**Prima domanda: modelli sempre più ‘grandi’ che richiedono grande quantità di dati.**

**Risposta: a suo parere è una domanda grande e dice che dipende. Infatti per lui gli algoritmi di deep learning hanno ancora possibilità di crescita, ma questo è valido solo per certi task, una serie di problemi come quelli di computer vision che richiedono più risorse per la loro risoluzione mentre per altri problemi è necessario l’utilizzo di data set molto più piccoli.**

**Seconda domanda: sui foundation model.**

**Risposta: si parla di grandi modelli, trained su grandi set di dati e che possono essere tarati per specifiche applicazioni. Offrono tante promesse per ML applications ma sono difficili in termini di giustizia e privi di bias.**

**Terza domanda: usare foundation model per i video**

**Risposta: problema di scalabailità dato dal dover processare un grande volume di immagini per i video. Motivo per cui si è sviluppato prima e principalmente per gli NLP. Infine lui parla del fatto che il modo per cui si è sviluppato il DL negli ultimi anni era per soddisfare esigenza di aziende con tanti utenti e tanti grandi data sets. Questo ha portato ad una grande economia in questi software ma tale ricetta di scala non funziona per altre aziende.**

**Nella domanda successiva gli viene fatto notare che lui lavorò per aziende del primo tipo e lui risponde che quando ha lavorato per google brain gli era stato avvisato che forse si doveva focussare su l’iinovazione dell architettura. Infatti gli criticarono anche l’utilizzo di CUDA mentre sviluppava un paper su NeurIPS.**

**Successivamente fa notare come si è ritrovato nel parlare negli ultimi anni alle persone del movimento IA data centrico come in passato 10-15 anni prima parlava della scalabilità e del deep learning. Nell’ultimo anno si è piu concentrato su un mix di c’è qualche cosa di nuovo qua e sembra che si sia intrapreso la direzione sbagliata.**

**Settima domanda: come definisci il data-centric AI e se lo consideri un movimento**

**Risposta: è la disciplina di ingegnerizzare i dati che successivamente vanno usato per costruire un sistema di AI. Spiega come per un sistema come questo bisogna implementare un algoritmo e poi addestrarlo con un data set. Il paradigma nell’ultimo decennio era di scaricare il data set e migliorare il codice. Grazie a ciò le architetture delle reti neurali sono migliorate significativamente che il problema del codice è risolto. Per alcune è più produttivo tenere l architettura fissata e invece cercare modi per migliorare i dati. Bisogna portare avanti questa cosa fatta da pochi individui.**

**Ottava domanda: parli di istituzione che hanno solo pochi dati con cui lavorare. Come può data centric aiutarli?**

**Risposta: vision system che usano 350 milioni di immagini per la face recognition. Architetturre costruite per centinaia di milioni di immagini non funziona con solo 50. Ma risulta che se hai 50 esempi buini puoi costruire qualcosa di valore come un sistema di rilevazione di difetti. In molte industrie non esistono giganti data sets per cui bisogna spostarsi da Big Data a Good Data.**

**Nona domanda: paarli di un modello addestrato prima con milioni e poi solo con 50 buoni? O un modello differente che può basarsi su solo 50;**

**Risposta: In big data risposta comune: se il dato è rumoroso prendine tanti e l’algoritmo medierà tra tutti. Ma se riesci a sviluppare un tool che etichetta dove i dati sono inconsistenti e ti da una vera via per migliorare la consistenza dei dati questo è molto piu efficiente per creare sistemi ad alte performance. Ad esempio in un sistema con 10000 immagini e 30 consistenti bisognerebbe concentrarsi sul sottoinsieme che è inconsistente così che puoi relabel per essere piu consisyenti e migliorare le performance.**

**Nona domanda: questo aiuta con il bias dei dataset?**

**Risposta: molto molti ricercatori hanno parlato che dati con bias portano a sisytemi con bias. Data centric AI è un pezzo di soluzione, non l’intera soluzione, anche DataSheets for datasets sembrano essere importanti. Cambiare l’intera architettura per un sistema che ha buone prestazione per tutto il data set tranne che per un sottoinsieme è difficile ma se si possono engignerizzare i dati si possono risolvere meglio i problemi.**

**Decima domanda: parli di ingegnerizzare i dati, che intendi?**

**Risposta: data cleaning in AI è importante ma solitamente viene fatto manualmente. In cv si spotta manualmente il problema e poi risolverlo. Ci sono tool che permettono su un data set enorme di individuare i label rumorosi e permettere di concentrarti su di essi o concentrarti in quella classe tra 100 per cui serve collezionare più dati. Spesso collezionarne di piu aiuta ma a volte è troppo costoso. Esempio speech recognition riconosciuto il car noise in the background.**

**Undicesima domanda: synthetic data è spesso una buona soluzione?**

**Risposta: è un importante tool per il data centric AI, ha anche utilizzi che vanno oltre il preprocessing per aumentare il data set dei algoritmi di learning.**

**Dodicesima: questo può permetterti di usa il modello su più data sets?**

**Risposta: no veramente, i synthetic data possono aiutare a concentrarti su migliorare il sistema per migliorare le prestazioni per una certa classe. Puoi generare più dati per quella categoria. Utile ma ci sono altri tool usabili prima come data augmentation, improving labelling consistency o collect more data.**

**I find it really important to empower manufacturing customers to correct data, retrain, and update the model. So you’re saying that to make it scale, you have to empower customers to do a lot of the training and other work.**

**In the last decade, the biggest shift in AI was a shift to deep learning. I think it’s quite possible that in this decade the biggest shift will be to data-centric AI. With the maturity of today’s neural network architectures, I think for a lot of the practical applications the bottleneck will be whether we can efficiently get the data we need to develop systems that work well. The data-centric AI movement has tremendous energy and momentum across the whole community. I hope more researchers and developers will jump in and work on it.**