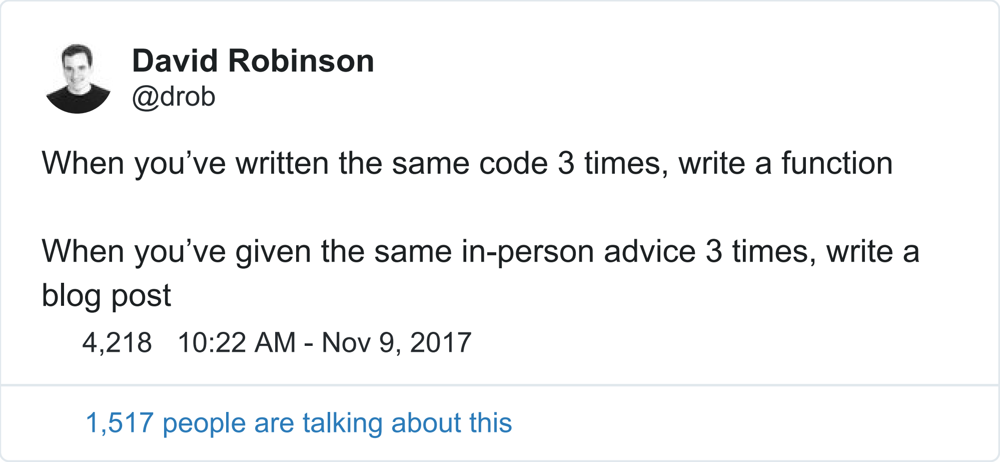
VARIANCE EXPLAINED

menu

What's the difference between data science, machine learning, and artificial intelligence?

When I introduce myself as a data scientist, I often get questions like “What’s the difference between that and machine learning?” or “Does that mean you work on artificial intelligence?” I’ve responded enough times that my answer easily qualifies for my “rule of three”:



David Robinson
@drob

When you’ve written the same code 3 times, write a function

When you’ve given the same in-person advice 3 times, write a blog post

4,218 10:22 AM - Nov 9, 2017

1,517 people are talking about this

The fields do have a great deal of overlap, and there’s enough hype around each of them that the choice can feel like a matter of marketing. But **they’re not interchangeable**: most professionals in these fields have an intuitive understanding of how particular work could be classified as data science, machine learning, or artificial intelligence, even if it’s difficult to put into words.

So in this post, I’m proposing an *oversimplified* definition of the difference between the three fields:

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

To be clear, this isn’t a *sufficient* qualification: not everything that fits each definition is a part of that field. (A fortune teller makes predictions, but we’d never say that they’re doing machine learning!) These also aren’t a good way of determining someone’s role or job title (“Am I a data scientist?”), which is a matter of focus and experience. (This is true of any job description: I write as part of my job but I’m not a professional writer).

But I think this definition is a useful way to *distinguish* the three types of work, and to avoid sounding silly when you’re talking about it. It’s worth noting that I’m taking a descriptivist rather than a prescriptivist approach: I’m not interested in what these terms “should mean”, but rather how people in the field typically use them.

Data science produces insights

Data science is distinguished from the other two fields because its goal is an especially human one: to gain insight and understanding. Jeff Leek has an excellent definition of the types of insights that data science can achieve, including descriptive ("the average client has a 70% chance of renewing") exploratory ("different salespeople have different rates of renewal") and causal ("a randomized experiment shows that customers assigned to Alice are more likely to renew than those assigned to Bob").

Again, not everything that produces insights qualifies as data science (the classic definition of data science is that it involves a combination of statistics, software engineering, and domain expertise). But we can use this definition to distinguish it from ML and AI. The main distinction is that in data science there's always a human in the loop: someone is understanding the insight, seeing the figure, or benefitting from the conclusion. It would make no sense to say "Our chess-playing algorithm uses data science to choose its next move," or "Google Maps uses data science to recommend driving directions".

This definition of data science thus emphasizes:

- Statistical inference
- Data visualization
- Experiment design
- Domain knowledge
- Communication

Data scientists might use simple tools: they could report percentages and make line graphs based on SQL queries. They could also use very complex methods: they might work with distributed data stores to analyze trillions of records, develop cutting-edge statistical techniques, and build interactive visualizations. Whatever they use, the goal is to gain a better understanding of their data.

Machine learning produces predictions

I think of machine learning as the field of **prediction**: of "Given instance X with particular features, predict Y about it". These predictions could be about the future ("predict whether this patient will go into sepsis"), but they also could be about qualities that aren't immediately obvious to a computer ("predict whether this image has a bird in it"). Almost all Kaggle competitions qualify as machine learning problems: they offer some training data, and then see if competitors can make accurate predictions about new examples.

There's plenty of overlap between data science and machine learning. For example, logistic regression can be used to draw insights about relationships ("the richer a user is the more likely they'll buy our product, so we should change our marketing strategy") and to make predictions ("this user has a 53% chance of buying our product, so we should suggest it to them").

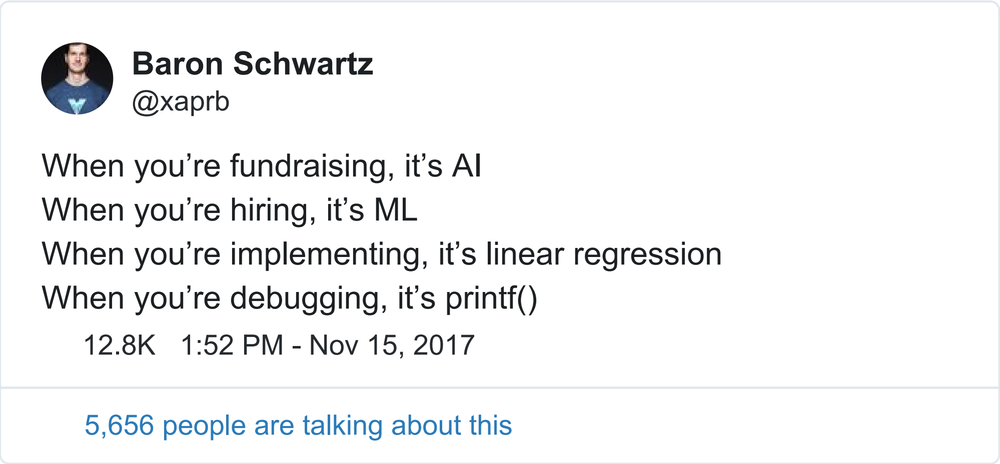
Models like random forests have slightly less interpretability and are more likely to fit the "machine learning" description, and methods such as deep learning are notoriously challenging to explain. This could get in the way if your goal is to extract insights rather than make predictions. We could thus imagine a "spectrum" of data science and machine learning, with more interpretable models leaning towards the data science side and more "black box" models on the machine learning side.

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Most practitioners will switch back and forth between the two tasks very comfortably. I use both machine learning and data science in my work: I might fit a model on Stack Overflow traffic data to determine which users are likely to be looking for a job (machine learning), but then construct summaries and visualizations that examine why the model works (data science). This is an important way to discover flaws in your model, and to combat algorithmic bias. This is one reason that data scientists are often responsible for developing machine learning components of a product.

Artificial intelligence produces actions

Artificial intelligence is by far the oldest and the most widely recognized of these three designations, and as a result it's the most challenging to define. The term is surrounded by a great deal of hype, thanks to researchers, journalists, and startups who are looking for money or attention.



Baron Schwartz
@xaprb

When you're fundraising, it's AI
When you're hiring, it's ML
When you're implementing, it's linear regression
When you're debugging, it's printf()

12.8K 1:52 PM - Nov 15, 2017

5,656 people are talking about this

This has led to a backlash that strikes me as unfortunate, since it means some work that probably *should* be called AI isn't described as such. Some researchers have even complained about the AI effect: "AI is whatever we can't do yet".[1] So what work can we fairly describe as AI?

One common thread in definitions of “artificial intelligence” is that an autonomous agent executes or recommends **actions** (e.g. Poole, Mackworth and Goebel 1998, Russell and Norvig 2003). Some systems I think should described as AI include:

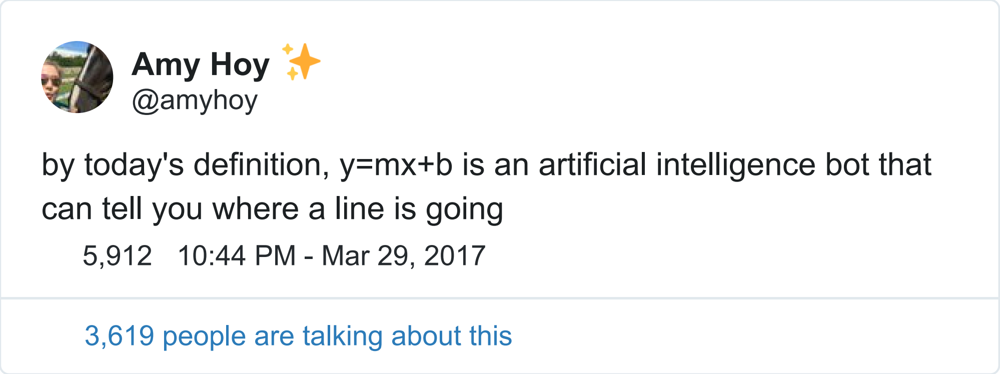
- Game-playing algorithms (Deep Blue, AlphaGo)
- Robotics and control theory (motion planning, walking a bipedal robot)
- Optimization (Google Maps choosing a route)
- Natural language processing (bots[2])
- Reinforcement learning

Again, we can see a lot of overlap with the other fields. Deep learning is particuarly interesting for straddling the fields of ML and AI. The typical use case is training on data and then producing predictions, but it has shown enormous success in game-playing algorithms like AlphaGo. (This is in contrast to earlier game-playing systems, like Deep Blue, which focused more on exploring and optimizing the future solution space).

But there are also distinctions. If I analyze some sales data and discover that clients from particular industries renew more than others (extracting an insight), the output is some numbers and graphs, not a particular action. (Executives might use those conclusions to change our sales strategy, but that action isn’t **autonomous**) This means I’d describe my work as data science: it would be cringeworthy to say that I’m “using AI to improve our sales.”

> *please*
>
> *please*
>
> *please do not write that someone who trained an algorithm has "harnessed the power of AI"*
>
> *— Dave Gershgorn (@davegershgorn) September 18, 2017*

The difference between artificial intelligence and machine learning is a bit more subtle, and historically ML has often been considered a subfield of AI (computer vision in particular was a classic AI problem). But I think the ML field has largely “broken off” from AI, partly because of the backlash described above: most people who to work on problems of prediction don’t like to describe themselves as AI researchers. (It helped that many important ML breakthroughs came from statistics, which had less of a presence in the rest of the AI field). This means that if you can describe a problem as “predict X from Y,” I’d recommend avoiding the term AI completely.



Amy Hoy ✨
@amyhoy

by today's definition, y=mx+b is an artificial intelligence bot that can tell you where a line is going

5,912 10:44 PM - Mar 29, 2017

3,619 people are talking about this

Case study: how would the three be used together?

Suppose we were building a self-driving car, and were working on the specific problem of stopping at stop signs. We would need skills drawn from all three of these fields.

- **Machine learning**: The car has to recognize a stop sign using its cameras. We construct a dataset of millions of photos of streetside objects, and train an algorithm to **predict** which have stop signs in them.

- **Artificial intelligence**: Once our car can recognize stop signs, it needs to decide when to take the **action** of applying the brakes. It's dangerous to apply them too early or too late, and we need it to handle varying road conditions (for example, to recognize on a slippery road that it's not slowing down quickly enough), which is a problem of control theory.

- **Data science**: In street tests we find that the car's performance isn't good enough, with some false negatives in which it drives right by a stop sign. After analyzing the street test data, we gain the **insight** that the rate of false negatives depends on the time of day: it's more likely to miss a stop sign before sunrise or after sunset. We realize that most of our training data included only objects in full daylight, so we construct a better dataset including nighttime images and go back to the machine learning step.

1. It doesn't help that AI is often conflated with **general AI**, capable of performing tasks across many different domains, or even **superintelligent AI**, which surpasses human intelligence. This sets unrealistic expectations for any system described as "AI". ↩
2. By "bots" here I'm referring to systems meant to interpret natural language and then respond in kind. This can be distinguished from *text mining*, where the goal is to extract insights (data science) or *text classification*, where the goal is to categorize documents (machine learning) ↩

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

✉ Email Twitter Github Stack Overflow

Subscribe

Your email

Subscribe to this blog

Recommended Blogs

- DataCamp
- R Bloggers
- RStudio Blog
- R4Stats
- Simply Statistics

What's the difference between data science, machine learning, and artificial intelligence? was published on January 09, 2018.

Comments Community 1 Login

♡ Recommend 40 Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

twoElectric • a year ago

When you've read three articles on a blog, subscribe to it.

17 ^ | ⌄ • Reply • Share ›

Chipmonkey75 • a year ago

I've struggled with defining Data Science for a while, and here's where I find most definitions falling short: what's the name to describe the people who _invent the algorithms_? To me, those are the true data scientists, and that's the work I do when I say I'm doing Data Science. Most people I see in business applying algorithms seem to me to be "Applied Data Scientists", perhaps, but much more often statisticians or analysts with up-to-date toolsets.

When we talk about scientists in other fields, we almost never use the term to apply to people who simply use the tools which other scientists have created. Lab Tests are invented by scientists but applied by doctors, nurses, phlebotomists... demanding, respectable, and advanced careers, but usually not labeled "scientist".

Even on the front lines, people with jobs named "data scientist" don't seem to do a lot of science -- not in a way that

see more

2 ^ | ⌄ • Reply • Share ›

Julio Trecenti • a year ago

Great post. Unfortunately, it seems like you've just relabeled descriptive, predictive and prescriptive analytics. I think that it is not so important to define what those things ARE. We should focus on how to DO data science. And IMO it is well described in Hadley/gareth's data science cycle (r4ds book).

2 ^ | ⌄ • Reply • Share ›

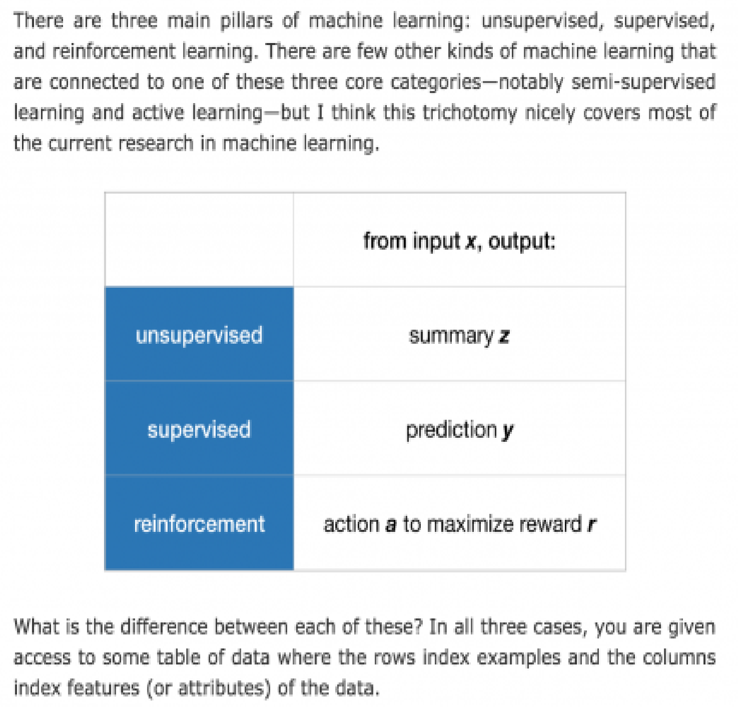
David Robinson Mod → Julio Trecenti • a year ago

I don’t think it’s accurate to say that the above definition of AI is equivalent to prescriptive analytics- no one would say that AlphaGo and Deep Blue were using prescriptive analytics to choose their next move. The algorithms, community and context simply differ.

Similarly, while ML does overlap with data science in predictive analytics, I think the former has grown much wider and deeper than that. (One could say “it should have been called predictive analytics from the start”, but I think treating it as a sub domain of DS would be underestimating the field.)

2 ⌃ | ⌄ • Reply • Share ›

Bernardo F. Nunes ➦ Julio Trecenti • a year ago



There are three main pillars of machine learning: unsupervised, supervised, and reinforcement learning. There are few other kinds of machine learning that are connected to one of these three core categories—notably semi-supervised learning and active learning—but I think this trichotomy nicely covers most of the current research in machine learning.

	from input *x*, output:
unsupervised	summary *z*
supervised	prediction *y*
reinforcement	action *a* to maximize reward *r*

What is the difference between each of these? In all three cases, you are given access to some table of data where the rows index examples and the columns index features (or attributes) of the data.

I agree with Julio T. Data science contains ML, which by definition contains DL. A much better alternative would be to follow Chris Wiggins' taxonomy of ML. Also, see the discussion here by Benjamin Recht. He uses two ways to define reinforcement learning: predictive analytics and control theory. In both, predictive analytics is responsible for actions.

1 ⌃ | ⌄ • Reply • Share ›

Athos Petri Damiani ➦ Julio Trecenti • a year ago

I agree.

⌃ | ⌄ • Reply • Share ›

Dean Abbott • a year ago

Artificial Intelligence
Statistical Learning
Pattern Recognition
Data mining
Predictive Analytics
Machine Learning
Data science
Artificial Intelligence

That’s what I’ve called what I’ve done over the past 30 years. I’ve essentially done the same kind of thing over my entire career

1 ⌃ | ⌄ • Reply • Share ›

Pranav • a year ago

How is data science different from Statistics?

1 ⌃ | ⌄ • Reply • Share ›

> **Pawan Khatri** ➦ Pranav • 3 months ago
>
> I guess a statistician who can program is a data scientist.
>
> 16 ⌃ | ⌄ • Reply • Share ›

Dan Ofer • a year ago

Really good summarization of the three "Domains" and some differences! Nice post :) . (I like the Leek slides too, haven't seen them before)

1 ⌃ | ⌄ • Reply • Share ›

godsaw • a year ago

very nice description of a touchy subject.

1 ⌃ | ⌄ • Reply • Share ›

Vishal Pouras • 6 months ago

A well crafted blog with easy language that gives precise knowledge. Example of self driving car made it even more interesting.

⌃ | ⌄ • Reply • Share ›

RobertWF • 7 months ago

In addition to David Robinson's comments, I'd add that data science (or statistical inference) is often concerned with identifying how much of the variance in the data is explained by the model, for example via R-squared statistics, and obtaining *unbiased* estimates of statistics (like regression coefficients) while minimizing the mean squared error. This means you need to correctly specify the probability distribution of the data, and then correctly interpret the p-values, both of which can be a major headache. Meanwhile, machine learning is more concerned with minimizing prediction error by tweaking both bias and variance in the MSE. Biased estimates of regression coefficients are OK as long as better predictions are obtained. No worries about fitting the correct probability distributions and getting correct standard errors in order to make inferences. Galit Shmueli wrote a helpful paper on predictive vs. explanatory modeling: https://www.stat.berkeley.e....

⌃ | ⌄ • Reply • Share ›

Vikas Prasad • 9 months ago

This page needs to appear higher in Google search.

⌃ | ⌄ • Reply • Share ›

MITDGreenb • 9 months ago

I'd pick up on your comment "It helped that many important

ML breakthroughs came from statistics, which had less of a presence in the rest of the AI field." This, to me, is the essence of the difference between AI and ML. Consider Natural Language Processing (NLP). When I first studied it, it was all symbolic analysis and programmed (in LISP!). That is, the system was quite good, but it "learned" only by having humans teach it the rules. Solutions for NLP are now quite varied, from Bayes to DL to distributed vectors (e.g., GloVe, and even more advanced things soon to come from one of my companies). What these new solutions have in common is two-fold: first, they are all based on numerical analysis - statistics - instead of symbols. And second, they all learn from examples (perhaps with human tweaking) rather than through human programming. That is, the machine learns... it is trained, not taught.

^ | ⌄ • Reply • Share ›

Carl Kruse • 10 months ago

What twoElectric said.

^ | ⌄ • Reply • Share ›

Thomas Haslam • 10 months ago

Dave, great closing example and a systematic, thoughtful discussion.

^ | ⌄ • Reply • Share ›

Praful Hambarde • a year ago

Awesome Artical

^ | ⌄ • Reply • Share ›

Zeca Rocha • a year ago

Is this structure ok in your point of view?
http://www.xmind.net/m/ANgy

^ | ⌄ • Reply • Share ›

Marwa Romdhan • a year ago

Big Thanks ! it's a great post

^ | ⌄ • Reply • Share ›

Francisco Aparicio • a year ago

Thank you, it will help me a lot. I really did not know how to answer the question because I had bad examples.

^ | ⌄ • Reply • Share ›

anonymousunblocked • a year ago

I couldn't see how to send you a message otherwise, but this article is quite relevant.

I've made a new tool (https://tree4pc.com) that lets people input CSV and interactively train the app (behind the scenes using neural networks) to find lines they want in it -- only by saying 'yes' or 'no' to keeping an example line. I think the simplicity, power, and wide applicability of this app might (should? :)) be something you'd appreciate and enjoy. I'd be interested to hear what you think -- and any of your readers too.

Currently you can buy the app for $2 through this link: https://tree4pc.com/buy.htm...

Reply • Share ›

Roberto Palloni • a year ago

Excellent.

Reply • Share ›

Tim Martin • a year ago

Great post, I like these distinctions. Tangentially, I would suggest you link to the original xkcd or cite it.

Reply • Share ›

buggyfunbunny • a year ago

If your old enough, or have read the history of IT, then you'll see that Data Science in 2018 is following the same path as Computer Science in 1968. To wit: Comp Sci was invented to satisfy the demands of not quite smart enough for EE guys who wanted to "do computers". Data Science was invented to satisfy the demands of not quite smart enough for Math Stat guys who want to "do quant". And get rich on Wall Street, of course.

paint me cynical, if you wish. But I've been there and seen it up close and personal.

Reply • Share ›

hrzafer • a year ago

IMO any system that mimics human intelligence is artificial intelligence. Be it a super simple rule based spellchecker or a very complex system for autonomous driving. It this sense ML, NLP, RL etc. all go under AI.

Reply • Share ›

Somcho • a year ago

Great article! I agree with all except one minor thing: How does "Google Maps choosing a route" qualify as something that "produces actions"
(Artificial intelligence)? ... What is the action in that scenario? I would argue that Gmaps provides a suggested route that it has calculated to me most optimal, then it is up to the user to make decisions about how closely to follow that suggestion given other domain knowledge they have at that moment (e.g. visible unreported constructions or obstructions) or from their personal experience. However I wouldn't go so far as to claim that the user, by doing so, is doing data science anymore than somebody reading a thermometer is doing Thermodynamic Physics.. So in my opinion Google maps is providing *insite* about the quickest route(s) and given a specific route, Google

maps is providing a *prediction* as to when one will arrive. I do not see any AI aspect (such as reinforcement learning, control theory and others you described) being involved. Perhaps rather than calling General Optimization, "Artificial Intelligence" let's just call it Optimization! ☺ (Or maybe, to be pedantic, 'Operations Research'?)

⌃ | ⌄ • Reply • Share ›

Saurabh Sharma • a year ago

We need a term which encompasses all the 3 field. Something more specific than computer science. Is there such a term?

⌃ | ⌄ • Reply • Share ›

Tiago D'Agostini ➦ Saurabh Sharma • 6 months ago

Why we NEED this term? Some things are better when kept separated enough to keep themselves meaningful. Generalization is not always productive for better communication. Identity comes with time only...

⌃ | ⌄ • Reply • Share ›

Riinu ➦ Saurabh Sharma • a year ago

Informatics?

⌃ | ⌄ • Reply • Share ›

Andrew Sila • a year ago

As promised you have given an oversimplified descriptive for the three terms which in many times we see being used interchangeably. Well written!

⌃ | ⌄ • Reply • Share ›

Ed • a year ago

They're all buzzwords with vague meanings, you could have easily said

Data science produces actions
Machine learning produces insights
Artificial intelligence produces predictions

and it would still make sense

⌃ | ⌄ • Reply • Share ›

YOU MIGHT ALSO ENJOY

(VIEW ALL POSTS)