# Stacking With Auxiliary Features: Improved Ensembling for Natural Language and Vision

Nazneen Rajani

PhD Proposal

November 7, 2016

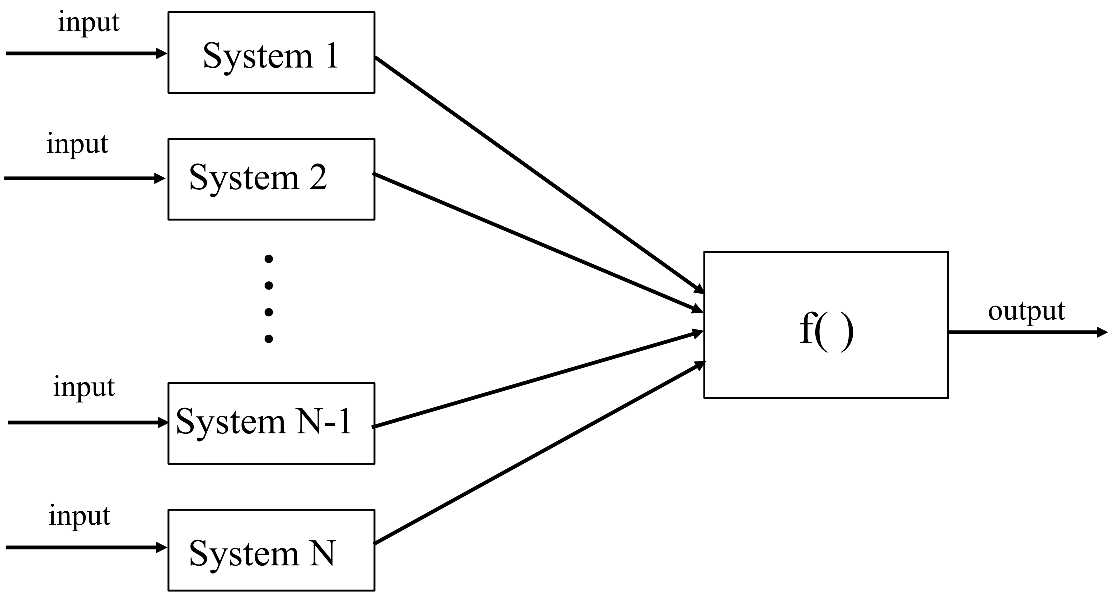
Committee members: Ray Mooney, Katrin Erk, Greg Durrett and Ken Barker

# Outline

- Introduction
- Background & Related Work
- Completed Work
  - Stacked Ensembles of Information Extractors for Knowledge Base Population (ACL 2015)
  - Stacking With Auxiliary Features (Under review)
  - Combining Supervised and Unsupervised Ensembles for Knowledge Base Population (EMNLP 2016)
- Proposed Work
  - Short-term proposals
  - Long-term proposals

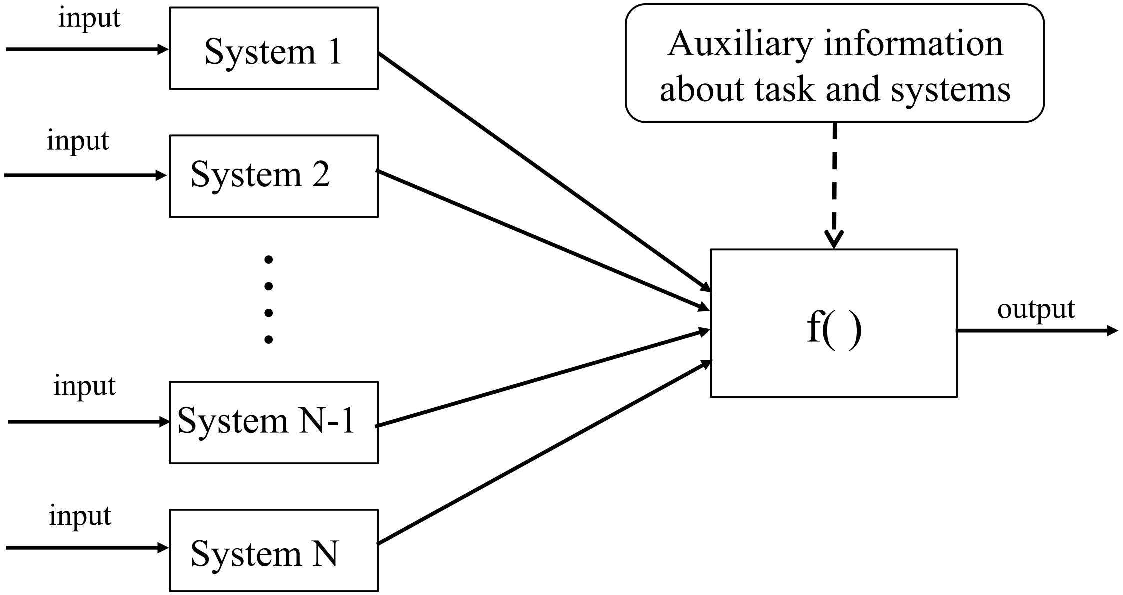
# Introduction

- Ensembling: Used by the $1M winning team for the Netflix competition



# Introduction

- Make auxiliary information accessible to the ensemble



# Background and Related Work

# Cold Start Slot Filling (CSSF)

- Knowledge Base Population (KBP) is a task of discovering entity facts and adding to a KB
- Relation extraction, a KBP sub-task, using fixed ontology is slot filling
- CSSF is an annual NIST evaluation of building KB from scratch
  - query entities and pre-defined slots
  - text corpus

# Cold Start Slot Filling (CSSF)

- Some slots are single-valued (per: age) while some are list-valued (per: children)
- Entity types: PER, ORG, GPE
- Along with fills, systems must provide
  - confidence score
  - provenance — *docid*: *startoffset-endoffset*

# Cold Start Slot Filling (CSSF)

org: Microsoft

1. city_of_headquarters:
2. website:
3. subsidiaries:
4. employees:
5. shareholders:

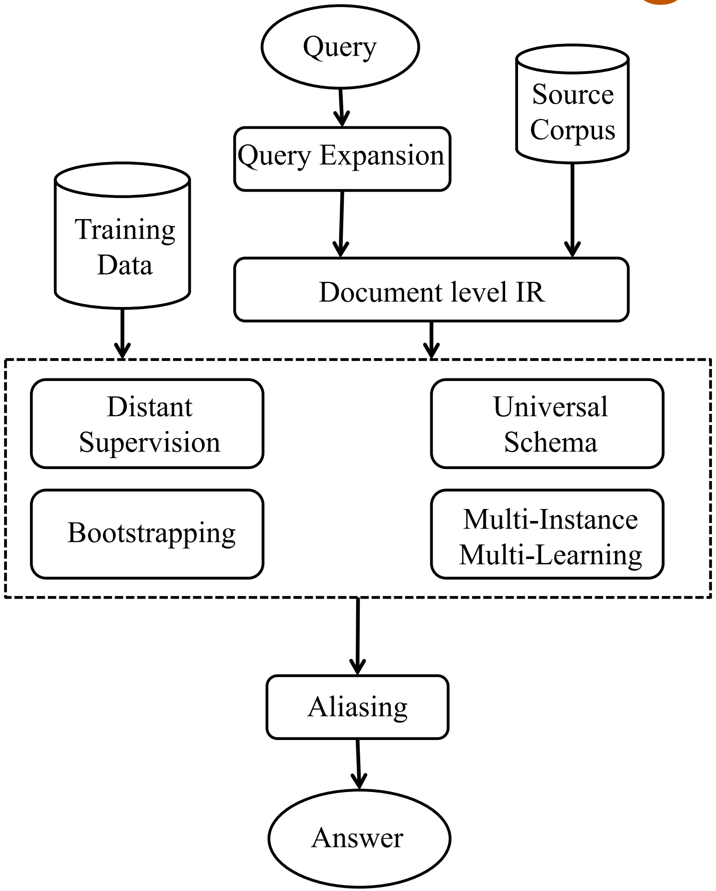
⋮

Microsoft is a technology company, headquartered in Redmond, Washington that develops …

**city_of_headquarters:**
Redmond
**provenance:**

**confidence score:**
1.0

# Cold Start Slot Filling (CSSF)



# Entity Discovery and Linking (EDL)

- KBP sub-task involving two NLP problems
  - Named Entity Recognition (NER)
  - Disambiguation
- EDL is an annual NIST evaluation in 3 languages: English, Spanish and Chinese
- Tri-lingual Entity Discovery and Linking (TEDL)

# Tri-lingual Entity Discovery and Linking (TEDL)

- Detect all entity mentions in corpus
- Link mentions to English KB (FreeBase)
- If no KB entry found, cluster into a NIL ID
- Entity types — PER, ORG, GPE, FAC, LOC
- Systems must also provide confidence score

# Tri-lingual Entity Discovery and Linking (TEDL)

**Source Corpus Document:**
*Hillary Clinton* Not Talking About ’92 *Clinton*-Gore Confederate Campaign Button..

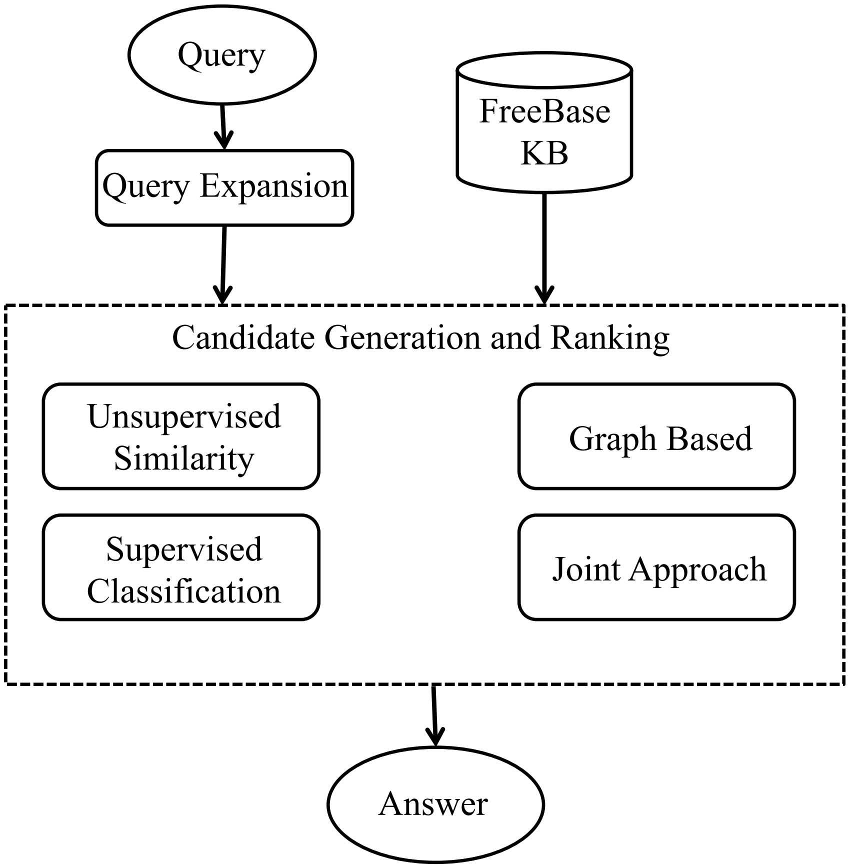
**FreeBase entry:**

Hillary Diane Rodham Clinton is a US Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate.

**FreeBase entry:**

William Jefferson "Bill" Clinton is an American politician who served as the 42nd President of the United States from 1993 to 2001. Clinton was Governor of Arkansas from 1979 to 1981 and 1983 to 1992, and Arkansas Attorney General from 1977 to 1979.

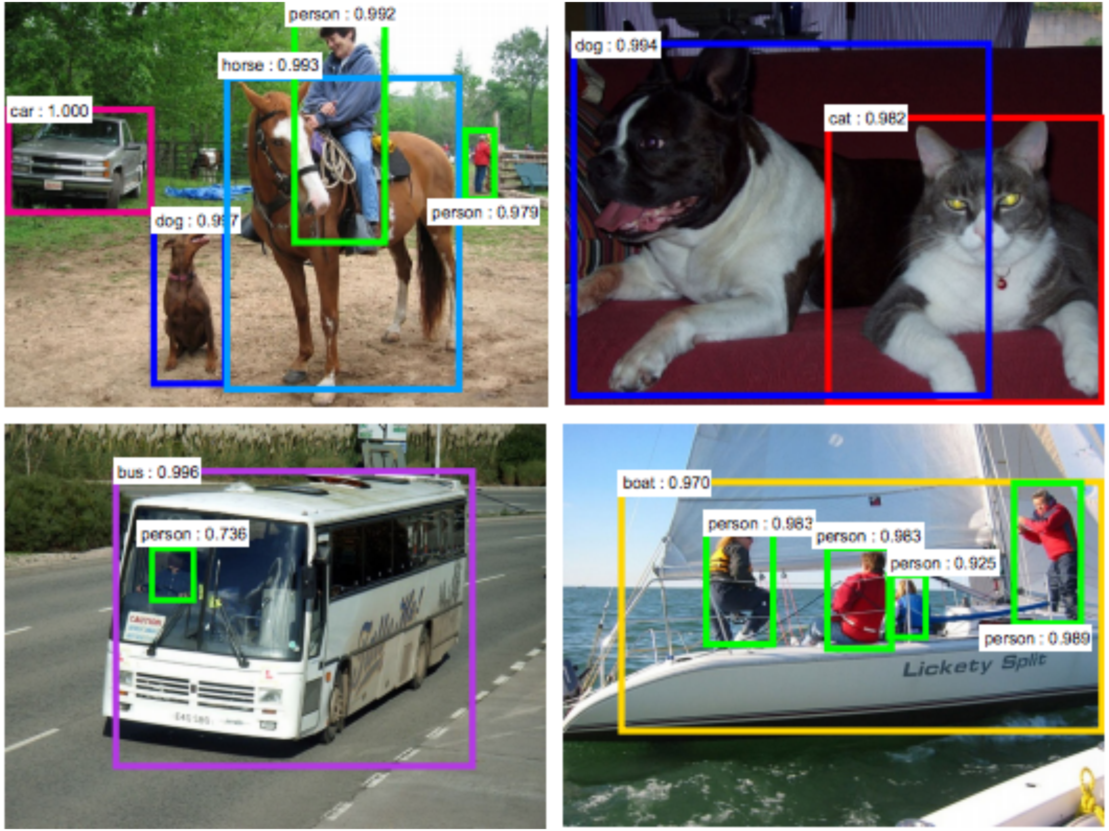
# Tri-lingual Entity Discovery and Linking (TEDL)



# ImageNet Object Detection

- Widely known annual competition in CV for large-scale object recognition
- Object detection
  - detect all instances of object categories (total 200) in images
  - localize using axis-aligned Bounding Boxes (BB)
- Object categories are WordNet synsets
- Systems also provide confidence scores

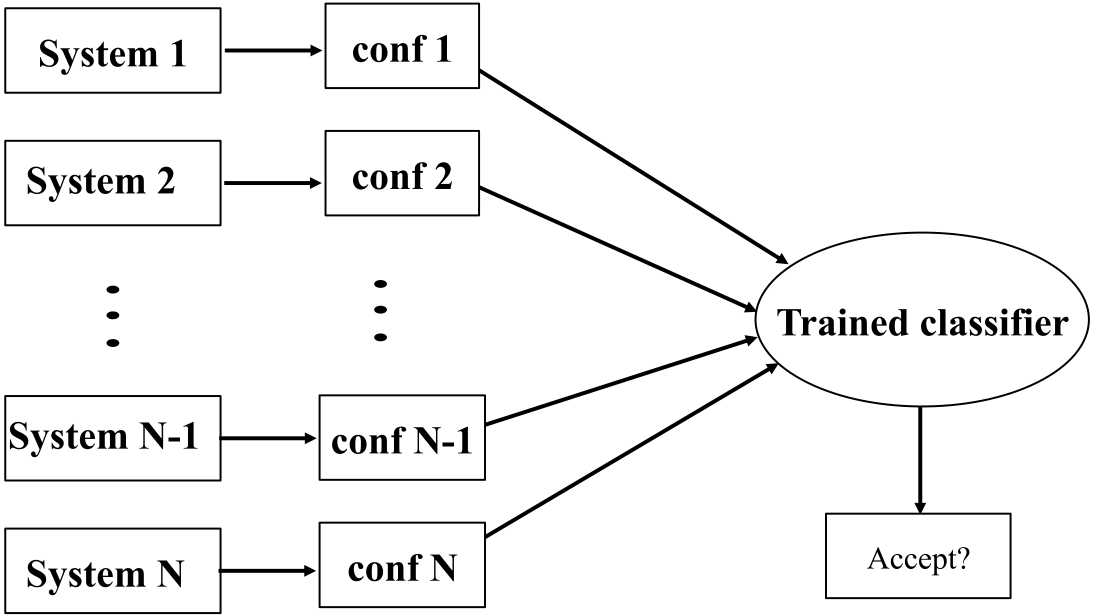
# ImageNet Object Detection



# Ensemble Algorithms

(Wolpert, 1992)

- Stacking



# Ensemble Algorithms

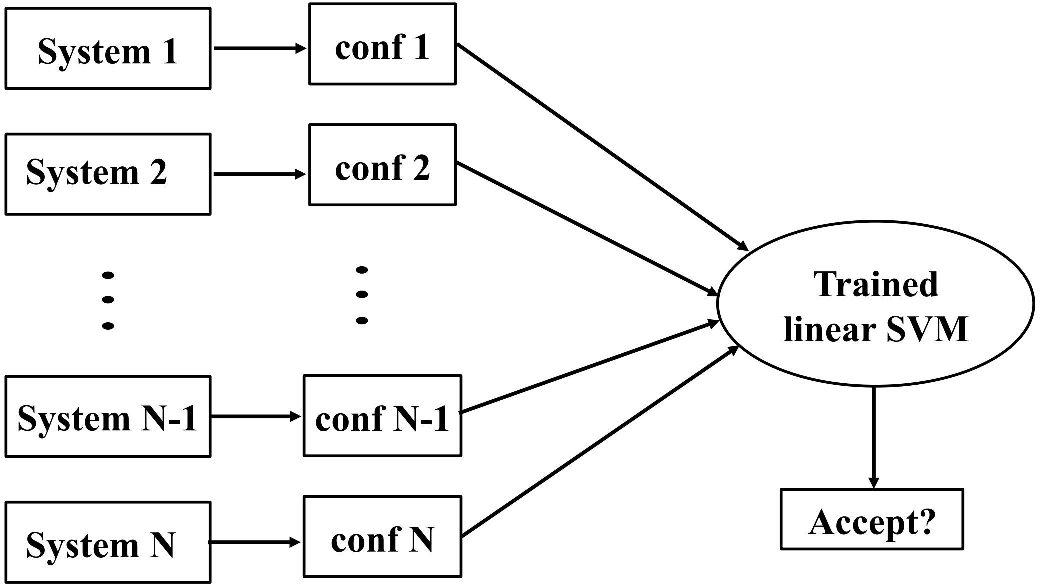
- Bipartite Graph-based Consensus Maximization (BGCM) (Gao et al., 2009)
  - ensembling -> optimization over bipartite graph
  - combining supervised and unsupervised models
- Mixtures of Experts (ME) (Jacobs et al., 1991)
  - partition the problem into sub-spaces
  - learn to switch experts based on input using a gating network
  - Deep Mixtures of Experts (Eigen et al., 2013)

# Completed Work:
# I. Stacked Ensembles of Information Extractors for Knowledge Base Population (ACL2015)

# Stacking

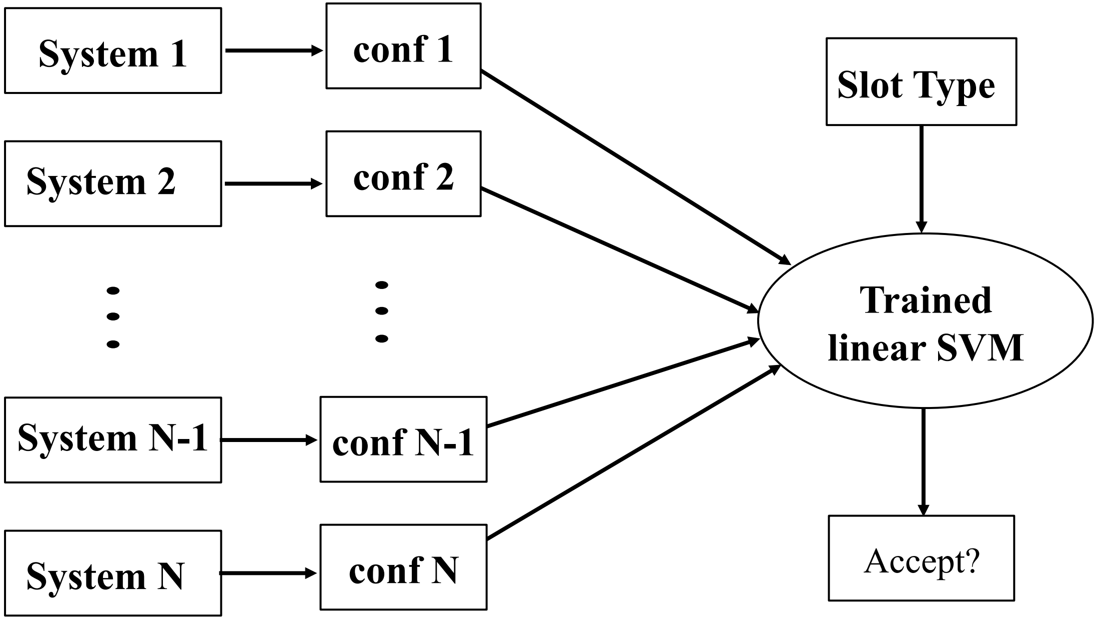
(Wolpert, 1992)

For a given proposed slot-fill, e.g. spouse(Barack, Michelle), combine confidences from multiple systems:



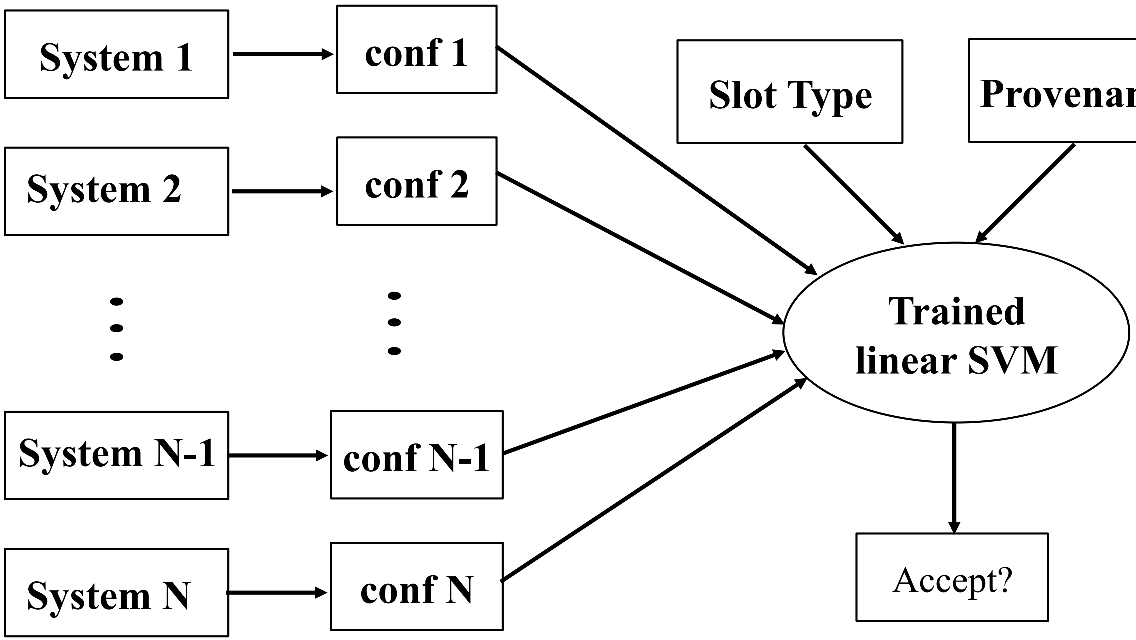
# Stacking with Features

For a given proposed slot-fill, e.g. spouse(Barack, Michelle), combine confidences from multiple systems:



# Stacking with Features

For a given proposed slot-fill, e.g. spouse(Barack, Michelle), combine confidences from multiple systems:



# Document Provenance Feature

- For a given query and slot, for each system, *i*, there is a feature $DP_i$:
  - $N$ systems provide a fill for the slot.
  - Of these, $n$ give same provenance *docid* as *i*.
  - $DP_i = n/N$ is the document provenance score.
- Measures extent to which systems agree on document provenance of the slot fill.

# Offset Provenance Feature

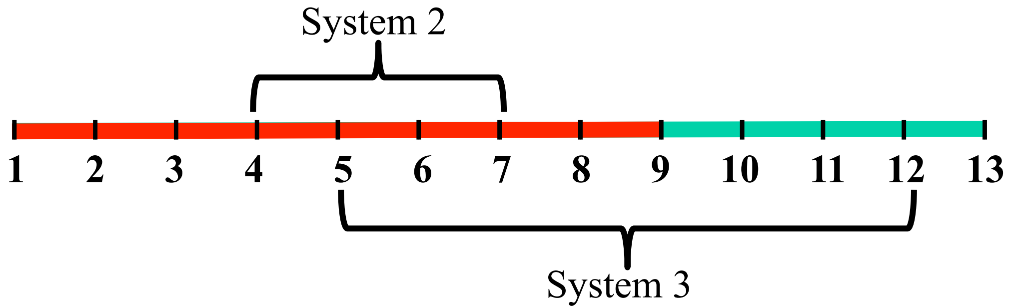
- Degree of overlap between systems' provenance strings.
- Uses Jaccard similarity coefficient.

$$PO(n) = \frac{1}{|N|} \times \sum_{i \in N, i \neq n} \frac{|\text{substring(i)} \cap \text{substring(n)}|}{|\text{substring(i)} \cup \text{substring(n)}|}$$

- Systems with different *docid* have zero OP

# Offset Provenance Feature

| Offsets | System 1 | System 2 | System 3 |
|---|---|---|---|
| Start Offset | 1 | 4 | 5 |
| End Offset | 9 | 7 | 12 |




$$OP_1 = \frac{1}{2} \times \left( \frac{4}{9} + \frac{5}{12} \right)$$

# Results

- Using the 10 common systems between 2013 and 2014

| Approach | Precision | Recall | F1 |
|---|---|---|---|
| Union | 0.176 | **0.647** | 0.277 |
| Voting (>=3) | **0.694** | 0.256 | 0.374 |
| Best ESF system in 2014 (Stanford) | 0.585 | 0.298 | 0.395 |
| Stacking | 0.606 | 0.402 | 0.483 |
| Stacking + Relation | 0.607 | 0.406 | 0.486 |
| Stacking + Provenance + Relation | 0.541 | 0.466 | **0.501** |

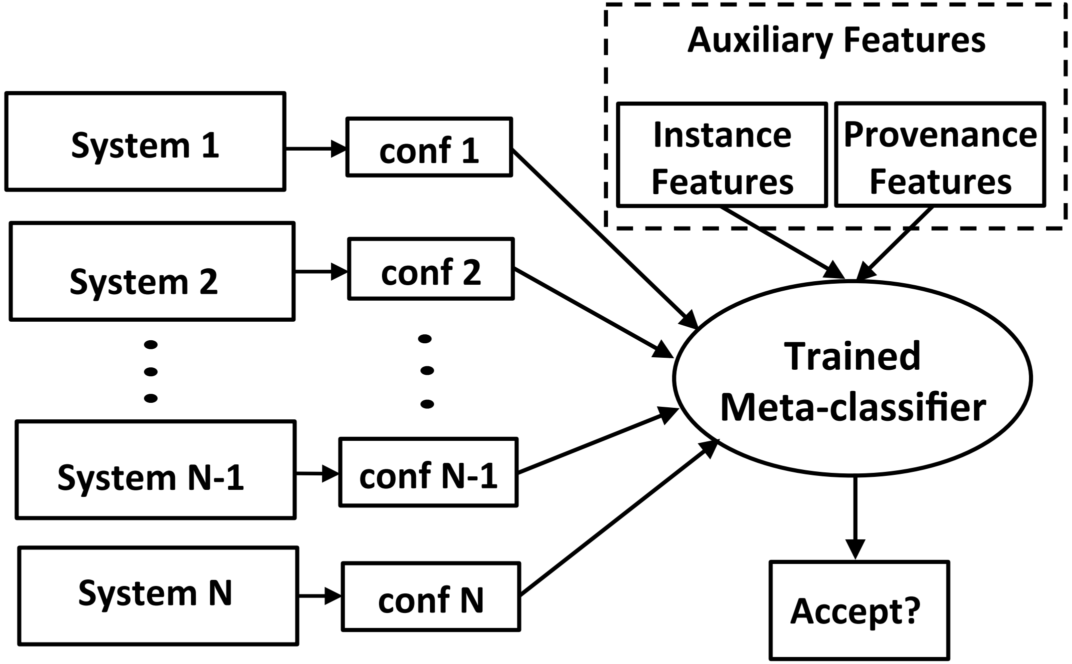
# Takeaways

- Stacked meta-classifier beats the best performing 2014 KBP SF system by an F1 gain of **11** points.
- Features that utilize auxiliary information improve stacking performance.
- Ensembling has clear advantages but naive approaches such as voting do not perform as well.
- Although systems change every year, there are advantages in training on past data.

# Completed Work:
# II. Stacking With Auxiliary Features (under review)

# Stacking With Auxiliary Features (SWAF)

- Stacking using two types of auxiliary features:



# Instance Features

- Enables stacker to discriminate between input instance types
- Some systems are better at certain input types
- CSSF — slot type (per: age)
- TEDL — entity type (PER/ORG/GPE/FAC/LOC)
- Object detection — object category and SIFT feature descriptors

# Provenance Features

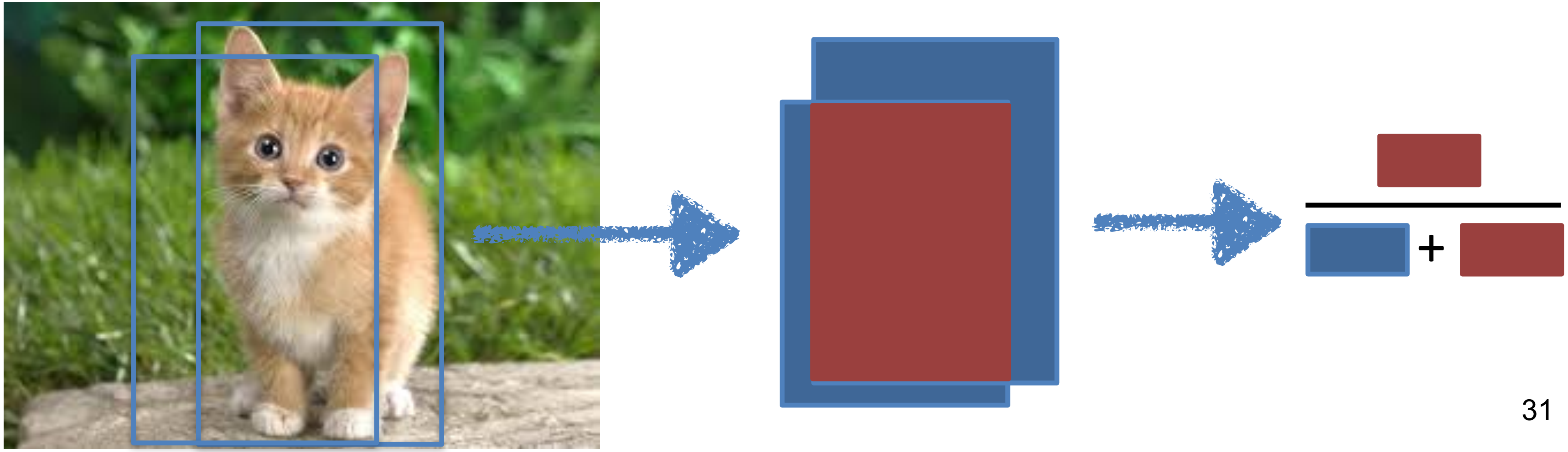
- Enables the stacker to discriminate between systems
- Output is reliable if systems agree on source
- CSSF same as slot filling
- TEDL — measures overlap of a mention

$$PO(n) = \frac{1}{|N|} \times \sum_{i \in N, i \neq n} \frac{|\text{substring(i)} \cap \text{substring(n)}|}{|\text{substring(i)} \cup \text{substring(n)}|}$$

# Provenance Features

- Object detection — measure BB overlap

$$BBO(n) = \frac{1}{|N|} \times \sum_{i \in N, i \neq n} \frac{|\text{Area}(i) \cap \text{Area}(n)|}{|\text{Area}(i) \cup \text{Area}(n)|}$$



# Post-processing

- CSSF
  - single valued slot fills — resolve conflicts
  - list valued slot fills — always include
- TEDL
  - KB ID — include in output
  - *NIL ID — merge across systems if at least one overlap
- Object detection
  - For each system, measure maximum sum overlap with other systems
  - Union/intersection — penalized by evaluation metric

# Results

- 2015 CSSF — 10 shared systems

| Approach | Precision | Recall | F1 |
|---|---|---|---|
| ME (Jacobs et al., 1991) | 0.479 | 0.184 | 0.266 |
| Oracle voting (>=3) | 0.438 | 0.272 | 0.336 |
| Top ranked system (Angeli et al., 2015) | 0.399 | 0.306 | 0.346 |
| Stacking | 0.497 | 0.282 | 0.359 |
| Stacking + instance features | 0.498 | 0.284 | 0.360 |
| Stacking + provenance features | **0.508** | 0.286 | 0.366 |
| SWAF | 0.466 | **0.331** | **0.387** |

# Results

- 2015 TEDL — 6 shared systems

| Approach | Precision | Recall | F1 |
|---|---|---|---|
| Oracle voting (>=4) | 0.514 | **0.601** | 0.554 |
| ME (Jacobs et al., 1991) | 0.721 | 0.494 | 0.587 |
| Top ranked system (Sil et al., 2015) | 0.693 | 0.547 | 0.611 |
| Stacking | 0.729 | 0.528 | 0.613 |
| Stacking + instance features | 0.783 | 0.511 | 0.619 |
| Stacking + provenance features | 0.814 | 0.508 | 0.625 |
| SWAF | **0.814** | 0.515 | **0.630** |

# Results

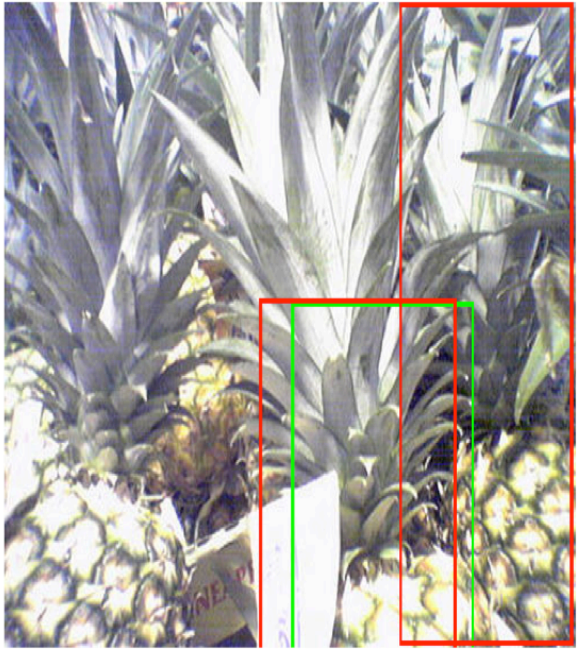
- 2015 ImageNet object detection— 3 shared systems

| Approach | Mean AP | Median AP |
|---|---|---|
| Oracle voting (>=1) | 0.366 | 0.368 |
| Best standalone system (VGG + selective search) | 0.434 | 0.430 |
| Stacking | 0.451 | 0.441 |
| Stacking + instance features | 0.461 | 0.45 |
| Mixtures of Experts (Jacobs et al., 1991) | 0.494 | 0.489 |
| Stacking + provenance features | 0.502 | 0.494 |
| SWAF | **0.506** | **0.497** |

# Results on object detection

object category: ping-pong ball

object category: pineapple

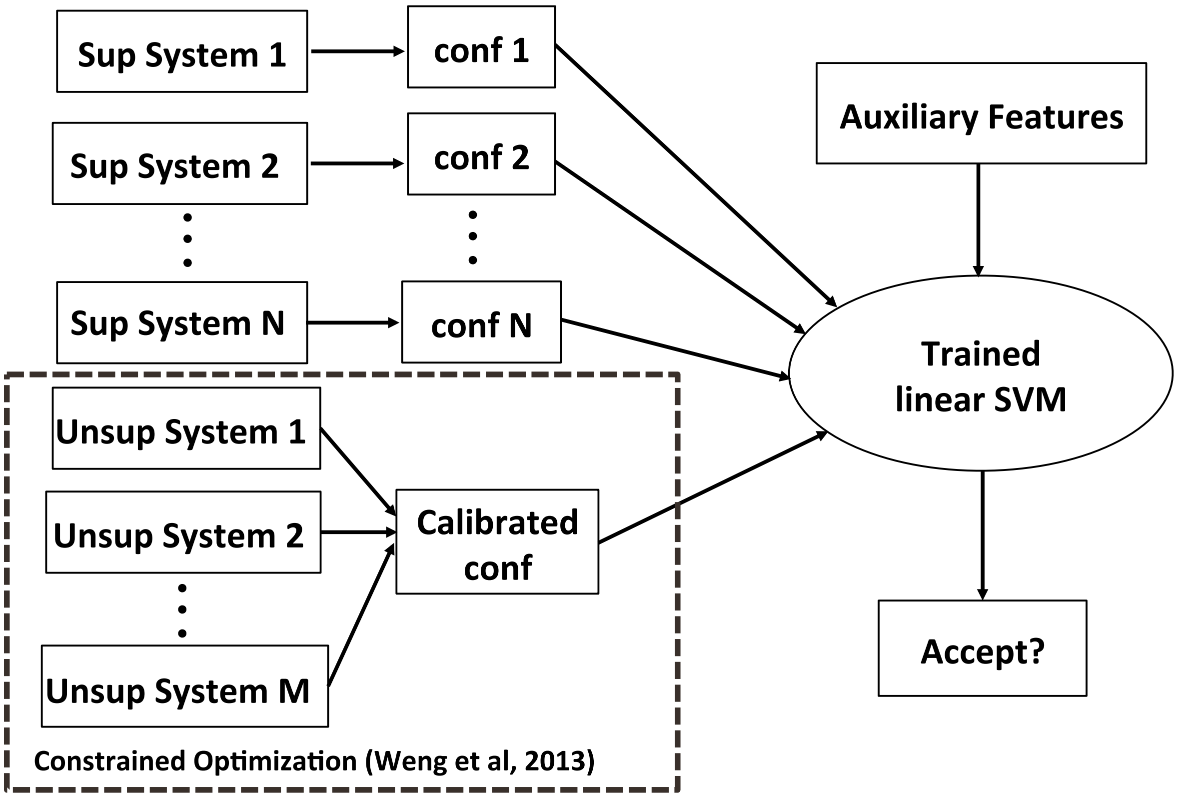


# Takeaways

- SWAF produced SOTA on CSSF and TEDL; significant improvements on object detection
- Our approach is more robust than ME in terms of number of component systems
- Works well for images with multiple instances of the same object

# Completed Work:
# III. Combining Supervised and Unsupervised Ensembles for Knowledge Base Population (EMNLP2016)

# Combining supervised & unsupervised ensembles



# Constrained Optimization

(Wang et al., 2013)

- Approach to aggregate raw confidence values
- Re-weight the confidence score of an instance
  - number of systems that produce it
  - rank of those systems
- Uniform weights for all systems
- Our work extends to entity linking

# Results

- 2015 CSSF —#sup systems=10, #unsup systems=13

| Approach | Precision | Recall | F1 |
|---|---|---|---|
| Constrained optimization | 0.1712 | 0.3998 | 0.2397 |
| Oracle voting (>=3) | 0.4384 | 0.2720 | 0.3357 |
| Top ranked system (Angeli et al., 2015) | 0.3989 | 0.3058 | 0.3462 |
| SWAF | 0.4656 | 0.3312 | 0.3871 |
| BGCM for combining sup + unsup | 0.4902 | 0.3363 | 0.3989 |
| Stacking for combining sup + unsup (BGCM) | **0.5901** | 0.3021 | 0.3996 |
| Stacking for combining sup + unsup (constrained optimization) | 0.4676 | **0.4314** | **0.4489** |

# Results

- 2015 TEDL —#sup systems=6, #unsup systems=4

| Approach | Precision | Recall | F1 |
|---|---|---|---|
| Constrained optimization | 0.176 | 0.445 | 0.252 |
| Oracle voting (>=4) | 0.514 | 0.601 | 0.554 |
| Top ranked system (Sil et al., 2015) | 0.693 | 0.547 | 0.611 |
| SWAF | **0.813** | 0.515 | 0.630 |
| BGCM for combining sup + unsup | 0.810 | 0.517 | 0.631 |
| Stacking for combining sup + unsup (BGCM) | 0.803 | 0.525 | 0.635 |
| Stacking for combining sup + unsup (constrained optimization) | 0.686 | **0.624** | **0.653** |

# Takeaways

- Many high ranking systems w/o training data
- Approximately 1/3 of possible outputs produced by unsupervised ensemble
- Combination improves recall substantially

# Proposed Work:
# I. Short-term proposals — Semantic Instance-level Features

# Instance-level features

- Completed work included only superficial instance features
- Focus more on the instance features — task specific
- Specifically, more semantic features
- Based on the results, these features:
  - help improve performance by themselves,
  - used along with provenance

# EDL instance-level features

(Francis et al., 2016)

- Used contextual information to disambiguate entity mentions using CNNs for EDL
- Computes similarities between a mention's source document and its potential entity targets at multiple granularities.
- CNNs: text block → topic vector

# EDL instance-level features

- Example source and target granularities for an instance in the 2016 NIST KBP dataset.

**Mention**
Hillary Clinton

$s_{mention}$

**Context**
For a host of reasons, Hillary Clinton somehow has failed to develop the rhetorical or interpersonal skills that made her husband, and Barack Obama, so appealing on the campaign trail.

$s_{context}$

**Document**
I don't disagree with the gist of Dowd's article. For a host of reasons, Hillary Clinton somehow has failed to develop the rhetorical or interpersonal skills that made her husband, and Barack Obama, so appealing on the campaign trail. Clinton has also, for reasons good and bad, made a number of errors in judgement in her run-up to her current campaign...

$s_{document}$

$t_{title}$

**Title**
Hillary Diane Rodham Clinton

$t_{article}$

**Article**
Hillary Clinton is a former United States Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate. Before that, as the wife of President Bill Clinton, she was First Lady from 1993 to 2001. In the 2008 election, Clinton was a leading candidate for the Democratic presidential nomination. A native of Illinois, Hillary Rodham was the first student commencement speaker at Wellesley College in 1969. She then earned a J.D. from Yale Law School in 1973. After a brief stint as a Congressional legal counsel, she moved to Arkansas and married Bill Clinton in 1975. Rodham cofounded the Arkansas Advocates for Children and Families in 1977..

# Object detection instance-level features

- ImageNet provides attributes dataset for certain categories
- Annotated with pre-defined sets of attributes:
  - **Color:** black, blue, brown, gray, green, orange, pink, red, violet, white, yellow
  - **Pattern:** spotted, striped
  - **Shape:** long, round, rectangular, square
  - **Texture:** furry, smooth, rough, shiny, metallic, vegetation, wooden, wet

# Proposed Work:
# I. Short-term proposals — Improve Foreign Language KBP

# Foreign language features

- This work will only apply to the KBP tasks
- Results on the 2016 TEDL task

| Language | Precision | Recall | F1 |
|---|---|---|---|
| English | **0.805** | **0.508** | **0.623** |
| Spanish | 0.79 | 0.443 | 0.568 |
| Chinese | 0.792 | 0.495 | 0.609 |
| Combined | 0.789 | 0.481 | 0.597 |

# Foreign language features

- TEDL - foreign language training data
- Auxiliary features do not translate to Chinese and Spanish
- Straightforward feature — language indicator
- Use language independent features
  - non-lexical

# Language Independent Entity Linking (LIEL) solution to TEDL

(Sil and Florian, 2016)

- Entity category PMI
- Categorical relation frequency
- Title co-occurrence frequency

# Proposed Work:
# II. Long-term proposals — Visual Question Answering

# Visual Question Answering (VQA)

(Antol et al., 2015)

- Understand how DNNs do object detection

# Visual Question Answering (VQA)

- VQA involves both language and vision
- Demonstrate SWAF on VQA
- Ensemble based on the answers
  - Multiple choice questions
  - Open ended answers — 90% one-word answers
- Use explanations as auxiliary features

# Proposed Work:
# II. Long-term proposals — Explanations as auxiliary features

# Explanation as auxiliary features

- Completed work focused on using provenance
- Captured "where" aspect of the output
- Recent work on generating explanations to interpret DNNs:
  - Towards Transparent AI systems (Goyal et al., 2016)
  - Generating visual explanations (Hendricks et al., 2016)
  - Visual Question Answering (VQA) (Antol et al., 2015)
- DARPA program for explainable AI (XAI)

# Explanation as auxiliary features

- Use explanations as auxiliary features
- Capture “why” aspect of the output
- Two types of explanations:
  - Textual
  - Visual

# Text as Explanation

(Hendricks et al., 2016)

- Generating visual explanations
- Jointly predict visual class and generate text as explanation
- Uses descriptive properties visible in the image

# Text as Explanation

**Input image**


**System A (Berkeley)**

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown

**System B**

This is a Kentucky warbler because this is a yellow bird with a short tail

# Text as Explanation

- Trust agreement between systems with similar explanations
- MT metrics — BLEU/METEOR for similarity
- Minimum Bayes Risk (MBR) decoding
- Embeddings of words in the explanation

# Images as Explanation

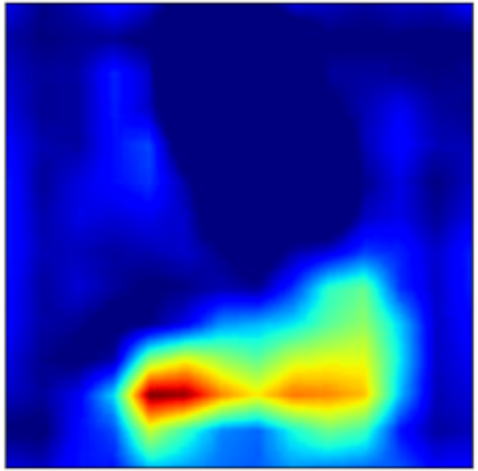
- DNNs attend to relevant parts of image while doing VQA (Goyal et al., 2016)
- Heat-map to visualize attention in images
- Humans trust systems with better explanations more even when they all predict the same output (Selvaraju et al., 2016)
- Enable the stacker to learn to rely on systems that "look" at the right region of the image while predicting the answer

# Images as Explanation

**Input image**

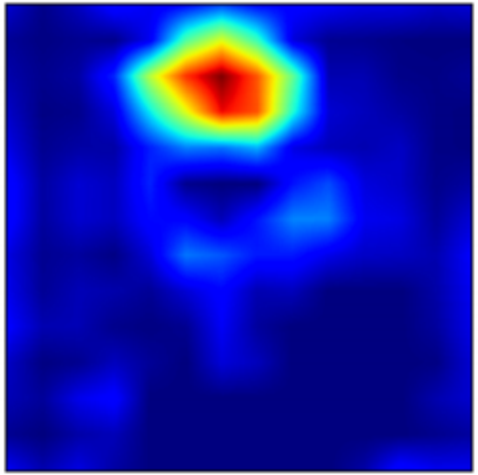
**Q: What color is the cat?**

**System A**



**A: Brown**

**System B**



**A: Brown**

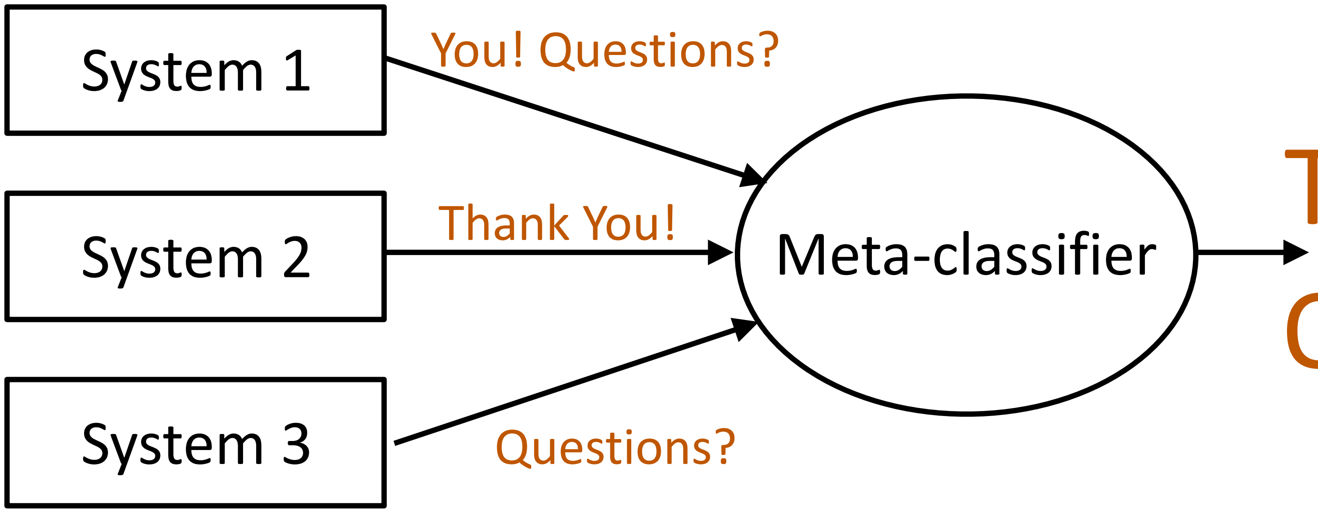
# Images as Explanation

- Use visual explanation to improve VQA
- Measure agreement between systems’ heat-maps
  - KL-divergence
  - Measure correlation
- Using visual explanation
  - improve performance
  - model with better explanations

# Conclusion

# Conclusion

- General problem of combining outputs from diverse systems
- SWAF on three difficult tasks
- Provenance captures “where” of the output
- Combining supervised and unsupervised ensembles improves recall
- Short-term: better auxiliary features
- Long-term: focus on “why” of the output




# Thank You!
# Questions?