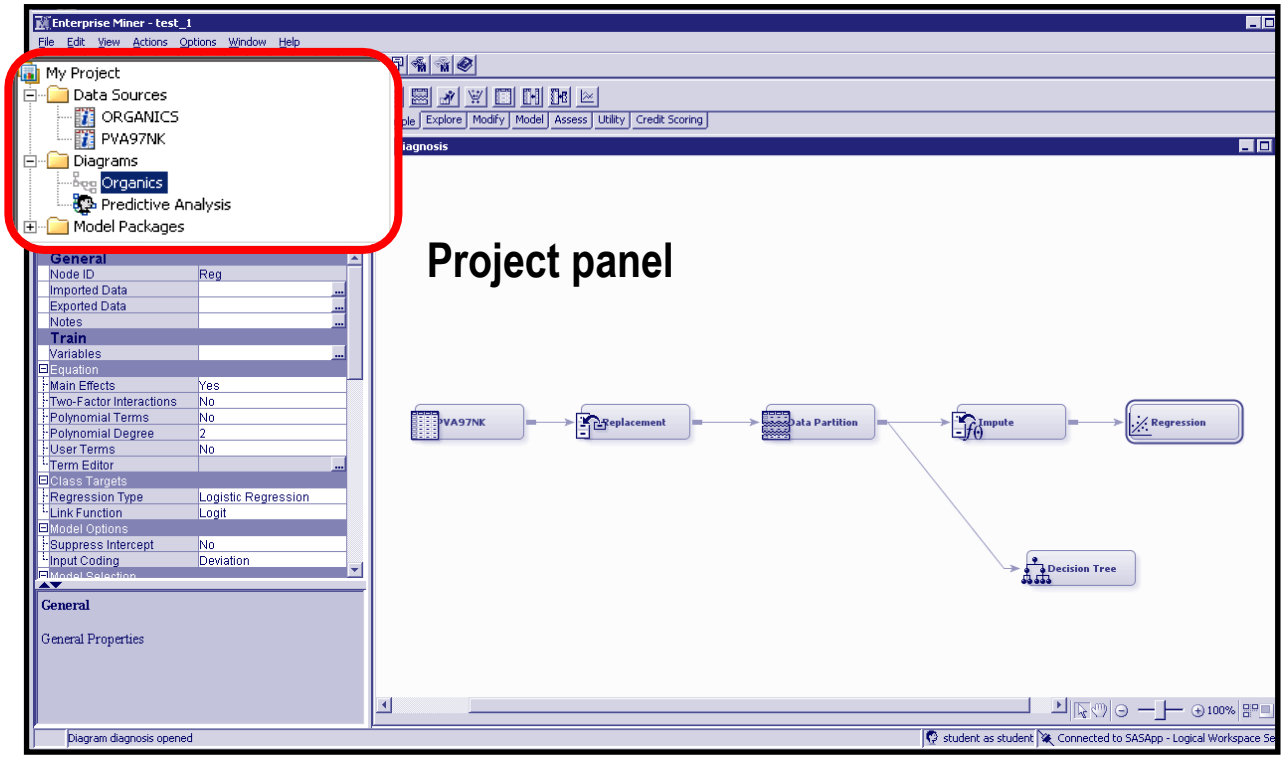


1. Создать библиотеку (File->new->library):
   - Связать с файловой директорией, где лежат наборы данных в формате SAS
   - Связать с соединением с БД, с хранилищем или с Name узлом Hive сервисом в Hadoop
   - Двойное имя всех источников <libname>.<dataset>
2. Создать набор данных и определить метаданные:
   - Роли переменных (ID, Input, Target и другие)
   - Типы переменных (Interval, Ordinal, Nominal)
   - Определить роль источника (Raw, Train/Test/Validate/Score, Transaction)

SAS Enterprise Miner – Обзор интерфейса

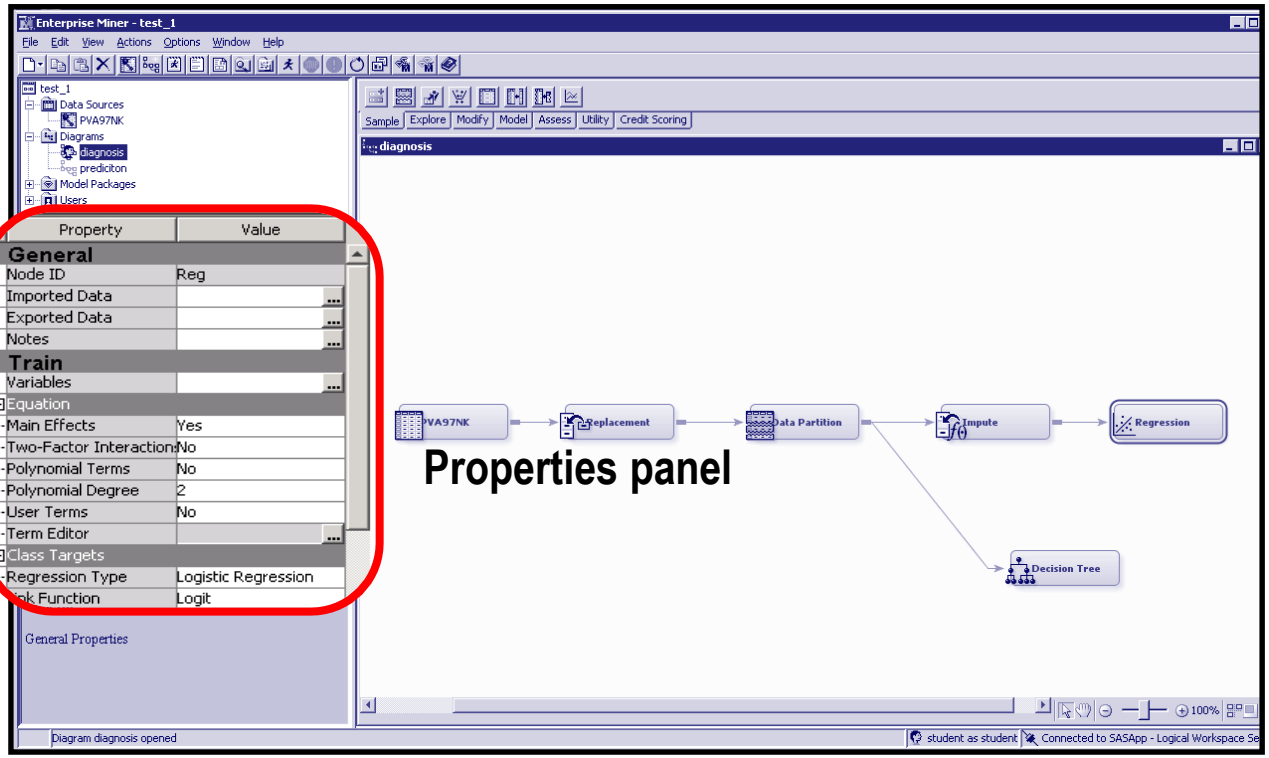


SAS Enterprise Miner – Обзор интерфейса



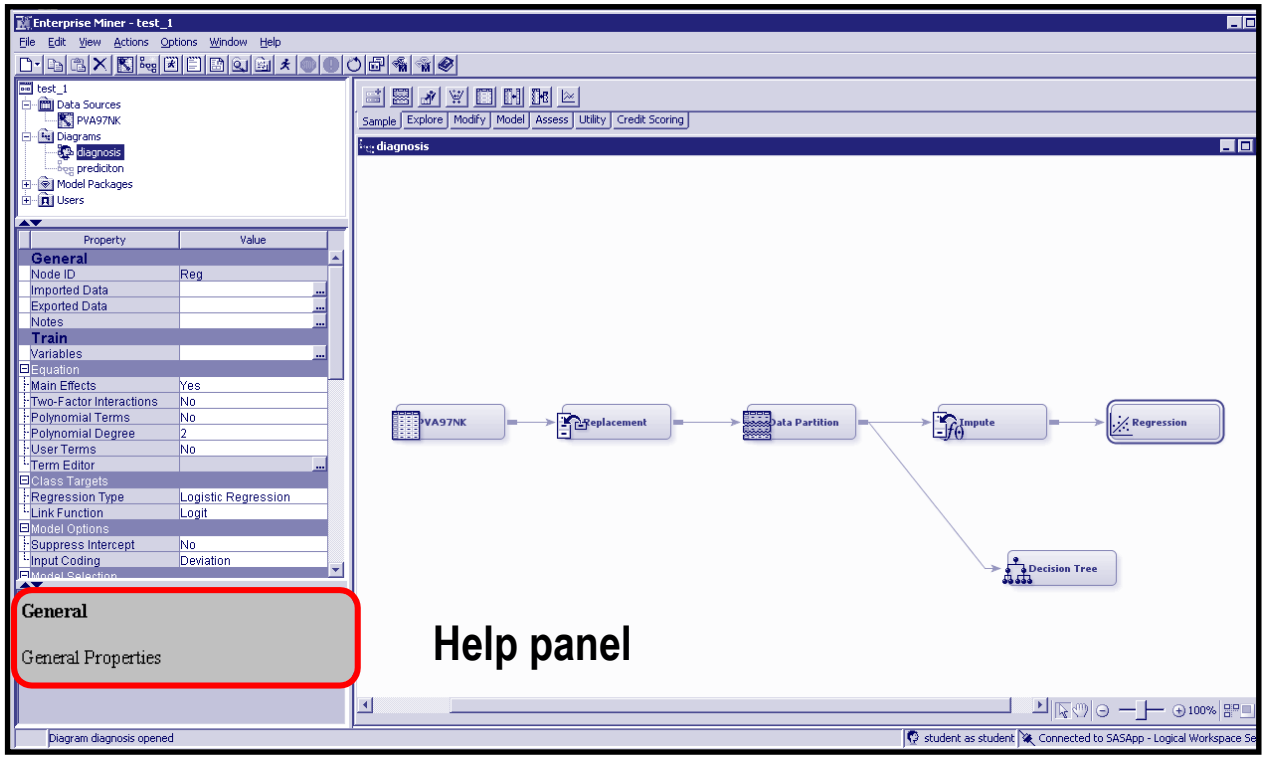
Быстрая навигация и управление источниками данных, рабочими диаграммами пакетами моделей

SAS Enterprise Miner – Обзор интерфейса



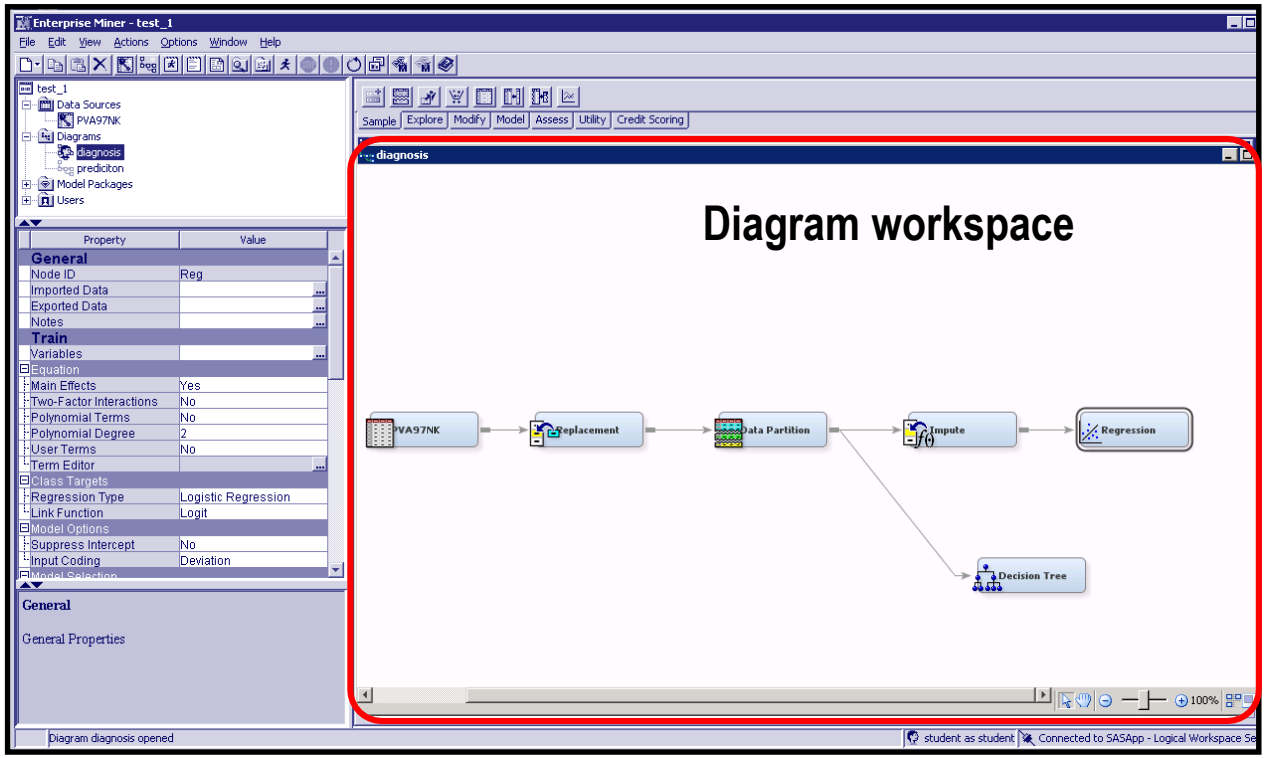
Свойства и настройки всех компонентов проекта (узлов, связей, источников данных)

SAS Enterprise Miner – Обзор интерфейса



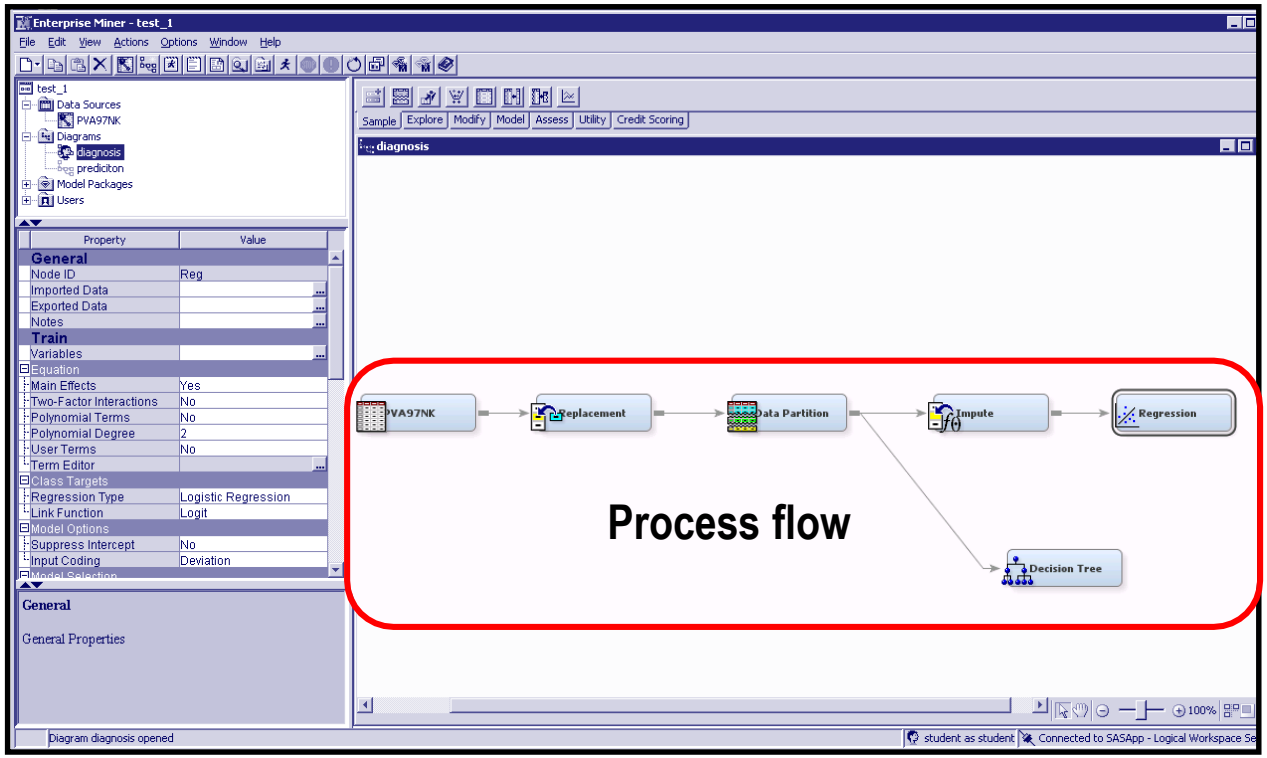
Контекстная помощь для всех компонент проекта (узлов, связей, источников данных) и значений настроек

SAS Enterprise Miner – Обзор интерфейса



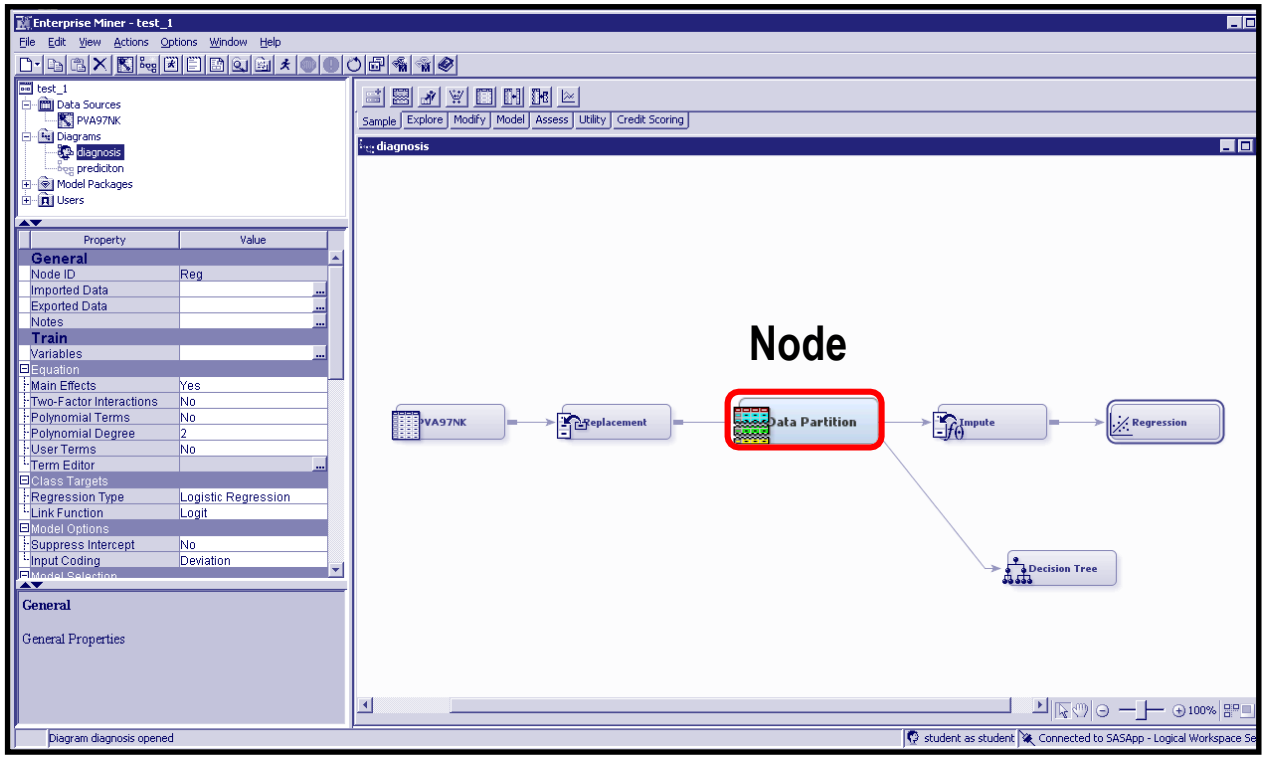
Основная рабочая область, их может быть много, содержит один или несколько блоков взаимосвязанных компонент

SAS Enterprise Miner – Обзор интерфейса



Каждый блок компонент – отдельный иад процесс:
Узел – алгоритм обработки данных
Стрелка – путь движения набора данных

SAS Enterprise Miner – Обзор интерфейса

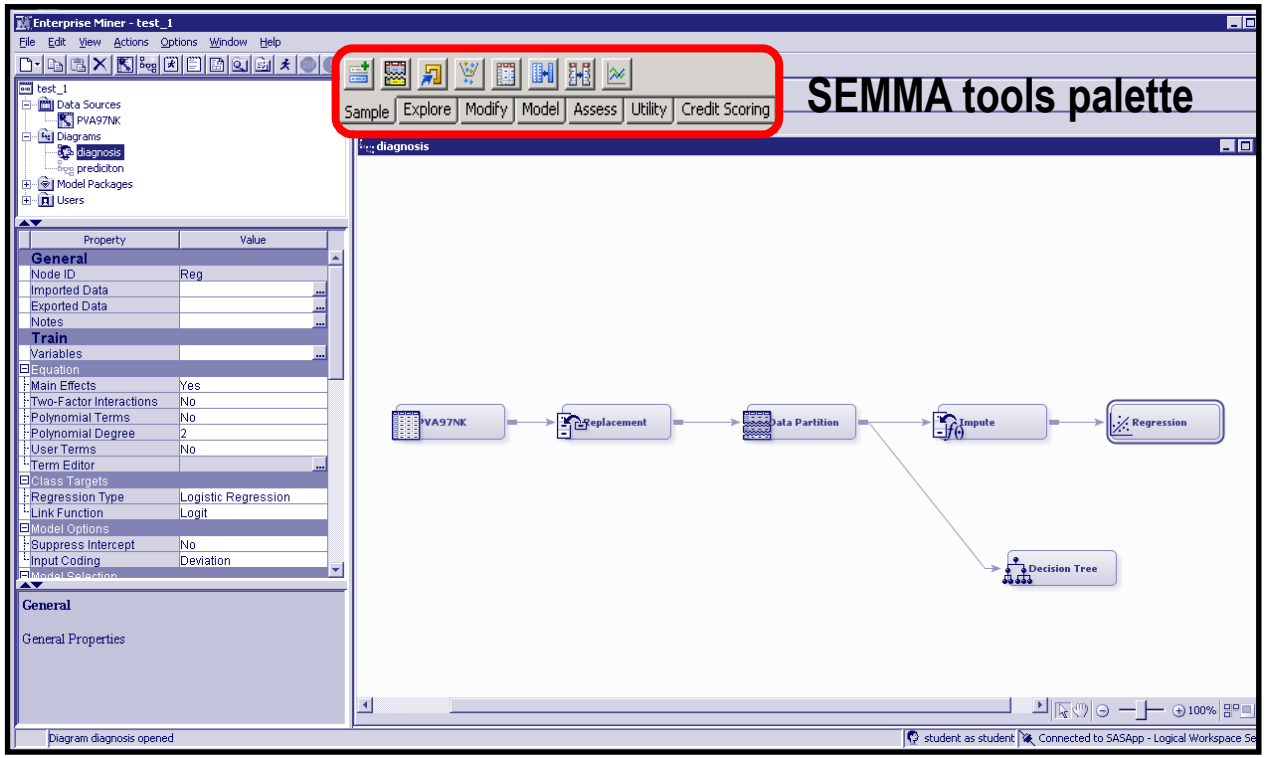


С каждым узлом неявно связано две программы (по сути первая генерирует вторую):
Первая строит модель, используя входной тренировочный набор (меню “Run”)
Вторая применяет построенную модель ко всем входным наборам данных

SAS Enterprise Miner – Обзор интерфейса

- Технически результатом применения модели может быть:
  - Журнал работы узла (есть всегда)
  - Дописывание переменных (например с результатом прогноза или номером кластера) во все входящие наборы данных
  - Создание новых наборов данных (например с параметрами модели или со статистикой)
  - Визуализация модели, статистики ее применения или просто построение интерактивных графиков
  - Изменение метаданных обрабатываемых наборов

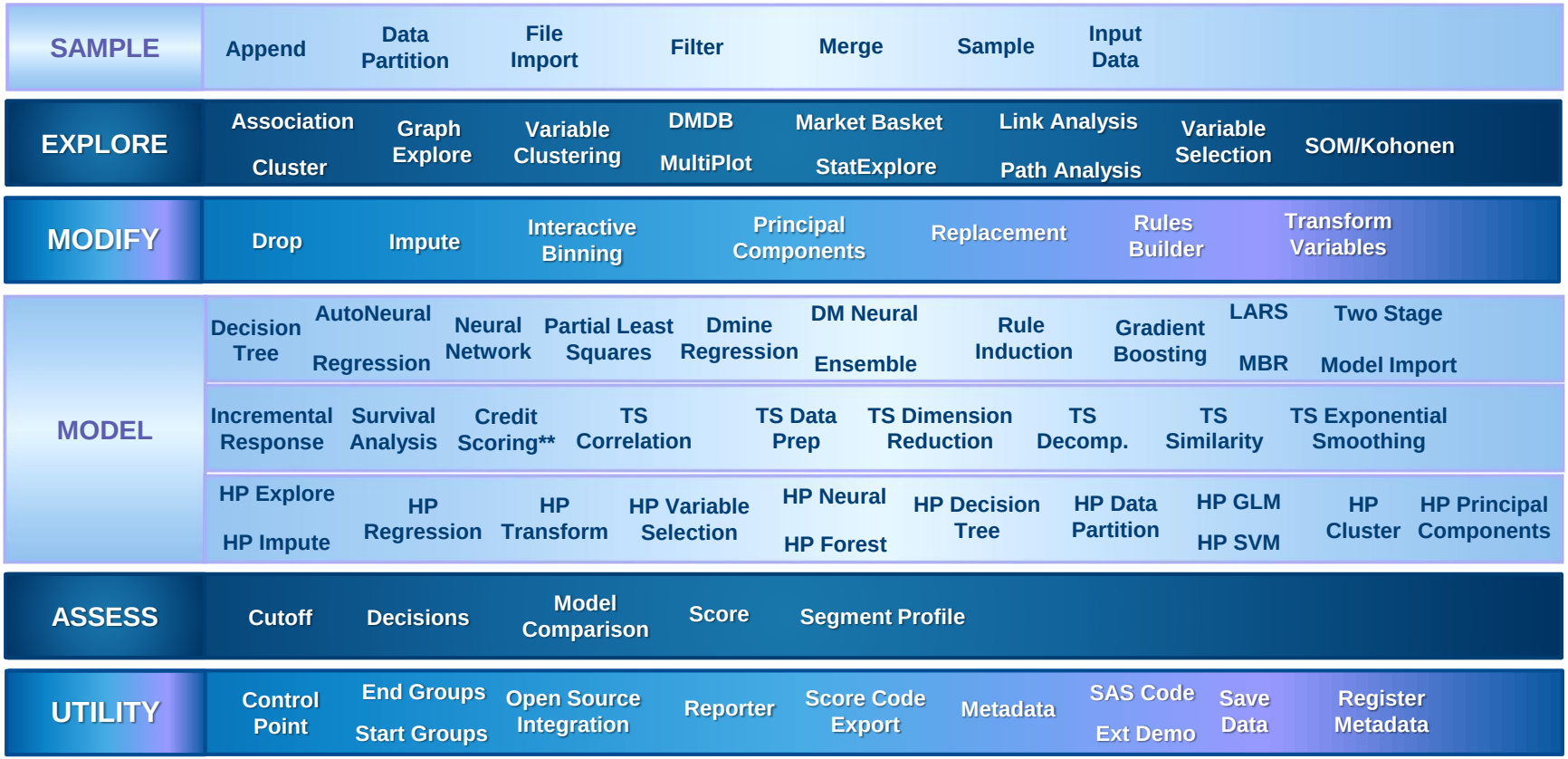
SAS Enterprise Miner – Обзор интерфейса



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

Основные узлы SAS Enterprise Miner



HP узлы (терминология)

Appliance	Кластер с высокопроизводительным хранилищем (чаще всего не на SAS)
Access Engine	Многопоточная библиотека доступа к даным (может параллельно читать и писать распределенные наборы данных, в том числе в HDFS)
Threaded Kernel (TK) Extensions	Механизм распределенных вычислений для SMP и MMP платформ (нити, MPI, MapRedcue)
HPDM nodes	Высокопроизводительные параллельные алгоритмы, использующие TK для распределенных вычислений