(出國類別：開會)

# 參加美國舊金山
# 開放資料科學研討會
## (Open Data Science Conference, ODSC)

| 服務機關 | 姓名職稱 |
| --- | --- |
| 數位發展部 | 曾華伶分析師 |

派赴國家：美國

出國期間：113年10月27日至11月1日

報告日期：114年01月14日

# 摘 要

2024年開放資料科學會議 (Open Data Science Conference, ODSC) 是全球頂尖的資料科學與人工智慧應用盛會之一，吸引來自世界各地的專家學者與相關領域的先驅，涵蓋人工智慧產業領袖、資料科學工程師以及技術創新者，該會議不僅提供最前沿的技術分享與實務經驗，還聚焦於生成式人工智慧、深度學習、大資料分析等領域的最新進展，透過多元化的主題演講、工作坊及交流活動，與會者得以深入了解技術的應用潛力及其對社會、產業和公共政策的深遠影響，是全球資料科學領域不可錯過的年度盛會。

該年度主題聚焦於生成式人工智慧 (Generative AI, GenAI)的應用，涵蓋生成合成資料、強化決策支援能力及複雜流程自動化，探討其在改變產業模式、推動創新與解決社會挑戰方面之潛力，本次會議深入討論於大型語言模型 (Large Language Model, LLM) 與大型多模態模型 (Large Multimodal Model, LMM) 的最新進展，及相關的倫理議題與資料治理挑戰，同時分享生成式人工智慧在多領域創新應用之實際案例與成功經驗。

# 目 次

## 壹、 目的

2024年開放資料科學會議 (ODSC) 是國際知名的資料科學與人工智慧盛會，匯集資料科學家、人工智慧專家及業界領袖，共同探討最新技術發展與實務應用，以主題演講、深度工作坊、實際案例分享及互動交流等多元形式，幫助與會者掌握生成式人工智慧 (GenAI) 的技術方向及應用潛力，並提供專家指導，強化其實務操作及技術應用能力，學習如何應用AI與資料科學技術推動創新、解決社會挑戰並創造新價值，為未來技術應用帶來重要啟發。

數位發展部為推動全國數位發展之主管機關，為深入了解生成式人工智慧應用於公共服務的國際趨勢與實務經驗，爰派員參加 2024年開放資料科學會議，汲取具實務經驗之機構與專家的專業洞見，包括AI訓練資料集規劃、資料治理及產業應用發展等領域，俾利我國推動開放資料政策及相關標準更加契合國際趨勢與生成式人工智慧應用需求。本次會議探討人工智慧於產業或服務等相關應用案例，為我國應用AI輔助公共事務提供實質參考，透過生成式人工智慧進一步強化公共服務創新能力及行政效率，促進資料驅動的政策發展，挖掘資料潛力，創造社會及公益價值。

## 貳、 過程

2024年於美國舊金山舉辦開放資料科學研討會，自10月28日開始展開至10月31日止為期4天研討會，並著重參與10月29日至10月31日有關生成式AI專題演講及應用案例介紹，參與議程摘陳如下：

### 一、 10月29日

(一)AI X Keynote: Searching for Meaning in the Age of AI

(二)Solution Showcase: Building an Open, Governed Lakehouse with Aapche Iceberg and Apache Polaris

(三)Synthetic Data for Anonymization, Efficiency and Insights

### 二、 10月30日

(一)ODSC Keynote: Infusing and Scaling Genterative AI into Business Differentiation

(二)ODSC Keynote: From AI to Data Processing: The Next Phase of Accelerated Computing

(三)Data Science is Dead

(四)Data Science in the Age of Generative AI

(五)Solution Showcase: Revolutionizing Data Management

### 三、 10月31日

(一)An Introduction to Data Contracts

(二)AI for Work: How GenAI Improve the Way We Work and Collabrate

(三)From Data Mess to Data Mesh- Data Management in the Age of Big Data and Gen AI

## 參、 人工智慧時代的意義探索（Searching for Meaning in the Age of AI）

Bryan McCann，you.com 的共同創辦人兼技術長，曾任 Salesforce 研究部的首席研究科學家，專注於深度學習及其在自然語言處理（National Language Processing, NLP）中的應用，在生成式人工智慧領域擁有豐富經驗，深入探討AI對未來十年的影響及應用前景，強調 AI技術在語言處理和語意探索方面的巨大潛力，並對未來社會的適應力與學習能力提出深刻見解，啟發大家共同思索人工智慧時代的深層意義。

### 一、「創造意義」及「尋找意義」

Bryan McCann 與 Richard Socher（現為 you.com 的共同創辦人）於 2013 年開始合作，將哲學理論融入計算模型，並致力於研究如何讓機器學習並「創造意義」。隨著研究的深入，逐步拓展至自然語言處理的多任務學習與統一模型領域，第一項研究是針對多任務學習的統一模型，探討如何以單一模型處理多種 AI 任務，而非為每個任務量身打造專屬神經網路架構的傳統方式，這一具前瞻性的理念為 AI 領域開啟了新的方向，並成為他加入 Salesforce 的契機，進而專注於語言技術的更深層次研究與創新。

### 二、 上下文的重要性

早期的自然語言研究主要是透過詞向量（Word Vectors）代表單詞的意義，但這種方法僅限於單一模型，知識僅能從一個模型轉移到另外一個模型上，應用程度相當有限。而Bryan首度嘗試將神經網路整體架構從一個任務轉移到另一個任務，例如，從翻譯任務轉移到分類、問題回答以及其他自然語言任務。

在實驗的過程中的一點意外小插曲，讓Bryan意識到模型的表現更多依賴於符號之間的對齊，而非詞向量本身的「意義」，引發他對對齊（alignment）和注意力（attention）機制的興趣，所謂的注意力概念係指透過對齊不同序列的符號，讓這些符號按照正確的統計模式排列，並嘗試將所有任務轉化為問題回答或對話形式，提出將上下文整合進系統的觀點，並發展出能夠在不同任務間遷移的模型，這一理念最終促成轉換器（Transformer）架構的誕生，並成為當今AI語言模型的基礎。

### 三、 重新定義於AI時代下的角色

隨著 AI 技術的迅速發展，許多過去被視為「人類獨有」的領域，例如寫詩、作詞作曲等，已逐漸被 AI 所觸及，當 AI 能夠完成任何人類可以做的事情時，我們需要重新思考未來世界的樣貌。許多工作建立在知識之上，當這些知識被完全自動化後，世界又將如何運作？ Bryan鼓勵大家跳脫現有框架，不再僅僅將自己的身份與工作緊緊綁定，而是轉變視角，將焦點放在我們能為未來創造的改變上，非僅在既有框架中追求安全感，主動塑造未來的思維，將是應對快速變化時代的關鍵。

### 四、 下一波技術浪潮的關鍵

在 AI 領域，我們已經從機器學習邁向深度學習的時代，不再需要手動定義特徵，而是讓電腦自主學習，這種轉變的核心在於設定明確的目標，而非直接指導電腦如何解決問題。未來技術浪潮的關鍵將是設計更高效的學習系統，使我們和組織能夠快速適應變化，並不斷優化自身，這種理念也延伸到組織的運作模式中—設計如神經網路般的組織架構，讓資訊能在各層級間自由流動，促進更高效的學習與靈活應變，從而實現持續的創新與成長。

### 五、 未來展望

Bryan鼓勵人們在 AI 時代重新思考個人角色與未來世界，從追求既有框架中的安全感，轉向為未來創造實質改變，提倡組織未來應轉型為如神經網路般的高效架構，促進組織內部學習、創新與靈活應變，以迎接快速變化的時代挑戰，並強調未來 10 年將是每個人影響世界的最佳時機，這種影響力將來自於持續學習與適應的動態過程，如果我們像 AI 一樣，不斷學習、進化並適應變化，就能在這個充滿可能性的時代中，真正塑造世界的未來。

## 肆、 合成資料於資料匿名化、提升效率與洞察發掘 (Synthetic Data for Anonymization, Efficiency and Insights)

合成資料是一項重要的資料技術創新，其核心在於透過演算法生成具真實資料特性的人工資料，廣泛應用於隱私保護、資料共享以及機器學習訓練等場景。本場專題

演講以「合成資料在匿名化、效率提升及洞察發掘中的應用」為題，深入探討其技術發展、實際應用及未來前景。

## 一、 合成資料的三種生成方式

首先介紹合成資料的三種生成方式：隨機生成、規則驅動生成及基於機器學習的生成方法、相關優劣勢及其適合的應用情境。

### (一)隨機生成資料（Random Data Generation）

隨機生成資料以簡單的隨機數生成器為基礎，能快速生成資料，例如在Google Sheets中使用隨機數函數，此方法計算成本低、生成速度快，適合用於資料結構測試或軟體開發中的初步測試場景，然而，其生成的資料缺乏真實資料的統計特性與結構性訊息，僅能滿足基本測試需求。此方法適用於需要快速生成大批量測試資料的軟體測試環境。

### (二)規則生成資料（Rule-Based Data Generation）

基於規則生成需用戶提供特定規則與邏輯，透過定義資料分佈與結構生成資料，此方法能生成符合特定場景或需求的資料，適合模擬較複雜的系統或業務邏輯，並提供更高的控制靈活性，便於生成契合業務需求的資料分佈。然而，這種方法需要用戶投入更多時間設計規則，特別是在處理高複雜度場景時可能變得困難且費時。基於規則生成資料常應用於需要精確模擬的軟體測試場景，例如模擬特定用戶行為或業務流程。

### (三)機器學習生成資料（Machine Learning-Based Data Generation）

基於機器學習生成資料則利用生成對抗網路（Generative Adversarial Network, GAN）、變分自編碼器（Variational Autoencoder, VAE）等模型生成合成資料，能高度模擬真實資料並保留其統計特性與結構特性。此方法可生成大量高品質資料，支持機器學習模型的訓練與分析，並能靈活調整資料分佈，例如在不平衡資料集中增加少數類別樣本；然而，該方法計算成本高，對硬體資源需求大，且需要以原始資料為基礎進行模型訓練。適合應用於醫療領域的患者病歷模擬、金融領域的詐欺檢測與風險評估，以及機器學習模型的訓練與模擬環境構建。

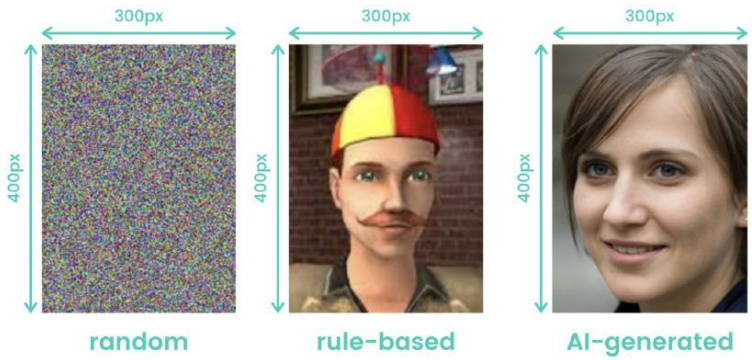


圖 1、三種不同樣態的合成資料

來源：MOSTLY AI 執行長 Tobias Hann 在 ODSC 2024 分享的簡報

## 二、 合成資料的發展與趨勢

在實務應用層面，合成資料展現出強大的靈活性與潛力，可被廣泛應用於下面各種領域，包含醫療領域，合成資料可以模擬患者病歷，用於AI模型的開發和優化，且避免直接使用真實資料而可能造成的隱私洩露問題；在金融領域，合成資料被應用於詐欺行為檢測，透過生成平衡的樣本資料集，提升模型預測的準確性；而在零售與供應鏈管理中，合成資料則可模擬多種假設場景，幫助企業進行風險預測與決策等場景，更有效的支援跨部門或跨域的資料共享及應用。合成資料的重要性日益提升，Gartner 預測 2030 年，合成資料將在 AI 模型中將會取代真實資料，成為主要資料來源，因此許多企業也投入越來越多資源於結構化及非結構化的合成資料研發及應用。

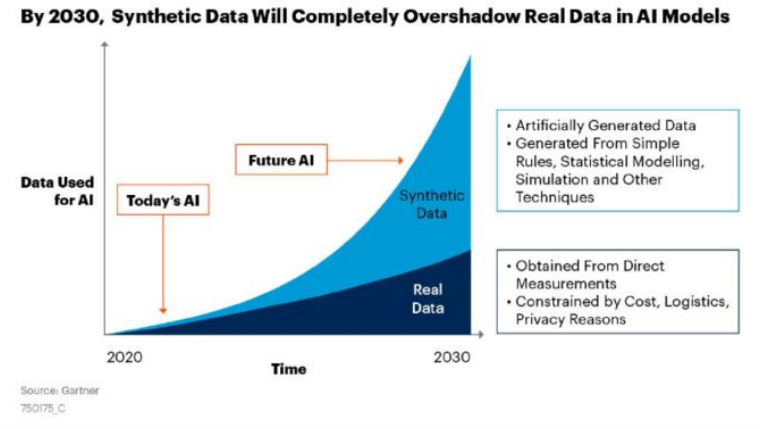


圖 2、合成資料應用於AI模型訓練趨勢圖

來源：Gartner

## 三、 巨量資料與傳統匿名化技術的挑戰

隨著資料收集規模的擴大及技術快速的進步，過去傳統的匿名化技術，如隨機化和資料遮罩，已被證實在多數情況下不足以防止重新識別風險，尤其當資料存在跨欄位關聯特徵時更為明顯，實現真正的匿名化變得愈來愈困難。為達到完全匿名化的目標，往往需捨棄更多重要欄位或資訊，這與充分挖掘和利用資料的需求形成強烈的矛盾，某部分人士甚至認為完全匿名化難以實現，建議放棄資料共享的構想，然而，這種觀點並非唯一的解決之道，也未全面考慮資料釋放的價值與應用的潛力。

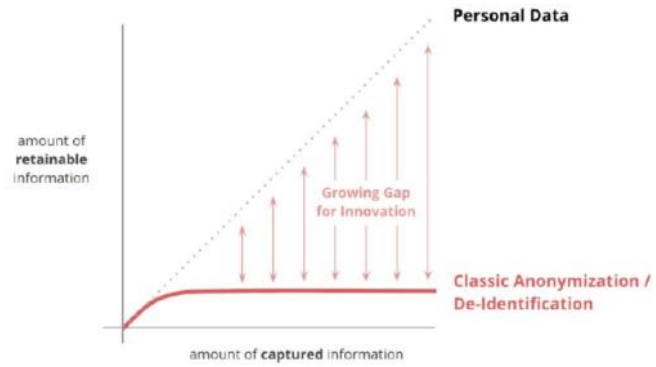


圖 3、巨量資料帶來的挑戰

來源：MOSTLY AI 執行長 Tobias Hann 在 ODSC 2024 分享的簡報

## 四、 合成資料的其他優勢

合成資料的一大特點是其強大的隱私保護能力，以人工生成的方式取代真實資料，並在保留統計特性的同時消除個人識別訊息，滿足如歐盟一般資料保護規則(General Data Protection Regulation, GDPR)和加州消費者隱私保護法(California Consumer Privacy Act, CCPA)等嚴格的隱私法規要求。除隱私保護外，合成資料還具有多項優勢：

(一)**填補缺失值：**在合成資料集中，智慧化的填補缺失值，比傳統技術（如簡單的平均值填補）能提供更好的效能。

(二)**調整資料分佈：**合成資料可以改變資料的分佈或變量特性，以模擬不同的場景，甚至可以生成真實資料中未出現的情境。

(三)**增強資料：**在金融詐欺檢測中，真實資料中詐欺案例的比例通常較少，導致資料集的不平衡，合成資料可以通過增強少數類別來平衡資料集，改善模型的學習效果。

| Method | Protection from re-identification risk | Feature statistics | Feature correlations | ML performance |
|---|---|---|---|---|
| Synthetic data | High | High | High | High |
| Randomization | Low | Medium | Low | Low |
| Permutation | Low | High | Very low | Very low |
| Generalization | Medium | Low | Low | Low |
| Pseudonymization | Very low | High | High | High |
| Data masking | Very low | Very low | Very low | Very low |

圖 4、合成資料與傳統匿名化技術的比較

來源：MOSTLY AI 執行長 Tobias Hann 在 ODSC 2024 分享的簡報

## 五、 三種生成方式實際應用與操作

(一)**隨機生成資料：**<u>Faker開源專案</u>

使用 Faker 生成隨機資料開源專案，以健康資料集為例，設定資料的變數（如 ID 和吸菸狀況），並快速生成 1000 條紀錄，雖然資料生成速度極快，幾乎沒有計算負擔，但隨機生成的資料缺乏統計意義，因此不適合應用於進一步分析。

(二)**規則生成資料：DataLens開源專案**

使用 DataLens 的專案，利用大型語言模型（Large Language Model, LLM）知識來生成資料的工具，此方法允許用戶以自然語言描述變數屬性，而不需要過多的技術定義。此方式在計算資源需求更高，但生成的資料能更有效的模擬真實場景。

(三)**機器學習生成資料：MOSTLY AI**

透過使用真實資料訓練模型後，利用訓練好的模型生成一個高品質的合成資料集，例如，利用模型來生成心臟病患者的資料，生成的資料保留了真實資料的統計特性且與原始資料極為相似，但完全沒有個人訊息，並成功模擬高血壓與心臟病之間的關聯性。此種方式的計算成本最高，但在資料分析和隱私保護方面的效果最佳。

**六、 結論**

本場演講深入探討合成資料的生成方式、實務應用及其未來前景，強調其在保護個人隱私、滿足法規要求及促進資料價值釋放上的關鍵作用，其中亦包含其潛在的應用風險，例如：涉及業務機密或極端類別資訊時，合成資料可能仍會反映出某些敏感特性，須謹慎處理模型偏差或極端情境的挑戰。

隨著資料需求快速增長及隱私法規日益嚴格，合成資料作為資料科學與人工智慧領域的重要創新工具，展現其在隱私保護、資料共享及應用潛力上的巨大優勢。使用者可根據應用需求及目的，選擇不同的合成方式，透過合成資料實現資料保護與應用的平衡，不僅能促進資料共享，更是AI應用的未來趨勢，過程中如何在隱私保護與資料共享之間取得最佳平衡，充分釋放資料的潛在價值，是值得深入探討課題。

## 伍、 資料科學已死（Data Science is Dead）

本場專題演講探討結構化資料的價值、現今資料科學的挑戰，以及透過人工智慧和圖形學習技術實現效率化建模的全新方法，講者分享如何從資料準備到模型開發，重新定義資料科學的未來發展。

## 一、 結構化資料的價值

結構化資料通常存儲在資料倉儲(Data Warehouse)中，包含多張具有關聯的資料表格，這些資料代表企業運作的藍圖，是用於業務分析及決策制定的關鍵資產。

過去在資料的應用上，僅利用這些資料進行歷史性的描述性分析，例如回顧銷售業績、分析流失客戶等問題，藉此回顧並總結過去的業務表現。然而，透過資料進行預測性分析才是成功留住顧客及增加銷售業績的關鍵成功因素，我們應該著重於分析預測性問題，例如「下個月的產品銷售如何？」或「哪些客戶即將流失？」等，與其事後補救，不如事前預防，這才是資料驅動策略及政策發展的核心價值所在。

## 二、 資料科學的現狀與挑戰

傳統資料科學的操作模式逐漸被視為過時且低效，正如「資料科學已死」的觀點所指出，當前的資料處理流程中，資料科學家往往耗費大量時間在資料收集、清理與特徵工程等繁瑣工作上，而非專注於模型的構建與優化。調查顯示，約80%的工作時間被用於這些前期準備工作，而真正用於高價值的模型開發與實驗的時間卻相對有限。

資料科學的現狀揭示傳統模式的瓶頸：資料準備環節不僅耗時，還需不斷處理來自多元來源的資料整合與聚合工作，甚至在模型完成後，由於特徵過期或特徵值不一致等問題，常需重新進行模型訓練與調整，這種冗長且低效的流程，不僅分散資源和精力，也限制分析成果的及時性與準確性。

在這樣的背景下，傳統資料科學的低效運作模式愈發顯得不合時宜，促使資料科學轉向更高效的技術與架構，透過新興技術自動化資料處理流程及高效的特徵工程，正在重塑資料科學的實踐方式，使資料科學家能更專注於價值創造和模型精進，從而推動資料應用的突破性發展。

## 三、 以深度學習與圖形學習重新定義資料科學

透過「關聯深度學習」（Relational Deep Learning），結合圖表示學習（Graph Representation Learning）和領域專用語言（Domain-Specific Language）的力量，簡化資料科學的工作流程。該方法的核心概念包括以下兩點：

(一)**預測查詢語言（Predictive Query Language，PQL）：** PQL 的核心在於將預測問題拆解為兩部分：標籤生成方式與標籤所屬的實體，以一個包含客戶、交易和產品的資料架構為例，PQL 可輕鬆定義如「預測未來 30 天內每位客戶的交易總額」這樣的查詢。系統會自動掃描資料並生成相應的訓練表格，取代傳統繁瑣的手動處理過程，使資料科學家的工作更加高效和便利。

(二)**將結構化資料視為圖形：** 多張結構化資料表格可被轉化為一張張異構時序圖（Heterogeneous Temporal Graph），每張表格的每一列則視為節點，並以其屬性（如產品描述、交易價格等）描述，節點之間則是以主外鍵相互連結。這種圖形結構為資料學習提供了一個自然的框架，允許利用圖形神經網路（Graph Neural Network, GNN）或圖形轉換等技術，直接在原始資料上進行學習，而無需傳統繁瑣的資料準備或特徵工程。

透過此方法不僅保留資料的屬性，並利用圖形化結構的關係特性，讓神經網路能自動學習資料的特性，擺脫過去資料準備需手動特徵工程的限制，為資料科學領域帶來了全新的發展與可能性。

## 四、 圖形學習技術的應用與優勢

透過圖形學習技術，資料科學家能夠直接處理多表格資料，無須將資料扁平化處理或手動設計特徵，神經網路可以學習資料表間的關聯，並自動發現具有預測性的特徵，極大簡化建模的過程。例如，該技術能夠：

(一)預測客戶流失的模式（如購物交易模式、店鋪訪問次數等）

(二)發現產品品質問題（如可能導致客戶投訴的問題）

(三)從屬性中學習特徵，並從大量資料中捕捉協作訊號

透過圖形學習技術，提高資料探索和模型建構的效率，其最大的特點是不需要手動特徵工程，只需重新訓練模型，神經網路便能自動識別新的行為模式，極大地減少資料科學家的工作量。同時，該技術將神經網路應用於結構化資料，直接從表格中挖掘資料的深層意義並進行精確預測，全面提升資料應用效能及決策能力。

## 五、 實務應用與案例分享

為實現這項技術，Jure Leskovec表示自身公司開發一個專為資料科學家設計的平臺(kumo)，以圖形化學習技術和自動化建模流程為基礎，有效簡化資料處理與特徵工程，減少資料準備的時間成本，幫助資料科學家能夠專注於模型建構與優化，加速模型開發與迭代，提升預測準確率，進一步將AI應用落實於業務決策，為企業帶來更大的效益與價值。

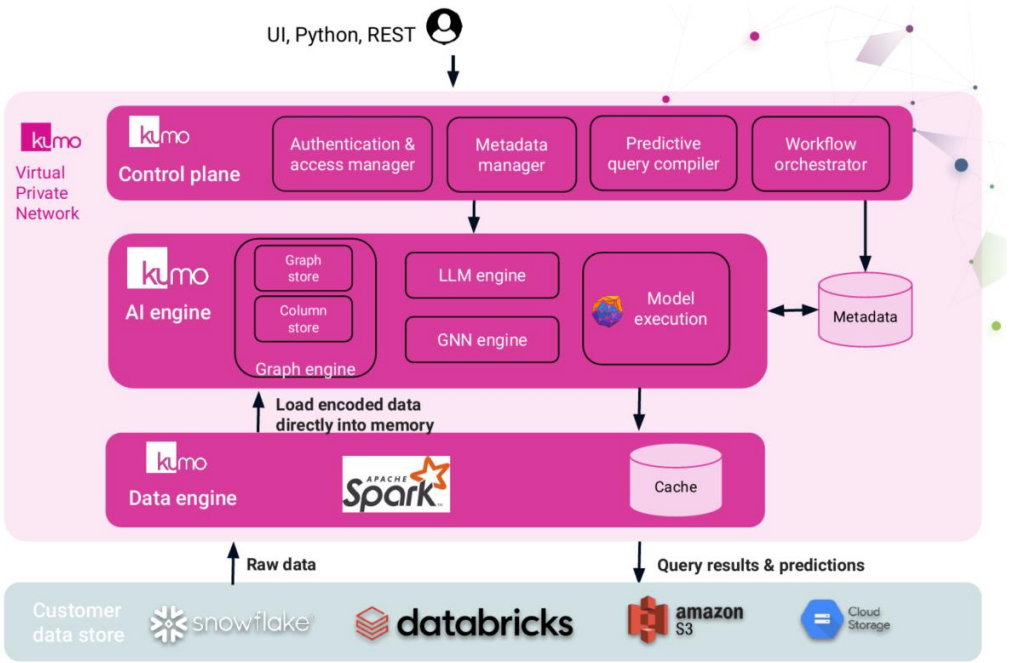


圖 5、Kumo平臺架構圖

來源：Kumo AI 首席科學家 Jure Leskovec 在 ODSC 2024 分享的簡報

此項技術在健康及金融等多個領域展現其應用潛力，涵蓋預測分析、網路詐欺、個人化推薦、銷售量優化等應用場景，並分享一些實際應用案例，展現該技術在多個領域的成功應用。例如，某公司使用11張資料表格，透過圖形學習技術快速生成模型，即可快速預測未來30天內的客戶流失和90天內的升級行為，縮短模型開發時間及提升企業決策支援能力。

### 六、 效益與未來展望

關聯深度學習技術不僅大幅減少資料準備和特徵工程的工作量，並幫助使用者快速迭代資料及預測問題，模型準確率亦從 0.7 提升至 0.93，從原本需要40天才能完成第一個模型的開發，縮短至僅需8天，這項技術的發展使資料科學家能將更多精力投入於建模與創新，充分釋放AI在產業應用中的價值，同時激發結構化資料的潛能，值得各界深入探討與推廣。

## 陸、 資料合約的介紹(An Introduction to Data Contracts)

本場專題演講聚焦於資料合約的核心價值、現代資料工程面臨的挑戰，以及如何透過資料合約實現從資料源頭到消費者的全流程品質保障，講者分享資料合約的實際應用案例與最佳實踐，並探討如何將其整合到持續整合與交付（CI/CD）流程中，重新定義資料管理與跨團隊合作的未來模式。

### 一、 資料科學及資料工程領域的現狀與挑戰

現今許多資料團隊面臨的共同問題包括資料湖的混亂、批次處理的低穩定性，以及無法信任資料品質等。Mark描述自己在職場上經常需要自行建立資料基礎設施來完成任務的經歷，這一現象反映出現有資料管理體系的缺失與不足。然而，這樣的困境不僅存在於個別公司，而是整個行業的普遍挑戰。

1980年代資料庫與建模的正規化至1990年代互聯網的興起，直到2000年代雲端基礎設施和大資料的應用，強調了現代資料堆疊（Modern Data Stack）崛起對行業的深遠影響。然而，隨著資料湖和ELT工具的廣泛使用，業界在快速採用雲端服務的同時，卻忽略了資料管理的規範性與一致性，導致資料品質下降、技術債務增加，許多公司儘管擁有龐大的資料儲備，卻因未能有效管理和應用而陷入混亂。

## 二、 資料合約的重要性

為解決上述問題，資料合約（Data Contracts）因應而生，成為重塑資料治理的一項關鍵工具，透過清晰的定義與自動化流程，為資料供應方和使用方建立了一套可操作的協作框架，促進業務邏輯與資料邏輯的一致性，解決資料品質問題、優化團隊間溝通，並有效防範資料變更帶來的風險，成為現代資料治理中不可或缺的基礎。

## 三、 資料合約的定義與核心原則

資料合約的核心在於針對資料及詮釋資料（metadata）設定明確的期望，並透過程式化的規範來確保這些期望的執行，例如，消費者可以定義其需求，如機器學習模型對資料穩定性的要求，並將這些需求以代碼形式保存為「合約規範」（contract spec）；資料生產者則需確認並遵守這些規範，以確保資料的一致性與可靠性。

資料合約是一種資料架構模式，旨在將軟體工程中的協作模式延伸至資料團隊，其運作方式類似於 API 的協議，透過程式化方式明確定義資料及詮釋資料的期望，並在持續整合/持續交付工作流程中實現自動化追蹤與執行，其核心包括以下幾點：

(一)**資料產品的合約要求：**每一個資料產品都必須附有合約，作為資料供應者與消費者之間的正式協議，明確彼此的需求及責任。

(二)**合約修改的規範化：**合約的任何修改都需遵循嚴格的流程，以確保變更的透明性及可追溯性，避免對下游資料應用造成意外影響。

(三)**詮釋資料的自動更新：**詮釋資料的變化需自動同步至資料目錄，確保相關團隊能夠即時獲取最新資訊，維持資料治理的效率及準確性。

(四)**資料品質變化的即時通知：**當資料品質發生變化時，需及時通知消費者，以便其快速應對，確保資料應用的穩定性及可靠性。

## 四、 資料合約組成要素及節點部署

資料合約是提升資料品質與管理效率的關鍵工具，透過明確的規範、監控與預防機制，保障資料在跨系統傳遞中的一致性與可靠性，以下整理構建資料合約的核心元素與最佳實踐，並說明其應部署的關鍵流程節點。

(一)**資料資產**：包括分析資料庫（Analytical Databases）、交易資料庫（Transactional Databases）、事件流資料（Event Streams），以及第一方在第三方平台（如 Salesforce）上的存放資料等。

(二)**合約定義**：包括資料合約規範、業務邏輯，以及管理詮釋資料的架構註冊表或資料目錄等。

(三)**檢測機制**：負責監控資料變更，例如資料變更捕捉（Change Data Capture）、資料流處理及靜態程式碼分析等技術。

(四)**預防機制**：確保在持續整合/持續交付流程中執行驗證，並輔以版本控制及監控功能。

資料合約應放置於哪些流程節點中？以下是一個典型的資料堆疊架構，包括交易資料庫、分析資料庫、資料湖、第三方資料及分析服務(圖7)，在資料從一個系統傳遞至另一個系統的每個交界點，皆可設置資料合約，但講者Mark建議盡量將資料合約部署在資料生成的最早階段，這樣能夠最大程度減少後續流程中潛在的問題，從源頭確保資料的品質及一致性。

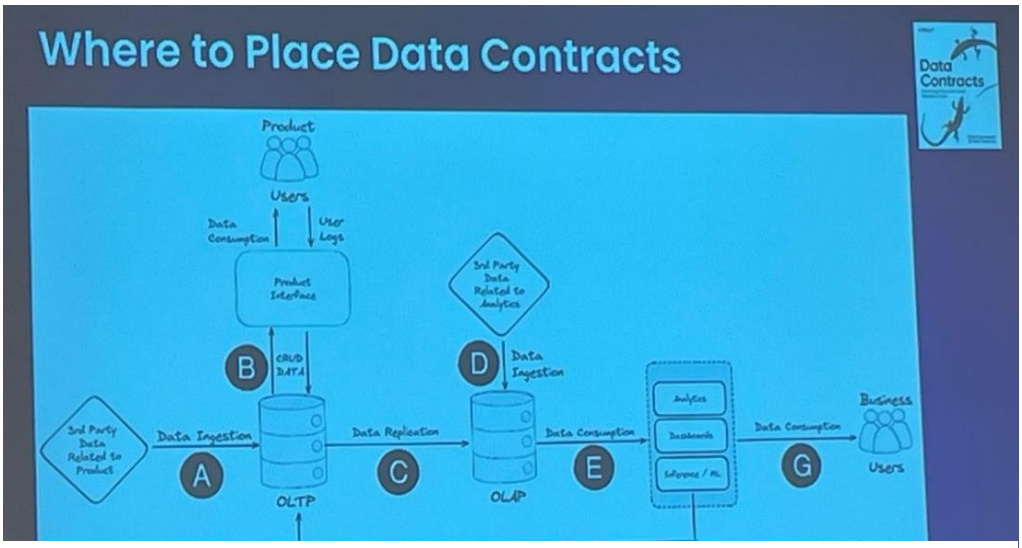


圖 6、資料合約放置節點

來源：拍攝於ODSC2024會場

## 五、 資料合約的運作流程

資料合約的核心在於將資料供應者與使用者的需求轉化為具體的約束條件，並以程式化方式定義為合約規範，當上游資料發生變更時，CI/CD 流程會檢查是否符合合約規範，若發現違規則會及時通知相關資料資產的負責人，根據變更的重要性，合約可以阻止合併請求或發送通知，維持開發效率的同時保障資料品質。

舉例來說，當一位開發人員在分支上更改資料庫結構，若該資料受資料合約約束，系統將提取更新後的資料庫結構，並與預期的架構進行比對，過程會參照架構註冊表進行核對，若發現不符之處，系統將阻止合併操作並生成合約違規警報，有效防止不符合要求的變更進入生產環境，此機制確保資料的一致性及穩定性，並促進供應者與使用者間的透明溝通與協作。

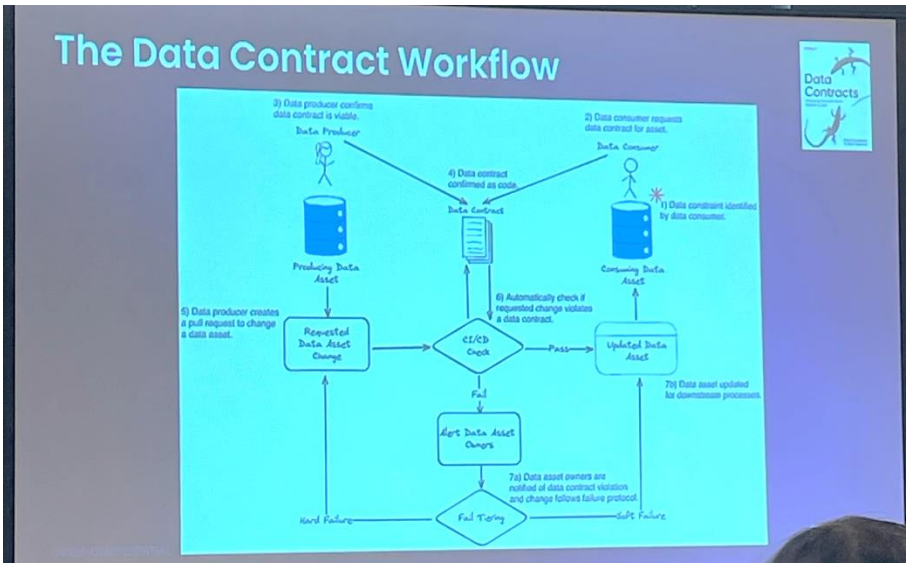


圖 7、資料合約運作流程

來源：拍攝於ODSC 2024會場

## 六、 資料合約與資料可觀察性(Data Observability)的差異

另一個常見疑問是，為何不直接使用資料可觀察性工具？資料合約與資料可觀察性並不衝突，而是互為補充的關係。資料合約專注於預防特定的資料品質問題，透過設定明確的約束條件來避免潛在風險；而資料可觀察性則聚焦於監測整體資料品質的趨勢，幫助團隊了解系統的健康狀況。

資料合約與持續整合/交付流程緊密結合，實現自動化的資料品質檢查；而資料可觀察性可以作為輔助工具，協助識別需要建立合約的關鍵資料區域。從功能上看，資料合約偏向於預防機制，能在問題發生前發出警報，避免不符合要求的

變更進入生產環境；資料可觀察性則更著重於事後的監測與分析，幫助團隊了解問題的原因與範圍。

例如，資料可觀察性好比手電筒，用於照亮整個資料系統的運行情況，而資料合約則像雷射光束，精確聚焦於關鍵的資料工作流程，確保最重要的問題得到優先處理，兩者的結合不僅提升資料品質的保障能力，還能優化整體資料管理流程。

## 七、 資料合約的實際應用案例

某線上招聘平台公司因資料品質問題導致廣告業務受到影響，無法保證廣告觀看資料的準確性，進而影響收入。為解決此問題，該公司引入資料合約，首先收集並分析所有品牌瀏覽事件的詮釋資料，清理重複資料，並與工程團隊合作制定明確的資料合約規範，並建立變更日誌與預警機制，確保任何資料變更都能及時通知相關人員。

透過這一系列步驟，該公司成功實現資料的標準化與高品質，同時顯著減少業務運作受到的負面影響，不僅如此，資料合約的引入還促進了跨團隊的合作文化轉變，將資料管理提升到更高的整合層次，使資料合約成為提升資料價值的重要工具。

## 八、 結論

資料合約是一種重要且實用的資料管理方法，致力於解決資料供應者與消費者之間的溝通障礙，並提升資料的準確性、一致性及品質，透過將需求轉化為具體的約束條件並程式化為合約規範，在資料生成、傳遞與應用每個環節提供明確的規範與保證，以循序漸進的方式，從整理詮釋資料開始到逐步實施合約，最終成功整合到持續集成工作流程中，為資料管理奠定了堅實的基礎。同時，資料合約不僅是一項技術解決方案，還促進了跨團隊的協作與文化轉型，讓資料管理從被動反應轉向主動規範化，成為驅動企業數位轉型及提升資料價值的重要策略之一。

## 柒、 從資料混亂到資料網格：大資料與生成式人工智慧時代的資料管理 (From Data Mess to Data Mesh-Data Management in the Age of Big Data and GenAI)

本場演講聚焦於資料網格（Data Mesh）的概念及其在大資料與生成式人工智慧（GenAI）時代的應用，幫助與會者深入了解資料網格的理論基礎與實踐策略。探討如何從過去中心化的資料管理架構，克服傳統資料湖與資料倉庫模式中的瓶頸問題，並深入解析資料網格的四大核心原則，透過結合實際案例，展示資料網格在多模式輸出、詮釋資料管理與生成式人工智慧應用中的實際價值，並強調資料治理、合規性及動態擴展能力對企業資料應用的重要性，啟發企業如何藉由資料網格實現高效協作與靈活應用，在快速變化的資料環境中達成更高效、更智慧化的資料管理與運營目標。

### 一、 資料管理的挑戰

隨著生成式人工智慧的興起，企業在處理資料時面臨諸多挑戰，這些挑戰包括資料來源多樣化、資料品質不穩定、以及資料合規性要求嚴苛等問題，例如，在金融與醫療產業中，必須遵守多項合規規範，確保資料使用符合隱私與授權要求，資料品質的好壞直接影響模型輸出的可靠性，不良資料可能導致錯誤的分析結果或模型表現下降，此外，企業在管理資料管道(pipeline)時，也面臨嵌入向量更新延遲、資料孤島以及跨部門協作不足等挑戰，這些問題大型企業資料擴展性尤其明顯。為解決這些核心難題，需要政策、技術與商業的多重驅動，從資料的生成、存儲、處理到應用的全流程進行重構，為組織的資料基礎設施建立更高效的管理架構。

### 二、 資料網格的概念介紹

資料網格是一種資料管理的新理念，旨在透過去中心化的方式，將資料管理的責任分散至各業務領域，此架構使每個領域能自主管理並共享自己的資料，有效解決傳統資料湖或資料倉庫模式下因中心化管理所導致的瓶頸問題，例如資料孤島與可擴展性限制，同時，資料網格促進了自主性、靈活性與可擴展性，進一步提升資料品質及組織內部的協作效率，此外，透過資料合約明確的約定來保障資料品質與一致性，為資料管理奠定堅實基礎。

資料網格的核心運作建立在以下四大原則之上：

(一)**資料即產品（Data as a Product）**：資料不僅是一種資源，更應被視為專為資料使用者設計的產品，每個資料產品應包含完整的資料內容、詳細的詮釋資料，以及清晰的存取與使用指南，同時強調對使用者的責任，確保資料具備高品質、可靠性和易用性，從而提升其價值與實用性。

(二)**領域導向的去中心化管理（Domain-Oriented Decentralized Data Ownership）**：資料管理責任被分散至各業務領域（Domain），讓每個領域能根據自身需求靈活管理和共享資料，並將資料管理與業務邏輯緊密結合，不僅更好地支持特定業務需求，還能減少對中心化團隊的依賴，避免因中心化管理帶來的瓶頸問題，進一步提升管理效率與靈活性。

(三)**自助式資料基礎設施（Self-Serve Data Infrastructure）**：提供技術支援平臺，包含自動化的資料管道和治理機制，降低技術門檻，讓各領域能輕鬆存取和操作資料，減少對中央團隊的依賴，使資料團隊能專注於提升資料價值。並促進資料的標準化和一致性。

(四)**聯邦式資料治理（Federated Computational Governance）**：在去中心化的基礎上，建立跨領域的統一治理框架，平衡自主性與整體性需求，並確保資料的合規性、隱私保護和安全性，透過自動化工具和政策維護資料的一致性與高品質。

透過四大原則，資料網格在保持去中心化靈活性的同時，兼顧了資料治理的標準化與一致性，為組織建立了一種高效且可持續運作的資料管理模式。

## 三、 資料網格的兩大核心支柱

Jörg介紹實踐資料網格技術的兩大核心支柱(圖9)，分別是資料產品容器化與控制平臺，這兩者共同構成資料網格成功運作的關鍵基礎建設：

(一)**資料產品容器化**：類似於軟體工程中的容器技術（如 Docker），資料產品被標準化為具有明確接口的單位，使得用戶能快速介接並使用，無需了解底層的複雜性。

(二)**控制平臺（Control Plane）：**控制平臺相當於編排工具（如 Kubernetes），用於管理多個資料產品的部署與使用，同時強化治理與合規性，確保資料使用符合公司或法律規範，此等架構使得團隊能夠快速回應新資料需求，並實現資料治理的靈活性與可追溯性。

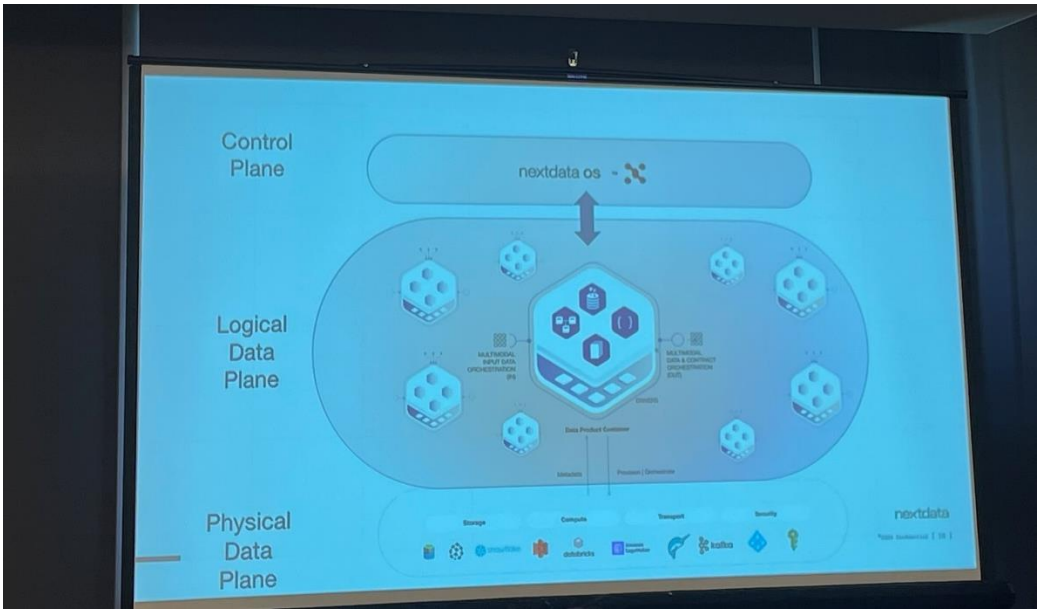


圖 8、資料網格核心基礎建設

來源：拍攝於ODSC 2024

## 四、 實際應用案例－音樂推薦系統

以基於生成式人工智慧的音樂推薦系統為例（圖10），說明資料網格在實際應用中的操作流程與優勢，該系統結合多個資料產品，如用戶播放清單、歌曲詮釋資料等等，生成個性化的推薦結果，透過資料網格技術：

(一)**整合多資料來源：**有效地將用戶聆聽記錄與歌曲詮釋資料等多個資料來源進行整合，形成系統化的輸入基礎。

(二)**動態嵌入向量更新：**資料網格提供策略化機制，決定何時更新嵌入向量或重新訓練模型，平衡即時性需求與系統效能，確保推薦結果的準確性。

(三)**多模式輸出：**同一組邏輯資訊可以以多種物理形式輸出，如用於機器學習的向量嵌入、支援商業智慧分析的 SQL 查詢或存儲為文件格式（如 Parquet 文件），以滿足不同應用場景與使用者需求。

(四)**靈活性與整合性：**資料網格架構簡化了新資料的整合與維護，允許新資料或模型的快速介接，減少技術門檻。

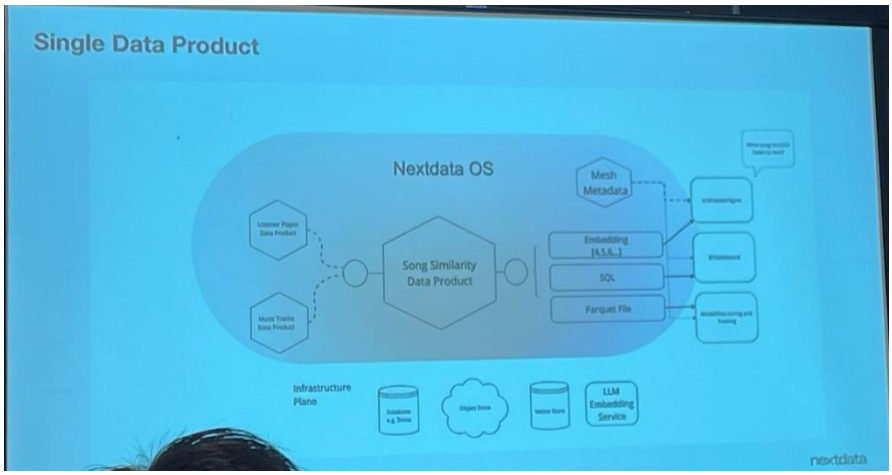


圖 9、音樂推薦系統示範案例

來源：拍攝於ODSC2024會場

整體而言，此示範案例充分展示了資料網格如何從資料取得、處理到應用，全方位提升系統的靈活性與效率，並為複雜的資料需求提供了一套可持續的解決方案。

## 五、 詮釋資料的重要性

詮釋資料在資料網格中扮演關鍵角色，展現其在多模式輸出與動態資料管理中的重要性。資料網格能夠將相同的邏輯資訊以多種實體形式（如嵌入向量、SQL 查詢、Parquet 文件）呈現，以滿足不同場景和使用者的需求，而詮釋資料則是實現這一過程的核心支柱，不僅協助使用者快速掌握資料的存取權限、內容範圍及應用方式，還能在動態環境中自動整合新增或更新的資料來源，從而確保團隊間的高效協作與資料應用的一致性。

在生成式人工智慧與檢索增強生成（Retrieval-Augmented Generation, RAG）中，即時查詢詮釋資料內容，例如資料的新鮮度、語義模型及使用歷史，企業能夠快速篩選出最適合的資料產品，並進行精準分析，有效避免因上下文過載而導致的效能瓶頸與增加查詢成本，亦大幅提升了資料管理與應用的效率。

### 六、 未來發展與建議

資料網格的動態性、靈活性與標準化特性，為企業實現智慧化資料管理和高效協作奠定了堅實基礎，資料管理的下一步應更加智慧化與自動化，包括：

(一)強化資料網格的詮釋資料管理，讓使用者能快速了解資料的使用情況與適用範圍。

(二)優化生成式人工智慧的應用，通過即時分析與推薦，幫助企業更高效地利用資料。

(三)進一步完善資料治理與合規機制，應對跨國資料流動帶來的挑戰。

## 捌、 開源資料目錄的興起：實現資料網格的新契機(The Rise of Open-Source Data Catalogs: A New Opportunity For Implementing Data Mesh)

這篇刊載於ODSC 2024的文章探討開源資料目錄的興起如何為實現資料網格架構帶來全新機遇，並說明其在大規模企業中的價值和實踐策略。

誠如上述章節所提，資料網格是一種以去中心化為核心的資料管理架構，適合大型或快速成長的中型企業，這些企業通常擁有數百甚至數千名跨部門和業務單位的資料使用者，透過資料網格，資料管理責任得以分散至各個領域，改善資料存取效率並減少對中心化資料團隊的依賴。

### 一、 開源目錄的湧現

資料目錄是一份詳細的清單，列出組織內所有的資料資產，可以幫助資料專業人員快速找到最適合任何分析或業務需求的資料，過去資料目錄工具因高昂成本成為實踐中的障礙，而開源資料目錄的興起則有效打破了這一侷限。開源資料目錄如 Apache Atlas 、LinkedIn DataHub、 Snowflake Polaris Catalog 及 Databricks Unity Catalog，為實現資料網格技術提供關鍵支持，這些工具提供集中式的詮釋資料儲存庫，是實現去中心化資料治理與高效資料共享的關鍵基礎，透過促進資料的發現、強化治理與合規性、支持動態更新與跨領域協，這些功能對於資料網格的核心原則，如資料作為產品和去中心化管理，具有重要意義。

開源資料目錄的靈活性和可擴展性讓企業能夠擺脫廠商綁定，並從全球開發者社群的支持中獲益，這些工具還提供自定義插件的開發能力，拓展資料網格在不同場景中的應用，然而，開源資料目錄的實施需要較高的技術門檻，特別是對於小型企業或多雲環境中的公司來說，可能面臨技術上的挑戰。

## 二、 開源目錄如何實現資料網格

開源資料目錄為資料網格提供了幾項關鍵功能，其中包括集中化的詮釋資料儲存庫，能夠跨去中心化的資料領域進行資料資產的發掘。此外，這些工具還有助於執行治理政策、追蹤資料血緣、確保資料品質，並透過單一控制層深入了解資料資產，無論其存放於何處，更支援基於角色的存取控制（Role-based Access Control, RBAC）和基於屬性的存取控制（Attribute-Based Access Control, ABAC），開源資料目錄在資料網格中的具備以下優勢：

(一)開源資料目錄開啟資料網格的新時代，因為這些工具既免費又靈活，並提供多種選擇，使用者可整合多個平臺和工具，而不會被特定供應商所綁定。

(二)開源工具由數百萬開發者與工程師組成的社群支持，他們樂於分享經驗並協助解決日常問題

(三)允許社群開發新的插件或擴展功能，解鎖更多的機會與新的應用場景。

(四)實現資料可觀察性，例如透過細粒度的存取控制、資料血緣追蹤，以及特定領域的操作日誌，幫助工程師與開發者全面掌握系統運作流程。不僅能更清晰地監控資料品質，還能高效地追蹤並解決潛在問題，為系統穩定性與資料治理奠定堅實基礎。

儘管這些工具擁有眾多優勢，資料目錄未必適用所有組織，構建此類架構並整合多個領域需要高度且專業技術能力，重要的是對可用選項進行深入研究，確保所使用的平臺與組織的現狀及未來目標相容。

## 三、 資料網格的實踐建議

組織中構建資料網格架構時，首先參與相關社群並參考其他組織如何使用這些工具的成功案例，儘管每個組織的實踐路徑不同，但一般需遵循以下步驟：

(一)**需求評估**：評估現有的資料基礎設施並定義組織的關鍵領域，探索每個領域的結構，並判斷組織是否具有足夠的規模和需求來進行重構。

(二)**試用不同的資料目錄**：比較不同開源資料目錄的功能和特性，確保選擇適合測試的工具。在安裝、配置和自訂工具時，可向社群尋求建議。

(三)**領域團隊(Domain Onboard)**：經比較後選擇最合適的工具，即可建立領域團隊，並分配資料的管理權限。

(四)**定義與實施治理政策**：與領域團隊負責人、法務及其他相關團隊合作，確定資料治理標準並制定相應政策。

(五)**整合現有資料基礎設施**：當已清楚定義領域並相關資料治理規範後，將資料目錄與資料來源、資料管道及商業智慧工具進行連接。

(六)**團隊培訓**：為該領域團隊和資料使用者提供培訓，確保每個團隊具備充分的知識來完全掌握自己的領域。

(七)**維護資料網格基礎設施**：當所有內容完成設置後，定期檢視政策並更新詮釋資料及實踐資料治理。

作者建議從單一領域小規模試行，確保相關流程與技術應用充分驗證後，再逐步擴展至更多領域，這不僅降低實施風險，也為組織提供靈活性及經驗累積的機會，為資料網格落地奠定基礎。

## 四、 結論

隨著機器學習與資料分析技術的快速發展，能否迅速存取資料將成為企業與組織成功的關鍵因素，在推進資料網格的同時，組織必須注重相關技能的培養，以應對資料基礎設施快速轉型所帶來的挑戰。開源資料目錄的出現為資料網格的實踐提供了新的可能性，透過優異的互操作性、多樣的整合選項以及可自訂的插件功能，各領域能更高效地構建其專屬的資料基礎設施，實現靈活性與擴展性。

然而，資料安全、隱私保護與合規性仍是推進過程中的首要考量，必須同步制定明確的保障措施。

資料目錄作為資料網格的核心工具，可以幫助組織清楚定義部門間的資料資產結構、存取權限與應用範圍，過這項技術應用，各機關可更加快速地尋找和共享所需資料，減少手動處理負擔，提升效率與透明度，同時，資料目錄支持聯邦式治理機制，確保分散式資料管理中的隱私保護和合規性，為資料的高效流通提供可靠保障。

## 玖、 生成式人工智慧如何改善我們的工作與協作方式(How Gen AI Improve the Way We Work and Collaborate)

本場專題演講聚焦於生成式人工智慧如何改善協作與提升工作效率，探討從數位化會議記錄到多方參與的未來 AI 模型應用，講者分享了生成式 AI 在提升記憶力與資訊分享效率上的突破，以及如何促進資訊透明化與組織內部的高效溝通。同時，演講還剖析生成式 AI 的實際應用場景與未來發展方向，強調透過隱私保護與文化轉型，實現個人、團隊與企業運作模式的全面革新。

### 一、 AI協作會議的優勢

AI 最大優勢的在於能夠捕捉並數位化所有對話內容，不論是通過 Google Meet、Microsoft Teams、Zoom，還是面對面的會議，AI 可以將所有內容轉錄下來進行摘要，讓所有內容變得更容易搜尋。透過數位化會議內容的記錄、摘要與查詢功能，人工智慧為個人及組織帶來了顯著的革新。例如，像 Otter.ai 這類工具能自動記錄會議內容，生成關鍵摘要，並提供即時查詢功能，顯著提升會議後的回顧效率與協作效果，無論身處何地，使用者皆可隨時搜尋過去的會議紀錄，針對特定問題快速獲得解答。

### 二、 解決資訊孤島與促進資訊流通

生成式人工智慧不僅解決人們記憶力有限的問題，也讓資訊能夠在組織內部自由流動，改善了資訊的分享與傳遞效率，讓更多人能更便利地獲取與運用關鍵資訊。Sam 強調，會議的核心目的是資訊的分享與溝通，但現實中會議資訊通常侷限於少數參與者之間，導致組織內的資訊孤島問題，為解決這一挑戰，他主張

重新定義溝通系統的運作模式，藉由 AI 技術模仿 Slack 的公開溝通機制，實現會議資訊透明化並提供高效的搜尋與查詢功能，減少因資訊不對稱導致的錯誤與重複溝通，並建議將90%的會議內容公開，僅對少數保密或敏感議題例外，這有助於促進整體資訊流動，並提升決策效率。

## 三、 從被動記錄者逐步演變為主動參與協作者

生成式人工智慧的角色正從被動的記錄者與資訊提供者，逐步演變為主動參與討論的協作者。現今AI 工具如 Otter 已能記錄和摘要會議內容，並提供便捷的查詢功能，大幅提升資訊整理與回溯效率；未來的 AI不僅能記錄和摘要會議內容，還能進一步成為討論中的協作者，例如，當團隊討論遇到瓶頸時主動介入，提供建議與解決方案，協助團隊快速凝聚共識。

未來的 AI 將具備深度分析與多方協作能力，分析說話者的背景、過去表達的意見及專業知識，精準的參與多方討論，同時，延伸至多方協作領域，結合每位成員的背景與專業知識，進一步助力團隊進行跨域決策與提升執行效率。

## 四、 數位孿生與個人化助理

Sam提出「數位孿生」（Digital Twin）的概念，這是一種基於AI的個人化助理，能整合用戶的知識與過往經驗，透過數位孿生，用戶可以在無法參與的會議中派出自己的數位化身代表出席，完成簡單結構化任務，甚至回答問題或參與討論。這樣的應用場景適合如初步篩選求職者、記錄銷售會談要點等重複性工作，讓使用者將精力集中在更具創造性和戰略性的核心業務上，為現代會議過載提供實際的解決方案。

## 五、 面臨的挑戰與未來發展

生成式人工智慧的推廣仍需克服隱私與安全性方面的挑戰，並解決人們對新技術的接受度問題，建立完善的安全機制以保護私密性資料是推動技術應用的基礎，並促進組織內部的文化轉型，使人們更願意接納並應用這些技術。人工智慧在未來將更進一步發展為具備情感智商的語音代理（Voice Agent），不僅需要實現多模態功能，例如能夠「聽」與「看」，還必須具備情感智能，能理解肢體語言與情緒，並根據情境做出適當回應和交流，能應對更複雜的溝通需求，為個人、團隊和整體企業的運作模式創造更多可能性。

生成式人工智慧為改善協作和提升工作效率帶來了前所未有的可能性，透過數位化記錄、摘要和查詢功能，AI 能解決記憶不足和資訊分享的難題，並使組織內部的資訊流動更高效透明，更是為個人與團隊開啟了全新的運作模式，然而，技術的推廣也伴隨著挑戰，尤其是在隱私、安全性及文化接受度方面，唯有通過完善的安全機制和內部文化轉型，才能充分發揮生成式 AI 的潛力，讓個人、團隊和企業在高效協作的基礎上，邁向更加智慧化與創新的未來。

# 壹拾、 心得

## 一、 生成式人工智慧的趨勢與應用

生成式人工智慧在當前資料科學與人工智慧領域中，已成為不容忽視的核心技術。本次ODSC會議深入探討GenAI的多樣化應用，從資料匿名化技術、強化決策支援，到複雜流程的自動化，展現其在推動產業創新中的巨大潛力。例如，在金融領域，生成式AI被應用於詐欺檢測，通過動態分析行為模式提高偵測精準度；在醫療領域，透過模擬患者資料生成安全且具代表性的資料集，支持研究和模型訓練；在供應鏈管理中，則透過即時資料整合和預測分析，優化物流與生產效率，這些應用不僅拓展GenAI的技術發展，也彰顯對社會與經濟結構的深刻影響。

另一個重要應用場景，是其對組織內部協作效率的顯著提升，會議中提到，應用AI工具數位化記錄與摘要會議內容，減少訊息遺失的可能性，未來可透過語音代理與數位孿生技術，實現更為靈活的協作方式，例如，AI可以作為數位孿生出席會議，完成基礎性任務或提供實時建議，使決策者能更專注於核心領域，此外，生成式AI能夠快速從多種資料來源中提取核心訊息，生成具備策略價值的分析報告，縮短跨部門協作的時間，實現更高效的知識分享與決策支持。

面對這波AI技術的浪潮，我國政府可透過生成式AI協助內部運作，以提升行政效率與流程優化，然而，在運用AI技術時，需謹慎處理倫理與隱私議題，確保科技應用符合法規與社會價值。同時，應制定全面策略，涵蓋技術創新與法規規範雙重層面，推動生成式AI在產業界的廣泛應用，透過技術支持、政策引導與跨部門協作，政府能為產業創新提供有力支撐，進一步強化臺灣在全球AI競爭力，實現產業升級與社會效益的最大化。

### (一)人工智慧時代下的資料治理與合作模式

本次會議的另一個亮點，是生成式人工智慧與資料治理結合的深入探討，隨著資料規模與複雜度的增加，傳統資料管理架構正逐漸走向瓶頸，而資料網格成為解決這些問題的重要框架，透過去中心化的管理方式，將資料管理責任下放至各業務領域（Domain），並通過詮釋資料實現多模式輸出（如嵌入向量、SQL查詢、Parquet文件）來滿足不同場景需求，不僅改善資料治理的靈活性與透明度，也為生成式AI的應用提供了高質量的資料基礎。

在資料共享與隱私保護的平衡中，合成資料技術成為突破的關鍵，它能在保留資料統計特性的同時滿足隱私法規（如GDPR與CCPA）的嚴格要求，降低資料共享的風險。從國際法規發展的經驗來看，隱私強化技術（PET）和資料治理機制的發展將成為全球趨勢。

我國政府的資料來源多元且結構複雜，涵蓋文本、統計資訊及政策文件等多種類型，這些資料雖具備豐富性與廣度，但在治理與應用上面臨跨機關整合、資料標準化以及資料價值釋放的挑戰。為應對這些問題，建議在資料治理規劃結合並落實資料網格與合成資料技術，並充分借鑒國際實踐案例，為資料共享與應用創造更多價值。

透過資料網格技術確保資料共享的安全性與合規性，從而解決資料孤島問題，提升跨機關協作效率；合成資料技術則為隱私保護與資料共享提供了創新解決方案，減少對敏感資料的直接依賴，降低資料濫用與隱私洩露的風險，結合政策引導與技術創新，不僅能有效提升資料管理與共享的效率，還可在全球資料驅動的經濟中建立臺灣競爭優勢。

### (二)人工智慧與未來工作模式的深遠影響

會議引導與會者對人工智慧時代工作模式轉型的深入思考，生成式AI所帶來的突破，不僅在於技術上的創新，更在於它對傳統工作流程的重塑。在知識密集型產業中，AI幫助組織快速識別並應對關鍵挑戰，實現從資料管理到戰略制定的全面升級，這不僅為組織的競爭力帶來全新的驅動力，也開啟了人類與人工智慧協作的嶄新篇章。生成式AI正重新定義資料治理與工作模式的邊界，透過技術創新與治理框架的結合，不僅能更高效地運用資料，還能實現更廣泛的協作與創新。

隨著AI技術的持續進步，政府應探索如何應用AI技術創新及解決問題，例如，透過快速地識別並解決公共治理的關鍵問題，並將 AI技術應用於智能客服、行政資料整合與分析等場景，促進跨部門的協作與創新，提升公共服務的效率與品質。同時，積極調整政府內部運作流程與工作模式，以更具彈性與靈活性應對科技發展帶來的挑戰與機遇，從而在人工智慧時代中實現永續發展，並為社會創造更大的公共價值。

## 二、 資料治理的重要性

除分享當前AI最新技術發展及應用，會中更強調現今在巨量資料與AI發展下資料治理的重要性，特別是在隱私法規日益嚴格的背景下，如何平衡隱私保護與資料價值釋放成為關鍵議題。從早期強調開放資料（Open Data）的推動，到更注重個人自主權的個人資料管理（MyData），再到如今以資料共享為核心的發展方向，臺灣在資料治理上的進展展現出逐步深化的趨勢，然而，這一過程面臨著高隱私要求與技術挑戰，這不僅對國家政策規劃提出更高要求，也對技術創新與實踐能力形成巨大考驗，會議中分享許多資料治理的工具、技術及方法，例如：

### (一) 資料合約與合成資料：推動資料治理與共享的基礎

資料合約作為資料治理的重要工具，透過清楚的規範框架促進跨組織協作，確保資料流動的透明性與可追溯性，不僅解決資料孤島與重複處理問題，也提升資料品質和應用規範性，借鑑國際實踐經驗，制定符合臺灣本地需求的資料合約標準，並結合政策支持與技術創新，持續深化及推動資料治理。此外，合成資料技術作為資料共享與隱私保護的創新解決方案，在醫療、金融等隱私需求高的領域展現強大潛力，能有效平衡隱私保護與資料價值釋放，促進資料流通與應用，政府應持續探索合成資料技術的研究與應用，突破跨機關資料共享瓶頸，助力智慧醫療、金融科技等領域的創新發展。

### (二)重構資料治理模式：資料網格與文化轉型並進

當前的資料治理已從傳統框架進化為涵蓋資料品質管理、動態更新能力與多模態輸出的綜合性模式，並透過去中心化的資料管理架構進一步重新定義資料治理，透過分散資料管理責任並結合聯邦式治理機制，提升資料治理的合規性與靈活性。會議中強調資料治理不僅是技術挑戰，更需文化與組織架構的轉

型，臺灣應積極參考國際成功案例，制定與國際接軌的資料治理政策，推動資料透明性、隱私保護及跨部門協作，同時，透過技術創新與文化轉型雙軌並進，打造具競爭力的資料治理體系，實現資料的價值最大化，為資料驅動經濟時代奠定厚實基礎。

本次會議的內容不僅為資料治理提供全新視角，也強調技術、法律與文化交融的重要性，在資料經濟快速變化的時代，唯有藉由資料治理框架的深化與技術創新，才能確保資料應用的合規性與價值最大化，為政府及社會創造更多可能性。

## 壹拾壹、 建議

我國政府的資料來源多元且結構複雜，這些資料大多來自不同機關單位依職務需求產出，如文本、統計資訊、政策文件等多種類型，這些資料雖然具備豐富性與廣度，但也面臨著管理與應用上的問題，特別是在資料標準化、跨機關整合以及資料價值釋放等方面，本次會議深入探討資料治理與生成式人工智慧技術的應用，提供以下幾點建議以供未來制定資料策略發展與技術應用之參考：

### 一、 強化資料管理策略

面對我國各政府機關分散式架構的資料管理需求，建議評估導入資料目錄工具，實現資料網格管理概念，並透過導入資料合約機制與引用領域資料標準，促進跨部門資料的有效流通，例如，工具如 Apache Atlas 和 DataHub，具有為分散式資料資產建立統一管理平臺的潛力，並透過自動化的詮釋資料管理功能，提供資料追蹤(Data Lineage)、版本控制和即時更新的能力，可為資料治理提供技術支援。

資料合約可作為資料目錄應用的補充工具，明確規範資料流通過程的一致性與安全性，例如，資料合約可規範資料的更新頻率、使用範圍及責任歸屬，以避免因資料版本不一致或定義模糊導致的合作障礙，考量資料合約尚在發展階段，建議政府應持續關注更多國際應用案例與經驗，結合我國需求，進行可行性評估，逐步測試並穩健推進相關應用，以確保實施效果與長期可持續性。

## 二、 跨機關資料共享與安全性

為促進跨機關資料整合與再利用，數位發展部鼓勵各機關開放更多高價值資料，加速資料共享與應用，為資料驅動的決策提供支持。同時，在技術層面，強調隱私強化技術的重要性，在資料共享與隱私保護之間取得平衡點。建議政府機關在推進相關應用時，可參考國際經驗與成功案例，評估生成式人工智慧與合成資料技術的可行性，逐步探索如何模擬真實資料的統計特性，生成安全且具代表性的資料集，以降低資料開放的風險，並促進醫療、金融、交通等多領域的資料整合與應用。

在新技術應用的推動過程中，應審慎考量倫理問題，雖然生成式人工智慧與合成資料技術能有效提升資料再利用與共享效率，但若缺乏明確的倫理框架與監督機制，可能帶來資料濫用、隱私洩露或偏見擴大的風險。最後，應提升社會對生成式AI及相關技術應用的認知，讓民眾理解技術帶來的效益與可能的風險，形成對話與監督的公共基礎，唯有在新技術的應用中充分考量倫理挑戰，才能確保技術發展與社會價值相輔相成，實現科技向善的長遠目標。

## 三、 提升資料品質，以利AI發展及應用

高品質的資料是AI模型訓練與部署的核心，常因資料不一致、遺漏值或重複資料而影響模型效能，建議政府機關在進行AI模型訓練過程，結合標準化流程並參考國際相關技術經驗，提升訓練資料品質，以增強模型訓練的準確性與效能。

在政策制定與分析方面，善用自然語言處理與機器學習技術，提升文本分析與資料運用的精準度，例如，運用生成式AI技術自動化地從海量資料中提取有價值的見解，生成政策報告或綜合性議題分析，幫助決策者快速掌握資訊核心，減少重複性勞動，提高政策研擬與執行效率。

為確保AI技術的合規應用與穩定發展，建議政府應持續關注國際間資料隱私保護與倫理議題。同時，透過公務人員的專業技能培訓，逐步提升生成式AI的實務應用能力，使其在資料價值挖掘與政策效能提升中發揮關鍵作用，為公共治理創造更大價值。

## 四、 逐步推進資料治理，強化政府部門資料科學素養與AI專業能力

ODSC專注於技術前沿與未來趨勢的探討，提供未來應用新技術的啟示，建議可逐步透過小範圍試行，驗證相關技術於各領域應用的可行性，為未來更大規模的應用奠定基礎。

此外，可透過舉辦工作坊與實務培訓，提升各機關在資料科學與人工智慧領域的素養，逐步培養內部專業能力。同時，建議派遣具技術專長的人員參與國際相關研討會與工作坊，汲取最新技術知識與成功經驗，促進國內技術與國際趨勢的接軌。

在經費與資源有限的情況下，逐步推行這些策略，能協助政府更高效地管理和運用多元的資料，進一步促進資料治理與價值釋放，助力智慧政府的發展，並在全球技術競爭中保持競爭力。

# 壹拾貳、 附錄

## 一、會議議程

### (一)10月29日

| ODSC West Talks - October 29, 2024 | | | |
|---|---|---|---|
| Time | Speaker | Organization | Session Title |
| 9:00 am - 9:25 am | Bryan McCann | Co-founder at You.com | Ai X KEYNOTE: Searching for Meaning in the Age of AI |
| 9:00 am - 9:25 am | Yohei Nakajima | General Partner at Untapped Capital | ODSC KEYNOTE: The Current and Future State of Autonomous Agents |
| 9:30 am - 10:00 am | Lin Qiao | CEO at Fireworks AI | Compound AI Systems and the Future of AI Integration |
| 9:35 am 10:05 am | GiorgioFrancesco Clauser | Data Scientist at Moneyfarm | Explainability Explained: From Beta Coefficients to SHAPly Values |
| 10:00 am - 10:30 am | Cody Coleman | CEO at Coactive AI | Tackling Socioeconomic Bias in Machine Learning |
| 10:15 am - 10:45 am | Lintang Sutawika | Researcher at EleutherAI | Challenges and Considerations in Language Model Evaluation |
| 11:00 am - 11:30 am | Shreya Rajpal | Founder at Guardrails AI | Managing the Volatility of AI Applications |
| 11:00 am - 11:30 am | Anoop Sinha | AI Research Scientist at Google | Larg e Model Quality and Evaluation |
| 11:35 am - 12:05 pm | Deepak Kanungo | Founder at Hedged Capital LLC | Probabilistic Machine Learning for Finance and Investing |
| 11:35 am - 12:05 pm | Sinan Ozdemir | CTO at Directly | Building High-Quality Domain-Specific Small Language Models |
| 12:10 pm - 12:40 pm | Matt Harrison | CEO at MetaSnake | Scaling Deep Learning Training in PyTorch |
| 12:10 pm - 12:40 pm | Jay Alammar | Director of Engineering at Cohere | Building with Llama 3.2 |
| 2:00 pm - 2:30 pm | Anjali Chourdia | Principal Researcher at Microsoft | Uncertainty Quantification: Approaches and Methods |

| | | | |
|---|---|---|---|
| 2:00 pm - 2:30 pm | Vino Duraisamy | Data Engineer at Snowflake | Mastering Web Data Acquisition Techniques |
| 2:35 pm - 3:05 pm | Geeta Shankar | Data Science Lead at Salesforce | Introduction to Containers for Data Science / Data Engineering |
| 2:35 pm - 3:05 pm | Dominic Bohan | Co-founder at StoryIQ | A Gentle Introduction to Vector Databases and Their Implementation |
| 3:10 pm - 3:40 pm | Alison Cossette | Data Scientist at Neo4j, Inc. | Enhancing AI Accuracy with Advanced Data Augmentation Techniques |
| 3:10 pm - 3:40 pm | Joep Kokkeler | Software Engineer at Dataworkz NL | Idiomatic Polars |
| 3:45 pm - 4:15 pm | Jeff Tao | CEO at TDengine | Machine Learning with CatBoost |
| 3:45 pm - 4:15 pm | Vino Duraisamy | Data Engineer at Snowflake | How to Make LLMs Fit Into Commodity Hardware Again: A Practical Guide |
| 4:20 pm - 4:50 pm | Geeta Shankar | Data Science Lead at Salesforce | Anomaly Detection for CRM Production Data |
| 4:20 pm - 4:50 pm | Alison Cossette | Data Scientist at Neo4j, Inc. | Bridging the Gap: Light Code Solutions to Uniting Social Science and Modern Knowledge Graphs |

(二)10月30日

| ODSC West Talks - October 30, 2024 | | | |
|---|---|---|---|
| **Time** | **Speaker** | **Organization** | **Session Title** |
| 9:00 am - 9:25 am | Dr. Alfred Spector | Visiting Scholar at MIT \| Senior Advisor at Blackstone | ODSC KEYNOTE: Beyond Models - Applying AI and Data Science Effectively |
| 9:00 am - 9:25 am | Dr. Ali Arsanjani | Director of Applied AI Engineering \| Head of AI Center of Excellence at Google Cloud | ODSC KEYNOTE: Infusing and Scaling Generative AI into Business Differentiation |
| 9:30 am - 9:55 am | Nick Becker | Product Leader in GPU-accelerated Data Science at NVIDIA | ODSC KEYNOTE: From AI to Data Processing: The Next Phase of Accelerated Computing |
| 9:35 am - 10:05 am | Ines Chami | Co-founder and Chief Scientist at Numbers Station AI | Operationalizing AI Agents in Data Analytics Workflows |
| 10:05 am - 10:30 am | Gary Marcus, PhD | Scientist, Best-selling Author, and Serial Entrepreneur | Virtual Keynote Fireside Chat |

| | | | |
|---|---|---|---|
| 10:30 am - 11:00 am | Chip Huyen | VP of AI & OSS at Voltron Data | TRACK KEYNOTE: From ML Engineering to AI Engineering |
| 10:25 am - 10:55 am | Dr. Einat Orr | Co-Founder & CEO at Treeverse | Don't Go Over the Deep End: Building an Effective OSS Management Layer for Your Data Lake |
| 10:25 am - 10:55 am | Jure Leskovec | Co-founder and Chief Scientist at Kumo.AI | Data Science is Dead |
| 11:00 am - 11:30 am | Tim Shi | Co-Founder at Cresta | Scaling GenAI at Cresta |
| 11:00 am - 11:30 am | Narendra Lakshamana Gowda/Narendra Lakshamana Gowda | Walmart Global Tech | The AI Advantage: Transforming Software Development for Operational Excellence |
| 11:00 am - 11:30 am | Jeremy Zhang, PhD | Head of Advanced Analytics at Gilead Sciences | Solution Showcase: Visualizing AI-Driven Clinical Trial Planning |
| 11:00 am - 11:30 am | Charles Frye, PhD | AI Engineer at Modal Labs | DIY LLMs: Rolling and LLM Inference Service from GPUs to olly |
| 11:00 am - 11:30 am | Jiwei Liu | Kaggle Grandmaster and Data Scientist at NVIDIA | Accelerating Data Agents with cuDF Pandas |
| 11:10 am - 11:40 am | Yan Liu, PhD | Professor at University of Southern California | Frontiers of Foundation Models for Time Series |
| 11:35 am - 12:05 pm | Benjamin Bengfort | CEO & Co-Founder at Rotational Labs | Privacy and Security in the Age of Generative AI |
| 11:35 am - 12:05 pm | Arzav Jain | Member of Technical Staff at OpenAI | Data Exfiltration Attacks in LLM Products |
| 11:35 am - 12:05 pm | Juan Guzman/ Jennifer Locke | Optimization Engineer at Gurobi/Manager - Technical Account Management at Gurobi | Solution Showcase: Mastering Complexity: Optimize your decision making for 500% ROI |
| 11:35 am - 12:05 pm | Paige Bailey | DevRel Lead, GenAI at Google | Data Science in the Age of Generative AI |
| 11:35 am - 12:05 pm | Rehgan Bleile | Co-Founder & CEO at AlignAI \| Founder at Women in Analytics (WIA) | Quantifying the Value of AI: Going Beyond Cost Savings |

| | | | |
|---|---|---|---|
| 11:50 am - 12:05 pm | Zipeng Fu | PhD Student at Stanford University | Towards Deployable Robot Learning Systems |
| 12:10 pm - 12:40 pm | Laurie Voss | VP, Developer Relations at LlamaIndex | RAG in 2024: Advancing to Agents |
| 12:10 pm - 12:40 pm | Mary Vue | VP of Partnerships and Marketing at Syncari | Solution Showcase: Revolutionizing Data Management |
| 12:10 pm - 12:40 pm | Cal Al-Dhubaib | Head of AI and Data Science at Further | Intro to AI Auditing - a guide for executives to navigate risk |
| 12:10 pm - 12:40 pm | Jun He | Staff Software Engineer at Netflix | Efficient Incremental Processing with Apache Iceberg and Netflix Maestro |
| 12:10 pm - 12:40 pm | Rafael Levi | Senior Solutions Architecture Expert at Bright Data | Powering AI with Endless Data from the Web |
| 12:35 pm - 1:05 pm | Dustin Dorsey | Principal Data Architect at Onix | Dimensional Data Modeling in the Modern Era |
| 2:00 pm - 2:30 pm | Ramesh Periyathambi | Distinguished Engineer at eBay | Data Pipeline for Retrieval Augmented Generation and Model Training at eBay |
| 2:00 pm - 2:30 pm | Eno Reyes | CTO at Factory | Building Reliable Coding Agents |
| 2:00 pm - 2:30 pm | Yashas Roy | Learning Solutions Architect at DataCamp | Solution Showcase: Building an AI-Ready Workforce at Scale with Data Camp |
| 2:00 pm - 2:30 pm | Jisheng Wang, PhD | VP of Engineering & Head of AI/ML at Traceable AI | Securing the Horizon: Safeguarding Large Language Models |
| 2:00 pm - 2:30 pm | Steven Hillion | Head of Data and AI at Astronomer | Building AI Applications with Airflow |
| 2:00 pm - 3:00 pm | Dr. Shelby Heinecke | Senior AI Research Manager at Salesforce | New Frontiers in GenAI: From Multi-Agent Systems to On-Device LLMs and Beyond |
| 2:00 pm - 3:00 pm | Micaela Kaplan | Machine Learning Evangelist at HumanSignal | Evaluating LLM Evaluators with a Human in the Loop |
| 2:00 pm - 3:00 pm | Swagata Ashwani | Principal Data Scientist/Data Science Lead at Boomi | Mind Mechanics: From Spacey States to Transformer Tech |
| 2:00 pm - 3:00 pm | Maria Lupetini | CEO and Chief Data Scientist at InfoMaker Inc | Mix Integer Programming for Good, Not Just Profit |

| | | | |
|---|---|---|---|
| 2:00 pm - 3:00 pm | Kristy Hollingshead, PhD | Senior Data Science Lead at Further | From NLPer to AI Engineer: The Life of a Hipster Data Scientist |
| 2:25 pm - 2:55 pm | Pablo Vega-Behar | Director of Machine Learning Engineering at Fitch Group | RAG Pipelines Letting You Down? How The Fitch Group Handles High-Similarity, Frequently Updated Document Sets in Financial Services |
| 2:35 pm - 3:05 pm | Lior Gavish | CTO and Co-Founder | Monte Carlo | "Day Two" Problems: 5 Hidden Hurdles to GenAI Success and How to Overcome Them |
| 2:35 pm - 3:05 pm | Amanpreet Singh | CTO and Co-Founder at Contextual AI | RAG on the Edge |
| 2:35 pm - 3:05 pm | Yixin Tang | Engineer Manager at DoorDash | Safeguarding App Health and Consumer Experience with Metric-aware Rollouts |
| 2:35 pm - 3:05 pm | Dr. Helen Gu | Founder and CEO at InsightFinder Inc. | Solution Showcase: Product Launch: The Answer for AI Observability |
| 2:35 pm - 3:05 pm | Jaeman An | Co-founder & CEO at VESSL AI | Mastering Enterprise-Grade LLM Deployment: Overcoming Production Challenges |
| 3:30 pm - 4:00 pm | Ville Tuulos | Co-founder and CEO at Outerbounds | Solution Showcase: Scaling AI/ML with Outerbounds: How Metaflow Powers Our Open-Source Foundation |
| 3:30 pm - 4:00 pm | Sarah Johnson | Developer Relations at Coiled | Making Cloud Computing Boring Again: Lessons Learned from Deploying Billions of Python Functions |
| 3:30 pm - 4:00 pm | Megha Jhunjhunwala | Senior AI Engineer at Glean | Considerations for Building Enterprise-Safe AI |
| 4:05 pm - 4:35 pm | Bharath Ramsundar, PhD | CEO at Deep Forest Sciences | Can self-supervised models make a difference in drug discovery? |
| 4:05 pm - 4:35 pm | Bo Lei | Co-founder and CTO at Fleak | Extending Data Pipelines to Workflow Microservices |
| 4:05 pm - 4:35 pm | Soomin Chun | Software Engineer at FriendliAI | Solution Showcase: Accelerate AI Agents with FriendliAI’s GPU-optimized Service |
| 4:05 pm - 4:50 pm | Ben Wilde, Cameron Turner, Anne Dwane, Igor Tabber | Various (Georgian, Oxonian Ventures, Village Global, Cortical Ventures) | Identifying the Next Generation of AI Startups |
| 4:40 pm - 5:10 pm | Jeremy Miller | Product Manager, Academic AI Platform at Clarivate | Bitter lessons learned while building production-quality RAG systems for professional users of academic data |
| 4:40 pm - 5:10 pm | Dr. Helen Gu | Founder and CEO at InsightFinder Inc. | Unsupervised Machine Learning for Responsible and Robust AI Models |

| | | | |
|---|---|---|---|
| 4:40 pm - 5:10 pm | Nick Carrick & Clement Wong | Solutions Consultants at OpenText | Solution Showcase: Intelligence Aviator: Your Data’s New Best Friend |
| 4:40 pm - 5:10 pm | Geetha Anne | Sr Manager, Customer Solutions at PureStorage | Disaster Recovery Options Running Apache Kafka in Kubernetes |

(三)10月31日

| ODSC West Talks - October 31, 2024 | | | |
|---|---|---|---|
| **Time** | **Speaker** | **Organization** | **Session Title** |
| 9:00 am - 9:25 am | Sergey Levine, PhD | Associate Professor, Computer Science | UC Berkeley | ODSC KEYNOTE: Reinforcement Learning with Large Datasets: a Path to Resourceful Autonomous Agents |
| 9:30 am - 9:55 am | Benjamin Mann | Co-founder at Anthropic | ODSC KEYNOTE: Lessons Learned while Building Anthropic’s LLM, Claude |
| 9:35 am - 10:05 am | Kaxil Naik | Sr. Director of Engineering at Astronomer | Building and Deploying LLM applications with Apache Airflow |
| 10:00 am - 10:30 am | Jay Alammar | Director, Engineering Fellow (NLP) at Cohere | Large Language Models as Building Blocks |
| 10:15 am - 10:45 am | Brij Kishore Pandey | Principal Engineer at ADP | AI-Powered ETL Pipeline Orchestration: Multi-Agent Systems in the Era of Generative AI |
| 10:55 am - 11:25 am | Ryan Boyd | Co-founder at MotherDuck | Small Data Manifesto: Data infrastructure to build bigger with less |
| 11:00 am - 11:30 am | Sharon Zhou, PhD | CEO & Co-Founder at Lamini | TRACK KEYNOTE: Removing Hallucinations by 95% with Memory Tuning: A technical deep dive |
| 11:00 am - 11:30 am | Mark Freeman | Tech Lead, GTM Engineering at Gable | An Introduction to Data Contracts |
| 11:00 am - 11:30 am | Dhruv Nathawani | Applied Research Scientist at Gretel | Preserving Privacy in LLM Training: Transforming Sensitive Data into High-quality Synthetic Data |
| 11:35 am - 12:05 pm | Mabel Geronimo | Senior Solutions Engineer at GitHub | Gen AI in Software Development. What should you be looking for? |
| 11:35 am - 12:05 pm | Danica Fine | Staff Developer Advocate at Confluent | Brick-by-Brick: Exploring the Elements of Apache Kafka® |

| | | | |
|---|---|---|---|
| 11:35 am - 12:05 pm | Aarushi Kansal | AI Engineering at AutoGPT | From Paper to Production: Implementing Gen AI Research |
| 11:40 am - 12:40 pm | Christina Zhu | Developer Relations & Community Manager at Visier | Unlocking the Potential of People Analytics with Data |
| 11:40 am - 12:40 pm | Afrozy Ara | Co-founder & CEO at LuminaData | Designing Human-Centric AI Interfaces |
| 11:40 am - 12:40 pm | Shubham Goel | Senior Machine Learning Scientist at ZEFR | Labelling Sparse Data at Scale Using Semantic Search |
| 11:40 am - 12:40 pm | Arpita Vats | Senior AI Engineer at Linkedin | Leveraging LLMs for Next-Generation Recommender Systems |
| 11:40 am - 12:40 pm | Ajay Jain & Paras Jain | CTO and Co-founder \| CEO and Co-founder at Genmo | Mastering High-Fidelity AI Video Generation |
| 12:10 pm - 12:40 pm | Stephen Hood & Justine Tunney | Open Source AI Lead \| Lead Developer at Mozilla Llamafile | Llamafile: Democratizing Open Source AI |
| 12:10 pm - 12:40 pm | Kevin Noel | AI/ML Lead at Uzabase/Edge Japan/US | Practical Fine Tuning Strategies for Language Models and Large Language Models |
| 12:10 pm - 12:40 pm | Sam Liang, PhD | Co-founder and CEO at Otter.ai | AI for Work: How GenAI Improve the Way We Work and Collaborate |
| 12:10 pm - 12:40 pm | Jörg Schad | Head of Engineering \| Nextdata | From Data Mess to Data Mesh - Data Management in the Age of Big Data and Gen AI |
| 2:00 pm - 2:30 pm | Christine Long | Software Engineering Manager (Machine Learning Engineering Team in Reality Labs) at Meta | Wearable AI in Meta: On Device ML with Neural Interface System |
| 2:00 pm - 2:30 pm | Jay Sen | Director, Data Engineering at PayPal | Practical Data Mesh for Enterprises |
| 2:00 pm - 2:30 pm | Kamyar Azizzadenesheli, PhD | Research Staff at NVIDIA | Neural Operators: A new era of scientific computing |
| 2:00 pm - 2:30 pm | Fatih Nayebi | Vice President, Data & AI at ALDO Group | Scaling AI Initiatives in Retail |
| 2:35 pm - 3:05 pm | Vivek Natarajan | Research Lead at Google | How LLMs Might Help Scale World Class Healthcare to Everyone |
| 2:35 pm - 3:05 pm | Timothy Chan, PhD | Head of Data Science at Statsig | Beyond Simple A/B Testing: Advanced Experimentation Tactics |

| | | | |
|---|---|---|---|
| 2:35 pm - 3:05 pm | Varant Zanoyan | Software Engineer at Airbnb | Chronon - Open Source Data Platform for AI/ML |
| 2:35 pm - 3:05 pm | Rebekah Westerlind | Full-stack Software Engineer | Snorkel AI | Productionizing GenAI with AI Data Development |
| 3:10 pm - 3:40 pm | Arnav Garg | ML Team Lead at Predibase | The Future is Fine-tuned: Training and Serving Task-specific LLMs |
| 3:10 pm - 3:40 pm | Kevin Van Gundy | CEO at Hypermode | Prev. COO at Vercel | The Business of Open Source AI |
| 3:45 pm - 4:15 pm | Avinash Sooriyarachchi | Senior Applied AI Engineer at Mistral AI | Innovating with Multimodality and Reasoning with Mistral AI |