智能化软件系统与工程

AI系统开发过程与团队合作

马 郓

人工智能研究院

北京大学 PEKING UNIVERSITY

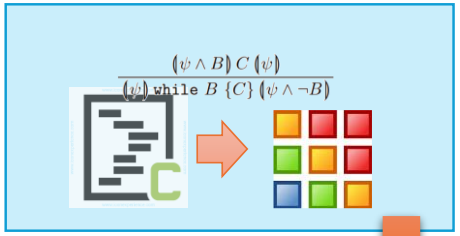
资料整理
火龙果科技

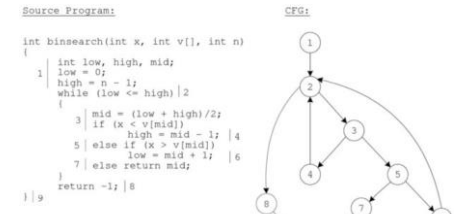
上讲回顾：什么是智能化软件？

智能化软件的内涵：融合了程序员编写的确定性符号计算程序部件和机器学习生成的非确定性概率计算模型部件，具有复杂网络结构，并可能呈现涌现行为的软件系统

由人或机器辅助人编写程序完成

代码部件

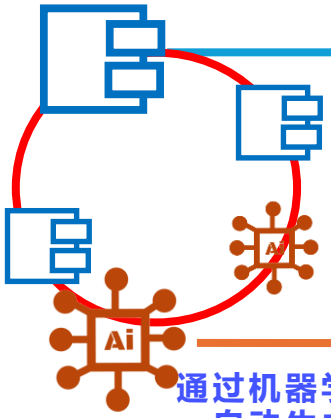




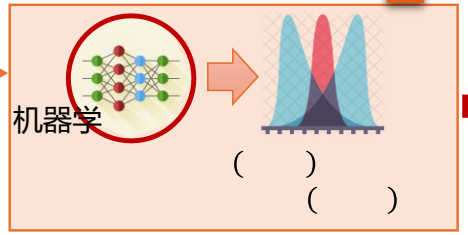
- **数值计算**
- **排序**
- **搜索引擎**

- **形态**：符号化的代码实现图灵机，数据作为处理对象
- **行为**：功能由可解释的程序代码描述的明确的计算过程实现

二者混合交织，协同工作



通过机器学习自动生成





- **分类（图像/语音识别）**
- **回归（预测）**
- **内容生成**

- **形态**：对神经网络的拟，权重数据是关键，框架代码相对简单固
- **行为**：功能由主要由不可解释的权重数据决定；更强的功能意味着更深的网络和更多权重数据

模型部件

资料整理 火龙果科技 荐

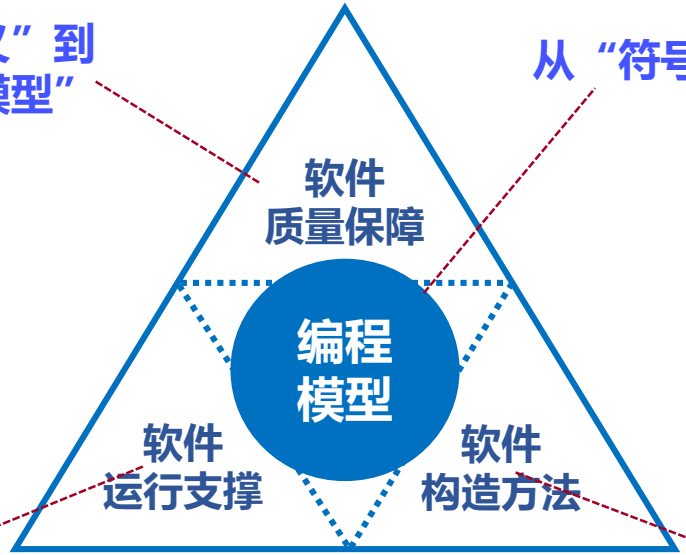
上讲回顾：智能化软件为何开发难？

北京大学 PEKING UNIVERSITY

软件范型四要素均在发生转变

从“确定性行为和可解释语义”到
兼有“非确定行为和难解释模型”

从“符号软件”到“神经-符号融合软件”



从“相对可控的软件个体及其交互”到
“[illegible]的自主协同和群智涌现”

从“知识驱动、人工开发”到
“数据/知识双驱动、自动合成

资料整理
火龙果科技 荐

团队实践组队

- **4~5人组队，进行团队合作**
- **自由组队，课间登记**
  - 旁听的同学请不要与选课同学共同组队
  - 无法找到队伍的同学由助教协助组队
- **本周任务**
  - 给团队起个名字
  - 建立团队公约和追踪机制
    - 本节课会具体介绍

资料整理
火龙果科技 荐

AI系统开发过程

资料整理
火龙果科技 荐

什么是过程?

北京大学 PEKING UNIVERSITY

软件生命周期

- 软件产品或系统的一系列活动的全周期。从形成概念开始，历经开发、交付使用、在使用中不断修订和演化，直到最后被淘汰

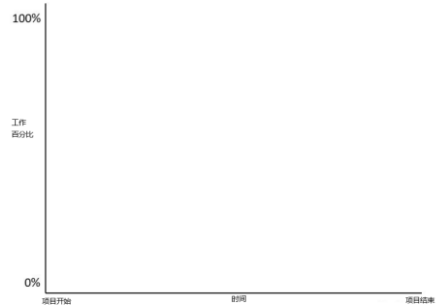
软件过程

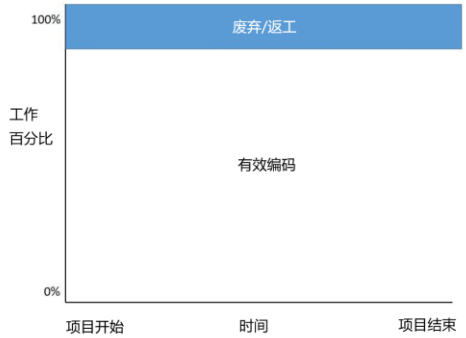
- 在软件生命周期内，为实现特定目标而进行的所有工程活动的集合
- 其本质是将用户需求转化为软件产品的系统性方法
- 过程、活动、任务
  - 过程是活动的集合
  - 活动是任务的集合
  - 任务是把输入转换成输出的操作

资料整理
火龙果科技 荐

什么是过程?

开发者通常讨厌过程管理

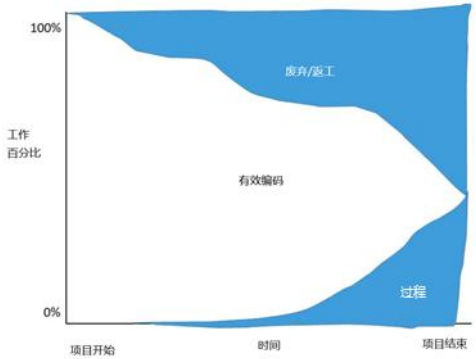




开发者的愿望

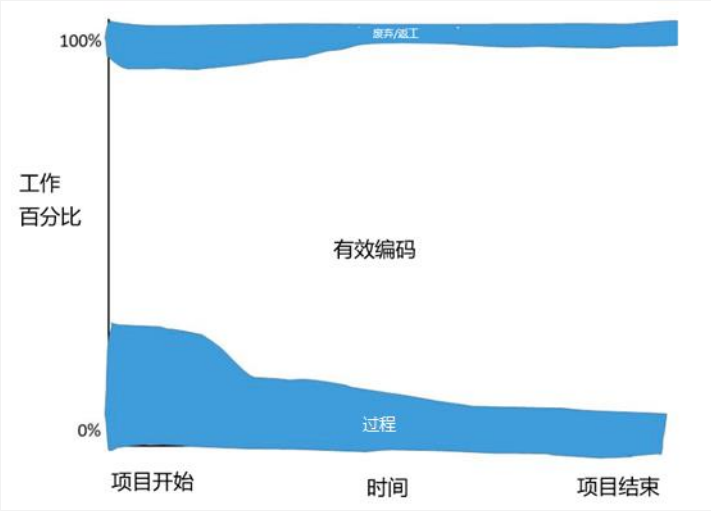


开发者认为的过程



最终实际发生的情况

- **过程能够提升软件开发的灵活性与效率**
- **提前做好过程规划能在项目后期获到巨大收益**



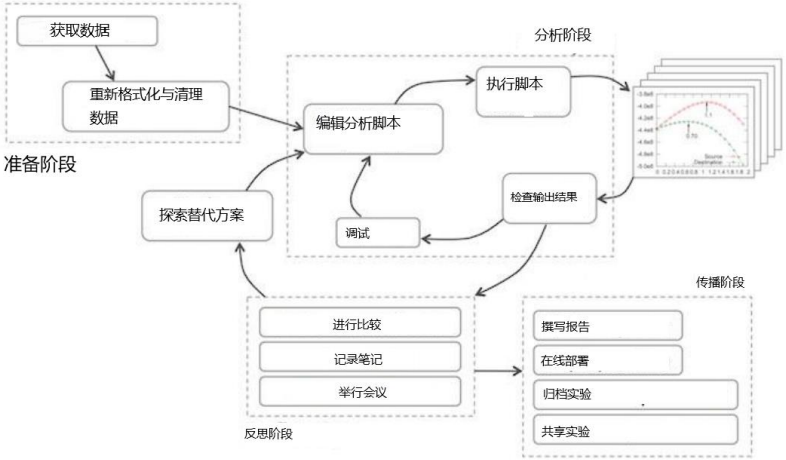
资料整理
火龙果科技

数据科学的工作过程

北京大学 PEKING UNIVERSITY

迭代与探索

- 目标粗略、需求模糊、能否成功未知
- 经验与启发式驱动
- 试错、细化、假设检验
- 可能需要回退到数据收集/清洗、修订目标



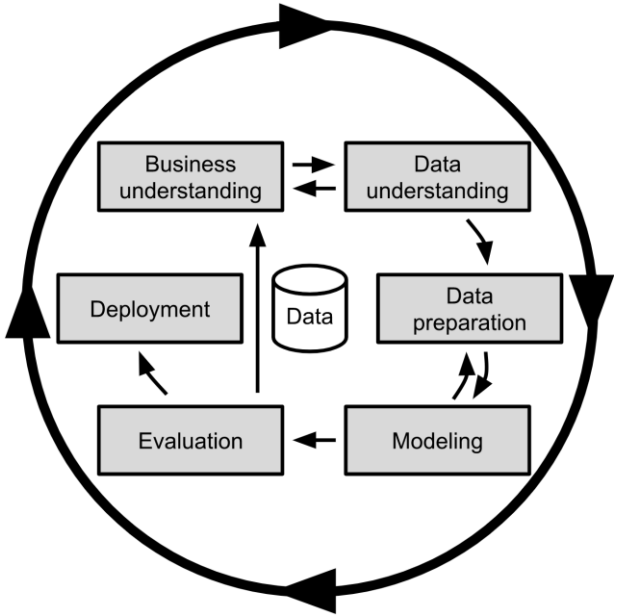
	First 2 Hours	Second 2 Hours	Final Accuracy
TAP1			84.7%
TAP2	X	X	75.3%
TAP3			78.3%
TAP4			82.9%
TAP5			84.7%
TAP6			78.0%
TAP7			56.9%
TAP8			22.8%
TAP9			78.8%
TAP10			84.4%

10位数据科学工作者训练手写体识别模型的过程记录

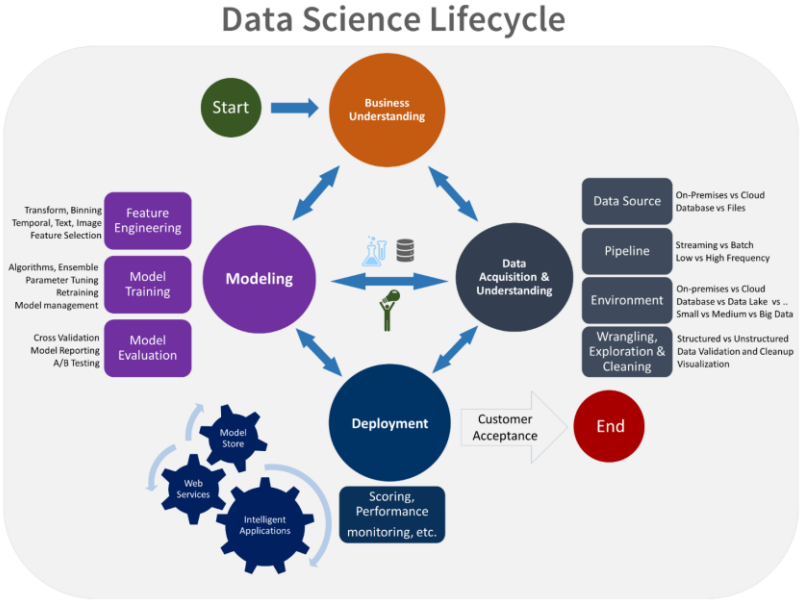
资料整理 火龙果科技 荐

数据科学的工作过程

业界案例



IBM的CRISP-DM过



微软的数据科学过

资料整理
火龙果科技 荐

数据科学的工作过程

北京大学 PEKING UNIVERSITY

可计算笔记本（Computational Notebooks）

- 源于“文学化编程”
- 文本+代码单元，结果即时展示
- 目前最为流行的是Jupyter Notebook



可计算笔记本的优点和局限?

```
# load data collected from team1
import pandas as pd

url = 'http://128.2.25.78:8080/private/log1.clean'
df = pd.read_csv(url)
df.head()
```

	dayIdx	user	userAvgTime	location	dow	isWeekend	time
0	0	Pittsburgh66Correy	7.045001	Pittsburgh	6	True	0.000000
1	1	Pittsburgh66Correy	7.045001	Pittsburgh	7	True	6.883333
2	2	Pittsburgh66Correy	7.045001	Pittsburgh	1	False	6.816667
3	3	Pittsburgh66Correy	7.045001	Pittsburgh	2	False	7.383333
4	4	Pittsburgh66Correy	7.045001	Pittsburgh	3	False	0.000000

Data was preprocessed externally, identifying the time at a given day when the light was first turned on 12pm). Weather and sunrise information is not included here, though that'd be important. If the light w this morning (quite common), 0 is recorded.

```
[ ] # just data encoding and splitting X and Y

    X = df.drop(['time'], axis=1)
    YnonZero = df['time'] > 0
    Y = df['time']

    from sklearn import preprocessing
    # leDate = preprocessing.LabelEncoder()
    # leDate.fit(X['date'])
    # leDate.transform(X['date'])

    X=X.apply(preprocessing.LabelEncoder().fit_transform)
    X
```

资料整理 火龙果科技 荐

软件开发过程

北京大学 PEKING UNIVERSITY

一个随意的开发过程

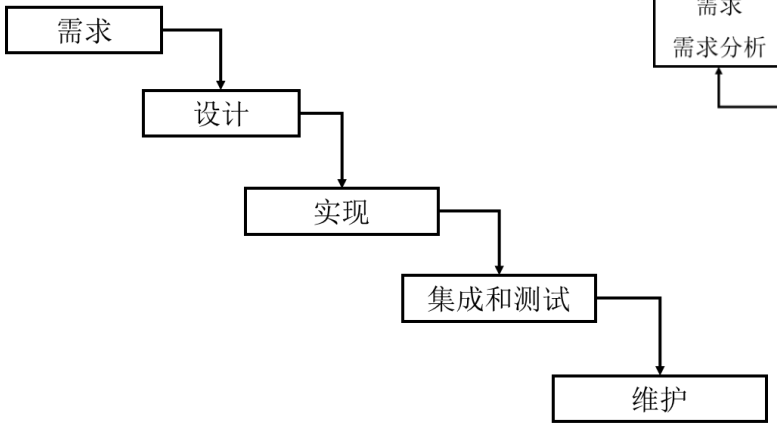
1. 讨论需求
2. 写代码
3. 测试找缺陷
4. 调试
5. 修复缺陷
6. 未完成则返回 1

资料整理
火龙果科技 荐

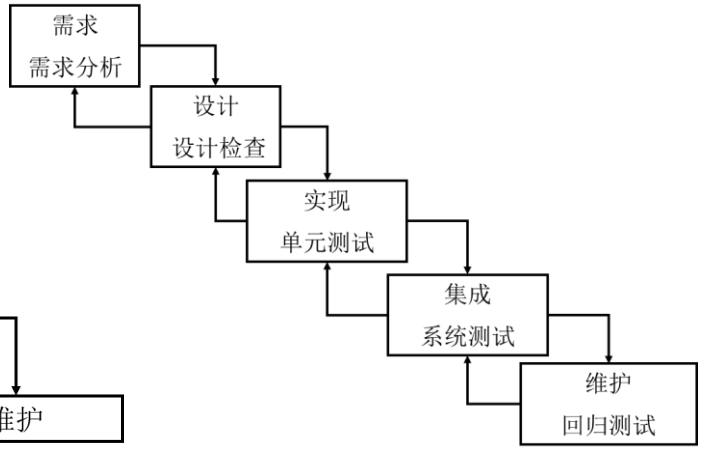
软件开发过程：计划驱动

瀑布模型

- 将软件生命周期的各项活动规定为依固定顺序连接的若干阶段工作
- 规定了每一阶段的输入，以及本阶段的工作成果，作为输出传入下一阶段



理论瀑布模型



修订版瀑布模型



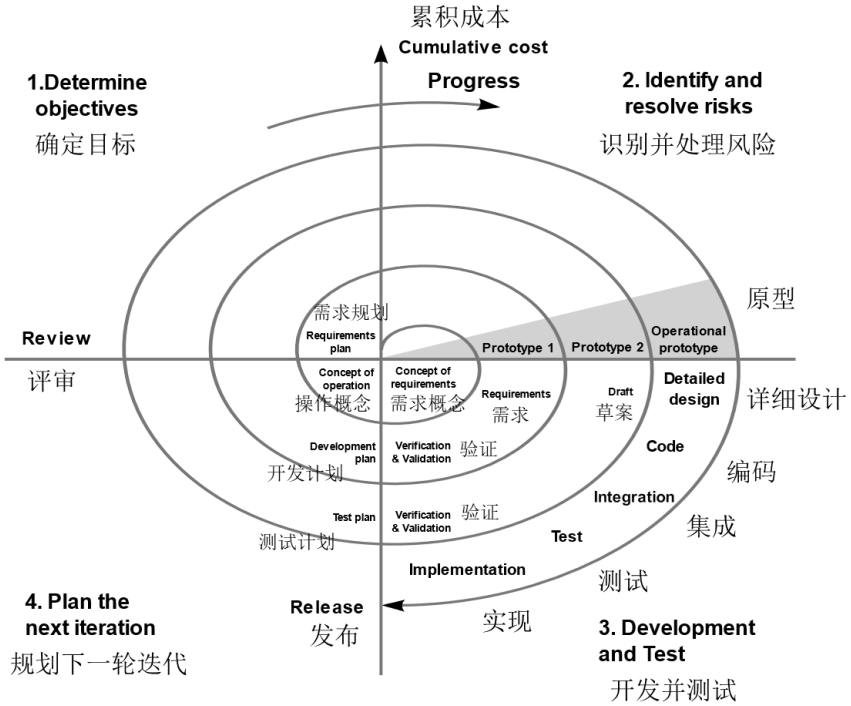
瀑布模型的缺点?

资料整理
火龙果科技 荐

软件开发过程：计划驱动

螺旋模型

- 增量的逐步开发原型，优先开发风险最大的部分



资料整理
火龙果科技 荐

软件开发过程：敏捷开发

Scrum：管理软件项目的一个轻量级敏捷方法

- 规划纲要阶段：建立大致的项目目标和设计软件体系结构
- 一系列的冲刺（Sprint）循环，每个冲刺开发出一个系统增量
- 项目结束阶段总结项目，完善需要的文档，如系统帮助和用户手册，并总结从项目中获得的经验。



资料整理
火龙果科技 荐

软件开发过程：敏捷开发

北京大学 PEKING UNIVERSITY

Scrum的3个主要工件

- 产品任务清单：需求、故事或特性等组成的列表，按照重要性的级别进行排

序号　优先级，重要程度　需求描述(story)

Item #	Priority	Product Backlog Item	Description	Issued by
3	1	Design web page look and feel, flow	Create and review with customer look and feel of the page.	JK
5	2	Create graphics & banners	Hotel logo, animated advertisment banner, background images	JK
7	3	Visitor and admin mode	Admin should have possibility to easily change content of the page	IT
4	4	Add address information including interactive map	User should be able to zoom in and zoom out map of the location of the hotel	JK
1	5	Create subpage with pricing	Pricing information available at front desk	JK
8	6	Create subpage with possible events handled in hotel	Contact John Kowalsky for the list of events	JK
6	7	Create gallery subpage		-
2	8	Create references subpage	Contact John Kowalsky for the references	JK

发布人

- 功能方面的需求，功能点
- 功能方面的需求，如性能改进等
- 需要修改的bug，上一版本已知的问题
- 新技术，如支持新的操作系统或平台
- 问题， 日后可能新增的项，新功能

- 产品任务清单是不断完善的

资料整理 火龙果科技 荐

软件开发过程：敏捷开发

Scrum的3个主要工件

- 冲刺任务清单：从产品任务清单中挑选出高优先级的任务，确定本次迭代的任务目标
  - 能提取多少产品任务清单中的任务取决于团队能承诺完成多少



Item #	Priority	Product Backlog Item	Size	Task	Owner	Est. [h]	Status
3	1	Design web page look and feel	2	Check requirements with customer	John	1	Done
				Discuss on page layout	Erik, Pavel, Marian, John	8	Done
				Create web page template	Pavel	6	In progress
				Create design document	Marian	3	In progress
				Review design with customer		4	Not started
5	2	Create graphics & banners	5	Task	Owner	Est. [h]	Status
				Create Hotel logo		2	Not started
				Create animated advertisment banner	Erik	2	Impeded
				Create background images		6	Not started

资料整理 火龙果科技 荐

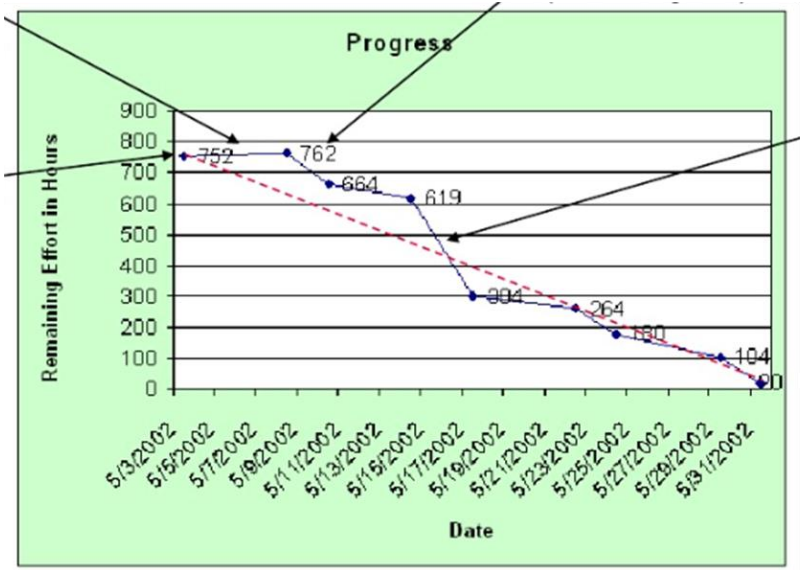
软件开发过程：敏捷开发

Scrum的3个主要工件

- 燃尽图：显示了冲刺中积累剩余的工作量，反应工作量完成状况的趋势

剩余工作量增加

剩余工作量汇总



在冲刺开始时预先估算的时间

快速下降：进度快

资料整理
火龙果科技 荐

软件开发过程：敏捷开发

北京大学 PEKING UNIVERSITY

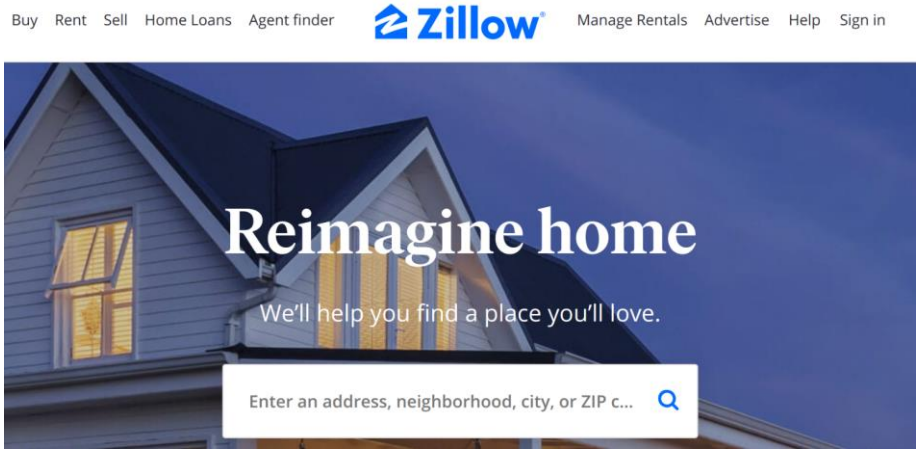
Scrum的4个主要活动

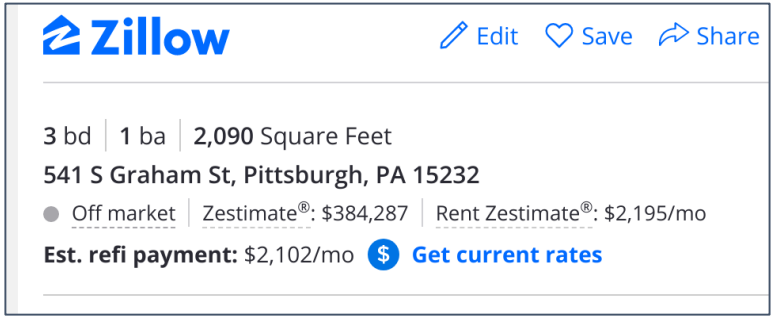
- 冲刺计划会议：冲刺开始前召开，团队共同选择和理解在即将到来的冲刺中要完成的工作，确定冲刺任务清单
- 每日站立会议：在整个迭代过程中每天定时召开，了解项目的真实进展情况，主要回答三个问题
  - 上次会议之后做了什么？
  - 下次会议之前准备做什么？
  - 有没有困难？
- 冲刺评审会议：冲刺结束时评审产出，确保成果与预期的一致
- 冲刺回顾会议：评价本次冲刺并酝酿改进，使得下一轮冲刺进行得更好

资料整理
火龙果科技 荐

软件开发过程

- **讨论：开发一个预测房屋价格的网站，选用哪种开发过程?**
  - 1. 瀑布模型；2. 螺旋模型；3. 敏捷开发





资料整理
火龙果科技 荐

过程视角的数据科学 vs 软件开发

北京大学 PEKING UNIVERSITY

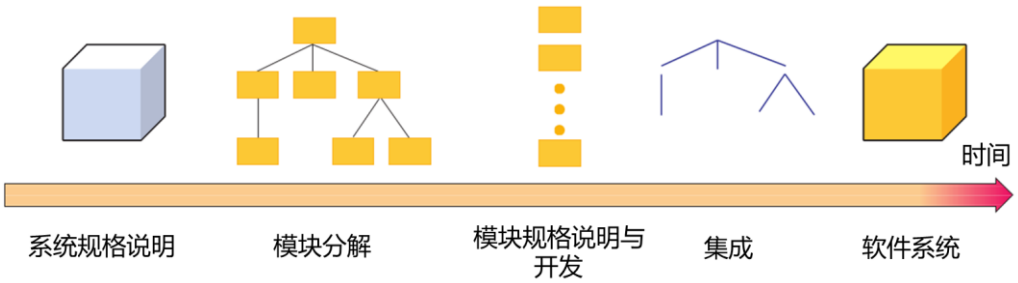
- **数据科学的迭代和软件开发的迭代有所不同**
  - 螺旋模型强调每一轮迭代都有提升，但AI模型的每次迭代不一定会有更好的效果
  - 敏捷开发强调通过迭代应对需求的不确定性，但智能化软件的需求通常明确，只是不确定AI模型是否能有效解决问题
- **智能化软件的开发不要盲目采用“软件工程最佳实践”**
  - 在模型探索阶段，文档/测试并非高优先级
  - 模型达到预期后，再开展工程化过渡
    - 从计算型记事本到流水线
    - 维护与迭代
    - 提供基础设施与工具

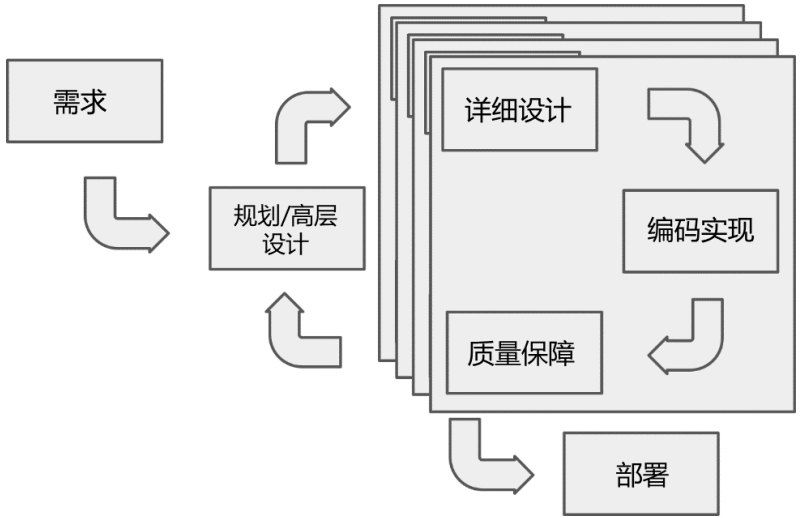
资料整理
火龙果科技 荐

一种可能的智能化软件开发过程

北京大学 PEKING UNIVERSITY

整体遵循软件系统开发的整体流程



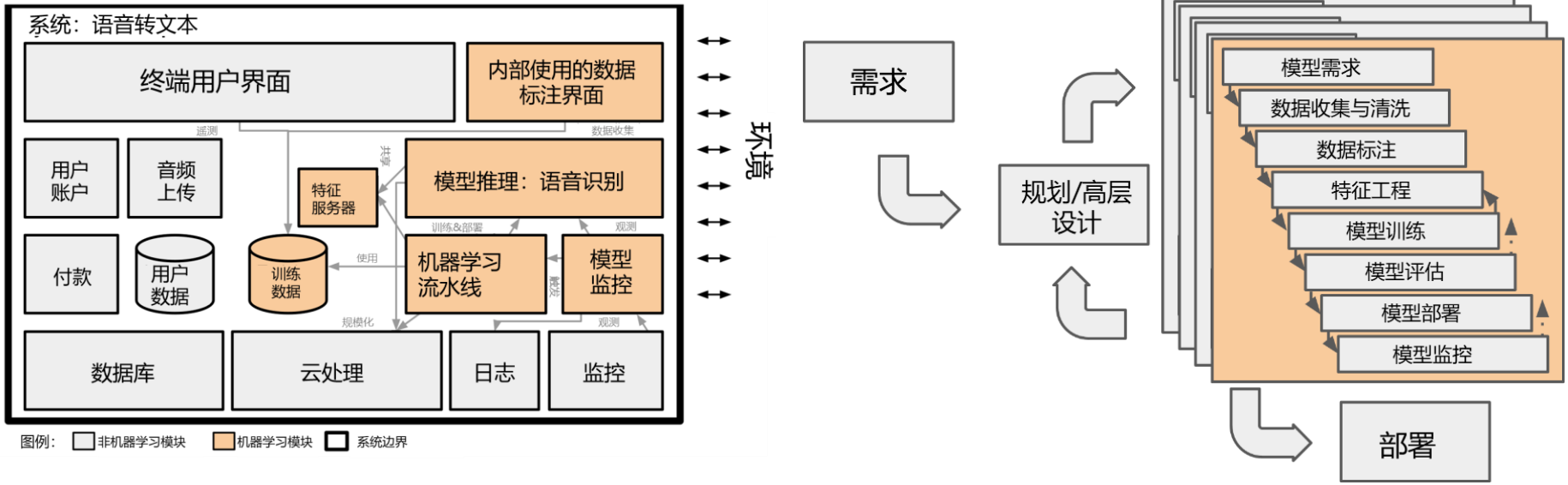


资料整理
火龙果科技 荐

一种可能的智能化软件开发过程

北京大学 PEKING UNIVERSITY

考虑AI模型作为系统的一个部分



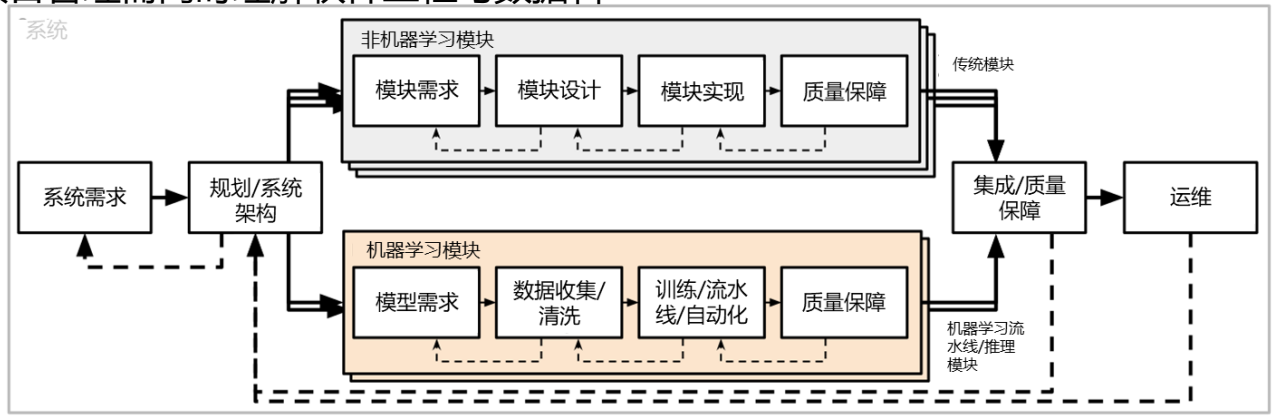
一种可能的智能化软件开发过程

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　　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

尚无“最佳实践”或过程模型



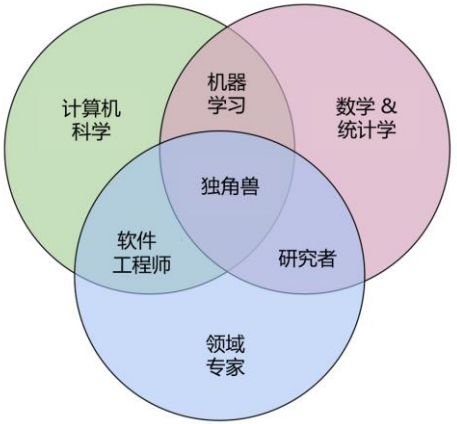
资料整理 火龙果科技 荐

跨学科团队合作

资料整理
火龙果科技 荐

团队合作

- **软件开发中，团队合作是必然**
  - 项目过于庞大，单人无法完成（工作分工）
  - 项目过于复杂，单人无法完全理解（分而治之）
  - 项目所需技能过多，单人无法全部掌握（专业分工）
- **智能化软件开发中，跨学科团队合作是必然**
  - 独角兽是不够的
    - 人才稀缺、薪酬极高
    - 认知负荷限制，无法在多个领域保持深度
    - 任何复杂的系统都需要分工
  - 放弃寻找独角兽，转而组建跨学科团队



资料整理
火龙果科技 荐

团队合作的常见问题与应对方法

北京大学 PEKING UNIVERSITY

- 过往的团队合作中失败/让人崩溃的经历?



资料整理
火龙果科技 荐

团队合作的常见问题与应对方法

北京大学 PEKING UNIVERSITY

从众思维

- 原因
  - 团队组成同质化
  - 结构性缺陷（与外界隔绝、领导偏见、缺乏方法论）
  - 环境因素（外部压力、近期的失败）
- 应对方法
  - 团队成员多样性
  - 对冲突持开放态度的文化
  - 在讨论中指定“反方辩手”，调节并轮换发言顺序，领导者在讨论中隐藏自己的观点
  - 引入外部专家
  - 始终要求备选方案
  - 监控与过程测量

资料整理
火龙果科技 荐

团队合作的常见问题与应对方法

北京大学 PEKING UNIVERSITY

- **社会惰化：个体作为团队成员进行群体活动时，会降低自己的努力和表现水平，所付出的努力比单独完成时偏少**
  - 原因
    - 责任分散
    - 缺少动力
    - 努力的可有可无 / 缺乏认可
    - 避免拖累所有人 / “傻瓜效应”
  - 应对方法
    - 让团队所有成员参与，集中办公
    - 分配具体任务，明确个人责任
      - 提高个人的可识别性
      - 团队契约，绩效衡量
    - 分配任务时提供选择的权利
    - 促进参与，激励开发者
    - 评审与反馈
    - 开展团队建设活动，提高凝聚力
    - 小型团队

资料整理
火龙果科技 荐

学生团队合作

常见问题

- 直到最后一刻都没有可见的进展
- 延期
- 集成时解决方案不完整或质量低下
- 团队成员没有回应
- 团队成员被动、缺乏兴趣、没有主动性
- 需要大量提醒和帮助

原因

- 优先级差异（“另一门课快把我逼疯了，我得先做那个”）
- 志向差异（“及格就够毕业了”）
- 能力差异（团队中“无能”的学生）
- 做事风格差异（DDL斗士 vs 计划型）
- 沟通偏好差异（避免干扰 vs 时刻在线）
- 团队内部围绕成绩的竞争（互相超越，对立式同伴评分）

资料整理
火龙果科技 荐

学生团队合作

北京大学 PEKING UNIVERSITY

讨论：遇到下面的情况，你们会怎么做?

有一位团队成员几乎没有技术经验，在处理基本的Python脚本和命令行操作时遇到了困难。其他团队成员接手任务比帮助他完成要快得多。

你们分配了工作，但在截止日前一天晚上尝试整合时，才发现有一位团队成员没有完成他的部分。他前一天尝试过，但遇到了环境配置依赖问题而卡住了。

在上一次作业最后时刻的压力之后，你们团队同意提前开始，并在截止日前几天的一个里程碑节点进行整合，以留出缓冲时间。然而，你在GitHub上几乎看不到一半团队成员的进展，微信上也几乎没人回应。在约定的里程碑节点时，完成的工作很少。虽然最终在截止日前完成了任务，但压力和上一次作业时一样大。

对于一位团队成员来说，这次作业优先级很低。他很少在最后时刻之外做出超过最低要求的贡献。团队其他成员勉强补足，最终获得了作业的全部分数。你不愿意批评这位同学，因为这可能会对他的成绩产生负面影响。

资料整理
火龙果科技 荐

本课程的团队合作要求

北京大学 PEKING UNIVERSITY

- **团队可以自行设定优先事项和具体策略**
  - 选择适合你们的方法，大胆尝试
  - 不要求每个人在每项任务中都能做出同等的贡献
  - 团队成员各有所长，也有各自的弱点
    - 这是好事情!

黄金法则：尽量在约定的时间内完成你承诺的事情。如果做不到，请及时寻求帮助并提前清楚沟通。

资料整理
火龙果科技 荐

本课程的团队合作要求

北京大学 PEKING UNIVERSITY

团队义务

- 保持积极响应和负责任的态度
- 按时参加会议，积极参与
- 遵守承诺，完成分配的任务
- 遇到问题时，主动联系，重新规划，提前沟通，积极应对
  - 重新规划和处理错误是非常正常的

如果出现以下情况，将影响成绩评定：

- 缺乏沟通
- 不尊重或轻视的沟通方式
- 无故不参加团队会议
- 未沟通就错过内部截止日期
- 未及时沟通而未完成约定任务

资料整理
火龙果科技
荐

团队合作建议

北京大学 PEKING UNIVERSITY

建立沟通与会议模式——建立团队公约

- 就团队内部如何沟通达成一致：使用电子邮件？微信群？
- 就沟通期望达成一致。不同的人有不同的习惯和期望。请明确说明！
  - 每天查看电子邮件？周末也查看吗？
  - 在3小时内回复紧急聊天消息？阅读之前的聊天消息吗？
  - 在特定时间段内保持在线以便聊天？
- 会议时间：提前计划还是按需开会？
- 为集成设定中间的内部截止日期
- 设定切合实际的期望：每个人都有其他课程和干扰；请坦诚沟通可用时间
- **将期望写下来！**

资料整理
火龙果科技 荐

团队合作建议

北京大学 PEKING UNIVERSITY

共享工作

- 团队成员各有所长和短板——这是好事
- 利用团队成员的个人优势（分工、结对、互助等）
- 通常会有人负责团队管理任务（如安排会议、主持、会议记录、跟踪工作、提醒、检查提交）或汇报
  - 团队管理也是一项工作
  - 可以考虑轮流负责

资料整理
火龙果科技 荐

团队合作建议

北京大学 PEKING UNIVERSITY

跟踪与审计——建立跟踪机制

- 写下明确的可交付成果：谁在什么时间前做出什么
  - 明确预期结果，结果应可验证
  - 跟踪完成情况，完成后进行标记
  - GitHub issues、Jira、Trello看板、在线文档等带有历史记录追踪的单一来源
- 在会议期间完成可交付成果清单：每个人写下自己的可交付成果，其他人阅读所有可交付成果以检查理解情况
  - 如果在会议期间未完成，或团队成员未参加会议，会后通过邮件将任务分配发送给所有人；24小时内无人提出异议则视为同意任务分配
- 在第一次团队项目汇报时介绍任务跟踪情况
- 与团队的责任助教共享此文档

资料整理
火龙果科技 荐

团队合作建议

北京大学 PEKING UNIVERSITY

沟通技巧

- 全体成员参加的会议聚焦于规划和反思，针对具体工作单独讨论
- 合理使用群聊工具
  - 将聊天视为临时交流，不要期望每个人都能补看所有的历史消息
  - 将社交沟通与工作沟通分开，将紧急事项与非紧急事项分开
  - 如需他人参与，请明确标记相关人员
  - 在“工作时间”内开启群通知
  - 将非紧急、长期事项放在与主题相关的聊天之外讨论（如GitHub issues、在线文档等）
- 预留时间用于团建和庆祝！

资料整理
火龙果科技 荐

本讲小结

- 数据科学和软件工程具有不同的工作过程
- 智能化软件系统的开发需兼顾两者的过程需求
- 团队合作是不可避免的，团队成员很少能完全自我选择，优秀的团队合作是有趣的
- 团队合作很难，需要学习相关技能
  - 要对沟通和工作分配设定明确的期望

资料整理
火龙果科技 荐

谢 谢



北京大学
PEKING UNIVERSITY

资料整理
火龙果科技
荐