4.1 Présentation

Dans les approches abordées jusqu'ici dans ce livre - observer le comportement (chapitre 2) et poser des questions (chapitre 3) - les chercheurs recueillent des données sans changer intentionnellement et systématiquement le monde. L'approche abordée dans ce chapitre - expériences en cours - est fondamentalement différente. Lorsque les chercheurs mènent des expériences, ils interviennent systématiquement dans le monde pour créer des données parfaitement adaptées à la réponse aux questions sur les relations de cause à effet.

Les questions de cause à effet sont très courantes dans la recherche sociale, et les exemples incluent des questions telles que: L'augmentation des salaires des enseignants augmente-t-elle l'apprentissage des élèves? Quel est l'effet du salaire minimum sur les taux d'emploi? Comment la course d'un demandeur d'emploi affecte-t-elle sa chance d'obtenir un emploi? En plus de ces questions explicitement causales, des questions de cause à effet sont parfois implicites dans des questions plus générales sur la maximisation de certaines mesures de performance. Par exemple, la question «De quelle couleur devrait être le bouton de don sur le site Web d'une ONG?» Est vraiment beaucoup de questions sur l'effet de différentes couleurs de bouton sur les dons.

Une façon de répondre aux questions de cause à effet est de rechercher des modèles dans les données existantes. Par exemple, pour revenir à la question sur l'effet des salaires des enseignants sur l'apprentissage des élèves, vous pourriez calculer que les élèves apprennent davantage dans les écoles qui offrent des salaires élevés aux enseignants. Mais cette corrélation montre-t-elle que des salaires plus élevés *incitent les*étudiants à en apprendre davantage? Bien sûr que non. Les écoles où les enseignants gagnent plus peuvent être différentes de plusieurs façons. Par exemple, les étudiants qui fréquentent des écoles où les salaires des enseignants sont élevés peuvent provenir de familles plus aisées. Ainsi, ce qui ressemble à un effet des enseignants pourrait simplement provenir de la comparaison de différents types d'élèves. Ces différences non mesurées entre les étudiants sont appelées *facteurs de confusion* et, en général, la possibilité que les facteurs confondants agissent sur la capacité des chercheurs à répondre aux questions de cause à effet en cherchant des modèles dans les données existantes.

Une solution au problème des facteurs de confusion est d'essayer de faire des comparaisons justes en ajustant les différences observables entre les groupes. Par exemple, vous pourriez être en mesure de télécharger des données sur l'impôt foncier à partir d'un certain nombre de sites Web gouvernementaux. Ensuite, vous pourriez comparer la performance des élèves dans les écoles où les prix des maisons sont similaires, mais les salaires des enseignants sont différents, et vous pourriez toujours constater que les élèves apprennent plus dans les écoles avec une rémunération plus élevée des enseignants. Mais il y a encore beaucoup de facteurs de confusion possibles. Peut-être que les parents de ces étudiants diffèrent dans leur niveau d'éducation. Ou peut-être que les écoles diffèrent par leur proximité avec les bibliothèques publiques. Ou peut-être que les écoles avec une rémunération plus élevée des enseignants ont aussi un salaire plus élevé pour les directeurs, et le salaire principal, et non la rémunération des enseignants, est vraiment ce qui augmente l'apprentissage des élèves. Vous pourriez aussi essayer de mesurer et d'ajuster ces facteurs, mais la liste des facteurs de confusion possibles est essentiellement infinie. Dans de nombreuses situations, vous ne pouvez pas mesurer et ajuster pour tous les facteurs de confusion possibles. En réponse à ce défi, les chercheurs ont développé un certain nombre de techniques pour faire des estimations causales à partir de données non expérimentales - j'en ai discuté quelques-unes au chapitre 2 - mais pour certains types de questions, ces techniques sont limitées. alternative.

Les expériences permettent aux chercheurs d'aller au-delà des corrélations dans les données naturelles afin de répondre de manière fiable à certaines questions de cause à effet. À l'ère analogique, les expériences étaient souvent difficiles d'un point de vue logistique et coûteuses. Maintenant, à l'ère numérique, les contraintes logistiques disparaissent progressivement. Non seulement il est plus facile de faire des expériences comme celles du passé, mais il est maintenant possible de faire de nouvelles expériences.

Dans ce que j'ai écrit jusqu'à présent, j'ai été un peu lâche dans mon langage, mais il est important de distinguer deux choses: les expériences et les expériences contrôlées randomisées. Dans une *expérience* , un chercheur intervient dans le monde et mesure ensuite un résultat. J'ai entendu cette approche décrite comme «perturber et observer». Dans une *expérience contrôlée randomisée,* un chercheur intervient pour certaines personnes et pas pour d'autres, et le chercheur décide quelles personnes reçoivent l'intervention par randomisation (p. Ex. Les expériences contrôlées randomisées créent des comparaisons équitables entre deux groupes: un qui a reçu l'intervention et un qui n'a pas reçu. En d'autres termes, les expériences contrôlées randomisées sont une solution aux problèmes des facteurs de confusion. Les expériences Perturb-and-observer, cependant, impliquent seulement un seul groupe qui a reçu l'intervention, et donc les résultats peuvent mener les chercheurs à la mauvaise conclusion (comme je le montrerai bientôt). Malgré les différences importantes entre les expériences et les expériences contrôlées randomisées, les chercheurs en sciences sociales utilisent souvent ces termes de façon interchangeable. Je vais suivre cette convention, mais, à certains points, je vais rompre la convention pour souligner la valeur des expériences contrôlées randomisées sur des expériences sans randomisation et un groupe de contrôle.

Les expériences contrôlées randomisées se sont avérées être un moyen puissant d'apprendre sur le monde social, et dans ce chapitre, je vais vous montrer plus sur la façon de les utiliser dans votre recherche. Dans la section 4.2, j'illustrerai la logique de base de l'expérimentation avec un exemple d'expérience sur Wikipédia. Ensuite, dans la section 4.3, je vais décrire la différence entre les expériences en laboratoire et les expériences sur le terrain et les différences entre les expériences analogiques et les expériences numériques. En outre, je ferai valoir que les expériences de terrain numériques peuvent offrir les meilleures caractéristiques des expériences de laboratoire analogique (contrôle serré) et des expériences de terrain analogiques (réalisme), le tout à une échelle qui n'était pas possible auparavant. Ensuite, dans la section 4.4, je vais décrire trois concepts - la validité, l'hétérogénéité des effets du traitement, et les mécanismes - qui sont essentiels pour concevoir des expériences riches. Dans ce contexte, je vais décrire les compromis impliqués dans les deux principales stratégies pour mener des expériences numériques: le faire vous-même ou en partenariat avec les puissants. Enfin, je terminerai par quelques conseils de conception sur la façon de tirer parti de la puissance réelle des expériences numériques (section 4.6.1) et décrirai certaines des responsabilités qui découlent de cette puissance (section 4.6.2).

## 4.2 Quelles sont les expériences?

**Expériences contrôlées randomisées ont quatre ingrédients principaux:** le **recrutement des participants, randomisation du traitement,** la **livraison du traitement, et** de **mesure des résultats.**

Les expériences contrôlées randomisées ont quatre ingrédients principaux: le recrutement des participants, la randomisation du traitement, l'administration du traitement et la mesure des résultats. L'ère numérique ne change pas la nature fondamentale de l'expérimentation, mais elle facilite la logistique. Par exemple, dans le passé, il aurait peut-être été difficile de mesurer le comportement de millions de personnes, mais cela se produit couramment dans de nombreux systèmes numériques. Les chercheurs qui savent comment tirer parti de ces nouvelles opportunités pourront mener des expériences impossibles auparavant.

Pour rendre le tout un peu plus concret - à la fois ce qui est resté le même et ce qui a changé - considérons une expérience de Michael Restivo et Arnout van de Rijt (2012) . Ils voulaient comprendre l'effet des récompenses informelles des pairs sur les contributions éditoriales à Wikipedia. En particulier, ils ont étudié les effets de barnstars , un prix que tout Wikipédien peut donner à tout autre Wikipédien pour reconnaître le travail acharné et la diligence raisonnable. Restivo et van de Rijt ont donné des barnstars à 100 Wikipédiens méritants. Ensuite, ils ont suivi les contributions ultérieures des destinataires à Wikipedia au cours des 90 prochains jours. À leur grande surprise, les gens à qui ils décernaient des barnstars avaient tendance à faire moins de modifications après en avoir reçu un. En d'autres termes, les barnstars semblaient décourager plutôt que d'encourager la contribution.

Heureusement, Restivo et van de Rijt ne menaient pas une expérience de «perturbation et observation»; ils couraient une expérience contrôlée randomisée. Donc, en plus de choisir 100 meilleurs contributeurs pour recevoir une barnstar, ils ont également choisi 100 meilleurs contributeurs à qui ils n'ont pas donné un. Ces 100 servaient de groupe de contrôle. Et, de manière critique, qui était dans le groupe de traitement et qui était dans le groupe de contrôle a été déterminé au hasard.

Lorsque Restivo et van de Rijt ont examiné le comportement des personnes du groupe de contrôle, ils ont constaté que leurs contributions diminuaient également. De plus, lorsque Restivo et van de Rijt ont comparé des personnes du groupe de traitement (c.-à-d. Des barnars reçus) à des personnes du groupe témoin, ils ont trouvé que les personnes du groupe de traitement contribuaient environ 60% de plus. En d'autres termes, les contributions des deux groupes diminuaient, mais celles du groupe témoin le faisaient beaucoup plus vite.

Comme l'illustre cette étude, le groupe témoin dans les expériences est critique d'une manière quelque peu paradoxale. Afin de mesurer avec précision l'effet des barnstars, Restivo et van de Rijt devaient observer des personnes qui n'avaient pas reçu de barnstars. Plusieurs fois, les chercheurs qui ne sont pas familiers avec les expériences ne parviennent pas à apprécier la valeur incroyable du groupe de contrôle. Si Restivo et van de Rijt n'avaient pas eu de groupe de contrôle, ils auraient tiré la mauvaise conclusion. Les groupes de contrôle sont si importants que le PDG d'une grande entreprise de casino a déclaré qu'il n'y avait que trois façons de licencier des employés de son entreprise: pour vol, harcèlement sexuel ou pour mener une expérience sans groupe de contrôle (Schrage 2011) .

L'étude de Restivo et van de Rijt illustre les quatre principaux ingrédients d'une expérience: le recrutement, la randomisation, l'intervention et les résultats. Ensemble, ces quatre ingrédients permettent aux scientifiques d'aller au-delà des corrélations et de mesurer l'effet causal des traitements. Plus précisément, la randomisation signifie que les personnes dans les groupes de traitement et de contrôle seront similaires. Ceci est important car cela signifie que toute différence dans les résultats entre les deux groupes peut être attribuée au traitement et non à un facteur de confusion.

En plus d'être une belle illustration de la mécanique des expériences, l'étude de Restivo et van de Rijt montre également que la logistique des expériences numériques peut être complètement différente de celle des expériences analogiques. Dans l'expérience de Restivo et van de Rijt, il était facile de donner le barnstar à n'importe qui, et il était facile de suivre le nombre de modifications sur une longue période (parce que l'historique des modifications est automatiquement enregistré par Wikipedia). Cette capacité à délivrer des traitements et à mesurer les résultats sans frais est qualitativement différente des expériences passées. Bien que cette expérience ait impliqué 200 personnes, elle aurait pu être menée avec 2 000 ou même 20 000 personnes. La principale chose qui empêchait les chercheurs de multiplier leur expérience par un facteur de 100 n'était pas le coût; c'était l'éthique. C'est-à-dire que Restivo et van de Rijt ne voulaient pas donner de mots-clés à des éditeurs non méritants, et ils ne voulaient pas que leur expérience perturbe la communauté Wikipedia (Restivo and Rijt 2012, 2014) . Je reviendrai sur certaines des considérations éthiques soulevées par les expériences plus loin dans ce chapitre et dans le chapitre 6.

En conclusion, l'expérience de Restivo et van de Rijt montre clairement que si la logique de base de l'expérimentation n'a pas changé, la logistique des expériences numériques peut être radicalement différente. Ensuite, afin d'isoler plus clairement les opportunités créées par ces changements, je vais comparer les expériences que les chercheurs peuvent faire maintenant avec le genre d'expériences qui ont été faites dans le passé.

4.3 Deux dimensions d'expériences: laboratoire-terrain et analogique-numérique

**Expériences** de **laboratoire offrent** un **contrôle,** des **expériences sur** le **terrain offrent réalisme, et** des **expériences de terrain numériques combinent** le **contrôle et** le **réalisme à** l' **échelle.**

Les expériences viennent dans beaucoup de différentes formes et tailles. Dans le passé, les chercheurs ont trouvé utile d'organiser des expériences sur un continuum entre les *expériences en laboratoire* et *les expériences sur le terrain* . Maintenant, cependant, les chercheurs devraient également organiser des expériences le long d'un second continuum entre *les expériences analogiques* et *les expériences numériques* . Cet espace de conception bidimensionnel vous aidera à comprendre les forces et les faiblesses des différentes approches et à mettre en évidence les domaines les plus prometteurs (figure 4.1).

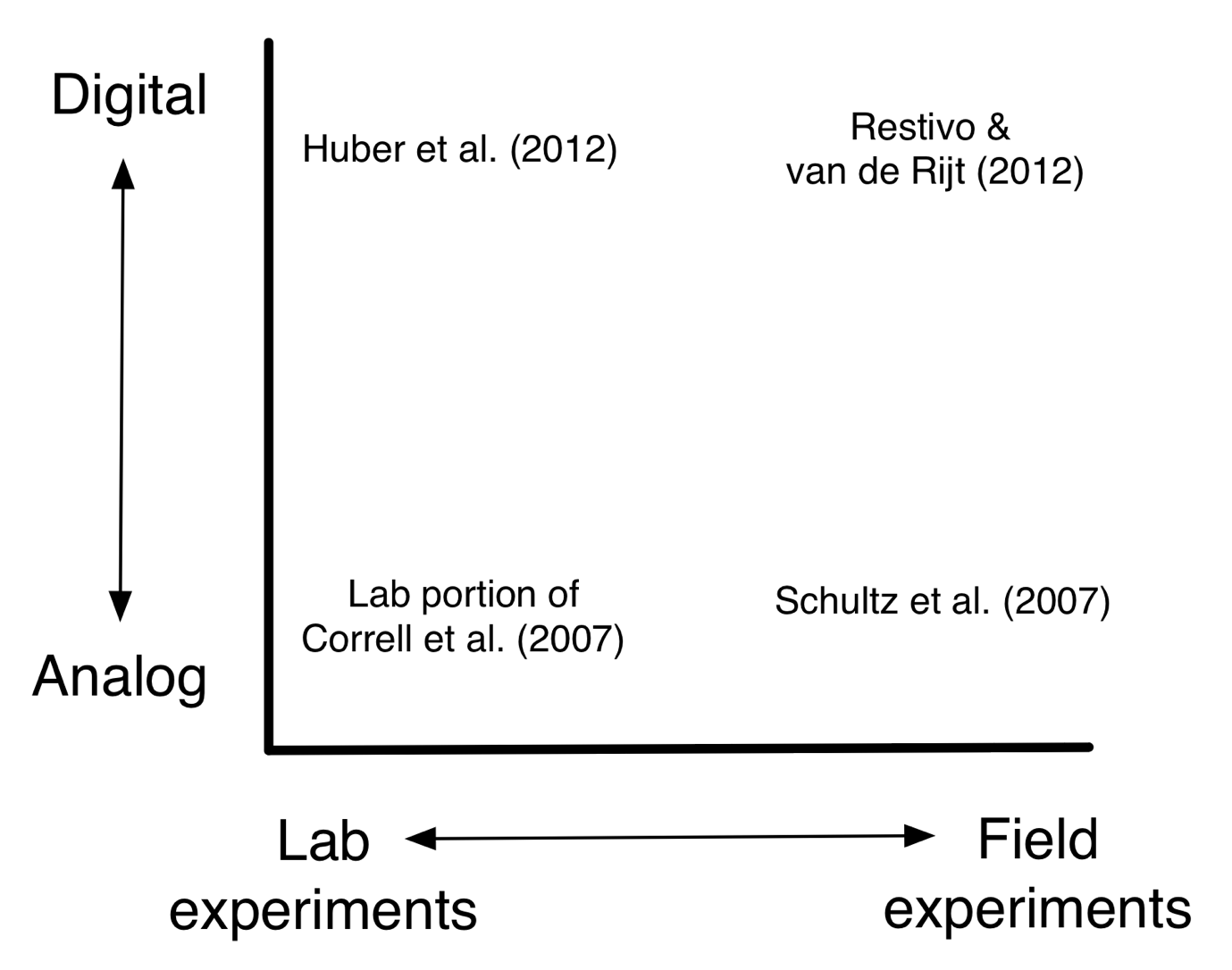


Figure 4.1: Schéma de l'espace de conception pour les expériences. Dans le passé, les expériences variaient selon la dimension du champ de laboratoire. Maintenant, ils varient également sur la dimension analogique-numérique. Cet espace de conception bidimensionnel est illustré par quatre expériences que je décris dans ce chapitre. À mon avis, le domaine de la plus grande opportunité est celui des expériences numériques sur le terrain.

Une dimension sur laquelle les expériences peuvent être organisées est la dimension lab-champ. Beaucoup d'expériences dans les sciences sociales sont des *expériences de laboratoire* où les étudiants de premier cycle effectuent des tâches étranges dans un laboratoire pour le crédit de cours. Ce type d'expérience domine la recherche en psychologie, car il permet aux chercheurs de créer des paramètres hautement contrôlés pour isoler et tester précisément des théories spécifiques sur le comportement social. Pour certains problèmes, cependant, quelque chose semble un peu étrange de tirer des conclusions solides sur le comportement humain de ces personnes inhabituelles effectuant des tâches aussi inhabituelles dans un cadre aussi inhabituel. Ces préoccupations ont conduit à un mouvement vers *des expériences* sur le *terrain* . Les expériences sur le terrain combinent la conception solide d'expériences de contrôle randomisées avec des groupes plus représentatifs de participants effectuant des tâches plus communes dans des environnements plus naturels.

Bien que certaines personnes considèrent les expériences de laboratoire et de terrain comme des méthodes concurrentes, il est préférable de les considérer comme complémentaires, avec des forces et des faiblesses différentes. Par exemple, Correll, Benard, and Paik (2007) utilisé une expérience de laboratoire et une expérience sur le terrain pour tenter de trouver les sources de la «pénalité de la maternité». Aux États-Unis, les mères gagnent moins d'argent que les femmes sans enfant. comparer des femmes ayant des compétences similaires et travaillant dans des emplois similaires. Il existe de nombreuses explications possibles à cette tendance, dont l'une est que les employeurs sont biaisés à l'égard des mères. (Fait intéressant, le contraire semble être vrai pour les pères: ils ont tendance à gagner plus que les hommes sans enfants comparables.) Afin d'évaluer les biais possibles contre les mères, Correll et ses collègues ont mené deux expériences: une en laboratoire et une sur le terrain.

D'abord, dans une expérience de laboratoire, ils ont dit aux participants, qui étaient des étudiants de premier cycle, qu'une entreprise menait une recherche d'emploi pour une personne pour diriger son nouveau département de marketing de la côte Est. On a dit aux étudiants que l'entreprise voulait leur aide dans le processus d'embauche et qu'on leur a demandé d'examiner les curriculum vitæ de plusieurs candidats potentiels et de noter les candidats sur plusieurs dimensions, comme leur intelligence, leur chaleur et leur engagement au travail. De plus, on a demandé aux étudiants s'ils recommanderaient d'embaucher le candidat et ce qu'ils recommanderaient comme salaire de départ. À l'insu des étudiants, cependant, les CV ont été spécifiquement construits pour être similaires, sauf pour une chose: certains d'entre eux ont signalé la maternité (en énumérant la participation à une association parent-enseignant) et certains ne l'ont pas fait. Correll et ses collègues ont constaté que les étudiants étaient moins susceptibles de recommander l'embauche des mères et qu'ils leur offraient un salaire de départ inférieur. De plus, grâce à une analyse statistique des notes et des décisions liées à l'embauche, Correll et ses collègues ont constaté que les désavantages des mères s'expliquaient en grande partie par le fait qu'elles étaient moins bien cotées en termes de compétence et d'engagement. Ainsi, cette expérience en laboratoire a permis à Correll et ses collègues de mesurer un effet causal et de fournir une explication possible pour cet effet.

Bien sûr, on peut être sceptique quant à tirer des conclusions sur l'ensemble du marché du travail américain sur la base des décisions de quelques centaines d'étudiants de premier cycle qui n'ont probablement jamais eu d'emploi à temps plein et encore moins d'embauche. Par conséquent, Correll et ses collègues ont également mené une expérience de terrain complémentaire. Ils ont répondu à des centaines d'offres d'emploi annoncées avec de fausses lettres d'accompagnement et des CV. Semblable aux matériaux montrés aux étudiants de premier cycle, certains ont signalé la maternité et certains n'ont pas. Correll et ses collègues ont constaté que les mères étaient moins susceptibles d'être rappelées pour des entrevues que les femmes sans enfants tout aussi qualifiées. En d'autres termes, les vrais employeurs qui prennent des décisions conséquentes dans un cadre naturel se comportent comme les étudiants de premier cycle. Ont-ils pris des décisions similaires pour la même raison? Malheureusement, nous ne savons pas. Les chercheurs n'ont pas été en mesure de demander aux employeurs d'évaluer les candidats ou d'expliquer leurs décisions.

Cette paire d'expériences révèle beaucoup de choses sur les expériences en laboratoire et sur le terrain en général. Les expériences de laboratoire offrent aux chercheurs un contrôle presque total de l'environnement dans lequel les participants prennent des décisions. Ainsi, par exemple, dans l'expérience en laboratoire, Correll et ses collègues ont pu s'assurer que tous les CV étaient lus dans un environnement calme; dans l'expérience sur le terrain, certains CV n'ont peut-être même pas été lus. De plus, étant donné que les participants au laboratoire savent qu'ils sont étudiés, les chercheurs sont souvent en mesure de recueillir des données supplémentaires qui peuvent aider à expliquer pourquoi les participants prennent leurs décisions. Par exemple, Correll et ses collègues ont demandé aux participants à l'expérience en laboratoire d'évaluer les candidats selon différentes dimensions. Ce type de *données* de *processus* pourrait aider les chercheurs à comprendre les mécanismes à l'origine des différences dans la façon dont les participants traitent les CV.

D'un autre côté, ces mêmes caractéristiques que je viens de décrire comme des avantages sont aussi parfois considérées comme des inconvénients. Les chercheurs qui préfèrent les expériences sur le terrain affirment que les participants aux expériences de laboratoire pourraient agir très différemment parce qu'ils savent qu'ils sont étudiés. Par exemple, dans l'expérience en laboratoire, les participants ont pu deviner le but de la recherche et modifier leur comportement afin de ne pas paraître biaisé. En outre, les chercheurs qui préfèrent les expériences sur le terrain pourraient argumenter que de petites différences dans les CV ne peuvent ressortir que dans un environnement de laboratoire stérile très propre, et donc l'expérience en laboratoire surestimera l'effet de la maternité sur les décisions d'embauche réelles. Enfin, de nombreux partisans des expériences sur le terrain critiquent la dépendance des expériences de laboratoire sur les participants à WEIRD: principalement des étudiants des pays occidentaux, éduqués, industrialisés, riches et démocratiques (Henrich, Heine, and Norenzayan 2010a) . Les expériences de Correll et ses collègues (2007) illustrent les deux extrêmes du continuum lab-field. Entre ces deux extrêmes, il y a aussi une variété de conceptions hybrides, y compris des approches telles que le fait d'amener des non-étudiants dans un laboratoire ou de se rendre sur le terrain tout en demandant aux participants d'accomplir une tâche inhabituelle.

En plus de la dimension laboratoire-champ qui existait dans le passé, l'ère numérique signifie que les chercheurs ont maintenant une deuxième dimension majeure au cours de laquelle les expériences peuvent varier: analogique-numérique. De même qu'il existe des expériences de laboratoire pur, des expériences sur le terrain pur et une variété d'hybrides entre les deux, il existe des expériences purement analogiques, des expériences numériques pures et une variété d'hybrides. Il est difficile d'offrir une définition formelle de cette dimension, mais une définition de travail utile est que *les expériences entièrement numériques* sont des expériences qui utilisent l'infrastructure numérique pour recruter des participants, randomiser, fournir des traitements et mesurer les résultats. Par exemple, l'étude de Restivo et van de Rijt (2012) sur les barnstars et Wikipédia était une expérience entièrement numérique, car elle utilisait des systèmes numériques pour ces quatre étapes. De même, *les expériences entièrement analogiques* n'utilisent pas l'infrastructure numérique pour l'une de ces quatre étapes. Beaucoup d'expériences classiques en psychologie sont des expériences entièrement analogiques. Entre ces deux extrêmes, il existe *des expériences partiellement numériques* qui utilisent une combinaison de systèmes analogiques et numériques.

Quand certaines personnes pensent à des expériences numériques, elles pensent immédiatement à des expériences en ligne. C'est regrettable parce que les possibilités d'exécuter des expériences numériques ne sont pas seulement en ligne. Les chercheurs peuvent effectuer des expériences partiellement numériques en utilisant des appareils numériques dans le monde physique afin de délivrer des traitements ou de mesurer les résultats. Par exemple, les chercheurs pourraient utiliser des téléphones intelligents pour fournir des traitements ou des capteurs dans l'environnement bâti afin de mesurer les résultats. En fait, comme nous le verrons plus loin dans ce chapitre, les chercheurs ont déjà utilisé des compteurs d'énergie domestique pour mesurer les résultats d'expériences sur la consommation d'énergie impliquant 8,5 millions de ménages (Allcott 2015) . À mesure que les appareils numériques s'intègrent de plus en plus dans la vie des gens et que les capteurs s'intègrent dans l'environnement bâti, ces occasions d'exécuter des expériences partiellement numériques dans le monde physique augmenteront considérablement. En d'autres termes, les expériences numériques ne sont pas seulement des expériences en ligne.

Les systèmes numériques créent de nouvelles possibilités d'expériences partout dans le continuum de laboratoire. Dans les expériences de laboratoire pur, par exemple, les chercheurs peuvent utiliser des systèmes numériques pour une mesure plus fine du comportement des participants; Un exemple de ce type de mesure améliorée est un équipement de suivi de l'œil qui fournit des mesures précises et continues de la localisation du regard. L'ère numérique crée également la possibilité d'exécuter des expériences de laboratoire en ligne. Par exemple, les chercheurs ont rapidement adopté Amazon Mechanical Turk (MTurk) pour recruter des participants pour des expériences en ligne (figure 4.2). MTurk correspond aux «employeurs» qui ont des tâches à accomplir avec les «travailleurs» qui souhaitent accomplir ces tâches pour de l'argent. Contrairement aux marchés du travail traditionnels, cependant, les tâches impliquées ne nécessitent habituellement que quelques minutes et l'interaction complète entre l'employeur et le travailleur est en ligne. Parce que MTurk imite les aspects des expériences de laboratoire traditionnelles - payer les gens pour accomplir des tâches qu'ils ne feraient pas gratuitement - il est naturellement adapté à certains types d'expériences. Pour l'essentiel, MTurk a créé l'infrastructure de gestion d'un groupe de participants - recrutant et rémunérant des personnes - et les chercheurs ont profité de cette infrastructure pour puiser dans un bassin de participants toujours disponible.

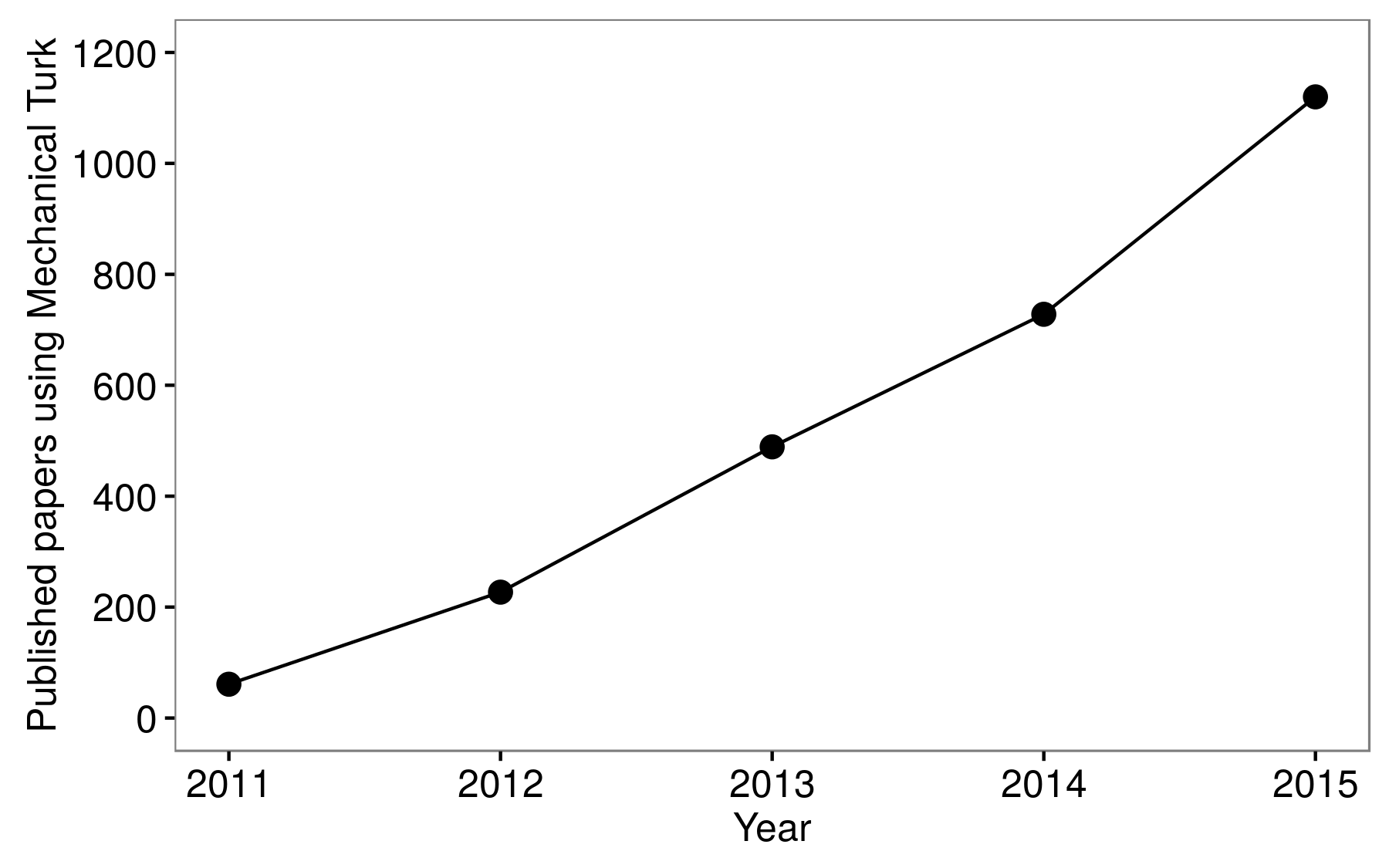


Figure 4.2: Documents publiés à l'aide des données d'Amazon Mechanical Turk (MTurk). MTurk et d'autres marchés du travail en ligne offrent aux chercheurs un moyen pratique de recruter des participants pour des expériences. Adapté de Bohannon (2016) .

Les systèmes numériques créent encore plus de possibilités pour des expériences de type champ. En particulier, ils permettent aux chercheurs de combiner les données de contrôle et de processus rigoureuses associées aux expériences en laboratoire avec les participants plus diversifiés et les paramètres plus naturels associés aux expériences en laboratoire. En outre, les expériences sur le terrain numérique offrent également trois opportunités qui ont tendance à être difficile dans les expériences analogiques.

Premièrement, alors que la plupart des expériences analogiques en laboratoire et sur le terrain ont des centaines de participants, les expériences sur le terrain peuvent avoir des millions de participants. Ce changement d'échelle est dû au fait que certaines expériences numériques peuvent produire des données à un coût variable nul. Autrement dit, une fois que les chercheurs ont créé une infrastructure expérimentale, augmenter le nombre de participants n'augmente généralement pas le coût. Augmenter le nombre de participants d'un facteur de 100 ou plus n'est pas seulement un changement *quantitatif* ; c'est un changement *qualitatif* , car il permet aux chercheurs d'apprendre différentes choses à partir d'expériences (par exemple, l'hétérogénéité des effets du traitement) et d'exécuter des conceptions expérimentales entièrement différentes (par exemple, des expériences en grands groupes). Ce point est si important, j'y reviendrai vers la fin du chapitre quand je donnerai des conseils sur la création d'expériences numériques.

Deuxièmement, alors que la plupart des expériences analogiques en laboratoire et sur le terrain traitent les participants comme des gadgets indiscernables, les expériences numériques sur le terrain utilisent souvent des informations générales sur les participants aux étapes de conception et d'analyse de la recherche. Ces informations de base, appelées informations de *pré-traitement* , sont souvent disponibles dans les expériences numériques car elles sont exécutées en plus des systèmes de mesure toujours actifs (voir chapitre 2). Par exemple, un chercheur de Facebook a beaucoup plus d'informations avant le traitement sur les personnes dans son expérience de terrain numérique qu'un chercheur universitaire a sur les personnes dans son expérience de terrain analogique. Ce prétraitement permet des conceptions expérimentales plus efficaces - telles que le blocage (Higgins, Sävje, and Sekhon 2016) et le recrutement ciblé de participants (Eckles, Kizilcec, and Bakshy 2016) - et une analyse plus perspicace - comme l'estimation de l'hétérogénéité des effets du traitement (Athey and Imbens 2016a) et ajustement des covariables pour une meilleure précision (Bloniarz et al. 2016) .

Troisièmement, alors que de nombreuses expériences de laboratoire et de terrain analogiques fournissent des traitements et mesurent les résultats dans un temps relativement réduit, certaines expériences sur le terrain se déroulent sur des échelles de temps beaucoup plus longues. Par exemple, les résultats de l'expérience de Restivo et van de Rijt ont été mesurés quotidiennement pendant 90 jours, et l'une des expériences dont je parlerai plus tard dans le chapitre (Ferraro, Miranda, and Price 2011) suivi les résultats sur trois ans. Coût. Ces trois possibilités - taille, informations de prétraitement, données longitudinales de traitement et de résultats - apparaissent le plus souvent lorsque les expériences sont effectuées sur des systèmes de mesure permanents (voir le chapitre 2 pour plus de détails sur les systèmes de mesure permanents).

Alors que les expériences sur le terrain numérique offrent de nombreuses possibilités, elles partagent également certaines faiblesses avec les expériences de laboratoire analogique et de terrain analogique. Par exemple, les expériences ne peuvent pas être utilisées pour étudier le passé, et elles ne peuvent qu'estimer les effets des traitements qui peuvent être manipulés. En outre, bien que les expériences soient sans aucun doute utiles pour orienter la politique, les indications précises qu'elles peuvent offrir sont quelque peu limitées en raison de complications telles que la dépendance environnementale, les problèmes d'observance et les effets d'équilibre (Banerjee and Duflo 2009; Deaton 2010) . Les expériences numériques sur le terrain amplifient également les préoccupations éthiques créées par les expériences sur le terrain - un sujet que je traiterai plus loin dans ce chapitre et dans le chapitre 6.

3.1 Présentation

Les chercheurs qui étudient les dauphins ne peuvent pas leur poser de questions et sont donc obligés d'essayer d'en apprendre davantage sur les dauphins en observant leur comportement. D'autre part, les chercheurs qui étudient les humains ont plus de facilité: leurs répondants peuvent parler. Parler aux gens était une partie importante de la recherche sociale dans le passé, et je m'attends à ce que ce soit aussi le cas à l'avenir.

Dans la recherche sociale, parler aux gens prend généralement deux formes: des enquêtes et des entretiens approfondis. Grosso modo, la recherche utilisant des enquêtes implique le recrutement systématique d'un grand nombre de participants, des questionnaires très structurés et l'utilisation de méthodes statistiques pour généraliser les participants à une plus grande population. D'un autre côté, la recherche utilisant des entretiens approfondis implique généralement un petit nombre de participants, des conversations semi-structurées et débouche sur une description riche et qualitative des participants. Les sondages et les entrevues approfondies sont deux approches puissantes, mais les enquêtes sont beaucoup plus touchées par le passage de l'ère analogique à l'ère numérique. Par conséquent, dans ce chapitre, je vais me concentrer sur la recherche par sondage.

Comme je le montrerai dans ce chapitre, l'ère numérique crée de nombreuses opportunités pour les chercheurs de recueillir des données plus rapidement et à moindre coût, de poser différents types de questions et d'augmenter la valeur des données d'enquête avec des sources de données volumineuses. Cependant, l'idée que la recherche par sondage peut être transformée par un changement technologique n'est pas nouvelle. Vers 1970, un changement similaire se produisait grâce à une technologie de communication différente: le téléphone. Heureusement, comprendre comment le téléphone a changé la recherche par sondage peut nous aider à imaginer comment l'ère numérique va changer la recherche par sondage.

La recherche par sondage, telle que nous la connaissons aujourd'hui, a commencé dans les années 1930. Au cours de la première phase de la recherche par sondage, les chercheurs échantillonnaient au hasard des régions géographiques (comme les pâtés de maisons) et se rendaient ensuite dans ces régions afin d'avoir des conversations face à face avec des ménages échantillonnés au hasard. Ensuite, un développement technologique - la diffusion généralisée des téléphones fixes dans les pays riches - a finalement mené à la deuxième ère de la recherche par sondage. Cette deuxième ère différait à la fois dans la manière dont les gens étaient échantillonnés et dans la façon dont les conversations avaient lieu. À la deuxième époque, plutôt que d'échantillonner les ménages dans des zones géographiques, les chercheurs ont échantillonné au hasard des numéros de téléphone dans une procédure appelée *numérotation aléatoire* . Et plutôt que de voyager pour parler aux gens face à face, les chercheurs les ont plutôt appelés au téléphone. Ceux-ci peuvent sembler être de petits changements logistiques, mais ils ont rendu la recherche par sondage plus rapide, moins coûteuse et plus flexible. En plus d'être habilitants, ces changements ont également été controversés parce que de nombreux chercheurs craignaient que ces nouvelles procédures d'échantillonnage et d'entrevue puissent introduire une variété de biais. Mais finalement, après beaucoup de travail, les chercheurs ont trouvé comment collecter des données de manière fiable en utilisant la numérotation aléatoire et les interviews téléphoniques. Ainsi, en cherchant comment exploiter avec succès l'infrastructure technologique de la société, les chercheurs ont pu moderniser la façon dont ils ont mené leurs enquêtes.

Maintenant, un autre développement technologique - l'ère numérique - nous conduira éventuellement à une troisième ère de recherche par sondage. Cette transition est en partie motivée par la décroissance graduelle des approches de la deuxième ère (BD Meyer, Mok, and Sullivan 2015) . Par exemple, pour diverses raisons technologiques et sociales, les taux de non-réponse - c'est-à-dire la proportion de personnes échantillonnées qui ne participent pas aux enquêtes - augmentent depuis de nombreuses années (National Research Council 2013) . Ces tendances à long terme signifient que le taux de non-réponse peut maintenant dépasser 90% dans les enquêtes téléphoniques standard (Kohut et al. 2012) .

D'un autre côté, la transition vers une troisième ère est également motivée en partie par de nouvelles opportunités passionnantes, dont certaines seront décrites dans ce chapitre. Bien que les choses ne soient pas encore réglées, je m'attends à ce que la troisième ère de la recherche par sondage soit caractérisée par l'échantillonnage non probabiliste, les entrevues informatisées et le couplage des enquêtes aux grandes sources de données (tableau 3.1).

| Tableau 3.1: Trois ères de la recherche par sondage fondée sur les Groves (2011) | | | |
| --- | --- | --- | --- |
|  | **Échantillonnage** | **Interviewer** | **Environnement de données** |
| Première époque | Échantillonnage probabiliste de zone | Face à face | Enquêtes autonomes |
| Deuxième époque | Échantillonnage aléatoire de numérotation aléatoire (RDD) | Téléphone | Enquêtes autonomes |
| Troisième ère | Échantillonnage non probabiliste | Administré par ordinateur | Enquêtes liées aux grandes sources de données |

La transition entre la deuxième et la troisième ère de la recherche par sondage n'a pas été complètement lisse, et il y a eu des débats féroces sur la façon dont les chercheurs devraient procéder. En revenant sur la transition entre la première et la deuxième époque, je pense qu'il y a une idée clé pour nous maintenant: *le début n'est pas la fin* . C'est-à-dire qu'au départ beaucoup de méthodes téléphoniques de la deuxième ère étaient ad hoc et ne fonctionnaient pas très bien. Mais, grâce à un travail acharné, les chercheurs ont résolu ces problèmes. Par exemple, les chercheurs effectuaient des numérotations aléatoires depuis de nombreuses années avant que Warren Mitofsky et Joseph Waksberg ne développent une méthode d'échantillonnage par numérotation aléatoire ayant de bonnes propriétés pratiques et théoriques (Waksberg 1978; **???** ) . Ainsi, nous ne devons pas confondre l'état actuel des approches de la troisième ère avec leurs résultats finaux.

L'histoire de la recherche par sondage montre que le domaine évolue, entraîné par les changements technologiques et sociétaux. Il n'y a aucun moyen d'arrêter cette évolution. Au contraire, nous devrions l'embrasser, tout en continuant à tirer la sagesse des époques antérieures, et c'est l'approche que je vais prendre dans ce chapitre. Premièrement, je soutiendrai que les grandes sources de données ne remplaceront pas les enquêtes et que l'abondance des sources de données volumineuses augmente (et non diminue) la valeur des enquêtes (section 3.2). Compte tenu de cette motivation, je résumerai le cadre d'erreur total de l'enquête (section 3.3) qui a été développé durant les deux premières époques de la recherche par sondage. Ce cadre nous permet de comprendre de nouvelles approches de la représentation - en particulier des échantillons non probabilistes (section 3.4) - et de nouvelles approches de mesure - en particulier de nouvelles façons de poser des questions aux répondants (section 3.5). Enfin, je vais décrire deux modèles de recherche pour lier les données d'enquête aux grandes sources de données (section 3.6).

## 3.2 Demander ou observer

**Nous allons toujours avoir besoin de poser des questions** aux **gens.**

Étant donné que de plus en plus de notre comportement est capturé dans les sources de données volumineuses, telles que les données administratives gouvernementales et administratives, certaines personnes pourraient penser que poser des questions appartient au passé. Mais ce n'est pas si simple. Il y a deux raisons principales pour lesquelles les chercheurs continueront à poser des questions aux gens. Premièrement, comme je l'ai indiqué au chapitre 2, il existe de réels problèmes d'exactitude, d'exhaustivité et d'accessibilité de nombreuses sources de données volumineuses. Deuxièmement, en plus de ces raisons pratiques, il y a une raison plus fondamentale: il y a certaines choses qui sont très difficiles à apprendre à partir de données comportementales, même des données comportementales parfaites. Par exemple, certains des résultats sociaux les plus importants et les prédicteurs sont les états internes, tels que les émotions, les connaissances, les attentes et les opinions. Les états internes existent dans la tête des gens, et parfois la meilleure façon d'apprendre sur les états internes est de demander.

Moira Burke et Robert Kraut (2014) ont illustré les limites pratiques et fondamentales des sources de données volumineuses et la façon dont elles peuvent être surmontées grâce à des enquêtes sur la façon dont la force de l'amitié a été influencée par l'interaction sur Facebook. À l'époque, Burke travaillait sur Facebook pour avoir un accès complet à l'un des enregistrements les plus massifs et détaillés du comportement humain jamais créé. Mais, même ainsi, Burke et Kraut ont dû utiliser des enquêtes pour répondre à leur question de recherche. Leur résultat d'intérêt - le sentiment subjectif de proximité entre l'intimé et son ami - est un état interne qui n'existe que dans la tête du répondant. En outre, en plus d'utiliser une enquête pour recueillir leurs résultats d'intérêt, Burke et Kraut ont également dû utiliser une enquête pour en apprendre davantage sur les facteurs potentiellement confondants. En particulier, ils voulaient séparer l'impact de la communication sur Facebook de la communication par d'autres canaux (par exemple, courriel, téléphone et face à face). Même si les interactions par e-mail et par téléphone sont enregistrées automatiquement, ces traces n'étaient pas disponibles pour Burke et Kraut, elles ont donc dû les collecter avec un sondage. En combinant leurs données d'enquête sur la force de l'amitié et l'interaction non-Facebook avec les données du journal Facebook, Burke et Kraut ont conclu que la communication via Facebook a en fait conduit à un sentiment accru de proximité.

Comme l'illustrent les travaux de Burke et Kraut, les grandes sources de données n'élimineront pas la nécessité de poser des questions aux gens. En fait, je tirerais la leçon inverse de cette étude: les grandes sources de données peuvent en fait augmenter la valeur de poser des questions, comme je le montrerai tout au long de ce chapitre. Par conséquent, la meilleure façon de penser à la relation entre demander et observer est que ce sont des compléments plutôt que des substituts; ils sont comme le beurre de cacahuète et la gelée. Quand il y a plus de beurre de cacahuète, les gens veulent plus de gelée; quand il y a plus de grosses données, je pense que les gens voudront plus de sondages.