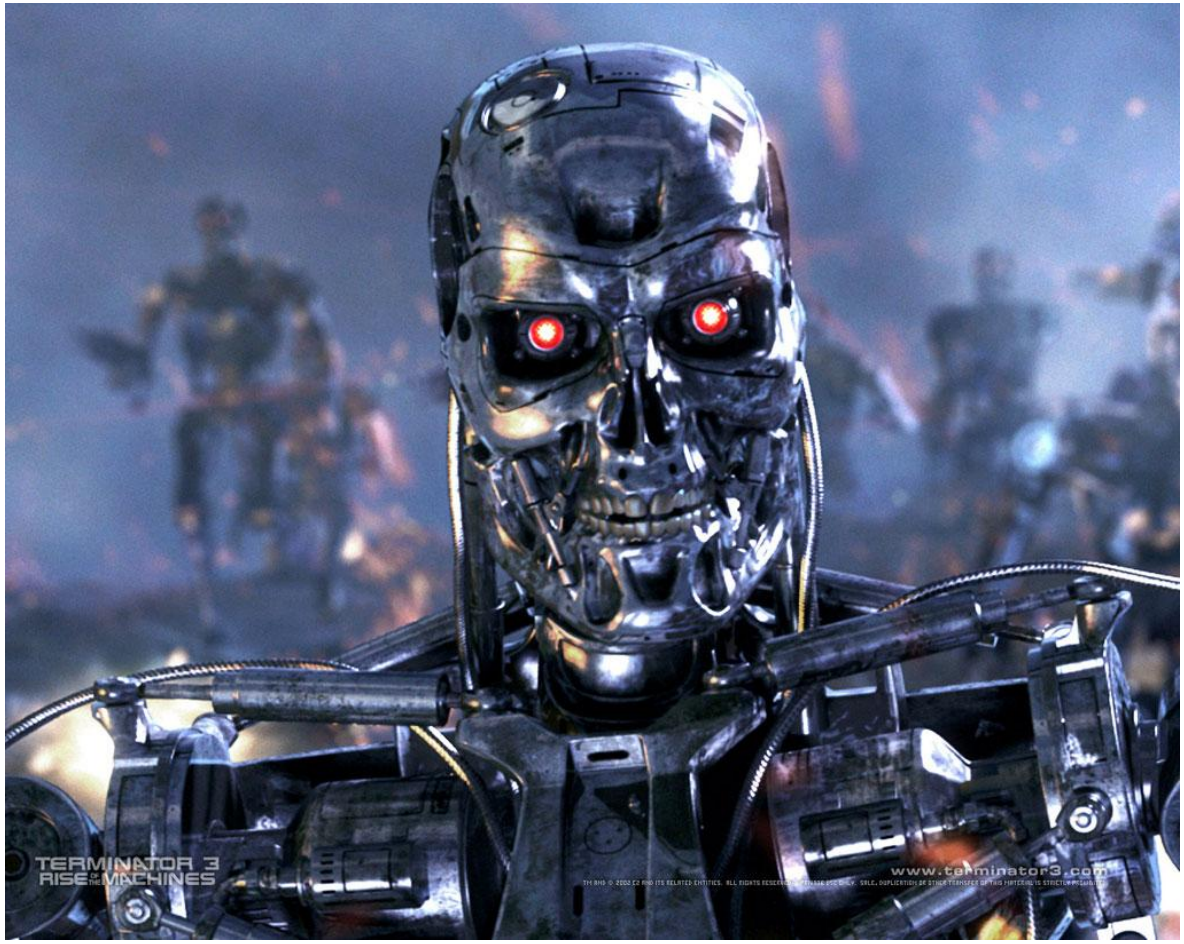
Big Data

And Other Fancy Buzzwords for Pretty
Simple Ideas

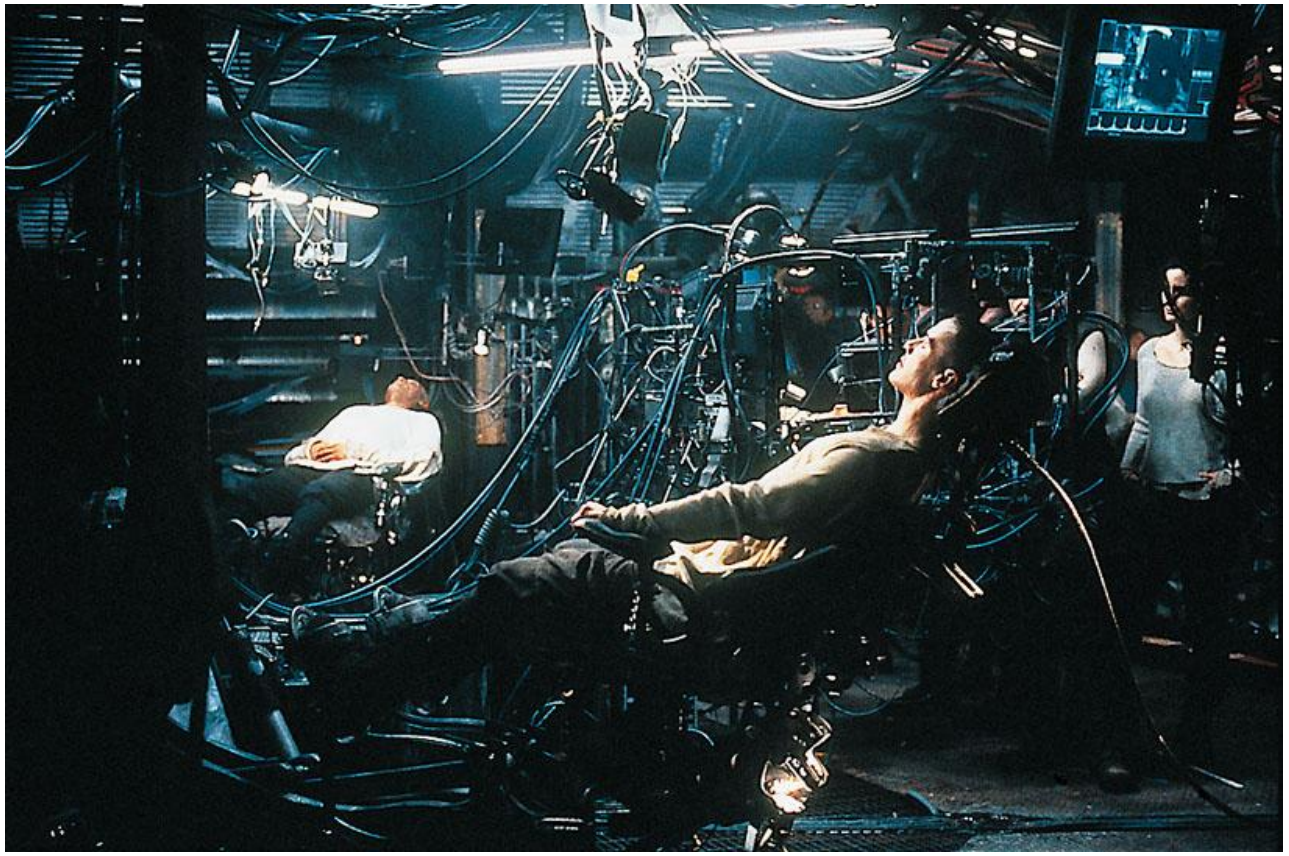
Acronyms and Buzzwords Galore

Big Data
Data Analytics
Machine Learning
Artificial Intelligence
Data-Mining
Hadoop
Business Intelligencezzzzzzzzz.....

Spoiler: We're not dealing with....



Or this...



And certainly not this...

Getting to the Point

Take the data we've already have :

and transform it into insights we can leverage.

It's a hot topic right now

Conference chatter, blogs, other companies.

Not just real estate, everywhere.

Beware, lots of snakeoil.

Some Definitions

Data Mining
(roughly interchangeable with data analytics)

Applying statistical analysis

to a collection of data

searching for trends, anomalies, and associations

That weren't previously known.

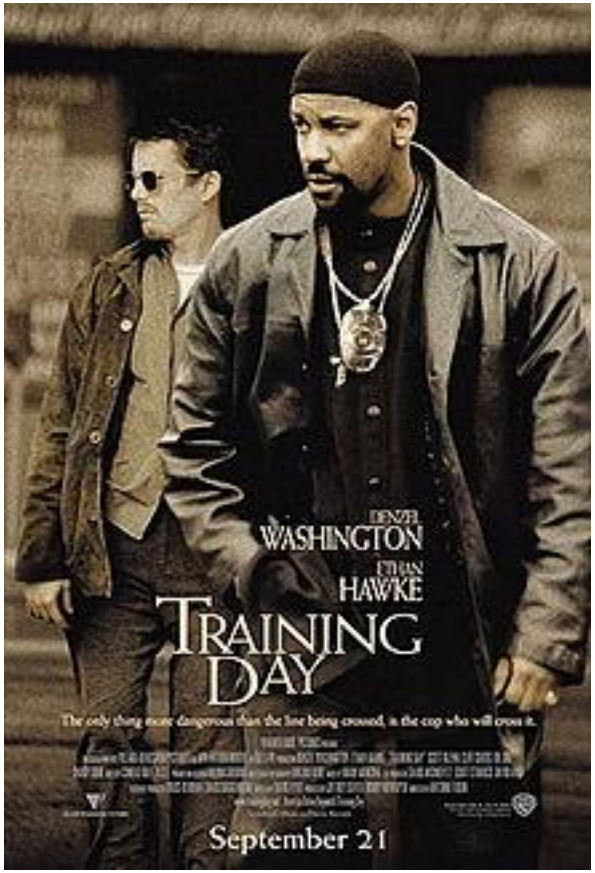
Machine Learning

Applying automated analysis

to a collection of data with a known pattern or association

to build a **model** to **predict** or **classify** future data.

The two are closely related



Unsupervised machine learning performs a data mining task first then trains the computer to categorize new data based on the discovered rules without (much) human interaction.

Big Data

Really overloaded term, depending on who is using it can mean:

- Any of the previous terms
- Infrastructure required to handle large data sets.
- the data itself
- The trend in the software industry of leveraging large data sets

Small Examples
(contrived)

Listing Price	Listing City	Year Built	Bedrooms
350,700	Mt. Pleasant	1996	4
176,000	Summerville	2006	3
128,000	Summerville	2003	2
450,000	Kiawah	2004	4
380,000	Mt. Pleasant	2001	4
186,000	North Charleston	2002	3
150,000	North Charleston	1997	3
155,000	Hanahan	1988	4
165,000	Hanahan	1992	3
575,000	Kiawah	2002	5
427,000	Mt. Pleasant	2006	3

Small Examples
(contrived)

Similar
Price range!

Similar
Price range!

Listing Price	Listing City	Year Built	Bedrooms
350,700	Mt. Pleasant	1996	4
176,000	Summerville	2006	3
128,000	Summerville	2003	2
450,000	Kiawah	2004	4
380,000	Mt. Pleasant	2001	4
186,000	North Charleston	2002	3
150,000	North Charleston	1997	3
155,000	Hanahan	1988	4
165,000	Hanahan	1992	3
575,000	Kiawah	2002	5
427,000	Mt. Pleasant	2006	3

Small Examples
(contrived)

More Avg. Bedrooms

Less avg. bedrooms.

Listing Price	Listing City	Year Built	Bedrooms
350,700	Mt. Pleasant	1996	4
176,000	Summerville	2006	3
128,000	Summerville	2003	2
450,000	Kiawah	2004	4
380,000	Mt. Pleasant	2001	4
186,000	North Charleston	2002	3
150,000	North Charleston	1997	3
155,000	Hanahan	1988	4
165,000	Hanahan	1992	3
575,000	Kiawah	2002	5
427,000	Mt. Pleasant	2006	3

Small Examples
(contrived)

Newer on average.

Listing Price	Listing City	Year Built	Bedrooms
350,700	Mt. Pleasant	1996	4
176,000	Summerville	2006	3
128,000	Summerville	2003	2
450,000	Kiawah	2004	4
380,000	Mt. Pleasant	2001	4
186,000	North Charleston	2002	3
150,000	North Charleston	1997	3
155,000	Hanahan	1988	4
165,000	Hanahan	1992	3
575,000	Kiawah	2002	5
427,000	Mt. Pleasant	2006	3

Older on average

And that tells us?

It's not too big of a logical leap to say that someone interested in Mt. Pleasant would also be interested in Kiawah given the choices in our table.

Someone looking for newer houses (less maintenance?) would be better off in Summerville than Hannahan

That's nice in theory

Each of our boards has thousands of listings.

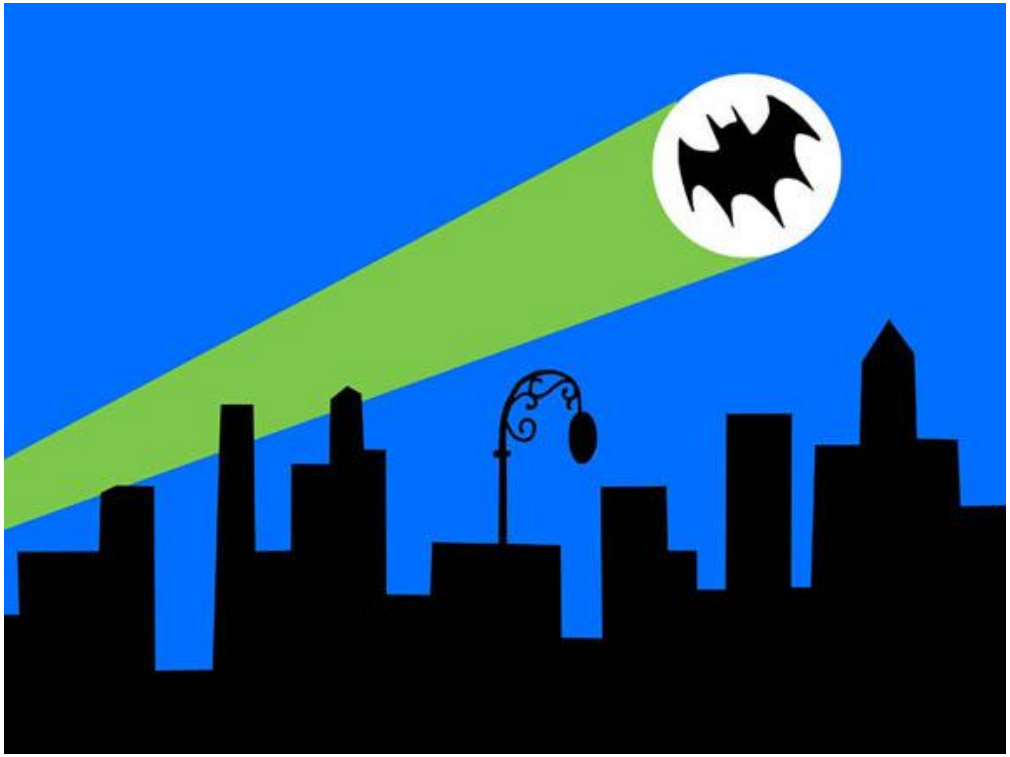
Each tenant has thousands of leads.

Clearly eyeballing it isn't going to cut it.

The Answer

We need to find the groups of listings/leads/etc automatically, we can't manually sort through all this data.

Data-Mining to the Rescue!



So...

All Hail Our Robot Overlords?

No?

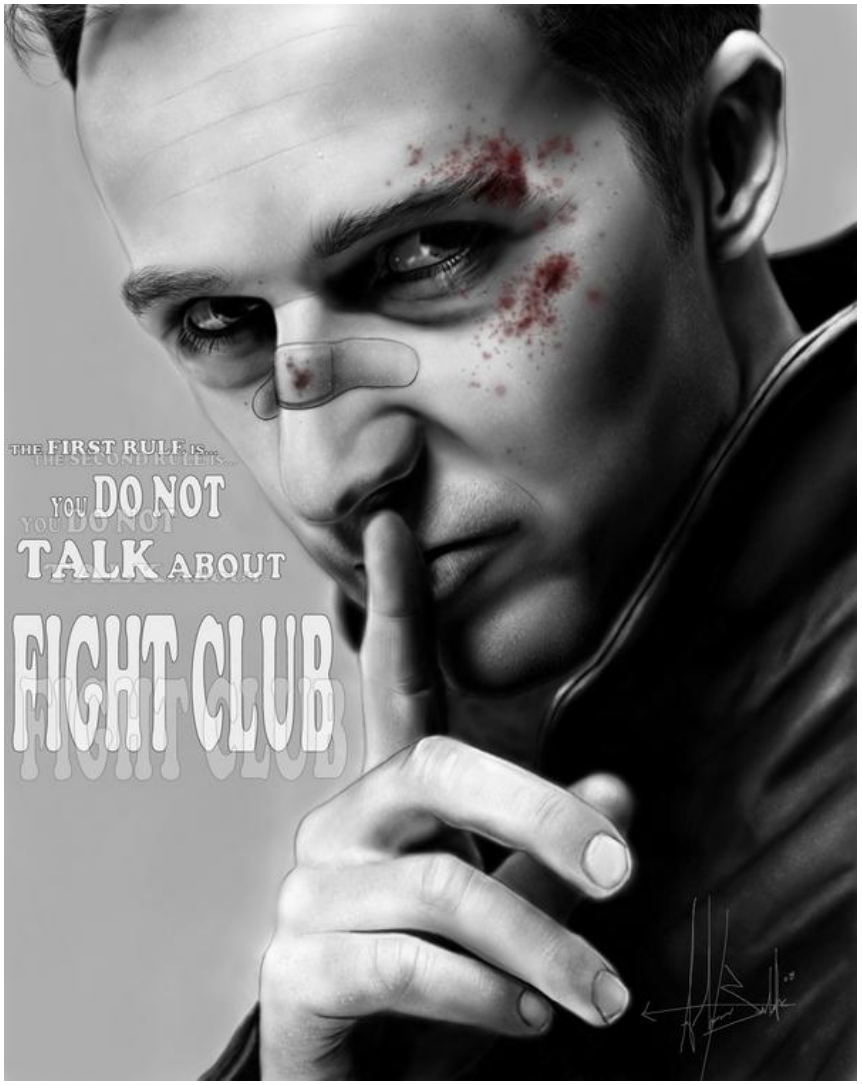
Then what does it mean?
(for BoomTown)

Well, We're already doing it!

Development team has several prototypes using these types of techniques.

I won't spoil the demos

Shhhh. Still very much in alpha, don't tell clients, or outsiders



So what else could we do?

Lets Go By Department - Support

Ideas:

Automatically reading incoming tickets and grouping the similar ones.

Automatically detecting potential system issues by ticket keyword patterns

Lets Go By Department - Production

Ideas:

Analyzing log files to catch weird failure conditions.

Actively predicting when new servers are needed.

Lets Go By Department - SEO/SEM

Ideas:

Correlating keywords with highest ROI buyers.

Relating keywords with buyers relative purchase horizon.

These are just a few examples

It's all on you

Everyone here is a domain expert.

Ask questions like:

"I wonder how many leads are ..."

Don't guess, calculate

(or ask the devs to calculate)

Questions?

Attributions

All images lifted (stolen) from google image search.

Wikipedia provided basis for some of the definition language.

The log analysis idea came out of a session at Velocityconf