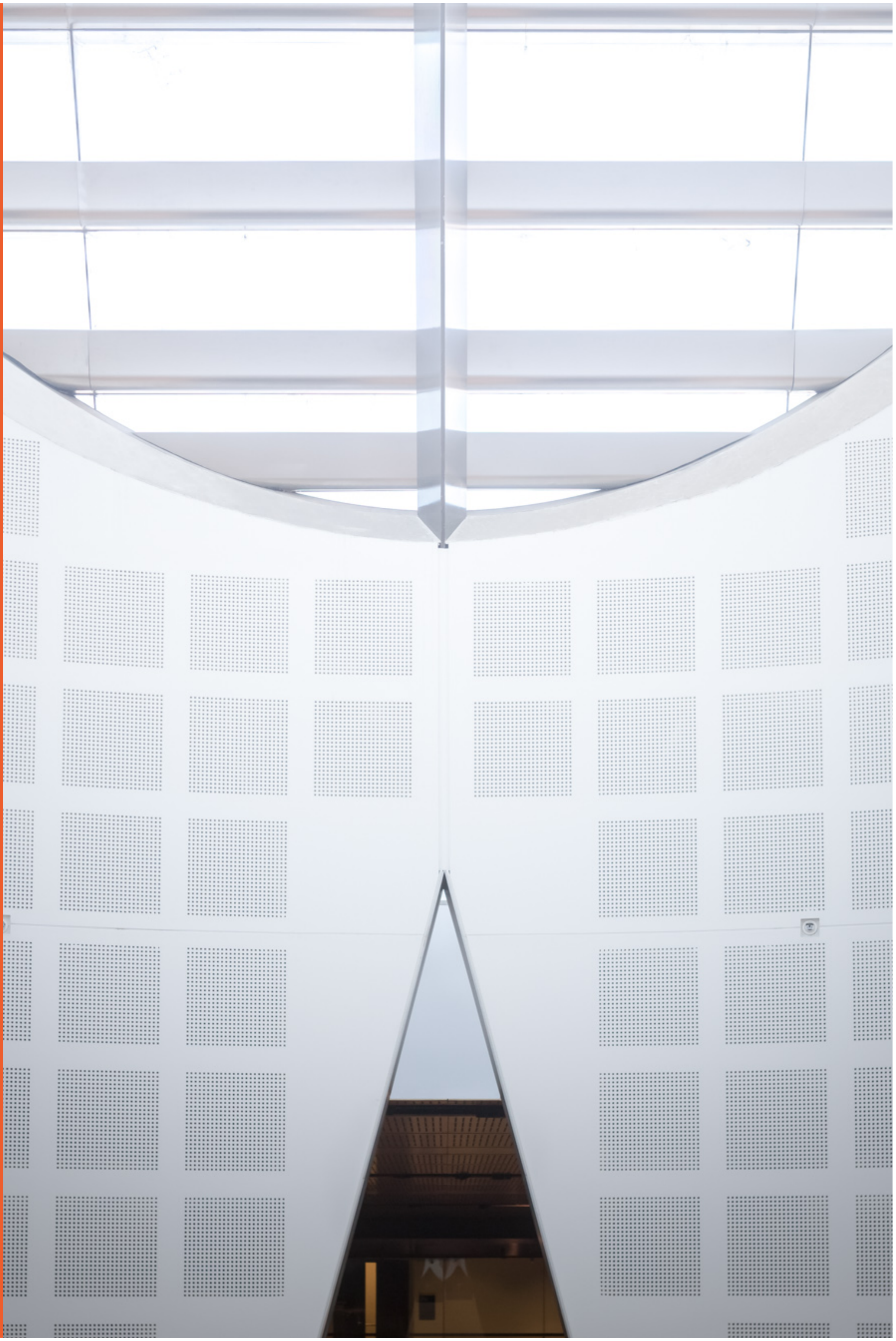
Computational News Project Proposal

Presented by

Chang Li

Ph.D. @ UBTECH Sydney AI Center,
University of Sydney,
CMCRC

THE UNIVERSITY OF SYDNEY



Summary

- Low Business Value: Data Infrastructure

- High Business Value: New Information Generation

- Invest in Deep Learning NOT Software Engineering

- Proposed Projects:
  - Specific, Light Area; Fast, Adaptive Iteration
  - Event Driven Trading
  - Sentimental Analysis

Topics

- Nature of Knowledge Base
- Business Value
- Computational News Project
- Our Proposal

Nature of Knowledge Base

Relationships
between
Entities

Supergraph

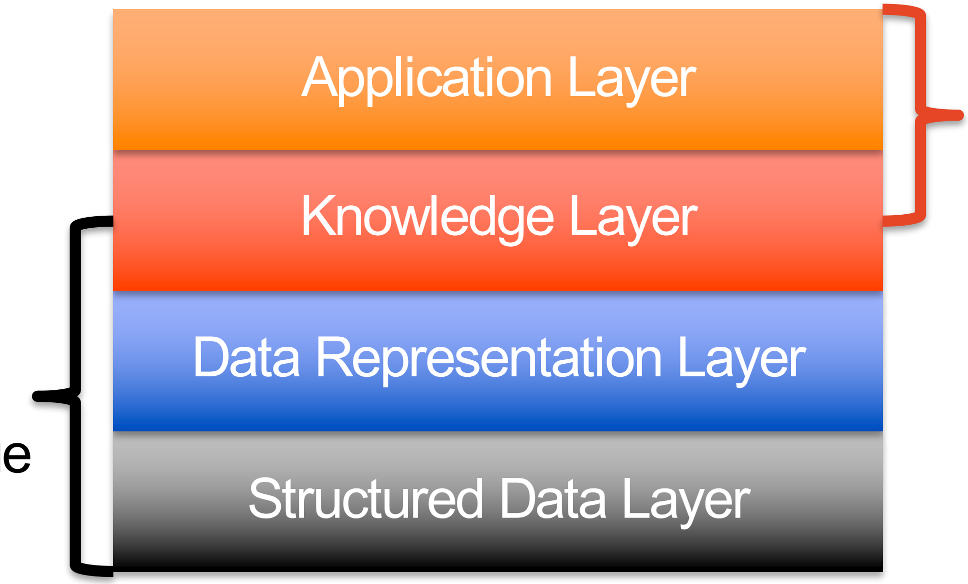
- Frequencies
- Sentiment
- Logical Inference
- Reasoning

Business Value of Knowledge Base

35+ Companies

MEMECT
Quandl
alphasense
KENSHO
iSENTIUM
Dataminr
CEREBELLUM CAPITAL
sentient

Business Value of Knowledge Base



Application Layer

Knowledge Layer

Data Representation Layer

Structured Data Layer

New Information

High Business Value

Domain Invariance

Extending Human Capabilities

Efficiency

Low Business Value

Low Motivation

Sensitive to Domains

Customer Education

Business Value: Efficiency

Difficulties

- Very Large Scale
- Unstructured Dataset

Solutions

- High Performance Computing
  - Hardware / Software Infrastructure
  - Highly Optimized Engineering
- Human Annotation

Business Value: Efficiency

Quandl

- Large Scale Datafeed
- Consumer transactions, cargo movement, employment trends
- BV: Unstructured Data -> Structured

AlphaSense

- Linguistic Search Engine
- Synonyms; Summary
- Over 10k data sources (financial reports & business terms)
- Strong sales team
- BV: Efficiency

MEMECT (China)

- Financial Knowledge Base
- Corp Info; Report Generation
- BV: Efficiency

Business Value: New Information

Impossibilities:

Extending Human Capabilities

- Very Large Scale
- Very Short Time Interval
- Very Complex Relationships

Difficulties:

- Knowledge Representation
- Logical Inference
- Reasoning

Solution:

- Deep Learning

Business Value: New Information

Dataminr

- High impact events from twitter etc.
- PR; Corp Alerts; Fin Info
- 230+ Engineers
- BV: Info before in news

Yu Qing Tong (Weibo)

iSentium

- Sentiment Indicator from twitter, stocktwits etc.
- Indicator for hedge funds
- BV: Complex Relationships

PS: Structured Dataset

Kensho

- Financial Knowledge Base
- PR; Corp Alerts; Fin Info
- Acquisited by S&P @ 550M
- 600 Engineers @ 120K / year
- BV: Undiscovered events – asset price relationship

PS: Quandl is a datafeed

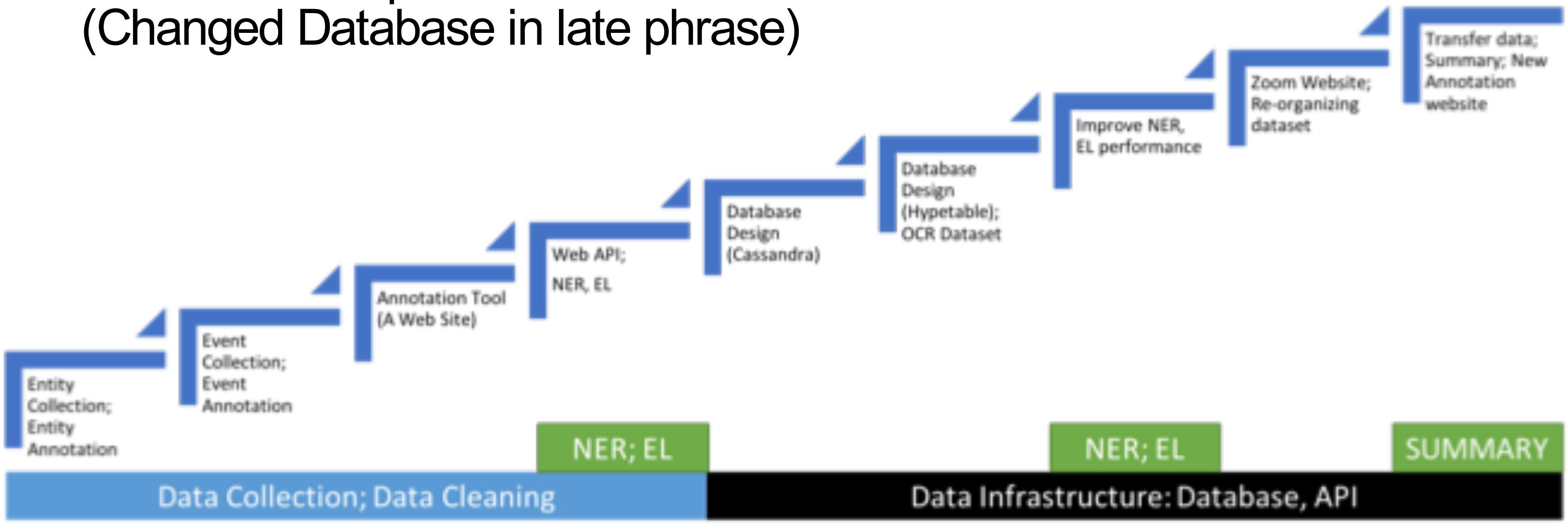
Topics

- Nature of Knowledge Base
- Business Value
- Computational News Project
- Our Proposal

Computational News Project

Waterfall Development Cycle:

- Clear (Static) Business Requirements
- Long delivery term
- ~70% Effort Spent on Data Infrastructure
- Hard to be adapted to new data source (Changed Database in late phrase)



Entity Collection; Entity Annotation

Event Collection; Event Annotation

Annotation Tool (A Web Site)

Web API; NER, EL

Database Design (Cassandra)

Database Design (Hypetable); OCR Dataset

Improve NER, EL performance

Zoom Website; Re-organizing dataset

Transfer data; Summary; New Annotation website

NER; EL

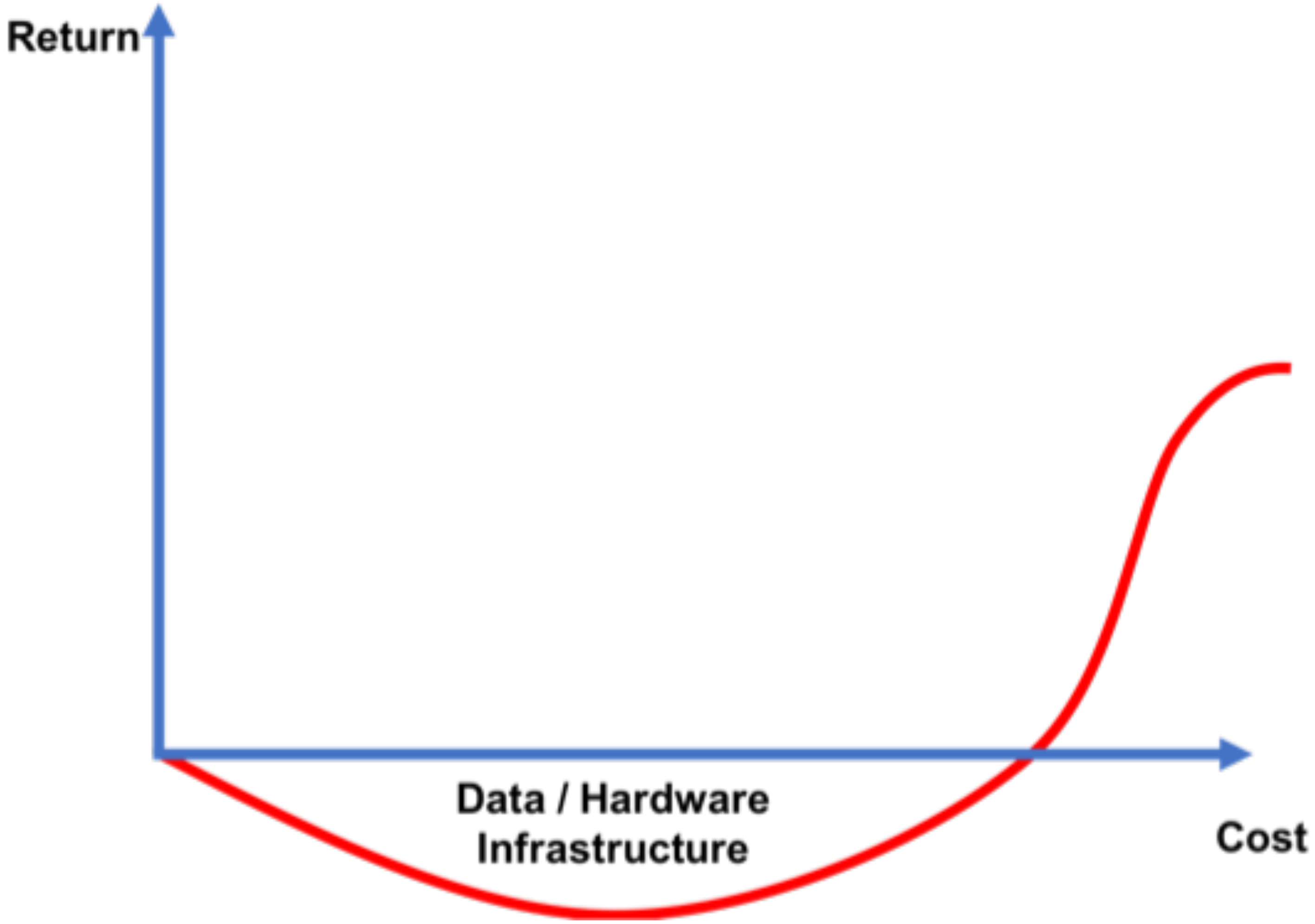
NER; EL

SUMMARY

Data Collection; Data Cleaning

Data Infrastructure: Database, API

Computational News Project



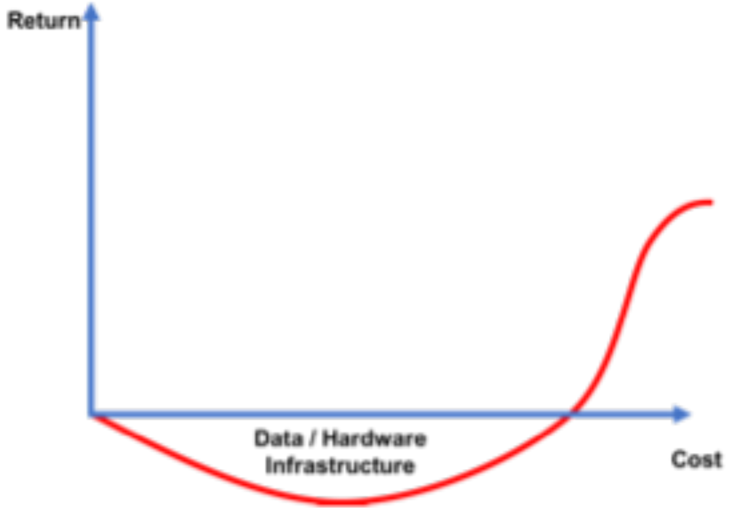
Computational News Project

- Long product delivery term (2~3 years)

- High cost
  - Massive human resources
    - $150000 = 1 Big Data Engineer
    - Annotation
  - Massive hardware cost
  - Large selling team

- Unclear business model
  - Customer requirements are elusive yet
  - Hard to be adaptive to changes

- Low business value
  - Expensive selling cost
  - Low product return



Topics

- Nature of Knowledge Base
- Business Value
- Computational News Project
- Our Proposal

Our Proposal

Impossibilities

- Very Large Scale
- Very Short Time Interval
- Very Complex Relationships

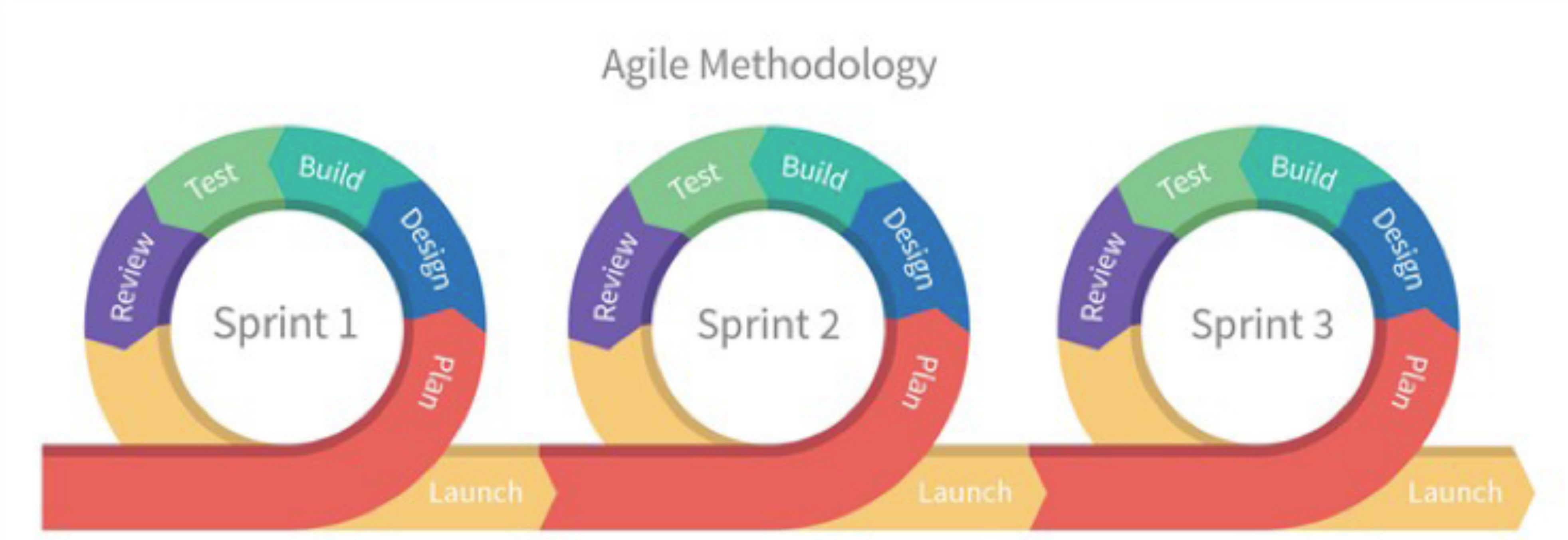
Objectives

- Minimal Data Infrastructure
  - Minimal Training Data (Annotated Data)
  - Minimal Data Sources (Structured Data Source)
  - Minimal Software Engineering
- Maximal Business Value
  - Maximal Relationship Complexity

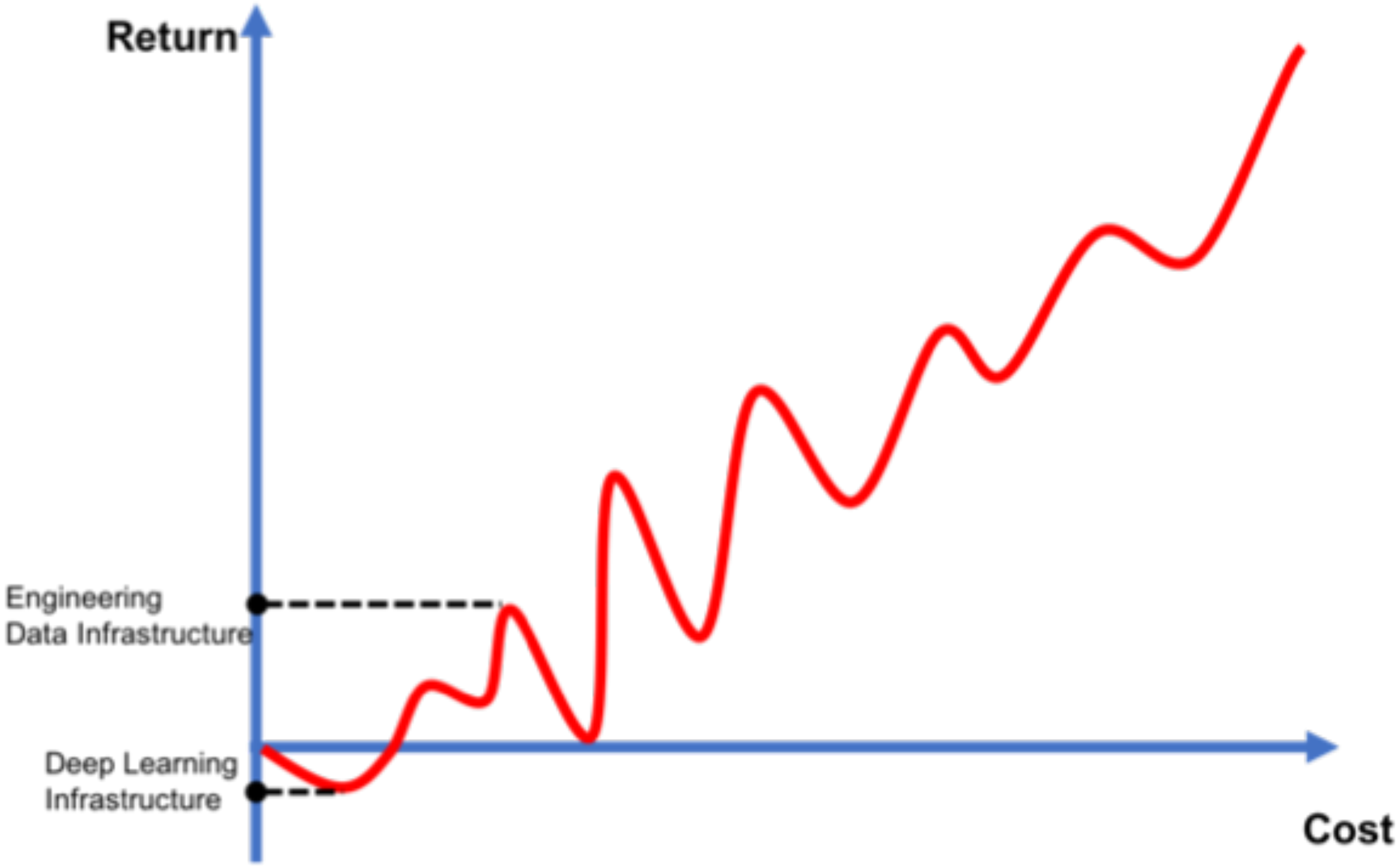
Projects

- Event Driven Trading
- Sentimental Analysis

Our Proposal: Agile Development Cycle



Our Proposal: Agile Development Cycle



Our Proposal: Event Driven Trading

Pros:

- ALTA 2015 1st Place (En – Fr Cognates)
- ICDM Business Chain Prediction Paper
- 4~6 Months Proof of Concepts = MVP
- Structured / Single Data Source
- Self-Annotated Data; Semi / Unsupervised Learning
- Results Transferable to Other NLP Tasks

Cons:

- 4~6 Months Proof of Concepts

Our Proposal: Sentimental Analysis

Pros:

- Proven Business Model
- Potential MQD Product
- Clear Customer Requirements
- Structured / Single Data Source
- Self-Annotated Data; Semi / Unsupervised Learning

Cons:

- Competitors
- Hard to differentiate

Appendix: ICDM 2017 Paper

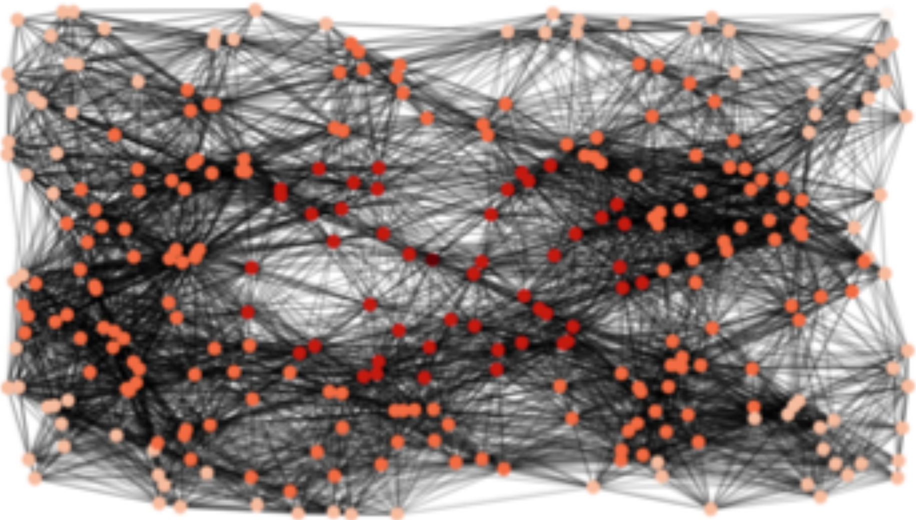
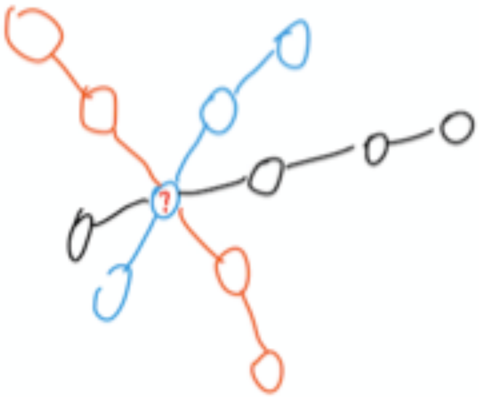


Figure 3: Markov Random Fields generated by data and algorithm described in section 3.1.



Supergraph:

Cooperation Relationship (Static)

Price Movement:

Joint Inference in Business chain (Dynamic)