Applied Research in Information Science and Technology

Sarana Nutanong

VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY

NSTDA Chair Professor Grants
June 29, 2018

What This Talk is and is Not

- Not an Elon Musk worshiping session
  - Approach me after the talk if you wish to join the Church of Elon

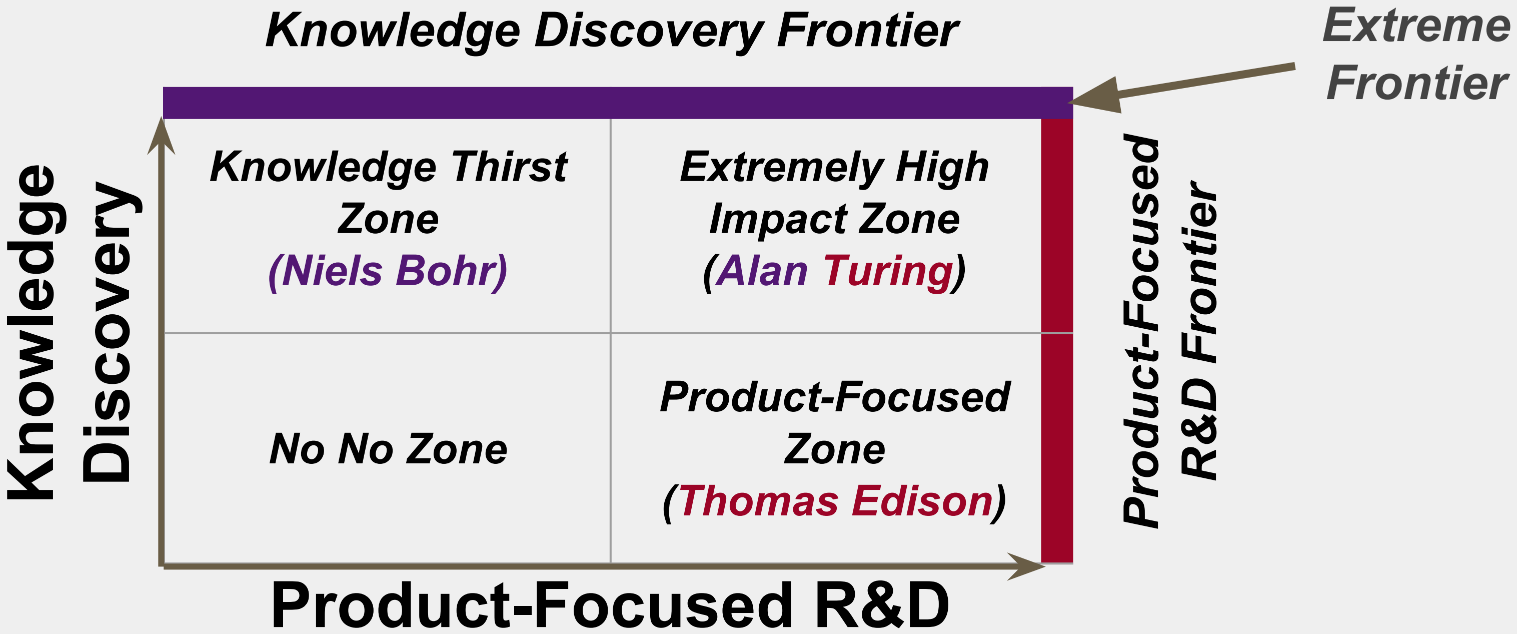
Is:

- Comparing and contrasting upstream, midstream, and downstream research in Information Science and Technology
- Pointing out connections between upstream, midstream and downstream research
- Advocating midstream research as our starting point

Knowledge Discovery VS Product-Focused R&D

	Knowledge Discovery	Product-Focused R&D
Target	Creating new generalizable knowledge	Creating commercial competitive advantage
Problem too easy?	Make it more difficult or find a new problem to work on	No one thought of this before. Really!? Let’s solve it and get rich.

Knowledge Discovery VS Product-Focused R&D



Knowledge Discovery Frontier

Extreme Frontier

Knowledge Discovery

Knowledge Thirst Zone (Niels Bohr)

Extremely High Impact Zone (Alan Turing)

No No Zone

Product-Focused Zone (Thomas Edison)

Product-Focused R&D Frontier

Product-Focused R&D

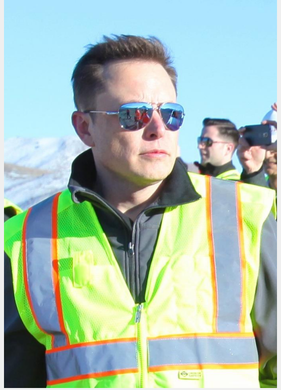
Product-Focused R&D

Without Knowledge Discovery Research

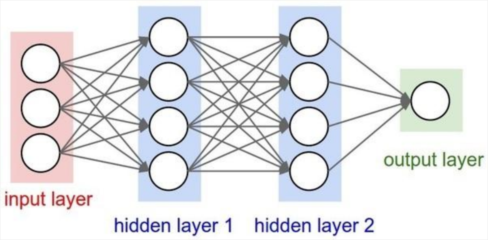
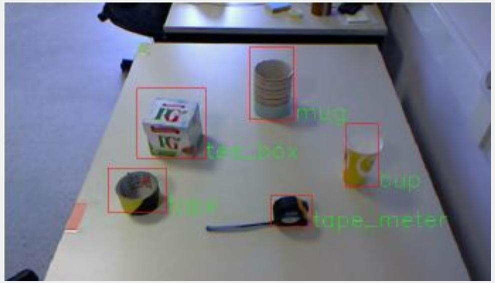
- Think of an *easy* problem *no one has thought of before*
- Sove it
- Get rich

With Knowledge Discovery Research (Deep Tech)

- Gain competitive advantage through results obtained from frontier knowledge discovery research
- ***Hire experts graduated from research labs with the required knowledge to work for you***
- Solve the problem with even more research
- ***Get bloody rich and look bloody smart!***



Driver Assistance System

Upstream			Midstream			Downstream		
1	2	3	4	5	6	7	8	9
Neural Networks			Object Recognition			Drive Assistance		
								

input layer
hidden layer 1
hidden layer 2
output layer

mug
tea_box
cup
tape
tape_meter

LEFT REARWARD VEHICLE CAMERA
MEDIUM RANGE VEHICLE CAMERA
RIGHT REARWARD VEHICLE CAMERA

Pictures: https://www.quora.com/What-is-a-simple-explanation-of-a-recurrent-neural-network; https://www.researchgate.net/figure/Real-time-object-recognition-with-disparity-based-segmentation_fig6_261844022

Interface Points

- Upstream-Midstream
  - Understand how to design and improve neural network architectures
  - Understand object recognition problem characteristics

- Midstream-Downstream
  - Understand the requirements of drive assistance problems
  - Understand how object recognition can be used to improve driving assistance systems

Applied ML Research

- Midstream
- Application of existing ML methods to domain-specific problems
- Still knowledge-focused
  - Regarded as basic research in academia
  - Generalizable to classes of problems in that domain
    - But less generalizable than core ML research
- Close to commercializability
  - Good platform for academic-industry collaborations

Examples: Applied ML VS Core ML

Examples:

- Faster Non-Convex Stochastic Optimization via Strongly Non-Convex Parameter
- A Simple Multi-Class Boosting Framework with Theoretical Guarantees and Empirical Proficiency
- A Closer Look at Memorization in Deep Networks
- An Alternative Softmax Operator for Reinforcement Learning

More Examples:

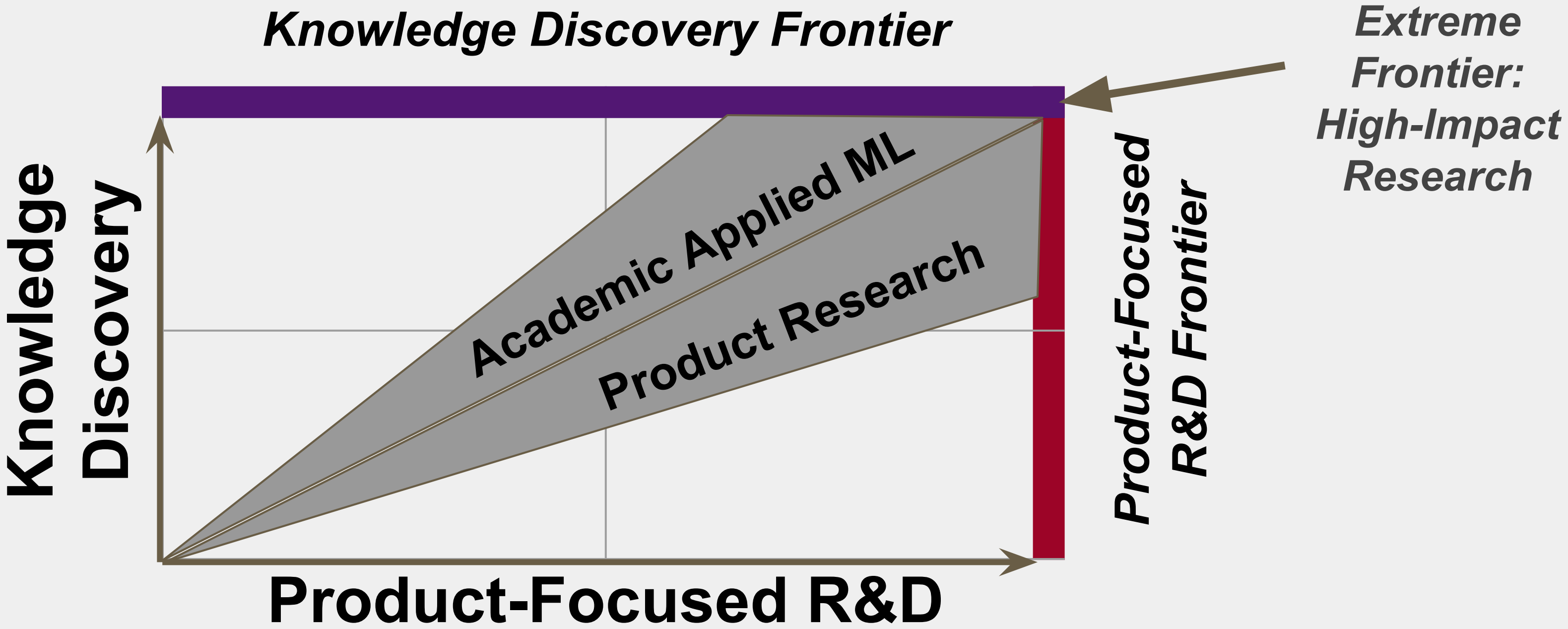
- Hierarchical Boundary-Aware Neural Encoder for Video Captioning
- Convolutional Random Walk Networks for Semantic Image Segmentation
- What do Neural Machine Translation Models Learn about Morphology
- Predicting Native Language from Gaze.

Frontier Applied ML with Industrial Collaboration

ECCV 2016

- *Learning to Refine Object Segments*, Pedro Pinheiro, EPFL; Tsung-Yi Lin, Cornell; Ronan Collobert, ***Facebook***; Piotr Dollar, ***Facebook***
- *Phase-based Modification Transfer for Video*, Simone Meyer, ETH Zurich; Alexander Sorkine-Hornung, ***Disney Research Zurich***; Markus Gross, ETH Zurich
- *Going Further with Point Pair Features*, Stefan Hinterstoisser, ***Google***; Vincent Lepetit, ; Kurt Konolige, ***Google***; Naresh Rajkumar, ***Google***
- *ActionSnapping: Motion-based Video Synchronization*, Jean-Charles Bazin, ETHZ; Alexander Sorkine-Hornung, ***Disney Research Zurich***
- *Globally Continuous and Non-Markovian Activity Analysis from Videos*, He Wang, ***Disney Research LA***; Carol O'Sullivan, Trinity College Dublin

Applied ML Research


Knowledge Discovery Frontier
Extreme Frontier: High-Impact Research
Knowledge Discovery
Academic Applied ML
Product Research
Product-Focused R&D Frontier
Product-Focused R&D

Hot Areas in Applied ML Research

Natural Language Processing	Speech Recognition
● Text summarization ● Sentiment analysis ● Native language identification ● Authorship attribution	● Speech to text ● Speech emotion recognition ● Speech biometric
Computer Vision	**Signal Processing**
● Object recognition & modeling ● Medical image analysis ● People counting ● Event detection	● EEG data analysis ● Brain-computer interaction ● IoT data processing ● Predictive maintenance

Case Studies: Applied ML Research Exp

- Case Study 1: Business Intelligence in Hospitality:
  - Modeling and analysing customer behaviors
  - Predict what each customer is gonna buy next
  - Product-Focused Business Consulting
  - Not academic research. Needed to justify to my boss BIG TIME!
- Case Study 2: Hardware-Accelerated Query Processing:
  - Designing a circuit board intercepting data from the storage
  - Reducing the amount of data going into the main memory
  - Basic research outcome
- Case Study 3: Stylometric Analysis: Academic Applied ML
  - Main Focus. Next Slide =>>>

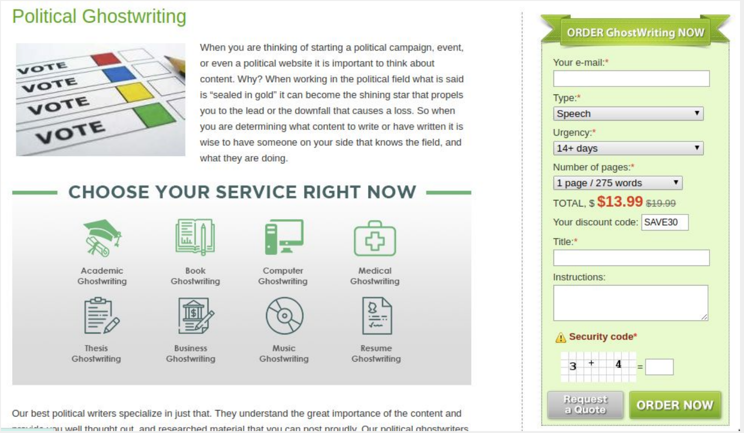
References

- C2Net: A Network-Efficient Approach to Collision Counting LSH Similarity Join. Hangyu Li; Sarana Nutanong; Hong Xu; Chenyun Yu; Foryu Ha. **IEEE Transactions on Knowledge and Data Engineering** Year: 2018, (Early Access)
- A Hardware-Accelerated Solution for Hierarchical Index-Based Merge-Join. Zimeng Zhou; Chenyun Yu; Sarana Nutanong; Yufei Cui; Chenchen Fu; Chun Jason Xue. **IEEE Transactions on Knowledge and Data Engineering** Year: 2018, (Early Access)

Case Study 3: Stylometric Analysis

Research Problems and Applications

- Authorship Attribution:
  - Intrinsic Plagiarism Detection
  - Ghost Writer Identification
- Multi-Author Authorship Attribution:
  - Bibliometrics
  - Scientometrics
- Cross-lingual Stylometric Analysis:
  - Native Language Identifications



Political Ghostwriting

When you are thinking of starting a political campaign, event, or even a political website it is important to think about content. Why? When working in the political field what is said is "sealed in gold" it can become the shining star that propels you to the lead or the downfall that causes a loss. So when you are determining what content to write or have written it is wise to have someone on your side that knows the field, and what they are doing.

CHOOSE YOUR SERVICE RIGHT NOW

Our best political writers specialize in just that. They understand the great importance of the content and provide you well thought out, and researched material that you can post proudly. Our political ghostwriters

ORDER GhostWriting NOW

Your e-mail:*

Type:* Speech

Urgency:* 14+ days

Number of pages:* 1 page / 275 words

TOTAL, $ $13.99 ~~$19.99~~

Your discount code: SAVE30

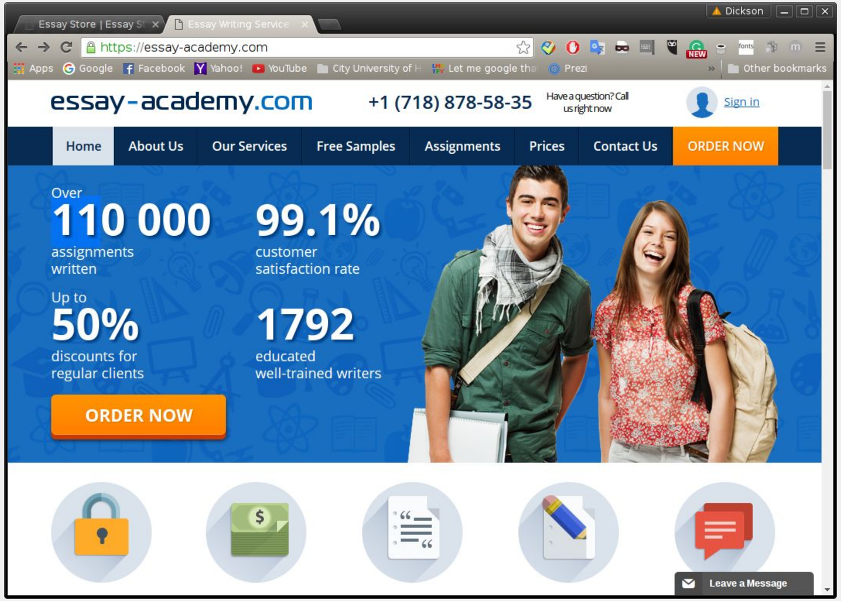
Title:*

Instructions:

Security code*

3 + 4 =

Case Study 3: Stylometric Analysis



- Plagiarism detection is a cat-and-mouse game
    - Students used to copy and paste
    - Turnitin is extremely effective at detecting this type of plagiarism
- Plagiarism gets more sophisticated
    - Students pay essay writing agencies to write essays for them
    - This form of plagiarism is still a serious academic offense
- Essay writing agencies, e.g.,
    - Essaytiger.com
    - Essay-academy.com
    - Essaystore.com
    - Essayscam.org
    - Advancedwriters.com

Case Study 3: Stylometric Analysis

Assumption: It is impossible for a student to get the same writer for the entire 4 years program

Research Problems:

- ***Authorship Verification***: Whether the students wrote the essay by themselves?
- ***Authorship Identification***: Identifying the ghost writer

Case Study 3: Stylometric Analysis

What we did:

- Design a new data representation method
- Develop a new ML technique for this problem based on a classical method
- Design experimental studies to demonstrate the superiority of our proposed method

Extensions:

- Cross-lingual Authorship Attribution
- Native Language Identification
- Multi-author Authorship Attribution

Refs:

- Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Urailertprasert, Nattapol Vannaboot, Thanawin Rakthanmanon: A Scalable Framework for Stylometric Analysis of Multi-author Documents. DASFAA (1) 2018: 813-829
- Sarana Nutanong, Chenyun Yu, Raheem Sarwar, Peter Xu, Dickson Chow: A Scalable Framework for Stylometric Analysis Query Processing. ICDM 2016: 1125-1130

Case Study 3: Stylometric Analysis

- Domain Specific: Digital library data, e.g.,
  - http://www.gutenberg.org/
  - https://arxiv.org/
- Knowledge Discovery
  - Deriving methods to solve a class of stylometric problems
- Close to real-world applications (but still need more product research)
  - Intrinsic plagiarism detection services
  - Bibliometric web services

Concluding Remarks and Recommendations

- Propelling research in Information Science and Technology through midstream research, e.g., applied ML
    - Linking upstream and downstream
- Encouraging collaborative research projects
    - We are already pretty well here
- Encouraging publishing research findings at high impact venues to utilize the rigorous peer review systems
    - Conferences recognized by csrankings.org
    - A* conferences at core.edu.au
    - High IF Journals and Transactions: http://data-mining.philippe-fournier-viger.com/the-top-journals-and-conferences-in-data-mining-data-science/