# Вступление

**1.** Если говорить о сфере действия больших информационных систем, к которым относится и Яндекс, то видно, что они достигли такого высокого уровня, который требует применения точных методов мышления и, прежде всего, общематематических методов, в частности, вероятностно-статистических.

Мы хорошо знаем, что зарождение собственно *теории вероятностей* было связано с желанием получить ответ на разного рода вопросы, естественно возникающие в связи с интересом к азартным играм. (Например, известна задача Галилея: бросаются три шестигранные кости с нанесенными на них цифрами 1, 2, …, 6; спрашивается, какова вероятность того, что сумма выпавших очков будет равна 10. Другими известными задачами являются задачи, возникшие в ходе переписки Паскаля и Ферма.) Мы также знаем, что зарождение *статистического мышления* и *статистических методов* было связано с необходимостью разрешения вопросов геодезии и астрономии. (Так, согласно первоначальному определению *меры длины*, принятому во Франции, *метр* — это есть одна десятимиллионная часть ($1 \cdot 10^{-7}$) четверти длины парижского географического меридиана. Тем самым, по многочисленным измерениям надо было как можно точнее определить арифметическую величину этой четверти меридиана.) Именно в связи с этими геодезическими, а также астрономическими запросами (определение параметров орбит планет и комет) возник *метод наименьших квадратов* (Лежандр, Лаплас), начала строиться количественная технология обработки эмпирических данных, стали вырабатываться логика и методология измерений в условиях неопределенности, создаваться «исчисление наблюдаемых данных». (В связи со сказанным, было бы хорошо иметь в рамках *Школы по анализу данных* специальный курс относительно таких базисных идей математической статистики, как метод наименьших квадратов, метод максимального правдоподобия, байесовский метод, непараметрическая статистика, … Для наилуч-

шего понимания всего этого полезно было бы вести изложение этого предмета в его *историческом* развитии.)

**2.** Настоящий курс посвящен изложению некоторых современных методов *теории принятия решений* в условиях неопределенности, нацеленных на решение *конкретных задач общего интереса*[1], возникающих при динамическом анализе (в режиме реального времени) статистических данных, получаемых, например, в финансовой инженерии, в теории обнаружения сигналов на фоне помех, … Для большинства информационных систем весьма актуальна разработка методов успешного обнаружения нежелательных внедрений в информационные системы («network intrusions») и методов создания систем защиты от кибер-атак («cyber-terrorism»).

**3.** В литературе описаны разнообразные методы обнаружения нежелательных «внедрений», основанные на технике «искусственного интеллекта», включая экспертные системы, нейронные сети, «pattern matching», и др. [4]. Существующие системы обнаружения «внедрений» (IDS — Intrusion Detection Systems) обычно классифицируются или как Signature Detection Systems, или как Anomaly Detection Systems [4], [5].

Signature Detection Systems обнаруживают атаки путем сравнения наблюдаемых шаблонов (pattern) сетевого трафика с известными образцами (signature) атак, хранящимися в базе данных [6].

Мы уделяем значительное время изложению второго метода обнаружения, основанного на Anomaly Detection Systems. Обычно, внедрение в сети (например, Denial-of-Service [DOS] attacks, Address Resolution Protocol Men-in-the-Middle [ARP MiM] attacks, …) происходит в неизвестный заранее момент времени $\theta$ и сопровождается *изменением вероятностно-статистических свойств некоторых характеристик наблюдаемого процесса* (например, количества отправленных и принятых сервером пакетов). Поэтому естественно возникает идея математически сформулировать задачу обнаружения атаки как *задачу* («*$\theta$-задачу*») *скорейшего обнаружения момента* ($\theta$) *появления разладки* в ходе наблюдаемого процесса. Наша цель будет состоять в том, чтобы, начиная с простых моделей

[1] Есть несколько, уже давно изданных, классических книг, посвященных теории принятия решений. Отметим в первую очередь монографии [1], [2].

и затем переходя к более сложным, изложить те методы скорейшего обнаружения, которым уделялось и поныне уделяется большое внимание.

Наряду с моментом $\theta$ появления разладки важной характеристикой рассматриваемых систем будет *момент подачи тревоги*, который мы обозначаем через $\tau$. Этот момент должен строиться по прошлым данным, т. е. быть моментом «без упреждения». Такие моменты называются моментами остановки или марковскими моментами (точное определение будет приведено ниже). Стремиться мы будем к тому, чтобы минимизировать (в некотором усредненном смысле) время запаздывания $\tau - \theta$ в обнаружении момента $\theta$ (когда $\tau \geqslant \theta$) при соблюдении условия, что ложное обнаружение (когда $\tau < \theta$) имеет малую вероятность.

Основным аппаратом решения таких задач является «последовательный анализ принятия решений». Такие задачи, как мы увидим, удобно формулировать как задачи об оптимальной остановке. Первые задачи такого типа были рассмотрены еще в сороковых годах А. Вальдом в рамках теории последовательного различения двух (а также многих) статистических гипотез. К настоящему времени теория оптимальных правил остановки получила значительное развитие. Основной литературой для нас будут книги [7], [8].

**4.** Мы рекомендуем читателям хотя бы бегло ознакомиться с книгами [1], [2], являвшимися, в сущности, первыми переводами на русский язык книг по вопросам принятия решений, применениям к теории игр, экономике, исследованию операций. Математические методы во многом были основаны на новых тогда методах линейного программирования. Возникшие затем методы динамического программирования (Р. Беллман), истоками которого были работы по последовательному анализу А. Вальда, Д. Блекуэлла, М. А. Гиршика и др., дали возможность исследовать динамические постановки задач типа задач оптимального управления.

Рассматриваемые нами методы решения задач скорейшего обнаружения в значительной мере опираются на современный аппарат теории случайных процессов, стохастического исчисления, теории мартингалов, нелинейной фильтрации и т. д. Хотелось бы подчеркнуть, что многие эти теоретические методы были выработаны именно на пути решения задач скорейшего обнаружения. Это слу-

жит хорошей иллюстрацией того, как происходит развитие теории, когда она нацелена на решение конкретных задач, имеющих практический интерес.

В небольшой статье [9] Клод Шеннон, говоря о теории информации, подчеркивает, что хотя она и «является сильнейшим средством решения проблем теории связи (и в этом отношении ее значение будет возрастать), нельзя забывать, что она не является панацеей для инженера-связиста и *тем более* для представителей всех других специальностей. Очень редко удается открыть одновременно несколько тайн природы одним и тем же ключом».

Мы приводим эти слова с тем, чтобы подчеркнуть, что излагаемые нами постановки задач и известные методы их решения должны побуждать читателей на формулирование *новых* задач, в том числе непосредственно интересных для Яндекса, для решения которых будут найдены *новые подходы*, стимулирующие развитие и собственно теоретических исследований.

Автор приносит свою благодарность руководству Яндекса и профессору И. Б. Мучнику за приглашение прочитать курс лекций, многочисленные советы и поддержку. Большая помощь при подготовке рукописи к печати была оказана автору Е. В. Бурнаевым.