智能化软件系统与工程

AI系统需求工程

马 郓

人工智能研究院

北京大学 PEKING UNIVERSITY

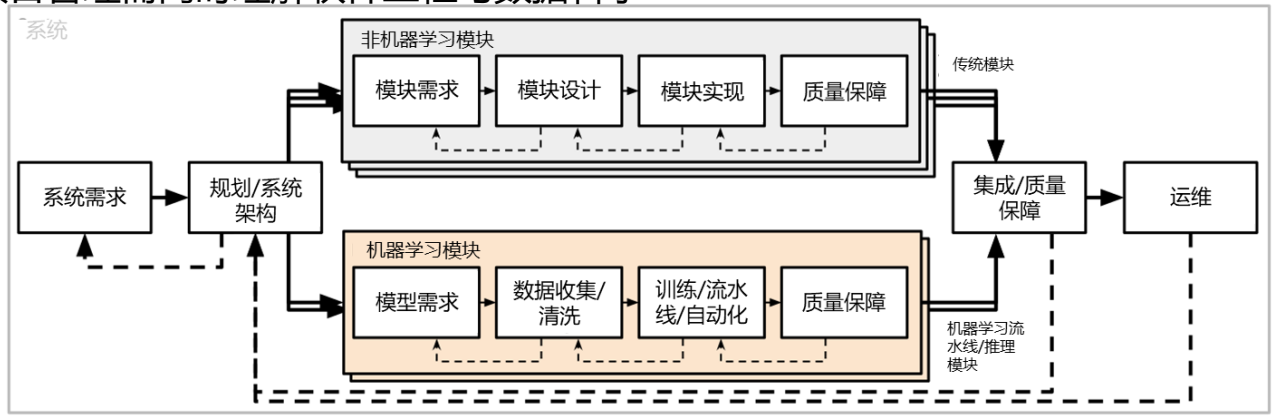
上讲回顾：一种可能的智能化软件开发过程

北京大学 PEKING UNIVERSITY

融合软件工程与数据科学的开发过程

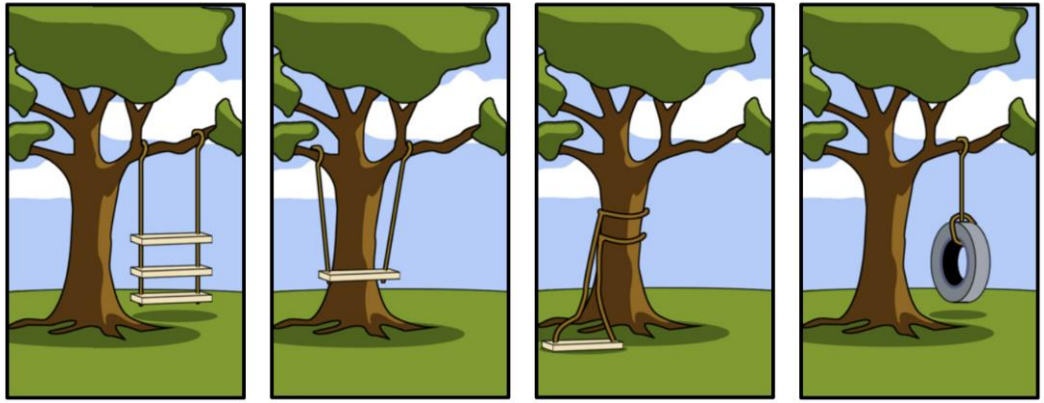
- 先确立系统级需求（用户、安全、公平...）
- 系统需求反向指导数据科学建模（隐私、公平...）
- 风险优先：先开发高风险部分（通常是机器学习模块；螺旋模型）
- 增量开发原型，获得用户反馈（敏捷开发）
- 支持持续迭代改进
- 按AI模块的特性设计系统（UI、安全防护）
- 在开发和运行的全过程都规划测试
- 项目管理需同时理解软件工程与数据科学

尚无“最佳实践”或过程模型



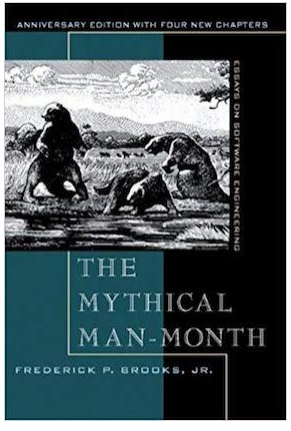
软件需求


客户需求VS实际
代码
程序员
产品经理
客户需求
扎心了


How the customer explained it
How the project leader understood it
How the programmer wrote it
What the customer really needed

软件需求

需求的重要性

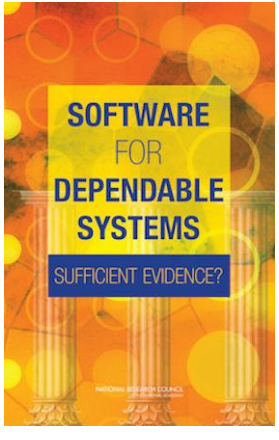


"构建软件系统最困难的部分是精确地决定要构建什么……如果做得不好，工作中的其他任何部分都不会如此严重地损害最终系统。"

——Fred Brooks，Mythical Man Month (1975)

代码中的错误仅占致命软件事故的3%，大多数故障是由于对需求理解不足或可用性问题造成的

——美国国家研究委员会对软件相关故障的调查 (2007)



软件需求

北京大学 PEKING UNIVERSITY

软件需求是指在软件开发过程中，为了满足特定用户或利益相关者的期望和目标，软件系统必须具备的功能和质量特性

- 功能需求
  - 系统提供的具体功能
  - 输入和输出之间的关系
- 质量需求（非功能需求）
  - 响应时间、吞吐量
  - 开发时间和成本投入
  - 代码质量，可维护性
  - 用户友好性
  - 安全、隐私、公平性
  - ……

例如：

① 系统必须有能力支持100个以上的并发用户，每个用户可以处理附录A中操作任务的任选组合，平均响应时间应该小于1秒，最大响应时间应小于5秒。

其中：功能-可以处理附录A中操作任务的任选组合

质量-有能力支持100个以上的并发用户

平均响应时间应小于1秒，最大响应时间应小于5秒

② 必须在对话窗口的中间显示错误警告，其中使用红色的、字号为14的加粗Arial字体。

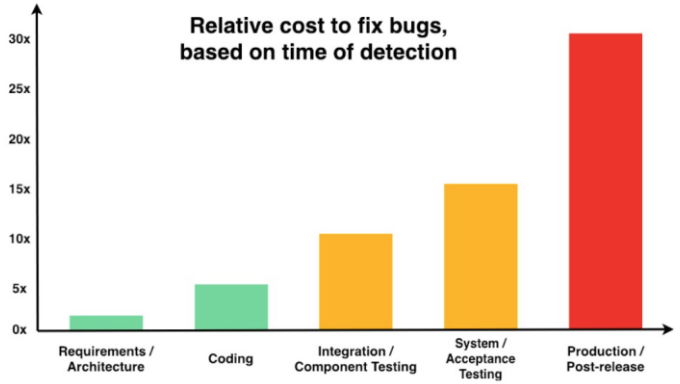
其中：功能-能显示错误警告

质量-在对话窗口的中间显示

使用红色的、字号为14的加粗Arial字体

软件需求

- **开发者倾向于直接投入编码，而忽略了对问题本身的理解**
  - 如果没有清晰的需求，我们很可能在构建一个“无法解决正确问题”的系统
  - 需求阶段的错误修复成本最低，越往后修复成本呈指数级增长

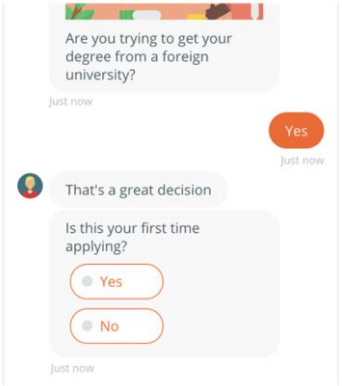


- **在机器学习项目中，对用户需求、交互方式和潜在风险的预先规划尤为重要**

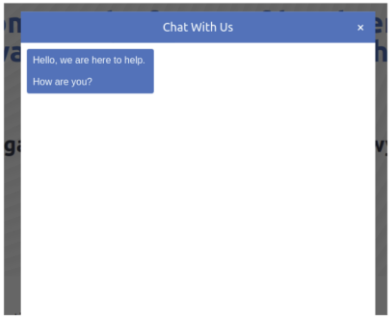
设定目标

案例分析：自助式法律聊天机器人

- **一家为律师和律师事务所提供营销服务的企业开发了一款聊天机器人**
  - 通过聊天，可以帮助用户对接律师
  - 既有系统——引导式聊天
    - 局限：无法列举所有问题，难以匹配开放式表述
  - 做一个更好的聊天机器人
    - 自然语言聊天
    - 帮助用户处理简单任务
    - 在用户需要时为其对接律师
    - 更新外观设计；打造 “数字营销的未来” 形象



既有系统



新系统

Pass, S., & Sengers, P. (2020). Making data science systems work. Big Data & Society, 7 (2).

案例分析：自助式法律聊天机器人

北京大学 PEKING UNIVERSITY

数据科学家面临的挑战

- 基础设施：了解自然语言聊天机器人的基础设施及其功能
- 了解主题：确定用户谈论的内容，利用过往聊天记录训练 / 测试概念
- 引导对话：支持开放式对话需要检测相关主题并给出恰当回应；意图 - 主题建模
- 数据收集 / 标注极具挑战性 —— 特殊情况太多

存在“太多边缘情况”，导致对话无法按计划进行

Pass, S., & Sengers, P. (2020). Making data science systems work. Big Data & Society, 7 (2).

案例分析：自助式法律聊天机器人

北京大学 PEKING UNIVERSITY

与客户的内部会议

客户：*或许我们可以用80/20法则来考虑这个问题。在某些情况下，它表现不错，但在另一些情况下，就比较棘手了。80%的情况都没什么问题，剩下的20%，我们会尽力处理。*

数据科学负责人：*问题在于如何自动识别哪些属于80%的情况，哪些属于20%的情况。*

数据科学家：*这比听起来要难。其中一个模型是匹配模型，用法律问答对训练而成。有6万个问答对。看似数量不少，但对于机器学习来说还是太少了。*

客户：*这已经很多了。它能回答比如签证续签相关的问题吗？*

数据科学家：*如果训练数据中有类似问题，就能回答。但仅靠6万个数据，模型很容易过拟合，遇到数据之外的内容就会失效。*

客户：*我明白你的意思了。从学术角度来看，边缘情况很有意思，但对于企业而言，首要的是价值。我知道你们在努力解决一个有趣的问题。但我觉得，你们可能已经解决了足够多的问题，能够带来商业价值了。*

Pass, S., & Sengers, P. (2020). Making data science systems work. Big Data & Society, 7 (2).

案例分析：自助式法律聊天机器人

系统目标

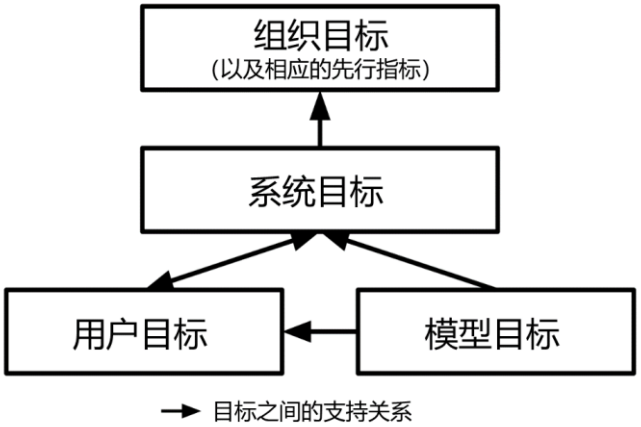
- 收集用户数据并出售给律师
- 向律师展示技术能力
- 可接受的失败情况：自助服务过于复杂时，为用户对接律师
- 解决边缘案例并不重要

“边缘案例固然重要，但最终目标是获取用户信息、将用户数据货币化。我们正在打造一个法律自助聊天机器人，但一个主要的商业用例是告诉人们：‘来，和这位律师聊聊。’我们确实希望为他们对接律师。即便是在那20%的情况下，当我们的机器人无法处理时，我们会告诉用户这个问题无法通过自助方式解决。那我们就为您联系一位律师，对吧？这从一开始就是我们的目标。”

Pass, S., & Sengers, P. (2020). Making data science systems work. Big Data & Society, 7 (2).

目标

- **目标：关于意图的规定性陈述**
- **目标的层次**
  - 组织目标：组织/企业/机构自身的整体目标
    - 先行指标：用于度量目标完成情况的短期量化指标
  - 系统目标：系统所产出的具体结果
  - 用户目标：用户使用系统所达成的目标
  - 模型目标：从模型视角来看系统中的模型质量
- **目标之间并不总是一致的**



模型目标 vs 系统目标

北京大学 PEKING UNIVERSITY

更精准的预测或许没那么重要

- “足够好”或许就已足够
  - 预测对系统成功而言是关键，还是仅仅是噱头?
- 更精准的预测可能要付出过高成本
  - 需要多得多的数据、长得多的训练时间
  - 隐私问题
- 优化用户界面（“体验”）或许能缓解诸多问题
  - 例如，向用户解释决策依据

模型目标 vs 系统目标

- **发表在ICML2012的论文就已经对过度关注算法改进和基准测试表达了担忧**
  - 专注于基准数据集，却不深入研究实际问题，例如鸢尾花分类、数字识别等
  - 注重抽象指标，而非衡量现实世界影响，如准确率、ROC曲线
  - 与现实世界的关切脱节
  - 缺乏后续行动，没有部署，也没有产生影响
- **论文中的研究成果难以复现和投入实际应用是常态**
- **忽视在数据收集方式、待解决问题的选择、人机智能界面设计、影响评估等方面的设计决策……**
- **文章主张：应当专注于产生实际影响——这需要构建实际的系统**

Machine Learning that Matters

Kiri L. Wagstaff KIRI.L.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

Abstract

Much of current machine learning (ML) research has lost its connection to problems of import to the larger world of science and society. From this perspective, there exist glaring limitations in the data sets we investigate, the metrics we employ for evaluation, and the degree to which results are communicated back to their originating domains. What changes are needed to how we conduct research to increase the impact that ML has? We present six Impact Challenges to explicitly focus the field's energy and attention, and we discuss existing obstacles that must be addressed. We aim to inspire ongoing discussion and focus on ML that matters.

1. Introduction

At one time or another, we all encounter a friend, spouse, parent, child, or concerned citizen who, upon learning that we work in machine learning, wonders "What's it good for?" The question may be phrased more subtly or elegantly, but no matter its form, it gets at the motivational underpinnings of the work that we do. Why do we invest years of our professional lives in machine learning research? What difference does it make, to ourselves and to the world at large?

Much of machine learning (ML) research is inspired by weighty problems from biology, medicine, finance, astronomy, etc. The growing area of computational sustainability (Gomes, 2009) seeks to connect ML advances to real-world challenges in the environment, economy, and society. The CALO (Cognitive Assistant that Learns and Organizes) project aimed to integrate learning and reasoning into a desktop assistant, potentially impacting everyone who uses a computer (SRI International, 2003–2009). Machine learning has effectively solved spam email detection (Zdziarski, 2005) and machine translation (Koehn et al., 2003), two problems of global import. And so on.

And yet we still observe a proliferation of published ML papers that evaluate new algorithms on a handful of isolated benchmark data sets. Their "real world" experiments may operate on data that originated in the real world, but the results are rarely communicated back to the origin. Quantitative improvements in performance are rarely accompanied by an assessment of whether those gains matter to the world outside of machine learning research.

This phenomenon occurs because there is no widespread emphasis, in the training of graduate student researchers or in the review process for submitted papers, on connecting ML advances back to the larger world. Even the rich assortment of applications-driven ML research often fails to take the final step to translate results into impact.

Many machine learning problems are phrased in terms of an objective function to be optimized. It is time for us to ask a question of larger scope: what is the field's objective function? Do we seek to maximize performance on isolated data sets? Or can we characterize progress in a more meaningful way that measures the concrete impact of machine learning innovations?

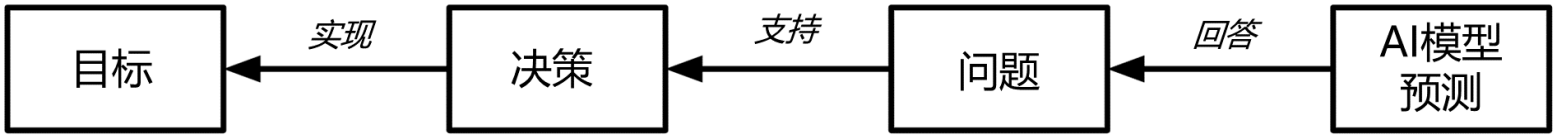
This short position paper argues for a change in how we view the relationship between machine learning and science (and the rest of society). This paper does not contain any algorithms, theorems, experiments, or results. Instead it seeks to stimulate creative thought and research into a large but relatively unaddressed issue that underlies much of the machine learning field. The contributions of this work are 1) the clear identification and description of a fundamental problem: the frequent lack of connection between machine learning research and the larger world of scientific inquiry and humanity, 2) suggested first steps towards addressing this gap, 3) the issuance of relevant Impact Challenges to the machine learning community, and 4) the identification of several key obstacles to machine learning

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.
Copyright 2012 California Institute of Technology.

Wagstaff, Kiri. Machine learning that matters. ICML 2012.

从目标到需求

- **设定项目的目标通常是获取系统需求的第一步**
- **目标有助于帮助组织识别使用AI的机会**
  - 探索组织内部人员为实现组织目标而经常做出的决策
  - 思考哪些决策可以通过预测来得到信息支持
  - 哪些预测可以通过AI来实现



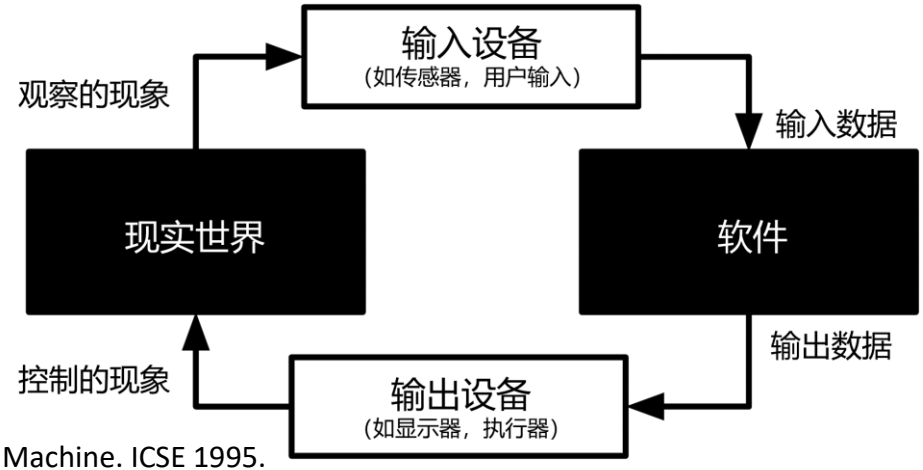
作为 [xx角色]，我需要做出 [xx决策] 来实现 [xx目标]

作为 [xx角色]，我需要知道 [xx问题] 的信息才能做出 [xx决策]

收集需求

需求的本质：“世界与机器”模型

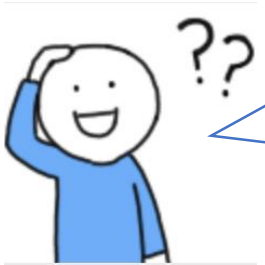
- **软件的最终目标是影响现实世界（The Real World）**
  - 软件并非在“真空”中，每个系统都是世界（环境）的一部分
  - **需求描述的是现实世界（环境）的期望状态**
  - 软件通过输入设备感知世界（环境）的状态，并通过输出设备将现实世界（环境）调控至期望状态



Michael Jackson. The World and the Machine. ICSE 1995.

讨论：基于智能手表的跌倒检测

- 开发一款在智能手表上运行的跌倒检测软件，根据手表的加速度计和陀螺仪检测手表佩戴者是否跌倒；当跌倒时通过连接的手机联系紧急救助服务



在这个案例中：

1. 环境的组成部分有哪些?
2. 系统在现实世界中的目标/需求是什么?
3. 什么情况会使得系统无法达到其目标?

需求、规约、假设、实现

- **系统需求 (System requirements, REQ)**
  - 通过对环境的期望效果描述系统必须确保什么
  - 系统需求捕捉的是现实世界中应该发生的事情，而不是软件应该如何处理数据
- **软件规约 (Specifications, SPEC)**
  - 也称作软件需求，通过输入/输出设备能够观测和操控的数据来描述软件必须实现什么
  - 仅指软件世界中的概念，例如输入和输出数据，而不指现实世界中的概念
- **假设 (Assumptions, ASM)**
  - 关于环境行为/属性的假设，弥合REQ和SPEC之间的差距
- **实现 (Implementations, IMPL)**
  - 提供软件系统的实际行为，该行为应与规约一致，通常以一些代码或可执行模型给出

系统目标能够被实现的条件：

$$ASM \wedge SPEC \models REQ$$

$$IMPL \models SPEC$$

案例分析：汉莎航空2904号班机跑道坠毁事故

北京大学 PEKING UNIVERSITY

- **汉莎航空2904号班机于1993年在华沙坠毁，原因是飞行员在着陆后未能及时启动反推力装置，导致飞机冲出跑道**
  - 反推力装置：着陆后使飞机减速
  - REQ：当且仅当飞机在地面上时，才启用反推力
  - SPEC：当且仅当车轮转动时，才启用反推力
    - 如果 (a) 每个起落架感应到6.3吨的重量 或(b) 传感器指示车轮转速超过72节
  - ASM：当且仅当飞机在地面上时，车轮才转动。
  - ASM：只有当飞机在地面上时，两个起落架上才有大的重量
- **事故当天，跑道因雨水湿滑**
  - 即使飞机在地面上，车轮也无法转动（假设被违反）；又因为风力，起落架承重不到6.3吨。
  - 飞行员尝试启用反推力，被软件阻止，最终飞机冲出跑道并坠毁！

审视AI系统的假设

北京大学 PEKING UNIVERSITY

- **不仅要审视关于系统输入和输出的假设，还要审视训练模型所用训练数据的假设**
  - 所使用的假设在世界变化时可能不再成立
- **假设不成立的情况**
  - 不现实或缺失的假设
  - 概念漂移
  - 攻击
  - 反馈循环

建立需求的过程

- 识别环境的组成和软件/机器学习模块
- 声明对环境的期望需求 (REQ)
- 识别环境与软件之间的接口
- 识别环境假设 (ASM)
- 建立满足系统需求（REQ）的规约 (SPEC)
- 检查是否 ASM $\wedge$ SPEC $\vdash$ REQ
- 如果不是，回到开始并重复

收集需求的挑战

北京大学 PEKING UNIVERSITY

用户侧问题

- 用户通常不知道自己真正想要什么，直到他们看到
- 用户描述模糊，且容易改变主意

开发者侧问题

- 工程师自认为已理解需求，用技术直觉填补空白
- 过分受限于技术可行性，而非用户真实需求
- 偏爱优雅、通用的抽象，而忽略现实世界的“脏活累活”

系统侧问题

- 容易忽略非功能性需求（如公平性、隐私）
- 只关注部分用户，而忽略了其他受影响的群体

收集需求的方法

第一步：识别利益相关者

- 利益相关者（Stakeholder）指所有对待开发系统有兴趣或可能受到影响的个人和实体
  - 开发项目的组织
  - 参与项目开发的所有人员
  - 项目的所有客户和用户
  - 间接受到正面或负面影响的人员
- 识别不同的利益相关者及其目标，有助于理解项目中的不同关注点，并就权衡利弊、目标和需求进行审议

收集需求的方法

北京大学 PEKING UNIVERSITY

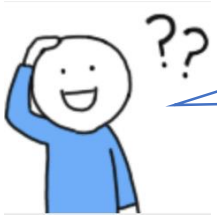
第二步：针对每个利益相关者，捕获其需求

- 背景研究：了解组织，阅读文档
- 访谈利益相关者
  - 提出关于问题、需求、可能的解决方案、偏好、担忧等方面的开放式问题
  - 使用问卷清单收集质量需求（可用性、隐私、延迟等）
- 调查、小组会议、研讨会： 与多个利益相关者同时互动，探讨冲突
- 用户志研究： 融入用户，被动观察或主动参与
- 建立需求分类法和清单：确保至少考虑了常见的需求
- 换位思考： 转换视角以探索未访谈的利益相关者的需求

收集需求的方法

第三步：需求协商

- 许多需求相互冲突/矛盾
- 不同的利益相关者目标不同，有不同的优先级、偏好和担忧
- 对相关需求进行分组，识别相互冲突的关注点和替代方案，并最终就优先事项和冲突解决方案做出决策
  - 卡片分类、亲和图、重要性-难度矩阵等



跌倒检测系统可能会有什么冲突需求?

收集需求的方法

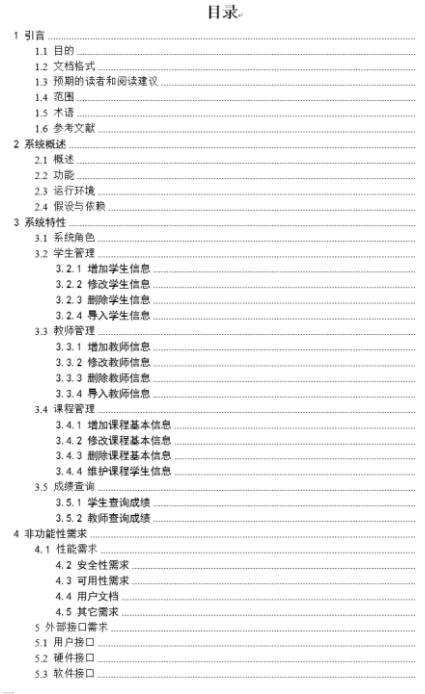
第四步：需求记录

- 将最终确定的需求、假设、以及决策理由写成文档
- 重量级方法：软件需求规格说明书
- 轻量级方法：用户故事、Wiki、IssueTracker
  - 用户故事：形式为“作为一个[用户类型]，我想要[一个功能]，以便[一个好处/价值]”的需求陈述

作为一名有健康顾虑的用户，我希望智能手表能够在检测到跌倒后30秒内无法做出反应时自动呼叫紧急援助，确保在我可能丧失行动能力的情况下派遣救援人员

作为一名喜欢散步的活跃老年人，我希望智能手表在检测到跌倒时将我的GPS位置发送给紧急联系人或服务，这样即使我无法沟通，也可以在我家外快速找到我并获得帮助。

作为不懂技术的用户，我希望智能手表的设置和配置过程简单直观，这样我就可以无需他人帮助即可完成。


目录

收集需求的方法

北京大学 PEKING UNIVERSITY

第五步：需求评审

- 人工检查（类似代码审查）
- 向利益相关者展示需求，询问是否有误解和遗漏
- 向利益相关者展示系统原型
- 建立清单以涵盖重要质量需求
- 批判性地检查假设的完整性和真实性
- 寻找不切实际的机器学习相关假设

关于需求工程的质疑

- **传统软件的开发，需求工程被广为诟病**
- **智能化软件的开发，需求工程越来越重要**
- **需求工程要做到什么程度？何时开展？**
  - 需求工程在构建高风险的系统时非常重要
  - 需求作为合同的基础（外包、分配责任）
  - 需求很少能完全预先确定且稳定，要预见变化
    - 利益相关者在原型中发现问题，改变主意
    - 机器学习需要大量的探索来确定可行性
  - 低风险的问题通常采用轻量级、敏捷的方法

分析与应对风险

回顾：软件中的“问题”

- Problem
- Mistake
- Defect
- Bug
- Fault
- Error
- Failure
- Exception
- Anomaly

核心理念：错误（Mistake）总会发生

北京大学 PEKING UNIVERSITY

- **将AI模型视为系统中的不可靠模块**
  - 它可能在未知的时间、因未知的原因而犯错
  - 模型失效的原因
    - 相关性 vs. 因果关系：模型可能通过背景（雪地）来识别动物（狼）
    - 混杂变量：误认为“秃顶”导致新冠重症，而忽略了真正的混杂因素“年龄”
    - 反向因果：看到酒店在高需求时提价，就错误地认为“提价可以创造需求”
    - 分布外数据：成年人横穿轨道的列车模型，可能无法识别出轨道上的儿童
  - 错误通常不是随机的
    - 物理系统中的故障通常可以建模为潜在随机过程，可进行概率推理
    - AI模型的错误通常是由特定输入触发，可复现、可被利用
- **设计一个即使AI模型犯错、整体也不会出现严重问题的系统**

容错设计

- **既然错误不可避免，我们就要在系统层面进行容错设计**
- **五大核心策略**
  - 人在回路（Human-in-the-Loop）
  - 可撤销的操作（Undoable Actions）
  - 安全护栏（Guardrails）
  - 错误检测与恢复（Detection & Recovery）
  - 遏制与隔离（Containment & Isolation）

容错设计：人在回路

- **将人类的判断力引入系统，作为监督者和最终决策者**
- **三种常见交互模式**
  - 自动化：系统自动执行，人类不参与
    - 适用于低风险或高置信度场景
    - 示例：垃圾邮件自动过滤
  - 提示：系统建议，用户确认后执行
    - 适用于高风险或低置信度场景，但需注意“通知疲劳”
    - 示例：欺诈交易确认提示
  - 组织/标记/增强：系统提供信息，辅助决策
    - 侵入性小，决策权完全在用户
    - 示例：Gmail的回复建议，搜索引擎结果排序

容错设计：可撤销的操作

- **如果一个操作可以被轻易逆转，那么错误预测带来的危害就是暂时和可控的**
- **实现方式**
  - 直接撤销：提供“撤销”功能
  - 流程可逆：设计明确的申诉或回退流程
    - 示例: 被错误封禁的账号可以申诉； 个性化服装推荐服务提供免费退货
- **局限性**
  - 并非所有操作都可逆，尤其是在物理世界中

容错设计：安全护栏

- **在模型的预测输出到最终执行之间，增加额外的检查和限制层**
- **护栏的形式**
  - 硬编码规则：限制预测值的范围，或使用禁用词列表过滤不当内容
  - 硬件保护：物理设备防止危险发生
    - 示例：自动驾驶列车的专用轨道 、车站的站台屏蔽门 、车门的压力传感器
  - 模型护栏：使用另一个独立的、更简单的模型来校验主模型的输出
    - 示例：使用检测模型过滤文本

容错设计：错误检测与恢复

- **设计独立机制来检测问题，并启动恢复预案**
- **执行者-检查者模式**
  - 执行者：复杂的AI模型，负责执行任务（如控制列车速度）
  - 检查者：独立的、更简单的监控模块，通过间接信号评估执行者的行为是否安全
  - 示例：检查者通过陀螺仪传感器发现列车转弯倾斜过大，即可判断执行者（速度控制模型）指令有误，并强制刹车
- **优雅降级**
  - 当检测到故障时，系统切换到功能受限但安全的操作模式
  - 示例：列车激光雷达故障后，降速并仅依靠视觉系统行驶

容错设计：遏制与隔离

- **确保低关键性模块的故障，不会传播并影响高关键性模块**
- **物理/网络隔离**
  - 示例：飞机的飞行控制系统和乘客的娱乐系统必须在物理上和网络上完全隔离
- **数据流隔离**
  - 谨慎考虑系统各部分对模型预测的依赖关系
  - 防止错误的预测数据通过数据流污染下游模块
  - 反例：因数据库输入错误导致整艘军舰的控制系统瘫痪

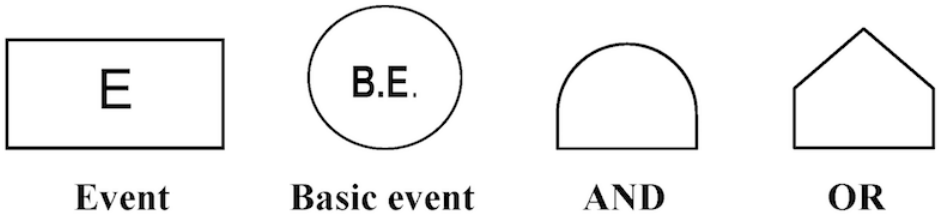
风险分析

北京大学 PEKING UNIVERSITY

- **预见错误及其后果，指导缓解措施的设计**
  - 风险 = 可能性 * 影响
- **三种经典方法**
  - 故障树分析（Fault Tree Analysis，FTA）
  - 失效模式与影响分析（FMEA）
  - 危险与可操作性研究（HAZOP）

风险分析：故障树分析

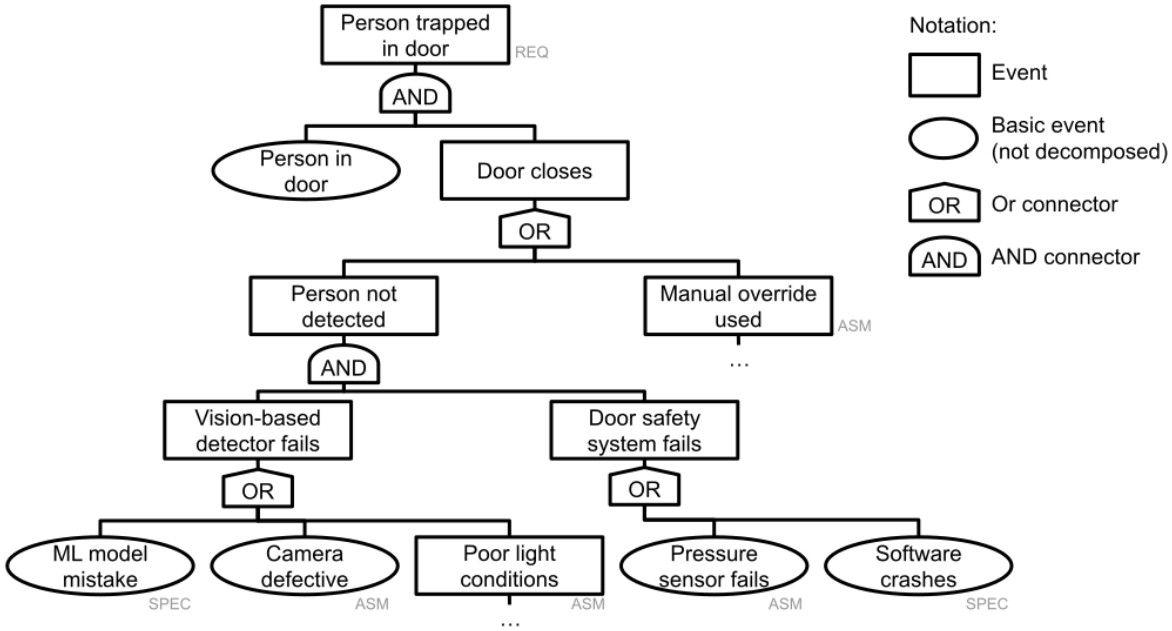
- **一种自上而下的演绎分析法，从一个不希望的“顶层事件”（系统故障）开始，逐层向下分析其所有可能的原因**
  - 事件：违反系统需求的故障
    - 如“列车与障碍物相撞”
  - 逻辑门：使用与门AND、或门OR连接事件
  - 基本事件：无需进一步分解的根本原因
    - 如“摄像头硬件故障”



风险分析：故障树分析

示例：列车门夹人故障

- **顶事件 (REQ)**: 乘客被门夹住 。
- **一级分解 (AND)**: [人仍在门内] **并且** [门执行了关闭动作] 。
- **二级分解**:
  - [门执行了关闭动作]
    可能是因为 (OR): [安全系统失效] **或者** [操作员手动覆盖] 。
  - [安全系统失效]
    可能是因为 (AND): [视觉系统未检测到人] **并且** [压力传感器也失效了] 。
- **基本事件**:
  - 视觉系统失效 (OR):
    模型错误 (SPEC), **摄像头缺陷 (ASM)**, **光线差 (ASM)** 。
  - 压力传感器失效 (OR):
    传感器故障 (ASM), **软件崩溃 (SPEC)** 。



风险分析：故障树分析

结合故障树设计缓解措施

- 识别最小割集 (Minimal Cut Set)：找到导致顶事件发生的最简根本原因组合
  - {摄像头故障, 压力传感器故障, 人在门内} 是一个最小割集
- 设计缓解措施的目标
  - 增加割集大小：增加更多的冗余或检查，让故障发生需要满足更多条件
    - 缓解: 将单个压力传感器升级为两个独立的压力传感器，现在需要两个都坏掉才行
  - 消除基本事件：通过重新设计来移除某个故障路径
    - 缓解: 重新设计软件，使其在崩溃时默认阻止车门关闭，而不是执行最后指令

风险分析： FMEA

- **一种自下而上的归纳分析法，从单个模块出发，系统地枚举其所有可能的“失效模式”，并分析每种模式对整个系统的“影响”**
  - FMEA 表格
    - 模块：分析的对象（如：视觉系统）
    - 失效模式：可能的出错方式（如：未检测到人，或错误检测到人）
    - 失效影响：对系统的最终后果（如：伤害乘客，或造成延误）
    - 严重度(S), 发生率(O), 探测度(D)：对风险进行量化评级
    - 建议措施：提出缓解方案

风险分析： FMEA

北京大学 PEKING UNIVERSITY

示例

模块	故障模式	故障影响	潜在原因	建议措施
摄像头	能见度差	碰撞风险	夜晚、天气、镜头遮挡	优雅降级、冗余、警报人工操作员
摄像头	硬件故障	碰撞风险	制造缺陷、寿命终止	半年检查
障碍物检测模型	漏检障碍物	碰撞风险	不可靠的机器学习模型	冗余感知系统
障碍物检测模型	误检障碍物	操作受阻	不可靠的机器学习模型	警报远程操作员、允许操作员覆盖
激光雷达	激光雷达干扰	碰撞风险、操作受阻	区域内其他车辆使用激光雷达	通过ID对激光信号进行编码以防止干扰

风险分析： HAZOP

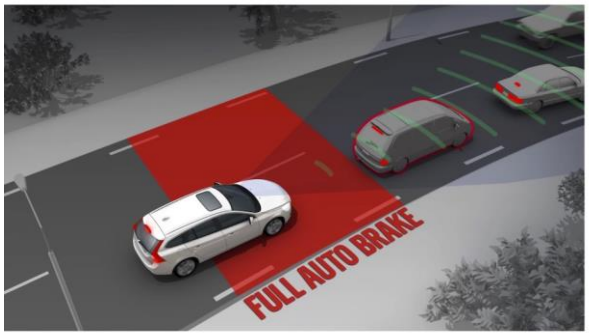
- **使用一系列标准引导词来系统地检查设计意图的各种可能偏差**
- **常用引导词:**
  - 无/否 (No/Not)：功能完全缺失
  - 更多/更少 (More/Less)：数量上的偏差
  - 晚/早 (Late/Early)：时间上的偏差
  - 反向 (Reverse)：逻辑上的相反
  - 其他/替代 (Other/Instead)：完全被替换
- **针对机器学习的扩展引导词: 错误(Wrong), 无效(Invalid), 不完整(Incomplete), 受扰动(Perturbed)**

风险分析： HAZOP

北京大学 PEKING UNIVERSITY

示例：汽车的紧急制动系统（EB）

- 无（NO 或 NOT）： EB未产生任何制动指令
- 不足（LESS）： EB发出的制动力小于最大制动力
- 延迟（LATE）： EB虽然发出最大制动指令，但延迟了2秒
- 反向（REVERSE）： EB生成了加速指令而不是制动指令
- 过早（BEFORE）： EB 在可能发生碰撞之前就提前发出了最大制动指令



Guide Word	Meaning
NO OR NOT	Complete negation of the design intent
MORE	Quantitative increase
LESS	Quantitative decrease
AS WELL AS	Qualitative modification/increase
PART OF	Qualitative modification/decrease
REVERSE	Logical opposite of the design intent
OTHER THAN / INSTEAD	Complete substitution
EARLY	Relative to the clock time
LATE	Relative to the clock time
BEFORE	Relating to order or sequence
AFTER	Relating to order or sequence

本讲小结

- **需求陈述了利益相关者的目标，并以现实世界中的概念来表达**
  - 软件/AI模型对现实世界的影响有限
  - 环境假设在建立需求方面同样重要
- **收集需求的方法：识别利益相关者，访谈他们，解决冲突**
- **AI模型是不可靠的，错误不可避免，且通常不是随机的**
- **应对错误的重点在于系统层面的容错设计，而非仅仅优化模型本身**
  - 五大设计策略来构建健壮的系统
  - 三种风险分析方法

团队实践：第一次汇报要求

北京大学 PEKING UNIVERSITY

- **时间：10月10日（周五）正课时间**
  - 每个小组10分钟汇报+10分钟交流
- **汇报内容**
  - 团队基本信息（成员，名称及由来，团队公约内容）
  - 团队追踪情况（第一阶段分工）
  - 项目信息（项目目标，需求分解，风险分析及应对方式）
- **项目信息**
  - 项目目标（组织目标/系统目标/用户目标/模型目标，尝试给出量化指标）
  - 需求分解（功能需求，非功能需求，对每个需求从ASM/SPEC/REQ方面分解，至少1个功能需求+1个非功能需求）
  - 风险分析（对应于需求，故障树分析，可以有多个）
  - 应对方式（至少两个）

团队实践：第一次汇报要求

示例：智能行车记录仪

- 背景：行车记录仪得到广泛使用，作为开发商希望与一个非营利组织在儿童安全方面合作，通过行车记录仪影像寻找失踪儿童，并实时报告给警方系统。
- 项目目标（示例）
  - 组织目标：开发的智能行车记录仪在 1 年内解决 20% 的儿童失踪案
  - 系统目标：开发的行车记录仪系统报告给警方的数据错误率<0.1%
  - 用户目标：行车记录仪的录制功能不受影响
  - 模型目标：模型在真实场景识别儿童准确率大于90%

团队实践：第一次汇报要求

北京大学 PEKING UNIVERSITY

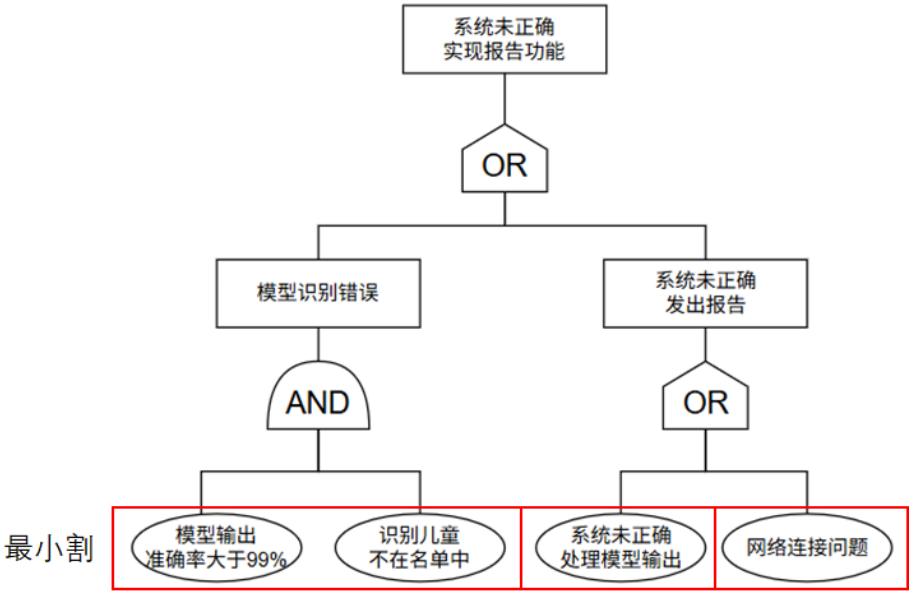
示例：智能行车记录仪

- 需求分解（示例）
  - REQ: 系统在识别到失踪儿童时，可以报告给警方系统（功能需求）
  - ASM1: 模型输出相似度大于99%时，认定两张图片描述的是一个人
  - ASM2: 系统在获取模型输出后，能够正确做出报告的行为
  - ASM3: 系统具有网络连接与较低的延时
  - ……
  - SPEC1: 系统获取到模型输出后，判断相似度，如果大于99%，报告警方
  - ……
- 非功能需求（示例）
  - REQ1: 用户隐私信息受到保护（安全性）
  - REQ2: 系统识别响应延迟<2s（性能）
  - ……

团队实践：第一次汇报要求

示例：智能行车记录仪

- 风险分析
  - 工具：draw.io或https://www.fault-tree-analysis-software.com/



- 应对方式
  - 二次模型验证：使用多种模型综合判断输出结果，提高可信度，可用于缓解发生“模型识别错误”的后果
  - 系统沟通回执：接受警方系统提供的信息回执，确保上报信息准确传达，可用于缓解发生“网络连接问题”的后果

谢 谢

欢迎在填写问卷反馈

北京大学
PEKING UNIVERSITY