Probability Theory and Machine Learning in Science

Ji-Hun Kim

Seoul National University

ROSAEC Workshop 2013

Outline

1. Probability in Science
2. Machine Learning
3. Machine Learning in Action
4. Opportunities in Machine Learning

Plan

1. Probability in Science
2. Machine Learning
3. Machine Learning in Action
4. Opportunities in Machine Learning

What is Probability?

Two school of thoughts

Frequentist probability

Traditional definition.

$$P(\text{event}) \equiv \lim_{N \to \infty} \frac{N_{event}}{N_{total}}$$

Bayesian probability

Became popular in the 20th century.
Probability is generalization of logic in presence of uncertainty.

$$P(\text{event}) = \text{Degree of plausibility}.$$

What "a 30% chance of rain tomorrow" means?

Probability Theory as Extended Logic

from Probability Theory by E.T. Jaynes

Probability theories come from the three assumptions :

1. Degrees of plausibility are represented by real numbers.
2. As new information supporting the truth of a proposition is supplied, the number will increase continuously and monotonically. The deductive limit must be obtained where appropriate.
3. Consistency.
   - If a conclusion can be reasoned out in more than one way, every possible way must lead to the same result.
   - Propriety : The theory must take account of all information, provided it is relevant to the question.
   - Jaynes consistency : Equivalent states of knowledge must be represented by equivalent plausibility assignments.

From these, we derive

$$0 \leq p(x) \leq 1$$

$$p(A, B|C) = p(A|C)p(B|A, C) = p(B|C)p(A|B, C)$$

$$p(A|B) + p(\bar{A}|B) = 1$$

Resolving Ambiguities/Aesthetics

Given clues, there are many possible guesses:

Principle of maximum entropy

Given constraints, prefer a least informative conclusion;
i.e. probability distribution function of a maximum entropy.
For example,

- No prior information : Uniform distribution.
- Mean value : Poisson distribution.
- Mean and variance : Gaussian distribution.

Given data, there are many possible theory:

Occam's Razor

- Among the consistent theories, prefer a simplest one.
- It automatically follows from Bayesian model selection.

Rise of Probability in Science

- The law of nature itself is non-deterministic;
  e.g. Quantum mechanics is probabilistic theory.
- Bell's theorem : no latent variable can remove the randomness of QM.
- The uncertainty principle of QM: impossible to measure position and speed simultaneously and exactly.
- Non-linear dynamics : tiny difference in initial condition can yield huge deviation in the long-term future.
- For a complex system, what we can study is its statistical property.
- The present is a unique event. It is often impossible to repeat experiments.

Fortunately, it is not an exact state of system(each proton in a cell), but its statistical property(a human) that we are usually interested in.

When to use Machine Learning?

Probability theory alone is often insufficient to solve interesting problems.

- Many problems are too difficult to solve from first principles:
  - Listen/speak Korean
  - Real estate price(such as an apartment or the 63 building).
  - Higgs jets vs gluon jets (in particle physics).
  - Definition of life from quantum mechanics and chemistry.
- Induction, which requires data, complements deductive reasoning.
- Abstraction is lossy data compression; efficient encoding builds on sample distribution.
- Insufficient data : when it is high dimensional, it is impossible to collect sufficient sample, say 2^{100}. How to conclude from incomplete data?
- Human also needs experience to learn something :
  e.g. human walks. Who can write a text on how to walk, by the way?

Machine learning is about learning from data.

Plan

1. Probability in Science
2. Machine Learning
3. Machine Learning in Action
4. Opportunities in Machine Learning

What is Machine Learning?

Supervised machine learning

- From training samples of (label, data), assign a label to a new data.
- Applications : classification, spam filtering, real estate pricing, movie rating, book recommendation, ...
- Algorithms : k-nearest neighbor, neural network, support vector machine, ...

Unsupervised Learning

- Assign a label to unlabelled data; *i.e.* no training samples.
- Applications :
  - Structure finding: life vs things, animal vs plants, , ...
  - Text topic modeling : extract topics from texts.
  - Lossy data compression.
  - Label generation for supervised learning.
- Algorithms : Clustering, Auto-encoder, ...

Three Elements of Machine Learning

Machine Learning = Representation + Evaluation + Optimization

0. Data gathering : remember we often short of data, not algorithms!
   - Data pre-processing : noise correction, parsing, ...
1. Representation
   - Feature extraction : how to represent data?
   - Hypothesis space : models or classifier we consider.
2. Evaluation
   - Evaluation function : a measure of goodness of models
   - Parametric vs non-parametric models.
3. Optimization
   - Strategy to find a best model within the hypothesis space.
   - Monte-Carlo method/parallel tempering, stochastic gradient, ...

Supervised Learning

Decision Tree

0. We have N-dimensional data.
1. Select a coordinate and slice it for best discrimination.
2. For each segment, repeat the step 1 until getting a desired data purity or the tree reaches a depth limit.

Random Forests

1. Randomly select a subset of coordinates.
2. Make a decision tree with the subset.
3. Repeat the step 1 and 2, N times.
4. Do majority rule for a final decision.

Neural networks, support vector machine, k-nearest neighbor algorithm, ...

Unsupervised Learning

Clustering

Grouping similar items. Need a distance measure.

- k-means clustering : Starts with K random means. Assign each point to its nearest mean and move the mean to a center of its cluster.
- Hierarchical clustering : Do merging a pair of points of a minimum distance until every clusters are apart.
- ...

Blind signal separation

Lossy compression by dimensional reduction.

- Principal component analysis : define axis so that align points with the axis.
- Independent component analysis : each axis is statistically independent .
- ...

Unsupervised Learning and Hierarchical Structure in Nature

Life in nature :

- Life := Plant, Animal.
- Animal := Vertebrate, Invertebrate.
- Vertebrate := Reptilia, Mammal, ...
- Mammal := Dog, Cat, ...
- Dog := Jindo, Shepard, Poodle, ...

How human brain(visual cortex) recognize image? A current guess is :

- Images $\equiv$ cells('pixels') in retina.
- Pixels $\rightarrow$ basic shapes such as lines, points, brightness, colors
- Lines, points, brightness, colors $\rightarrow$ edges, contrasts
- Edges, contrasts $\rightarrow$ objects
- Objects $\rightarrow$ complex objects(on a background)

How we can learn these from high dimensional data?
Discover representation(*e.g.* dimensional reduction),
and structure finding(Hierarchical clustering).

Parametric vs non-parametric algorithms

Parametric algorithms

- has a fixed number of model parameter
- easier to make it computationally efficient.
- learning capacity is limited.
- neural network, K-means clustering, ...

Non-parametric algorithms

- a number of model parameter grows as data size increases
- a way to go for human-like learning system.
- Support vector machine, Hierarchical clustering, ...

Choosing Machine Learning Algorithm

Each algorithm has many options and parameters to adjust.
It is like medicines :

- no best, general solution
- many (imperfect) solutions to problems,
- difficult for non-expert to use them correctly and efficiently.
- we need experts(or expert systems?) to configure and combine them.

Diversity helps for better solutions.

- Which is the best algorithm? Depends on problems.
- No Free Lunch theorem : Optimization/Supervised Machine Learning
- Which algorithm to use? All of them!

Recall that for complex problems in real world, each expert often have a distinct opinion, too. Many machine learning system in real world often employ multiple algorithms with majority rule.

Plan

1. Probability in Science
2. Machine Learning
3. Machine Learning in Action
4. Opportunities in Machine Learning

Distance measures matter

Point

x = (age, height, weight)

To cluster points, we need a distance measure.

Distance - I

$$d^2 = (a_1 - a_2)^2 + (h_1 - h_2)^2 + (w_1 - w_2)^2$$

Distance measures matter

Point

x = (age, height, weight)

To cluster points, we need a distance measure.

Distance - I

$$d^2 = (a_1 - a_2)^2 + (h_1 - h_2)^2 + (w_1 - w_2)^2$$

In fact, a_i, h_i, w_i are different types.
Lesson : dimensions (or types) are important.

Distance measures matter

Point

x = (age, height, weight)

To cluster points, we need a distance measure.

Distance - II

Let's normalize them correctly.

$$d^2 = \frac{(a_1 - a_2)^2}{\sigma_a^2} + \frac{(h_1 - h_2)^2}{\sigma_h^2} + \frac{(w_1 - w_2)^2}{\sigma_w^2}$$

What normalization scales should be?

Distance measures matter

Point

x = (age, height, weight)

To cluster points, we need a distance measure.

Distance - III

Scaling is a global transformation. What about a local transformation?

$$d^2 = \frac{(a_1 - a_2)^2}{\sigma_a(a_1, a_2)^2} + \frac{(h_1 - h_2)^2}{\sigma_h(h_1, h_2)^2} + \frac{(w_1 - w_2)^2}{\sigma_w(w_1, w_2)^2}$$

Conceptually simple, but symbolically complex. How to simplify this?

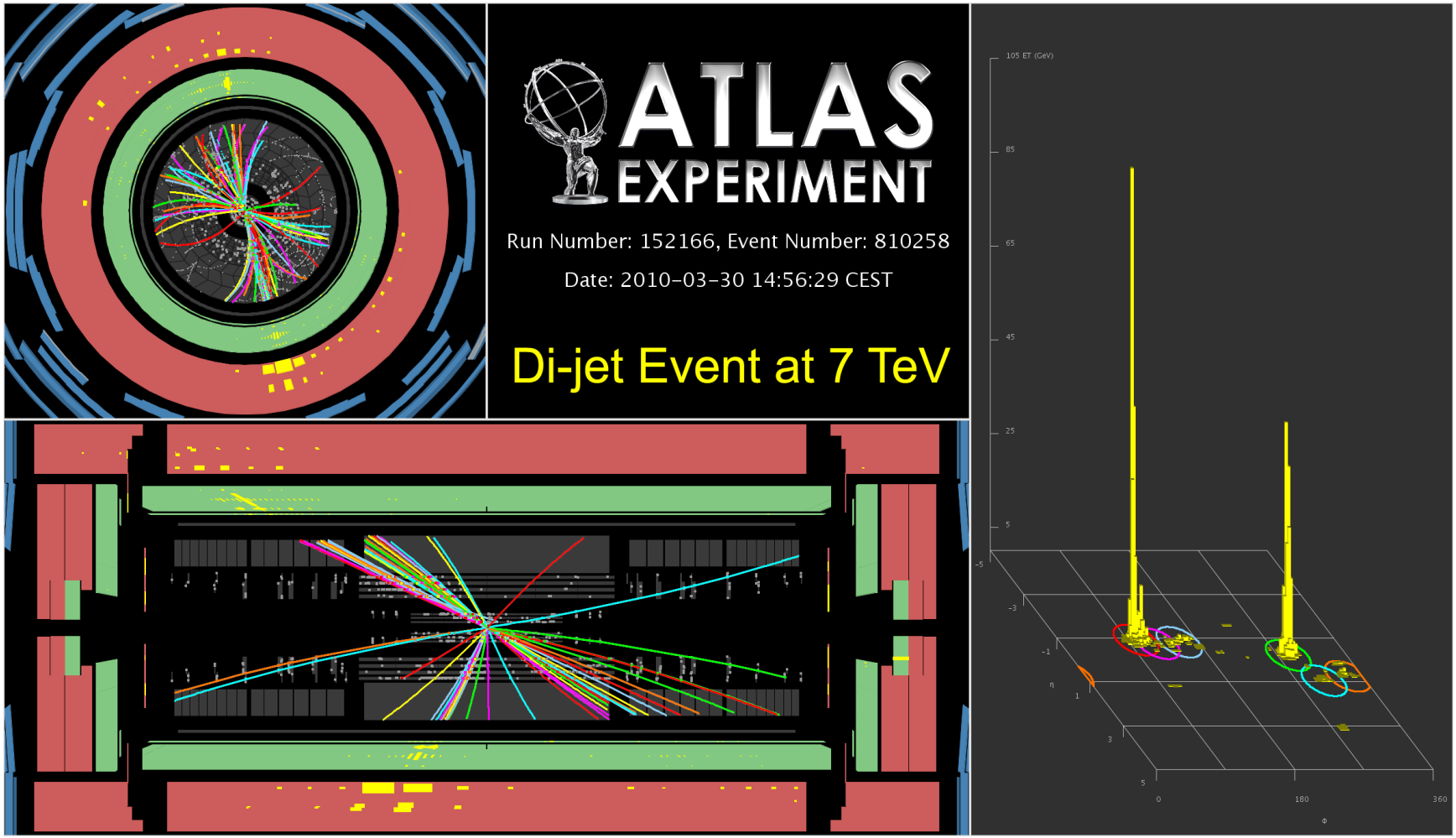
CERN LHC

A proton-proton collider.

Four detectors are connected with a large ring .

Clustering examples from high energy physics - I

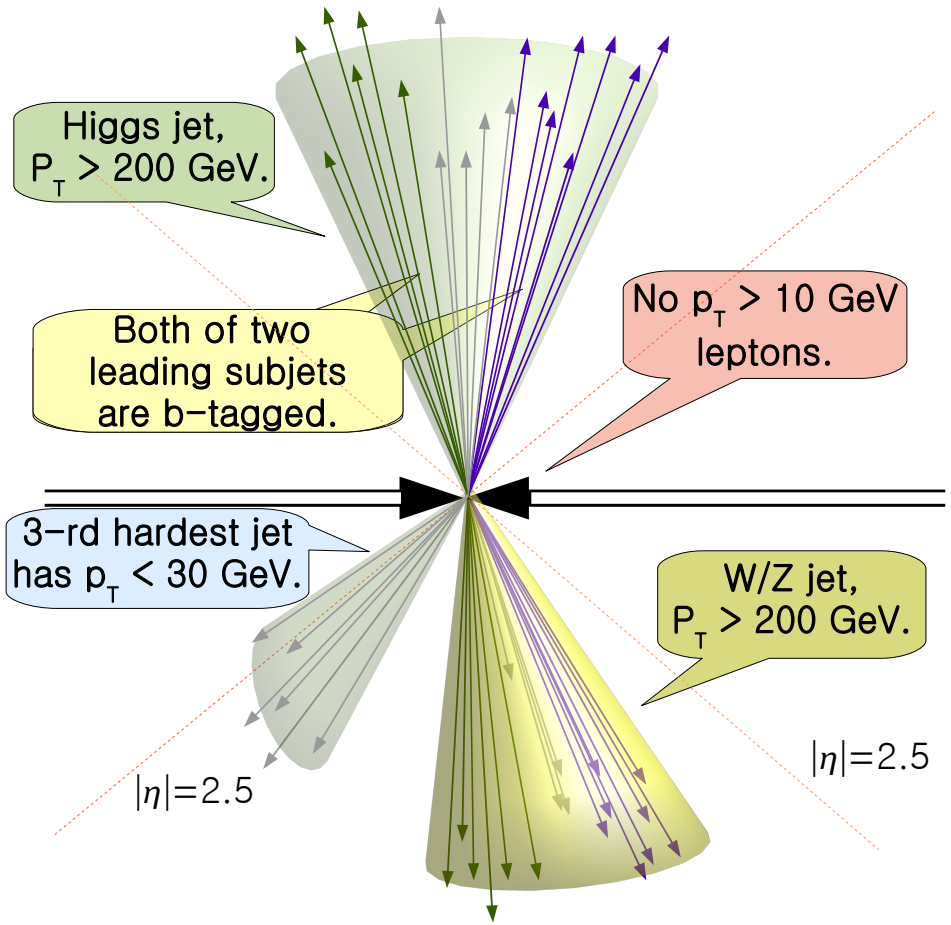
A proton-proton collision event



We call a collimated bunch of hadrons as a hadron jets. A hadron jet comes from a colored particle.

Clustering examples from high energy physics - II

Finding events with specific clusters



Both of two leading jets are required to satisfy :

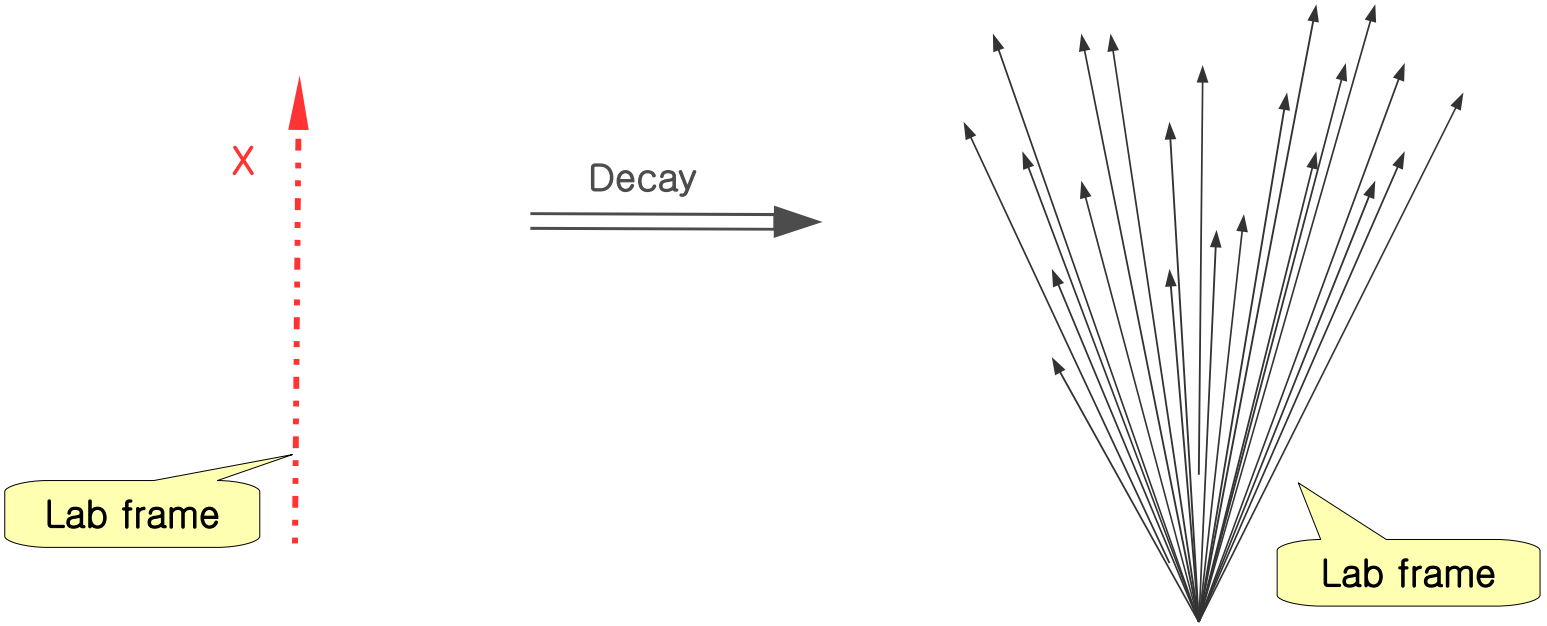
- $\tau_2^{rest} < 0.08$ cut.
- $\cos\theta_s < 0.8$ cut.

And two leading subjets of the Higgs candidate jets are required to be b-tagged.

Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

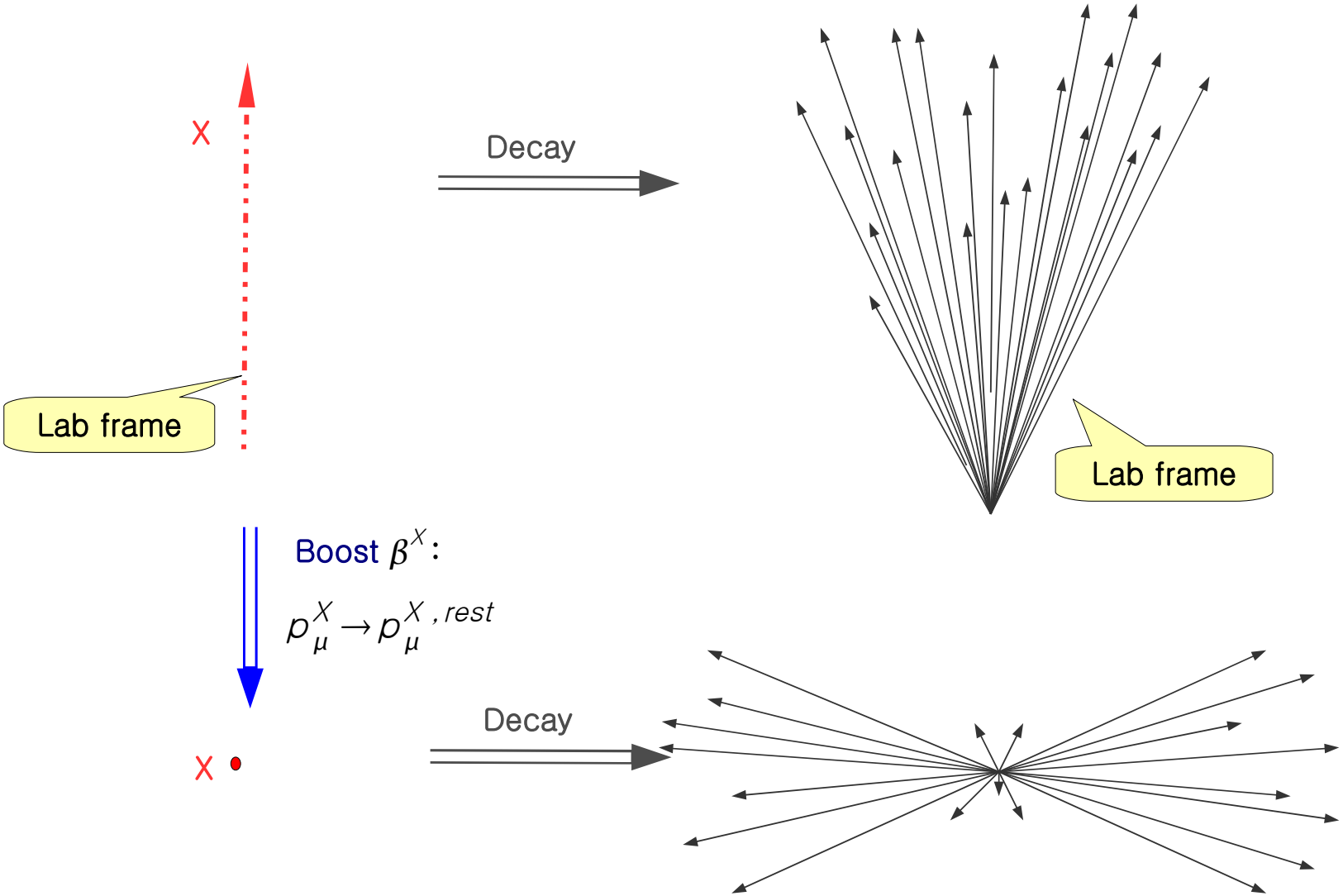
To find deformed object, define a local, deformed distance measure.



Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

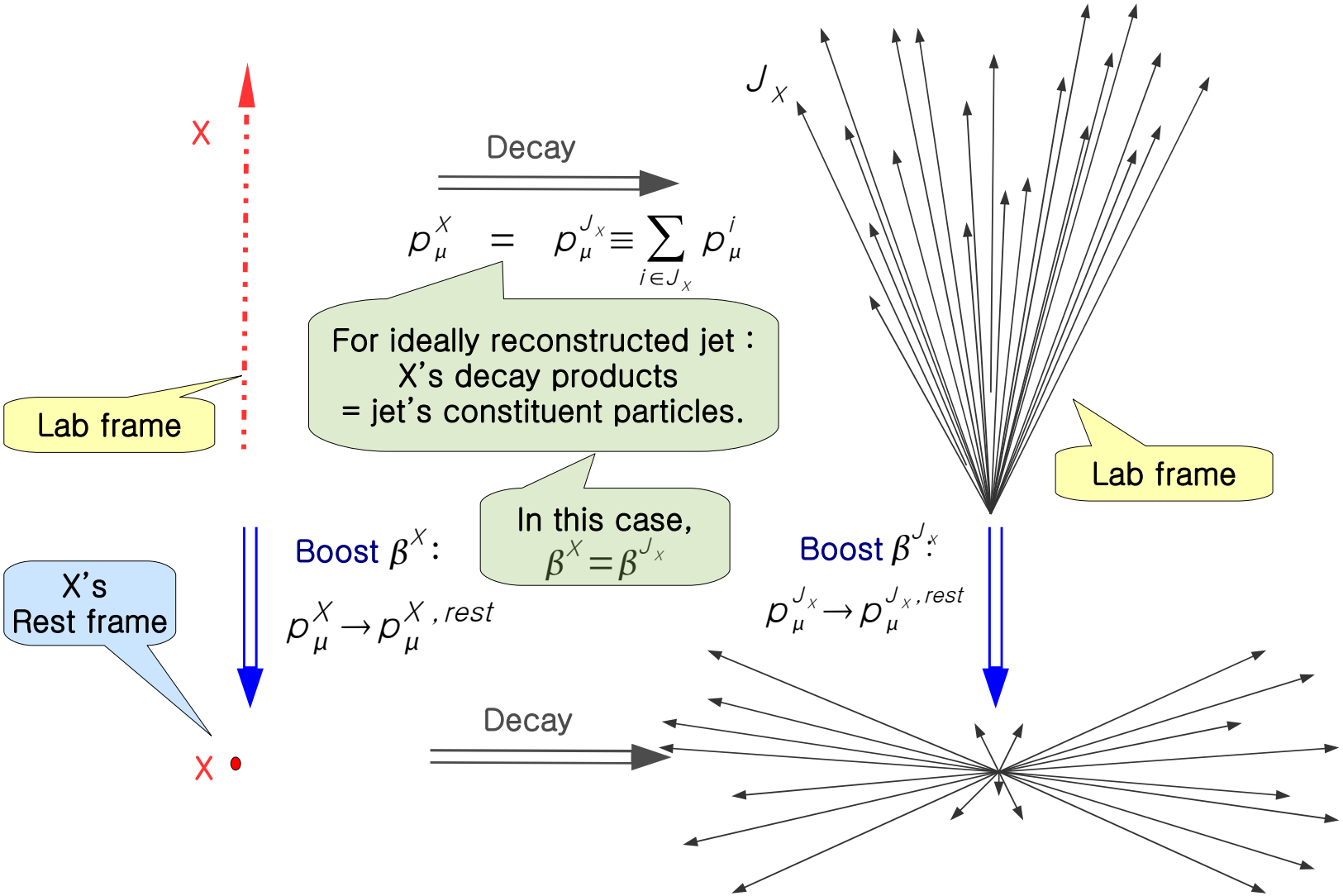
To find deformed object, define a local, deformed distance measure.



Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

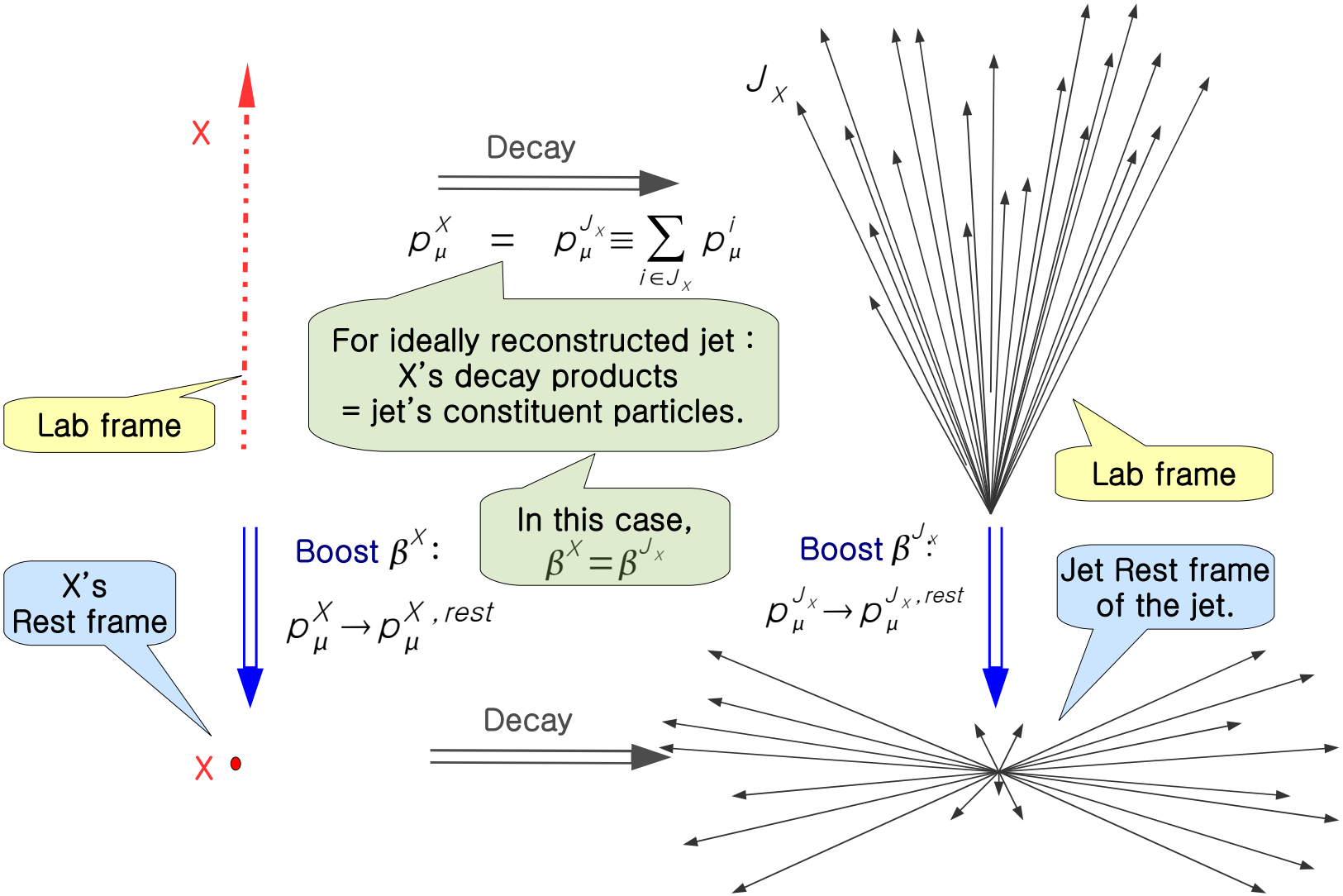
To find deformed object, define a local, deformed distance measure.



Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

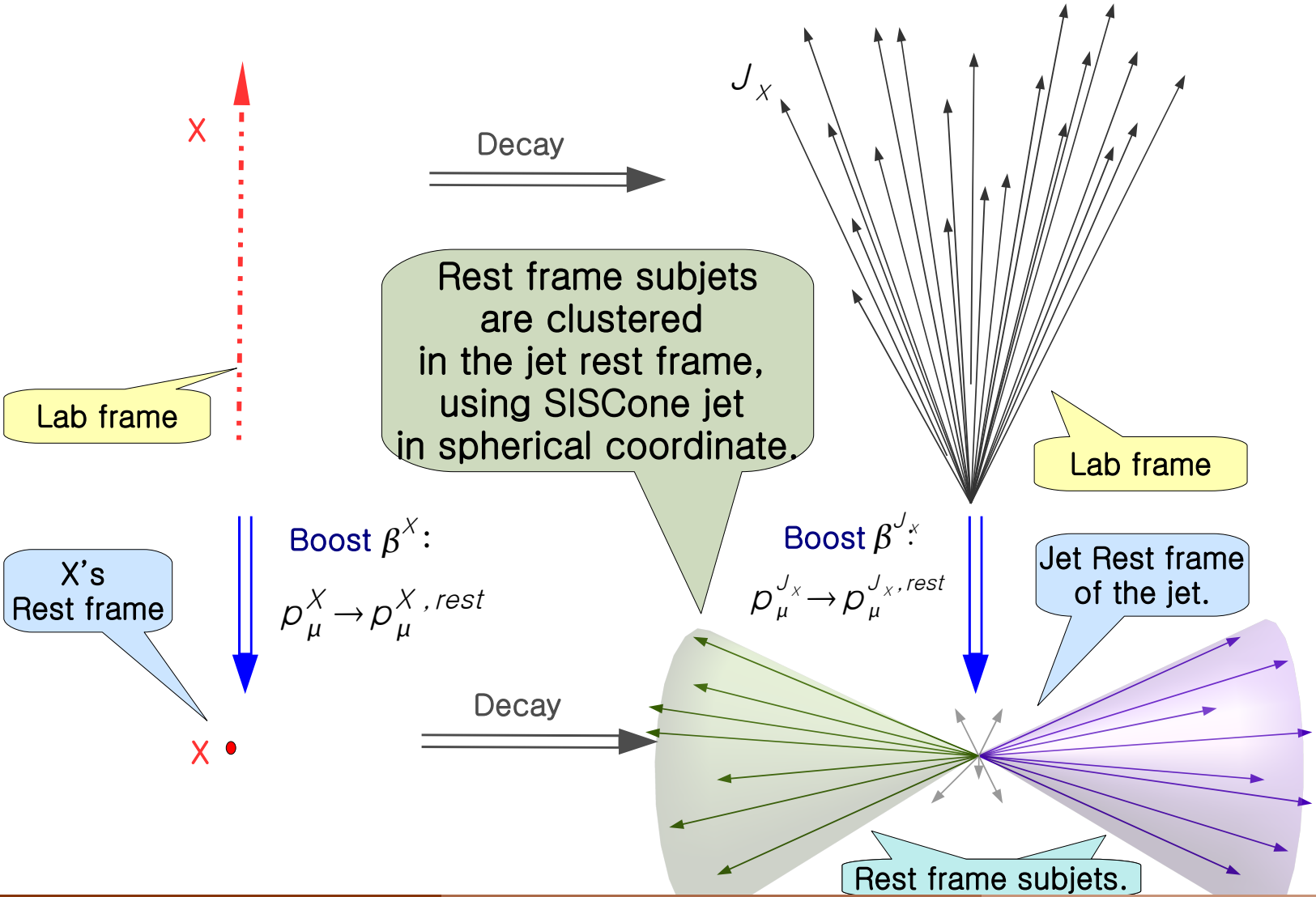
To find deformed object, define a local, deformed distance measure.



Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

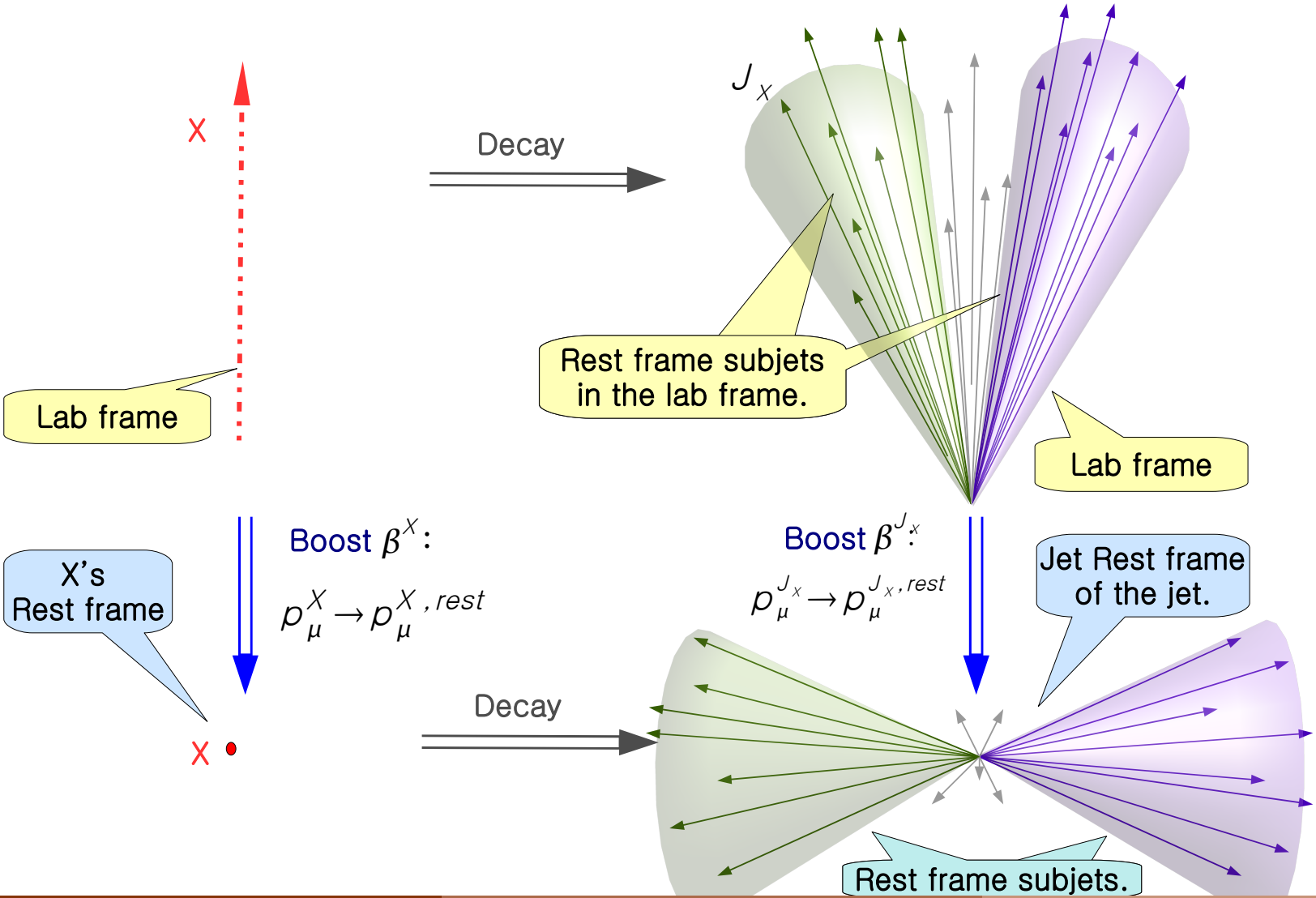
To find deformed object, define a local, deformed distance measure.



Clustering examples from high energy physics - III

Finding a cluster with N sub-clusters

To find deformed object, define a local, deformed distance measure.



Plan

1. Probability in Science
2. Machine Learning
3. Machine Learning in Action
4. Opportunities in Machine Learning

Computer Science Topics in Machine Learning

- Distributed file system : Hadoop, Apache Spark, ...
- Distributed computing :
  MPI(too low level) vs PGAS models(IBM X10, Cray Chapel, ...).
- Database : SQL query? Distributed? ACID?
- Avoid over-specification : Declarative over imperative. Critical for optimization, auto-tuning, auto-configuration.
- Optimization for concurrency :
  e.g. addition + association rule = map-reduce addition.
- Vectorization : For now, human (energy) resource is too expensive to use(not use) SIMD.
- Heterogeneous computing : CPU and/or GPU?
- Scaling at large : minimizing communication cost,
  topology aware algorithms.
- How computer language can help to check numerical program?

Open Questions in Machine Learning

- How to decide whether there are clusters?
- How to embed symmetries of data?
  - Data have various symmetries : rotation, re-scale, ...
  - How we can compare, say, images of different size?
  - In fact, humans are not very good at this.
    (can you read a book upside down without difficulty?).
- How to select a correct algorithm for a given question?
- A human brain does not need million images to learn 'animal'. How?
- Why a brain(~ 10 Watts) simulator would consume a few GW?
- How to devise parallel and scalable algorithms, instead of parallelizing algorithms.

Opportunities in Machine Learning

Science needs data.
Historically, theory innovations often follow experimental data explosion.

- Data :
  - Internet traffic, storage, contents : doubled in every few years.
  - Rise of public data sets.
- Demands :
  - Data size will continue to grow exponentially; people want to use them.
  - Directory/file will obsolete soon; what will be the successor? Key-value store. How to generate keys?
- Tools :
  - Computing power and efficiency are also improving exponentially.
  - Rise of distributed computing
  - Self-check and self-management tools are rapidly evolving.
- Theory :
  - The only thing missing at this point.
  - machine learning?!