

ORACLE MATRIX

La Recette de l'Algorithme

Comment fonctionne notre moteur d'analyse des marchés de prédiction,
la théorie économique qui le sous-tend, et pourquoi ça marche.

Document confidentiel — Oracle Signals

Février 2026

SOMMAIRE

1. La théorie : pourquoi les prediction markets sont battables
2. Vue d'ensemble : le pipeline en 8 étapes
3. Étape 1 — Scan des marchés (Polymarket CLOB API)
4. Étape 2 — Enrichissement profond (Order Book + Spread)
5. Étape 3 — Classification par domaine (mots-clés pondérés)
6. Étape 4 — Pré-filtrage (Top 10 par domaine)
7. Étape 5 — Données externes (cross-referencing)
8. Étape 6 — Analyse IA (Claude / GPT-4o)
9. Étape 7 — Score composite (0-100)
10. Étape 8 — Routage des signaux (VIP / Free / Academy)
11. Les leviers d'amélioration

1. LA THÉORIE

Pourquoi les prediction markets sont battables

Les marchés de prédiction (Polymarket, Kalshi, Metaculus) fonctionnent sur un principe simple : le prix d'un contrat YES reflète la probabilité implicite que l'événement se produise. Si "Trump gagne 2028" se trade à 42¢, le marché estime la probabilité à 42%.

L'Hypothèse des Marchés Efficients (EMH)

Eugene Fama (Nobel 2013) a théorisé que les marchés financiers intègrent instantanément toute l'information disponible dans les prix. En théorie, il est impossible de "battre le marché" de manière consistante. C'est la version forte de l'EMH.

Mais les prediction markets ne sont PAS des marchés financiers classiques. Ils souffrent de plusieurs inefficiences structurelles :

- **Liquidité faible** — Polymarket a ~\$500M de volume mensuel (vs \$50B/jour sur le NYSE). Peu de participants = prix moins précis = plus d'opportunités.
- **Biais comportementaux** — Les parieurs surestiment les événements spectaculaires (anchoring), suivent la foule (herding), et sur-réagissent aux dernières nouvelles (recency bias). Kahneman & Tversky ont démontré ces biais dans Prospect Theory (Nobel 2002).
- **Asymétrie d'information** — Un trader qui croise les données Polymarket avec des sources externes (données macro, order books, news en temps réel) a un avantage sur le parieur moyen qui trade au feeling.
- **Frictions d'exécution** — Les spreads (écart bid/ask) sur Polymarket sont larges (3-10% vs 0.01% sur les marchés financiers). Celui qui comprend la microstructure du carnet d'ordres peut entrer à de meilleurs prix.
- **Absence de market makers professionnels** — Sur les marchés financiers, des algos de Citadel ou Jane Street maintiennent des prix efficaces. Sur Polymarket, c'est encore le Far West.

"Les marchés de prédiction sont efficaces à 80-90%. C'est dans les 10-20% restants que se cache l'alpha."

La Loi des Grands Nombres + Expected Value

Le principe fondamental de notre algo repose sur l'Expected Value (EV) positif. Si on identifie un marché où le prix est à 40¢ mais que notre estimation de la vraie probabilité est de 55%, alors :

$$EV = (0.55 \times 60¢ \text{ de gain}) - (0.45 \times 40¢ \text{ de perte}) = 33¢ - 18¢ = +15¢ \text{ par contrat}$$

Sur un grand nombre de trades à EV positive, la Loi des Grands Nombres garantit que le résultat converge vers le profit espéré. C'est exactement comme ça que les casinos gagnent — sauf qu'ici, c'est nous le casino.

Le critère de Kelly

Pour dimensionner les positions, on utilise la formule de Kelly (1956) :

$$f^* = (bp - q) / b$$

Où b = cote nette, p = probabilité estimée de gagner, q = probabilité de perdre. Kelly donne la fraction optimale du capital à miser pour maximiser la croissance long terme sans risquer la ruine. En pratique, on utilise un "demi-Kelly" (50% de la taille suggérée) pour être plus conservateur.

2. VUE D'ENSEMBLE

Le pipeline en 8 étapes

Étape	Nom	Input	Output
1	Scan Polymarket	CLOB API (paginated)	~200-500 marchés actifs
2	Enrichissement profond	Order book, spread, prix 7j	Liquidité, momentum, spread
3	Classification domaine	Question + tags + keywords	1 domaine parmi 5
4	Pré-filtrage	Prob 10-90%, <60j, tri score	Top 10 par domaine (~50)
5	Données externes	APIs gratuites (crypto, éco...)	Contexte cross-référencé
6	Analyse IA	Claude Haiku / GPT-4o-mini	Position, mispricing, confiance
7	Score composite	6 facteurs pondérés + bonus	Score 0-100
8	Routage signal	Score + qualité	VIP / Free / Academy / Skip

Le pipeline complet tourne en ~5-10 minutes. Il scanne tous les marchés Polymarket, enrichit les plus intéressants avec des données de marché profondes, les classe par domaine, puis utilise une IA pour analyser chaque opportunité selon une méthodologie macro → méso → micro. Le résultat est un score composite qui détermine si le signal mérite d'être envoyé, et dans quel canal.

3. ÉTAPE 1 — Scan Polymarket

Source de données primaire

On utilise l'API CLOB (Central Limit Order Book) de Polymarket, pas l'ancienne API Gamma qui a été dépréciée. Le CLOB donne accès aux marchés actifs avec leurs carnets d'ordres.

Ce qu'on récupère pour chaque marché :

- **Question** : l'intitulé du marché ("Will Trump win 2028?")
- **Prix YES / NO** : le prix actuel en centimes (ex: 42¢ / 58¢)
- **Probabilité implicite** : $\text{prix YES} \times 100 = \%$ de probabilité
- **Date de résolution** : quand le marché ferme
- **Jours restants** : combien de temps avant la résolution
- **Tokens IDs** : identifiants pour requêter l'order book
- **Rewards** : si Polymarket incentive les market makers (signe de liquidité)
- **Tags** : catégories assignées par Polymarket

Filtres appliqués :

- Marché actif (pas fermé, pas archivé)
- Accepte les ordres (accepting_orders = true)
- Résolution dans les 90 jours
- Prix YES valide (pas null)

On pagine les résultats (max 5 pages) pour récupérer ~200-500 marchés actifs.

4. ÉTAPE 2 — Enrichissement Profond

Order Book + Spread + Momentum

Pour les 50 marchés les plus intéressants (probabilité entre 10% et 90%), on va chercher des données de microstructure :

- **Order Book (carnet d'ordres)** : combien d'argent est posé en bid (achat) et en ask (vente). Un carnet profond = plus facile d'entrer/sortir sans bouger le prix.
- **Spread (écart bid-ask)** : la différence entre le meilleur prix d'achat et de vente. Un spread serré (<2%) = marché liquide et efficient. Un spread large (>5%) = opportunité mais risque d'exécution.
- **Midpoint** : le prix moyen entre bid et ask — le "vrai" prix du marché.
- **Historique 7 jours** : le momentum. Le prix monte ou descend ? Un mouvement de +10% en 7 jours sur un marché qui résout dans 14 jours est un signal fort.

La liquidité est estimée comme : $\text{bidDepth} + \text{askDepth}$ (total des \$ dans le carnet). C'est un proxy imparfait mais efficace.

5. ÉTAPE 3 — Classification par Domaine

Mots-clés pondérés + anti-mots-clés

Chaque marché est classifié dans un des 5 domaines. Le système utilise un scoring par mots-clés pondérés (poids 1 à 3) avec des mots-clés négatifs qui réduisent le score.

Domaine	Poids global	Exemples mots-clés (poids)	Anti mots-clés
Politique	35%	trump(3), election(3), ceasefire(3), nato(3)	bitcoin, nfl, nasdaq
Crypto	25%	bitcoin(3), ethereum(3), defi(3), halving(3)	election, president, nfl
Économie	20%	fed(2), inflation(3), s&p 500(3), tariff(3)	bitcoin, nfl, election
Sports	15%	nfl(3), premier league(3), champions league(3)	bitcoin, fed, inflation
Culture	5%	oscar(3), openai(2), elon musk(3), tiktok(2)	bitcoin, election, nfl

Les tags Polymarket comptent double (répétés 2x dans le texte de matching) car ils sont plus fiables que la question seule. Si un marché parle de "Bitcoin ETF approval", les mots-clés crypto (bitcoin×3, etf×2) battent largement l'anti-mot-clé "etf" du domaine économie.

6. ÉTAPE 4 — Pré-filtrage

Sélection des meilleurs candidats

Avant d'envoyer les marchés à l'IA (qui coûte des tokens), on filtre :

- **Probabilité 10-90%** : les extrêmes (1% ou 99%) n'offrent pas assez de upside
- **Résolution ≤ 60 jours** : on veut du ROI rapide, pas des paris à 6 mois
- **Tri par score composite** : rewards (bonus +100) + liquidité/100 - distance à 50% - jours restants
- **Top 10 par domaine** : on analyse max 50 marchés au total

La logique : les marchés proches de 50% sont les plus intéressants car ils offrent le meilleur ratio risque/récompense symétrique. Un marché à 90% YES n'a que 10¢ de upside mais 90¢ de downside.

7. ÉTAPE 5 — Données Externes

Cross-referencing multi-sources

C'est ici que l'algo prend son avantage. Au lieu de trader uniquement sur le prix Polymarket, on croise avec des données réelles pour valider ou invalider la cote.

Domaine	Sources	Données récupérées
Crypto	CoinGecko, Fear & Greed Index, DeFi Llama, NewsAPI	Prix BTC/ETH/SOL, variation 24h, market cap, indice peur/avidité (7j+30j), TVL DeFi, coins trending, headlines
Économie	Treasury.gov, FRED (Fed), Alpha Vantage, NewsAPI	Taux obligataires US, CPI (inflation), taux de chômage, fed funds rate, BTC comme indicateur macro
Politique	Congress.gov, NewsAPI	Projets de loi récents, votes, headlines politiques
Sports	The Odds API, Football-Data.org, NewsAPI	Cotes sportives (bookmakers), classements ligues, headlines sportives
Culture	NewsAPI	Headlines pertinentes

Toutes ces APIs sont gratuites (free tier). Le Fear & Greed Index par exemple donne le sentiment crypto sur 30 jours — si l'indice dit "Extreme Fear" (≤ 25) mais qu'un marché Polymarket price un événement bullish à seulement 30%, il y a peut-être un mispricing causé par la panique.

8. ÉTAPE 6 — Analyse IA

Le cerveau de l'Oracle

Chaque marché pré-filtré est analysé par une IA (Claude Haiku en priorité, GPT-4o-mini en fallback) avec un prompt structuré en 3 niveaux :

Méthodologie MACRO → MÉSO → MICRO

- **MACRO (contexte global)** : situation du domaine dans son ensemble. Ex pour crypto : cycle de marché, sentiment, régulation, flux institutionnels.
- **MÉSO (facteurs spécifiques)** : quels événements concrets peuvent faire bouger cette cote ? Précédents historiques, données quantitatives.
- **MICRO (le trade)** : la cote reflète-t-elle la vraie probabilité ? Si non, quel catalyseur va corriger le mispricing ? Quel est le timing ?

L'IA renvoie un JSON structuré avec :

- **estimated_probability** : sa propre estimation (0-100%)
- **mispricing_direction** : UNDERVALUED_YES, OVERVALUED_YES, ou FAIR
- **mispricing_pct** : l'écart en points de % entre le prix et la "vraie" proba
- **position** : YES, NO, ou PASS (si pas assez de conviction)

- **entry / target / stop_loss** : prix d'entrée, objectif, stop loss
- **risk_reward_ratio** : combien on gagne vs combien on risque
- **confidence** : niveau de confiance de l'IA (0-100)
- **signal_quality** : A (excellent), B (bon), C (médiocre)
- **key_catalyst** : le déclencheur principal qui va faire bouger le prix

Règle stricte : si le mispricing estimé est $< 5\%$, l'IA doit répondre PASS. On ne trade que quand il y a un vrai edge.

10. ÉTAPE 8 — Routage des Signaux

3 canaux de distribution

Canal	Critère	Fréquence	Format
VIP Daily	Score \geq 65	1-5 signaux/jour	Concis, actionnable : Entry / Target / Stop / R:R
Free Weekly	Score \geq 80 + Qualité A	1 signal/semaine (le meilleur)	Signal complet + analyse + catalyseur
Academy	Score \geq 55 + Confiance \geq 60	1 signal/semaine	Éducatif : macro \rightarrow méso \rightarrow micro + leçon à retenir

La logique : le canal Free ne publie que le meilleur signal de la semaine (Score \geq 80, qualité A). C'est le "teaser" qui donne envie de passer VIP. Le VIP publie 1-5 signaux/jour à partir de Score 65. L'Academy est éducatif — on explique le raisonnement complet pour apprendre à trader les prediction markets.

Format des signaux :

VIP : "■ YES 'Bitcoin above \$120K by March' — Entry 42¢ / Target 65¢ (+55%) / Stop 30¢ / R:R 1.9 / Score 75"

Free : signal complet avec analyse, catalyseur, risques, score détaillé.

Academy : analyse complète macro \rightarrow méso \rightarrow micro + "Ce qu'on apprend : comment repérer un mispricing de X% sur un marché de prédiction."

11. LEVIERS D'AMÉLIORATION

Ce qu'on peut optimiser

A. Données

- **Ajouter des sources payantes** : Nate Silver's model, 538 data, Metaculus probabilities — croiser avec Polymarket pour détecter les divergences
- **Social sentiment** : analyser Twitter/X en temps réel (mentions, tone, volume) sur les sujets liés aux marchés. Un spike de tweets négatifs peut précéder un mouvement de prix
- **Google Trends** : le volume de recherche sur un sujet corrèle avec l'intérêt des parieurs retail. Un spike de recherche = afflux de volume désinformé = mispricing potentiel
- **On-chain data** : pour les marchés crypto, analyser les flux de whales (grands portefeuilles) sur Bitcoin/Ethereum comme indicateur avancé

B. Modèle de scoring

- **Backtesting** : rejouer l'algo sur les marchés résolus des 6 derniers mois pour calibrer les poids. Le 35% sur le mispricing est-il optimal ? Peut-être que 40% serait mieux
- **Apprentissage adaptatif** : suivre les résultats réels (win/loss) et ajuster les poids automatiquement. Si la politique sur-performe, augmenter son poids
- **Score de calibration** : mesurer si nos estimations de probabilité sont fiables. Si on dit 70%, est-ce que l'événement arrive 70% du temps ? (Brier Score)
- **Détection de régime** : le marché est en mode "risk-on" ou "risk-off" ? Adapter l'agressivité des signaux selon le régime macro

C. Exécution

- **Timing d'entrée** : analyser à quelle heure de la journée les prix sont les plus favorables. Sur les marchés US, les prix bougent souvent à l'ouverture NYSE (15h30 Paris)
- **Position sizing (Kelly)** : implémenter la formule de Kelly pour recommander la taille de position optimale en % du capital
- **Multi-plateforme** : comparer les prix Polymarket vs Kalshi vs Metaculus. Si le même événement est à 42% sur Polymarket mais 55% sur Kalshi, c'est un signal d'arbitrage pur
- **Alertes en temps réel** : ne pas attendre le scan quotidien. Si un marché bouge de >10% en 1h, déclencher un scan immédiat et alerter

D. Track record transparent

Le facteur #1 de crédibilité : publier un track record vérifiable. Chaque signal envoyé doit être timestamped et tracké publiquement. Après 100 signaux, on peut calculer : win rate, ROI moyen, max drawdown, Sharpe ratio. C'est ça qui vend le VIP.