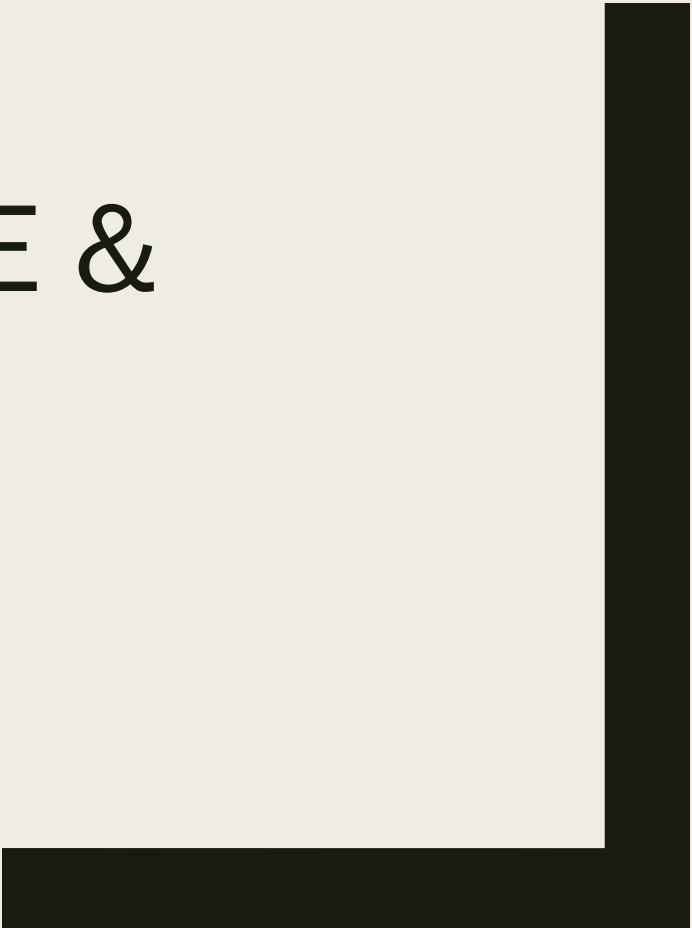




WEEK 1: WELCOME & INTRODUCTION

INSC 592
Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025



About me

- Instructor: Kai Li
- kli16@utk.edu



- <https://sis.utk.edu/about/directory/kai-li>
- <https://orcid.org/0000-0002-7264-365X>

My Research

Scientometrics

🌐 21 languages ▾

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

For the journal, see [Scientometrics \(journal\)](#).

Scientometrics is the field of study which concerns itself with measuring and analysing [scholarly literature](#). Scientometrics is a sub-field of [informetrics](#). Major research issues include the measurement of the impact of research papers and academic journals, the understanding of scientific citations, and the use of such measurements in policy and management contexts.^[1] In practice there is a significant overlap between scientometrics and other scientific fields such as [information systems](#), [information science](#), [science of science policy](#), [sociology of science](#), and [metascience](#). Critics have argued that over-reliance on scientometrics has created a system of [perverse incentives](#), producing a [publish or perish](#) environment that leads to low-quality research.

It is a field focusing on **using large-scale data sources** (such as the Web of Science and Scopus) and various **quantitative and network methods** to understand the scientific system and the production and communication of scientific knowledge.

My Research

- Scholarly communication
 - *Roles played by research data and software in knowledge production*
 - *Software and data citation practice*
 - *How do researchers publish and reuse data and software*
- Library bibliographic data
 - *Connecting library bibliographic data to other data sources (such as scholarly databases)*
 - *Understanding the intellectual space in library catalogs, such as the topics being covered and book authors' publication profiles*
- I am hiring an RA till the end of summer. Let me know if you are interested.

My contact information

- Email: kli16@utk.edu
 - *Sending emails to me directly is the most convenient way to find me.*
 - *I will try to respond your message in the same day.*
 - *Try NOT to use the message function in Canvas.*
- My office hour: 1:00-2:30 PM of every Tuesday.
 - *Please [reserve a time slot](#) (30 mins) before attending.*
 - *You are still welcome to use my office time without reservation!*
 - *If the time does not work for you, I am more than glad to arrange a meeting to talk to you!*

Self-introduction

- I am going to call you names and please share:
 - *Who are you?*
 - *Why are you taking this class?*
 - *What do you expect to learn from this class?*
 - Especially, how can this class contribute to your future career?

Introduction to the Canvas site

- <https://utk.instructure.com/>

Format of the course

- We are meeting every Thursday night!
- However:
 - *This is the first time I am teaching it synchronously (it was previous taught as an asynchronous class).*
 - *And this class requires more practice and doing than teaching.*
- → We will have shorter lectures (*normally 45-90 minutes) every week and will have **optional** lab time after the lecture.
- In lab time, we will either do some technical demonstration and/or you can finish some tasks in the class.

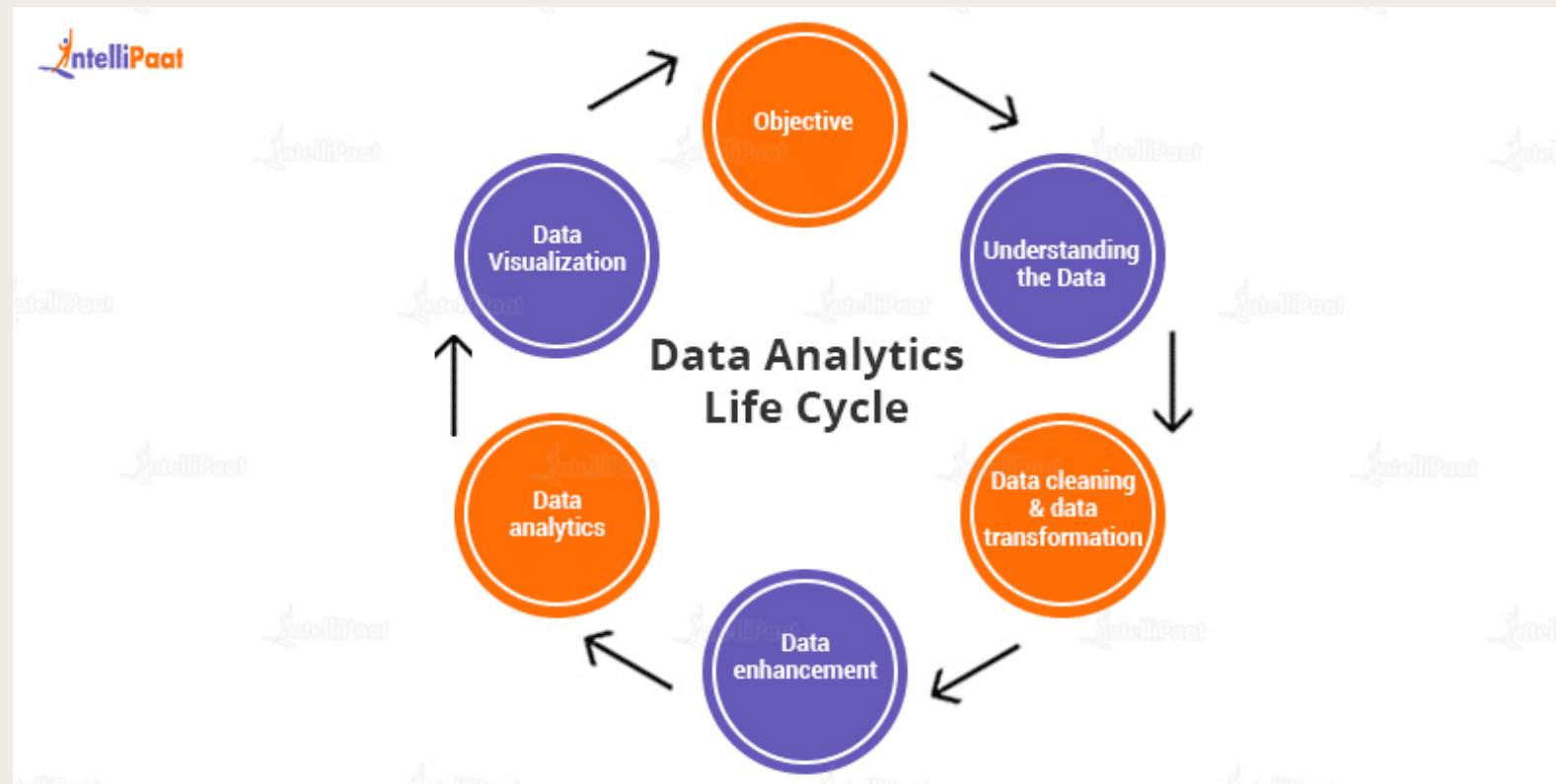
Purpose of the Course

- The following major topics will be covered by this course:
 - Concepts of *big data and data analytics* in settings such as academia and professional organizations.
 - Basic techniques of the whole *lifecycle of data analysis: data collection, processing, analysis, and visualization*.
- I am going to open a new 590 class on the topic of data visualization from the next fall semester. But I still try to cover the basic concepts of data visualization in this class.

Learning Outcomes

- Understand the needs and importance of data analytics, both quantitative and qualitative, in various contexts;
- Understand the lifecycle and tasks for completing a research project using data;
- Conduct data collection and understand major approaches;
- Practice data wrangling (data cleaning);
- Conduct exploratory analysis and visualization;
- Conduct text analysis;
- Develop and evaluate statistical models.

Data analytics lifecycle



<https://intellipaate.com/blog/tutorial/data-analytics-tutorial/data-analytics-lifecycle/>

What we are going to do

- Doing is as important as (if not more important than) just listening to me.
 - *Our regular weekly lecture will be about 45-90 minutes long.*
 - *We will have regular assignments as well as classroom activities, which require you do “get your hands dirty.”*
 - *In doing the assignments and activities, you will also need to solve a lot of the technical challenges. I will try to offer some help. But in most cases, Google or ChatGPT are probably more helpful sources!*

What we are going to do

- I will upload slides and class recording (and any other materials) to Canvas after the lecture.
 - *Let me know if you prefer reading the slides before the class.*
- Reading materials will be generally made available two weeks ahead of the schedule.

Technical skills?

- This course is not about programming language.
 - *That said, it is impossible to teach this class without having a default tool.*
 - *We will use R (and to a lesser degree, Python) as the default demonstration tool.*
- Ideally, I hope you could learn about one programming language.
 - *Excel and SPSS are useful but extremely limited.*
- We will talk about what tools are available to you next week and I will share my experience of using some tools.

Technical skills -- Survey

- A quick survey:

- *How many of you have used any programming language before (R, Python...)?*
- *How many of you have learned statistics before?*

Technical skills vs. understanding

- But we will try to develop a more balanced view of all topics, especially between technical skills and the general understanding of the bigger picture of data analysis and storytelling.
 - The *socio-technical view* is very important in information science and more broadly, social science.
 - Also, learning these tools and acquiring technical skills are a life-long project. I am just hoping to give you the general capacities to learn by yourself in the future.

The Real Challenges

- How to turn data into information, knowledge, and insights?
 - *Data analysis and visualization are two critical, interconnected yet distinct steps in the pipeline.*
- It is hard to deal with real data, which tends to be *messy* and *difficult* to collect (in many cases) and process.
 - *Not every dataset is available or easily findable and not every available dataset is usable or valuable.*
 - *No one knows for sure what we will get from an analysis in the beginning.*
 - *Interpreting data requires business context.*

Communication matters!

- Communication is an essential part of data science.
 - *All data analysis results are supposed to be communicated to someone.*
- It is also critical for an online course like ours.
 - *Don't hesitate to share any question or concern with me, so that I can help you as much and as soon as I can.*
 - *We will also have a mid-term evaluation so that any issue in the first half of the course can be identified and fixed.*

Policies on ChatGPT

- Use of any GenAI to generate/write research papers, essays, or other materials in the course is considered plagiarism.
 - *Please don't use it to create the whole homework.*
 - ***The report should be from you instead of the algorithm.***
- GenAI is a very useful source for you to solve programming issues and get ideas of data analysis:
 - *I am OK if you use GenAI to get inspiration or to solve technique problems.*
 - *If we have time, we can talk about using ChatGPT API.*

Structure of the Course

- Weeks 2-3: basic concepts about data science and analysis
 - Weeks 4-5: data collection and processing
 - Weeks 6-8: statistical methods
 - Weeks 10-12: text analysis & visualization
 - Weeks 14-15: final project presentations
-
- This is a tentative plan, and I expect to make some changes during the semester!

Course Assessment

- 2 solo regular assignments (30 points)
- Final project (50 points)
 - *It was designed to be a solo assignment.*
 - *This project includes a class presentation in the final weeks.*
- Participation (20 points)
 - *It will be measured by your attendance and participation in discussion questions.*
 - *Expect weekly discussion questions as part of the class content.*

Grading Policy

- You should expect 90% of points in a regular assignment if the report meets all my expectation.
- I'll accept late assignments, but 10% of points will be deducted within the first 48 hours, 20% within the first week, and 30% afterwards.
 - *For discussion questions, occasional late submissions won't affect your grade, but please try not to be late.*
- If something's come up or you're having any issue with the assignment or grading, talk to me.

Textbooks

- We will rely on a series of textbooks.
 - [R for Data Science](#) (R4DS)
 - *Introduction to Statistics for the Social Sciences (I3S)*
 - *Modern data visualization with R (DataVis with R)*
 - *Text Mining for Social Scientists (TM2S)*
 - *Data collection: key debates and methods in social research (DC):
Our library has this book here.*

Some other resources

- Many other MOOCs, especially on Coursera.
- For most of the technical questions (especially those about programming language and statistics), you can find many good instruction, Q&A websites, and blog posts.
- But why should we have this class? – **Our discussion question**
 - *Veritasium has this very interesting video about the relationship between technologies and education:*
<https://www.youtube.com/watch?v=GEmuEWjHr5c>.

Questions, Problems, etc.

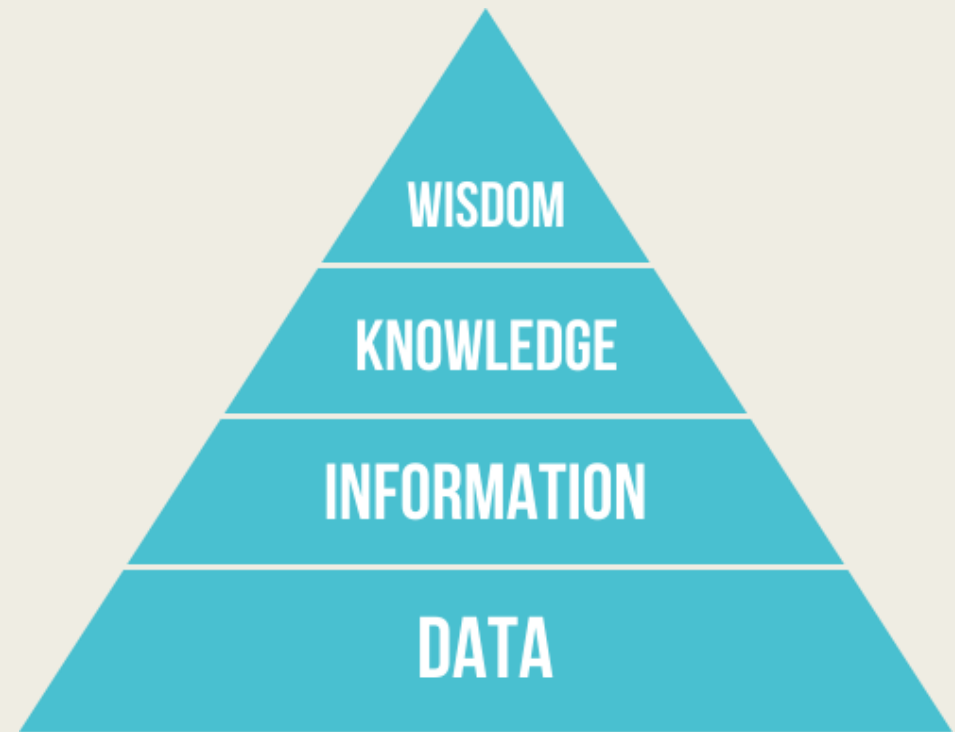
- If you have any question or suggestion, please don't hesitate to let me know!
- You are the co-owner of this course, and I am here to support your learning and achieving success!

Some basic concepts we will talk about

- Data & Big data
- Data analysis
- (We will talk more about data science next week.)

Data

- Different views about data:
 - *Data as facts, raw materials*
 - *Data as constructed objects*
- Both views are (not in-)consistent with the fact that we can process and analyze the data to draw more insights from it.



Big data

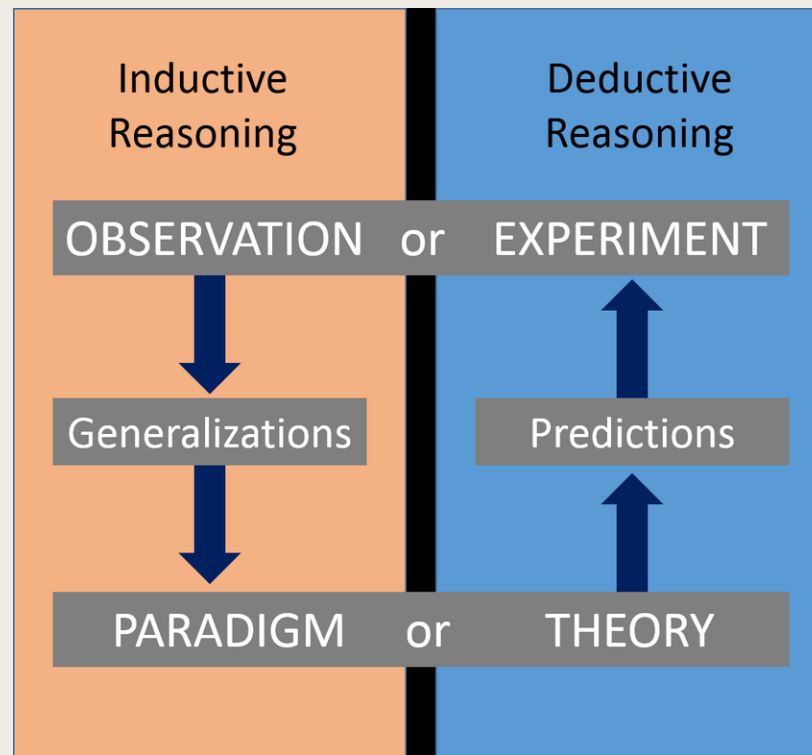
- The size of data created by human beings significantly increased after the 21st century, which further calls for new methods and technologies to store, manage, and analyze the data.
 - *Theoretical and technical challenges...*
- Three aspects of bigness:
 - *volume: quantity*
 - *velocity: speed of growth*
 - *variety: types (multimodal data)*

Data vs. Theory

- A major argument about big data is that bigness of data will render theories less useful (or even useless) in scientific investigation, because we can draw "data-driven" conclusions from data.
 - *For example, Chris Andersen's famous article "The end of theory"*
<https://www.wired.com/2008/06/pb-theory/>
- This argument has been debated by many social science researchers, such as Boyd & Crawford (2012).

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

Inductive vs. deductive reasoning



Issues with bigness (or over-reliance on it)

- We should acknowledge that not everything is measurable; and in many cases, what is measurable may not be important or relevant or even real.
 - *Relevance to the realities is a justification we very frequently need to give in quantitative social science research.*
 - *Such as using IQ to measure one's intelligence*
 - *Veritasium (again) has a very good video on the history of IQ:*
<https://www.youtube.com/watch?v=FkKPsLxgpuY>.
- So, please critically understand the data before you use it, especially when it was collected by other people!

Issues with bigness (or over-reliance on it)

- Bigness is not necessarily better.
 - *There are many other important criteria:*
 - Reliability / Quality
 - Representativeness
 - Documentation
 - Relevance
 - *Methodological issues with big data: spurious correlation → just relying on data can be misleading.*
 - *In many cases, a small dataset is enough and more feasible / economic.*

Another aspects: Quantitative vs. Qualitative

- Even though this course is focused on quantitative analysis methods, please remember that no research method alone can solve all problems.
 - *Every method has its advantages and disadvantages.*
- Question:
 - *What do you think are the pros and cons for quantitative vs. qualitative methods?*

Another aspects: Quantitative vs. Qualitative

- **Quan** methods can produce broader (more data points) but shallower (much fewer relationships) conclusions than **Qual** methods.
- Both types of methods have clear pros, cons, and biases.
- Mixed-method design is very useful when it is possible.

Discussion activity

- Please watch the Veritasium video and reflect on the following questions:
 - What do you expect from this class?
 - Or any other ideas from the video that you want to discuss.
- For the question, you can post an answer with 150-200 words that covers your key points.

Discussion activity (cont.)

- Please try to submit your answer to Week 1 Discussion board by the end of Jan 29, so that I can address your answers in the next class.
 - *Like I said, being late would not be a big problem that affects your grading, especially in the discussion activity.*
 - *But please try to catch up and let me know if you have any questions or concerns.*

Assignment 1: Tasks & skills of data scientists

- Deadline: 2/9
- Compose a report to summarize the some of the following information about data scientists (you don't have to answer all of them and look at the next slide for details):
- Please visit the assignment page on Canvas to get more information.

Assignment 1: Tasks & skills of data scientists

- You don't have to address all the questions below, but feel free to focus on some of them in your answer:
 - *Explore the current landscape of data science*
 - *What are the job titles?*
 - *What are the basic competences in the job postings?*
 - *What are the desired skills?*
 - *What are the salary ranges*
 - *How to learn or obtain these skills?*
 - *Academic programs*
 - *Certificates*