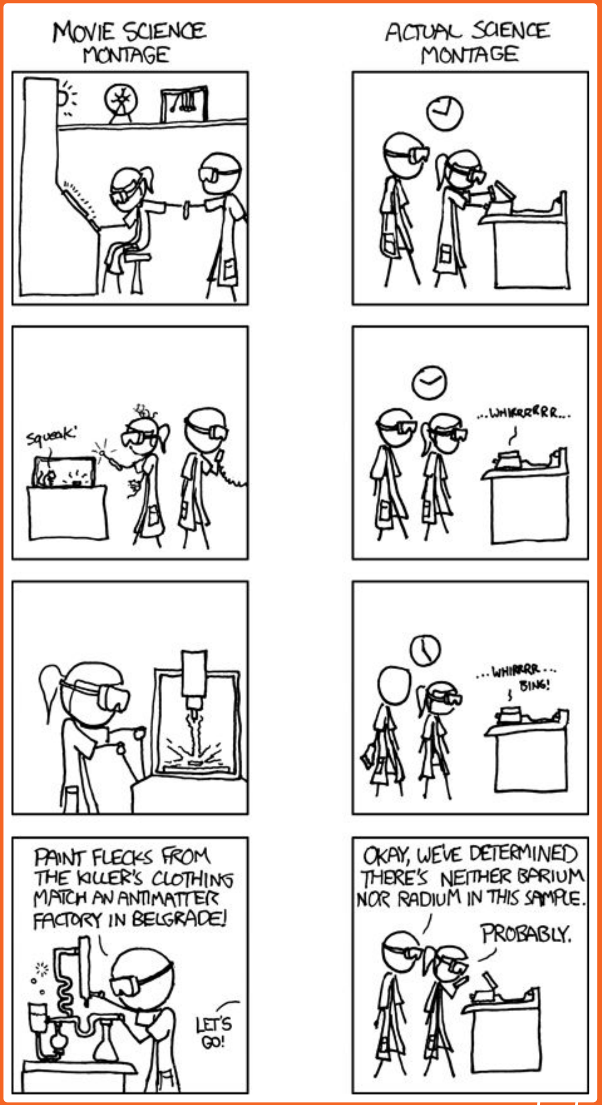
How to design experiments in NLU


MOVIE SCIENCE MONTAGE
ACTUAL SCIENCE MONTAGE
squeak!
...WHIRRRRRR...
...WHIRRRR...
BING!
PAINT FLECKS FROM THE KILLER'S CLOTHING MATCH AN ANTIMATTER FACTORY IN BELGRADE!
LET'S GO!
OKAY, WE'VE DETERMINED THERE'S NEITHER BARIUM NOR RADIUM IN THIS SAMPLE.
PROBABLY.

xkcd

Sam Bowman

NYU

(With pieces from Chris Potts and Bill MacCartney's Stanford CS 224U)

Today

The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876.
Vol. 7--No. 2.

- If you want to publish an influential paper in NLP, the engineering is only half the work.
- This is the first of three lectures about the other half:
  - How to find and understand related work on your problem
  - How to design effective experiments and analyze their results
  - How to make sure your work has the impact you want (and avoid ethical trouble)
  - How to write and publish your work

Today

The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876.
Vol. 7--No. 2.

- If you want to publish an influential paper in NLP, the engineering is only half the work.
- This is the first of three lectures about the other half:
  - **How to find and understand related work on your problem**
  - **How to design effective experiments and analyze their results**
  - How to make sure your work has the impact you want (and avoid ethical trouble)
  - How to write and publish your work

Today

- This is not specific to the class project, but almost all of it applies.
- For a successful project, you must:
  - Identify an open problem, present a hypothesis about it, and survey the relevant literature.
  - Design and run an experiment to test that hypothesis.
  - Analyze the results to reveal what your experiment tells us about your hypothesis.

The California Agriculturist AND LIVE STOCK JOURNAL.

SUBSCRIPTION PRICE, $1.50 a Year.
Single Copies, 15 Cents.

Vol. 7--No. 2. SAN JOSE, CAL., FEBRUARY, 1876.

A Valentine.

Oh, little loveliest lady mine,
What shall I send for your valentine?
Summer and flowers are far away;
Gloomy old Winter is king to-day;
Buds will not blow and sun will not shine;
What shall I do for a valentine?

Prithee, St. Valentine, tell me here,
Why do you come at this time o' year?
Plenty of days when lilies are white;
Plenty of days when sunbeams are bright.
But now, when everything's dark and drear,
Why do you come, St. Valentine dear?

I've searched the garden through and through,
For a bud to tell of my love so true;
But buds were asleep and blossoms were dead,
And the falling snow came down on my head.
So, little loveliest lady mine,
Here is my heart for your valentine!
—[St. Nicholas.

RALPH'S VALENTINE.

A Story as Aunt Abby Told It.

BY NELL VAN.

One rainy February day Dr. Pemberton's children, with their cousin George, were all gathered around the dining-room table preparing valentines for the approaching festivity of St. Valentine. No one could excel Cousin George, so the young folks thought, in the making of comic valentines. He had the faculty of producing burlesque likenesses of any one he knew, with bodies of birds or or beasts arranged in the most killing positions, and causing the most intense satisfaction among the young folks, while Bertha selected suitable verses from a book of promiscuous poems. The sedate Arthur was arranging sentimental valentines of gilt and perforated paper, profusely ornamented with decalcomanie birds and flowers and tinted tissue paper trimming. They were having a most delightful time, when, as is often the case, an altercation arose between the younger ones because George could not draw fast enough for Susy and Jano to color, and Tom declared it was his turn to have the next to paint, and so on, when Aunt Abby stepped in to quell the tumult.

She was a kind, motherly soul, without kith or kin in the world, but known everywhere as Aunt Abby. Whenever there was a murmur of discontent among the children, Aunt Abby invariably appeared upon the scene and took her station by the fireside with her everlasting knitting-work, and before she was aware of it she would be telling a story of bygone days to the listening crowd.

"Flutter, flutter! what's the buzz now little folks?" said she, cheerily, as she entered the room just as Arthur was trying to pacify the youngsters. "What's going on here, with all these picters and papers? Makin' valentines, do ye say? Well, well, when I was young"—

"Yes, that's right, Aunt Abby," said Susy, brightening up, for well she knew those to be the words used to preface a story.

"Yes, a story!" cried another of the group.

"We'll be good if you'll tell us one," said little Jane, quite penitently. Aunt Abby replied:

"Well, I was jest sayin', when I was young we never had nothin' of that sort for valentines. Real flesh and blood valentines were none too good for us in them days. On the mornin' of the fourteenth of February up we would jump and peep out the window, for the first lad we clapped eyes upon was to be our valentine for the year; and though we felt a little shy about naming it, yet somehow or other 'twas sure to come out, and there was no end to the fun we had at the spellin' match, quiltin' bee, or any of the frolics where we young folks were sure to meet."

"And was it the same rule for the boys, Aunt Abby?" adding "for in these days a fellow might watch out for half a day and never see a girl go past"

"Lor' bless you, no, child," said Aunt Abby. "The lads had much the best chance to choose their valentines. They used to go about the town, peeping up at the chamber windows of the gals they knew, or into open doors, a'most sure to see somebody they was lookin' for; and it was droll to see how gallant they'd be to the gal of their choice, or ill-mannered and rude if the wrong one appeared first."

"How well I mind the year that Master Ralph, then a lad of fourteen, passed the Winter at his Cousin Susan's house in Amherst. He had been left an orphan when a little boy, but was a prime favorite among his folks. He was always a studious lad, and so old-fashioned in his talk and ways that it was curious to watch and listen to him. There was quite a tumult in the house on Valentine's day mornin' that year, owin' to the arrival, durin' the night, of a first baby to Miss Susan, as I always called her. You see, when but a slip of a gal, I had gone to live at Miss Susan's mother's house, and after Miss Susan got married to a Colton, who was second cousin to my brother's wife's uncle, I became Aunt Abby to the hull tribe. About this time Brother Tom was took with the janders and died, leavin' me alone in the world, for his wife, Polly Ann, died of grief afore the year was out; so you see nothin' was left for me but to go out nursin'. I did amazin' well, and brought three families safely through the measles and mumps, and no end to the cases of chicken-pox and canker rash I was called in to tend. Well, I was with Miss Susan (Mrs. Colton, I should say) with this first baby, and as I was going down to breakfast that first mornin', who should I meet on the stairs but Ralph. He had heard the news, and seemed proper glad, and mighty anxious to take a peep at the little stranger.

"'Mayn't I see Cousin Susan and the baby, Aunt Abby?' says he to me.

"'Well, yes,' says I, 'certingly; you can walk right in and I'll show you the purty creatur.'

"How softly he stepped in! and after sayin' good mornin' to his Cousin Susan, he watched me pick up the little mite of a baby from its mother's side and lay it across my knee in front of the fire.

"'Oh,' said he, 'what tiny fingers and finger-nails, too!' Then comparing the length of the little hand with his own brown one, he said, 'Do you think mine were ever so small as these?'

"Then I uncovered the feet, and such a time as he made over the funny little toes, and he measured the length of its foot with his fingers, sayin', in his quaint fashion, 'Well, there will be a wife for somebody, eh, Aunt Abby?'

"'Who's your valentine to-day, Ralph?' asked his cousin, from the bed.

"'Why,' said the lad, 'I had not thought of that. This young lady here is the first one I've seen to-day. She must be my valentine this year, and I'll take a kiss to seal the compact;' so, pressing his lips to the soft baby cheek, he said, 'You'll be my valentine, eh, baby? silence gives consent, you see,' and with a murmur of applause from baby's mamma, Ralph kissed his hand to her, said 'bye-bye' to baby and was gone.

"Well, as the years passed on I was at Mrs. Colton's off and on, till that pretty baby had grown into a blooming lass, with dark, curling hair and eyes like night. They named her Clarindy, and she was right smart with the six younger brothers and sisters, who gave their mother a sight of trouble. The year that Clarindy was sixteen, the measles broke out among the children, followed by the whoopin'-cough, and nothin' would do but that Aunt Abby must be sent for to carry them through. Besides, Clarindy had a beau, and folks said she was to be married in the Spring. True, he was a'most double her age, bein' close on to thirty, and a medical man, having gone through his studies, graduated and traveled some, but come to Amherst to settle into practice. I never mistrusted who the Doctor, as they called him, was, till one day I chanced to open the door for him, and there stood before me, as nat'ral as life, Master Ralph! only the bits of whisker half-way down his cheeks made him appear more manly looking. His eyes lit up when he see me and he said:

"'Why Aunt Abby, are you here? Do you remember me?'

"'Indeed I do, Master Ralph,' says I, with clear wonderment.

"'Is Clara in?' he asked; and says I to him, 'Well, yes; she's fixin' to go ridin' with'—

"Before I could say more one of the young ones came out, and seeing who it was, sung out, 'The doctor's come! The doctor's come! Now mayn't I ask him to let me go too?'

"And this, then, was Clarindy's lover; the same boy, Ralph, grown into a han'som' doctor, a goin' to claim for his wife the valentine of his boyhood! Oh, children, truth is stranger than fiction every time."

"Tell us more," said little Jane.

"What became of them?" asked sedate Bertha.

"Do tell us more, dear Aunt Abby," said Susie, with enthusiasm.

"Well, children," slowly resumed Aunt Abby, "the last time I see any of the doctor's family was one rainy day in February, when I stepped in to quiet the children, who were skirmishin' over the makin' of valentines."

"Just like me," said little Jane.

"So," continued the old lady, "I dropped into a chair with my knittin' work and set out to tell them of bygone days and the true story of Dr. Ralph Pemberton's valentine."

"Oh, that's a good one," said George. "I knew all along it was Uncle Ralph and Aunt Clara you were telling about."

"Was it Papa and Mamma, really, truly, Aunt Abby?" asked little Jane, while Bertha and Sue uttered little exclamations of surprise.

"Of course it's true, simpleton," spoke Arthur. "Don't you know Mamma's birthday is on St. Valentine's day, and her name is Clara while Pa's is Ralph?"

"Why no," said the persistent little Jane; "Papa is only Pa and Doctor, and Mamma is—well, Papa calls her Deary and we call her Mamma, you know."

FAVORABLE NOTICES.

We are under obligations to the local press generally over the State for many flattering notices of our new volume. We make room for but one this month, and give that because it comes wholly unexpected from one of our home papers. We value the notice all the more highly for, as a general rule, there is a sort of jealousy existing between papers published in the same town, which prevents the expression of a good will. But the *Patriot*, under its present management, is able to be independent of and above petty meanness, and is conducted in a manner to gain the respect and confidence of its readers.

"THE CALIFORNIA AGRICULTURIST.—The first number of the seventh volume of this magnificent periodical is before us, and we are proud of it. THE CALIFORNIA AGRICULTURIST AND LIVE STOCK JOURNAL, to use its proper title, is a publication which reflects honor on the taste, intellect and progressive spirit of San Jose. It is a magazine, no more of the special interests indicated by its name, than a journal of general art, knowledge, literature and fine miscellany. Thus, in the splendid number on our table we find, besides a variety of able editorials, whole departments well filled with essays on many important subjects, such as gardening, hygiene, stock raising, education, pisciculture, etc. A history of the State Normal School (with illustrations) will be found, together with a week's reading of the most interesting matter for old and young, for the farmer, artisan, scholar, housekeeper; in fact it is a magazine for everybody, and better worth the subscription price—$1.50—than any similar work in the country. Office over the San Jose Savings Bank.—*San Jose Patriot, Jan.18.*

STONE jars which have become offensive and unfit for use, may be rendered perfectly sweet by packing them full of earth and letting them stand two or three weeks.

Reading Related Work

The literature review

Google Scholar

- Do this early!
- Why?
  - Make sure that what you're doing hasn't already been done before.
    - If it has, and the paper is easy to find, you won't get full credit.
  - Learn about common methods, datasets, and libraries that will make your life easier.
  - Buy yourself more time to think about the questions that *haven't* been answered in the literature.

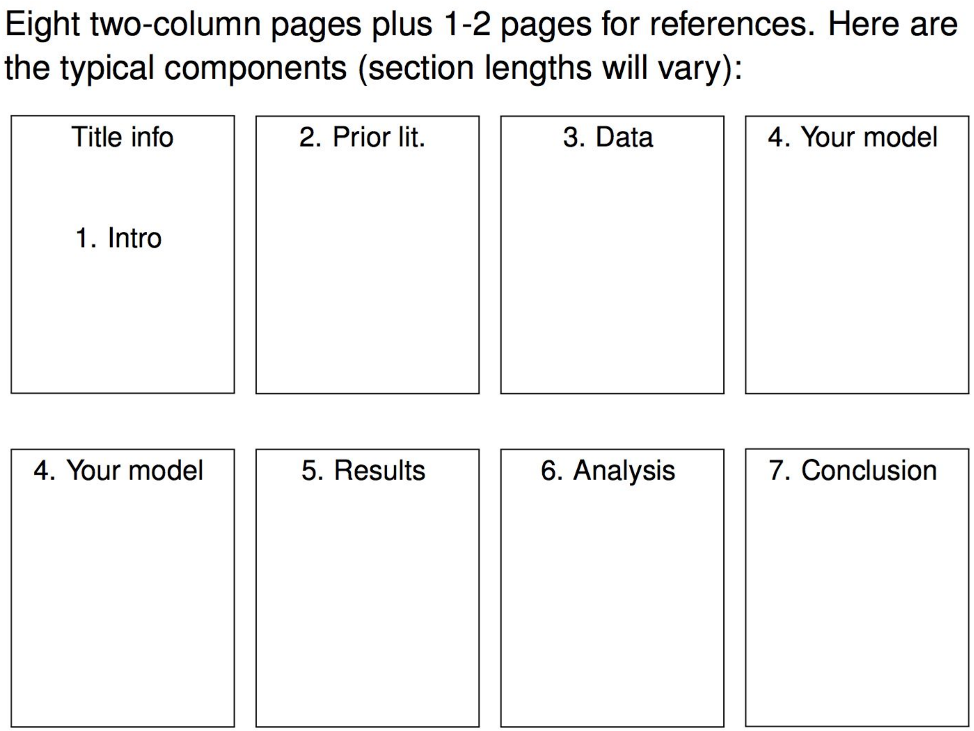
Identifying papers

1. Do a search on Google Scholar (or the ACL Anthology)
2. Download the papers that seem most relevant
3. Skim the abstracts, intros, & previous work sections
4. Identify papers that look relevant, appear often, & have lots of citations on Google Scholar
5. Download those papers 6.
6. Return to step 3

Credit: Bill MacCartney

Anatomy of an NLP paper

Google Scholar



Eight two-column pages plus 1-2 pages for references. Here are the typical components (section lengths will vary):

Title info 1. Intro	2. Prior lit.	3. Data	4. Your model
4. Your model	5. Results	6. Analysis	7. Conclusion

- Your class project gets only four pages, but it only needs to contain one experiment.

The literature review

- Where to find the most trustworthy papers:
  - NLP and Computational Linguistics:
    - Proceedings of ACL conferences (ACL, NAACL, EACL, AACL, EMNLP, CoNLL)
    - TACL
  - Machine Learning/AI:
    - Proceedings of NeurIPS, ICML, ICLR, AAAI, IJCAI
  - Crowdsourcing:
    - Proceedings of CHI, HComp
- See also: arXiv*

* Official (reasonable) ACL policy: arXiv papers *are* prior work and should be in literature reviews and comparison tables, but you can still treat claims from those papers more skeptically.

The literature review

Google Scholar

- You’ll have to use your judgment in deciding which papers to read, and which papers to trust.
  - Papers can disagree with one another.
  - You might just find too many papers.

The literature review

- Which papers will be most useful?
  - Newer ones, especially if they cite the older papers that you're interested in.
    - The newer paper might contain a good summary of the older one!
  - Papers published in top conferences and journals, rather than arXiv papers or papers published elsewhere.
    - Reviewers have carefully looked at these papers for mistakes or inconsistencies.
  - *Published* papers with negative results (method X *doesn't work*, method X *doesn't* do what you think it does, ...), rather than papers with positive results.
    - Negative results are usually held to a higher standard in for publishing.

The literature review

Google Scholar

- Finding negative results
  - Negative results are often reported as brief mentions in a paper that has mostly positive results.
    - A paper on transformers for summarizations might meniton that CNNs didn't work for them.
    - Often there will be real results on this!
    - ...but these results often won't appear in the abstract.
  - Look for *gaps in the literature.*
    - Generative adversarial networks (GANs) are a huge topic in machine learning, with thousands of papers each year.
    - NLP is a major area of machine learning, with thousands of papers a year.
    - There are very few papers on GANs for NLP. Why?

Organizing a Project
& Doing Good Science

The Hypothesis

- Bad/unfalsifiable:
  - *Neural networks are more elegant and principled than feature-based systems for sentiment analysis.*
- Falsifiable but uninformative:
  - *My convolutional neural network (CNN) model outperforms the Socher et al. (2015) baseline on the Stanford Sentiment Treebank data.*
- Typical good paper:
  - *CNN models outperform feature-based systems on sentiment analysis for reviews.*
- Ambitious paper:
  - *Non-professional web users tend to express sentiment using a set of common* fixed phrases *when they discuss products. These phrases vary enough that simple symbolic features don't capture them well, but they're structured enough that filter-based neural networks like CNNs can capture them very efficiently.*

Controlled experiments

Your experiment needs to test your hypothesis. Everything that you've learned about *controlled* experimentation applies in NLP.

Example results on SNLI:

- Paper A: LSTM gets 77.6%
- Paper B: CBOW gets 80.6%

What can you conclude?

Controlled experiments

Your experiment needs to test your hypothesis. Everything that you’ve learned about *controlled* experimentation applies in NLP.

Example results on SNLI:

- Paper A: LSTM gets 77.6%
- Paper B: CBOW gets 80.6%
- Paper C: GRU gets 80.7%

What can you conclude?

Controlled experiments

Your experiment needs to test your hypothesis. Everything that you've learned about *controlled* experimentation applies in NLP.

Example results on SNLI:

- Paper A: LSTM gets 77.6%
- Paper B: CBOW gets 80.6%
- Paper C: GRU gets 80.7%
- Paper C: BiLSTM gets 84.5%
  - With the same input preprocessing, hyperparameter tuning method, etc.

What can you conclude?

Baselines

- *Baseline* can mean any of three things, and you usually want all three:
  - Anyone's performance numbers for *simplest reasonable approach* to your problem (CBOW, logistic regression, plan seq2seq).
  - *Your* numbers for a reasonably competitive system based on existing ideas. This can be based on a published system or your own work, but you should make a precise, controlled comparison with it.
  - The *best published number* for your problem. It's common to not beat this, but you should compare with it. This is what could give you the right to claim the 'state of the art'.

Upper bounds

- Less necessary, but often helpful: Measure (or find out) the accuracy of a human annotator!
  - You can’t use the same annotations that you used to create the data.
  - It’s better if you don’t even use the same people. (Why?)
- 77% accuracy doesn’t look so bad if a human only gets 81%!

Datasets

Datasets

- You’ll want real natural language data for any NLP project. For a publication, you’ll usually want multiple datasets.
- You’ll almost always want to use at least one dataset that appeared in related prior work, even if you’re the first one to solve a specific problem.
  - Working on sentiment analysis in Portuguese?
    - Show that your method is competitive on English datasets like SST or IMDB.
    - Collect or create a dataset for Portuguese.
    - Show results on the new data with your method and several existing baselines.

Getting datasets

- Find them!
  - The ACL anthology!
  - The Linguistic Data Consortium (free membership through NYU)!
- Build them!
  - Write *very* detailed notes on what you do and why: Readers and reviewers will assume that all of your decisions are biased to that unfairly favors the systems you're interested in. To convince them that you didn't, you'll need to show that you made fair and reasonable decisions.
- Scrape them?
  - It's easy to break the law this way. Or get NYU temporarily banned from Twitter/eBay/Reddit/Tinder/etc.

Getting datasets

- Have someone else build them!
  - Write *simple* annotation guidelines that non-NLPers can follow.
    - You still need to convince your reviewers these guidelines aren't unfairly advantaging your method.
  - Pay some friends or Mechanical Turk (MTurk) workers to follow your guidelines.
  - This is often quick and cheap in practice. Ballpark costs (at $10–20/hr):
    - Writing: ~$0.25–$1 per sentence written
    - Labeling: ~$0.05–$0.5 per sentence read
  - Some advice:
    - Try your annotation task yourself for half an hour.
    - Communicate with your annotators! Even on MTurk, they'll often be trying hard to understand your task, and you want to be available if they have questions or if they see issues.

Getting datasets

- Have someone else build them!
    - For this class:
        - You’re welcome to use MTurk *if* you have access to research funding (mostly PhD students).
        - Otherwise, consider trading data with another team, and having the members of the other team annotate your data.
        - Don’t tell them too much about your project—just share the annotation guidelines.

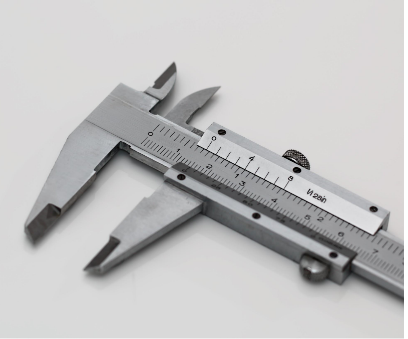
Artificial datasets

- Artificial (machine-generated) datasets can allow you to run preliminary experiments in minutes, minimizing the risk that you'll spend hours or days running something that won't work.
- They can also help you with error analysis, by revealing what kinds of pattern your system can/can't learn.
- Artificial datasets ***won't*** convince NLP readers that your system works on language.
  - Case study BaBi QA:

```
1 Mary moved to the bathroom.
2 John went to the hallway.
3 Where is Mary?          bathroom
```

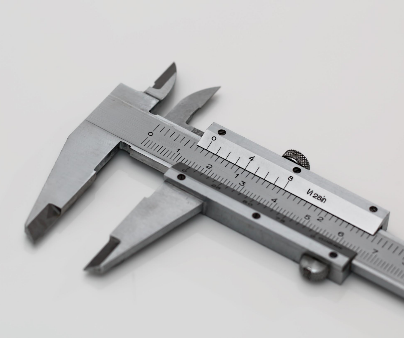
Organizing a Project & Doing Good Science

Quantitative evaluation



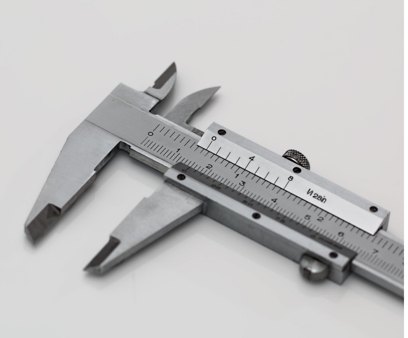
- This is a huge topic, but in short:
  - Follow prior work ***precisely*** in how you choose and implement your main evaluation metric.
  - *Do* show metrics as many variants of your model as you can (ablation analysis).
    - Example:
      - **CNN sentiment classifier with GloVe embeddings**
      - CNN sentiment classifier with random embeddings
      - LSTM sentiment classifier with GloVe embeddings
      - CBOW classifier with GloVe embeddings
      - CBOW classifier with random embeddings
    - Should you use the dev set or the test set?
  - *Do* use carefully-designed human evaluations for tasks where this is standard (dialog, summarization, etc.).

Quantitative evaluation



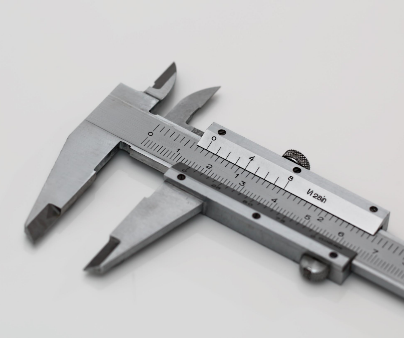
- This is a huge topic, but in short:
    - *Do* invent new analysis metrics if they help you make your point.
        - If your system good at classifying long sentences, also report accuracy on the subset of sentences of length >20.
        - If your baseline for a text generation task only uses very common words, and your system fixes that, measure the *average frequency* of the words that each system generates.
        - *Don't* talk about the 'state of the art' for things that nobody else measures.
    - *Do* perform extrinsic evaluations on downstream tasks if you expect the output of your system to be used as the input to any other system.

Quantitative evaluation



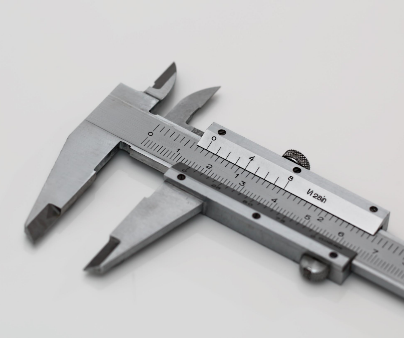
- This is a huge topic, but in short:
  - *Do* explicitly test for statistical significance, especially when your hypothesis depends on a small difference or when model performance is highly variable. (See Berg-Kirkpatrick et al. '12)
    - Methods here vary widely within NLP, but don't expect readers to take you seriously if the main claim of your paper is that you get an 0.2% accuracy improvement on the state of the art.

Quantitative evaluation



- If your system doesn't beat your baselines, or if the differences between your results aren't significant: Say so! And explain what you found!
    - Well-presented negative results do move the field forward, and are in common in projects like class papers that have to happen in a short time.

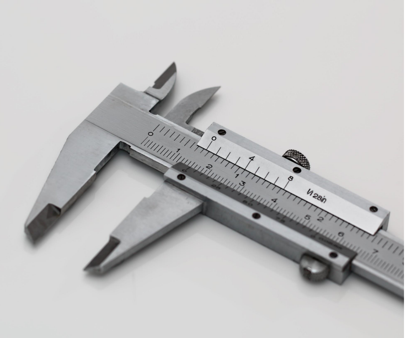
Qualitative evaluation and error analysis



- NLP papers generally have an *analysis* section.
  - This may be called something else!
- Your goal here: Convince the reader of your hypothesis.
- If your hypothesis is interesting, it'll be hard to evaluate with standard/intuitive quantitative metrics.
  - *Non-professional <u>web users tend to express sentiment using a set of common fixed phrases</u> when they discuss products. These phrases vary enough that simple symbolic features <u>don't capture them well</u>, but they're structured enough that filter-based neural networks like <u>CNNs can capture them</u> very efficiently.*

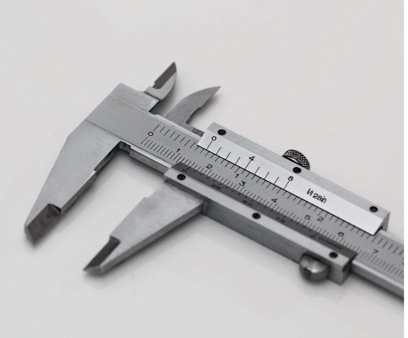
Qualitative evaluation and error analysis

- So, give the reader qualitative evidence, too!
- Places to start:
  - Look to prior work, and do what they do.
  - Show examples of system output.
  - Come up with *categories* to describe system errors and count them.
  - Visualize the internal state of your model with tools like LSTMVis or BertVis.
  - Plot how your model performance varies with amount of data.
  - Build an online demo!



Formative vs. summative evaluation

When the cook tastes the soup, that's formative; when the customer tastes the soup, that's summative.



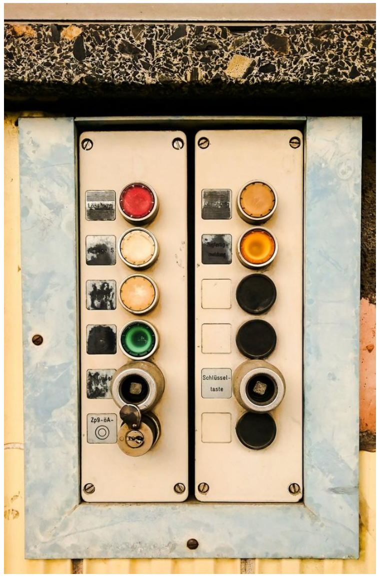
- Formative evaluation: guiding further investigations
  - Typically: lightweight, automatic, intrinsic
  - Compare design option A to option B
  - Tune hyperparameters: smoothing, weighting, learning rate
- Summative evaluation: reporting results
  - Compare your approach to previous approaches
  - Compare different *major* variants of your approach
  - Only use the test set here
  - Generally only bother with human or extrinsic evaluations here
- Potential serious mistake: Don't save all your qualitative evaluation for the summative evaluation!

Poll:

Formative and summative evaluation.

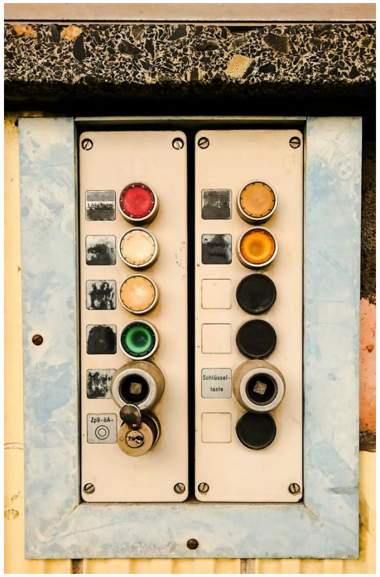
Misc.

Hyperparameter tuning



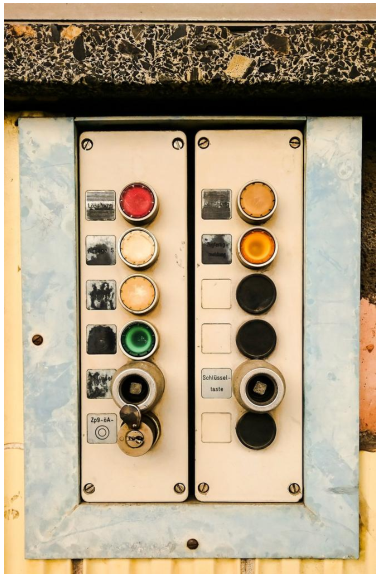
- **Crucial point:**
  - You must tune the hyperparameters of your baselines just as thoroughly as you tune them for any new model you propose!
  - Failure to do this invalidates your comparisons, and depending on how you write the paper, could border on academic misconduct. (Related reminder: Don't tune on the test set!)
- Read the fine print while you're doing your literature review to get a sense of what hyperparameters to worry about and what kinds of value to expect.
- If you're not sure whether to tune a hyperparameter, you probably should.

Hyperparameter tuning



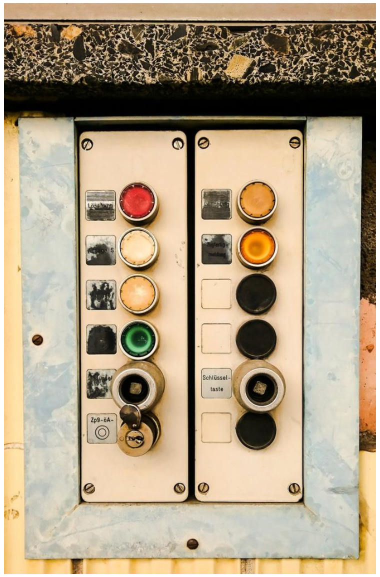
- Grid search: Inefficient.
- Bayesian optimization: Optimal, but public packages can be difficult to use.
- Random search (Bergstra and Bengio ‘12): Easy, and near-optimal.

Hyperparameter tuning

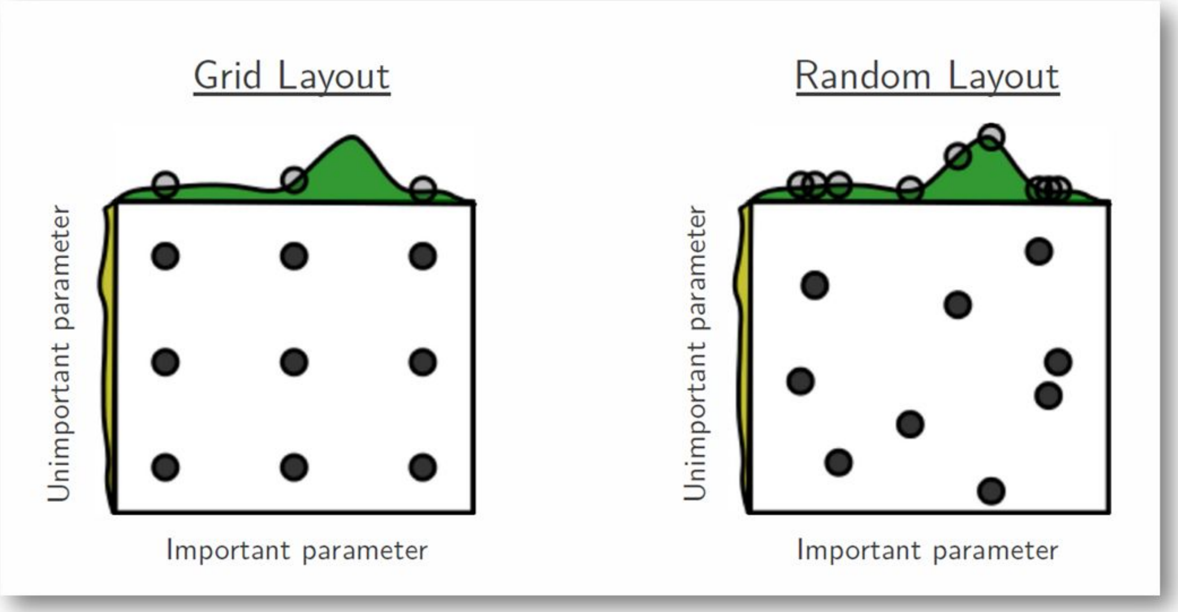


- Random search (Bergstra and Bengio '12):
  - Define distributions over all your hyperparameters.
  - Sample N times for N experiments.
  - Look for patterns in your results.
  - Adjust the distributions and repeat until you run out of resources or performance stops improving.

Hyperparameter tuning



- Random search (Bergstra and Bengio ‘12):
- Why?



Iterative development

- As soon as you start:
    - Create a git repo for your project.
    - Find or build code to load your data.
    - Find (try not to build) code to evaluate results.
    - Find or build a *very simple* baseline.
- After that, keep track of:
    - Commands and git checkpoint identifiers for each of your experiments.
    - Saved model checkpoint files for all reasonably effective/interesting experiments.
    - Notes on what each experiment was meant to test.
- Done this way, research is an *anytime algorithm*.
    - All of these are good practices for independent reasons, and they ensure that you'll be able to write a paper on time even if your more ambitious plans don't work out.

Recommended reading

- Resnik and Lin on evaluation
- Gardner slides on engineering
- Jason Eisner’s advice posts on NLP/ML research