hyadata lab

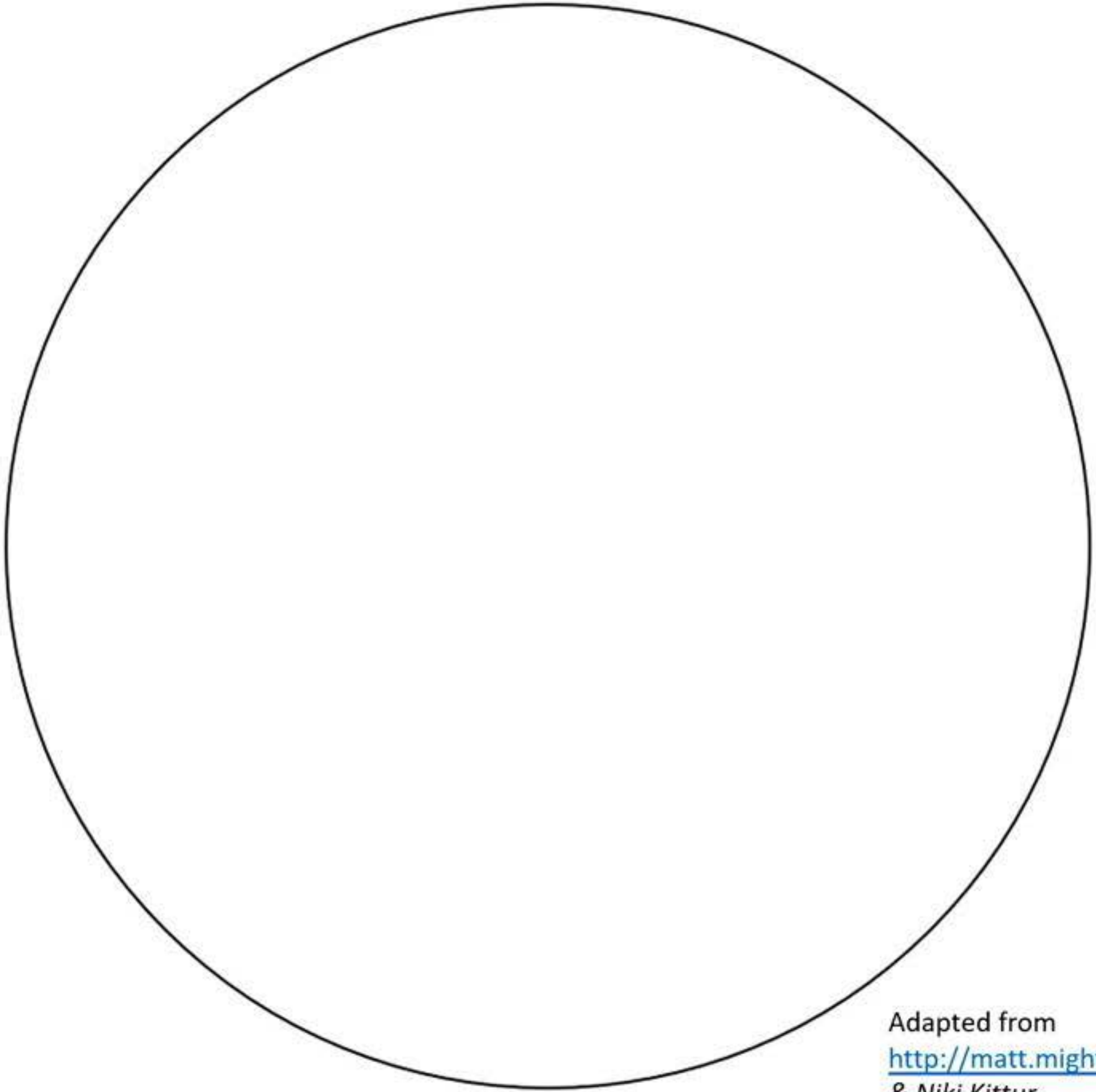
THE HEBREW UNIVERSITY OF JERUSALEM

Automating Inn💡vation and Discovery with Machine Learning

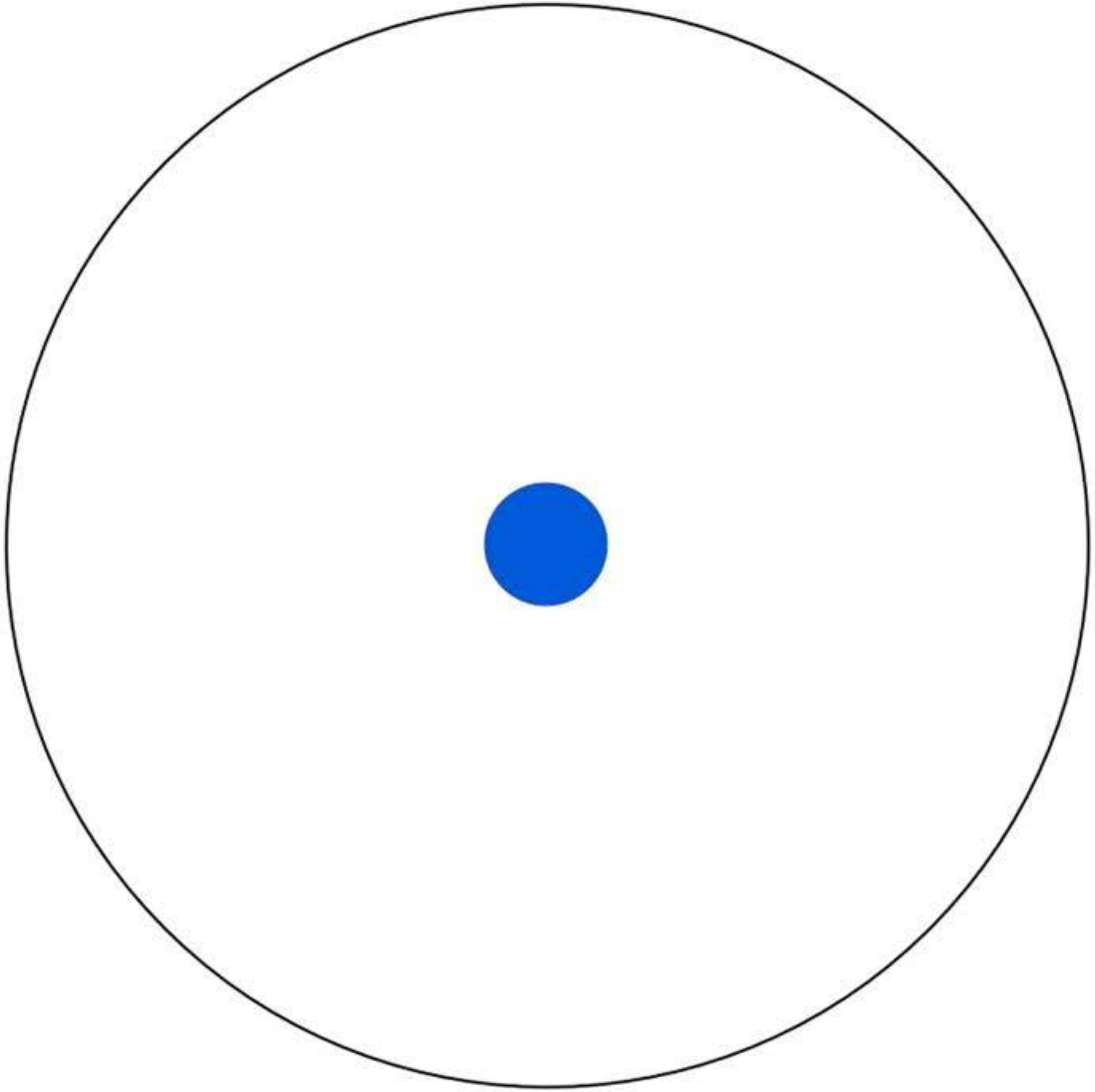
My PhD (so far)

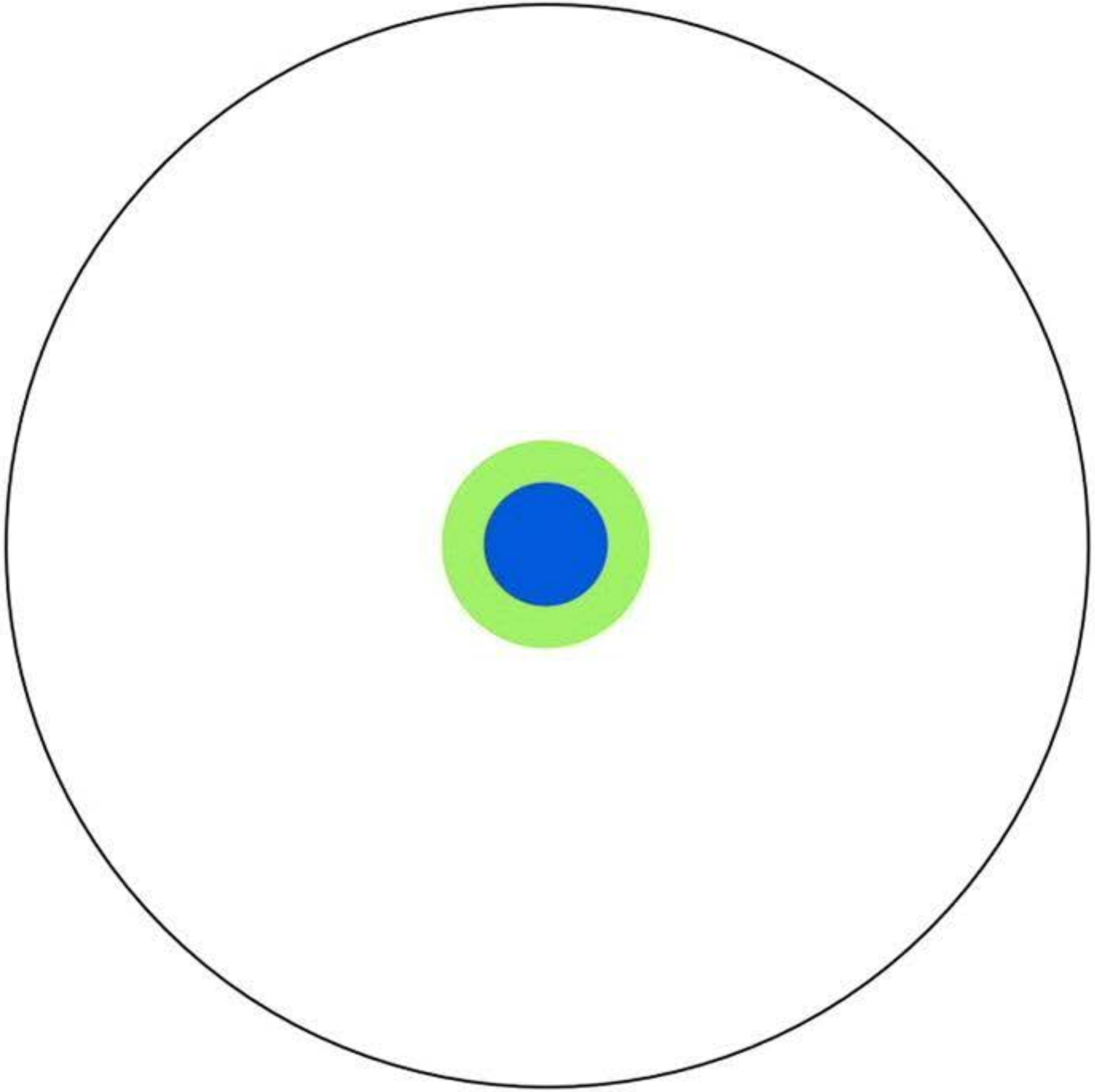
Tom Hope

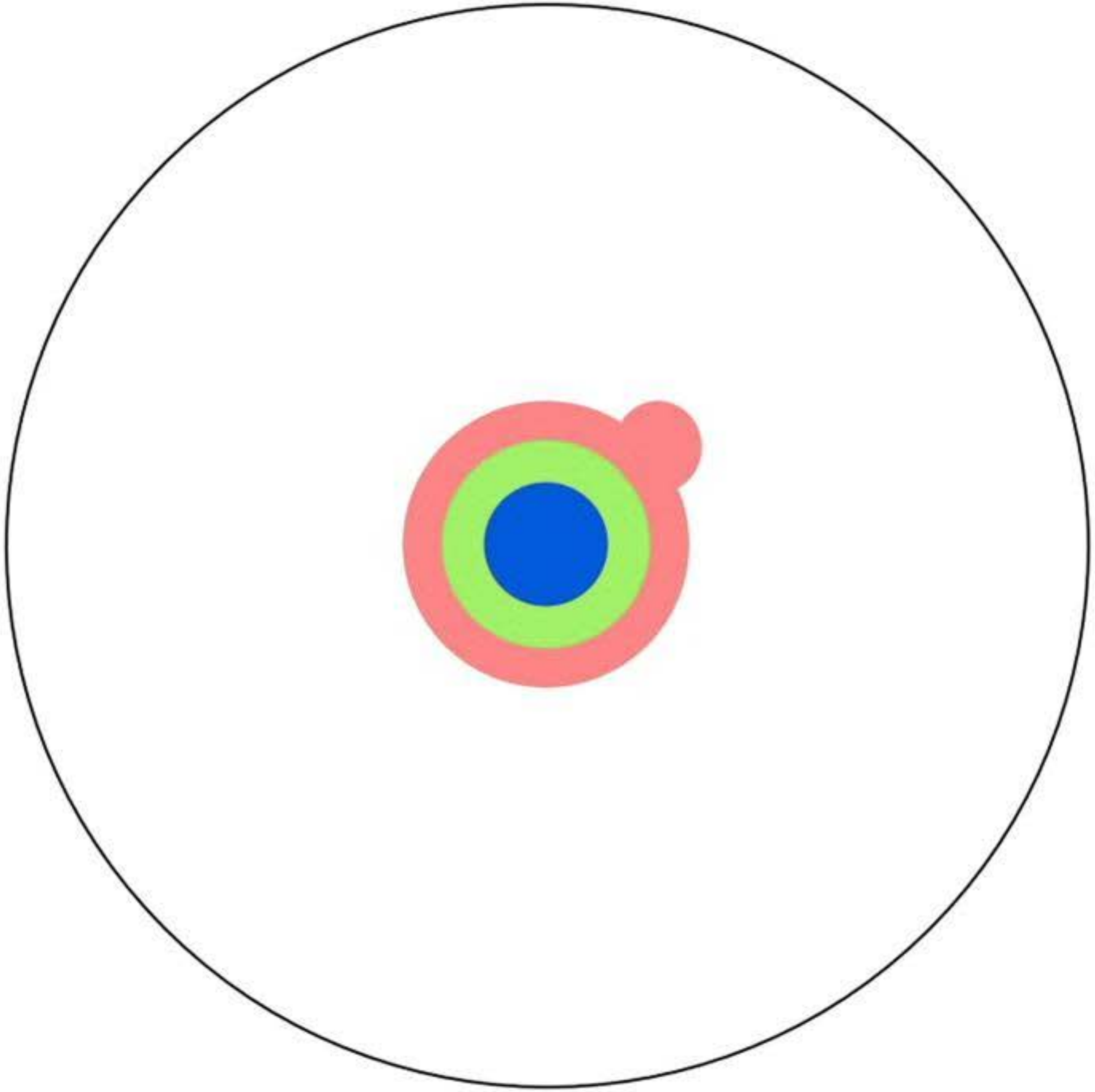
PhD Advisor: Dafna Shahaf

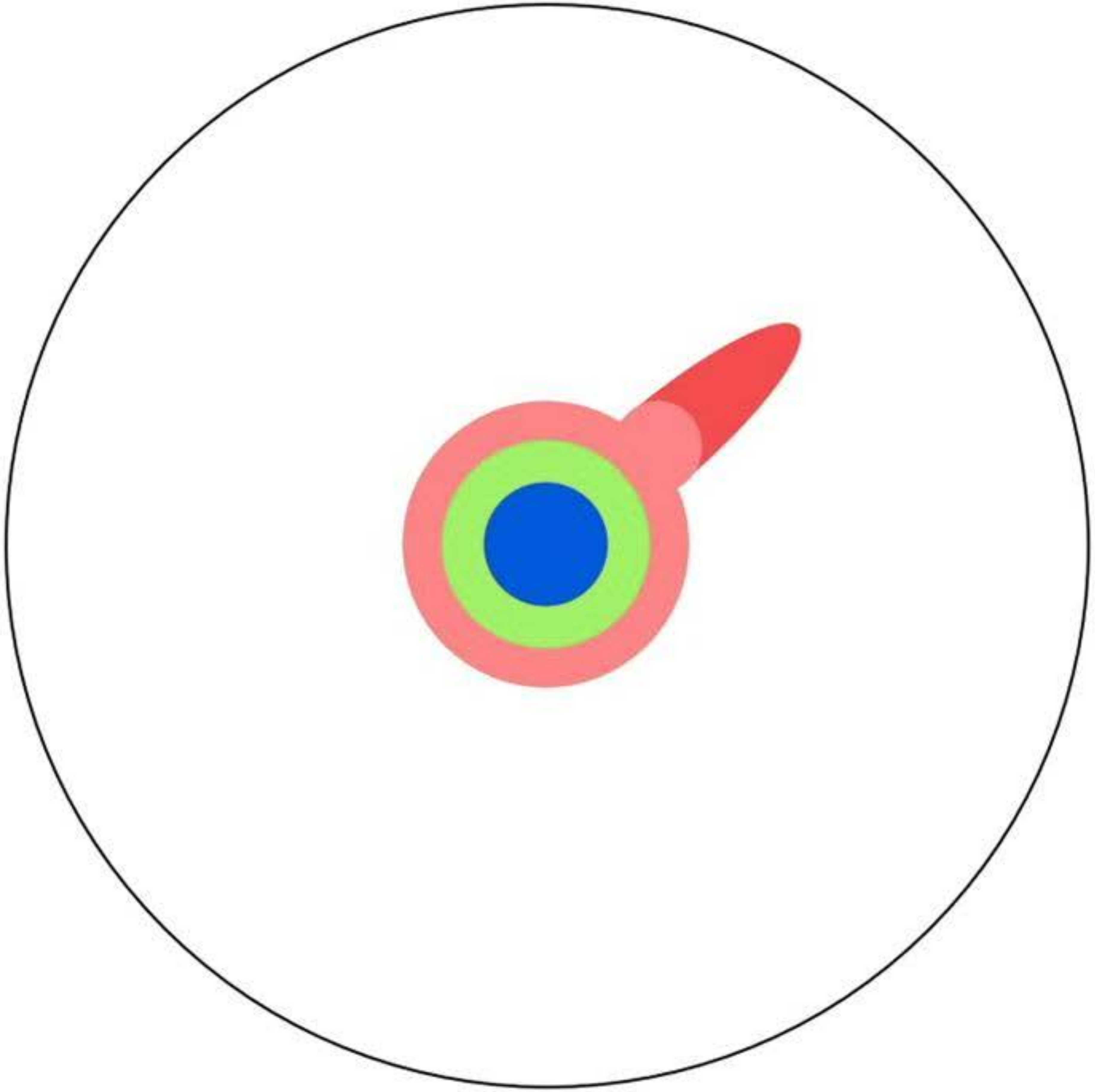


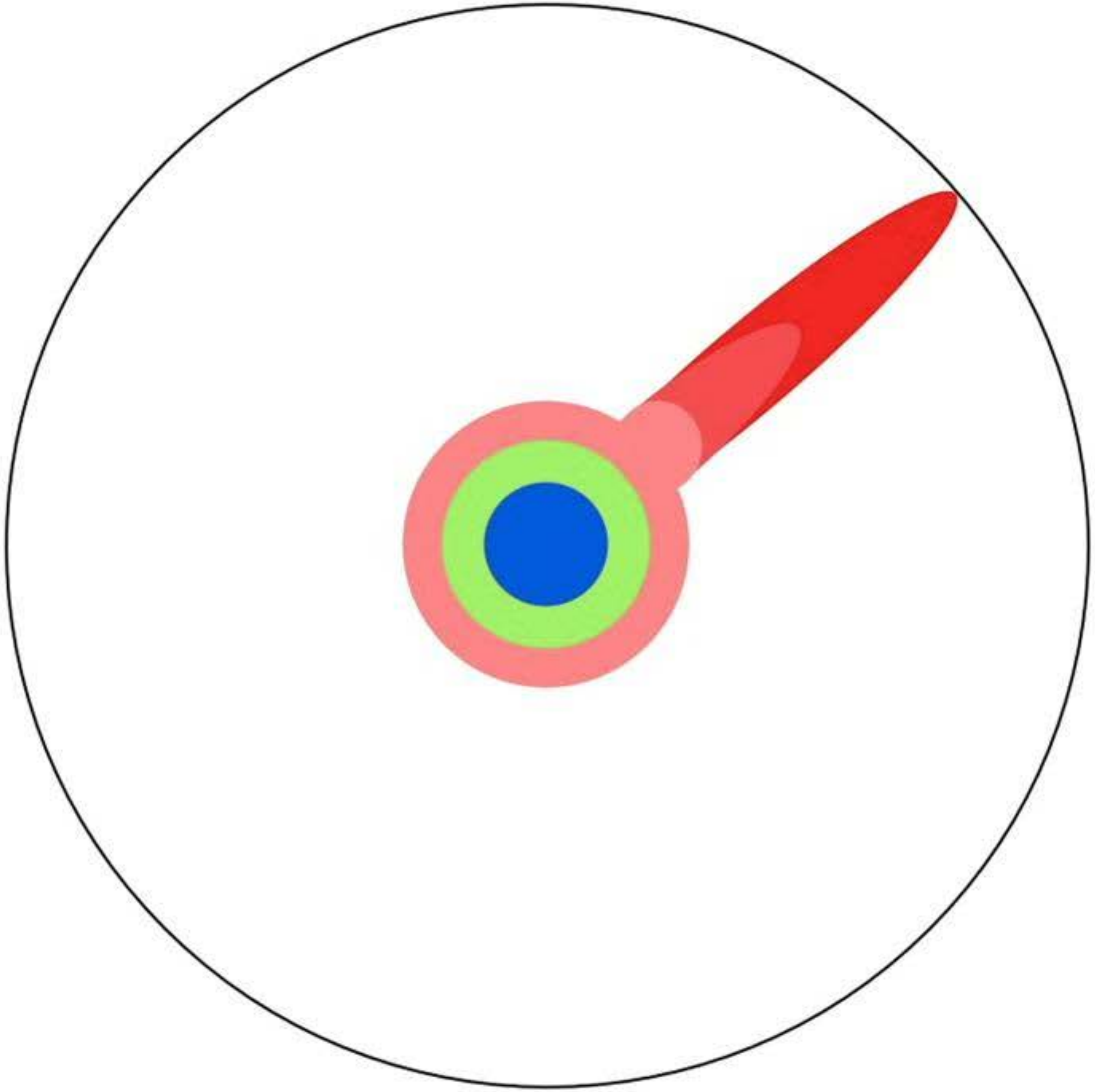
Adapted from
http://matt.might.net/articles/phd-school-in-pictures/
& Niki Kittur

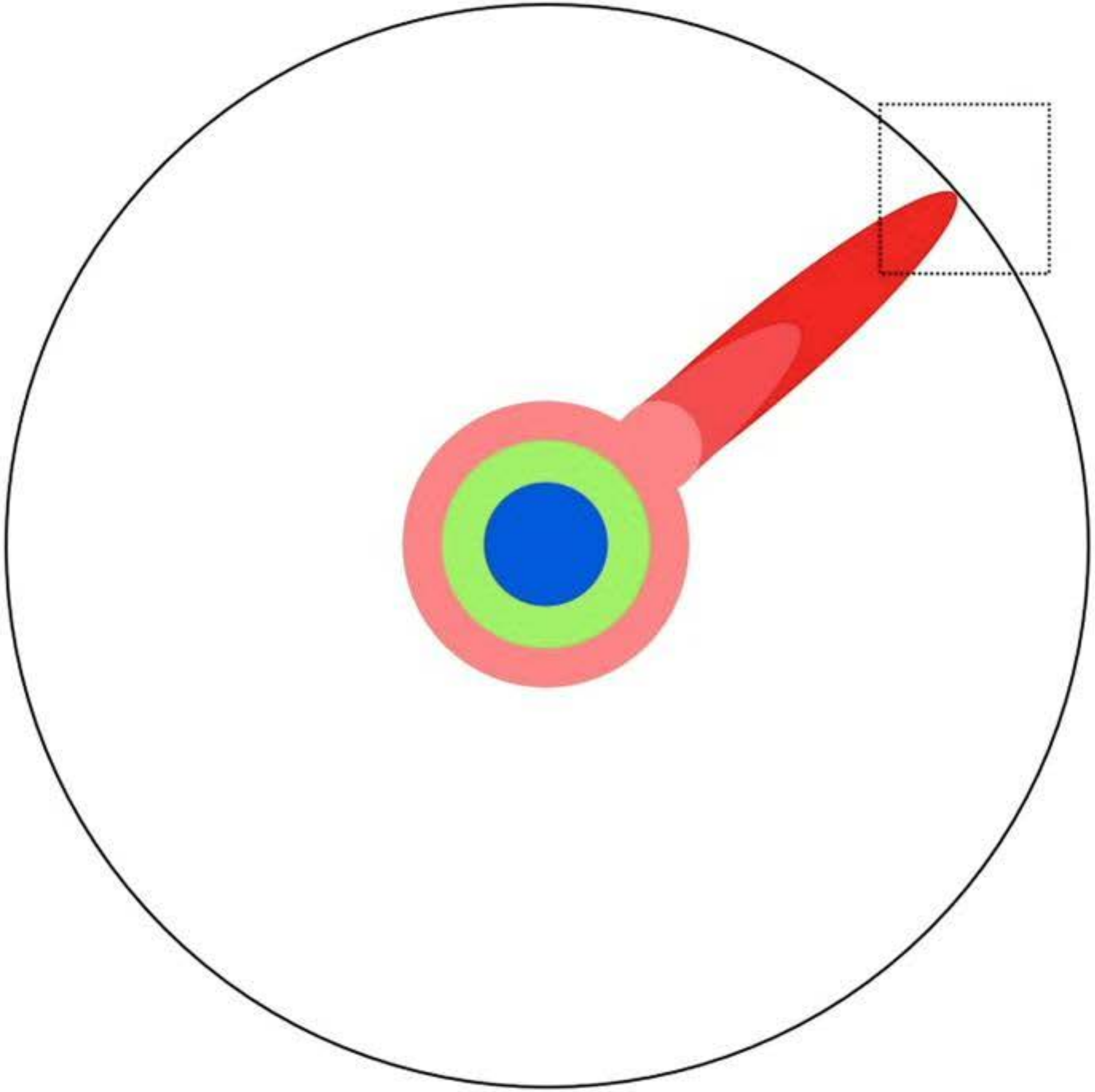


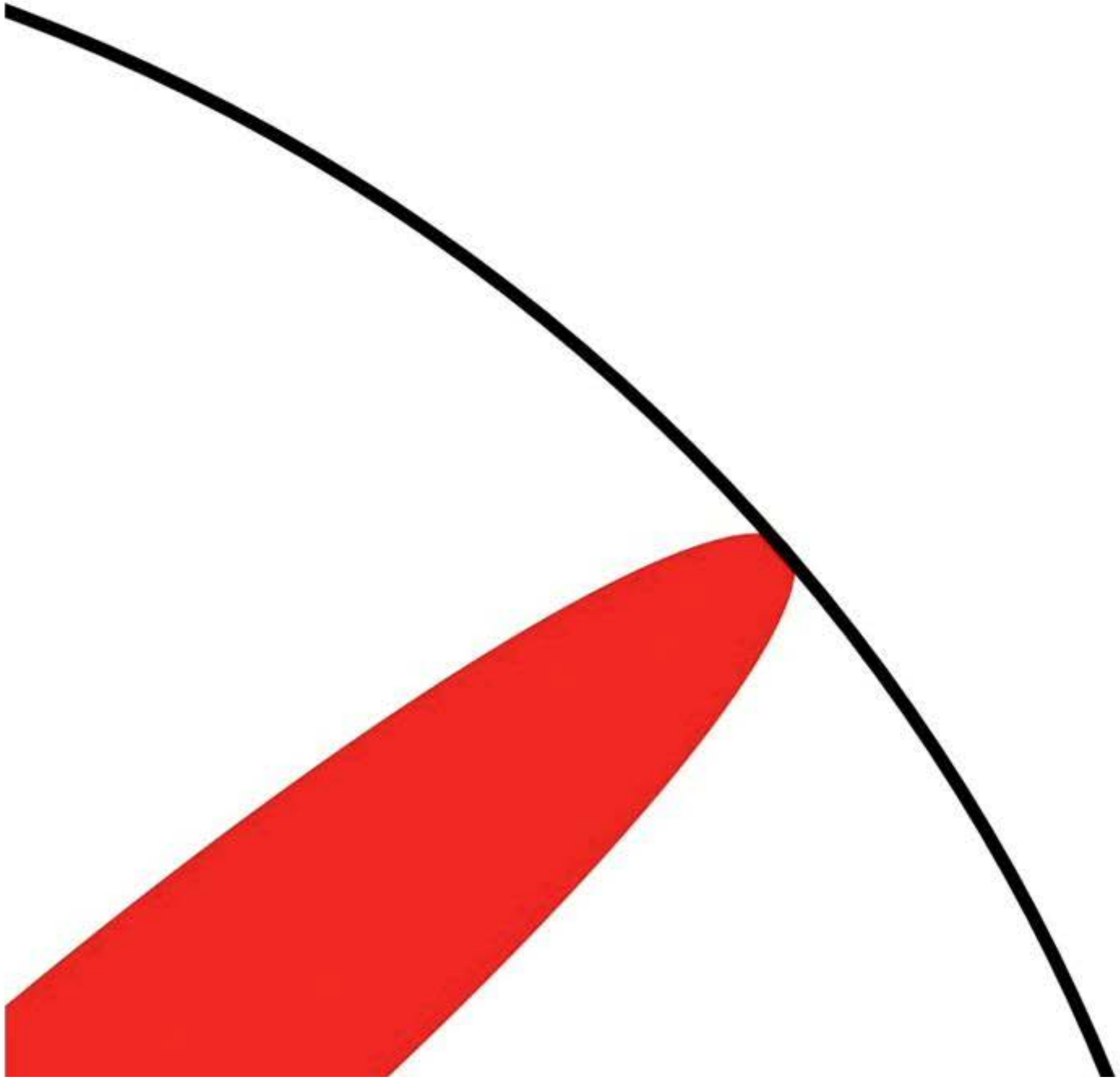


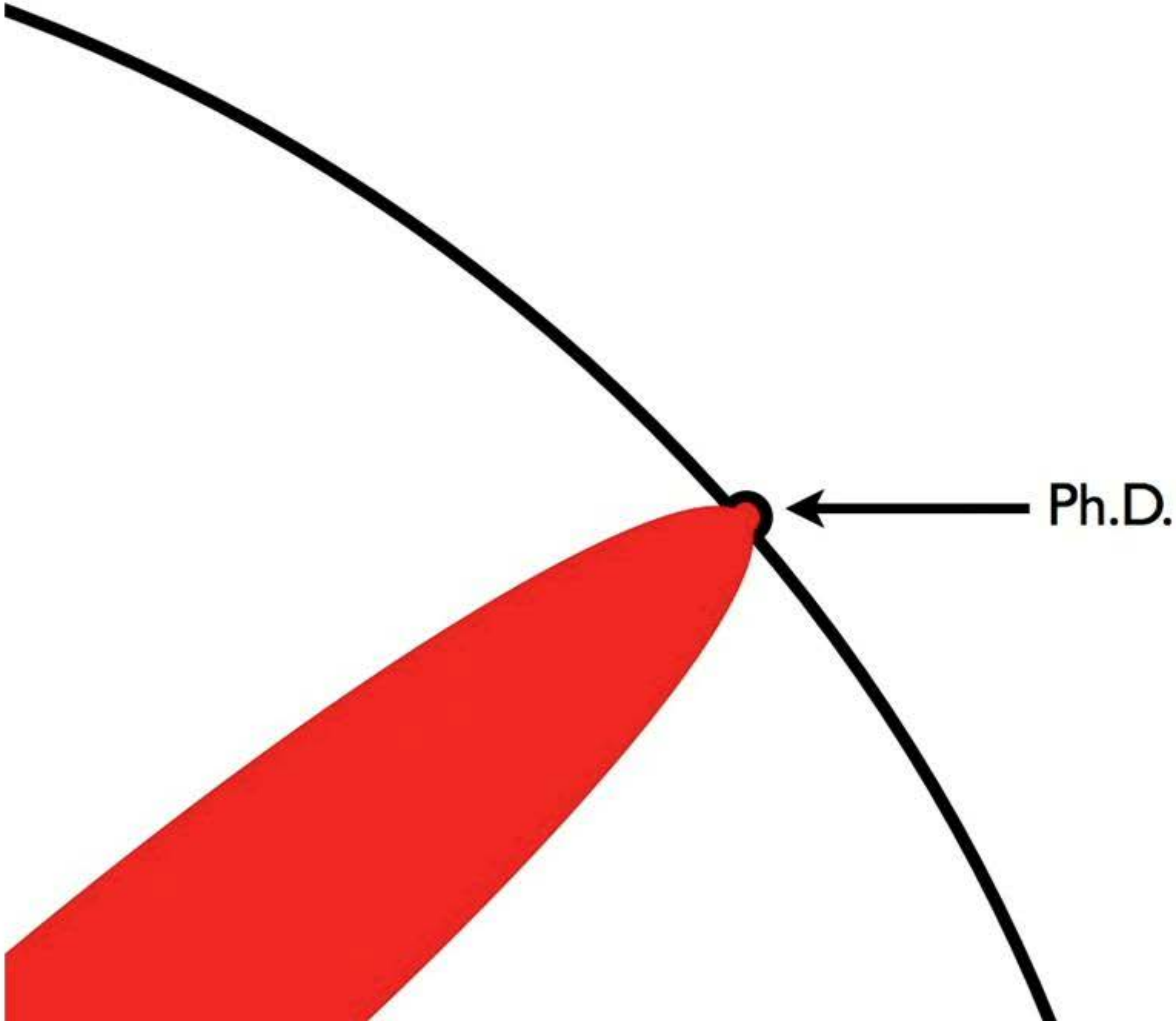


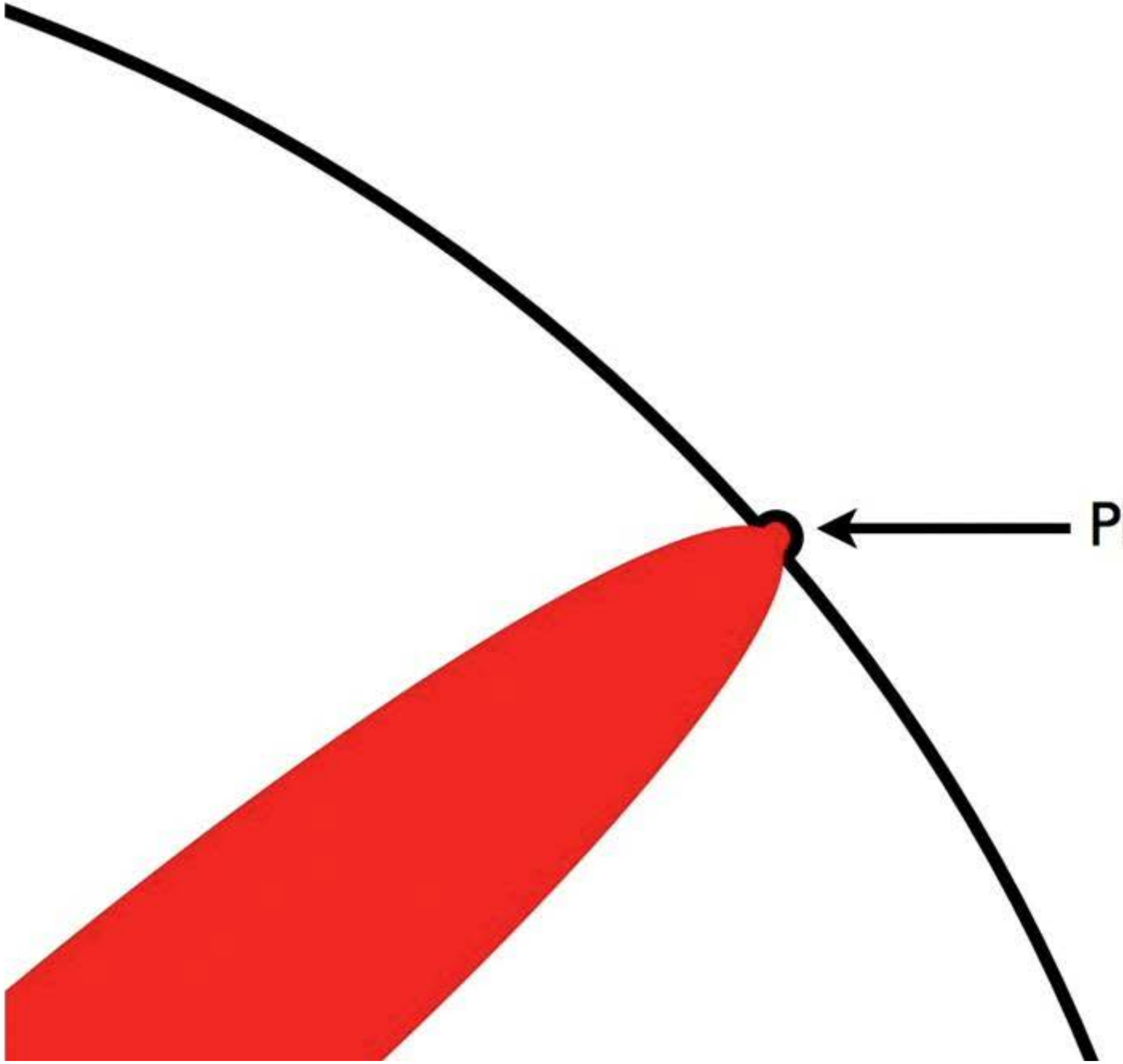




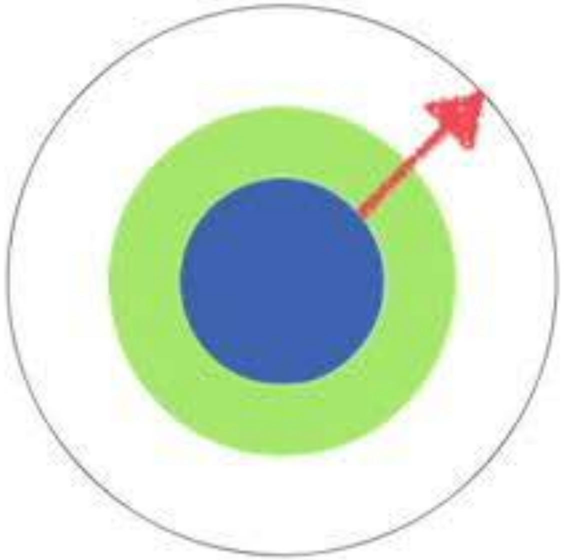


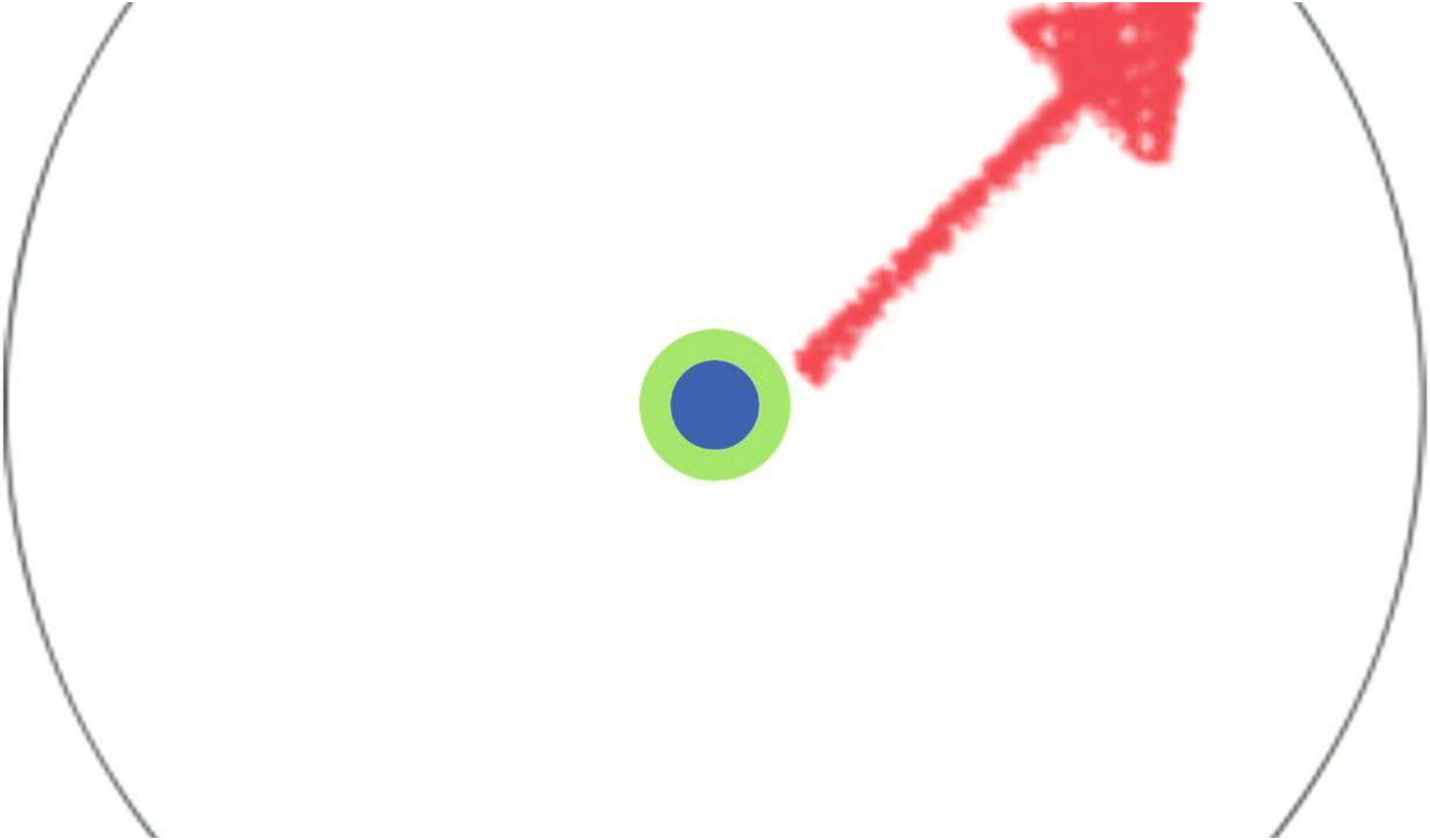


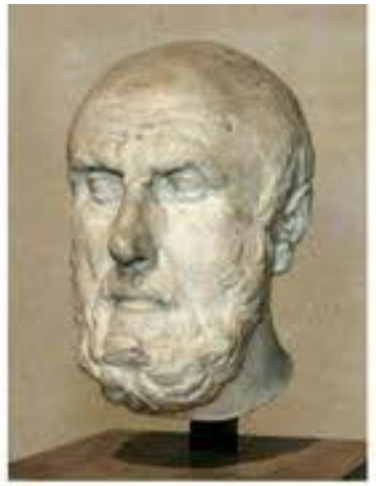

Ph.D.



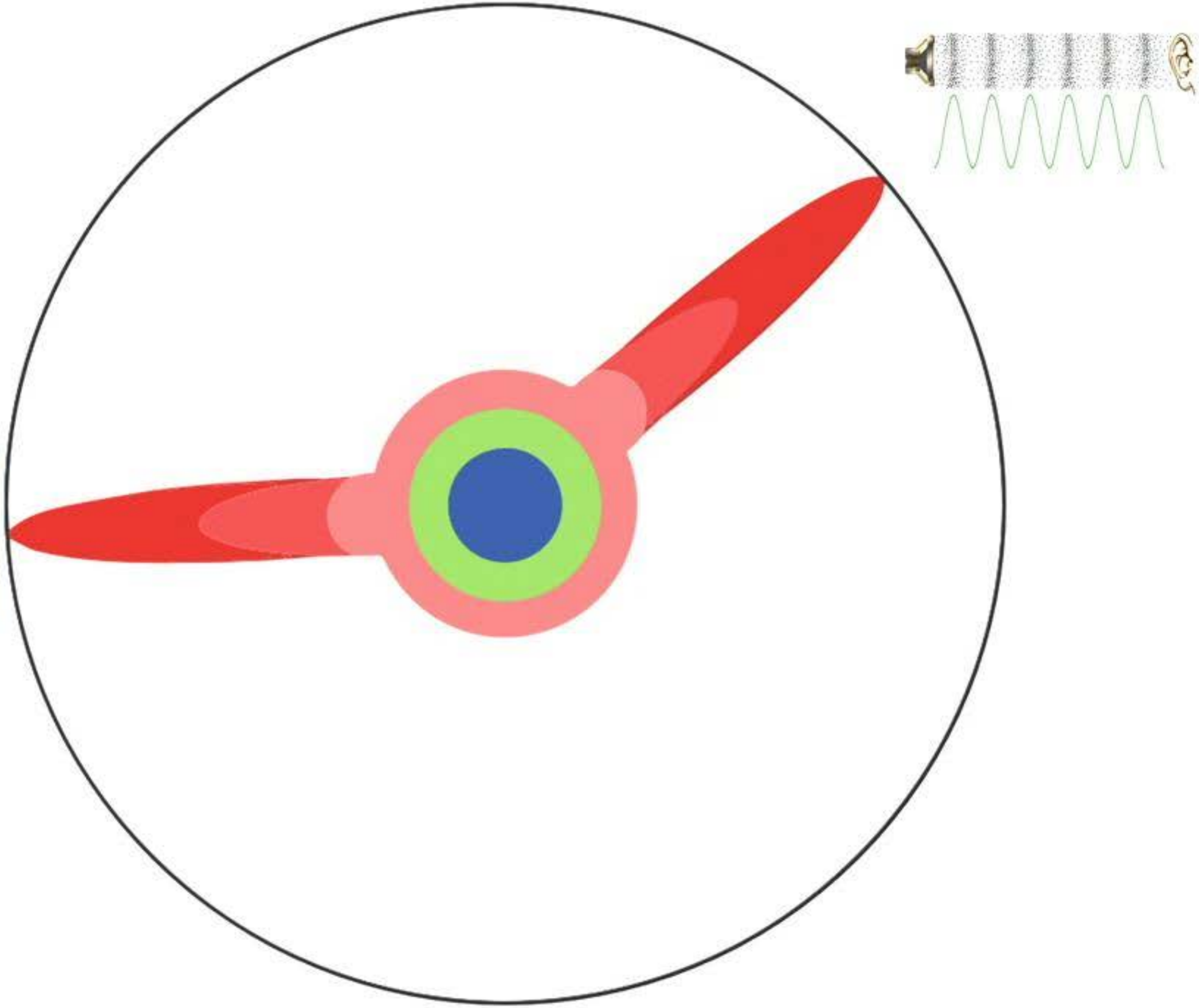
Ph.D.

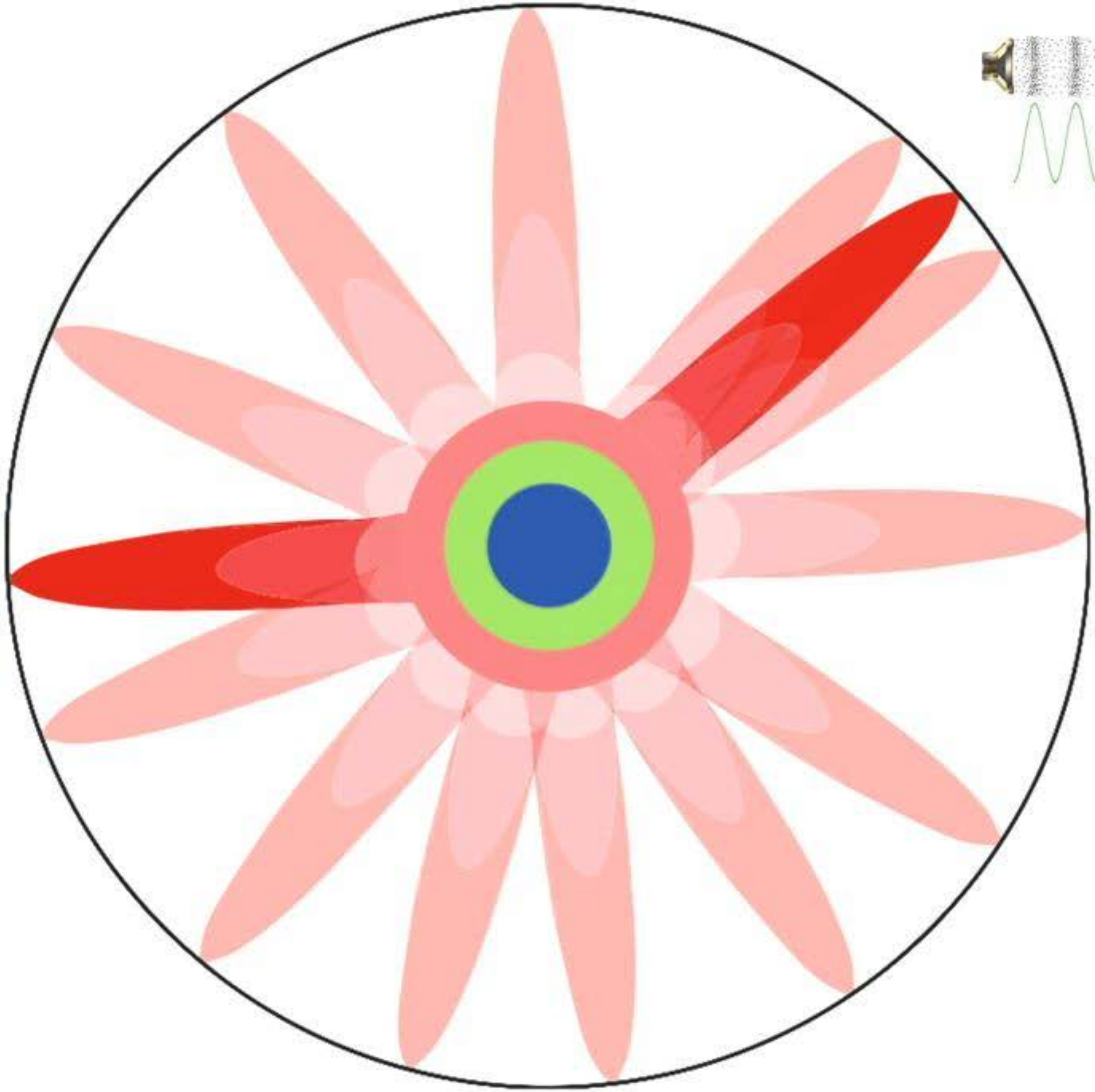


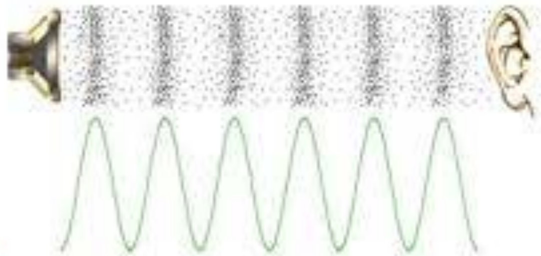


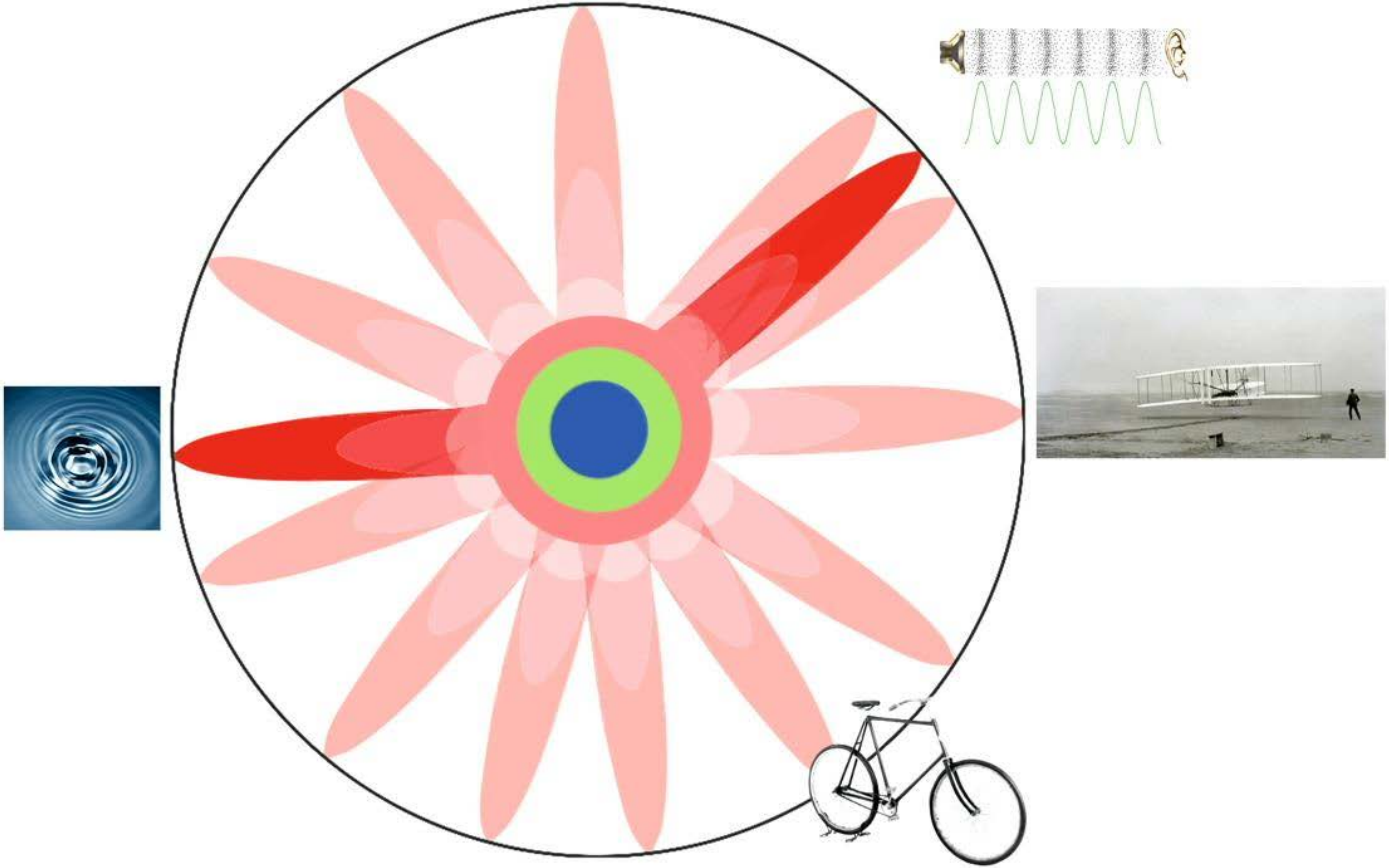


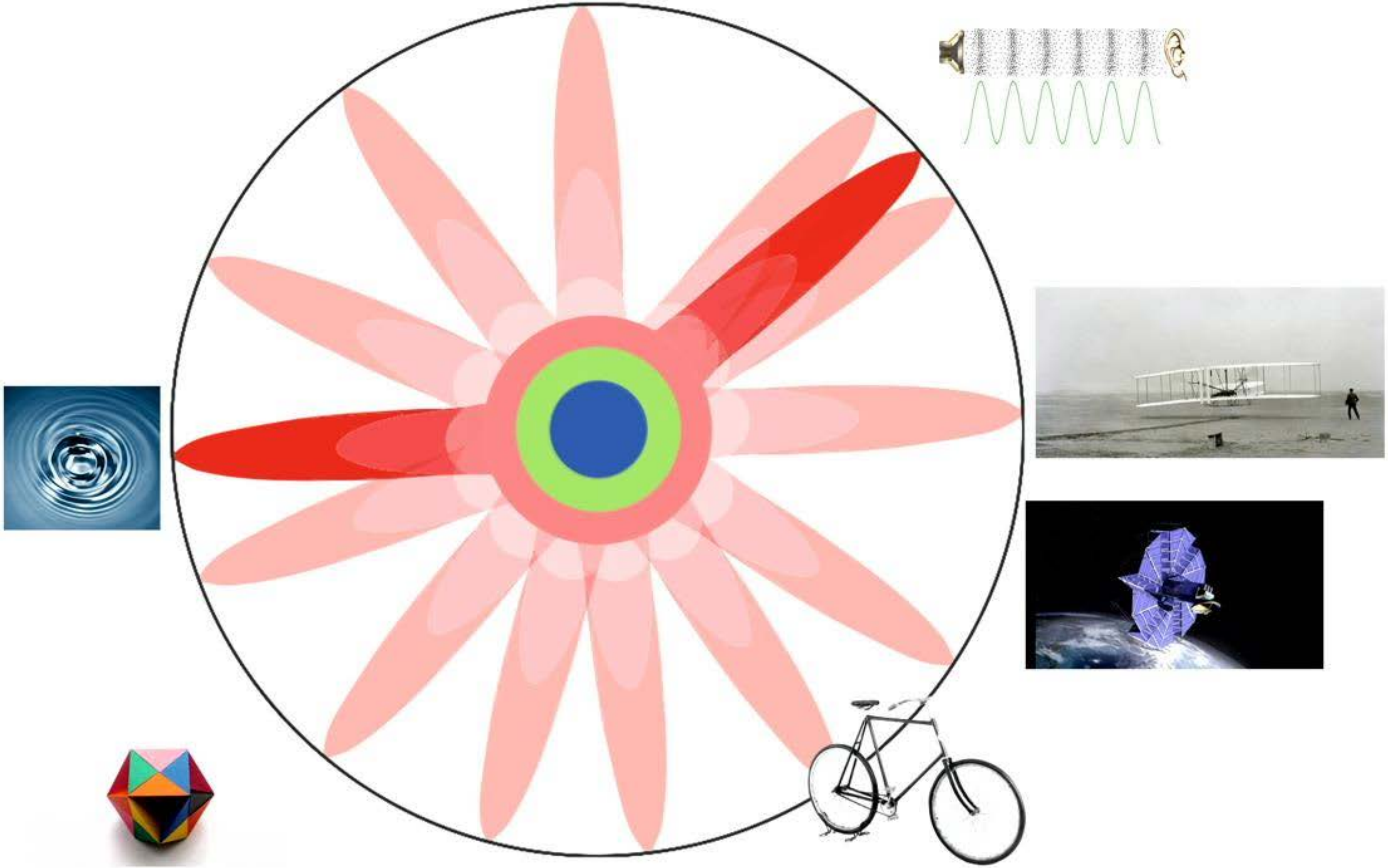
Chrysippus, 279 B.C.





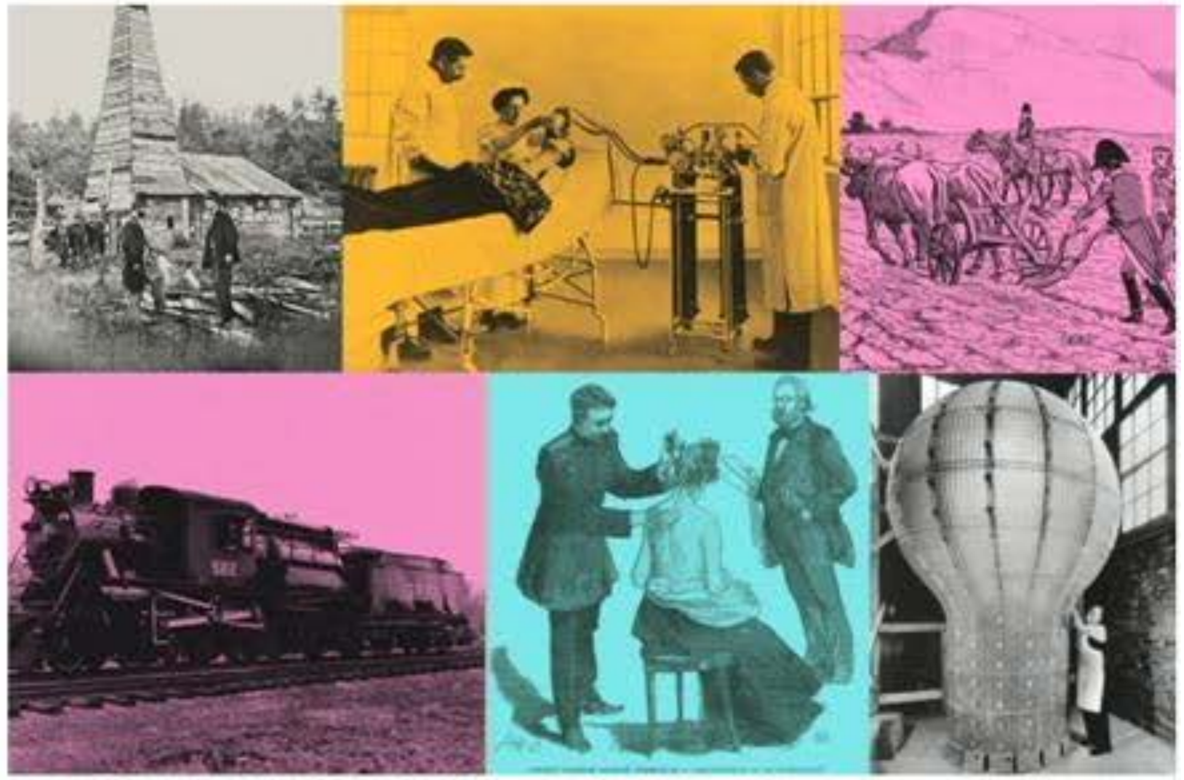






Online data: Big opportunity

- Large idea repositories
  - USPTO (millions of patents)
  - Scientific publications
  - Kickstarter, InnoCentive, Quirky
  - …



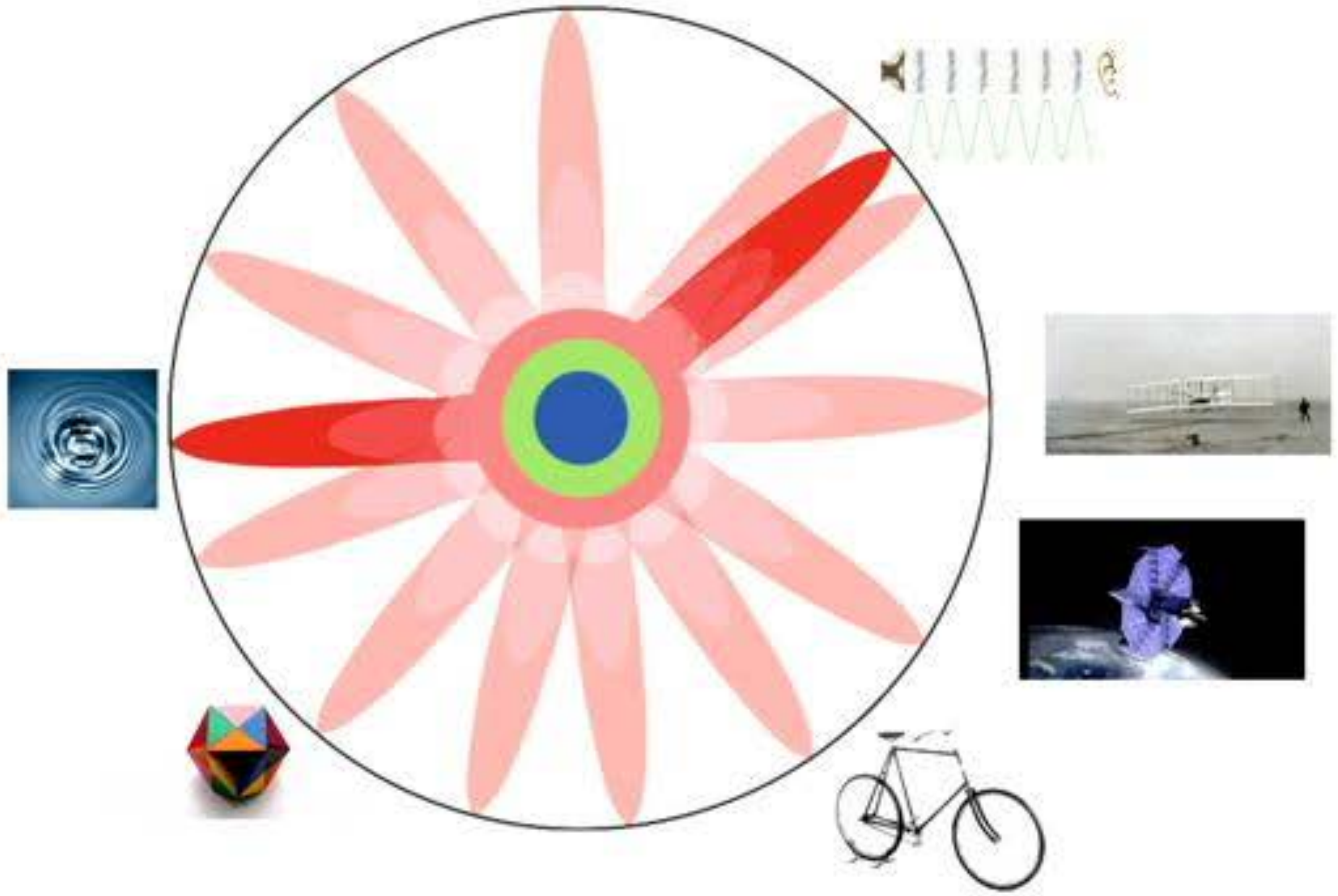
- **Opportunity ***
  - Accelerate problem solving, innovation and discovery
  - Inspirations, analogies, identify promising new directions…
- **Challenge**
  - Large, messy datasets – daunting for humans
  - How can we help automate discovery and innovation?

hyadata lab PNAS

***** *Scaling up analogical innovation with crowds and AI, PNAS 2019***

In this talk

- Boosting creativity with analogies
- Identifying novel ideas with weak supervision



Innovation through analogies

- Lots of important inventions involved analogies
  - Thomas Edison: *"A logical mind that sees analogies"*
  - Salvador Luria: Slot machines -> bacteria mutations (Nobel)
  - Genetic algorithms, simulated annealing…

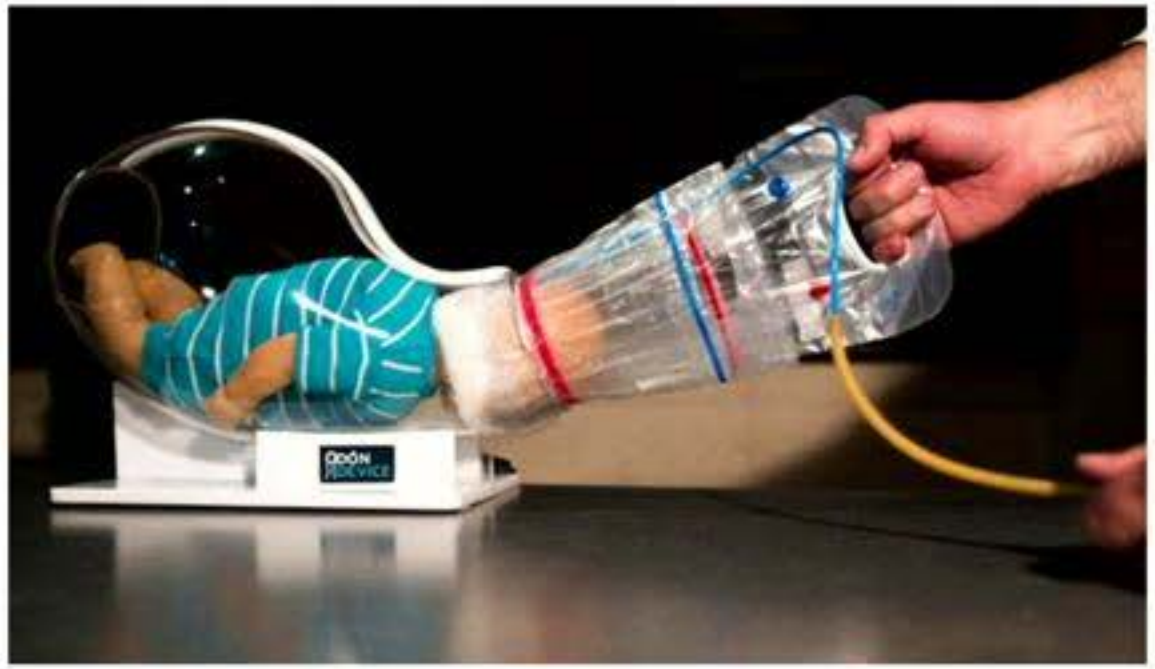
http://www.nytimes.com/2013/11/14/health/new-tool-to-ease-difficult-births-a-plastic-bag.html

Innovation through analogies

- Lots of important inventions involved analogies
  - Thomas Edison: *"A logical mind that sees analogies"*
  - Salvador Luria: Slot machines -> bacteria mutations (Nobel)
  - Genetic algorithms, simulated annealing…
- A car mechanic: extracting babies stuck in birth canal

"For many years, there has been no innovation in this area"

World Health Organization



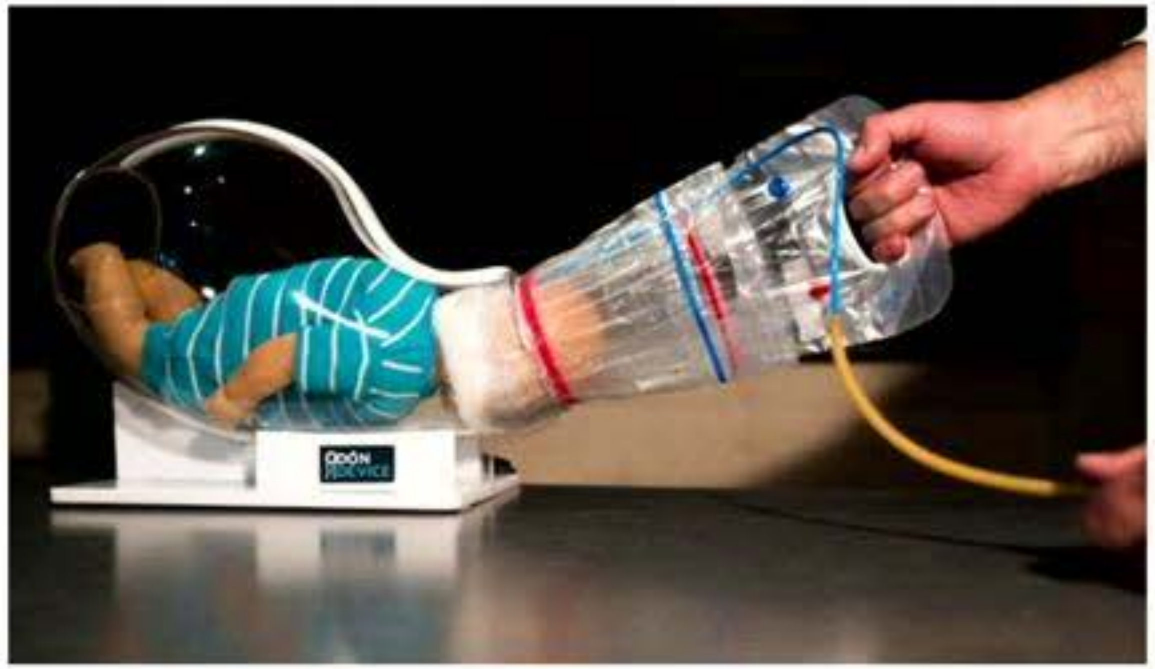
http://www.nytimes.com/2013/11/14/health/new-tool-to-ease-difficult-births-a-plastic-bag.html

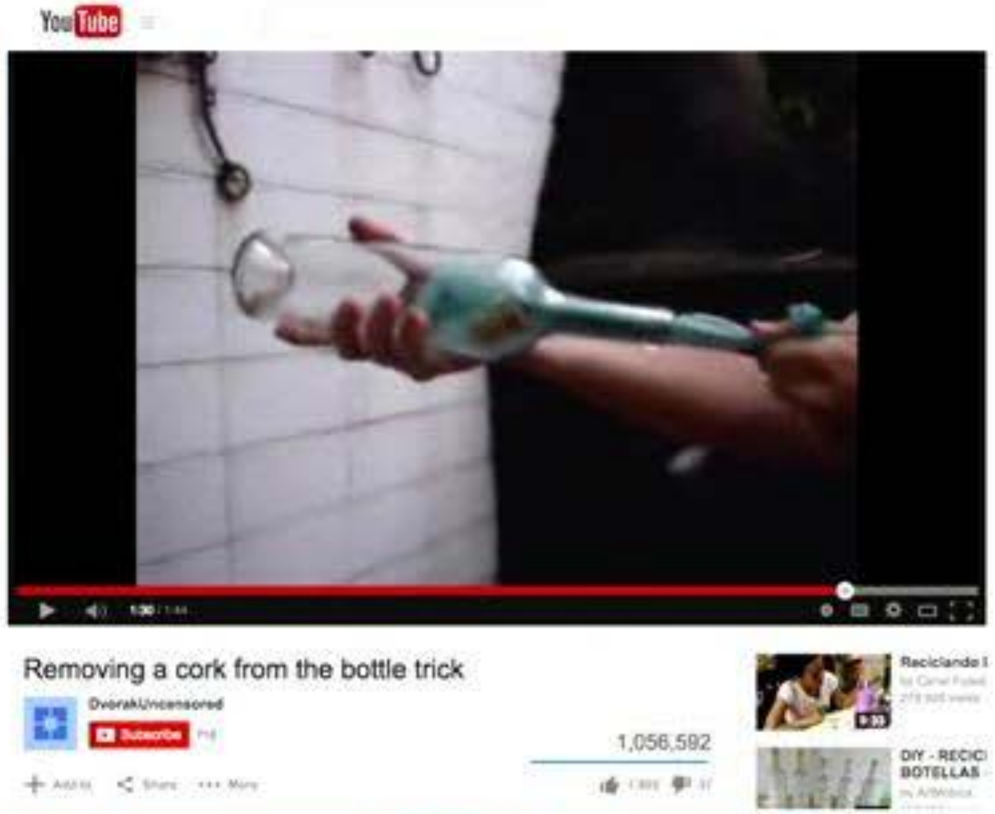
Innovation through analogies

- Lots of important inventions involved analogies
  - Thomas Edison: *“A logical mind that sees analogies”*
  - Salvador Luria: Slot machines -> bacteria mutations (Nobel)
  - Genetic algorithms, simulated annealing…
- A car mechanic: extracting babies stuck in birth canal

“For many years, there has been no innovation in this area”

World Health Organization





Removing a cork from the bottle trick

DvorakUncensored

1,056,592

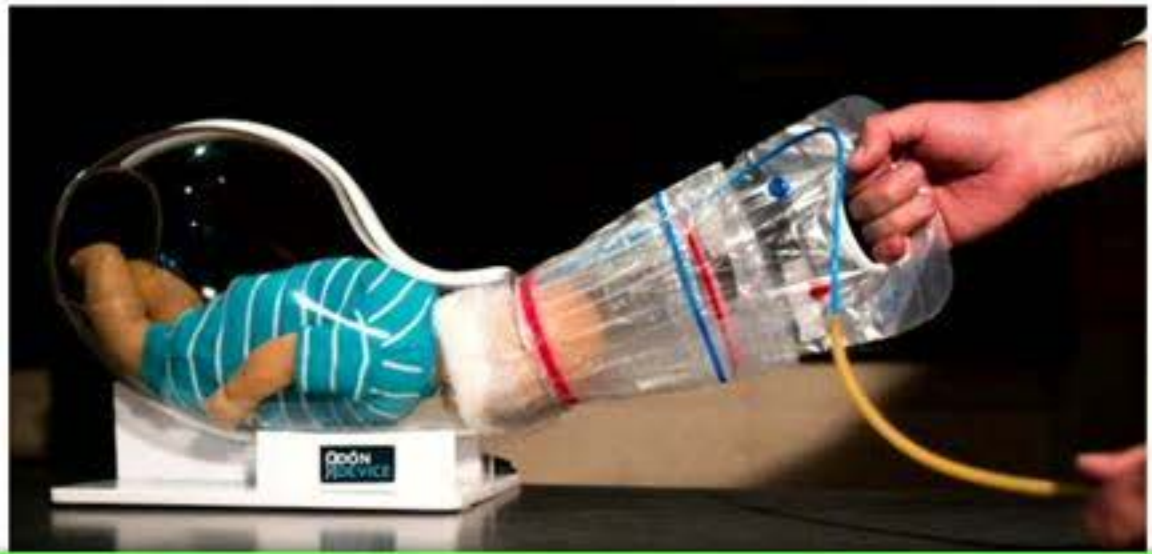
http://www.nytimes.com/2013/11/14/health/new-tool-to-ease-difficult-births-a-plastic-bag.html

Innovation through analogies

- Lots of important inventions involved analogies
  - Thomas Edison: *“A logical mind that sees analogies”*
  - Salvador Luria: Slot machines -> bacteria mutations (Nobel)
  - Genetic algorithms, simulated annealing…
- A car mechanic: extracting babies stuck in birth canal

“For many years, there has been no innovation in this area”

World Health Organization



Our Goal: Automatically discover analogies in large, unstructured data sets

Analogies: Hard for machines (1)

- NLP/IR methods excel at surface similarity
  - Especially with same source and target domains

Analogies: Hard for machines (1)

- NLP/IR methods excel at surface similarity
  - Especially with same source and target domains

Heavy duty and hard ice scraper blade is capable of pushing the wettest, deepest snow from your entire vehicle…

Burnt on oil can attach itself to the pan in a permanent fashion. Remove the oil easily with…

Analogies: Hard for machines (2)

- Hand-created databases, high relational structure

```
ENABLE( ENABLE( ENABLE(
      SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
      INSERT_INTO(person, plastic_bag, bottle)),
      CAUSE(CAUSE(
            INSERT_INTO(person,air, plastic_bag),
            EXPAND(air, plastic_bag))
      ENCLOSE(plastic_bag, cork)) ),
      CAUSE(
      PULL(person, plastic_bag),
      ¬IN(cork, bottle)))
```

- Difficult to obtain, does not scale
  - 40-100 person-hours for complex products [Vattam ’11]

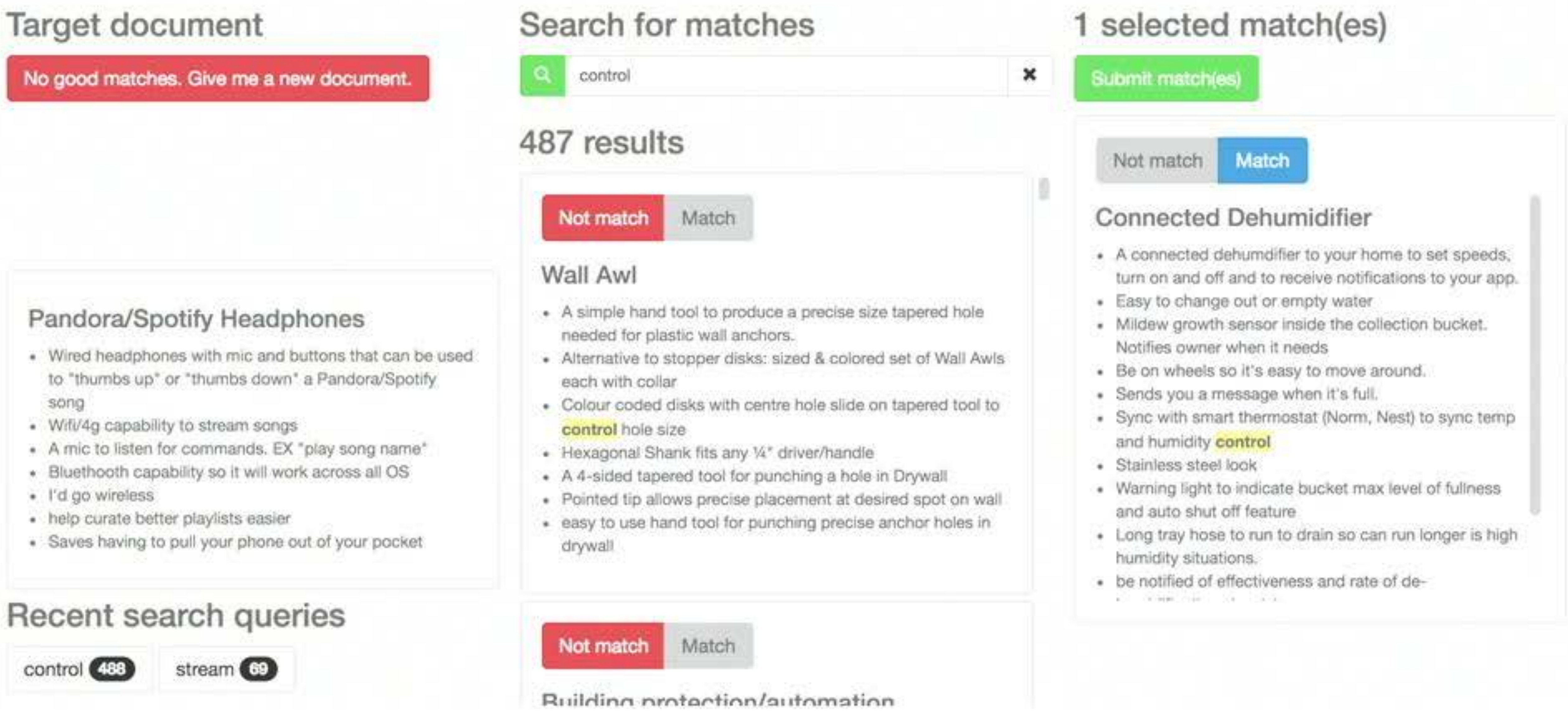
Attempt 1: Supervised learning

- Data: ~10,000 crowdsourced product ideas from Quirky.com
  - Many categories
  - Natural (messy) language

123 4FUN .

> *An Educational Math Mat 4 Kids.*
> *numbers in any configuration 4 learning to De / Composing Numbers.*
> *Folds Up Perfect For Carrying.*
> *you can walk-on , put your mouth on and or hands on.*
> *Learn & play games on with 4 single or multiple players hands on.*
>
> *A mix between "Twister", "Simon Says", and "Whac-A-Mole".*
> *A mat with Number Blocks.*
> *A mat with numbers in any configuration 4 learning Place Values.*
> *Give commands like "what 2 numbers added together will give you 10".*
> *Hypo-allergenic.*
> *Have a whac-a-mole type mallet to hit the numbers.*
> *A mat with any shape and color with a number.*
> *A mat that lights up numbers in any type of order and or formula.*

Search interface



Target document

Pandora/Spotify Headphones

- Wired headphones with mic and buttons that can be used to "thumbs up" or "thumbs down" a Pandora/Spotify song
- Wifi/4g capability to stream songs
- A mic to listen for commands. EX "play song name"
- Bluethooth capability so it will work across all OS
- I'd go wireless
- help curate better playlists easier
- Saves having to pull your phone out of your pocket

Recent search queries

control 488 stream 69

Search for matches

control

487 results

Wall Awl

- A simple hand tool to produce a precise size tapered hole needed for plastic wall anchors.
- Alternative to stopper disks: sized & colored set of Wall Awls each with collar
- Colour coded disks with centre hole slide on tapered tool to control hole size
- Hexagonal Shank fits any ¼" driver/handle
- A 4-sided tapered tool for punching a hole in Drywall
- Pointed tip allows precise placement at desired spot on wall
- easy to use hand tool for punching precise anchor holes in drywall

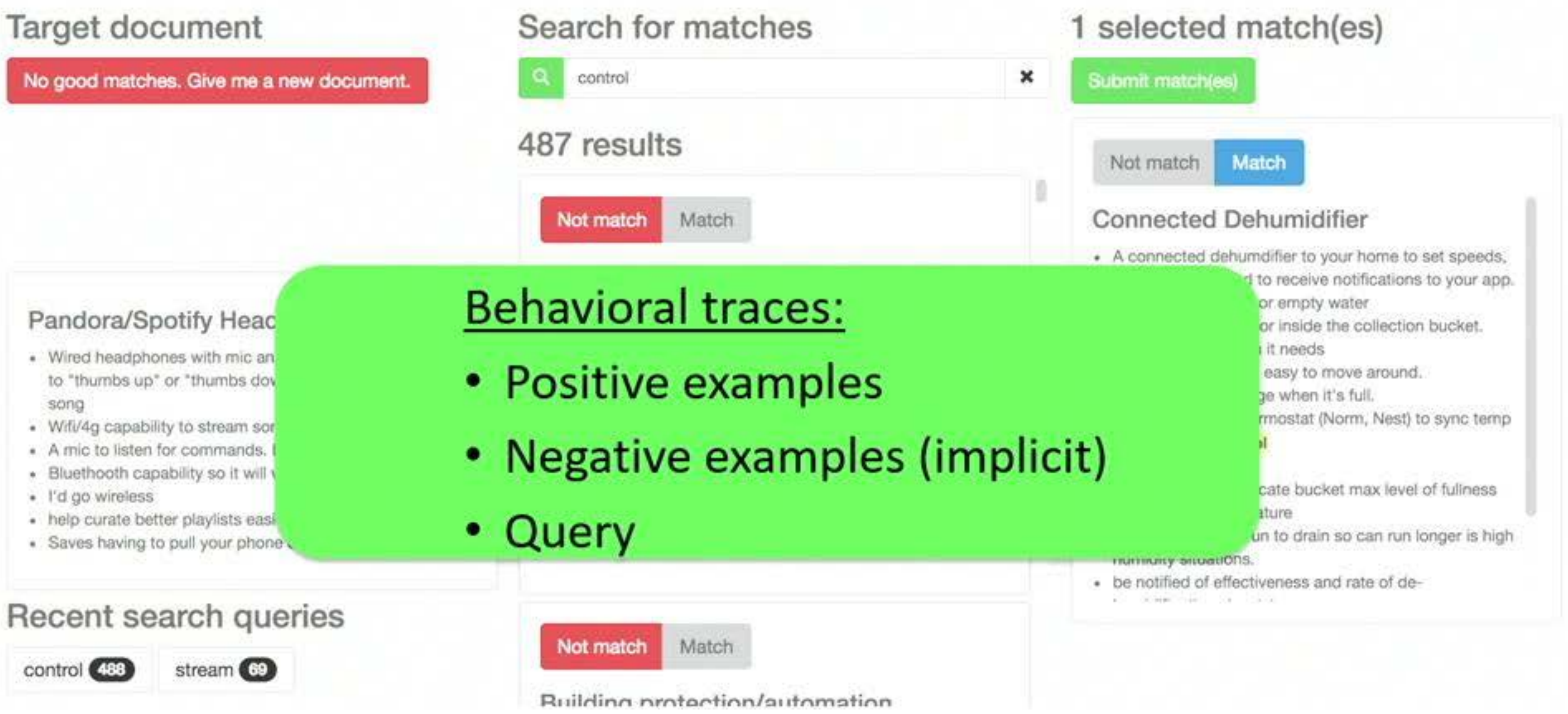
Building protection/automation

1 selected match(es)

Connected Dehumidifier

- A connected dehumdifier to your home to set speeds, turn on and off and to receive notifications to your app.
- Easy to change out or empty water
- Mildew growth sensor inside the collection bucket. Notifies owner when it needs
- Be on wheels so it's easy to move around.
- Sends you a message when it's full.
- Sync with smart thermostat (Norm, Nest) to sync temp and humidity control
- Stainless steel look
- Warning light to indicate bucket max level of fullness and auto shut off feature
- Long tray hose to run to drain so can run longer is high humidity situations.
- be notified of effectiveness and rate of de-

Search interface



Target document

No good matches. Give me a new document.

Pandora/Spotify Heac

- Wired headphones with mic an to "thumbs up" or "thumbs dov song
- Wifi/4g capability to stream sor
- A mic to listen for commands.
- Bluethooth capability so it will
- I'd go wireless
- help curate better playlists easi
- Saves having to pull your phone

Recent search queries

control 488 stream 69

Search for matches

control

487 results

Not match Match

Not match Match

Building protection/automation

1 selected match(es)

Submit match(es)

Not match Match

Connected Dehumidifier

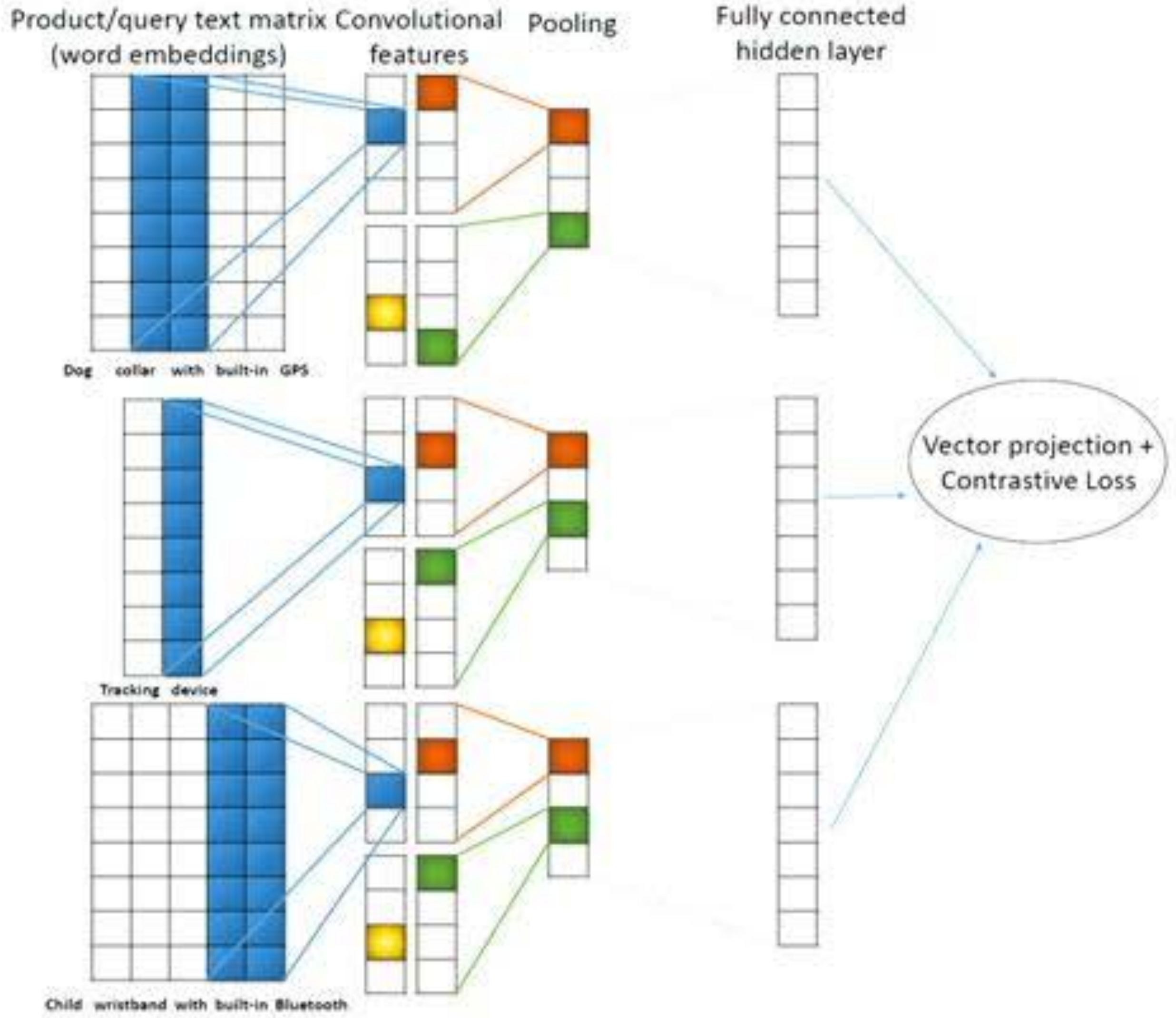
- A connected dehumdifier to your home to set speeds, d to receive notifications to your app. or empty water or inside the collection bucket. it needs easy to move around. ge when it's full. rmostat (Norm, Nest) to sync temp cate bucket max level of fullness ature un to drain so can run longer is high humidity situations.
- be notified of effectiveness and rate of de-

<u>Behavioral traces:</u>

- Positive examples
- Negative examples (implicit)
- Query

Supervised model

- Learn a similarity metric reflecting analogy
  - CNN-based network, contrastive Loss
- Incorporate user queries
  - Helps model focus on non-surface features


Product/query text matrix (word embeddings)
Convolutional features
Pooling
Fully connected hidden layer
Dog collar with built-in GPS
Tracking device
Child wristband with built-in Bluetooth
Vector projection + Contrastive Loss

** Scaling up **Analogy** with Crowdsourcing and Machine Learning, ICCBR-16*

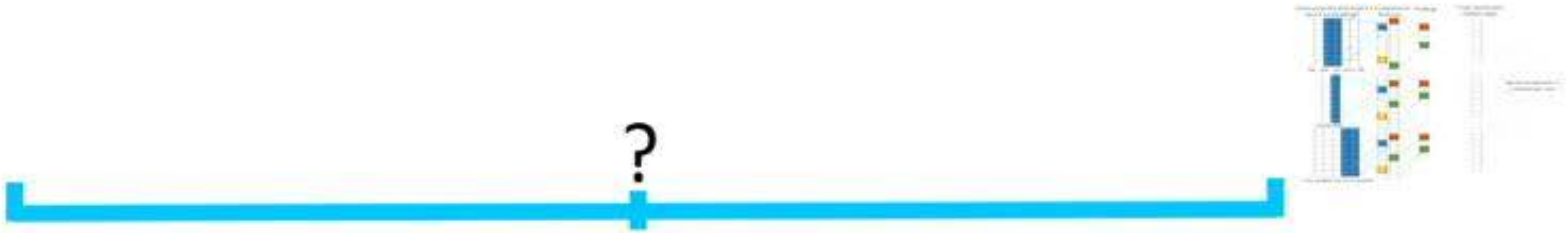
Results: The good and the bad

- Precision, Recall@K: beat standard retrieval baselines
  - Baselines: TFIDF-based, embedding based
  - With query: better!

- Labels costly, noisy, slow (hard for humans)
  - Superficial matches, near analogies
  - Data scarcity
- Model blind to rich structures
  - Hard to learn from data

Attempt 2: Weak structural representations

- Goal: find weak structural representations
  - Expressive enough for analogy mining
  - Can be learned



```
ENABLE( ENABLE( ENABLE(
            SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
            INSERT_INTO(person, plastic_bag, bottle)),
            CAUSE(CAUSE(
                    INSERT_INTO(person,air, plastic_bag),
                    EXPAND(air, plastic_bag))
            ENCLOSE(plastic_bag, cork)) ),
        CAUSE(
            PULL(person, plastic_bag),
            ¬IN(cork, bottle)))
```

**Accelerating Innovation Through Analogy Mining*

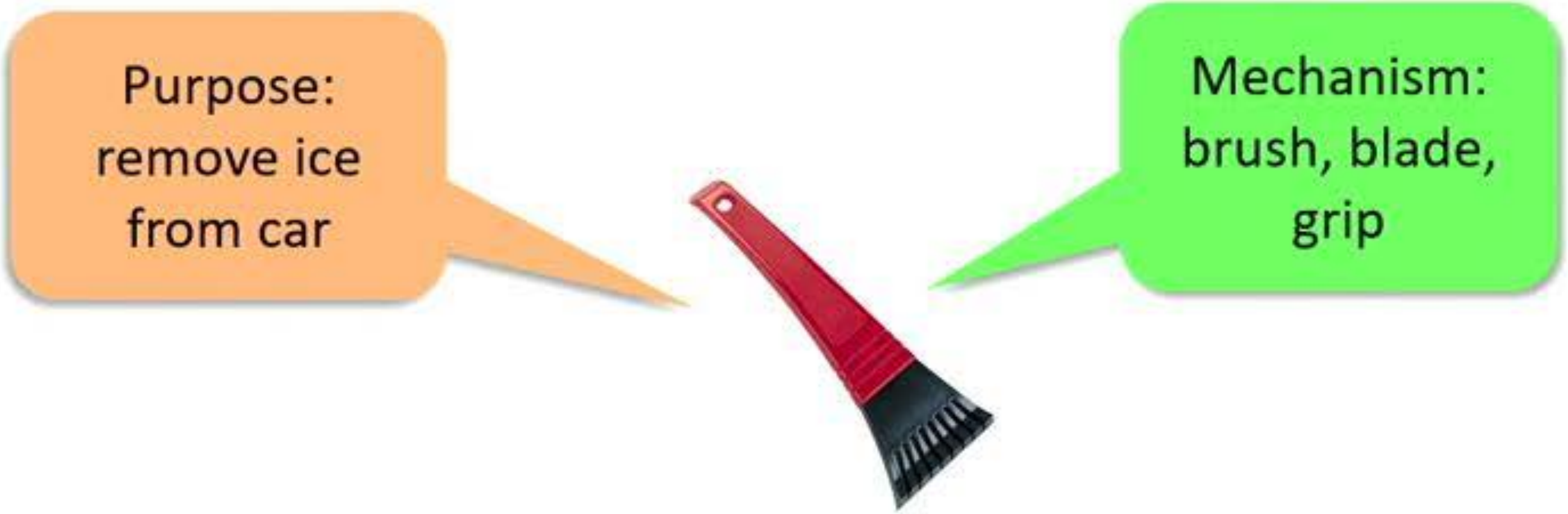
KDD '17 Best Research Paper + Best Student Paper

Purpose and mechanism

- Rooted in cognitive psychology (schema induction)
- Analogies deeply related to **purpose/mechanism**
- Purpose: what it does, what it is used for
- Mechanism: how it does it, how it works

Purpose and mechanism

- Rooted in cognitive psychology (schema induction)
- Analogies deeply related to **purpose/mechanism**
  - Purpose: what it does, what it is used for
  - Mechanism: how it does it, how it works



Purpose: remove ice from car

Mechanism: brush, blade, grip

Purpose and mechanism

- Enables core analogical tasks:
    - *Same* purpose, *different* mechanism

$$\underset{\tilde{\mathbf{i}}\in\mathcal{P}}{\operatorname{argmin}}\, d_p(\mathbf{p}_i, \mathbf{p}_{\tilde{i}})$$
$$s.t. d_m(\mathbf{m}_i, \mathbf{m}_{\tilde{i}}) \geq \text{threshold}$$

    - *Same* mechanism, *different* purpose (Repurposing)

$$\underset{\tilde{\mathbf{i}}\in\mathcal{P}}{\operatorname{argmin}}\, d_m(\mathbf{m}_i, \mathbf{m}_{\tilde{i}})$$
$$s.t. d_p(\mathbf{p}_i, \mathbf{p}_{\tilde{i}}) \geq \text{threshold}$$

Purpose and mechanism

- Enables core analogical tasks:
  - *Same* purpose, *different* mechanism

$$\underset{\tilde{\mathbf{i}} \in \mathcal{P}}{\operatorname{argmin}}\ d_p(\mathbf{p}_i, \mathbf{p}_{\tilde{i}})$$

$$s.t. d_m(\mathbf{m}_i, \mathbf{m}_{\tilde{i}}) \geq \text{threshold}$$

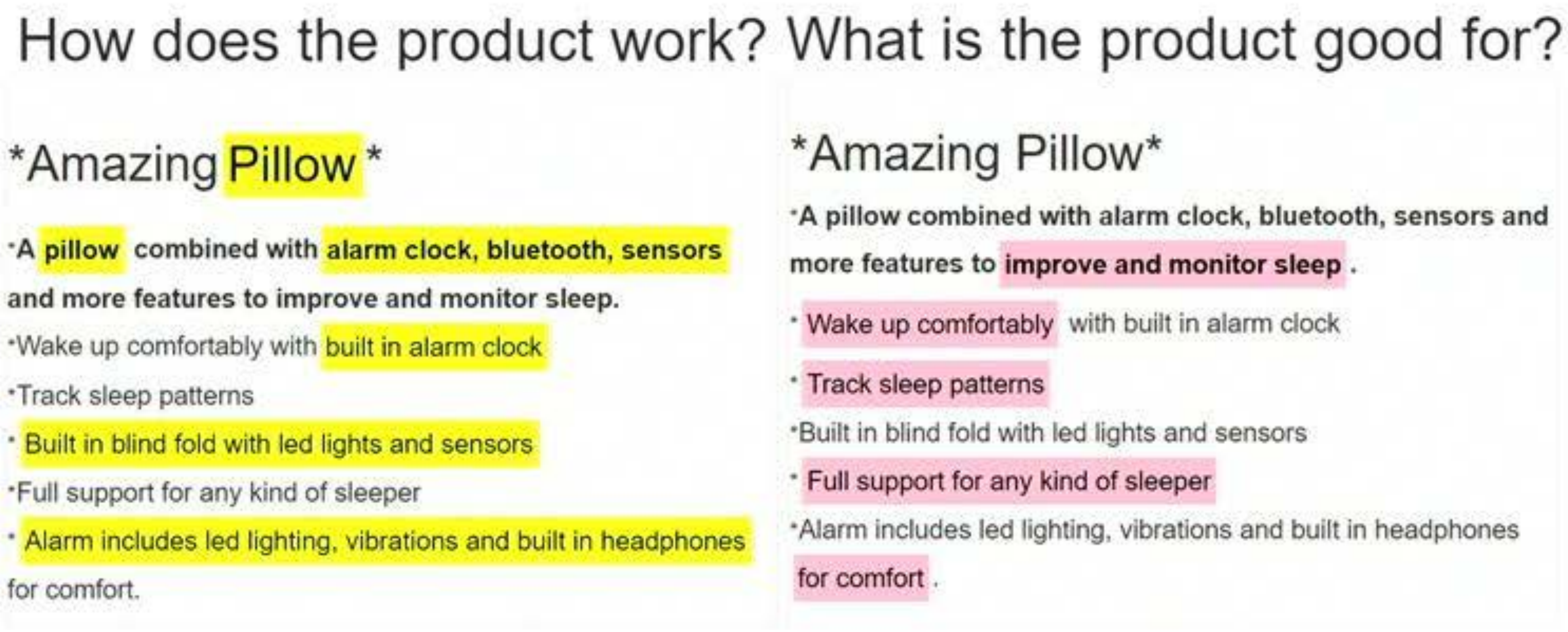
  - *Same* mechanism, *different* purpose (Repurposing)

Our Goal: Learn vector representations capturing the **purposes** and **mechanisms** of inventions

$$s.t. d_p(\mathbf{p}_i, \mathbf{p}_{\tilde{i}}) \geq \text{threshold}$$

Collecting purpose/mechanism

- Mechanical Turk annotations
- 4 workers per product



How does the product work?

Amazing Pillow

·A pillow combined with alarm clock, bluetooth, sensors and more features to improve and monitor sleep.

·Wake up comfortably with built in alarm clock

·Track sleep patterns

· Built in blind fold with led lights and sensors

·Full support for any kind of sleeper

· Alarm includes led lighting, vibrations and built in headphones for comfort.

What is the product good for?

Amazing Pillow

·A pillow combined with alarm clock, bluetooth, sensors and more features to improve and monitor sleep .

· Wake up comfortably with built in alarm clock

· Track sleep patterns

·Built in blind fold with led lights and sensors

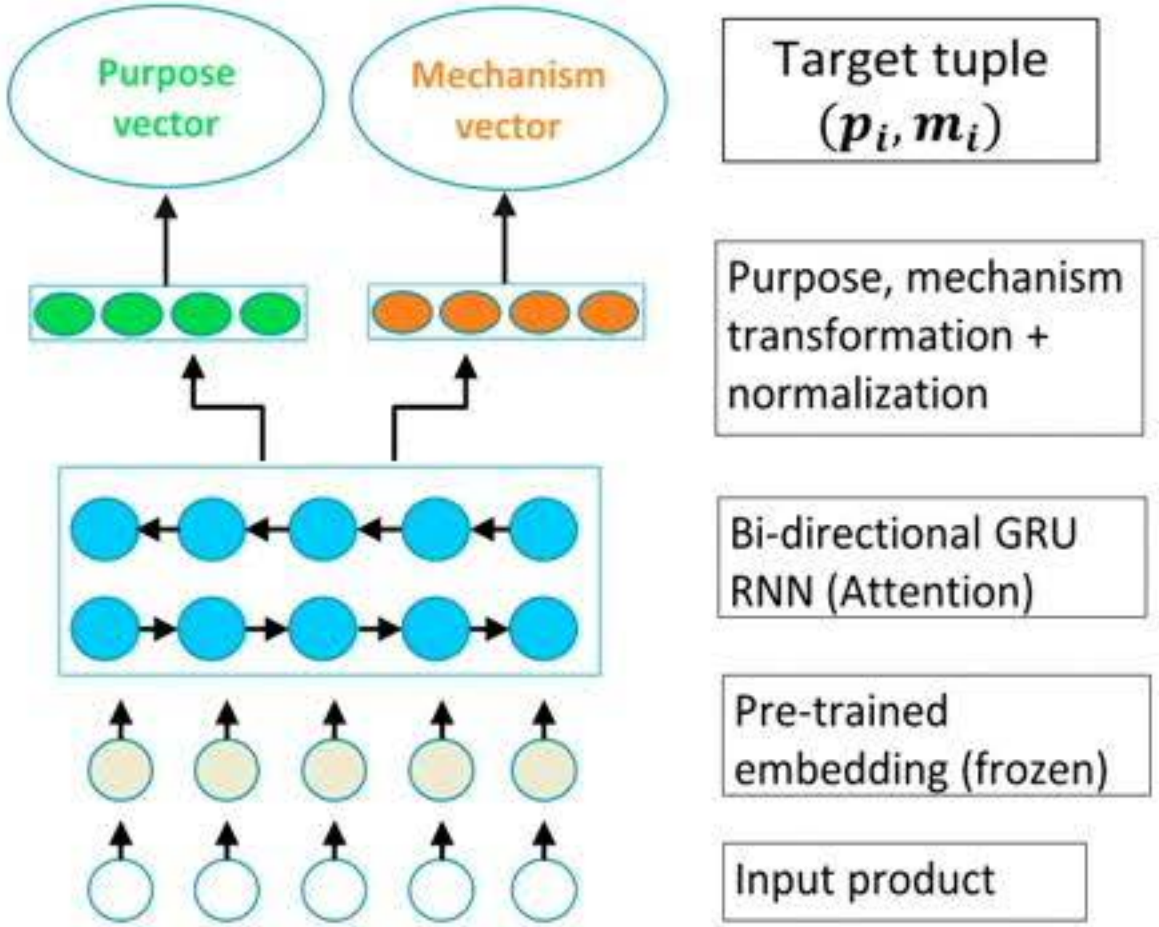
· Full support for any kind of sleeper

·Alarm includes led lighting, vibrations and built in headphones for comfort .

Purpose/mechanism targets: Soft vectors

- Aggregate across K annotations:
  - Purpose/mechanism TF-IDF weights
  - Weighted average of pre-trained word vectors (GloVe) *

    *Purpose: [1.5*sleep + 0.9*support…], Mechanism: [2.3*pillow + 1.1*sensors…]*
- Train model to predict soft purpose/mechanism vectors

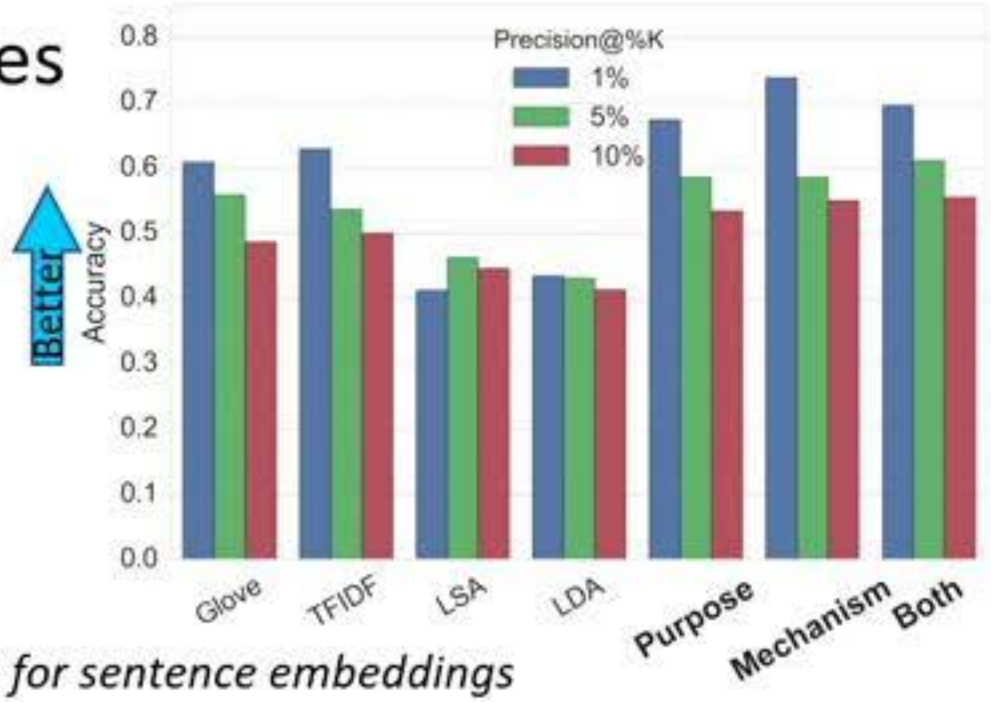

Purpose vector
Mechanism vector
Target tuple $(\boldsymbol{p_i}, \boldsymbol{m_i})$
Purpose, mechanism transformation + normalization
Bi-directional GRU RNN (Attention)
Pre-trained embedding (frozen)
Input product

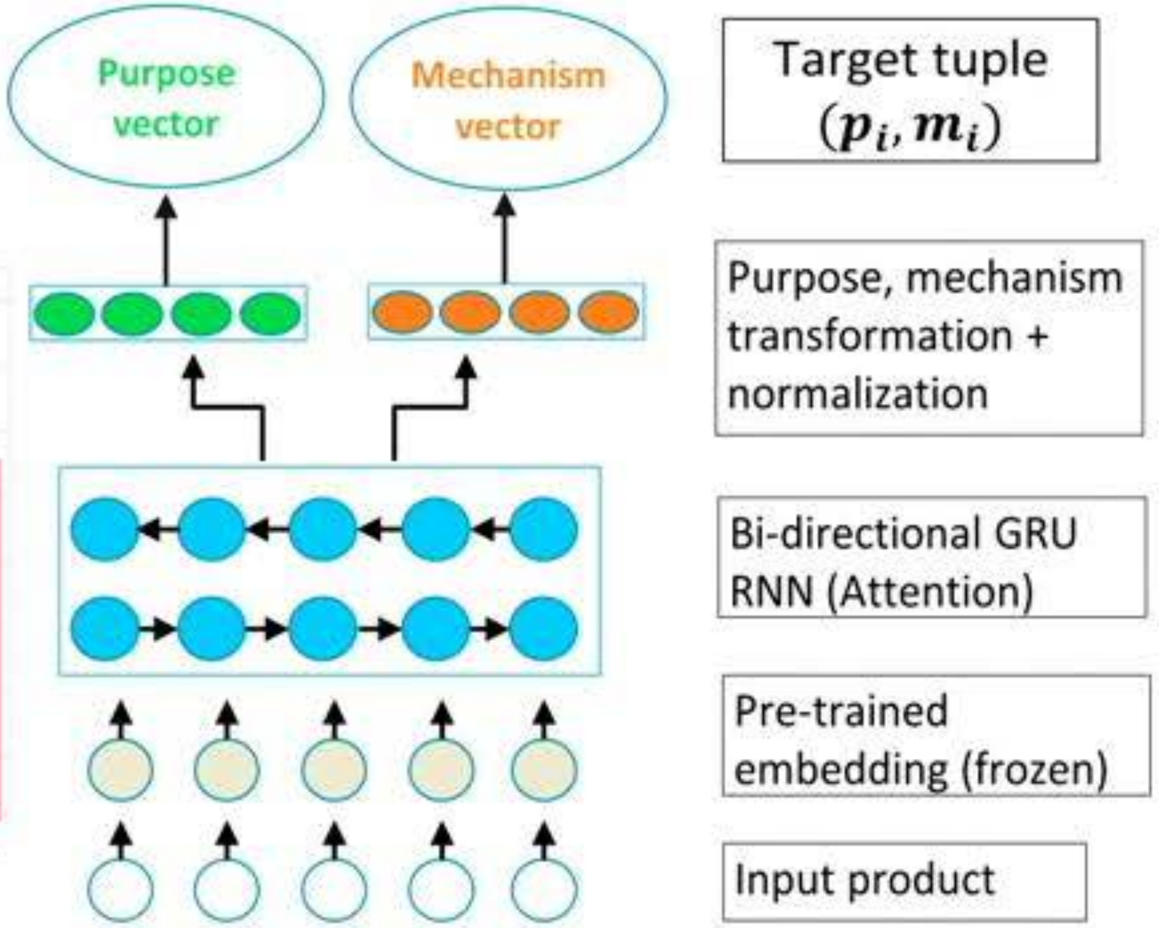
Arora et al. (ICLR 2017), A simple but tough-to-beat baseline for sentence embeddings

Purpose/mechanism targets: Soft vectors

- Aggregate across K annotations:
  - Purpose/mechanism TF-IDF weights
  - Weighted average of pre-trained word vectors (GloVe) *

    *Purpose: [1.5*sleep + 0.9*support…], Mechanism: [2.3*pillow + 1.1*sensors…]*
- Train model to predict soft purpose/mechanism vectors
- Outperforms standard retrieval
  - Despite noisy annotations and surface analogies favoring baselines





Target tuple $(\boldsymbol{p_i}, \boldsymbol{m_i})$

Purpose, mechanism transformation + normalization

Bi-directional GRU RNN (Attention)

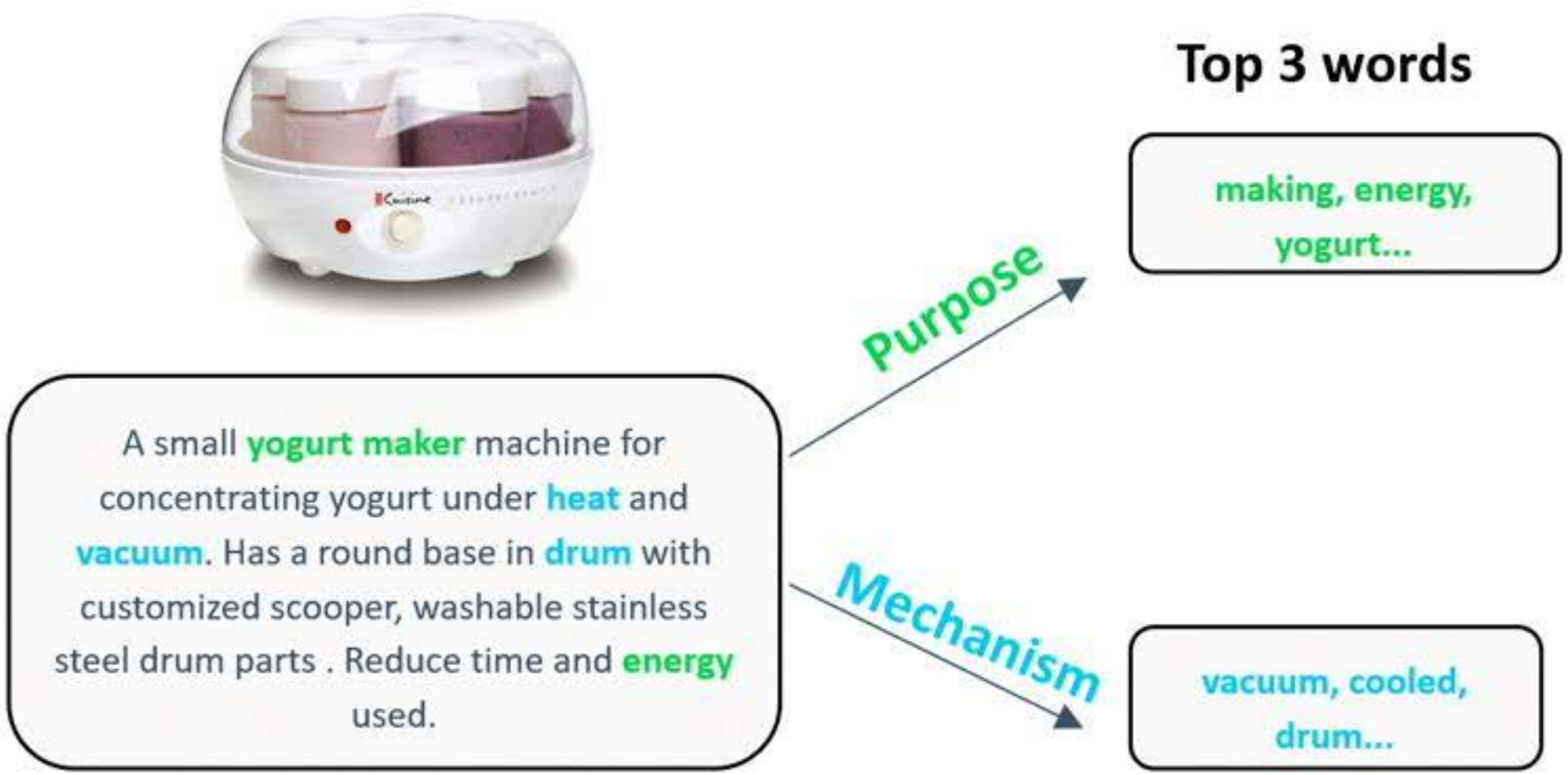
Pre-trained embedding (frozen)

Input product

Arora et al. (ICLR 2017), A simple but tough-to-beat baseline for sentence embeddings

Reality Check: Interpreting Vectors

- Sparse linear model of output vectors as combination of words *
  - Mechanism words indeed of more mechanical nature



A small **yogurt maker** machine for concentrating yogurt under **heat** and **vacuum**. Has a round base in **drum** with customized scooper, washable stainless steel drum parts . Reduce time and **energy** used.

Purpose

Mechanism

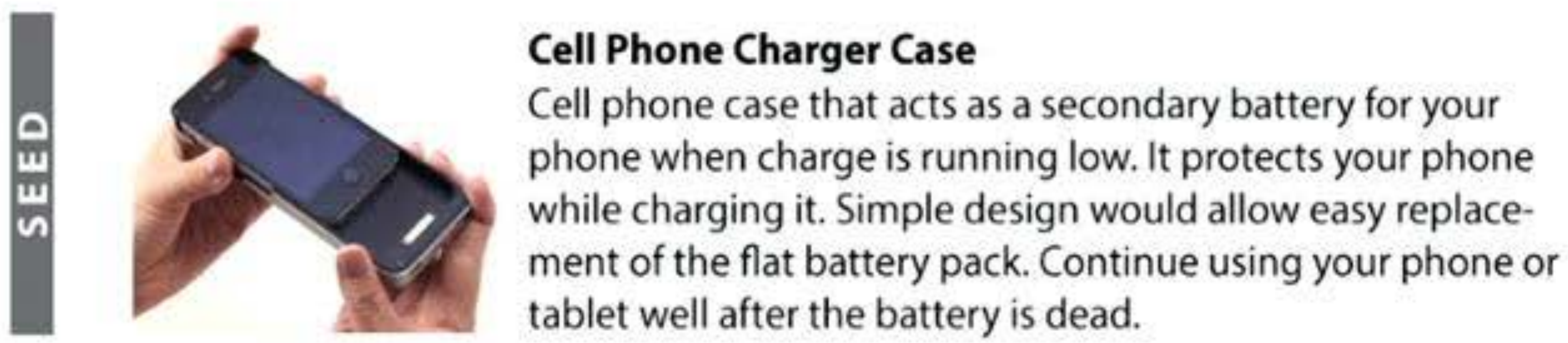
Top 3 words

making, energy, yogurt...

vacuum, cooled, drum...

* *Arora et al. (TACL 2018), Linear Algebraic Structure of Word Senses, with Applications to Polysemy*

Application: Ideation!



SEED

Cell Phone Charger Case

Cell phone case that acts as a secondary battery for your phone when charge is running low. It protects your phone while charging it. Simple design would allow easy replacement of the flat battery pack. Continue using your phone or tablet well after the battery is dead.

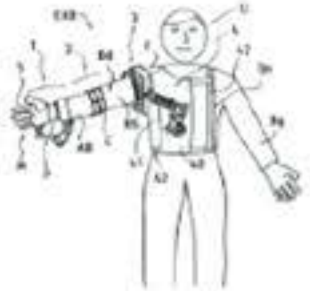
- Ideation common creative task - redesign an existing product
- Find other ways to solve same problem
- Inspiration: random, surface (TFIDF-based), ours
  - Ours: Near-purpose, far-mechanism (MAX-MIN diversification)
  - Assumption: our approach will help explore more diverse parts of the design space
    - Random: highly diverse, non-relevant
    - Surface: highly relevant, non-diverse
    - Ours: diverse, still relevant?

Ideation task: Inspirations

OUR APPROACH

Flash Charge Carabiner

USB tower with backup battery



Human pulley-powered generator suit

SURFACE (TF-IDF)

Multi-adapter case

Cell phone battery with GPS

Cell phone case with GPS

RANDOM

Solar pool skimmer

Shampoo pods

Dog meetup app

Results

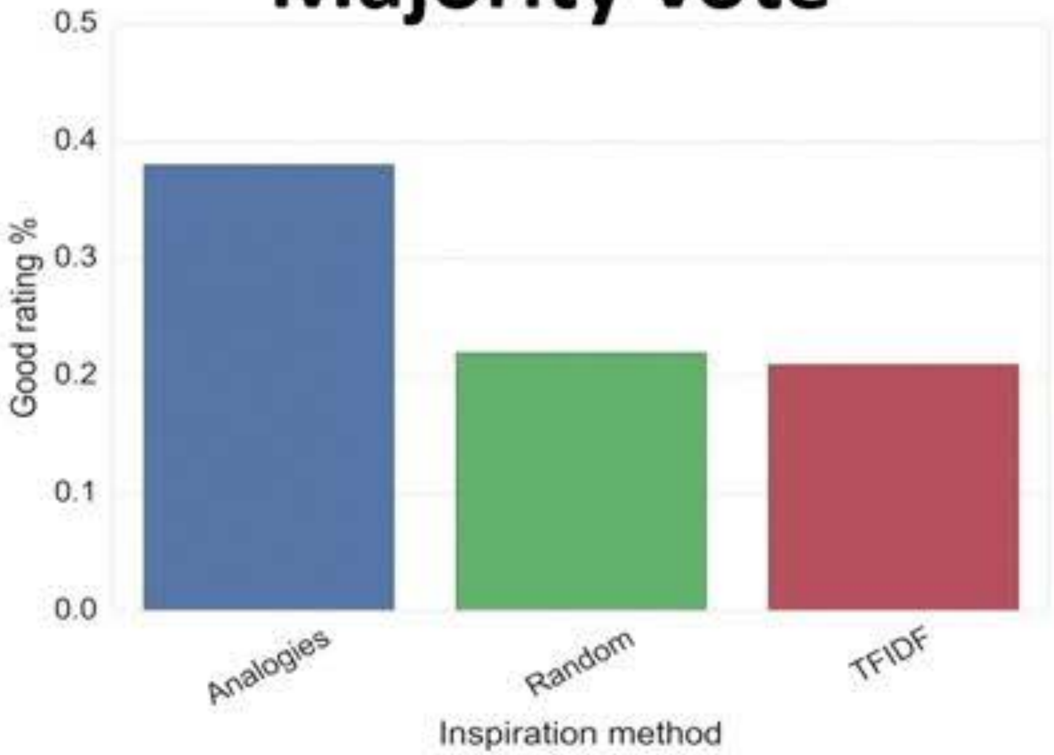
- Ideas externally rated by 5 judges. Criteria:
  - **Novelty:** Different technology than original product
  - **Quality:** Technology achieves same purpose as original product
  - **Feasibility:** Could be implemented, does not defy physics

- Substantial judge agreement (Fleiss kappa 0.51)

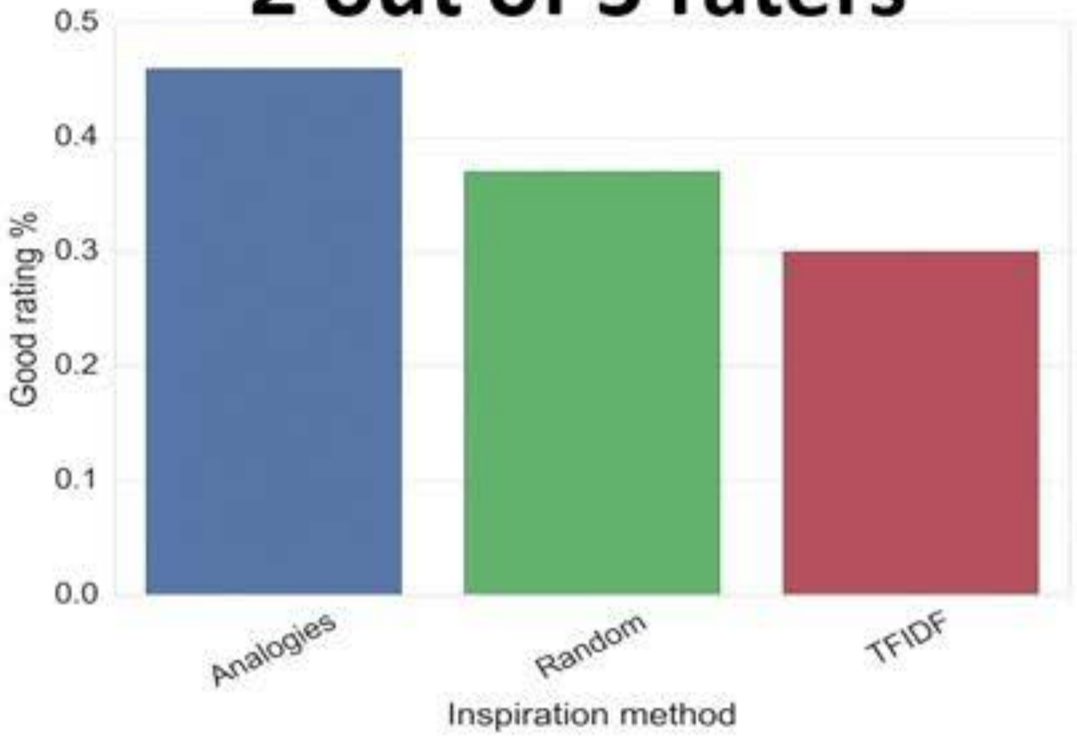
Results

- Both in terms of proportions and absolute number, our approach generates a considerably large relative positive effect

Majority vote

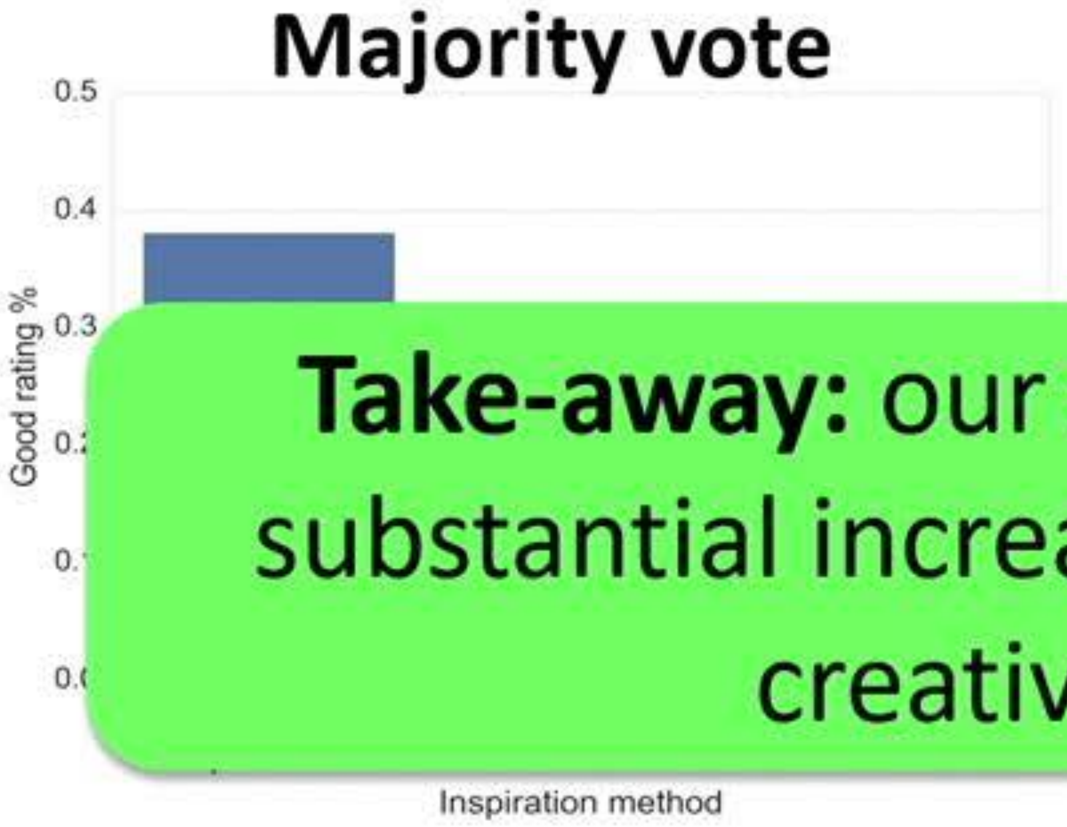


2 out of 5 raters

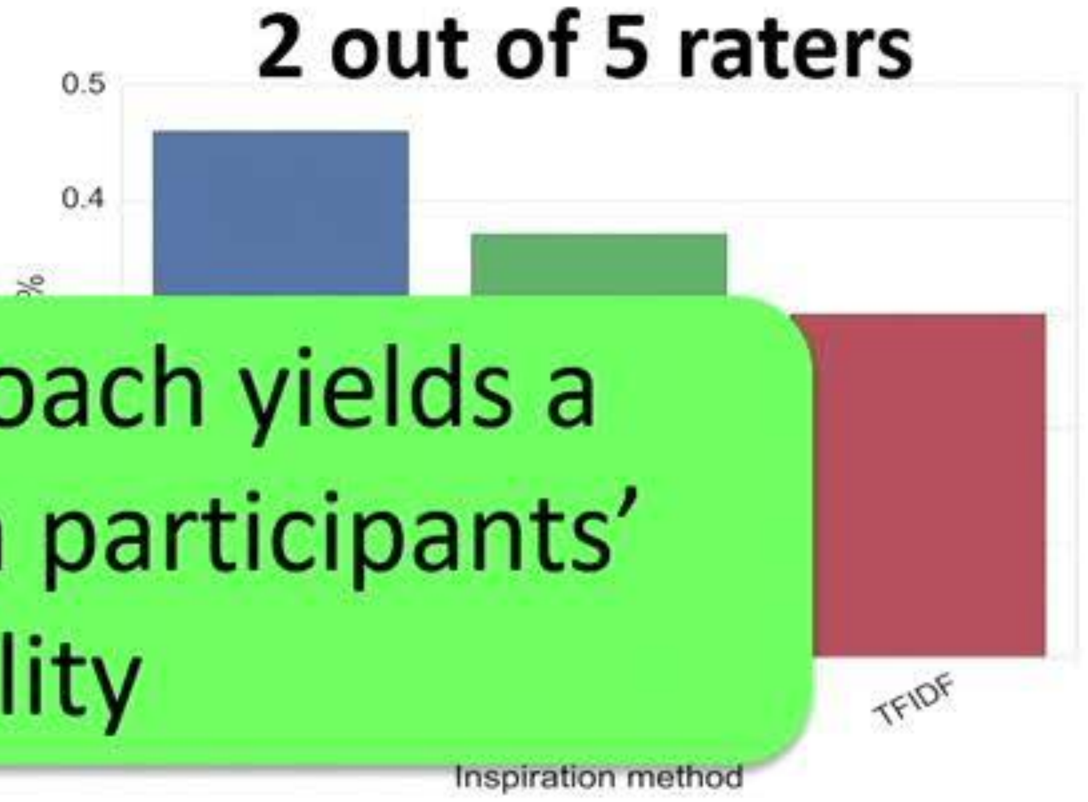


Results

- Both in terms of proportions and absolute number, our approach generates a considerably large relative positive effect



Majority vote

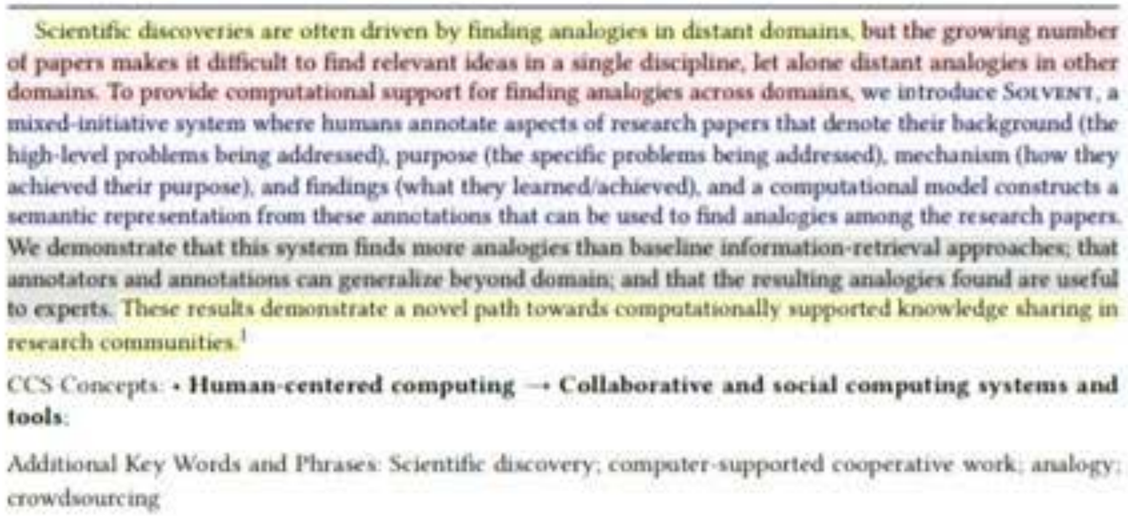


2 out of 5 raters

Take-away: our approach yields a substantial increase in participants' creative ability

Application: Boosting science

- Find analogies between scientific papers
- Extend annotation scheme: background, findings
- Help biomechanical engineering research group
  - Find inspirations, novel applications
- Most relevant matches found by **standard search**:
  - Near-direct replication of lab's work
- Most relevant matches with **our approach**:
  - Our detected analogies judged as relevant *and* new
  - Out-of-domain, alternative approaches never seen by the lab



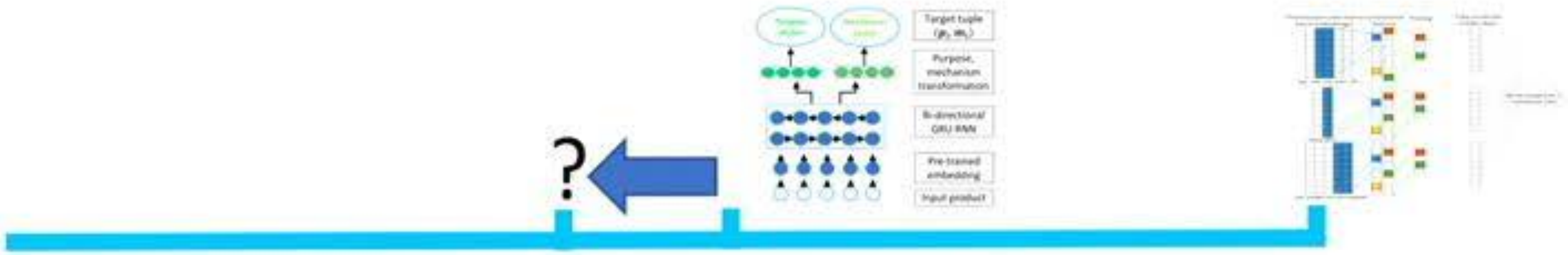
Scientific discoveries are often driven by finding analogies in distant domains, but the growing number of papers makes it difficult to find relevant ideas in a single discipline, let alone distant analogies in other domains. To provide computational support for finding analogies across domains, we introduce SOLVENT, a mixed-initiative system where humans annotate aspects of research papers that denote their background (the high-level problems being addressed), purpose (the specific problems being addressed), mechanism (how they achieved their purpose), and findings (what they learned/achieved), and a computational model constructs a semantic representation from these annotations that can be used to find analogies among the research papers. We demonstrate that this system finds more analogies than baseline information-retrieval approaches; that annotators and annotations can generalize beyond domain; and that the resulting analogies found are useful to experts. These results demonstrate a novel path towards computationally supported knowledge sharing in research communities.[1]

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**;

Additional Key Words and Phrases: Scientific discovery; computer-supported cooperative work; analogy; crowdsourcing

** SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers, CSCW '18*

A stronger structural representation

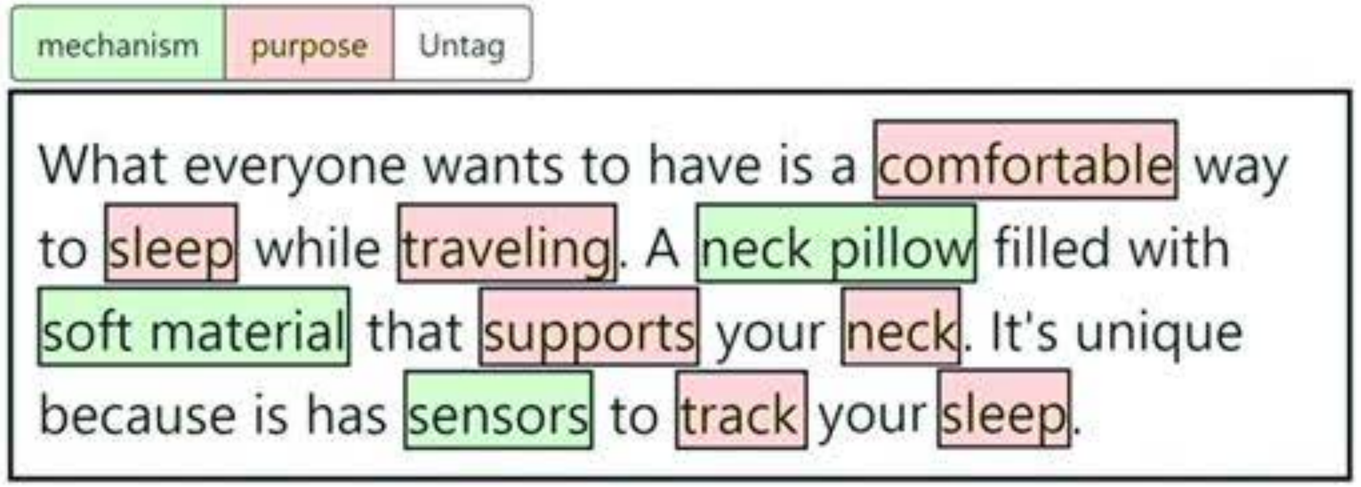


Soft vectors

Raw text

```
ENABLE( ENABLE( ENABLE(
            SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
            INSERT_INTO(person, plastic_bag, bottle)),
            CAUSE(CAUSE(
                        INSERT_INTO(person,air, plastic_bag),
                        EXPAND(air, plastic_bag))
            ENCLOSE(plastic_bag, cork)) ),
        CAUSE(
            PULL(person, plastic_bag),
            ¬IN(cork, bottle)))
```

From soft vectors to fine-grained functional models



What everyone wants to have is a comfortable way to sleep while traveling. A neck pillow filled with soft material that supports your neck. It's unique because is has sensors to track your sleep.

- Objective: Extracting spans of text corresponding purposes, mechanisms
- Enables new applications:
  - Fine-grained faceted search
  - Ontology
  - Interpretability (for search users)

hyadata lab

*** How Things Work: Large-scale Functional Modeling of Ideas (under review)***

Patent data – a (messy) treasure trove of innovation

- Important, large-scale source of engineering innovation
- Common guideline for patent examiners/writers:
  - *“The subject matter of the invention should be described in one or more* clear, concise sentences *or paragraphs…”*
  - *One patent* ***sentence:***

> *An absorptive article containing a surface material comprising a combined non-woven fabric comprising at least two layers of a long fiber non-woven fabric and a short fiber non-woven fabric joined together and an absorbing body for retaining a body fluid is disclosed in which the short fiber non-woven fabric is composed of hot-melt-adhesive composite short fibers having at least two kinds of thermoplastic resin components of a high melting point component and low melting point component, and the hot-melt-adhesive composite short fibers are hot-melt-adhered together, the crossing angle of the short fibers at least preferably at least 45%, preferably at least 50% of the total contact points in the short fiber non-woven fabric are occupied by an angle of 60 degree to 90 degree in the analysis of the distribution of the crossing angle at the contact points of the fibers.*

Patent 2001100012 (circa 2001)

- “circular transportation facilitation device”…

Patent 2001100012 (circa 2001)

- “circular transportation facilitation device”…

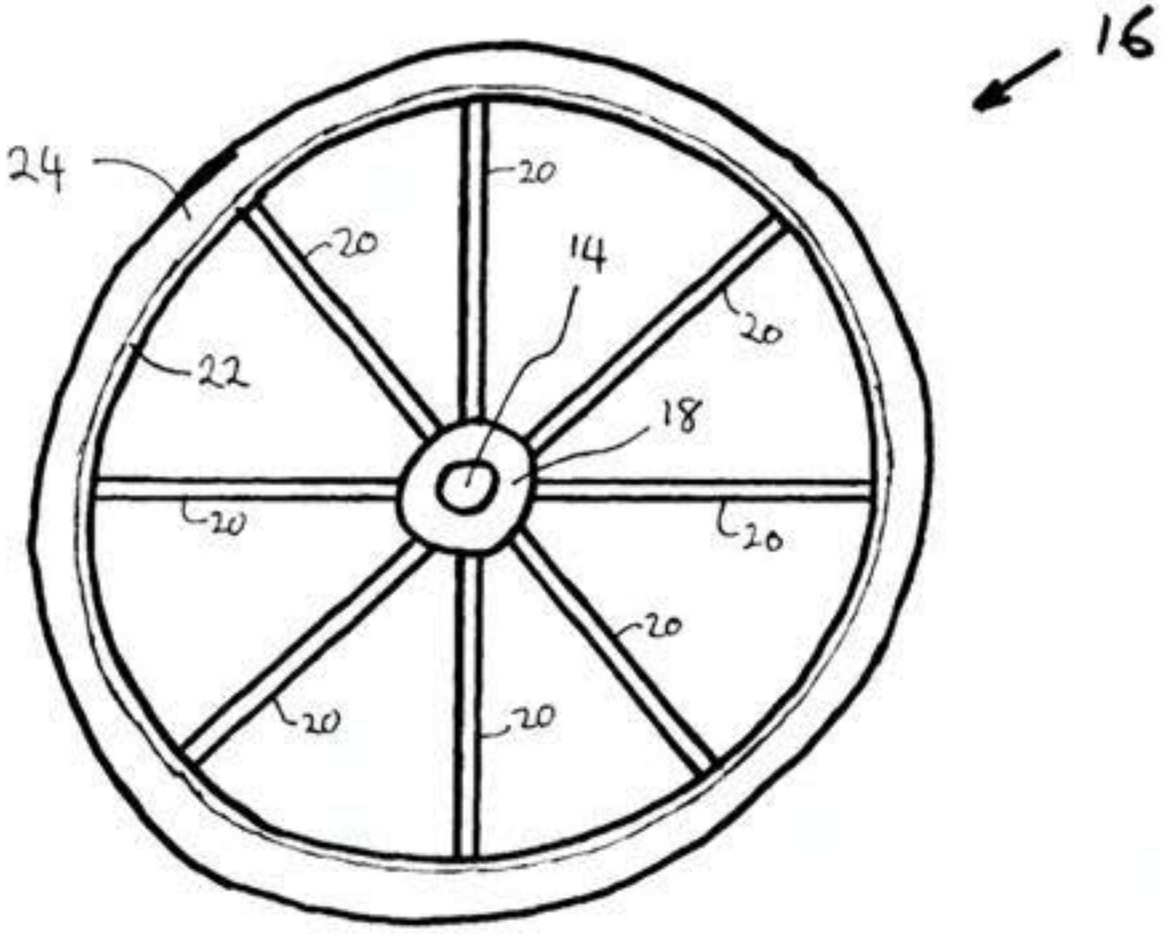


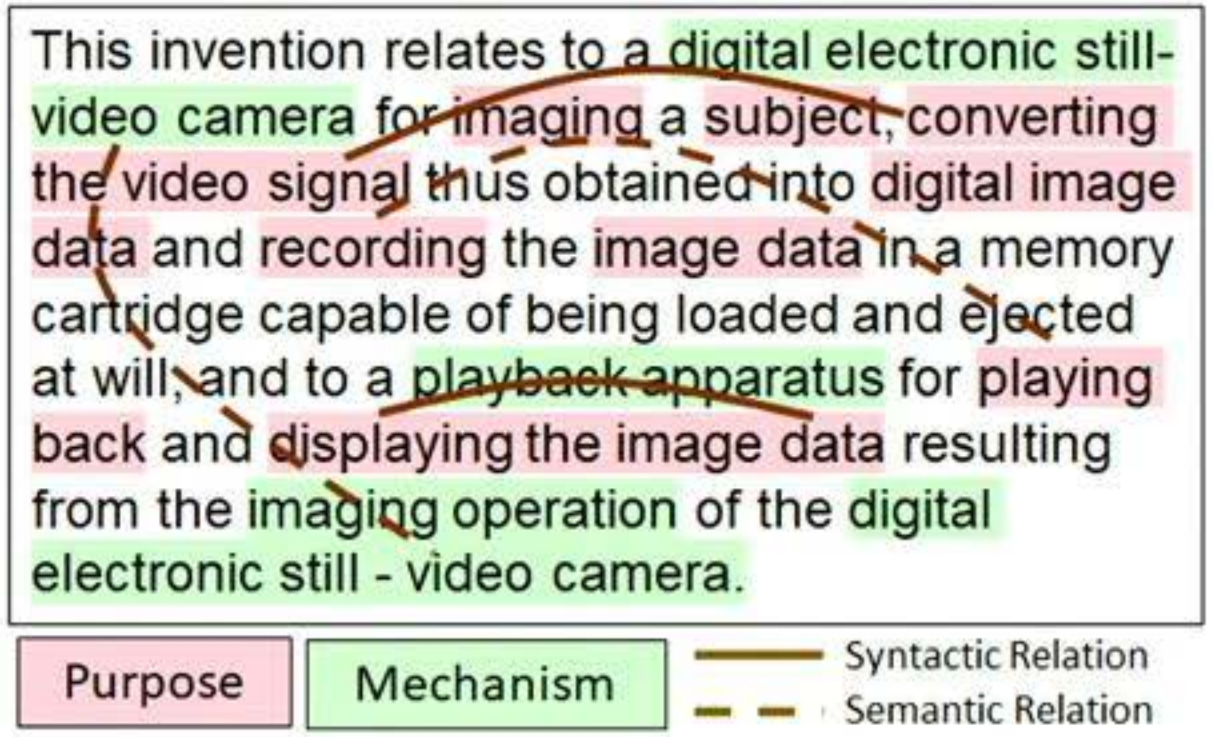
FIGURE 2

Annotating purposes and mechanisms in patents

- Hard task for workers
  - Long, obfuscated sentences
  - Complex technical jargon
  - Often skirt around the purpose of the invention
  - Median completion time: 1 minute 40 seconds
- Sample from domains easier for crowds to understand (relatively)
  - Vehicles, games, television, music, exercise, surgery, robotics…
- Noisy annotations
  - Partial tagging – workers skip sentences

Building a graph representation

- **Syntactic relations**
  - Capture purpose/mechanism syntactic patterns*
  - Dependency parse tree
- **Semantic relations**
  - Information propagation across long spans
  - Patents: long sentences with repetitions
  - Embedding similarity (tuned threshold)

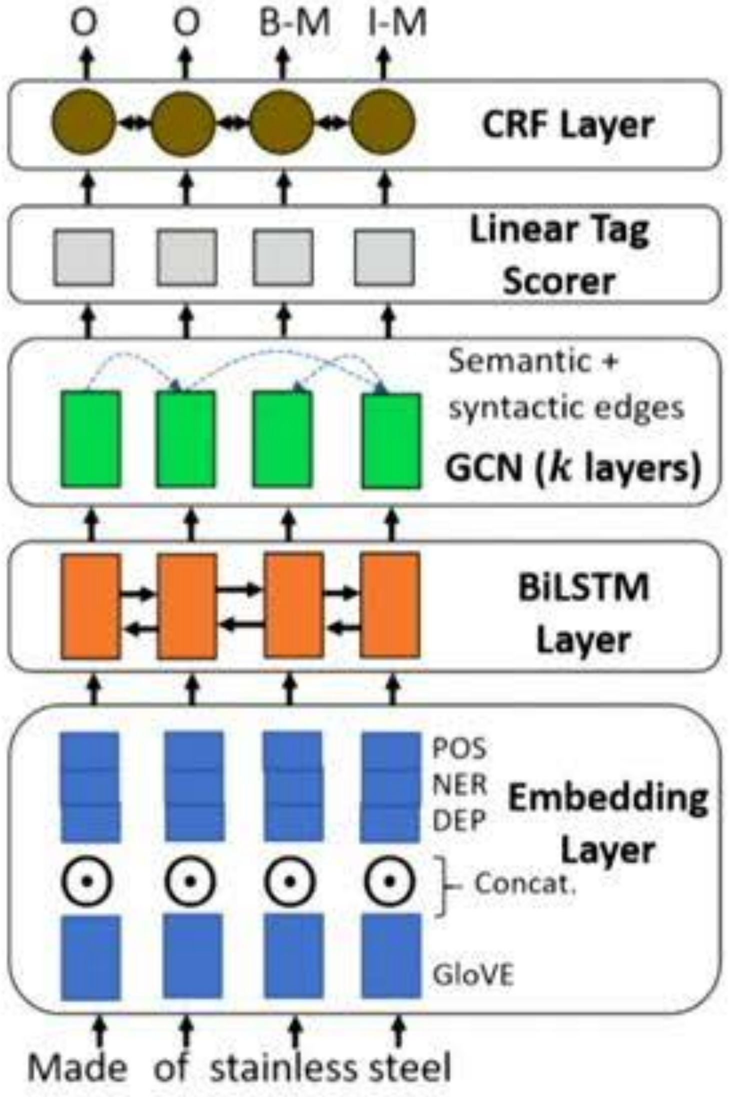


This invention relates to a digital electronic still-video camera for imaging a subject, converting the video signal thus obtained into digital image data and recording the image data in a memory cartridge capable of being loaded and ejected at will, and to a playback apparatus for playing back and displaying the image data resulting from the imaging operation of the digital electronic still - video camera.

* *K. Fu, Discovering structure in design databases through functional and surface based mapping,* Journal of mechanical Design '13

Model - contextualized graph convolution net

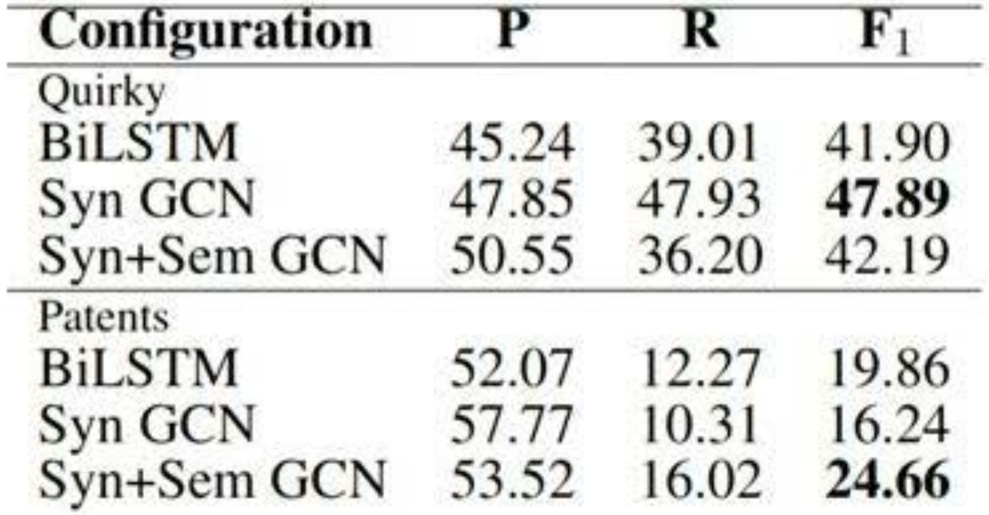
- Multi-channel input embeddings* (GloVE, POS, NER, DEP)
- BiLSTM to capture sequential context
- Relational GCN over graph
  - Add edge-wise gating for attention
- CRF layer to capture tag dependencies
- Baseline:
  - BiLSTM-CRF enriched with same multi-channel input
  - Ablation study for different edge types in graph



* *Zhang, Qi, Manning (2018), Graph Convolution over Pruned Dependency Trees Improves Relation Extraction*

Evaluation: Accuracy

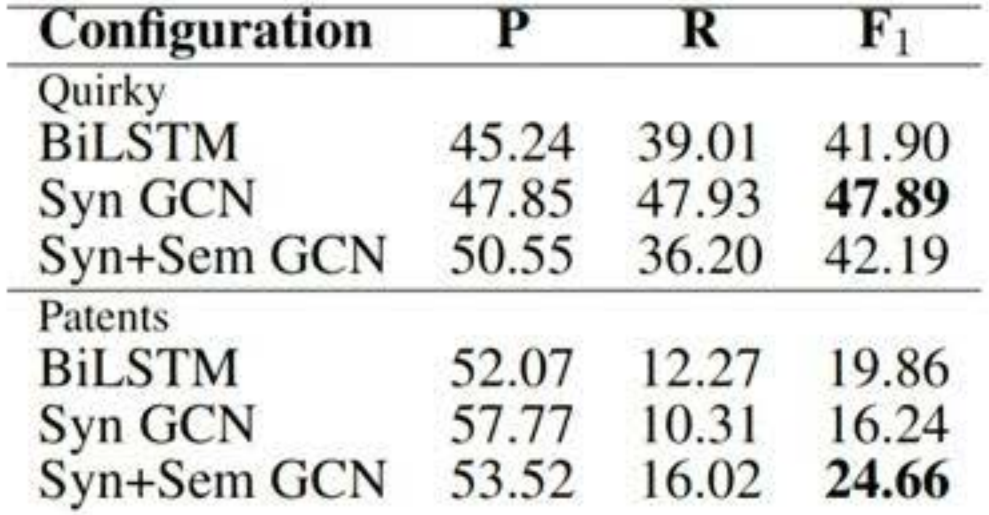
- Quirky(~23K sentences), patents (~10K)
- Graph model significantly improves results
  - Longer-range semantic edges helps in patents
  - Overall accuracy is low – noisy training annotations!
  - Model predictions often better than workers
  - Gold-standard test set



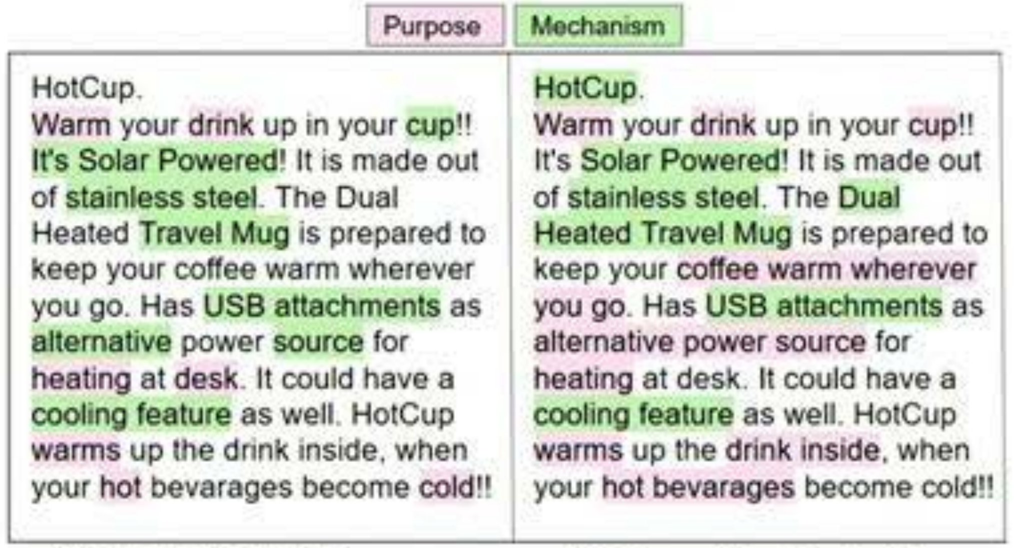
Configuration	P	R	F_1
Quirky			
BiLSTM	45.24	39.01	41.90
Syn GCN	47.85	47.93	**47.89**
Syn+Sem GCN	50.55	36.20	42.19
Patents			
BiLSTM	52.07	12.27	19.86
Syn GCN	57.77	10.31	16.24
Syn+Sem GCN	53.52	16.02	**24.66**

Evaluation: Accuracy

- Quirky(~23K sentences), patents (~10K)
- Graph model significantly improves results
  - Longer-range semantic edges helps in patents
  - Overall accuracy is low – noisy training annotations!
  - Model predictions often better than workers
  - Gold-standard test set



Configuration	P	R	F_1
Quirky			
BiLSTM	45.24	39.01	41.90
Syn GCN	47.85	47.93	**47.89**
Syn+Sem GCN	50.55	36.20	42.19
Patents			
BiLSTM	52.07	12.27	19.86
Syn GCN	57.77	10.31	16.24
Syn+Sem GCN	53.52	16.02	**24.66**


Purpose
Mechanism

HotCup. Warm your drink up in your cup!! It's Solar Powered! It is made out of stainless steel. The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go. Has USB attachments as alternative power source for heating at desk. It could have a cooling feature as well. HotCup warms up the drink inside, when your hot bevarages become cold!!

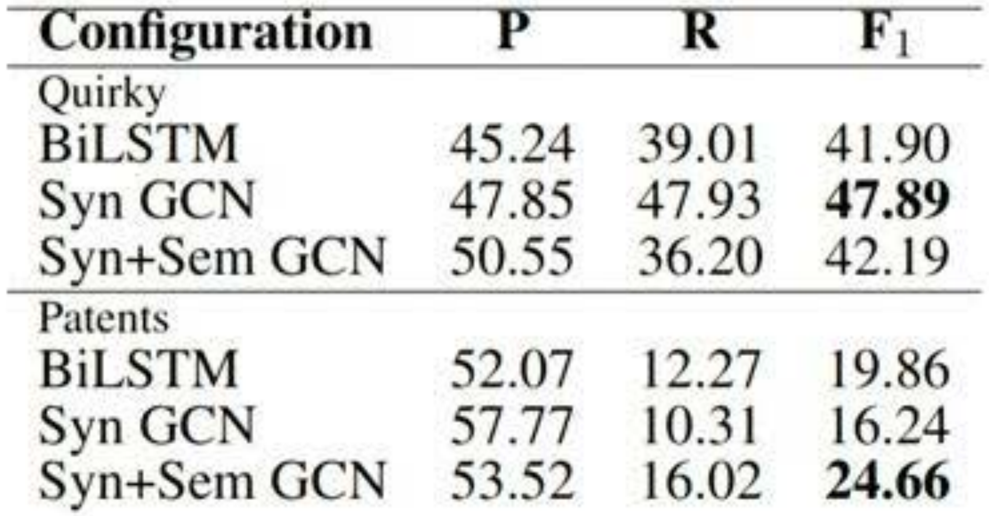
HotCup. Warm your drink up in your cup!! It's Solar Powered! It is made out of stainless steel. The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go. Has USB attachments as alternative power source for heating at desk. It could have a cooling feature as well. HotCup warms up the drink inside, when your hot bevarages become cold!!

(a) Mechanical Turk Annotation

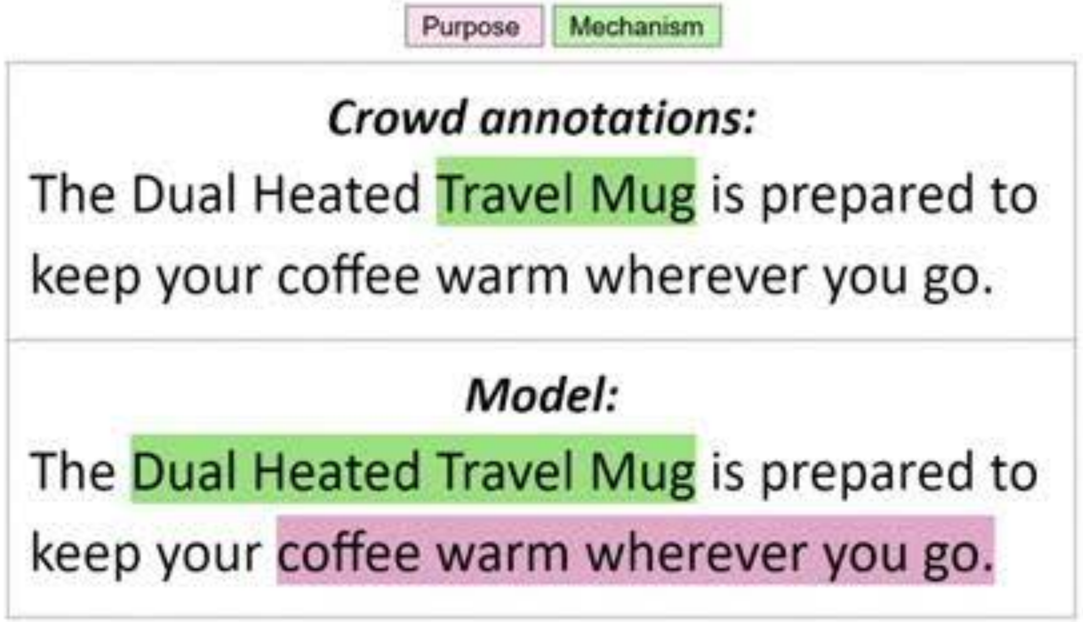
(b) Semantic + Syntactic relations model

Evaluation: Accuracy

- Quirky(~23K sentences), patents (~10K)
- Graph model significantly improves results
  - Longer-range semantic edges helps in patents
  - Overall accuracy is low – noisy training annotations!
  - Model predictions often better than workers
  - Gold-standard test set



Configuration	P	R	F_1
Quirky			
BiLSTM	45.24	39.01	41.90
Syn GCN	47.85	47.93	**47.89**
Syn+Sem GCN	50.55	36.20	42.19
Patents			
BiLSTM	52.07	12.27	19.86
Syn GCN	57.77	10.31	16.24
Syn+Sem GCN	53.52	16.02	**24.66**


Purpose
Mechanism

Crowd annotations:

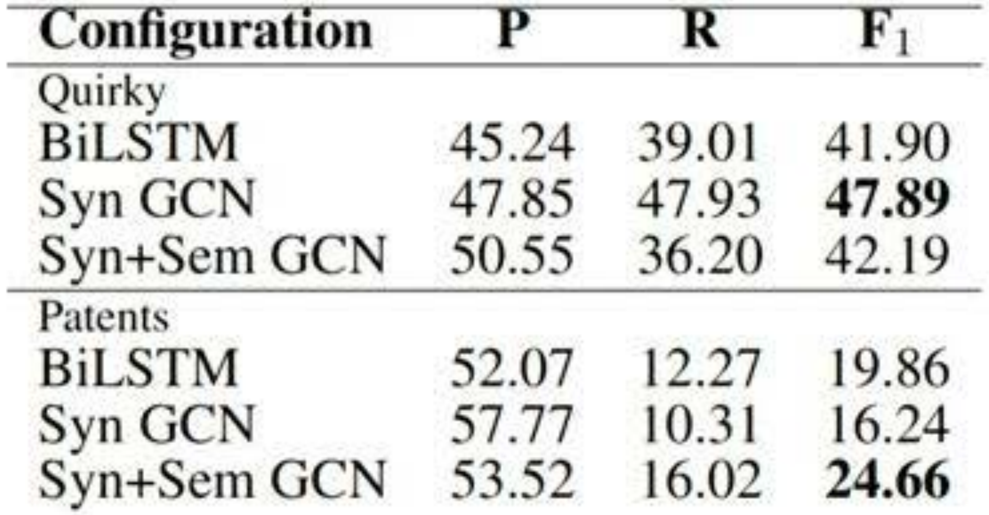
The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go.

Model:

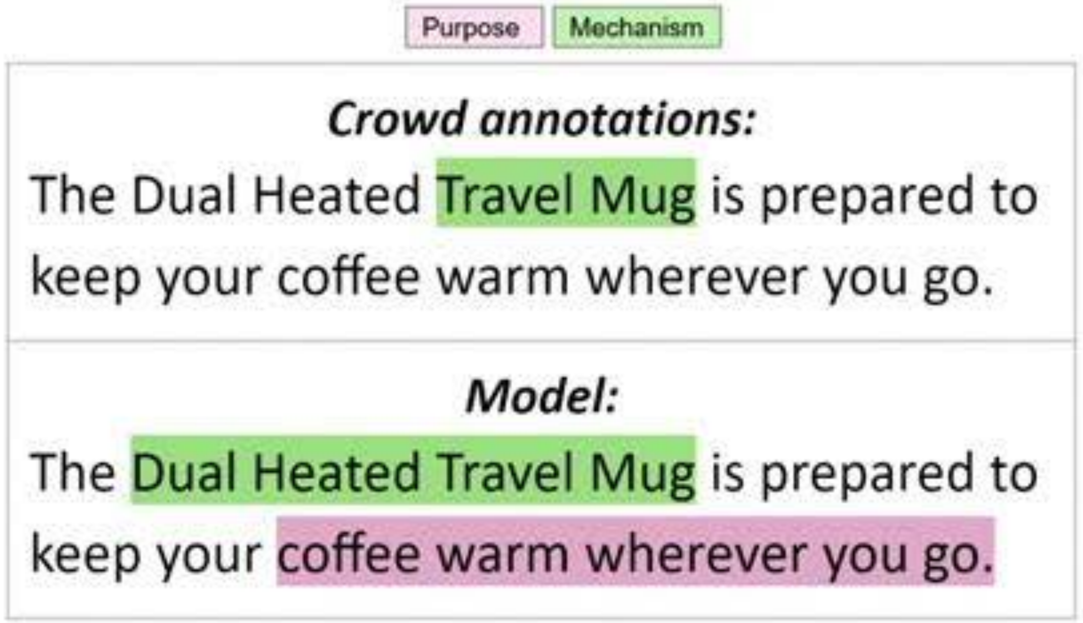
The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go.

Evaluation: Accuracy

- Quirky(~23K sentences), patents (~10K)
- Graph model significantly improves results
  - Longer-range semantic edges helps in patents
  - Overall accuracy is low – noisy training annotations!
  - Model predictions often better than workers
  - Gold-standard test set
- Adding self-training boosts results
  - Many training sentences (erroneously) un-annotated
  - New patents **F1 32.5** (up from 24.6, recall up from 16 to 23); Quirky **50.5** (47.9)



Configuration	P	R	F_1
Quirky			
BiLSTM	45.24	39.01	41.90
Syn GCN	47.85	47.93	**47.89**
Syn+Sem GCN	50.55	36.20	42.19
Patents			
BiLSTM	52.07	12.27	19.86
Syn GCN	57.77	10.31	16.24
Syn+Sem GCN	53.52	16.02	**24.66**



Purpose Mechanism

Crowd annotations:

The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go.

Model:

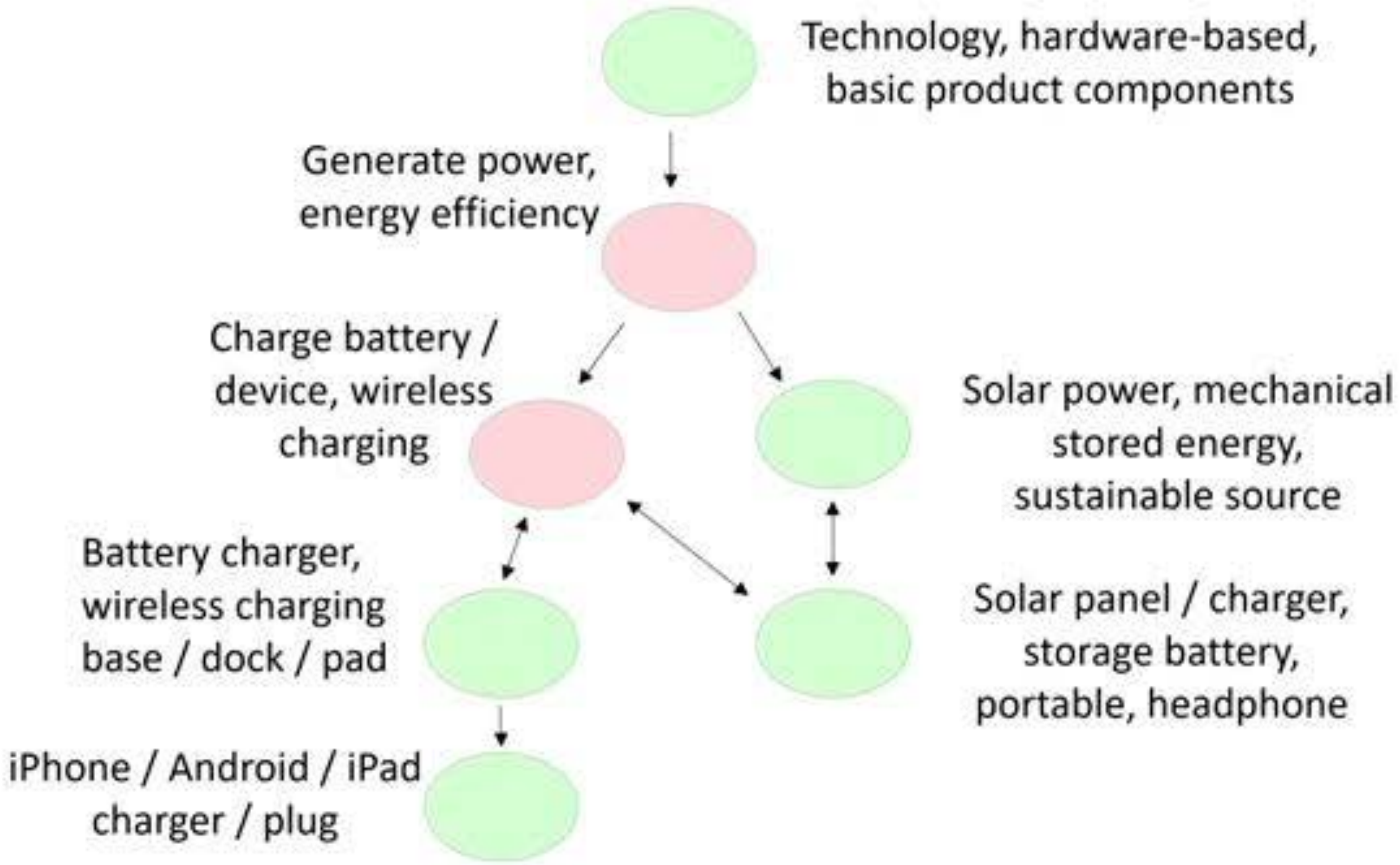
The Dual Heated Travel Mug is prepared to keep your coffee warm wherever you go.

The advantage of purpose/mechanism spans

- Unlike soft aggregate vectors – finer granularity, interpretability
- Applications:
    - Purpose-mechanism ontology
    - Expressive search

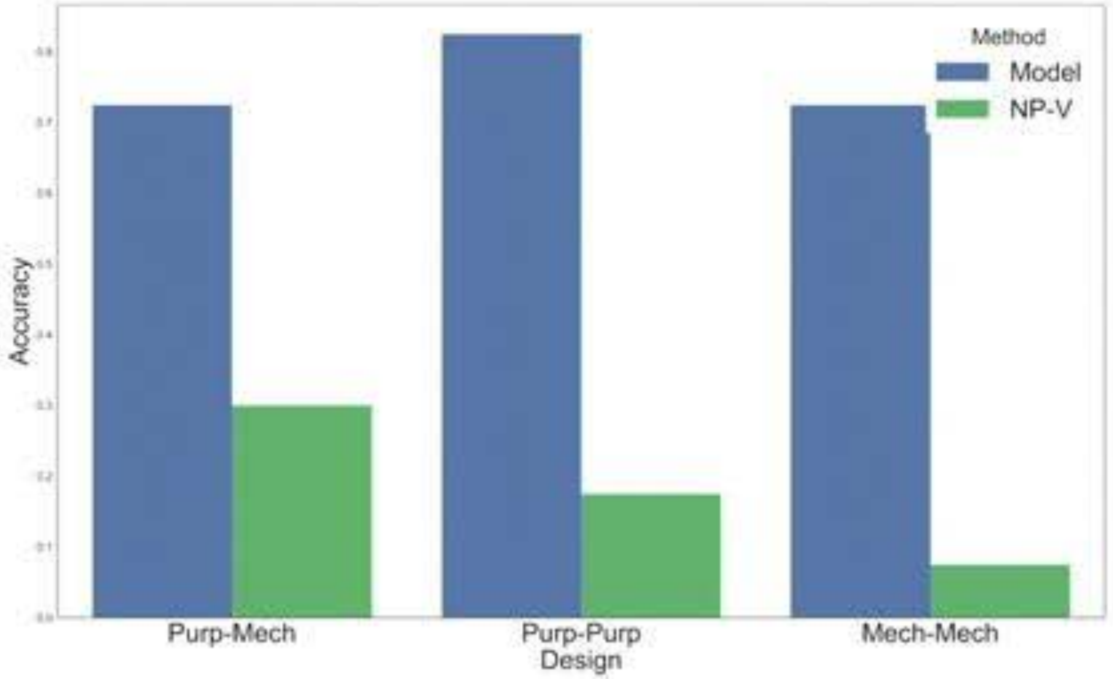
Application: Commonsense functional ontology

- Mapping the landscape of ideas with a purpose-mechanism hierarchy
- Implications for engineering – functional ontologies (handcrafted)
    - Abstraction: Allow problem-solvers to "break out" of fixation
    - Reasoning: Understanding the inter-relatedness of purposes and mechanisms


Technology, hardware-based, basic product components
Generate power, energy efficiency
Charge battery / device, wireless charging
Solar power, mechanical stored energy, sustainable source
Battery charger, wireless charging base / dock / pad
Solar panel / charger, storage battery, portable, headphone
iPhone / Android / iPad charger / plug

Ontology construction

- Discrete representation of *concepts*
  - Cluster purpose/mechanism chunks
    - Example cluster: *Solar power, stored energy, sustainable energy source*
- Hierarchical *relations*
  - Rule-mining (*Antecedent => Consequent)*
  - "protect head" => "safety", *"charge" => "charger"*
- Baseline: Using clusters of based on POS *
- Large improvement over baseline
  - Good performance in absolute terms
  - Substantial judge agreement (47%)



* *K. Fu, Discovering structure in design databases through functional and surface based mapping,* Journal of mechanical Design '13

Application: Expressive search engine to boost innovation

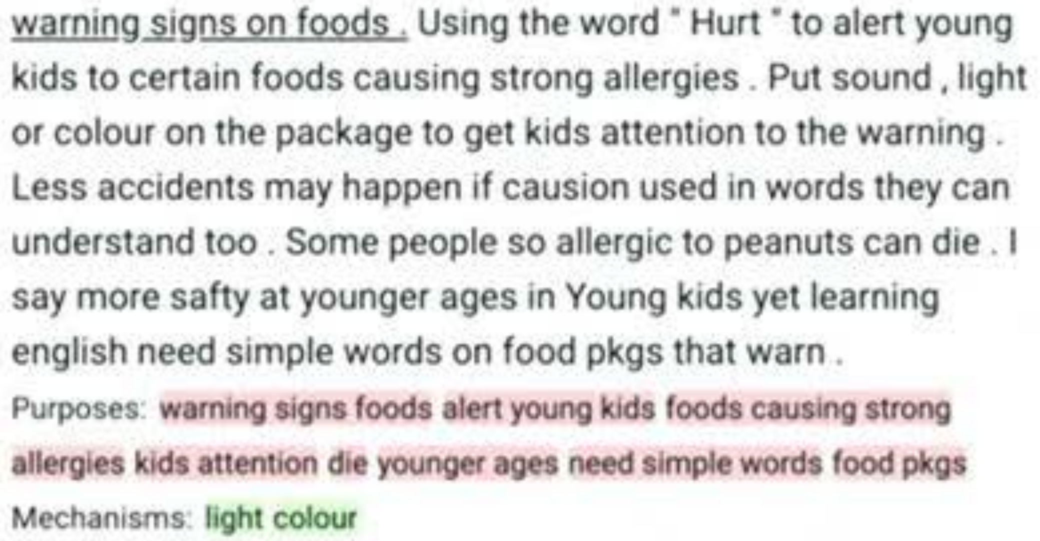
- Example: lightbulb manufacturer, wants to expand to other markets
  - Find products using light, but where light is **not** the main purpose

Application: Expressive search engine to boost innovation

- Example: lightbulb manufacturer, wants to expand to other markets
  - Find products using light, but where light is **not** the main purpose

Mechanisms	Light
Must not include	-

Purposes	**-**
Must not include	Light



warning signs on foods . Using the word " Hurt " to alert young kids to certain foods causing strong allergies . Put sound , light or colour on the package to get kids attention to the warning . Less accidents may happen if causion used in words they can understand too . Some people so allergic to peanuts can die . I say more safty at younger ages in Young kids yet learning english need simple words on food pkgs that warn .

Purposes: warning signs foods alert young kids foods causing strong allergies kids attention die younger ages need simple words food pkgs

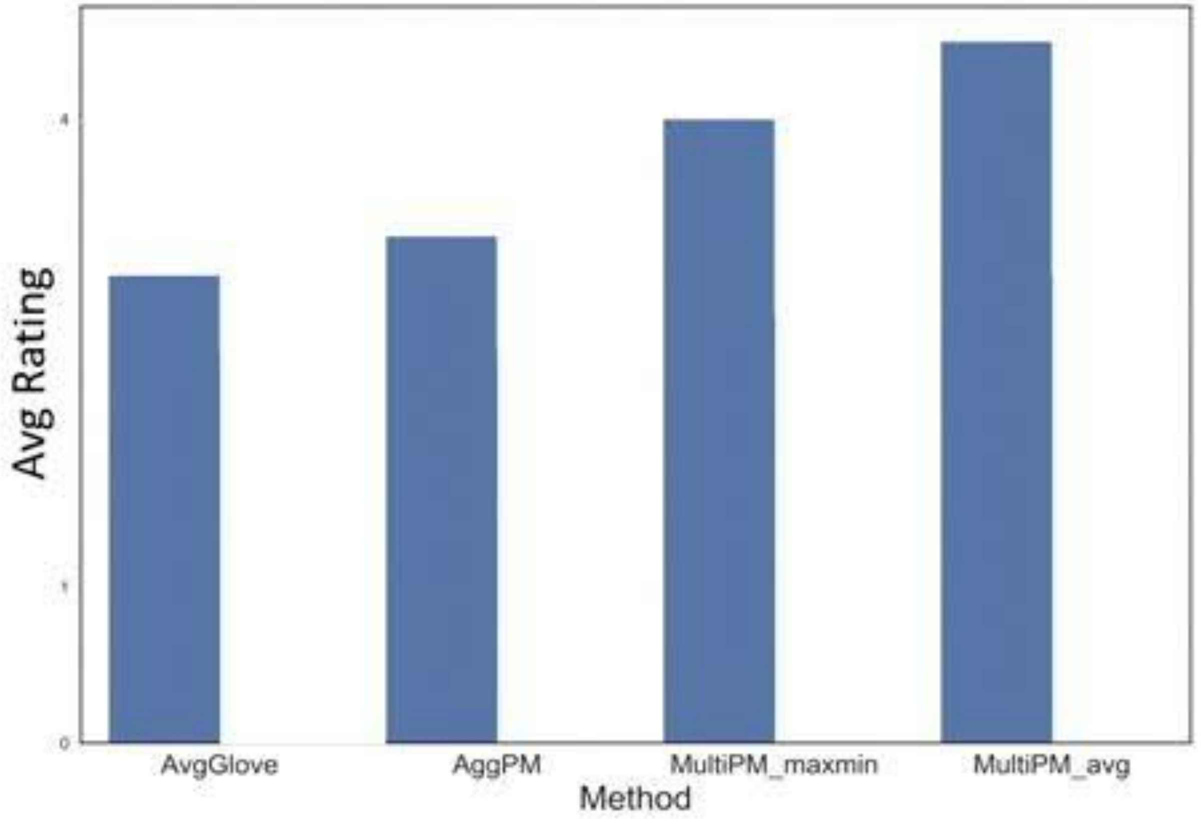
Mechanisms: light colour

- Sets of purposes (*P*), mechanisms (*M*), query (*Q*) terms
- Distance metric over sets

$$\begin{aligned}\text{argmin}_{\,i}\ & d_p(\{\mathbf{q}_{\text{“locate dog”}}\}, \mathcal{P}_{\tilde{i}}) \\ s.t.\ & d_m(\{\mathbf{q}_{\text{“GPS”}}\}, \mathcal{M}_{\tilde{i}}) \geq \text{threshold} \\ & d_m(\{\mathbf{q}_{\text{“RFID”}}\}, \mathcal{M}_{\tilde{i}}) \leq \text{threshold}\end{aligned}$$

Search evaluation

- Four search scenarios
  - Example: Use light for the purpose of cleaning
- Compare against standard search, our own previous work
- Retrieval (distance) metrics based on purpose/mechanism chunks
  - Average over terms, MinMax (query matches small subset of chunks)
- Finer-grained purposes/mechanisms lead to better search expressivity



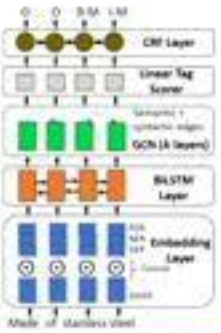
Summary: Analogies

```
ENABLE( ENABLE( ENABLE(
            SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
            INSERT_INTO(person, plastic_bag, bottle)),
            CAUSE(CAUSE(
                        INSERT_INTO(person,air, plastic_bag),
                        EXPAND(air, plastic_bag))
            ENCLOSE(plastic_bag, cork)) ),
        CAUSE(
            PULL(person, plastic_bag),
            ¬IN(cork, bottle)))
```

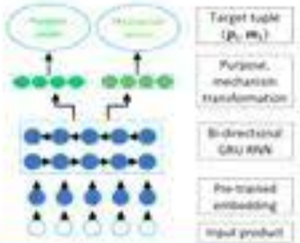
Raw text

Summary: Analogies

```
ENABLE( ENABLE( ENABLE(
            SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
            INSERT_INTO(person, plastic_bag, bottle)),
            CAUSE(CAUSE(
                    INSERT_INTO(person,air, plastic_bag),
                    EXPAND(air, plastic_bag))
            ENCLOSE(plastic_bag, cork)) ),
        CAUSE(
            PULL(person, plastic_bag),
            ¬IN(cork, bottle)))
```



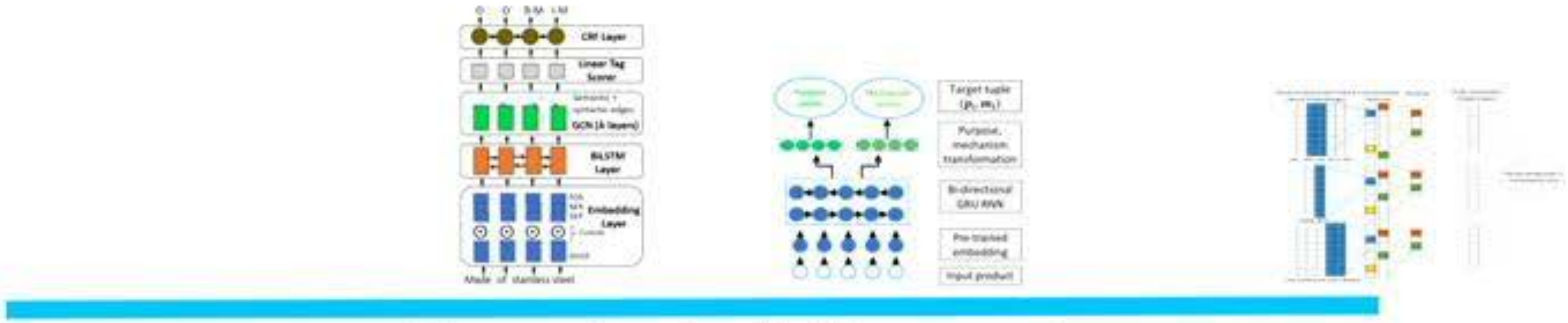
Fine-grained


Target tuple (p_i, m_i)
Purpose, mechanism transformation
Bi-directional GRU RNN
Pre-trained embedding
Input product

Soft aggregate vectors

Raw text

Summary: Analogies



```
ENABLE( ENABLE( ENABLE(
            SMALLER_THAN(UNINFLATED(plastic_bag), bottle_opening),
            INSERT_INTO(person, plastic_bag, bottle)),
            CAUSE(CAUSE(
                    INSERT_INTO(person,air, plastic_bag),
                    EXPAND(air, plastic_bag))
            ENCLOSE(plastic_bag, cork)) ),
        CAUSE(
            PULL(person, plastic_bag),
            ¬IN(cork, bottle)))
```

Fine-grained

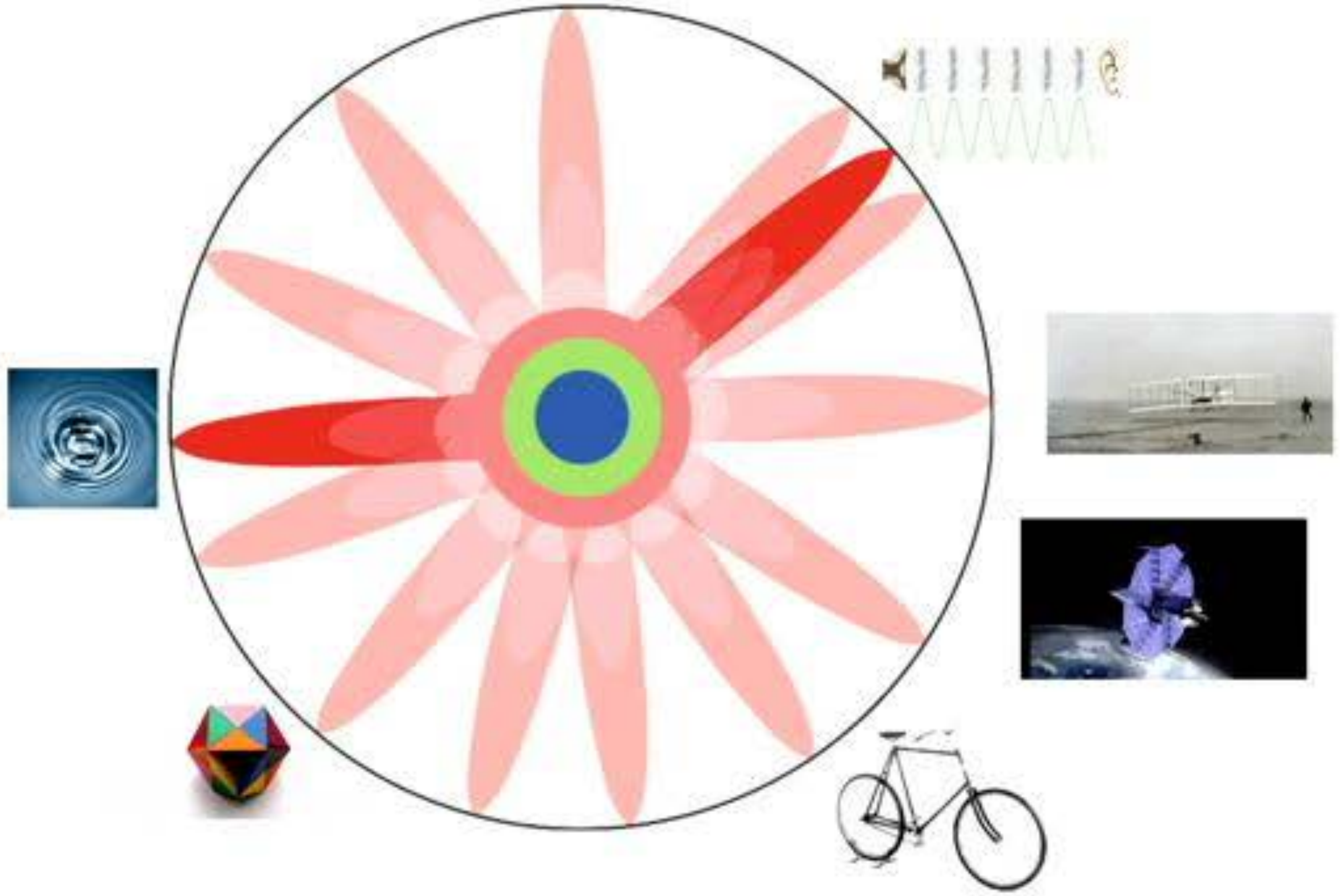
Soft aggregate vectors

Raw text

- Applications:
  - Ideation – boosting creativity with analogies
  - Boosting science
  - Purpose/mechanism commonsense functional ontology
  - Expressive search for inspiration queries
- With Joel Chan, Ronen Tamari, Daniel Hershcovich, Hyeonsu Kang, Niki Kittur, Dafna Shahaf

In this talk

- Boosting creativity with analogies
- **Identifying novel ideas with weak supervision**



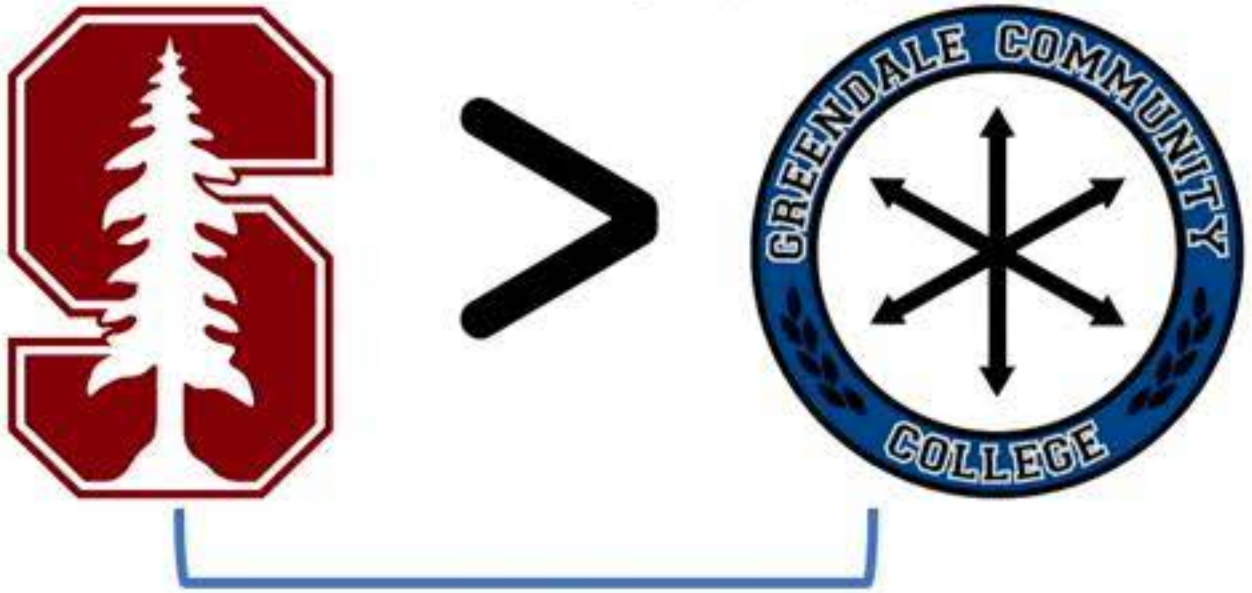
Identifying novel patents with weak supervision
(work in progress)

- Problem: Labels are extremely hard to obtain
    - Citations are a (very) weak proxy… *

**R. Patton (2016), Measuring Scientific Impact Beyond Citation Counts*

Identifying novel patents with weak supervision
(work in progress)

- Problem: Labels are extremely hard to obtain
  - Citations are a (very) weak proxy… *
- Weak supervision over aggregate **patent portfolio** information:
  - Average proportion of novel patents < %1
  - Some organizations are more innovative than others!
  - Higher average proportion of novel patents in their portfolio

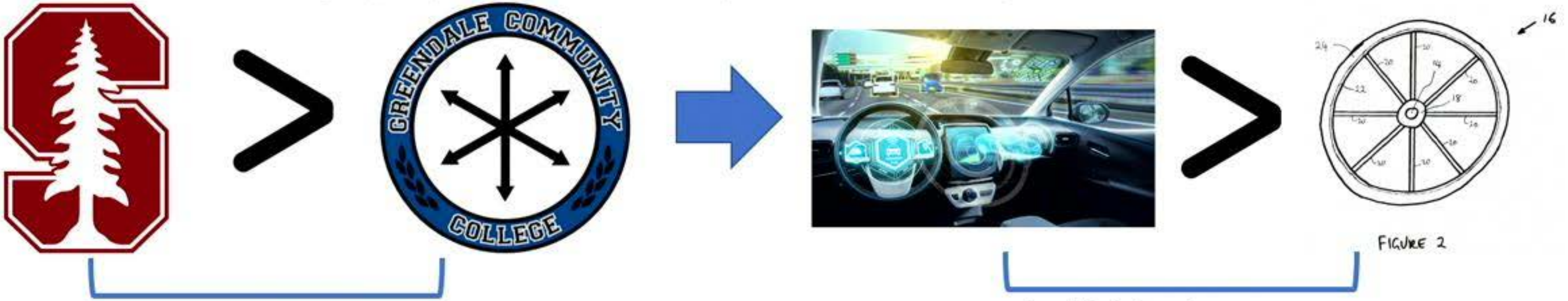


Across patent portfolios

*R. Patton (2016), *Measuring Scientific Impact Beyond Citation Counts*

Identifying novel patents with weak supervision
(work in progress)

- Problem: Labels are extremely hard to obtain
  - Citations are a (very) weak proxy… *
- Weak supervision over aggregate **patent portfolio** information:
  - Average proportion of novel patents < %1
  - Some organizations are more innovative than others!
  - Higher average proportion of novel patents in their portfolio



Across patent portfolios

Individual patents

*R. Patton (2016), *Measuring Scientific Impact Beyond Citation Counts*

Obtaining labels is often hard

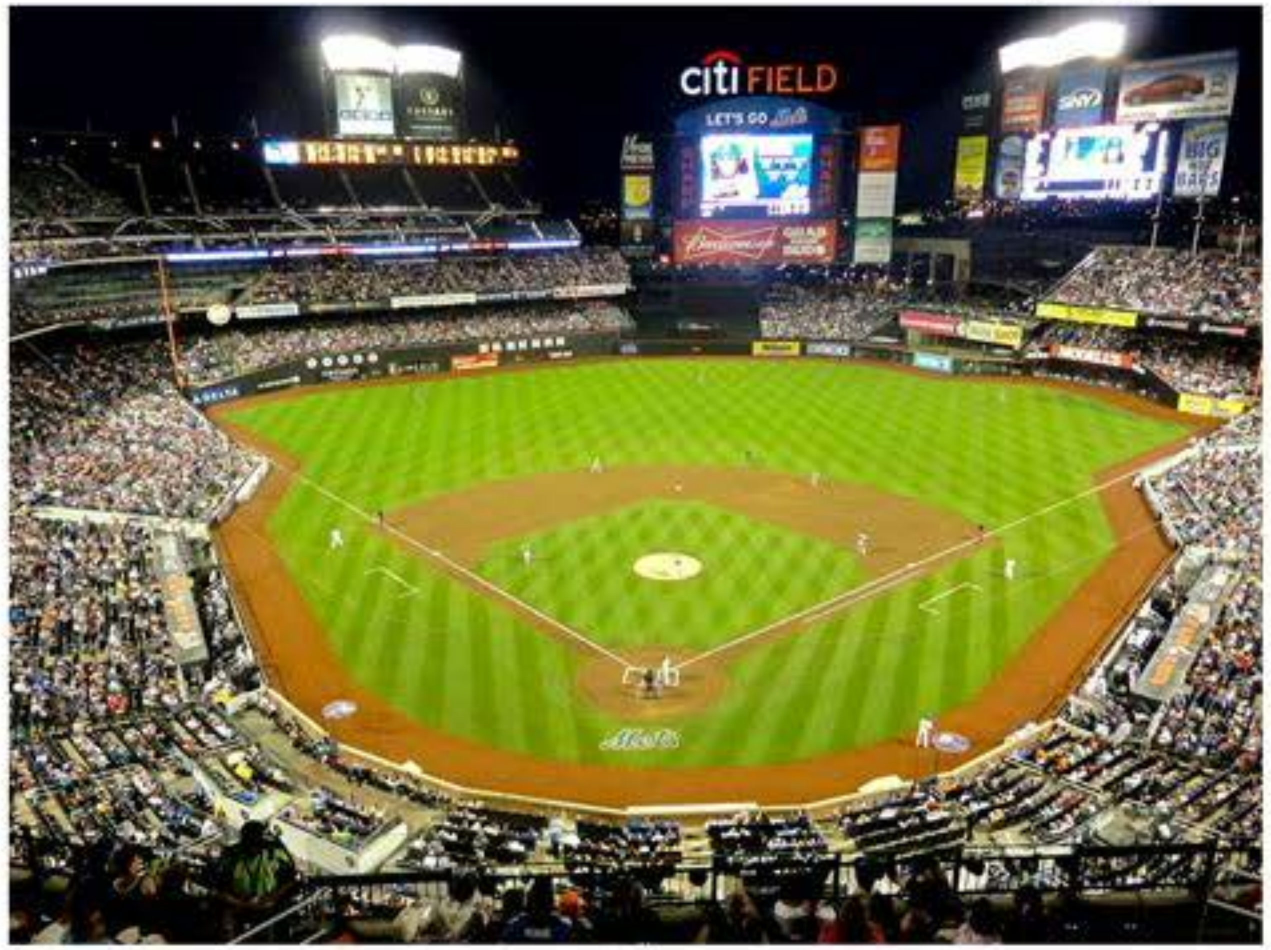
- No access to ground-truth information
  - Examples: Who did a user vote for? User health conditions?
- **Privacy** issues
- Crowdsourcing: Not a cure-all
  - Workers are not domain experts
  - Can be a difficult, noisy, **costly** process
- But rough aggregate information is often much easier to obtain!

Ballpark Learning

Ballpark figure: An educated guess or estimation within acceptable bounds.

*"No more stalling. Give me a **ballpark figure** of our projected losses."*

https://en.wiktionary.org/wiki/ballpark_figure



Ballpark Learning – learning from rough constraints on groups

- Training instances $\mathrm{X} = \{\boldsymbol{x}_1,\ldots\boldsymbol{x}_N\}$ with **unknown** labels y_i
- Our instances are divided into K bags $\{B_1, B_2, \ldots, B_K\}$
  - Example: Patent portfolios
- Constraints on **unknown** label averages within bags $\{y_i: \boldsymbol{x}_i \in B_k\}$
  - Bag **upper/lower** bounds $l_k \leq Avg(B_k) \leq u_k$
    - *0.01* ≤ *Avg*(Stanford) ≤ 0.05
  - **Differences** (additive) $l_{k_{12}} \leq Avg(B_{k_1}) - Avg(B_{k_2}) \leq u_{k_{12}}$
    - *Avg*(Stanford) - *Avg*(Greendale) ≥ 0.03
  - **Differences** (multiplicative) $l_{k_{12}} \leq Avg(B_{k_1}) / Avg\left(B_{k_2}\right) \leq u_{k_{12}}$
- Learning goal: Predict instance-level labels from rough constraints on label averages given for each bag / pairs of bags

Problem formulation – classification *

- Discrete (binary) label space $y_i \in \{-1,1\}$
- Bi-convex optimization problem:

$$\begin{aligned}
&\underset{\mathbf{y},\mathbf{w},\xi}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{N}\sum_{i=1}^{N}\max(0, 1 - y_i\mathbf{w}^T\varphi(\mathbf{x}_i)) + \frac{C_L}{L}\sum_{j=N+1}^{N+L}\xi_j \\
&\quad s.t. -1 \leq y_i \leq 1 \quad \forall i \in 1,\ldots,N \\
&\qquad y_j\mathbf{w}^T\varphi(\mathbf{x}_j) \geq 1 - \xi_j \quad \forall j \in \{N+1,\ldots,N+L\} \\
&\qquad \xi_j \geq 0 \quad \forall j \\
&\qquad l_k \leq \hat{p}_k \leq u_k \quad \forall\{k : \mathcal{B}_k \in \mathcal{R}\} \\
&\qquad l_{k_{12}} \leq \hat{p}_{k_1} - \hat{p}_{k_2} \leq u_{k_{12}} \quad \forall\{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\},
\end{aligned}$$

where $\hat{p}_k = \frac{1}{2|\mathcal{B}_k|}\sum_{i \in B_k} y_i + \frac{1}{2}$

hyadata lab

*** Ballpark Learning: Estimating Labels from Rough Group Comparisons, ECML-PKDD ’16***

“most thought-provoking paper of the week” -- *MIT Technology Review*

Problem formulation – classification *

- Discrete (binary) label space $y_i \in \{-1,1\}$
- Bi-convex optimization problem:

$$\operatorname*{argmin}_{\mathbf{y},\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \boxed{\frac{C}{N}\sum_{i=1}^{N}\max(0, 1 - y_i\mathbf{w}^T\varphi(\mathbf{x}_i)) + \frac{C_L}{L}\sum_{j=N+1}^{N+L}\xi_j}$$

$$\begin{aligned}
s.t. -1 \le y_i \le 1 \quad & \forall i \in 1,\dots,N \\
y_j\mathbf{w}^T\varphi(\mathbf{x}_j) \ge 1 - \xi_j \quad & \forall j \in \{N+1,\dots,N+L\} \\
\xi_j \ge 0 \quad & \forall j \\
l_k \le \hat{p}_k \le u_k \quad & \forall \{k : \mathcal{B}_k \in \mathcal{R}\} \\
l_{k_{12}} \le \hat{p}_{k_1} - \hat{p}_{k_2} \le u_{k_{12}} \quad & \forall \{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\},
\end{aligned}$$

where $\hat{p}_k = \frac{1}{2|\mathcal{B}_k|}\sum_{i \in B_k} y_i + \frac{1}{2}$

Objective:
1. Hinge Loss: Matching $\mathbf{y}, \mathbf{w}$
2. Slack vars: For known y_j

hyadata lab

*** Ballpark Learning: Estimating Labels from Rough Group Comparisons, ECML-PKDD '16**

"most thought-provoking paper of the week" -- *MIT Technology Review*

Problem formulation – classification *

- Discrete (binary) label space $y_i \in \{-1,1\}$
- Bi-convex optimization problem:

$$\underset{\mathbf{y},\mathbf{w},\xi}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \boxed{\frac{C}{N}\sum_{i=1}^{N}\max(0, 1 - y_i\mathbf{w}^T\varphi(\mathbf{x}_i)) + \frac{C_L}{L}\sum_{j=N+1}^{N+L}\xi_j}$$

$$s.t. -1 \le y_i \le 1 \quad \forall i \in 1,\ldots,N$$

$$y_j\mathbf{w}^T\varphi(\mathbf{x}_j) \ge 1 - \xi_j \quad \forall j \in \{N+1,\ldots,N+L\}$$

$$\xi_j \ge 0 \quad \forall j$$

$$\boxed{\begin{array}{l} l_k \le \hat{p}_k \le u_k \quad \forall\{k : \mathcal{B}_k \in \mathcal{R}\} \\ l_{k_{12}} \le \hat{p}_{k_1} - \hat{p}_{k_2} \le u_{k_{12}} \quad \forall\{k_1 \ne k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\}, \end{array}}$$

where $\hat{p}_k = \frac{1}{2|\mathcal{B}_k|}\sum_{i\in B_k} y_i + \frac{1}{2}$

Objective:
1. Hinge Loss: Matching $\mathbf{y}, \mathbf{w}$
2. Slack vars: For known y_j

Proportion constraints
1. Bag upper/lower bounds
2. Bag difference bounds

- Objective helps find assignment to latent $\mathbf{y}$ that accurately "matches" $\mathbf{w}$, vice versa
- Linear constraints ensure a "correct" assignment to $\mathbf{y}$

hyadata lab

** **Ballpark Learning: Estimating Labels from Rough Group Comparisons, ECML-PKDD '16***

"most thought-provoking paper of the week" -- *MIT Technology Review*

Problem formulation – regression

- Continuous label space $y_i \in \mathbb{R}$

$$\underset{\mathbf{y},\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \boxed{\frac{C_N}{N}\sum_{i=1}^{N} ||y_i - \mathbf{w}^T\varphi(\mathbf{x}_i))||_2^2}$$

MSE loss

$$+ \frac{C_L}{L}\sum_{j=N+1}^{N+L} ||y_j - \mathbf{w}^T\varphi(\mathbf{x}_j))||_2^2$$

$$\boxed{\begin{aligned} s.t. \quad & l_k \le \hat{\bar{y}}_k \le u_k \quad \forall\{k : \mathcal{B}_k \in \mathcal{R}\} \\ & l_{k_{12}} \le \hat{\bar{y}}_{k_1} - \hat{\bar{y}}_{k_2} \le u_{k_{12}} \quad \forall\{k_1 \ne k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\} \end{aligned}}$$

Average constraints

where $\hat{\bar{y}}_k = \frac{\sum_{i \in \mathcal{B}_k} y_i}{|\mathcal{B}_k|}$

- **Convex** (quadratic w.r.t y) problem

Formulation as feasibility problem

- Alternative formulation: Optimize only for $\mathbf{w}$ under constraints

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C_L}{L}\sum_{j=N+1}^{N+L} ||y_j - \mathbf{w}^T\varphi(\mathbf{x}_j))||_2^2$$

$$\boxed{\begin{aligned} s.t. \quad & l_k \leq \frac{\sum_{i\in\mathcal{B}_k} \mathbf{w}^T\varphi(\mathbf{x_i})}{|\mathcal{B}_k|} \leq u_k \quad \forall\{k : \mathcal{B}_k \in \mathcal{R}\} \\ & l_{k_{12}} \leq \frac{\sum_{i\in\mathcal{B}_{k_1}} \mathbf{w}^T\varphi(\mathbf{x_i})}{|\mathcal{B}_{k_1}|} - \frac{\sum_{i\in\mathcal{B}_{k_2}} \mathbf{w}^T\varphi(\mathbf{x_i})}{|\mathcal{B}_{k_2}|} \leq u_{k_{12}} \end{aligned}}$$

$$\forall\{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\}$$

Feasibility constraints

- No optimization for latent labels – less parameters, faster
- PAC formulation
- More details, discussion in paper *

hyadata lab

* ***Ballpark Crowdsourcing: The Wisdom of Rough Group Comparisons, WSDM '18***

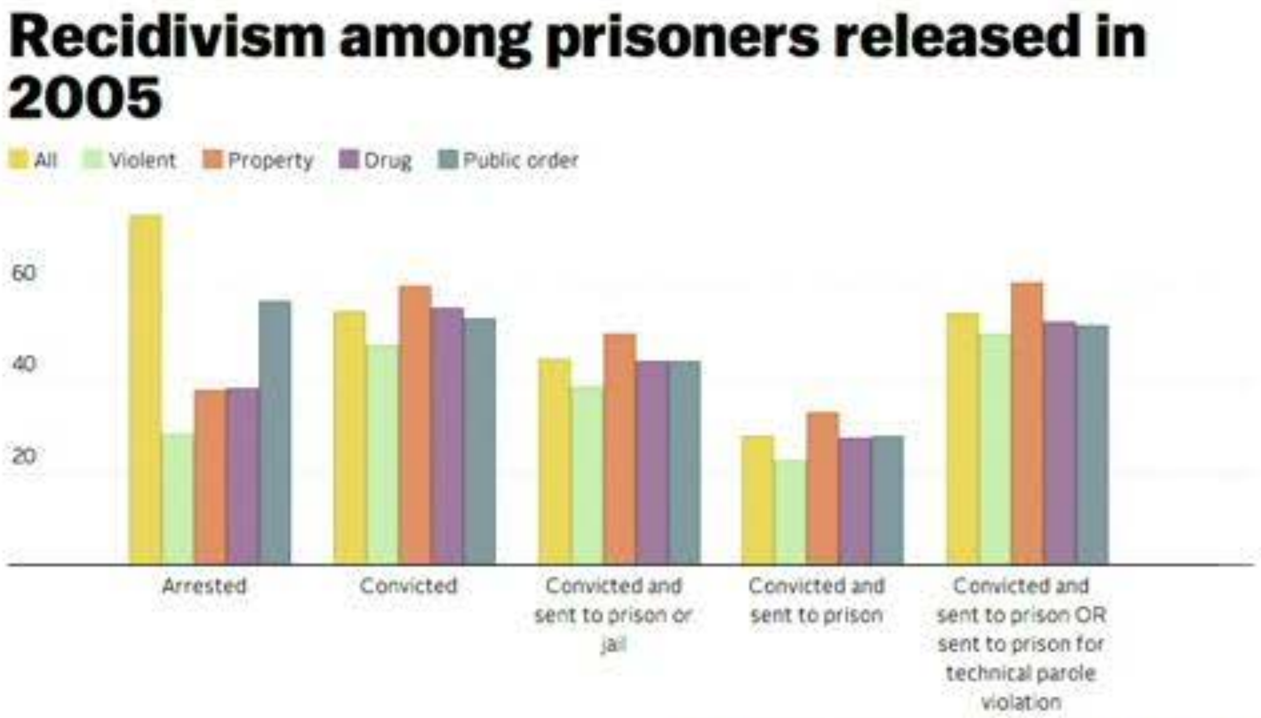
Hyperparameter optimization

- How do we find regularization **hyperparameter**?
- No labeled examples, so standard cross-validation (CV) grid search won't work

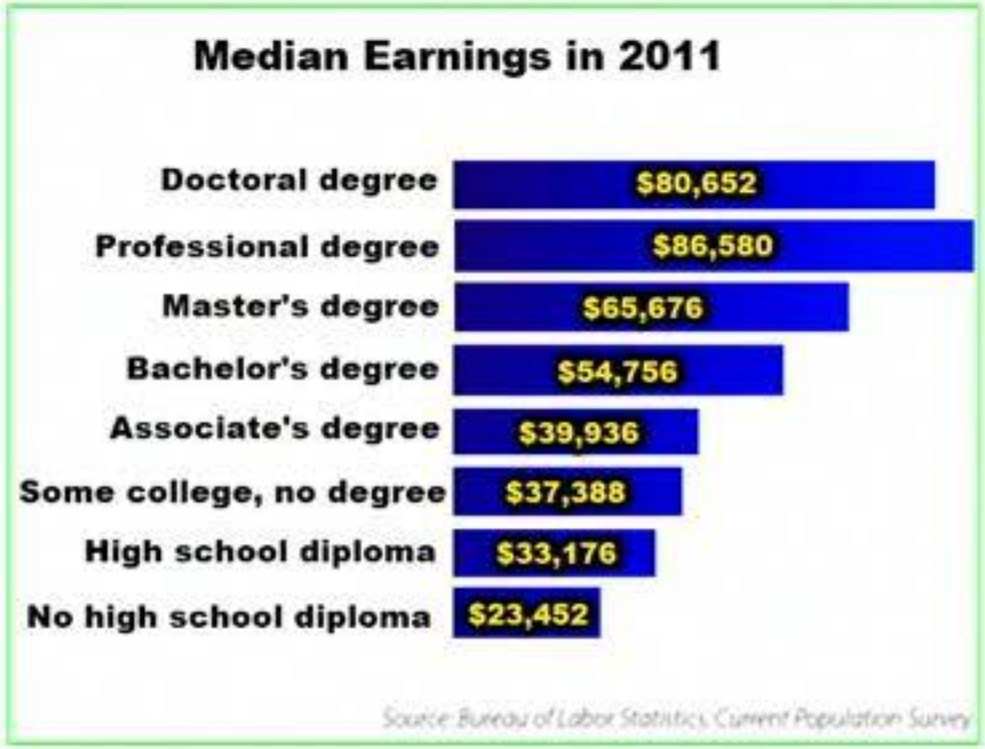
Constraint-violation CV (CVCV) grid search

- Run K-fold CV, splitting each bag B_k into training and held-out subsets
- Due to uniform sampling, assume approximately unchanged proportions
- We measure **constraint violation** on the held-out data
- Pick C with minimal average violation

Evaluation: Learning from rough group constraints



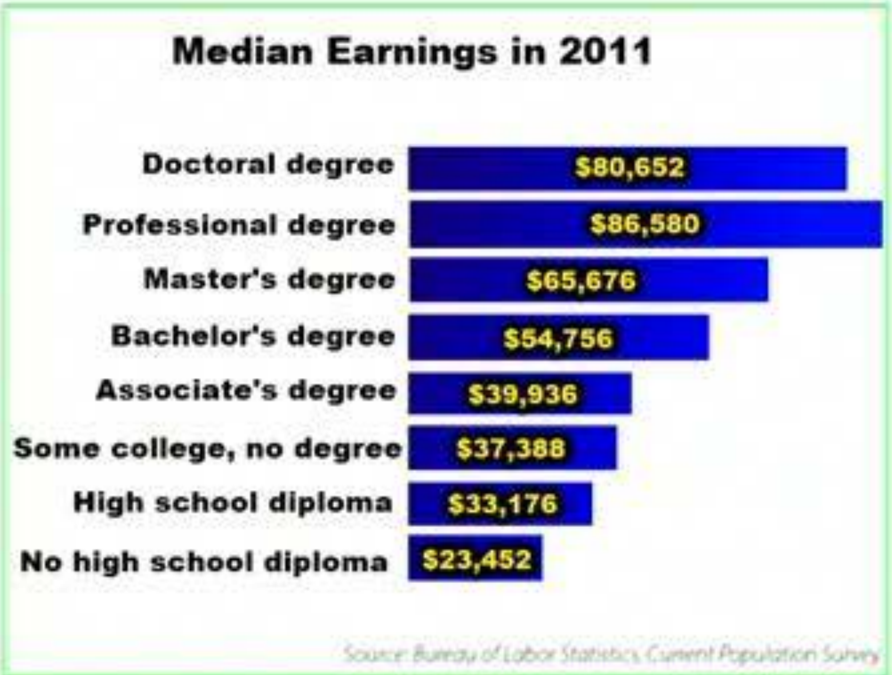
Source: Bureau of Justice Statistics



Source: Bureau of Labor Statistics, Current Population Survey

Evaluation example: Predicting income from basic groups

- Predict income = {high,low} in census dataset
- No labels, but can build bags of people based on **education level + gender**
- Ballpark constraints (surveys, previous census, sampling…):
  - $p_{PhD} \geq u$ *[**PhDs** have higher income than **overall** population]*
  - $p_{Some\ college} \leq p_{Bachelor's} \leq p_{PhD}$
  - …
- Bags are not used as features
  - Model can predict label without knowing education level



Source: Bureau of Labor Statistics Current Population Survey

Evaluation example: Learning from basic intuition

- Need 900 labels in SVM to match ballpark model
  - We use no labels!

Ballpark	**0.77**
Supervised SVM	900 labels (0.77)

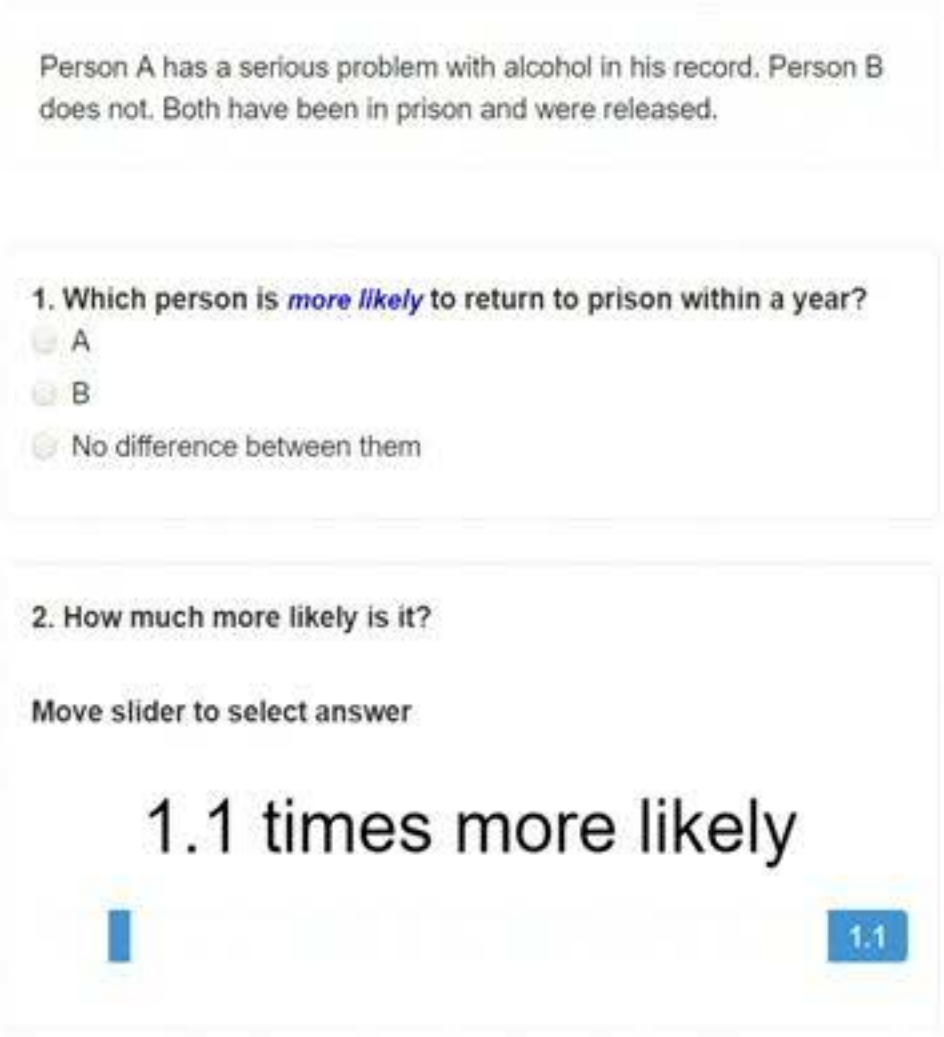
- Comparison to **Learning from Label Proportions** baselines *
  - **Exact** label proportions assumed known!

Mean Map	**Kernel Density Estimation**	**Discriminative Sorting**	**MCMC Sampling**
0.81	0.75	0.77	0.81

* Quadrianto et al., *Estimating Labels from Label Proportions*, JMLR 2009

Where do we get constraints?

- Ballpark crowdsourcing: Pooling noisy crowd guesses on intuitive groups



Person A has a serious problem with alcohol in his record. Person B does not. Both have been in prison and were released.

1. Which person is *more likely* to return to prison within a year?

○ A

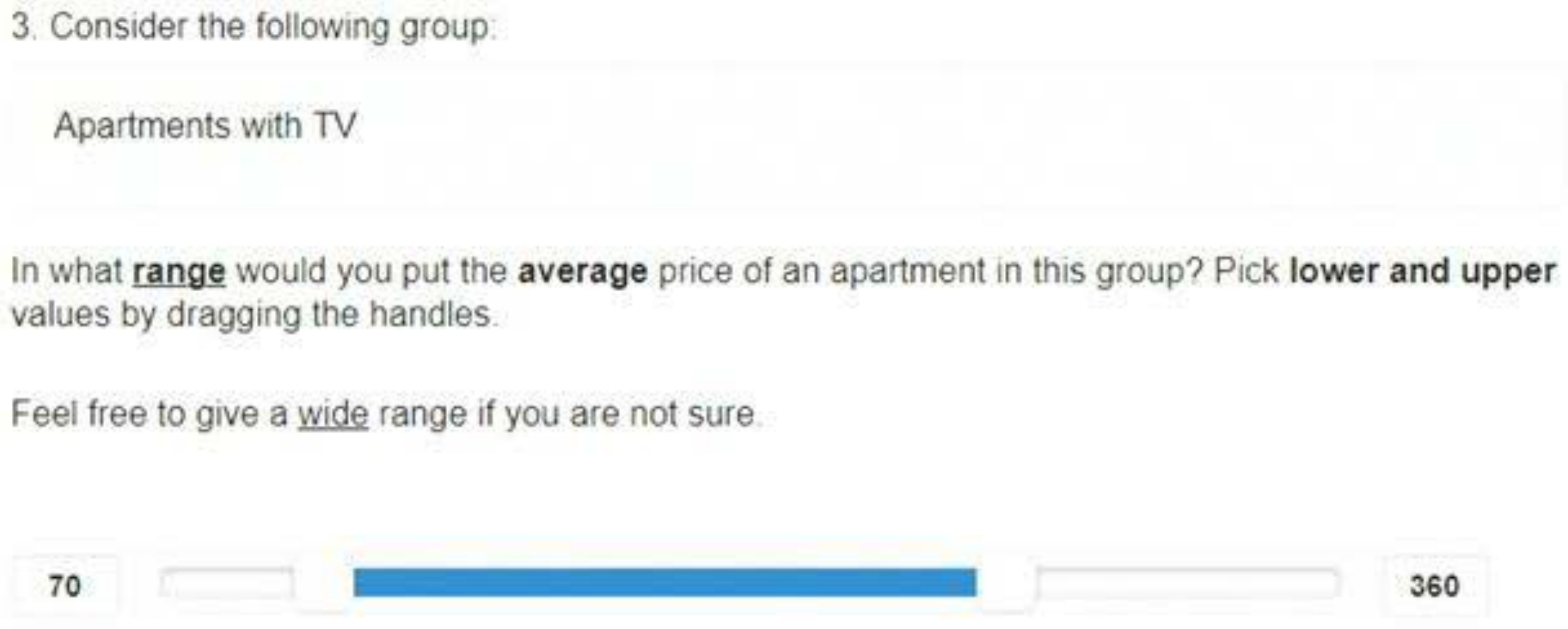
○ B

○ No difference between them

2. How much more likely is it?

Move slider to select answer

1.1 times more likely



3. Consider the following group:

Apartments with TV

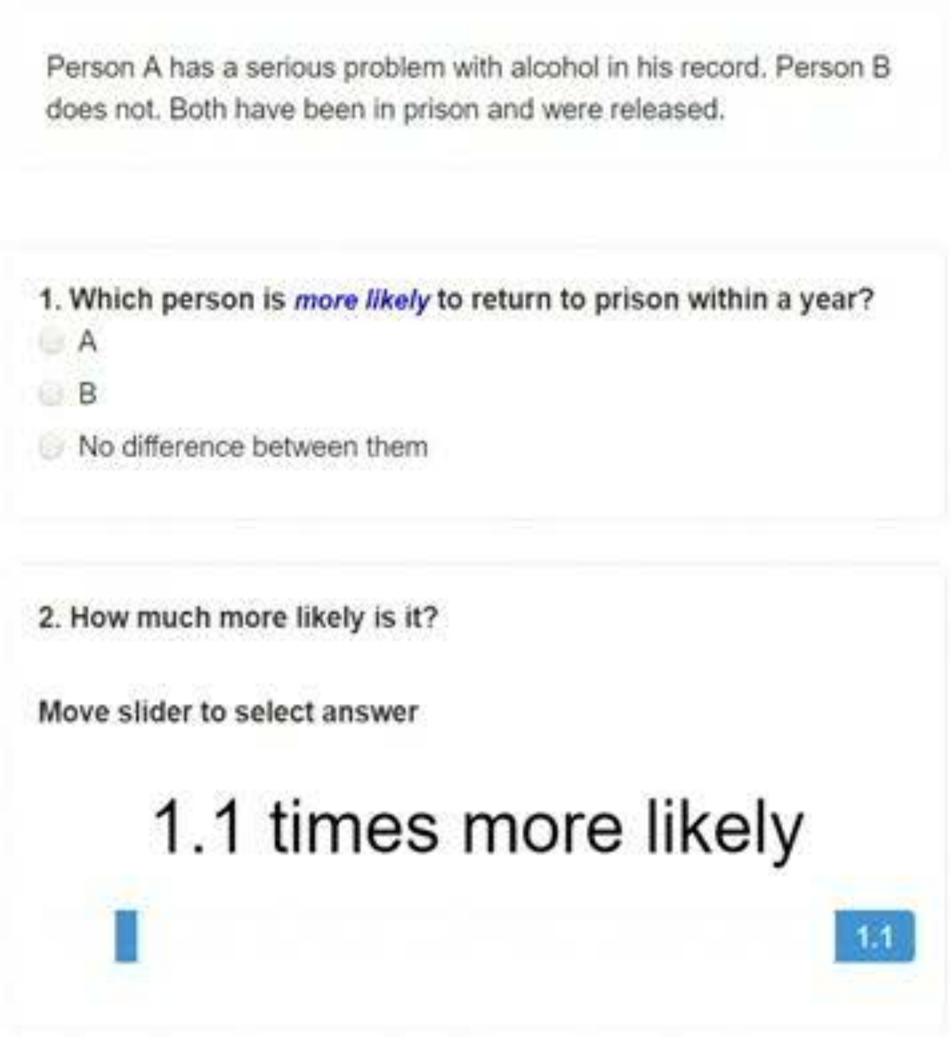
In what <u>**range**</u> would you put the **average** price of an apartment in this group? Pick **lower and upper** values by dragging the handles.

Feel free to give a <u>wide</u> range if you are not sure.

**** Ballpark Crowdsourcing: The Wisdom of Rough Group Comparisons, WSDM '18***

Where do we get constraints?

- Ballpark crowdsourcing: Pooling noisy crowd guesses on intuitive groups



Person A has a serious problem with alcohol in his record. Person B does not. Both have been in prison and were released.

1. Which person is ***more likely*** to return to prison within a year?

○ A

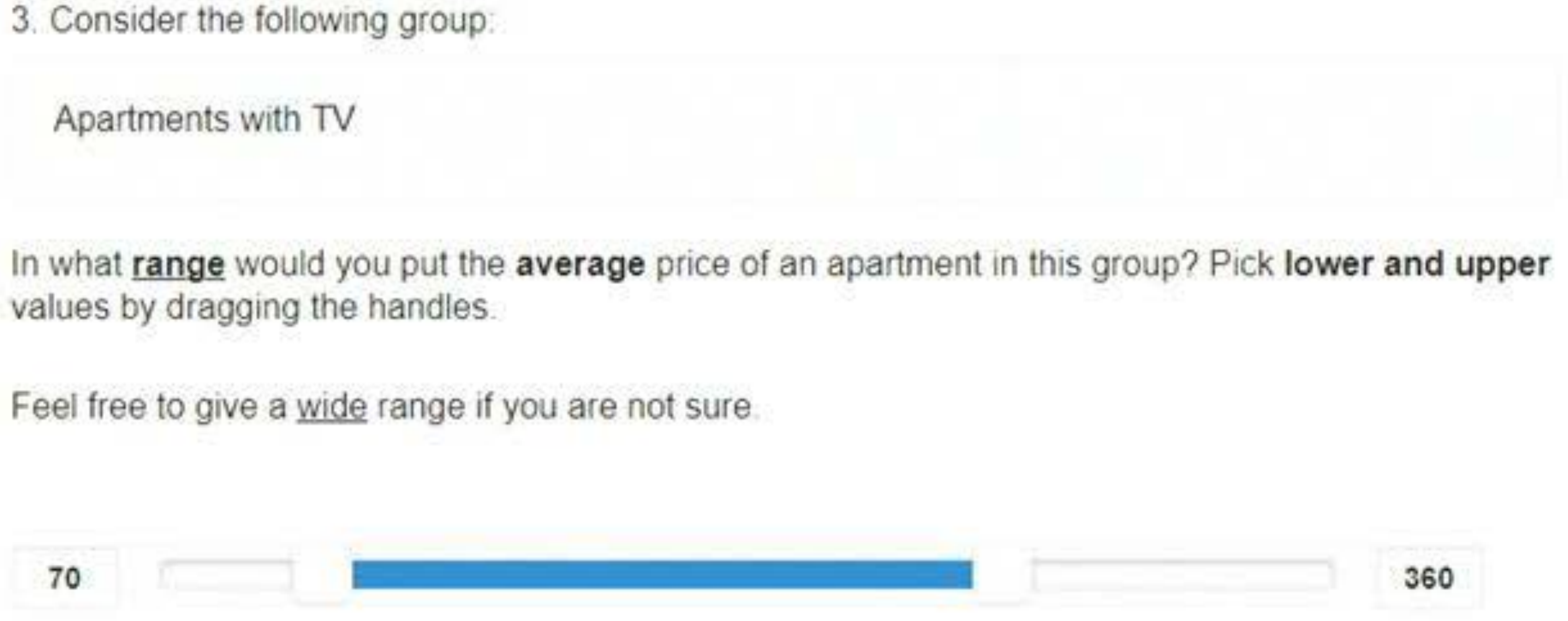
○ B

○ No difference between them

2. How much more likely is it?

Move slider to select answer

1.1 times more likely



3. Consider the following group:

Apartments with TV

In what <u>**range**</u> would you put the **average** price of an apartment in this group? Pick **lower and upper** values by dragging the handles.

Feel free to give a <u>wide</u> range if you are not sure.

- Results rival supervised models that use many **true** labels
- Crowd guesses on **individual** instances yield poor results
- Total cost **3X** less than standard labeling
  - Questions on groups “cover” many instances

hyadata lab **** Ballpark Crowdsourcing: The Wisdom of Rough Group Comparisons, WSDM ‘18***

When to go to the Ballpark

- When we have no labels, only rough information on groups
- When crowdsourcing for individual labels is hard
    - Less need for domain **experts**: Human intuition on groups & comparisons *
    - Fewer questions – cutting **cost**
    - No **privacy** concerns
- Also helps with:
    - **Continuous** targets**,** crowd **bias**, **outliers**…
- Can **complement** standard crowdsourcing

** Thurstone, L.L. (1927). A law of comparative judgement. Psychological Review*

Ballpark for discovery (work in progress)

- Goal: Identify interesting/novel/groundbreaking patents
- Where do get our aggregate constraints?

Moving to higher levels of innovation

Level 5: Discovery (1%)

Level 4: Invention outside the paradigm (4%)

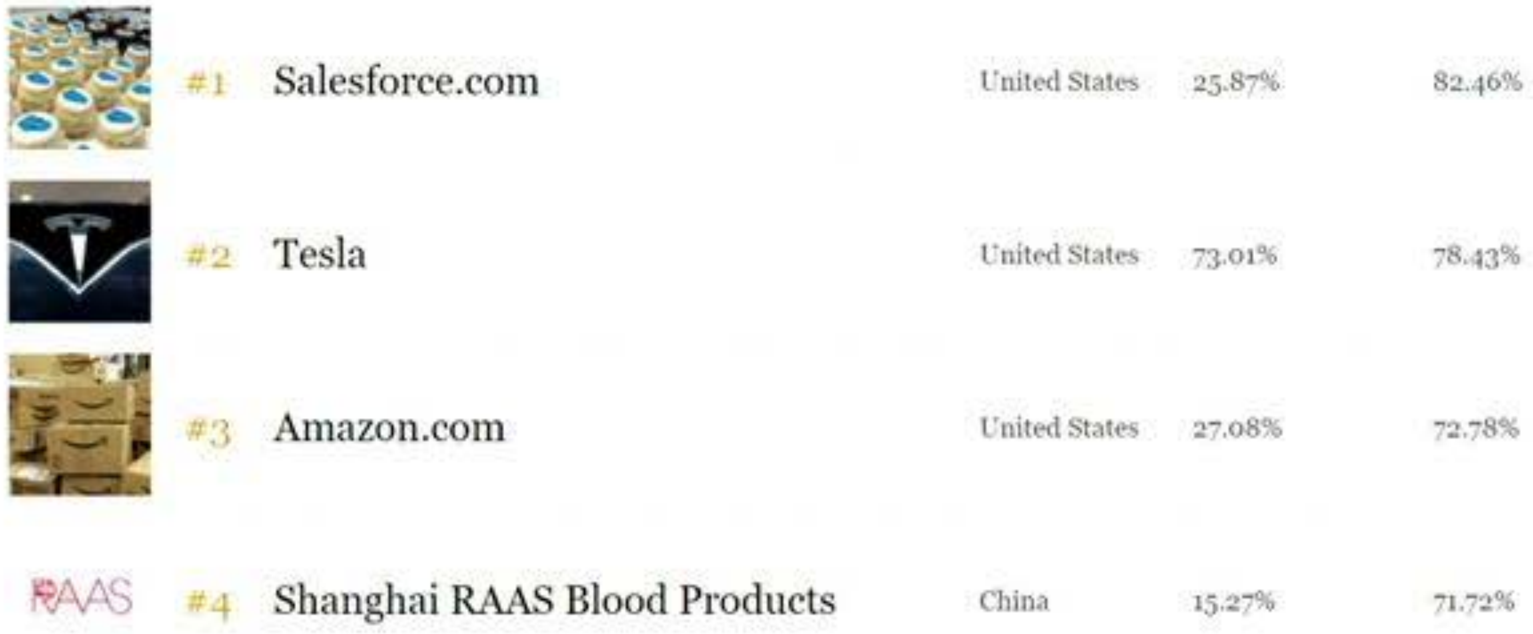
Level 3: Invention inside the paradigm (18%)

Level 2: Improvement (45%)

Level 1: Apparent solution (no innovation) (32%)

TRIZ Level of Invention (Fey & Rivlin 2005; Savransky 2000)

Company rankings



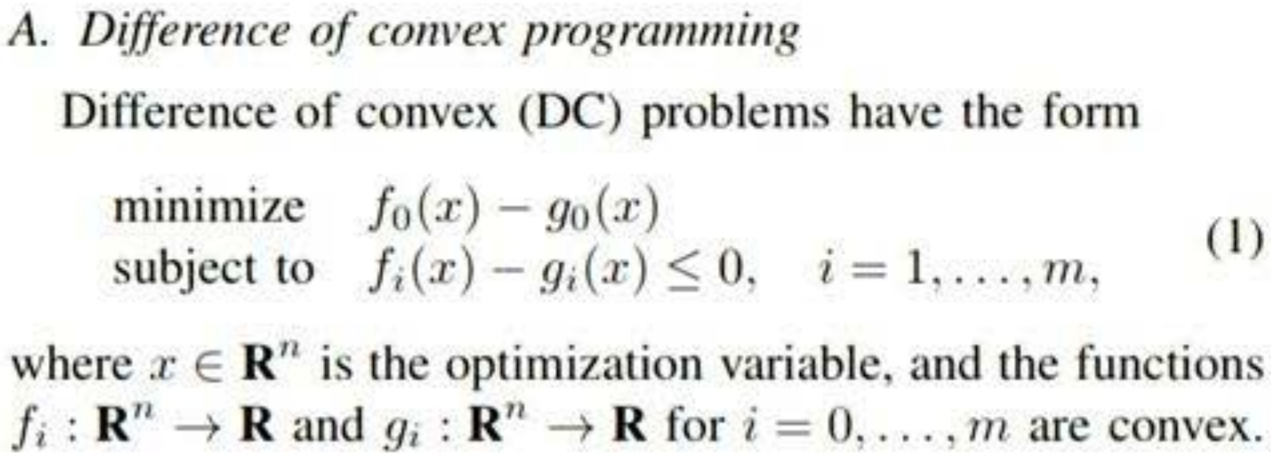
	#1	Salesforce.com	United States	25.87%	82.46%
	#2	Tesla	United States	73.01%	78.43%
	#3	Amazon.com	United States	27.08%	72.78%
RAAS	#4	Shanghai RAAS Blood Products	China	15.27%	71.72%

Ballpark + analogy for discovery

- Combining distant mechanisms, mechanisms for new/far purposes…
  - Rich economic literature: Recombination + innovation…
- Enrich with patent citation network/text embeddings…

Non-convex constraints to control distribution

- Challenge: Only top-1% of patents by companies might be novel…
- We want to gain finer control on distribution
  - Constrain *sum_of_top_k* (convex function)
  - *sum_top_k (Microsoft) > sum_top_k (ACME)*
- Leads to violation of convex programming



A. Difference of convex programming

Difference of convex (DC) problems have the form

$$\begin{array}{ll} \text{minimize} & f_0(x) - g_0(x) \\ \text{subject to} & f_i(x) - g_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{1}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and the functions $f_i : \mathbf{R}^n \to \mathbf{R}$ and $g_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 0, \ldots, m$ are convex.

- Initial eyeballing:
  - Higher recombination indicators and citation counts (not used as features)
    - Better than anomaly detection on same features
  - Next-up: Patent examiners evaluation

Looking forward

Future work and broader interests

Looking forward: Future work

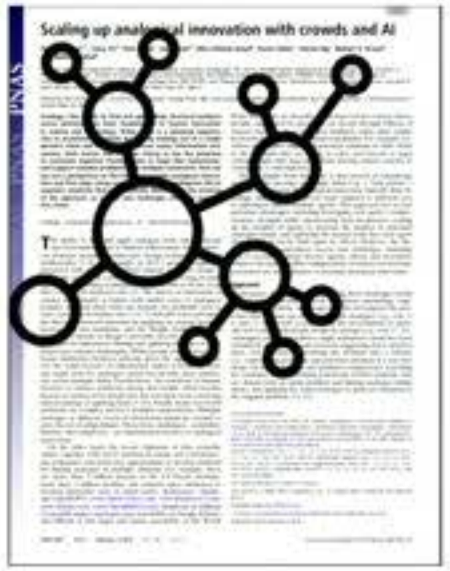
- Analogies:
  - Richer NLP models/representations for analogies
    - Graph extraction
    - External commonsense KG
  - Live search engine for researchers, engineers, designers, public
- Ballpark:
  - Non-convex extension for patents (in progress)
  - Learning deep neural networks from rough aggregate constraints

Looking forward: Broader interests

Developing new models for complex texts and behavioral data in the domains of scientific knowledge discovery, health, social science, psychology

Looking forward: Modelling complex texts

- Modelling complex long texts with hierarchical graph neural nets *
  - Scientific, patents, conversations, medical, books, legal, websites, film scripts…



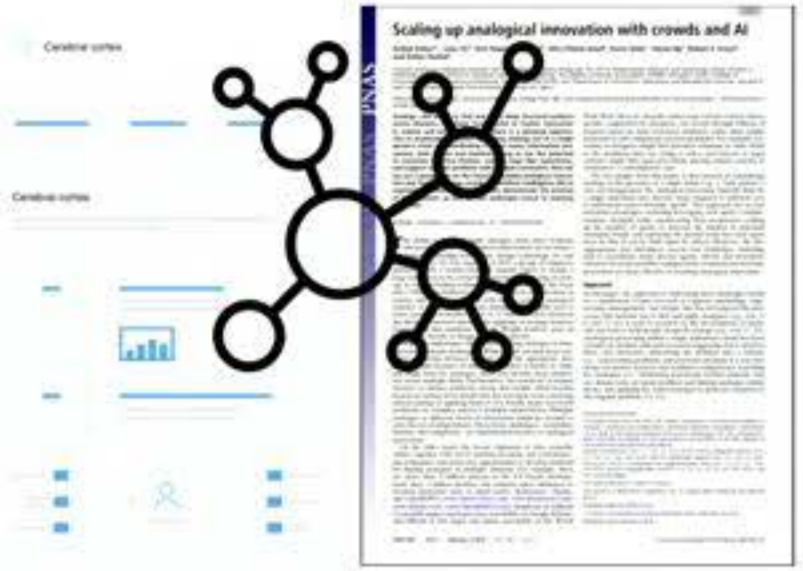
PNAS

Scaling up analogical innovation with crowds and AI

* *Peng, Poon et al. (2017), Cross-Sentence N-ary Relation Extraction with Graph LSTMs*

Looking forward: Modelling complex texts

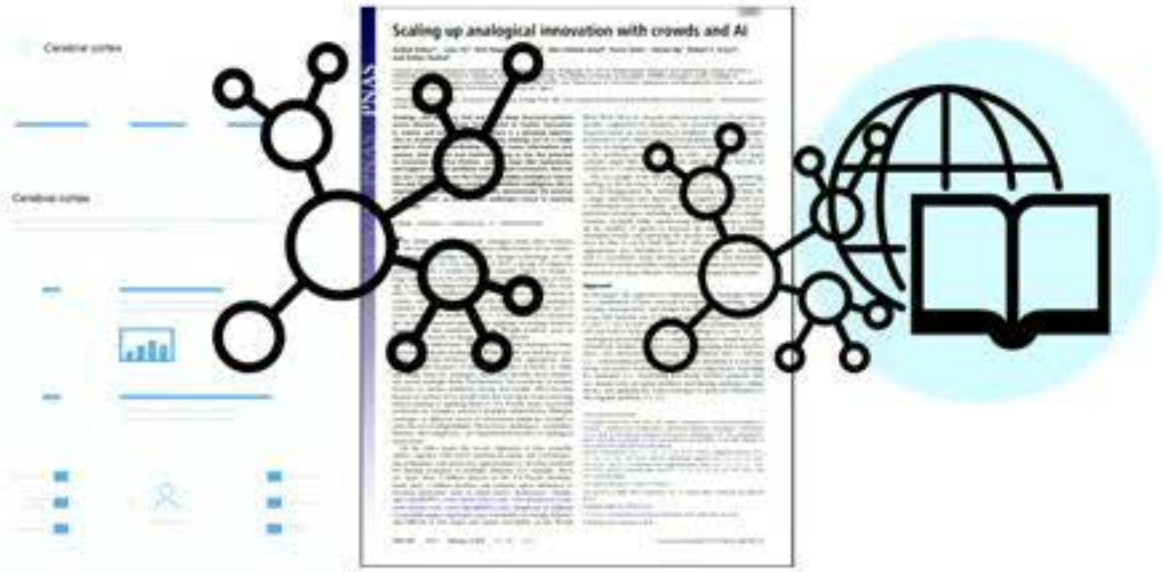
- Modelling complex long texts with hierarchical graph neural nets *
  - Scientific, patents, conversations, medical, books, legal, websites, film scripts…
  - **Multimodal** fusion of **search behavior** with **scientific paper content**
    - Example: How do we search for new ideas? Idea search beyond scientific domain
    - Graphs: Matching behavior with fine-grained information



** Peng, Poon et al. (2017), Cross-Sentence N-ary Relation Extraction with Graph LSTMs*

Looking forward: Modelling complex texts

- Modelling complex long texts with hierarchical graph neural nets *
  - Scientific, patents, conversations, medical, books, legal, websites, film scripts…
  - **Multimodal** fusion of **search behavior** with **scientific paper content**
    - Example: How do we search for new ideas? Idea search beyond scientific domain
    - Graphs: Matching behavior with fine-grained information
  - Incorporating external world knowledge graphs / databases into NLP models
    - Medical KG / EHR data **



** Peng, Poon et al. (2017), Cross-Sentence N-ary Relation Extraction with Graph LSTMs*

*** Nordon, Horvitz et al. (2019) Separating Wheat from Chaff: Joining Biomedical Knowledge and Patient Data for Repurposing Medications*

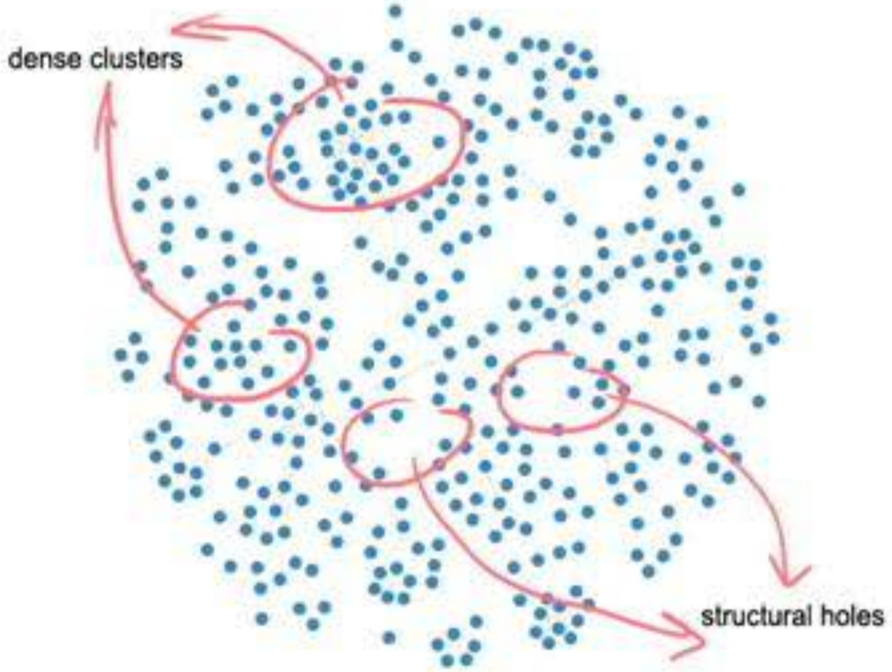
Looking forward: Scientific discovery

- Predicting the trajectories of science
  - Emerging trends across dynamic heterogenous graph of scientific knowledge*
  - Predicting the path of experts *("what will your next paper be about"?*)

** Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Looking forward: Scientific discovery

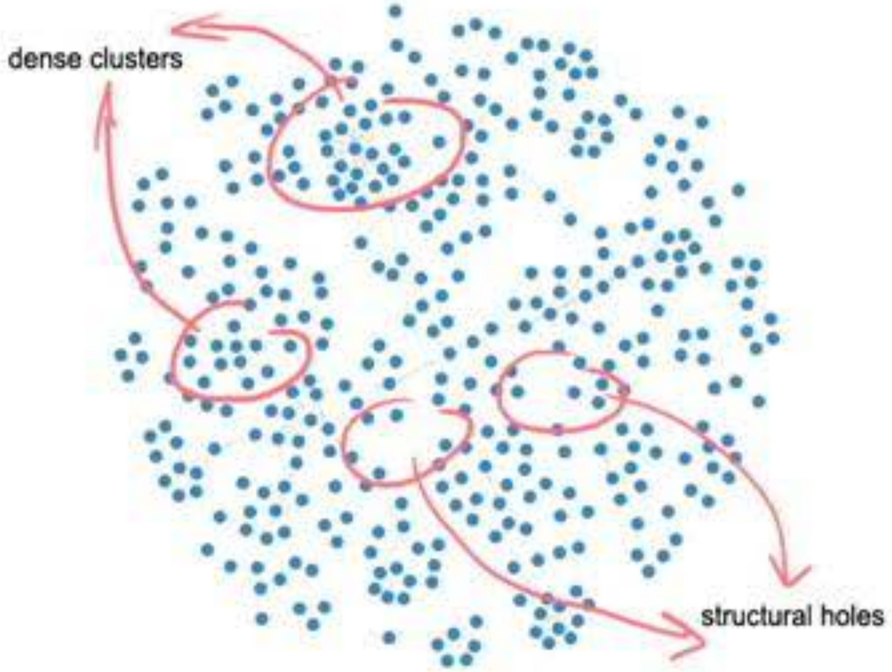
- Predicting the trajectories of science
  - Emerging trends across dynamic heterogenous graph of scientific knowledge*
  - Predicting the path of experts *("what will your next paper be about"?)*
  - Can we identify structural holes with **high-value** research potential?



* *Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Looking forward: Scientific discovery

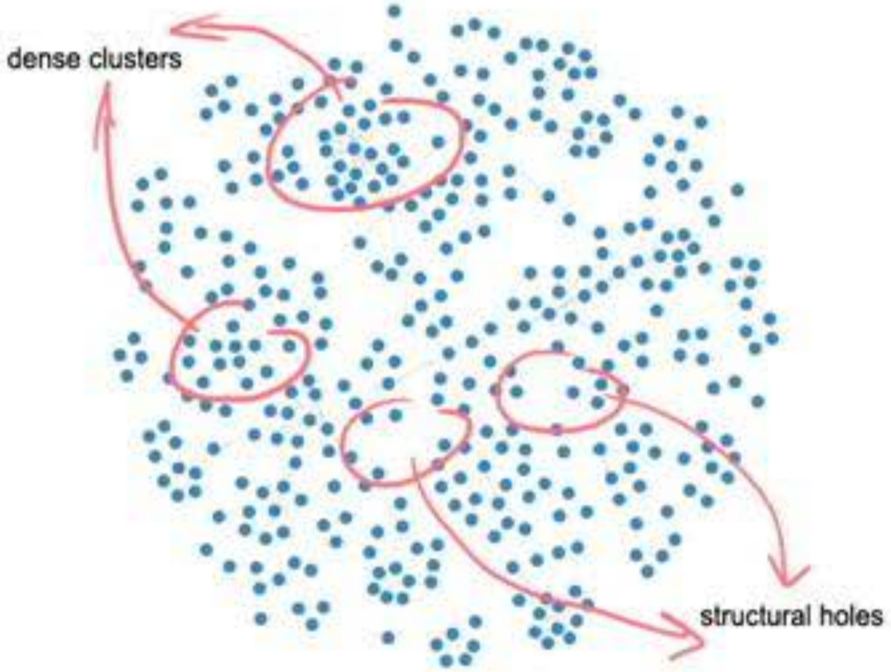
- Predicting the trajectories of science
  - Emerging trends across dynamic heterogenous graph of scientific knowledge*
  - Predicting the path of experts *("what will your next paper be about"?*)
  - Can we identify structural holes with **high-value** research potential?



* *Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Looking forward: Scientific discovery

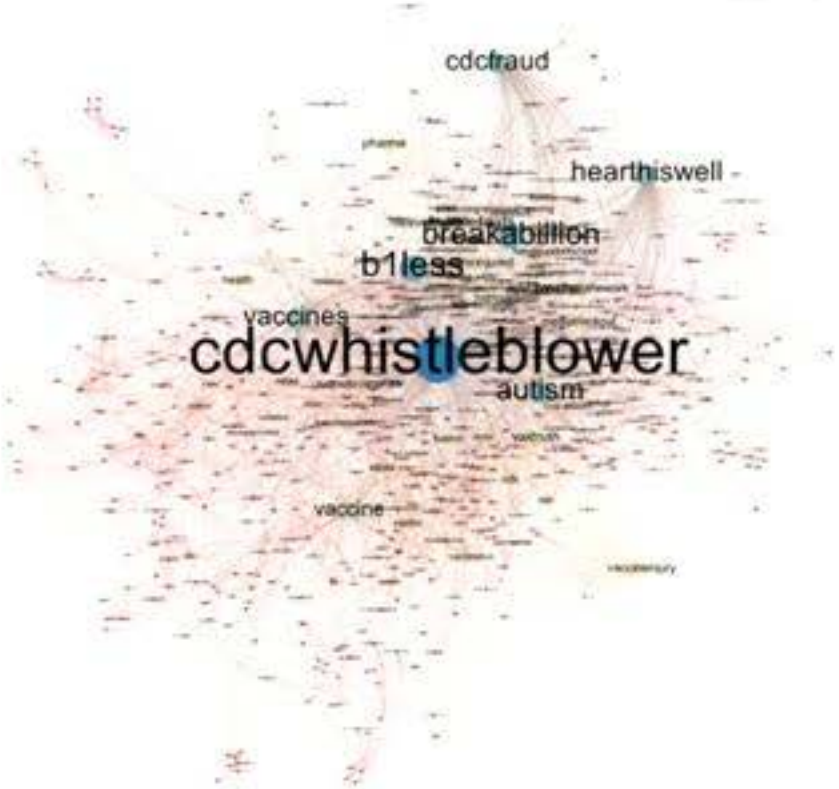
- Predicting the trajectories of science
  - Emerging trends across dynamic heterogenous graph of scientific knowledge*
  - Predicting the path of experts *("what will your next paper be about"?*)
  - Can we identify structural holes with **high-value** research potential?
  - Can we learn generative models of ideas, and then "sample" from them?



* *Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Looking forward: Social & health insights

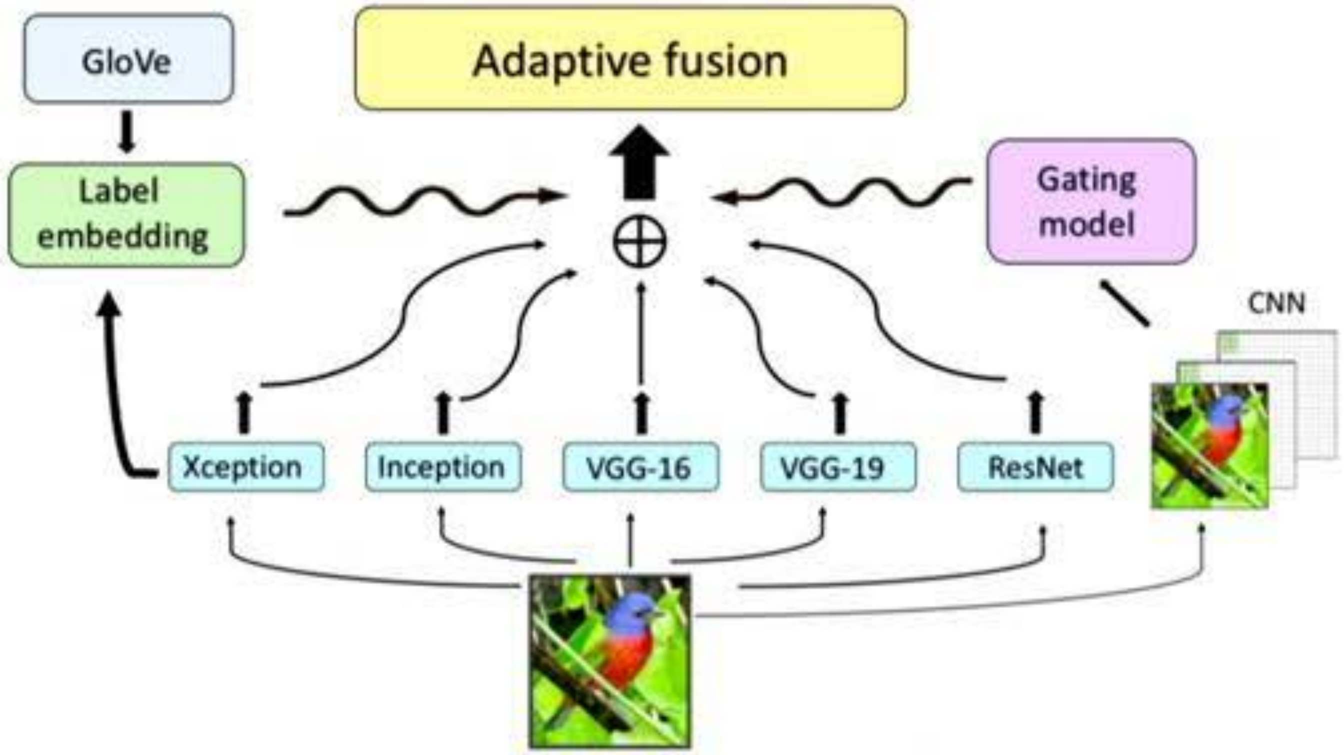
- Emergent/novel social phenomena, new concepts, misconceptions
  - Anti-vaccination movement from online discussions
  - Predicting collective behavior and decision-making from group conversations
  - Can we identify group dynamics and latent individual roles? [*]
- Discover new health/psychological behaviors, symptoms, causes
  - Learn knowledge graph of health-related activities, psychological states in context



[*] Roles People Play in Groups, Stanford

Research in health and social knowledge discovery

- Lead applied research team – NLP, CV, graph ML
  - Healthcare, knowledge discovery from social media and web
- Adaptive fusion of pre-trained vision + NLP models for transfer learning



Altogether now! The Benefits of Adaptively Fusing Pre-trained Deep Representations, ICPRAM '19 Best Research Paper

Conclusion

- Boosting innovation with analogies
  - Extract purposes/mechanisms from noisy real-world texts
  - Products, patents, science
  - Enhancing creativity, expressive search, commonsense functional ontology

- Weak supervision with rough group information
  - New learning framework based on coarse label average constraints
  - Evaluate across many domains
  - Identifying novel ideas: Non-convex extension for patents (in progress)

Conclusion

- Boosting innovation with analogies
  - Extract purposes/mechanisms from noisy real-world texts
  - Products, patents, science
  - Enhancing creativity, expressive search, commonsense functional ontology

- Weak supervision with rough group information
  - New learning framework based on coarse label average constraints
  - Evaluate across many domains
  - Identifying novel ideas: Non-convex extension for patents (in progress)

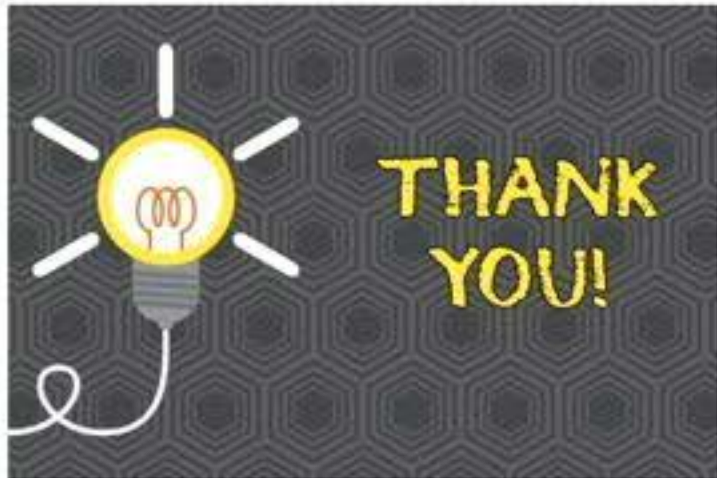
- Broader interests:
  - Modelling complex texts + behavioral data for knowledge discovery in science, health, novel social phenomena

Conclusion

- Boosting innovation with analogies
  - Extract purposes/mechanisms from noisy real-world texts
  - Products, patents, science
  - Enhancing creativity, expressive search, commonsense functional ontology

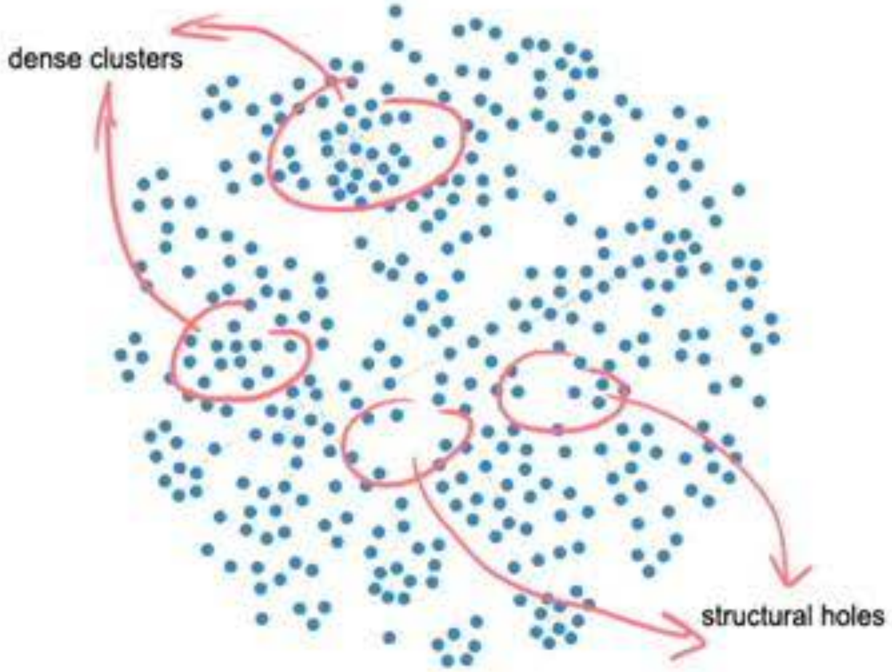
- Weak supervision with rough group information
  - New learning framework based on coarse label average constraints
  - Evaluate across many domains
  - Identifying novel ideas: Non-convex extension for patents (in progress)

- Broader interests:
  - Modelling complex texts + behavioral data for knowledge discovery in science, health, novel social phenomena



Looking forward: Scientific discovery

- Predicting the trajectories of science
    - Emerging trends across dynamic heterogenous graph of scientific knowledge*
    - Predicting the path of experts *("what will your next paper be about"?*)
    - Can we identify structural holes with **high-value** research potential?
    - Can we learn generative models of ideas, and then "sample" from them?



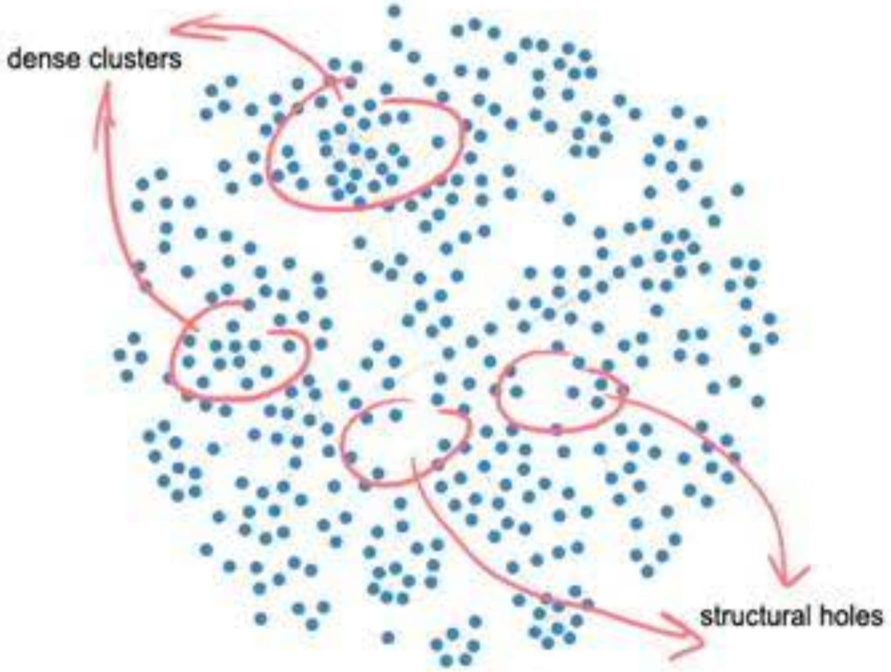
* *Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Ballpark + analogy for discovery

- Combining distant mechanisms, mechanisms for new/far purposes…
  - Rich economic literature: Recombination + innovation…
- Enrich with patent citation network/text embeddings…

Looking forward: Scientific discovery

- Predicting the trajectories of science
  - Emerging trends across dynamic heterogenous graph of scientific knowledge*
  - Predicting the path of experts *("what will your next paper be about"?*)
  - Can we identify structural holes with **high-value** research potential?
  - Can we learn generative models of ideas, and then "sample" from them?



** Sinha et al. (2015), An Overview of Microsoft Academic Service (MAS) and Applications*

Ballpark + analogy for discovery

- Combining distant mechanisms, mechanisms for new/far purposes…
  - Rich economic literature: Recombination + innovation…
- Enrich with patent citation network/text embeddings…