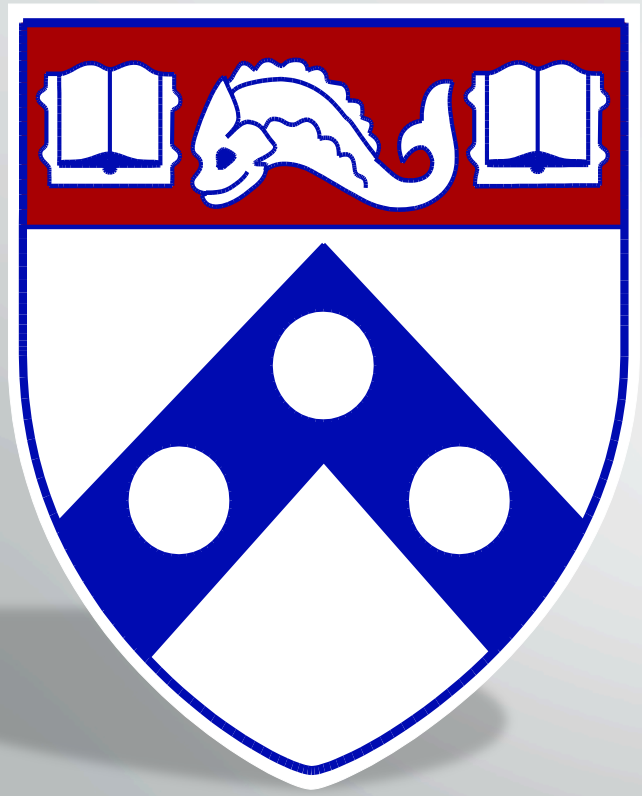


# Writing a compelling conference/journal paper

Susan B. Davidson

CIS 700: Advanced Topics in Databases

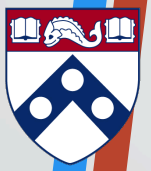
MW 1:30-3

Towne 309

Material drawn from Black, Leen & Maier course on Scholarship Skills.

# Rules for better writing

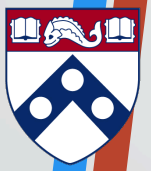
- Good writing is clear, concise and simple.
- Good writing is easy to read.  A good paper educates the reader without frustrating them.
- **Write to be understood, not to impress.**
- Note that to do this you have to know who the audience is.



# How do people read papers?

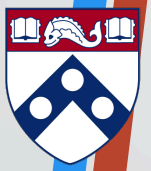
- **Title:** to decide if the paper is potentially relevant.
- **Abstract:** to determine relevance and contributions of paper.
- **Pictures:** tables, graphs, diagrams to understand concepts.
- **References:** do I recognize them or know what they are about?

- **Introduction, Section beginnings, Examples, Conclusions: to determine organization and content.**
  - Might decide on the basis of this to read only parts of the paper.
  - Decide whether or not the authors are good writers.
  - Help decide whether the paper is worth reading more carefully.



# How to get a paper rejected (quickly)

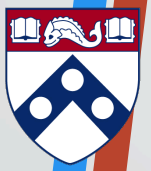
- Distract the referee with misspellings, bad grammar and poor layout.
- Make stupid mistakes: forget to turn off “screams” in your paper, don’t generate the references (or forget to include author names or title)
- Don’t give any examples and use overly technical notation
- Don’t include one of the standard sections (e.g. related works, experiments)
- Don’t explain the problem or solution anywhere in the abstract or introduction
- Require the referee to read two other papers to understand this one.



# These things help too…

- Fail to cite relevant work of the program committee.
- Submit a paper that is all text, without figures or tables.
- Include absolutely every experiment or bit of data that you gathered.
- Other stupid mistakes:  ignore the page limit or formatting instructions, miss the deadline.

# 1. Format  of journal or conference paper.



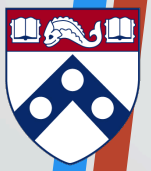
# Title

- Make the title precise:

Some problems on graphs

Finding Hamiltonian circuits in directed graphs

Parallel algorithms for finding Hamiltonian circuits in directed graphs



# Title, cont.

- State your result if you have one

A complexity result for coding

**Better:** Maximal prefix compression is NP-complete

- Use an action verb if you can

Techniques for agent implementation

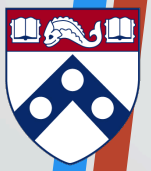
**Better**: Implementing agents in Java

- Sometimes a witty title can be effective.

Nineteen dubious ways to compute the exponential of a matrix

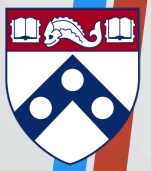
- But avoid cliches

Agents considered harmful



# Titles from this semester

- Answering queries using views: A survey
- Big data analytics with Datalog queries on Spark
- Data citation: Giving credit where credit is due
- Discretized streams: Fault-tolerant streaming computation at scale
- PrivateClean: Data cleaning and differential privacy



# Authors

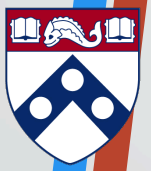
- Pick a professional name and stick with it

Susan B. Davidson

Susan Davidson

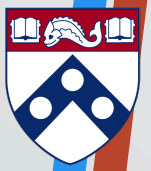
S.B. Davidson

- Order of authors
  - Alphabetical is common in CS when contribution is roughly equal
  - Sometimes authors are grouped, then alphabetical within each group
  - Discuss it when you start the paper!

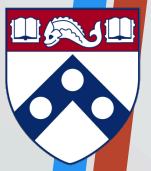


# Abstract

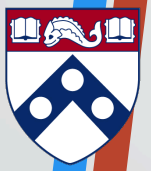
- The abstract should be a “mini-paper”, stating the motivation, problem, approach and results.
  - Be specific
  - No citations
  - Don’t make it a condensation of the introduction
- It explains the whole paper
  - Recall that readers often “shop” by looking at the abstract
- Its primary purpose is to pull people into your paper
  - Get it into the “A” pile of a PC member.
  - Entice people to read the full paper.



Analysts often clean dirty data iteratively–cleaning some data, executing the analysis, and then cleaning more data based on the results. We explore the iterative cleaning process in the context of statistical model training, which is an increasingly popular form of data analytics. We propose ActiveClean, which allows for progressive and iterative cleaning in statistical modeling problems while preserving convergence guarantees. ActiveClean supports an important class of models called convex loss models (e.g., linear regression and SVMs), and prioritizes cleaning those records likely to affect the results. We evaluate ActiveClean on five real-world datasets UCI Adult, UCI EEG, MNIST, IMDB, and Dollars For Docs with both real and synthetic errors. The results show that our proposed optimizations can improve model accuracy by up-to 2.5x for the same amount of data cleaned. Furthermore for a fixed cleaning budget and on all real dirty datasets, ActiveClean returns more accurate models than uniform sampling and Active Learning.



Analysts often clean dirty data iteratively–cleaning some data, executing the analysis, and then cleaning more data based on the results. ~~We explore the iterative cleaning process in the context of statistical model training, which is an increasingly popular form of data analytics.~~ ~~We propose~~ ActiveClean, which allows for progressive and iterative cleaning in statistical modeling problems while preserving convergence guarantees. ActiveClean supports an important class of models called convex loss models (e.g., linear regression and SVMs), and prioritizes cleaning those records likely to affect the results. We evaluate ActiveClean on five real-world datasets UCI Adult, UCI EEG, MNIST, IMDB, and Dollars For Docs with both real and synthetic errors. The results show that our proposed optimizations can improve model accuracy by up-to 2.5x for the same amount of data cleaned. Furthermore for a fixed cleaning budget and on all real dirty datasets, ActiveClean returns more accurate models than uniform sampling and Active Learning.



Analysts often clean dirty data iteratively–cleaning some data, executing the analysis, and then cleaning more data based on the results.
ActiveClean is an iterative cleaning technique for statistical model training which preserves convergence guarantees. It supports an important class of models called convex loss models (e.g., linear regression and SVMs) by prioritizing cleaning those records likely to affect the results.
We evaluate ActiveClean on five real-world datasets (UCI Adult, UCI EEG, MNIST, IMDB, and Dollars For Docs) with both real and synthetic errors.
The results show that our method can improve model accuracy by up-to 2.5x for the same amount of data cleaned. Furthermore for a fixed cleaning budget and on all real dirty datasets, ActiveClean returns more accurate models than uniform sampling and Active Learning.

# Introduction

- The introduction should be accessible to a wider audience than the paper, and should catch the interest of the reader.
- A good first sentence is essential.
- Fit the work into the larger context of the field.
- Avoid introducing lots of notation and definitions in the introduction.
- Ends with a summary of sections, which can be folded into the contributions.

# Structure of an introduction

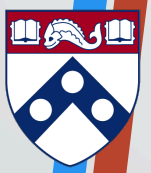
- **Problem statement**
- **Reading preparation**

# Problem statement

- *What* is the problem that you solve
- *Why* is the problem important
- What are the *benefits* and *characteristics* of a good solution
- What is the *current status* of the problem
- What *form* have previous solutions taken
- The *approach* you're taking and how it's motivated
- What the approach *accomplishes*

# Reading preparation

- Gives the "lay of the land" of the rest of the paper
  - Explains what the reader can expect the paper to accomplish
  - Lists the topics that the paper discusses, in the order in which they are presented
  - May be used by the reviewer for organizing their critique of the paper.



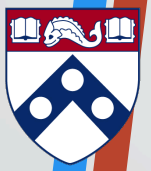
**Contributions** of this paper include:

(1) A semantics for citations to general queries using citation views based on covering sets of mappings between the views and the input query (Section 3.2).

(2) Three approaches to constructing citations for general queries, in which the reasoning progressively shifts from the tuple level to the schema level (Sections 4.1- 4.3). Two of the approaches generate citations to *individual tuples* in the query result, while the last approach generates a citation to the *entire* result. We also describe optimizations that were used in implementing each of the approaches.

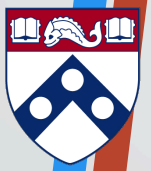
(3) Alternative *policies* for the joint, alternate, and aggregated use of citation views, and a description of how the policies are integrated into each of the approaches (Section 4.4).

(4) Extensive experiments using *synthetic* citation views and queries as well as *realistic* citation views and queries for two different choices of policies (Section 5). The experiments show the tradeoffs between the approaches in terms of (i) the time to generate citations as well as (ii) the size of the resulting citation. The experiments show that, for the syn-



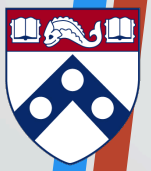
# Prior Work

- A discussion of prior work can sometimes be folded into the introduction
- However, a careful review of prior *and related* work has become increasingly important so typically this becomes its own section placed at the beginning or end of the paper.
- What constitutes relatedness?
  - Similar problem or approach
- Some conferences (e.g. SIGMOD) do not include references in the page limit to allow you to cite as much related work as possible.



# Structure of the rest of the paper

- Most common form: Problem-Solution-Defense
- **Experimental paper** (e.g. the Data Citation paper)
  - **Problem:** citing the results of a query over a data repository
  - **Solution:** specify citations for common queries (citation views) and use these to construct citation to general queries
  - **Defense:** performance measurements over real and synthetic data sets and queries.
- Theory paper (we didn't cover any in class so we'll make one up!)
  - Problem: need for increased concurrency
  - Solution: new locking protocol
  - **Defense**: correctness proof, show that it subsumes other techniques

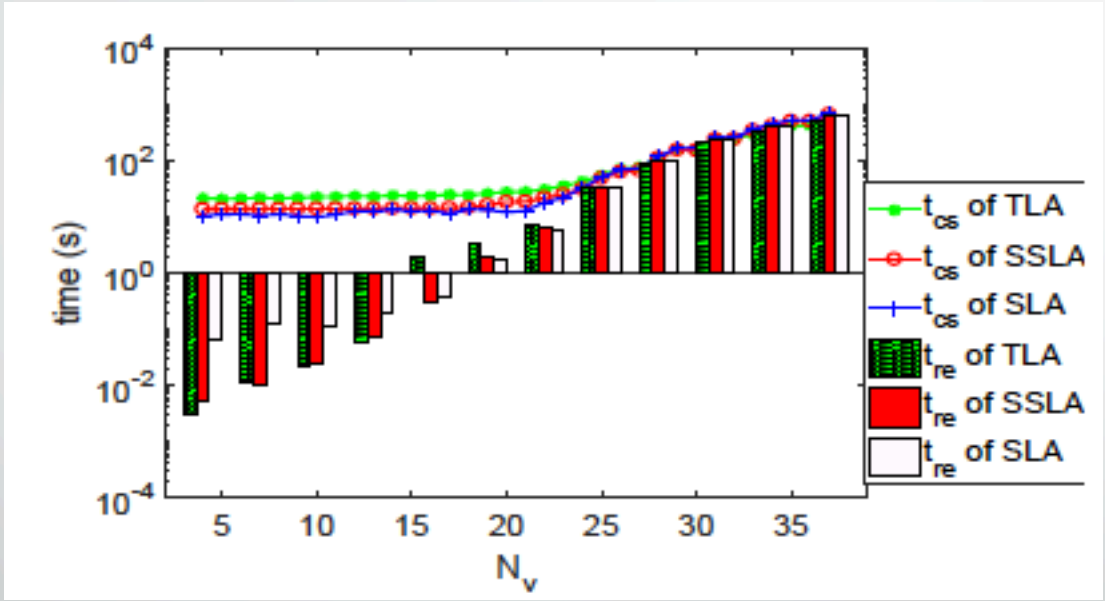


# Performance section

- State what the *purpose* of the performance study is
  - Sensitivity analysis
  - Model validation
  - Comparison of methods
  - Algorithm tuning
  - Determining range of utility
- Describe the *procedure*: environment, software used, test data, test procedure.
  - For replication
  - So people can understand why they got different results

# Figures and Tables

- Use them!  A picture is worth a thousand words
- Make them self-explanatory:  readers may be skimming
  - Use captions to explain what you might say in text
- Say what you want the reader to see in the figure or table
- Find the best way of capturing the information you wish to convey succinctly


time (s)
$10^4$
$10^2$
$10^0$
$10^{-2}$
$10^{-4}$
5
10
15
20
25
30
35
$N_v$
$t_{cs}$ of TLA
$t_{cs}$ of SSLA
$t_{cs}$ of SLA
$t_{re}$ of TLA
$t_{re}$ of SSLA
$t_{re}$ of SLA

# Discussion and Conclusion

- Sometimes discussion and conclusion are in a separate sections, sometimes they are merged.
- The **conclusion** restates major results and their implications.
  - Should point out any limitations of the work, and directions for extension (future work).
- The **discussion** is generally harder to write, and may include
  - Lessons learned, consequences
  - Applicability elsewhere
  - Follow-on work in progress, or perhaps published
- Take a step back and ask what you most want the reader to remember.
  - Speculation is OK

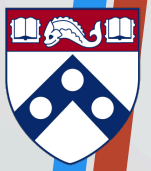
# Acknowledgements

- **Be generous.** Acknowledge:
  - Colleagues you've discussed the work with.
  - Non-authors who provided materials (data, graphics, simulators, code)
  - Referees who made comments and helped you improve the manuscript
  - Grant agencies supporting the work (include grant numbers)
  - Initiator of the idea

# 2. Tips for better writing

# Core rules for better writing

- Use **active** voice and verbs.
- Put **key ideas in lead position** of sections and paragraphs.
- Don't make unsubstantiated claims.
- Be **concise**.
- Be **simple**.
- Use a **consistent** terminology.
- **Define terms** when first used.
- Avoid single sentence paragraphs.
- **Rewrite** with an intent to make things simpler, more concise, and clearer.



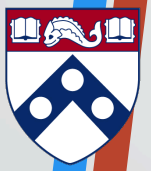
# Use the active voice

- The syntactic correctness of each command is checked by the parser.
- Subsequent look-up times are reduced by caching the directory nodes.
- The passive voice often results from avoiding personal pronouns (we, I, you).
- Learning effects were minimized by randomized order of tasks.

# Why use the active voice?

- It is more specific, and therefore more informative and clearer, than the passive voice (in the passive voice, the subject is often missing).
- It is frequently more concise than the passive voice.
- It is more direct and forceful.

- **If you do nothing more than concentrate on using the active voice, your writing will improve!**



# Use active verbs

- If we make a comparison of execution times with and without logging…
- The need to minimize user interaction is in contradiction with the requirement to provide many search options.

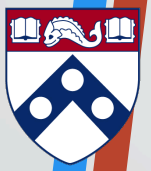
# Organize to help the reader

- Within *sections* of the paper (except the introduction):

**Put key ideas in lead position.** The introductory paragraph should summarize the key ideas in the section. The following paragraphs get more specific about the details of the key ideas. Thus the paragraphs move from the general to the more specific, from the most important to the least important.

- Within each *paragraph*:

**Put key ideas in lead position.** The first sentence in a paragraph should carry the most important ideas. The following sentences flesh out the particulars of the ideas. Thus sentences within a paragraph move from general to more specific.



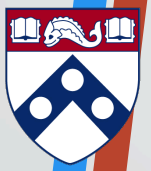
# Of course, there are exceptions…

- When you're trying to persuade the audience, you can lead them along and give the key idea, or punchline, at the end. Mathematical developments frequently use this structure.

- **But don't overuse it, particularly if the argument is long!**

# Why put key ideas in the lead position?

- Helps prepare the reader for what's coming.
- Allows the reader to skim efficiently.

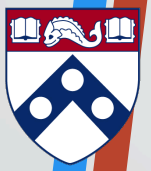


# Don’t make unsubstantiated claims

- Statements of belief or fact should be backed up by either 1) a specific result of your own work or 2) a citation to the literature

Data-stream engines have become an important component of most system-monitoring applications.

Data analytics tasks increasingly require cluster- and cloud-based parallelism.



# Be concise

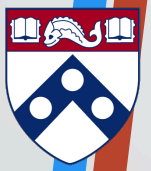
- Avoid wordiness.

In a situation in which updates outnumber reads, index costs can outweigh the benefits.

The Tri-Max algorithm has the capacity to handle noisy data.

The classification algorithm must recluster in the event that clusters start to overlap.

The reason to consider an alternative to using a touch-based interface is the absence of support on desktop and some laptop computers.



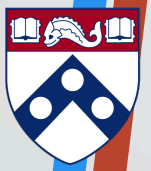
# Be simple

- Avoid “fancy” words

Insertion of a key commences by a lookup of the key value in the index.

Consider the steps involved when a user conceptualizes an information need.

- Even if the reader knows the fancier word, he or she will understand the plain word more quickly.



# Use a consistent lexical set

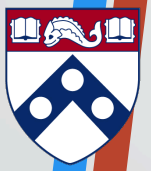
- Use the same word to refer to a concept throughout the paper.

Our system has four main modules. The parsing component analyzes the command and checks that the target file exists.

The analyzer locates the search phrase within the command. The expression is then passed to the term-expansion routine.

- Warn the reader about synonymous terms.

We use buffer and page slot interchangeably in this section.



# Define terms when first introduced

- Don't make the reader guess what you mean until the last section of the paper, where you finally get specific about the meaning of a term.

Sec. 2: The optimizer uses a calibration constant to accommodate the differences in processor speeds.

Sec. 4: The calibration constant is the ratio of its measured I/O speed to that of a reference processor.

**Better:** The caller must supply a convergence criteria (a bound on the residual error).

# Avoid single-sentence paragraphs

- Single sentence paragraphs are usually an indication that there's a problem with organization. Figure out where the idea belongs.
- If the sentence doesn't fit in an existing paragraph, and is not important enough to develop, then remove it!

# Re-write

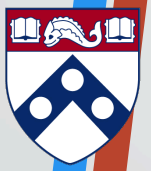
- When you re-write a section, paragraph or sentence that you know is difficult to read, keep asking yourself "what do I really mean?" Allow yourself several passes to get rough spots really concise, simple and clear.

- A good way to get more concise is to ask yourself if words in a sentence, or sentences in a paragraph are helping you make the key points or whether they can be discarded.

- **But also consider whether you've left out important information that the reader needs to know.**

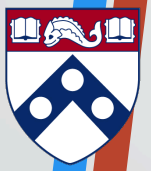
# Some stylistic comments

- Be consistent in American (rather than British) spelling
  - Don't use "color" and "colour", "modeling" and "modelling"
- Avoid contractions
  - Doesn't → does not
- Numbers
  - Spell out whole numbers less than 10 ("three choices" rather than "3 choices")
- Avoid Latin (so they say!)
  - E.g. → for example, for instance
  - i.e. → that is
  - Etc. → and so forth
- Avoid jargon
  - We sought user input on the design…



# Citations

- Don't use citations as nouns
  - In [15], the authors extend the method…
- Instead, treat them as parenthetical remarks
  - McDonnel and Slington [15] extend the method…
- Don't use citations in titles, section headings or abstracts.



# Avoid non-referential “This”

Reducing the number of service queues increases average delay and reduces the number of idle periods. This affects the recovery subsystem.

- Almost always clearer if you put a noun after “this”

This reduction affects the recovery subsystem.

# Past vs. present tense

- Be consistent in use of tense throughout a discussion
- Past tense
  - Use to indicate that results apply only to the particular study or experiment, or to say what you did

  "The survey showed that this population believes…"

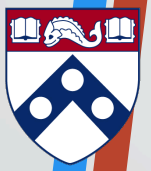
  "We removed all personal information from the data…"
- Present tense
  - Use to indicate that results generalize:

  "The experiments show that loops degrade the performances…"
  - Use to say what you are going to discuss later in the paper

  "In the next section, we show…"

  "In the next section, we will show…"

# 3. Conclusions



# Take-aways

- Having a good idea is only half the challenge of getting a paper accepted to a top conference or journal – writing is the other half.
- Your goal should be to explain the idea to the reader as simply as possible while convincing them of its technical depth.
- **Plan ahead:** start early so you are done a week before the deadline with a first draft of the paper.
- **Rewrite** to clarify and simplify.