Principles for Data Science Process

Duncan Temple Lang
Dept. of Statistics
UC Davis

Core Concepts

- Statistics and Machine Learning.
- Computing and Technologies.
- Domain knowledge of each problem.
- Most important is to be able to reason about each of these and the data in each step of the process.
  - More than templates/methods/algorithms.
  - Experience and exposure.
  - Quite different from many university programs in each field.

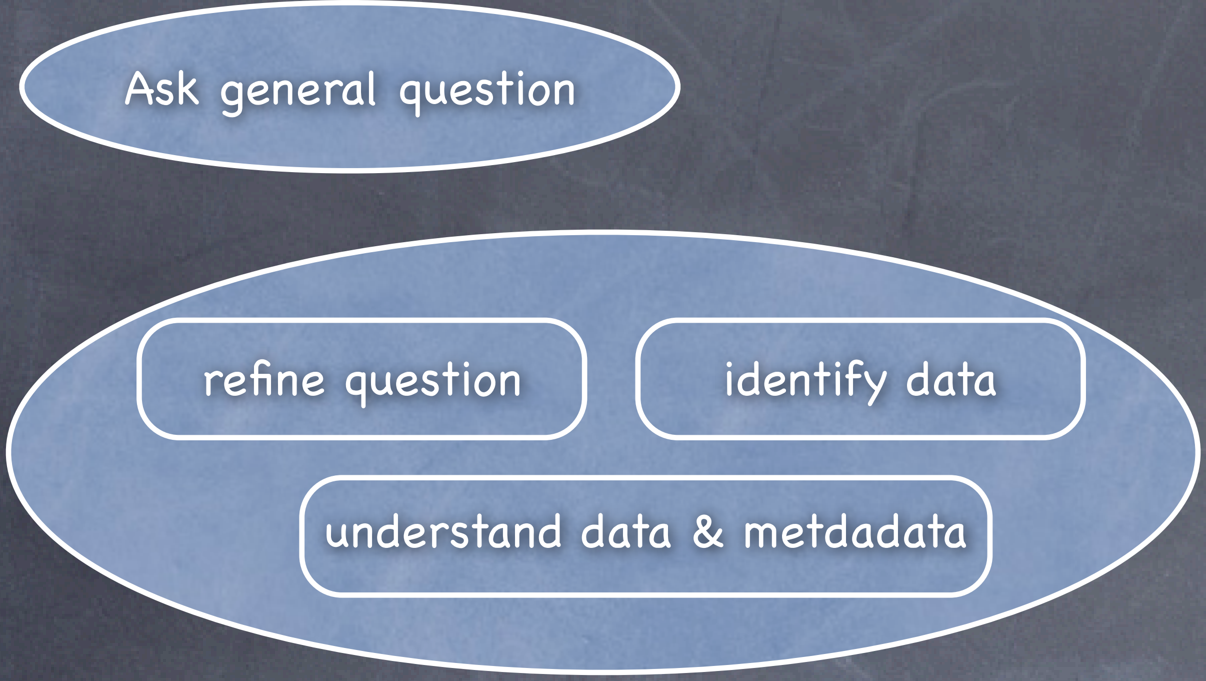
Combination of Skills & Areas

- How many students have more than one of these skills?
  - Understand randomness and uncertainty.
  - Understand statistical methods and have experience using them.
  - Programming skills and knowledge of multiple technologies.
- In statistics, most students have 1 quarter/semester of computing!
- In CS, how many statistics/probability/machine learning courses?

Data Science/Analysis Pipeline

- Might be useful to identify the common steps in a "typical" process of data analysis and exploration.
- Then identify the associated key data science concepts statistical and machine learning, and computing/technology concepts
- Just one hopefully representative and common pipeline of many
  - different types of student/researchers, types of analyses.
- Focusing here on data scientist working on applied problem. Consumer of statistical/machine learning methods and high-level technologies.

Ask general question

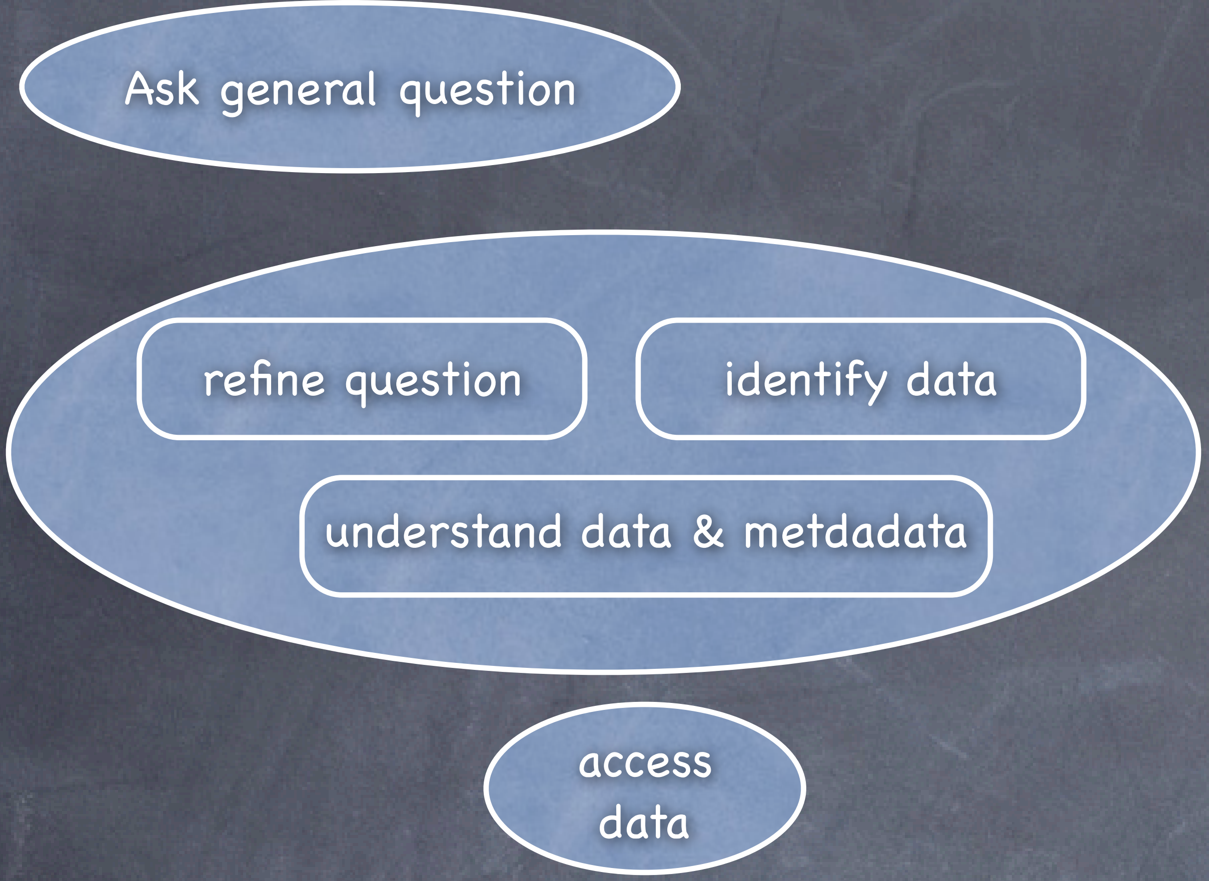


Ask general question

refine question

identify data

understand data & metdadata



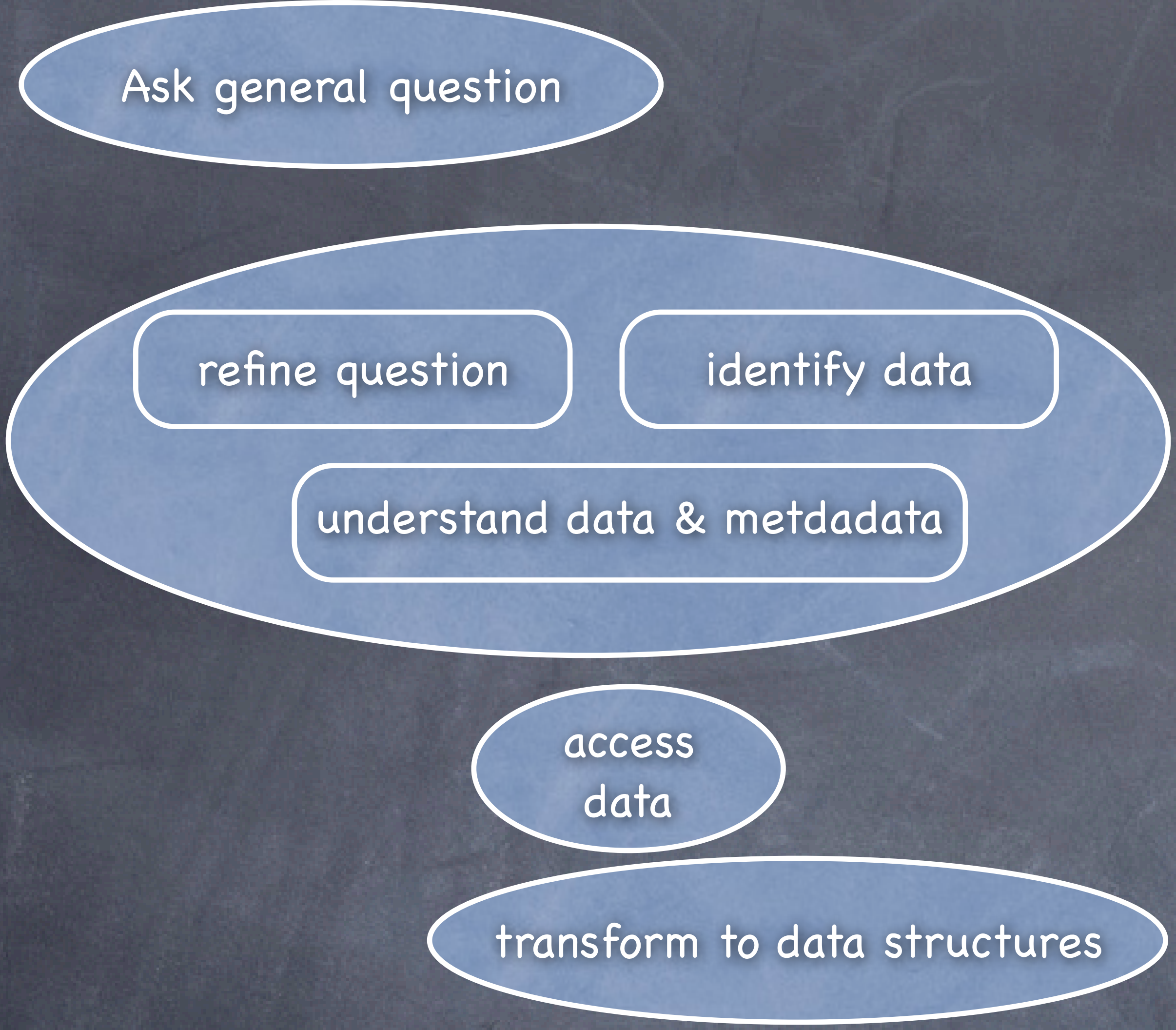
Ask general question

refine question

identify data

understand data & metdadata

access data



Ask general question

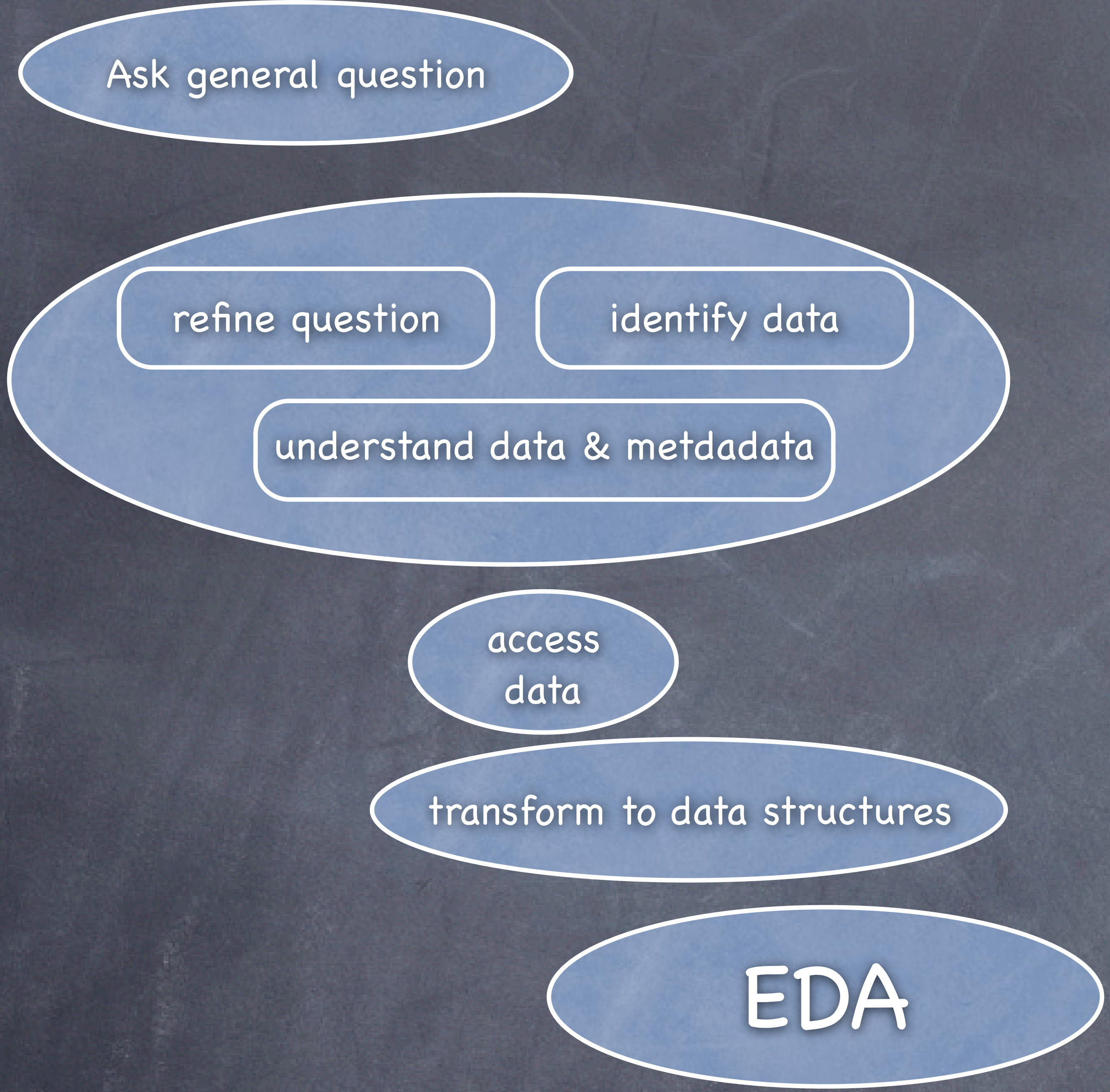
refine question

identify data

understand data & metdadata

access data

transform to data structures



Ask general question

refine question

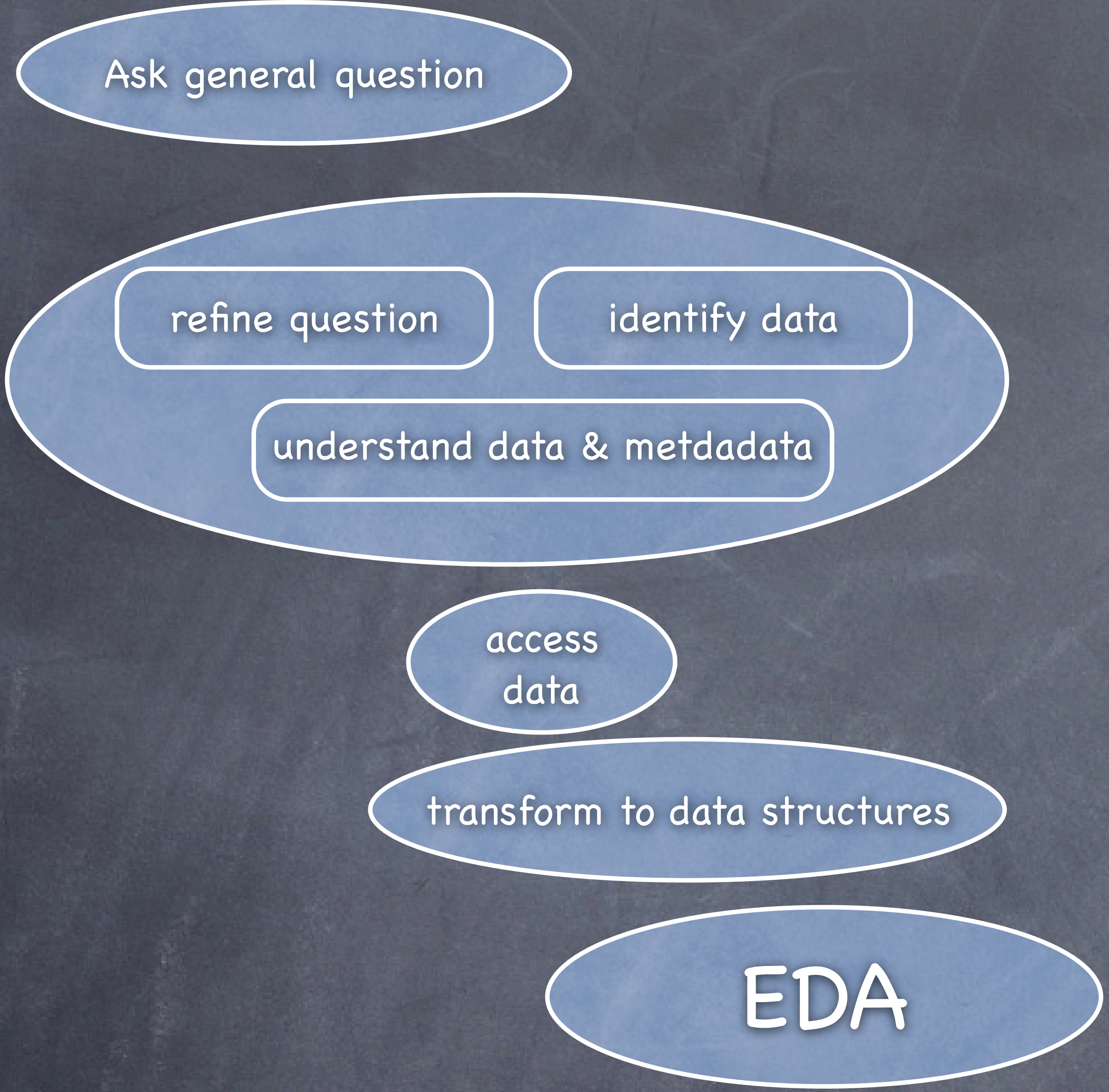
identify data

understand data & metdadata

access data

transform to data structures

EDA



Ask general question

refine question

identify data

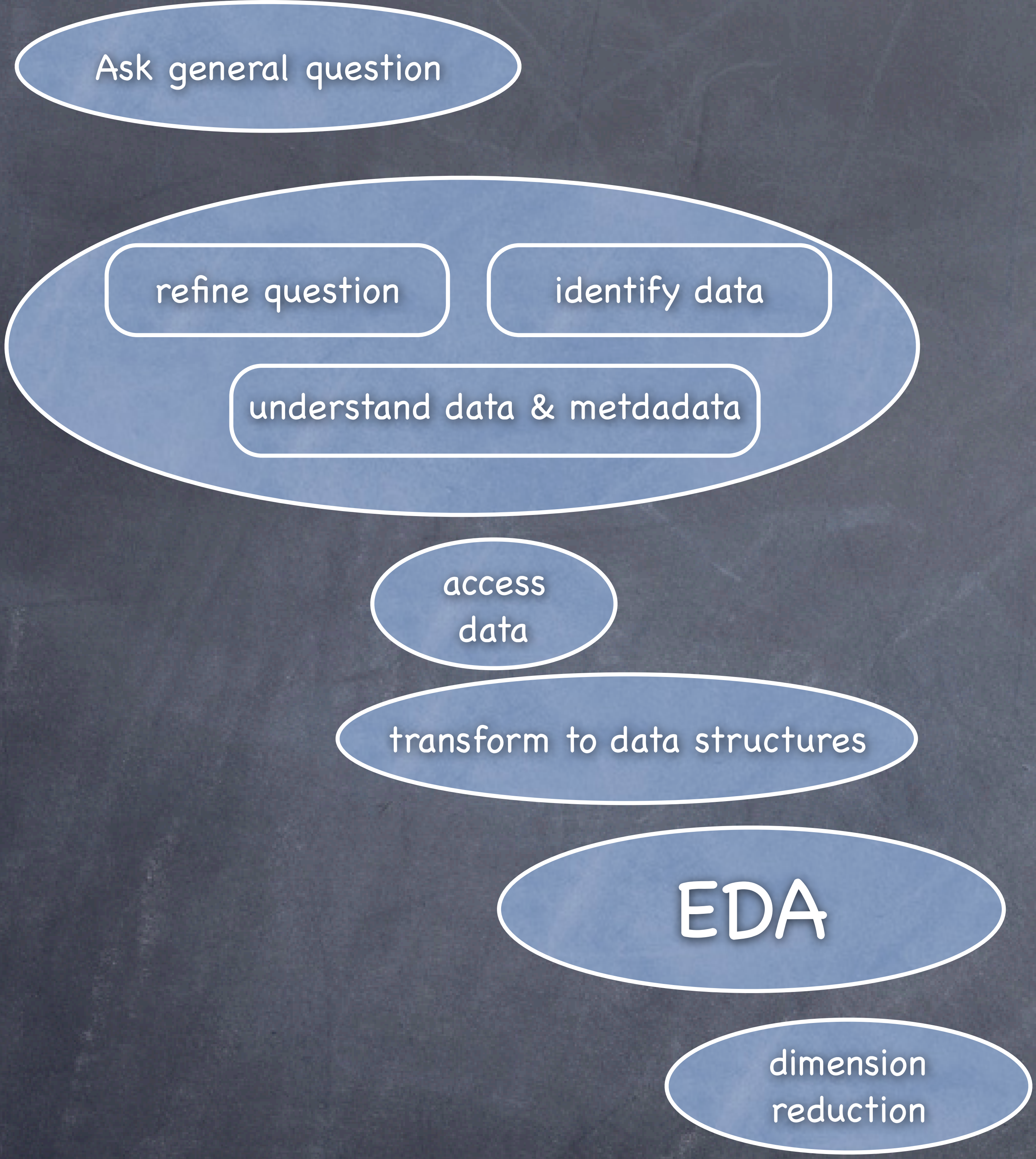
understand data & metdadata

access data

transform to data structures

EDA

Understanding the distributions (single, joint, conditional), outliers & anomalies,
new features of interest, transformations of variables,
new derived variables.



Ask general question

refine question

identify data

understand data & metdadata

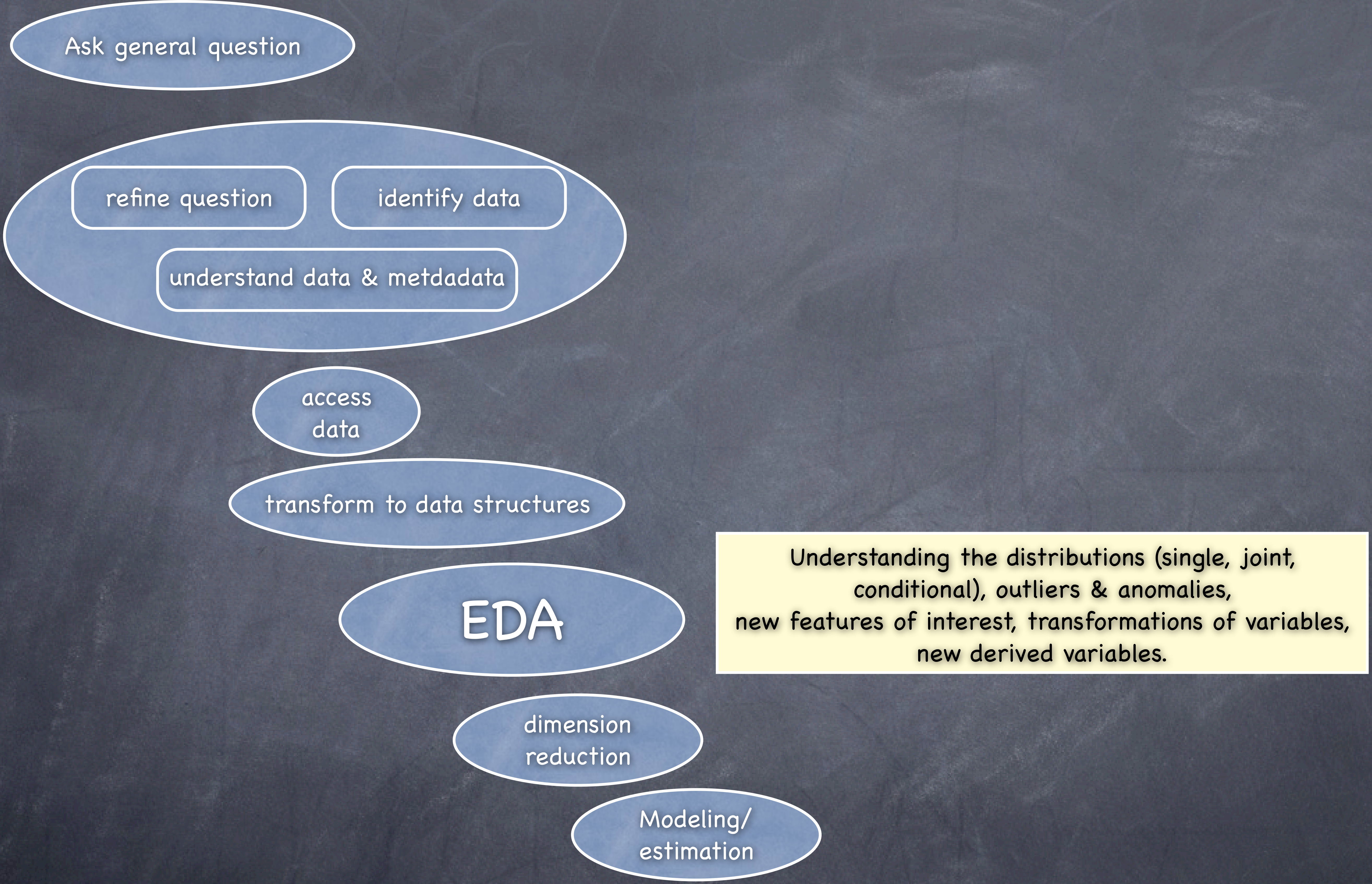
access data

transform to data structures

EDA

dimension reduction

Understanding the distributions (single, joint, conditional), outliers & anomalies,
new features of interest, transformations of variables,
new derived variables.



Ask general question

refine question

identify data

understand data & metdadata

access data

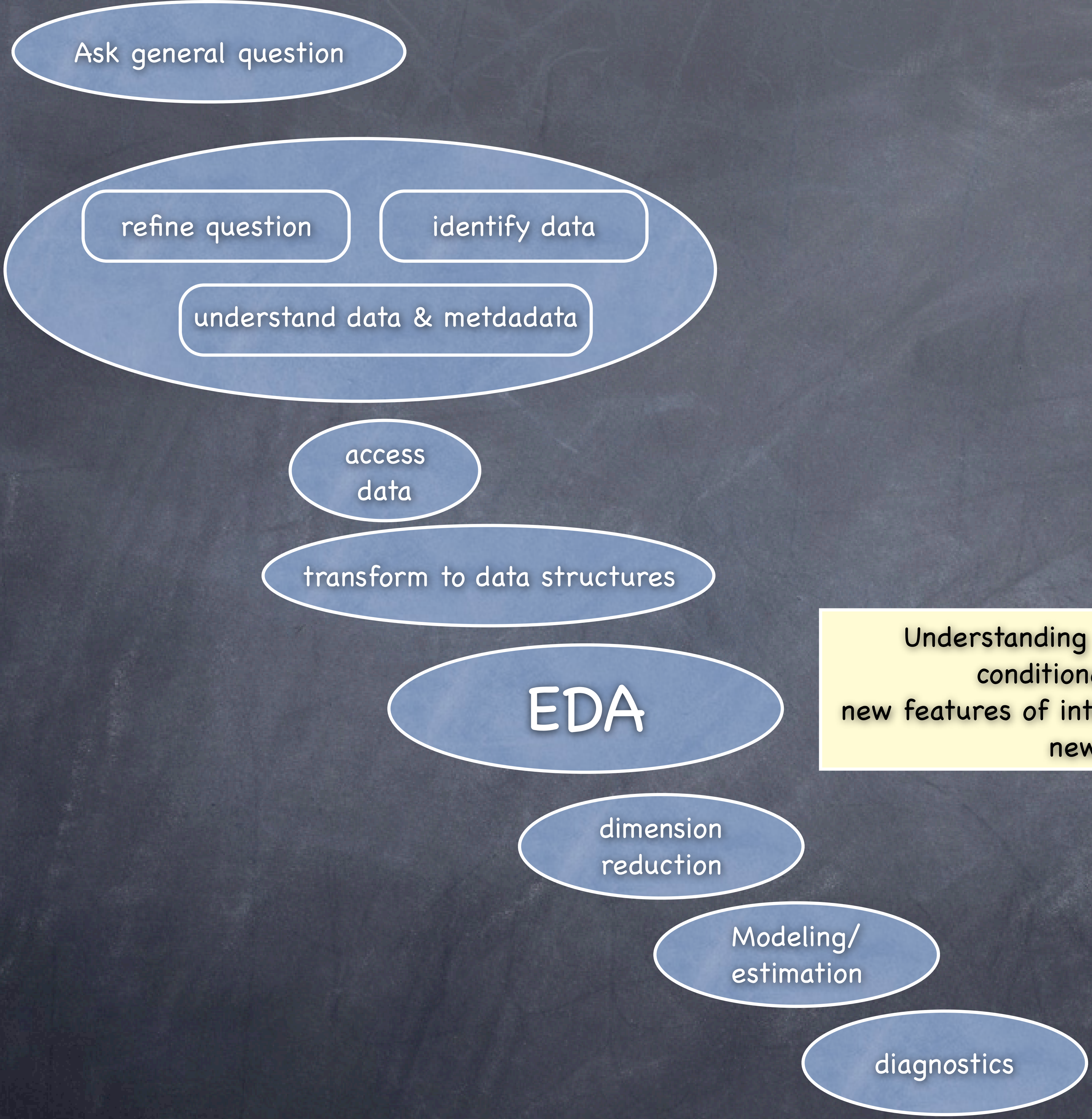
transform to data structures

EDA

Understanding the distributions (single, joint, conditional), outliers & anomalies, new features of interest, transformations of variables, new derived variables.

dimension reduction

Modeling/ estimation



Ask general question

refine question

identify data

understand data & metdadata

access data

transform to data structures

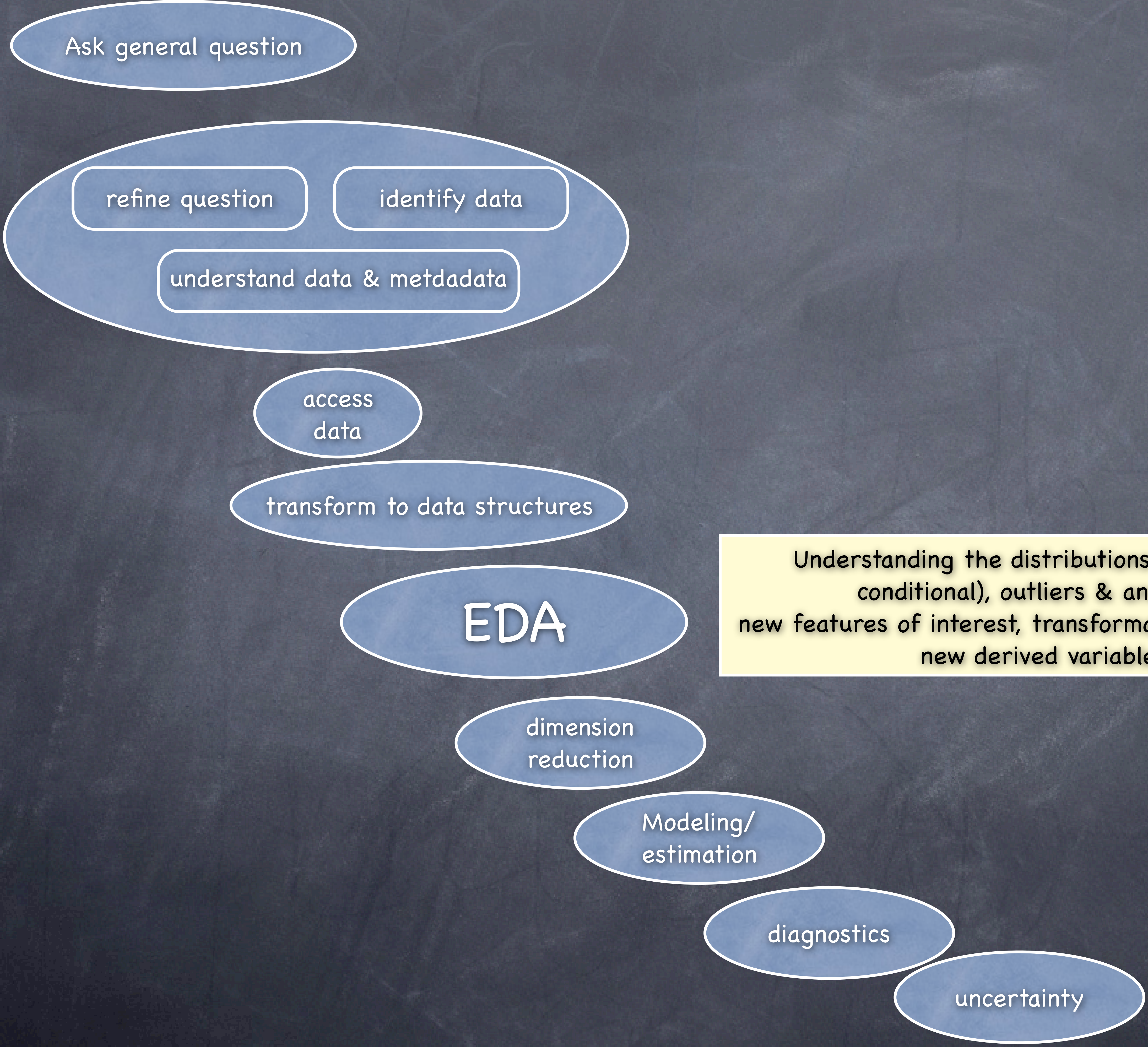
EDA

dimension reduction

Modeling/ estimation

diagnostics

Understanding the distributions (single, joint, conditional), outliers & anomalies,
new features of interest, transformations of variables,
new derived variables.



Ask general question

refine question

identify data

understand data & metdadata

access data

transform to data structures

EDA

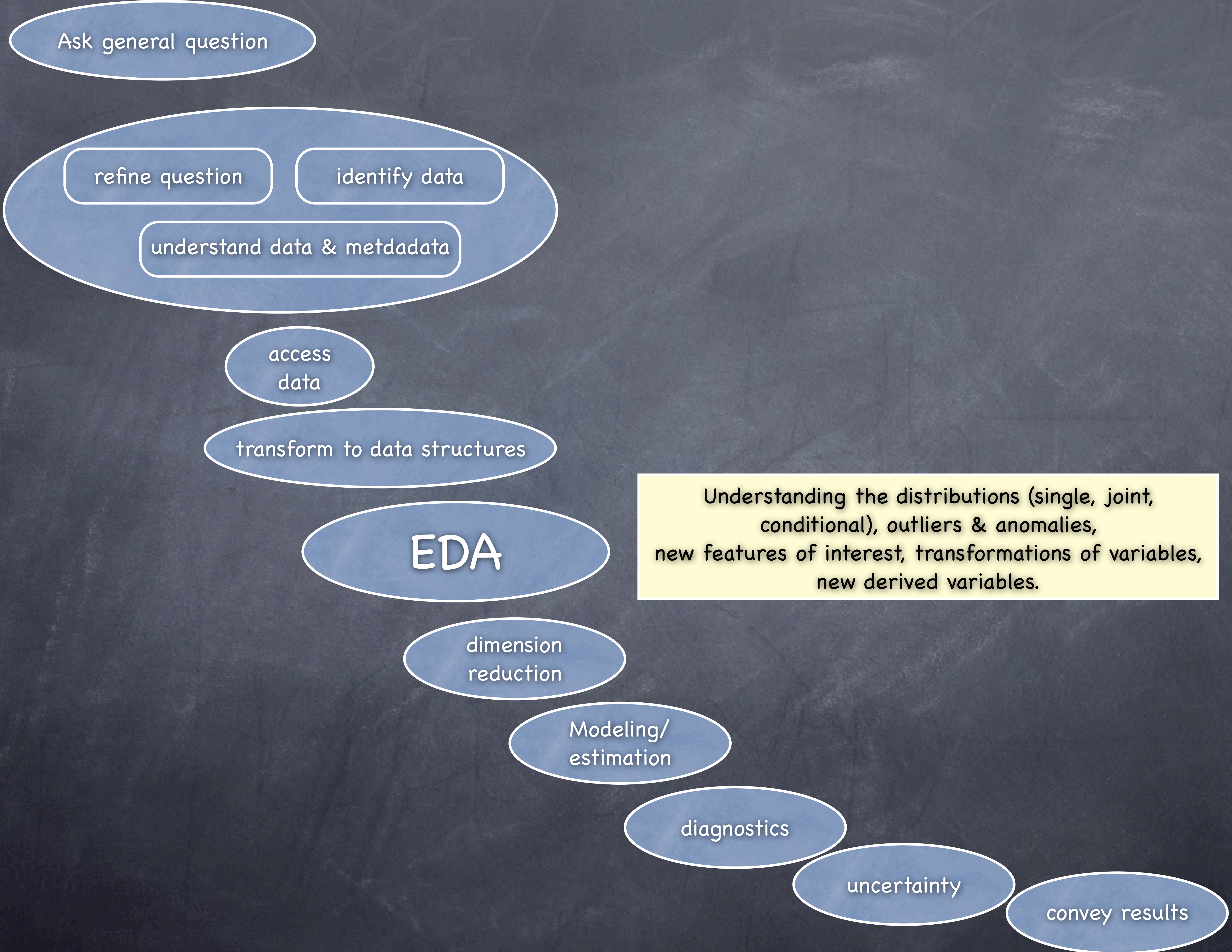
dimension reduction

Modeling/ estimation

diagnostics

uncertainty

Understanding the distributions (single, joint, conditional), outliers & anomalies,
new features of interest, transformations of variables,
new derived variables.



Ask general question

refine question

identify data

understand data & metdadata

access data

transform to data structures

EDA

Understanding the distributions (single, joint, conditional), outliers & anomalies, new features of interest, transformations of variables, new derived variables.

dimension reduction

Modeling/ estimation

diagnostics

uncertainty

convey results

Nature of the Process

- Process is highly
  - interactive
  - iterative
- Don’t necessarily know what we will do next until we see results of current step
- Return to find new data, EDA steps, modeling and diagnostics.
- Find the unexpected iteratively by learning the expected.
- Also, changes the set of skills we need in the process.

Abstract Skills & Concepts

- For many, most essential skills are actually
  - exposure and experience to complete data science process
  - statistical reasoning - identifying biases, scope of inference and conjecturing alternative explanations and exploring those.
  - gaining confidence and familiarity with going beyond methods alone.
    - methods & computing are not the goal, but the means. But typically significant obstacles.

Abstract Skills & Concepts

- basic programming and computational problem solving
- statistical concepts
- statistical problem solving & “sleuthing”

Statistical Concepts

Statistical Concepts

- mapping the general question to a statistical framework

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
    - data collection provenance & metadata

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
  - data collection provenance & metadata
- EDA and “knowing the data”
  data quality and cleaning, data matching & fusing
  missing values (NAs)
  identifying & investigating alternative explanations

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
  - data collection provenance & metadata
- EDA and “knowing the data”

  data quality and cleaning, data matching & fusing

  missing values (NAs)

  identifying & investigating alternative explanations
- Understanding randomness, variability & uncertainty

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
  - data collection provenance & metadata
- EDA and “knowing the data”
  data quality and cleaning, data matching & fusing
  missing values (NAs)
  identifying & investigating alternative explanations
- Understanding randomness, variability & uncertainty
- Conditional dependence & heterogeneity

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
  - data collection provenance & metadata
- EDA and "knowing the data"
  data quality and cleaning, data matching & fusing
  missing values (NAs)
  identifying & investigating alternative explanations
- Understanding randomness, variability & uncertainty
- Conditional dependence & heterogeneity
- Dimension reduction & variable selection & sparsity

Statistical Concepts

- mapping the general question to a statistical framework
- scope of inference, sampling, biases, limitations.
  - data collection provenance & metadata
- EDA and “knowing the data”
  data quality and cleaning, data matching & fusing
  missing values (NAs)
  identifying & investigating alternative explanations
- Understanding randomness, variability & uncertainty
- Conditional dependence & heterogeneity
- Dimension reduction & variable selection & sparsity
- spurious relationships & multiple testing

Statistical Concepts

Statistical Concepts

- Parameter estimation versus “black box” prediction/classification

Statistical Concepts

- Parameter estimation versus “black box” prediction/ classification
- diagnostics: residuals & comparing models and understanding their similarities and differences
  - EDA on sets of residuals

Statistical Concepts

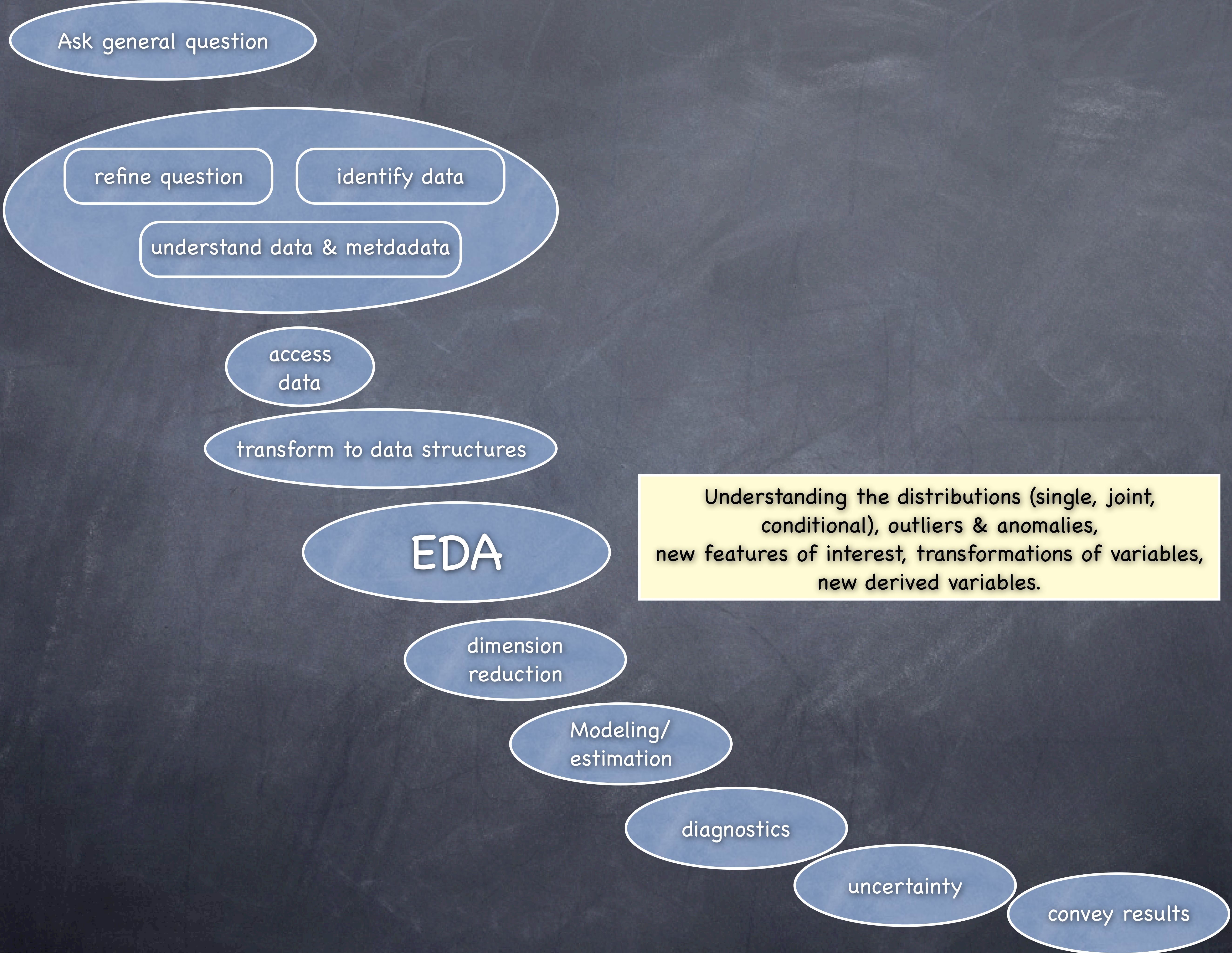
- Parameter estimation versus “black box” prediction/ classification
- diagnostics: residuals & comparing models and understanding their similarities and differences
  - EDA on sets of residuals
- quantifying the uncertainty of our “model”, e.g. bootstrapping, asymptotics, ...

Statistical Concepts

- Parameter estimation versus “black box” prediction/ classification
- diagnostics: residuals & comparing models and understanding their similarities and differences
  - EDA on sets of residuals
- quantifying the uncertainty of our “model”, e.g. bootstrapping, asymptotics, …
- sampling structure and dependence for data reduction & bootstrapping (e.g. spatio-temporal)

Statistical Concepts

- Parameter estimation versus “black box” prediction/ classification
- diagnostics: residuals & comparing models and understanding their similarities and differences
    - EDA on sets of residuals
- quantifying the uncertainty of our “model”, e.g. bootstrapping, asymptotics, ...
- sampling structure and dependence for data reduction & bootstrapping (e.g. spatio-temporal)
- statistical accuracy versus computational complexity/ efficiency - approximate answers.



Ask general question

refine question

identify data

understand data & metdadata

access data

transform to data structures

EDA

Understanding the distributions (single, joint, conditional), outliers & anomalies, new features of interest, transformations of variables, new derived variables.

dimension reduction

Modeling/ estimation

diagnostics

uncertainty

convey results

Computational Concepts

- Accessing data
  - Web services, REST, OAuth,
  - XML, JSON, Databases & SQL
- Raw data manipulation (data munging)
  - Text manipulation, regular expressions
  - Shell tools & Pipes
  - Stemming, stop words, NLP
  - Unicode

Computational Concepts

- Data structures and storage
  - R/Python data frames
  - Ragged, unstructured objects
  - Databases
    - relational model & SQL
    - NoSQL for semi-structured instances
    - Lucene (Solr/elasticsearch) for text search
    - Hive, HBase, Pig
    - XQuery?
- Depends on what we are going to do with the data and where?
- Filtering, summaries, conditional plots, ...

Computing Concepts

- Visualization at different stages
  - Exploratory Data Analysis
  - Diagnostic analysis
  - Conveying final results
- Static plots
- Interactive, animated dashboards
  - SVG, Javascript, D3, JS libraries, CSS, HTML
    - Create in, e.g. R, and annotate or

Computational Concepts

- Computationally intensive methods
  - modeling and uncertainty
    - Statistical & Machine Learning algorithms
    - Bootstrapping
- Parallel & Distributed Computing
  - Specialized - e.g. R multicore/cluster
  - Hadoop & Map Reduce, Pig, Hive, Spark, ...
  - Mahout
  - Threads - POSIX, Java, ...
  - GPUs

Computing Concepts

- Data-locality and scheduling in parallel computations is critical.
- Transition from local to HDFS can be significant bump.
- Many choices and technical issues in a complex, changing landscape of technologies.

- Important to keep computations high-level and say "what to do, not how to do it"
    - Leave "smarts" to infrastructure - compilation, job and data allocation & scheduling.
- We have been increasingly successful/productive using high-level languages such as R, MATLAB, Python. We can make these and new languages smarter to gain the efficiency, with prescriptive code that can be adapted to new platforms and computational models.

Choices and Alternatives

- At each step, there are many different choices.
- Evaluating which approach to use requires
  - experience and familiarity with the technologies
  - their pros and cons,
  - understanding of the lifetime and goals of the project - e.g. maintainability, one-off/ad hoc, skills of other participants.
- Starting to evolve into software development and engineering principles.

Additional Concepts

- Combining statistical reasoning and computational skills
    - So students need to explicitly learn this
- Often, collaboration is vital
    - Again, students need to learn and practice this.

Take-aways

- Need to teach computing/programming much earlier to equate with mathematics.
- Basic principles of statistical inference are essential to interpreting data.
- Students need authentic experiences in “big data” problems.
  - Need good case studies, internships, novel courses that combine techniques and applications.

Paths & Reproducability

- Many different paths of exploration.
  - The process is not just of the final results.
    - alternative ways of exploring the same ideas that confirmed or contradicted results.
    - dead ends.
- An important part of this process is reproducability of the process itself
  - the reasons for different decisions
  - different tasks that were not part of the “final” result

- THE END

- Do Statistics students have sufficient computational skills
  - How much computing is explicitly in the curricula? How effective is it in practice?
- Do Comp Sci. students adequately learn statistical concepts of randomness, uncertainty, scope of inference and answering questions with data?
  - or is the focus more on deterministic algorithms
- Do either of these groups of students get sufficient experience with authentic problem solving for data analysis and computing?

- Students don’t necessarily get to experience streaming data (velocity).
- Students don’t necessarily get to experience pure prediction/classification problems, but rather estimation and model selection problems.
- Many students don’t have basic programming skills or education, so starting from a significant void
  - programming concepts in high school (or earlier)
  - consistent exposure to high-level, practical/authentic computational reasoning and problem solving.

Inference & Sampling

- It is essential to know how the data were obtained relative to the population of interest in order to know the scope of inference.
    - Many uses of the data are biased. Many consumers of the data are not “focused” (aware) of this (even with small data) and are often distracted by the computational burden.

- Need to understand how much time to spend thinking about how to manage the data based on how it will be used.
  - If it is a one time analysis, generic solutions are fine.
  - If multiple passes of the data for different purposes, then it is important to determine how best to organize, store, structure the data for these purposes.

- Most students don’t have understanding of basic statistical concepts - i.e. inference, bias, across sample variability/randomness.
- Most students know “at most” one language and so try to force all computations into that framework.
- So “is big data” the real focus immediately or is it just “data” - period!

Computational Core

- Access data
- Ability to manipulate raw data into a form that is amenable to further modeling
- Visualization to explore data
- Be able to reason about how to organize data
- Familiarity with major paradigms and frameworks for parallel computing - MapReduce, Hadoop, high-level mechanism for distributed computing e.g. in R, thread model(s),

- Exploratory Data Analysis and Diagnostics are still as important as in “small” data.
    - Need means for interactive, exploration of data to understand what to explore next - iterative process.
- Estimation and model selection need to be developed for big data in addition to prediction/classification.