How To Do Research?

June 2018 @ UCAS

Dongwon Lee

Penn State / IST
Timeline in US PhD Programs

- **Year 1:**
  - Taking courses
  - Passing Qualification Exam

- **Year 2-3:**
  - Taking more courses: per individual needs/tastes
  - Publishing 1-2 papers as first-author
  - 1-2 Internships
  - Defending a proposal → a PhD candidate (ABD)

- **Year 4-5:**
  - Publishing $k$ papers as first-author in *top* venues
  - Defending dissertation
  - Have a job offer
Timeline

advisor

student
Reading Papers

- #1 most important skill for scholars
- Survey paper
- Reviews for conferences or journals
Personal Research Log

- Maintain personal research log
  - Sketch your research ideas into a writing
  - Update your ideas as time passes
  - Occasionally go back to old writings
- Prepare a short review for each paper that you read
  - Summary
  - Pros and cons
  - Limitations or problems
  - If needed, contact authors and ask questions
    Usually authors are willing to discuss with their readers
Collaboration

- Collaboration are encouraged
  - If collaborated significantly, co-author papers
  - Else, acknowledgement in papers
- First author
  - One who contributed the most
    - Idea, design/implementation, experiment, analysis, writing
  - Can use materials in his/her dissertation
- CANNOT use the same materials in two dissertations
Research Tools

● Work
  ● Implementation: Python, Java, R, JavaScript
  ● Visualization: Excel, Plotting tool
  ● Repository: Google Drive

● Writing papers
  ● \textit{LaTeX}: Overleaf + ShareLatex
  ● Google Doc, MS Word

● Communication
  ● Email: SEND-ACK protocol with time window
  ● Skype, Zoom, Hangout, WhatsApp, WeChat, …
Dissemination of Your Work

- Have personal homepage
  - http://sites.psu.edu/
  - Share your papers

- Presentation
  - Grad research presentation

- Make datasets (and codes) publicly available
  - Replicable research becomes more important
  - By default, after publication, share all datasets/codes
Target Venues

- **Data Science**
  - SIGMOD, VLDB, ICDE, EDBT
  - KDD, ICDM, CIKM, WSDM
- **AI and ML**
  - AAAI, IJCAI, ICML, NIPS, ICLR
- **Cybersecurity**
  - CCS, Oakland, NDSS, USENIX
- **Social Computing**
  - ICWSM, WWW, HT, WebSci
Attending Conferences

- Networking
- Goal of presentation is NOT to have others understand ideas → have others get interested in your ideas (so that they read your paper later)
  - Make your presentation as SIMPLE as possible
  - Make it INTERESTING
- Find other interesting topics and papers
What *NOT* to do

- Research Misconduct
  - Plagiarism
  - Research manipulation
  - Double submission/publication
  - Submission/publication w/o discussion
- Sexual Misconduct
  - Big NO NO at Penn State and US institutions
  - Respect others’ culture, religion, political view, …
  - Be sensitive to gender issues
    “When a user presses Like... It is hard to know why *she* …”
WHAT IS “RESEARCH”?
What is “Ph.D.”?

http://gizmodo.com/5613794/what-is-exactly-a-doctorate
Publish or Perish

- has to be written first
- has to be validated as novel
- has to be published

To get a good job, have to have many & good papers...
The Goal of Research Papers

- Disseminate your ideas to others so that people appreciate/use/cite them

- **Graduate**… Of course
  - MS: need to write a thesis to graduate…
  - Ph.D.: “Publish or Perish”

- Without good publications…
  - No good job, no good career
  - And possibly no good life either

- GPA: nobody cares about PhD’s GPA
  - Maintain about 3.0/4.0
Where to Start?

- Given that you have acquired
  - basic theory/knowledge/tools from classes and books…

- Next, first thing to learn:
  - Read others’ papers
  - Critique and evaluate them

- Which paper to read?
  - Start from good ones
    - Classical ones
    - Ones from good journals or conferences
Differences in Disciplines

- **Computer Science**
  - Peer-reviewed conferences
  - Top conferences have 5-20% acceptance rates
  - Specialized and small conferences (attendance of 400+)
  - Often value conferences > journals

- **Pure Sciences (eg, Math, Physics)**
  - Pre-print at Arxiv.org
  - Rigorous reviews for journals
  - Huge flagship conference (ICM 98 attracted ~4,000)

- **Social Sciences**
  - Often value journals > conferences
  - Conferences are mostly for gathering or short abstract based screening
  - Rigorous reviews for journals
Where to Start: eg, Databases

<table>
<thead>
<tr>
<th>DB Conferences/Symposiums/Workshops (81)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADB, ADBIS, ADBT, ADC, ARTDB, Berkeley Workshop, BNCOD, CDB, CIDR, CIKM, CISM, CISMOD, COMAD, COODBSE, CoopIS, DAISD, DANTE, DASFAA, DaWaK, DBPL, DBSEC, DDB, DDW, DEXA, DIWeB, DMDW, DMKD, DNIS, DOLAP, DOOD, DPDS, DS, EDBT, EDS, EFIS/EFDBS, ER, EWDW, FODO, FoIKS, FQAS, Future Databases, GIS, HPTS, IADT, ICDE, ICDM, ICDT, ICOD, IDA, IDC(W), IDEAL, IDEAS, IDS, IGIS, IWDM, IW-MMDBMS, JCDKB, KDD, KR, KRDB, LID, MDA/MDM, MFDBS, MLDM, MSS, NLDB, OODBS, OOIS, PAKDD, PKDD, PODS, RIDE, RIDS, RTDB, SBBD, SDM-SIAM, Semantics in Databases, SIGMOD, SSD, SSDBM, SWDB, TDB, TSDM, UIDIS, VDB, VLDB, WebDB, WIDM, WISE, XP, XSym</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DB Journals (19)</th>
</tr>
</thead>
</table>

The list excludes Information Retrieval and Digital Library
Reference Chase

- Don’t trap into the “Exponential Reference Chase” problem

\[
\begin{align*}
\text{Papers to read} & \quad \uparrow \\
\text{in queue} & \quad \downarrow \\
\text{References that you use} & \quad \rightarrow
\end{align*}
\]
Symptoms

- After chasing relevant works that are increasing super-exponentially fast, you might feel…
  - All relevant problems are ALREADY studied by someone else
    Others have 1000+ history: Mathematics, Art, …
  - Problem is too BROAD for me to tackle
    Divide-n-conquer
Finding DARN Research Problems?

- Easy but non-helpful answer:
  - Read and think and read and think and…
- Subjective but MAYBE-helpful answer
  - MAP approach
  - MATRIX approach
  - DELTA approach
  - DROP approach

What I Call M2D2
1. MAP Approach

- To start a research, initially, you have to read a lot of papers anyway
- While reading those, why don’t you analyze and summarize what you’ve read and put them into your own wording?
  - Good for a survey paper – a MAP for future readers
- To be publishable, your survey must have novel view-point, taxonomy, comprehensive analysis, or all of them
2. MATRIX Approach

- Now, You have read a lot of papers
- Draw a MATRIX on a specific problem, and map the paper that you read to cells of matrix
- At the end, non-filled cell is the missing work that no one has done
- But wait… first make sure that:
  - The hole is worthwhile to fill in
    - Doable (good as my dissertation topic?)
    - Value (what’s good?)
Eg, XML-Relational Conversion Problem

Around 2000 when I was a Ph.D. student

<table>
<thead>
<tr>
<th></th>
<th>Schem a</th>
<th>Const raint</th>
<th>Query</th>
<th>View</th>
<th>Trigger s</th>
<th>Securit y</th>
<th>Top-K</th>
<th>Temporal</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML → Relatio nal</td>
<td>O (40+)</td>
<td>O (5+)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatio nal → XML</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. DELT Approach

- Arguably easiest...
  - Pick one paper of your interest
  - Read a lot – more than 10 times
  - Find limitations and Extend it by DELTA
  - Prove or demonstrate that
    The limitation that you pointed out is valid
    Your suggestion improved the problem by DELA
- The more well-known work you choose, the harder to improve, but the better for your reputation...
  - Eg, “E.F. Codd’s relational model is insufficient to handle semi-structured model because…”
- The bigger the DELTA is, the better your paper gets
Eg, The optimal wedding problem

- When a person has a chance to date $K$ persons, the optimal wedding algorithm is:
  - Date upto $K/3$ persons
    - Let the best person among $K/3$ as $B$ using a criteria $C$
  - Start dating again from $K/3+1$ person, $p$
  - If $p$ is better than $B$ using $C$
    - Stop and Marry $p$
  - Otherwise, keep dating till $K$-th person

- How many ways can we improve this algo?
Possible DELTAs

- Parameters fitting:
  - How to determine $K$? Estimate?
  - How to determine $C$? Comparison?

- Scalability? $K=10$ vs. $K=100,000$? Sub-optimal?

- Question the assumptions:
  - Monogamy vs. Polygamy vs. N-gamy? (How to find $n^{th}$ best spouse fast?)
  - Data distribution? Uniform/Poisson/Scale-free

- Application to another domain?
- System building?
Which DELTA to Choose

- Pick the DELTA that is the most **significant**

- Some criteria are:
  - Have practical values
    - Has motivational scenario as of NOW, or
    - Predicted to be useful in $N$ years
  - Non-trivial
  - Hot topic $<>$ good topic
4. DROP Approach

- Pick a simple but fundamental assumption of existing theory/model/systems/methods
  - Drop It

- Reconsider to see how the drop affects all aspects of the existing theory/model/systems/methods

From http://www-db.stanford.edu/~widom/stream.ppt
Eg, Two Stanford DB Projects

- The **LORE** Project
  - Dropped assumption:
    "Data has a fixed schema declared in advance"
  - Semi-structured data (→ XML)

- The **STREAM** Project
  - Dropped assumption:
    "First load data, then index it, then run queries"
  - Continuous data streams (+ continuous queries)

From http://www-db.stanford.edu/~widom/stream.ppt
Facts on Paper Reviews
(adopted from J. Cho’s slides @ UCLA)

- 3-4 reviewers per paper
- 5-20% acceptance rates for top-tier venues
  - Very competitive
- Criteria
  - Accept/Weak Accept
  - Neutral
  - Weak Reject/Reject
- One reject kills a paper
  - At least Accept or Weak Accept
About Reviewers

- 10-15 papers per reviewer (for top conferences)
- Reviewer cannot spend 5-10 hours per paper
  - 20 \times 10 = 200 \text{ hours} = (40 \text{ hours} \times 5) = \text{5 weeks!}
  - No reviewers can afford this
- Give a good impression in 1-2 hours!
  - Impression matters the most
  - Content comes next!
- Reviewer do NOT get paid \Rightarrow no motivation to do extra work

WARNING: Of course, to start with, your main idea must be good to get into top-tier…
Good Impression in 1-2 hours?

1. **Good introduction**
   - Everyone reads it
   - If not interesting, people stop reading

2. **Easy to read**
   - People should understand what you say
   - Easy to confuse, difficult to understand

3. **Build an excitement and a strong case**
   - What is good?

4. **Broad reference**
   - Sometimes kills a paper
   - Program committee members
Good Introduction Sells

I have often said reviewers make an initial impression on the first page and don’t change 80% of the time

Mike Pazzani

This idea, that first impressions tend to be hard to change, has a formal name in psychology, Anchoring.

Excerpt from “How to do good research, get it published” by Eamonn Keogh
Good Introduction Sells

The Most Important Part of Your Paper: the Introduction

- The 1/3 – 2/3 Rule from a reviewer’s perspective:
  - 1/3 time to read your introduction and make a decision
  - Remaining 2/3 time to find evidence for the decision
- [Take-Home Message #6] A good introduction with a good motivation is half of your success!

Excerpt from “How to do good research, get it published” by Eamonn Keogh
Good Introduction Sells

The First Page as an Anchor

The introduction acts as an anchor. By the end of the introduction the reviewer must know.

- What is the problem?
- Why is it interesting and important?
- Why is it hard? Why do naive approaches fail?
- Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions?)
- What are the key components of my approach and results? Also include any specific limitations.
- A final paragraph or subsection: “Summary of Contributions”. It should list the major contributions in bullet form, mentioning in which sections they can be found. This material doubles as an outline of the rest of the paper, saving space and eliminating redundancy.

Excerpt from “How to do good research, get it published” by Eamonn Keogh
How to Write an Introduction

1. Start with 5 bullets
   - What’s the problem?
   - Why is it interesting?
   - ...

2. 1-2 sentence answer to each question

3. Add more content

4. Spend enough time on introduction
   1. Bullet points enough
Easy-to-Read Paper

- You can always make it complicate later
  1. Lots of examples
  2. Figures & Tables – Figure speaks !!
     - Summary of notations
  3. Define assumptions/models/architecture precisely
     - Explicitly write down assumptions
     - Input, output, property, goal function
  4. Make a connection
     - Why this experiment?
Paper Organization (10 pages)

1. Introduction (2 pages)
2. Related Work (half page)
3. Prelim. & Framework (2 pages)
4. Main Ideas (3 pages)
5. Experiments (2 pages)
6. Conclusion (half page)
7. References (half page)

- Actual idea – only 3 pages !!!
- Even tiny idea can turn into a good paper if you DEVELOP it well
Watson & Crick’s Nature paper on double helical structure of DNA is only 1 page (+ 1 paragraph) long.
Start Writing Early On…

● Even if you feel you are NOT ready yet
  ● Your advisor will throw away your initial draft anyway
  ● Your initial submission will be rejected anyway

● But you get:
  ● (good or bad) Experiences and learn from that
    Having such experiences early is important
  ● Writing sharpens your ideas and gives more ideas
  ● Writing can be improved only via more writing