LOG680
Introduction à l'approche DevOps

Analyser la télémétrie pour mieux anticiper les problèmes et atteindre les objectifs

The DevOps Handbook

Part IV, Chap 15

Objectifs d'apprentissage

- Expliquer en quoi consiste la technique de détection des valeurs aberrantes ("outlier detection") utilisé par Netflix. Comment Netflix utilisait-il cette technique? Quel a été le résultat de l'utilisation de la technique?
- Expliquer en quoi consiste l'utilisation de moyennes et des écarts types pour analyser une métrique de production
- Expliquer ce qu'est une distribution non-gaussienne et ce qu'on peut faire pour détecter les variances lorsque les données n'ont pas une distribution Gaussienne
- Expliquer en quoi consiste la technique de lissage. Expliquer aussi l'effet de son utilisation

Sujets

- Introduction
- Utilisation de moyennes et des écarts types
- Instrumentation et alerte
- Problèmes de distribution non-gaussienne
- Détection d'anomalies
- Conclusion

- **Introduction**
- Utilisation de moyennes et des écarts types
- Instrumentation et alerte
- Problèmes de distribution non-gaussienne
- Détection d'anomalies
- Conclusion

Introduction

- Comme nous l'avons vu dans le chapitre précédent, nous avons besoin de suffisamment de télémétrie de production dans nos applications et infrastructures pour voir et résoudre les problèmes à mesure qu'ils surviennent

- Dans ce chapitre, nous allons **créer des outils nous permettant de découvrir les écarts et les signaux de défaillance qui se trouvent dans notre télémétrie de production**, de manière à **éviter les défaillances catastrophiques**
- De nombreuses techniques statistiques seront présentées, ainsi que des études de cas démontrant leur utilisation

Exemple Netflix

- Netflix constitue un bon exemple d'analyse de la télémétrie pour rechercher et résoudre de manière proactive les problèmes avant que les clients ne soient impactés
  - Fournisseur mondial de films en streaming et de séries télévisées
  - Chiffre d'affaires (2015): $6,2 milliards USD – 75M d'abonnés
  - Objectif: fournir la meilleure expérience possible à ceux qui regardent des vidéos en ligne dans le monde entier, ce qui nécessite une infrastructure de diffusion robuste, évolutive, et résiliente

- Roy Rapoport (Manager, Insight Engineering) décrit l'un des défis de la gestion du service de diffusion vidéo dans le nuage Netflix:

  «Dans un troupeau de bovins qui doivent tous avoir l'air identique et se comporter de la même manière, quels sont les animaux qui ont une apparence différente?

  Ou plus concrètement, si nous avons un cluster de calcul "stateless" de 1 000 nœuds exécutant tous le même logiciel et soumis à la même charge de trafic approximative, notre défi est de trouver les nœuds qui ne ressemblent pas au autres nœuds. »

Exemple Netflix

- L'une des techniques statistiques utilisées par l'équipe chez Netflix en 2012 a été la **détection des valeurs aberrantes ("outlier detection")**, définies par Victoria J. Hodge et Jim Austin (University of York) comme visant à

  « **détecter les conditions de fonctionnement anormales pouvant entraîner une dégradation importante des performances**, comme un défaut de rotation du moteur d'un avion ou un problème d'écoulement dans un pipeline. »

- Rapoport explique

  « Netflix utilisait la détection des valeurs aberrantes d'une manière très simple, c'était d'abord de calculer ce qu'était la «normale actuelle», en fonction de la population de nœuds dans un cluster de calcul. Nous identifions ensuite les nœuds qui ne correspondaient pas à ce modèle et les avons retirés de la production. »

Introduction

- Tout au long de ce chapitre, nous explorerons de nombreuses **techniques statistiques et de visualisation**, y compris la **détection des valeurs aberrantes**, que nous pourrons utiliser **pour analyser notre télémétrie afin de mieux anticiper les problèmes**
- Cela nous **permet de résoudre les problèmes plus rapidement**, **à moindre coût**, et **plus tôt** que jamais, **avant que notre client ou quiconque au sein de notre organisation ne soit touché**
- En outre, nous allons également **créer plus de contexte pour nos données afin de nous aider à prendre de meilleures décisions et à atteindre nos objectifs organisationnels**

- Introduction
- **Utilisation de moyennes et des écarts types**
- Instrumentation et alerte
- Problèmes de distribution non-gaussienne
- Détection d'anomalies
- Conclusion

Utilisation de moyennes et des écarts types

- L'une des techniques statistiques les plus simples que nous puissions utiliser pour analyser une métrique de production consiste à calculer sa **moyenne** et ses **écarts-types**
- Créer un filtre qui détecte quand cette métrique est significativement différente de sa norme, et même configurer notre alerte afin que nous puissions prendre des mesures correctives
  - Par exemple, avertir le personnel de production sur appel à 2 heures du matin pour enquêter sur les requêtes de base de données lorsqu'elles sont nettement plus lentes que la moyenne
  - Lorsque les services de production critiques rencontrent des problèmes, il peut être judicieux de se faire réveiller à 2 heures du matin
  - Cependant, lorsque nous créons des alertes qui ne peuvent donner lieu à aucune action ou sont des faux positifs, nous réveillons inutilement des personnes au milieu de la nuit
- Comme l'a observé John Vincent (un des premiers leaders du mouvement DevOps)

  «La fatigue liée aux alertes est le plus gros problème que nous ayons à l'heure actuelle…

  Nous devons être plus intelligents en ce qui concerne nos alertes, sinon nous allons tous devenir fous. »

Utilisation de moyennes et des écarts types

- Nous créons de meilleures alertes en **augmentant le rapport signal-bruit**, en nous **concentrant sur les variances** ou les **valeurs aberrantes** qui importent
  - Supposons que nous analysions le nombre de tentatives de connexion non autorisées par jour
  - Nos données collectées ont une distribution gaussienne (c'est-à-dire une distribution normale ou une courbe en forme de cloche) qui correspond au graphique de la figure 29
  - La ligne verticale au centre de la courbe en cloche est la moyenne
  - Les premier, deuxième et troisième écarts types sont indiqués pour les autres lignes verticales
  - Les autres lignes verticales contiennent respectivement 68%, 95% et 99,7% des données

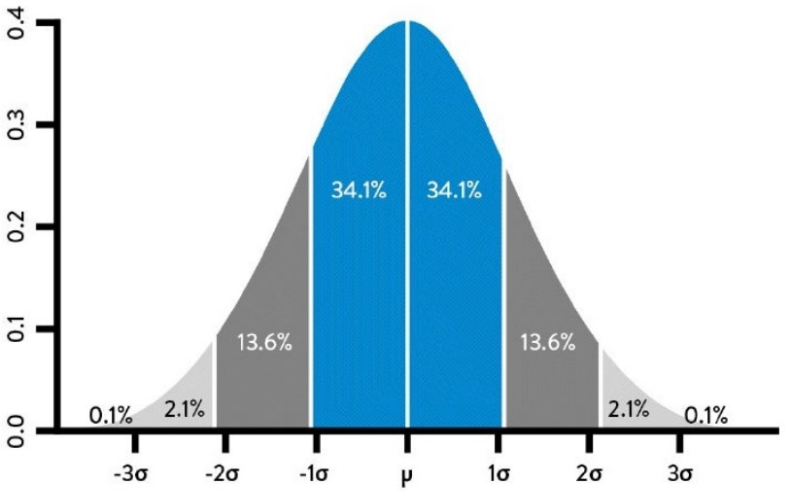

Figure 29: *Standard deviations (σ) & mean (μ) with Gaussian distribution (Source: Wikipedia's "Normal Distribution" entry, https://en.wikipedia.org/wiki/Normal_distribution.)*

Utilisation de moyennes et des écarts types

- Une utilisation courante des écarts-types consiste à **inspecter périodiquement l'ensemble de données pour une métrique donnée** et **générer une alerte si les valeurs diffèrent de manière significative de la moyenne**
  - Par exemple, nous pouvons définir une alerte lorsque le nombre de tentatives de connexion non autorisées par jour est supérieur de trois écarts types à la moyenne
  - À condition que cet ensemble de données ait une distribution gaussienne, nous nous attendrions à ce que seulement 0,3% des points de données déclenchent l'alerte
- Même ce type d'analyse statistique simple est utile, car personne n'a dû définir une valeur seuil statique, ce qui est impossible si nous suivons des milliers ou des centaines de milliers de métriques de production
- Les termes *télémétrie*, *métrique* et *ensembles de données* de manière interchangeable.
  - En d'autres termes, une métrique (e.g. «temps de chargement d'une page») correspondra à un ensemble de données (par exemple, 2 ms, 8 ms, 11 ms, etc.), terme utilisé par les statisticiens pour décrire une matrice de points de données dans laquelle chaque colonne représente une variable sur laquelle des opérations statistiques sont effectuées

- Introduction
- Utilisation de moyennes et des écarts types
- **Instrumentation et alerte**
- Problèmes de distribution non-gaussienne
- Détection d'anomalies
- Conclusion

Instrumentation et alerte

- Tom Limoncelli (co-auteur de "The Practice of Cloud System Administration: Designing and Operating Large Distributed Systems" et ex-SRE chez Google), raconte l'histoire suivante sur la surveillance:

  «Quand les gens me demandent des recommandations sur ce qu'il faut surveiller, je plaisante que dans un monde idéal, nous supprimerions toutes les alertes que nous avons actuellement dans notre système de surveillance

  Ensuite, après chaque panne visible par l'utilisateur, nous nous demanderions quels indicateurs auraient pu prédire cette panne,

  puis nous les ajouterions à notre système de surveillance, en alertant si nécessaire.

  Répéter.

  Maintenant, nous n'aurions plus que des alertes qui préviennent les pannes, au lieu d'être bombardées par des alertes indiquant une panne déjà survenue. »

Instrumentation et alerte

- L'une des manières les plus simples de procéder consiste à **analyser nos incidents les plus graves du passé récent** (par exemple, 30 jours) et à **créer une liste de télémétries** qui aurait pu permettre une **détection et un diagnostic plus rapides du problème**, ainsi qu'une **confirmation plus rapide qu'un correctif efficace a été implanté**
- Par exemple, si nous rencontrons un problème où notre serveur Web NGINX cessait de répondre aux demandes, nous devrions nous pencher sur les métriques qui auraient pu nous avertir plus tôt que nous commencions à nous écarter des opérations standard:
  - **Niveau d'application**: temps de chargement des pages Web croissant, etc.
  - **Niveau du système d'exploitation**: mémoire disponible insuffisante sur le serveur, espace disque insuffisant, etc.
  - **Niveau base de données**: les temps de transaction de base de données prennent plus de temps que la normale, etc.
  - **Niveau réseau**: nombre de serveurs en fonction qui tombent derrière le "load balancer", etc.
- Chacune de ces métriques est un **précurseur potentiel d'un incident de production**

Instrumentation et alerte

- Nous devons **configurer nos systèmes d'alerte** de manière à les **informer lorsque ces métriques s'écartent suffisamment de la moyenne** pour pouvoir **prendre des mesures correctives**
- En répétant ce processus sur des **signaux de défaillance de plus en plus faibles**, nous **détectons des problèmes de plus en plus tôt dans le cycle de vie**, ce qui **réduit le nombre d'incidents**, et les **quasi-incidents ("near misses")**, ayant des répercussions sur les clients
- En d'autres termes, nous **évitons les problèmes et permettons une détection et une correction plus rapides**

- Introduction
- Utilisation de moyennes et des écarts types
- Instrumentation et alerte
- **Problèmes de distribution non-gaussienne**
- Détection d'anomalies
- Conclusion

Problèmes de distribution non-gaussienne

- L'utilisation de moyennes et d'écarts-types pour détecter la variance peut s'avérer extrêmement utile
- Cependant, l'utilisation de ces techniques sur de nombreux jeux de données de télémétrie utilisés dans Ops ne générera pas les résultats souhaités
  - Comme le remarque Dr. Toufic Boubez, «Non seulement nous recevrons des appels à 2h00 du matin, mais nous les aurons aussi à 2h37, 4h13 et 5h17. Cela se produit lorsque les données sous-jacentes que nous surveillons n'ont pas une distribution gaussienne
- En d'autres termes, **lorsque les données n'ont pas une distribution Gaussienne, les propriétés associées aux écarts-types ne s'appliquent pas**
  - Par exemple, considérons le scénario dans lequel nous surveillons le nombre de téléchargements de fichiers par minute à partir de notre site Web
  - Nous souhaitons détecter les périodes au cours desquelles nous avons un nombre de téléchargements anormalement élevé, par exemple lorsque notre taux de téléchargement est supérieur à trois écarts types par rapport à notre moyenne, de manière à pouvoir augmenter de manière proactive notre capacité

Problèmes de distribution non-gaussienne

- La figure 30 montre notre nombre de téléchargements simultanés par minute dans le temps, ainsi qu'un graphique (barre) identifiant les alertes résultantes
- Lorsque la barre est noire, le nombre de téléchargements au cours d'une période donnée (parfois appelée «fenêtre glissante») est d'au moins trois écarts types par rapport à la moyenne. Sinon, il est gris

- Le problème évident que montre le graphique est que nous alertons presque tout le temps. En effet, dans presque toute période donnée, nous avons des cas où le nombre de téléchargements dépasse notre seuil de trois écarts-types

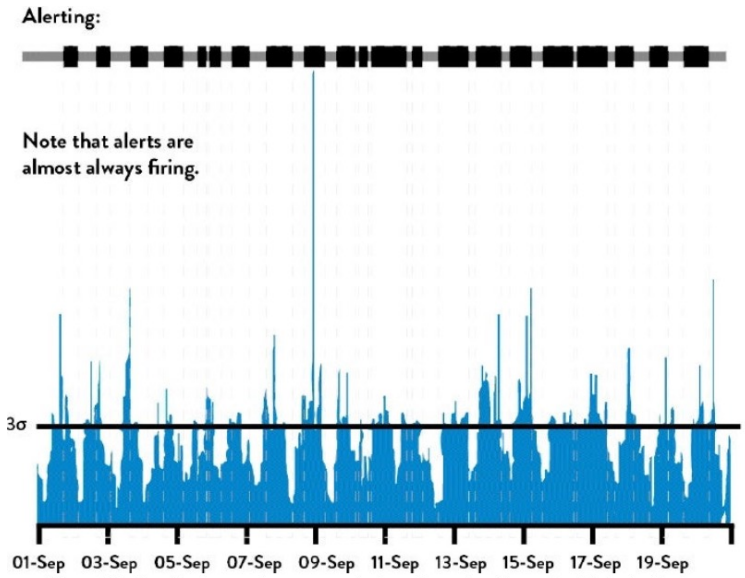

Figure 30: *Downloads per minute: over-alerting when using "3 standard deviation" rule (Source: Dr. Toufic Boubez, "Simple math for anomaly detection.")*

Problèmes de distribution non-gaussienne

- Pour confirmer cela, lorsque nous créons un histogramme (voir la figure 31) qui indique la fréquence de téléchargements par minute, nous constatons qu'il n'a pas la forme classique de la distribution gaussienne (courbe en cloche symétrique)
- Au lieu de cela, il est évident que la distribution est biaisée vers le bas, ce qui montre que la plupart du temps, nous effectuons très peu de téléchargements par minute, mais que leur nombre compte souvent trois écarts-types plus élevés

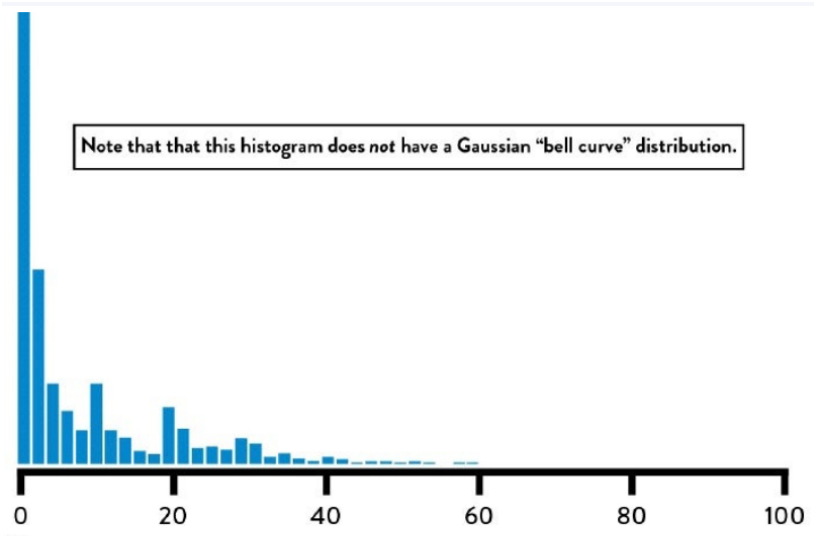

Figure 31: *Downloads per minute: histogram of data showing non-Gaussian distribution (Source: Dr. Toufic Boubez, "Simple math for anomaly detection.")*

Problèmes de distribution non-gaussienne

- De nombreux ensembles de données de production ont une distribution non gaussienne
- Dr. Nicole Forsgren:

  «Dans les opérations, beaucoup de nos ensembles de données ont ce que nous appelons la distribution «chi carré ».

  L'utilisation d'écarts-types pour ces données entraîne non seulement une alerte excessive ou insuffisante, mais également des résultats absurdes. »

  « Lorsque vous calculez un nombre de téléchargements simultanés inférieur de trois écarts-types à la moyenne, vous vous retrouvez avec un nombre négatif, ce qui n'a évidemment aucun sens. »

Problèmes de distribution non-gaussienne

- Les alertes trop fréquentes ont pour effet de réveiller les ingénieurs des opérations au milieu de la nuit pendant de longues périodes, même s'ils n'ont que peu d'actions à prendre
- **Le problème associé à la sous-alerte est tout aussi important**
  - Par exemple, supposons que nous surveillons le nombre de transactions terminées et que le nombre de transactions terminées chute de 50% au milieu de la journée en raison d'une défaillance d'un composant logiciel
  - Si cela reste toujours à moins de trois écarts-types de la moyenne, aucune alerte ne sera générée, ce qui signifie que nos clients découvriront le problème avant nous
  - À ce stade, le problème risque d'être beaucoup plus difficile à résoudre

- Heureusement, il existe des techniques que nous pouvons utiliser pour détecter des anomalies dans des ensembles de données, même non-gaussiens, qui sont décrites ci-après

Étude de cas : Netflix (2012)

Étude de cas : Netflix (2012)

- Étude de cas : Capacité de mise à l'échelle automatique de Netflix (2012)
- **Scryer**
  - Outil développé par Netflix pour améliorer la qualité de service
  - Corrige certaines des faiblesses d'Amazon Auto Scaling (AAS)
  - **Augmente et réduit de manière dynamique le nombre de serveurs de calcul AWS** en fonction des données de la charge de travail.
  - **Prévoit les demandes des clients** en se basant sur les schémas d'utilisation historiques et fournit la capacité nécessaire
- Scryer a abordé trois problèmes avec AAS
  - Le premier concernait les pics rapides de la demande
    - Les temps de démarrage des instances AWS pouvant aller de 10 à 45 minutes, une capacité de calcul supplémentaire a souvent été fournie trop tard pour faire face à des pics de demande
  - Le deuxième problème était qu'après les pannes, la diminution rapide de la demande des clients avait conduit AAS à supprimer une trop grande capacité de calcul pour gérer la demande entrante future
  - Le troisième problème était que le SAA n'a pas pris en compte les modèles de trafic d'utilisation connus lors de la planification de la capacité de calcul

Étude de cas : Netflix (2012)

- Netflix a profité du fait que leurs "customer viewing patterns" étaient étonnamment cohérentes et prévisibles, bien qu'elles n'aient pas une distribution gaussienne
  - Vous trouverez ci-dessous un tableau reflétant les demandes des clients par seconde tout au long de la semaine de travail, illustrant les habitudes de visionnage des clients, régulières et cohérentes, du lundi au vendredi

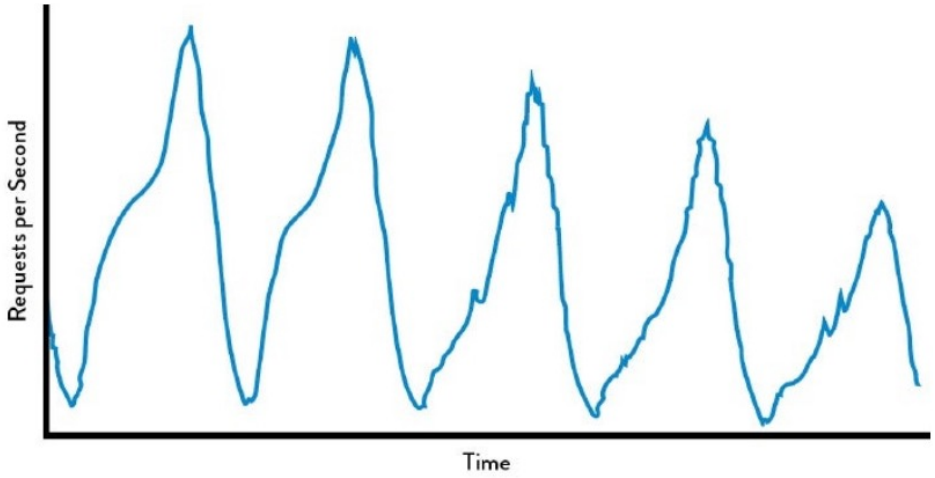

Figure 32: *Netflix customer viewing demand for five days (Source: Daniel Jacobson, Danny Yuan, and Neeraj Joshi, "Scryer: Netflix's Predictive Auto Scaling Engine,"* The Netflix Tech Blog, *November 5, 2013, http://techblog.netflix.com/2013/11/scryer-netflixs-predictive-auto-scaling.html.)*

Étude de cas : Netflix (2012)

- Scryer a utilisé une **combinaison de détections de valeurs aberrantes** pour **éliminer les points de données parasites**, puis des techniques telles que la **transformée de Fourier rapide (FFT)** et la **régression linéaire** pour **lisser les données** tout en préservant les pointes de trafic légitimes qui se reproduisent dans leurs données
- Résultat: **Netflix peut prévoir la demande de trafic avec une précision surprenante**
- Quelques mois seulement après avoir utilisé Scryer pour la première fois en production, Netflix a considérablement amélioré l'expérience de visionnage de ses clients, la disponibilité de ses services, et la réduction des coûts Amazon EC2

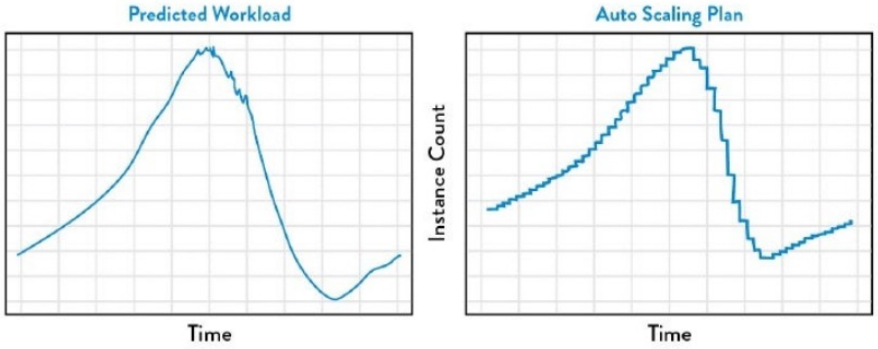

Figure 33: *Netflix Scryer forecasting customer traffic and the resulting AWS schedule of compute resources (Source: Jacobson, Yuan, Joshi, "Scryer: Netflix's Predictive Auto Scaling Engine.")*

- Introduction
- Utilisation de moyennes et des écarts types
- Instrumentation et alerte
- Problèmes de distribution non-gaussienne
- **Détection d'anomalies**
- Conclusion

Détection d'anomalies

- Lorsque nos données n'ont pas de distribution gaussienne, nous pouvons toujours trouver des variances remarquables en utilisant diverses méthodes
- Ces techniques sont en gros classées dans la **détection des anomalies**, souvent définies comme «la **recherche d'éléments ou d'événements qui ne se conforment pas à un modèle attendu**»
- Certaines de ces fonctionnalités peuvent être trouvées dans nos outils de surveillance, d'autres peuvent nécessiter l'aide de personnes ayant des compétences en statistiques
- Tarun Reddy (VP développement et opérations chez Rally Software) plaide activement en faveur de cette **collaboration active entre les opérations et les statistiques**:

  «Pour améliorer la qualité de service, nous intégrons toutes nos métriques de production dans "Tableau", un logiciel d'analyse statistique.

  Nous avons même un ingénieur des opérations formé en statistique qui écrit du code R (un autre progiciel statistique).

  Cet ingénieur a son propre backlog, qui regorge de demandes d'autres équipes de l'entreprise qui veulent trouver des variances plus tôt, avant que les variances augmentent, ce qui pourrait affecter les clients. »

Technique de lissage

- L'une des techniques statistiques que nous pouvons utiliser est le **lissage**
- Convient particulièrement si nos données sont une série chronologique, c'est-à-dire que chaque point de données possède un horodatage (« timetsamp »)
  - Par exemple, événements de téléchargement, événements de transaction terminés, etc.
- Le lissage implique souvent l'utilisation de moyennes mobiles ("moving average"), qui transforment nos données en faisant la moyenne de chaque point avec toutes les autres données de notre fenêtre glissante
- Cela a pour effet d'atténuer les fluctuations à court terme et de mettre en évidence les tendances ou les cycles à plus long terme
- La figure 34 montre un exemple de cet effet de lissage.
  - La ligne bleue représente les données brutes, tandis que la ligne noire indique la moyenne mobile à trente jours (c'est-à-dire la moyenne des trente derniers jours)

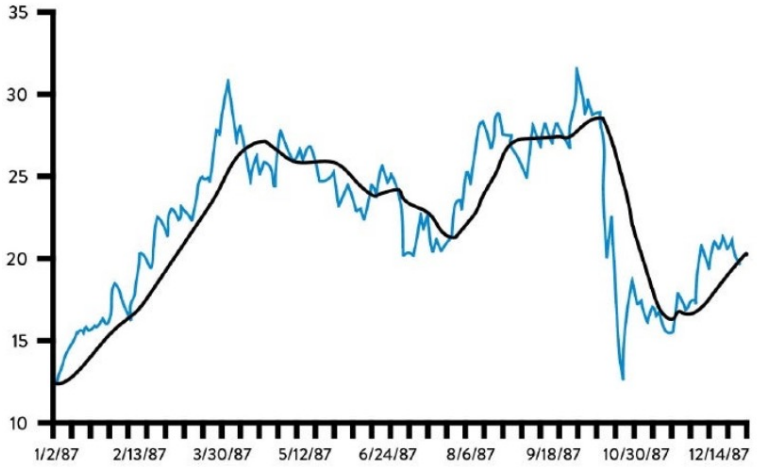

Figure 34 : *Autodesk share price and thirty day moving average filter (Source: Jacobson, Yuan, Joshi, "Scryer: Netflix's Predictive Auto Scaling Engine.")*

Détection d'anomalies

- Des techniques de filtrage plus exotiques existent, telles que
  - Transformées de Fourier rapides (FFT), qui ont été largement utilisées dans le traitement d'images
  - Test de Kolmogorov-Smirnov (trouvé dans Graphite et Grafana), qui est souvent utilisé pour trouver des similitudes ou des différences dans les données métriques périodiques/saisonnières

- On peut s'attendre à ce qu'un grand pourcentage de télémétrie concernant les données des utilisateurs présente des similitudes périodiques/saisonnières
  - Par exemple, trafic Web, transactions de vente au détail, visionnage de films et de nombreux autres comportements des utilisateurs ont des habitudes quotidiennes, hebdomadaires et annuelles étonnamment prévisibles
- Cela nous permet de détecter des situations qui diffèrent des normes historiques, par exemple lorsque le taux de traitement des commandes le mardi après-midi tombe à 50% de nos normes hebdomadaires

Étude de cas : Détection avancée d'anomalies (2014)

Étude de cas : Détection avancée d'anomalies (2014)

- En 2014, à Monitorama, le Dr Toufic Boubez a décrit la puissance des techniques de détection des anomalies
  - En soulignant en particulier l'efficacité du test de Komogorov-Smirnov, technique souvent utilisée en statistiques pour déterminer si deux ensembles de données diffèrent de manière significative et qui se retrouvent dans le populaire outil Graphite/Grafana
- Le but de cette étude de cas n'est pas de faire un tutoriel, mais une démonstration de la manière dont une classe de techniques statistiques peut être utilisée dans notre travail, ainsi que de la manière dont elle est probablement utilisée dans nos organisations dans des applications complètement différentes

Étude de cas : Détection avancée d'anomalies (2014)

- La figure 35 indique le nombre de transactions par minute sur un site de commerce électronique
- Notez la périodicité hebdomadaire du graphique, le volume des transactions diminuant le week-end
- Une inspection visuelle permet de constater que quelque chose de particulier semble se produire la quatrième semaine lorsque le volume des transactions normales ne revient pas à des niveaux normaux le lundi
- Cela suggère un événement sur lequel nous devrions enquêter
- L'utilisation de la règle des trois écarts-types ne nous alerterait que deux fois, manquant ainsi la baisse critique du lundi du volume de transactions
- Idéalement, nous souhaiterions également être avertis que les données ont dévié de notre modèle de lundi prévu

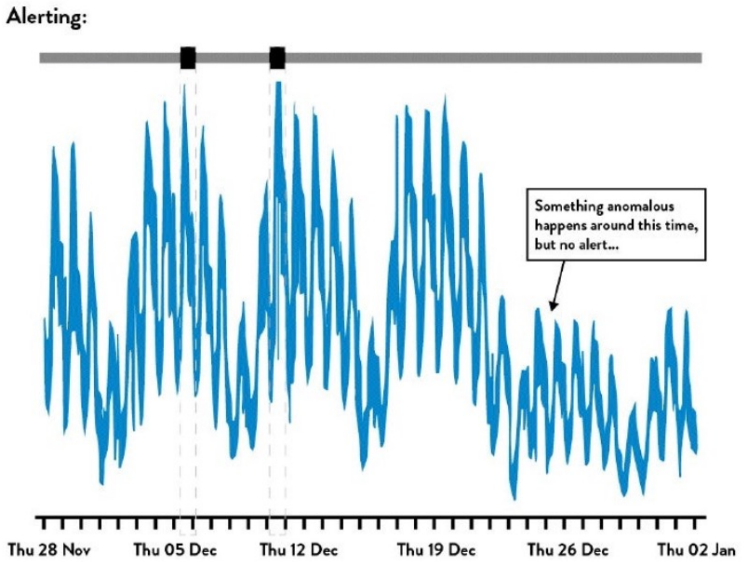

Figure 35: *Transaction volume: under-alerting using "3 standard deviation" rule (Source: Dr. Toufic Boubez, "Simple math for anomaly detection.")*

Étude de cas : Détection avancée d'anomalies (2014)

- La figure 36 montre le même ensemble de données auquel le filtre K-S est appliqué
- La troisième zone soulignant le lundi anormal où le volume des transactions n'est pas revenu à des niveaux normaux
- Cela nous aurait alerté d'un problème dans notre système qu'il aurait été pratiquement impossible de détecter à l'aide d'un contrôle visuel ou d'un écart type
- Dans ce scénario, cette détection précoce pourrait empêcher un événement ayant un impact sur le client et nous aider à mieux atteindre nos objectifs organisationnels

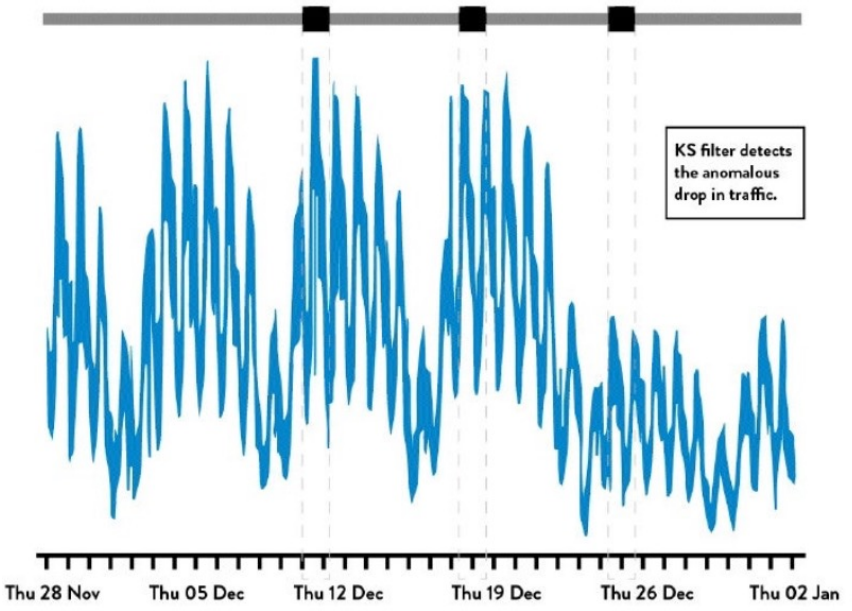

Figure 36: *Transaction volume: using Kolmogorov-Smirnov test to alert on anomalies (Source: Dr. Toufic Boubez, "Simple math for anomaly detection.")*

- Introduction
- Utilisation de moyennes et des écarts types
- Instrumentation et alerte
- Problèmes de distribution non-gaussienne
- Détection d'anomalies
- **Conclusion**

Conclusion

- Dans ce chapitre, nous avons exploré **différentes techniques statistiques pouvant être utilisées pour analyser notre télémétrie de production** afin de pouvoir **détecter et résoudre les problèmes plus tôt**, souvent lorsqu'ils sont encore petits et bien avant qu'ils ne provoquent des conséquences catastrophiques
- Cela nous **permet de détecter des signaux d'échec toujours plus faibles** sur lesquels nous pouvons agir, créant ainsi un système de travail toujours plus sûr et augmentant notre capacité à atteindre nos objectifs
- Des études de cas spécifiques ont été présentées, notamment sur la manière dont Netflix a utilisé ces techniques pour supprimer de manière proactive les serveurs de calcul de la production et mettre à l'échelle automatiquement leur infrastructure de calcul
- Nous avons également discuté de l'utilisation d'une moyenne mobile et du filtre de Kolmogorov-Smirnov, tous deux disponibles dans les outils graphiques de télémétrie courants
- Dans le chapitre suivant, nous décrirons comment intégrer la télémétrie de production au travail quotidien de Développement afin de rendre les déploiements plus sûrs et d'améliorer le système dans son ensemble