This question seems subjective, but I'll try to answer it:

1. Deep learning [5] seems to be getting the most press right now. It is a form of a Neural Network (with many neurons/layers). Articles are currently being published in the New Yorker [1] and the New York Times[2] on Deep Learning.

2. Combining Support Vector Machines (SVMs) and Stochastic Gradient Decent (SGD) is also interesting. SVMs are really interesting and useful because you can use the kernel trick [10] to transform your data and solve a non-linear problem using a linear model (the SVM). A consequence of this method is the training runtime and memory consumption of the SVM scales with size of the data set. This situation makes it very hard to train SVMs on large data sets. SGD is a method that uses a random process to allow machine learning algorithms to converge faster. To make a long story short, you can combine SVMs and SGD to train SVMs on larger data sets (theoretically). For more info, read this link[4].

3. Because computers are now fast, cheap, and plentiful, Bayesian statistics is now becoming very popular again (this is definitely not "new"). For a long time it was not feasible to use Bayesian techniques because you would need to perform probabilistic integrations by hand (when calculating the evidence). Today, Bayesist are using Monte Carlo Markov Chains[6], Grid Approximations[7], Gibbs Sampling[8], Metropolis Algorithm [13], etc. For more information, watch the videos on Bayesian Networks on Coursera. or a read these books [11], [12] (They're da bomb!!!)

4. Any of the algorithms described in the paper "Map Reduce for Machine Learning on a Multicore"[3]. This paper talks about how to take a machine learning algorithm/problem and distribute it across multiple computers/cores. It has very important implications because it means that all of the algorithms mentioned in the paper can be translated into a map-reduce format and distributed across a cluster of computers. Essentially, there would never be a situation where the data set is too large because you could just add more computers to the Hadoop cluster. This paper was published a while ago, but not all of the algorithms have been implemented into Mahout yet.

Machine learning is a really large field of study. I am sure there are a lot more topics but these are four I definitely find interesting.

[1] [Is “Deep Learning” a Revolution in Artificial Intelligence?](http://www.newyorker.com/online/blogs/newsdesk/2012/11/is-deep-learning-a-revolution-in-artificial-intelligence.html)

[2] [Scientists See Advances in Deep Learning, a Part of Artificial Intelligence](http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?_r=0)

[3] [http://www.cs.stanford.edu/peopl...](http://www.cs.stanford.edu/people/ang//papers/nips06-mapreducemulticore.pdf)

[4] [Kernel Approximations for Efficient SVMs (and other feature extraction methods) [update]](http://peekaboo-vision.blogspot.com/2012/12/kernel-approximations-for-efficient.html)

[5] [Deep learning](http://en.wikipedia.org/wiki/Deep_learning)

[6] [Markov chain Monte Carlo](http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo)

[7] [http://www.people.fas.harvard.ed...](http://www.people.fas.harvard.edu/~plam/teaching/methods/grid/grid_print.pdf)

[8] [Gibbs sampling](http://en.wikipedia.org/wiki/Gibbs_sampling)

[9] [Coursera](https://www.coursera.org/course/pgm)

[10] [Kernel trick](http://en.wikipedia.org/wiki/Kernel_trick)

[11] [Doing Bayesian Data Analysis](http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/)

[12] [Amazon.com: Probability Theory: The Logic of Science (9780521592710): E. T. Jaynes, G. Larry Bretthorst: Books](http://www.amazon.com/Probability-Theory-Science-T-Jaynes/dp/0521592712)

[13] [Metropolis–Hastings algorithm](http://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm)

I can answer this question from the other side of the interviewing table since I’m one of Palantir’s machine learning interviewers. One thing to understand going into this is that there are many different skill-sets associated with “data science” or “machine learning” and so making a fair and complete set of questions is challenging. In general, Palantir likes its ML candidates to be well-rounded.

Our interview process is constantly evolving but it currently involves three main facets: coding, machine learning theory, and end-to-end problem solving.

Coding is the most straightforward. You will be given some real data and asked to implement, from scratch, a machine-learning algorithm on that data. You should be comfortable coding several different kinds of algorithms depending on the task you are given. The point of this interview is to assess how well you would perform in a practical setting and to ensure that you are comfortable with ML outside of just using common libraries.

The ML theory section tests if you paid attention in your ML courses and whether you understand how that material connects to the real world. You should have a good breadth of knowledge on things like: statistical assumptions of models, why certain models work the way they do and how to deal with data irregularities.

Finally end-to-end interviews test your practicality. You will be given a hypothetical scenario with some input data and some goal. You will be asked to design in detail a machine-learning pipeline going from raw data to usable predictions. You will have to make practical assumptions and reasonable choices. There is no single right answer to these questions. but people who have designed ML systems in a practical setting will have a distinct advantage here.

For context on what this team’s goals are why we interview the way we do I’ll direct you to another quora response written by one of my coworkers:

[Anirvan Mukherjee's answer to How important is machine learning to what Palantir is doing?](https://www.quora.com/How-important-is-machine-learning-to-what-Palantir-is-doing/answer/Anirvan-Mukherjee?srid=iOkn&share=1)

In [Matt](https://www.quora.com/Matt-Gordon-5)'s earlier answer to this question, he wrote that “Palantir doesn’t do much ML…yet.” That was true in late 2012, but things have changed. As the company has grown and expanded into new domains, **machine learning has become absolutely central to the work Palantir is doing with many of its biggest customers**. As a member of Palantir’s machine learning team, I wanted to write an updated answer that sheds some light on how and why Palantir is concentrating and scaling its machine learning capabilities.

**The primary driver of Palantir’s focus on machine learning is its expanded footprint at commercial enterprises**. These businesses are trying to discover creative new ways to leverage their existing data to:

* Generate new streams of revenue
* Improve operations
* Protect themselves against risk

In many cases, the customer’s existing data was not collected for any of these three purposes. (Much of an organization’s useful data is simply a byproduct of specific business operations—what we sometimes call “data exhaust.”) And in all cases the problems themselves are fundamentally ambiguous—there are no straightforward, fully automate-able ways to monetize transactional data, diagnose supply chain inefficiencies, or identify and stop fraud, insider threats, or unauthorized activity (e.g., rogue trading).

Palantir has found that machine learning methods are especially effective at removing friction between business stakeholders, users of our products, and their data. When the friction is gone, our customers can more easily use their data to realize new business outcomes.

We are actively building machine learning capabilities into our core products, as well as deploying these enhancements against problems in health care, law enforcement, and other areas.

## Connecting top-down and bottom-up analysis

Palantir’s customers know they have a lot of micro-level data that they haven’t tapped. They want to leverage it to tackle macro-level problems. But it’s not always obvious what the “right” level of granularity is for an underspecified problem. Should a user be looking at entities? Categories? Clusters? In highly ambiguous contexts such as these, users need a way to identify patterns of *local* relationships, trends, or anomalies that have some meaningful *global* consequence. The question isn’t “Do we want top-down or bottom-up analysis?” but “**What’s the right resolution for *this* problem?**”—and in some cases “**Just what *is* the problem, anyway?**”

Machine learning provides ways to identify and surface structure at various resolutions in a unified way. By unifying analysis on multiple levels, it becomes easier for analysts to quickly scope out a problem, identify possible factors of interest, and estimate what the RoI of tackling the problem might be.

This is especially valuable to us in the face of dynamic use cases or adaptive adversaries, where analytical unification via machine learning gives analysts the flexibility and dexterity to work with their data in evolving ways.

## Lead generation

As my colleague [Yael](https://www.quora.com/Yael-Shrager) wrote in her answer to “[How will you improve the Palantir Gotham platform?](https://www.quora.com/Palantir-Technologies/How-will-you-improve-the-Palantir-Gotham-platform/answer/Yael-Shrager?srid=u1Pb&share=1), ” automated and semi-automated analysis is becoming a more important area of focus for our product team. In particular, automated lead generation is now a major component of many use cases, such as detecting healthcare provider fraud or unauthorized trading. Here, semisupervised learning works alongside rule-based business logic to triage data and event streams, pointing analysts towards known-suspicious behavior, anomalies, and unexpected change points.

From a product perspective, lead generation allows analysts to focus their attention on a manageable set of high-signal data so they can cover more ground quickly. Over time, collected explicit and implicit feedback both improves the lead generator and enriches the data.

We’re still never going to make a “find terrorist” button, but it’s becoming much easier to use machine learning for lead generation—to build a “find suspicious bank transfers” button that surfaces suspicious cases for further investigation by human experts.

## Integrating legacy data

Integrating disparate data into a unified analytic environment is a critical first step to deriving insights from it. But data integration is especially difficult when customers are trying to leverage their legacy data—including their “data exhaust”—against novel problems. **Higher impedance between the dataset and the use case makes data integration more complex**.

For example, one of our customers is a retailer that wanted to optimize its inventory to improve its demand forecasting and reduce wastage by using insights from its sales data. This data comes from a database designed for sales transaction record keeping, so information critical to the problem (e.g. which shipment an item was a part of, an item’s expiry date) is often missing or sparsely documented. Useful metadata, such as the product categorization hierarchy, were modeled in a way that was optimized for procurement rather than demand forecasting.

In situations like this, machine learning can help in several ways.

**Entity resolution and data enrichment:** We use machine learning to perform entity resolution, which allows us to jointly leverage multiple datasets that weren’t designed to be joined, construct canonical identifiers based on noisy information, and identify natural hierarchy. Entity resolution not only facilitates better *integration* of existing data, it can also *enrich* existing data with new datasets or algorithmically inferred structure and characteristics. In some cases, automated entity resolution even helps us detect and fix errors in source data, like when a record with missing/incorrect/out of date information is correctly resolved with another record containing the right information.

We only employ automated entity resolution, however, where it’s appropriate (e.g., when some false positives are acceptable). Whenever the costs of false positives are intolerably high, as is often the case in the public sector, users review and resolve entities manually.

**Value Interpretation:** Customers often want to integrate data into Palantir, but are missing adequate documentation on the meanings and types of each field. This can be problematic—an arbitrary integer field with a cryptic name could be a code (categorical variable), a rounded continuous value (e.g., monetary amount), a count, an identifier, a phone number, a postal code, or something else. Until its true type is de-obfuscated, the data can’t be correctly identified and used by higher-level logic.

While humans are good at recognizing possible meanings of data, machine learning helps us perform value interpretation at scale. The result is that users can get off the ground with less upfront effort.

## What makes machine learning at Palantir different

This section is my shameless plug to anyone who’s read this far.

As an engineer on our machine learning team, I get to work with new kinds of data on a regular basis across a variety of domains. I personally have worked on problems in health care, consumer credit, law enforcement, insurance, energy, and more, and each customer’s data and perceived use case is different.

To me, this is an opportunity to **find structural similarities among many superficially different problems**, and develop algorithms and tools for solving the underlying problems. This also enables some amount of rigor—anything we develop can be assessed and refined in multiple environments with different datasets and assumptions.

I also see this as an opportunity to **add value quickly**. Our customers are some of the most significant institutions in the world, and we are exposed to their most critical data problems. Most of our customers have never successfully deployed effective machine learning at scale before, so our initial efforts often make a big difference.

If you’re interested in learning more about what we do, feel free to ask here or shoot me a message.

Our team is relatively small, but we’re actively hiring. If our work sounds interesting and you’re up to the challenge, we’d love to talk. You can apply/submit your resume[here](http://www.palantir.com/careers/OpenPosDetail?id=a0m80000005LTN1AAO).