

2024MCM Problem C: 网球运动中的动力



在 2023 年温布尔登网球公开赛男子组决赛中，20 岁的西班牙新星卡洛斯-阿尔卡拉斯击败了 36 岁的诺瓦克-德约科维奇。这是德约科维奇自 2013 年以来首次在温布尔登输掉比赛，也结束了这位大满贯历史上最伟大球员之一的辉煌战绩。

比赛本身就是一场出色的战斗。1 德约科维奇似乎注定要轻松获胜，因为他在第一盘以 6-1 的比分占据优势（7 局比赛中赢了 6 局）。然而，第二盘比赛却十分紧张，最终阿尔卡拉斯在决胜盘中以 7-6 获胜。第三盘的情况与第一盘相反，阿尔卡拉斯以 6-1 的比分轻松获胜。第四盘开始后，年轻的西班牙人似乎完全控制了局面，但不知何故，比赛的走势再次发生了变化，德约科维奇完全控制了局面，以 6-3 的比分赢得了这一盘。第五盘也是最后一盘比赛开始后，德约科维奇延续了第四盘的优势，但比赛的走向再次发生了变化，阿尔卡拉斯取得了控制权，并以 6-4 的比分赢得了胜利。本场比赛的数据在提供的数据集中，"match_id" 为 "2023-wimbledon-1701"。您可以使用 "set_no" 列（等于 1）查看第一盘德约科维奇占优时的所有得分情况。似乎占优的一方出现了令人难以置信的波动，有时是多分甚至是多局的波动，这通常归因于 "势头"。

在字典中，"动量" 的定义是 "通过运动或一系列事件获得的力量或作用力"。"[2] 在体育运动中，一支球队或一名球员可能会觉得他们在比赛中拥有动量或 "力量/作用力"，但很难衡量这种现象。此外，如果存在 "势" 的话，比赛中的各种事件是如何产生或改变 "势" 的，也不是一目了然的。

提供 2023 年温布尔登网球公开赛前两轮之后所有男子比赛中每一分的数据。您可以自行决定加入其他球员信息或其他数据，但必须完整记录数据来源。使

用这些数据：

(1) 建立一个模型，捕捉赛点发生时的比赛流程，并将其应用到一场或多场比赛中。您的模型应能确定哪位球员在比赛中的某个特定时间段表现更好，以及他们的表现好到什么程度。根据您的模型提供可视化的比赛流程描述。注意：在网球比赛中，发球的一方赢得赛点/比赛的概率要高得多。您可能希望以某种方式将这一因素考虑到您的模型中。

(2) 一位网球教练对 "势头" 在比赛中的作用持怀疑态度。相反，他假设比赛中的波动和一名球员的成功是随机的。请使用您的模型/度量来评估这一说法。

(3) 教练们很想知道，是否有一些指标可以帮助判断比赛的流程何时会从偏向一名球员变为偏向另一名球员。

①利用提供的至少一场比赛的数据，建立一个模型来预测比赛中的这些波动。哪些因素似乎最有关系（如果有的话）？

② 鉴于过去比赛中 "势头" 波动的差异，你如何建议球员在新的比赛中与不同的球员交手？

(4) 在一场或多场其他比赛中测试您开发的模型。您对比赛中的波动预测得如何？如果模型有时表现不佳，您是否能找出未来模型中可能需要包含的任何因素？您的模型对其他比赛（如女子比赛）、锦标赛、球场表面和其他运动（如乒乓球）的通用性如何？

(5) 撰写一份不超过 25 页的报告，介绍您的研究结果，并附上一至两页的备忘录，总结您的研究结果，并就 "动力" 的作用以及如何让球员做好准备，应对网球比赛中影响比赛进程的事件，向教练提出建议。

您的 PDF 解决方案总页数不超过 25 页，其中应包括：

- 一页摘要表。
- 目录
- 您的全套解决方案
- 一至两页的备忘录
- 参考文献列表
- 人工智能使用报告（如已使用，则不计入 25 页限制）。

注意：对于提交的完整材料，没有具体的最低页数要求。您可以用最多 25 页的篇幅来完成所有的解答工作，以及您想要包含的任何其他信息（例如：图纸、图表、计算、表格）。我们允许谨慎使用 ChatGPT 等人工智能，但没有必要为这一问题创建解决方案。如果您选择使用生成式人工智能，则必须遵守 COMAP 人工智能使用政策。这将导致一份额外的人工智能使用报告，您必须将其添加到 PDF 解决方案文件的末尾，并且不计入解决方案的 25 页总页数限制中。

提供的文件：

- Wimbledon_featured_matches.csv - 2023 年温布尔登网球公开赛第二轮之后的男子单打比赛数据集。
- data_dictionary.csv - 数据集说明。
- data_examples - 帮助理解所提供数据的示例。

术语表：

大满贯：网球大满贯是指在一个日历年度内赢得一个项目的全部四项主要锦标赛。四项大满贯赛事是澳大利亚网球公开赛、法国网球公开赛、温布尔登网球公开赛和美国网球公开赛，每项赛事为期两周。

关键术语/概念词汇表：：

计分

比赛：五局三胜制（温布尔登网球公开赛的男子组比赛）

一局：一组对局；6 局为一局，但棋手必须以两局的优势获胜，直到 6 - 6 打平时再进行决胜局（见下文）。

游戏：积分收集；玩家达到 4 分时获胜，但必须以 2 分获胜。见下文 "游戏得分"。

比赛得分：

- 0 points = Love
- 1 point = 15
- 2 points = 30
- 3 points = 40
- Tied score = All (e.g., "30 all")
- 40 – 40 = Deuce (players have won the same number of points, at least 3 points each)
- Server wins a deuce point = Ad-in (or "advantage in")
- Receiver wins a deuce point = Ad-out

发球：球员交替担任 "发球员"（击球的球员）和 "回击球员"。在职业网球比赛中，发球员往往占有很大优势。发球员有两次发球的机会，每次发球都要将球送入发球区，如果两次发球都未能将球送入发球区，即为 "双误"，发球员将获得这一分。

破发--回球选手赢得一局比赛。

破发点 - 如果回击者获胜，他们将赢得比赛的点。

保发--发球方赢得比赛。

决胜局：每局比赛在一方赢得 6 局后结束，只要他们至少领先两局（即 6 - 4）。否则，比赛继续进行，直到出现 6 - 6 的平局。此时将进行平局决胜。温布尔登网球赛的决胜局为先得 7 分（必须以 2 分优势获胜），但在第 5 盘比赛中，决胜局为先得 10 分（必须以 2 分优势获胜）。

休息时间/场地两侧：球员在第一场比赛后交换场地两侧，然后每两场比赛后交换场地两侧。从第 3 局开始，每次换边都有 90 秒的休息时间。在决胜局中，球员每六分换边一次。每局比赛结束后，球员还需休息至少 2 分钟。允许医疗暂停和一次卫生间休息。

References:

[1] Braidwood, J. (2023), Novak Djokovic has created a unique rival – is Wimbledon defeat the beginning of the end, The Independent,

<https://www.independent.co.uk/sport/tennis/novak-djokovic-wimbledon-final-carlos-alcarazb2376600.html>.

[2] <https://www.merriam-webster.com/dictionary/momentum>

[3] Rivera, J. (2023), Tennis scoring, explained: A guide to understanding the rules terms & point system at Wimbledon, The Sporting News,

<https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-system-pointterms/7uzp2evdhbd11obdd59p3p1cx>.

帮助理解数据集的示例

例 1：第 5 行

Column(s)	Value(s)	Description
match_id	"2023-wimbledon-1301"	The 3 in "1301" indicates a round 3 match and the "01" indicates the first match listed from that round.
elapsed_time	"0:01:31"	The point begins with a serve 1 minute and thirty-one seconds after the start of the first point of the match.
point_no, game_no, set_no ("no" is an abbreviation for number)	4, 1, 1	The point played is the 4 th point of the 1 st game of the 1 st set of the match.
p1_sets, p2_sets, p1_games, p2_games	0, 0, 0, 0	Since this is the first game of the match neither player has won a game or set yet.
p1_score, p2_score	15, 30	The score when the point is played is 15 (player 1), to 30 (player 2). Thus, player 1 won one of the previous points and player 2 won two points.
server	1	Player 1 (Alcaraz) is serving on this point.
serve_no	1	The point was played on the first serve meaning Alcaraz hit his first serve in play.
point_victor	1	Alcaraz wins this point (player 1).
p1_points_won, p2_points_won	2, 2	Player 1 (Alcaraz) is the point victor so his total is now 2 for the match (it was previously 1). For player 2 the value remains 2 since player 2 lost the point.
game_victor, set_victor	0, 0	Alcaraz winning the point makes the score in the game 30 – 30 (2 points each) so neither a game or set was won by either player on this point (both = 0).
Columns U – AC		Allow us to determine how the point was won:
p1_winner	1	Alcaraz won the point by hitting an "unreturnable shot."
p1_ace	0	The shot was not a serve (since = 0).
winner_shot_type	F	The shot was a forehand (as opposed to a backhand).
p2_net_pt	1	Player 2 (Jarry) positioned himself near the net somewhere during the point.
p2_net_pt_won	0	Since Alcaraz won the point, although Jarry was at the net during the point this value is 0.
Columns AH – AM	All = 0	Even had player 2 won the point, the game would not have been over so the point was not a "break point" and these are all 0.
p1_distance_run, p2_distance_run	51.408, 75.631	The distance each player ran (in meters) on this point.
rally_count	13	Number of shots hit during the point by both players combined.
speed_mph, serve_width, serve_depth, return_depth	130, BW, CTL, D	Alcaraz (the server) hit a 130 serve "Body/Wide" of the returner (we saw it was a first serve previously) and close to the line denoting in or out of play. Jarry (the returner) returned the ball "Deep" in the court (so near the other end of the court).

例 2：第 8 - 12 行

第一局的最后四分说明了平分 ("deuce") 和优势 ("ad") 的概念。每一行都是比赛中的后续时间点。

Row	Column(s)	Value(s)	Description
Row 8	p1_score, p2_score	40, 40	The score is 40 – 40 meaning each player has won 3 previous points (this is also called "deuce").
	point_victor	1	Alcaraz wins point 7 (in row 8).
Row 9	p1_score, p2_score	AD, 40	Since Alcaraz won the previous point (point 7) the score on point 8 is now "AD" for Alcaraz and "40" for Jarry meaning Alcaraz has won one more point and could win the game on the next point.
	point_victor	2	Jarry (player 2) wins point 8 (in row 9).
Row 10	p1_score, p2_score	40, 40	The score returns to 40 – 40 ("deuce") meaning each player has won the same number of previous points although now it is 4 points each.
	point_victor	1	Alcaraz wins point 9 (in row 10).
Row 11	p1_score, p2_score	AD, 40	Alcaraz again has the advantage having won point 9.
	point_victor	1	Alcaraz wins point 10 (in row 11) which means he has won the game (has score 2 more points now).
Row 12	game_no	2	This is now the first point of game 2.
	p1_games	1	Alcaraz won game 1.

例 3：第 51 行

比赛的第 51 分是 "破发点"--非发球方（回发球方）有机会赢得比赛的得分点。

Row	Column(s)	Value(s)	Description
Row 51	<i>p1_score,</i> <i>p2_score</i>	40, 30	The score is 40 – 30 meaning player 1 (Alcaraz) is ahead.
	<i>server</i>	2	Jarry (player 2) is serving.
	<i>p1_break_pt</i>	1	If Alcaraz wins the point he will win the game; since he is not serving this is a "break point."
	<i>point victor</i>	1	Alcaraz wins the point (and therefore the game).
	<i>p1 break pt won</i>	1	Alcaraz won the game and was not serving on the point.

在 COMAP 竞赛中使用大型语言模型和生成式人工智能工具

这项政策是由大型语言模型（LLM）和生成式人工智能辅助技术的兴起所推动的。技术的兴起。该政策旨在为团队、顾问和评委提供更高的透明度和指导、和评委提供更高的透明度和指导。本政策适用于学生工作的各个方面，从模型的研究和开发（包括代码创建模型（包括代码创建）到书面报告的所有方面。由于这些新兴技术发展迅速，COMAP 将适时完善本政策。

团队必须对其使用的所有 AI 工具保持公开和诚实。团队和其提交的透明度越高，其工作就越有可能被他人充分信任、赏识并正确使用。这些披露有助于理解知识作品的发展，并对贡献进行适当的认可。如果没有对 AI 工具的作用进行公开和清晰的引用和参考，那么可能更容易将具有疑问的部分和工作识别为剽窃并被取消资格。

解决问题并不要求使用 AI 工具，尽管允许其负责任的使用。COMAP 承认 LLMs 和生成式 AI 作为提高团队准备提交的效率工具的价值；例如，在生成初步结构的想法时，或者在总结、改写、语言润色等方面。在模型开发的许多任务中，人类创造力和团队合作至关重要，而依赖 AI 工具则存在风险。因此，在诸如模型选择和构建、协助代码创建、解释数据和模型结果以及得出科学结论等任务时，我们建议在使用这些技术时保持谨慎。人工智能工具会带来风险。因此，我们建议在以下任务中使用这些技术时要谨慎。

重要的是要注意，LLMs 和生成式 AI 存在局限性，无法替代人类的创造力和批判性思维。COMAP 建议团队在选择使用 LLMs 时要意识到这些风险：

- 客观性：LLMs 生成的文本中可能包含先前发布的带有种族主义、性别歧视或其他偏见的内容，而且一些重要观点可能没有得到充分体现。

- **准确性：**LLMs 可能会产生“幻觉”，即生成虚假内容，特别是在超出其领域范围或处理复杂或模糊主题时。它们可能生成在语言上合理但在科学上不合理的内容，可能错误地获取事实，并且已经显示它们能够生成并不存在的引用。某些 LLMs 只在特定日期之前发布的内容上进行训练，因此呈现的图片可能不完整。

- **上下文理解：**LLMs 无法将人类理解应用于文本的背景，特别是在处理惯用表达、讽刺、幽默或隐喻语言时。这可能导致生成的内容中出现错误或误解。

- **训练数据：**LLMs 需要大量高质量的训练数据才能实现最佳性能。然而，在某些领域或语言中，这样的数据可能不容易获得，从而限制了任何输出的实用性。

团队指南

要求参赛队：

- 在报告中清晰地指明使用 LLMs 或其他 AI 工具的情况，包括使用的具体模型以及用途。请使用内文引用和参考文献部分。此外，在 25 页解决方案之后附上 AI 使用报告（下文有描述）。
- 验证语言模型生成的内容和引用的准确性、有效性和适当性，并纠正任何错误或不一致之处。
- 提供引文和参考文献，按照提供的指导进行。仔细检查引文，确保其准确并得到正确引用。
- 考虑到 LLMs 可能会复制其他来源的大量文本，要警觉可能存在的抄袭情况。检查原始来源，确保你没有抄袭他人的工作。

当 COMAP 发现可能是使用未公开的工具准备的提交时，将采取适当的措施。

引用和参考文献的指导方针：

认真考虑如何记录和引用团队可能选择使用的任何工具。越来越多的样式指南开始纳入关于引用和参考 AI 工具的政策。在你的 25 页解决方案中，使用内文引用，并在参考文献部分列出所有使用的 AI 工具。

无论团队选择是否使用 AI 工具，主要解决方案报告仍然受到 25 页的限制。

如果团队选择利用 AI，在报告结束后添加一个名为“Report on Use of AI”的新部分。这个新部分没有页数限制，且不计入 25 页解决方案的篇幅。

示例（这不是详尽无遗的，根据你的情况进行适应）：

AI 使用报告

1. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4)

Query1: <insert the exact wording you input into the AI tool>

Output: <insert the complete output from the AI tool>

2. OpenAI Ernie (Nov 5, 2023 version, Ernie 4.0)

Query1: <insert the exact wording of any subsequent input into the AI tool>

Output: <insert the complete output from the second query>

3. Github CoPilot (Feb 3, 2024 version)

Query1: <insert the exact wording you input into the AI tool>

Output: <insert the complete output from the AI tool>

4. Google Bard (Feb 2, 2024 version)

Query: <insert the exact wording of your query>

Output: <insert the complete output from the AI tool>