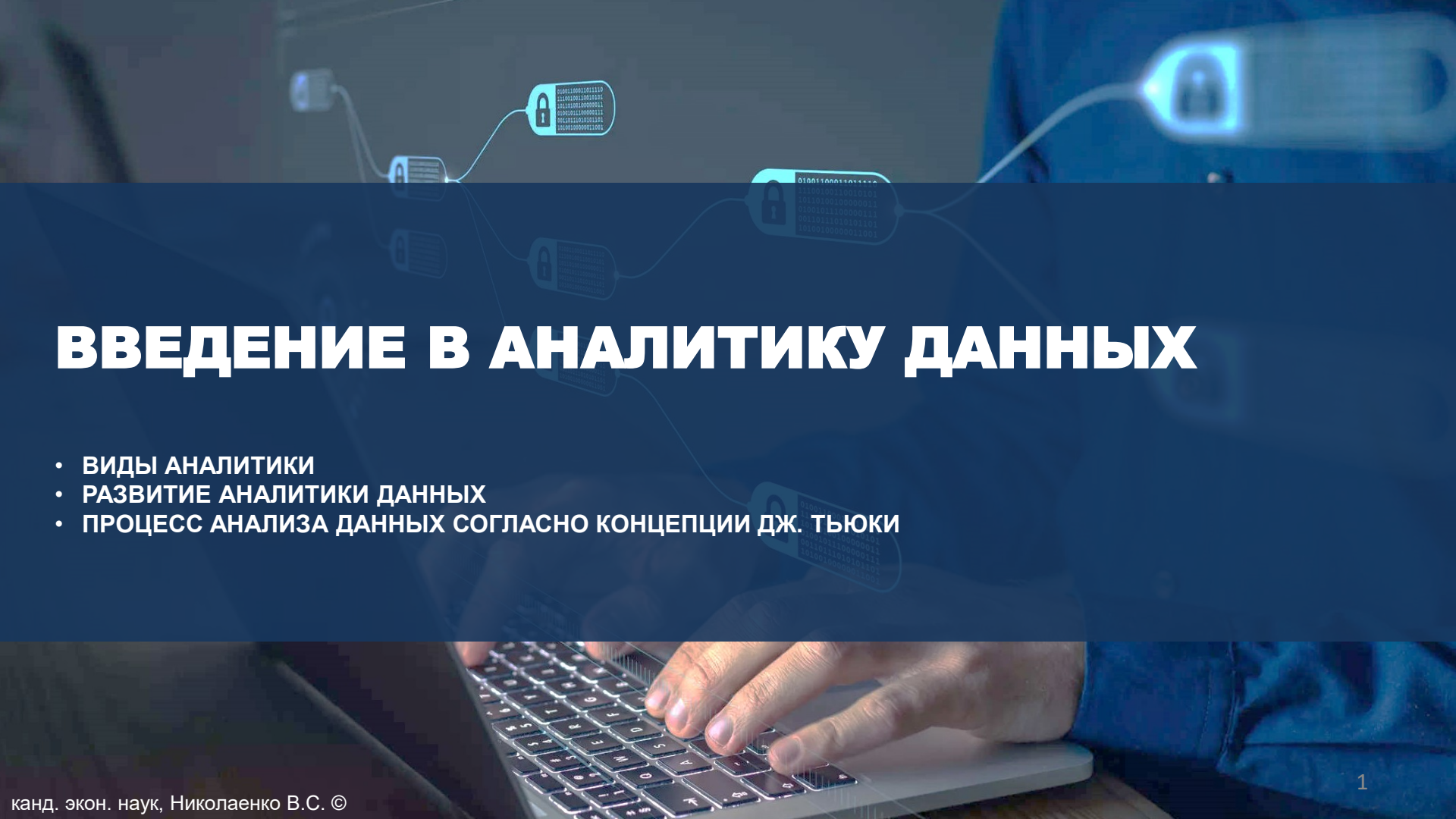


ВВЕДЕНИЕ В АНАЛИТИКУ ДАННЫХ

- ВИДЫ АНАЛИТИКИ
- РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ
- ПРОЦЕСС АНАЛИЗА ДАННЫХ СОГЛАСНО КОНЦЕПЦИИ ДЖ. ТЬЮКИ

канд. экон. наук, Николаенко В.С. ©

ВИДЫ АНАЛИТИКИ

В 2011 году вышел фильм **«Moneyball»** с Брэдом Питтом в главной роли. В российском прокате фильм известен под названием **«Человек, который изменил все»**.

Это история о спортивном менеджере, который пытается добиться больших высот с заурядной бейсбольной командой. **Билл Бин** (Б.Питт) разочаровывается в классическом подходе к подбору игроков – на основе наблюдения скаутов и их экспертных оценок. При обсуждении кандидатур Бин слышит от коллег аргументы в духе: *«Он не нравится моей жене»* – и прочие **непрофессиональные суждения**.

Однажды в офисе у знакомого Бин встречает молодого специалиста по **статистике** и **анализу данных**. Вместе они начинают анализировать большие объемы информации о технических действиях сотен игроков. По результатам статистического анализа Бин собирает команду из спортсменов, которые хороши только в одном или двух игровых действиях: **броске**, **приеме**, **перемещениях между базами** и др. Как менеджер команды Билли Бин использует игроков лишь в наиболее выгодных для них ситуациях. Таким образом, он становится первым кто отказывается от субъективного поиска игроков. Взяв в союзники **статистику** и **аналитику**, Бин за меньшие деньги формирует бейсбольный коллектив. Впоследствии такой подход перенимают и другие команды.

ВИДЫ АНАЛИТИКИ

Фильм **«Moneyball»** является ярким примером который показывает суть аналитики – **помочь в принятии решений на основе данных**. Данные используемые аналитиком в работе, могут представлять собой все что угодно: **статистические таблицы**, **карты бизнес-процессов**, **техническую документацию** и др.

Соответственно, задача аналитика состоит в том, чтобы собрать необходимые данные, проанализировать их и представить руководителю для принятия взвешенного и обоснованного управленческого решения.

ВИДЫ АНАЛИТИКИ

Анализ – это логический прием, с помощью которого субъект мысленно разбивает предметы и явления, выделяя их отдельные части и свойства. Обратный анализу логический прием называется **синтез**.

На основании данных определений можно заключить, что **аналитик стремится разложить на составные части сложный предмет или явление, чтобы понять как они устроены**. Например, если аналитик работает в организации, занимающейся разработкой программ для ЭВМ или автоматизацией деятельности, то объектом его анализа будут бизнес-процессы заказчика.

ВИДЫ АНАЛИТИКИ

Рассмотрим наиболее распространенные виды аналитики.

- Бизнес-анализ
- Системный анализ
- Продуктовая аналитика
- Финансовый анализ
- Аналитика данных (Business Intelligence, BI)

ВИДЫ АНАЛИТИКИ: БИЗНЕС-АНАЛИЗ

Согласно **Business Analysis Body of Knowledge (BABOK)** под **бизнес-анализом** понимается деятельность, которая позволяет внедрять изменения в организацию путем определения потребностей заинтересованных сторон.

С позиции бизнес-анализа у организации есть два состояния: **текущее состояние** и **целевое состояние**. **Текущее состояние (AS IS)** – это положение дел, которое по тем или иным причинам не устраивает руководство организации. Проблемы могут касаться длительности выполнения процессов, высоких заработных плат, неясных зон ответственности и др. **Целевое состояние (TO BE)** – это состояние, при котором организация проанализировала свои слабые места и оптимизировала бизнес-процессы: повысила их прозрачность и эффективность.

Роль бизнес-аналитика заключается **в оказании помощи идентификации проблемных зон организации** в ее текущем состоянии и предложении вариантов перехода в целевое состояние.

ВИДЫ АНАЛИТИКИ: СИСТЕМНЫЙ АНАЛИЗ

Если бизнес-аналитик отвечает на вопрос **«что нужно сделать?»**, то системный аналитик определяет, **«как сделать то, что нужно заказчику»**. Задача системного аналитика – предложить заказчику варианты реализации требований, полученных от бизнес-аналитика.

Рассмотрим работу системного аналитика в сфере ИТ (06.022 https://clck.ru/PaFVa).

Системный аналитик, как правило, является экспертом, который разбирается в работе конкретного программного продукта или информационной системы. Он знает основные технологические процессы системы, принципы хранения информации в ней, нюансы интеграции с другими системами и др. Системный аналитик формирует **техническое задание (ТЗ)**, в котором на более детальном техническом уровне, нежели бизнес-аналитик, описывает требования к будущей программе для ЭВМ. Если бизнес-аналитик говорит о том, что должно произойти при нажатии на определенную кнопку, то системный аналитик фиксирует, за счет чего достигается необходимый результат: как **происходит обращение к серверу, как обрабатывает ответ, как хранится информация в базе данных** и др.

ВИДЫ АНАЛИТИКИ: ПРОДУКТОВАЯ АНАЛИТИКА

Продуктовая аналитика посвящена разработке и развитию какого-либо продукта.

При работе с программными продуктами главным инструментом продуктового аналитика становится **A/B-тестирование**. **A/B-тестирование** – это маркетинговый метод, использующийся для оценки и управления эффективностью веб-страницы. Суть метода заключается в том, что создается **веб-страница A**, копируется и в копии (**страница B**) меняется какой-либо параметр, например, заголовок. Затем половине пользователей показывается страница A, а другой половине – страница B и оценивается какая из двух страниц имеет большую **конверсию**.

Рассмотрим работу A/B-тестирования на примере. Допустим у заказчика есть интернет-магазин и он хочет добавить к нему пару новых функций – красивую кнопку оформления заказа и 3D-обзор товара. Для удобства назовем их «Функция 1» и «Функция 2». Текущее состояние интернет-магазина без новых функций будет служить контрольным показателем. Подрядчик создает несколько версий интернет-магазина: **только с «Функция 1»**, **только с «Функция 2»** и **обеими «Функциями» сразу**. Для того, чтобы понять как новые изменения и их комбинации повлияли на работу магазина по сравнению с контрольным показателем подрядчик демонстрирует новые версии интернет-магазина фокус-группе пользователей. Во время работы с фокус-группой подрядчик устанавливает, что смена цвета кнопки с синего на красный приводит к оттоку пользователей, а смена на зеленый цвет повышает **кликабельность**.

ВИДЫ АНАЛИТИКИ: ФИНАНСОВЫЙ АНАЛИЗ

Финансовый анализ связан с изучением **финансово-хозяйственной деятельности организации**. Финансовый аналитик постоянно сталкивается в своей работе с:

- планами развития организации;
- разработкой и контролем ключевых показателей эффективности деятельности;
- управлением запасами;
- ассортиментной и ценовой политикой;
- подготовкой отчетов для контролирующих органов;
- формированием бюджетов организации и др.

ВИДЫ АНАЛИТИКИ: АНАЛИТИКА ДАННЫХ (Business Intelligence, BI)

Аналитик данных решает следующие задачи:

- Построение прозрачных моделей на основе больших массивов данных. Например, прогнозов оттока клиентов на основе анализа данных об их активности с момента появления в клиентской базе.
- Разработка механизмов персональных рекомендаций на основе анализа больших объемов данных.
- Выявление скрытых аномалий и закономерностей в данных.

Для решения подобных задач аналитику данных нужны глубокие знания в области **математики**, **статистики** и эксплуатации **аналитических информационных систем поддержки принятия решений** (**OnLine Analytical Processing, OLAP**).

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ

По одной из классификаций анализ данных проистекает из задач **прикладной математики**. Кроме анализа данных, в данной классификации выделяют еще два направления: **вычислительная математика** и **идентификация моделей**. Так как исторически они возникли первыми, их еще называют классическими подходами прикладной математики.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: ВЫЧИСЛИТЕЛЬНАЯ МАТЕМАТИКА

Вычислительная математика решает задачу вычисления одних характеристик изучаемого объекта или явления по известным значениям других его характеристик. При этом модель объекта считается известной, а зависимости между характеристиками описываются аналитическим выражением в виде уравнения или системы уравнений. Проблемы, возникающие при решении таких задач, связаны, в основном, с большими объемами вычислений и с защитой от погрешностей, накапливающихся из-за округления чисел.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: ИДЕНТИФИКАЦИЯ МОДЕЛИ

Задача **идентификации модели** формулируется по-другому. Известен набор переменных, влияющих на целевую характеристику, известен также общий вид зависимости между характеристиками, но коэффициенты, показатели степени и другие параметры модели неизвестны, и, чтобы их определить, используются протоколы наблюдений, отражающие значения одних характеристик при разных значениях других.

Делается серия предположений о значениях неизвестных параметров модели, и эти предположения проверяются на протоколах. В результате выбираются такие значения параметров, при которых модель с заданной точностью позволяет по одним (входным) характеристикам определять другие (выходные или целевые) характеристики. К таким задачам принято относить дедуктивные процедуры математической статистики: **корреляционный** и **регрессионный анализы**, **факторный анализ**, **численные методы оптимизации** и др.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: «МОДЕЛЬ» И «МОДЕЛИРОВАНИЕ»

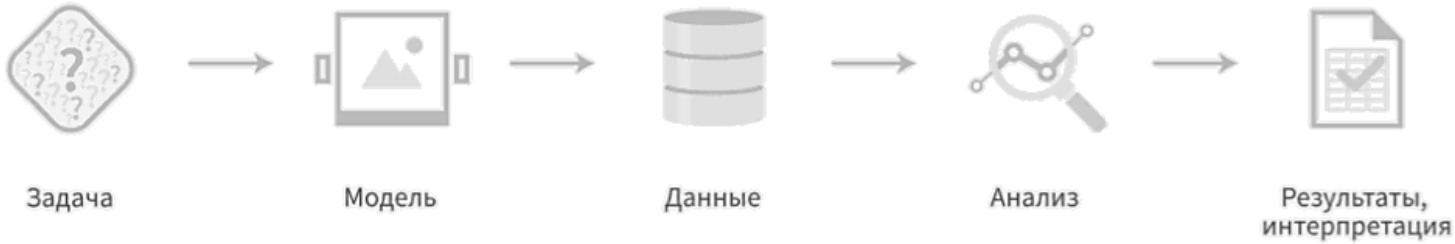
Рассмотрим термины **«модель»** и **«моделирование»** подробнее.

Слово **модель** *(лат.: modelium)* означает «меру», «способ», «сходство с какой-то вещью». **Построение моделей** – универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других задач.

Модель – это объект или описание объекта, системы для замещения (при определенных условиях, предположениях, гипотезах) одной системы (то есть оригинала) другой системой для лучшего изучения оригинала или воспроизведения каких-либо его свойств.

Моделирование – универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности. В прикладной математике имеют дело, как правило, с математическими моделями. Основная цель моделирования в том, что модель должна достаточно хорошо отображать функционирование моделируемой системы.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: ПРОЦЕСС АНАЛИЗА В ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ



На рисунке изображен **процесс анализа в вычислительной математике**, так называемый **классический подход**. Для решения задач в данном подходе выбираются готовые математические модели с известными параметрами. Модели проверяются на основе имеющихся данных. Расчеты осуществляются с помощью специализированного программного обеспечения.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: ПРОЦЕСС АНАЛИЗА В ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

Сразу видны минусы подхода. **Риск сделать ошибку в выборе модели и ее параметров остается вне поля внимания.**

Конечно же, исследователь, решая современные задачи анализа данных, практически никогда не знает ни точный вид «истинной» модели, ни характер связей между переменными, ни исчерпывающий перечень самих переменных. Здесь становится актуально высказывание известного статистика Дж. Бокса *«…все модели неправильные, но некоторые из них полезные»*.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

На практике часто возникают задачи, для решения которых нет готовых математических моделей, поэтому в середине XX века разрабатываются новые подходы к анализу данных. В частности, в 1962 году **Джон Тьюки** (John Tukey) предложил концепцию **разведочного анализа данных** (англ.: Exploratory Data Analysis, **EDA**). Дж. Тьюки был убежден, что можно многое узнать из данных, просто **визуализируя** их. Этот первичный этап анализа он и назвал **разведочным**, а важнейшим его элементом определил **широкое использование визуального представления многомерных данных**. Для этого данные представляются в виде **графиков**, **схем**, **условных рисунков**, **таблиц**, особенностью которых является **наглядность** – возможность увидеть признаки каких-либо закономерностей.

Тьюки, Джон
(1915-2000)

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

Всего в своей концепции Дж. Тьюки выделил **три** этапа анализа данных:

- **разведочный**;
- **подтверждающий** (конфирматорный);
- **итоговый**.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

Разведочный анализ – это синтез стохастических и эвристических подходов к анализу выборочных наблюдений. На этом этапе **цель исследователя** – выявить внутренние вероятностные и геометрические закономерности в данных для формирования и верификации тех или иных рабочих гипотез о связях между переменными, когда отсутствуют априорные представления о природе этих связей.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

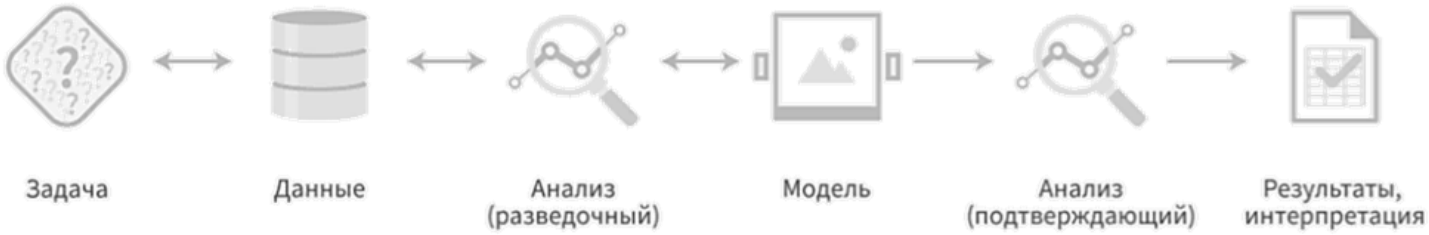
На следующем этапе **подтверждающего анализа** ставится задача проверки соответствия сформулированных гипотез полученным эмпирическим данным. Вычисляются итоговые статистические оценки моделей и определяются их погрешности.

РАЗВИТИЕ АНАЛИТИКИ ДАННЫХ: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

На **итоговом анализе** данных проводится экспертный анализ результатов и их обобщение. В случае необходимости, на всех этапах исследования возможны итерационные уточнения и обобщения.

Можно смело утверждать, что концепция разведочного анализа Дж. Тьюки воплотилась позже в таких подходах аналитики данных как **обнаружение знаний в базах данных (KDD), Data Mining , Big Date** и др.

ПРОЦЕСС АНАЛИЗА ДАННЫХ СОГЛАСНО КОНЦЕПЦИИ ДЖ. ТЬЮКИ



Что касается самого процесса анализа данных, то по Дж. Тьюки он выглядит по-другому: вместо последовательности действий **Модель-Анализ-Данные** идет **Данные-Анализ (разведочный)-Модель-Анализ (подтверждающий)**.

Отправной точной в процессе анализа согласно концепции Дж. Тьюки служат данные, характеризующие исследуемый объект или явление. Модель «следует» за данными, а не наоборот, как в классическом подходе. Двунаправленные стрелки показывают, что анализ данных носит циклический характер: **на первых итерациях выдвинутые гипотезы могут потребовать дополнительных экспериментальных данных или наблюдений, уточнений**. Это существенно облегчает подбор способов более глубокой обработки данных на этапах подтверждающего анализа.

ПРОЦЕСС АНАЛИЗА ДАННЫХ СОГЛАСНО КОНЦЕПЦИ ДЖ. ТЬЮКИ

Однако концепция «моделей от данных» требует тщательного подхода к качеству самих исходных данных, поскольку ошибочные, зашумленные данные могут привести к моделям и выводам, не имеющим никакого отношения к действительности. Поэтому в анализе данных важную роль играют **интеграция, подготовка** и **очистка данных.**

ПРОЦЕСС АНАЛИЗА ДАННЫХ СОГЛАСНО КОНЦЕПЦИИ ДЖ. ТЬЮКИ

Благодаря концепции Дж. Тьюки был создан современный анализ данных, где решается задача анализа явлений, для которых еще нет математических моделей. Есть только наборы экспериментальных данных входы-выходы и даже только входы, представленные в виде массивов или таблиц.