*8th Annual*
*Global TechMining(GTM) Conference*
*September, 2018*
*Leiden, Netherlands*

# Visualizing Dispersed Risk Signals for A Specific Emerging Technology: A Novel Approach of Keywords Aggregation across Topics (KAAT)

School of Business,
South China University of Technology, China
Prof. Munan Li
2018-09
Email: limn@scut.edu.cn

# Research Background

- A typical interdisciplinary topic query:
- "*Risk for Graphene, Additive Manufacturing*" could involve social science, management, business, environment science, engineering and so forth.

- Research Question:
- For a specific emerging technology, Can we timely and efficiently discovery the risk signal / relevant works, especially in early period?

- Research Design

- Goal: Attempt to find a simple algorithm of machine learning to improve the capability and efficiency of discovery the risk signals of a specific emerging technology from publications, patents, and even Internet space.

- Keywords Aggregation across Topics (KAaT):
- KAaT could be or could be similar with a simple algorithm of machine learning

- Basic Philosophy (Components) of Machine Learning:

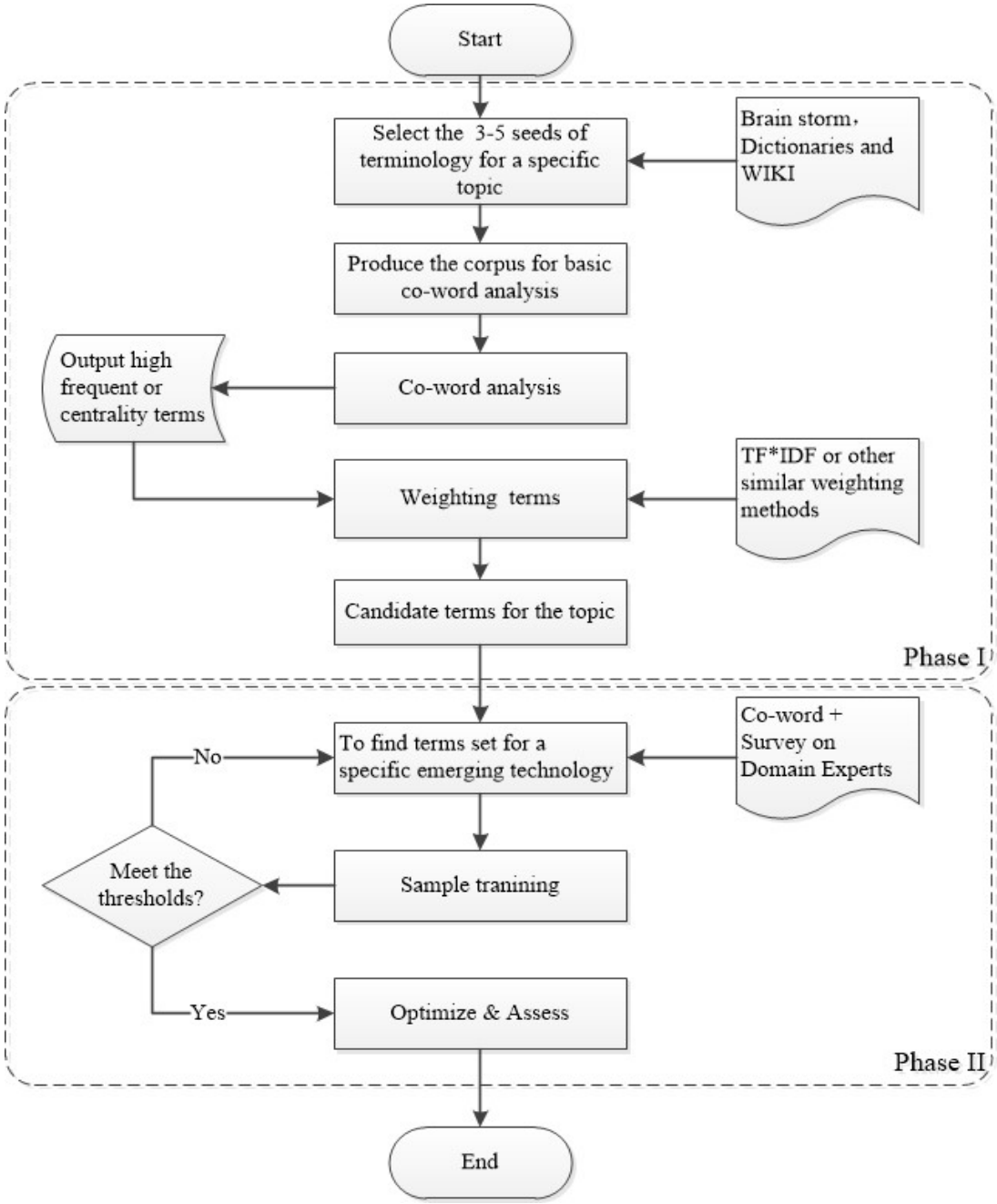
- (1)Computation logic (e.g. Non-linear programing model)

- (2) Training  & Optimize parameters

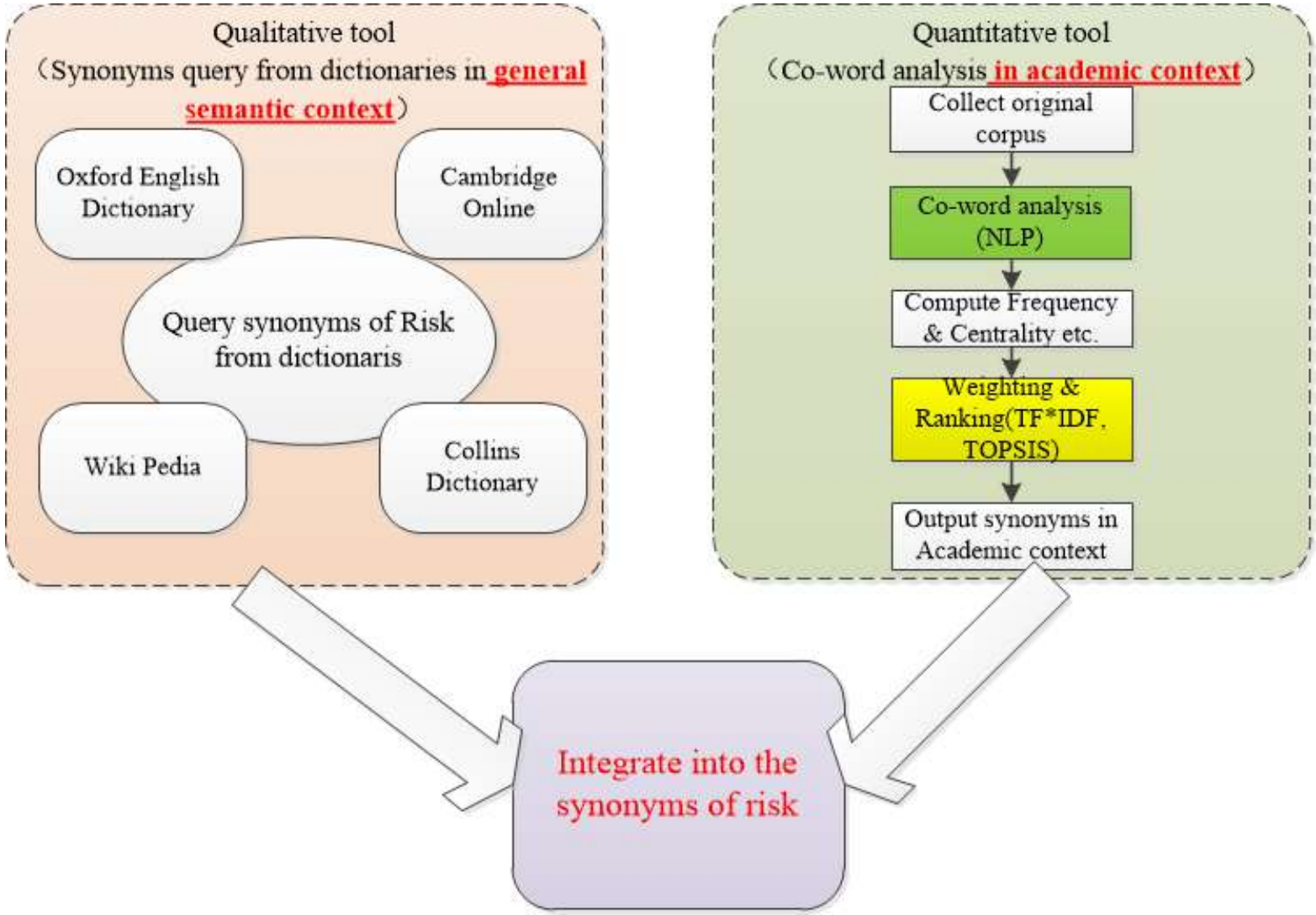
- (3) Run Algorithm

- (4) Verification & Feedback

- KAaT could be taken into account a naïve machine learning.
- Why is it naïve ?
- Because this algorithm only utilize the basic philosophy of machine learning, and not involve such complicated topic modeling: LDA, LSA and etc.



- Phase I:  To find the synonyms of risk



- KAaT: Produce Training Sample in Phase II

- TS="emerging technolog*" AND

- TS=(risk* OR unsafe OR uncertainty OR danger* OR peril OR threat* OR menace OR fear OR unpredictab* OR precarious* OR instability* OR insecurity* OR perilousness OR venture OR jeopardy OR loss OR chancy OR toxic OR poison* OR vulnerability OR injury OR hazard* OR misfortune OR endanger OR jeopardize OR imperil) AND TS =(environment* OR health* OR security OR safety OR ecosystem OR "air pollution" OR "soil contaminat*" OR "water pollution" OR "water cotaminat*")

- AND DOCUMENT TYPES: (Article)
- Indexes=SCI-EXPANDED, SSCI, A&HCI, ESCI, CCR-EXPANDED Timespan=2007-2009

| Training Sample | Signal | Noise | Accuracy(%) | Presumed Recall(%) |
|---|---|---|---|---|
| 83 | 15 | 68 | 18.07 | 100 |

Algorithm training is to find the optimized keywords combinations that can efficiently identify signal and noise.

# Training Results:

| Training Times ( Attempt the different combinations of Terms | Recall >0 & Accuracy >0 | Recall > 20% | Accuracy > 50% |
|---|---|---|---|
| 11372 | 170 | 13 | 126 |

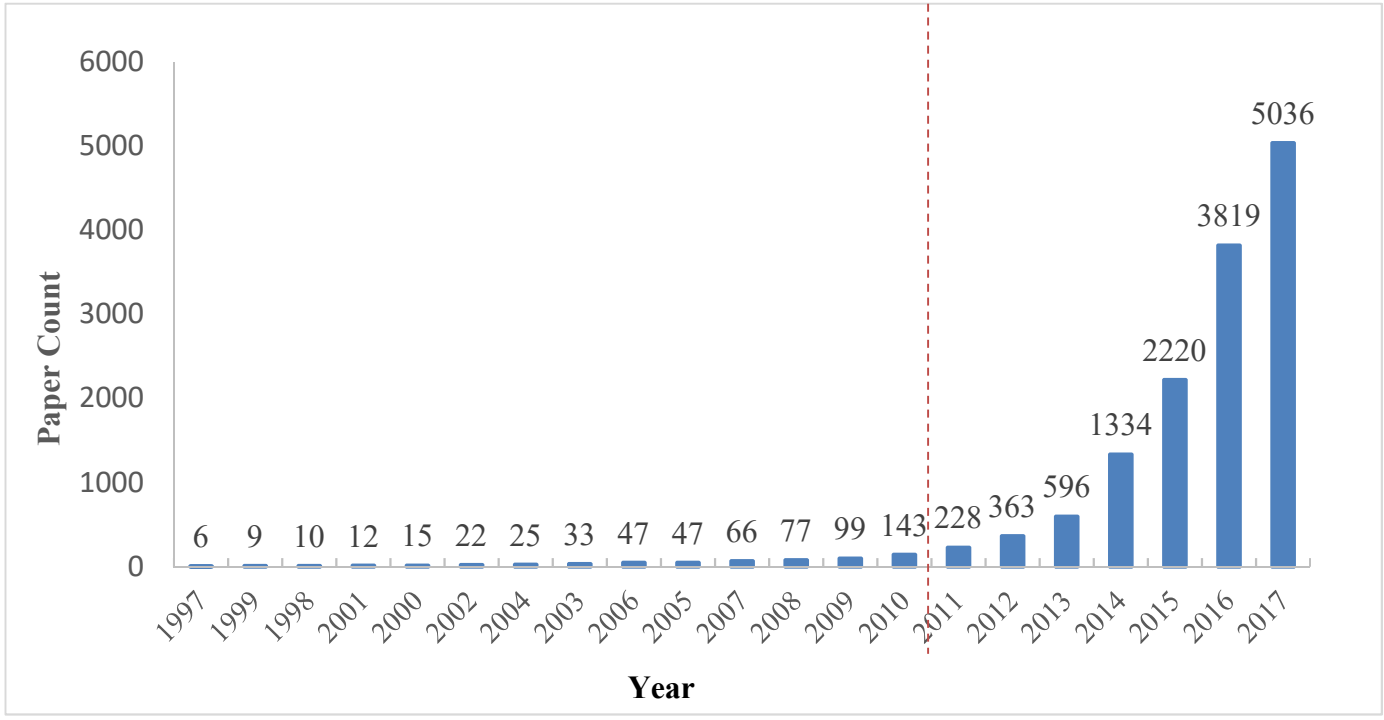
If the Recall (%) is prior to Accuracy (%), such keywords as: **risk, health, environment, toxicity** ; and those 126 combination of keywords whose Accuracy is larger than 50% are selected to the next compuation.

# Case Study: The Risk Discovery of 3D printing/Additive Manufacturing

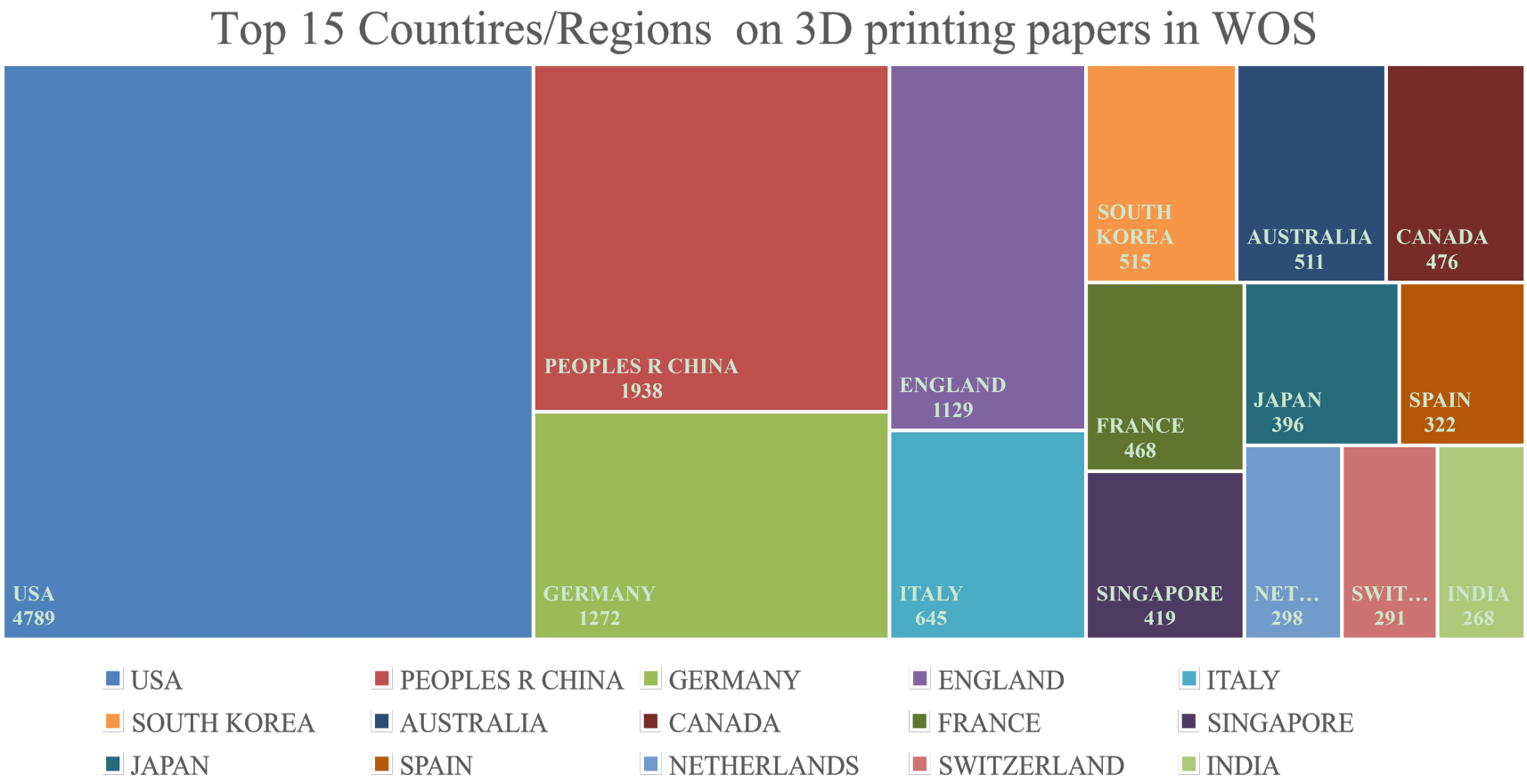
| | Query Formula | Result | Type | Refine Rules |
|---|---|---|---|---|
| #1 | TS=((3D OR 3-D OR "3 dimension*" OR "three dimension*" OR additive) NEAR/2 (print* OR fabricat* OR manufactur* OR product*)) | 8477 | ARTICLE (5,321)<br>PROCEEDINGS PAPER (3,043)<br>REVIEW (345) | DOCUMENT TYPES: ( ARTICLE OR PROCEEDINGS PAPER OR REVIEW )<br>Timespan: 2015-2016. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. |
| #2 | TS=((3D OR 3-D OR "3 dimension*" OR "three dimension*" OR additive) NEAR/2 (print* OR fabricat* OR manufactur* OR product*)) AND TS=(risk* OR pathogen OR "Scenario planning" OR environment* OR health OR toxic*) | 995 | ARTICLE (609)<br>PROCEEDINGS PAPER (340)<br>REVIEW (72) | |

(#1) query formula refers the work of Yin Huang et al. , which is published in 2017. and (2#) query formula combines 1# with training results.

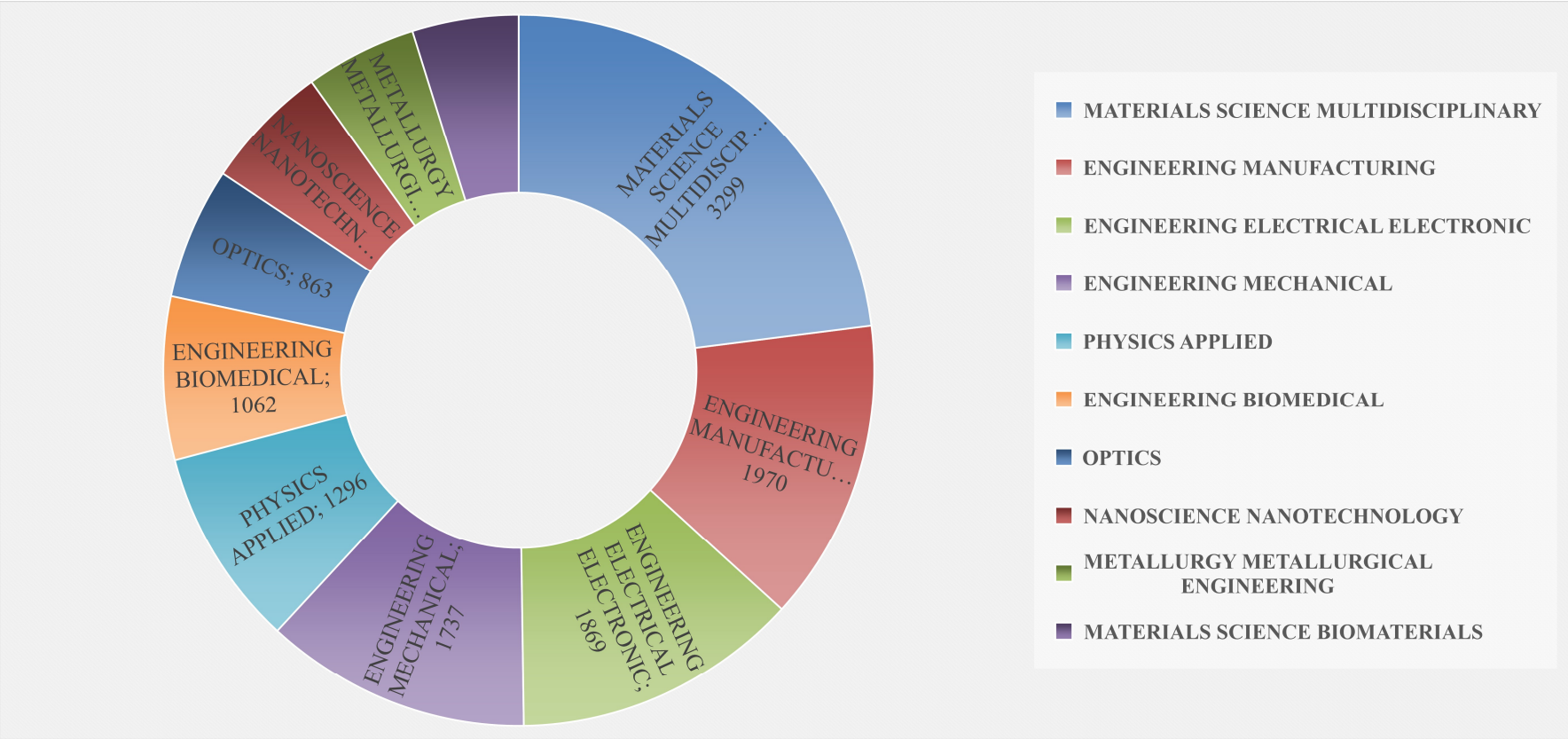
# Descriptive results on 3D printing studies:



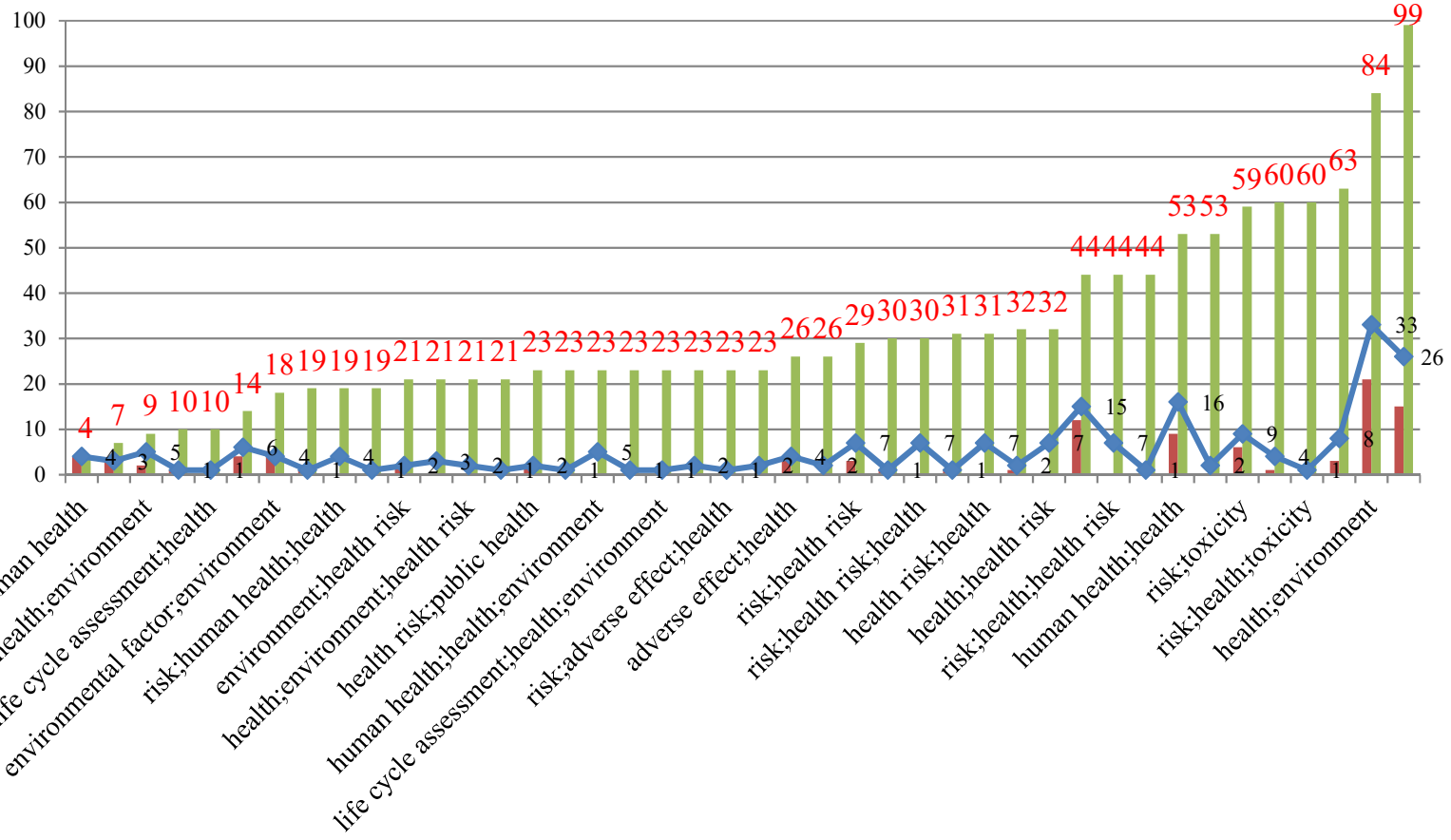
# Descriptive results on 3D printing studies:



# Descriptive results on 3D printing studies:



# Naïve Machine Learning Result:



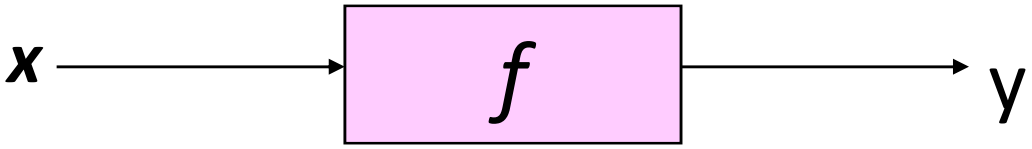

**Finally, based on the 126 rules, 99 publications are selected from 995 samples; and 23 are matched signal, another are noise; therefore, the recall is 100%, accuracy is just 23%.**

# KAaT: a naïve algorithm of machine learning
# Conclusion or Implication

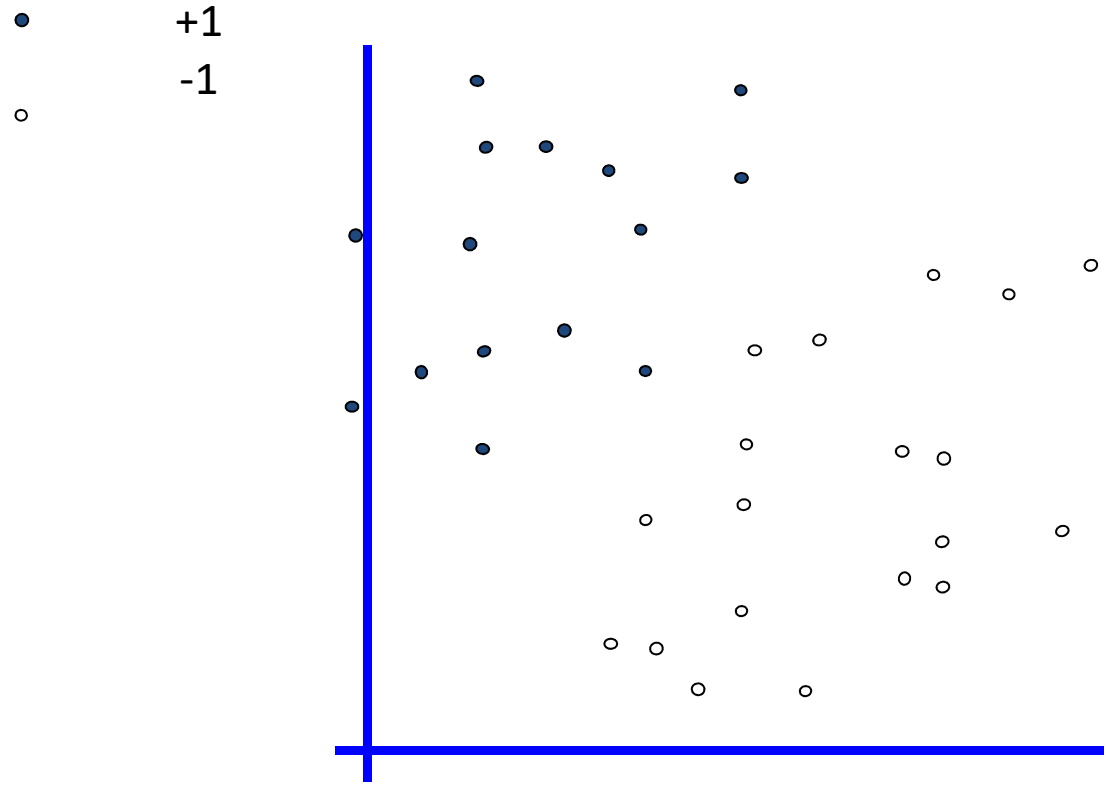
**In summary, based on the introduced NML (Naïve Machine Learning), a method or algorithm for discovery the risk signal of a specific emerging technology are explored. And NML can compress the noise space, and bring a moderate accuracy of identification.**

**Also, in the future research, more complicated semantic modeling can be integrated into NML to further improve the accuracy.**

**Research Extension:** Can the above question be transformed into a linear classifier question?



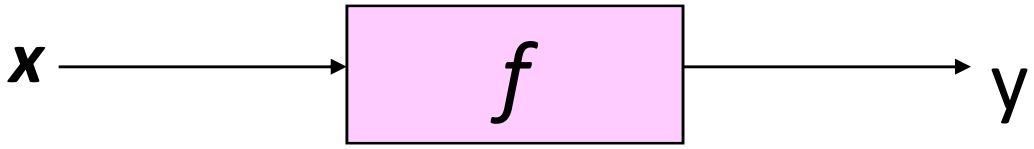

$f(\mathbf{x}, \mathbf{w}, b) = sign(\mathbf{w}.\ \mathbf{x} - b)$



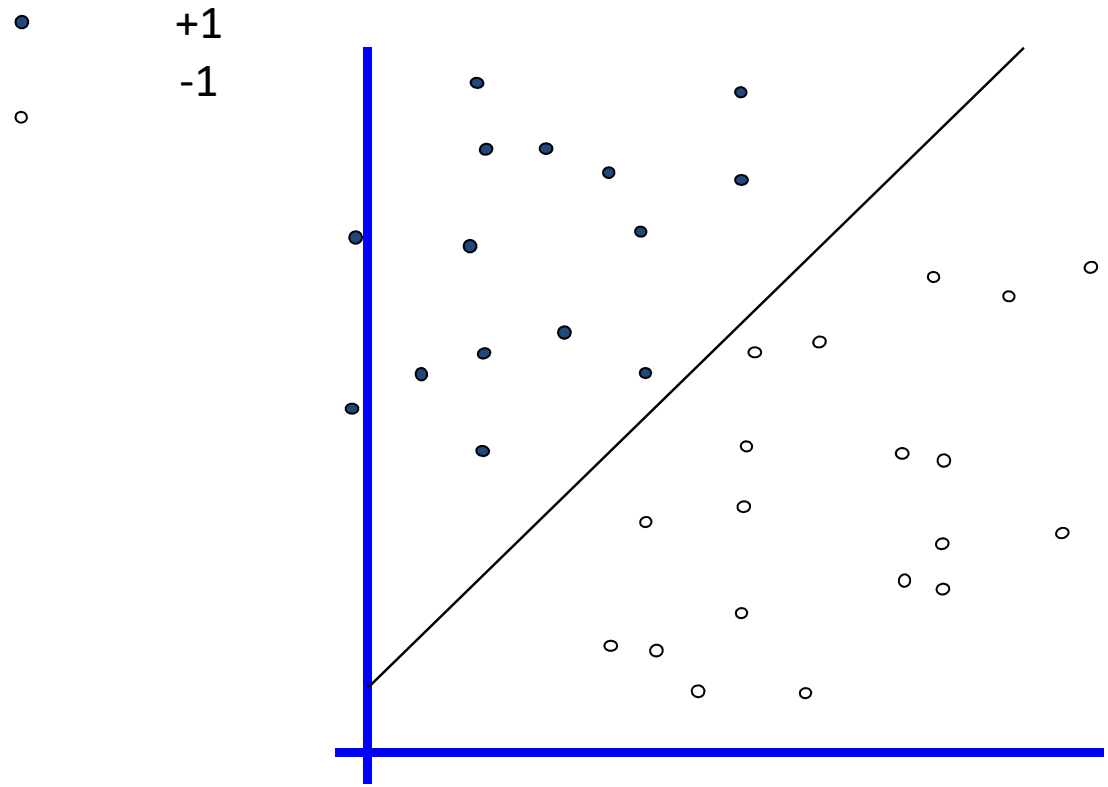

How would you classify this data?

**A classic linear classifier:**

Support Vector Machine (SVM)



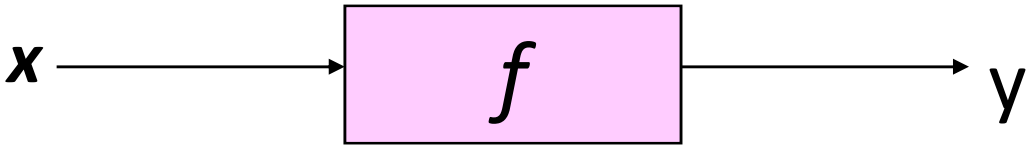

$f(\mathbf{x}, \mathbf{w}, b) = sign(\mathbf{w}.\ \mathbf{x} - b)$

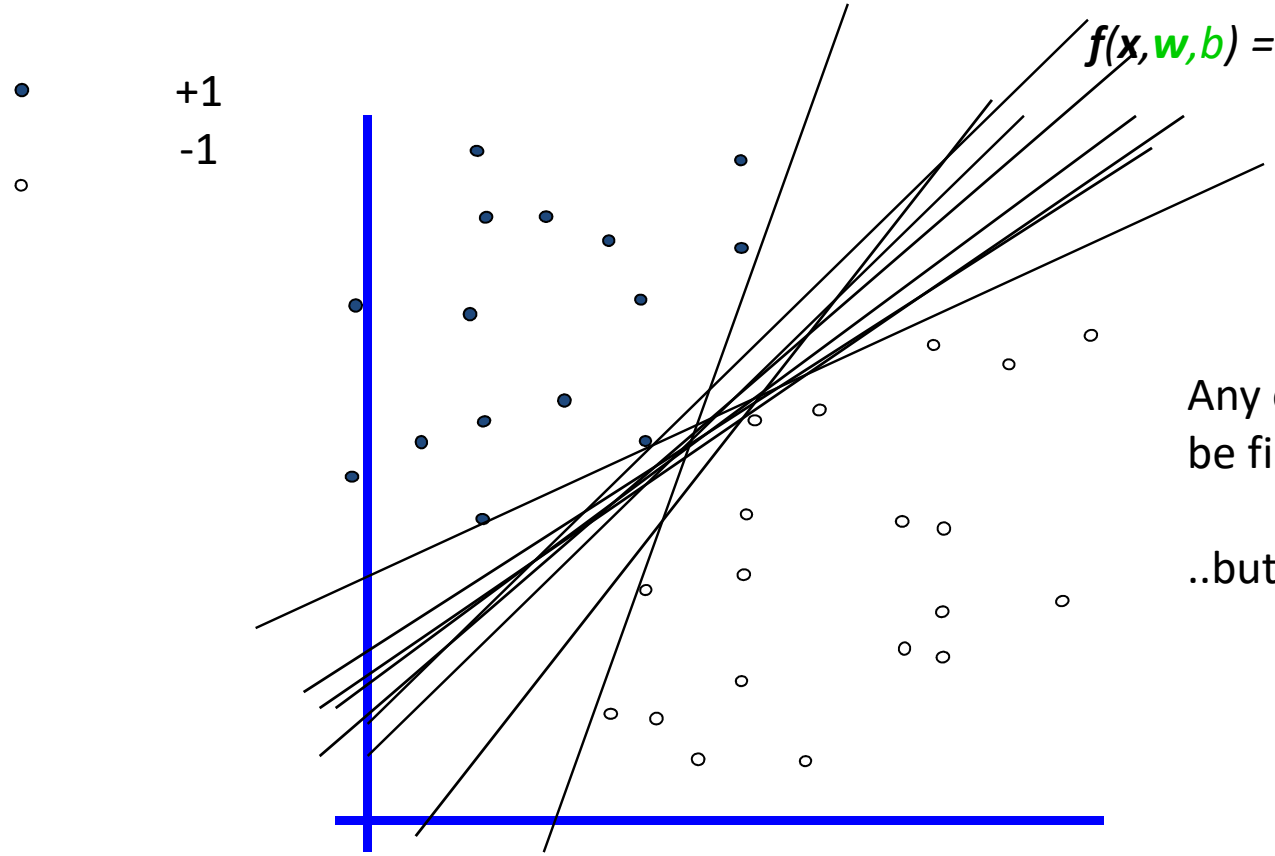



How would you classify this data?

**Support Vector Machines**

# Support Vector Machines

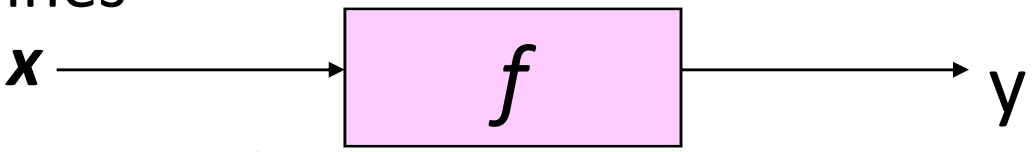


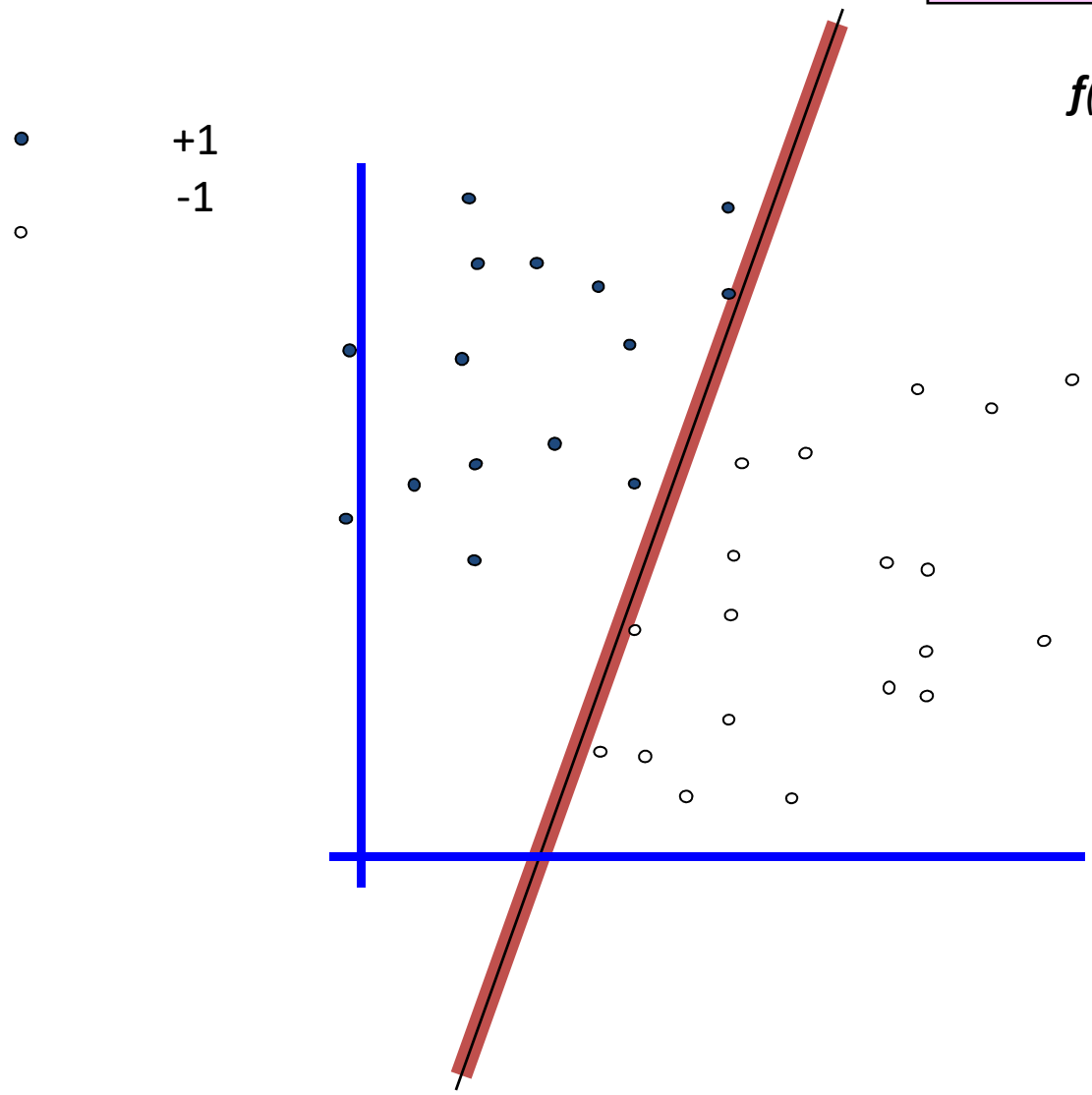




Any of these would be fine..

..but which is best?

# Support Vector Machines

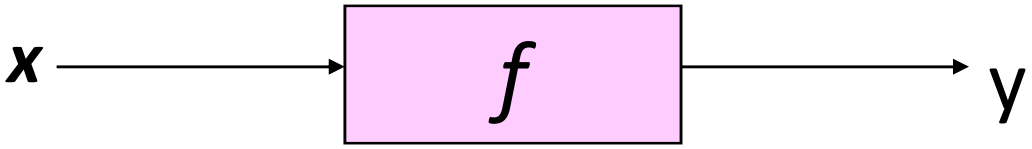


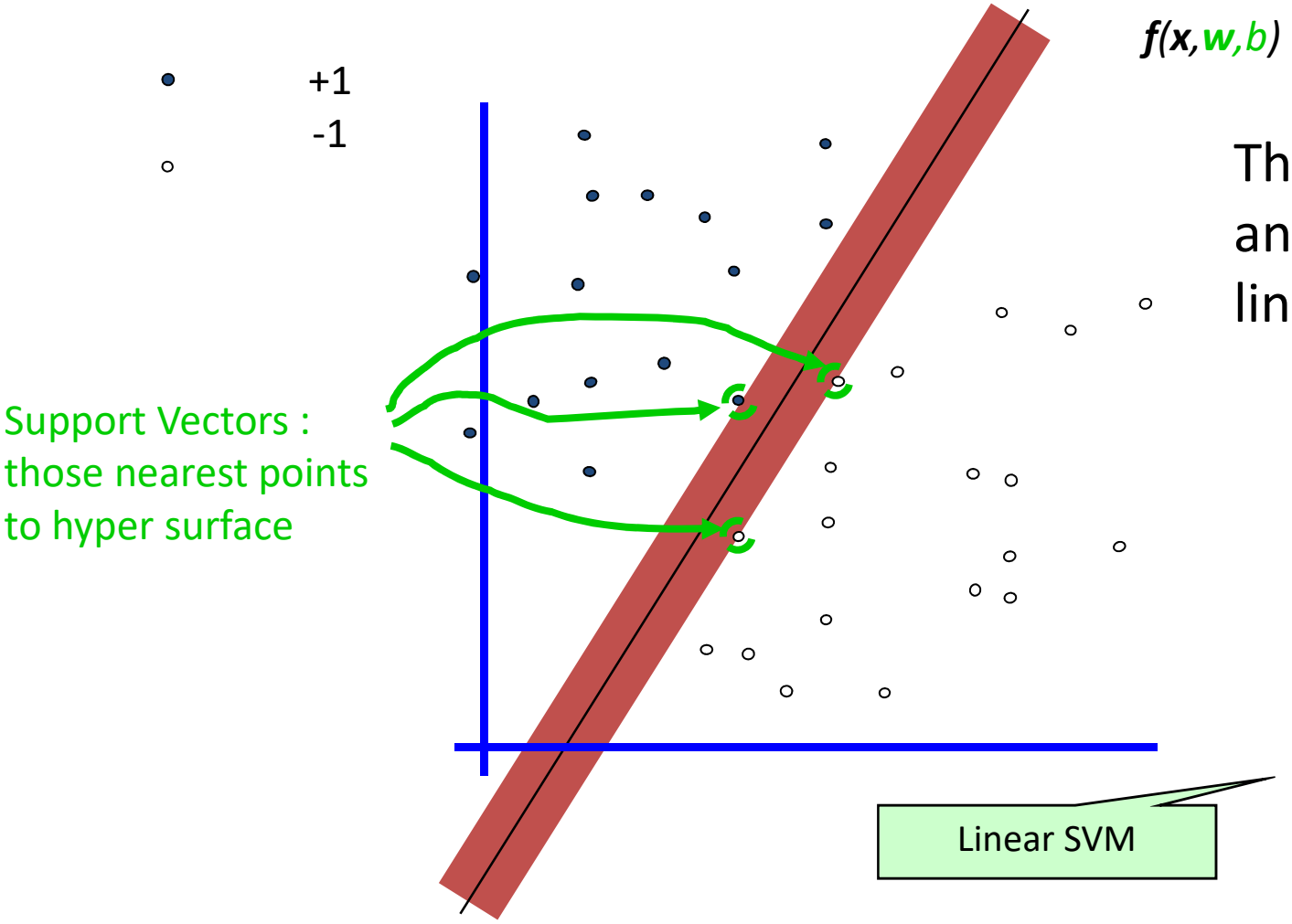

$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$

Linear Classifier margin:
the distance between hyper surface and the nearest samples (dots)

# Find the biggest margin

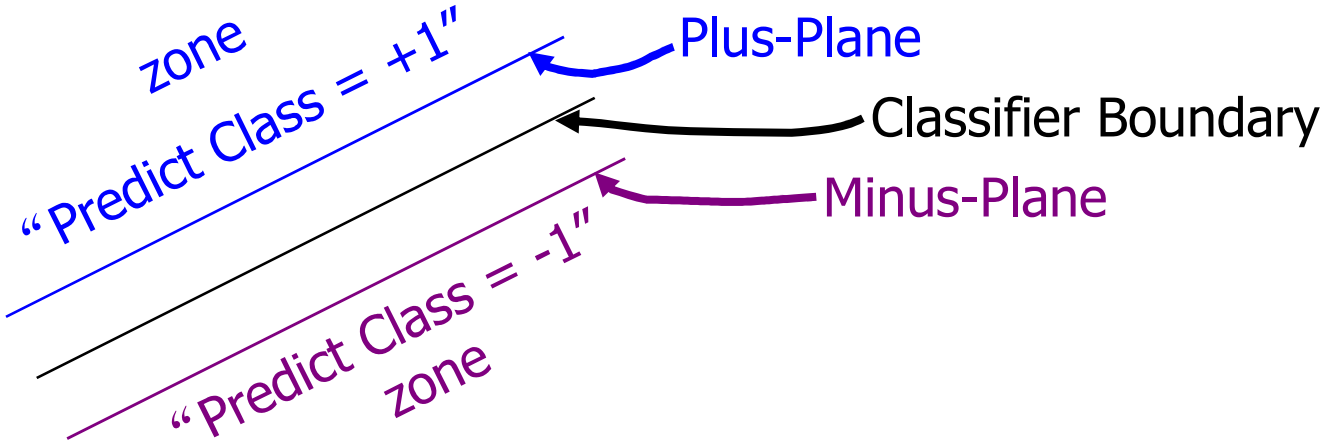


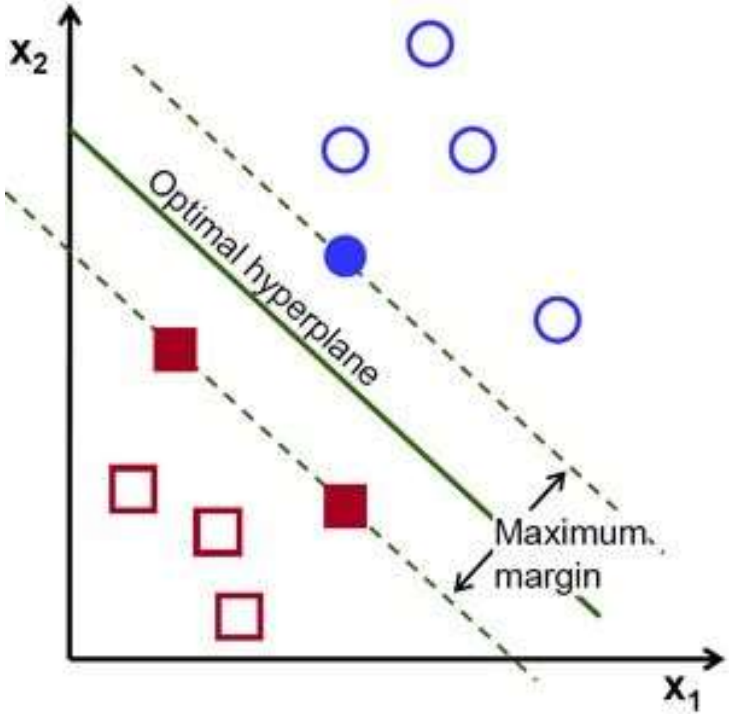

$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$



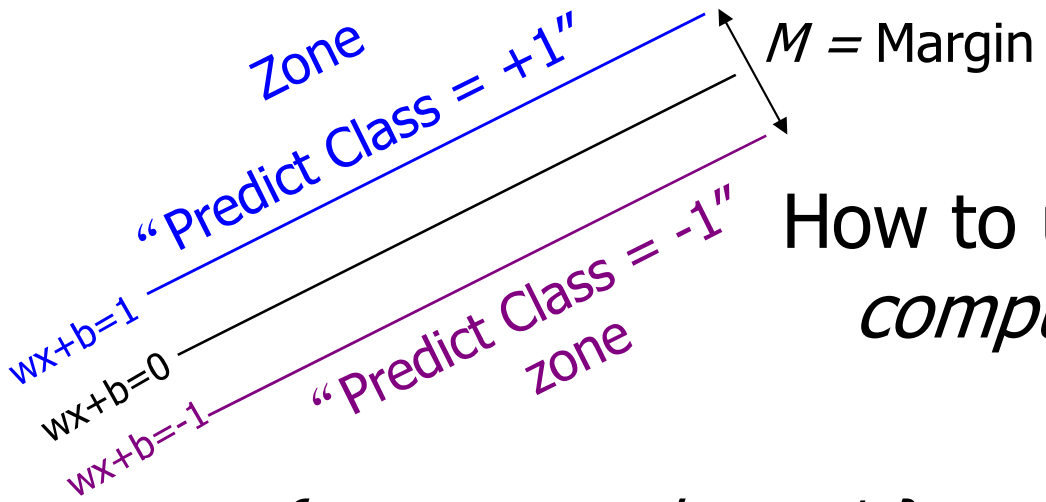

This is a simple SVM, and also named as linear SVM

# Hyperplane and the margin



# Compute margin




How to use ***w and*** *b compute margin?*

- Plus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} . \boldsymbol{x} + b = +1 \}$
- Minus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} . \boldsymbol{x} + b = -1 \}$

**So, the above question could be transformed into**
**Optimization question (Quadratic Programming)**

Minimize $\frac{1}{2}\mathbf{w}.\mathbf{w}$

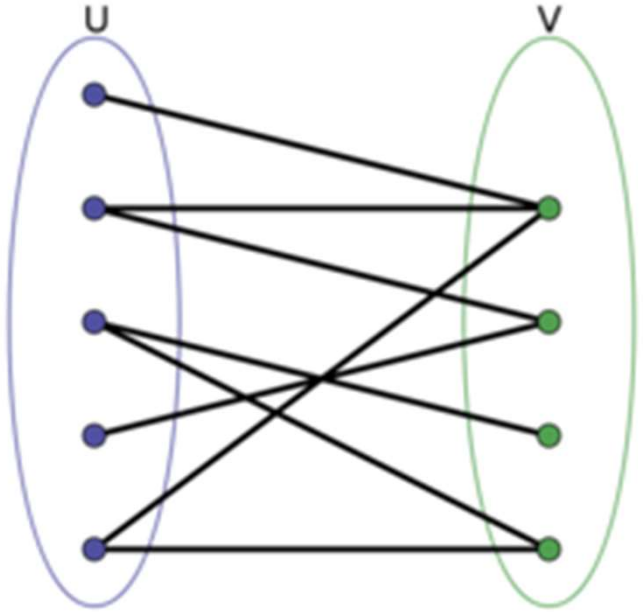
subject to $y_k\,(\boldsymbol{w} \,.\, \boldsymbol{x}_k + b\,) >= 1$

$k=1,2,\ldots,n$

## Research Extension II: Can be transformed into Bipartite Graph ?

**( And then, graph theory and those related algorithms could be helpful)**

**U:** the **relevant** papers on risk analysis (**E**nvironment, **H**ealth and **S**afety etc.)for a specific emerging technology




**V:** the **irrelevant** papers on risk analysis (**E**nvironment, **H**ealth and **S**afety etc.) for a specific emerging technology

**U+V:** All papers on the specific emerging technology(e.g. 3D printing, synthetic biology, Graphene, etc.)

# Thank you!