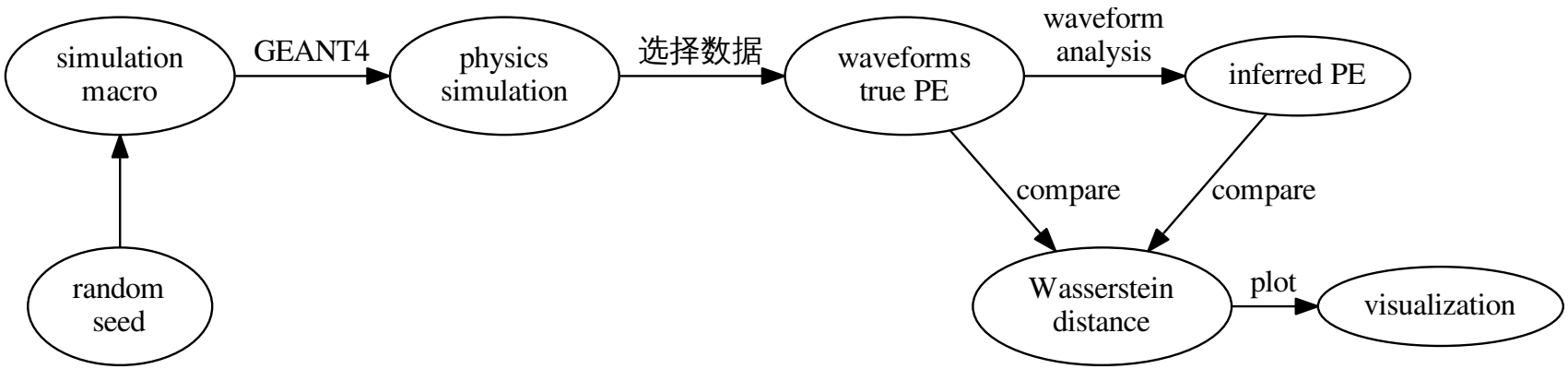
GNU Make 驱动的数据编译与流水线构建

续本达

清华大学

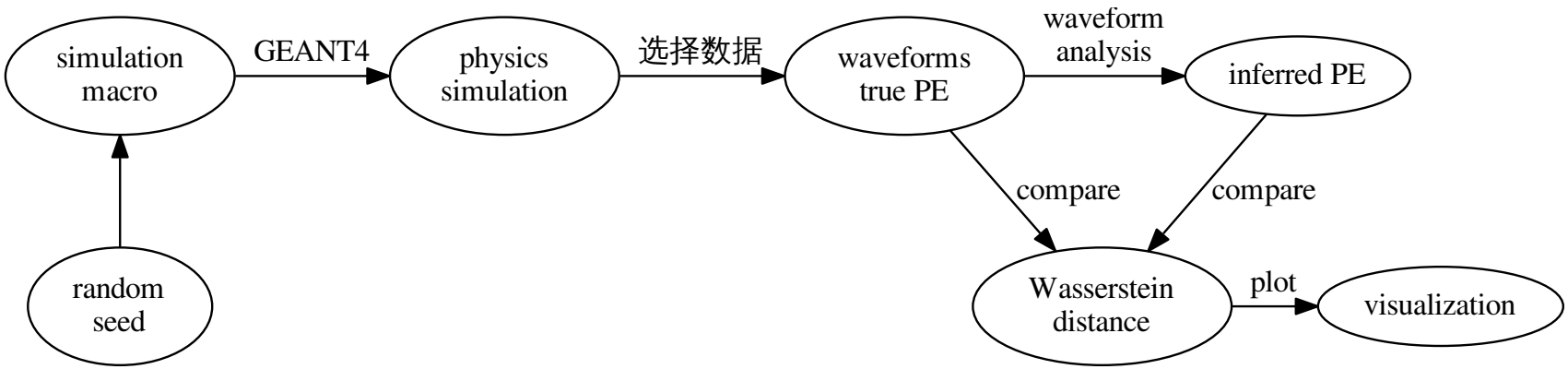
2019 年 4 月 13 日金枪鱼之夜

如何驱动这条数据流水线？



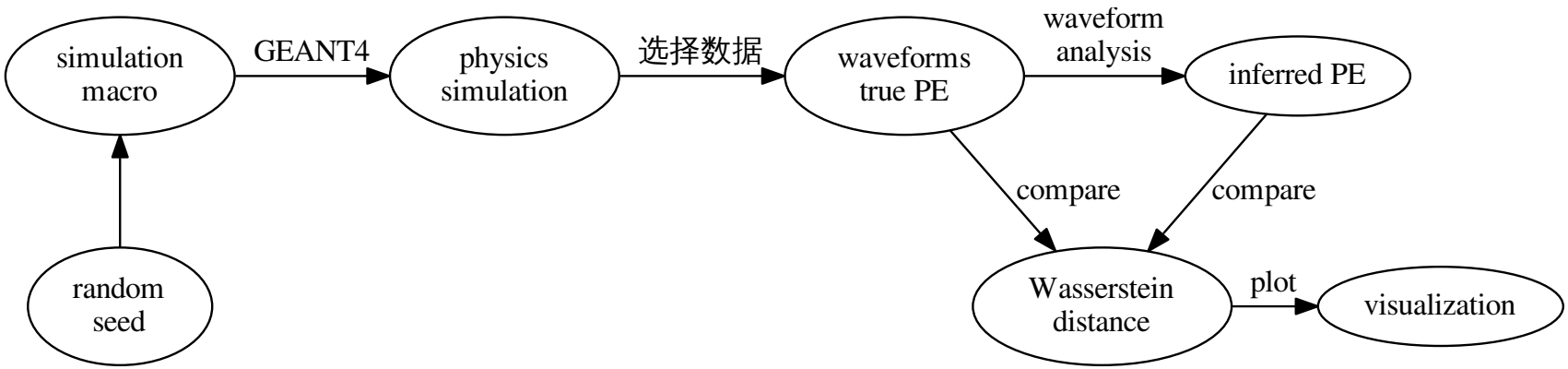
1. 手动：数据量增大 100 倍？改参数？

如何驱动这条数据流水线？



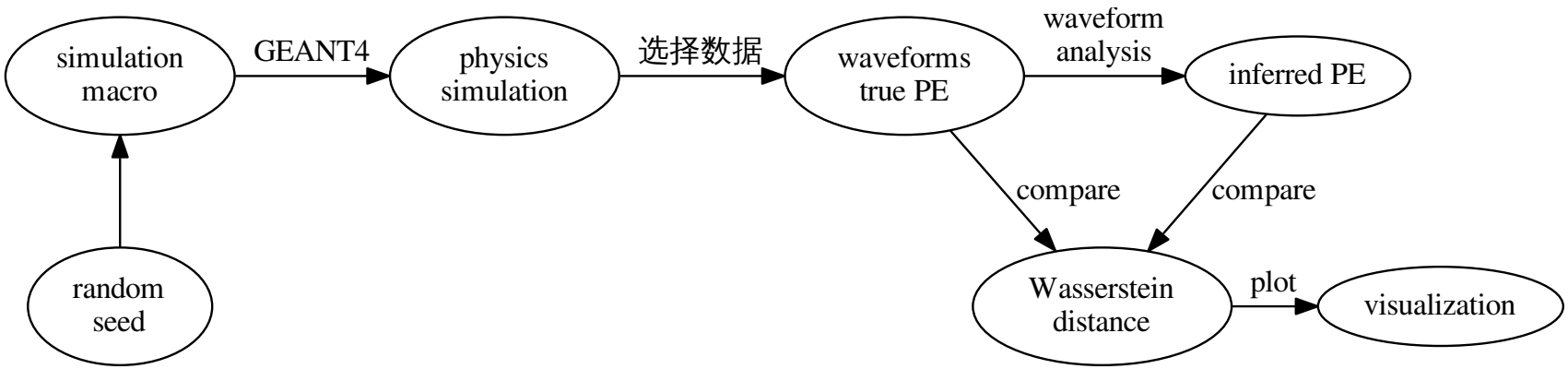
1. 手动：数据量增大 100 倍？改参数？
2. 全都写在同一个程序里：C++？用 Python 统合？多人合作？

如何驱动这条数据流水线？



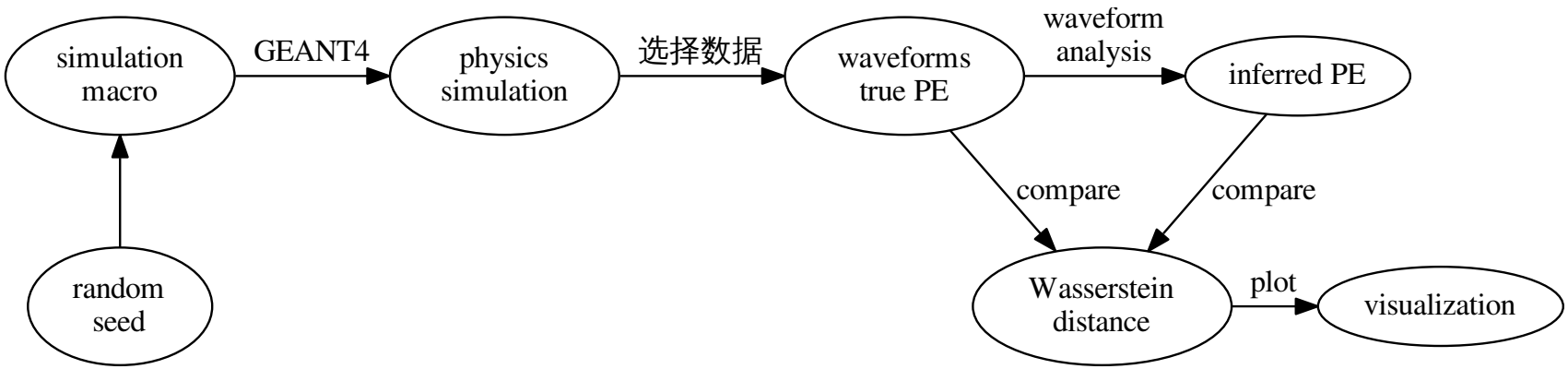
1. 手动：数据量增大 100 倍？改参数？
2. 全都写在同一个程序里：C++？用 Python 统合？多人合作？
3. Shell 脚本，Perl 脚本：中间出错了，重来？

如何驱动这条数据流水线？



1. 手动：数据量增大 100 倍？改参数？
2. 全都写在同一个程序里：C++？用 Python 统合？多人合作？
3. Shell 脚本, Perl 脚本：中间出错了，重来？
4. Hadoop, Spark：只能用 Java 和 scala?

如何驱动这条数据流水线？



1. 手动：数据量增大 100 倍？改参数？
2. 全都写在同一个程序里：C++？用 Python 统合？多人合作？
3. Shell 脚本, Perl 脚本：中间出错了，重来？
4. Hadoop, Spark：只能用 Java 和 scala?
5. GNU Make：解决以上所有问题，并有额外功能！

数据分析的原则

1. Transparency
   每一步的数据都应尽量可以被人类直接阅读。如果不得不用二进制模式，一定是使用最普遍最开放的格式。e.g., csv, hdf5.

数据分析的原则

1. Transparency
   每一步的数据都应尽量可以被人类直接阅读。如果不得不用二进制模式，一定是使用最普遍最开放的格式。e.g., csv, hdf5.
2. SPOT and DRY
   Single Point of Truth, Don't Repeat Yourself. 不可在分析做任何重复，任何有意义的信息都应该被共享。

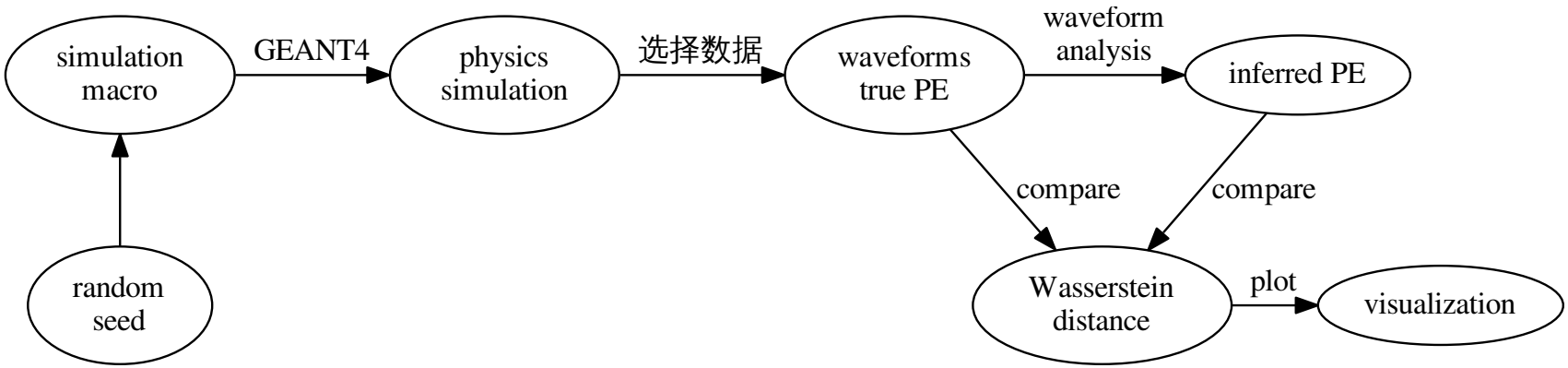
数据分析的原则

1. Transparency
   每一步的数据都应尽量可以被人类直接阅读。如果不得不用二进制模式，一定是使用最普遍最开放的格式。e.g., csv, hdf5.
2. SPOT and DRY
   Single Point of Truth, Don't Repeat Yourself. 不可在分析做任何重复，任何有意义的信息都应该被共享。
3. Simplicity and Economy
   尽量使用高级语言和语法糖，为每个子任务选择合适的工具。只有在性能分析之后，才在必要时使用低级语言进行性能加速。

数据分析的原则

1. Transparency
   每一步的数据都应尽量可以被人类直接阅读。如果不得不用二进制模式，一定是使用最普遍最开放的格式。e.g., csv, hdf5.
2. SPOT and DRY
   Single Point of Truth, Don't Repeat Yourself. 不可在分析做任何重复，任何有意义的信息都应该被共享。
3. Simplicity and Economy
   尽量使用高级语言和语法糖，为每个子任务选择合适的工具。只有在性能分析之后，才在必要时使用低级语言进行性能加速。
4. Reproducible Research and Literate Programming
   同时以人类语言和计算机语言的形式，详细记录每一步计算。

第一轮：初识 Makefile



- 后三项依赖：
  - waveforms, true PE: `ftraining-6.h5`
  - inferred PE: `ft-ans-6.h5`
  - Wasserstein distance: `wdist-6.h5`
  - visualization: `wdist-6.pdf`

第二轮：并行计算

- Waveform analysis 一步运行比较慢
- 可以把文件分成 10 份来处理
- Make 的 *pattern matching* 可以巧妙表达每一份的依赖关系

第三轮：多语言协作

- Make 与 shell 紧密结合。只要定义好数据文件的输入输出，就可以调用任何语言：
  - `awk`
  - `sed`
  - `perl`
  - `java`
  - `R`
  - `Python`
  - `C++`
  - …

总结

- 还没有来得及讨论的
  1. 与 slurm 的集成: sync, async 皆可
  2. guile scheme 高级编程
  3. Make template
- 使用 GNU Make 构建的数据流水线满足所有要求
  1. Transparency
  2. SPOT and DRY
  3. Simplicity and Economy
  4. Reproducible Research and Literate Programming
- 缺点：门槛较高
  - 参考书：John Graham-cumming, The GNU Make Book
  - 参考资料：GNU Make Texinfo