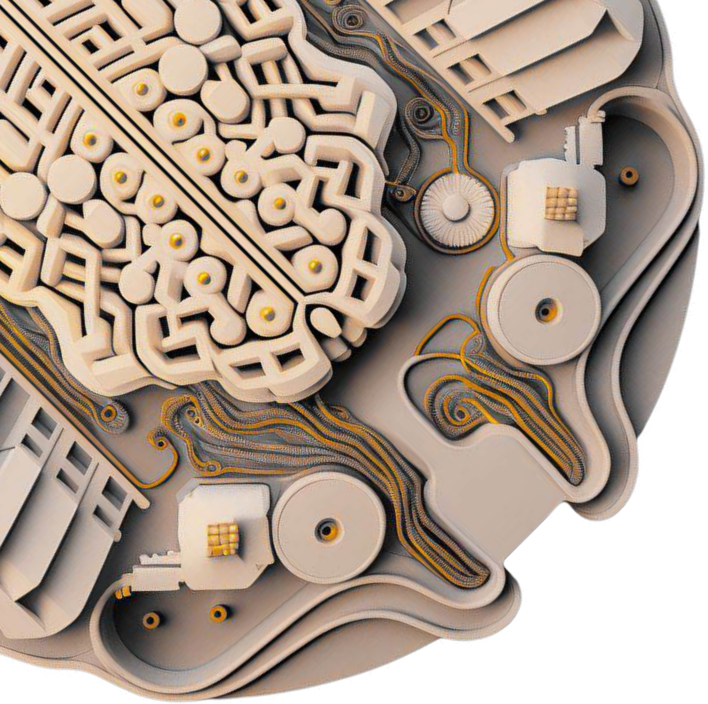


EQT MOTHERBRAIN

Decoding AI for Entrepreneurial Finance:
Applied Research Projects from EQT Motherbrain

Lele Cao

Ph.D., Principal AI Researcher

lele.cao@eqtpartners.com

EQT & Motherbrain

- EQT: Private Equity Fund
  - Global investment fund w €200Bn+ AUM
  - Buys, grows & sells companies
  - Venture Capital, Buyout, Infrastructure, etc.

- Motherbrain: Data & Machine Learning Platform
  - Support investment professionals globally
  - Merging and enriching data using AI (on GCP)

- Read more about us → https://motherbrain.ai

About me and this talk

About me

Lele Cao is a **Principal AI Research Scientist** in **EQT Motherbrain**. He holds a **Ph.D.** specialized in **AI & Robotics** from **Tsinghua University**.

He has published **over 30 academic papers/patents** on Applied Machine Learning, including in many renowned conferences and journals. Lele has 16 years of occupational experience from **EQT, Microsoft (King), Alibaba, Elisa, Ericsson and The University of Melbourne**.

Lele supervises Master Thesis students and serves as reviewers in many well-known AI conferences and journals.

About this talk

This presentation introduces the innovative applications of AI within a broad scope of entrepreneurial finance, highlighting the some research projects conducted by EQT Motherbrain. We will explore the following practical topics consecutively:

- Startup Success Prediction with Deep Learning;
- Sector Prediction with Large Language Model;
- Revenue Forecasting using classic and state-of-the-art methods;
- Deal Document Mining using Knowledge Graph.

References:

[1] Senane, Z.*, Cao, L.*, Buchner, V. L., Tashiro, Y., You, L., Herman, P., Nordahl, M., Tu, R., & von Ehrenheim, V. "Self-Supervised Learning of Time Series Representation via Diffusion Process and Imputation-Interpolation-Forecasting Mask." *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, under review, ACM, 2024.

[2] Cao, L., Buchner, V. L., Senane Z., & Yang, F. K. "Introducing GenCeption for Multimodal LLM Benchmarking: You May Bypass Annotations." *NAACL Workshop on Trustworthy Natural Language Processing (TrustNLP)*, to appear, ACL, 2024.

[3] Buchner, V. L.*, Cao, L.*, Kalo, J. C., & von Ehrenheim, V. "Prompt Tuned Embedding Classification for Industry Sector Allocation." In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, to appear, ACL, 2024.

[4] Cao, L., von Ehrenheim, V., Berghult, A., Henje, C., Stahl, R. A., Wandborg, J., ... & Ingelhag, H. "A Scalable and Adaptive System to Infer the Industry Sectors of Companies: Prompt+ Model Tuning of Generative Language Models." *IJCAI Workshop of Financial Technology and Natural Language Processing (FinNLP)*, pp. 55-62. ACL, 2023.

[5] Cao, L., von Ehrenheim, V., Granroth-Wilding, M., Stahl, R. A., McCornack, A., Catovic, A., & Cavalcanti Rocha, D.D. "CompanyKG: A Large-Scale Heterogeneous Graph for Company Similarity Quantification." *IEEE Transactions on Big Data*, to appear, IEEE, 2024.

[6] Cao, L., Halvardsson, G., McCornack, A., von Ehrenheim, V., & Herman, P. "Sourcing Investment Targets for Venture and Growth Capital Using Multivariate Time Series Transformer." *International Conference on Artificial Neural Networks (ICANN)*, under review, European Neural Network Society (ENNS), 2024.

[7] Cao, L., von Ehrenheim, V., Krakowski, S., Li, X., & Lutz, A. "Using Deep Learning to Find the Next Unicorn: A Practical Synthesis on Optimization Target, Feature Selection, Data Split and Evaluation Strategy." *IJCAI Workshop of Multimodal AI For Financial Forecasting (MuFFin)*, pp. 63-73. ACL, 2023.

[8] Cao, L., von Ehrenheim, V., Krakowski, S., Li, X., & Lutz, A. "Using Deep Learning to Find the Next Unicorn: A Practical Synthesis." *arXiv preprint arXiv:2210.14195,* 2022.

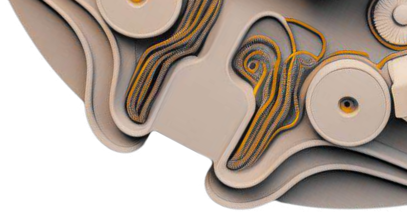
[9] Cao, L., Horn, S., von Ehrenheim, V., Anselmo Stahl, R., & Landgren, H. "Simulation-Informed Revenue Extrapolation with Confidence Estimate for Scaleup Companies Using Scarce Time-Series Data." In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 2954-2963. ACM, 2022.

[10] Cao, L., Larsson, E., von Ehrenheim, V., Cavalcanti Rocha, D.D., Martin, A., & Horn, S. "PAUSE: Positive and Annealed Unlabeled Sentence Embedding." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10096-10107. ACL, 2021.

Entrepreneurial finance :

The study of value and resource allocation, applied to **new ventures**. It addresses key questions which challenge all **entrepreneurs**: how much money can and should be raised; when should it be raised and from whom; what is a reasonable valuation of the startup; and how should funding contracts and exit decisions be structured.

The study and practice of financial **management and decision-making** in **new ventures, startups, and growth companies**. It deals with questions around sourcing, allocating, and managing capital within an entrepreneurial context. It differs from traditional corporate finance primarily due to the high uncertainty and limited historical performance data associated with new ventures.



Deal Process:
a general and super simplified view

Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

Due Diligence

- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?

Agenda

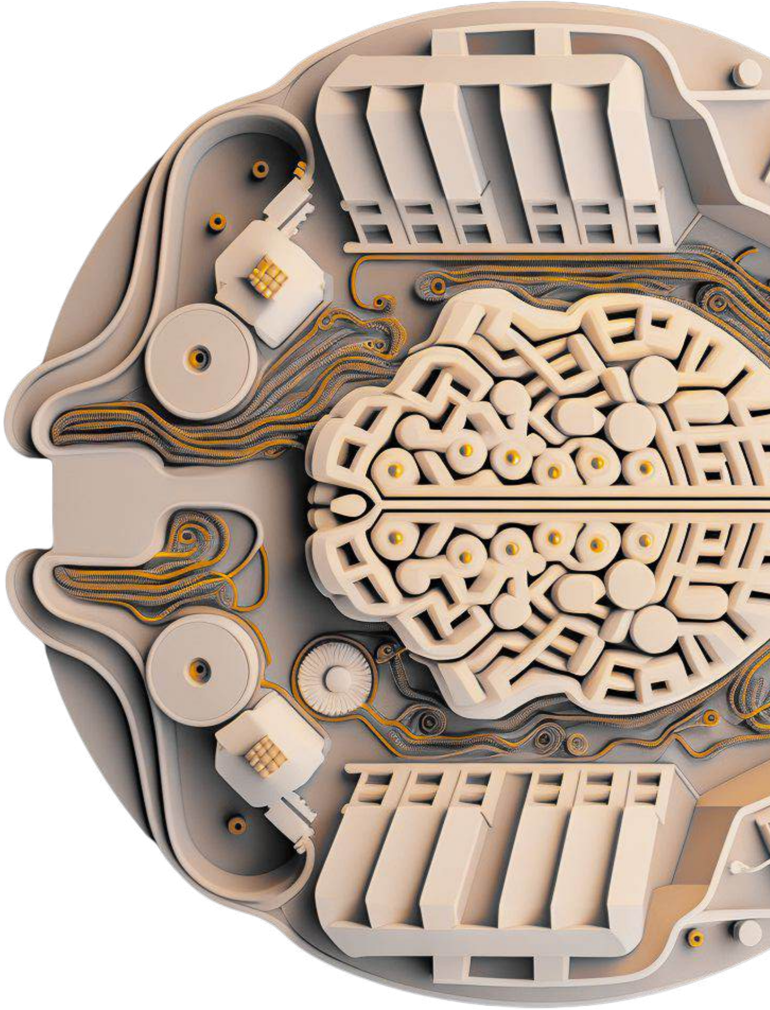
Success Prediction: Deep Learning

Sector Prediction: Large Language Model

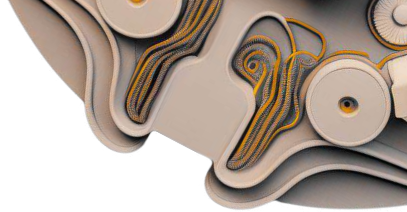
Revenue Forecasting: Classic and State-of-The-Art

Document Mining: Knowledge Graph

Summary



Success Prediction



Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

Due Diligence

- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?

Success is Rare

The successful startup is rare like "**finding a unicorn in the wild**". [7][8]

- On average, only around 60% of new companies stay in business for more than 3 years;
- Startups occupies only a few percentage of the firms' population, but tended to create about 60% of new jobs across most countries and sectors;
- Top 2% of VC funds receive 95% of the returns in the industry;
- VC traditionally has 10% success rate with startups.

Success Prediction: human

Define Success, what is your choice?

IPO or Acquired

Has > 100 employees

Founder obtains an MBA degree

Win an entrepreneurial competition

Get series A fund

Becomes profitable

Success Prediction: human

Is it a founders, investors, or policy maker's view?

IPO or Acquired

Investors

Has > 100 employees

policy maker

Founder obtains an MBA degree

Win an entrepreneurial competition

founders

Get series A fund

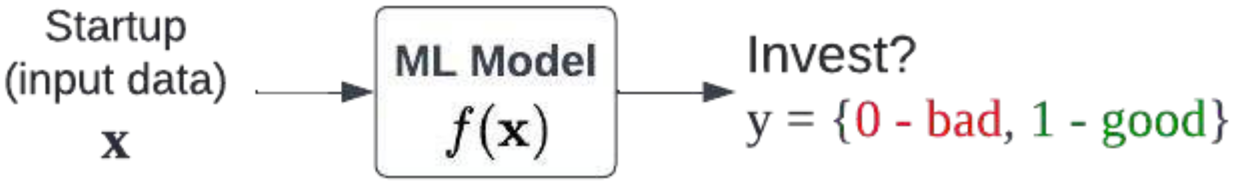
founders

Becomes profitable

founders

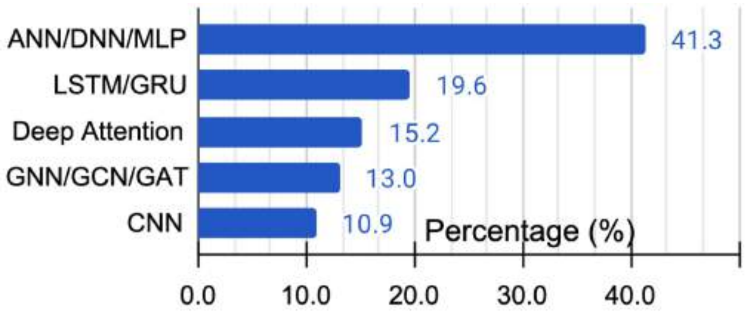
Success Prediction: model

- VC strives to identify and invest in unicorn startups as early as possible, hoping to gain a high return.
- This work is traditionally manual and empirical, making it inherently biased and hard to scale. [7][8]
- Recently, the rapid growth of data volume and variety is quickly ushering in **deep learning (DL)** as a potentially superior approach in this domain.



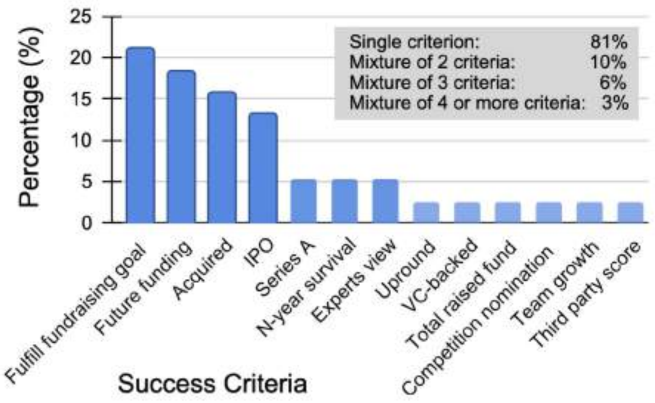
Success Prediction: deep learning models

- Over 40% of the surveyed papers adopt an ANN/DNN/MLP due to its wide applicability to many data types.[7][8]
- LSTM/GRU almost dominates the cases when time-series are used. [7][8]
- Deep attention and graph based models (GNN/GCN/GAT) have a rising trend of adoption due to increasing introduction of text and graph input. [7][8]
- Images and videos are relatively least used, leading to only around 10% adoption rate for CNN (convolutional NN). [7][8]



Optimization Target

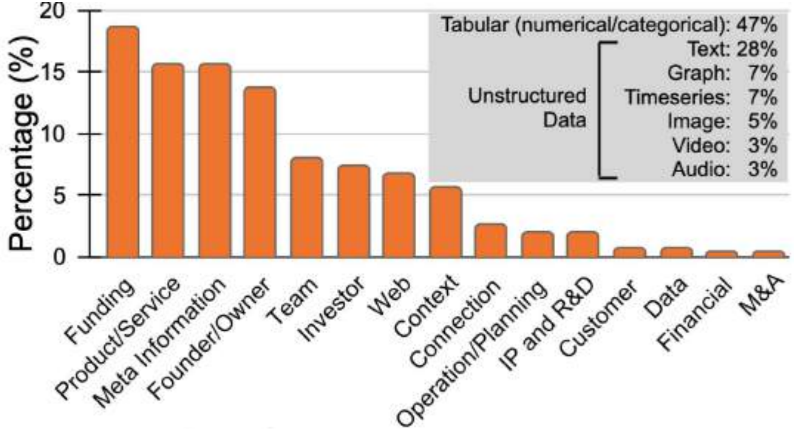
- So far there is no universally agreed definition of “true success”. – Find the Unicorn? [7][8]
- For investors, it is relatively straightforward: a profitable exit, often in the form of acquisition or IPO, which incur high ROI. [7][8]
- Short-term events like successive funding rounds have a higher adoption rate than longer-term acquisition/IPO. [7][8]
- Possible to combine multiple criteria.



Feature Selection: categories

- Most popular categories: Funding, Product, Founder, and the meta information of the company. [7][8]

- Noticeable Trends:
  - Single-modal→multi-modal
  - Structured(aggregated)→unstructured(raw)
  - Proprietary→paid→free
  - Intrinsic(independent)→extrinsic(contextual)
  - Average dataset size is 35,621 and keeps increasing



Feature Category

Feature Selection: the complete list

- Most popular categories: Funding, Product, Founder, and the meta information of the company.

- Noticeable Trends:
  - Single-modal→multi-modal
  - Structured(aggregated)→unstructured(raw)
  - Proprietary→paid→free
  - Intrinsic(independent)→extrinsic(contextual)
  - Average dataset size is 35,621 and keeps increasing

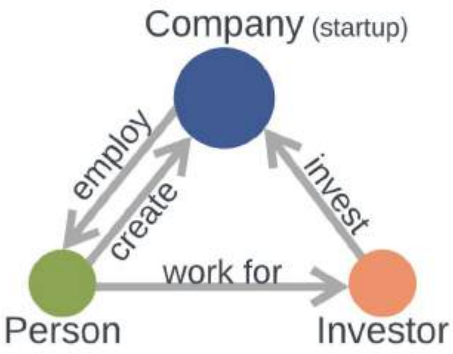
- **Complete list: see ref [7] and [8].**
- **Motherbarin's list: see ref [6].**

Category	Description of Common Features	The Reference[ref] of Example Work	#ref
Funding	Total number of funding rounds and amount raised	(Allu and Padmanabhuni, 2022; Yin et al., 2021; Horn, 2021; Stahl, 2021) ...	10
	Funding type (e.g., angel and series A/B/C)	(Dellermann et al., 2021; Stahl, 2021; Yeh and Chen, 2020; Sharchilev et al., 2018) ...	8
	Elapsed time since latest funding	(Garkavenko et al., 2022; Ang et al., 2022; Gastaud et al., 2019) ...	6
	Size and type of the latest funding	(Ang et al., 2022; Garkavenko et al., 2022; Ross et al., 2021; Gastaud et al., 2019)	4
	Size and type of seed funding	(Dellermann et al., 2021; Bai and Zhao, 2021; Lyu et al., 2021)	3
	Average per-round statistics	(Garkavenko et al., 2022; Ang et al., 2022; Garkavenko et al., 2021)	3
	Average time between consecutive rounds	(Ross et al., 2021; Garkavenko et al., 2021; Sharchilev et al., 2018)	3
	The raw time-series of funding rounds	(Chen et al., 2021; Stahl, 2021; Horn, 2021)	3
	Accumulated amount for different funding types	(Ross et al., 2021; Sharchilev et al., 2018)	2
	Total amount raised from VC	(Dellermann et al., 2021; Ross et al., 2021)	2
	Post-money valuation of rounds	(Garkavenko et al., 2021)	1
Product/ Service	Industry/sector/sub-sector	(Ang et al., 2022; Ghassemi et al., 2020; Sharchilev et al., 2018; Yu et al., 2018) ...	11
	Textual product description	(Chen et al., 2021; Kim et al., 2020; Cheng et al., 2019; Lee et al., 2018) ...	9
	Project specification on crowdfunding platforms	(Yeh and Chen, 2020; Cheng et al., 2019; Yu et al., 2018; Kim and Park, 2017) ...	7
	Image, video or audio of the product/service	(Tang et al., 2022; Shi et al., 2021; Kaminski and Hopp, 2020; Cheng et al., 2019) ...	5
	Time to market, novelty and differentiation	(Bai and Zhao, 2021; Dellermann et al., 2021; Sharchilev et al., 2018)	3
	Technology maturity, novelty and differentiation	(Allu and Padmanabhuni, 2022; Dellermann et al., 2021; Bai and Zhao, 2021)	3
	Customer focus (e.g., B2B/B2C/B2B2C)*	(Stahl, 2021; Dellermann et al., 2021)	2
	Quality, market penetration and traction	(Bai and Zhao, 2021)	1
	Business models[1] and scalability	(Dellermann et al., 2021)	1
	The number of product varieties	(Sharchilev et al., 2018)	1
	Textual product review and comment	(Lee et al., 2018)	1
Meta Info.	Founded date and geographical location	(Chen et al., 2021; Garkavenko et al., 2021; Sharchilev et al., 2018; Yu et al., 2018) ...	16
	Has Facebook/Linkedin/Twitter account	(Shi et al., 2021; Dellermann et al., 2021; Ross et al., 2021; Kim and Park, 2017) ...	5
	Domain name or homepage URL	(Ross et al., 2021; Srinivasan et al., 2020; Kim and Park, 2017)	3
	Company legal name and aliases	(Ross et al., 2021; Srinivasan et al., 2020)	2
	Office count and age	(Garkavenko et al., 2022; Sharchilev et al., 2018)	2
	Registered address, email and phone number	(Ross et al., 2021)	1
	Incubator or accelerator support	(Dellermann et al., 2021)	1
Founder Owner	Founding team size (number of co-founders)	(Garkavenko et al., 2021; Ross et al., 2021; Gastaud et al., 2019) ...	11
	Founders' (successful) founding/industry experience	(Bai and Zhao, 2021; Shi et al., 2021; Yeh and Chen, 2020; Srinivasan et al., 2020) ...	11
	Gender, ethnicity or education (uni., major and year)	(Lyu et al., 2021; Ross et al., 2021; Kaiser and Kuhn, 2020; Corea, 2019) ...	8
	Founder ID and score from 3rd-party data sources	(Shi et al., 2021; Yeh and Chen, 2020; Srinivasan et al., 2020; Sharchilev et al., 2018)	4
	Skill (e.g., leadership, sales, law, finance, marketing)	(Bai and Zhao, 2021; Ghassemi et al., 2020; Pasayat et al., 2020; Bento, 2018)	4
	Social capital‡	(Shi et al., 2021; Srinivasan et al., 2020)	2
	Founders' biography (text) and photo	(Srinivasan et al., 2020; Kim and Park, 2017)	2
	Founders' entrepreneurial vision and dedication	(Bai and Zhao, 2021; Dellermann et al., 2021)	2
Team	Team size of all or different functions	(Ang et al., 2022; Garkavenko et al., 2022; Ross et al., 2021; Kim et al., 2020) ...	6
	Completeness and capability of managers and board	(Garkavenko et al., 2021; Bai and Zhao, 2021; Sharchilev et al., 2018)	3
	The time-series of team size	(Stahl, 2021; Horn, 2021)	2
	Statistics of new hire or leavers	(Garkavenko et al., 2021; Sharchilev et al., 2018)	2
	Team composition (e.g., diversity and gender)	(Ross et al., 2021; Sharchilev et al., 2018)	2
	Educational degrees, vocational skill and experience	(Garkavenko et al., 2021; Ross et al., 2021)	2
	3rd-party team score and person ID	(Ghassemi et al., 2020; Sharchilev et al., 2018)	2
	Employees from renowned organizations	(Chen et al., 2021)	1
	Balance/empowerment/competence of the project team	(Yeh and Chen, 2020)	1
Investor	The number of total/distinct investors	(Ferrati et al., 2021; Chen et al., 2021; Kim et al., 2020; Sharchilev et al., 2018) ...	8
	Investor rank by reputation, experience and performance	(Stahl, 2021; Yin et al., 2021; Ferrati et al., 2021; Sharchilev et al., 2018)	4
	VC syndicate (e.g., advantage, diversity and centrality)	(Gastaud et al., 2019; Shin, 2019; Hochberg et al., 2007; Nahata, 2008)	4
	Share and involvement time of each investor	(Sharchilev et al., 2018)	1
Web	Rank/count/duration/bounce rate of website visit	(Garkavenko et al., 2022; Dellermann et al., 2021; Stahl, 2021) ...	5
	The count (**aggregated or timeseries**) of published news	(Yin et al., 2021; Garkavenko et al., 2021; Gastaud et al., 2019; Sharchilev et al., 2018)	4
	Topic or sentiment of **news/articles**	(Garkavenko et al., 2022; Kim et al., 2020; Sharchilev et al., 2018)	3
	Twitter statistics (e.g., followers, tweets and sentiment)	(Garkavenko et al., 2022, 2021; Dellermann et al., 2021)	3
	Count of web pages and domain names	(Garkavenko et al., 2022; Dellermann et al., 2021; Sharchilev et al., 2018)	3
Context	The number of direct competitors	(Allu and Padmanabhuni, 2022; Pasayat and Bhowmick, 2021; Xiang et al., 2012) ...	8
	Funding raised by competitors	(Stahl, 2021; Gastaud et al., 2019)	2
	Per-industry prosperity of the hosting geo-location	(Yin et al., 2021; Gastaud et al., 2019)	2
	Country/state/sector economy and financing env.	(Ross et al., 2021; Yin et al., 2021)	2
	Market/industry size and growth rate	(Allu and Padmanabhuni, 2022)	1
Connection	**The raw company-person-investor graph**	(Allu and Padmanabhuni, 2022; Pasayat and Bhowmick, 2021; Xiang et al., 2012)	3
	Pre-calculated graph features (e.g., betweenness)	(Bonaventura et al., 2020; Liang and Yuan, 2016; Hochberg et al., 2007)	3
Operation/ Planning	Planned revenue model	(Allu and Padmanabhuni, 2022; Dellermann et al., 2021; Bai and Zhao, 2021)	3
	Global exposure and internationalization	(Sharchilev et al., 2018)	1
	Market positioning and go-to-market strategy	(Bai and Zhao, 2021)	1
	Technological surveillance	(Allu and Padmanabhuni, 2022)	1
IP and/ R&D	The number, category and growth of patents	(Kinne and Lenz, 2021; Ferrati et al., 2021; Ross et al., 2021; Kim et al., 2020)	4
	University partnership	(Dellermann et al., 2021)	1
Customer	Customer satisfaction/loyalty	(Chen et al., 2021)	1
	The number of pilot customers	(Dellermann et al., 2021)	1
Financial	Revenue and/or turnover	(Kim et al., 2020; Cao et al., 2022a)	2
M&A	The number of acquisitions	(Ross et al., 2021)	1
Data	The total number of events/records	(Kim et al., 2020)	1

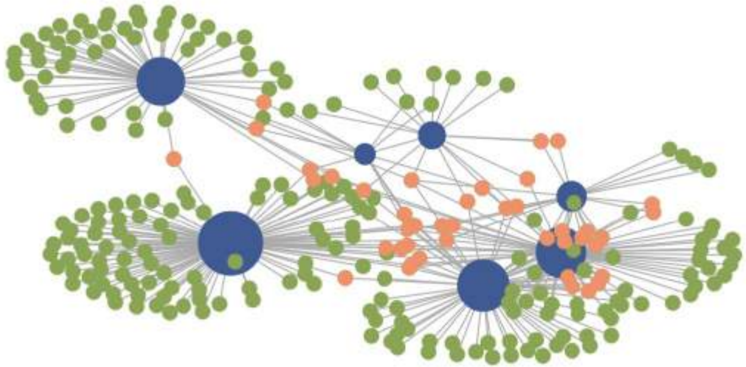
Feature Selection: Connection Category

This is a trending feature category!

- Usually extracted from a graph that encodes connections between different entities: startup, person and investor. [7][8]



(a) Entity connections.



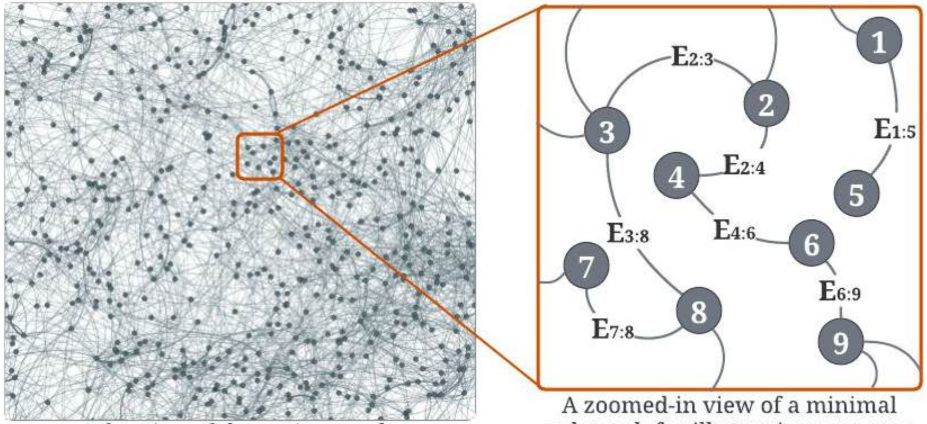
(b) An example graph.

Connection Feature: Motherbrain's example [5]

Relations: 15 relation types in 6 categories

- competitive landscape
- industry sector
- M&A transactions
- people's affiliation
- news/event engagement
- and product positioning

51.06 million weighted edges



A fraction of the entire graph

A zoomed-in view of a minimal subgraph for illustrative purposes

Edge weights: 15-dim vector, where the *i*-th dimension is the weight of the *i*-th edge type (ET*i*)

$E_{1:5}$ [0.0, 1.0, 0.0, 2.0, ... 5.1, 0.0, 1.6] $\in \mathbb{R}^{15}$

$E_{2:3}$ [0.0, 1.0, 1.0, 0.0, ... 0.0, 0.0, 2.3] $\in \mathbb{R}^{15}$

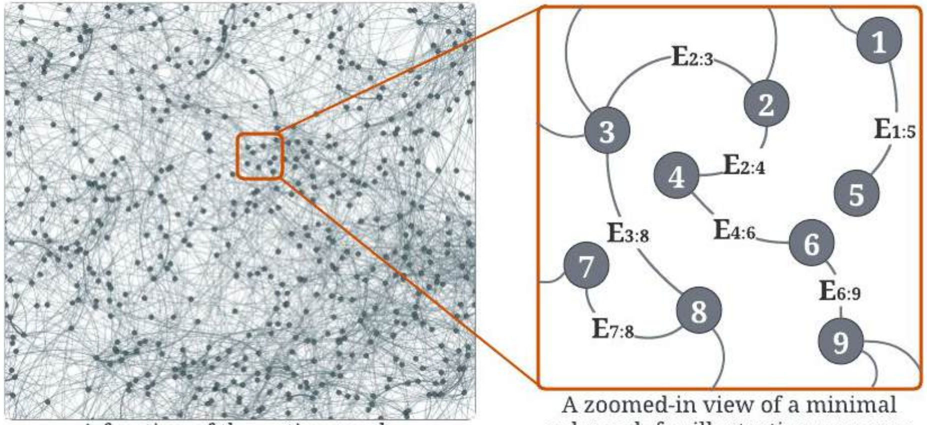
⋮

Connection Feature: Motherbrain's example [5]

Nodes: 1.17 million companies

Node feature: description/keywords embeddings:

- multilingual BERT
- ADA2 (GPT3.5)
- SimCSE https://doi.org/10.18653/v1/2021.emnlp-main.552
- PAUSE [10] https://doi.org/10.18653/v1/2021.emnlp-main.791



A fraction of the entire graph

A zoomed-in view of a minimal subgraph for illustrative purposes

Node features: company description embeddings from 4 language models

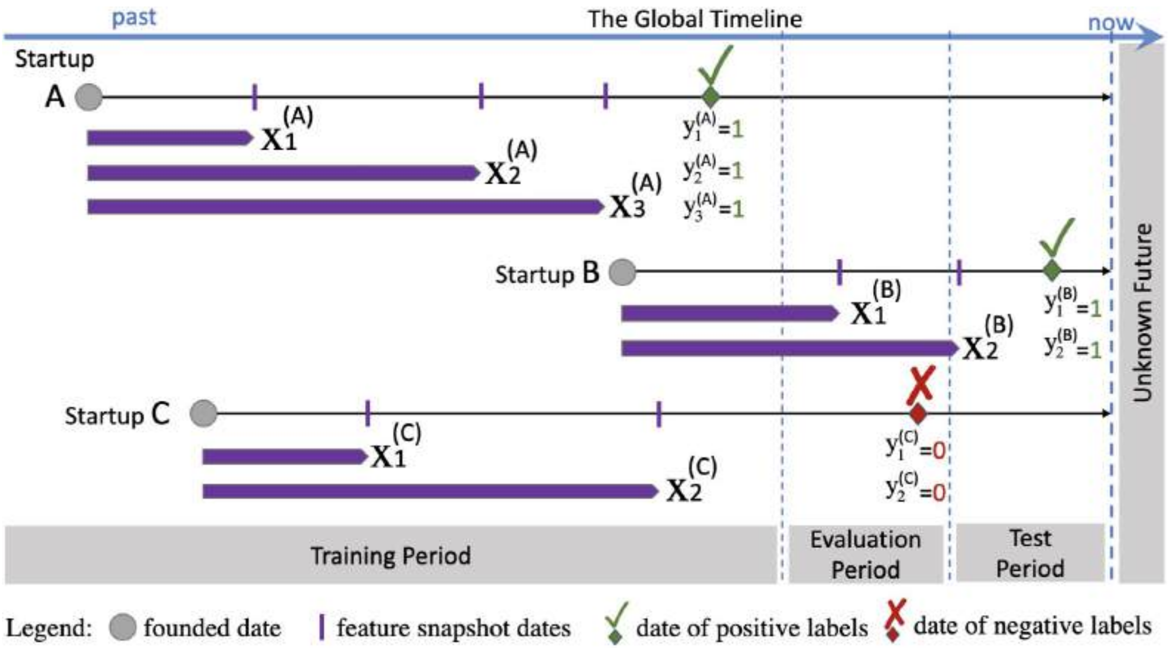
1
mSBERT: $[0.32, 0.01, \ldots -0.49] \in \mathbb{R}^{512}$
ADA2: $[0.05, 0.20, \ldots\ 0.35] \in \mathbb{R}^{1536}$
SimCSE: $[0.29, 0.16, \ldots -0.24] \in \mathbb{R}^{768}$
PAUSE: $[0.02, 0.73, \ldots\ 0.88] \in \mathbb{R}^{32}$

2
mSBERT: $[0.90, 0.53, \ldots\ 0.05] \in \mathbb{R}^{512}$
ADA2: $[0.44, -0.10, \ldots\ 0.35] \in \mathbb{R}^{1536}$
SimCSE: $[0.83, 0.01, \ldots\ 0.54] \in \mathbb{R}^{768}$
PAUSE: $[-0.22, 0.06, \ldots\ 0.90] \in \mathbb{R}^{32}$

⋮

Data Split

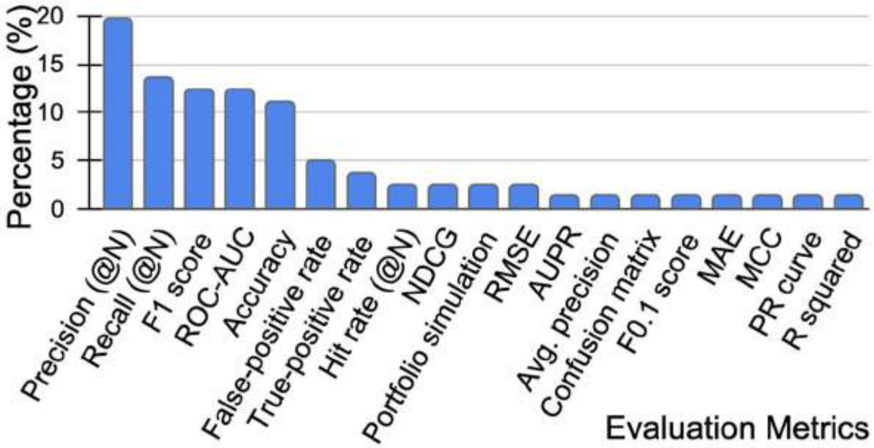
- Investor-centric view
- Two steps:
  - Augmentation
  - Split
- In this example:
  - Training: A (x3)
  - Evaluation: C (x2)
  - Test: B (x2)



Evaluation Strategy

- Realistically, human professionals are only able to assess a limited amount of startups.

- Evaluation metric should aim for high-precision (corresponding to high-certainty and low-recall) [7][8]

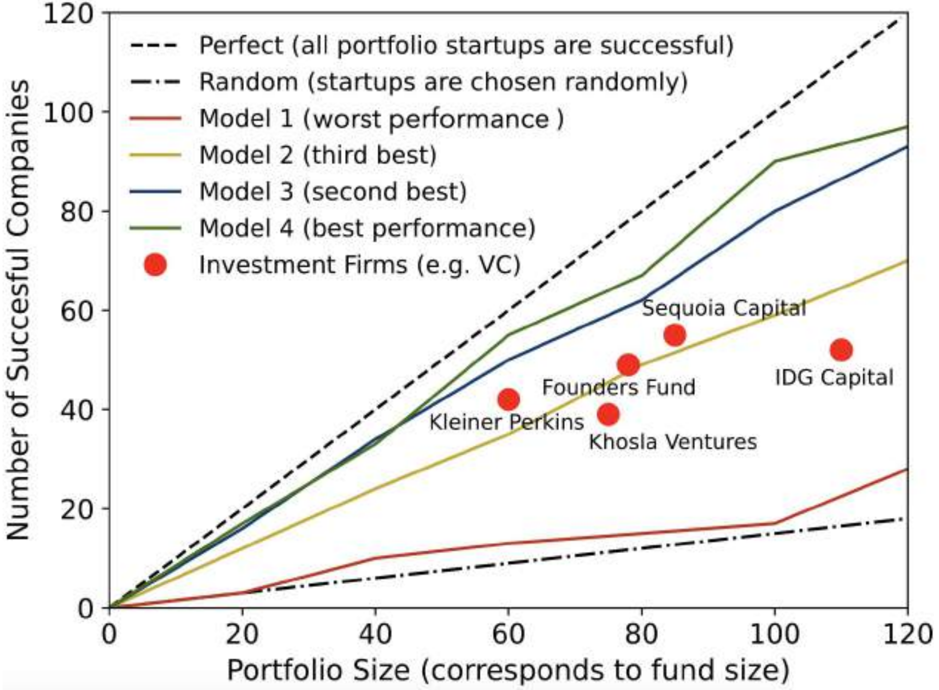


Evaluation Strategy

Portfolio Simulations:

- The trained model is used to form portfolios of size k (x-axis); the number of eventually successful startups is plotted over y-axis.

- Questions to be answered:
    - What is the expected success ratio?
    - How is it comparing to real funds?
    - How much better than random policy?
    - How far from perfect case?



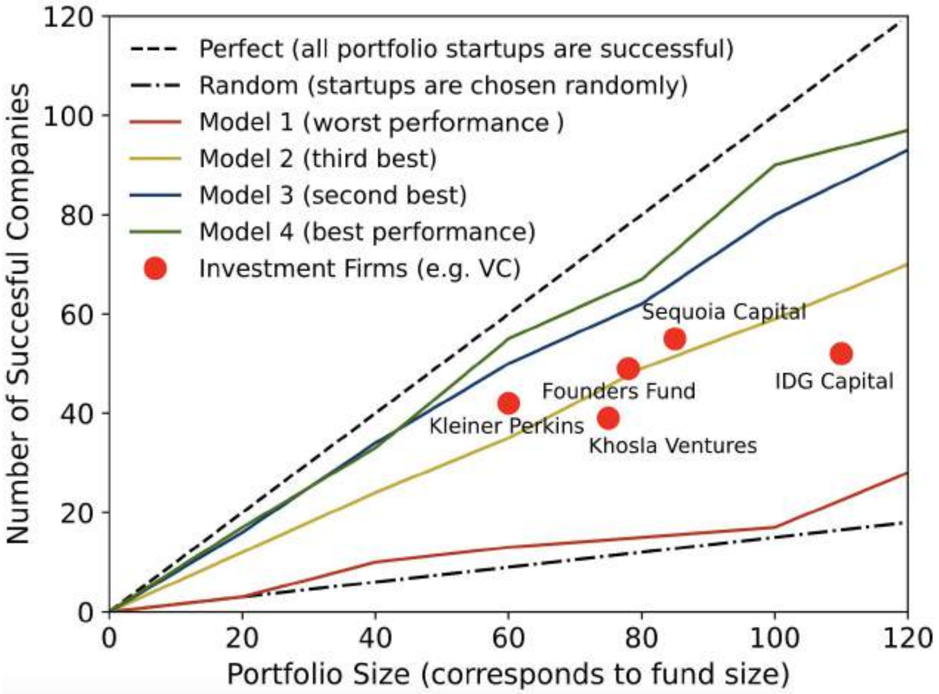
Evaluation Strategy

Portfolio Simulations:

- The trained model is used to form portfolios of size k (x-axis); the number of eventually successful startups is plotted over y-axis.

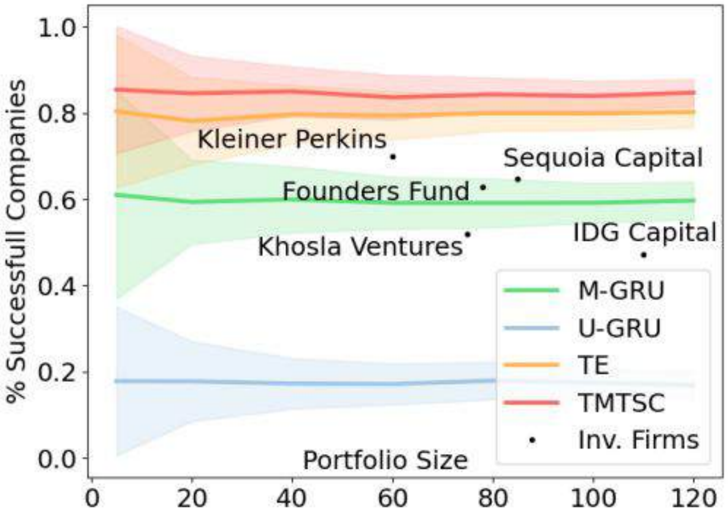
- Questions to be answered:
    - What is the expected success ratio?
    - How is it comparing to real funds?
    - How much better than random policy?
    - How far from perfect case?

- In practice, investment firms are more constrained than simulation: they can not invest in any startup due to many reasons like founders preference, portfolio conflict and investment mandate.

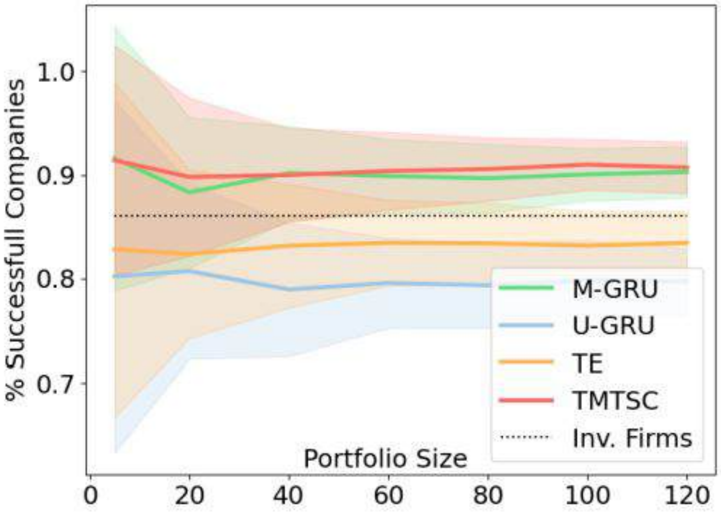


Our Simulation for EQT Ventures and Growth Fund [6]

success rate (slope) vs. portfolio size



(a) VC dataset



(b) GC dataset

Note: TMTSC is a Transformer-based model [6] developed by Motherbrain Research.

Agenda

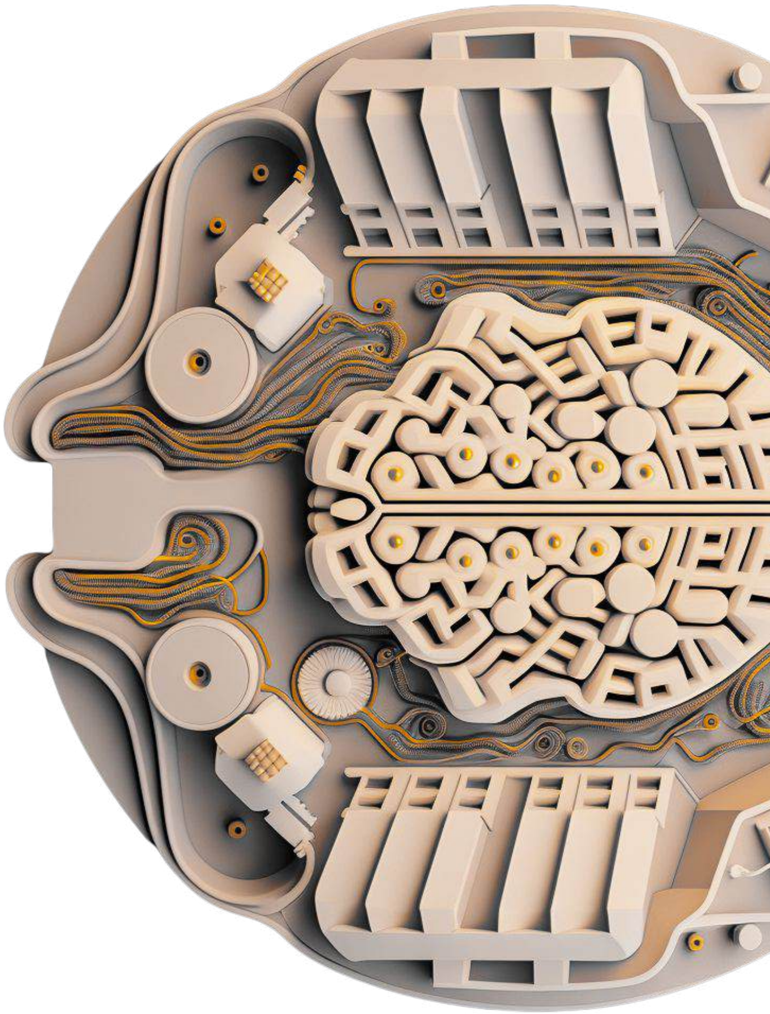
Success Prediction: Deep Learning

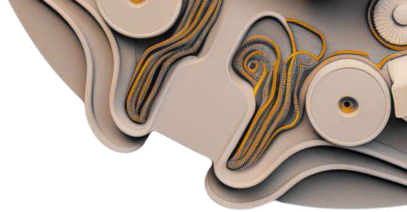
Sector Prediction: Large Language Model

Revenue Forecasting: Classic and State-of-The-Art

Document Mining: Knowledge Graph

Summary





Sector Prediction

YOU ARE HERE

Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

YOU ARE HERE

Due Diligence

- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

YOU ARE HERE

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?

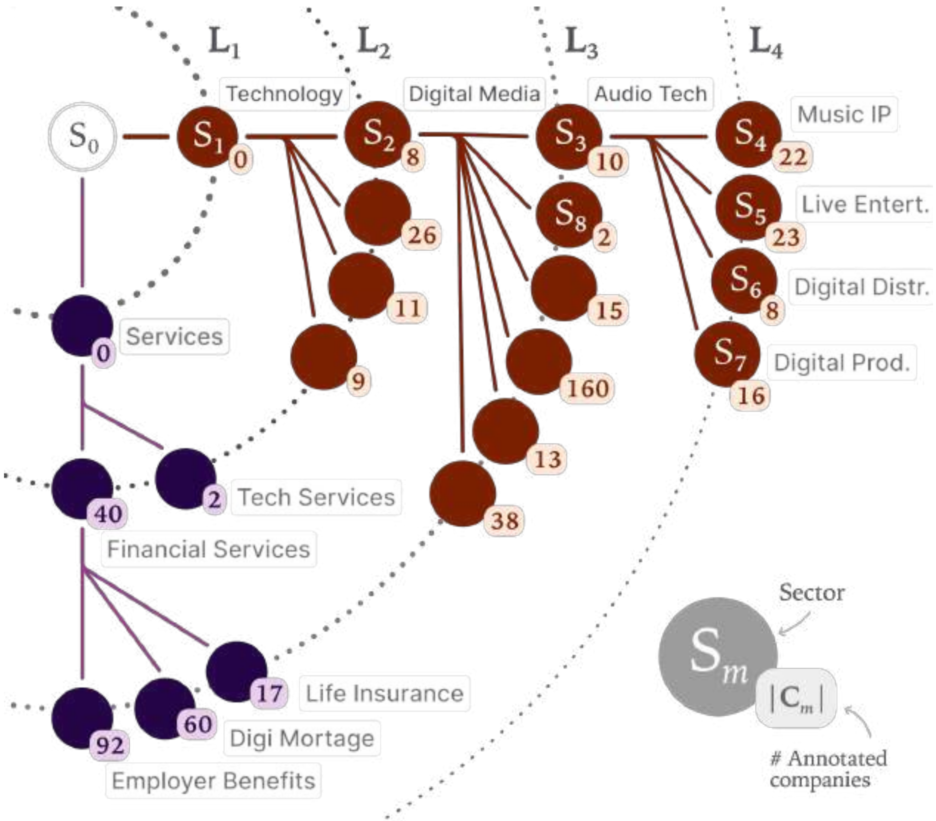
Sector Framework

Why?

- Identifying promising macro-trends
  - E.g. renewable energy, circular economy
- Finding investments within these macro-trends

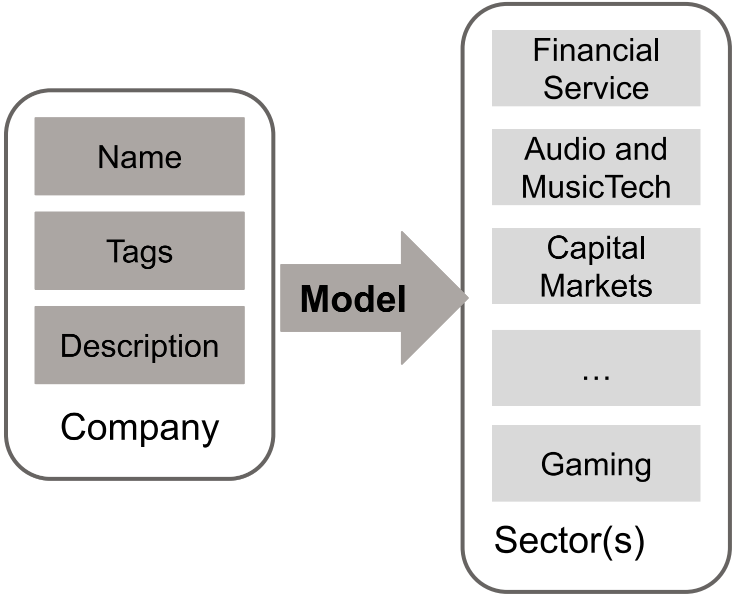
What? [4]

- **Customized**: no one-size-fit-all
- **Hierarchical**: a tree structure
- **Dynamic**: prone to change
- **Imbalanced**: varying granularity
- **Low-resource:** few labels available



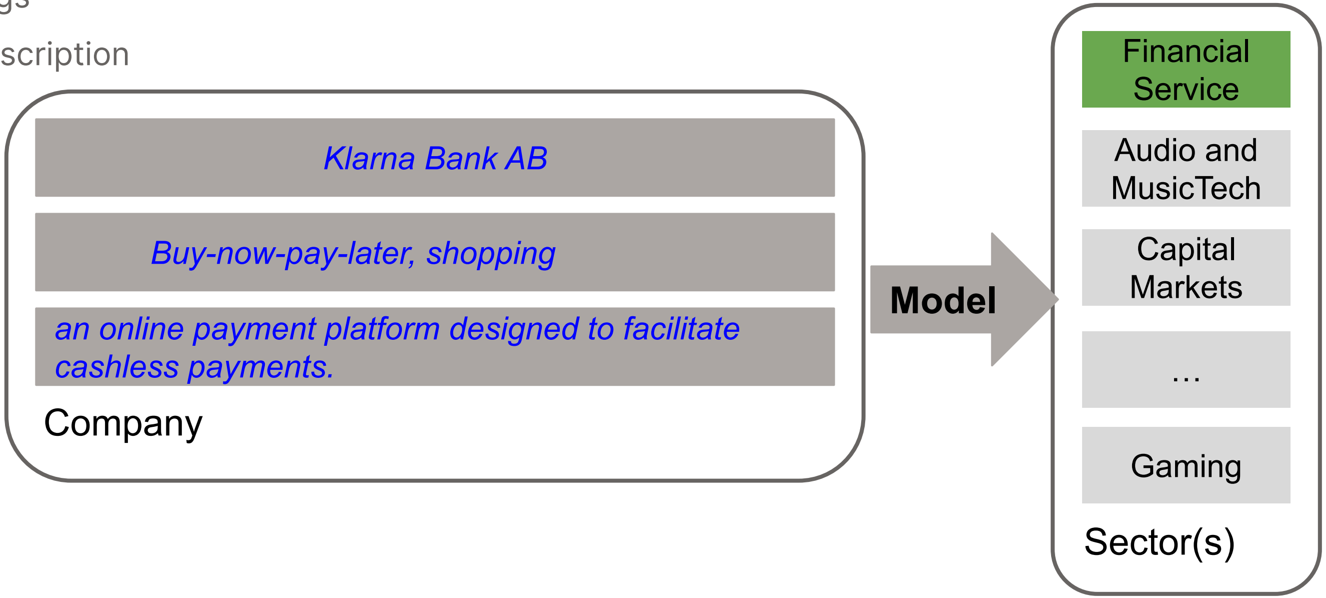
Method

- **Task:** Assign each company to the most relevant sector.
- Input Features originating from multiple sources:
  - Company Name
  - Company Tags
  - Company Description



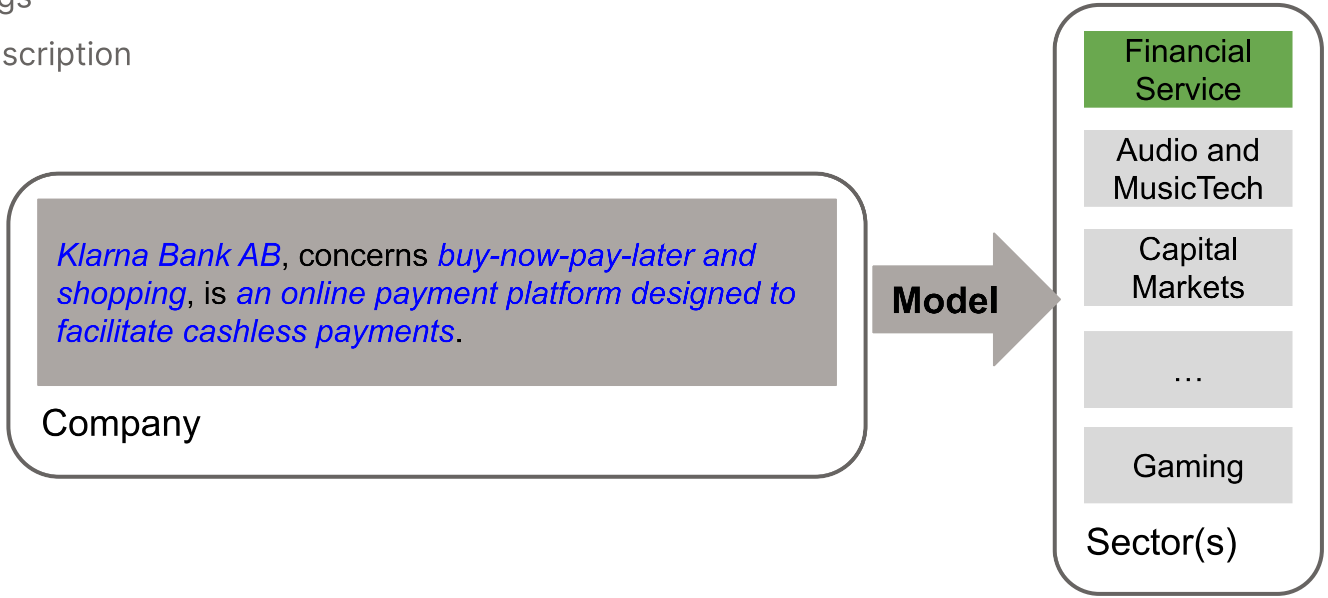
Method

- **Task:** Assign each company to the most relevant sector.
- Input Features originating from multiple sources:
  - Company Name
  - Company Tags
  - Company Description



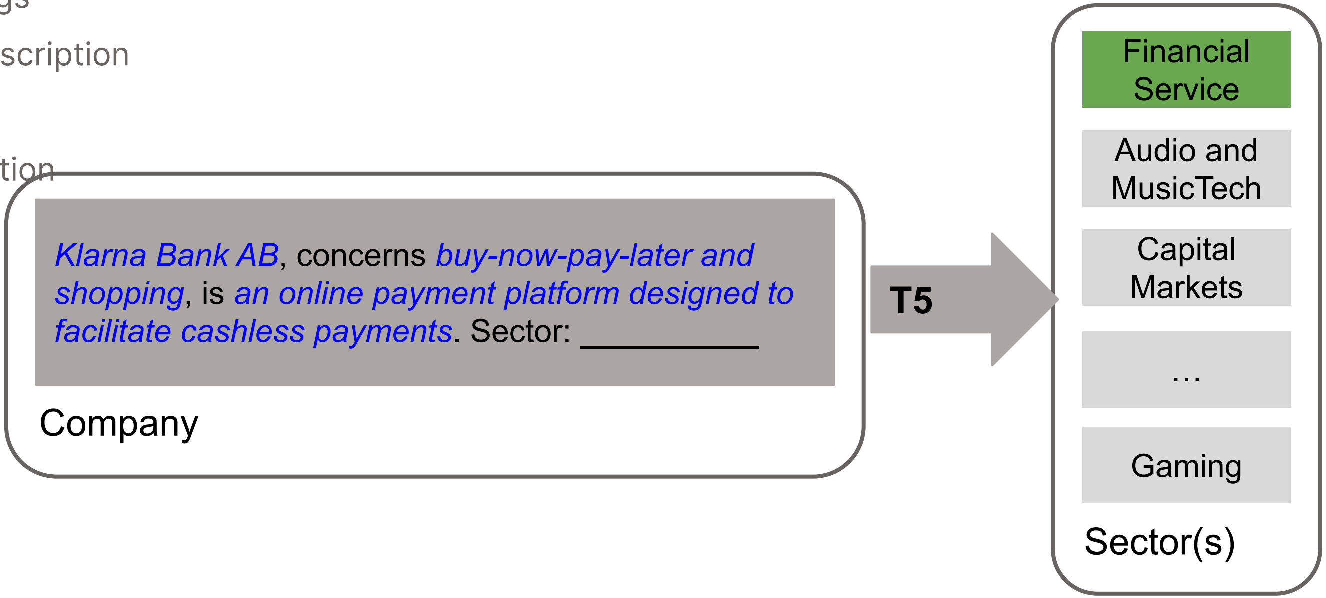
Method

- **Task:** Assign each company to the most relevant sector.
- Input Features originating from multiple sources:
  - Company Name
  - Company Tags
  - Company Description
- Input Template



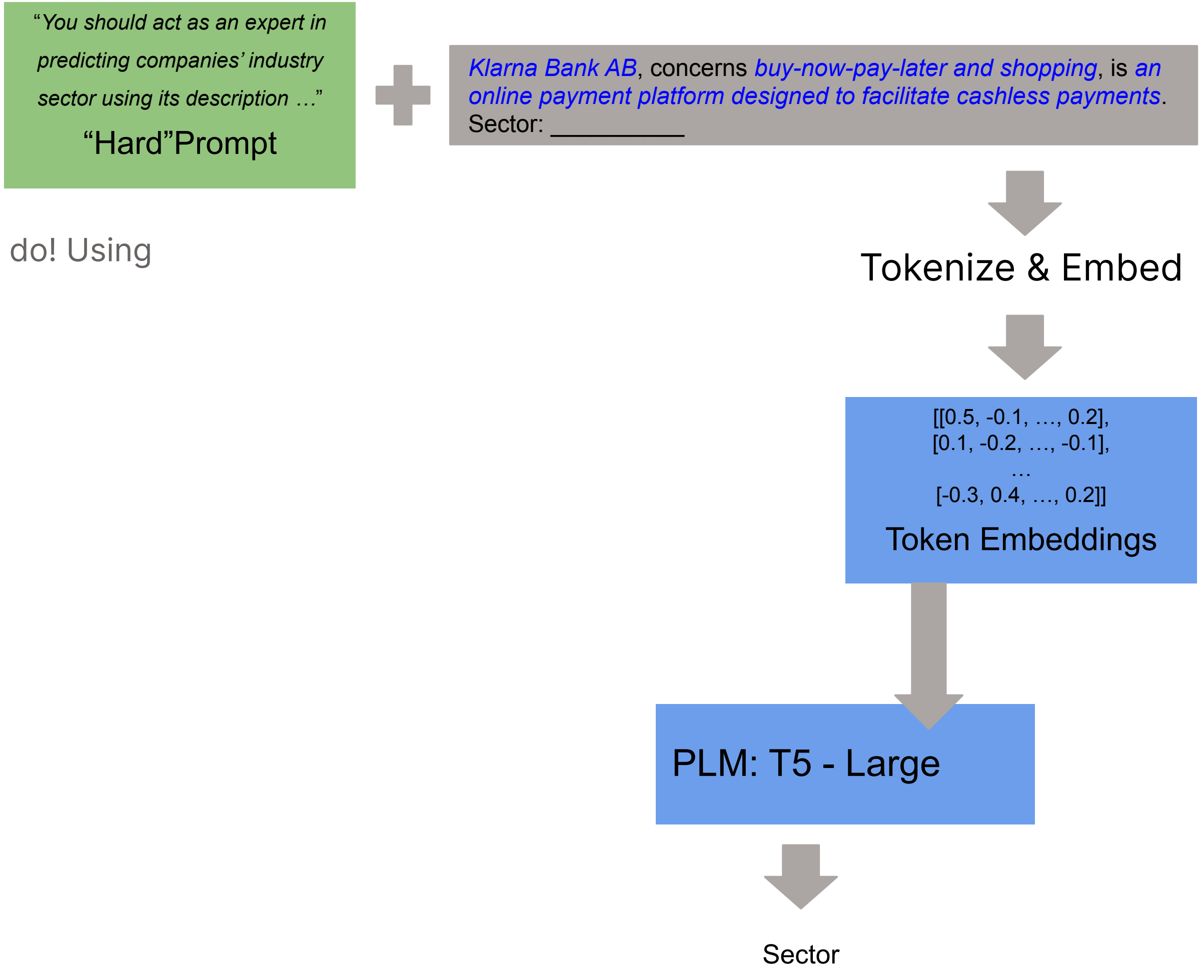
Method

- **Task:** Assign each company to the most relevant sector.
- Input Features originating from multiple sources:
  - Company Name
  - Company Tags
  - Company Description
- Input Template
- Generative Completion



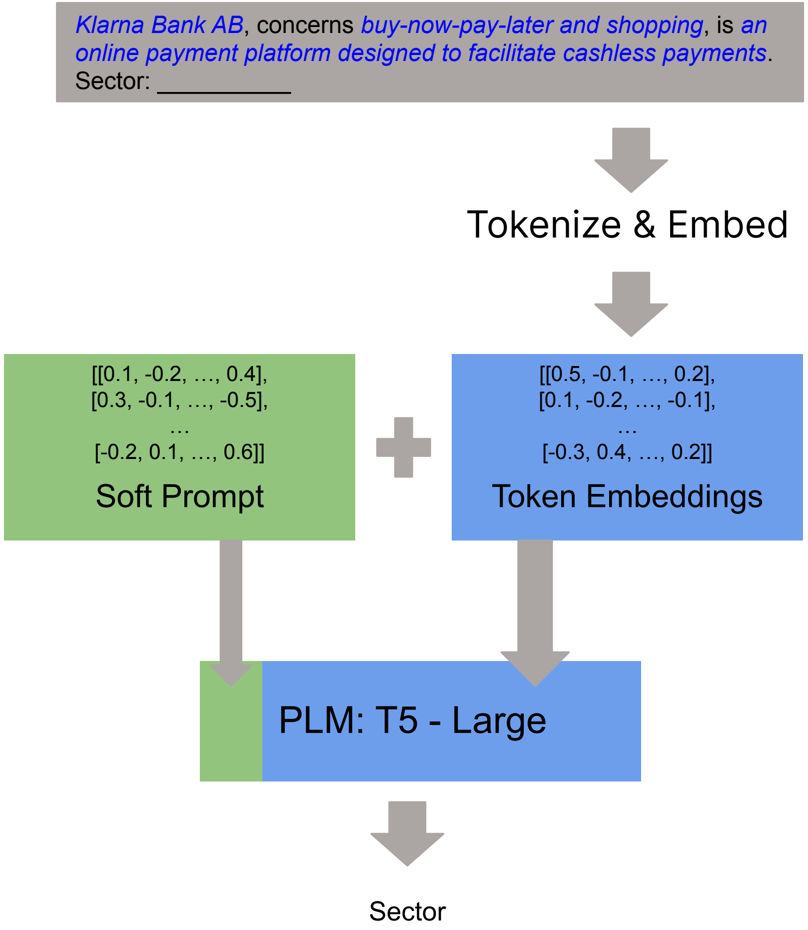
Training

- We need to tell the model what to do! Using fixed “hard” prompt??



Training

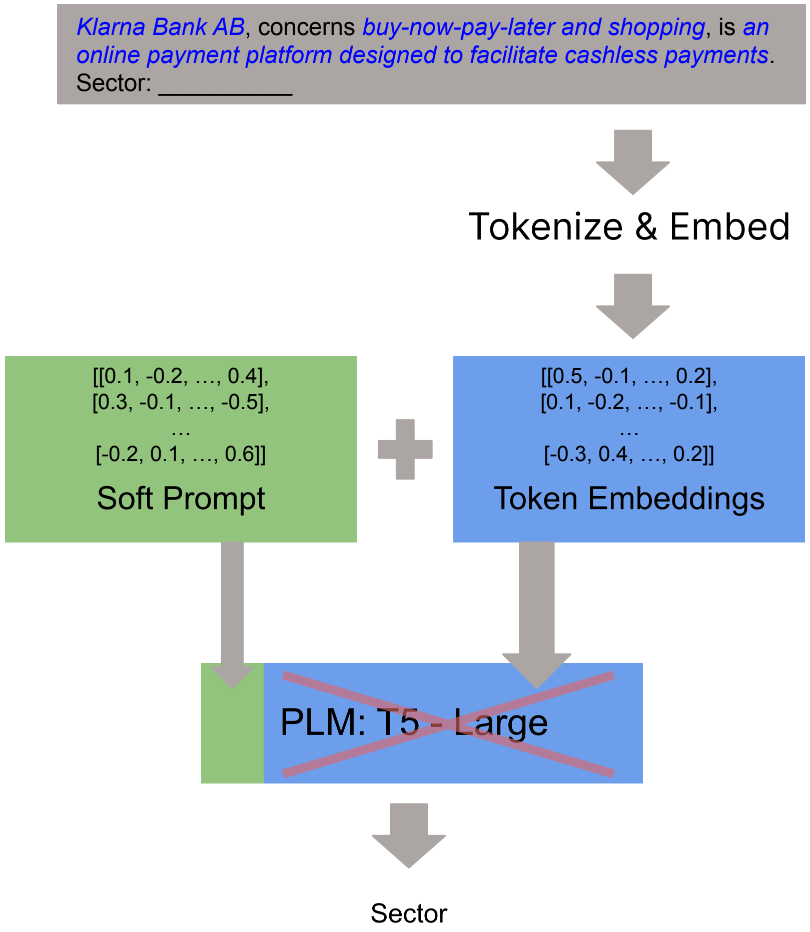
- We need to tell the model what to do! Using fixed “hard” prompt?? **NO!**
- Add random **soft prompt** to represent the task to be learned.



Training

- We need to tell the model what to do! Using fixed “hard” prompt?? **NO!**
- Add random **soft prompt** to represent the task to be learned.

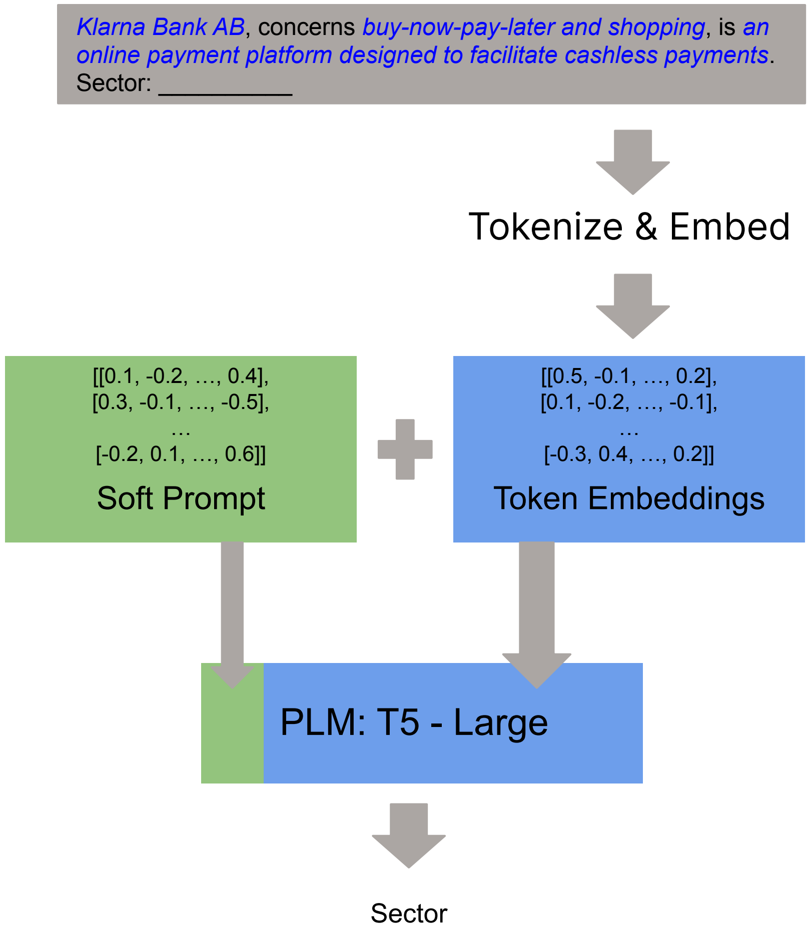
- First t' steps: only tune **soft prompt** part.



Training

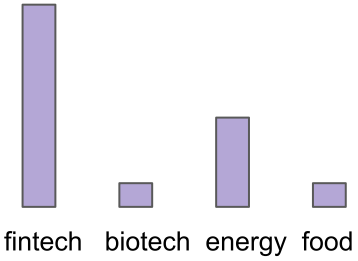
- We need to tell the model what to do! Using fixed “hard” prompt?? **NO!**
- Add random **soft prompt** to represent the task to be learned.

- First t' steps: only tune **soft prompt** part.
- Then, tune the entire architecture jointly until convergence. [4]



Training: data balancing

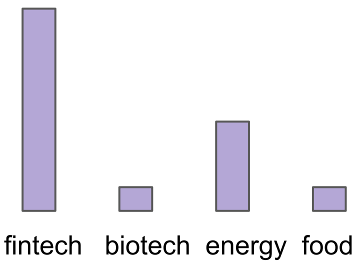
The annotations are heavily unbalanced.

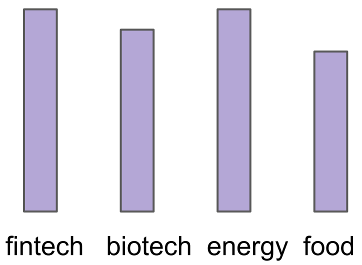


Training: data balancing

The annotations are heavily unbalanced.

- Augment classes with few labels using *EDA*[1], which randomly does:
    - Synonym Replacement
    - Random Insertion
    - Random Deletion
    - Random Swap



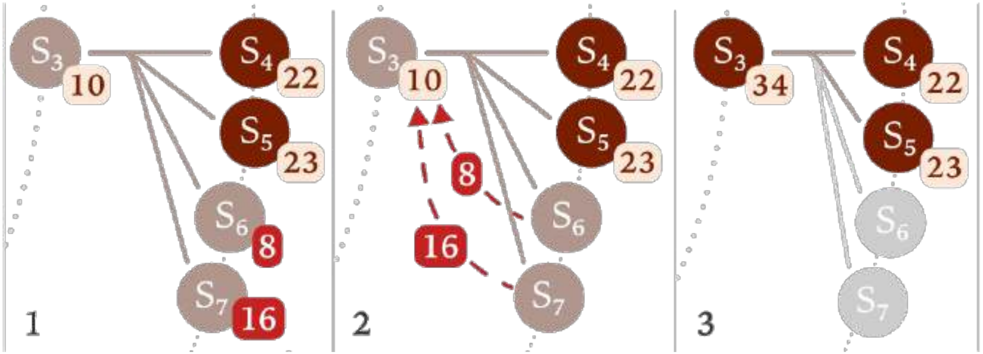


1: EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks (Wei & Zou, EMNLP-IJCNLP 2019)

Training: label attribution

Some sectors have extremely few annotations. To maximize utilization of labels, we collapse these sectors into their parents. [4]

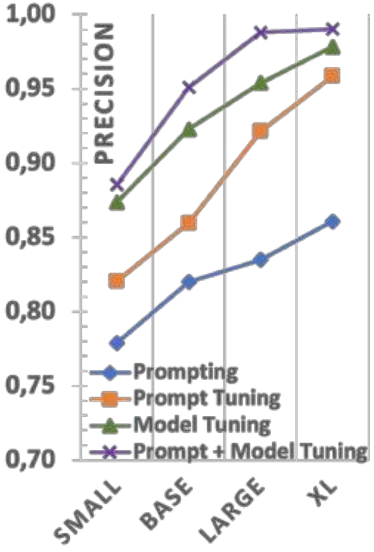
Keep the granularity if we have enough labels.



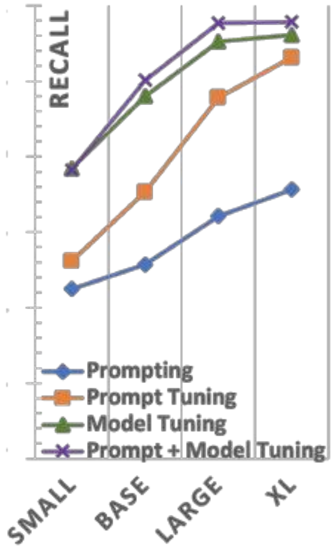
Experiments

Impact of **Methods** and **Model Sizes** [4]

- Our method (**Prompt + Model Tuning**) performs the best.
- All methods performs better with larger model size.
- Our method (**Prompt + Model Tuning**) is able to achieve better performance with a smaller model size.



(a) Average Precision



(a) Average Recall

Experiments

A snapshot of confusion matrix. [4]

- Since L3 sectors are more fine-grained requiring less (than L2) annotations, a generally better performance is observed for L3 than L2.

- Sectors on L3 levels have an accuracy of over 90% except **horizontal software** and **vertical software**.

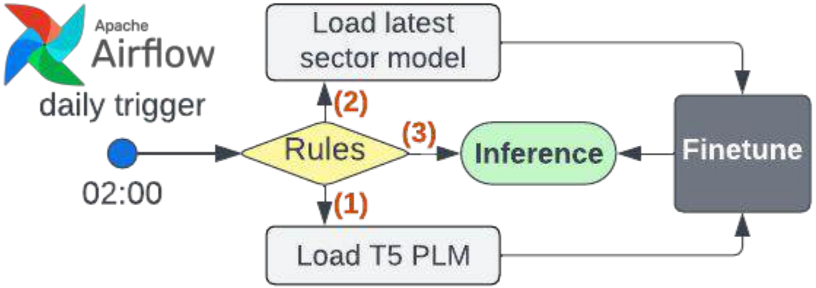
L2	diginal media	12	0	0	0	0	2	1
	deep tech.	0	11	0	3	0	1	1
L3	game	1	0	14	0	0	0	1
	cyber security	0	0	0	15	0	0	0
	market place	0	0	0	0	13	0	2
	horizontal software	0	0	0	1	1	11	2
	vertical software	2	0	1	0	0	3	10
	SECTOR	digi. media	deep tech.	game	cyber sec.	market place	horiz. soft.	vert. soft.
LAYER		L2		L3				

System [4]

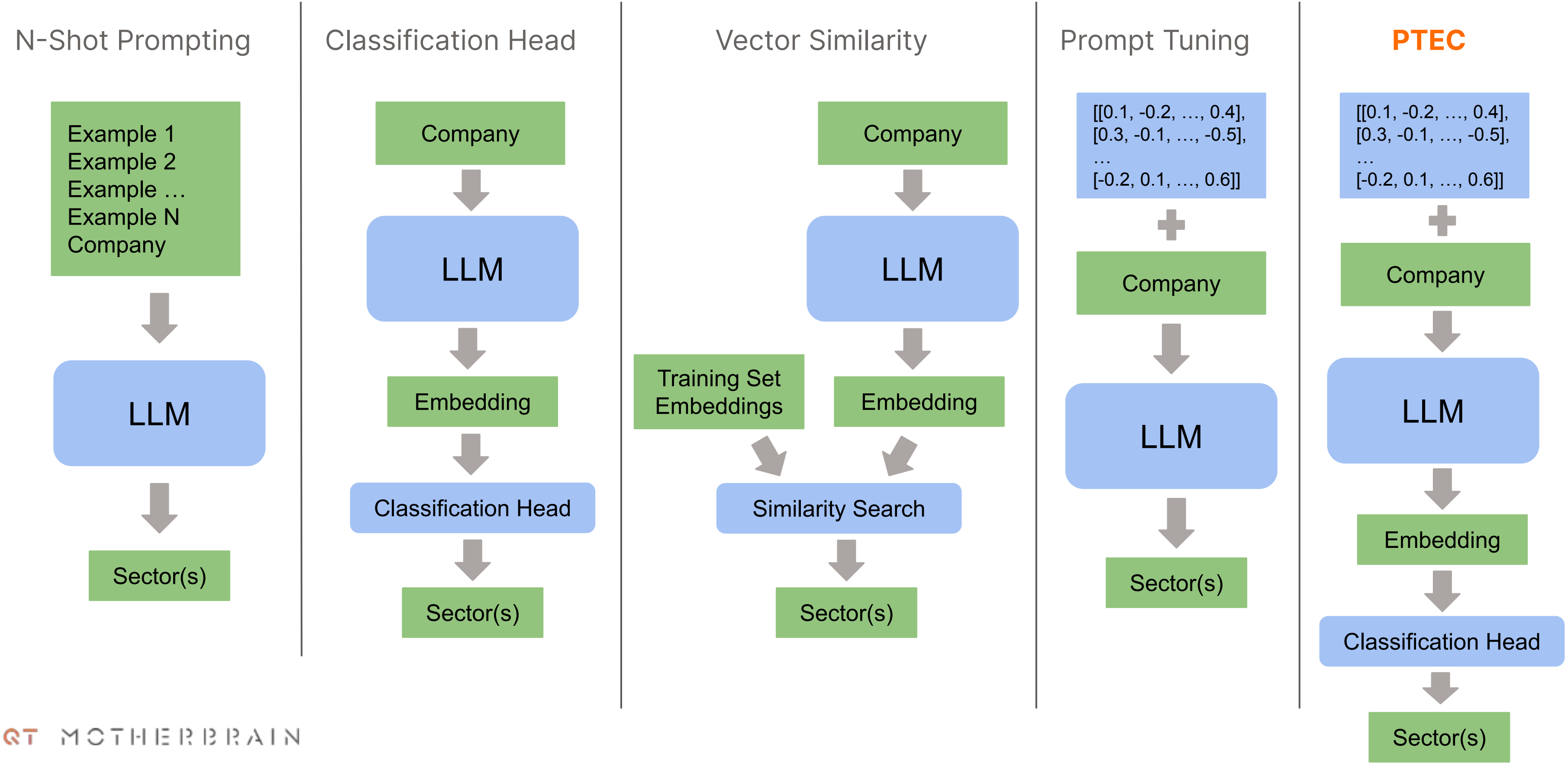
1) **Retrain entirely**: when the sector framework is changed or the annotation for any existing sector has evolved significantly. Rerun inference for all.

2) **Finetune from last model**: when the sector annotation only changed marginally.

3) **Inference directly**: skip finetune and only run inference for changed companies. Greatly reducing the daily inference load (by approx 95%).



Recent advancement [3]



Agenda

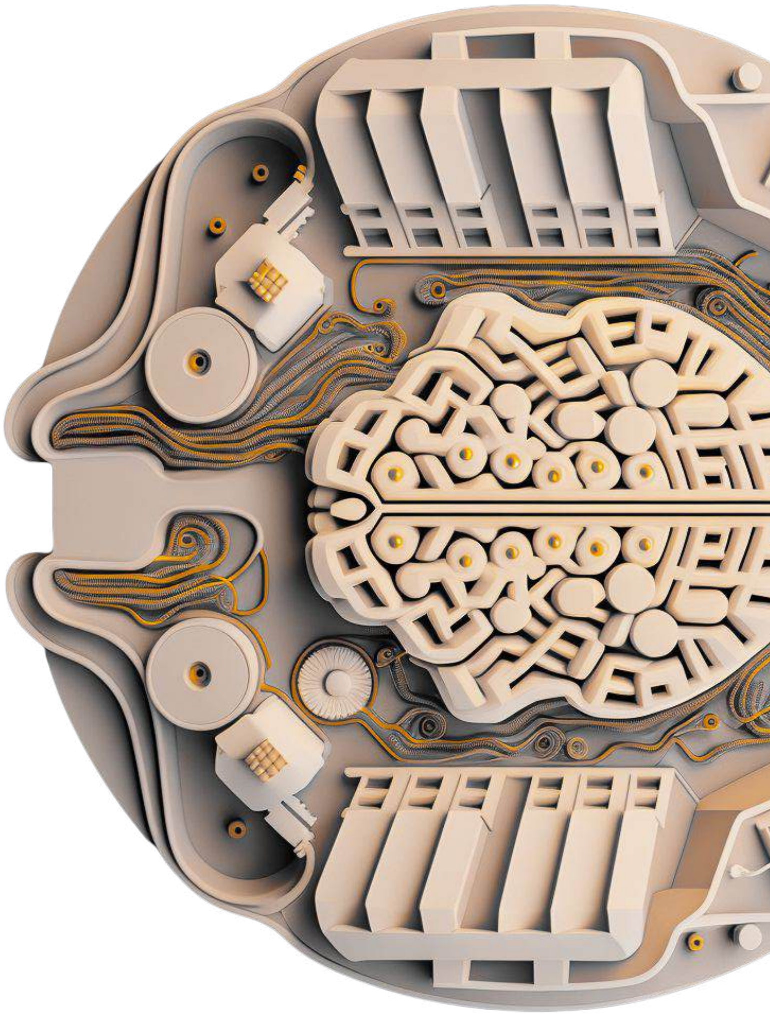
Success Prediction: Deep Learning

Sector Prediction: Large Language Model

Revenue Forecasting: Classic and State-of-The-Art

Document Mining: Knowledge Graph

Summary



Financial Forecasting

YOU ARE HERE

YOU ARE HERE

YOU ARE HERE

Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

Due Diligence

- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?

Revenue and Scaleup

Revenue: total income from generated from main business, indicating performance of a company’s performance.

Scaleups: companies with proven scalability, viability and accelerated revenue growth.

Revenue is a highly relevant metric to evaluate a scaleup company!

Revenue Forecast

Investment Professionals(IP) rely on **extrapolating company revenue** into the future to **approximate the valuation of companies** and inform their investment decision

Financial data on scaleups is typically **proprietary, costly and scarce,** forming a huge **obstacle** for directly **applying data-driven methodologies**

Forecasting typically done **manually** and **empirically** leaving the quality **heavily dependent** on the investment professionals' experiences and insights

Promise of Data-Driven Approach

Level of **automation, objectiveness, consistency** and **adaptability** for empirical revenue forecasting is **far from optimal**

Highly desirable for investment professionals evaluating scaleups to have a data-driven method that performs revenue extrapolation on scarce data in an automated way

- A **quick way** to **assess companies' revenue potential** with little information needed
- **Benchmarking** of a **manually produced revenue forecasting**

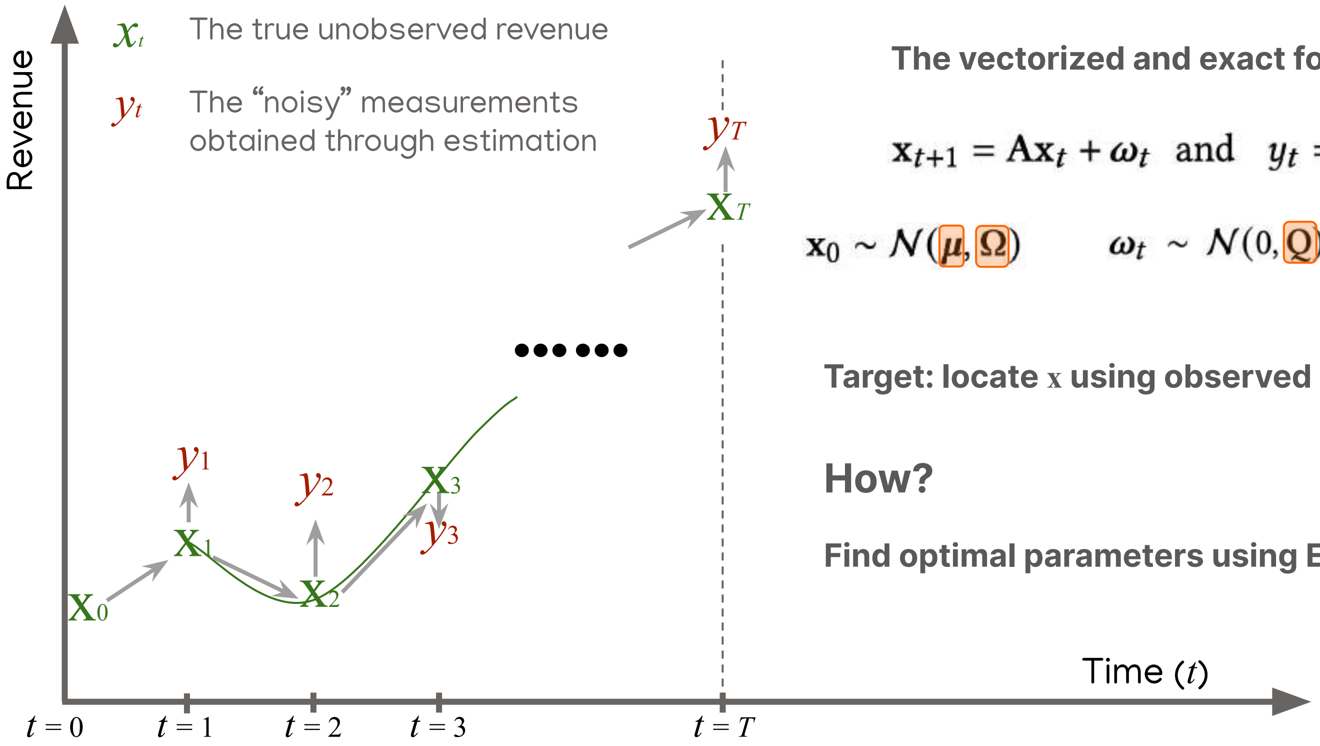
Data-Driven Revenue Forecast

The algorithm should [9]

- work for multiple business sectors,
- work on a small dataset,
- commence from short time-series,
- extrapolate for long term (e.g. 3 years),
- estimate confidence,
- have low requirement on auxiliary information,
- be easy to explain.

This is the first work that meets all practical requirements simultaneously.

Revenue Model: LDS (Linear Dynamical System) [9]



The vectorized and exact form:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \boldsymbol{\omega}_t \quad \text{and} \quad y_t = \mathbf{c}\mathbf{x}_t + \epsilon_t$$

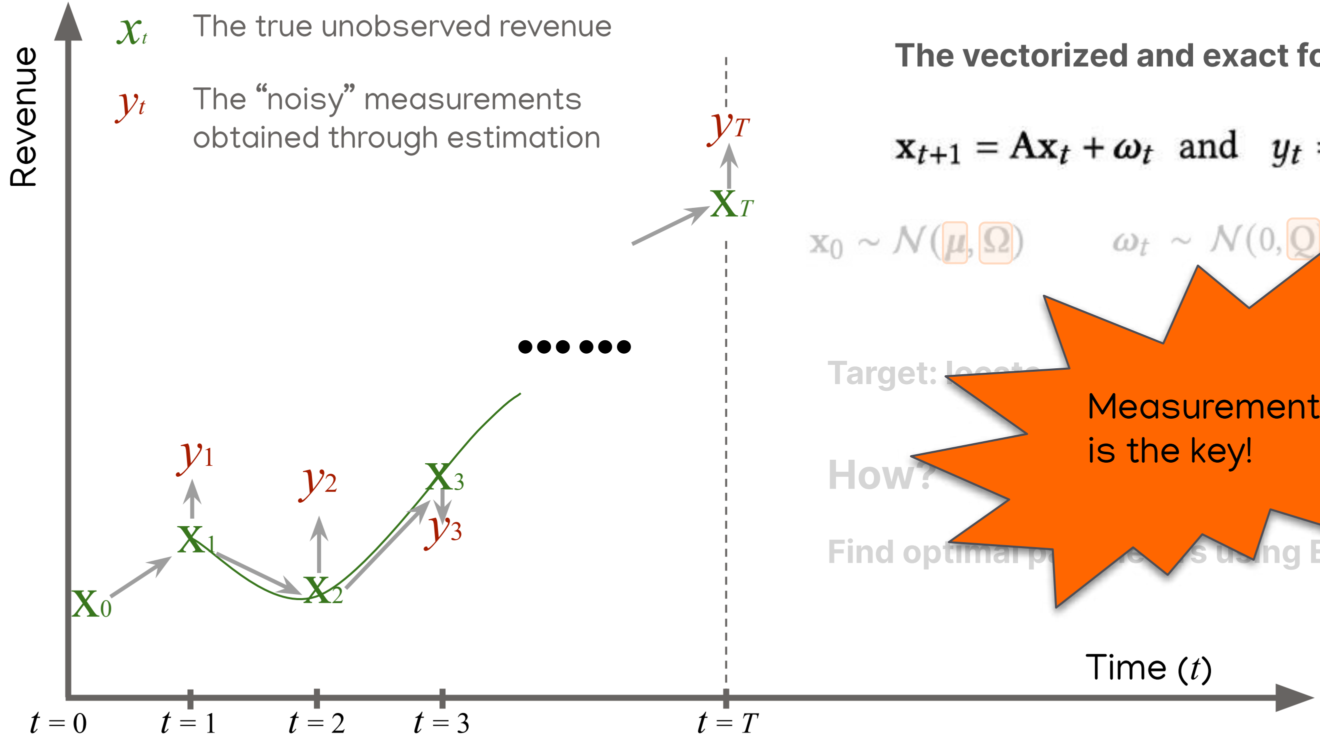
$$\mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}) \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(0, \mathbf{Q}) \qquad \epsilon_t \sim \mathcal{N}(0, \mathbf{R})$$

Target: locate x using observed measurement y

How?

Find optimal parameters using EM algorithm!

Revenue Model : Optimization!



The vectorized and exact form:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \boldsymbol{\omega}_t \quad \text{and} \quad y_t = \mathbf{c}\mathbf{x}_t + \epsilon_t$$

$$\mathbf{x}_0 \sim \mathcal{N}(\mu, \Omega) \qquad \omega_t \sim \mathcal{N}(0, \mathrm{Q}) \qquad \epsilon_t \sim \mathcal{N}(0, \mathrm{R})$$

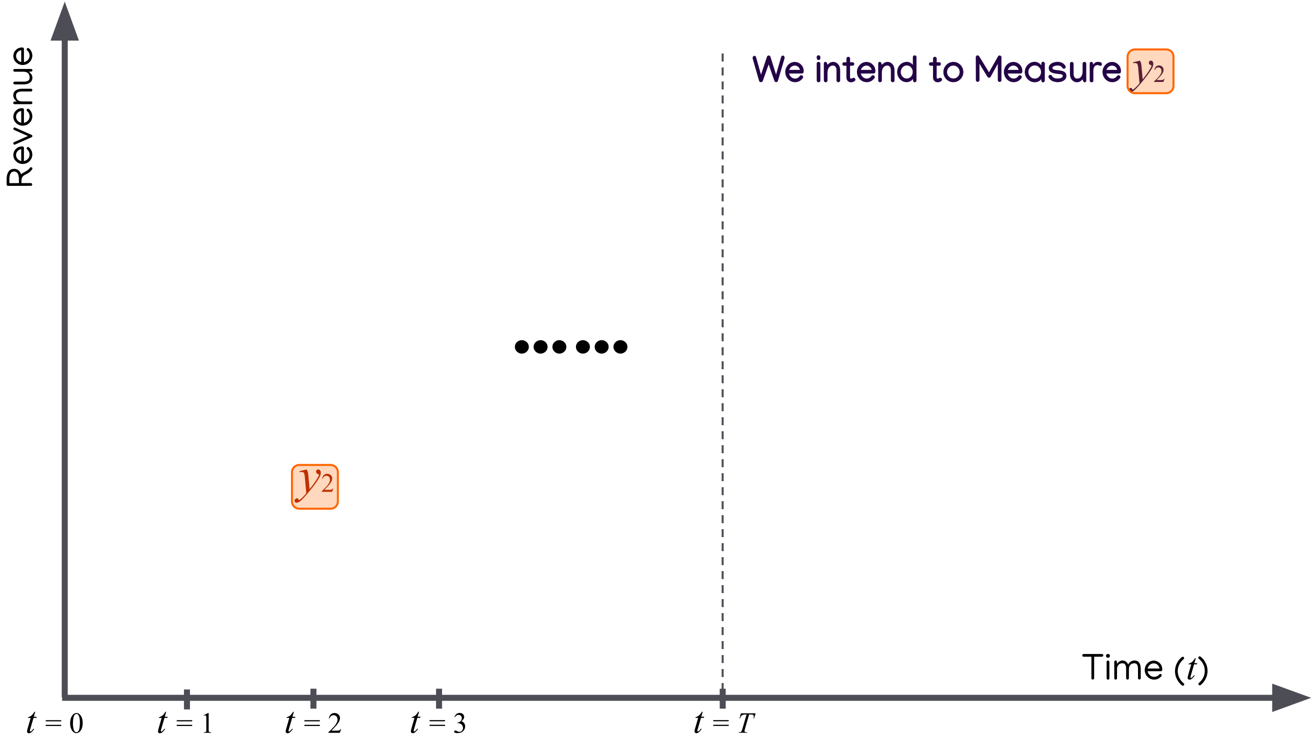
Target: locate ... ment y

How?

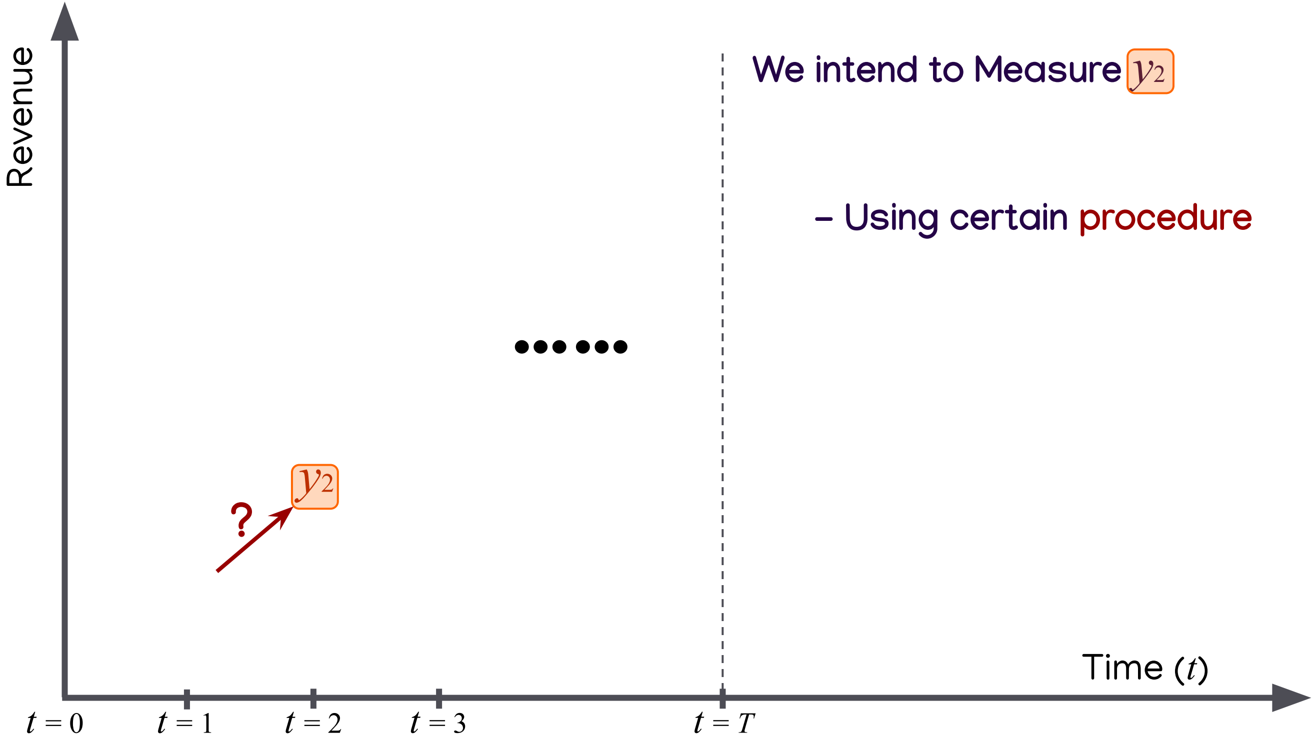
Find optimal p... s using EM algorithm!

Measurement is the key!

Measurements



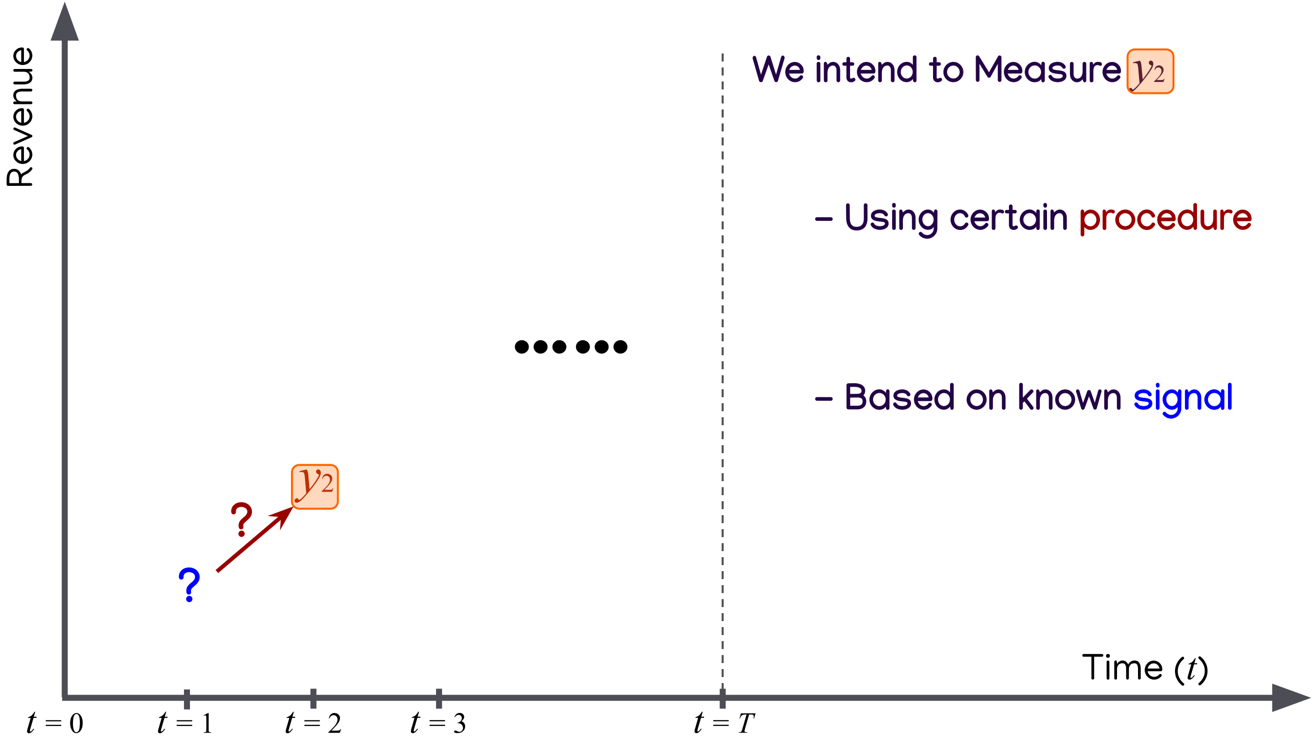
Measurements



We intend to Measure $\boxed{y_2}$

– Using certain procedure

Measurements

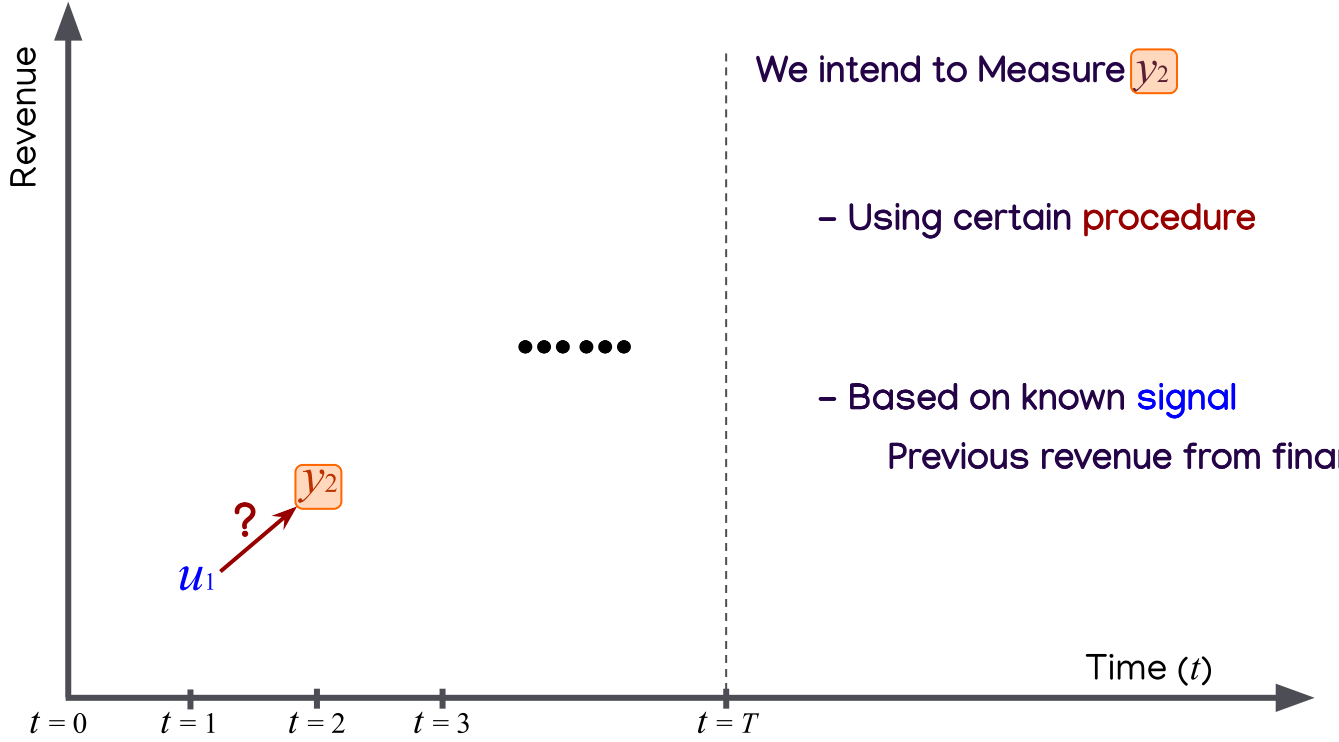


We intend to Measure $\boxed{y_2}$

– Using certain procedure

– Based on known signal

Measurements



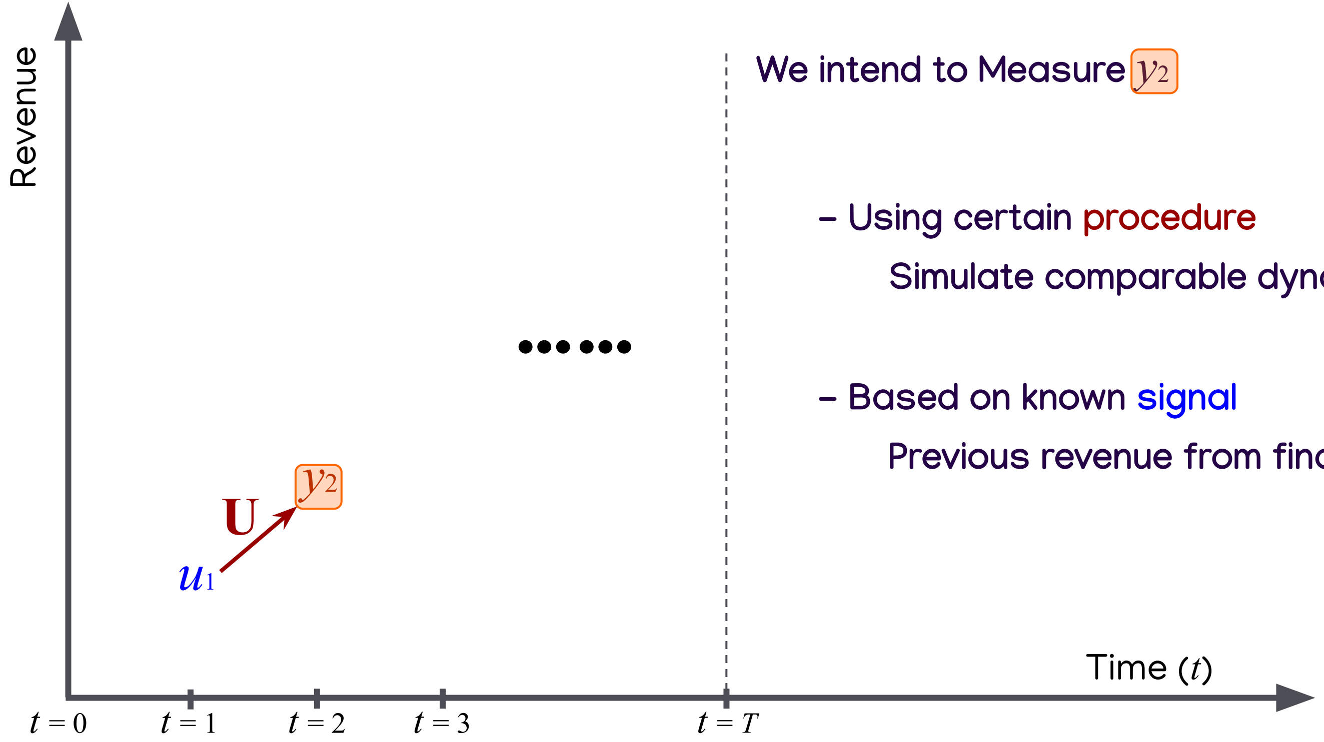
We intend to Measure $\boxed{y_2}$

– Using certain procedure

– Based on known signal

Previous revenue from financial report u_1

Measurements



We intend to Measure $\boxed{y_2}$

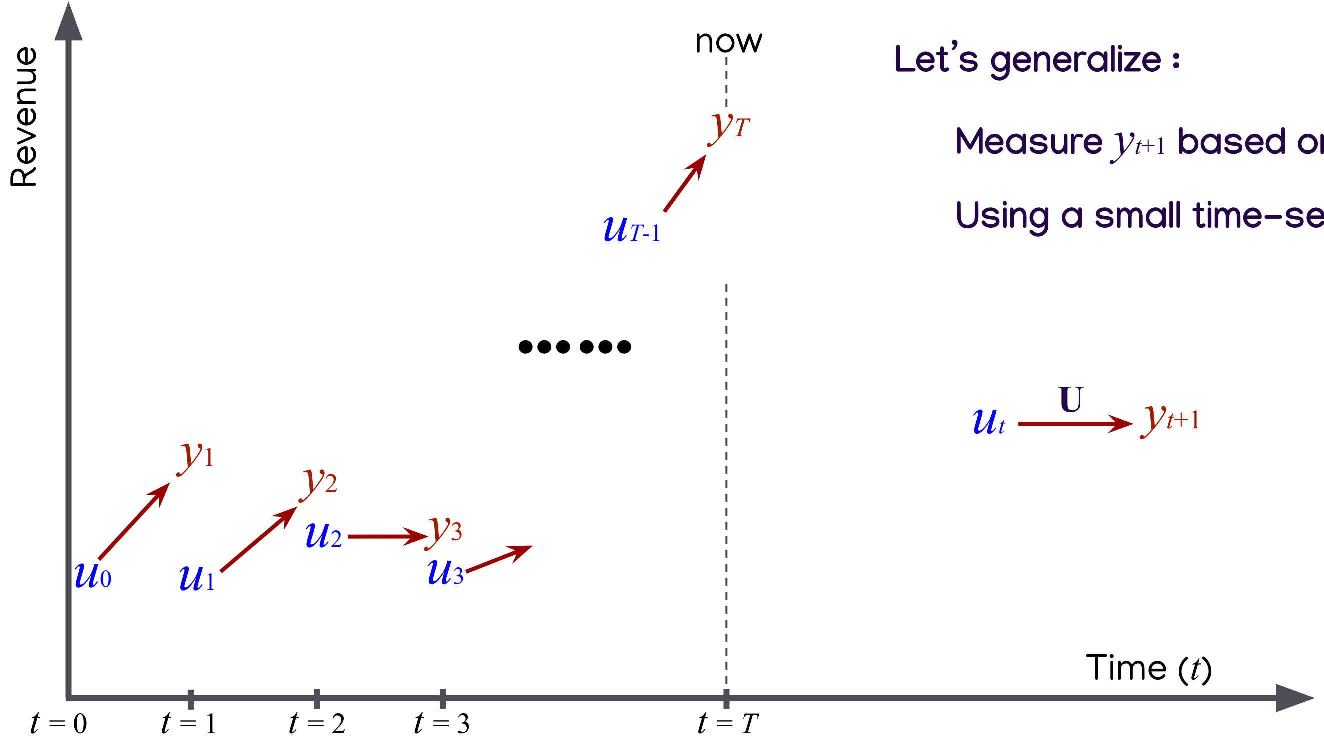
- Using certain procedure

  Simulate comparable dynamics in a dataset $\mathbf{U}$

- Based on known signal

  Previous revenue from financial report u_1

Measurements



Let's generalize :

Measure y_{t+1} based on u_t

Using a small time-series dataset $\mathbf{U}$

$$u_t \xrightarrow{\mathbf{U}} y_{t+1}$$

Measurements

$$u_t \xrightarrow{\mathbf{U}} y_{t+1}$$

Measurements

$u_t \xrightarrow{\mathbf{U}} y_{t+1}$

Dataset $\mathbf{U}$

timeline

companies

(8.7, Jan 18, 1.5, 1.7), ……

(4.5, Aug 20, 1.2, 2.0), ……

⋮

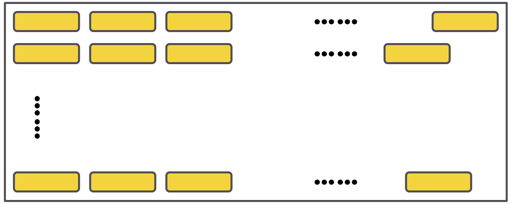
Explain a Tuple **(8.7, Jan 18, 1.5, 1.7)**

- **Jan 18** – the time when obtaining this tuple
- **8.7** – the revenue obtained in **Jan 18** is **8.7**
- **1.5** (current YoY growth)
  – the revenue of **Jan 17** is **8.7/1.5=5.8**
- **1.7** (next YoY growth)
  – the revenue ratio: **Feb 18 / Feb 17 = 1.7**

Measurements

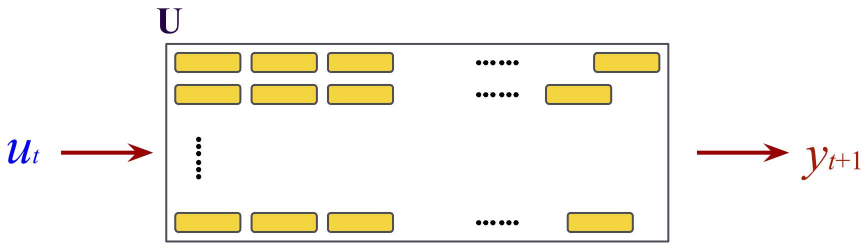
$$u_t \xrightarrow{\mathbf{U}} y_{t+1}$$

Dataset $\mathbf{U}$



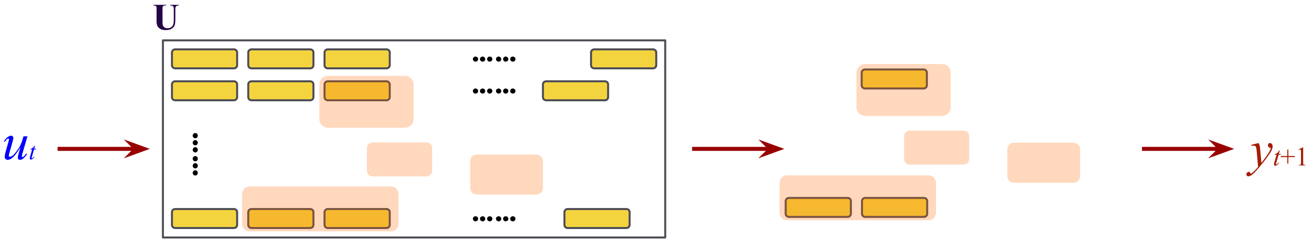
Measurements

$$u_t \xrightarrow{\mathbf{U}} y_{t+1}$$



Measurements

$$u_t \xrightarrow{\mathbf{U}} y_{t+1}$$

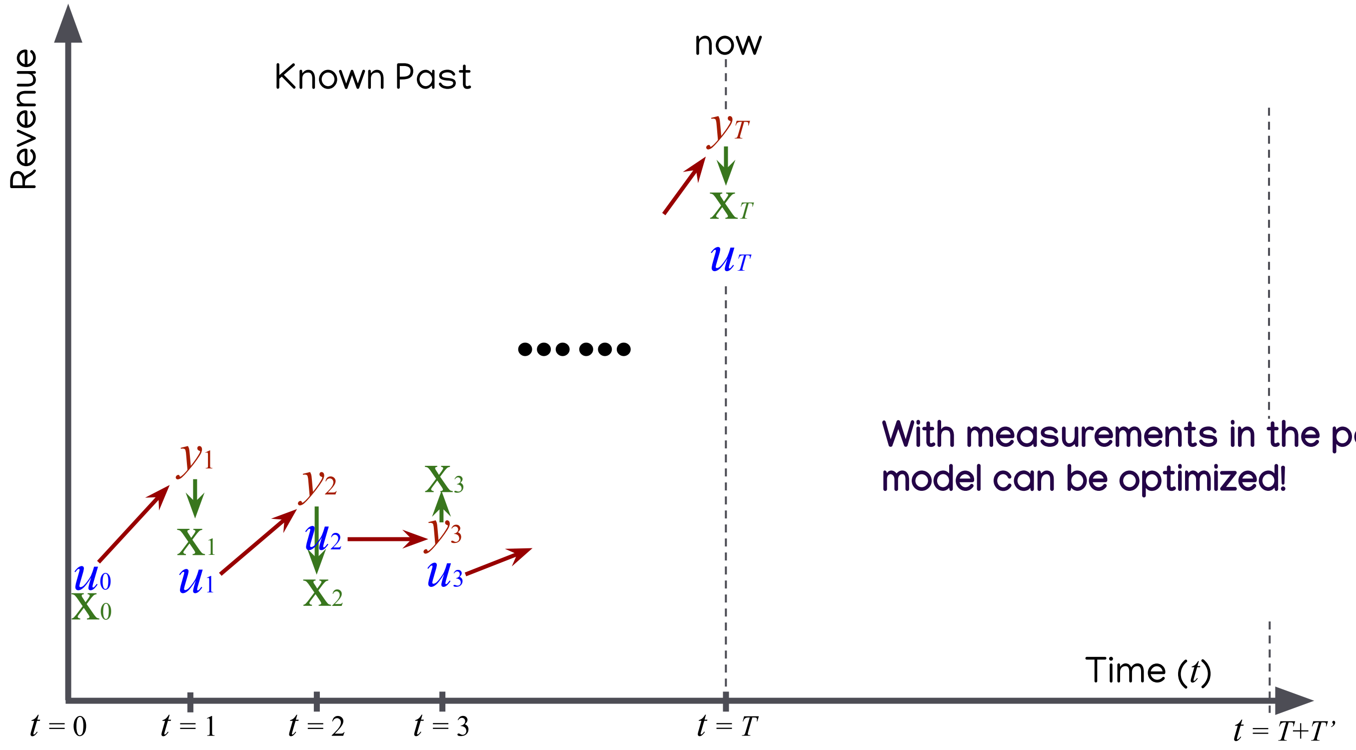


Filtering by :

- business,
- year-month,
- revenue,
- YoY revenue growth

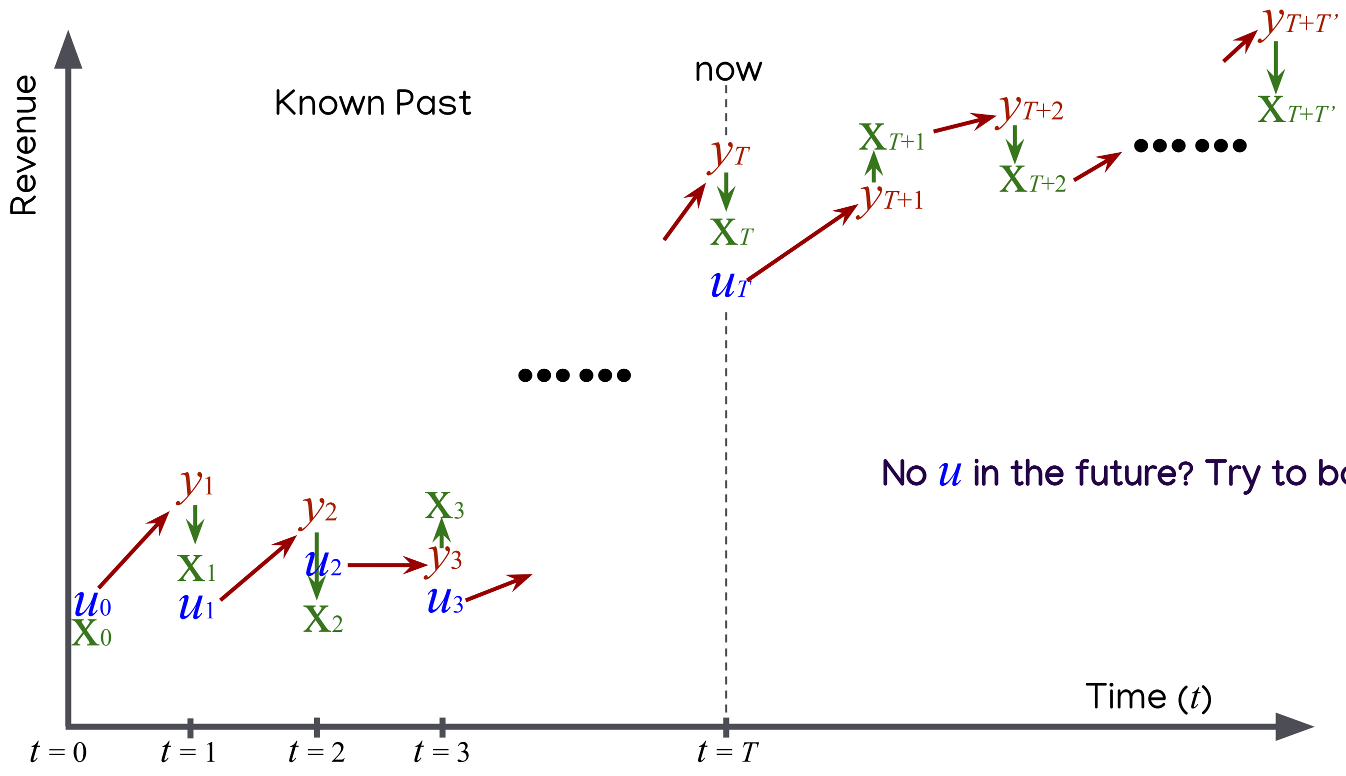
Sampling:
A **stochastic** approach

Measurements



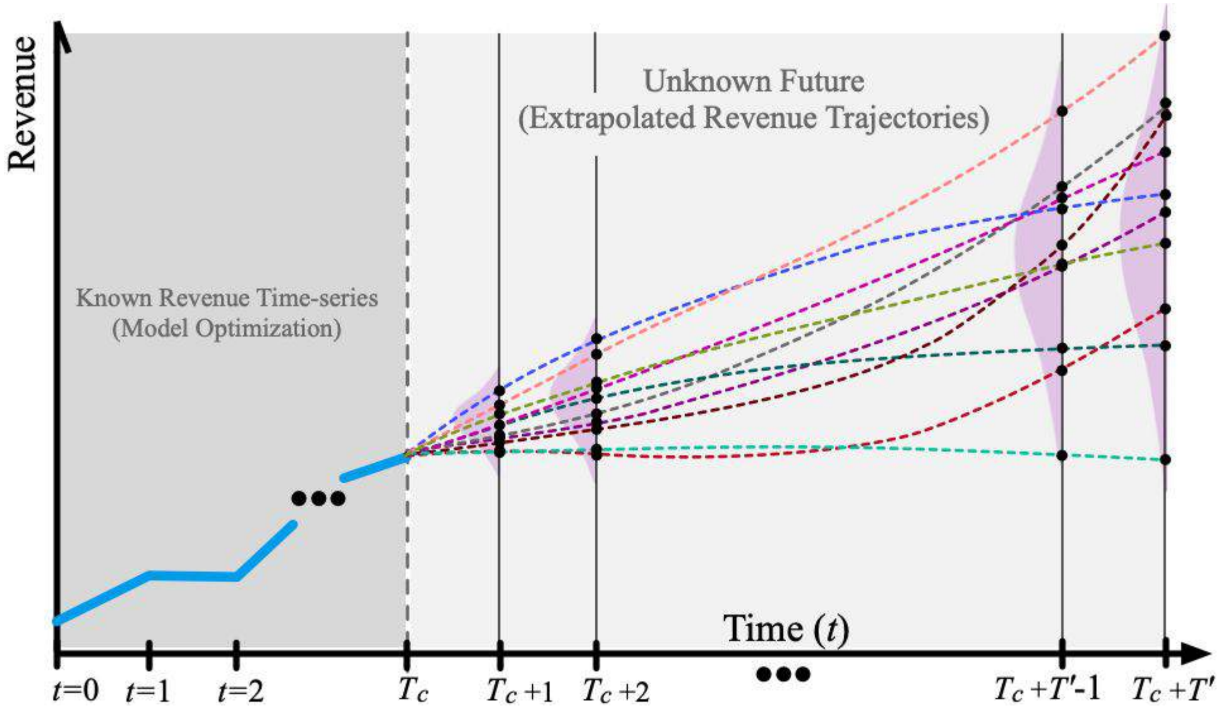
With measurements in the past, revenue model can be optimized!

Forecast



No u in the future? Try to base on x instead!

Forecast with Confidence

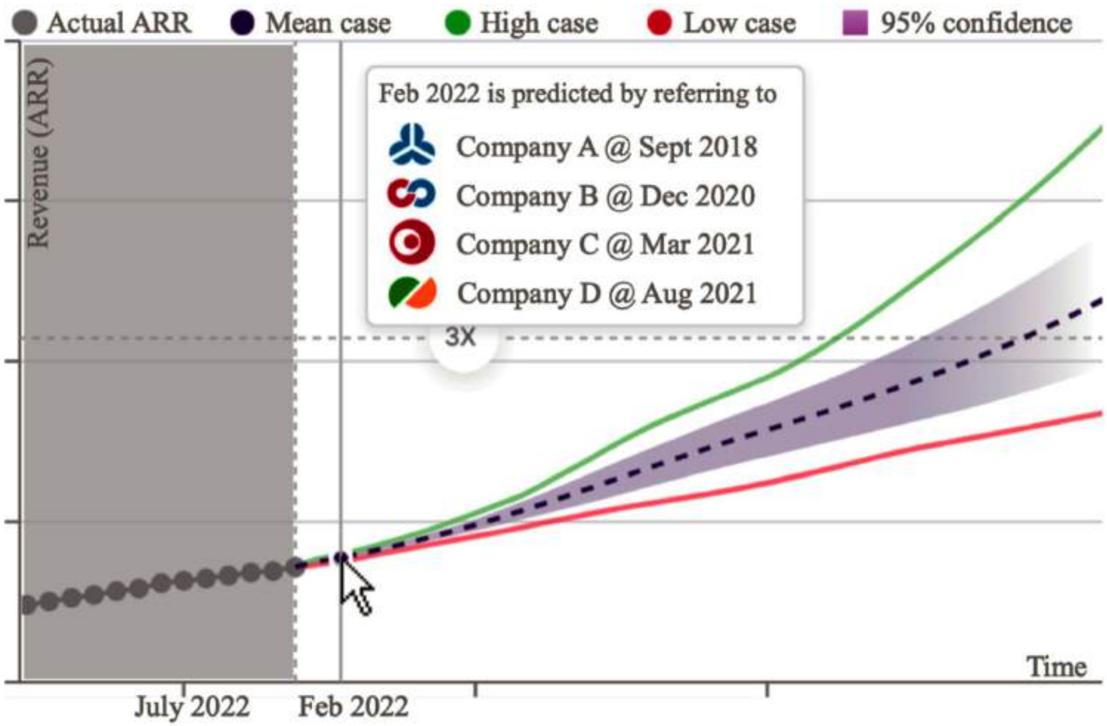


Extrapolate multiple times

&

Assume Gaussian

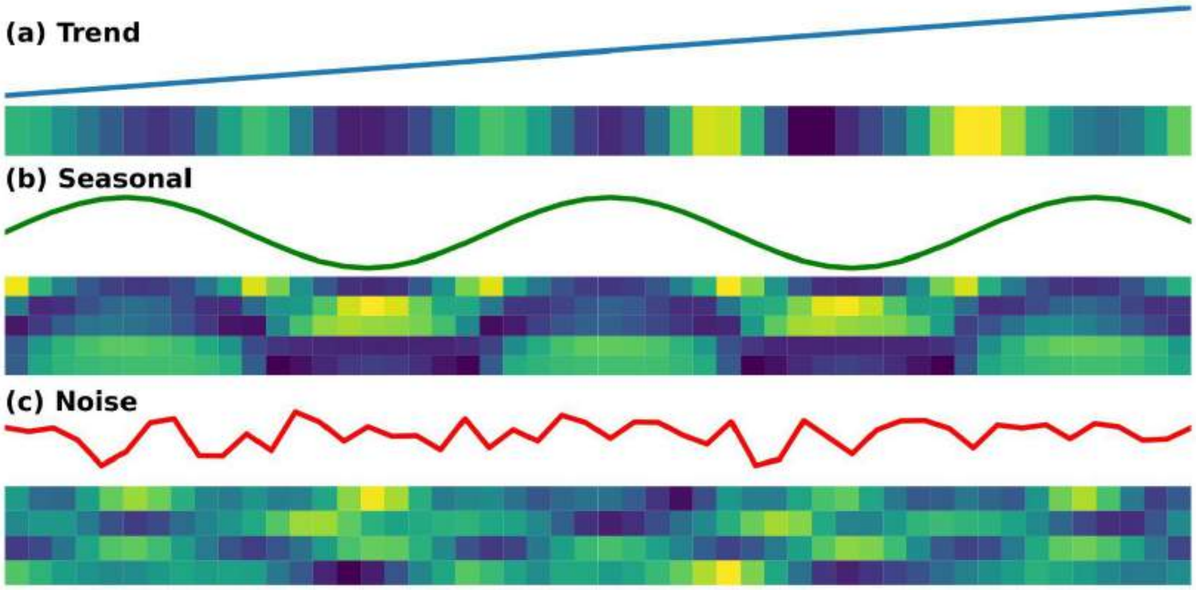
In Production!



Adapted from
EQT Motherbrain
Platform

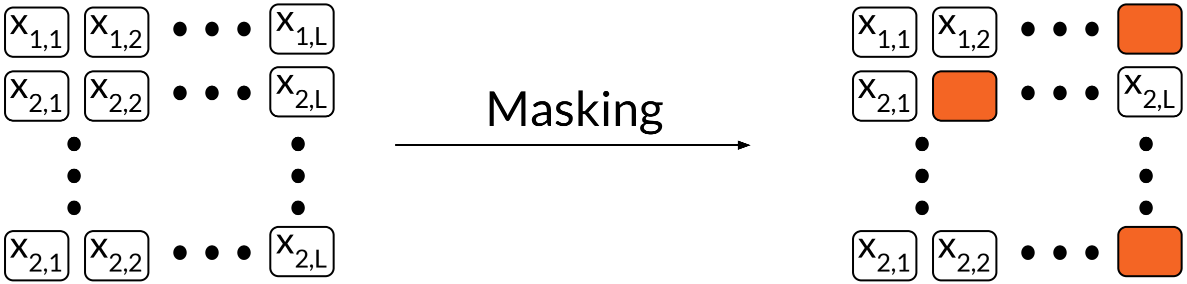
The State-of-The-Art: TSDE [1]

- Learn an **embedding** for (multiple) time series.



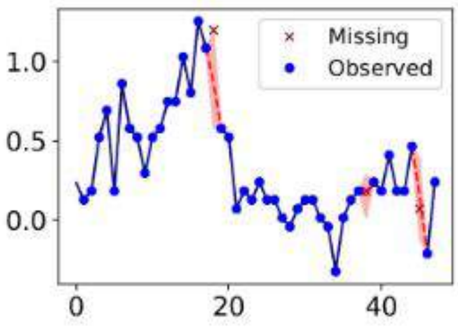
The State-of-The-Art: TSDE [1]

- Learn an **embedding** for (multiple) time series.
- It is achieved by learning to recover the **masked** part.

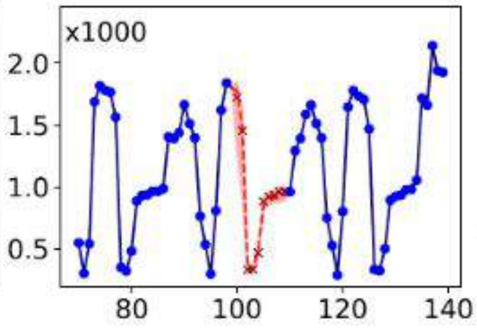


The State-of-The-Art: TSDE [1]

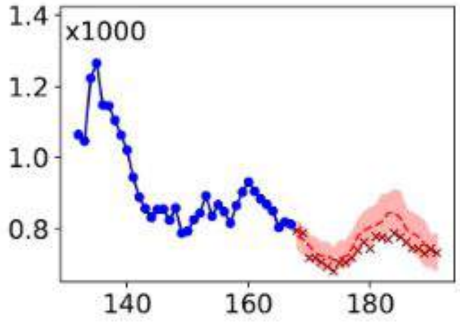
- Learn an **embedding** for (multiple) time series.
- It is achieved by learning to recover the **masked** part.
- The embedding can be used for many tasks such as **forecasting, interpolation, imputation, anomaly detection, classification, clustering** ...



imputation



interpolation



forecasting

The State-of-The-Art: TSDE [1]

- Learn an **embedding** for (multiple) time series.
- It is achieved by learning to recover the **masked** part.
- The embedding can be used for many tasks such as **forecasting, interpolation, imputation, anomaly detection, classification, clustering** ...

Paper:
https://arxiv.org/pdf/2405.05959

Github (source code):
https://github.com/EQTPartners/tsde

Agenda

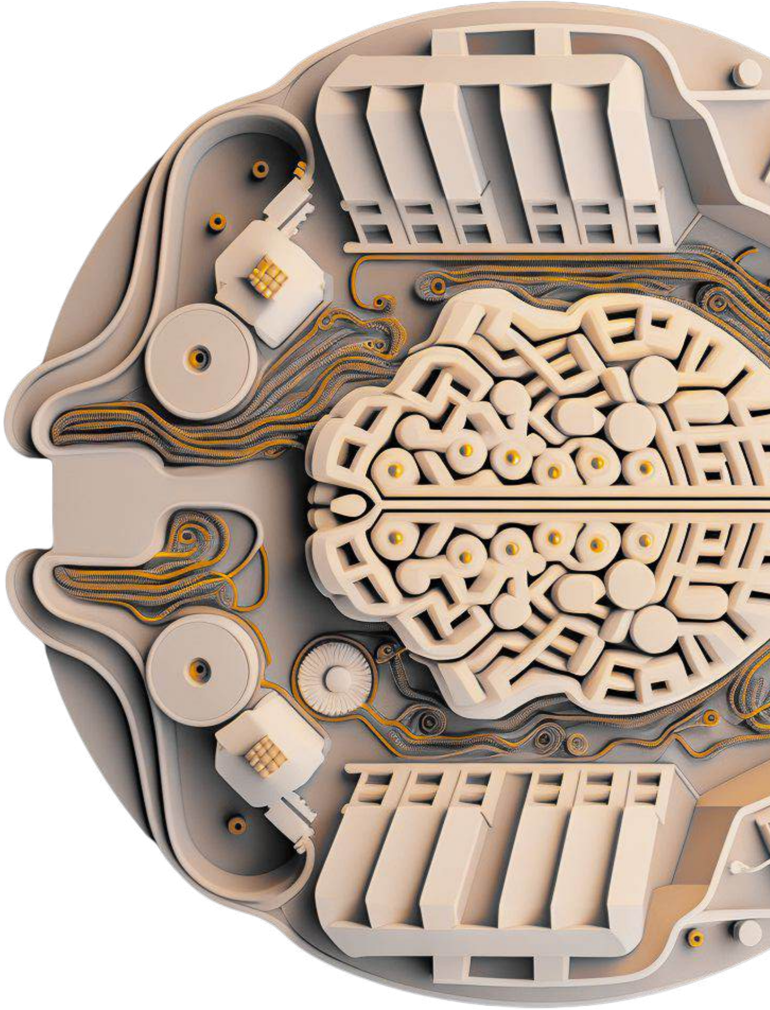
Success Prediction: Deep Learning

Sector Prediction: Large Language Model

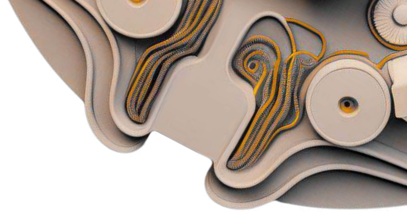
Revenue Forecasting: Classic and State-of-The-Art

Document Mining: Knowledge Graph

Summary



Document Mining



YOU ARE HERE — YOU ARE HERE — YOU ARE HERE — YOU ARE HERE

Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

Due Diligence

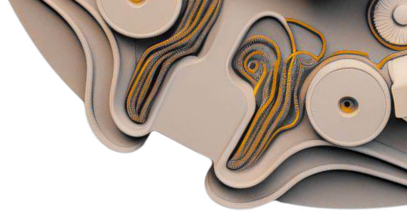
- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?



Document Mining

Sourcing

- Which companies exists within my mandate?
- How do I make sure to spend time on interesting opportunities?
- What does the market look like around a deal?

Due Diligence

- What are the risks and attractions in this opportunity?
- What is the full potential this company can achieve?
- Who do we know we can ask for insights?

Holding

- How can we make sure to fulfil the full potential plan?
- What are the potential disruptions and market trends?
- How are we actually tracking on the plan?

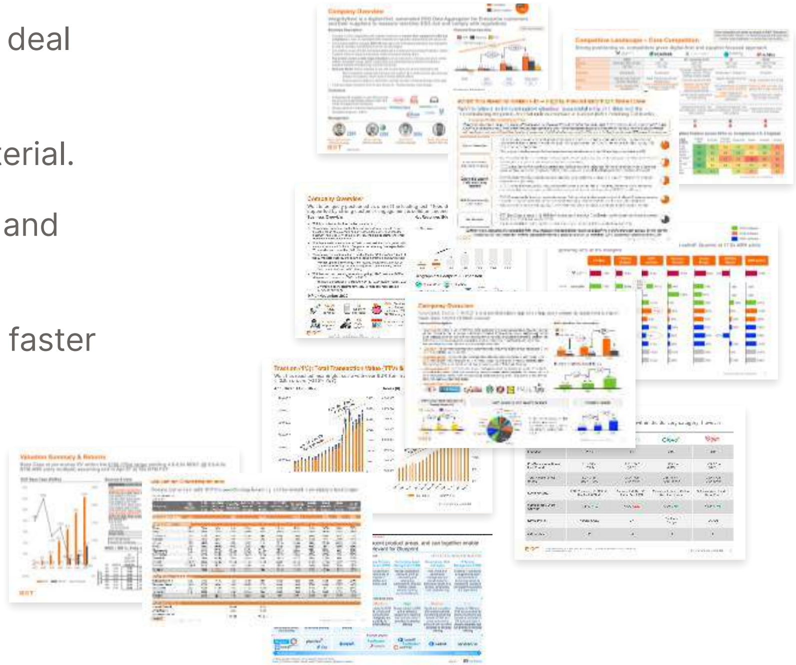
Exit

- How do we position the deal up for a successful sale?
- How can we make sure the company is in good health and can continue to grow?
- Which bankers and advisors should I talk to?

Motivation: leverage proprietary knowledge

Proprietary Knowledge

- So much interesting information is gathered throughout a deal process.
- EQT has been doing this since 1994... there is a lot of material.
- Reading through material is pretty intense as most decks and market reports are very dense. And long...
- But understanding historical deals is very important to be faster and smarter in future deal assessments.



Approach: objective and applications

Develop an Intelligent and Knowledgeable Agent/Service to help deal professionals in many applications:

Ask a question about

- Mergers & Acquisitions
- Competitor Analysis
- Market Mapping
- Due Diligence Support
- Portfolio Company Growth Strategy
- Investment Thesis Formulation
- … …

Answer
here are some suggestions for deal advisers … …

Approach: a concrete example

Develop an Intelligent and Knowledgeable Agent/Service to help deal professionals in many applications:

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Approach: LLM prompting

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

→

LLMs
ChatGPT
LLaMA
Gemini
Mistral
Claude
...

→

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Approach: LLM prompting

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

LLMs
ChatGPT
LLaMA
Gemini
Mistral
Claude
...

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Problem: Lack domain-specific and up-to-date knowledge.

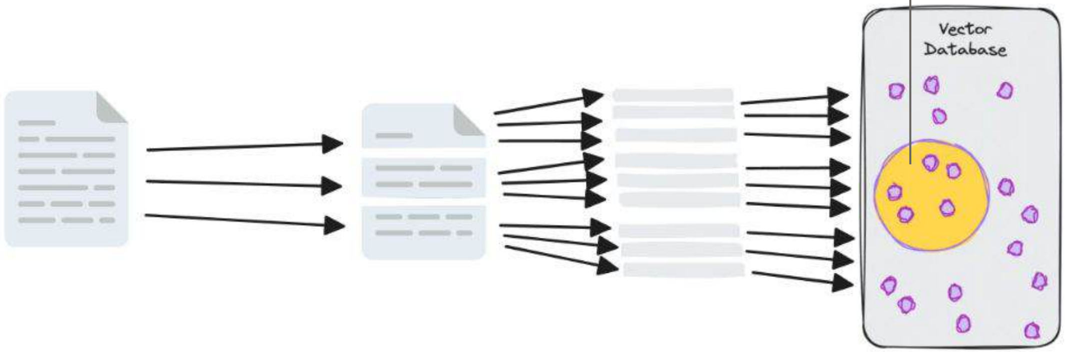
Approach: RAG - retrieval augmented generation

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

LLMs
ChatGPT
LLaMA
Gemini
Mistral
Claude
...

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Retrieve document slices



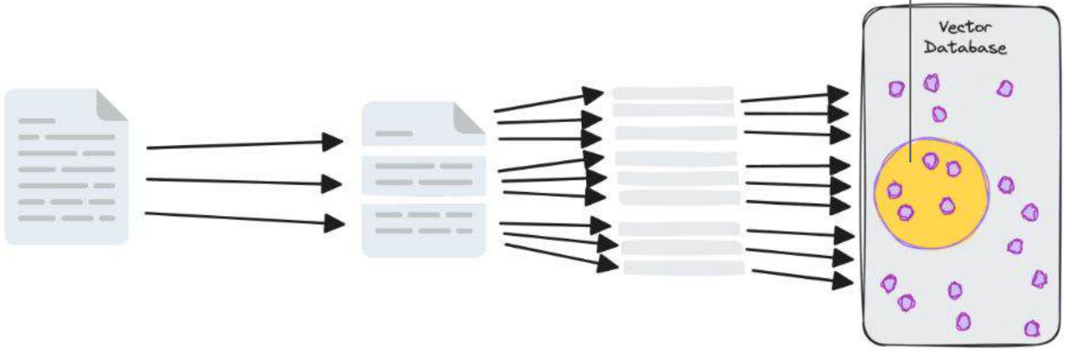
Approach: RAG - retrieval augmented generation

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

LLMs
ChatGPT
LLaMA
Gemini
Mistral
Claude
...

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Retrieve document slices



Problem:
- not modeling for PE-specific entities and their relationships;
- hard to integrate the third-party knowledge.

Approach: RAG over KG (knowledge graph)

Ask a question
We are looking at a deal of a company named ManyPets (manypets.com). Can you suggest some deal experts in related sectors and also some direct competitors in EU?

→

LLMs
ChatGPT
LLaMA
Gemini
Mistral
Claude
...

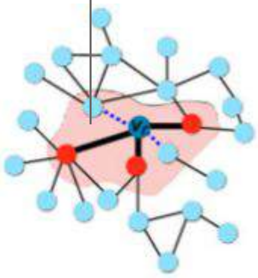
→ **Reasoning** →

Answer
here are some suggestions for deal advisers in the pet insurance sector, as well as some direct competitors in the EU:
...

Retrieve relevant sub-graph

Proprietary Knowledge (Unstructured Docs)

Construct →



Knowledge Graph

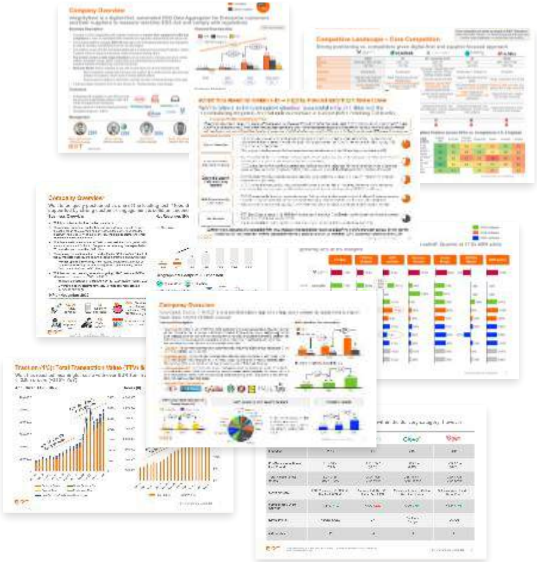
Approach: RAG over KG (knowledge graph)

Three Key Components:

1. **KG Construction** - Extract relevant entities, relations, and attributes from **proprietary documents** and **third-party data**.
2. **Contextual Retrieval** - According to the context provided by the query/question, retrieve the relevant sub-KG for LLM to reason about.
3. **Reasoning** - With the query/question and the retrieved sub-KG, generate the response/answer.

KG Construction

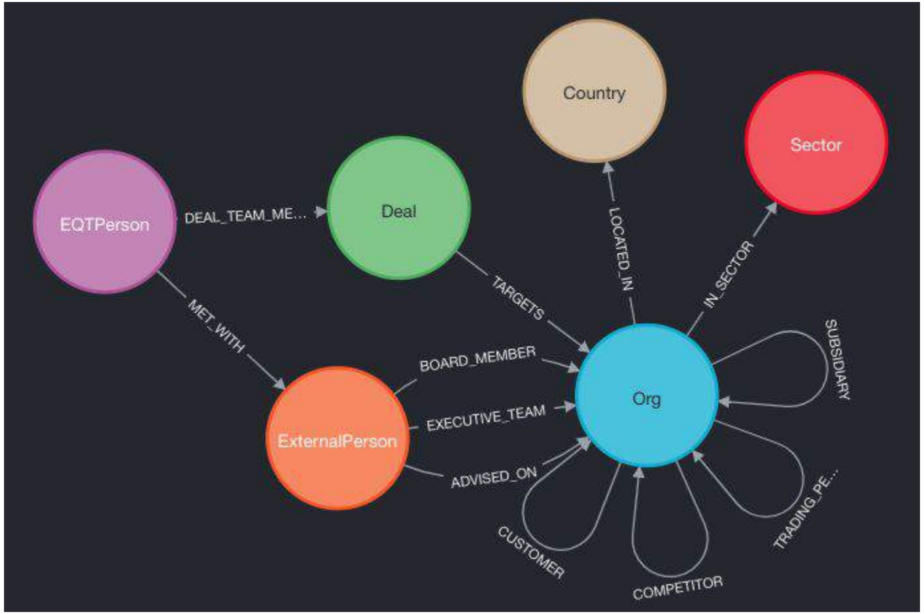
- **Purpose**: Construct Knowledge Graph for EQT's deals.
- **Data source**: EQT's proprietary deal related documents. Each document is about a specific company (a.k.a., target company) in scope.



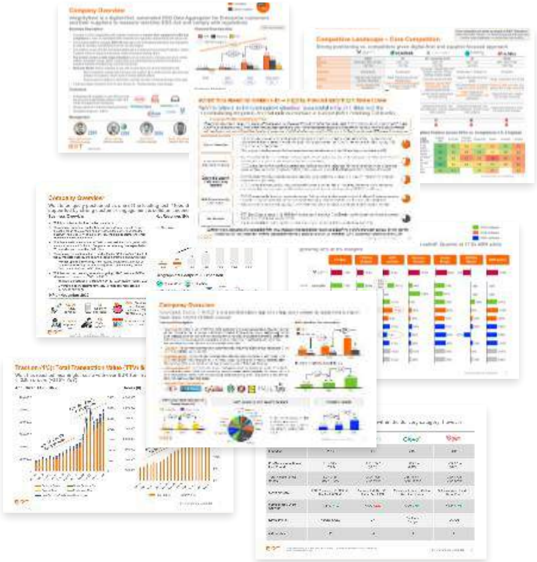
EQT's deal docs

KG Construction

- What entities and relations we extract and build into PEKG?



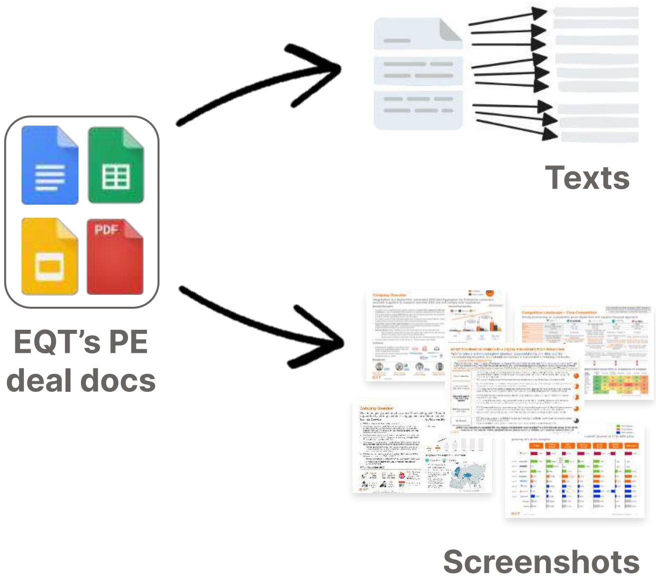
Extraction Guide
(Meta Graph)



EQT's
deal docs

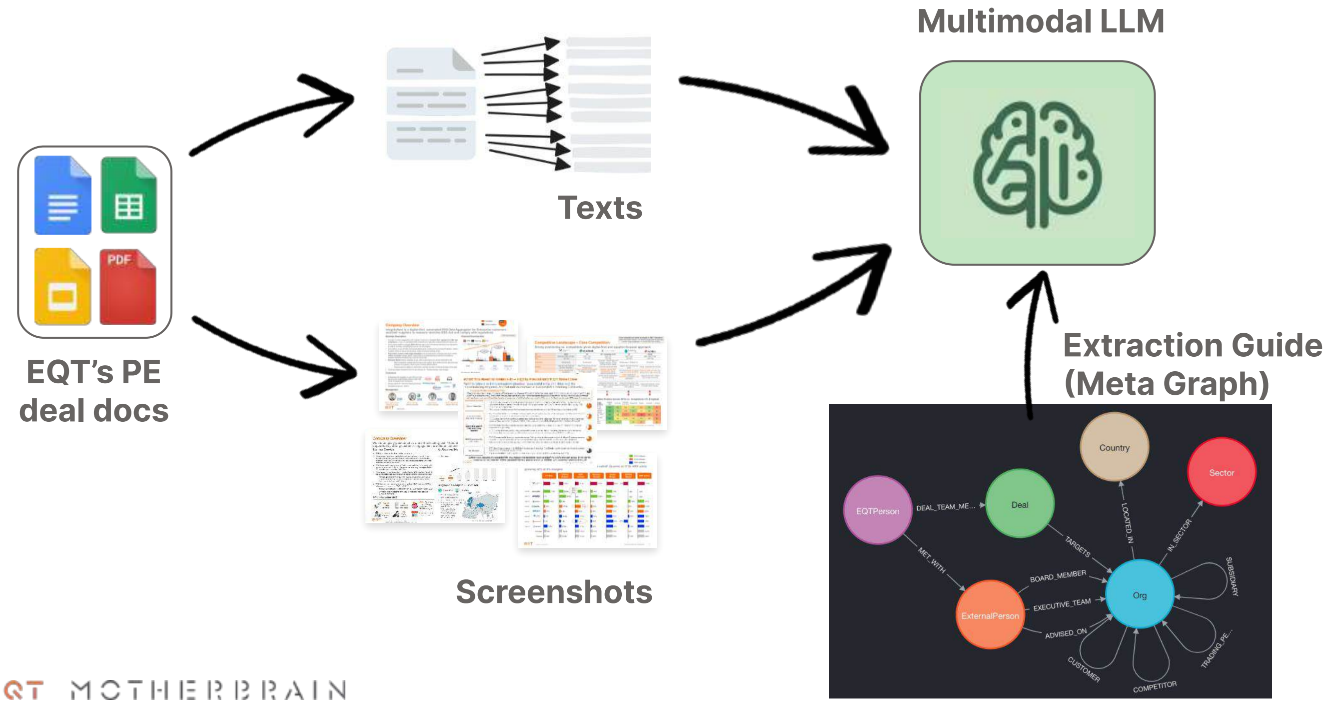
KG Construction

- **How do we automate PEKG construction?**



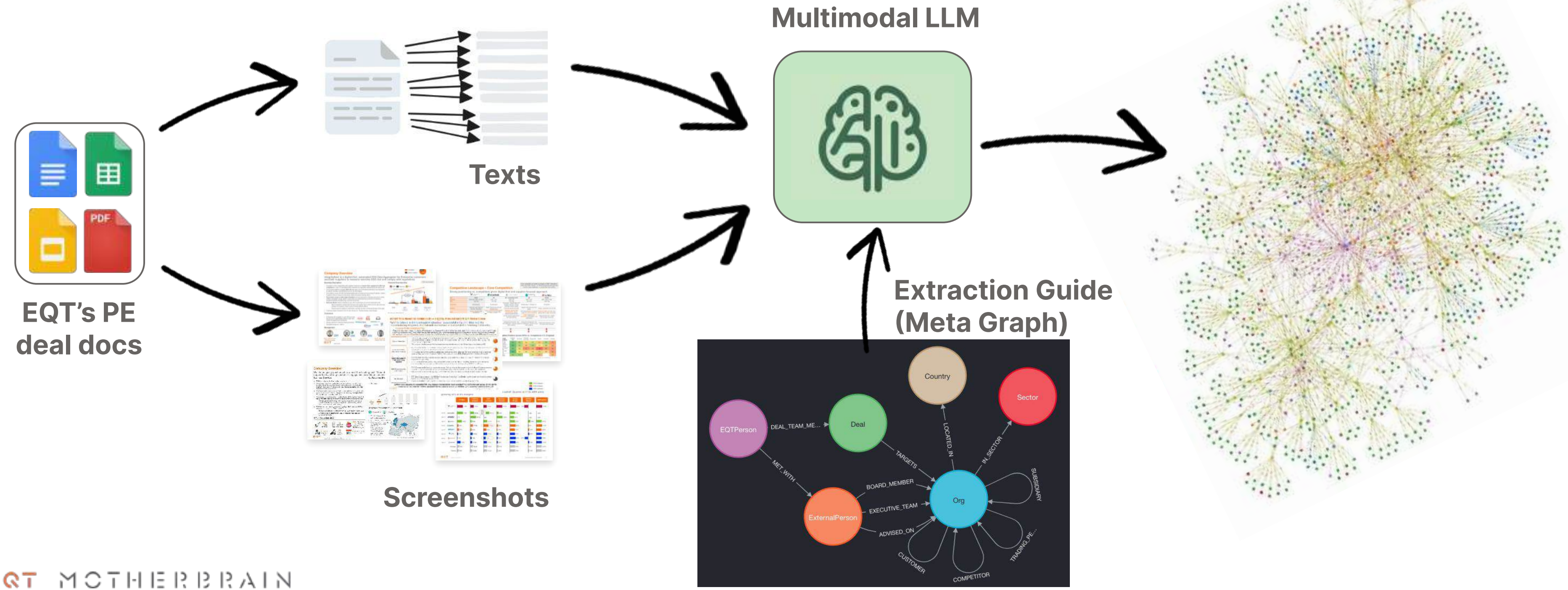
KG Construction

- How do we automate PEKG construction?



KG Construction

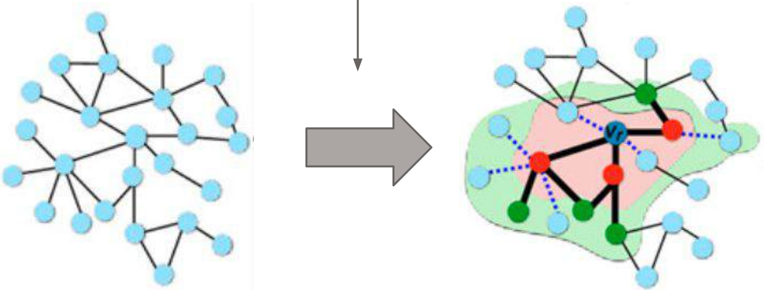
- How do we automate PEKG construction?



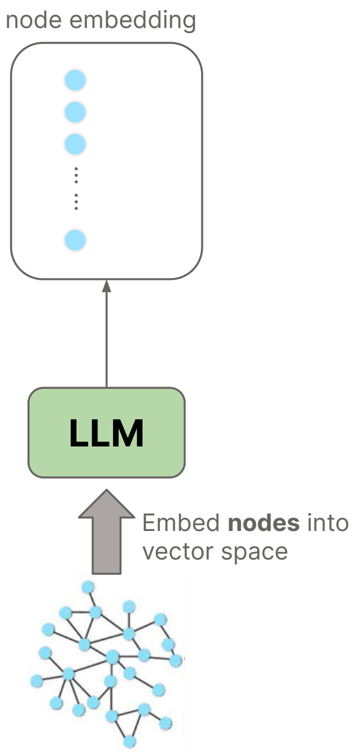
Contextual Retrieval

Objective: retrieve a sub-KG from the entire KG according to the context of the input query/question.

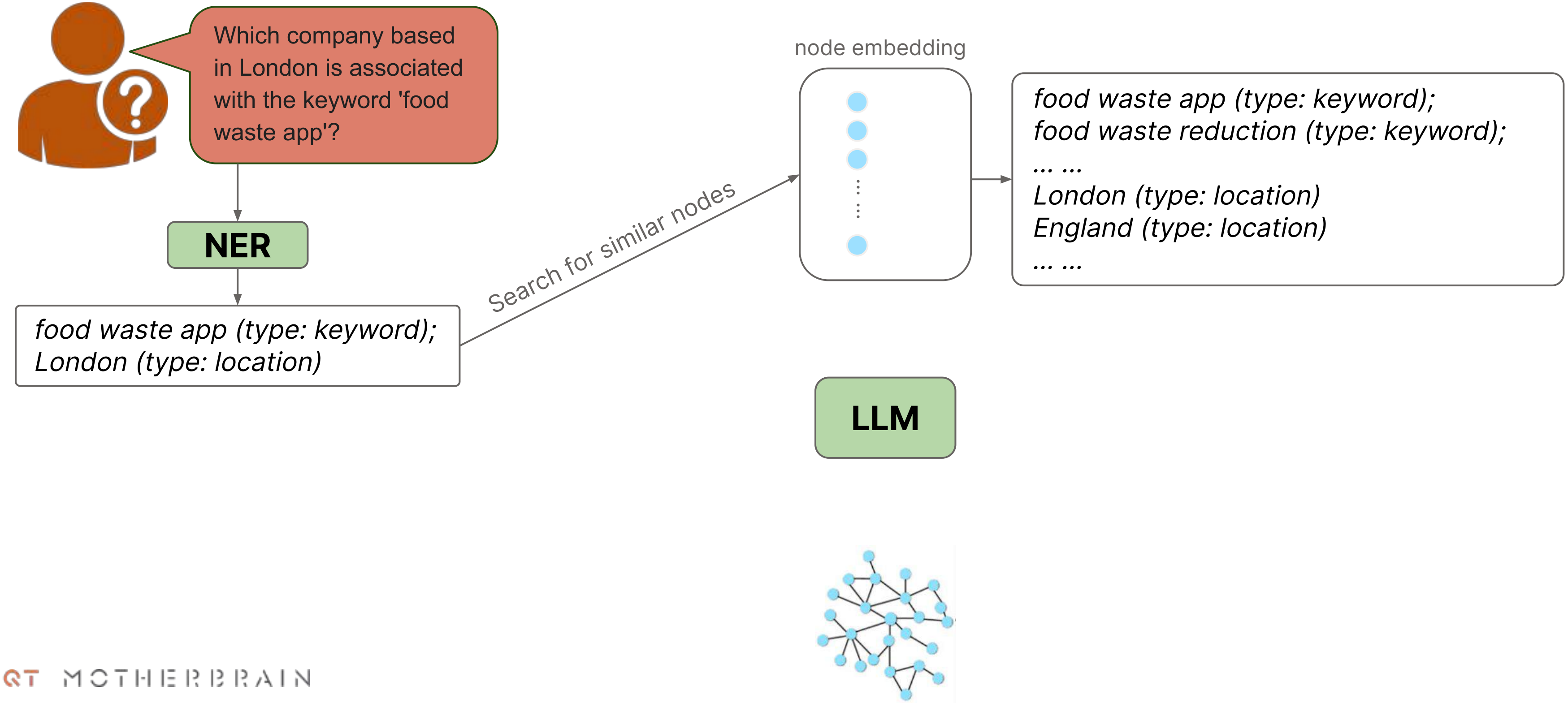
- The existing work mostly
  - assumes the availability of a contextual sub-KG, which is not true in reality;
  - or adopt a overly simplified approach, such as randomly expand 2 steps from a center node.



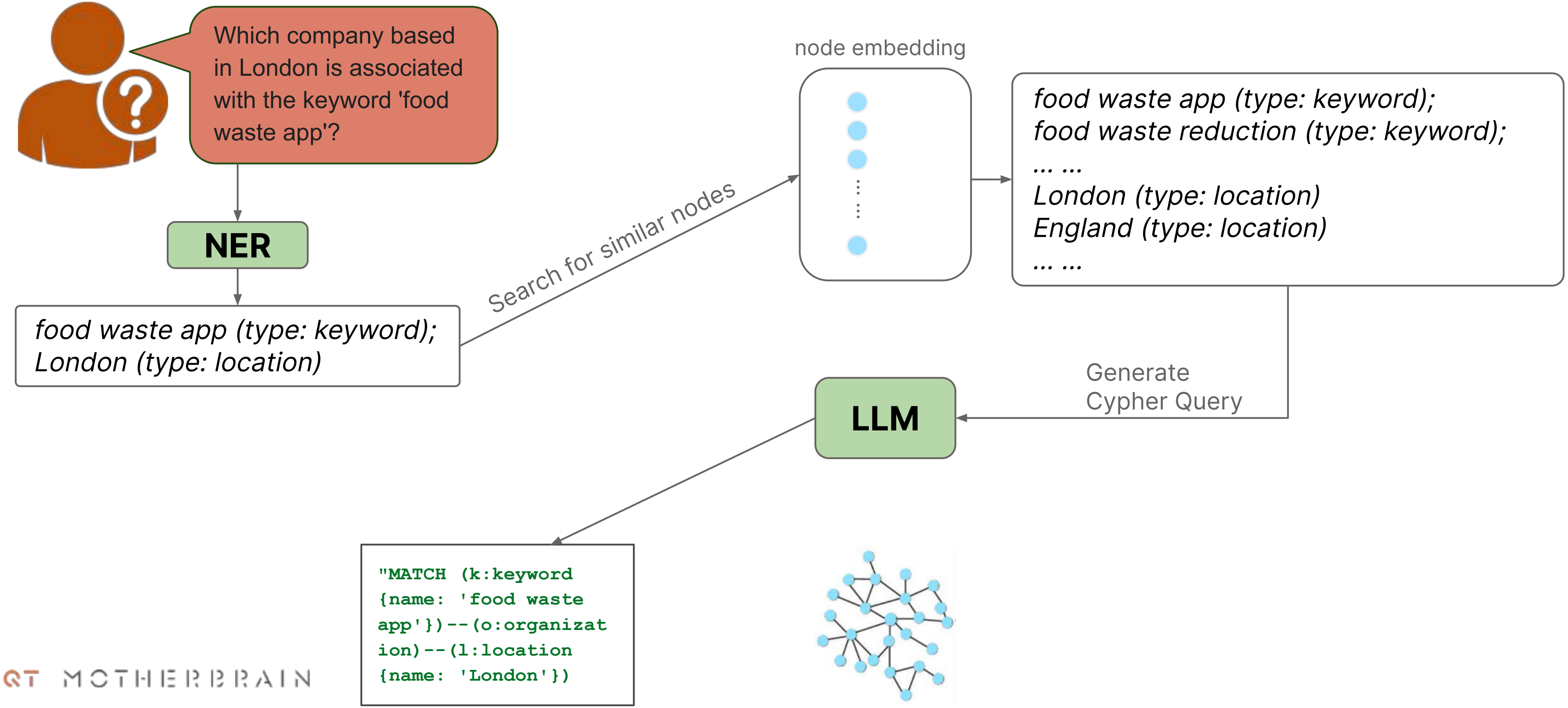
Contextual Retrieval



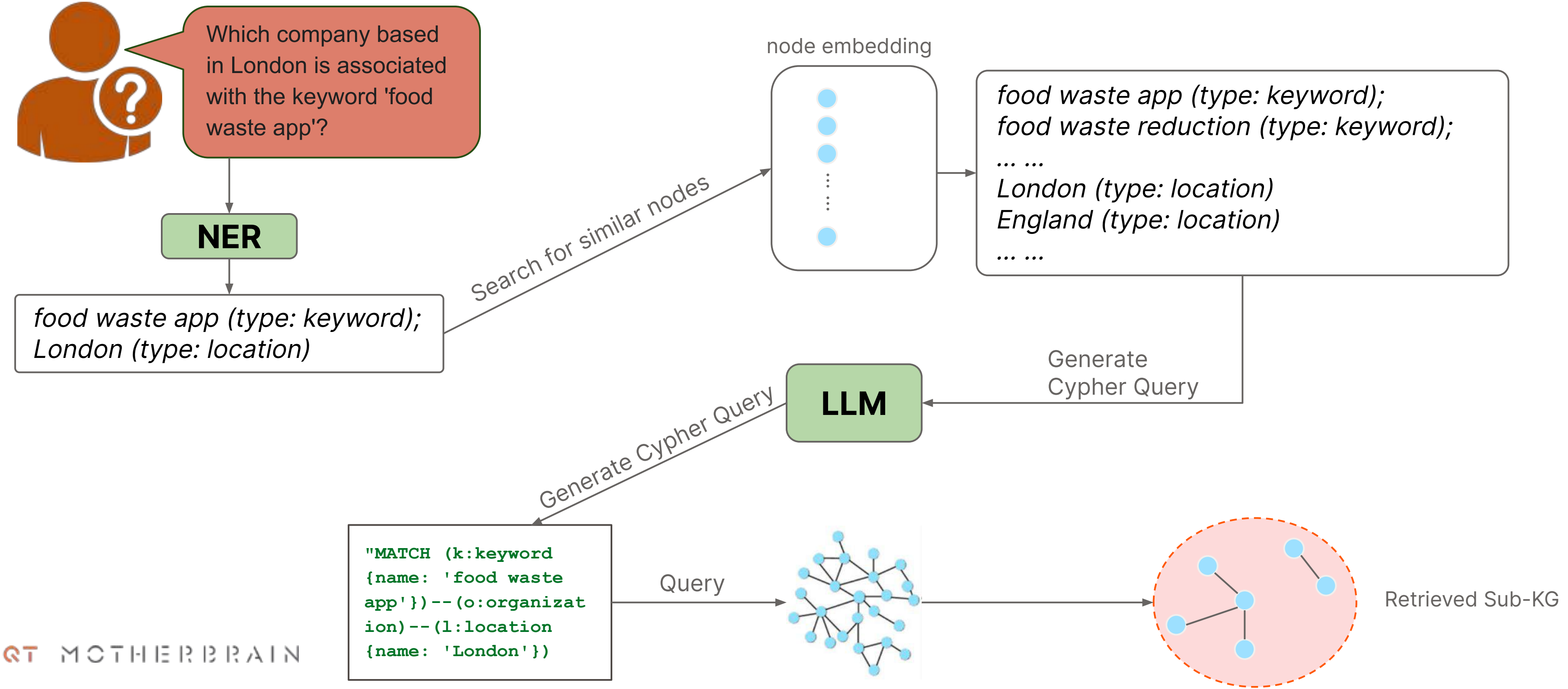
Contextual Retrieval



Contextual Retrieval

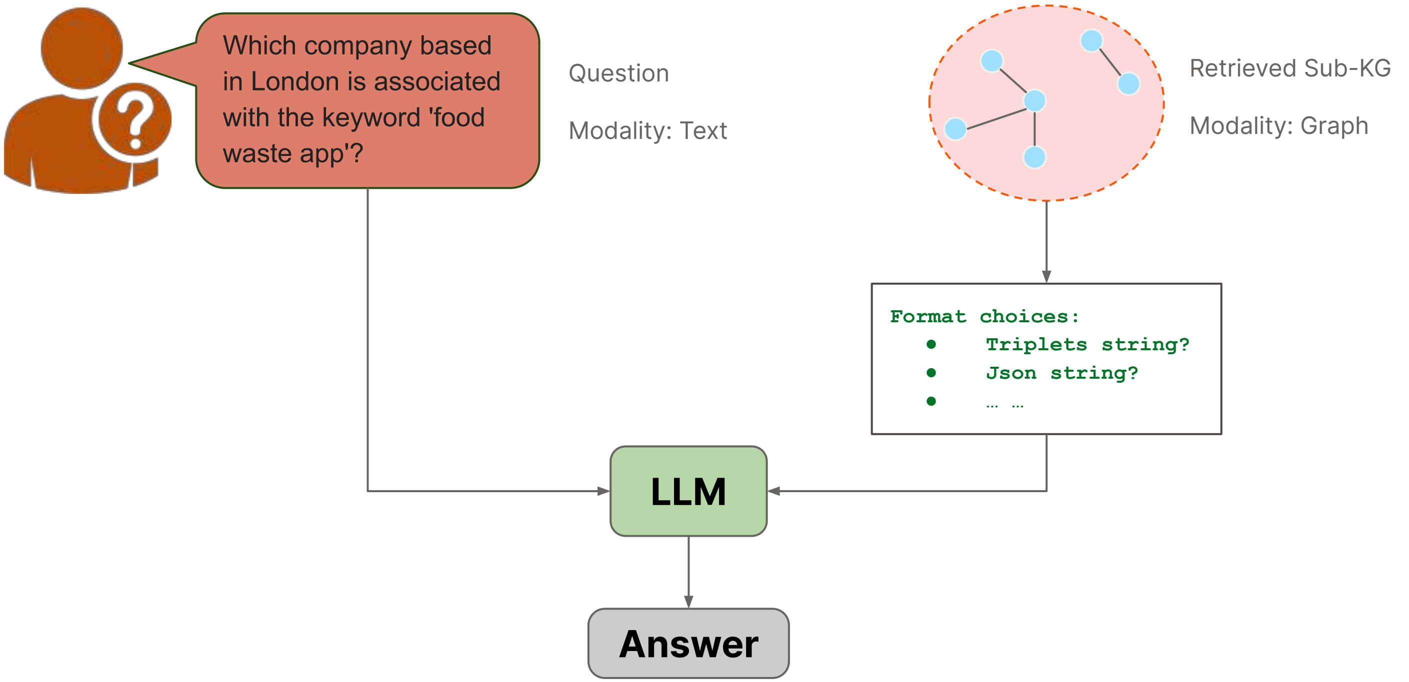


Contextual Retrieval



Reasoning: one simple solution

Objective: generate answer to the textual question using the retrieved sub-KG as input context.



Agenda

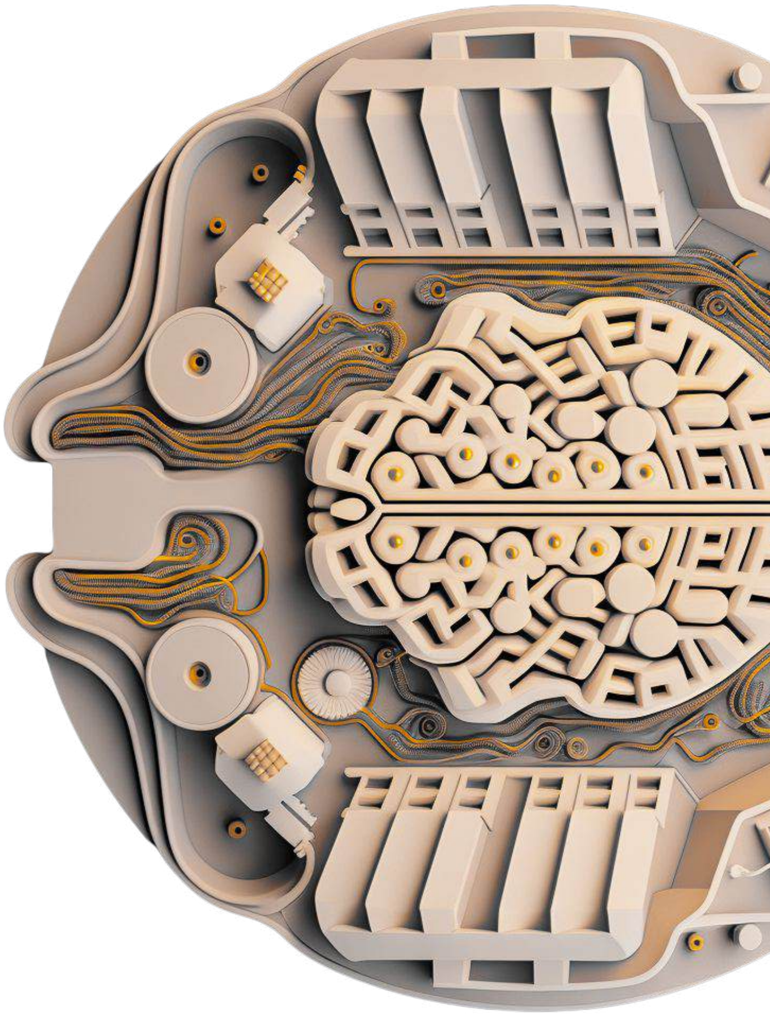
Success Prediction: Deep Learning

Sector Prediction: Large Language Model

Revenue Forecasting: Classic and State-of-The-Art

Document Mining: Knowledge Graph

Summary



Thanks!

- For further interaction, feel free to ping me on LinkedIn:
  www.linkedin.com/in/caolele

- Or, email us: tech_motherbrain-research@eqtpartners.com

- Learn more about EQT Motherbrain at:
  https://eqtgroup.com/motherbrain
  https://motherbrain.ai/