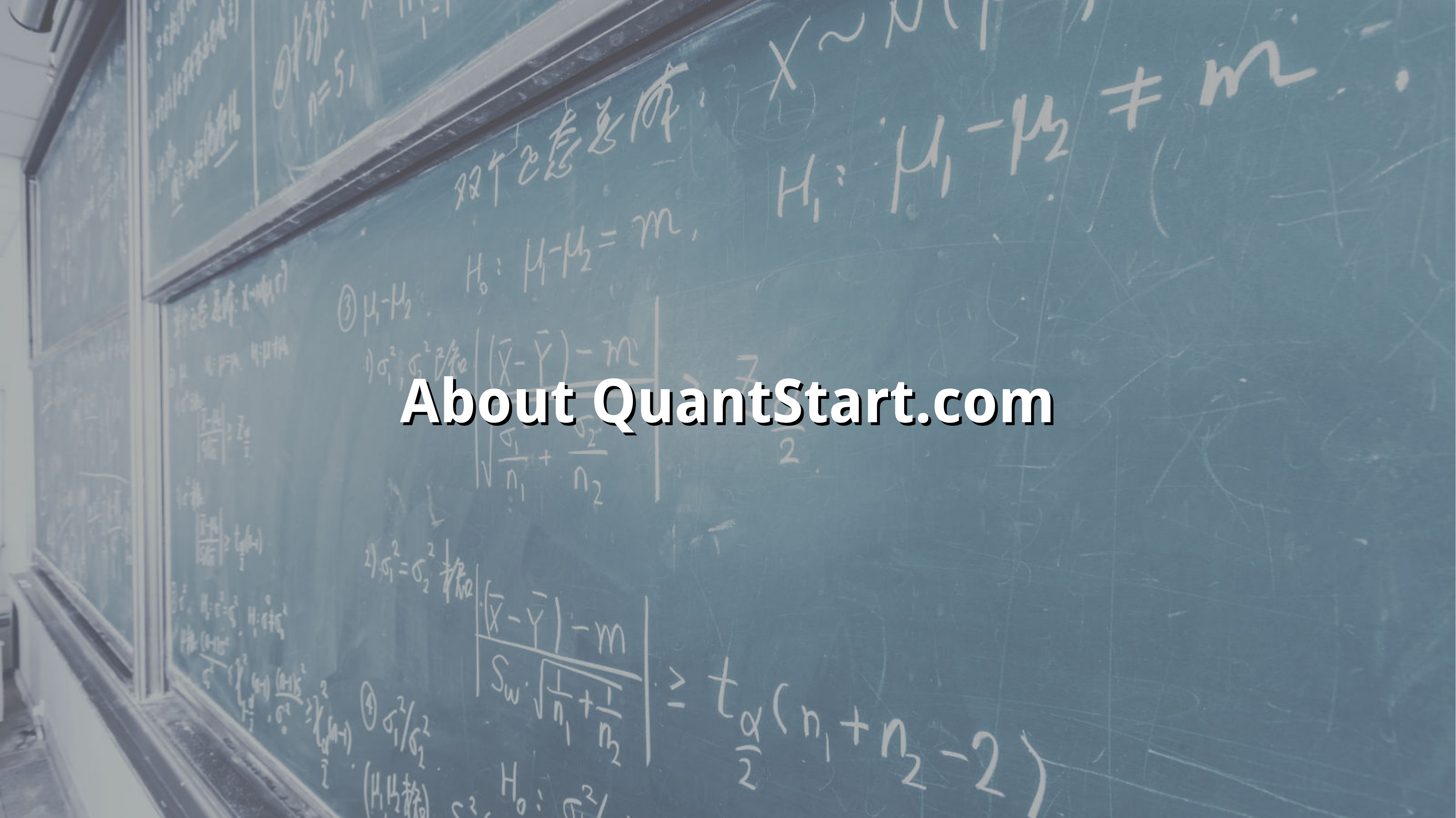
The Hunt For Alpha Among Alternative Data Sources

QuantCon Singapore – 11th November 2016

Michael Halls-Moore
QuantStart.com

Talk Outline

- About **QuantStart**
- Our **goal** as quant traders
- The problem of **Alpha Decay**
- Alpha from **new data sources**
- *Which* new data sources?
- **Tools** to quantify new data sources
- **Alpha-generating strategies** based on new data
- Where to go **from here**?



About QuantStart.com

About QuantStart.com

- QuantStart was founded in 2012
- **Educational portal** for quantitative trading
- Talks about **algo trading**, **backtesting** and **machine learning**
- Mainly **Python** and **open-source backtesting**
- My background is originally in:
  - Computational Fluid Dynamics (CFD) research
  - Quantitative development at small London quant fund

Our Goal as Quant Traders

The Hunt for Alpha

- Our goal is to **search for “alpha”**
- Alpha is a **new stream of returns** uncorrelated with other “known” sources of returns
- Purely, it is a function applied to a time series that produces **predictions/weights** of assets for the next time-period/rebalance → Roughly the “strategy”
- The main idea is to look for **approaches that others don't know about** otherwise it's not “alpha”

The Problem – Alpha Decay

Alpha Decay

- **Very cheap** to get quality asset pricing and fundamentals data
- Easy to **“wrangle”** data into the correct format
- Can analyse **thousands of strategies** with cloud computing
- **Diffusion of information** and “democratisation” of technology ensures faster “alpha decay”
- Need to look for alpha **elsewhere**
  - **Alternative data sources!**

The Solution – Alternative Data

Alternative Data

- New alpha can be found in **alternative data**
- Quant funds, family offices and prop trading desks are **already using it successfully**
- **Standard practice** for retail quants within next five years
- Those who don't use it will be on the **wrong side of the informational edge**

What Data Sources are Available?

- **Satellite data** - Visual, IR
- **Aerial/drone data** – Visual, LiDAR, IR
- **Social media data** – Blogs, FB, Twitter, Instagram, Reddit…
- **Internet-of-Things data** – Smartphones, car logs, sensors
- **Energy supply/demand data** – Oil, natural gas, consumer demand
- **Weather data** – Wind, temperature, rainfall
- **Automated email receipts** - E-commerce purchases
- **Geolocation monitoring** - Shipping, airline and freight locations
- Many, many more…

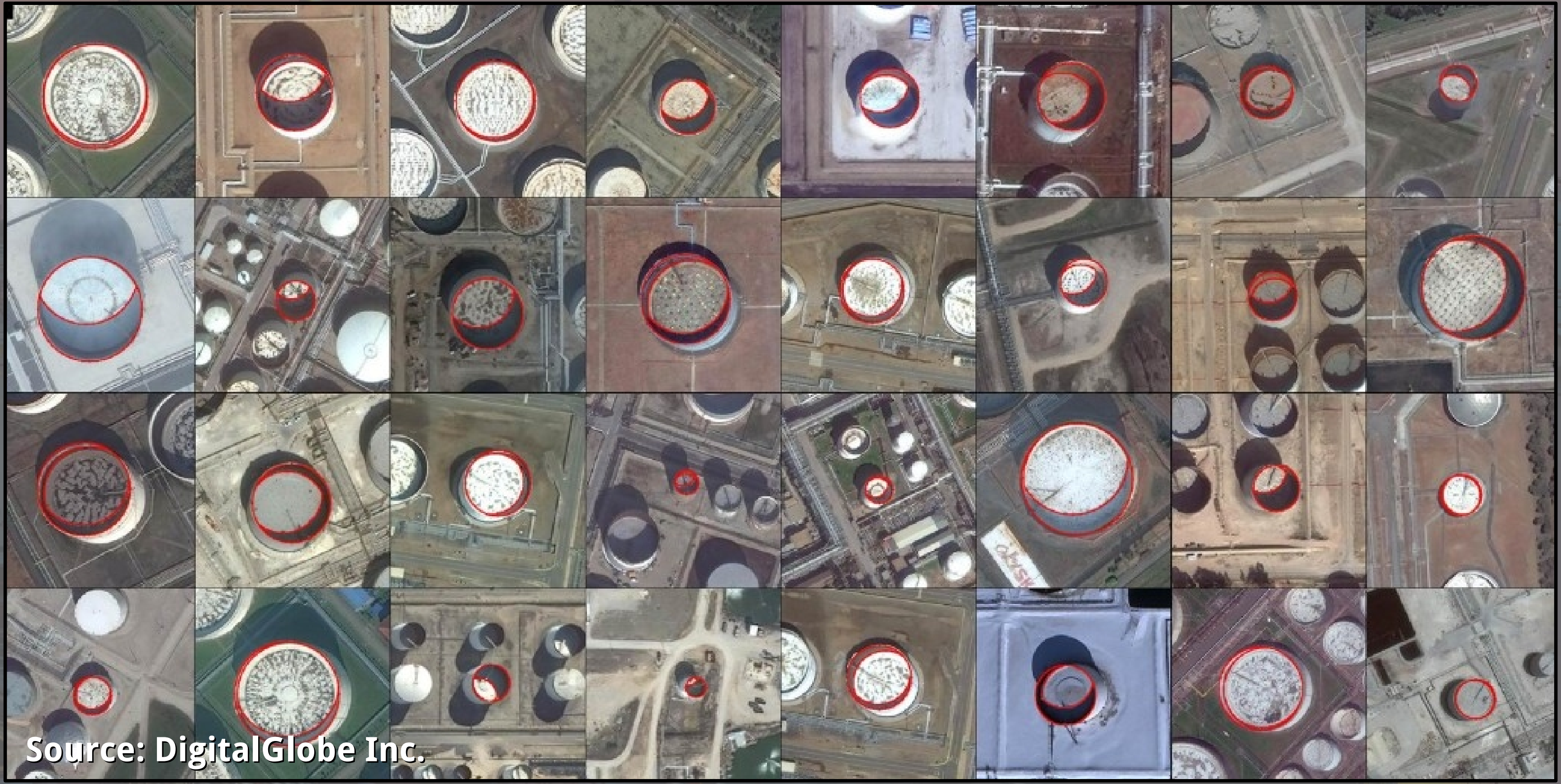
Alternative Data Examples

Remote Observation Data Abundance

- **Satellite imagery** and **aerial drones**
- Multiple EM wavelengths → **“Hyperspectral”**
- **Microsats** becoming cheaper to develop and launch
- **Drones** are cheap to build, fly and collect data with
- Vendors offering **frequent high-resolution observation data** from both at low(ish) cost

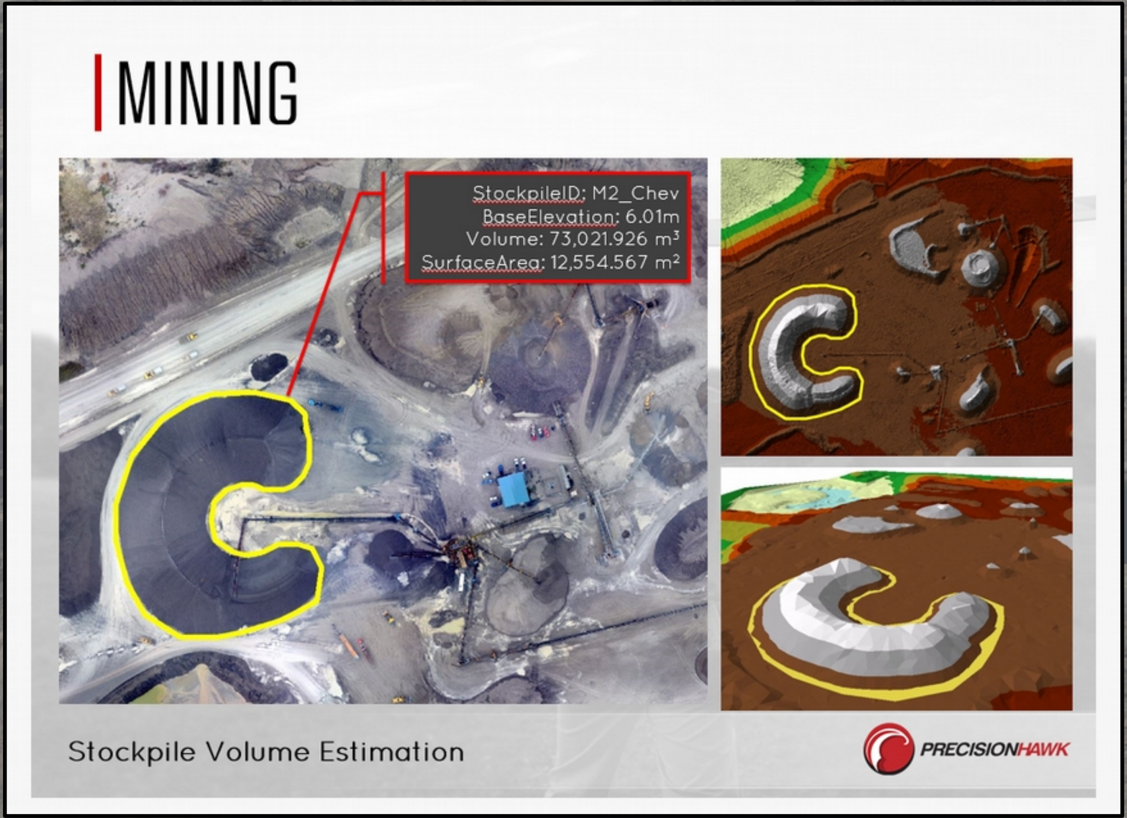
Remote Observation Data Uses

- Estimating **oil volume** by calculating oil storage floating-tank height with their **shadows**
- **Air** and **marine freight traffic** location determination
- **Counting cars** in retail car parking lots to **estimate sales**
- Hyperspectral **crop yield estimation** for “softs” trading
- Estimating **mining yields** via LiDAR volume calculation
- Previously this data had to be collected *in-person, by hand*



Source: DigitalGlobe Inc.

Oil Depot Floating Tank Shadow Height

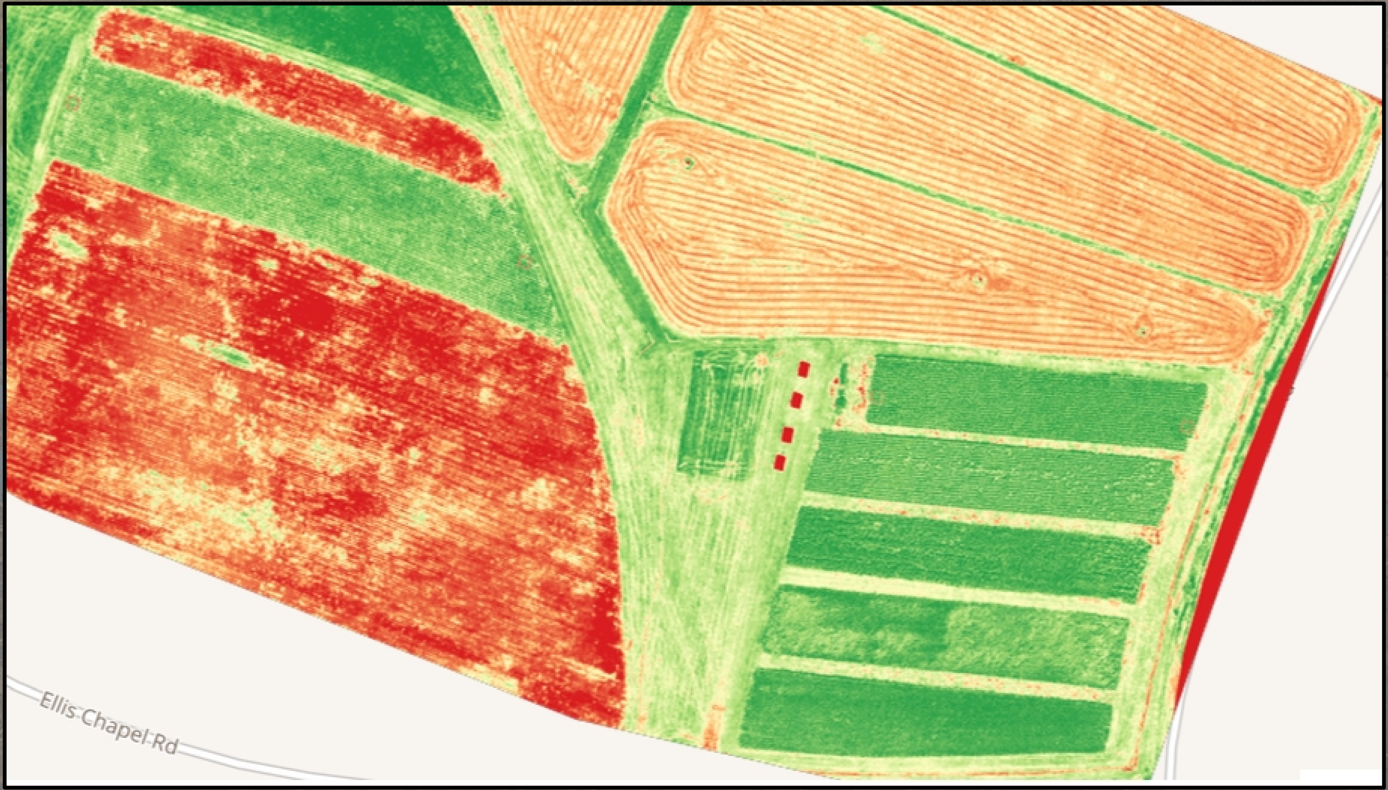


MINING

StockpileID: M2_Chev
BaseElevation: 6.01m
Volume: 73,021.926 m^3
SurfaceArea: 12,554.567 m^2

Stockpile Volume Estimation

Mining Yields from Raw Material Stockpiles



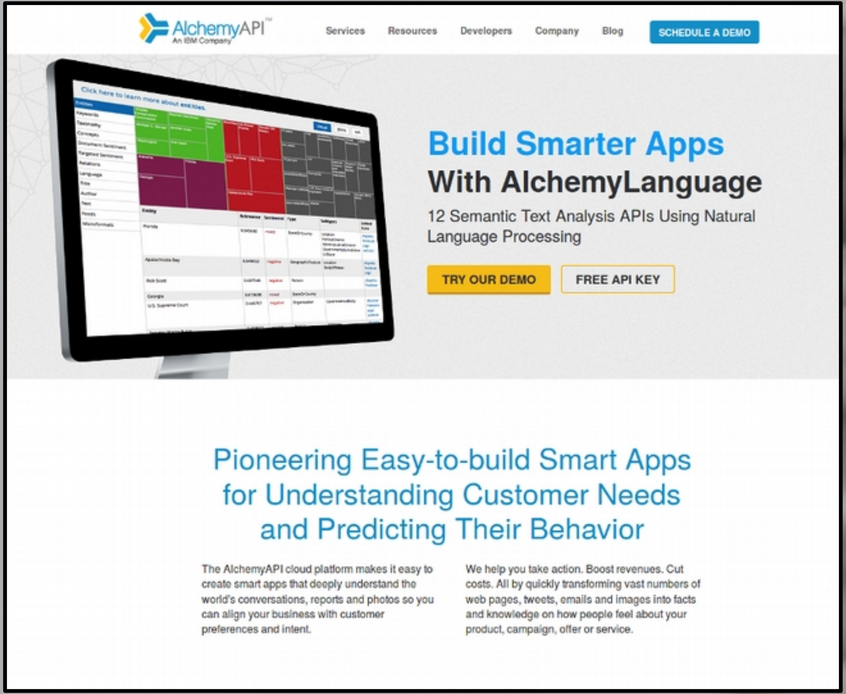
Crop Yields via “AgTech” Drone Usage

Sentiment Analysis

- **Numerous vendors** – Gnip, DataSift, Quandl, AlchemyAPI
- Provides access to **thousands of news sources** as well as Disqus, FB, Instagram, Reddit, Twitter, YouTube and more
- Datasets are large → YouTube added 1PB *per day* in 2015
- Often used for **equities returns prediction** through news, tweets and earnings reports
- Challenging to make effective strategy!

Sentiment Analysis Challenges

- **Rapidity:** Requires fast trade execution after receipt of news
- **Relevance:** Which equities does news affect and how much?
  - e.g. new Tesla car release impacts Ford, GM, Google
- **Categorisation:** Each category has variable market response
  - e.g. surprise earnings vs legal battle
- **Novelty:** Market only moves if news not “priced in”
  - Must account for *relative* value of news



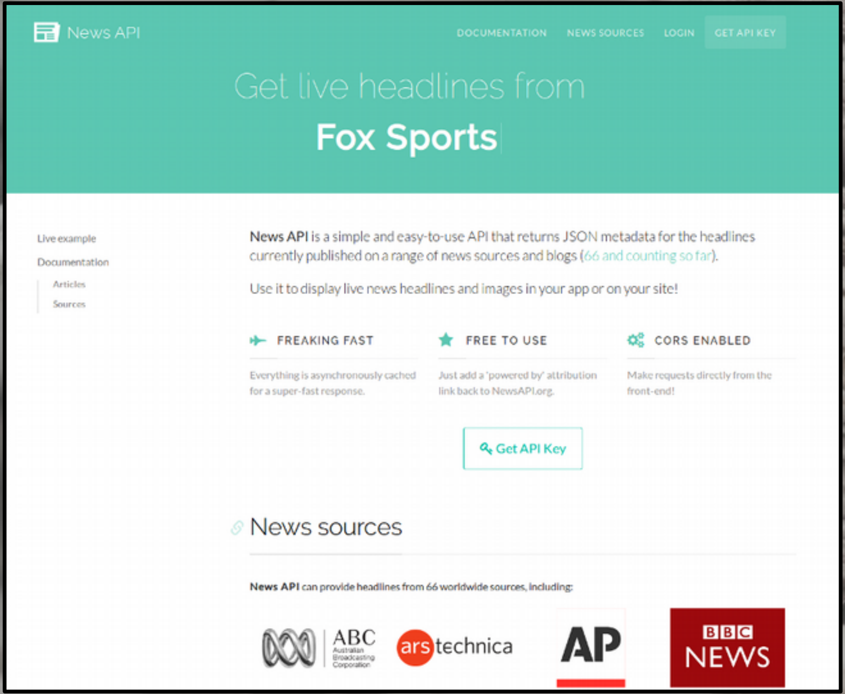
Build Smarter Apps With AlchemyLanguage

12 Semantic Text Analysis APIs Using Natural Language Processing

Pioneering Easy-to-build Smart Apps for Understanding Customer Needs and Predicting Their Behavior

The AlchemyAPI cloud platform makes it easy to create smart apps that deeply understand the world's conversations, reports and photos so you can align your business with customer preferences and intent.

We help you take action. Boost revenues. Cut costs. All by quickly transforming vast numbers of web pages, tweets, emails and images into facts and knowledge on how people feel about your product, campaign, offer or service.



Get live headlines from

Fox Sports

News API is a simple and easy-to-use API that returns JSON metadata for the headlines currently published on a range of news sources and blogs (66 and counting so far).

Use it to display live news headlines and images in your app or on your site!

FREAKING FAST

Everything is asynchronously cached for a super-fast response.

FREE TO USE

Just add a 'powered by' attribution link back to NewsAPI.org.

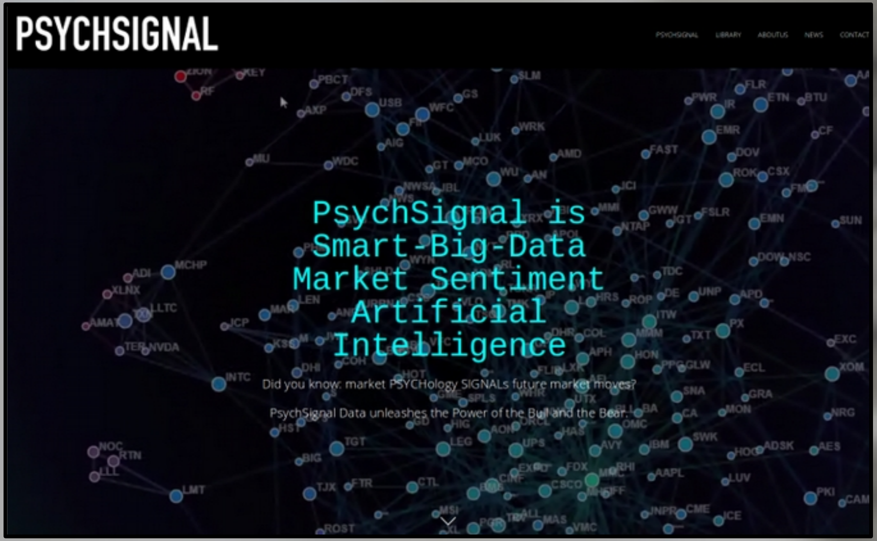
CORS ENABLED

Make requests directly from the front-end!

News sources

News API can provide headlines from 66 worldwide sources, including:

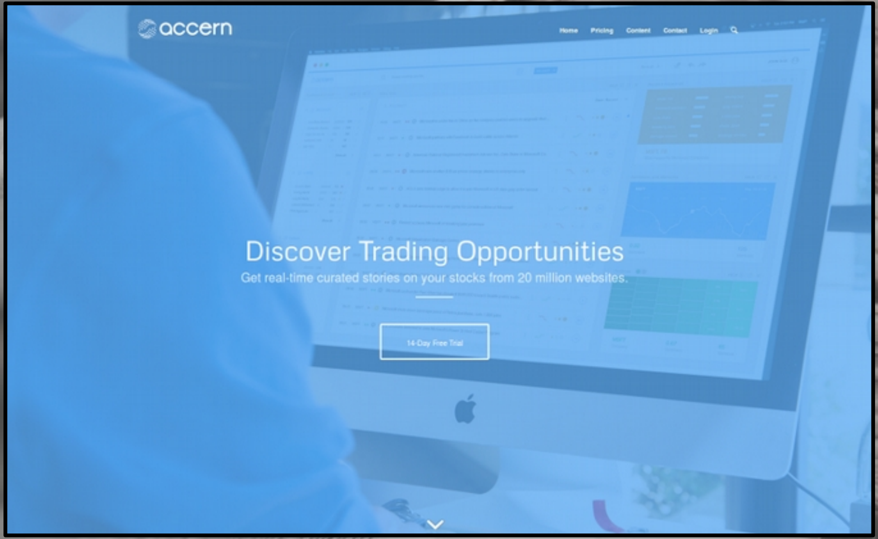
News API Vendors



PsychSignal is Smart-Big-Data Market Sentiment Artificial Intelligence

Did you know: market PSYCHology SIGNALs future market moves?

PsychSignal Data unleashes the Power of the Bull and the Bear.



Discover Trading Opportunities

Get real-time curated stories on your stocks from 20 million websites.

Sentiment Analysis API Vendors

Internet-of-Things (IoT) Data

- Smartphones, GPS, sensors → All **internet-connected**
- **Huge impact** in O&G/energy, AgTech, healthcare and insurance
- Vendors beginning to **anonymise** and **sell** data
- Hedge funds are first to **exploit alpha** in these datasets
  - e.g. Consumer footfall via GPS/smartphones for **retail sales estimation** ahead of analyst expectations

Energy and Weather Data

- **Physical weather** data and **energy supply/demand**
- Funds/banks use this to trade **commodity futures**, **cat bonds** and **weather derivatives**
  - One example is London-based **Cumulus** fund
  - Reported to be able to predict weather **better than Met Office**
- Many companies rely on **favourable weather for revenue**
  - Retail, adventure sports, agriculture, energy
  - Motivates earnings-based trading ideas

E-Commerce Purchase Receipts Data

- Some startups have **indirect visibility into email** inboxes
  - Gmail, productivity apps, to-do apps
- Vendors now provide millions of anonymised emails as data
- Trading strategy **estimates quarterly revenues from email purchase receipts** and trades when expectations differ
- Quandl.com talks about this at length in blog posts

Pros and Cons

Advantages of Alternative Data

- Good **signal-to-noise ratio** compared to pricing data
- Often **uncorrelated** to other financial data sources
- Many off-the-shelf techniques available to **quantify the data**
- **Competitive advantage** once 'data pipeline' is built and tested
- New data sources **appear frequently**
- Retail traders **can compete** with funds in niches
  - Open source data science tools freely available
  - Compute power in the cloud is cheap

Disadvantages of Alternative Data

- Often **non-quantitative** – Video, imagery, text
- Extremely **high-dimensional** – Video, imagery, text
- **Unstructured/hierarchical** – no key-value schema
- **Missing values** – Interpolation or imputation required
- Data vendors all have **differing formats**
- Data vendor **quality** is highly variable
- Some datasets can be **prohibitively expensive** for retail

Alternative Data for Quant Trading

- **Prediction:** Volume, volatility, returns?
- **Liquidity:** Can you actually trade on it?
- **Timeframe:** HF microstructure or longer-term macro trends?
- **Exclusivity:** Too many users causes **alpha decay**
- **Domain Expertise:** Can data be used “out of the box”?
- **Consistency:** Does the data format **change** over time?

Overcoming Alternative Data Challenges

- Alternative data can be **terabytes** or **petabytes** in size
- Often requires **quantification** through **vectorisation**
- Software and algorithms need to be **highly parallelisable**
- “Big Data” era requires new **data science** tools
  - **Storage/Processing:** AWS S3, Hadoop, HDF5, MapD
  - **Analysis:** Machine Learning

Machine Learning

Machine Learning

- A mechanism for **extracting useful signals** from alternative data
- Learns model **from the data**
  - Not pre-programmed “if-then-else” rules
- Main goals are **prediction** and **classification**
- Machine learning is **pervasive in quant finance**
- Three main areas:
  - **Supervised Learning:** Asset Price Prediction, Trade Parameter Optimisation
  - **Unsupervised Learning:** Factor Analysis, Portfolio Clustering
  - **Reinforcement Learning:** Optimising execution algos

Supervised Learning

- Attempt to **match inputs** with **known outputs**
  - Predicting tomorrow's stock price from the previous ten days of prices
  - Classifying a text document into a set of known categories
- **Advantage:**
  - **State-of-the-art** for classification tasks in alternative data
- **Disadvantages:**
  - Data must be **labelled**, which is costly
  - Prone to **overfitting** – performance might not generalise
  - Requires substantial **training data** to perform well

Unsupervised Learning

- Find **useful structure** in the data – no “outputs”
  - Which equity returns tend to **cluster together**?
  - Which **factors** drive equity returns?
- **Advantages:**
  - Most data in the world is unlabelled so UL is widely applicable
  - Used to reduce dimensionality of high-dimensional alternative data
- **Disadvantage:**
  - Lack of **consistent evaluation mechanism** makes it hard to know if algorithm is effective

Reinforcement Learning

- **Agent** interacting with **environment** via **actions** and **rewards**
- **More challenging** than supervised and unsupervised learning
- Has recently become very famous due to **DeepMind** success on **Atari 2600 games** and **AlphaGo** competition
- Recent promise has prompted many to apply it to quant trading
  - **Stochastic environment** and **noisy reward signal** make it tricky
  - Is used in execution algo optimisation (discussed here at QuantCon!)

Deep Learning

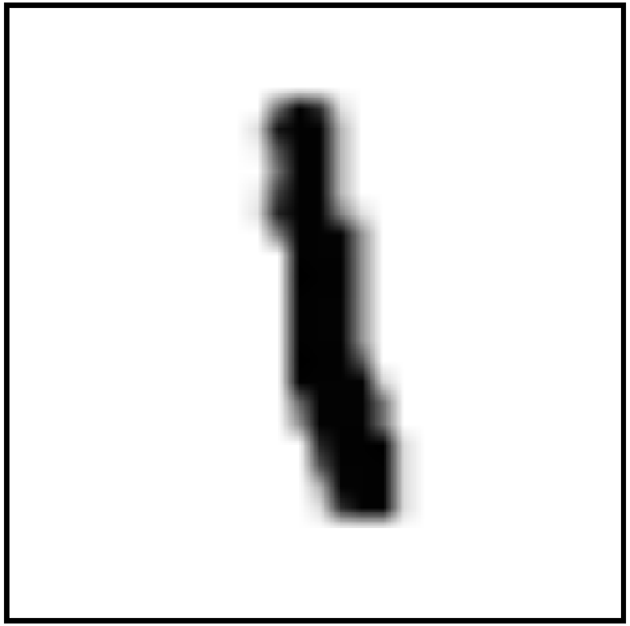
- Deep learning is a **state-of-the-art** machine learning technique
- It involves 'deep' **neural networks** with many 'hidden' layers
- Allows **feature extraction** that other ML methods can't achieve
- Primary method for **extracting signal** from alternative data
- **Advantages:**
  - Usually the 'best' method to extract signal for image, text or audio datasets
- **Disadvantages:**
  - Steep learning curve, requires a good background in ML
  - Significant trial-and-error needed to achieve best results

Analysing Alternative Data

Quantification of Alternative Data

- **Quantification Steps:**
  - **Vectorise** the data into numerical form
  - **Reduce the dimensionality** of the data
  - **Scale** the data to make it comparable across different datasets
- **Image/Video:**
  - Convert each pixel into grayscale [0, 1] intensity value vector
- **Text:**
  - Each word is a dimension representing weighted frequency in a document (TF-IDF)

Image Vectorisation



$$\simeq \begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

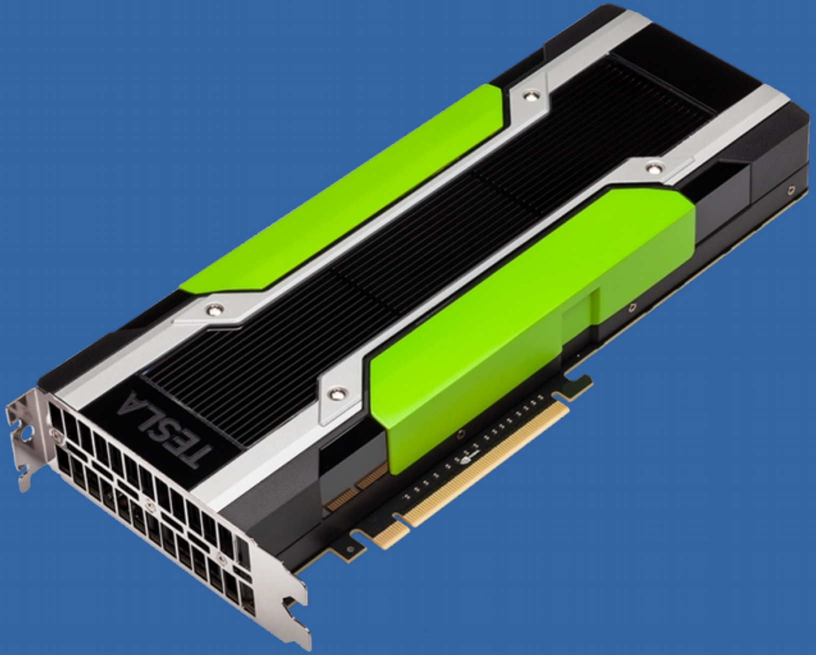
- 14x14 greyscale image converted into 196-dimensional vector

Data Science Tools for Exploratory Analysis

- Freely-available **open-source tools** are **best for the job**
  - Top-tier quant funds, big Silicon Valley firms, data scientists and retail traders
- **Python**
  - **Anaconda** → Research environment
  - **NumPy/Pandas** → Data wrangling
  - **Scikit-Learn** → Unified SL and UL API
  - **TensorFlow** → Deep Learning
- Goal: Check data for **alpha**!

Compute Power via The Cloud

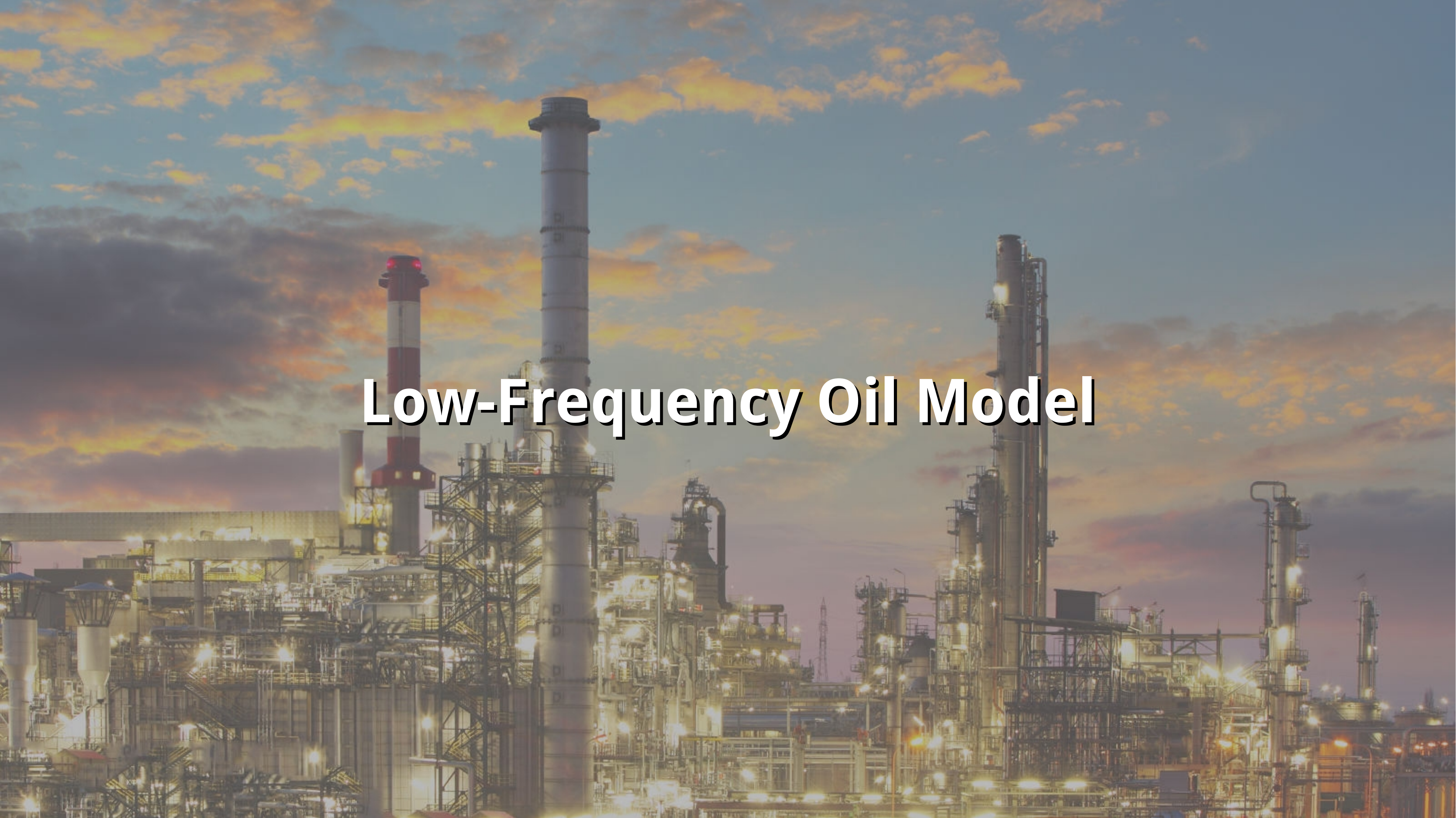
- Previously it was **expensive** to get access to highly-parallelised supercomputing
- Required complex HPC machines with **many CPU cores**
- GPUs and cloud vendors have **changed the economics** significantly
- GPU compute power in **the cloud**
  - Amazon EC2 p2.xlarge instance - $0.90/hr
  - Amazon EC2 p2.16xlarge instance - $14.40/hr



Quant Trading on Alternative Data

Quant Trading on Alternative Data

- Must have **underlying economic rationale** for strategy
- **Model the factors** that move asset prices:
  - **Supply/Demand** → Physical, statistical, network/graph models
  - **Market Sentiment** → Text, news, social sentiment analysis models
- Generate **better estimates** than “the market”
- Ensure model produces **alpha-generating predictions**
  - Accounting for liquidity constraints and transaction costs



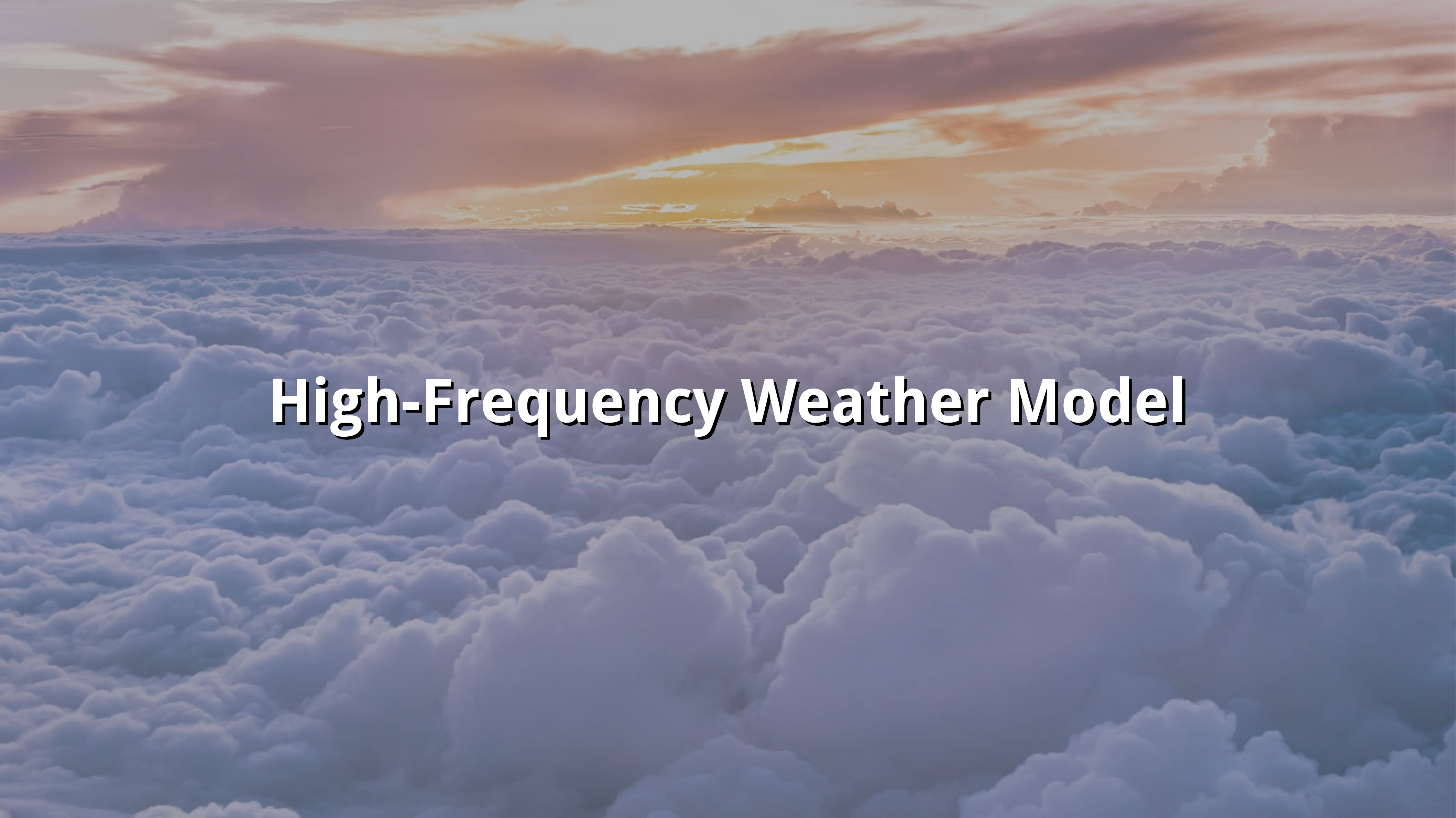
Low-Frequency Oil Model

Oil Model Sketch

- Attempt to model **major drivers** of the oil price via alternative data sources
  - Specifically **supply/demand imbalance** and **market sentiment**
  - **Alpha should decay slowly** as model will be tricky to replicate
- Trading strategy **is likely to work**:
  - Current oil inventory data is based on **estimates**
  - Estimates have **varying levels of quality** and **truthfulness** across regions
  - We can generate **better estimates** via alternative data
- Trade weekly when **our predictions differ** from market expectations
  - Oil futures → CL
  - Oil ETFs → USO, XOP, UCO

Oil Price Drivers Estimation

- **Estimating Supply:**
  - Satellite: Global oil depot tank **classification** and **volume**
  - Satellite: **US domestic fracking output** → Indirectly via transportation data (e.g. counting tanker-wagons on freight trains via satellite)
  - Geolocation: **MarineTraffic.com** for oil tanker locations/destinations
- **Estimating Demand:**
  - Economics: Population models, cars per household, freight truck usage, avg miles driven, efficiency of cars, local gasoline taxation
- **Estimating Sentiment:**
  - **OPEC/trading sentiment** via Twitter, media and research reports



High-Frequency Weather Model

Weather Model Sketch

- Attempt to model **major drivers** of **weather derivatives** via alternative data
  - Alpha is generated through **better predictions** at **intraday frequencies**
  - Must be able to **predict local weather** to an **extremely high accuracy**
  - Strategy likely to require a **small data-science/quant/developer team**
- For **accurate temperature/rainfall prediction** at major cities we can combine:
  - **Numerical Weather Prediction** (NWP) model and **statistical ensemble** of forecasts
  - **Entity extraction/sentiment analysis** from social/text sources in geo-referenced posts
- Can create **portfolio of weather derivatives** to bet on predictions
  - CMEGroup provides futures/options for larger US cities as well as London and Amsterdam

Weather Derivatives Model Details

- Backtesting will be **challenging**:
  - Potential **illiquidity** of weather derivatives
  - **Market impact** is tough to simulate
  - Combining NWP with statistical ensemble intraday will require **sophisticated HPC infrastructure**
- Advantages:
  - **Capacity constraint** of assets limit it to smaller funds or small team
  - **Alpha will likely decay slowly** as it requires expertise in many areas

Where To Go From Here?

Where To Go From Here?

- **Beginner Data Science Tutorials:**
  - Scikit-Learn: http://scikit-learn.org/stable/tutorial
  - TensorFlow: https://www.tensorflow.org/tutorials
  - Kaggle/Quantopian: Practice, practice, practice!
- **Data Vendors:**
  - Quandl, Gnip, DataSift, AlchemyAPI, PyschSignal
  - Forecast.io, NOAA, FlightRadar24, MarineTraffic
- **Compute Power:**
  - Buy Nvidia Titan X GPU → $1200
  - Rent p2.xlarge Amazon EC2 instance → ~$670/month

Thank you!

Q&A?